Methods in Molecular Biology 2453

Springer Protocols

Anton W. Langerak *Editor*

Immuno-genetics

Methods and Protocols

OPEN ACCESS

Humana Press

METHODS IN MOLECULAR BIOLOGY

Series Editor
John M. Walker
School of Life and Medical Sciences
University of Hertfordshire
Hatfield, Hertfordshire, UK

For further volumes:
http://www.springer.com/series/7651

For over 35 years, biological scientists have come to rely on the research protocols and methodologies in the critically acclaimed *Methods in Molecular Biology* series. The series was the first to introduce the step-by-step protocols approach that has become the standard in all biomedical protocol publishing. Each protocol is provided in readily-reproducible step-by-step fashion, opening with an introductory overview, a list of the materials and reagents needed to complete the experiment, and followed by a detailed procedure that is supported with a helpful notes section offering tips and tricks of the trade as well as troubleshooting advice. These hallmark features were introduced by series editor Dr. John Walker and constitute the key ingredient in each and every volume of the *Methods in Molecular Biology* series. Tested and trusted, comprehensive and reliable, all protocols from the series are indexed in PubMed.

Immunogenetics

Methods and Protocols

Edited by

Anton W. Langerak

Department of Immunology, Erasmus MC, Rotterdam, The Netherlands

Humana Press

Editor
Anton W. Langerak
Department of Immunology
Erasmus MC
Rotterdam, The Netherlands

ISSN 1064-3745 ISSN 1940-6029 (electronic)
Methods in Molecular Biology
ISBN 978-1-0716-2114-1 ISBN 978-1-0716-2115-8 (eBook)
https://doi.org/10.1007/978-1-0716-2115-8

This Humana imprint is published by the registered company Springer Science+Business Media, LLC, part of Springer Nature.
The registered company address is: 1 New York Plaza, New York, NY 10004, U.S.A.

Preface

Adaptive immune cells (lymphocytes) are equipped with unique antigen receptors, termed immunoglobulins (IG) and T cell receptors (TR), which collectively form a highly diverse repertoire. In the lymphocytes, IG/TR diversity is actually created at the DNA level, thus giving rise to an enormous adaptive immune receptor repertoire (also known as the *immunome*) that can be studied in healthy and diseased subjects in the context of research questions and clinical applications. This field of (fundamental and translational) research is known as *immunogenetics*.

The immunogenetics domain has rapidly evolved in the last ten years or so, mainly through the introduction of high-throughput technologies. With these new technologies, unprecedented insight into the adaptive immune receptor repertoire could be obtained with much more sequencing depth and coverage of the repertoire than ever before. In this volume, many chapters are dedicated to lab protocols, bioinformatics, and immunoinformatics analysis of this high-resolution immunome analysis, exemplified by many different applications. Additionally, the newest technological variations on these protocols are discussed, including non-amplicon, single-cell, and cell-free strategies. Collectively, the chapters illustrate the impact that immunogenetics has achieved and will further expand in all fields of medicine, from infection and (auto)immunity, to vaccination, to lymphoid malignancy and tumor immunity.

As the guest editor of this volume on immunogenetics in the *Methods in Molecular Biology* book series, I am very pleased with the content and quality of this book. I am grateful to all authors who contributed to the success of this book volume with their valuable and informative chapters that collectively cover a broad spectrum of methodologies for applications in research and clinical diagnostics. I sincerely hope that readers will find the protocols and the method descriptions as useful as I did, for their own laboratory studies. Enjoy reading!

Rotterdam, The Netherlands ***Anton W. Langerak***

Contents

Contributors

Chrystelle Abdo • *Hôpital Necker Enfants-Malades, Laboratoire d'Onco-Hématologie, Assistance Publique– Hôpitaux de Paris, Paris, France*

Andreas Agathangelidis • *Institute of Applied Biosciences, Centre for Research and Technology Hellas, Thessaloniki, Greece; Department of Biology, School of Science, National and Kapodistrian University of Athens, Athens, Greece*

Beatrice Alessandria • *Hematology Division, Department of Molecular Biotechnologies and Health Sciences, University of Torino, Torino, Italy*

Safa Aouinti • *IMGT®, the international ImMunoGenetics information system®, Laboratoire d'ImmunoGénétique Moléculaire LIGM, Institut de Génétique Humaine, (IGH), Centre National de la Recherche Scientifique (CNRS), Université de Montpellier (UM), Montpellier, France; Clinical Research and Epidemiology Unit, CHU Montpellier, Univ Montpellier, Montpellier, France*

Marine Armand • *AP-HP, Pitié-Salpêtrière Hospital, Laboratory of Hematology, Paris, France; Sorbonne Université, Paris, France*

Zakia Azmani • *Center for Biomics, Erasmus MC University Medical Center, Rotterdam, The Netherlands; Department of Cell Biology, Erasmus MC University Medical Center, Rotterdam, The Netherlands*

Lmar Babrak • *Institute of Biomedical Engineering and Medical Informatics, School of Life Sciences, FHNW University of Applied Sciences and Arts Northwestern Switzerland, Muttenz, Switzerland*

Davide Bagnara • *Department of Experimental Medicine, University of Genoa, Genoa, Italy*

Lorenz Bastian • *Department of Hematology, University of Schleswig-Holstein, Kiel, Germany*

Matthieu Bertrand • *Hôpital Necker Enfants-Malades, Laboratoire d'Onco-Hématologie, Assistance Publique– Hôpitaux de Paris, Paris, France*

Michiel G. H. Betjes • *Erasmus MC Transplant Institute, Division of Nephrology and Transplantation, Department of Internal Medicine, Erasmus MC University Medical Center, Rotterdam, The Netherlands*

Magnolia L. Bostick • *Takara Bio USA, Inc., San Jose, CA, USA; PACT Pharma, Inc., South Francisco, CA, USA*

Myriam Boudjoghra • *AP-HP, Pitié-Salpêtrière Hospital, Laboratory of Hematology, Paris, France*

Clotilde Bravetti • *AP-HP, Pitié-Salpêtrière Hospital, Laboratory of Hematology, Paris, France; Sorbonne Université, Paris, France*

Rutger W. W. Brouwer • *Center for Biomics, Erasmus MC University Medical Center, Rotterdam, The Netherlands; Department of Cell Biology, Erasmus MC University Medical Center, Rotterdam, The Netherlands*

Monika Brüggemann • *Department of Hematology, University of Schleswig-Holstein, Kiel, Germany; Medical Department II, University Hospital Schleswig-Holstein, Kiel, Germany*

Christian E. Busse • *Division of B Cell Immunology, German Cancer Research Center (DKFZ), Heidelberg, Germany*

Giovanni Cazzaniga • *Centro Ricerca Tettamanti, Fondazione Tettamanti, Centro Maria Letizia Verga, Monza, Italy; Genetics, Department of Medicine and Surgery, University of Milan Bicocca, Monza, Italy*

Anastasia Chatzidimitriou • *Institute of Applied Biosciences, Centre for Research and Technology Hellas, Thessaloniki, Greece; Department of Molecular Medicine and Surgery, Karolinska Institute, Stockholm, Sweden*

Elaine C. Chen • *Department of Pathology, Microbiology, and Immunology, Vanderbilt Vaccine Center, Vanderbilt University Medical Center, Nashville, TN, USA*

Scott Christley • *Department of Population and Data Sciences, UT Southwestern Medical Center, Dallas, TX, USA*

Cheng-Yu Chung • *Department of Pharmaceutical Chemistry, The University of Kansas, Lawrence, KS, USA; Department of Chemical Engineering, The University of Kansas, Lawrence, KS, USA*

Aurora Maria Civita • *Hematology Division, Department of Molecular Biotechnologies and Health Sciences, University of Torino, Torino, Italy*

Brian D. Corrie • *Biological Sciences, Simon Fraser University, Burnaby, BC, Canada*

Lindsay G. Cowell • *Department of Population and Data Sciences, UT Southwestern Medical Center, Dallas, TX, USA; Department of Immunology, UT Southwestern Medical Center, Dallas, TX, USA*

Nikos Darzentas • *Department of Hematology, University of Schleswig-Holstein, Kiel, Germany; Medical Department II, University Hospital Schleswig-Holstein, Kiel, Germany*

Frederic Davi • *AP-HP, Pitié-Salpêtrière Hospital, Laboratory of Hematology, Paris, France; Sorbonne Université, Paris, France*

Reno Debets • *Laboratory of Tumor Immunology, Department of Medical Oncology, Erasmus MC-Cancer Institute, Rotterdam, The Netherlands*

Brandon J. DeKosky • *Department of Pharmaceutical Chemistry, The University of Kansas, Lawrence, KS, USA; Department of Chemical Engineering, The University of Kansas, Lawrence, KS, USA; Bioengineering Graduate Program, The University of Kansas, Lawrence, KS, USA*

Xander den Dekker • *Center for Biomics, Erasmus MC University Medical Center, Rotterdam, The Netherlands; Department of Cell Biology, Erasmus MC University Medical Center, Rotterdam, The Netherlands*

Daniela Drandi • *Hematology Division, Department of Molecular Biotechnology and Health Sciences, University of Torino, Torino, Italy*

Patrice Duroux • *IMGT®, the international ImMunoGenetics information system®, Laboratoire d'ImmunoGénétique Moléculaire LIGM, Institut de Génétique Humaine, (IGH), Centre National de la Recherche Scientifique (CNRS), Université de Montpellier (UM), Montpellier, France*

Cornelia Eckert • *Department of Pediatric Oncology Hematology, Charité - Universitätsmedizin Berlin, Berlin, Germany; German Cancer Consortium, and German Cancer Research Center, Heidelberg, Germany*

Anne Eugster • *Center for Regenerative Therapies Dresden, Faculty of Medicine, TU Dresden, Dresden, Germany*

Ahmed S. Fahad • *Department of Pharmaceutical Chemistry, The University of Kansas, Lawrence, KS, USA*

Andrew Farmer • *Takara Bio USA, Inc., San Jose, CA, USA*

Simone Ferrero • *Hematology Division, Department of Molecular Biotechnologies and Health Sciences, University of Torino, Torino, Italy; Hematology Division, AOU "Città della Salute e della Scienza di Torino", Torino, Italy*

Géraldine Folch • *IMGT®, the international ImMunoGenetics information system®, Laboratoire d'ImmunoGénétique Moléculaire LIGM, Institut de Génétique Humaine, (IGH), Centre National de la Recherche Scientifique (CNRS), Université de Montpellier (UM), Montpellier, France*

Eva Fronkova • *CLIP - Childhood Leukaemia Investigation Prague, Department of Paediatric Haematology and Oncology, Second Faculty of Medicine, Charles University and University Hospital Motol, Prague, Czech Republic*

Chrysi Galigalidou • *Institute of Applied Biosciences, Centre for Research and Technology Hellas, Thessaloniki, Greece; Department of Molecular Biology and Genetics (MBG), Democritus University of Thrace, Alexandroupolis, Greece*

Jana Gazdova • *Patrick G Johnston Centre for Cancer Research, Queen's University Belfast, Belfast, UK*

Elisa Genuardi • *Hematology Division, Department of Molecular Biotechnologies and Health Sciences, University of Torino, Torino, Italy*

Mathieu Giraud • *Université de Lille, CNRS, UMR 9189—CRIStAL, Inria, Lille, France*

Véronique Giudicelli • *IMGT®, the international ImMunoGenetics information system®, Laboratoire d'ImmunoGénétique Moléculaire LIGM, Institut de Génétique Humaine, (IGH), Centre National de la Recherche Scientifique (CNRS), Université de Montpellier (UM), Montpellier, France*

David Gonzalez • *Patrick G Johnston Centre for Cancer Research, Queen's University Belfast, Belfast, UK*

Victor Greiff • *Department of Immunology, University of Oslo, Oslo University Hospital, Oslo, Norway*

Patricia J. T. A. Groenen • *Department of Pathology, Radboud University Medical Center, Nijmegen, The Netherlands*

Nidhi Gupta • *Takara Bio USA, Inc., San Jose, CA, USA*

Matías Gutiérrez-González • *Department of Pharmaceutical Chemistry, The University of Kansas, Lawrence, KS, USA; Department of Chemical Engineering, The University of Kansas, Lawrence, KS, USA*

Dora Hammerl • *Laboratory of Tumor Immunology, Department of Medical Oncology, Erasmus MC-Cancer Institute, Rotterdam, The Netherlands*

Louise Harewood • *Patrick G Johnston Centre for Cancer Research, Queen's University Belfast, Belfast, UK*

Alina M. Hartmann • *Department of Hematology, University of Schleswig-Holstein, Kiel, Germany*

Kenneth B. Hoehn • *Department of Pathology, Yale School of Medicine, New Haven, CT, USA*

Kalisse I. Horne • *Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA*

Hanna IJspeert • *Department of Immunology, Laboratory Medical Immunology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands; Academic Center for Rare Immunological Diseases (RIDC), Erasmus University Medical Center, Rotterdam, The Netherlands*

Joumana Jabado-Michaloud • *IMGT®, the international ImMunoGenetics information system®, Laboratoire d'ImmunoGénétique Moléculaire LIGM, Institut de Génétique*

Humaine, (IGH), Centre National de la Recherche Scientifique (CNRS), Université de Montpellier (UM), Montpellier, France

Mouhamad Khouja • *Medical Department II, University Hospital Schleswig-Holstein, Kiel, Germany*

Dian Kortleve • *Laboratory of Tumor Immunology, Department of Medical Oncology, Erasmus MC-Cancer Institute, Rotterdam, The Netherlands*

Sofia Kossida • *IMGT®, the international ImMunoGenetics information system®, Laboratoire d'ImmunoGénétique Moléculaire LIGM, Institut de Génétique Humaine, (IGH), Centre National de la Recherche Scientifique (CNRS), Université de Montpellier (UM), Montpellier, France*

Michaela Kotrova • *Medical Department II, University Hospital Schleswig-Holstein, Kiel, Germany*

Gloria Kraus • *Center for Regenerative Therapies Dresden, Faculty of Medicine, TU Dresden, Dresden, Germany*

Anton W. Langerak • *Laboratory Medical Immunology, Department of Immunology, Erasmus MC, University Medical Center, Rotterdam, The Netherlands*

Anne Langlois de Septenville • *AP-HP, Pitié-Salpêtrière Hospital, Laboratory of Hematology, Paris, France*

William D. Lees • *Institute of Structural and Molecular Biology, Birkbeck College, University of London, London, UK*

Gérard Lefranc • *IMGT®, the international ImMunoGenetics information system®, Laboratoire d'ImmunoGénétique Moléculaire LIGM, Institut de Génétique Humaine IGH, UMR 9002, CNRS, Université de Montpellier, Montpellier cedex 5, France*

Marie-Paule Lefranc • *IMGT®, the international ImMunoGenetics information system®, Laboratoire d'ImmunoGénétique Moléculaire LIGM, Institut de Génétique Humaine IGH, UMR 9002, Centre National de la Recherche Scientifique (CNRS), Université de Montpellier (UM), Montpellier cedex 5, France*

Ida Lindeman • *Department of Immunology, Oslo University Hospital and K.G. Jebsen Coeliac Disease Research Centre, University of Oslo, Oslo, Norway*

Nicolle H. R. Litjens • *Erasmus MC Transplant Institute, Division of Nephrology and Transplantation, Department of Internal Medicine, Erasmus MC University Medical Center, Rotterdam, The Netherlands*

Sheila N. López Acevedo • *Department of Pharmaceutical Chemistry, The University of Kansas, Lawrence, KS, USA; Department of Chemical Engineering, The University of Kansas, Lawrence, KS, USA*

Eline T. Luning Prak • *Department of Pathology and Laboratory Medicine, Perelman School of Medicine, The University of Pennsylvania, Philadelphia, PA, USA*

Elizabeth Macintyre • *Hôpital Necker Enfants-Malades, Laboratoire d'Onco-Hématologie, Assistance Publique– Hôpitaux de Paris, Paris, France*

Encarnita Mariotti-Ferrandiz • *INSERM, Immunology-Immunopathology-Immunotherapy (i3), Sorbonne Université, Paris, France*

Susanna Marquez • *Department of Pathology, Yale School of Medicine, New Haven, CT, USA*

Kristen Martins-Taylor • *10x Genomics, Pleasanton, CA, USA*

Manisha Maurya • *Patrick G Johnston Centre for Cancer Research, Queen's University Belfast, Belfast, UK*

Wenzhao Meng • *Department of Pathology and Laboratory Medicine, Perelman School of Medicine, The University of Pennsylvania, Philadelphia, PA, USA*

Enkelejda Miho • *Institute of Biomedical Engineering and Medical Informatics, School of Life Sciences, FHNW University of Applied Sciences and Arts Northwestern Switzerland, Muttenz, Switzerland; SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland; aiNET GmbH, Basel, Switzerland*

Kevin Mujangi-Ebeka • *INSERM, Immunology-Immunopathology-Immunotherapy (i3), Sorbonne Université, Paris, France*

Rogier J. Nell • *Department of Ophthalmology, Leiden University Medical Center, Leiden, The Netherlands*

Kira C. M. Neller • *Health Sciences, Simon Fraser University, Burnaby, BC, Canada*

Mats Ohlin • *Department of Immunotechnology, Lund University, Lund, Sweden*

Christiane Pott • *Medical Department II, University Hospital Schleswig-Holstein, Kiel, Germany*

Fotis Psomopoulos • *Institute of Applied Biosciences, Centre for Research and Technology Hellas, Thessaloniki, Greece; Department of Molecular Medicine and Surgery, Karolinska Institute, Stockholm, Sweden*

Susanne Reinhardt • *DRESDEN-concept Genome Center, DFG NGS Competence Center, c/o Center for Molecular and Cellular Bioengineering (CMCB), Technische Universität Dresden, Dresden, Germany*

Julia Revolta • *Patrick G Johnston Centre for Cancer Research, Queen's University Belfast, Belfast, UK*

Maël Rollin • *IMGT®, the international ImMunoGenetics information system®, Laboratoire d'ImmunoGénétique Moléculaire LIGM, Institut de Génétique Humaine, (IGH), Centre National de la Recherche Scientifique (CNRS), Université de Montpellier (UM), Montpellier, France*

Aaron M. Rosenfeld • *Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA*

Florian Rubelt • *Roche Sequencing Solutions, Roche, Pleasanton, CA, USA*

Mikaël Salson • *Université de Lille, CNRS, UMR 9189—CRIStAL, Inria, Lille, France*

Blanca Scheijen • *Department of Pathology, Radboud University Medical Center, Nijmegen, The Netherlands; Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, The Netherlands*

Chaim A. Schramm • *Vaccine Research Center, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD, USA*

Nicholas Schwab • *Department of Neurology with Institute of Translational Neurology, University of Muenster, Muenster, Germany*

Cinque Soto • *The Vanderbilt Vaccine Center and Department of Pediatrics, Vanderbilt University Medical Center, Nashville, TN, USA*

Shambhavi Srivastava • *Patrick G Johnston Centre for Cancer Research, Queen's University Belfast, Belfast, UK*

Kostas Stamatopoulos • *Institute of Applied Biosciences, Centre for Research and Technology Hellas, Thessaloniki, Greece; Department of Molecular Medicine and Surgery, Karolinska Institute, Stockholm, Sweden*

Irene Della Starza • *Hematology, Department of Translational and Precision Medicine, "Sapienza" University of Rome, Rome, Italy; GIMEMA Foundation, Rome, Italy*

Ulrik Stervbo • *Center for Translational Medicine, Immunology, and Transplantation, Medical Department, and Immundiagnostik, Marien Hospital Herne, University Hospital of the Ruhr-University Bochum, Herne, Germany*

James Peter Stewart • *Patrick G Johnston Centre for Cancer Research, Queen's University Belfast, Belfast, UK*
Michael Svaton • *CLIP - Childhood Leukaemia Investigation Prague, Department of Paediatric Haematology and Oncology, Second Faculty of Medicine, Charles University and University Hospital Motol, Prague, Czech Republic*
Florian Thonier • *Inria, Rennes, France*
Jan Trka • *CLIP - Childhood Leukaemia Investigation Prague, Department of Paediatric Haematology and Oncology, Second Faculty of Medicine, Charles University and University Hospital Motol, Prague, Czech Republic*
Johannes Trück • *Division of Immunology and Children's Research Center, University Children's Hospital Zurich, University of Zurich (UZH), Zurich, Switzerland*
Diede A. G. van Bladel • *Department of Pathology, Radboud University Medical Center, Nijmegen, The Netherlands; Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, The Netherlands*
Mandy van Brakel • *Laboratory of Tumor Immunology, Department of Medical Oncology, Erasmus MC-Cancer Institute, Rotterdam, The Netherlands*
Mirjam van der Burg • *Department of Pediatrics, Laboratory for Pediatric Immunology, Willem-Alexander Children's Hospital, Leiden University Medical Center, Leiden, The Netherlands*
Jessica L. M. van der Last-Kempkes • *Department of Pathology, Radboud University Medical Center, Nijmegen, The Netherlands*
Pieter A. van der Velden • *Department of Ophthalmology, Leiden University Medical Center, Leiden, The Netherlands*
Vincent H. J. van der Velden • *Department of Immunology, Laboratory Medical Immunology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands*
Wilfred F. J. van IJcken • *Center for Biomics, Erasmus MC University Medical Center, Rotterdam, The Netherlands; Department of Cell Biology, Erasmus MC University Medical Center, Rotterdam, The Netherlands*
Pauline A. van Schouwenburg • *Department of Pediatrics, Laboratory for Pediatric Immunology, Willem-Alexander Children's Hospital, Leiden University Medical Center, Leiden, The Netherlands*
Mieke Versluis • *Department of Ophthalmology, Leiden University Medical Center, Leiden, The Netherlands*
Patrick Villarese • *Hôpital Necker Enfants-Malades, Laboratoire d'Onco-Hématologie, Assistance Publique– Hôpitaux de Paris, Paris, France*
Corey T. Watson • *Department of Biochemistry and Molecular Genetics, University of Louisville School of Medicine, Louisville, KY, USA*
Rebecca Wijers • *Laboratory of Tumor Immunology, Department of Medical Oncology, Erasmus MC-Cancer Institute, Rotterdam, The Netherlands*
Laura Zaragoza-Infante • *Institute of Applied Biosciences, Centre for Research and Technology Hellas, Thessaloniki, Greece; First Department of Medicine, Aristotle University of Thessaloniki, Thessaloniki, Greece*
Willem H. Zoutman • *Department of Dermatology, Leiden University Medical Center, Leiden, The Netherlands*

Chapter 1

The Advent of Precision Immunology: Immunogenetics at the Center of Immune Cell Analysis in Health and Disease

Anton W. Langerak

Abstract

Adaptive immune cells (i.e., lymphocytes of the B and T lineage) are equipped with unique antigen receptors, which collectively form a highly diverse repertoire. Within the lymphocytes, the antigen receptor diversity is created at the DNA level through recombination processes in the immunoglobulin (IG) and T cell receptor (TR) genes that encode these receptors. This gives rise to an enormous immune repertoire (a.k.a. the "immunome") that can be studied in health and disease, both in a scientific and clinical context. In fact, the inherent distinctiveness of the IG/TR rearrangements on a per cell basis allows their usage as unique DNA fingerprints, which enables precision medicine, or for that matter "precision immunology." The field of (fundamental and translational) research on IG/TR repertoire diversity is the topic of the *Immunogenetics* volume in the *Methods in Molecular Biology* series.

Key words Immunoglobulin, T cell receptor, Immunogenetics, Immunome, Precision immunology

1 Introduction

Our current understanding of the diversity of antigen receptors started with the publication on "Somatic generation of antibody diversity" by Susumu Tonegawa in 1983 [1], which resulted in the Nobel Prize in Physiology for the author in 1989. In this seminal publication, Tonegawa introduced the concept of genetic recombination mechanisms of V (variable), D (diversity), and J (joining) genes in the loci encoding the immunoglobulin (IG) chains, which—as was subsequently discovered—also applies to the T cell receptor (TR) loci. These recombinations lead to an enormous repertoire diversity of B and T lymphocytes, referred to as the "immunome." The research into the genetics of the immune cell repertoire has been termed "immunogenetics." Besides IG/TR gene diversity, the field of immunogenetics formally also includes diversity in the human leukocyte antigens (HLA), but this is largely beyond the scope of the current *Immunogenetics* volume in the *Methods in Molecular Biology* series.

Anton W. Langerak (ed.), *Immunogenetics: Methods and Protocols*, Methods in Molecular Biology, vol. 2453, https://doi.org/10.1007/978-1-0716-2115-8_1,

2 Immunogenetics in the Hematology-Immunology Domain

B and T lymphocyte populations and their respective IG/TR repertoires are mostly studied in the context of immune diseases (autoimmune diseases, allergies, immune deficiencies) and immune responses (infections, inflammation, vaccinology, cancer), but also frequently in the context of hematological malignancies of immune cells (leukemias and lymphomas).

Irrespective of the application, it is important, when evaluating IG and TR repertoire diversity in B and T cell populations, to consider the repertoire data as being part of a spectrum ranging from broadly diverse (polyclonal), to restricted (oligoclonal), to dominant (clonal +/− poly/oligoclonal background) (Fig. 1). This spectrum reflects the minimal to moderate to dominant outgrowth of B or T lymphocytes of a particular specificity, which are selected based on their antigen reactivity.

Immunogenetic analysis can provide in-depth insight into the diversity of immune cells and immune responses in the context of different research questions. Additionally, the diversity or clonality of the immune repertoire can also help to address clinical and diagnostic questions. In the hematological domain, this relates to the distinction between reactive lymphoproliferations (poly- to oligoclonality) and malignantly transformed lymphocytes

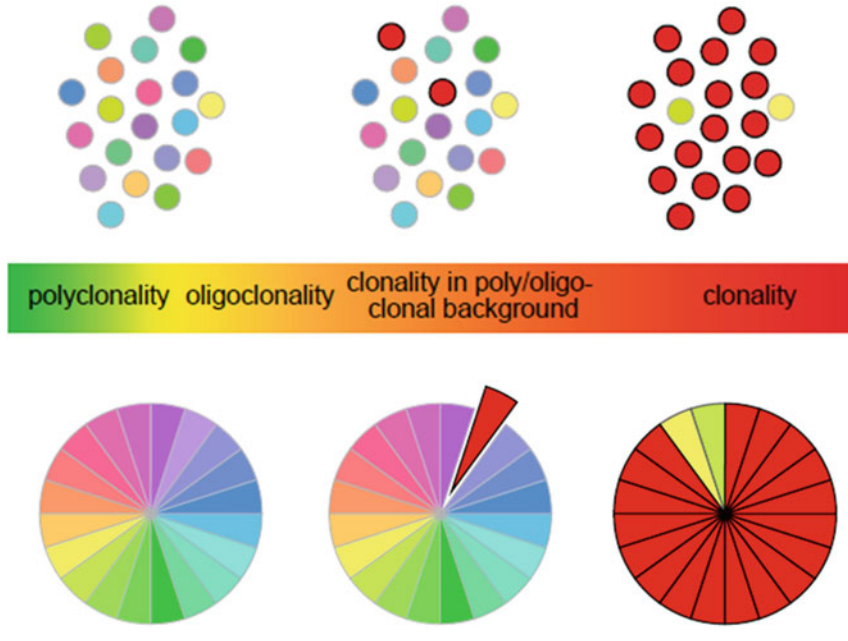


Fig. 1 Spectrum of IG/TR immune repertoire diversity, ranging from diverse (polyclonal) to highly restricted (clonal), which can be disclosed using high-throughput sequencing technologies. (Adapted from *Langerak, J Immunol 2017;198:3765* [2])

(clonality) or to detection of minimal residual disease of a clone upon treatment (weak clonality in background). In other areas of medicine, immunogenetic analysis can shed light on proper or defective immune responses in infected and vaccinated individuals and/or can help to distinguish between disease entities (e.g., in due time for particular autoimmune IG/TR profiles).

3 Immunogenetics Methods

Historically, immunogenetic analysis has been performed using low-resolution methodologies, such as Southern blot analysis, fragment analysis or spectratyping, and Sanger sequencing of cloned, rearranged IG/TR genes [3]. Even though these approaches enabled us to grasp the diversity of antigen receptors to some extent, they suffered from limitations in completely disclosing the depth and broadness of the IG/TR immune repertoire. The introduction of high-throughput technologies some 15 years ago allowed for a more high-resolution immune repertoire analysis via massively parallel sequencing (Fig. 2). These next-generation

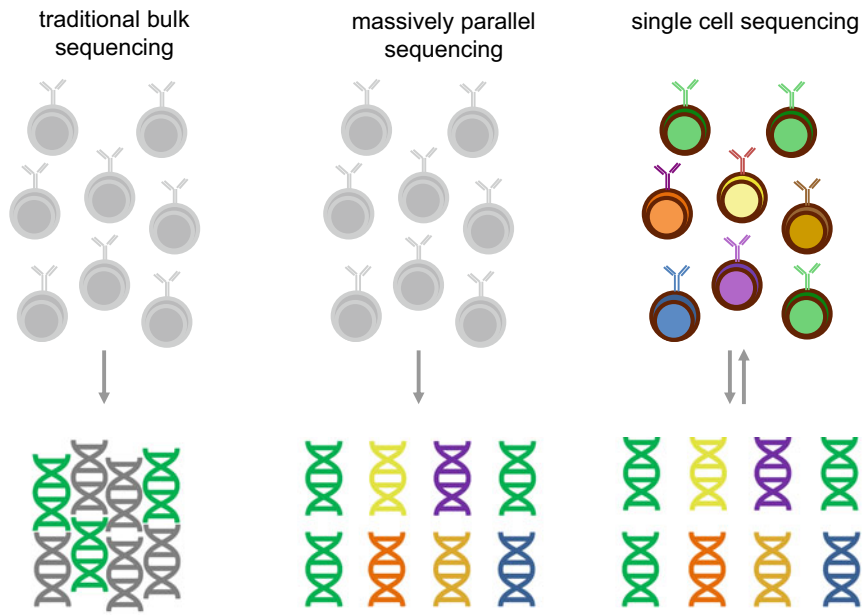


Fig. 2 Graphical representation of different sequencing approaches for IG/TR repertoire analysis. By means of traditional (Sanger) bulk sequencing, only the dominant immune repertoire (in green) can be identified over the background (grey), which strongly contrasts with the high-resolution output of many individual IG/TR rearrangements (represented by the different colors) through massively parallel sequencing. The additional advantage of single-cell sequencing technologies is that the high-resolution IG/TR repertoire analysis can be traced back to individual cells, which allows evaluation of paired IG or TR chains at the single-cell level and/or combination of immune repertoire and differentiation or maturation stage features

sequencing methods have the advantage that thousands to millions of IG/TR rearrangement sequences can be analyzed in parallel, thus approximating the true IG/TR repertoire diversity much more closely. A further development has been the introduction of single-cell sequencing technologies (Fig. 2), allowing paired analysis of different IG or TR chains at the single-cell level and the combination of immune repertoire analysis with RNA sequencing-based cell characteristics (e.g., naïve vs. memory, activated or exhausted cells).

4 (Pre- and Post-)Analytical Aspects of Immunogenetics

As with any experimental method, immune repertoire analysis also entails pre-analytical, analytical, and post-analytical phases. For immune repertoire studies, the pre-analytical considerations specifically focus around the choice of sample type, nucleic acid type, IG/TR targets, etc., whereas the analytical phase relates to the pros and cons of the applied method (next-generation sequencing, quantitative PCR, droplet digital PCR). Finally, the post-analytical phase involves the readouts and tools for data analysis, but also the immuno-informatics to accurately annotate the IG/TR sequences and the bioinformatic pipelines and platforms that allow sophisticated analysis of the IG/TR data and all of their characteristic features (gene usage, CDR3, somatic mutations, clustering, and clonal evolution and competition).

In this volume of the *Methods in Molecular Biology* series, all of the above aspects of the pre-analytical, analytical, and post-analytical phases of IG/TR repertoire analysis are addressed in different methodological chapters that together cover a spectrum of technologies, ranging from quantitative and droplet digital PCR approaches to various NGS methodologies such as amplicon-based, capture-based, and single-cell NGS. Additionally, bioinformatic approaches are discussed that allow for extraction of IG/TR repertoire sequences from -omics data sets, i.e., RNA sequencing, whole genome sequencing, and whole exome sequencing. Finally, several novel approaches in the immunogenetic domain are covered, concerning cell-free IG/TR analysis, analysis of germline areas of the TR loci, analysis of aberrantly rearranged IG genes leading to IG translocations, and engineering of TR sequences in view of adoptive therapy.

5 Immunogenetics at the Basis of Precision Immunology

Collectively, the chapters in this volume are a perfect illustration of the central position that immunogenetics has obtained in the hematology-immunology domain in both health and disease

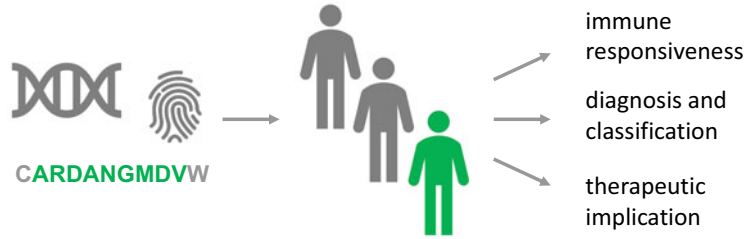


Fig. 3 Precision immunology through immunogenetic analysis. Characteristic IG/TR CDR3 profiles allow identification of individual patients. These profiles have implications to define immune responsiveness, to make diagnosis and/or sub-classification, or even support therapeutic choices

[2]. Immunogenetic profiles constitute physiological and pathophysiological signatures of cell populations, thereby allowing a more personalized approach in terms of immune responsiveness, diagnostics and classification, and even therapeutic choices [4]. This form of precision medicine involving immunogenetics could therefore best be referred to as "precision immunology" (Fig. 3). The future of immunogenetics is bright!

References

1. Tonegawa S (1983) Somatic generation of antibody diversity. Nature 302(5909):575–581
2. Langerak AW, Brüggemann M, Davi F, Darzentas N, van Dongen JJM, Gonzalez D et al (2017) High-throughput immunogenetics for clinical and research applications in immunohematology: potential and challenges. J Immunol 198:3765–3774
3. Van Dongen JJ, Langerak AW, Bruggemann M, Evans PA, Hummel M, Lavender FL et al (2003) Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 concerted action BMH4-CT98–3936. Leukemia 17:2257–2317
4. Arnaout RA, Prak ETL, Schwab N, Rubelt F, Adaptive Immune Receptor Repertoire Community (2021) The future of blood testing is the immunome. Front Immunol 12:626793

Chapter 2

Next-Generation Sequencing-Based Clonality Detection of Immunoglobulin Gene Rearrangements in B-Cell Lymphoma

Diede A. G. van Bladel, Jessica L. M. van der Last-Kempkes, Blanca Scheijen, Patricia J. T. A. Groenen, and on behalf of the EuroClonality Consortium

Abstract

Immunoglobulin (IG) clonality assessment is a widely used supplementary test for the diagnosis of suspected lymphoid malignancies. The specific rearrangements of the immunoglobulin (IG) heavy and light chain genes act as a unique hallmark of a B-cell lymphoma, a feature that is used in clonality assessment. The widely used BIOMED-2/EuroClonality IG clonality assay, visualized by GeneScanning or heteroduplex analysis, has an unprecedented high detection rate because of the complementarity of this approach. However, the BIOMED-2/EuroClonality clonality assays have been developed for the assessment of specimens with optimal DNA quality. Further improvements for the assessment of samples with suboptimal DNA quality, such as from formalin-fixed paraffin-embedded (FFPE) specimens or specimens with a limited tumor burden, are required. The EuroClonality-NGS Working Group recently developed a next-generation sequencing (NGS)-based clonality assay for the detection of the IG heavy and kappa light chain rearrangements, using the same complementary approach as in the conventional assay. By employing next-generation sequencing, both the sensitivity and specificity of the clonality assay have increased, which not only is very useful for diagnostic clonality testing but also allows robust comparison of clonality patterns in a patient with multiple lymphoma's that have suboptimal DNA quality. Here, we describe the protocols for IG-NGS clonality assessment that are compatible for Ion Torrent and Illumina sequencing platforms including pre-analytical DNA isolation, the analytical phase, and the post-analytical data analysis.

Key words Clonality analysis, Next-generation sequencing, B-cell lymphoma, Immunoglobulin gene rearrangements, ARResT/Interrogate

1 Introduction

Clonality assessment of the immunoglobulin (IG) or T-cell receptor genes is a useful supplementary tool for the diagnosis of B-cell and T-cell lymphoid malignancies. Cancer cells have a unique feature that they originate from a single transformed cell. The malignant cells of a B-cell lymphoma all have the same rearranged IG

Anton W. Langerak (ed.), *Immunogenetics: Methods and Protocols*, Methods in Molecular Biology, vol. 2453, https://doi.org/10.1007/978-1-0716-2115-8_2,

DNA sequences encoding for a unique antigen-receptor molecule, also called the B-cell receptor (BCR). Clonality assessment makes use of this feature. In patients suspected for having a B-cell lymphoma, clonality assessment enables demonstration of a clonal expansion of clonally related B cells, all having the identical molecular footprint of the antigen receptor encoded by the IG genes.

1.1 Immunoglobulin Gene Rearrangements

The BCR consists of two IG heavy chains (IGH) and two light chains, IG kappa (IGK) or IG lambda (IGL), with unique nucleotide sequences at the antigen binding region that are generated during lymphoid development. The proper assembly of a functional BCR is controlled by several checkpoints at different stages of B-cell development [1–3]. Once a mature B cell has encountered an antigen, it will undergo somatic hypermutation (SHM) in the germinal center. During this process that is mediated by the enzyme activation-induced cytidine deaminase (AID), random sequence alterations [mostly point mutations, but deletions or insertions can occur as well] are introduced to improve antigen binding, a phenomenon called affinity maturation [2, 3].

The BCR is generated by a stepwise process involving rearrangements of the different germline variable (V), diversity (D), and joining (J) IG genes, called V(D)J recombination. This process is initiated by the recombination-activating gene (RAG) products RAG1 and RAG2 [4, 5], which relies on the recognition of recombination signal sequences (RSSs) flanking the individual genes. V(D)J recombination starts with the IG heavy chain, by the recombination of one of the D genes with one of the J genes, followed by the subsequent joining of one of the V genes to the rearranged DJ gene (Fig. 1). This random recombination of V, D, and J genes generates the so-called combinatorial diversity. Imprecise joining of the genes by the activation of exonucleases, as well as the addition of non-template DNA nucleotides by the enzyme terminal deoxynucleotidyl transferase (TdT), results in junctional diversity, on top of the combinatorial diversity. As a consequence of the combinatorial and junctional diversity, only one out of three VDJ rearrangements will be able to express a functional BCR. This high frequency of out-of-frame rearrangements may explain why many of the B lymphocytes have rearranged both their IGH genes, so-called biallelic IGH gene rearrangements. Lymphomas with biallelic gene rearrangements occur frequently, whereas lymphomas that are truly bi-clonal are rare [7].

For the light chain (IGK or IGL), a direct V to J gene rearrangement takes place, where the IGK locus will first undergo gene rearrangement. When there is no productive IGKV-IGKJ rearrangement, additional rearrangements will occur that inactivate the IGK locus by removal of the IGKC region and the enhancers.

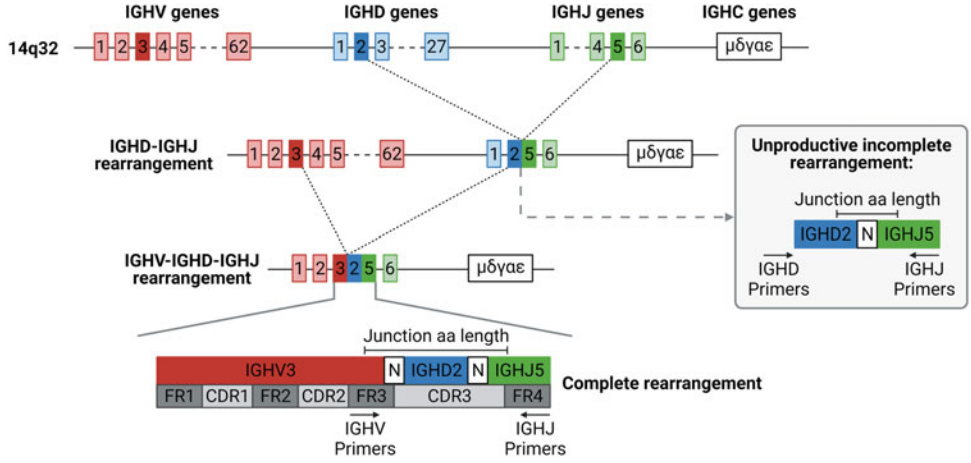


Fig. 1 Detection of V(D)J gene rearrangements at the immunoglobulin heavy chain locus. After a functional DJ rearrangement has been generated, a V gene is joined to this DJ fragment. Each B cell may generate one (productive rearrangement) or two (an unproductive and productive rearrangement) specific clonotypes that consist of one IGHV, IGHD, and IGHJ gene segment. The locations of the primers used for IG-NGS clonality assessment are indicated by arrows. For detection of IGHV-IGHD-IGHJ gene rearrangements, the forward primers are located in framework region 3 (VH FR3), which are combined with IGHJ reverse primers. The detection of unproductive, incomplete IGHD-IGHJ rearrangements makes use of forward IGHD primers (located 5′ of the IGHD genes) and reverse IGHJ primers, hence enabling detection of incompletely rearranged IGHD-IDHJ joining. Once an IGHV gene is recombined to the IGHD-IGHJ segment, the IGHD primer binding site will be removed. Successful amplification will result in DNA fragments that cover the junctional region with a specific amino acid length. Figure adapted from Scheijen et al., 2019 [6]

These rearrangements involve the KDE sequence that can rearrange to one of the kappa V genes and thereby delete the initial IGKV-IGKJ rearrangement, resulting in an IGKV-KDE rearrangement or to an isolated recombination signal sequence (RSS) that is located in the J kappa-C kappa intron (intron RSS), resulting in an Intron RSS-KDE rearrangement [8] (Fig. 2). If there is no proper IGK rearrangement, the IGL genes will rearrange. Theoretically, all mature B-cell malignancies should possess IGK rearrangements, regardless of the light chain expression [9]. Based on the amount of functional genes, the estimated number of unique BCRs generated by combinational diversity of both the heavy and light chain is 4.6×10^6 [10]. However, the actual number of unique receptors is lower, since not all genes are used at the same frequency, and not every heavy and light chain can pair to form a functional BCR.

The junctional diversity further increases BCR diversity by a factor 10.

B cells that assembled a functional BCR will further diversify by undergoing somatic SHM to extend the IG repertoire upon antigen recognition within the germinal center of a lymph node [2, 11]. When B cells fail or become autoreactive during this process, they will be silenced and eliminated [1, 3].

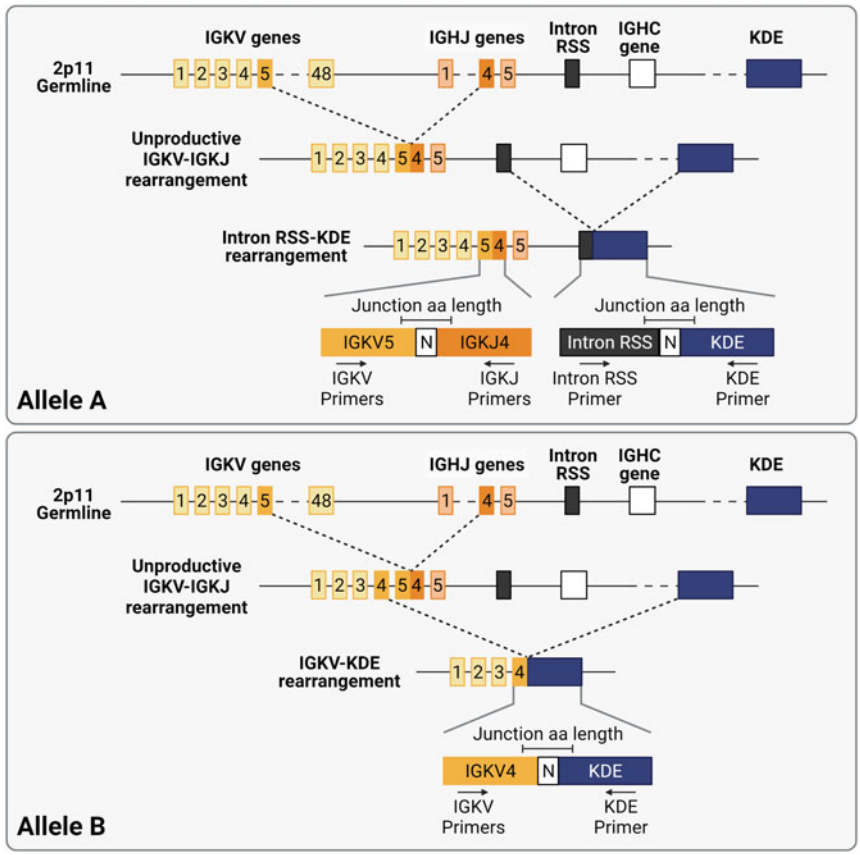


Fig. 2 IGK rearrangements involving Kappa deleting element. IGK gene rearrangement starts with an initial IGKV-IGKJ recombination. If this results in a productive rearrangement, no subsequent recombination events will occur within the IGK locus. However, in case there is an unproductive IGKV-IGKJ rearrangement, this may lead to inactivation of the IGK locus involving rearrangements with the Kappa deleting element (KDE) sequence. This can include a rearrangement between KDE and Intron RSS-KDE recombination on the same allele (Allele A). The initially formed unproductive IGKV-IGKJ segment remains present on that allele. Both the unproductive IGKV-IGKJ and Intron RSS-KDE rearrangements are detectable with clonality analysis. The second option, involves a recombination of an upstream IGKV gene with the KDE sequence, thereby deleting the preexisting unproductive IGKV-IGKJ rearrangement on that allele (Allele B). Potentially, up to four distinct IGK rearrangements can be generated that go along in one B cell clone. The locations of the primers used for IG-NGS clonality assessment are indicated by arrows. Figure adapted from Scheijen et al., 2019 [6]

1.2 Clonality Detection in B-Cell Lymphoma Based on BIOMED-2/ EuroClonality Assays

Clonality assessment by detecting IG gene rearrangements is widely used for diagnostics, and multiple assays have been developed over the years, which differ in the level of sensitivity [12]. The current gold standard are the PCR-based BIOMED-2/EuroClonality assays, visualized with either GeneScan fragment analysis or heteroduplex analysis [13, 14]. In this assay, standardized PCR protocols are used that cover IGH and IGK gene rearrangements. These include complete IGHV-IGHD-IGHJ rearrangements but also

incomplete IGHD-IGHJ rearrangements, which are not affected by somatic hypermutation either. For IGK gene rearrangements, not only IGKV-IGKJ rearrangements are included but also rearrangements involving KDE, which are not affected by somatic hypermutation. Notably, these occur on one or both alleles in virtually all IgLambda-positive B-cell malignancies and in one-third of the IgKappa-positive B-cell malignancies. The primers and protocols of the BIOMED-2/EuroClonality PCR assays allow detection of virtually all clonal B-cell proliferations, and the primer design has been based on family primers and consensus primers relevant for the IG genes. A clonal cell population gives rise to one or two dominant PCR products of a given size on GeneScan. A polyclonal cell population will result in a range of differently sized PCR fragments, corresponding to the presence of different V(D)J gene rearrangements showing Gaussian distribution with respect to the amount of inserted or deleted nucleotides in the junctional region.

The BIOMED-2/EuroClonality assays are used worldwide and have resulted in increased clonality detection of lymphoid malignancies [15, 16]. However, there are still some drawbacks that could potentially yield (mainly) false-negative results. The BIOMED-2/EuroClonality assays have been designed for high-quality DNA samples generating amplicons in the range of 150–400 bp. However, formalin-fixed paraffin-embedded (FFPE) tissue specimens, which are mostly used in a diagnostic setting, may yield DNA samples of inferior quality. Clonal rearrangements that correspond to relatively longer amplicons may therefore potentially be missed [13, 15, 17]. Furthermore, detection of minor clones in a background of nonmalignant B cells is highly dependent on the position of the clonal product within the Gaussian curve of the polyclonal background, where it can be difficult or even impossible to detect these minor clones.

1.3 NGS-Based Clonality Detection in B-Cell Lymphomas

To further improve the application potential of clonality assessment, the EuroClonality-NGS Working Group has developed a novel next-generation sequencing (NGS)-based clonality assay for detection of IG gene rearrangements (IG-NGS) [6], together with the bioinformatics tool ARResT/Interrogate [18]. New primers were designed for the incomplete and complete IGH gene rearrangements, the complete IGK rearrangements as well as for the IGK rearrangements involving KDE, again making use of the complementary approach that is one of the strengths of the conventional BIOMED-2/EuroClonality assays. The primer design for the NGS-based clonality assay is based on gene-specific primers for the relevant genes and, importantly, on the generation of shorter amplicon sizes, which makes it more suitable for clonality detection in samples of inferior DNA quality. Furthermore, the IG-NGS assay immediately provides the nucleotide sequences of the identified

clonotypes from both the malignant clone and the nonmalignant background B cells. Using this sequence information, reliable detection of minor clones is possible, resulting in a high sensitivity of the clonality analysis as recently described by Scheijen et al. [6]. Clonal rearrangements of lymphomas with a high tumor load still can be traced back when diluted in a concentration of 5% and 2.5% in a polyclonal background of tonsil DNA. The detection rate of 2.5% is not possible by the conventional assay combined with GeneScanning or heteroduplex, because the clonal product will be blurred by the polyclonal background [13]. Furthermore, the sequence information, the design for suboptimal DNA specimen, and the sensitivity are extremely valuable for comparison of sequential lesions or multiple lymphomas at different locations in a single patient.

1.4 Different NGS Platforms for Clonality Testing: Ion Torrent Versus Illumina

Similar to the BIOMED-2 approach, the IG-NGS clonality assay is based on a multiplex PCR to amplify the target regions and by subsequent ligation of adaptors for sequencing. The targets detected in the NGS clonality assay include IGH (IGHV-IGHD-IGHJ and IGHD-IGHJ) and IGK (IGKV-IGKJ, IGKV-KDE, and Intron RSS-KDE) gene rearrangements. After purification of the PCR products, the library preparation is performed, followed by sequencing on Ion Torrent or Illumina platforms (Fig. 3).

The initial IG-NGS workflow described the protocols for the Ion Torrent platform [6], a technique that makes use of electrochemical detection of hydrogen ions that are released during DNA synthesis [19]. The Illumina platform represents also a widely used NGS application in diagnostic laboratories, and both are very suitable for high-throughput NGS-based molecular assays. In contrast to Ion Torrent-based sequencing, Illumina employs fluorescently labeled nucleotides that are incorporated during complementary DNA strand synthesis [20]. Depending on the type of Illumina sequencer, this can be a 2-channel (e.g., MiniSeq, NextSeq, NovaSeq) or 4-channel chemistry (e.g., MiSeq, HiSeq).

The Ion Torrent and the Illumina sequencing technologies require specific adapters for sequencing and barcodes for sample identification. In the workflow that was developed for Ion Torrent sequencing, the adapters and barcodes are ligated to the amplicons (adapter ligation protocol). For Illumina sequencing platforms, the sequencing adapters need to be incorporated in the amplicon primers. Recently, the EuroClonality-NGS Working Group described a two-step PCR assay for minimal residual disease (MRD) target identification using an Illumina-compatible workflow [21]. With this approach, the barcoded adapter sequences are incorporated in the second PCR of the two-step PCR assay with universal barcoded M13-tailed primers. The workflow for clonality detection using the Illumina sequencing platform that will be described in this chapter

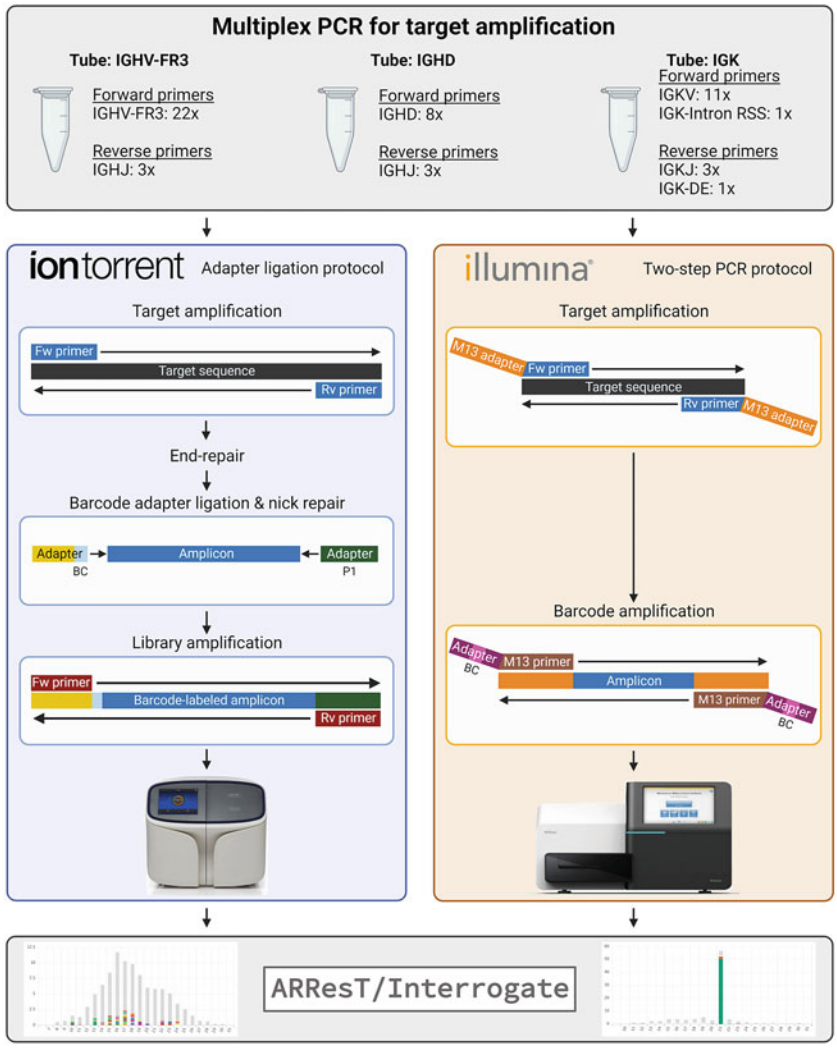


Fig. 3 Schematic workflow for IG-NGS clonality assay. A multiplex PCR is performed on extracted DNA of specimens suspect for lymphoproliferations to amplify IGHV-IGHD-IGHJ, IGHD-IGHJ, IGKV-IGKJ, and IGKV/Intron RSS-KDE gene rearrangements. Library preparation for sequencing on Ion Torrent (left panel) or Illumina (right panel) is shown. The Ion Torrent library preparation is an adapter ligation protocol, requiring end repair of the obtained amplicons and the ligation of barcode and adapters to them and nick repair, followed by a final library amplification step. Library preparation for Illumina is a two-step PCR protocol in which the target-specific amplicons are generated using primers containing an M13 adapter, which is used in the second PCR to add specific barcodes to them. Obtained sequencing data is analyzed using the bio-informatics tool ARResT/Interrogate

is based on this previously described two-step PCR protocol [21], with some minor modifications in the first PCR reaction and purification steps of the amplicons as well as the PCR conditions of the Ion Torrent protocol (Table 1) (*see* **Note 1**).

Table 1 note in the bookversion; Table 1 was split over 2 pages in not a nice way; if this Table should be split, please do so starting the second page with the row: PCR program (all targets)

PCR conditions for target amplification comparing different EuroClonality NGS protocols. The components of the PCR mixes and programs are shown for the Ion Torrent protocol for clonality detection, the first PCR step of the two-step Illumina protocol for clonality detection as described in this paper, and the previously published two-step Illumina protocol for marker identification [21]. Primer sequences and final concentrations are provided in Tables 2, 3, and 4

	Ion Torrent protocol for clonality detection			Two-step Illumina protocol for clonality detection			Two-step Illumina protocol for marker identification			
PCR mix	IGH-FR3	IGHD	IGK (IGKV-IGKJ + IGKV/intron RSS-KDE)	IGH-FR3	IGHD	IGK (IGKV-IGKJ + IGKV/intron RSS-KDE)	IGH-FR1	IGHD	IGKV-IGKJ +IGKV-KDE	Intron RSS-KDE
Input DNA	40 ng	40 ng	40 ng	40 ng	40 ng	40 ng	100 ng	100 ng	100 ng	100 ng
Buffer	1× GeneAmp Gold[a]	1× GeneAmp Gold[a]	1× GeneAmp Gold[a]	1× GeneAmp Gold[a]	1× GeneAmp Gold[a]	1× GeneAmp Gold[a]	1× PCR buffer II	1× PCR buffer II	1× PCR buffer II	1× PCR buffer II
Primers (μM)	0.2/0.4[b]	0.2[b]	0.2	0.2/0.4[b]	0.2[b]	0.2	0.1/0.2[b]	0.2/0.4[b]	0.1	0.1
dNTP (mM)	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
MgCl2 (mM)	1.5	2.0	1.5	1.5	2.0	1.5	2.5	3	1.5	1.5
Polymerase	0.5 U AmpliTaq[c]	0.5 U AmpliTaq[c]	0.5 U AmpliTaq[c]	0.5 U AmpliTaq[c]	0.5 U AmpliTaq[c]	0.5 U AmpliTaq[c]	1 U AmpliTaq[c] /EagleTaq	1 U AmpliTaq[c] /EagleTaq	1 U AmpliTaq[c] /EagleTaq	1 U AmpliTaq[c] /EagleTaq
Total reaction volume[d]	*25 μl*	*25 μl*	*25 μl*	*25 μl*	*25 μl*	*25 μl*	*50 μl*	*50 μl*	*50 μl*	*50 μl*

PCR program (all targets)	Time	Temp.	# cycles	Time	Temp.	# cycles	Time	Temp.	# cycles
Initial denaturation	10 min	94 °C	1	10 min	94 °C	1	10 min	94 °C	1
Denaturation	30 s	92 °C	30	30 s	92 °C	30	60 s	94 °C	35
Annealing	40 s	60 °C		40 s	60 °C		60 s	63 °C	
Extension	40 s	72 °C		40 s	72 °C		30 s	72 °C	
Final extension	30 min 10 min	72 °C 20 °C	1 1	30 min 10 min	72 °C 20 °C	1 1	30 min ∞	72 °C 20 °C	1 1

[a]GeneAmp buffer Gold
[b]IGH-FR3 reaction of two-step Illumina protocol for marker identification uses 2 IGHJ reverse primers and the IGHD reaction only 1 IGHJ reverse primer, where the two-step Illumina protocol for clonality detection and the Ion Torrent protocol use 3 IGHJ reverse primers for both the IGH-FR3 and IGHD reaction
[c]AmpliTaq Gold DNA polymerase
[d]Adjust the total reaction volume with MQ

In the subsequent paragraphs, we present a complete overview of the different steps of IG-NGS clonality analysis in suspected B-cell malignancies that are compatible for either Ion Torrent or Illumina sequencing platforms. For complete IGH rearrangements, in this NGS approach, framework-3 (FR3) primers are used in contrast to the BIOMED-2/EuroClonality assay that employs additional FR1 and FR2 primers, generating larger-sized products. Data analysis with ARResT/Interrogate and the technical interpretation and reporting of the obtained results will be addressed. It is of utmost importance that molecular clonality results are eventually interpreted in the context of available clinical, morphological, and immunophenotypic data. Also detailed knowledge of the immunobiology of IG gene rearrangements is mandatory to be able to correctly interpret the different molecular patterns.

2 Materials

2.1 General Materials and Equipment

1. Volume-adjustable single-channel and multichannel pipettes.
2. Filter tips.
3. Eppendorf tubes (0.5 ml, 1.5 ml, and 2 ml).
4. 0.2 ml PCR tubes/strips (and caps).
5. Vortex.
6. Centrifuge, e.g., Eppendorf Centrifuge 5420 or equivalent equipment.
7. Microcentrifuge, e.g., MiniStar/MiniStar blueline with tube and PCR rotors (VWR).
8. Ethanol 99%, absolute pro analyse (molecular biology quality grade).
9. Low TE-buffer ($T_{10}E_{0.1}$): 10 mM Tris-HCl pH 8.0 and 0.1 mM EDTA.
10. Nuclease-free water/Milli-Q (MQ).
11. Qubit dsDNA High Sensitivity Assay Kit (Thermo Fisher Scientific).
12. Qubit Assay tubes (Thermo Fisher Scientific).
13. Qubit Fluorometer (Thermo Fisher Scientific).
14. Thermal cyclers, e.g., Veriti 96-Well Thermal Cycler (Thermo Fisher Scientific), PTC-0200 (MJ Research) or equivalent equipment.

2.2 DNA Isolation

1. Xylene (molecular biology quality grade).
2. TET lysis buffer: 10 mM Tris/HCl pH 8.5, 1 mM EDTA pH 8.0, 0.01% Tween-20.

3. Chelex-100 (Bio-Rad).
4. TSE: 10 mM Tris–HCl pH 7.5, 0.4 M NaCl, 2 mM EDTA pH 8.0.
5. SDS 20%.
6. Proteinase K (QIAGEN).
7. Genomic DNA isolation kit, e.g., QIAamp DNA FFPE Tissue Kit (QIAGEN), QIAamp DNA Micro Kit (QIAGEN) (*see* **Note 2**).

2.3 IG-NGS Clonality Assays

2.3.1 Target Amplification and Purifications

1. dNTPs.
2. AmpliTaq Gold DNA Polymerase, kit with GeneAmp 10× Gold Buffer and $MgCl_2$ (Applied Biosystems).
3. Standard purified primers to be dissolved in Low TE-buffer at 300 pmol/μl (300 μM) (e.g., from Sigma-Aldrich; *see* Tables 2, 3, and 4 and **Note 3**).
4. DNA LoBind, Deepwell plate 96/500 μl, white border (Eppendorf).
5. Agencourt AMPure XP Beads (Beckman Coulter).
6. Dyna Mag-96side Magnet (Thermo Fisher Scientific).

2.3.2 Ion Torrent Library Preparation and Sequencing

1. Ion Plus Fragment Library kit (Life Technologies).
2. Ion Xpress Barcode Adapters kit (Life Technologies).
3. Ion 318™ Chip Kit v2 BC (Thermo Fisher Scientific) or Ion 520™ Chip Kit (Thermo Fisher Scientific).
4. Ion Torrent template preparation (Ion One Touch system, Ion Chef system) and sequencing equipment (Ion PGM, Ion GeneStudio S5).

2.3.3 Illumina Library Preparation and Sequencing

1. dNTPs.
2. HPLC purified Illumina M13-barcoded primers to be dissolved in Low TE-buffer at 100 pmol/μl (100 μM stock solution, dilute to 5 μM work solution) (e.g., from Sigma-Aldrich; *see* Table 5).
3. FastStart High Fidelity PCR system, dNTPack (Roche).
4. Sequencing equipment and associated Illumina Reagent Kit (e.g., MiniSeq sequencer and MiniSeq Mid Output Kit).

3 Methods

3.1 Samples and Quality Controls

IG-NGS clonality analysis can be performed on DNA extracted from any preserved human lymphoid tissue. However, each sample type requires a specific extraction procedure for DNA isolation. We here describe DNA extraction methods for formalin-fixed paraffin-

Table 2
Primers included in the multiplex PCR reaction for NGS-based clonality assessment: Tube IGHV-FR3

Primer nomenclature	Final concentration	Primer direction	Primer Sequence (with M13 adapter forward/reverse) 5' to 3'*
IGH-V-FR3-A-1	0.4 µM	5'	GTAAAACGACGGCCAGTAAGTTCCAGGGCAGAGTCAC
IGH-V-FR3-B-1	0.4 µM	5'	GTAAAACGACGGCCAGTGTCCATCAGCACAGCCTACA
IGH-V-FR3-C-1	0.4 µM	5'	GTAAAACGACGGCCAGTGACATGTCCACAAGCACAGC
IGH-V-FR3-D-1	0.2 µM	5'	GTAAAACGACGGCCAGTTCTCCAAGGACACCTCCAAGA
IGH-V-FR3-E-1	0.2 µM	5'	GTAAAACGACGGCCAGTCAGGCTCACCATCTCCAAGG
IGH-V-FR3-F-1	0.2 µM	5'	GTAAAACGACGGCCAGTCCATCTCTGAAGAGCAGGCT
IGH-V-FR3-G-1	0.4 µM	5'	GTAAAACGACGGCCAGTTGAAGGGCCGATTCACCATC
IGH-V-FR3-H-1	0.4 µM	5'	GTAAAACGACGGCCAGTAGGCAGATTCACCATCTCAAGA
IGH-V-FR3-I-1	0.4 µM	5'	GTAAAACGACGGCCAGTAGCGCCGATTCATCATCTCC
IGH-V-FR3-J-1	0.2 µM	5'	GTAAAACGACGGCCAGTCCAAAAGCATCACCTATCTGCA
IGH-V-FR3-K-1	0.2 µM	5'	GTAAAACGACGGCCAGTGAAGGGCCGGTTCACCATC
IGH-V-FR3-L-1	0.2 µM	5'	GTAAAACGACGGCCAGTACCTCCAGAGATAACGCCAAG
IGH-V-FR3-M-1	0.2 µM	5'	GTAAAACGACGGCCAGTCAGGAAGGGCAGATTCACCA
IGH-V-FR3-N-1	0.2 µM	5'	GTAAAACGACGGCCAGTGAAGGGCCGATTGACCATCTC
IGH-V-FR3-O-1	0.4 µM	5'	GTAAAACGACGGCCAGTCTCCGTGAAGGGCAGATTCA
IGH-V-FR3-P-1	0.2 µM	5'	GTAAAACGACGGCCAGTGATGATTCAAAGAACACGGCGT
IGH-V-FR3-Q-1	0.4 µM	5'	GTAAAACGACGGCCAGTCCGTCCCTCAAGAGTCGAGT
IGH-V-FR3-R-1	0.4 µM	5'	GTAAAACGACGGCCAGTCCGTCCCTCAAGAGTCGAAT
IGH-V-FR3-S-1	0.2 µM	5'	GTAAAACGACGGCCAGTGTCACCATCTCAGCCGACAA
IGH-V-FR3-T-1	0.2 µM	5'	GTAAAACGACGGCCAGTCAAGTCCATCAGCACTGCCT
IGH-V-FR3-U-1	0.4 µM	5'	GTAAAACGACGGCCAGTCAGTTCTCCCTGCAGCTGAA
IGH-V-FR3-V-1	0.4 µM	5'	GTAAAACGACGGCCAGTGGCTTCACAGGACGGTTTGT
IGH-J-A-1	0.2 µM	3'	TAATACGACTCACTATAGGGCTTACCTGAGGAGACGGTGACC
IGH-J-B-1	0.2 µM	3'	TAATACGACTCACTATAGGGCTCACCTGAGGAGACAGTGACC
IGH-J-C-1	0.2 µM	3'	TAATACGACTCACTATAGGGCTCACCTGAGGAGACGGTGACC

* For the Ion Torrent protocol, primers without the M13 sequence are used

Table 3
Primers included in the multiplex PCR reaction for NGS-based clonality assessment: Tube IGHD

Primer nomenclature	Final concentration	Primer direction	Primer Sequence (with M13 adapter forward/reverse) 5' to 3'*
IGH-D-A-1	0.2 µM	5'	GTAAAACGACGGCCAGTGATTCYGAACAGC CCCGAGTCA
IGH-D-B-1	0.2 µM	5'	GTAAAACGACGGCCAGTGATTTTGTGGGGG YTCGTGTC
IGH-D-C-1	0.2 µM	5'	GTAAAACGACGGCCAGTGTTTGRRGTGAGG TCTGTGTCA
IGH-D-D-1	0.2 µM	5'	GTAAAACGACGGCCAGTGTTTRGRRTGAGG TCTGTGTCACT
IGH-D-E-1	0.2 µM	5'	GTAAAACGACGGCCAGTCTTTTTGTGAAGG SCCCTCCTR
IGH-D-F-1	0.2 µM	5'	GTAAAACGACGGCCAGTGTTATTGTCAGGS GRTGTCAGAC
IGH-D-G-1	0.2 µM	5'	GTAAAACGACGGCCAGTGTTATTGTCAGGG GGTGYCAGRC
IGH-D-H-1	0.2 µM	5'	GTAAAACGACGGCCAGTGTTTCTGAAGSTG TCTGTRTCAC
IGH-J-A-1	0.2 µM	3'	TAATACGACTCACTATAGGGCTTACCTGAG GAGACGGTGACC
IGH-J-B-1	0.2 µM	3'	TAATACGACTCACTATAGGGCTCACCTGAG GAGACAGTGACC
IGH-J-C-1	0.2 µM	3'	TAATACGACTCACTATAGGGCTCACCTGAG GAGACGGTGACC

* For the Ion Torrent protocol, primers without the M13 sequence are used

embedded (FFPE) and fresh frozen tissue using the Chelex method (FFPE), TSE (fresh frozen), and column-based extraction procedure of QIAGEN; equivalent isolation systems are also possible (*see* **Note 2**).

To perform reliable clonality assessment it is important to determine whether a representative tissue section is used, whether obtained DNA is of sufficient quality (*see* **Note 4**) and using a standardized DNA input per PCR. Furthermore, robust performance of the multiplex PCR reaction should be assessed by including control samples such as a polyclonal control sample (e.g., tonsil or mononuclear peripheral blood cells) and negative control (water), while preparing the samples for IG-NGS clonality assessment (*see* **Note 5**).

3.2 DNA Isolation

For isolation of genomic DNA from FFPE tissue, different methods are available. Here two of such protocols are described, a commercially available DNA isolation kit (QIAGEN) and the Chelex method. Both protocols use a microcolumn purification of the extracted DNA; this is an important step in preparing DNA samples for clonality assays and is described in Subheading 3.2.3. Finally, a protocol for isolation of genomic DNA from fresh frozen tissue is described.

Table 4
Primers included in the multiplex PCR reaction for NGS-based clonality assessment: Tube IGK

Primer nomenclature	Final concentration	Primer direction	Primer Sequence (with M13 adapter forward/reverse) 5' to 3'*
IGK-V-A-1	0.2 μM	5'	GTAAAACGACGGCCAGTAAGTGGGGTCCC ATCAAGGTTCAG
IGK-V-B-1	0.2 μM	5'	GTAAAACGACGGCCAGTAGTCCCATCTCG GTTCAGTGGCAG
IGK-V-C-1	0.2 μM	5'	GTAAAACGACGGCCAGTGAAACAGGGGTC CCATCAAGGTTC
IGK-V-D-1	0.2 μM	5'	GTAAAACGACGGCCAGTTCCCAGACAGAT TCAGTGGCAGTG
IGK-V-E-1	0.2 μM	5'	GTAAAACGACGGCCAGTCTGGAGTGCCAG ATAGGTTCAGTG
IGK-V-F-1	0.2 μM	5'	GTAAAACGACGGCCAGTCCCTGGAGTCCC AGACAGGTTCAG
IGK-V-G-1	0.2 μM	5'	GTAAAACGACGGCCAGTGCATCCCAGCCA GGTTCAGTG
IGK-V-H-1	0.2 μM	5'	GTAAAACGACGGCCAGTGTCCCTGACCGA TTCAGTGGCA
IGK-V-I-1	0.2 μM	5'	GTAAAACGACGGCCAGTAATCCCACCTCG ATTCAGTGGC
IGK-V-J-1	0.2 μM	5'	GTAAAACGACGGCCAGTCTCAGGGGTCCC CTCGAGGTT
IGK-V-K-1	0.2 μM	5'	GTAAAACGACGGCCAGTAGACACTGGGGT CCCAGCCA
IGK-INTR-A-1	0.2 μM	5'	TAATACGACTCACTATAGGGGAGTGGCTTT GGTGGCCATGC
IGK-DE-A-1	0.2 μM	3'	TAATACGACTCACTATAGGGGCAGCTGCA GACTCATGAGGAG
IGK-J-A-1	0.2 μM	3'	TAATACGACTCACTATAGGGACGTTTGATC TCCACCTTGGTCCC
IGK-J-B-1	0.2 μM	3'	TAATACGACTCACTATAGGGACGTTTGATA TCCACTTTGGTCCC
IGK-J-C-1	0.2 μM	3'	TAATACGACTCACTATAGGGACGTTTAATC TCCAGTCGTGTCCC

* For the Ion Torrent protocol, primers without the M13 sequence are used

3.2.1 DNA Extraction from Formalin-Fixed Paraffin-Embedded Tissue with Genomic DNA Isolation Kit

All steps are performed at room temperature, unless specified otherwise.

1. Place two to five 10 μm sections of FFPE tissue (of approximately 1 cm^2 in size) into a 1.5 ml Eppendorf tube (*see* **Note 4**).
2. Add 1000 μl xylene and vortex thoroughly for 10 s (work in a protective cabinet) (*see* **Note 6**).
3. Centrifuge for 5 min, 20,000 × *g*.
4. Remove the supernatant carefully and dispense it in specific waste containers.
5. Add 1000 μl 99% ethanol.
6. Centrifuge for 5 min, 20,000 × *g*.
7. Remove carefully the ethanol.

Table 5
M13 adapter barcoded primers for the Second PCR of the Illumina protocol

	Primer nomenclature	Final concentration	Primer direction	Barcoded primer sequence with M13 adapter (forward/reverse) 5′ to 3′
Forward	Ill-D501-F	0.2 μ M	5′	AATGATACGGCGACCACCGAGATCTACACTATAGCCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGTAAAACGACGGCCAGT
	Ill-D502-F	0.2 μ M	5′	AATGATACGGCGACCACCGAGATCTACACATAGAGGCACACTCTTTCCCTACACGACGCTCTTCCGATCTGTAAAACGACGGCCAGT
	Ill-D503-F	0.2 μ M	5′	AATGATACGGCGACCACCGAGATCTACACCCTATCCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGTAAAACGACGGCCAGT
	Ill-D504-F	0.2 μ M	5′	AATGATACGGCGACCACCGAGATCTACACGGCTCTGAACACTCTTTCCCTACACGACGCTCTTCCGATCTGTAAAACGACGGCCAGT
	Ill-D505-F	0.2 μ M	5′	AATGATACGGCGACCACCGAGATCTACACAGGCGAAGACACTCTTTCCCTACACGACGCTCTTCCGATCTGTAAAACGACGGCCAGT
	Ill-D506-F	0.2 μ M	5′	AATGATACGGCGACCACCGAGATCTACACTAATCTTAACACTCTTTCCCTACACGACGCTCTTCCGATCTGTAAAACGACGGCCAGT
	Ill-D507-F	0.2 μ M	5′	AATGATACGGCGACCACCGAGATCTACACCAGGACGTACACTCTTTCCCTACACGACGCTCTTCCGATCTGTAAAACGACGGCCAGT
	Ill-D508-F	0.2 μ M	5′	AATGATACGGCGACCACCGAGATCTACACGTACTGACACACTCTTTCCCTACACGACGCTCTTCCGATCTGTAAAACGACGGCCAGT
Reverse	Ill-D701-R	0.2 μ M	3′	CAAGCAGAAGACGGCATACGAGATCGAGTAATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTAATACGACTCACTATAGGG
	Ill-D702-R	0.2 μ M	3′	CAAGCAGAAGACGGCATACGAGATTCTCCGGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTAATACGACTCACTATAGGG
	Ill-D703-R	0.2 μ M	3′	CAAGCAGAAGACGGCATACGAGATAATGAGCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTAATACGACTCACTATAGGG
	Ill-D704-R	0.2 μ M	3′	CAAGCAGAAGACGGCATACGAGATGGAATCTCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTAATACGACTCACTATAGGG
	Ill-D705-R	0.2 μ M	3′	CAAGCAGAAGACGGCATACGAGATTTCTGAATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTAATACGACTCACTATAGGG
	Ill-D706-R	0.2 μ M	3′	CAAGCAGAAGACGGCATACGAGATACGAATTCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTAATACGACTCACTATAGGG
	Ill-D707-R	0.2 μ M	3′	CAAGCAGAAGACGGCATACGAGATAGCTTCAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTAATACGACTCACTATAGGG
	Ill-D708-R	0.2 μ M	3′	CAAGCAGAAGACGGCATACGAGATGCGCATTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTAATACGACTCACTATAGGG
	Ill-D709-R	0.2 μ M	3′	CAAGCAGAAGACGGCATACGAGATCATAGCCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTAATACGACTCACTATAGGG
	Ill-D7010-R	0.2 μ M	3′	CAAGCAGAAGACGGCATACGAGATTTCGCGGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTAATACGACTCACTATAGGG
	Ill-D7011-R	0.2 μ M	3′	CAAGCAGAAGACGGCATACGAGATGCGCGAGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTAATACGACTCACTATAGGG
	Ill-D7012-R	0.2 μ M	3′	CAAGCAGAAGACGGCATACGAGATCTATCGCTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTAATACGACTCACTATAGGG

8. Add again 1000 μl 99% ethanol and mix carefully by inverting the tube.
9. Centrifuge for 5 min, 20,000 × *g*.
10. Remove all ethanol and air-dry the remaining tissue by leaving the tube open, and incubate for 10–15 min at 37 °C in a thermo block.
11. Resuspend the pellet in 180 μl Buffer ATL.
12. Add 20 μl proteinase K, and mix by vortexing.
13. Incubate overnight at 56 °C (*see* **Note 7**).
14. Incubate at 90 °C for 1 h (*see* **Note 8**).
15. Add 200 μl Buffer AL to the sample, and mix thoroughly by vortexing (*see* **Note 9**).
16. Add 200 μl ethanol (96–100%), and mix again thoroughly by vortexing (*see* **Note 9**).
17. Continue with the extraction and DNA purification as described in Subheading 3.2.3, before starting the sample preparations for clonality assessment.

3.2.2 DNA Extraction from Formalin-Fixed Paraffin-Embedded (FFPE) Tissue Starting with the Chelex Method

This Chelex-based DNA extraction protocol is developed as common workflow that is suitable for the majority of the molecular tests used in diagnostics. However, for clonality assessment, it is important to purify the DNA obtained with this protocol before use in the clonality assay in order to obtain good quality results.

All steps are performed at room temperature, unless specified otherwise.

1. Deparaffinize two to five 10 μm tissue sections as described in Subheading 3.2.1 until **step 10**.
2. Add 200 μl of 5% Chelex-100 homogeneously mixed in TET lysis buffer.
3. Incubate for 5 min at 95 °C in a thermo shaker at 350 rpm.
4. Cool down for 5 min at room temperature.
5. Add 20 μl of proteinase K and incubate o/n at 56 °C in a thermo shaker at 350 rpm (*see* **Note 7**).
6. Incubate for 10 min at 95 °C in a thermo shaker at 350 rpm (*see* **Note 8**).
7. Centrifuge for 10 min, 20,000 × *g* at room temperature.
8. Transfer the supernatant to a clean 1.5 ml Eppendorf tube.
9. Centrifuge for 10 min, 20,000 × *g* at room temperature.
10. Transfer the supernatant to a clean 1.5 ml Eppendorf tube.
11. Add 180 μl ATL buffer (QIAamp DNA Micro Kit), vortex for 10 s, and incubate at room temperature for 30 min.
12. Incubate at 80 °C for 10 min, and then allow to cool down to room temperature.
13. Briefly centrifuge the 1.5 ml tube to remove drops from the inside of the lid.
14. Add 200 μl buffer AL and vortex briefly.
15. Incubate at 70 °C for 10 min, and then allow to cool down to room temperature.
16. Add 250 μl of ethanol 96%, vortex briefly.
17. Continue the procedure with the DNA purification as described in Subheading 3.2.3, before starting the sample preparations for clonality assessment.

3.2.3 DNA Purification with QIAamp DNA Microcolumn

All steps are performed at room temperature.

1. Carefully transfer the entire lysate to the QIAamp MinElute column (in a 2 ml collection tube) and centrifuge at 6000 × *g* for 1 min (*see* **Note 10**).
2. Place the QIAamp MinElute column in a clean 2 ml collection tube, and discard the collection tube containing the flow-through.

3. Add 500 µl Buffer AW1 on the column and centrifuge at 6000 × *g* for 1 min (*see* **Note 11**).
4. Discard the flow-through and add 500 µl Buffer AW2 to the column (*see* **Note 11**).
5. Centrifuge at 6000 × *g* for 1 min and discard the flow-through.
6. Centrifuge at full speed (20,000 × *g*) for 3 min to dry the membrane completely (*see* **Note 12**).
7. Place the QIAamp MinElute column in a clean 1.5 ml Eppendorf tube, and discard the collection tube containing the flow-through.
8. Apply 20–100 µl Buffer ATE (QIAamp DNA FFPE Kit) or 20–100 µl Buffer AE (QIAamp DNA Micro Kit) to the center of the column membrane (*see* **Note 13**).
9. Incubate at room temperature for 5 min.
10. Centrifuge at full speed (20,000 × *g*) for 1 min.
11. Discard the column and keep the 1.5 ml tube containing the DNA solution.
12. Determine the DNA concentration and dilute if necessary to a working solution of 20–40 ng/µl with the used elution buffer or MQ (*see* **Note 14**).

3.2.4 DNA Extraction from Fresh Frozen Tissue: TSE Method

All steps are performed at room temperature, unless specified otherwise.

1. Place five to ten 10 µm tissue sections in a 1.5 ml tube with 400 µl TSE (*see* **Note 15**).
2. Add 21 µl SDS 20% and 37.5 µl proteinase K (20 mg/ml) and mix manually.
3. Incubate o/n at 56 °C on a thermo shaker at 350 rpm (*see* **Note 7**).
4. Keep the lysate in the 1.5 ml Eppendorf tube.
5. Add 168 µl NaCl 5 M and shake the tube for 15 s.
6. Centrifuge for 15 min at 18,000 × *g*.
7. Transfer the supernatant to a clean 2 ml Eppendorf tube.
8. Centrifuge for 15 min at 18,000 × *g*.
9. Transfer the supernatant to a clean 2 ml Eppendorf tube.
10. Add 1.2 ml ethanol 100% and turn around the tube a few times so the DNA will precipitate (*see* **Note 16**).
11. Centrifuge for 10 min at 18,000 × *g*.
12. Remove the supernatant carefully using a pipette.
13. Wash the pellet with 1 ml ethanol 70%.
14. Centrifuge for 10 min at 18,000 × *g*.

15. Remove the supernatant carefully using a pipette.
16. Air-dry the pellet for at least 1 min until the pellet is completely dry.
17. Add Low TE-buffer to the pellet (20–50 μl when there is not a clearly visible pellet present; use larger volumes when the DNA pellet is bigger).
18. Incubate for at least 2 h at room temperature on a roller bank or incubate for longer time (o/n or longer) at 4 °C to allow the DNA to dissolve completely.
19. Determine the DNA concentration and dilute if necessary to a working solution of 20–40 ng/μl with Low TE-buffer or MQ (*see* **Note 14**).

3.3 Ion Torrent Protocol for IG-NGS Clonality Assessment

3.3.1 Multiplex PCR for Amplification of IGH-FR3, IGHD, and IGK

1. Prepare three different 0.2 ml PCR tubes per sample: IGH-FR3, IGHD, and IGK (*see* Table 1, panel "Ion Torrent protocol for clonality detection").
2. Add 40 ng DNA (Qubit measured; *see* **Note 14**) and the other components of the reaction, according to Table 1 (panel "Ion Torrent protocol for clonality detection"), and adjust the total PCR volume to 25 μl with MQ (*see* **Note 17**).
3. Use a pipette to mix the prepared PCR reaction with sample thoroughly while avoiding air bubbles in the reaction mix and perform a quick spin to collect all liquid to the bottom of the PCR tube.
4. Run the PCR in a thermocycler with heated lid, according to the program in Table 1 (panel "Ion Torrent protocol for clonality detection").
5. After completing the PCRs, combine the tree targets IGH-FR3, IGHD, and IGK per sample (~75 μl total volume).

3.3.2 Cleanup of IGH-FR3, IGHD, and IGK Amplicons

1. Allow the AMPure XP magnetic beads to warm to room temperature for at least 30 min before use. Ensure that the beads are homogeneous prior to use by mixing the tube by hand for 20 s (*see* **Note 18**).
2. Pipette the pooled samples in a DNA LoBind plate, and add 1.8 times (135 μl) volume Agencourt AMPure XP magnetic beads per sample (*see* **Note 19**).
3. Use a pipette to mix the solution thoroughly (avoid air bubbles), until the beads and sample are homogeneously mixed, and incubate for 5 min at room temperature.
4. Place the samples for 2–5 min in a magnetic stand until the solution is clear (*see* **Note 20**).
5. Carefully remove the supernatant using a 200 μl pipette (*see* **Note 21**).
6. Add 150 μl freshly made 70% ethanol per sample (*see* **Note 22**).

7. Move the plate in the magnetic stand approximately 4 times from left to right, and make sure the bead pellet migrates and is washed clean.
8. Carefully remove the supernatant using a 200 μl pipette (*see* **Note 21**).
9. Repeat **steps 6–8** once.
10. Carefully remove any remaining supernatant using a 10 μl pipette, and air-dry the beads for 5 min to allow complete evaporation of residual ethanol (*see* **Note 23**).
11. Resuspend the samples in 25 μl Low TE-buffer.
12. Use a pipette to mix the solution thoroughly (avoid air bubbles), to generate a homogeneously mixed solution, and incubate for 2 min at room temperature.
13. Place the samples for 2 min in the magnetic stand until the solution is clear (*see* **Note 20**).
14. Collect the purified DNA by pipetting the solution (~25 μl) into a new PCR strip (*see* **Note 24**).

3.3.3 End Repair of Amplicons

1. Measure the DNA concentration of every individual sample using Fluorometric Quantitation (using 2 μl of the sample for the Qubit high sensitivity assay).
2. Transfer max. 40 ng DNA to a 0.2 ml PCR tube for the end repair step. In case the total yield is less than 40 ng, use as much as possible. Adjust the volume to 39.5 μl with Low TE-buffer.
3. Add the reagents for the end repair reaction from the Ion Plus Fragment Library kit to the amplicons according to Table 6 (*see* **Note 25**).
4. Use a pipette to mix the suspension thoroughly, perform a quick spin to collect all liquid to the bottom, and incubate 30 min at room temperature (*see* **Note 26**).
5. Perform a cleanup as described in Subheading 3.3.2 with 1.8 times volume Agencourt AMPure XP magnetic beads (90 μl) and elution volume of 25 μl Low TE-buffer.

Table 6
Protocol of the end repair reaction for Ion Torrent

Component	Volume (μl)
Pooled amplicons (40 ng) adjusted to 39.5 μl with low TE-buffer	39.5
5× End Repair Buffer	10
End Repair Enzyme	0.5
Total volume	*50*

Table 7
Protocol of the adapter ligation/nick repair reaction for Ion Torrent

Component	Volume (μl)
Pooled amplicons (40 ng)	24.5
10× Ligase Buffer	5
Ion P1 adapter from the Barcode Kit (not the one within the Library Kit)	1
dNTP mix	1
Nuclease-free water	12.5
DNA Ligase	1
Nick Repair Polymerase	4
Ion Xpress barcode X	1
Total volume	*50*

Table 8
Adapter ligation program for Ion Torrent

1 cycle	25 °C	15 min
1 cycle	72 °C	5 min
1 cycle	4 °C	10 min
	4 °C	∞

3.3.4 Adapter Ligation

1. To ligate adapters to the amplicon and to perform nick repair, for each sample, add the amplicons and reagents from the Ion Plus Fragment Library kit and Ion Xpress Barcode Adapter kit to a 0.2 ml PCR tube according to Table 7 (*see* **Note 27**). Make sure that for each sample a different barcode is used (*see* **Note 28**).
2. Run the adapter ligation program according to Table 8.
3. Perform a cleanup as in Subheading 3.3.2 with 1.8 times volume Agencourt AMPure XP magnetic beads (90 μl) and elution volume of 13 μl Low TE-buffer (*see* **Note 29**).

3.3.5 Library Amplification

1. To amplify the libraries, for each sample, add the purified adapter-ligated amplicons (*see* **Note 29**) and reagents from the Ion Plus Fragment Library kit to a 0.2 ml PCR tube according to Table 9.
2. Use a pipette to mix the suspension thoroughly and perform a quick spin to collect all liquid from the sides of the tube.

Table 9
Composition of the library amplification reaction for Ion Torrent

Component	Volume (μl)
Pooled amplicons (unamplified library)	12.5
Platinum PCR SuperMix High Fidelity	50
Library amplification primer mix	2.5
Total volume	*65*

Table 10
Library amplification PCR program for Ion Torrent

1 cycle	Initial denaturation	95 °C	5 min
8 cycles	Denaturation Annealing Extension	95 °C 58 °C 70 °C	15 s 15 s 1 min
1 cycle	Hold	4 °C	∞

3. Run the PCR in a thermocycler with heated lid, according to the program in Table 10.
4. Perform a cleanup as in Subheading 3.3.2 with 1.4 times volume Agencourt AMPure XP magnetic beads per sample (90 μl) and elution volume of 25 μl Low TE-buffer.

3.3.6 Ion Torrent Sequencing Run

1. Measure the DNA concentration of all samples using Fluorometric Quantitation.
2. Pool all samples at an equivalent DNA amount and measure the total pool DNA concentration with Fluorometric quantitation.
3. Dilute each pooled sample to a final DNA concentration of 12 ng/ml with Low TE-buffer (Qubit measured). Alternatively, library quantification can be performed with the Ion Library TaqMan Quantification Kit (220–250 pM final concentration).
4. Run Ion Torrent on a 318-chip (Ion PGM) or 5S Chip (Ion GeneStudio S5) for a total of 24–32 samples, according to your local Sequence Facility (*see* **Note 30**).

3.4 Illumina Protocol for IG-NGS Clonality Assessment

This two-step PCR protocol is based on a previously published protocol for marker identification for MRD [21], with some modifications for the first PCR reaction (Table 1). Furthermore, the protocol described below is optimized for sequencing on a MiniSeq (Illumina), but other equipment may be used according to the instructions of the local Sequence Facility.

3.4.1 Multiplex PCR for Amplification of IGH-FR3, IGHD, and IGK

1. Prepare three different 0.2 ml PCR tubes per sample: IGH-FR3, IGHD, and IGK (*see* Table 1, panel "Two-step Illumina protocol for clonality detection").
2. Add 40 ng DNA (Qubit measured; *see* **Note 14**) and the other components of the reaction, according to Table 1 (panel "Two-step Illumina protocol for clonality detection"). Adjust the total PCR volume to 25 μl with MQ.
3. Use a pipette to mix the prepared PCR reaction with sample thoroughly while avoiding air bubbles in the reaction mix and perform a quick spin to collect all liquid to the bottom of the PCR tube.
4. Run the PCR in a thermocycler with heated lid, according to the program in Table 1 (panel "Two-step Illumina protocol for clonality detection").
5. After completion of the PCR protocol, combine tube IGH-FR3, IGHD, and IGK per sample (~75 μl total volume).

3.4.2 Cleanup of IGH-FR3, IGHD, and IGK Amplicons

1. Allow the AMPure XP magnetic beads to warm to room temperature for at least 30 min before use. Ensure that the beads are homogeneous prior to use by mixing the tube by hand for 20 s (*see* **Note 18**).
2. Pipette the pooled samples in a DNA LoBind plate and add 1.8 times (135 μl) volume Agencourt AMPure XP magnetic beads per sample (*see* **Note 19**).
3. Use a pipette to mix the solution thoroughly (avoid air bubbles), until the beads and sample are homogeneously mixed and incubate for 5 min at room temperature.
4. Place the samples for 2–5 min in a magnetic stand until the solution is clear (*see* **Note 20**).
5. Carefully remove the supernatant using a 200 μl pipette (*see* **Note 21**).
6. Add 150 μl freshly made 70% ethanol per sample (*see* **Note 22**).
7. Move the plate in the magnetic stand approximately 4 times from left to right, and make sure the bead pellet migrates and is washed clean.
8. Carefully remove the supernatant using a 200 μl pipette (*see* **Note 21**).
9. Repeat **steps 6–8** once.
10. Carefully remove any remaining supernatant using a 10 μl pipette, and air-dry the beads for 5 min to allow complete evaporation of residual ethanol (*see* **Note 23**).
11. Resuspend the samples in 25 μl Low TE-buffer.

Table 11
Composition of the barcode amplification reaction for Illumina. Primer sequences and final concentrations are provided in Table 5

Component	Concentration
Purified amplicons (1:50 diluted)	1 µl
FastStart buffer 10× + 18 mM $MgCl_2$	1×
FastStart dNTP's 10 mM	0.2 mM
Fast start High Fidelity polymerase 5 U/µl	0.8 U
M13 adapter barcoded forward primer X	0.2 µM
M13 adapter barcoded reverse primer X	0.2 µM
MQ	Adjust to 25 µl
Total volume	*25*

12. Use a pipette to mix the solution thoroughly (avoid air bubbles), to generate a homogeneously mixed solution, and incubate for 2 min at room temperature.
13. Place the samples for 2 min in the magnetic stand until the solution is clear (*see* **Note 20**).
14. Collect the purified DNA by pipetting the solution (~25 µl) into a new PCR strip (*see* **Note 24**).

3.4.3 Second PCR to Generate Barcoded Amplicons

1. Dilute the purified amplicons 1:50 in Low TE-buffer.
2. Use a Roche FastStart™ High Fidelity PCR kit (Sigma-Aldrich) to prepare a PCR mix for each sample according to Table 11. Make sure that a unique barcode combination is used for each sample (*see* **Note 31**).
3. Perform the PCR reaction in a total reaction volume of 25 µl. Mix thoroughly and spin down. Make sure that all reagents are at the bottom of the tube and avoid air bubbles.
4. Run the PCR in a thermocycler with heated lid, according to the program in Table 12.
5. After completion of the second PCR, continue with a double purification as described in Subheading 3.4.4.

3.4.4 Cleanup of Barcode-Labeled Amplicons

The cleanup procedure described below is based on a double purification procedure, where the first step (0.6× volume beads) is a negative selection and the second step (0.25× volume beads) a positive selection. This double purification can be replaced by a single purification protocol, as described in Subheading 3.4.2. In that case, use 1.0× volume beads for the products from tubes IGHV-FR3 and IGK and 0.9× volume beads for the products from tube IGHD.

Table 12
PCR program for the barcode amplification reaction for Illumina. A header of this table is missing, please include a Header, which should be: Cycle (1st column), PCR step (column 2) Temperature (column 3) Time (column 4)

1 cycle This row is a normal row, however the header of this Table is missing, please include a header.	Initial denaturation	95 °C	2 min
20 cycles this row should not be in bold	**Denaturation** **Annealing** **Extension**	**94 °C** **63 °C** **72 °C**	**30 s** **30 s** **30 s**
1 cycle	Final extension	72 °C	5 min
1 cycle	Hold	12 °C	∞

1. Allow the AMPure XP magnetic beads to warm to room temperature for at least 30 min before use. Ensure that the beads are homogeneous prior to use by mixing the tube by hand for 20 s (*see* **Note 18**).
2. Transfer the PCR reaction from each sample (25 μl) in a DNA LoBind plate (one sample per well), and add 0.6× volume (15 μl) Agencourt AMPure XP magnetic beads per sample (*see* **Note 19**).
3. Use a pipette to mix the solution thoroughly (avoid air bubbles), until the beads and sample are homogeneously mixed, and incubate for 5 min at room temperature.
4. Put the plate for 2–5 min in a magnetic stand, or until the solution is clear (*see* **Note 20**).
5. Carefully transfer the supernatant (40 μl) into a new well (*see* **Note 21**).
6. Remove the plate from the magnetic stand, and add 0.25× volume (10 μl) magnetic beads per sample (*see* **Note 18**).
7. Use a pipette to mix the solution thoroughly (avoid air bubbles), until the beads and sample are homogeneously mixed, and incubate for 5 min at room temperature.
8. Put the plate on the magnetic rack and incubate for 5 min.
9. Remove and discard the supernatant.
10. Add 200 μl freshly made 70% ethanol to each sample to wash the beads (*see* **Note 22**).
11. Move plate in the magnetic stand approximately 4 times from left to right, and make sure the bead pellet migrates and is washed clean.
12. Carefully remove and discard the supernatant (*see* **Note 21**).
13. Repeat **steps 10–12** once.

14. Carefully remove any remaining supernatant using a 10 μl pipette, and air-dry the beads for 5 min to allow complete evaporation of residual ethanol (*see* **Note 23**).
15. Remove the plate from the magnet and add 10 μl Low TE-buffer per sample.
16. Mix thoroughly (avoid air bubbles) and incubate for 2 min at room temperature.
17. Put the plate on the magnetic stand for 2 min or until the solution is clear (*see* **Note 20**).
18. Collect the purified DNA by pipetting the solution (~10 μl) into a new PCR strip.
19. Continue with setting up the Illumina sequencing run.

3.4.5 Illumina Sequencing Run

1. Measure the DNA concentration of all samples using Fluorometric Quantitation.
2. Pool all samples equimolar and measure the concentration of the total sample pool using fluorometric quantitation.
3. Perform sequencing on the Illumina instrument employing corresponding reagent Kit, according to the manufacturer's instructions and your local Sequence Facility (*see* **Note 32**).

3.5 Post-Analytical Data Analysis

The obtained sequencing results can be analyzed with the bioinformatics tool ARResT/Interrogate (http://arrest.tools/interrogate) [18] (*see* **Note 33**). FastQ data files are uploaded and processed for analysis. Subsequently, the results can be visualized for further analysis using the "reporting" or "questions" sections. Here, the different rearrangements are referred to as clonotypes that include information about the 5′ gene, junction, and 3′ gene, as shown in Fig. 4.

For basic clonality analysis, it is recommended to use the reporting function. Here, a complete overview of the clonality results is generated automatically for all IG loci, i.e., IGHV-IGHD-IGHJ (FR3), IGHD-IGHJ, and IGKV-IGKJ and a combined result for IGKV/Intron RSS-KDE, comparable with conventional clonality testing using BIOMED-2/EuroClonality assays. Select a sample, make sure the correct target is selected (i.e., choose "IG" under cell type for B-cell clonality assessment), the filter is set on 0–100% to include all detected clonotypes, and click on "report". The following information will be shown in the report that is generated:

1. First, an overview of some quality parameters is shown. The most important is the QC status: "Pass" when the data meets all quality criteria or "Fail" when the data does not meet all quality criteria. Under "QC report" it can be found why the QC failed and which target failed; please interpret these targets with caution.

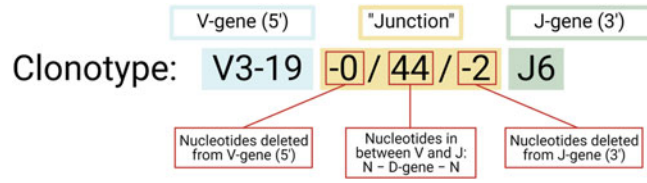


Fig. 4 Clonotype annotation. A rearrangement (complete IGH and IGK rearrangements) is referred to as a clonotype notated as an immunoglobulin nucleotide sequence with a 5′ gene (V-gene), the junction, and the 3′ gene (J-gene). The junction consists of three parts: the first and last numbers are the amount of nucleotides that is removed from the 5′- or 3′-genes, respectively. The middle number is the amount of nucleotides that is present between the 5′- and 3′-genes and includes the so-called N-nucleotides that are added by the enzyme terminal deoxynucleotidyl transferase (TdT) during the V(D)J recombination process, as well as the D-gene in case of a complete VDJ rearrangement. Incomplete, nonfunctional IGH and IGK rearrangements (IGHD-IGHJ, IGKV-KDE, Intron RSS-KDE) are "artificially" described as clonotypes in ARResT/Interrogate in a similar way as shown here for complete IGH rearrangements, using the corresponding 5′- and 3′-genes and their junctions

2. A bar chart for each IG locus is created with the abundancy of detected clonotypes on the y-axis and the junction amino acid (aa) length on the x-axis. Note IGHD-IGHJ and all rearrangements involving KDE are not expressed and are "artificially" described as clonotypes in ARResT/Interrogate. In this way, each bar represents all clonotypes with the same junction aa length, whereas each clonotype with a unique nucleotide sequence is depicted with a specific color. Please note that only the top 50 most abundant clonotypes are colored; all other clonotypes are merged and represented by gray bars.
3. A table with more detailed information about the top 15 clonotypes is generated and shown next to the bar chart. This table includes the genes and segmentation of the clonotypes, the amino acid sequence of the junction, functionality of the rearrangement (pro, productive; pop, potentially productive, for incomplete rearrangements; unp, unproductive; unk, unknown, for special rearrangements), the length of the junction in amino acids, and the number and percentage of reads with the specific clonotype and length.
4. In a small-sized table, some general information ("stats for junction class set") is shown, of which especially the total number of reads (for normalization) is important.

   With the "PDF" button that is present in the reporting function, the total report can be exported as a PDF file.

More advanced analyses can be performed using the "questions" function. In addition to the standard parameters (i.e., junction aa length and clonotype), also other ones can be chosen, like

amplicon length or the 5′- or 3′-genes/primers to analyze the data in more detail. Furthermore, in contrast to the reporting section, it is possible to select 2 or more samples at the same time in the questions section, to directly compare the nucleotide sequences in either a bar chart or table. This is especially of added value when a clonal comparison has to be made for a patient with multiple tumors, for example. Using the questions function, the data can be visualized as follows:

1. The "table" subsection can be used to create a table of all detected clonotypes with information about the chosen parameters. This complete table can be exported using the small download button (download full table in .xlsx format). Again, it is possible to select one or more samples for the analysis.
2. Using the "bars" subsection, bar charts can be created with the parameters of interest. This can be done for a single sample or for the analysis of multiple samples simultaneously in case of a clonal comparison for example. The generated bar charts can be downloaded as image by using the button at the upper right corner next to the legend.

Within ARResT/Interrogate, all bar charts (created within both the reporting and questions function) are "interactive" meaning that by clicking on one or more colored parts of a bar, the corresponding clonotypes are selected. A so-called minitable pops up at the top of the page, with the general information about the clonotype(s), but also the most popular full nucleotide sequence of the corresponding clonotype. This information can be downloaded using the download button. Further analysis of the most popular nucleotide sequence, but also all other sequences belonging to the same clonotype, can be done within the "forensics" function. By using the green button "run tests," this forensics section will open automatically or go manually to this section. Here, the following analyses can be performed:

1. When forensics is opened, automatically the subsection "tests" is shown. By selecting "interrogate" and choosing "run the test on minitable," more detailed information will appear for the most popular nucleotide sequence, like the segmentation of the clonotype and alignment to the germline sequences of the corresponding genes. Also additional tools (i.e., IMGT/V-QUEST and Vidjil) are available for further analysis of the selected clonotypes.
2. By going to the subsection "sequences," a table with all nucleotide sequences, including the number of reads, is shown that corresponds to the selected clonotype which can be retrieved and downloaded. Please note that when working via the "reporting" section, first "reporting panel features" needs to be selected before sequences can be retrieved.

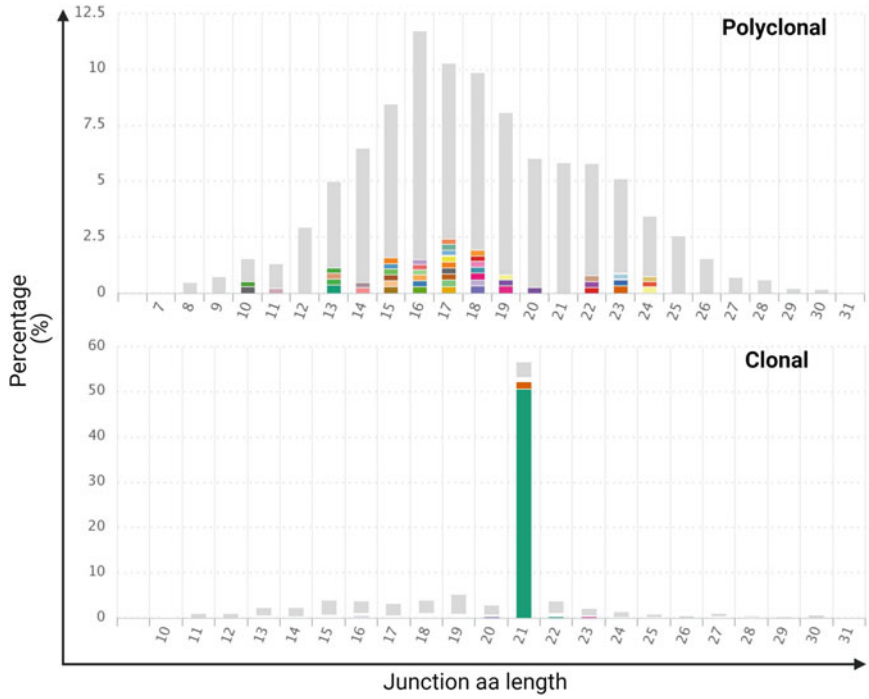


Fig. 5 Output data generated by ARResT/Interrogate with IG-NGS clonality assay. IG-NGS clonality profiles of a polyclonal (upper panel) and monoclonal (lower panel) sample are shown in bar charts generated by ARResT/Interrogate (target IGHV-IGHD-IGHJ FR3). On the y-axis, the abundancy of detected clonotypes is shown in percentage and on the x-axis, the junction length is shown in amino acids (aa). Each bar represents clonotypes with the same junction aa length, and each color indicates a unique clonotype based on their nucleotide sequence. Only the 50 most abundant clonotypes are colored, and all other, less frequent gene rearrangements are merged and represented by the gray bars

After visualization of the results, the obtained clonality patterns of each sample, which is run under standardized conditions (input of DNA and number of samples per run), can be interpreted. It is strongly advised to include a polyclonal control sample in the run. A standardized input of an FFPE-derived polyclonal control sample under standardized run conditions should demonstrate a Gaussian curve with differently sized junctions of the gene rearrangement and a high variety of clonotypes represented by the presence of gray bars as shown in the top panel of Fig. 5, as well as the detection of the V/D/J gene families. Skewing of the curve to either short or long amplicon lengths could imply that the library preparation was not optimal and may interfere with the analysis of samples prepared within the same run. The same holds true for too few reads and/or clonotypes. Depending on the tumor load of a clonal sample, as well as the input of DNA (per PCR-library), a dominant clonotype will be present, as shown in the lower panel of Fig. 5.

For correct interpretation of the clonality assay per sample, several steps should be followed:

1. Make a technical interpretation of the obtained results per locus and rearrangement type (i.e., IGHV-IGHD-IGHJ, IGHD-IGHJ, IGKV-IGKJ, and IGKV/Intron RSS-KDE). This includes the number of obtained reads, but also the number of different clonotypes detected, the run conditions, the DNA input, and the tumor load. Based on this protocol, the data can be analyzed reliably when at least 1000 reads are available for a PCR-target as evaluated in the EuroClonality-NGS biological validation study [22]. The technical interpretation per PCR target uses the terminology: "clonal," "polyclonal," or "no specific product," with the possibility to add a more detailed information, similar as described in the guidelines for conventional clonality testing [17].
2. Evaluation of the technical interpretation of the individual rearrangement types (PCRs) into a molecular clonality conclusion, according to the EuroClonality guidelines for conventional clonality testing [17].

Guidelines for the technical interpretation of the obtained result per locus and rearrangement type, as well as for the molecular clonality conclusion, are under development. Furthermore, it should be stressed that the clonality results should be integrated with the clinical, morphological, and immunophenotypic data to make a final diagnosis.

4 Notes

1. The design of the two-step Illumina protocol for clonality detection is based on two previously developed assays: the Ion Torrent protocol for clonality detection [6] and the two-step Illumina protocol for marker identification [21]. The first step, i.e., target amplification, is based on the Ion Torrent protocol, since the analyzed targets are identical. Therefore, the PCR conditions and primer sequences for target amplification are the same for both protocols, except the M13 sequence on the forward and reverse primers. Due to the occurrence of SHM in most B-cell lymphomas within the IGHV genes, the annealing temperature for IGHV-FR3 PCR is lower (60 °C) in the two-step Illumina clonality detection protocol compared to the MRD marker identification protocol (63 °C). To create uniformity in the PCR programs for all targets, the PCR programs are similar for all targets.
2. The chosen DNA extraction method should significantly reduce protein and RNA contamination, and thus a procedure

that includes column-based purification is strongly recommended. Extraction methods that isolate both DNA and RNA in parallel are not suitable. RNA present in the DNA solution negatively influences the PCR reaction resulting in an abnormal, disturbed polyclonal pattern. The DNA should be quantified to enable standardized DNA input in the PCR.

3. Different primer pools need to be prepared for each of the three multiplex PCR reactions (IGHV-FR3, IGHD, and IGK). Each primer pool should preferably contain all forward and reverse primers for that specific target. It is recommended to prepare 25× concentrated primer pools, where 1 μl primer pool can be used for each PCR reaction. Within the pool, 5 μM or 10 μM concentrations of each primer should be added, according to Tables 2, 3, and 4, which yields 0.2 μM or 0.4 μM final concentration within the reaction, respectively. For instance, if 600 μl primer pool is prepared, add the different primer volumes from 300 μM stock concentrations, i.e., 10 μl for 5 μM primer pool concentration and 20 μl for 10 μM primer pool concentration, and adjust the final volume to 600 μl with Low TE-buffer. Please note that for the Ion Torrent protocol, the primer sequences without the M13 adapter (blue or green sequences) should be used.
4. To assess whether the tissue slices used for DNA extraction is representative for the disease, 4 μm haematoxylin-eosin sections just before and after these slices should be evaluated by an experienced hematopathologist. The tissue fixation protocol may affect the degradation of the extracted DNA and thus the DNA quality in terms of amplifiability. It is important that neutral-buffered formalin is used. Prolonged fixation should be avoided as this induces too much cross-linking between DNA and other biomolecules resulting in inferior DNA quality for molecular analysis. In case the sample surface of the paraffin block has been exposed to air, it is advised to discard the first 2–3 sections before cutting sections for DNA isolation to avoid contamination. It is strongly advised to assess the quality of the purified DNA from FFPE samples by a quality control procedure using a size ladder PCR [13] and gel system, TapeStation (Agilent) or Bioanalyzer (Agilent).
5. A negative control sample should be included to monitor possible contaminations throughout the entire procedure. To this end, H_2O can be used as non-template negative control. A polyclonal positive control sample is essential as it allows to evaluate whether the multiplex PCR reaction was successful and to identify the V and J genes in the polyclonal sample. To obtain a complete, polyclonal IG pattern, tonsil or reactive lymph node DNA is preferred because of the higher B-cell numbers compared to peripheral blood samples.

6. For deparaffinization of FFPE tissue sections, Deparaffinization Solution (QIAGEN) can be used instead of xylene. In that case, replace **steps 2** until **14** for the manufacturer's protocol supplied with the deparaffinization solution, and continue the protocol with **step 15**.
7. It is important that the tissue sample is completely lysed for optimal DNA yields. If the tissue is not yet lysed completely after overnight incubation, add additional proteinase K (15–20 μl), and incubate at 56 °C for a few hours on a thermomixer until all tissue is dissolved.
8. In case one heating block is used, leave the sample(s) at room temperature after the 56 °C incubation until the temperature within the heating block has reached 90 °C or 95 °C. Also, be aware that longer incubation at 90 °C/95 °C may result in more fragmented DNA.
9. The DNA sample, buffer AL, and ethanol should be mixed immediately and thoroughly by vortexing or pipetting. When processing multiple samples, buffer AL and ethanol can be premixed and added together in one step.
10. If the lysate has not completely passed through the membrane after centrifugation, centrifuge again at a higher speed until the QIAamp MinElute column is empty.
11. Buffers AW1 and AW2 are provided as concentrated solutions. Make sure that ethanol has been added to prepare Buffer AW1 and Buffer AW2 in a correct way before use.
12. When buffer AW2, containing ethanol, is not removed completely, this will end up in the eluate and may interfere with downstream applications.
13. Ensure that the elution buffer is equilibrated to room temperature, and add the elution buffer onto the center of the membrane to ensure complete elution of bound DNA. This is especially important when elution volumes <50 μl are used.
14. For the subsequent PCR step (both Ion Torrent and Illumina protocol), 40 ng input DNA is standardly used per reaction as measured by Qubit, which yields optimal results. With limited DNA stock available, it is possible to go as low as 10 ng input DNA per PCR reaction [6].
15. DNA extraction can be performed at a later time point. To do so, tissue sections should be placed in an empty 1.5 ml Eppendorf tube and stored at −80 °C until further processing.
16. After ethanol is added and mixed, the sample can also be incubated for up to a few hours in the fridge. This could potentially increase the DNA yield obtained after the complete procedure.

17. When multiple samples are prepared at the same time, a master mix can be made containing the shared components of the different reactions. This will save time and generates uniform reaction mixtures. Please note that such master mix should be prepared for the number of samples +10% or one extra sample, kept cool (not frozen), and the polymerase is added as a final component.
18. Temperature can alter the behavior of magnetic beads. These are tested and optimized for use at room temperature. Also make sure that the beads are homogenized before use, so the DNA to bead ratio is correct.
19. Agencourt AMPure XP magnetic beads can also be pipetted into the DNA LoBind plate first for each sample, after which the pooled samples are added to the beads.
20. Make sure that the beads form a compact pellet before removing the supernatant. If not, move the plate a little bit around the magnet, and allow them to form a more distinctive compact pellet.
21. Be aware that as little beads as possible are taken up. This will lower the target specific amplicon yield after purification, as these molecules are bound to the beads. In Subheading 3.4.4, **step 5**, the target specific amplicons are in the supernatant, while undesired molecules are bound to the magnetic beads. Pipetting too many beads along with the supernatant can disturb downstream applications. To be sure no beads are transferred together with the supernatant, it is advised to transfer a smaller sample volume (e.g. 39 μl) and adjust with the appropriate beads volume in the next step.
22. It is recommended to use a freshly prepared 70% ethanol solution to wash the samples/beads, to ensure that the ethanol concentration is correct.
23. It is important that all residual ethanol is evaporated, but make sure not to overdry the beads as this will lower the DNA yield after recovery.
24. After each cleanup, samples can be stored at 4 °C for up to 1 week before continuing to the next step.
25. This protocol uses half the amount of reagents per sample as advised with the Ion Xpress kit. This has been tested extensively and works fine for this specific assay.
26. To ensure a constant room temperature, it is recommended to perform this incubation step in a thermocycler at 20 °C.
27. It is very important to use the Ion P1 Adapter from the barcode kit and not the adapters that are provided in the Ion Plus Fragment Library kit (green lid). In case the incorrect

adapters are used, the barcode ligation will be very inefficient and up to 85% of the generated reads will not be barcoded and are useless.

28. For the preparation of samples for one Ion Torrent run containing 24–32 samples, the same number of different barcodes is required. Each kit of Ion Xpress™ Barcode Adapters contains 16 different barcodes, so at least 2 kits are required for a library of 24–32 samples.
29. When collecting the supernatant containing the amplicon products, only 12.5 μl is collected into new PCR strip for this step, to ensure that the pellet is not disturbed. Furthermore, in this way the samples are ready to be used for the next step which requires 12.5 μl of pooled amplicons.
30. For sequencing and sample pool preparation, always follow the protocols and instructions from the local Sequence Facility. There are different Ion Torrent sequencers with associated kits and chips (e.g., Ion PGM Template OT2 200 Kit, Ion 510TM & Ion 520TM & Ion 530TM Kit Chef, Ion Chef [or Ion OneTouch 2 System]), which may require different concentrations and volumes of the prepared sample pool for optimal results. For running a 318 chip on the Ion Chef, it is recommended to include 24–32 samples per run.
31. When preparing a set of multiple samples, PCR master mixes can be used. Depending on the choices of the barcodes primers and the number of samples, one of the primers (i.e., forward or reverse) can eventually be included in the master mix. Otherwise, add each barcoded primer separately to each sample. Please note that each sample analyzed in a single sequencing run should get a unique combination of a forward and reverse barcode.
32. For sequencing and sample pool preparation, always follow the protocols and instructions from the local Sequence Facility. There are different Illumina sequencers that can be used, with their associated kits and chips for sequencing (e.g., MiniSeq sequencer and MiniSeq Mid Output Kit, or MiSeq sequencer and MiSeq Reagent Kit v2). Depending on the sequencing equipment, different concentrations and volumes of the prepared sample pool may be required for optimal results. For example, sequencing on a MiniSeq instrument using a mid-output chip requires a sample pool of 4 nM in a volume of 20 μl. Also the number of samples that can be analyzed in a single sequencing run depends on the sequencing instrument and chip. For running a mid-output chip on a MiniSeq instrument, it is recommended to include 24–32 samples per run.
33. ARResT/Interrogate is best viewed using Google Chrome or Firefox. The availability of the below described functions (i.e.,

"reporting" and "questions") depends on the user mode. However, in each user mode, at least one of these functions is available. For specific questions regarding an ARResT/Interrogate account, please contact the ARResT team (contact@arrest.tools).

Acknowledgments

The development of the NGS-based protocols for clonality detection was executed by laboratories within the EuroClonality-NGS Working Group, part of the EuroClonality consortium. A special thanks to Jos Rijntjes and Jeroen Luijks (Department of Pathology, Radboud University Medical Center, Nijmegen, The Netherlands) for technical assistance during the development of the two-step Illumina protocol for clonality detection. This project was funded by EuroClonality and the Dutch Cancer Society (KWF-11137). Figures are created with BioRender.com.

References

1. Melchers F (2015) Checkpoints that control B cell development. J Clin Invest 125(6): 2203–2210. https://doi.org/10.1172/jci78083
2. Pieper K, Grimbacher B, Eibel H (2013) B-cell biology and development. J Allergy Clin Immunol 131(4):959–971. https://doi.org/10.1016/j.jaci.2013.01.046
3. Rajewsky K (1996) Clonal selection and learning in the antibody system. Nature 381(6585):751–758. https://doi.org/10.1038/381751a0
4. Kim DR, Park SJ, Oettinger MA (2000) V(D)J recombination: site-specific cleavage and repair. Mol Cells 10(4):367–374
5. Gellert M (2002) V(D)J recombination: RAG proteins, repair factors, and regulation. Annu Rev Biochem 71:101–132. https://doi.org/10.1146/annurev.biochem.71.090501.150203
6. Scheijen B, Meijers RWJ, Rijntjes J, van der Klift MY, Möbs M, Steinhilber J, Reigl T, van den Brand M, Kotrová M, Ritter JM, Catherwood MA, Stamatopoulos K, Brüggemann M, Davi F, Darzentas N, Pott C, Fend F, Hummel M, Langerak AW, Groenen P (2019) Next-generation sequencing of immunoglobulin gene rearrangements for clonality assessment: a technical feasibility study by EuroClonality-NGS. Leukemia 33(9): 2227–2240. https://doi.org/10.1038/s41375-019-0508-7
7. Sanchez ML, Almeida J, Gonzalez D, Gonzalez M, Garcia-Marcos MA, Balanzategui A, Lopez-Berges MC, Nomdedeu J, Vallespi T, Barbon M, Martin A, de la Fuente P, Martin-Nuñez G, Fernandez-Calvo J, Hernandez JM, San Miguel JF, Orfao A (2003) Incidence and clinicobiologic characteristics of leukemic B-cell chronic lymphoproliferative disorders with more than one B-cell clone. Blood 102(8): 2994–3002. https://doi.org/10.1182/blood-2003-01-0045
8. Collins AM, Watson CT (2018) Immunoglobulin light chain gene rearrangements, receptor editing and the development of a self-tolerant antibody repertoire. Front Immunol 9:2249. https://doi.org/10.3389/fimmu.2018.02249
9. van Zelm MC, van der Burg M, de Ridder D, Barendregt BH, de Haas EF, Reinders MJ, Lankester AC, Révész T, Staal FJ, van Dongen JJ (2005) Ig gene rearrangement steps are initiated in early human precursor B cell subsets and correlate with specific transcription factor expression. J Immunol 175(9):5912–5922. https://doi.org/10.4049/jimmunol.175.9.5912
10. Rees AR (2020) Understanding the human antibody repertoire. MAbs 12(1):1729683. https://doi.org/10.1080/19420862.2020.1729683

11. Neuberger MS, Milstein C (1995) Somatic hypermutation. Curr Opin Immunol 7(2): 248–254. https://doi.org/10.1016/0952-7915(95)80010-7
12. Gazzola A, Mannu C, Rossi M, Laginestra MA, Sapienza MR, Fuligni F, Etebari M, Melle F, Sabattini E, Agostinelli C, Bacci F, Sagramoso Sacchetti CA, Pileri SA, Piccaluga PP (2014) The evolution of clonality testing in the diagnosis and monitoring of hematological malignancies. Ther Adv Hematol 5(2):35–47. https://doi.org/10.1177/2040620713519729
13. van Dongen JJ, Langerak AW, Bruggemann M, Evans PA, Hummel M, Lavender FL, Delabesse E, Davi F, Schuuring E, Garcia-Sanz R, van Krieken JH, Droese J, Gonzalez D, Bastard C, White HE, Spaargaren M, Gonzalez M, Parreira A, Smith JL, Morgan GJ, Kneba M, Macintyre EA (2003) Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 concerted action BMH4-CT98-3936. Leukemia 17(12): 2257–2317. https://doi.org/10.1038/sj.leu.2403202
14. Boone E, Heezen KC, Groenen P, Langerak AW (2019) PCR GeneScan and heteroduplex analysis of rearranged immunoglobulin or T-cell receptor genes for clonality diagnostics in suspect lymphoproliferations. Methods Mol Biol 1956:77–103. https://doi.org/10.1007/978-1-4939-9151-8_4
15. Evans PA, Pott C, Groenen PJ, Salles G, Davi F, Berger F, Garcia JF, van Krieken JH, Pals S, Kluin P, Schuuring E, Spaargaren M, Boone E, Gonzalez D, Martinez B, Villuendas R, Gameiro P, Diss TC, Mills K, Morgan GJ, Carter GI, Milner BJ, Pearson D, Hummel M, Jung W, Ott M, Canioni D, Beldjord K, Bastard C, Delfau-Larue MH, van Dongen JJ, Molina TJ, Cabecadas J (2007) Significantly improved PCR-based clonality testing in B-cell malignancies by use of multiple immunoglobulin gene targets. Report of the BIOMED-2 concerted action BHM4-CT98-3936. Leukemia 21(2):207–214. https://doi.org/10.1038/sj.leu.2404479
16. van Krieken JH, Langerak AW, Macintyre EA, Kneba M, Hodges E, Sanz RG, Morgan GJ, Parreira A, Molina TJ, Cabeçadas J, Gaulard P, Jasani B, Garcia JF, Ott M, Hannsmann ML, Berger F, Hummel M, Davi F, Brüggemann M, Lavender FL, Schuuring E, Evans PA, White H, Salles G, Groenen PJ, Gameiro P, Pott C, Dongen JJ (2007) Improved reliability of lymphoma diagnostics via PCR-based clonality testing: report of the BIOMED-2 concerted action BHM4-CT98-3936. Leukemia 21(2):201–206. https://doi.org/10.1038/sj.leu.2404467
17. Langerak AW, Groenen PJ, Bruggemann M, Beldjord K, Bellan C, Bonello L, Boone E, Carter GI, Catherwood M, Davi F, Delfau-Larue MH, Diss T, Evans PA, Gameiro P, Garcia Sanz R, Gonzalez D, Grand D, Hakansson A, Hummel M, Liu H, Lombardia L, Macintyre EA, Milner BJ, Montes-Moreno S, Schuuring E, Spaargaren M, Hodges E, van Dongen JJ (2012) EuroClonality/BIOMED-2 guidelines for interpretation and reporting of Ig/TCR clonality testing in suspected lymphoproliferations. Leukemia 26(10):2159–2171. https://doi.org/10.1038/leu.2012.246
18. Bystry V, Reigl T, Krejci A, Demko M, Hanakova B, Grioni A, Knecht H, Schlitt M, Dreger P, Sellner L, Herrmann D, Pingeon M, Boudjoghra M, Rijntjes J, Pott C, Langerak AW, Groenen P, Davi F, Bruggemann M, Darzentas N (2017) ARResT/Interrogate: an interactive immunoprofiler for IG/TR NGS data. Bioinformatics 33(3):435–437. https://doi.org/10.1093/bioinformatics/btw634
19. Merriman B, Rothberg JM (2012) Progress in ion torrent semiconductor chip based sequencing. Electrophoresis 33(23):3397–3417. https://doi.org/10.1002/elps.201200424
20. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IM, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DM, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Chiara ECM, Chang S, Neil Cooley R, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo KV, Scott Furey W, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson

MQ, James T, Huw Jones TA, Kang GD, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ling Ng B, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Pike AC, Chris Pinkard D, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Chiva Rodriguez A, Roe PM, Rogers J, Rogert Bacigalupo MC, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Ernest Sohna Sohna J, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Rogers J, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456(7218):53–59. https://doi.org/10.1038/nature07517

21. Bruggemann M, Kotrova M, Knecht H, Bartram J, Boudjogrha M, Bystry V, Fazio G, Fronkova E, Giraud M, Grioni A, Hancock J, Herrmann D, Jimenez C, Krejci A, Moppett J, Reigl T, Salson M, Scheijen B, Schwarz M, Songia S, Svaton M, van Dongen JJM, Villarese P, Wakeman S, Wright G, Cazzaniga G, Davi F, Garcia-Sanz R, Gonzalez D, Groenen P, Hummel M, Macintyre EA, Stamatopoulos K, Pott C, Trka J, Darzentas N, Langerak AW (2019) Standardized next-generation sequencing of immunoglobulin and T-cell receptor gene recombinations for MRD marker identification in acute lymphoblastic leukaemia; a EuroClonality-NGS validation study. Leukemia 33(9):2241–2253. https://doi.org/10.1038/s41375-019-0496-7

22. van den Brand M, Rijntjes J, Mobs M, Steinhilber J, van der Klift MY, Heezen KC, Kroeze LI, Reigl T, Porc J, Darzentas N, Luijks J, Scheijen B, Davi F, ElDaly H, Liu H, Anagnostopoulos I, Hummel M, Fend F, Langerak AW, Groenen P, EuroClonality NGSWG (2021) Next-generation sequencing-based clonality assessment of Ig gene rearrangements: A multicenter validation study by euroClonality-NGS. J Mol Diagn 23(9):1105–1115

Chapter 3

One-Step Next-Generation Sequencing of Immunoglobulin and T-Cell Receptor Gene Recombinations for MRD Marker Identification in Acute Lymphoblastic Leukemia

Patrick Villarese, Chrystelle Abdo, Matthieu Bertrand, Florian Thonier, Mathieu Giraud, Mikaël Salson, and Elizabeth Macintyre

Abstract

Within the EuroClonality-NGS group, immune repertoire analysis for target identification in lymphoid malignancies was initially developed using two-stage amplicon approaches, essentially as a progressive modification of preceding methods developed for Sanger sequencing. This approach has, however, limitations with respect to sample handling, adaptation to automation, and risk of contamination by amplicon products. We therefore developed one-step PCR amplicon methods with individual barcoding for batched analysis for IGH, IGK, TRD, TRG, and TRB rearrangements, followed by Vidjil-based data analysis.

Key words Next-generation sequencing, One step, T cell receptor, B cell receptor

1 Introduction

Recombination of the V (D) J genes of immunoglobulin (IG) and T cell receptor (TR) loci is an essential step in the differentiation of B and T cells, allowing the production of a unique antigen receptor which is present in all clonal progeny. As such, acute lymphoblastic leukemias (ALLs) are characterized by clonal, homogeneous IG/TR rearrangement patterns that are widely used for clonal tracking during evaluation of response to treatment, commonly referred to as quantification of minimal (or measurable) residual disease (MRD) [1]. The EuroMRD group has played a seminal role in developing, standardizing, and accompanying optimized use of IG/TR clonal markers in lymphoid malignancies, essentially using CDR3 clone-specific quantitation by PCR. Initial IG/TR target identification was based predominantly on EuroClonality/BIOMED-2 multiplex PCR-based protocols for IG/TR targets combined with heteroduplex analysis or fragment length (GeneScan) analysis, followed by Sanger sequencing and design of

Anton W. Langerak (ed.), *Immunogenetics: Methods and Protocols*, Methods in Molecular Biology, vol. 2453, https://doi.org/10.1007/978-1-0716-2115-8_3,

CDR3-specific PCR primers [2–4]. With the development of NGS immunogenetics [5–9], the EuroClonality-NGS working group developed a standardized two-step multiplex amplicon approach to IG/TR target identification in ALL that enabled switching of sequencing adaptors and a reduction of the total number of primers required for individual sample identification in mixed libraries [10].

Two-step PCR approaches, however, have several limitations, particularly in MRD laboratories, where contamination by PCR products can be a risk of false-positive results. These include more extensive sample handling with consequent increased overall cost and risk of contamination and reduced suitability for automation. We therefore developed a single-step PCR approach to screening for IG/TR rearrangements in lymphoid malignancies, as described here.

2 Materials

2.1 Sample Preparation

1. 15 mL polypropylene tubes.
2. Phosphate-buffered saline (1xPBS) without Ca2+ and Mg2+ pH 7.4 (Invitrogen).
3. Sysmex XE 2100.
4. Maxwell RSC instrument (Promega).
5. Maxwell RSC Buffy Coat DNA kit (Promega).
6. Nanodrop ND2000 (Thermo Fisher Scientific).
7. Centrifuge (1000 × *g*).
8. 2 mL tubes (Eppendorf).

2.2 PCR Amplification

1. UltraPure Distilled Water DNase-/RNase-Free (Invitrogen).
2. SafeSeal Microcentrifuge Tubes (Sorenson).
3. 0.2 mL Thin-walled Tubes with Flat Caps (Thermo Fisher Scientific).
4. Kit FastStart High Fidelity PCR System, dntPack (Roche Diagnostic).
5. Thermocycler BioRad T100 or Applied Biosystem Veriti 96.

2.3 Sample Purification

1. Agencourt AMPure XP (Beckman Coulter).
2. 0.8 mL 96-well storage plate (Thermo Fisher Scientific).
3. TE buffer pH 8 (Invitrogen).
4. MicroAmp Optical Adhesive Film (Thermo Fisher Scientific).
5. DynaMag-96 Side Skirted Magnet (Thermo Fisher Scientific).

2.4 Sample Assay

1. Assay plate, 96 well (Costar).
2. Kit QuantiFluor ONE dsDNA System (Promega).
3. GLOMAX (Promega).
4. Qubit 4 fluorometer (Thermo Fisher Scientific).
5. Qubit assay tubes (Thermo Fisher Scientific).
6. 2100 Bioanalyzer Instrument (Agilent Technologies).
7. Agilent High Sensitivity DNA Kit (Agilent Technologies).

2.5 Pool Sample (2 nM)

1. TE buffer (Invitrogen).
2. DNA low bind tubes 1.5 mL (Eppendorf).

2.6 Denaturation Step Before Sequencing

1. Sodium Hydroxide solution, 10 M in H_2O (Sigma Aldrich).
2. PhiX Control v3 (Illumina).
3. UltraPure Distilled Water DNase-/RNase-Free (Invitrogen).

2.7 Sequencing

1. UltraPure Distilled Water DNase-/RNase-Free (Invitrogen).
2. Tween 20 (Euromedex).
3. Precision wipes (KIMTECH Science).
4. MiSeq Reagent Kit V2 2x250pb (Illumina).
5. MiSeq System (Illumina).

2.8 Bioinformatic Analysis

Access to a Vidjil server allowing hosting of patient data [11] (see http://www.vidjil.org/doc/healthcare/).

3 Methods

3.1 Sample Preparation

1. Use blood or bone marrow cells.
2. Enumerate white blood cells, e.g., with theSysmex XE2100 system.
3. Extract DNA from ten million white blood cells with the Maxwell RSC Buffy Coat DNA kit.
4. After extraction, quantify DNA by Nanodrop.
5. If necessary, adjust DNA concentration to 100 ng/μL with TE buffer pH 8.

3.2 PCR Amplification

3.2.1 Prepare a Mix of Primers for each Target of Interest (See Notes Below)

1. Prepare the primer mix for IGH VDJ FR2 (*see* **Note 1**).
2. Prepare the primer mix for IGH DHJH (*see* **Note 2**).
3. Prepare the primer mix for IGK (*see* **Note 3**).
4. Prepare the primer mix for TRG (*see* **Note 4**).
5. Prepare the primer mix for TRD (*see* **Note 5**).

6. Prepare the primer mix for TRB DJ (*see* **Note 6**).
7. Prepare the primer mix for TRB VDJ (*see* **Note 7**).

Importantly, each primermix should be prepared with the same index.

3.2.2 PCR Amplification

1. Prepare the PCR mix for each reaction on ice (*see* Table 1).
2. First, mix H_20, buffer, and $MgCl_2$ on ice.
3. Then prepare a 0.1× dilution of Taq polymerase with H_2O.
4. Add primer indexes to the mix.
5. Lastly, add 100 ng of patient DNA to each PCR (or 250 ng DNA for the TRG reaction).
6. Run amplification protocol in a thermocycler (*see* Table 2).

3.3 Sample Purification

Remark: TRG samples do not need to be purified, but other targets must be purified with double purification ratio 0.6×/0.25×. Take out the AMPure XP Kit at least 30 min before use.

1. Take a storage plate and add 28.8 μL of Agencourt beads per well.
2. Add 48 μL of sample to the beads per well.
3. Cover with an adhesive film.
4. Centrifuge at 280 × *g* for 1 min.
5. Put the plate on a microplate shaker at 200 × *g* for 2 min.
6. Incubate the plate for 5 min at room temperature.
7. Centrifuge at 280 × *g* for 1 min.
8. Put the plate to the side skirted magnet for 5 min.
9. Transfer 76 μL of supernatant to a new storage plate.
10. Add to the new wells 19 μL beads for the IGH VDJ/IGK/TRD/TRB DJ/TRB VDJ reactions and 15.2 μL for the IGH-DJ reaction.
11. Cover with adhesive film.
12. Centrifuge at 280 × *g* for 1 min.
13. Put the plate to the side skirted magnet for 5 min.
14. Discard the supernatant and wash the beads twice with 190 μL 70% ethanol.
15. Shift the plate on the side skirted magnet and wait 1 min.
16. Discard all supernatant and wait 1 min.
17. Leave the plate on the side skirted magnet and add 10 μL TE.
18. Centrifuge at 280 × *g* for 1 min.
19. Put the plate to the side skirted magnet for 5 min.
20. Collect 8.5 μL of each sample.

Table 1
PCR reaction mixes for different IG and TR targets

IGH V-J FR2			**TRG**			**TRB V-J**		
	Final conc.	*µL/1×*		*Final conc.*	*µL/1×*		*Final conc.*	*µL/1×*
H_2O	n.a.	20	H_2O	n.a.		H_2O	n.a.	30.75
10× buffer	1×	5	10× buffer	1×	5	10× buffer	1×	5
$MgCl_2$ 25 mM	2 mM	4	$MgCl_2$ 25 mM	1.5 mM	3	$MgCl_2$ 25 mM	4 mM	8
dNTPs 10 mM	0.2 mM	1	dNTPs 10 mM	0.2 mM	1	dNTPs 10 mM	0.2 mM	1
Primer mix FR2V – 10 µM pool A	0.1 µM	11	Primer mix A 10 µM	0.1 µM	2.5	Primer mix 100 µM VB	Variable	0.5
Primer mix FR2V – 10 µM pool B	0.2 µM	5	Primer mix B 10 µM	0.05 µM	0.5	Primer mix 10 µM JB	0.025 µM each	1.75
Primer mix T7JH – 10 µM pool C	0.1 µM	1	Primer mix TRGV11 10 µM	0.2 µM	1			
			Primer mix C 20 µM	0.2 µM	2			
Taq high Fidelity; 5 U/µL; dil. 1:10	1 U	2	Taq high Fidelity; 5 U/µL; dil. 1:10	2.5 U	5	Taq high Fidelity; 5 U/µL; dil. 1:10	1 U	2
IGH D-J			**TRD**					
H_2O	n.a.	33	H_2O	n.a.	37.25			
10× buffer with 18 mM $MgCl_2$	n.a.	5	10× buffer with 18 mM $MgCl_2$	n.a.	5			
dNTPs 10 mM	0.2 mM	1	dNTPs 10 mM	0.2 mM	1			
Primer mix DH 10 µM	0.1 µM	4	Primer mix V DD2	0.1 µM	0.45			
Primer mix JH 10 µM	0.2 µM	1	Primer mix J DD	0.1 µM	0.3			

(continued)

Table 1
(continued)

Taq high Fidelity; 5 U/μL; dil. 1:10	2.5 U	5	Taq high Fidelity; 5 U/μL; dil. 1:10	2.5 U	5
IGK			**TRB D-J**		
H_2O	n.a.	22	H_2O	n.a.	30.25
10× buffer with 18 mM $MgCl_2$	n.a.	8	10× buffer	1×	5
dNTPs 10 mM	0.4 mM	2	$MgCl_2$ 25 mM	4 mM	8
Primer mix VK 10 μM	0.2 μM	13	dNTPs 10 mM	0.2 mM	1
Primer mix intron new 10 μM	0.2 μM	1	Primer mix 10 μM DB	0.1 μM each	1
Primer mix Kde 10 μM	0.2 μM	1	Primer mix 10 μM JB	0.025 μM each	1.75
Taq high Fidelity; 5 U/μL; dil. 1:10	1 U	2	Taq high Fidelity; 5 U/μL; dil. 1:10	1 U	2

Table 2
Amplification protocols for different IG and TR targets

IGH V-J FR2 (35 cycles)	IGH D-J (35 cycles)	IGK (30 cycles)	TRD (35 cycles)	TRG (35 cycles)	TRB V-J (35 cycles)	TRB D-J (35 cycles)
94 °C 10′	94 °C 10′	94 °C 10′	94 °C 8′	94 °C 8′	94 °C 10′	94 °C 10′
94 °C 1′	94 °C 1′	92 °C 30″	94 °C 45″	94 °C 45″	94 °C 1′	94 °C 1′
63 °C 1′	63 °C 1′	61 °C 40″	62 °C 1′	57 °C 1′	65 °C 1′	63 °C 1′
72 °C 30″	72 °C 30″	72 °C 40″	72 °C 1′30″	72 °C 1′30″	72 °C 30″	72 °C 30″
72 °C 30′	72 °C 30′	72 °C 30′	72 °C 10′	72 °C 10′	72 °C 30′	72 °C 30′
15 °C final	15 °C final	15 °C final	15 °C final	15 °C final	15 °C final	15 °C final

3.4 Sample Assay Quantification

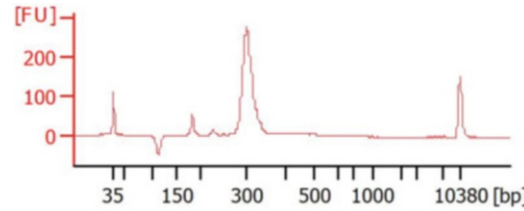
1. Take an assay plate.
2. Prepare a dilution of 199 µL ds DNA Dye buffer +1 µL purified library.
3. Measure the concentration in ng/µL of samples at GLOMAX.
4. Transform ng/µL into nM with this formula:

$$= (\text{conc ng}/\mu\text{L} \times 10^6)/(\text{size of library in base pairs} \times 660).$$

Option: One can verify the size of each library by electrophoresis on a Bioanalyzer 2100. Analyze 1 µL sample with the DNA High Sensitivity Agilent kit.

After migration, profiles and sizes should be as illustrated below (example is shown for TRB VDJ).

Locus	IGH R2	IGH DJ	IGK	TRG	TRD	TRB DJ	TRB VDJ
Median size (pb)	430	250	300	260	300	300	300



3.5 Pool Preparation (2 nM)

1. Make a dilution at 2 nM of each sample with TE.
2. Make an equimolar pool at 2 nM with 5 µL of each sample.
3. Measure the concentration of the pool by Qubit.

4. Transform ng/μL into nM with this formula:

$$= \left(\text{conc ng}/\mu\text{L} \times 10^6\right)/(\text{size of library in base pair} \times 660).$$

3.6 Denaturation Step

1. Normalize the library pool to 2 nM in Resuspension Buffer.
2. Prepare 0.1 N NaOH: 5 μL 2 N NaOH +95 μL H_2O (or 1 μL 10 N NaOH +99 μL H_2O).
3. Vortex.
4. Put the HT1 tube in ice.
5. Add in an Eppendorf tube: 5 μL library pool 2 nM + 5 μL 0.1 N NaOH.
6. Vortex.
7. Centrifuge quickly.
8. Incubate for 5 min at room temperature (DNA denaturation).
9. Add 823 μL ice-cold HT1 to prepare a 12 pM denatured library pool.
10. Vortex.
11. Centrifuge quickly.
12. Place the Eppendorf tube on ice until it settles in the cartridge.
13. Add in another Eppendorf tube 120 μL 20 pM denatured PhiX library +80 μL ice-cold HT1 to prepare a 12pM denatured PhiX library.
14. Vortex.
15. Centrifuge quickly.
16. Place the Eppendorf tube on ice.

Adding 10% of PHIX control in pool library:

(a) In 2 mL low bind tube: 540 μL 12pM denatured library pool +60 μL 12pM denatured PhiX library.
(b) Vortex.
(c) Centrifuge quickly.
(d) Place the Eppendorf tube in ice.
(e) Load 600 μL into the "load sample" well of the MiSeq cartridge V2.

3.7 Bioinformatic Analysis with the Vidjil Platform

1. Copy each of the FASTQ files in the folder MiSeq Output.
2. Connect to the Vidjil server [11] with a personal login and password.
3. Create a "run" and as many "patients" as necessary (*see* Fig. 1).

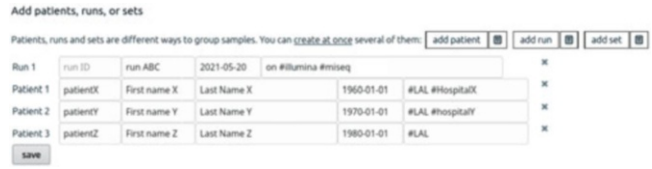


Fig. 1 Adding patients and runs in Vidjil

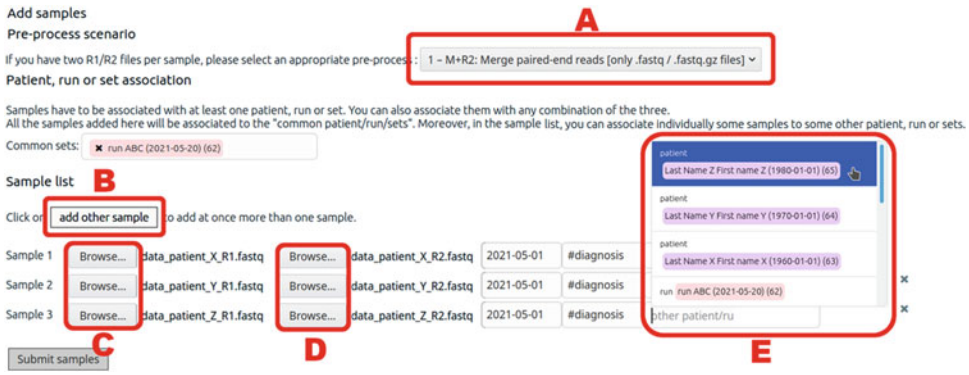


Fig. 2 Adding samples in Vidjil. The rectangles refer to the different steps described in the main text

(a) Click on *runs* and then on *new runs*.

(b) Fill information on the run (date, metadata on the sample using tags prefixed with a #). Afterwards, the samples can be searched by tags.

(c) Add as many *patients* as required and specify a first and last name for each case.

4. Open the created run and click on the *Add samples* button.

(a) Select the pre-process *M + R2: Merge paired-end reads* (A in Fig. 2).

(b) Click on *Add other sample* to have as many sample lines as required (B in Fig. 2).

(c) Add each sample one by one.

- Select the FASTQ file for the R1 reads in the first field (C in Fig. 2).
- Select the FASTQ file for the R2 reads in the second field (D in Fig. 2).
- Enter the sampling date.
- In the last field, type the last name of the patient and select the corresponding one in the list that appears (E in Fig. 2). This will associate the sample to the patient, which will then be available from *run* or *patient*.

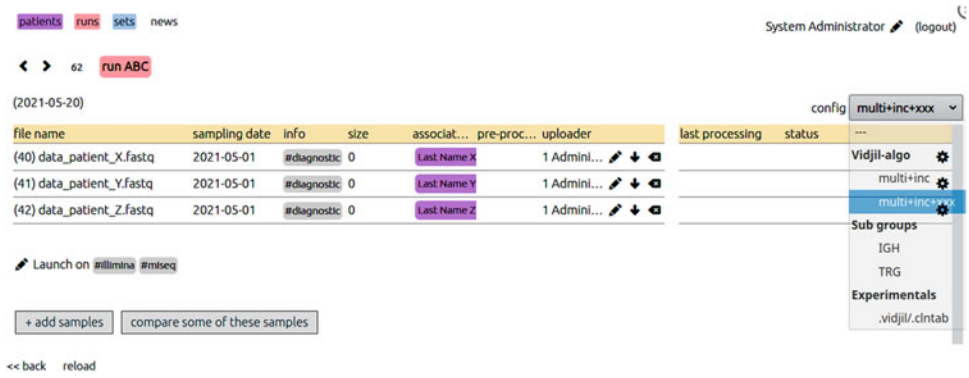


Fig. 3 Selecting configuration and launching Vidjil-algo processes

5. Submit the samples.
6. Choose the configuration of the algorithm: "multi+inc+xxx." This is the advised configuration for target identification as it will detect both complete and incomplete recombinations (Fig. 3).
7. Launch the analysis with the selected configuration for each sample.
8. Click on *reload*, at the bottom left, to see the job status going through the different steps: QUEUED → ASSIGNED → RUNNING → COMPLETED. It is possible to launch several processes at the same time (some will wait in the QUEUED/ASSIGNED states).
9. Once the jobs are completed, return to the patient list to visualize the results by clicking on the configuration name.
10. Analyze the sample to determine the markers of interest (Fig. 4).
    (a) The percentage of analyzed reads should normally be above 90%; otherwise the sequencing run may be of poor quality (A in Fig. 4).
        - In case this percentage is too low, investigate the reason why by clicking on the info button in the upper left panel (B in Fig. 4).
        - Specifically, check the percentage of reads that are classified as:
            - UNSEG only V/5′ (reads only matching V genes).
            - UNSEG only J/3′ (reads only matching J genes).
            - UNSEG too few V/J (reads matching no V or J gene).

Fig. 4 Analyzing the clonotypes in the Vidjil client. Clonotypes are viewed at the same time in a Genescan-like view, a grid view (depending on V/J genes) and in a list. Moreover, the sequences of the selected clonotypes appear at the bottom

(b) Identify the loci of interest, with at least 10,000 reads (C in Fig. 4).

11. Study each clonotype of interest one by one.
12. Switch in order to each of those loci.
13. Cluster all sub-clonotypes linked to the clonotype being studied.
    (a) Select all the clonotypes with the same V and J genes as the studied clonotype.
    (b) Align the sequences (D in Fig. 4).
    (c) Remove the sequences that do not align properly with the studied clonotype.
    (d) Realign the sequences.
    (e) Restart steps c and d until all the sequences align with only few differences.
    (f) Cluster the aligned sequences (button *cluster*, E in Fig. 4).
14. Send the clonotypes to IMGT/V-QUEST [12–15], by clicking on the IMGT button (F in Fig. 4). Next the V, D, and J genes as computed by IMGT/V-QUEST are underlined. This must be taken into account for the design of the patient-specific primer in case of MRD analysis by qPCR.
15. Save the analysis by going to the menu at the top left corner and click on *save*.

4 Notes

1. Primer mix for IGH VDJ FR2. Each primer should be mixed with the same index; mixing of primers needs to be repeated for each unique index. Prepare a 1.5 mL low bind tube for primer mixes A, B, and C:
   (a) Tube A: combine primer for index D502.
   - Add 2 μL of each primer at 100 μM + 396 μL H_2O; each primer is at 10 μM.

IGHVFR2-5	IGHVFR2-21	IGHVFR2-33	IGHVFR2-44
IGHVFR2-10	IGHVFR2-22	IGHVFR2-38	F93
IGHVFR2-11	IGHVFR2-23	IGHVFR2-39	F83
IGHVFR2-12	IGHVFR2-27	IGHVFR2-41	F88
IGHVFR2-13	IGHVFR2-28	IGHVFR2-42	F75
	IGHVFR2-32	IGHVFR2-43	

   (b) Tube B: combine primer for index D502.
   - Add 2 μL of each primer at 100 μM + 90 μL H_2O; each primer is at 10 μM.

IGHVFR2-18	IGHVFR2-36
IGHVFR2-19	IGHVFR2-45
IGHVFR2-20	

   (c) Tube C: combine primer for index D701.
   - Add 2 μL of each primer at 100 μM + 36 μL H_2O; each primer is at 10 μM.

T7-JH consensus
T7-IGJH-137(faham)

2. Primer mix for IGH DHJH. Each primer should be mixed with the same index; mixing of primers needs to be repeated for each unique index. Prepare a 1.5 mL low bind tube for primer mixes of DH primers and JH primers:
   (a) Tube DH primer: combine primer for index D502.
   - Add 2 μL of each primer at **10 μM**; each primer is at 10 μM.

DH1	DH4
DH2	DH5a
DH3a	DH5b
DH3b	DH6

   (b) Tube JH primer: combine primer for index D701.

- Add 2 μL of each primer at 100 μM + 36 μL H_2O; each primer is at 10 μM.

T7-JH consensus
T7-IGJH-137(faham)

3. Primer mix for IGK. Each primer should be mixed with the same index; mixing of primers needs to be repeated for each unique index. Prepare a 1.5 mL low bind tube for primer mix Vkappa, Intron, Jkappa, and Kde:
   (a) Tube Vkappa: combine primer for index D502.
   - Add 5 μL of each primer at 100 μM + 585 μL H_2O; each primer is at 10 μM.

VK1-A	VK5
VK1-D	VK7
VK1-E	VK2-A
VK1-F	VK2-Bdef
VK6-D	VK2-D
VK4	VK3-B
	VK3-C

   (b) Tube Intron: primer for index D502.
   - Dilute 5 μL of each primer at 100 μM + 45 μL H_2O.

   (c) Tube Jkappa: combine primer for index D701.
   - Add 4 μL of each primer at 100 μM + 108 μL H_2O; each primer is at 10 μM.

JK1-4
JK5
JK3

   (d) Tube Kde: primer for index D701.
   - Dilute 5 μL of each primer at 100 μM + 45 μL H_2O.

4. Primer mix for TRG. Each primer should be mixed with the same index; mixing of primers needs to be repeated for each unique index. Prepare a 1.5 mL low bind tube for primer mix A TCRGV, mix B TCRGV, and mix C TCRGJ, TCRGV11:
   (a) Tube mix A TCRGV: combine primer for index D502.
   - Add 2 μL of each primer at 100 μM + 90 μL H_2O; each primer is at 10 μM.

TRGV2	TRGV3/5
TRGV4	TRGV10
TRGV8	

(b) Tube mix B TCRGV: combine primer for index D502.

- Add 2 µL of each primer at 100 µM + 36 µL H_2O; each primer is at 10 µM.

TRGV7
TRGV9

(c) Tube TCRGV11: primer for index D502.

- Dilute 1 µL of each primer at 100 µM + 36 µL H_2O.

(d) Tube mix C TCRGJ: primer for index D701.

- Mix 7 µL of each primer at 20 µM.

TRGJ1/2	TRGJP1
TRGJP01	TRGJP2

5. Primer Mix for TRD. Each primer should be mixed with the same index; mixing of primers needs to be repeated for each unique index. Prepare a 1.5 mL low bind tube for primer mix VDD2 and mix JDD:

(a) Tube mix VDD2: combine primer for index D502.

- Mix 5 µL of each primer at 100 µM.

VD1	VD5
VD2	VD6
VD3	VD8
VD4	VD7
	DD2-5'

(b) Tube mix JDD: combine primer for index D701.

- Mix 5 µL of each primer at 100 µM.

DD3-3'	JD4
JD1	Jalpha29
JD2	
JD3	

6. Primer mix for TRB DJ. Each primer should be mixed with the same index; mixing of primers needs to be repeated for each unique index. Prepare a 1.5 mL low bind tube for primer mix TRB DB and mix TRB JB:

(a) Tube mix TRB DB: combine primer for index D502.

- Mix 2 µL of each primer at 10 µM + 36 µL H_2O.

TRBDB1
TRBDB2

(b) Tube mix TRB JB: combine primer for index D701.

- Mix 2 μL of each primer at 10 μM + 252 μL H_2O.

TRBJ1.1	TRBJ2.1
TRBJ1.2	TRBJ2.2
TRBJ1.3	TRBJ2.3
TRBJ1.4	TRBJ2.4
TRBJ1.5	TRBJ2.5
TRBJ1.6	TRBJ2-6_1
	TRBJ2-6_2
	TRBJ2.7

7. Primer mix for TRB VDJ. Each primer should be mixed with the same index; mixing of primers needs to be repeated for each unique index. Prepare a 1.5 mL low bind tube for primer mix TRB VB and mix TRB JB:

(a) Tube mix TRB VB: combine primer for index D502.

- Mix each primer at 100 μM with the volume below:

primer	volume (μl)
TRBV2	2
TRBV4	2
TRBV5-5	2
TRBV5-3	2
TRBV5-4	2
TRBV5-8	2
TRBV6-2	2
TRBV6-6	2
TRBV6-7	2
TRBV7-3	2
TRBV7-5	2
TRBV7-8	2
TRBV9	2
TRBV10-2	2
TRBV12-5	2
TRBV13	2
TRBV14	2
TRBV15	2
TRBV16	2
TRBV18	2
TRBV21-1	2
TRBV24-1	2
TRBV25-1	2
TRBV27	2
TRBV29-1	2

primer	volume (μl)
TRBV3-1	4
TRBV5-1univ	4
TRBV6-4	4
TRBV7-7	4
TRBV28	4
TRBV30	4

primer	volume (μl)
TRBV5-6	8
TRBV10-3	8
TRBV11-1	8
TRBV12-3	8
TRBV19	8
TRBV20-1	8
TRBV23-1	8

(b) Tube mix TRB JB: combine primer for index D701.

- Mix 2 µL of each primer at 10 µM + 252 µL H_2O.

TRBJ1.1	TRBJ2.1
TRBJ1.2	TRBJ2.2
TRBJ1.3	TRBJ2.3
TRBJ1.4	TRBJ2.4
TRBJ1.5	TRBJ2.5
TRBJ1.6	TRBJ2-6_1
	TRBJ2-6_2
	TRBJ2.7

References

1. Tonegawa S (1983) Somatic generation of antibody diversity. Nature 302(5909): 575–581
2. Evans PAS, Pott C, Groenen PJTA, Salles G, Davi F, Berger F et al (2007) Significantly improved PCR-based clonality testing in B-cell malignancies by use of multiple immunoglobulin gene targets. Report of the BIOMED-2 concerted action BHM4-CT98-3936. Leukemia 21(2):207–214
3. Langerak AW, Groenen PJTA, Brüggemann M, Beldjord K, Bellan C, Bonello L et al (2012) EuroClonality/BIOMED-2-guidelines for interpretation and reporting of Ig/TCR clonality testing in suspected lymphoproliferations. Leukemia 26(10):2159–2171
4. van Dongen JJM, Langerak AW, Brüggemann M, Evans PA, Hummel M, Lavender FL et al (2003) Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 concerted action BMH4-CT98-3936. Leukemia 17(12):2257–2317
5. Langerak AW, Brüggemann M, Davi F, Darzentas N, van Dongen JJM, Gonzalez D et al (2017) High-throughput Immunogenetics for clinical and research applications in immunohematology: potential and challenges. J Immunol 198(10):3765–3774
6. Scheijen B, Meijers RWJ, Rijntjes J, van der Klift MY, Möbs M, Steinhilber J et al (2019) Next-generation sequencing of immunoglobulin gene rearrangements for clonality assessment: a technical feasibility study by EuroClonality-NGS. Leukemia 33(9): 2227–2240
7. Kotrova M, Trka J, Kneba M, Brüggemann M (2017) Is next-generation sequencing the way to go for residual disease monitoring in acute lymphoblastic leukemia? Mol Diagn Ther 21(5):481–492
8. Logan AC, Gao H, Wang C, Sahaf B, Jones CD, Marshall EL et al (2011) High-throughput VDJ sequencing for quantification of minimal residual disease in chronic lymphocytic leukemia and immune reconstitution assessment. Proc Natl Acad Sci U S A 108(52):21194–21199
9. Wu D, Sherwood A, Fromm JR, Winter SS, Dunsmore KP, Loh ML et al (2012) High-throughput sequencing detects minimal residual disease in acute T lymphoblastic leukemia. Sci Transl Med 4(134):134ra63
10. Brüggemann M, Kotrová M, Knecht H, Bartram J, Boudjogrha M, Bystry V et al (2019) Standardized next-generation sequencing of immunoglobulin and T-cell receptor gene recombinations for MRD marker identification in acute lymphoblastic leukaemia; a EuroClonality-NGS validation study. Leukemia 33(9):2241–2253
11. Giraud M, Salson M, Duez M, Villenet C, Quief S, Caillault A et al (2014) Fast multiclonal clusterization of V(D)J recombinations from high-throughput sequencing. BMC Genomics 15:409
12. Alamyar E, Duroux P, Lefranc M-P, Giudicelli V (2012) IMGT(®) tools for the nucleotide analysis of immunoglobulin (IG) and T cell receptor (TR) V-(D)-J repertoires, polymorphisms, and IG mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. Methods Mol Biol 882:569–604

13. Aouinti S, Giudicelli V, Duroux P, Malouche D, Kossida S, Lefranc MP (2016) IMGT/StatClonotype for pairwise evaluation and visualization of NGS IG and TR IMGT clonotype (AA) diversity or expression from IMGT/HighV-QUEST. Front Immunol 7: 339
14. Lefranc M-P (2014) Immunoglobulin and T cell receptor genes: IMGT® and the birth and rise of Immunoinformatics. Front Immunol 5: 22
15. Lefranc M-P, Giudicelli V, Duroux P, Jabado-Michaloud J, Folch G, Aouinti S et al (2015) IMGT®, the international ImMunoGeneTics information system® 25 years on. Nucleic Acids Res 43(Database issue):D413–D422

Chapter 4

Immunoglobulin/T-Cell Receptor Gene Rearrangement Analysis Using RNA-Seq

Vincent H. J. van der Velden, Lorenz Bastian, Monika Brüggemann, Alina M. Hartmann, and Nikos Darzentas

Abstract

Identification of immunoglobulin (IG) and T-cell receptor (TR) gene rearrangements in acute lymphoblastic leukemia (ALL) patients at initial presentation are crucial for monitoring of minimal residual disease (MRD) during subsequent follow-up and thereby for appropriate risk-group stratification. Here we describe how RNA-Seq data can be generated and subsequently analyzed with ARResT/Interrogate to identify possible MRD markers. In addition to the procedures, possible pitfalls will be discussed. Similar strategies can be employed for other lymphoid malignancies, such as lymphoma and myeloma.

Key words Minimal residual disease, Acute lymphoblastic leukemia, Immunoglobulin, T-cell receptor, Gene rearrangements, RNA-Seq, Whole exome sequencing, Whole genome sequencing, Marker identification

1 Introduction

Most clinical protocols for patients with acute lymphoblastic leukemia (ALL) nowadays include minimal residual disease (MRD)-based stratification [1–4]. Molecular MRD analysis is, at least in Europe, most commonly used and is generally based on analysis of rearranged immunoglobulin (IG) and T-cell receptor (TR) genes according to international guidelines [5–8]. In a diagnostic setting, IG/TR gene rearrangements are generally identified using DNA-based PCR analysis, followed by classical Sanger sequencing or next-generation sequencing (NGS) [5, 8]. In recent years, whole transcriptome RNA sequencing (RNA-Seq) is increasingly used to identify fusion genes and to assign patients into distinct molecular subgroups according to the WHO 2016 classification, or for protocol-based clinical decisions [9]. Clearly, it would be beneficial if RNA-Seq data could also be used for the identification of IG/TR gene rearrangements pertaining to the leukemic clone. A recent study already showed that RNA-Seq data allowed the identification

Anton W. Langerak (ed.), *Immunogenetics: Methods and Protocols*, Methods in Molecular Biology, vol. 2453, https://doi.org/10.1007/978-1-0716-2115-8_4,

of IG heavy chain (IGH) gene rearrangements in approximately 90% of B-ALL patients [10]. It should however be noted that the majority of ALL rearrangements is unproductive; this is in clear contrast to rearrangements present in normal B cells, which virtually all are functional. Therefore, caution is warranted in the analysis of RNA-Seq data for IG/TR marker screening in ALL (and in other lymphoproliferative disorders requiring multiple RNA/DNA analyses) [11, 12], and applying computational methods that only focus on productive rearrangements (e.g., like for most repertoire analyses) will clearly result in incomplete interpretation of IG/TR data for marker identification [13].

In this chapter, we describe how RNA-Seq data can be obtained and subsequently evaluated using the ARResT/Interrogate immunoprofiling platform [arrest.tools/interrogate] to identify possible IG/TR markers. Similar strategies can likely be employed for other lymphoid malignancies, such as lymphoma and myeloma. Finally, comparable data analysis tools may be used for whole genome sequencing and whole exome sequencing data.

2 Materials

The following equipment, materials, and reagents (or equivalents) should be available:

2.1 RNA-Input Quality Check

1. Agilent 2100 Bioanalyzer (Agilent Technologies).
2. Chip priming station (supplied with the Agilent 2100 Bioanalyzer).
3. IKA vortex mixer (supplied with the Agilent 2100 Bioanalyzer).
4. 16-pin bayonet electrode cartridge (supplied with the Agilent 2100 Bioanalyzer).
5. Agilent RNA 6000 Nano Kit (Agilent Technologies).
6. Microcentrifuge ($\geq 1300 \times g$).

2.2 Preparation of RNA-Seq Library

1. Illumina TruSeq® Stranded mRNA Library Prep Kit, 96 Sample.
2. TruSeq® RNA CD Index Plate (96 Indexes, 96 Samples) (Illumina).
3. 96-well storage plates, round well, 0.8 ml ("midi" plate) (Thermo Fisher Scientific).
4. Agencourt AMPure XP 60 ml kit (Beckman Coulter Genomics).
5. Agilent DNA 1000 Kit (Agilent Technologies).

6. Ethanol 200 proof (absolute) for molecular biology (500 ml) (Sigma-Aldrich).
7. Microseal "B" adhesive seals (Bio-Rad).
8. Nuclease-free ultrapure water.
9. RNaseZap (to decontaminate surfaces).
10. RNase-/DNase-free 8-tube strips and caps.
11. RNase-/DNase-free multichannel reagent reservoirs, disposable (VWR).
12. SuperScript II Reverse Transcriptase (1 vial per 48 reactions) (Thermo Fisher Scientific).
13. Tris–HCl 10 mM, pH 8.5.
14. Tween 20 (Sigma-Aldrich).
15. 96-well thermal cycler (with programmable heated lid).
16. Magnetic stand-96 (Thermo Fisher Scientific).
17. Microplate centrifuge.
18. Vortex.

The following supplies are specifically required for the "HS" workflow described in this protocol (*see* **Note 1**).

1. 96-well Hard-Shell 0.3 ml PCR plate (Bio-Rad).
2. Microseal "A" film (Bio-Rad).
3. High-Speed Microplate Shaker (VWR).
4. Midi plate insert for heating system (Illumina) (two inserts recommended for successive heating procedures).
5. Stroboscope.
6. SciGene TruTemp Heating System (Illumina) (115 V) or SC-60-504 (220 V).
7. Hybex Microsample Incubator (SciGene) 1057-30-0 (115 V) or 1057-30-2 (230 V) (two systems recommended for successive heating procedures).

3 Methods

Be careful when handling RNA samples (*see* **Note 2**).

Here we describe the workflow for RNA-Seq using the Illumina® TruSeq Stranded mRNA Kit (Illumina® Document # 1000000040498 v00) Library Kit chemistries and workflows [https://support.illumina.com//sequencing/sequencing_kits/truseq-stranded-mrna/documentation.html]. Before you proceed, please check carefully for any changes issued by the manufacturer regarding the kit or protocol (*see* **Note 3**).

3.1 RNA Isolation and Quality Assessment

Input RNA quality and quantity is essential to transcriptome sequencing. Illumina True Seq mRNA library Kit requires 0.1–1 μg total RNA as input (*see* **Note 4**). To assess the RNA quality, use the Agilent RNA 6000 Nano Kit:

1. Ensure a correct setup of the chip priming station (*see* **Note 5**).
2. The following kit components need to be prepared before first time use according to the manufacturer's instructions:
   (a) RNA ladder aliquots, stable at −70 °C for extended time periods.
   (b) Agilent RNA 6000 Nano gel matrix aliquots (65 μl), can be stored at 4 °C for 1 month (protect from light during use).
3. To prepare the gel-dye matrix, allow one aliquot Agilent RNA 6000 Nano gel matrix (65 μl) and RNA 6000 Nano dye concentrate (blue cap) to come to room temperature for 30 min.
4. Vortex the dye concentrate for 10 s and spin down.
5. Pipette 1 μl of dye concentrate to 65 μl of gel matrix (one aliquot) in a microcentrifuge tube.
6. Vortex thoroughly and check for proper mixing of gel and dye.
7. Spin at room temperature for 10 min at 13,000 × *g*, protect from light, and use within 1 day. Store at 4 °C if not used immediately.
8. To load the gel-dye matrix, check proper setup of the chip priming station, and place an unused RNA 6000 Nano chip on the chip priming station.
9. Pipette 9.0 μl of the gel-dye mix to the bottom of the well "G" with black background.
10. Position the plunger of the syringe in the chip priming station to 1 ml; then close the chip priming station and pressurize quickly by pressing down the plunger. Keep for exactly 30 s.
11. Release the plunger and pull back gently to the 1 ml mark.
12. Open the chip priming station and pipette 9.0 μl gel-dye matrix to the bottom of the two wells marked "G" without background.
13. To load ladder and sample to the chip, pipette 5 μl of RNA marker (green cap) to all sample wells and the well that is marked with the ladder symbol.
14. Add 1 μl of ladder (prepared in **step 2**, Subheading 3.1) to the well that is marked with the ladder symbol.
15. Add 1 μl of RNA of interest to all sample wells (*see* **Note 6**). If there are less than 12 samples to measure, put 1 μl of RNA marker (green cap) to wells that are not used.

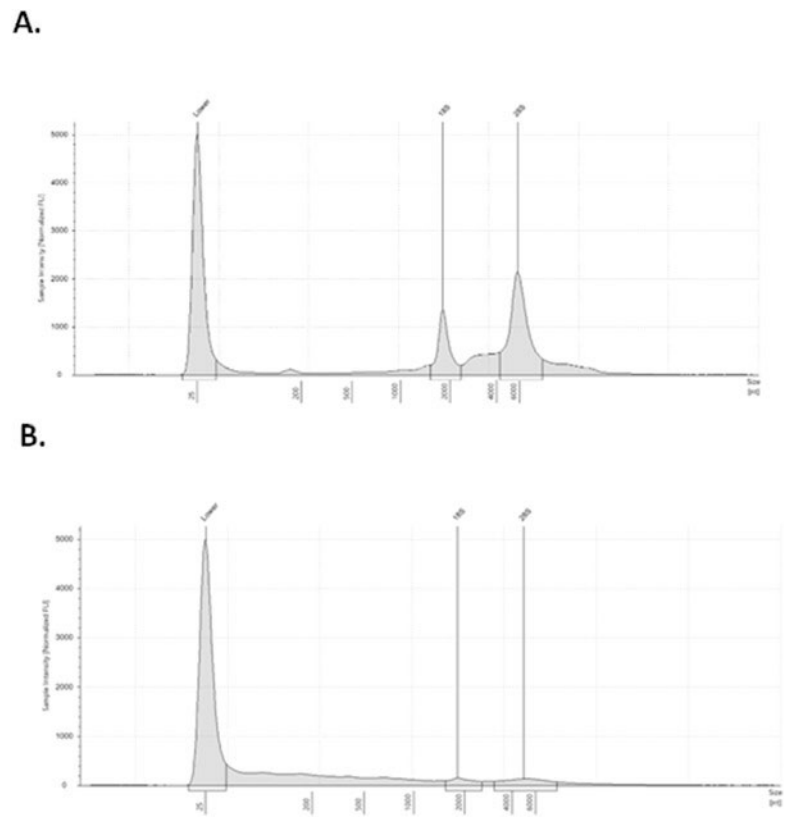


Fig. 1 RNA quality assessment by Bioanalyzer. RNA from bone marrow samples of patients with first diagnosis of ALL was isolated by silica columns (Qiagen, AllPrep) and subjected to microcapillary electrophoresis on the Agilent 2100 Bioanalyzer using the Agilent RNA 6000 Nano Kit as described. (**a**) Electropherogram representing a high-quality RNA sample (RIN 9.4). (**b**) Electropherogram representing a low-quality RNA sample with ongoing RNA degradation (RIN 4.2)

16. Vortex the chip horizontally in the IKA vortex (1 min, 2400 rpm), and insert to the Bioanalyzer 2100 station within 5 min to perform analysis using 2100 Expert Software.

Analysis will deliver a microcapillary electropherogram together with the RNA concentration measured and the RNA integrity number (RIN). The presence of a marker peak and two ribosomal RNA peaks (18S and 28S) will indicate a successful measurement of RNA with at least intermediate quality. Integrity of RNA is quantified on a scale from 1 (poor) to 10 (best) by RIN, based on a proprietary algorithm developed by Agilent © [14]. Figure 1 shows a good and a poor RIN example. In the poor RIN example, the ribosomal peaks are hardly detectable and RNA degradation is observed as a smear of RNA with decreasing size. Illumina True Seq protocols recommend a RIN of 8.0 or higher to be used for library preps (*see* **Note 7**).

3.2 Library Preparation

The following protocol will describe how to obtain a sequencing-ready RNA-Seq library using the Illumina TruSeq® Stranded mRNA library prep kit (October 2017) [https://support.illumina.com//sequencing/sequencing_kits/truseq-stranded-mrna/documentation.html]. Sequencing kit chemistries and protocols are subject to change by the manufacturer. Before beginning, please check with the current version of the Kit and protocol you are using.

The original protocol includes two workflow options based on the number of samples to be processed. We are here describing the "HS" option for >48 samples (*see* **Note 8**).

3.2.1 Purification and Fragmentation of mRNA

In this step, polyadenylated mRNA is pulled down from the total RNA sample using oligo dT-attached magnetic beads (*see* **Note 9**). The isolate is then purified and fragmented.

1. Bring 0.1–1.0 μg of total RNA to 50 μl volume using nuclease-free ultrapure water and transfer to the individual wells of the RNA bead plate.
2. Vortex RNA purification beads and transfer 50 μl to each well.
3. Seal the plate with Microseal "B" and shake plate (1 min, 1000 rpm).
4. Heat the plate in the microheating system (5 min, 65 °C, lid closed), and cool down on ice (1 min).
5. Incubate the plate at room temperature (5 min). Meanwhile bring the microheating system to 80 °C.
6. For magnetic pull-down, remove the seal and put the plate on the magnetic stand. Inspect visually until the liquid is clear (ca. 5 min).
7. Remove the supernatant from all wells. Then remove the plate from the stand.
8. Pipette 200 μl *Bead Washing* Buffer to all wells; seal the plate and mix by shaking (1 min, 1000 rpm).
9. Repeat **steps 6** and **7** of Subheading 3.2.1.
10. Pipette 50 μl *Elution* Buffer to all wells; seal the plate and mix by shaking (1 min, 1000 rpm).
11. Heat the plate in the microheating system (2 min, 80 °C, lid closed), and cool down on ice (1 min).
12. Place the plate at the bench and remove the seal.
13. To prepare for RNA fragmentation, pipette 50 μl *Bead Binding* Buffer to all wells; seal the plate and mix by shaking (1 min, 1000 rpm).
14. Incubate the plate at room temperature (5 min).

15. Repeat **steps 6–9** of Subheading 3.2.1 (magnetic pull-down, bead washing, another magnetic pull-down).
16. Pipette 19.5 µl *Fragment, Prime, Finish* Mix to all wells, seal the plate, and mix by shaking (1 min, 1000 rpm).
17. Remove the seal and transfer the samples well by well to the RNA fragmentation plate.
18. Seal the plate and run the following program on the thermocycler with preheated lid:
    (a) 94 °C for 8 min
    (b) Hold at 4 °C.
19. Quickly spin down.

3.2.2 First Strand cDNA Synthesis

Purified RNA fragments are reverse transcribed to first strand cDNA using random hexamer primers.

1. For magnetic pull-down, put the plate on the magnetic stand. Inspect visually until the liquid is clear (ca. 5 min).
2. Remove the seal and transfer 17 µl supernatant well by well to the cDNA plate.
3. Spin down the First *Strand Synthesis Act D* Mix (5 s, 600 × *g*).
4. Mix *SuperScript II reverse transcriptase* and the *First Strand Synthesis Act D* Mix at a 1:10 ratio (1 µl SuperScript II reverse transcriptase plus 9 µl First Strand Synthesis Act D Mix (can be stored for extended time periods at −20 °C).
5. Pipette 8 µl of this mix to each well of the cDNA plate and mix by shaking (20 s, 1600 rpm).
6. Spin down by centrifugation (1 min, 280 × *g*).
7. Run the following program on the thermocycler with the lid preheated to 100 °C:
    (a) 25 °C for 10 min.
    (b) 42 °C for 15 min.
    (c) 70 °C for 15 min.
    (d) Hold at 4 °C.

3.2.3 Second Strand cDNA Synthesis

To maintain strand specificity during cDNA synthesis and to remove the mRNA template, dUTP is replaced by dTTP in second strand cDNA synthesis. Second strand cDNA synthesis results in blunt-end double-stranded cDNA, which can be stored for 1 week (first safe stopping point).

1. Add 5 µl *Resuspension* buffer to each well.
2. Spin down *Second Strand Marking* Mix (5 s, 600 × *g*), and pipette 20 µl to each well, and mix by shaking (20 s, 1600 rpm).

3. Spin down the plate (1 min, 280 × *g*).
4. Place the plate on the thermocycler with lid preheated to 30 °C and run at 16 °C for 60 min. Allow to come to room temperature afterwards.
5. Pipette 90 μl of *AMPure XP* beads to new cDNA Clean-Up Plate, and transfer the content of the cDNA plate well by well to cDNA Clean-Up Plate. Mix by shaking (2 min, 1800 rpm).
6. Incubate at room temperature (15 min), and then spin down (1 min, 280 × g).
7. Put the plate on the magnetic stand. Inspect visually until the liquid is clear (ca. 5 min). Remove 135 μl supernatant from each well.
8. Wash the beads with the plate retained on the magnetic stand by adding 200 μl *ethanol 80%* to all wells (*see* **Note 10**). After 30 s, remove all supernatant from each well (*see* **Note 11**).
9. Repeat **step** 8 of Subheading 3.2.3 and carefully remove all remaining ethanol (using a small-volume pipette). Let the samples air dry on the magnetic stand (15 min).
10. Remove the plate from the stand and pipette 17.5 μl *Resuspension* buffer to all wells. Mix by shaking (2 min, 1800 rpm). Let stand at room temperature (2 min).
11. Spin down (1 min, 280 × *g*).
12. Put the plate on the magnetic stand. Inspect visually until the liquid is clear (ca. 5 min). Transfer 15 μl supernatant to the new plate (adapter ligation plate). This is the first safe stopping point, where the sealed plate can be stored at −20 °C for 1 week.

3.2.4 3′ End Adenylation and Adapter Ligation

During these steps, single adenine nucleotides are added to the 3′ fragment ends to prevent fragment ligation during addition of adapters. Next, indexing adapters are ligated to 3′ fragment ends. These will later hybridize fragments to the flow cell. The ligated fragments can be stored for 1 week (second safe stopping point).

1. Add 2.5 μl *Resuspension* buffer to all wells of the adapter ligation plate and spin down (5 s, 600 × g).
2. Pipette 12.5 μl *A-Tailing* Mix to each well and mix by shaking (2 min, 1800 rpm).
3. Cover plate with Microseal "B" and spin down (1 min, 280 × *g*).
4. Incubate on 37 °C microheating system (30 min), then transfer to 70 °C microheating system (5 min, lid closed), and cool down on ice (1 min).
5. Spin down *TruSeq® RNA CD Index* Plate (1 min, 280 × *g*).

6. Transfer to all wells of the adapter ligation plates in this order:
   (a) 2.5 μl *Resuspension* buffer
   (b) 2.5 μl *Ligation* mix
   (c) 2.5 μl *RNA Adapters* from the Index adapter Plate (to each corresponding well).

   And mix by shaking (2 min, 1800 rpm).
7. Spin down (1 min, 280 × *g*) plate and transfer to the micro-heating system (10 min, 30 °C, lid closed), and then cool down on ice.
8. Spin down the *Stop Ligation* Buffer (5 s, 600 × *g*) and add 5 μl to every well. Mix by shaking (2 min, 1800 rpm).
9. Spin down the plate (1 min, 280 × *g*).
10. Add 42 μl *AMPure XP beads* to every well. Mix by shaking (2 min, 1800 rpm).
11. Incubate at room temperature (15 min), and then spin down the plate (1 min, 280 × *g*).
12. Put the plate on the magnetic stand. Inspect visually until the liquid is clear (ca. 5 min). Remove all supernatant.
13. Wash the beads with the plate retained on the magnetic stand by adding 200 μl *ethanol 80%* to all wells. After 30 s remove ethanol.
14. Repeat **step 13** of Subheading 3.2.4 and carefully remove all remaining ethanol (using a low-volume pipette). Let samples air dry on the magnetic stand (15 min).
15. Remove plate form stand and pipette 52.5 μl of *Resuspension buffer* to every well, and mix by shaking (2 min, 1800 rpm).
16. Incubate at room temperature (2 min).
17. Spin down the plate (1 min, 280 × *g*) and place it on the magnetic stand. Inspect visually until the liquid is clear (ca. 5 Min).
18. Transfer 50 μl supernatant well by well to a novel plate (Clean Up ALP Plate).
19. Repeat **steps 10–17** of Subheading 3.2.4, but use 50 μl *AMPure XP* beads in **step 10** and 22.5 μl Resuspension buffer in **step 15**.
20. Transfer 20 μl supernatant well by well to a novel plate (PCR plate). Sealed plate can be kept at −20 °C for 1 week (this is the second safe stopping point).

3.2.5 DNA Fragment Enrichment

PCR is used to amplify the library and to select for DNA fragments with successful adapter ligation.

1. On ice, pipette 5 μl *PCR Primer Cocktail* and 25 μl PCR *Master Mix* to each well, mix by shaking (20 s, 1600 rpm), and spin down (1 min, 280 × *g*).
2. Transfer to thermocycler with the preheated lid set to 100 °C and run the program:
   (a) 98 °C for 30 s.
   (b) 15 cycles of:
      - 98 °C for 10 s
      - 60 °C for 30 s
      - 72 °C for 30 s.
   (c) 72 °C for 5 min.
   (d) Hold at 4 °C.
3. Spin down plate (1 min, 280 × *g*) and pipette 47.5 μl *AMPure XP* beads to every well. Mix by shaking (2 min, 1800 rpm).
4. Incubate at room temperature (15 min), and then spin down (1 min, 280 × *g*).
5. Put the plate on the magnetic stand. Inspect visually until the liquid is clear (ca. 5 min). Remove all supernatant.
6. Wash the beads with the plate retained on the magnetic stand by adding 200 μl *ethanol 80%* to all wells. After 30 s, remove ethanol.
7. Repeat **step 6** of Subheading 3.2.5 and carefully remove all remaining ethanol (using low-volume pipette). Let the samples air-dry on magnetic stand (15 min).
8. Resuspend the beads in 32.5 μl *Resuspension* buffer. Mix by shaking (2 min, 1800 rpm).
9. Incubate at room temperature (2 min), and then spin down (1 min, 280 × g).
10. Put plate on magnetic stand. Inspect visually until the liquid is clear (ca. 5 min), and transfer 30 μl supernatant to a novel plate (Target sample plate 1). Libraries can be kept at −20 °C for 1 week (this the third safe stopping point).

3.2.6 Library Quality Check, Normalization, and Pooling

Library quantity and fragment size are determined using the Bioanalyzer. Indexed libraries are pooled prior to sequencing (*see* **Note 12**).

1. Quantify library concentration by qPCR as outlined in manufacturers protocol (lluminaSequencing Library qPCR Quantification Guide (document # 11322363) [https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/qpcr/sequencing-library-qpcr-quantification-guide-11322363-c.pdf].

2. Run 1 μl of each sample on the Agilent Bioanalyzer 2100 using the DNA 1000 chip. The principle follows the outline given for RNA quality control (Subheading 3.1). Average fragment size of a typical library will be ca. 260 bp.
3. Pipette 10 μl library well by well to a novel plate (diluted cluster template plate), and adjust the concentration of the libraries to 10 nM using Tris–HCl 10 mM, pH 8.5 with 0.1% Tween 20. Mix by shaking (2 min, 1000 rpm) and spin down (1 min, 280 × *g*).
4. Pool the libraries according to the system guide of the Illumina sequencing platform that is used. Pooling and normalization should be performed immediately before sequencing to avoid index hopping.

3.2.7 Sequencing

Sequencing of the libraries is performed on an Illumina® sequencing system. Multiple configurations exist which provide the targeted output of 30 million reads per sample (e.g., Illumina's NextSeq 550, 1000 & 2000 and NovaSeq 6000 systems). Depending on the selected sequencing system, flow cell, and read length, between 4 and 132 samples (maximum: two S4 flow cells on NovaSeq 6000 with 2x100 bp sequencing) can be multiplexed in one sequencing run (*see also* **Notes 12** and **13**). For further details, please refer to standard protocol of the manufacturer for the preferred sequencing system.

3.2.8 Raw Data Processing

When sequencing multiple samples on the same sequencing run, outputs have to be demultiplexed to create individual FASTQ files for each respective sample. This is most commonly done using bcl2fastq, a tool developed by Illumina® and pre-installed on most Illumina® sequencers (https://support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software.html). Note that run outputs usually include already demultiplexed reads using default bcl2fastq options (mainly allowing one mismatch in the barcode sequence). Alternatively, bcl2fastq can be run on a Linux system using different (usually more stringent, to avoid in silico contamination) options.

3.2.9 Quality Control

FASTQ files include quality information per sequence base, which can be visualized either using the Illumina graphical software ("Sequencing Analysis Viewer") that is pre-installed on the sequencing machines or by external software that takes the demultiplexed FASTQ files as input. The most commonly used are FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), which outputs an HTML file per sample, and multiqc, which processes multiple FastQC-outputs into a single HTML file to analyze batches [15]. FastQC was developed for DNA applications and might output QC failures for some parameters when used

for RNA-Seq reads. Nonetheless, it can be used to quickly and easily check per base sequencing quality, the number of input reads, and adapter content in the reads.

3.2.10 Adapter Trimming

When read length exceeds DNA insert size, a run can sequence beyond the DNA insert and read bases from the sequencing adapter. To prevent these bases from appearing in FASTQ files, the adapter sequence is trimmed from the 3′ ends of reads. Trimming the adapter sequence improves alignment accuracy and performance in Illumina FASTQ generation pipelines [https://support-docs.illumina.com/SHARE/AdapterSeq/DNAandRNACDIndexes.html].

3.3 Bioinformatic Analysis of the Sequencing Data Using ARResT/Interrogate

We use the ARRest/Interrogate immunoprofiling platform for data analysis, which has been developed and validated within EuroClonality-NGS [16–18].

1. ARResT/Interrogate can be accessed at http://arrest.tools/interrogate and requires an account that can be created by emailing contact@arrest.tools (*see* **Note 14**).
2. Once logged in, select the "Interrogate.WholeMark" user mode at the top-left of the user interface.
3. Switch to the "processing" panel, and follow the instructions to upload the samples in compressed FASTQ format (the extension should be ".fastq.gz") (*see* **Note 15**).
4. Click on the blue "test it" button. If the test was OK, one should be able to click on the green "process" button. If not, check the "process output" tab for feedback; email contact@arrest.tools if necessary.
5. When the run is complete (*see* **Note 16**), follow the instructions on the user interface to retrieve the result files.
6. The main result file is a table in both tab-delimited and Microsoft Excel formats.
7. The table contains information on the analyzed samples and on the reported rearrangements. The "usable" column refers to the sum of fragments with rearrangements and should be in the 1000s. Other columns currently include the rearranged genes and their approximate genomic coordinates; the junction class (rearrangement type), amino acid and nucleotide sequence, and its segmentation; the absolute and relative abundances, in fragments and in reads; technical comments; and the consensus nucleotide sequence of the rearrangement. Column descriptions are provided.
8. The sorting and filtering to obtain a meaningful set of rearrangements depends on the context (*see* **Notes 17** and **18**). We will be releasing further guidelines for interpretation over time;

please stay updated via ARResT/Interrogate [arrest.tools/interrogate] and EuroClonality-NGS [euroclonalityngs.org] (*see* **Note 19**).

4 Notes

1. Alternatively, an "LS" workflow (for <48 samples) is also described in the manufacturer's protocol which requires no additional technical equipment.
2. RNA degrades at room temperature. Keep RNA on ice at all times. On warm days, ensure that the ice does not melt. Always wear gloves. Use RNAse-free materials and reagents. Clean surfaces of bench and flow cabinets with RNaseZap before use.
3. If you prepare an RNA-Seq library for the first time, we strongly advise to take note of the manufacturer's protocol, which includes further systematic considerations.
4. We have successfully performed transcriptome sequencing using RNA isolated by silica columns (Qiagen©), guanidinium thiocyanate-phenol-chloroform extraction (TRIzol©), or automated RNA-purification (Maxwell ©).
5. Note that syringes have to be replaced with each reagent kit. Make sure that the electrode cartridge is inserted to the Bioanalyzer and that the vortex mixer is adjusted to 2400 rpm. It is recommended to follow a daily electrode decontamination to avoid decomposition of the RNA samples.
6. Heat samples (70 °C, 2 min) before loading to minimize secondary structure.
7. Illumina® TruSeq Stranded mRNA Kit protocol requires a RIN of 8.0 or higher. RNA from primary patient samples is a precious and often irreplaceable material. Therefore, we have also used samples with RINs between 6.0 and 8.0 to allocate ALL samples to molecular disease subtypes based on gene expression profiling and calling of driver gene fusions. However, the frequency of samples which could not be allocated to a specific leukemia driver subtype increased substantially with decreasing RIN. While still more than half of the samples could be classified with a RIN between 6.00 and 8.00, this rate was markedly lower in a small sample set with RINs <6.0. This should be kept in mind when processing samples with low RINs. However, PCR-based approaches might still be successfully performed in samples with reduced RNA quality.
8. Safe stopping points are indicated throughout the protocol. Before proceeding to the next step, carefully read the protocol, and make sure that you have all reagents required brought to

the indicated temperatures, that heating systems are preheated to the temperatures required, and that thermal cyclers have been programmed according to the given programs.

9. Magnetic bead wash steps are critical for nucleic acid and sequencing quality. Make sure to vortex the beads thoroughly before use and then pipette immediately, to avoid settling of the beads within the container. It might be helpful to prepare aliquots in smaller tubes and vortex more often. When using multistep pipettes with beads, choose smaller volumes, and pipette quickly to avoid settling of the beads (concentration gradient) within the pipette tip.
10. Ethanol is hygroscopic (attracts and holds water from environment). Fresh aliquots of absolute alcohol should be used for every library preparation and dilutions, for example, 80% ethanol in water should be made on the day of use to ensure correct concentrations. Incorrect ethanol concentrations can decrease yield of DNA/RNA.
11. Remnants of wash buffer (typically including alcohol) can disrupt further library preparation steps. On the other hand, overdried beads can crack and destroy the bound nucleic acids. If unsure whether beads are dry, fanning and smelling the tubes can be an indicator for remaining ethanol.
12. We use RNA-Seq to simultaneously analyze IG/TR gene rearrangements in ALL samples and to perform molecular subgroup allocation based on gene expression profiles and driver gene fusions. For this approach, we have successfully used 75 bp or 100 bp paired end sequencing on Illumina HiSeq2000, NextSeq, and NovaSeq systems, aiming for 30 million reads per sample on all sequencers. Currently, we pool 66 samples on a NovaSeq S4 flow cell for 100 bp paired end sequencing, which yields on average 30–40 million reads per sample.
13. Sequencing systems offer different read lengths for RNA-Seq, typically ranging from 50 bp to 150 bp. Longer reads (100–150 bp) provide a better coverage of transcripts and splice sites, while shorter reads (50–75 bp) are typically used when gene expression profiling is the only intended use. In a small patient subset, molecular subgroup allocation and driver fusion calling was equally effective when the sequencing read length was 75 bp, compared to our standard of 100 bp. Also based on general considerations, we would recommend a read length of 100 bp or more for marker identification from transcriptome data sets. We typically aim for a sequencing depth of 30 million reads, which is suitable for gene expression profiling, fusion calling, and marker identification. Less sequencing depth (5–25 million reads) is required for gene

expression profiling only. Increasing sequencing depth to >30 million reads will improve the detection of subclonal and less covered markers.

14. Details for accessing the bioinformatics pipeline and its results may change. In any case, the latest information will always be available either through the user interface of ARResT/Interrogate at http://arrest.tools/interrogate or through the authors.
15. Dependent on the number of samples and quality of internet connection, uploading data can take time.
16. Progress of the bioinformatic pipeline can be followed in the "process output" tab. The user does not have to wait; one may even close the browser and either log in later or better make sure to provide an email address to receive email notifications.
17. In ALL patients, many rearrangements are incomplete or nonproductive, and therefore it is crucial to include such rearrangements in the analysis and not to filter on productive rearrangements only. In mature lymphoid malignancies, by definition, at least one productive IGH rearrangement should be present.
18. Incomplete and nonproductive rearrangements may not be transcribed or only at very low levels (Fig. 2). It is not yet known to what extent transcription levels of cross-lineage TR

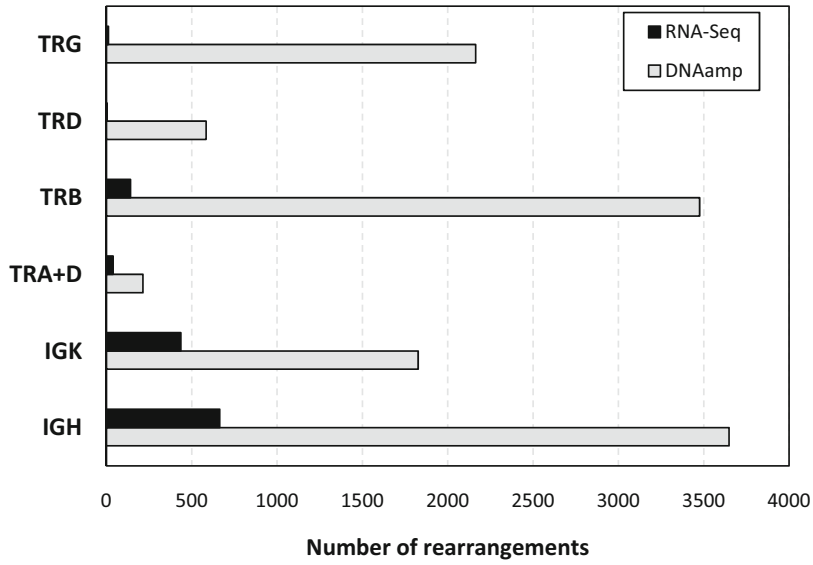


Fig. 2 Comparison between IG/TR rearrangements detected by RNA-Seq and amplicon-based assays ("DNAamp") in 165 ALL patients [13]. Average number of rearrangements detected for the various IG/TR loci per case

gene rearrangements in leukemic rearrangements differ from transcription levels in reactive T cells. Higher read depth may facilitate identification of lowly expressed IG/TR mRNA.

19. Of note, IG/TR rearrangements may also be derived from whole exome sequencing (WES) or whole genome sequencing (WGS) data sets that, in contrast to RNA-Seq data, do not depend on the transcriptional level of rearrangements. This creates a clear advantage as was recently showcased in work introducing IgCaller for WGS-derived IGH data [19].

Acknowledgments

We thank the members of EuroClonality and EuroMRD for their support, especially the participants of the WholeMark work package (Blanca Scheijen, Bastiaan Tops, Jan Trka, Karol Pál, Sonja Hänzelmann, Gianni Cazzaniga, Grazia Fazio, Simona Songia, and Anton W. Langerak).

References

1. Pieters R, de Groot-Kruseman H, Van der Velden V, Fiocco M, van den Berg H, de Bont E et al (2016) Successful therapy reduction and intensification for childhood acute lymphoblastic leukemia based on minimal residual disease monitoring: study ALL10 from the Dutch childhood oncology group. J Clin Oncol 34: 2591–2601
2. Stutterheim J, van der Sluis IM, de Lorenzo P, Alten J, Ancliffe P, Attarbaschi A et al (2021) Clinical implications of minimal residual disease detection in infants with KMT2A-rearranged acute lymphoblastic leukemia treated on the Interfant-06 protocol. J Clin Oncol 39:652–662
3. Flohr T, Schrauder A, Cazzaniga G, Panzer-Grumayer R, van der Velden V, Fischer S et al (2008) Minimal residual disease-directed risk stratification using real-time quantitative PCR analysis of immunoglobulin and T-cell receptor gene rearrangements in the international multicenter trial AIEOP-BFM ALL 2000 for childhood acute lymphoblastic leukemia. Leukemia 22:771–782
4. Van Dongen JJM, van der Velden VHJ, Brüggemann M, Orfao A (2015) Minimal residual disease diagnostics in acute lymphoblastic leukemia: need for sensitive, fast, and standardized technologies. Blood 125: 3996–4009
5. Van der Velden VHJ, Cazzaniga G, Schrauder A, Hancock J, Bader P, Panzer-Grümayer ER et al (2007) Analysis of minimal residual disease by Ig/TCR gene rearrangements: guidelines for interpretation of real-time quantitative PCR data. Leukemia 21: 604–611
6. Van der Velden VHJ, Panzer-Grümayer ER, Cazzaniga G, Flohr T, Sutton R, Schrauder A et al (2007) Optimization of PCR-based minimal residual disease diagnostics for childhood acute lymphoblastic leukemia in a multi-center setting. Leukemia 21:706–713
7. van der Velden VHJ, van Dongen JJM (2009) MRD detection in acute lymphoblastic leukemia patients using Ig/TCR gene rearrangements as targets for real-time quantitative PCR. Methods Mol Biol 538:115–150
8. Brüggemann M, Kotrova M, Knecht H, Bartram J, Boudjogrha M, Bystry V et al (2019) Standardized next-generation sequencing of immunoglobulin and T-cell receptor gene recombinations for MRD marker identification in acute lymphoblastic leukaemia; a EuroClonality-NGS validation study. Leukemia 33:2241–2253
9. Bastian L, Schroeder MP, Eckert C, Schlee C, Tanchez JO, Kampf S et al (2019) PAX5 biallelic genomic alterations define a novel subgroup of B-cell precursor acute lymphoblastic leukemia. Leukemia 33:1895–1909
10. Li Z, Jiang N, Lim EH, Chin WHN, Lu Y, Chiew KH et al (2020) Identifying IGH disease clones for MRD monitoring in childhood

B-cell acute lymphoblastic leukemia using RNA-Seq. Leukemia 34:2418–2429

11. Bueno C, Ballerini P, Varela I, Menendez P, Bashford-Rogers R (2020) Shared D-J rearrangements reveal cell of origin of TCF3-ZNF384 and PTPN11 mutations in monozygotic twins with concordant BCP-ALL. Blood 136: 1108–1111
12. Abdo C, Thonier F, Simonin M, Kaltenbach S, Valduga J, Petit A et al (2020) Caution encouraged in next-generation sequencing immunogenetic analyses in acute lymphoblastic leukemia. Blood 136:1105–1107
13. Van der Velden VHJ, Brüggemann M, Cazzaniga G, Scheijen B, Tops B, Trka J et al (2021) Potential and pitfalls of whole transcriptome-based immunogenetic marker identification in acute lymphoblastic leukemia; a EuroMRD and EuroClonality-NGS working group study. Leukemia 35:924–928
14. Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, Gassmann M et al (2006) The RIN: an RNA integrity number for assigning integrity values to RNA measurements. BMC Mol Biol 7:3
15. Ewels P, Magnusson M, Lundin S, Kaller M (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32:3047–3048
16. Bystry V, Reigl T, Krejci A, Demko M, Hanakova B, Grioni A et al (2017) ARResT/interrogate: an interactive immunoprofiler for IG/TR NGS data. Bioinformatics 33:435–437
17. Knecht H, Reigl T, Kotrova M, Appelt F, Stewart P, Bystry V et al (2019) Quality control and quantification in IG/TR next-generation sequencing marker identification: protocols and bioinformatic functionalities by EuroClonality-NGS. Leukemia 33:2254–2265
18. Scheijen B, Meijers RWJ, Rijntjes J, Van der Klift MY, Mobs M, Steinhilber J et al (2019) Next-generation sequencing of immunoglobulin gene rearrangements for clonality assessment: a technical feasibility study by EuroClonality-NGS. Leukemia 33:2227–2240
19. Nadeu F, Mas-de-Les-Valls R, Navarro A, Royo R, Martin S, Villamor N et al (2020) IgCaller for reconstructing immunoglobulin gene rearrangements and oncogenic translocations from whole-genome sequencing in lymphoid neoplasms. Nat Commun 11:3390

Chapter 5

Minimal Residual Disease Analysis by Monitoring Immunoglobulin and T-Cell Receptor Gene Rearrangements by Quantitative PCR and Droplet Digital PCR

Irene Della Starza, Cornelia Eckert, Daniela Drandi, and Giovanni Cazzaniga and on behalf of the EuroMRD Consortium

Abstract

Analysis of immunoglobulin and T-cell receptor gene rearrangements by real-time quantitative polymerase chain reaction (RQ-PCR) is the gold standard for sensitive and accurate minimal residual disease (MRD) monitoring; it has been extensively standardized and guidelines have been developed within the EuroMRD consortium (www.euromrd.org). However, new generations of PCR-based methods are standing out as potential alternatives to RQ-PCR, such as digital PCR technology (dPCR), the third-generation implementation of conventional PCR, which has the potential to overcome some of the limitations of RQ-PCR such as allowing the absolute quantification of nucleic acid targets without the need for a calibration curve. During the last years, droplet digital PCR (ddPCR) technology has been compared to RQ-PCR in several hematologic malignancies showing its proficiency for MRD analysis. So far, no established guidelines for ddPCR MRD analysis and data interpretation have been defined and its potential is still under investigation. However, a major standardization effort is underway within the EuroMRD consortium (www.euromrd.org) for future application of ddPCR in standard clinical practice.

Key words Minimal residual disease, Immunoglobulin, T-cell receptor, Rearrangement, RQ-PCR, ddPCR

1 Introduction

After a single lymphoid cell undergoes clonal neoplastic transformation, all progeny leukemic cells will contain the same rearranged clonal Immunoglobulin (IG) and T-cell receptor (TR) genes, thus representing highly specific molecular targets for minimal residual disease (MRD) detection in lymphoproliferative disorders [1].

MRD monitoring has been proven to be a compelling tool for advising therapeutic choices especially in acute lymphoblastic leukemia (ALL), the first neoplasm where MRD has been used to assess early response to therapy [2–6]. The availability of drug combinations capable of unprecedented complete clinical responses

Anton W. Langerak (ed.), *Immunogenetics: Methods and Protocols*, Methods in Molecular Biology, vol. 2453, https://doi.org/10.1007/978-1-0716-2115-8_5,

leads to a growing interest for MRD assessment also in other lymphoid malignancies over time, i.e., chronic lymphocytic leukemia, multiple myeloma, as well as mantle cell lymphoma [7–10].

Currently, antigen-receptor gene analysis by real-time quantitative polymerase chain reaction (RQ-PCR) is the gold standard for sensitive and accurate MRD monitoring and has been extensively standardized within the EuroMRD consortium (www.euromrd.org), which established guidelines for the analysis and interpretation of RQ-PCR data [11] to favor a homogeneous application of MRD studies within different lymphoid malignancies and treatment protocols all over the world. However, the measurement of a dynamic process, such as the rate of target amplification, carries some intrinsic fluctuations that cannot be fully eliminated. The digital PCR technology (dPCR) [12], the new generation of conventional PCR, is based on partitioning by nanofluidics and emulsion chemistries which allow performing a limiting dilution of DNA into individual (partitioned) PCR reactions. The DNA template can thus be randomly distributed and the Poisson statistics can be applied to quantify the DNA amount in positive partitions. In comparison with RQ-PCR, dPCR allows the quantification of nucleic acid targets without the need of calibration curves [13]. Moreover, it has the potential to overcome some of the limitations of RQ-PCR. Based on the dynamic nature of these two methods, dPCR appears more accurate than RQ-PCR with a greater amplification efficiency, since each sample is partitioned and each partition is analyzed individually, so small changes in fluorescence intensity are more readily detected [14, 15].

Recently, droplet digital PCR (ddPCR) technology, a type of dPCR characterized by partitioning the sample in droplets, has been applied in comparison to RQ-PCR in several hematologic malignancies, and its additional technical and clinical value to the gold-standard RQ-PCR was demonstrated [16–22]. However, no established guidelines for ddPCR MRD analysis and interpretation have been defined so far, and its potential is still under investigation. A major standardization effort is underway within the ddPCR group of the EuroMRD consortium (www.euromrd.org) for its future application in standard clinical practice.

The PCR approach for IG/TR screening and RQ-PCR MRD analysis have been recently described in this book series, on behalf of the EuroMRD consortium [23]. Briefly, to identify IG/TR markers at diagnosis, either a standard multiplex-PCR/Sanger sequencing [23] or the new and more efficient NGS-based approaches [24, 25] can be applied to define the unique V-(D-)J junctional regions. Complementary patient- and allele-specific oligonucleotide (ASO) primers and common fluorescent probes must be designed for each target of any patient for its MRD monitoring. To perform the MRD relative quantification by RQ-PCR,

amplification conditions and sensitivity testing for each ASO-primer are established on the diagnostic material serially diluted in normal mononuclear cells, before quantifying MRD in bone marrow samples collected during treatment. Interpretation guidelines developed and continuously refined within the EuroMRD group are fundamental for issuing comprehensive clinical reports and for comparing independent studies applying the IG/TR RQ-PCR MRD monitoring [11].

Since the PCR approach for IG/TR screening and RQ-PCR MRD analysis in lymphoproliferative disorders has been recently described [23], in this chapter we will focus on the ddPCR protocol.

2 Materials

1. Supermix for probes (no dUTP).
2. Albumin gene primers and probe.
3. *HINFI (optional) (see below and notes)*.
4. Target gene primers and probe.
5. Thermal cycler.
6. ddPCR droplet generator.
7. PCR plate sealer.
8. ddPCR droplet reader.
9. PC and software for analysis of ddPCR data (Quantasoft, Bio-Rad Laboratories, Hercules, CA, USA) (note: not available for Mac).
10. DG8 cartridges, DG8 gaskets, ddPCR droplet generation oil for probes, ddPCR 96-well plates, pierceable foil heat seals, ddPCR droplet reader oil (Bio-Rad Laboratories).
11. QX100 or QX200 System.
12. 96-well PCR plates and optical adhesive films or 0.2 ml strip tubes with cups (used for mix preparation and collection before droplets generation).
13. Oligo Analyzer 3.1 (www.eu.idtdna.com).
14. PrimerQuest (Integrated DNA Technologies, www.idtdna.com).
15. Primer3Plus (www.primer3plus.com).

Along this chapter, the Bio-Rad system (Bio-Rad Laboratories) is described. Alternative instruments will require adaptation of this protocol.

3 Methods

To identify IG/TR markers at diagnosis, either a standard multiplex PCR/Sanger sequencing [23] or the new and more efficient NGS-based approaches [24, 25] can be applied to define the unique V-(D-)J junctional regions. Complementary ASO primers and common fluorescent probes must be designed for each target of each patient, for MRD monitoring [23]. Several tools are available for assay design and optimization, such as Oligo Analyzer 3.1 (www.eu.idtdna.com), PrimerQuest (Integrated DNA Technologies, www.idtdna.com), Primer3Plus (www.primer3plus.com), or others.

3.1 ddPCR MRD Quantification for the Target Genes

No standard curve generation is needed for a ddPCR experiment setup. However, as for any kind of PCR experiments, a positive control is mandatory (i.e., either a 10-1 dilution or 10-4 dilution point performed in 2-wells could be used). Follow-up samples must be tested in triplicate (two replicates are acceptable only in cases with insufficient DNA or failed technical criteria in third replicate).

To check for unspecific amplifications, nonspecific DNA controls (PB-MNC) should be run in 3 or 6 replicates (*see* **Note 1**) and a no template control (NTC) at least in duplicate, for each specific target quantification, respectively.

The specific oligonucleotide primers and probe, as selected based on available IG/TR targets and sensitivity testing, must be used (*see* **Note 2**).

1. Prepare the reaction mixture for each sample/well as follows:

2× ddPCR Supermix for Probes (no dUTP)	11.0 µl
20× target primers/probe mix	1.1 µl
(HINFI (2 U/µl) - optional)	*(1.1 µl)*
H_2O	4.4 µl
(volume must be modified if the enzyme is used)	
Total volume	**16.5 µl** (*see* **Note 3**)

2. Dispense the mix in the plate or in 0.2 ml strip tubes.
3. Add 5.5 µl of the DNA (100 ng/µl) sample for each well.
4. Seal the plate or the strips with optical adhesive film or caps, mix and spin down briefly.
5. Proceed with droplets generation (*see* **Note 4**).
   (a) Load 20 µl of reaction mix and 70 µl of droplet generation oil into the proper DG8 cartridge wells.

(b) Carefully remove any bubble created into the DG8 cartridge "sample" well during sample loading.

(c) Put the DG8 gasket and start the droplets generation.

6. Carefully transfer 40 μl of generated droplets into a ddPCR 96-well plate.
7. Seal with a pierceable foil on PX1 PCR plate sealer.
8. Start the amplification using the default Bio-Rad thermal cycling protocol (95 °C, 10 min; 94 °C, 30 s; proper Tm°C, 1 min for 40 cycles; 98 °C for 10 min) adjusting the Tm according to the ASO-primers annealing temperature.
9. Load the post-PCR 96-wells plate into the QX100/QX200 droplet reader and follow the manufacturer's instructions. Figure 1 shows a schematic diagram of a ddPCR experiment.

3.2 DNA Quantification Using the Reference Gene

A reference gene must be tested to correct the MRD value in the actual follow-up sample based on the quantity of DNA loaded. Although no consensus has been reached on reference gene usage, the albumin gene is the most frequently used housekeeping control gene. Details on primers and probe concentrations to amplify a portion of the albumin gene as a reference are indicated in **Note 2**. The reference gene is recommended to be tested (in a single well) in the same ddPCR plate as for the target gene.

1. Prepare the reaction mixture for each sample/well as follows:

2× ddPCR Supermix for Probes (no dUTP)	11.0 μl
20× target primers/probe mix	1.1 μl
(HINFI (2 U/μl) -optional)	[*1.1 μl*]
H2O	8.8 μl
(volume must be modified if the enzyme is used)	
Total volume	**20.9 μl** (*see* **Note 3**)

2. Dispense the mix in the plate or 0.2 ml strip tubes.
3. Add 1.1 μl of the DNA (100 ng/μl) sample in each well.
4. Seal the plate or the strips with optical adhesive film or caps, mix, and spin down briefly.
5. Proceed with droplets generation (*see* **Note 4**).

(a) Load 20 μl of reaction mix and 70 μl of droplet generation oil into the proper DG8 cartridge wells.

(b) Carefully remove any bubble created into the DG8 cartridge "sample" well during sample loading.

(c) Put the DG8 gasket and start the droplets generation.

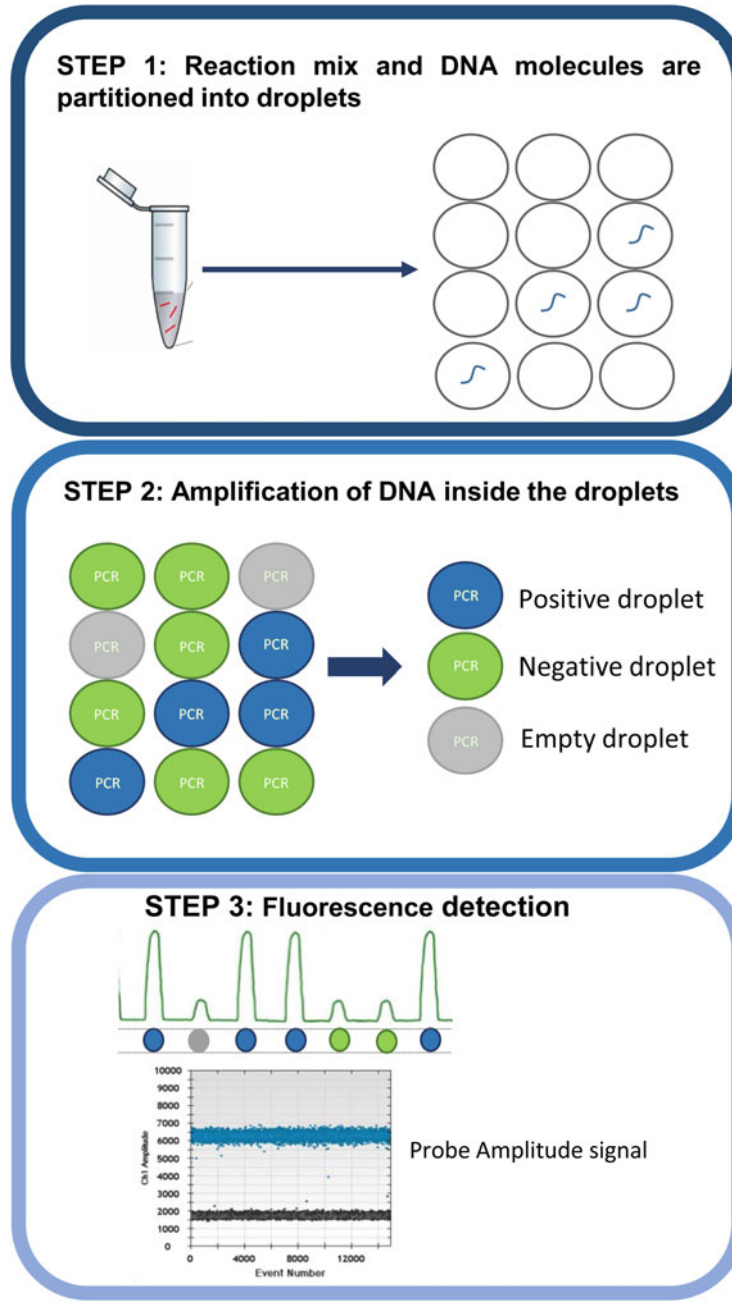


Fig. 1 ddPCR MRD quantification: schematic diagram of a ddPCR experiment. **Step 1**: the reaction mix is prepared with the same primer/probes as for the RQ-PCR assay. Both the reaction and the DNA samples are partitioned into 20,000 droplets of identical volume through a microfluidic system. **Step 2**: in a thermal cycler, 20,000 PCR reactions are amplified and fluorescence is the output during the reaction of polymerization. **Step 3**: a droplet reader analyzes each droplet individually and detects an increased fluorescence in positive droplets, which contain at least one copy of the target DNA

6. Carefully transfer 40 μl of generated droplets in ddPCR 96-well plates.
7. Seal with a pierce able foil on PX1 PCR plate sealer.
8. Start the amplification using the default Bio-Rad thermal cycling protocol (95 °C, 10 min; 94 °C, 30 s; annealing 60 °C, 1 min for 40 cycles; 98 °C for 10 min).
9. Load the post-PCR 96-wells plate in the QX100/QX200 droplet reader and follow the manufacturer's instructions.
10. At the end of the ddPCR reaction, analyze the plots. Check that no amplification is seen in the NTC and exclude samples with very low or high values, outside the range of 300–7500 copies/μl (corresponding theoretically to 20–500 ng) [21].

3.3 ddPCR Results Analysis

The analysis must be performed by QuantaSoft or QuantaSoft PRO according to the following criteria:

1. Only replicates with equal or more than 9000 droplets and equal or less than 20,000 must be considered for the analysis.
2. The threshold must be established manually. The threshold should be settled below the positive control cloud and as close as possible to the background signal (*see* **Note 5**).
3. Set a single threshold, based on the patient specific positive control, for all those samples that use the same set of ASO-primers. However, for those samples presenting unaligned amplitude signal (due to different background amplification signal related to the follow-up DNA quality), a sample specific threshold must be set.
4. In case of few positive events in follow-up samples, NTC or PB-MNC wells, verify the consistency of the amplification signal by checking for the presence of positive droplets in channel 2 (ch2). If a signal in ch2 is detected, in the same position of the ch1 signal, this represents an unspecific amplification (false-positive signal) and must be excluded from the analysis (Fig. 2).

3.4 Interpretation of ddPCR MRD Results

An excel sheet can be used to report all IG/TR target amplification values for all follow-up samples. In the process of setting an international standardization, ddPCR results have been interpreted so far with different guidelines [21, 22]. See Table 1 for the provisional EuroMRD guidelines.

Interpretations must be incorporated into the clinical report. Although it has not been standardized so far, just as for RQ-PCR, a clinical report ideally should contain the following information for each follow-up sample analyzed: date and type of sampling, the actual MRD value, and the corresponding quantitative limit (QL). If the MRD value is positive but below the quantitative limit, the

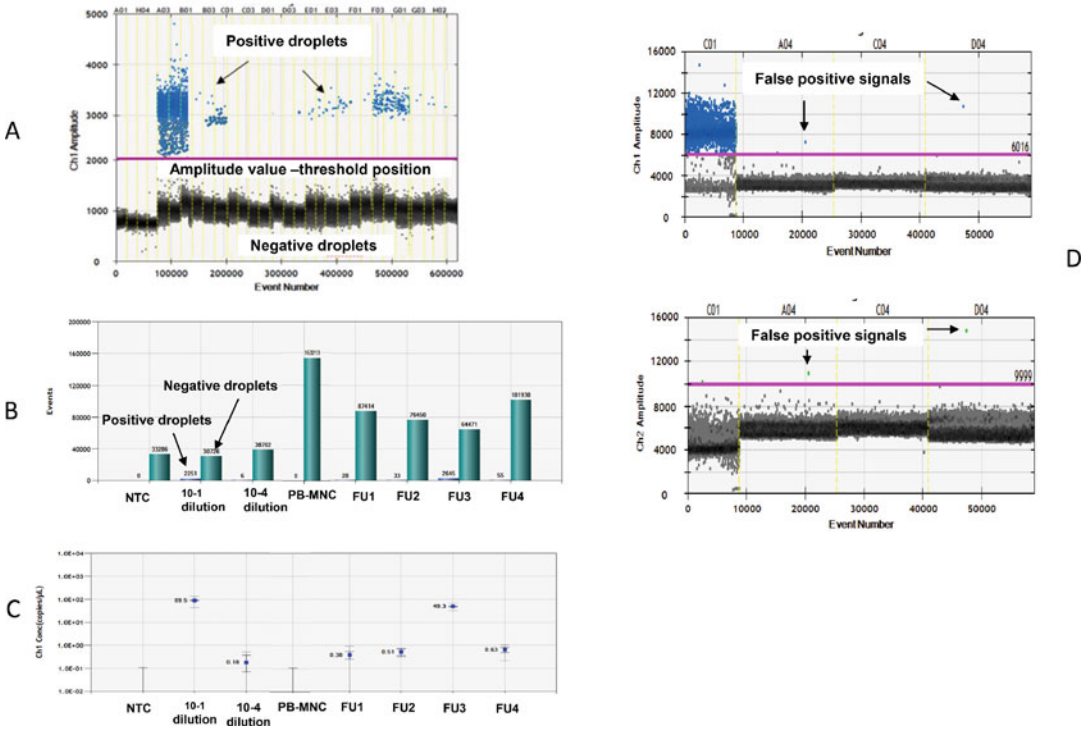


Fig. 2 ddPCR results analysis: each droplet is plotted on the graph of fluorescence intensity versus droplet number (**a**). The concentration is calculated on the fraction of empty droplets (green bar), which is the fraction that does not contain any target DNA (**b**). Fraction of positive droplets is fitted to a Poisson algorithm to determine the absolute copy number, and results are presented in copies per µL (**c**). In case of few positive events in follow-up samples, NTC or PB-MNC wells, verify the consistency of the amplification signal by checking for the presence of positive droplets in channel 2. If a signal in ch2 is detected, in the same position of ch1 signal, this represents an unspecific amplification (false-positive signal) and must be excluded from the analysis (**d**). (*Adapted from Della Starza I,* et al *Front Oncol. 2019 Aug 7;9:726*)

Table 1
Provisional EuroMRD guidelines for ddPCR

Data interpretation	Provisional EuroMRD guidelines [20]
MRD positive sample	A merge of events ≥3, regardless of number of positive replicates After verification for unspecific signals, in the presence of positive background, the calculation has to be: Copies MRD sample - copies background
MRD negative sample	All acceptable replicates with a merge of no or only one event
MRD positive below the quantitative limit	A merge of event = 2 After verification for unspecific signals, in the presence of positive background, the calculation has to be: Copies MRD sample - copies background
MRD quantification	Target copies in 500 ng of DNA are calculated as the mean values of replicates (mean copies/µl × 20 µl)

value can be reported as "POS < QL" (i.e., POS < 1.0×10^{-4}). As already established for RQ-PCR results, this qualitative result cannot be further interpreted: it only means that the sample is positive and lower than the QL, but it cannot be quantified precisely and should not be used for clinical decision, in particular not for upgrading the therapy, because of the intrinsic risk of false-positivity. In case of negative MRD, the actual QL and the specific time point need to be considered for clinical interpretation and decision-making.

3.5 Conclusion

During the last years, many publications have reported on the ddPCR application in different hematological diseases. Its intrinsic characteristics (accuracy, sensitivity, quantification without the need of a standard curve, etc.) make this method also attractive for MRD evaluation. However, at the moment, the use of ddPCR as a MRD molecular method in clinical protocols is prevented by the lack of published international guidelines for data interpretation, which is a fundamental requirement to ensure reproducibility and to compare MRD data in different clinical protocols. For this reason, a major standardization effort is underway within the EuroMRD consortium groups, and five ddPCR QC rounds have so far been performed, involving 24 laboratories around the world [21]. The further challenges will be to achieve this goal and to assess the prognostic relevance of ddPCR in large studies in the light of its future application in clinical practice.

4 Notes

1. Since ddPCR allows accurate quantification of rare events, we suggest to use the same number of replicates (=3) for patients and for PBMNC samples.
2. Primers and probe should be used at a final concentration of 500 nM and 200 nM, respectively. For 25 μl total volume of 20X target primers/probe mix: 1 μl probe (100 μM), 2.5 μl each primers (100 μM), and 19 μl H_2O. For ddPCR analysis, BHQ1 or MGB or Zen probes should be used. TAMRA probes must be avoided since they lead to high background and noise signals.
3. Prepare the ddPCR mix for a volume increased by 10% to be sure to have enough reaction mix volume for each replicate, and not to risk the generation of air bubbles into the DG8 cartridges, when loading the samples.
4. Based on the DNA extraction method, gDNA could be viscous and this characteristic can affect the droplets generation. In case of sticky DNA, add 2 U/μl of enzyme (1.1 μl) to the

ddPCR reaction mix, adjusting properly with water. Importantly, before using the enzyme, verify that target sequences or primers and probes will be not damaged.

5. In case of one or two positive droplets, in the PB-MNC wells, just above the background signal, threshold line could be settled just above these observed droplets of the PB-MNC. In case of positive droplets in the PB-MNC samples at higher amplitude respect to the cloud of positive control, these are unspecific signals and must be omitted from the analysis.

References

1. Cazzaniga G, Biondi A (2005) Molecular monitoring of childhood acute lymphoblastic leukemia using antigen receptor gene rearrangements and quantitative polymerase chain reaction technology. Haematologica 90: 382–390
2. Faderl S, O'Brien S, Pui C-H, Stock W, Wetzler M, Hoelzer D et al (2010) Adult acute lymphoblastic leukemia: concepts and strategies. Cancer 116:1165–1176
3. Gökbuget N, Raff R, Brüggemann M, Flohr T, Scheuring U, Pfeifer H et al (2004) Risk/MRD adapted GMALL trials in adult ALL. Ann Hematol 83:S129–S131
4. Conter V, Bartram CR, Valsecchi MG, Schrauder A, Panzer-Grumayer R, Moricke A et al (2010) Molecular response to treatment redefines all prognostic factors in children and adolescents with B-cell precursor acute lymphoblastic leukemia: results in 3184 patients of the AIEOP-BFM ALL 2000 study. Blood 115:3206–3214
5. Vora A, Goulden N, Wade R, Mitchell C, Hancock J, Hough R et al (2013) Treatment reduction for children and young adults with low-risk acute lymphoblastic leukaemia defined by minimal residual disease (UKALL 2003): a randomised controlled trial. Lancet Oncol 14: 199–109
6. Eckert C, Henze G, Seeger K, Hagedorn N, Mann G, Panzer-Grümayer R et al (2013) Use of allogeneic hematopoietic stem-cell transplantation based on minimal residual disease response improves outcomes for children with relapsed acute lymphoblastic leukemia in the intermediate-risk group. J Clin Oncol 31: 2736–2742
7. Pieters R, de Groot-Kruseman H, Van der Velden V, Fiocco M, van den Berg H, de Bont E et al (2016) Successful therapy reduction and intensification for childhood acute lymphoblastic leukemia based on minimal residual disease monitoring: study ALL10 from the Dutch childhood oncology group. J Clin Oncol 34: 2591–2501
8. Del Giudice I, Raponi S, Della Starza I, De Propris MS, Cavalli M, De Novi LA et al (2019) Minimal residual disease in chronic lymphocytic leukemia: a new goal? Front Oncol 9:689
9. Landgren O, Lu SX, Hultcrantz M (2018) MRD testing in multiple myeloma: the Main future driver for modern tailored treatment. Semin Hematol 55(1):44–50
10. Pott C, Brüggemann M, Ritgen M, van der Velden VHJ, van Dongen JJM, Kneba M (2019) MRD detection in B-cell non-Hodgkin lymphomas using Ig gene rearrangements and chromosomal translocations as targets for real-time quantitative PCR. Methods Mol Biol 1956:199–228
11. Van der Velden VH, Cazzaniga G, Schrauder A, Hancock J, Bader P, Panzer-Grumayer ER et al (2007) European study group on MRD detection in ALL (ESG-MRD-ALL). Analysis of minimal residual disease by Ig/TCR gene rearrangements: guidelines for interpretation of real-time quantitative PCR data. Leukemia 21:604–611
12. Huggett JF, Whale A (2013) Digital PCR as a novel technology and its potential implications for molecular diagnostics. Clin Chem 59: 1691–1693
13. Sanders R, Huggett JF, Bushell CA, Cowen S, Scott DJ, Foy CA (2011) Evaluation of digital PCR for absolute DNA quantification. Anal Chem 83:6474–6484
14. Vincent ME, Liu W, Haney EB, Ismagilov RF (2010) Microfluidic stochastic confinement enhances analysis of rare cells by isolating cells and creating high density environments for control of diffusible signals. Chem Soc Rev 39:974–984
15. Stahl T, Böhme MU, Kröger N, Fehse B (2015) Digital PCR to assess hematopoietic

chimerism after allogeneic stem cell transplantation. Exp Hematol 43:462–468

16. Drandi D, Kubiczkova-Besse L, Ferrero S et al (2015) Minimal residual disease detection by droplet digital PCR in multiple myeloma, mantle cell lymphoma, and follicular lymphoma: a comparison with real-time PCR. J Mol Diagn 17(6):652–660
17. Della Starza I, Nunes V, Cavalli M et al (2016) Comparative analysis between RQ-PCR and digital-droplet-PCR of immunoglobulin/T-cell receptor gene rearrangements to monitor minimal residual disease in acute lymphoblastic Leukaemia. Br J Haematol 174(4):541–549
18. Cavalli M, De Novi LA, Della Starza I, Cappelli LV, Nunes V, Pulsoni A et al (2017) Comparative analysis between RQ-PCR and digital droplet PCR of BCL2/IGH gene rearrangement in the peripheral blood and bone marrow of early stage follicular lymphoma. Br J Haematol 177:588–596
19. Coccaro N, Anelli L, Zagaria A, Casieri P, Tota G, Orsini P et al (2018) Droplet digital PCR is a robust tool for monitoring minimal residual disease in adult Philadelphia-positive acute lymphoblastic leukemia. J Mol Diagn 20:474–482
20. Drandi D, Ferrero S, Ladetto. (2018) Droplet digital PCR for minimal residual disease detection in mature lymphoproliferative disorders. Methods Mol Biol 1768:229–256
21. Drandi D, Alcantara M, Benmaad I, Söhlbrandt A, Lhermitte L, Zaccaria G et al (2020) Droplet digital PCR quantification of mantle cell lymphoma follow-up samples from four prospective trials of the European MCL network. Hemasphere 4(2):e347
22. Della Starza I, Nunes V, Lovisa F, Silvestri D, Cavalli M, Garofalo A et al (2021) Droplet digital PCR improves IG−/TR-based MRD risk definition in childhood B-cell precursor acute lymphoblastic leukemia. Hemasphere 5(3):e543
23. Cazzaniga G, Songia S, Biondi A, EuroMRD Working Group (2021) PCR technology to identify minimal residual disease. Methods Mol Biol 2185:77–94
24. Kotrova M, Darzentas N, Pott C, Brüggemann M, EuroClonality-NGS Working Group (2021) Next-generation sequencing technology to identify minimal residual disease in lymphoid malignancies. Methods Mol Biol 2185:95–111
25. Stewart JP, Gazdova J, Darzentas N, Wren D, Proszek P, Fazio G et al (2021) EuroClonality-NGS Working Group. Validation of the EuroClonality-NGS DNA capture panel as an integrated genomic tool for lymphoproliferative disorders. Blood Adv. 5(16):3188–3198. https://doi.org/10.1182/bloodadvances.2020004056.PMID:3442432

Chapter 6

Quality Control for IG/TR Marker Identification and MRD Analysis

Eva Fronkova, Michael Svaton, and Jan Trka

Abstract

Selection of the proper target is crucial for clinically relevant monitoring of minimal residual disease (MRD) in patients with acute lymphoblastic leukemia using the quantitation of clonal-specific immunoreceptor (immunoglobulin/T cell receptor) gene rearrangements. Consequently, correct interpretation of the results of the entire analysis is of utmost importance. Here we present an overview of the quality control measures that need to be implemented into the process of marker identification, selection, and subsequent quantitation of the MRD level.

Key words Minimal residual disease, Acute lymphoblastic leukemia, Quality control, Next-generation sequencing, PCR

1 Introduction

Minimal residual disease (MRD) monitoring became the standard tool for acute lymphoblastic leukemia (ALL) patient risk stratification. Development of the methodology, as started by the leading pediatric international consortia, has led to the wide acceptance of this approach by both pediatric and adult hematologists alike. Among all potentially available strategies for MRD follow-up analysis, detection and subsequent quantitation of immunoreceptor gene (immunoglobulin/T cell receptor; IG/TR) rearrangements have become the gold standard. IG/TR-based MRD monitoring is currently not only used in frontline treatment of ALL patients but also for the prediction of outcome after relapse of ALL and for follow-up analysis of patients before and after hematopoietic stem cell transplantation (SCT).

As really crucial treatment decisions are being made based on the results of MRD measurement, the accuracy of the method is critical. At particular time-points of treatment, both potential false-negative and false-positive results may have serious consequences.

Anton W. Langerak (ed.), *Immunogenetics: Methods and Protocols*, Methods in Molecular Biology, vol. 2453, https://doi.org/10.1007/978-1-0716-2115-8_6,

Therefore, quality controls must be an integral part of this approach throughout all the critical procedures of MRD marker identification, selection, and follow-up analysis.

Here, we summarize the critical steps in marker identification and MRD analysis together with the description of related quality control measures.

2 IG/TR Marker Identification

2.1 PCR-Based Marker Identification

A classical approach of clonal marker identification includes PCR amplification, clonality assessment, and Sanger sequencing of PCR products. The strategy for choosing IG/TR markers for amplification differs based on the type of malignancy. In CLIP laboratories, we prefer to use separate singleplex PCR reactions for ALL (25 for B-ALL, 20 for T-ALL), as described by the BIOMED-1 consortium [1, 2], with frozen premixes including primer pairs and polymerase for each rearrangement, complemented by T cell receptor beta (TRB) detection via three multiplex PCR reactions, as described by the BIOMED-2 consortium [3].

2.1.1 Control Samples

Cell line or patient samples with respective rearrangements are used as positive control, and water is used as negative control to check for possible contamination. Using 20–25 single reactions, it is not possible to add positive and negative controls to each mix. One positive control and one negative control are used for each marker screening, with positive control changing (rotating) for each screening round to control all the PCR premixes.

2.1.2 Distinguishing Between Monoclonal and Polyclonal PCR Products

In case of positive amplification, it is necessary to distinguish monoclonal PCR products from oligo/polyclonal ones. This was previously done using heteroduplex analysis on polyacrylamide gels [3]. Currently, GeneScan analysis or technologies of automated electrophoresis (Agilent Bioanalyzer or similar) are preferred due to significantly reduced hands-on time. We use Agilent Bioanalyzer on-a-chip electrophoresis for clonality detection, because it does not require fluorescently labeled primers as in GeneScan, while providing a similar degree of size distinction. Moreover, PCR products can be directly used for further analysis. The TRB multiplex interpretation is difficult due to possible unspecific bands. Therefore, polyclonal control samples consisting of a mix of at least ten healthy donor "buffy coat" samples should be used for each TRB multiplex tube, together with positive controls and water, to discern nonspecific bands. The monoclonal products are then sequenced and clone-specific primers are designed (see below).

2.2 NGS-Based Marker Identification

Alternatively—and currently more frequently—the methods used for the screening of IG/TR gene rearrangements as clonal markers in ALL are routinely based on next-generation sequencing (NGS), providing a rapid and full overview of the rearrangements present in the sample. These methods, usually based on amplicon sequencing for particular markers (IG/TR rearrangements), rely on multiplex PCR with a large number of specific primers, and thus a reliable and standardized quality control is needed in routine practice to obtain reliable results. When focusing on noncommercial and thus freely available solutions, EuroClonality-NGS assays and approaches that were developed to standardize routine diagnostic practice for both the wet lab and bioinformatic parts of marker identification are optimal [4].

2.2.1 Quality Control of the Library Preparation

To ensure that all possible IG/TR gene rearrangements that are present in the diagnostic DNA sample can be detected, a routine control of the PCR primer mixes should regularly be performed using a polyclonal quality control sample (PC-QC). A mixture of polyclonal DNA samples isolated from the PBMCs obtained from multiple healthy donors is easily accessible in routine laboratory practice and provides a diverse repertoire of IG/TR gene rearrangements. NGS library preparation from the PC-QC is required each time a new working dilution of the primer mix is prepared to test the correct performance of all primers and should be periodically repeated to assess stable primer mix composition over longer periods of time.

Standard quality control (QC) of the NGS library is required for each sequencing run and consists of gel electrophoresis of the final products to assess a good specific amplification of the library at the expected amplicon length and quantitation of the purified specific products.

For the purpose of assessing correct PCR amplification during each NGS library preparation, a central in-tube quality/quantitation control (cIT-QC) is used and added to the PCR reaction to undergo the whole process in parallel with the diagnostic sample. The cIT-QC consists of selected human B and T cell lines with defined IG/TR rearrangements [5] and serves as a positive control for all the IG/TR gene loci, including the ones that were not rearranged in the patient's malignant cells and would otherwise lack specific rearrangements. Reads from the cIT-QC are used during the bioinformatic analysis to confirm correct NGS library preparation and aid with the normalization of all the other reads to cell counts.

2.2.2 Data Analysis

A large number of specifically developed software tools exist for the analysis of IG/TR gene rearrangements, with the ARResT/Interrogate [6] and Vidjil [7] applications being developed in collaboration with the EuroClonality-NGS working group to be well suited

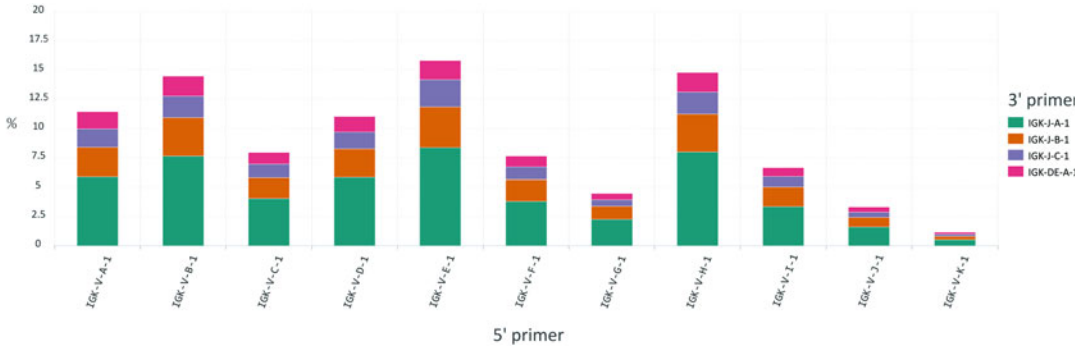


Fig. 1 Primer usage in a mixed polyclonal sample. Individual primers from the EuroClonality-NGS IGK-VJ-Kde primer mix are shown with 5′ primers on the x axis and 3′ primers in different colors. The y axis shows the relative abundance of reads identified with the respective primer sequence in the NGS library

for MRD marker identification including the automatic quality control of the libraries prepared according to the EuroClonality-NGS working group protocols. An essential prerequisite for the analysis is sufficient sequencing coverage of the NGS libraries with good base quality for reliable identification of all IG/TR rearrangements present in the DNA sample, including the cIT-QC. This is taken into consideration during the bioinformatic analysis with these tools.

Using correct primer annotation, the usage of specific 5′ and 3′ primers can be examined in each sample to assess their individual performance. An example of such analysis of IGK-VJ-Kde primer usage in a polyclonal sample is shown (Fig. 1). Although influenced by the gene usage in a healthy polyclonal repertoire, it is a reliable indicator of any errors that may have occurred during the primer mix preparation. Primer mix performance should be checked regularly using a PC-QC.

Reads corresponding to the cIT-QC are identified during the bioinformatic analysis and serve as an amplification control for each individual library. An automatic QC determines that all expected rearrangements of the cIT-QC are present in the respective libraries and a quantitation factor is calculated based on the DNA input of the cIT-QC as well as the patient's sample. A potential failure to detect some of the cIT-QC rearrangements may occur in a situation with a low coverage of the NGS library and a high infiltration of blasts in the patient's sample with monoclonal rearrangement. In such cases usually only some of the cIT-QC rearrangements are not covered and the MRD marker can still be clearly identified. In samples with limited polyclonal IG/TR background, the cIT-QC makes up a large proportion of reads.

2.3 Choosing Markers for MRD and Optimization of the Clonal-Specific RQ-PCR Systems

There have been many debates on the subject of (preferential) selection of the most specific and stable markers. However, in the real-life situation, prioritization of markers is not really an issue; for the sake of time of routine diagnostic throughput, usually all available (mono)clonal markers identified from Sanger sequencing are used for clonal-specific primer design and subsequent RQ-PCR optimization. Sequential testing of potential markers and primers is not preferred as the total time spent on the entire selection-optimization process must fit in the diagnostic window for MRD monitoring. Markers are therefore mostly selected based on their real, rather than predicted performance during the optimization process.

However, with the advent of NGS-based marker identification, more information is available on every marker. First, the real abundance of the clonal marker in the analyzed DNA sample can be estimated based on the cIT-QC and the background, and second, and perhaps most importantly, its specificity can be confirmed against a large dataset of IG/TR rearrangements from other patients and polyclonal samples. Detailed description of this is well beyond the scope of this chapter.

Ultimately, the real performance of the selected clonal marker-primer in RQ-PCR is the criterion for its use in MRD monitoring.

The EuroMRD (former ESG-MRD-ALL) consortium has established strict criteria for defining sensitivity and specificity of RQ-PCR systems [8].

Reaching adequate sensitivity and specificity based on EuroMRD criteria represents a QC of a well-designed RQ-PCR system per se. Similar rearrangements in normal B and T cells are the source of possible false-positivity, and background amplification is unavoidable in some markers. The extent of nonspecific amplification (NSA) depends on the involved genes and the number of inserted and deleted nucleotides in the junction. It has been estimated that NSA occurs in 35% of IGH markers and in more than 90% of TCRG markers [9]. IGK markers are also highly prone to NSA. IGK-KDE rearrangements are recommended as first-choice markers due to their stability, but based on our NGS data, the presence of highly similar rearrangements with resulting NSA is extremely high in polyclonal controls (unpublished data).

Therefore, it is mandatory to use adequate polyclonal controls. At least six wells of polyclonal DNA (preferentially from at least 10 healthy donors PB samples) should be used together with MRD samples in the RQ-PCR assay. Usually, 2–3 specific primers are tested for each monoclonal rearrangement, and two independent markers with the lowest NSA and sufficient sensitivity are selected and further optimized if needed. To reduce NSA, it is possible to slightly correct RQ-PCR conditions, i.e., to increase the annealing temperature by 2–4 °C or to titrate the primer concentration, usually by decreasing the clonal marker-specific primer concentration.

3 Interpretation of RQ-PCR MRD Analysis Results

The EuroMRD consortium has also defined and published guidelines for the correct interpretation of RQ-PCR MRD monitoring results. These criteria not only reflect the potential biological issues of the approach but also the clinical relevance of the result.

Consequently, the criteria for MRD positivity were defined more strictly for situations, where possible false-positivity would lead to unjustified treatment intensification. This is typically the situation of an emerging molecular relapse, most commonly during regular follow-ups after stem cell transplantation (SCT). In the opposite situation, i.e., when treatment reduction would be the outcome of false-negative MRD result, the criteria are intentionally stricter toward negativity [8].

In summary, sample is considered to be MRD positive in the context of therapy reduction (e.g., risk group stratification into lower risk group) if:

- The CT value of at least one of the three replicates is $\geq$1.0 CT lower than the lowest CT of background

  and.
- The CT value of at least one of the three replicates is within 4.0 CT from the highest CT value of the previously defined "sensitivity."

A sample is considered to be MRD positive in the context of therapy intensification (e.g., therapeutic intervention after SCT) if:

- The CT value of at least one of the three replicates is $\geq$3.0 CT lower than the lowest CT of background

  and.
- The CT value of at least one of the three replicates is within 4.0 CT from the highest CT value of the previously defined "sensitivity."

3.1 Identification of False-Positive and False-Negative Results

In an intra-laboratory setting, a newly emerged low MRD positivity remains a diagnostic challenge. Before the era of NGS methods, the extent of false-positivity was assessed only indirectly. Van der Velden et al. retested the low-positive samples in different timepoints of ALL using MRD assays designed for different (irrelevant) markers and concluded that the NSA differs between timepoints and markers and is mostly present in IGH markers with background amplification in PB (buffy coats) in post-maintenance treatment phases. Their study concluded that the background for IGH markers was lowest at the end of induction treatment (day 33) and that EuroMRD criteria sufficiently excluded most of the false-positives [10].

Our group focused on MRD positivity during the post-SCT period. Starting 140 days post-SCT, we frequently observed positive results fulfilling EuroMRD criteria for therapy intensification in patients who turned negative in the following examinations. Using indirect methods, we showed that the positives were nonspecific and their occurrence correlated with intense B cell regeneration, which is usually very intense post-SCT [11]. With the development of NGS-based MRD methods, we expanded the previous cohort and reanalyzed post-SCT RQ-PCR-positive samples by NGS. A vast majority of RQ-PCR positive samples in patients who subsequently did not progress into hematological relapse were negative using NGS. NGS sequences of amplified physiological rearrangements were highly similar to ASO primer sequences, suggesting that RQ-PCR amplification was not specific [12].

Based on these data, we decided to recheck every MRD result post-SCT that was concluded by RQ-PCR to be "positive, nonquantifiable." The size of the nonspecific RQ-PCR products is usually different from the expected size of the amplified marker. Therefore, it is helpful to keep RQ-PCR products and check their size using the Agilent Bioanalyzer together with products of the standard curve dilution (usually 10–1 and 10–4) as size standard and with buffy coats that previously showed positive signals. Based on our experience, up to 30–40% of low-positive (nonquantifiable) RQ-PCR results can be identified as false-positive, because the length of the "clonal-specific" product differs from its original size and overlaps with buffy coat amplification (unpublished data). In the remaining cases, the sizes of all products including buffy coat are in the same size range and thus cannot be distinguished. With IG/TR NGS available, it is possible to reevaluate the remaining positive RQ-PCR result via NGS. However, to ensure that NGS has the same (or better) sensitivity as RQ-PCR, it is crucial to test the sensitivity of NGS using the diluted sample (e.g., 10–4), preferentially in a separate NGS run to avoid sample cross-contamination.

4 Conclusion

MRD monitoring using an IG/TR-based quantitation method is an elegant and clinically relevant approach. However, as several steps are prone to technical and interpretational errors, adequate quality control measures must be included throughout the process. Some of the basic and more advanced tips have been listed in this chapter. On top of these, intra-laboratory procedures and interlaboratory measures can be introduced as well.

Acknowledgments

This work was supported by the Ministry of Health of the Czech Republic, grant NV20-03-00284.

References

1. Pongers-Willemse M, Seriu T, Stolz F, d'Aniello E, Gameiro P, Pisa P et al (1999) Primers and protocols for standardized detection of minimal residual disease in acute lymphoblastic leukemia using immunoglobulin and T cell receptor gene rearrangements and TAL1 deletions as PCR targets report of the BIOMED-1 CONCERTED ACTION: investigation of minimal residual disease in acute leukemia. Leukemia 13:110–118. https://doi.org/10.1038/sj.leu.2401245
2. Szczepański T, Pongers-Willemse MJ, Langerak AW, Harts WA, Wijkhuijs AJ, van Wering ER et al (1999) Ig heavy chain gene rearrangements in T-cell acute lymphoblastic leukemia exhibit predominant DH6-19 and DH7-27 gene usage, can result in complete V-D-J rearrangements, and are rare in T-cell receptor alpha beta lineage. Blood 93:4079–4085
3. Van Dongen JJM, Langerak AW, Brüggemann M, Evans PA, Hummel M, Lavender FL et al (2003) Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98-3936. Leukemia 17:2257–2317. https://doi.org/10.1038/sj.leu.2403202
4. Brüggemann M, Kotrová M, Knecht H, Bartram J, Boudjogrha M, Bystry V et al (2019) Standardized next-generation sequencing of immunoglobulin and T-cell receptor gene recombinations for MRD marker identification in acute lymphoblastic leukaemia; a EuroClonality-NGS validation study. Leukemia 33:2241–2253. https://doi.org/10.1038/s41375-019-0496-7
5. Knecht H, Reigl T, Kotrová M, Appelt F, Stewart P, Bystry V et al (2019) Quality control and quantification in IG/TR next-generation sequencing marker identification: protocols and bioinformatic functionalities by EuroClonality-NGS. Leukemia 33:2254–2265. https://doi.org/10.1038/s41375-019-0499-4
6. Bystry V, Reigl T, Krejci A, Demko M, Hanakova B, Grioni A et al (2017) ARResT/interrogate: an interactive immunoprofiler for IG/TR NGS data. Bioinformatics 33:435–437. https://doi.org/10.1093/bioinformatics/btw634
7. Giraud M, Salson M, Duez M, Villenet C, Quief S, Caillault A et al (2014) Fast multiclonal clusterization of V(D)J recombinations from high-throughput sequencing. BMC Genomics 15:409. https://doi.org/10.1186/1471-2164-15-409
8. Van der Velden VHJ, Cazzaniga G, Schrauder A, Hancock J, Bader P, Panzer-Grumayer ER et al (2007) Analysis of minimal residual disease by Ig/TCR gene rearrangements: guidelines for interpretation of real-time quantitative PCR data. Leukemia 21:604–611. https://doi.org/10.1038/sj.leu.2404586
9. Van der Velden VHJ, Hochhaus A, Cazzaniga G, Zczepanski T, Gabert J, van Dongen JJ (2003) Detection of minimal residual disease in hematologic malignancies by real-time quantitative PCR: principles, approaches, and laboratory aspects. Leukemia 17:1013–1034. https://doi.org/10.1038/sj.leu.2402922
10. Van der Velden VHJ, Wijkhuijs JM, van Dongen JJM (2008) Non-specific amplification of patient-specific Ig/TCR gene rearrangements depends on the time point during therapy: implications for minimal residual disease monitoring. Leukemia 22:641–644. https://doi.org/10.1038/sj.leu.2404925
11. Fronkova E, Muzikova K, Mejstrikova E, Kovac M, Formankova R, Sedlacek P et al (2008) B-cell reconstitution after allogeneic SCT impairs minimal residual disease monitoring in children with ALL. Bone Marrow Transplant 42:187–196. https://doi.org/10.1038/bmt.2008.122

12. Kotrova M, Van der Velden VHJ, van Dongen JJM, Formankova R, Sedlacek P, Brüggemann M et al (2017) Next-generation sequencing indicates false-positive MRD results and better predicts prognosis after SCT in patients with childhood ALL. Bone Marrow Transplant 52(7):962–968. https://doi.org/10.1038/bmt.2017.16

Chapter 7

cfDNA-Based NGS IG Analysis in Lymphoma

Christiane Pott, Michaela Kotrova, Nikos Darzentas, Monika Brüggemann, and Mouhamad Khouja and on behalf of the EuroClonality-NGS Working Group

Abstract

Liquid biopsy is a novel diagnostic approach at first developed to characterize the molecular profile of solid tumors by analyzing body fluids. For cancer patients, it represents a noninvasive way to monitor the status of the solid tumor with respect to representative biomarkers. There is growing interest in the utilization of circulating tumor DNA (ctDNA) analysis also in the diagnostic and prognostic fields of lymphomas. Clonal immunoglobulin (IG) gene rearrangements are fingerprints of the respective lymphoid malignancy and thus are highly suited as specific molecular targets for minimal residual disease (MRD) detection. Tracing of the clonal IG rearrangement patterns in ctDNA pool during treatment can be used for MRD assessment in B-cell lymphomas. Here, we describe a reproducible next-generation sequencing assay to identify and characterize clonal IG gene rearrangements for MRD detection in cell-free DNA.

Key words Cell-free DNA, Plasma, Immunoglobulin rearrangements, Therapy monitoring, Liquid biopsy, Minimal residual disease, Digital droplet PCR, Next-generation sequencing

1 Introduction

Circulating cell-free DNA (cfDNA) is fragmented extracellular DNA, which is released from apoptotic and necrotic cells in small fragments of <200 bp [1]. cfDNA is typically isolated from the blood stream; however, it is also possible to detect cfDNA in other biological fluids such as urine or cerebrospinal fluid [2–6]. Interestingly, in cancer patients, a fraction of 0.01–60% of the total cfDNA consists of circulating tumor DNA (ctDNA), which originates from neoplastic lesions [7].

The fact that ctDNA shares the same biological features as the cellular DNA of the tumor, such as point mutations, gene amplifications, and immunoglobulin (IG) and T-cell receptor (TR) gene rearrangements in lymphoma, makes utilizing ctDNA as a noninvasive biopsy in diagnostic approaches and monitoring the status of minimal residual diseases (MRD) very attractive.

Anton W. Langerak (ed.), *Immunogenetics: Methods and Protocols*, Methods in Molecular Biology, vol. 2453, https://doi.org/10.1007/978-1-0716-2115-8_7,

The ideal markers for MRD detection in B-cell lymphomas are clonal IG gene rearrangements. The IG heavy chain (IGH) gene rearrangements are frequently used as target due to their unique junctional regions. However, IGH rearrangements may not be that reliable because of the somatic hypermutations (SHM) mainly in the V gene regions taking place during B-cell development and maturation. This might result in mismatches in primer binding [8]. Alternatively, incomplete IGHD-*IGHJ* rearrangements and IGK gene rearrangements could be used as targets for MRD. Both rearrangement types are mainly unmutated. Incomplete rearrangements in the IGH locus do not contain SHM in the majority of cases, because transcription only starts from the promoters in the V genes [9]. The finding of hypermutation in a small proportion of incomplete DJH rearrangements suggests important biological implications concerning the process of SHM. The rearrangements of the IGK genes can also be an important complementary MRD target, as in rearrangements involving the kappa deletion element (Kde), no SHM can occur after Kde recombination, since the deletion of the JK-CK introns removes the IGK enhancer that is essential for SHM [10].

PCR-based methods like allele-specific real-time quantitative PCR or digital droplet (dd) PCR targeting the clonal IG rearrangements are currently the gold standard for MRD quantification in cfDNA but are limited by a sensitivity of 1xE-05 in a polyclonal B-cell background. Since ctDNA is present at a very low overall amount in the peripheral blood, highly sensitive technologies are needed to detect MRD in cfDNA. Next-generation sequencing (NGS) of IG rearrangements (IG-NGS) is the technology that can overcome the limitation of PCR-based approaches with a potential higher sensitivity.

MRD assessment in cfDNA by IG-NGS requires the identification of the lymphoma-associated clonotypes in diagnostic tumor tissue. Therefore, fresh or formalin-fixed paraffin-embedded (FFPE) lymph node material or diagnostic peripheral blood or bone marrow with sufficient tumor infiltration is required for the initial marker identification. The EuroClonality-NGS working group has recently shown that IGH and IGK rearrangements are highly suitable for detecting clonality in frozen and FFPE-embedded tissue specimens [11]. Due to the small fragment size of cfDNA (~166 bp) and the high frequency of SHM in the variable heavy framework region 3 (IGHV-FR3), the EuroClonality-NGS IGHV-FR3 multiplex PCR was redesigned and optimized for the specific requirements of MRD detection in cfDNA. IGHD-IGHJ and IGK (IGKV-IGKJ, IGKV-KDE, and intron RSS-KDE) primer sets remained unchanged; only the reaction conditions and primer concentrations were modified to facilitate balanced amplification of all rearrangements also in this type of material.

As illustrated in Fig. 1, using a one-step NGS PCR protocol, clonal IG rearrangements are amplified in cfDNA and combined with molecular barcodes and sequencing adapters. The amplicons bind the flow cell of the Illumina MiSeq through the introduced adapters and are sequenced by synthesis. A standardized bioinformatic analysis of the high-throughput sequencing data allows the verification of clonal IG rearrangements and their precise quantitation. The bioinformatic platform ARResT/Interrogate (at http://arrest.tools/interrogate/), developed within the EuroClonality-NGS working group, allows identification of clonotypes and MRD follow-up in the same workflow.

The EuroClonality-NGS "central intra-tube quality/quantification control" (cIT-QC), comprising of known copy numbers of clonal rearrangements, is added to each reaction to enable the quantification of ctDNA as fraction of cfDNA and the correction of potential amplification biases. The cIT-QC is used to calculate the coverage of each single rearrangement copy in order to calculate the read coverage per cell and to determine the MRD level. The utility of this approach has been published recently [12].

Here we provide detailed instructions on amplicon sequencing of clonal IG rearrangements in cfDNA using modified EuroClonality-NGS protocols for IGH (VJ + DJ) and IGK (VJ + intron-Kde/V-Kde) (http://www.euroclonality.org/protocols/). The process of marker identification in FFPE samples or diagnostic bone marrow or peripheral blood is not part of this chapter; for that we refer to the publication of Scheijen et al. [11].

2 Materials

Solutions must be prepared with double-distilled water (supplied as ultrapure water or purified by filtering of 18 MΩ-cm at 25 °C). All reagents should be stored at 18–25 °C unless otherwise indicated.

2.1 Sample Collection

1. Blood collection tubes (S-Monovette®EDTA, PAXgene®, Cell-free DNA BCT® Blood collection tubes [Streck]).

2.2 cfDNA Extraction

1. QIAamp® Circulating Nucleic Acid Kit (QIAGEN). Add 200 ml isopropanol (100%) to 300 ml buffer ACB concentrate to prepare ready-to-use buffer ACB. Add 25 ml ethanol (96–100%) to 19 ml buffer ACW1 concentrate to prepare ready-to-use buffer ACW1. Add 30 ml ethanol (96–100%) to 13 ml buffer ACW2 concentrate to prepare ready-to-use buffer ACW2.
2. QIAvac 24 Plus vacuum pump (QIAGEN) or any vacuum pump capable of producing a pressure of −800 to −900 mbar. Alternatively, use the Maxwell® RSC instrument (Promega) with the corresponding Maxwell® RSC ccfDNA Plasma Kit (Promega) for automated extraction.

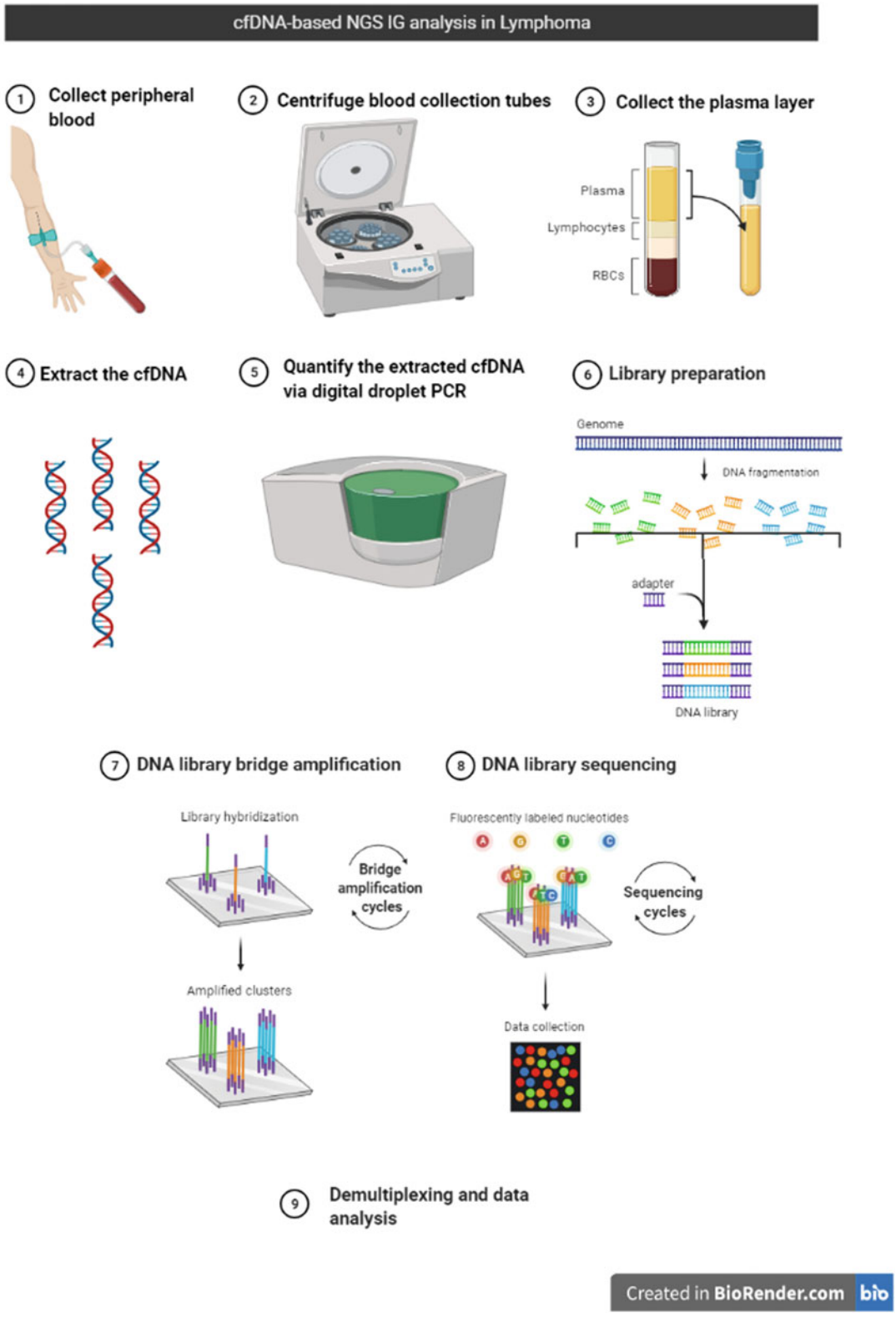


Fig. 1 Schematic representation of the cfDNA-based NGS IG rearrangement analysis in lymphoma. Adapted from "Next Generation Sequencing (Illumina)," by BioRender.com (2021). Retrieved from https:/app.biorender.com/biorender-templates

Table 1
List of target primers used for quantification using digital droplet PCR as described previously [14]

Target	Genome loci	Primer/probe sequence (5′-3′)	Hydrolysis probe fluorophore/ quencher
Albumin	4q13.3	F: CTGGAAGTCGATGAAACATACGTT R: CTCTCCTTCTCAGAAAGTGTGCATA P: TGCTGAAACATTCACCTTCCATG CAGA	FAM/TAM
GAPDH	12p13	F: AGGTTTACATGTTCCAATATGATTCCA R: ATGGGATTTCCATTGATGACAAG P: CCGTTCTCAGCCTTGACGGTGC	FAM/TAM
TERT	5p15.33	F: CCTCACATAAATGCTACCAAACGA R: TTCCAAGAAGGAGGCCATAGTC P: AAGAAATGAACAGACCCATCCCC CAGG	FAM/TAM

3. Vortexer.
4. Centrifuge.
5. Heat block.
6. Water bath.
7. Sterile 50 ml conical tubes.

2.3 Digital Droplet PCR (ddPCR)

1. 2× ddPCR™ Supermix (no dUTP, Bio-Rad).
2. ddH_2O.
3. Forward and reverse primers (*see* Table 1 for ddPCR primer list).
4. Extracted ccfDNA samples.
5. DG8 Cartridges for QX200 Droplet Generator (Bio-Rad).
6. DG8 Gaskets for QX200 Droplet Generator (Bio-Rad).
7. QX200™ Droplet Digital PCR system (Bio-Rad).
8. Twin.tec PCR Plate 96 (Eppendorf).
9. Pierceable Foil Heat Seal (Bio-Rad).
10. Thermal cycler.
11. PX1 PCR plate sealer (Bio-Rad).
12. Centrifuge.
13. Vortexer.

2.4 One-Step Next-Generation Sequencing (NGS) PCR

1. FastStart™ High Fidelity reaction buffer (Roche) w/o $MgCl_2$.
2. FastStart™ High Fidelity Taq polymerase (Roche).
3. $MgCl_2$ (Roche).

4. dNTPs.
5. ddH_2O (HPLC purified).
6. Library preparation target primers (desalted) (*see* Table 2).
7. High sensitivity dsDNA concentration measuring kit (i.e., Qubit® dsDNA HS Assay Kit (ThermoFisher Scientific) or QuantiT PicoGreen dsDNA Assay Kit (ThermoFisher Scientific)).
8. Qubit assay 0.5 ml tubes.
9. 2× 250 MiSeq Reagent Kits (Illumina) (i.e., v2, v2-Nano Kit (500 cycles), or v3 Kit (600 cycles), depending from number of samples and sequencing depths).
10. PhiX Control v3 (Illumina) (10 nM). Dilute the PhiX (10 nM) to 4 nM by mixing 2 μl 10 nM PhiX with 3 μl EBT buffer.
11. HT1 (Hybridization buffer) (Illumina).
12. NaOH. Prepare a fresh 0.2 N NaOH daily by mixing 20 μl 10 N NaOH with 980 μl ddH_2O.
13. EBT-Buffer (Illumina).
14. Tween 20®.
15. Thermal Cycler.
16. Low binding tubes.
17. Illumina MiSeq sequencer.
18. Light Cycler LC480/Qubit Fluorometer.

2.5 Purification of Subpools by Gel Extraction

1. Gel electrophoresis chamber.
2. Gel Red (VWR).
3. Gel loading dye.
4. MinElute Gel extraction kit (Qiagen).
5. Scalpel or X-Tracta tips.
6. Agarose.
7. TBE (or TAE) Buffer.

3 Methods

All experimental procedures should be carried out at room temperature unless otherwise indicated.

3.1 Sample Preparation

1. Centrifuge blood collection tubes at 2000 × *g* for 10 min. If using S-Monovette®EDTA tubes, processing time should not be longer than 4 h after sample taking.
2. Carefully move the supernatants (plasma) into a 5 ml tube with a conic bottom without damaging the buffy coat phase.

Table 2
List of target primer pools used for library preparation. Sequencing binding adapter (violet), barcoding variable sequence (XXXXXXXX), sequencing primer (blue), target-specific sequence (green)

Primer mix	Name	Sequence (5'-3')	µl in Pool	µM in PCR
IGHcfFR3_FP_mix	IGHcfFR3_VH1	AATGATACGGCGACCACCGAGATCTACACXXXXXXXXACACTCTTTCCCTACACGACGCTCTTCCGATCTTGAGGAGCCTGAGATCTGAVGACW	5	0.5
	IGHcfFR3_VH2	AATGATACGGCGACCACCGAGATCTACACXXXXXXXXACACTCTTTCCCTACACGACGCTCTTCCGATCTTGACCAACATGGACMYTGTGGACA	2	0.2
	IGHcfFR3_VH3_a	AATGATACGGCGACCACCGAGATCTACACXXXXXXXXACACTCTTTCCCTACACGACGCTCTTCCGATCTTGAACAGYCTGAGAGCYGAGGACA	1	0.1
	IGHcfFR3_VH3_b	AATGATACGGCGACCACCGAGATCTACACXXXXXXXXACACTCTTTCCCTACACGACGCTCTTCCGATCTTGAACAGCCTGARARCCGAGGACA	1	0.1
	IGHcfFR3_VH3_c	AATGATACGGCGACCACCGAGATCTACACXXXXXXXXACACTCTTTCCCTACACGACGCTCTTCCGATCTTGAACAGHCTGAGAGCHGAGGRCA	1	0.1
	IGHcfFR3_VH4	AATGATACGGCGACCACCGAGATCTACACXXXXXXXXACACTCTTTCCCTACACGACGCTCTTCCGATCTTGAGCTCTGTGACYGCCGCRGACA	1	0.1
	IGHcfFR3_VH5	AATGATACGGCGACCACCGAGATCTACACXXXXXXXXACACTCTTTCCCTACACGACGCTCTTCCGATCTAGCAGCCTGAAGGCYTCGGACR	1	0.1
	IGHcfFR3_VH6	AATGATACGGCGACCACCGAGATCTACACXXXXXXXXACACTCTTTCCCTACACGACGCTCTTCCGATCTTGAACTCTGTGACTCCCGAGGACA	10	1.0
	IGHcfFR3_VH7	AATGATACGGCGACCACCGAGATCTACACXXXXXXXXACACTCTTTCCCTACACGACGCTCTTCCGATCTTCWGCAGCCTAAAGGCTGAGGACA	10	1.0
	IGHcfFR3_VH	AATGATACGGCGACCACCGAGATCTACACXXXXXXXXACACTCTTTCCCTACACGACGCTCTTCCGATCTGHGGACACNGCYGTGTATTAC	1	0.1
IGHJ_A	IGH-J-A-1	CAAGCAGAAGACGGCATACGAGATXXXXXXXXGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTTACCTGAGGAGACGGTGACC	1	0.1
IGHDJ_FP_mix	IGHDJ_A	AATGATACGGCGACCACCGAGATCTACACXXXXXXXXACACTCTTTCCCTACACGACGCTCTTCCGATCTGATTCYGAACAGCCCCGAGTCA	1	0.2
	IGHDJ_B	AATGATACGGCGACCACCGAGATCTACACXXXXXXXXACACTCTTTCCCTACACGACGCTCTTCCGATCTGATTTTGTGGGGGYTCGTGTC	1	0.2
	IGHDJ_C	AATGATACGGCGACCACCGAGATCTACACXXXXXXXXACACTCTTTCCCTACACGACGCTCTTCCGATCTGTTTGRRGTGAGGTCTGTGTCA	1	0.2
	IGHDJ_D	AATGATACGGCGACCACCGAGATCTACACXXXXXXXXACACTCTTTCCCTACACGACGCTCTTCCGATCTGTTTRGRRTGAGGTCTGTGTCACT	1	0.2
	IGHDJ_E	AATGATACGGCGACCACCGAGATCTACACXXXXXXXXACACTCTTTCCCTACACGACGCTCTTCCGATCTCTTTTTGTGAAGGSCCCTCCTR	1	0.2
	IGHDJ_F	AATGATACGGCGACCACCGAGATCTACACXXXXXXXXACACTCTTTCCCTACACGACGCTCTTCCGATCTGTTATTGTCAGGSGRTGTCAGAC	1	0.2
	IGHDJ_G	AATGATACGGCGACCACCGAGATCTACACXXXXXXXXACACTCTTTCCCTACACGACGCTCTTCCGATCTGTTATTGTCAGGGGGTGYCAGRC	1	0.2
	IGHDJ_H	AATGATACGGCGACCACCGAGATCTACACXXXXXXXXACACTCTTTCCCTACACGACGCTCTTCCGATCTGTTTCTGAAGSTGTCTGTRTCAC	1	0.2
IGHJ_A	IGH-J-A-1	CAAGCAGAAGACGGCATACGAGATXXXXXXXXGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTTACCTGAGGAGACGGTGACC	1	0.1
IGK_FP_mix	IGK_V_A_1	AATGATACGGCGACCACCGAGATCTACACXXXXXXXXACACTCTTTCCCTACACGACGCTCTTCCGATCTAAGTGGGGTCCCATCAAGGTTCAG	1	0.1
	IGK_V_B	AATGATACGGCGACCACCGAGATCTACACXXXXXXXXACACTCTTTCCCTACACGACGCTCTTCCGATCTAGTCCCATCTCGGTTCAGTGGCAG	1	0.1
	IGK_V_C	AATGATACGGCGACCACCGAGATCTACACXXXXXXXXACACTCTTTCCCTACACGACGCTCTTCCGATCTGAAACAGGGGTCCCATCAAGGTTC	1	0.1
	IGK_V_D	AATGATACGGCGACCACCGAGATCTACACXXXXXXXXACACTCTTTCCCTACACGACGCTCTTCCGATCTTCCCAGACAGATTCAGTGGCAGTG	1	0.1
	IGK_V_E	AATGATACGGCGACCACCGAGATCTACACXXXXXXXXACACTCTTTCCCTACACGACGCTCTTCCGATCTCTGGAGTGCCAGATAGGTTCAGTG	1	0.1
	IGK_V_F	AATGATACGGCGACCACCGAGATCTACACXXXXXXXXACACTCTTTCCCTACACGACGCTCTTCCGATCTCCCTGGAGTCCCAGACAGGTTCAG	1	0.1
	IGK_V_G	AATGATACGGCGACCACCGAGATCTACACXXXXXXXXACACTCTTTCCCTACACGACGCTCTTCCGATCTGCATCCCAGCCAGGTTCAGTG	1	0.1
	IGK_V_H	AATGATACGGCGACCACCGAGATCTACACXXXXXXXXACACTCTTTCCCTACACGACGCTCTTCCGATCTGTCCCTGACCGATTCAGTGGCA	1	0.1
	IGK_V_I	AATGATACGGCGACCACCGAGATCTACACXXXXXXXXACACTCTTTCCCTACACGACGCTCTTCCGATCTAATCCCACCTCGATTCAGTGGC	1	0.1
	IGK_V_J	AATGATACGGCGACCACCGAGATCTACACXXXXXXXXACACTCTTTCCCTACACGACGCTCTTCCGATCTCTCAGGGGTCCCCTCGAGGTT	1	0.1
	IGK_V_K	AATGATACGGCGACCACCGAGATCTACACXXXXXXXXACACTCTTTCCCTACACGACGCTCTTCCGATCTAGACACTGGGGTCCCAGCCA	1	0.1
	IGK_INTRA_A	AATGATACGGCGACCACCGAGATCTACACXXXXXXXXACACTCTTTCCCTACACGACGCTCTTCCGATCTGAGTGGCTTTGGTGGCCATGC	1	0.1
IGK_RP_mix	IGK_J_A	CAAGCAGAAGACGGCATACGAGATXXXXXXXXGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTACGTTTGATCTCCACCTTGGTCCC	1	0.1
	IGK_J_B	CAAGCAGAAGACGGCATACGAGATXXXXXXXXGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTACGTTTGATATCCACTTTGGTCCC	1	0.1
	IGK_J_C	CAAGCAGAAGACGGCATACGAGATXXXXXXXXGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTACGTTTAATCTCCAGTCGTGTCCC	1	0.1
	IGK_DE_A	CAAGCAGAAGACGGCATACGAGATXXXXXXXXGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGCAGCTGCAGACTCATGAGGAG	1	0.1

3. Centrifuge the isolated plasma at 16,000 × *g*, 4 °C for 10 min (*see* **Note 1**).
4. Move supernatants to a clean 5 ml tube. Proceed immediately to cfDNA extraction (Subheading 3.2) or store at −80 °C (*see* **Note 2**).

3.2 cfDNA Extraction

cfDNA should be extracted from 4 ml purified blood plasma (set samples to 4 ml by adding phosphate-buffered saline if the volume is less than 4 ml).

1. Prior to starting the extraction the following should be done:
   (a) Equilibrate samples and buffers to room temperature (18–25 °C).
   (b) Heat a water bath or heating block to 60 °C for use with 50 ml tubes.
   (c) Heat a heating block to 56 °C for use with 2 ml Eppendorf tubes.
2. Pipet 400 µl QIAGEN Proteinase K into a pre-labeled 50 ml tube.
3. Add 4 ml plasma to the tube.
4. Add 3.2 ml buffer ACL to the tube. Mix well by vortexing for 30 s.
5. Incubate for 30 min at 60 °C.
6. Add 7.2 ml buffer ACB, mix well by vortexing for 15–30 s.
7. Incubate the mixture for 5 min on ice.
8. Insert the QIAamp Mini column into the VacConnector on the QIAvac 24 Plus. Insert a 20 ml tube extender into the open QIAamp Mini column. Make sure that the tube extender is firmly inserted into the QIAamp Mini column to avoid leakage of the sample.
9. Carefully pour the mixture from **step 6** into the tube extender of the QIAamp Mini column. Set the vacuum pump to produce a vacuum of −800 mbar to −900 mbar until all lysates are drawn through (takes up to 15 min) (*see* **Note 3**).
10. Release the pressure to 0 mbar, discard the tube extenders carefully and leave the QIAamp Mini columns attached to the VacConnector on the QIAvac 24 Plus.
11. Add 600 µl washing buffer ACW1 to the QIAamp Mini column. Switch on the vacuum pump (−800 mbar to −900 mbar) while the lid is open (*see* **Note 3**). After the entire washing buffer has been drawn through the column, switch the vacuum pump off and release the pressure to 0 mbar.
12. Add 750 µl washing buffer ACW2 to the QIAamp Mini column. Switch on the vacuum pump (−800 mbar to

−900 mbar) while the lid is open (*see* **Note 3**). After the entire washing buffer has been drawn through the column, switch the vacuum pump off and release the pressure to 0 mbar.

13. Add 750 µl ethanol (96–100%) to the QIAamp Mini column. Switch on the vacuum pump (−800 mbar to −900 mbar) while the lid is open (*see* **Note 3**). After the entire washing buffer has been drawn through the column, switch the vacuum pump off and release the pressure to 0 mbar.
14. Close the lid of the QIAamp Mini column. Remove it from the vacuum manifold and discard the VacConnector. Move the QIAamp Mini column to a clean 2 ml collection tube and centrifuge at full speed (16,000 × *g*) for 3 min.
15. Move the QIAamp Mini column to a new 2 ml collection tube, open the lid, and incubate at 56 °C for 10 min to dry the membrane completely.
16. Move the QIAamp Mini column to a clean 1.5 ml elution tube. Carefully pipet 20–150 µl of Buffer AVE onto the filter of the QIAamp Mini column. Close the lid and incubate at room temperature for 3 min (*see* **Note 4**).
17. Centrifuge for 1 min at 16,000 × *g*. Discard columns.
18. For downstream processing, isolated cfDNA can be stored at 4 °C for up to 24 h. For longer storage freeze at <−30 °C.
19. Use a fragment analyzer to accurately size and qualify the extracted cfDNA (*see* **Note 5**).

3.3 Digital Droplet PCR (ddPCR)-Mediated Copy Number Quantification

1. Prepare the reaction mixture by adding 10 µl 2× ddPCR Supermix, 0.3 µl forward primer (20 µM), 0.3 µl reverse primer (20 µM), and 0.1 µl probe (20 µM) to 1 µl DNA and fill with ddH_2O up to 20 µl.
2. Generate the droplets by pipetting 20 µl sample into the sample wells and 70 µl QX200 Droplet generation oil into the Oil well in the DG8 Cartridge, cover using the DG8 Gaskets and place into the QX200 Droplet generator.
3. Carefully transfer 40 µl of the generated droplets to the Twin.tec PCR Plate by slowly pipetting using a multichannel pipette. Seal plate using a pierceable Foil Heat Seal and a PX1™ PCR Plate sealer.
4. Perform the PCR Reaction on a thermal cycler by incubating the sample for 10 min at 95 °C for denaturation followed by 40 cycles of 30 s incubation at 94 °C for initial denaturation and 1 min incubation at 60 °C for annealing. Apply a final extension step by incubating for 10 min at 98 °C and store the sample at 4 °C.

5. Move the plates to the QX200 Droplet Reader to measure the amplified Droplets (*see* **Note 6**).
6. To measure the amplified fragments, start a new experiment in the QuantaLife™ Software, adjust Supermix to ddPCR™ Supermix w/o dUTP for the desired wells, name targets and select the appropriate channels as used in the probe (FAM, HEX, etc.).
7. Place the plate in the reader and run the experiment, choose DyeSet: FAM or FAM/HEX.
8. Save the analysis and load the exported file into the QuantaSoft™ Pro software.
9. Set the threshold in the 1D-Amplitude to the positive fraction (*see* **Notes 7** and **8**).
10. Export the results to Excel or any other spreadsheet application.
11. Calculate the mean of the copies/20 µl reads and divide it by 2 to determine the copy number/cell.
12. Divide the copy number/cell by 150 to calculate the concentration (ng/µl). Calculate the cfDNA concentration per 1 ml of plasma.

3.4 NGS Library Preparation

For library preparation, keep the reagents on ice until use. Use precooled racks for preparing the reaction mixture. Alternatively, the reaction could be prepared by placing the tubes on ice. Use the EuroClonality-NGS cIT-QC with known copy numbers of clonal rearrangements, in order to calculate the read coverage per cell (*see* **Note 9** and [12]).

1. Prepare the target-specific master mixes according to Table 3. Vortex and centrifuge the tubes.
2. Add 40 µl target-specific master mix to 1500 cell equivalents of a pre-quantified cfDNA. Adjust the volume to 50 µl with ddH_2O. Consider preparing one reaction for buffy coat (positive control) and another without added DNA (negative control).
3. Perform the targeted PCR Reactions on a thermal cycler as indicated in Table 4.
4. Verify the amplification of the target genes by agarose gel electrophoresis by loading 7 µl PCR product and 3 µl DNA loading dye into a 2% agarose gel (Fig. 2). Estimate the concentration of the amplified products by comparing the bands' intensity with these of the DNA ladder. To prepare a 2% agarose gel, weigh 8 g agarose and add it to 400 ml TBE buffer in a 1 l glass container. Dissolve the agarose by heating for intervals of 15–20 s in a microwave at 800 W, swirl the container gently,

Table 3
Targeted-PCR master mixes

		IGH-VJ (cfFR3)		IGH-DJ		IGK	
	Stock concentration	**Final concentration**	**Volume (µl)/ library**	**Final concentration**	**Volume (µl)/ library**	**Final concentration**	**Volume (µl)/ library**
FastStart™ high fidelity reaction buffer w/o $MgCl_2$	10×	1×	5	1×	5	1×	5
$MgCl_2$	25 mM	3 mM	6	3 mM	6	1.5 mM	3
Forward primer		Variable	1.65	Variable	0.8	0.1 µM each	0.6
Reverse primer	10 µM	0.1 µM	0.5	0.4 µM	2	0.1 µM	2
Aqua			23.65		22		26.2
dNTP-Mix	10 mM	0.2 mM	1	0.4 mM	2	0.2 mM	1
Spike-in DNA	40 copies	80 copies	2	80 copies	2	80 copies	2
FastStart™ high fidelity enzyme mix	5 U/µl	1 U	0.2	1 U	0.2	1 U	0.2

Table 4
Thermal cycler profiles of targeted PCR reactions

		IGH-VJ (cfFR3)		IGH-DJ and IGK	
1 cycle	Initial denaturation	94 °C	10 min	94 °C	10 min
35 cycles	Denaturation	94 °C	1 min	94 °C	1 min
1 cycle	Annealing	50 °C	1 min	63 °C	1 min
	Extension	72 °C	30 s	72 °C	30 s
1 cycle	Final extension	72 °C	10 min	72 °C	10 min
	Storage	4 °C	∞	4 °C	∞

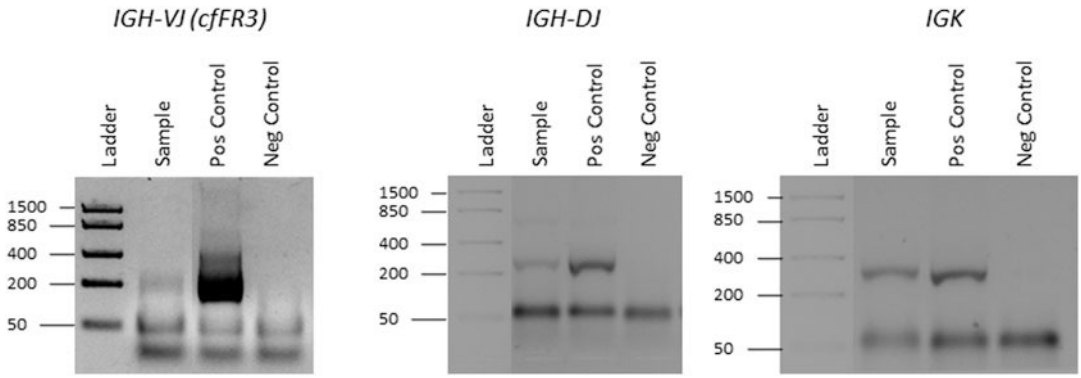


Fig. 2 PCR products analyzed by agarose gel electrophoresis. The target-specific band size is around 250 bp in IGH-VJ, 300 bp in IGH-DJ, and 300 bp in IGK. Amplification reactions using buffy coat and distilled water served as positive and negative controls, respectively

and repeat until the agarose completely dissolves. Allow the mixture to cool to <60 °C. Add 1:10,000 Gel-Red, mix gently and pour the dissolved agarose into the casting form, place the comb in place, and wait until the gel sets (around 30 min).

5. Pool 2–10 μl PCR products of the same target (i.e., cfFR3, IGHDJ, IGK). Consider the differences in the bands' intensities to pool equal amounts of the samples. Load the mixture on a new 2% agarose gel for separation by electrophoresis. After clear separation of the bands, extract the corresponding DNA fragments by excising the bands out of the gel using a clean scalpel or X-tracta tips. Using a gel extraction kit, extract the DNA from the agarose gel and elute in 10–20 μl elution buffer. Alternatively, pools could be purified using a magnetic beads-based purification kit.
6. Quantify the extracted DNA fragments using a Qubit® assay and use a fragment analyzer to identify the length of the extracted fragments. Fragment length should vary from

280 bp to 290 bp for IGH-VJ (cfFR3), 250–270 bp for IGH-DJ, and around 300 bp for IGK.

7. Using the fragment length and the DNA concentration (ng/μl), convert the DNA concentration (ng/μl) to nM using the following formula:

$$\text{Concentration (ng/μl)}/(\text{DNA fragment size} \times 650) \times 1,000,000$$

8. Dilute the amplified libraries to 2 nM or 4 nM when using a V3 Kit. According to the desired total reads, choose the corresponding MiSeq Reagent Kit (V2, V3, nano V2, etc.) (*see* **Note 10**).
9. Define the desired read count of each target and pool all targets in the corresponding ratio.

3.5 Next-Generation Sequencing

The Illumina MiSeq instrument must be set up correctly before use. Post-run and maintenance wash procedures must be performed after each run.

1. Before starting, thaw the HT1 Hybridization buffer on ice and the cartridge in water for at least 1 h. After the cartridge thaws, invert it 10× to mix the thawed reagents. Inspect that all positions are thawed. If air bubbles are present, gently tap the cartridge on the bench. Keep cool until use.
2. After preparing and pooling the sequencing libraries, denature the PhiX control using 0.2 N NaOH by mixing 5 μl 4 nM PhiX with 5 μl 0.2 N NaOH, centrifuge for 1 min at 300 × *g*, and incubate for 5 min at room temperature. Stop the denaturation by adding 990 μl precooled HT1 buffer. The PhiX end concentration of the mixture is 20 pM (*see* **Note 11**).
3. Denature a 2 nM library by mixing 5 μl 2 nM library with 5 μl 0.2 N NaOH, centrifuge for 1 min at 300 × *g*, and incubate for 5 min at room temperature. Stop the denaturation by adding 990 μl precooled HT1 buffer. The end concentration of the library is 10 pM (*see* **Note 12**).
4. Dilute both denatured PhiX and denatured library pool to the same concentration using precooled HT1 buffer. Combine 10% denatured PhiX (60 μl) to the denatured library pool (540 μl) for a low-diversity library (*see* **Note 13**). Mix well by inverting the tube and centrifuge for 1 min at 300 × *g*. Keep the mixture on ice until loading into the cartridge.
5. Create a new sample sheet using the Illumina Experiment Manager software by choosing the corresponding Illumina device and selecting "other analysis followed by FastQ Only." Define cycle reads for forward and reverse to 200 each. Select the desired Index primers and save the sheet.

6. Load 600 µl library pool mixture into the sample position in the cartridge. Load the cartridge and the flow cell into the MiSeq device and start the run according to the manufacturer's instructions (*see* **Note 14**) for monitoring data quality.
7. Export the data for demultiplexing and processing.

3.6 Bioinformatic Analysis

Output data are analyzed using the previously described bioinformatic platform ARResT/Interrogate [12, 13]; *see* below.

1. Sequencing output data are demultiplexed using the bcl2fastq tool (Illumina) with 0 mismatches allowed in the barcode sequence ("—mismatches 0").
2. ARResT/Interrogate can be accessed at http://arrest.tools/interrogate and requires an account which can be created by emailing contact@arrest.tools.
3. Switch to the "processing" panel, create a new analysis, and then upload the samples in compressed FASTQ format (the extension should be ".fastq.gz").
4. Make sure to select the "ARResT.cfDNA" pipeline scenario for IGHfcFR3 samples and "ARResT.Routine" for IGH-DJ and IGK samples.
5. Click on the blue "test it" button. If the test was OK, one should be able to click on the green "process" button. If not, check the "process output" tab for feedback, and email con tact@arrest.tools if necessary.
6. Progress of the bioinformatic pipeline can be followed in the "process output" tab. There is no need to wait; one may even close the browser; either log in later or better make sure to provide an email address to receive email notifications.
7. When the run is complete, switch to the "file" panel and select it from the drop-down menu, and click on "load results;" when looking for clonotypes of very low abundance, load the "not prefiltered" results.
8. Switch to the "questions" panel.
9. To access normalized values (i.e., number of cells instead of number of reads), it is also necessary to provide the number of cells (derived from the DNA amount) in the sample, e.g., ~15,000 cells from 100 ng of DNA (also *see* **Note 15**)—this will be used as the denominator for the percentage calculation. There is a widget in the "processing" panel and the same in the "questions" panel (Fig. 3), which sets the value for all samples; if one needs to set different values for different samples, this should be done in a sample sheet and its "cells" column. Do not include spike-ins in those numbers.

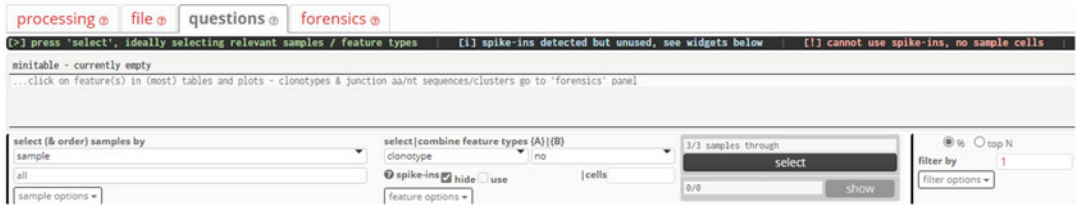


Fig. 3 Messages and widgets related to cIT-QC (spike-ins)

10. One should be able to see extra relevant widgets and messages in "questions" (and remember to hover over the "?" tooltip anchors) – to see normalized abundances, check the "use" box (Fig. 3).

4 Notes

1. Precool the centrifuges to 4 °C prior sample preparation.
2. Avoid freeze-thaw cycles of plasma as this may lead to DNA degradation.
3. In case of clogged QIAamp Mini column, close the VacValve, carefully remove the whole assembly, and transfer the remaining sample from the tube extender to a new 50 ml tube. Move the QIAamp Mini column from the assembly into a clean 2 ml collection tube and centrifuge it at full speed for 1 min or until the sample has completely passed through the membrane. Reassemble the QIAamp Mini column with Tube extender, VacConnector, and VacValve. Transfer the remaining lysate into the Tube extender, and proceed as described previously.
4. The recovered eluate volume will be up to 5 μl less than the elution volume applied to the column.
5. The expected fragment length is 130–200 bp.
6. The droplet measurements must be carried out within 3–4 days.
7. The threshold must be defined manually and set closely as possible above the fraction of negative droplets of the positive control.
8. In the case of positive droplets in the negative controls, channel 2 should be also checked for positive droplets. Positive droplets in channel 2 refer to nonspecific amplifications, and these should not be considered in the analysis.
9. Central intra-tube control (cIT-QC), also known as spike-in DNA, is a mixture of 9 cell lines, comprising 46 clonal rearrangements with known copy numbers. Spike-in DNA is added to

Fig. 4 Representation of the "minitable" with the selected clonotypes and full nt sequences

each reaction for the quantification of ctDNA as fraction of cfDNA and for correction of potential amplification biases [12].

10. MiSeq Reagent V2 Kit provides up to 1.5×10^7 Reads, V3 Kit provides up to 2.5×10^7 Reads, V2 Micro up to 4×10^6 Reads, and V2 Nano up to 1×10^6 Reads.
11. Denatured PhiX is stable for up to 3 weeks at −20 °C.
12. If using a V3 kit, start with a 4 nM library to end with 20 pM library pool.
13. PhiX concentration varies according to the level of diversity. For high diversity panels, use 1% PhiX.
14. Use the Illumina Sequencing analysis viewer (SAV) software for monitoring the quality parameters of the running assay. Check the cluster optimization overview for optimal control on cluster density.
15. Input copy numbers of cfDNA might be low due to availability. For retrieval of clonal IG rearrangements in cfDNA, we recommend using 1500 cfDNA copy numbers. Define the level of detection (lod) by dividing 1 on the amplified copy number (in this case 1500 copies). For MRD detection during treatment, the LOD has to be determined for each sample according to the variability of input copy numbers.
16. Because of the nature of the cfDNA primers that result in short amplicons and thus compromise the rearranged gene annotations, the presented clonotypes only feature the junction amino acid sequence. There are more feature types that can be accessed, including the rearranged genes and the full nt sequences themselves (click on any clonotype and check the "minitable" or through the "forensics" panel and its "sequences" tab) (Fig. 4).

Acknowledgments

FKZ 01KT1807 TRANSCAN V-NOVEL by BMBF.

References

1. Volckmar A-L, Sültmann H, Riediger A, Fioretos T, Schirmacher P, Endris V et al (2018) A field guide for cancer diagnostics using cell-free DNA: from principles to practice and clinical applications. Genes Chromosom Cancer 57:123–139
2. Li Y, Pan W, Connolly ID, Reddy S, Nagpal S, Quake S et al (2016) Tumor DNA in cerebral spinal fluid reflects clinical course in a patient with melanoma leptomeningeal brain metastases. J Neuro-Oncol 128:93–100
3. Pan W, Gu W, Nagpal S, Hayden Gephart M, Quake SR (2015) Brain tumor mutations detected in cerebral spinal fluid. Clin Chem 61:514–522
4. Swinkels DW, de Kok JB, Hanselaar A, Lamers K, Boerman RH (2000) Early detection of leptomeningeal metastasis by PCR examination of tumor-derived K-ras DNA in cerebrospinal fluid. Clin Chem 46:132–133
5. Salvi S, Gurioli G, Martignano F, Foca F, Gunelli R, Cicchetti G et al (2015) Urine cell-free DNA integrity analysis for early detection of prostate cancer patients. Dis Markers 2015: 574120
6. Xia Y, Huang CC, Dittmar R, Du M, Wang Y, Liu H et al (2016) Copy number variations in urine cell free DNA as biomarkers in advanced prostate cancer. Oncotarget 7:35818–35831
7. Leon SA, Shapiro B, Sklaroff DM, Yaros MJ (1977) Free DNA in the serum of cancer patients and the effect of therapy. Cancer Res 37:646–650
8. van Dongen JJM, Langerak AW, Brüggemann M, Evans PA, Hummel M, Lavender FL et al (2003) Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 concerted action BMH4-CT98-3936. Leukemia 17:2257–2317
9. González D, González M, Alonso ME, Lopez-Perez R, Balanzategui A, Chillon MC et al (2003) Incomplete DJH rearrangements as a novel tumor target for minimal residual disease quantitation in multiple myeloma using real-time PCR. Leukemia 17:1051–1057
10. Langerak AW, Nadel B, de Torbal A, Wolvers-Tettero ILM, Van Gastel-Mol EJ, Verhaaf B et al (2004) Unraveling the consecutive recombination events in the human IGK locus. J Immunol 173:3878–3888
11. Scheijen B, Meijers RWJ, Rijntjes J, van der Klift MY, Möbs M, Steinhilber J et al (2019) Next-generation sequencing of immunoglobulin gene rearrangements for clonality assessment: a technical feasibility study by EuroClonality-NGS. Leukemia 33:2227–2240
12. Knecht H, Reigl T, Kotrová M, Appelt F, Stewart P, Bystry V et al (2019) Quality control and quantification in IG/TR next-generation sequencing marker identification: protocols and bioinformatic functionalities by EuroClonality-NGS. Leukemia 33:2254–2265
13. Bystry V, Reigl T, Krejci A, Demko M, Hanakova B, Grioni A et al (2017) ARResT/interrogate: an interactive immunoprofiler for IG/TR NGS data. Bioinformatics 33:435–437
14. Devonshire AS, Whale AS, Gutteridge A, Jones G, Cowen S, Foy CA et al (2014) Towards standardisation of cell-free DNA measurement in plasma: controls for extraction efficiency, fragment size bias and quantification. Anal Bioanal Chem 406:6499–6512

Chapter 8

Targeted Locus Amplification as Marker Screening Approach to Detect Immunoglobulin (IG) Translocations in B-Cell Non-Hodgkin Lymphomas

Elisa Genuardi, Beatrice Alessandria, Aurora Maria Civita, and Simone Ferrero

Abstract

Although MRD monitoring by the classic polymerase chain reaction (PCR) approach is a powerful outcome predictor, about 20% of mantle cell lymphoma (MCL) and 50% of follicular lymphoma (FL) patients still lack a molecular marker and are thus resulting not eligible for MRD monitoring. Targeted locus amplification (TLA), a new NGS technology, has been revealed as a feasible marker screening approach able to identify uncommon B-cell leukemia/lymphoma 1 (BCL1) and B-cell leukemia/lymphoma 2 (BCL2) rearrangements in MCL and FL cases defined as having "no marker" by the classic PCR approach.

Key words Mantle cell lymphoma, Follicular lymphoma, Immunoglobulin, Translocations, Molecular marker, Next-generation sequencing, Targeted locus amplification

1 Introduction

Mantle cell lymphoma (MCL) and follicular lymphoma (FL) are non-Hodgkin lymphomas with an aggressive and indolent clinical course, respectively [1]. Despite the high rate of success of modern immunotherapies in the treatment of these patients, relapsing disease at variable time from disease presentation is still the rule, and the consequent acquisition of more aggressive behavior overtime is common [2, 3]. Therefore, it is crucial to track the disease course by highly sensitive minimal residual disease (MRD) approaches, in order to assess both the effective treatment efficacy and to early identify patients at risk of relapse [4]. In the last decade several prospective clinical trials revealed MRD as a strong outcome predictor both in MCL and FL [5–8].

Chromosomal translocations, which juxtapose oncogenes to the immunoglobulin (IG) regions, are ideal molecular markers for

Anton W. Langerak (ed.), *Immunogenetics: Methods and Protocols*, Methods in Molecular Biology, vol. 2453, https://doi.org/10.1007/978-1-0716-2115-8_8,

MRD in mature B-cell lymphoproliferative diseases. In detail, MCL and FL are characterized by chromosomal translocations that transpose the B-cell leukemia/lymphoma 1 (*BCL1*) and B-cell leukemia/lymphoma 2 (*BCL2*) genes, respectively, near the IG heavy chain (IGH) regions; t(11;14) and t(14;18) result in the overexpression of cyclin D1 (CCND1) and BCL2 proteins and to the constitutive activation of proliferative and antiapoptotic cellular pathways, respectively [9].

Actually, fluorescence in situ hybridization (FISH) revealed that almost 90% of MCL and 80% of FL harbor the translocation in the diagnostic tissues, but this technology is not sensitive enough to monitor MRD in follow-up samples [7, 10].

On the other hand, polymerase chain reaction (PCR), able to detect up to one clonal cell among 100,000 analyzed cells, can overcome this limitation. Currently, due to its high international standardization level, it represents the gold-standard approach employed for MRD monitoring purposes in MCL and FL [7, 11].

The well-known t(11;14) and t(14;18) breakpoints concern, respectively, (a) major translocation cluster (MTC), involving the *BCL1* region at 11q13 and the *IGH* locus at 14q32; (b), major breakpoint region (MBR), and minor cluster region (mcr), occurring between *BCL2* gene at 18q21 and the 14q32.3 IGH region. In FL MBR is most frequently involved (80% of the identified rearrangements), while mcr is less frequently identified (~15%) [8]; some rare (<5% of cases) "minor" breakpoints involving regions 3′ and 5′ of the MBR and mcr and named 3′MBR, 5′mcr, and distal MBR have also been described [12].

Moreover, between the juxtaposed chromosomic regions, nucleotides, also called N insertions, are randomly added, establishing the tumor "fingerprint-like sequences" essential for MRD monitoring allele-specific oligonucleotide (ASO) assay design [13].

Although classic PCR approaches for marker screening and MRD monitoring have been defined and standardized within the EuroClonality-NGS and EuroMRD working groups (/www.euroclonality.org), about 20% of MCL and 50% of FL patients still lack a molecular marker, thus resulting not eligible for MRD monitoring.

In the last few years, IGH amplicon-based next-generation sequencing (NGS) applications successfully provided new scenarios in several hematological diseases, such as acute lymphoblastic leukemia [14, 15], multiple myeloma [16, 17], and different lymphomas [18, 19] describing those NGS approaches as feasible tools for marker identification and MRD monitoring, allowing clinical correlations in large patient populations.

Also NGS capture panels appeared to be useful in the detection of multiple molecular targets, but their limited sensitivity hampers the application to the MRD context [20, 21].

Targeted locus amplification (TLA), a NGS-based technology firstly developed in 2014 by Cergentis B.V., allows the detection of structural variants not identified by classic PCR methods. TLA protocol differs from NGS capture approaches: actually, it is based on the principle of physiological cross-linking of genome regions placed in physical proximity. Moreover, employing the targeted enrichment of short, locus-specific sequences, it results in the sequencing of all single nucleotide variants (SNVs) and structural variants, such as chromosomal translocations [22, 23].

Since its first publication, TLA approach has been employed in different contexts such as transgene detection, vector design, and novel SNVs identification, thus resulting in a promising technology also for onco-hematology [24–26]. In this context, the application of a multiplex TLA, as a marker screening tool, showed promising results in acute leukemia through detection of cryptic rearrangements and multiple (un)known translocated genes involved in leukemia pathogenesis [27, 28].

Recently the implementation of TLA targeting the fusion partners of the IGH enhancer described the presence of novel, uncommon *BCL1* and *BCL2* rearrangements in MCL and FL patients lacking a MRD molecular marker by classic PCR marker screening approach [29]. The newly identified TLA rearrangements allowed the design of highly sensitive ASO MRD assays (up to 1E-05), thus priming the potential use of this NGS technology to increase the number of lymphoma patients eligible for MRD monitoring in clinical trials.

Here we provide a detailed description of the TLA protocol as marker screening tool in MCL and FL patients, followed by an ASO MRD assay based on the TLA sequence (Fig. 1).

2 Materials

2.1 Reagents and Kits

- Red blood cell lysis buffer (NH_4Cl).
- 0.9% NaCl solution.
- Maxwell® RSC Blood DNA Kit (Promega).
- Go Taq G2 Flexi DNA Polymerase (Promega).
- Agarose gel electrophoresis.
- TAE 1×.
- Targeted locus amplification gDNA library prep kit (Cergentis).
- QuantiFluor® ONE dsDNA System (Promega).
- AMPure XP Beads (Beckman Coulter).
- 80% ethanol.
- MilliQ.

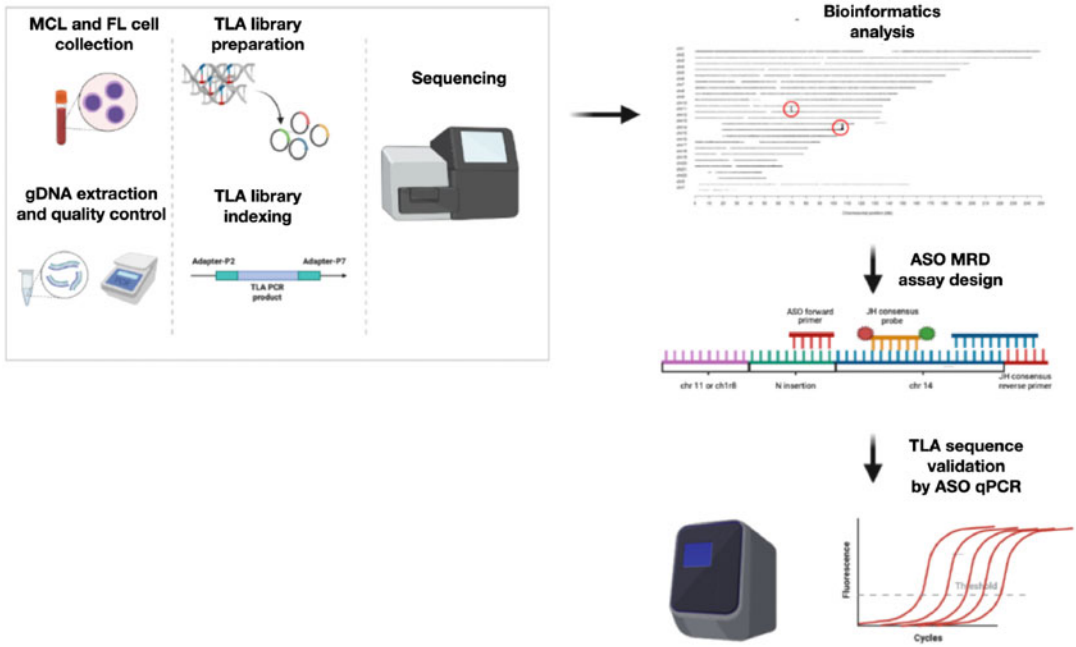


Fig. 1 TLA library preparation workflow

- 2-propanol.
- Nextera DNA Flex Library Prep kit (Illumina).
- Nextera DNA CD Indexes (24 Indexes, 24 sample-Illumina) (Illumina).
- High sensitivity D1000 ScreenTape (Agilent).
- NaOH.
- MiSeq Illumina v3 reagent kit (600 cycles).

2.2 Instruments and Software

- 50 ml conical tubes.
- Centrifuge with conical tube adapters and minispin centrifuge.
- 1–1.5 ml vials.
- 2 ml reaction tubes (with safe lock caps).
- 0.5 ml PCR tubes.
- Heating block.
- Maxwell RSC (Promega).
- NanoDrop2000 Spectrophotometer (Thermo Scientific).
- Thermal cycler with heated lid.
- Electrophoresis Systems.
- Quantus fluorometer (Promega).
- Magnetic rack fitting with 1.5–2 ml tubes.

- 2200 Tapestation system (Agilent).
- MiSeq sequencer (Illumina).
- PC, software for genome alignment and primer design.

3 Methods

3.1 Mantle Cell Lymphoma and Follicular Lymphoma Cell Collection

Collect bone marrow (BM) and peripheral blood (PB) samples in EDTA vacutainers, ranging from 2 to 7 ml and from 7 to 14 ml for BM and PB, respectively. Next, the red blood cell (RBC) lysis procedure is carried out to collect total white blood cells (WBC), as follows:

1. Mix 10–15 ml of BM or PB with lysis buffer 1× in 50 ml conical tubes up to a final volume of 45 ml.
2. Invert to mix and incubate at room temperature (RT) for 10–15 min.
3. Centrifuge at 226 × *g* for 15 min at RT.
4. Discard the supernatant and add lysis buffer up to a final volume of 20 ml.
5. Centrifuge at 226 × *g* for 10 min at RT.
6. Decant the supernatant and resuspend the pellet with 0.9% NaCl solution.
7. Aliquot 5–10 × 10^6 cells in 1–1.5 ml vials.
8. Centrifuge at 6440 × *g* for 1 min and discard the supernatant.
9. Store the samples at −80 °C.

3.2 gDNA Extraction and Quality Control

gDNA is extracted from 5–10 × 10^6 BM and PB dry cell pellets (*see* **Note 1**). High purity gDNA is obtained using semiautomated or automated DNA extraction procedures, avoiding DNA cross-contamination among the samples. Here we describe the Maxwell® Rapid Sample Concentrator (RSC) Blood protocol.

1. Set the heating block at 56 °C.
2. Resuspend 5–10 × 10^6 dry cell pellet using 300 μl of 0.9% NaCl solution or PBS.
3. Add 300 μl of Lysis Buffer and vortex for 10 s.
4. Add 30 μl of Proteinase K Solution and vortex.
5. Incubate at 56 °C for 20 min.
6. During sample incubation, prepare cartridges as described in the instrument operating manual.
7. Load the samples in the cartridge and proceed with the gDNA extraction program.

Table 1
Control gene PCR mix

	Concentration	μl/sample
Buffer 5×	5×	10
Primer Forward P8/5 5′-TATCCTGAGTAGTGGTAATC-3′	20 pM	1
Primer Reverse P8/3 5′- AAGTGAATCTGAGGCATAAC-3′	20 pM	1
$MgCl_2$	15 mM	5
dNTPs	2 μM	5
Go Taq G2 Flexi DNA polymerase	1.25u	0.25
H_2O		add to 50 μl

Table 2
Control gene thermal profile

Temperature	Time	n° cycles
94 °C	10 min	1
94 °C	40 s	40
56 °C	30 s	
72 °C	30 s	
72 °C	10 min	1
12 °C	Hold	Hold

After extraction, gDNA quantity (ng/μl) and quality (OD ratio A260/A280 and A260/A230) are evaluated by the NanoDrop2000 Spectrophotometer (Thermo Scientific, Waltham, MA, USA). gDNA is stored at 4 °C or −20 °C until library preparation.

Next, a control gene (P53 exon 8) amplification is performed to further qualitatively check the gDNA [30].

1. Prepare the PCR mix and set the thermal cycler as described in Tables 1 and 2, respectively.
2. Prepare the reaction mix on ice.
3. Add 100–500 ng as input gDNA.
4. To each run add a positive gDNA from a healthy subject and a no-template control (NTC).

Run the PCR products on a 2% agarose gel. The P53 exon 8 amplification signal should appear as a 150 bp band; samples without any signal should not be considered for TLA library preparation.

3.3 Targeted Locus Amplification (TLA) Library Preparation

TLA consists of different steps that allow gDNA cross-linking and circularization, followed by IGH target enrichment. The protocol outline takes 4 workdays; for more detailed information and technical support, please refer to Cergentis (www.Cergentis.com).

Day 1:

- *Assembly and fixation*: gDNA is assembled and fixed to fold, to cross-link and to connect regions placed in a very close proximity through the genome.
- *First enzymatic digestion, ligation,* and *reverse cross-linking*: long cross-linked genomic templates originating from the same locus are obtained using restriction enzymes, ligase, and proteinase.

Day 2:

- *Second enzymatic digestion and ligation*: the large gDNA fragments are digested to obtain molecules suited for PCR amplification and then circularized.

Day 3:

- *TLA PCR*: the circularized gDNA molecules are amplified using the IGH enhancer complementary primer.

Day 4:

- *TLA library indexing*: the TLA PCR products are fragmented and tagged with adapter sequences using the Bead-Linked Transposomes (BLT) kit.
- *TLA library sequencing on an Illumina platform.*

3.3.1 Assembly and Fixation

At least 5 μg of gDNA (*see* **Note 2**) is required for TLA library preparation

1. Measure the gDNA concentration using Quantus fluorometer (*see* **Note 3**) and dilute the sample in the concentration range of 50–100 ng/μl (*see* **Note 4**).
2. Add, in two separate 2 ml tubes, 290 μl Assembly Buffer (AB), 2.5 μg DNA, and dilution buffer (DB) until a 900 μl final volume.
3. Add 100ul Assembly Mix (AM) and incubate 15 min at 37 °C.
4. Add 90 μl Fixation Buffer (FB) and incubate exactly 10 min at 37 °C.
5. Add 90 μl Quencher Buffer (QB) and 800 μl thoroughly mixed AMPure XP Beads and incubate 15 min at RT.
6. Place the tubes on a magnetic rack until the beads are separated.
7. Discard the supernatant and add 1800 μl of fresh 80% ethanol to each tube and wait until the beads are completely separated.

8. Repeat **step 7**.
9. Remove residual ethanol and air dry the beads for 5 min.
10. Resuspend the beads in 520 µl 1× Restriction Buffer (RB), pool the samples in one 2 ml tube, and incubate 15 min at 55 °C (*see* **Note 5**).

3.3.2 First Enzymatic Digestion and Ligation

1. Incubate the samples on ice for 5 min.
2. Add 5 µl of Restriction enzyme 1 (RE1) and incubate for 2 h at 37 °C.
3. Inactivate RE1 by incubating the samples 25 min at 65 °C and then keeping them on ice for 5 min.
4. Add 125 µl 10× Ligation Buffer (10×LB) and 125 µl MilliQ (*see* **Note 5**).
5. Add 5 µl Ligase (LIG) and incubate 1 h at RT.
6. Reverse cross-linking by using 5 µl Proteinase K (ProtK) and incubating the samples overnight at 65 °C.

3.3.3 Second Enzymatic Digestion and Ligation

1. Mix 150 µl 10× Restriction Buffer (10×RB), 135 µl MilliQ, and 5 µl Restriction Enzyme 2 (RE2) and incubate for 1 h at 37 °C (*see* **Note 5**).
2. Inactivate RE2 for 20 min at 65 °C and then keep the samples on ice for 5 min.
3. Add 175 µl 10×LB, 70 µl MilliQ, and 5 µl LIG; then incubate the samples with the ligation mix for 1 h at RT.
4. Divide the samples in two 2 ml tubes, add 875 µl 2-propanol and 10 µl Magnetic Bead (MB), and incubate 1 h at RT.
5. Place tubes on the magnetic rack and remove supernatant when the beads are completely separated.
6. Wash the beads, with 1800 and 900 µl 80% ethanol. Between these two beads-wash steps, place the samples on the magnet, wait till the beads are separated, and then discard the supernatant. After the second wash, leave the beads to air dry for 15 min.
7. Resuspend the beads with 105 µl Elution Buffer (EB), place the samples on the magnet, and when the beads are completely separated, transfer 100 µl of the eluted sample in a 1.5 ml tube.
8. Measure the TLA template concentration using the Quantus fluorometer.

3.3.4 TLA PCR

1. Prepare 600 ng of TLA template and add MilliQ until a 90 µl final volume.
2. Set up the TLA PCR mix and thermal profile as described in Table 3.

Table 3
TLA PCR mix

	µl/sample
PCR mix 2×	100
Primer IGH enhancer mix 5′-AGCAATTAAGACCAGTTCCC-3′ 5′-CTCCACAACCTCTGAATGG-3′	10

Table 4
TLA PCR thermal profile

Temperature	Time	n° cycles
98 °C	30 s	1
98 °C 67 °C to 60 °C (0.5 °C/cycles) in 30 sec 65 °C	10 s 15 s 3.5 min	15
98 °C 60 °C 65 °C	10 s 30 s 3.5 min	25
65 °C	5 min	
12 °C	Hold	Hold

3. Divide the PCR mix in three PCR tubes and set up the thermal profile as described in Table 4.
4. Pool all the three PCR tubes into a 1.5 ml tube.
5. Add 200 µl of AMPure XP Beads and incubate 15 min at RT.
6. Place the samples on the magnet and remove the supernatant when the beads are completely separated.
7. Wash twice with 900 µl 80% ethanol and air dry the beads for 5 min.
8. Resuspend the beads with 55 µl EB and place on the magnet for 1 min.
9. Transfer 50 µl of the eluted TLA PCR products into a clean 1.5 tube and measure the concentration using Quantus fluorometer (*see* **Note 5**).

3.3.5 TLA Library Indexing

TLA library indexing is performed through Nextera DNA Flex Library Prep kit (Illumina), using a Bead-Linked Transposomes protocol to fragment and tag the TLA PCR products with adapter sequences, according to Illumina manual protocol procedures.

1. TLA library preparation is performed using at least 50–100 ng of TLA PCR products as starting material.

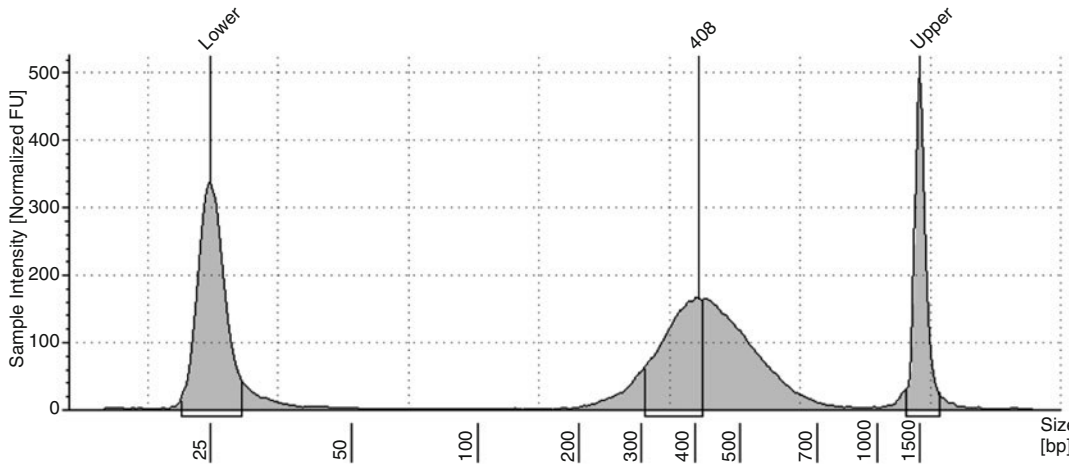


Fig. 2 TLA library profiles obtained using High sensitivity D1000 ScreenTape (Agilent)

2. After adapter ligation, quantify the libraries using Quantus fluorometer and check the quality using High sensitivity D1000 ScreenTape (Agilent). Figure 2 shows typical TLA library profiles with an average fragment size of 300–450 bp when analyzed with a size range of 150–1500 bp.

3.3.6 Sequencing

1. Pool 20 ng of each of the libraries and check the quantity and quality to obtain the concentration (nM) of the final TLA pool.
2. Dilute the final TLA pool to 4 nM and follow the Illumina's MiSeq sequencing protocol to denature the library.
3. Sequence the TLA final pool spiked with 1% PhIX on a MiSeq platform using a v3 chemistry (600 cycles, paired end-PE).

3.4 Bioinformatics Analysis

Bioinformatics analysis is performed by Cergentis. To identify break-spanning reads, FASTQ files are mapped against the human genome version hg19 using BWA-SW, which is a Smith-Waterman alignment tool. Then, the Integrated Genomic Viewer (IGV; http://software.broadinstitute.org/software/igv/) tool is used to confirm rearrangements, translocations, to identify TLA-BCL1 or TLA-BCL2 defined as breakpoint sequences mapping on chromosome 11 (for MCL), chromosome 18 (for FL), and the mate chromosome 14 (IGH locus).

3.5 TLA Sequence Validation by Allele-Specific Oligonucleotide (ASO) Approach

Once bioinformatically defined, TLA-BCL1 or TLA-BCL2 is further investigated to detect the breakpoint nucleotide sequences and to develop the ASO MRD assay based on the TLA sequence.

The free websource BlastN (https://blast.ncbi.nlm.nih.gov/Blast.cgi) aligns the query to the human genome version hg38, thus defining the t(11;14) and t(14;18) translocated regions and the N insertion nucleotides.

Then, the ASO MRD assay based on the TLA sequence is designed as follows:

1. An ASO forward primer is designed on the N insertions, which are randomly inserted between the translocated regions and which are not mapped with the Blast alignment.
2. The consensus probe and reverse primer are set to anneal the joining IGH region (JH) involved in the translocation.
3. The primer annealing temperature (Tm) is established both manually and using a primer design tool such as Integrated DNA Technologies (IDT free websource available at https://eu.idtdna.com/pages) following these criteria:
   (a) Primer Tm ranges between 58 °C and 62 °C.
   (b) Probe Tm is 10 °C higher than primer Tm.
   (c) It is recommended that the primer GC content is 40–60%.

Tables 5 and 6 show the list of JH primer available for TLA validation, while the ASO forward MRD assay design is detailed in Fig. 3.

Table 5
JH consensus reverse primer used in ASO forward MRD assay

JH reverse primer	Sequence (5′-3′)
R-JH1	CGCTATCCCCAGACAGCAGA
R-JH2	GGTGCCTGGACAGAGAAGACT
R-JH3	AGGCAGAAGGAAAGCCATCTTAC
R-JH4	CAGAGTTAAAGCAGGAGAGAGGTTGT
R-JH5 RP1	AGAGAGGGGGTGGTGAGGACT
R-JH5 RP2	CAAGCTGAGTCTCCCTAAGTGGA
R-JH6	GCAGAAAACAAAGGCCCTAGAGT

Table 6
JH consensus probe used in ASO forward MRD assay

JH probe	Sequence (5′-3′)
T-JH 1.2.4.5	CCCTGGTCACCGTCTCCTCAGGTG
T-JH3	CAAGGGACAATGGTCACCGTCTCTTCA
T-JH6	CACGGTCACCGTCTCCTCAGGTAAGAA

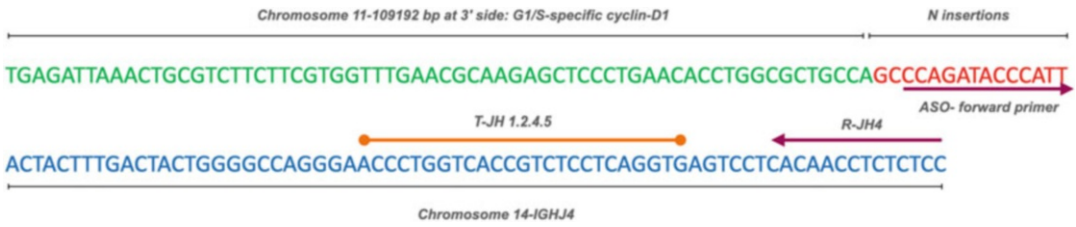


Fig. 3 ASO forward MRD assay design on TLA-BCL1 sequence

TLA-BCL1 or TLA-BCL2 assay validation is performed using highly sensitive quantification approaches as quantitative PCR (ASO qPCR) [13], setting a tenfold standard curve starting from 500 ng BM and or PB diagnostic sample serially diluted in pooled polyclonal healthy gDNAs or gDNA from a cell line not featuring any of the t(11;14) and t(18;14) translocations.

TLA-BCL1 or TLA-BCL2 is confirmed as MRD molecular markers if the validation experiment achieves a sensitivity level that allows the identification of 1 clonal cell within 100,000 analyzed cells, defined according to EuroMRD guidelines for qPCR data interpretation [11].

4 Notes

1. High cell amounts lead to beads carryover, which is affecting TLA library quality. On the other hand, the minimum gDNA yield requested by TLA library preparation is not obtained when starting from a too low cell amount. gDNA samples should have an OD 260/280 range 1.8–1.9 and OD 260/230 > 1.5.
2. gDNA quantification is performed using a fluorometer. Nanodrop could overestimate gDNA yield, thus affecting TLA library quality.
3. Verify the gDNA concentration after dilution and before starting TLA library preparation.
4. To check the quality of TLA library preparation, "undigested control," "digested control," and "ligation control" samples can be collected and processed with ProtK for 1 h at 65 °C. After AMPure XP beads purification, separate the samples on a 1% agarose gel. The "undigested control" is expected as a >10 kb signal; a smear between 0.3 and 2 kb and >5 kb are reported for the "digested control" and "ligation control."
5. Check TLA PCR products by separating the sample on a 1.2% agarose gel. Samples with a smear between 0.3 and 5 kb can proceed to the indexing phase.

References

1. Swerdlow SH, Campo E, Pileri SA, Harris NL, Stein H, Siebert R et al (2016) The 2016 revision of the World Health Organization classification of lymphoid neoplasms. Blood 127: 2375–2390
2. Dreyling M, Ferrero S, Hermine O (2014) How to manage mantle cell lymphoma. Leukemia 28:2117–2130
3. Freedman A (2014) Follicular lymphoma: 2014 update on diagnosis and management. Am J Hematol 89:429–436
4. Ferrero S, Dreyling M (2017) Minimal residual disease in mantle cell lymphoma: are we ready for a personalized treatment approach? Haematologica 102:1133–1136
5. Pott C, Hoster E, Delfau-Larue M-H, Beldjord K, Böttcher S, Asnafi V et al (2010) Molecular remission is an independent predictor of clinical outcome in patients with mantle cell lymphoma after combined immunochemotherapy: a European MCL intergroup study. Blood 115:3215–3223
6. Ladetto M, Lobetti-Bodoni C, Mantoan B, Ceccarelli M, Boccomini C, Genuardi E et al (2013) Persistence of minimal residual disease in bone marrow predicts outcome in follicular lymphomas treated with a rituximab-intensive program. Blood 122:3759–3766
7. Grimaldi D, Genuardi E, Ferrante M, Ferrero S, Ladetto M (2018) Minimal residual disease in indolent lymphomas: a critical assessment. Curr Treat Options in Oncol 19:71
8. Pott C, Brüggemann M, Ritgen M, van der Velden VHJ, van Dongen JJM, Kneba M (2019) MRD detection in B-cell non-Hodgkin lymphomas using Ig gene rearrangements and chromosomal translocations as targets for real-time quantitative PCR. Methods Mol Biol 1956:199–228
9. Seto M (2002) Molecular mechanisms of lymphomagenesis through transcriptional disregulation by chromosome translocation. Int J Hematol 76:323–326
10. Pott C, Sehn LH, Belada D, Gribben J, Hoster E, Kahl B et al (2020) MRD response in relapsed/refractory FL after obinutuzumab plus bendamustine or bendamustine alone in the GADOLIN trial. Leukemia 34:522–532
11. van der Velden VHJ, Cazzaniga G, Schrauder A, Hancock J, Bader P, Panzer-Grumayer ER et al (2007) Analysis of minimal residual disease by Ig/TCR gene rearrangements: guidelines for interpretation of real-time quantitative PCR data. Leukemia 21: 604–611
12. van Dongen JJM, Langerak AW, Brüggemann M, Evans PAS, Hummel M, Lavender FL et al (2003) Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98-3936. Leukemia 17:2257–2317
13. Donovan JW, Ladetto M, Zou G, Neuberg D, Poor C, Bowers D et al (2000) Immunoglobulin heavy-chain consensus probes for real-time PCR quantification of residual disease in acute lymphoblastic leukemia. Blood 95:2651–2658
14. Faham M, Zheng J, Moorhead M, Carlton VEH, Stow P, Coustan-Smith E et al (2012) Deep-sequencing approach for minimal residual disease detection in acute lymphoblastic leukemia. Blood 120:5173–5180
15. Brüggemann M, Kotrová M, Knecht H, Bartram J, Boudjogrha M, Bystry V et al (2019) Standardized next-generation sequencing of immunoglobulin and T-cell receptor gene recombinations for MRD marker identification in acute lymphoblastic leukaemia; a EuroClonality-NGS validation study. Leukemia 33:2241–2253
16. Oliva S, Genuardi E, Belotti A, Frascione PMM, Galli M, Capra A et al (2019) Minimal residual disease evaluation by multiparameter flow cytometry and next generation sequencing in the forte trial for newly diagnosed multiple myeloma patients. Blood 134:4322
17. Martinez-Lopez J, Lahuerta JJ, Pepin F, González M, Barrio S, Ayala R et al (2014) Prognostic value of deep sequencing method for minimal residual disease detection in multiple myeloma. Blood 123:3073–3079
18. Ladetto M, Brüggemann M, Monitillo L, Ferrero S, Pepin F, Drandi D et al (2014) Next-generation sequencing and real-time quantitative PCR for minimal residual disease detection in B-cell disorders. Leukemia 28: 1299–1307
19. Kurtz DM, Green MR, Bratman SV, Scherer F, Liu CL, Kunder CA et al (2015) Noninvasive monitoring of diffuse large B-cell lymphoma by immunoglobulin high-throughput sequencing. Blood 125:3679–3687
20. Wren D, Walker BA, Brüggemann M, Catherwood MA, Pott C, Stamatopoulos K et al (2017) Comprehensive translocation and clonality detection in lymphoproliferative disorders by next-generation sequencing. Haematologica 102:57–60

21. Stewart P, Gazdova J, Darzentas N, Wren D, Proszek P, Fazio G et al (2019) Euroclonality-NGS DNA capture panel for integrated analysis of IG/TR rearrangements, translocations, copy number and sequence variation in lymphoproliferative disorders. Blood 134:888
22. de Vree PJP, de Wit E, Yilmaz M, van de Heijning M, Klous P, Verstegen MJAM et al (2014) Targeted sequencing by proximity ligation for comprehensive variant detection and local haplotyping. Nat Biotechnol 32: 1019–1025
23. Hottentot QP, van Min M, Splinter E, White SJ (2017) Targeted locus amplification and next-generation sequencing. Methods Mol Biol 1492:185–196
24. Godet I, Shin YJ, Ju JA, Ye IC, Wang G, Gilkes DM (2019) Fate-mapping post-hypoxic tumor cells reveals a ROS-resistant phenotype that promotes metastasis. Nat Commun 10:4862
25. Chen YH, Pallant C, Sampson CJ, Boiti A, Johnson S, Brazauskas P et al (2020) Rapid lentiviral vector producer cell line generation using a single DNA construct. Mol Ther Method Clin Dev 19:47–57
26. Aeschlimann SH, Graf C, Mayilo D, Lindecker H, Urda L, Kappes N et al (2019) Enhanced CHO clone screening: application of targeted locus amplification and next-generation sequencing technologies for cell line development. Biotechnol J 14:e1800371
27. Alimohamed MZ, Johansson LF, De Boer EN, Splinter E, Klous P, Yilmaz M et al (2018) Genetic screening test to detect translocations in acute leukemias by use of targeted locus amplification. Clin Chem 64:1096–1103
28. Kuiper R, van Duin M, van Vliet MH, Broijl A, van der Holt B, el Jarari L et al (2015) Prediction of high- and low-risk multiple myeloma based on gene expression and the International Staging System. Blood 126:1996–2004
29. Genuardi E, Klous P, Mantoan B, Drandi D, Ferrante M, Cavallo F et al (2021) Targeted locus amplification to detect molecular markers in mantle cell and follicular lymphoma. Hematol Oncol 39(3):293–303
30. Gaidano G, Ballerini P, Gong JZ, Inghirami G, Neri A, Newcomb EW et al (1991) p53 mutations in human lymphoid malignancies: association with Burkitt lymphoma and chronic lymphocytic leukemia. Proc Natl Acad Sci 88: 5413–5417

Chapter 9

Immunoglobulin/T Cell Receptor Capture Strategy for Comprehensive Immunogenetics

James Peter Stewart, Jana Gazdova, Shambhavi Srivastava, Julia Revolta, Louise Harewood, Manisha Maurya, Nikos Darzentas, and David Gonzalez

Abstract

In the era of genomic medicine, targeted next generation sequencing strategies (NGS) are becoming increasingly adopted by clinical molecular diagnostic laboratories to identify genetic diagnostic and prognostic biomarkers in hemato-oncology. We describe the EuroClonality-NGS DNA Capture (EuroClonality-NDC) assay, which is designed to simultaneously detect B and T cell clonal rearrangements, translocations, copy number alterations, and sequence variants. The accompanying validated bioinformatics pipeline enables production of an integrated report. The combination of the laboratory protocol and bioinformatics pipeline in the EuroClonality-NDC minimizes the potential for human error, reduces economic costs compared to current molecular testing strategies, and should improve diagnostic outcomes.

Key words EuroClonality, Next generation sequencing, BIOMED-2, Immunoglobulin, T cell receptor, Copy number alteration, Translocation, Lymphoma

1 Introduction

Lymphoproliferative disorders (LPD) can be classified based on multiple parameters, including morphology, immunophenotyping, and genetic analysis. While a large number of lymphoproliferative disorders can be classified solely by assessment of morphology and immunophenotyping, there is an increasing role for the evaluation of genetic features as evidenced by the publication of the updated WHO guidelines in 2016 [1]. The updated guidelines include a vast array of genomic alterations that can significantly improve the diagnostic criteria and the prognostic relevance of existing entities and has led to the introduction of new disease entities. This may result in a change in practice in clinical laboratories with the validation of multiple molecular tests covering the required genetic alterations.

Anton W. Langerak (ed.), *Immunogenetics: Methods and Protocols*, Methods in Molecular Biology, vol. 2453, https://doi.org/10.1007/978-1-0716-2115-8_9,

As all cells within a tumor are presumed to arise from a common clonal progenitor, lymphoid malignancies should exhibit clonal rearrangements of the immunoglobulin (IG) and/or T cell receptor (TR) loci. Detection of a clonal IG/TR rearrangement, which can aid in differentiating between a clonal B/T cell proliferation and a reactive hyperplasia, is typically performed using a PCR-based method with primers designed, optimized, and validated during the EuroClonality BIOMED-2 study [2, 3]. PCR products are commonly analyzed using either capillary electrophoresis (i.e., GeneScan) or heteroduplex analysis on polyacrylamide gels. Analysis of clonality using these detection methods is prone to subjectivity, particularly in cases with low tumor infiltration, where it can be difficult to distinguish a clonal peak within a polyclonal background, or for targets with limited complementary determining region 3 (CDR3) variability (e.g., IGKV-J rearrangements). Interpretation of the results is also subject to confounding variables such as the impact of somatic hypermutation (SHM) on detection of IGH, IGK, and IGL rearrangements. The presence of mutations within the binding sites of the PCR primers can prevent annealing, leading to false-negative results which can be addressed in the majority of cases by performing PCR for alternate IG targets that are less prone to SHM such as IGH D-J and IGK Kde rearrangements [3].

Translocations are another genetic alteration tested for in molecular pathology laboratories as they are a hallmark of specific entities within non-Hodgkin lymphoma (NHL) as well as acute lymphoblastic leukemia (ALL) or plasma cell myeloma (PCM), among others. Translocations involving the IG/TR loci such as t(8;14)(q24;q32) in Burkitt lymphoma (BL), t(14;18)(q32;q21) in follicular lymphoma (FL), t(11;14)(q13;q32) in mantle cell lymphoma (MCL), and ALK translocations (to various partners such as *NPM1*, *ATIC* and *RANBP2*) in anaplastic large cell lymphoma (ALCL) are part of the diagnostic testing regime for those particular lymphomas. Currently, translocations are commonly assessed by FISH or PCR methods, although a number of tests are often required, particularly in B-NHL types to encompass different IG loci (IGH, IGK, and IGL) and to accommodate the large region where translocation breakpoints can occur. Multiple tests are often required to accurately define double/triple hit lymphomas as the recent 2016 WHO guidelines established a new classification of high-grade B-NHL based on the presence of a *MYC* translocation along with *BCL2* and/or *BCL6* translocations.

Single-nucleotide DNA alterations and/or small insertions or deletions, traditionally detected using Sanger sequencing and more recently by amplicon and capture targeted NGS, can aid in diagnostic and prognostic classification of the disease [4–6]. Mutations can often show a higher frequency in particular LPD subtypes such

as mutations of *TCF3* or *ID3* which have been reported in 70% of sporadic BL, mutations of *MYD88* in >90% of Waldenström's macroglobulinemia, or mutations of *TET2*, *IDH2*, and *RHOA* in a large percentage of angioimmunoblastic T cell lymphoma (AITL). From several sequencing studies, specific mutation profiles can define molecular subtypes such as in diffuse large B cell lymphoma (DLBCL) where specific mutations are associated with the germinal center B-cell (GCB) and activated B cell (ABC) subtypes which have pronounced survival differences with standard chemotherapy [7–9]. The presence of *TP53* mutations in chronic lymphocytic leukemia (CLL) has been shown to be an independent prognostic factor and predictor of chemotherapy refractoriness [10]; similarly, *NOTCH1* and *SF3B1* mutations can be independent prognostic markers in CLL [11, 12]. Activating mutations of *NOTCH1* are observed in approximately 60% of T-ALL cases and are reported to be associated with shorter survival in adults [13, 14].

Finally, copy number alterations (CNA) are also prevalent in LPD and can be associated with the underlying biology, with 17p deletion in CLL and PCM being associated with a less favorable outcome. The European Research Initiative on Chronic Lymphocytic Leukemia (ERIC) recommends analysis of del(17p) and *TP53* gene mutations as an integral part of routine diagnostics for CLL patients requiring treatment [15].

The overarching objective of the EuroClonality-NDC is to enable a single NGS test to integrate genomic analyses that are currently performed by a number of molecular testing strategies. As part of the EuroClonality-NGS working group, we have developed the EuroClonality-NGS DNA capture assay (EuroClonality-NDC) to detect clinically relevant genetic alterations in LPD using a capture-hybridization approach. To achieve this objective, EuroClonality-NDC was designed to capture all functional variable (V), diversity (D), and joining (J) genes of the IG and TR loci along with additional probes to identify structural variants (SV) in the form of chromosomal translocations and detect CNA and somatic mutations. The accompanying purpose-built bioinformatics pipeline, ARResT/Interrogate, which was originally developed for amplicon assays, was customized and validated for the EuroClonality-NDC [16]. An optimized standard operating procedure (SOP), which has undergone a multi-site validation, ensures robust assay performance [17]. The development and validation of both the EuroClonality-NDC capture panel and the bioinformatics platform provides an end-to-end workflow which minimizes subjective interpretation of results. The methods detailed in this chapter relate to an updated version of the SOP to reflect recent improvements in library preparation and target enrichment.

2 Materials

2.1 DNA Quantification

The following products and equipment from Thermo Fisher Scientific (Waltham, MA, USA) are required:

1. Qubit dsDNA broad-range (BR) assay.
2. Qubit dsDNA high-sensitivity (HS) assay.
3. Qubit Assay Tubes.
4. Qubit Fluorometer.

2.2 DNA Integrity Assessment

The following products and equipment from Agilent Technologies (Santa Clara, CA, USA) are required:

1. Genomic DNA Reagents.
2. Genomic DNA ScreenTape.
3. D1000 Reagents.
4. D1000 ScreenTape.
5. High Sensitivity D1000 Reagents.
6. High Sensitivity D1000 ScreenTape.
7. 4150/4200 TapeStation System.

2.3 DNA Library Preparation

The following products from Roche Sequencing Solutions (Pleasanton, CA, USA) are required:

1. KAPA HyperPlus Kit.
2. KAPA UDI Primer Mixes.
3. KAPA Universal Adapter.
4. KAPA HyperPure Beads.

The following items will be required for the multiple bead cleanup steps that are performed in both a pre- and post-PCR environment:

1. 96-well magnetic plate.
2. Magnetic stands for 0.2 mL PCR strips.
3. Magnetic stands for 1.5 mL microfuge tubes.

2.4 DNA Hybridization

The following products from Roche Sequencing Solutions (Pleasanton, CA, USA) are required:

1. KAPA HyperCapture Reagent Kit.
2. KAPA HyperCapture Bead.

The following product from Univ8 Genomics Ltd. (Belfast, UK) is required:

1. EuroClonality-NDC.

2.5 Sequencing of Enriched DNA Library

The following products from Illumina, Inc. (San Diego, CA, USA) are required:

1. PhiX Sequencing Control V3.
2. NextSeq 500/550 Mid Output Kit v2.5 (150 cycles).
3. NextSeq 500/550 Sequencing System.

3 Methods

3.1 Genomic DNA Evaluation and Preparation for DNA Library Generation

1. If the extraction of genomic DNA leads to the DNA being eluted into a buffer containing EDTA (*see* **Note 1**), a column or bead-based purification should be performed prior to performing any additional steps as the fragmentation enzyme is sensitive to EDTA.
2. The gDNA concentration is assessed using the Qubit broad range assay. Manufacturer guidelines are followed with two modifications: (1) the standard/sample is added to the Qubit assay tubes first followed by the Qubit working solution, and (2) the incubation time prior to reading the standard/sample is 20 min.
3. The gDNA integrity assessment is performed using the Genomic DNA ScreenTape Assay. Manufacturer guidelines are followed without any modifications.
4. For the EuroClonality-NDC protocol, a positive control, a no template control (NTC), and 22 samples are processed in each batch (*see* **Note 2**). In well A1 of a 96-well PCR plate, place 100 ng of the positive control in a total of 35 μL, and in well A2, place 35 μL of the NTC.
5. For the EuroClonality-NDC assay, 100 ng of high-molecular-weight genomic DNA is required or for genomic DNA extracted from formalin-fixed DNA 100 ng (average fragment size >1000 bp) or 200 ng (average fragment size <1000 bp) is used in a total of 35 μL. Each sample to be prepared should be placed into a separate well of a 96-well PCR plate (*see* **Note 3**).

3.2 DNA Library Generation

1. Remove the following products from the KAPA HyperPlus Kit and thaw on ice:
   (a) KAPA Frag Buffer (10×).
   (b) End Repair & A-Tailing Buffer.
   (c) Ligation Buffer.
   (d) KAPA HiFi HotStart ReadyMix (2×).
   (e) Library Amplification Primer Mix (10×).
2. Prepare a thermocycler by selecting the fragmentation program (Table 1) and pausing prior to the commencement of the first step to ensure the block is pre-cooled to 4 °C (*see* **Note 4**).

Table 1
Fragmentation program

Step	Temperature (°C)	Time (min)	Heated lid (°C)
Pre-cool block	4	2	50
Fragmentation	37	22	
Hold	4	∞	

Table 2
End repair and A-tailing buffer program

Step	Temperature (°C)	Time (min)	Heated lid (°C)
End repair and A-tailing	65	30	85
Hold	4	∞	

3. While keeping the reagents on ice, prepare a mastermix which contains 5 μL KAPA Frag Buffer (10×) and 10 μL KAPA Frag Enzyme for each reaction to be performed.
4. While on ice, add 15 μL of the fragmentation mastermix to the well containing 35 μL double-stranded genomic DNA to achieve a total volume of 50 μL (*see* **Note 5**). Vortex gently before spinning down briefly.
5. Place the reaction in the pre-cooled thermocycler and start the paused fragmentation program.
6. While samples are undergoing fragmentation, prepare the End Repair and A-Tailing Buffer mastermix which contains 7 μL KAPA End Repair & A-tailing Buffer and 3 μL HyperPlus ERAT Enzyme Mix for each reaction to be performed (*see* **Note 6**).
7. Following completion of the fragmentation reaction, place samples onto the plate cooler.
8. Add 10 μL of the End Repair and A-Tailing Buffer mastermix to the well containing 50 μL of fragmented genomic DNA to achieve a total volume of 60 μL. Vortex gently before spinning down briefly.
9. Incubate samples on a thermocycler using the selected End Repair and A-Tailing Buffer program (Table 2).
10. While the End Repair and A-Tailing program is underway, prepare the Ligation mastermix which contains 30 μL Ligation buffer and 10 μL DNA Ligase for each reaction to be performed.

Table 3
Adapter ligation program

Step	Temperature (°C)	Time (min)	Heated lid (°C)
Adapter ligation	20	15	50
Hold	4	∞	

11. Also, while the End Repair and A-Tailing program is underway, KAPA HyperPure Beads are removed from 4 °C to ensure they are equilibrated to room temperature in time for later paramagnetic bead clean up steps (*see* **Note 7**).
12. Following completion of the End Repair and A-Tailing reaction, place samples onto the plate cooler.
13. Add 10 μL of the universal adapter followed by 40 μL Ligation mastermix to each well to achieve a total volume of 110 μL (*see* **Note 8**). Vortex gently before spinning down briefly.
14. Incubate samples on a thermocycler using the selected Adapter Ligation program (Table 3).
15. While the Adapter Ligation program is running, remove the required number of UDI primer mixes from the freezer and thaw on ice.
16. Following completion of the Adapter Ligation thermocycler program, remove samples from the thermocycler.
17. Resuspend the room temperature KAPA HyperPure beads by vortexing vigorously.
18. Perform a 0.8× bead cleanup by adding 88 μL of KAPA HyperPure beads to each well to achieve a total volume of 198 μL before pipette mixing ten times taking care not to generate bubbles.
19. Incubate the bead/sample mixture for 15 min at room temperature to allow the DNA to bind to the beads (*see* **Note 9**).
20. While the bead/sample mixture is incubating, prepare 20 mL of fresh 80% ethanol by adding 4 mL PCR grade water to 16 mL molecular grade ethanol. Vortex and leave at room temperature until required.
21. While the bead/sample mixture is incubating, prepare 5 mL of fresh 10 mM Tris-HCl, pH 8.0, by adding 50 μL 1 M Tris-HCl, pH 8.0, to 4.95 mL PCR grade water. Vortex and leave at room temperature until required.
22. Place samples onto a magnetic stand and wait approximately 5 min for the solution to clear (*see* **Note 10**).
23. Carefully remove and discard the supernatant taking care not to disturb the pellet.

24. With the plate remaining on the magnetic stand, perform an ethanol wash by adding 200 μL of freshly prepared 80% ethanol (*see* **Note 11**).
25. Incubate the sample in 80% ethanol for 30 s.
26. Carefully remove the ethanol taking care not to disturb the pellet.
27. Repeat **steps 24–26** until a total of two ethanol washes have been performed.
28. Remove residual ethanol without disturbing the beads.
29. Air-dry the beads at room temperature to enable evaporation of any remaining ethanol (*see* **Note 12**).
30. Remove the sample from the magnetic stand.
31. Resuspend each bead pellet in 22 μL of 10 mM Tris-HCl, pH 8.0 (*see* **Note 11**).
32. Incubate the sample for 2 min to enable DNA to elute from the beads.
33. Place the sample on the magnetic stand to pellet the beads and for the solution to clear.
34. With the plate remaining on the magnetic stand, transfer 20 μL of the eluate to a new 200 μL PCR plate.
35. To the 20 μL of eluate, add 5 μL of KAPA UDI Primer mix to each individual sample library followed by 25 μL of KAPA HiFi HotStart ReadyMix. Vortex gently before spinning down briefly (*see* **Note 3**).
36. Incubate samples on a thermocycler using the selected Pre-Capture PCR Amplification program (Table 4).
37. Following completion of the Pre-Capture PCR Amplification program, remove samples from the thermocycler.

Table 4
Pre-capture PCR amplification program

Step	Temperature (°C)	Time (min)	Cycles	Heated lid (°C)
Initial denaturation	98	45	1	105
Denaturation	98	15	6	
Annealing	60	30		
Extension	72	30		
Final extension	72	60	1	
Hold	4	∞	1	

38. To the 50 μL PCR reaction, add 70 μL of KAPA HyperPure beads before pipette mixing ten times taking care not to generate bubbles.
39. Incubate the bead/sample mixture for 15 min at room temperature to allow the DNA to bind to the beads (*see* **Note 9**).
40. Place samples onto a magnetic stand and wait approximately 3 min for the solution to clear (*see* **Note 10**).
41. Carefully remove and discard the supernatant taking care not to disturb the pellet.
42. With the plate remaining on the magnetic stand, perform an ethanol wash by adding 200 μL of freshly prepared 80% ethanol (*see* **Note 11**).
43. Incubate the sample in 80% ethanol for 30 s.
44. Carefully remove the ethanol taking care not to disturb the pellet.
45. Repeat **steps 42–44** until a total of two ethanol washes have been performed.
46. Remove residual ethanol without disturbing the beads.
47. Air-dry the beads at room temperature to enable evaporation of any remaining ethanol (*see* **Note 12**).
48. Remove the sample from the magnetic stand.
49. Resuspend each bead pellet in 32 μL of 10 mM Tris-HCl, pH 8.0 (*see* **Note 11**).
50. Incubate the sample for 2 min to enable DNA to elute from the beads.
51. Place the sample on the magnetic stand to pellet the beads and for the solution to clear.
52. With the plate remaining on the magnetic stand, transfer 31 μL of the eluate to a new 200 μL PCR plate, labelled "Master Plate" which is to be retained for preparation of the hybridization.
53. From the 31 μL of the transferred eluate, remove 4 μL of the eluate and transfer to a new plate, labelled "QC Plate" for the purposes of quality control assessment. The Master Plate can be stored at −20 °C until ready to perform the hybridization. Proceed to Subheading 3.3 with the QC Plate.

3.3 Quality Control of DNA Libraries

1. The concentration of each individual library is assessed using the Qubit broad range assay (*see* **Note 13**). Manufacturer guidelines are followed with two modifications: (1) the standard/sample is added to the Qubit assay tubes first followed by the Qubit working solution, and (2) the incubation time prior to reading the standard/sample is 20 min.

2. The average fragment size of each individual library is assessed using the TapeStation D1000 assay. Manufacturer guidelines are followed without any modifications.

3.4 DNA Hybridization

1. Thaw reagents required for the DNA hybridization step which include COT Human DNA, Universal Enhancing Oligos, Hybridization Buffer, and Hybridization Component H.
2. Remove the KAPA HyperPure beads from 4 °C and allow to equilibrate to room temperature for 30 min.
3. For the EuroClonality-NDC protocol, 22 clinical samples are pooled, in equal amounts, into one hybridization reaction to achieve a total of 1.5 μg of DNA (i.e., 68.2 ng of each individual library). To achieve this, calculate the volume of each library to enable 68.2 ng of each library to be added to the hybridization reaction. For the NTC, which should not have a measurable DNA concentration, the average volume of library being added from the 22 samples is calculated to determine the amount of volume of the NTC library to add (*see* **Note 14**).
4. Label a LoBind DNA 1.5 mL tube and add the required volume of each of the 22 libraries and the NTC to this tube. Calculate the total volume of the 22 pooled libraries plus the NTC library. If the total volume of libraries is <45 μL (i.e., libraries), then PCR grade water is added to adjust volume to 45 μL.
5. To the pooled libraries, add 20 μL COT Human DNA. Vortex gently before spinning down briefly.
6. Calculate the total volume of the 22 pooled libraries, the NTC library plus the 20 μL of COT DNA. The volume of beads required in the next step is 2× this total volume (i.e., if the total volume was calculated to be 75 μL, then 150 μL KAPA HyperPure beads will be required).
7. Vortex the KAPA HyperPure beads until a homogenous solution is achieved.
8. To the pooled libraries, add the volume of KAPA HyperPure beads calculated in the **step 6**. Seal the tube and vortex vigorously for 10 s.
9. Incubate the bead/sample mixture for 10 min at room temperature to allow the pooled libraries and COT Human DNA to bind to the beads.
10. Place samples onto a magnetic stand and wait approximately 3 min for the solution to clear (*see* **Note 10**).
11. Carefully remove and discard the supernatant taking care not to disturb the pellet.

12. With the plate remaining on the magnetic stand, perform an ethanol wash by adding 200 μL of freshly prepared 80% ethanol.
13. Incubate the sample in 80% ethanol for 30 s.
14. Carefully remove the ethanol taking care not to disturb the pellet. Remove residual ethanol with an additional pipetting step without disturbing the beads.
15. Air-dry the beads at room temperature for approximately 5 min to enable evaporation of any remaining ethanol (*see* **Note 12**).
16. Remove the sample from the magnetic stand.
17. Add 13.4 μL of Universal Enhancing Oligos (UEO) to the tube, before sealing the tube and vortexing vigorously for 10 s to ensure a homogeneous mixture is achieved.
18. To the library pool and UEO mixture, add 43 μL of mastermix prepared using the following components: 28 μL Hybridization Buffer, 12 μL Hybridization Component H, and 3 μL PCR grade water.
19. Vortex before spinning down briefly. Incubate for 2 min at room temperature.
20. Place samples onto a magnetic stand and wait approximately 3 min for the solution to clear.
21. Transfer 56.4 μL of the eluate into a new well containing 4 μL of the EuroClonality-NDC panel.
22. Vortex vigorously before spinning down briefly.
23. Incubate samples on a thermocycler using the selected "Hybridization" program (Table 5).
24. Dilute wash buffers provided in the KAPA HyperCapture Reagent Kit using the volumes of stock buffer solution and PCR grade water detailed in Table 6.
25. Split the 400 μL of 1× Stringent Wash Buffer into two aliquots of 200 μL in 0.2 mL PCR tubes, and incubate on the thermocycler at 55 °C for at least 15 min.
26. Place the 100 μL aliquot of 1× Wash Buffer I into the thermocycler at 55 °C for at least 15 min.

Table 5
Hybridization program

Step	Temperature (°C)	Time (min)	Heated lid (°C)
Denaturation	95	5 min	105
Hybridization	55	16–20 h	

Table 6
Preparation of post-hybridization wash buffers

Step	Volume of stock buffer (μL)	Volume of PCR grade water (μL)	Temperature (°C)
10× stringent wash buffer	40	360	55
10× wash buffer I	10 20	90 180	55 RT
10× wash buffer II	20	180	RT
10× wash buffer III	20	180	RT
2.5× bead wash buffer	120	180	RT

27. Vortex the Capture Beads from the KAPA HyperCapture Bead kit thoroughly to ensure a homogenous solution.
28. Remove 50 μL of Capture Beads for each pool and place into a 1.5 mL tube, and equilibrate to room temperature for 30 min.
29. Following the 30-min incubation, place the tube containing Capture Beads onto a magnetic stand, and wait approximately 3 min for the solution to clear.
30. Carefully remove and discard the supernatant taking care not to disturb the pellet.
31. Add a volume of 1× Bead Wash Buffer which is twice the original volume of Capture Bead used in **step 28**, to the pelleted Capture Beads.
32. Remove from the magnetic stand and vortex for 10 s before spinning down briefly.
33. Place the tube back onto the magnetic stand and wait until the beads have pelleted and the solution is clear.
34. Carefully remove and discard the supernatant taking care not to disturb the pellet.
35. Perform a second wash of the Capture Beads by performing **steps 31–34** again.
36. Add a volume of 1× Bead Wash Buffer which is the same volume of Capture Beads used in **step 28** to the pelleted Capture Beads.
37. Remove from the magnetic stand and vortex for 10 s before spinning down briefly.
38. Aliquot 50 μL of the resuspended Capture Beads into a 0.2 mL tube for each capture to be performed.
39. Place the tube onto the magnetic stand and wait until the beads have pelleted and the solution is clear (*see* **Note 15**).

40. Carefully remove and discard the supernatant taking care not to disturb the pellet.
41. The Capture Beads are now ready to bind the hybridized DNA.
42. Proceed to the next step immediately to prevent the Capture Beads from drying out.
43. Transfer the hybridization sample (60.4 μL) to the tube containing the pelleted Capture Beads from the previous step (*see* **Note 16**).
44. Vortex for 10 s before spinning down briefly (*see* **Note 17**).
45. Incubate the tube now containing the hybridized DNA and the Capture Beads on the thermocycler at 55 °C for 15 min.
46. After the 15-min incubation, remove the samples and the tube containing the 100 μL aliquot of 1× Wash Buffer from the thermocycler.
47. Add the 100 μL of 55 °C 1× Wash Buffer I before vortexing for 10 s and spinning down briefly.
48. Place the tube onto the magnetic stand and wait until the beads have pelleted and the solution is clear.
49. Carefully remove and discard the supernatant taking care not to disturb the pellet.
50. Remove a tube containing a 200 μL aliquot of 1× Stringent Wash Buffer from the thermocycler, and add the 200 μL of 1× Stringent Wash Buffer to the sample.
51. Remove the sample from the magnet before vortexing for 10 s.
52. Incubate the sample on the thermocycler at 55 °C for 5 min.
53. After the incubation, spin the sample briefly before placing the sample onto the magnetic stand. Wait until the beads have pelleted and the solution is clear.
54. Carefully remove and discard the supernatant taking care not to disturb the pellet.
55. Repeat **steps 50–54** with the only remaining 200 μL aliquot of pre-warmed 1× Stringent Wash Buffer.
56. Add 200 μL of room temperature 1× Wash Buffer I to the sample followed by vortexing for 10 s.
57. Incubate the sample for 1 min at room temperature.
58. After the incubation, spin the sample briefly before placing the sample onto the magnetic stand. Wait until the beads have pelleted and the solution is clear.
59. Carefully remove and discard the supernatant taking care not to disturb the pellet.
60. Add 200 μL of room temperature 1× Wash Buffer II to the sample followed by vortexing for 10 s.

61. Incubate the sample for 1 min at room temperature.
62. After the incubation, spin the sample briefly before placing the sample onto the magnetic stand. Wait until the beads have pelleted and the solution is clear.
63. Carefully remove and discard the supernatant taking care not to disturb the pellet.
64. Add 200 μL of room temperature 1× Wash Buffer III to the sample followed by vortexing for 10 s.
65. Incubate the sample for 1 min at room temperature.
66. After the incubation, spin the sample briefly before placing the sample onto the magnetic stand. Wait until the beads have pelleted and the solution is clear.
67. Carefully remove and discard the supernatant taking care not to disturb the pellet.
68. Add 20 μL of PCR grade water to the sample followed by vortexing for 10 s and subsequently spin the sample briefly.
69. Remove KAPA HyperPure Beads (70 μL required for each capture) for use in later steps within a post-PCR area, and allow to equilibrate to room temperature.
70. For each hybridization, add to a fresh 0.2 mL tube 25 μL of KAPA HiFi HotStart ReadyMix and 5 μL of Post-Capture PCR Oligos.
71. Add the 20 μL of bead-bound captured DNA from **step 68** to the 0.2 mL tube containing the PCR reagents to achieve a total volume of 50 μL. Mix thoroughly by pipette mixing.
72. Within a post-PCR designated area, place samples on a thermocycler and run the selected Post-Capture PCR Amplification program detailed in Table 7 (*see* **Note 18**).

Table 7
Post-capture PCR amplification program

Step	Temperature (°C)	Time (min)	Cycles	Heated lid (°C)
Initial denaturation	98	45	1	105
Denaturation	98	15	11	
Annealing	60	30		
Extension	72	30		
Final extension	72	60	1	
Hold	4	∞	1	

73. While the Post-Capture amplification is underway, prepare 80% ethanol by adding 200 μL PCR grade water to 800 μL molecular grade ethanol in a 1.5 mL microfuge tube. Vortex gently before spinning down briefly.
74. Following completion of the Post-Capture amplification thermocycler program, vortex the KAPA HyperPure Beads which are now equilibrated to room temperature.
75. Add 70 μL KAPA HyperPure Beads to each 50 μL PCR reaction which contains the amplified and enriched DNA library pool before vortexing for 10 s and spinning down briefly.
76. Incubate the bead/sample mixture for 15 min at room temperature to allow the sample to bind to the beads.
77. Place samples onto a magnetic stand and wait approximately 3 min for the solution to clear.
78. Carefully remove and discard the supernatant taking care not to disturb the pellet.
79. With the plate remaining on the magnetic stand, perform an ethanol wash by adding 200 μL of freshly prepared 80% ethanol.
80. Incubate the sample in 80% ethanol for 30 s.
81. Carefully remove the ethanol taking care not to disturb the pellet.
82. Repeat **steps 79–81** for a total of 2 ethanol washes.
83. Carefully remove any residual ethanol with an additional pipetting step without disturbing the beads.
84. Air-dry the beads at room temperature for approximately 5 min to enable evaporation of any remaining ethanol (*see* **Note 12**).
85. Remove the sample from the magnetic stand.
86. Resuspend the bead pellet in 22 μL of PCR grade water before vortexing for 10 s and spinning down briefly (*see* **Note 17**).
87. Incubate the sample for 2 min to enable DNA to elute from the beads.
88. Place the sample on the magnetic stand to pellet the beads and for the solution to clear.
89. With the plate remaining on the magnetic stand, transfer 20 μL of the eluate to a new 200 μL PCR plate.
90. The amplified and enriched library is now ready for the final quality control steps prior to sequencing.

3.5 Quality Control of Enriched DNA Library

1. The concentration of the amplified and enriched library is assessed using the Qubit high sensitivity assay. Manufacturer guidelines are followed with two modifications: (1) the

standard/sample is added to the Qubit assay tubes first followed by the Qubit working solution, and (2) the incubation time prior to reading the standard/sample is 20 min.

2. The average fragment size of each individual library is assessed using the High Sensitivity D1000 TapeStation assay. Manufacturer guidelines are followed without any modifications.

3.6 Sequencing of Enriched DNA Library

1. Prepare the amplified and enriched pooled library for sequencing on the NextSeq 500/550 by adhering to the Illumina "Denature and Dilute Guidelines" using the following parameters (*see* **Note 19**):
   (a) Protocol A (Standard Normalization Method).
   (b) Final dilution of library is to 1.5 pM for Mid Output kits.
   (c) Final PhiX (sequencing control) spike-in percentage is 1% of the final library and PhiX composition.

3.7 Bioinformatic Analysis of the Sequencing Data Using ARResT/Interrogate

1. The bioinformatics pipeline, ARResT/Interrogate, can be accessed at arrest.tools/interrogate and requires an account which can be created by emailing contact@arrest.tools (*see* **Note 20**).
2. Once logged in to the ARResT/Interrogate bioinformatics website, switch to the "Interrogate.EC NDC" user mode at the top left of the user interface.
3. Switch to the "processing" tab, and follow the instructions to upload your samples in compressed FASTQ format; the extension should be ".fastq.gz" (*see* **Note 21**).
4. Following upload of the FASTQ files, no further options require selection.
5. Click on blue "test it" button and if the subsequent response is "OK," click on the green "process" button.
6. Progress of the bioinformatic pipeline can be followed using the user interface or an email can be sent to notify user of completion of the analysis.
7. When the run is complete, follow the instructions on the user interface to retrieve the result files.
8. The file with extension ".gathered.xlsx" contains results for samples contained within one analysis and is the main output. It contains panel-supported events, i.e., rearrangements, translocations, and somatic mutations (*see* **Note 22**).
9. Open the file with extension ".gathered.xlsx" with Microsoft Excel or a suitable equivalent. Column descriptions are provided.

4 Notes

1. An alternative method to address EDTA-containing genomic DNA samples is to add 5 μL of a conditioning solution, provided in the KAPA HyperPlus Kit to genomic DNA in a total volume of 30 μL. The conditioning solution, provided in the KAPA HyperPlus Kit, is diluted to a concentration dependent on the EDTA concentration in the DNA sample.
2. NTC is either PCR grade water or the same buffer used to elute the gDNA following DNA extraction. Users can use commercially sourced high-molecular-weight gDNA as a positive control to monitor the consistency in performance of library preparation across multiple batches.
3. It is worthwhile generating a template to record the order the samples are added to the 96-well plate and for later stages of library preparation to record the UDI primer mix assigned to each sample. Plates can be prepared the day before to minimize set up times on the day of library preparation.
4. Fragmentation is key to the size distribution of the final library and is impacted by fragmentation time and temperature. With the wide range of different thermocyclers on the market, optimization of the fragmentation time is advised to ensure the ideal size distribution profile following fragmentation is achieved.
5. Maintaining the temperature of the reaction at 4 °C during the setup of the fragmentation reaction is critical, and it is advised using a PCR cooler to ensure fragmentation does not being prior to loading the plate on the thermocycler.
6. After thawing, the End Repair and A-Tailing buffer may contain precipitates which may require incubation at 37 °C and thorough vortexing before use to ensure they have been completely resuspended.
7. Aliquots of KAPA HyperPure Beads can be made to reduce the number of times the KAPA HyperPure Beads are removed from storage at 4 °C.
8. Universal adapter stocks are aliquoted to avoid repeated freeze/thaw cycles.
9. We use a 15-min incubation period with beads to ensure maximal recovery of library in samples with a poor gDNA integrity.
10. Various plate magnets are available on the market with different locations of the magnets and variable magnetic strengths. This variation can impact on the time required for beads to pellet, and it is worthwhile determining optimal times for incubation of bead containing samples for specific magnetic plates.

11. Steps such as ethanol washes of beads and elution of beads should be performed using a multichannel pipette to ensure consistency and prevent beads drying out.
12. Air-drying of beads is dependent on room temperature. The bead pellet should still be dark brown and glossy but show little sign of excess liquid. Over-drying of beads can lead to poor elution of DNA from the beads and therefore lower yields.
13. With this version of the library preparation method, the concentration of individual libraries tends to lie within the range of the Qubit broad range assay.
14. Inclusion of the positive control for the hybridization steps is not required. While the NTC is included in the hybridization reaction, the NTC DNA concentration should be negligible and not factored into the calculations for the combined DNA mass of 1.5 μg.
15. Different magnetic stands will be required for the workflow and some will be required in both pre- and post-PCR environments. From our experience, sourcing good magnetic stands is essential to the success of the workflow.
16. Keep the "Hybridization" program on thermocycler running for incubations with the Capture Beads and subsequent wash steps.
17. This step is to ensure all the sample-bead mix is at the bottom of the well for the incubation steps on the thermocycler. Do not spin long enough for the beads to pellet at the bottom of the tube.
18. It is advised the Post-Capture PCR and all further steps are performed in an area designated for post-PCR activities to minimize contamination risks.
19. The EuroClonality-NDC protocol was developed and validated to achieve an optimal mean target coverage depth for the detection of clonal rearrangements, translocation, copy number alterations, and single nucleotide variant/indels in 22 samples, using a single hybridization reaction and sequenced on an Illumina NextSeq 500/550 system with the Mid Output kit sequencing reagents. It is at the readers' discretion if they want to adapt the protocol to:
    (a) Employ different Illumina sequencing platforms.
    (b) Utilize sequencing reagents with increased output.
    (c) Alter the number of samples being applied to the flow cell for sequencing.
20. The method of accessing the bioinformatics pipeline may change. The latest version will always be available either through the user interface of ARResT/Interrogate or through the authors.

21. Dependent on the number of samples and quality of internet connection, this can take time.
22. Additional files are available for in-depth analysis such as a comprehensive QC metrics summary, BAM files, CNV files and VCF files.

Acknowledgments

We would like to acknowledge the EuroClonality-NGS working group and, in particular, those members that directly or indirectly contributed to the validation of the EuroClonality-NDC assay. The EuroClonality-NGS Working Group is an independent scientific subdivision of EuroClonality that aims at innovation, standardization, and education in the field of diagnostic clonality analysis. The revenues of the previously obtained patent (PCT/NL2003/000690), which is collectively owned by the EuroClonality Foundation and licensed to InVivoScribe, are exclusively used for EuroClonality activities, such as for covering costs of the Working Group meetings, collective Work Packages, and the EuroClonality Educational Workshops. The EuroClonality consortium operates under an umbrella of ESLHO, which is an official EHA Scientific Working Group.

References

1. Swerdlow SH, Campo E, Pileri SA, Harris NL, Stein H, Siebert R et al (2016) The 2016 revision of the World Health Organization classification of lymphoid neoplasms. Blood 127(20): 2375–2390. https://doi.org/10.1182/blood-2016-01-643569
2. van Dongen JJ, Langerak AW, Bruggemann M, Evans PA, Hummel M, Lavender FL et al (2003) Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98-3936. Leukemia 17(12):2257–2317. https://doi.org/10.1038/sj.leu.2403202
3. Langerak AW, Groenen PJ, Bruggemann M, Beldjord K, Bellan C, Bonello L et al (2012) EuroClonality/BIOMED-2 guidelines for interpretation and reporting of Ig/TCR clonality testing in suspected lymphoproliferations. Leukemia 26(10):2159–2171. https://doi.org/10.1038/leu.2012.246
4. Pestinger V, Smith M, Sillo T, Findlay JM, Laes JF, Martin G et al (2020) Use of an integrated pan-cancer oncology enrichment next-generation sequencing assay to measure tumour mutational burden and detect clinically actionable variants. Mol Diagn Ther 24(3): 339–349. https://doi.org/10.1007/s40291-020-00462-x
5. Bratman SV, Newman AM, Alizadeh AA, Diehn M (2015) Potential clinical utility of ultrasensitive circulating tumor DNA detection with CAPP-Seq. Expert Rev Mol Diagn 15(6): 715–719. https://doi.org/10.1586/14737159.2015.1019476
6. Cheng DT, Mitchell TN, Zehir A, Shah RH, Benayed R, Syed A et al (2015) Memorial sloan kettering-integrated mutation profiling of actionable cancer targets (MSK-IMPACT): a hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. J Mol Diagn 17(3): 251–264. https://doi.org/10.1016/j.jmoldx.2014.12.006
7. Richter J, Schlesner M, Hoffmann S, Kreuz M, Leich E, Burkhardt B et al (2012) Recurrent mutation of the ID3 gene in Burkitt lymphoma identified by integrated genome, exome and transcriptome sequencing. Nat Genet 44(12):

1316–1320. https://doi.org/10.1038/ng.2469

8. Lopez C, Kleinheinz K, Aukema SM, Rohde M, Bernhart SH, Hubschmann D et al (2019) Genomic and transcriptomic changes complement each other in the pathogenesis of sporadic Burkitt lymphoma. Nat Commun 10(1):1459. https://doi.org/10.1038/s41467-019-08578-3
9. Karube K, Enjuanes A, Dlouhy I, Jares P, Martin-Garcia D, Nadeu F et al (2018) Integrating genomic alterations in diffuse large B-cell lymphoma identifies new relevant pathways and potential therapeutic targets. Leukemia 32(3):675–684. https://doi.org/10.1038/leu.2017.251
10. Gonzalez D, Martinez P, Wade R, Hockley S, Oscier D, Matutes E et al (2011) Mutational status of the TP53 gene as a predictor of response and survival in patients with chronic lymphocytic leukemia: results from the LRF CLL4 trial. J Clin Oncol 29(16):2223–2229. https://doi.org/10.1200/JCO.2010.32.0838
11. Rossi D, Rasi S, Spina V, Bruscaggin A, Monti S, Cresta S et al (2012) The genome of chemorefractory chronic lymphocytic leukemia reveals frequent mutations of NOTCH1 and SF3B1. Leukemia Suppl 1(Suppl 2):S26–S28. https://doi.org/10.1038/leusup.2012.16
12. Oscier DG, Rose-Zerilli MJ, Winkelmann N, Gonzalez de Castro D, Gomez B, Forster J et al (2013) The clinical significance of NOTCH1 and SF3B1 mutations in the UK LRF CLL4 trial. Blood 121(3):468–475. https://doi.org/10.1182/blood-2012-05-429282
13. Weng AP, Ferrando AA, Lee W, Morris JP, Silverman LB, Sanchez-Irizarry C et al (2004) Activating mutations of NOTCH1 in human T cell acute lymphoblastic leukemia. Science 306(5694):269–271. https://doi.org/10.1126/science.1102160
14. Zhu YM, Zhao WL, Fu JF, Shi JY, Pan Q, Hu J et al (2006) NOTCH1 mutations in T-cell acute lymphoblastic leukemia: prognostic significance and implication in multifactorial leukemogenesis. Clin Cancer Res 12(10):3043–3049. https://doi.org/10.1158/1078-0432.CCR-05-2832
15. Malcikova J, Tausch E, Rossi D, Sutton LA, Soussi T, Zenz T et al (2018) ERIC recommendations for TP53 mutation analysis in chronic lymphocytic leukemia-update on methodological approaches and results interpretation. Leukemia 32(5):1070–1080. https://doi.org/10.1038/s41375-017-0007-7
16. Bystry V, Reigl T, Krejci A, Demko M, Hanakova B, Grioni A et al (2017) ARResT/Interrogate: an interactive immunoprofiler for IG/TR NGS data. Bioinformatics 33(3):435–437. https://doi.org/10.1093/bioinformatics/btw634
17. Stewart JP, Gazdova J, Darzentas N, Wren D, Proszek P, Fazio G (2021) EuroClonality-NGS Working Group. Validation of the EuroClonality-NGS DNA capture panel as an integrated genomic tool for lymphoproliferative disorders. Blood Adv 5(16):3188–3198. https://doi.org/10.1182/bloodadvances.2020004056 PMID: 34424321; PMCID: PMC8405189

Chapter 10

Immunoglobulin Gene Mutational Status Assessment by Next Generation Sequencing in Chronic Lymphocytic Leukemia

Anne Langlois de Septenville, Myriam Boudjoghra, Clotilde Bravetti, Marine Armand, Mikaël Salson, Mathieu Giraud, and Frederic Davi

Abstract

B cell receptor (BcR) immunoglobulins (IG) display a tremendous diversity due to complex DNA rearrangements, the V(D)J recombination, further enhanced by the somatic hypermutation process. In chronic lymphocytic leukemia (CLL), the mutational load of the clonal BcR IG expressed by the leukemic cells constitutes an important prognostic and predictive biomarker. Here, we provide a reliable methodology capable of determining the mutational status of IG genes in CLL using high-throughput sequencing, starting from leukemic cell DNA or RNA.

Key words Chronic lymphocytic leukemia, Immunoglobulin genes, Next generation sequencing, Somatic hypermutation analysis, Mutational status

1 Introduction

Chronic lymphocytic leukemia (CLL) is a malignant clonal proliferation of mature B cells. It is the most frequent leukemia in adults in the Western world and is characterized by a marked clinical heterogeneity. For some patients, it is an indolent disease with no or only late need of treatment, while in others it displays an aggressive behavior requiring early initiation of therapy [1]. Many prognostic factors have been identified, and among them, the mutational status of the immunoglobulin heavy chain variable (IGHV) genes of the B cell receptor (BcR) has emerged as one of the most robust parameters [2]. It has several advantages as it is stable and can be evaluated at any time including at diagnosis and is independent of other clinical or biological factors [3]. In addition, it has also proved to be a predictive factor of response to chemoimmunotherapy [4, 5]. Therefore the recent guidelines from the International Workshop on CLL recommend that determination

Anton W. Langerak (ed.), *Immunogenetics: Methods and Protocols*, Methods in Molecular Biology, vol. 2453, https://doi.org/10.1007/978-1-0716-2115-8_10,

of the IGHV mutational status should be performed before treatment initiation both in clinical trials and in general practice [6].

The BcR IG display huge diversity in their variable regions which results from complex mechanisms: (1) assembly of variable (V), diversity (D), and joining (J) genes, (2) imprecise junction of these rearranged genes with random nucleotide insertion and deletions, and (3) pairing of heavy and light chains [7]. Further diversification occurs after antigen encounter by somatic hypermutation in the V regions coupled with affinity maturation of the BcR [8]. In tumors such as CLL, all leukemic cells bear the same clonal BcR which reflects the developmental stage from which they derive and constitutes a biomarker of the disease.

Determination of IGHV mutational status is achieved by sequencing the IGHV gene from the clonal IGH rearrangement of the leukemic cells, followed by its comparison with the closest germline counterpart from which it derives [9]. An identity <98% classifies the CLL as "mutated" which is associated with a favorable outcome, while an identity ≥98% defines "unmutated" CLL and confers a poor prognosis [10, 11]. Since this initial observation in 1999, numerous studies have confirmed that unmutated CLL have shorter time-to-first treatment and overall survival when compared to mutated cases [2, 3]. In addition, large-scale repertoire analyses have shown that CLL display a skewed IG repertoire with a sizeable fraction of patients sharing quasi-identical IG variable heavy chain regions sequences, a phenomenon termed BcR IG stereotypy [12]. Importantly, some of these CLL cases belonging to the same stereotypic groups (or subsets) may also share similar clinical and biological features, separating them from other patients with the same IGHV mutational status [13, 14]. Therefore, BcR IG stereotypy further refines the categorization into mutated or unmutated CLL.

The European Research Initiative on CLL (ERIC) has published methodological guidelines and recommendations on how to perform and interpret IGHV mutational status in CLL [15]. The first step consists in polymerase chain reaction (PCR) amplification of clonal IGH rearrangements. Importantly, as the whole IGHV gene sequence is necessary for accurate calculation of the somatic hypermutation load, 5′ primers need to be positioned upstream, e.g., on the leader peptide. Both genomic DNA (gDNA) and RNA extracted from leukemic cells can serve as templates, with gDNA having the advantage of being a more robust material, simpler to obtain and also a source for other genomic investigations. However, in a fraction of cases, amplification from gDNA is hampered by the presence of somatic hypermutation in the primer binding sites. Although starting from RNA requires an additional step of reverse transcription (RT), this can be a useful alternative or complementary approach as it allows the use of primers binding to sequences less or not targeted by somatic hypermutation upstream

and downstream of the IGHV-IGHD-IGHJ rearrangement, respectively, in the leader region L1 part and the constant regions.

Sequencing of the IGH rearrangements amplicons was traditionally performed by Sanger methodology. However, with the constant advance of next generation sequencing (NGS) in the diagnostic field, there is a need to adapt this technology to IGHV mutational determination [16, 17]. Here, we describe detailed protocols for NGS-based determination of the IGHV mutational in CLL, starting from either gDNA or cDNA templates.

2 Materials

2.1 Sample Preparation

1. 50 mL polypropylene tubes.
2. UNI-SEPMaxi (EUROBIO Scientific).
3. Phosphate-buffered saline (1×PBS) pH 7.4 (Thermo Fisher Scientific).
4. Centrifuge.
5. 1.5 mL microfuge tubes.
6. Blood and Cell Culture DNA kit (Qiagen).
7. RNAeasy Mini Kit (Qiagen).
8. Nanodrop ND1000 (Thermo Fisher Scientific).
9. Nuclease-free water (Promega).
10. Thermocycler.

2.2 Primer Preparation

1. Primers (Eurogentec).
2. Nuclease-free water (Promega).
3. 1.5 mL microfuge tubes (Eppendorf).

2.3 PCR Amplification

1. PCR primers (*see* Tables 1 and 2).
2. High Fidelity Platinum® Taq DNA Polymerase, 5 U/μL (with 10× High Fidelity PCR Buffer and 50 mM $MgSO_4$) (Thermo Fisher Scientific).
3. dNTP Mix, 10 mM (Thermo Fisher Scientific).
4. Nuclease-free water (Promega).
5. 0.2 mL 96-well PCR plate (AB-0600) (Thermo Fisher Scientific).
6. Adhesive sealing sheets (Dominique Dutscher).
7. Thermocycler (typically: Applied Biosystems Veriti 96).

2.4 PCR Product Purification

1. Agencourt AMPure XP beads (Beckman Coulter).
2. Ethanol absolute (VWR).

Table 1
Primers for gDNA template (sequence composition : flow cell binding adapter_[barcode]_sequencing primer site_gene-specific primer)

Forward primers (IGHV-leader)	
Flow cell binding adapter_[barcode]_sequencing sequence	
	aatgatacggcgaccaccgagatctacac_[barcode]_acactctttccctacacgacgctcttccgatct_
IGHV-L2 sequence	
IGHL2_1.1	GTGTTCTCTCCACAGGAGCC
IGHL2_1.2	GTGTCTTCTCTACAGGTGCCCA
IGHL2_1.3	GTGTTCTCTCCACAGGTGCC
IGHL2_1.4	GTGTCCTCTCCACAGGTGCC
IGHL2_1.5	CTGTCCTCTCCACAGGCACC
IGHL2_1.6	GTGTCCCCTCCACAGATGC
IGHL2_1.7	GTGTCCTCTCCGCAGGTG
IGHL2_1.8	GTGTCCTCTCCACAGGTGTCCAGTCC
IGHL2_1.9	TTCTCTTCTCCACAGGCACC
IGHL2_2.1	CTTATGTCTTCTCCACAGGGGTC
IGHL2_2.2	CTTATGCTTTCTCCACAGGGGT
IGHL2_3.1	TGTGTTTGCAGGTGTCCAGTG
IGHL2_3.2	TGTGTTTGCAGCTGTCCAGTG
IGHL2_3.3	TCTGTTTGCAGGTGTCCAGTG
IGHL2_3.4	TTTGTTTGCAGGTGTCCAGTG
IGHL2_3.5	TGTGTTTGCAGGTGTCCAATG
IGHL2_3.6	CGTGTTTGCAGGTGTCCAGT
IGHL2_4.1	GTCTCTCTGTTCACAGGGGTCC
IGHL2_4.2	GTTTCTCTGTTCACAGGGGTCC
IGHL2_4.3	GTTTTTCTGTTCACAGGGGTCC
IGHL2_5.1	TCTCCCCCACAGGAGTCTGT
IGHL2_5.2	TCTTCCATACAGGAGTCTGTGC
IGHL2_61	TGTCTCCAGGTGTCCTGTCAC
IGHL2_7.1	CTTCATGCACTCCCATCTCCT
Reverse primer (IGHJ)	
	caagcagaagacggcatacgagat_[barcode]_gtgactggagttcagacgtgtgctcttccgatct
IGHJ sequence	
IGHJ	CTTACCTGAGGAGACGGTGACC

Table 2
Primers for cDNA template (sequence composition : flow cell binding adapter_[barcode]_ sequencing primer site_gene-specific primer)

Forward primers (IGHV-leader)	
Flow cell binding adapter_[barcode]_sequencing sequence	
	aatgatacggcgaccaccgagatctacac_[barcode]_acactctttccctacacgacgctcttccgatct
IGHV-L1 sequence	
IGHL1_1	CTCACCATGGACTGSAYYTGGAG
IGHL1_2	ATGGACAYACTTTGYTMCACRCTCC
IGHL1_3	ATGGARTTKGGGCTKWGCTGGGTTT
IGHL1_4	CTGTGGTTCTTYCTBCTSCTGGTGG
IGHL1_5	CCTCCTCCTRGCTRTTCTCCAAG
IGHL1_6	CTGTCTCCTTCCTCATCTTCCTGCC
Reverse primer (IGHC)	
Flow cell binding adapter_[barcode]_sequencing sequence	
	caagcagaagacggcatacgagat_[barcode]_gtgactggagttcagacgtgtgctcttccgatct
IGHC sequence	
IGHC_mu	GGTTGGGGCGGATGCACT
IGHC_gamma	CGATGGGCCCTTGGTGGA

3. Magnetic Stand-96 (Invitrogen).
4. Nuclease-free water (Promega).
5. 0.2 mL 96-well PCR plate (AB-0600) (Thermo Fisher Scientific).
6. Microplate shaker (Eppendorf).
7. Microplate centrifuge.

2.5 Quantification of Purified PCR Products

1. Quant-iT™ dsDNA High-Sensitivity Assay Kit (Thermo Fischer Scientific).
2. Microplate fluorescence reader, such as Clariostar (BMG Labtech).
3. MicroPlate 96-well, F-bottom (chimney well), black (Greiner).
4. Adhesive PCR sealing foil sheets (Thermo Fisher Scientific).
5. Microplate shaker (Eppendorf).
6. Microplate centrifuge.

2.6 Library Preparation

1. 0.2 mL 96-well PCR plate (AB-0600) (Thermo Fisher Scientific).
2. 1.5 mL microfuge tubes (Eppendorf).

2.7 Library Denaturation and Illumina MiSeq Sequencing

1. Sodium hydroxide (NaOH) 1 N (VWR).
2. Nuclease-free water (Promega).
3. 1.5 mL microfuge tubes (Eppendorf).
4. EBT: 10 mM Tris-Cl, pH 8.5 (Buffer EB Qiagen) with 0.1% Tween 20 (Euromedex).
5. PhiX control (10 nM) (Illumina).
6. MiSeq Reagent Kit v3 (600 cycles) (Illumina) including cartridge, HT1 buffer, flow cell and sequencing buffer.
7. MiSeq System (Illumina).

2.8 Bioinformatics Analysis

1. Vidjil platform account (support@vidjil.org).

3 Methods

3.1 Template Preparation (See Note 1)

3.1.1 Lymphocyte Isolation from Peripheral Blood Using Density Gradient Separation

1. Slowly add 10–17.5 mL of blood to a UNI-SEP Maxi tube.
2. Centrifuge at 1000 × *g* for 15 min.
3. Collect the mononuclear cell ring above the membrane and transfer to a 50 mL tube; fill up to 50 mL with PBS 1×.
4. Centrifuge at 600 × *g* for 10 min, and then discard the supernatant.
5. Resuspend the cell pellet in 1 mL of PBS and then fill the tube with PBS 1×.
6. Centrifuge at 600 × *g* for 10 min, and then discard the supernatant.
7. Repeat **steps 5** and **6** of this section.
8. Transfer the cell pellet into a 1.5 mL microfuge tube and remove all remaining supernatants.

3.1.2 Genomic DNA Extraction

1. Extract gDNA from cell pellets or tissue biopsy with the Qiagen DNA kit following the manufacturer's instructions.
2. Quantify DNA by spectrophotometry (Nanodrop) and adjust to a final working concentration of 20 ng/mL with nuclease-free water.

3.1.3 RNA Extraction and cDNA Synthesis

1. Extract RNA with the Qiagen RNAeasy Mini kit according the manufacturer's instructions and then quantify on a Nanodrop spectrophotometer.

2. Dilute 1 μg of RNA in a microcentrifuge tube with nuclease-free water in a 10 μL total volume and incubate 10 min at 70 °C.
3. Add 10 μL of the RT mix containing: 2 μL RT buffer (10×), 0.8 μL dNTP mix (25×), 2 μL random primers (10×), 1 μL MultiScribe reverse transcriptase (50 U/μL), 1 μL RNAase inhibitor (20 U/μL), 3.2 μL nuclease-free water.
4. Place in a thermocycler with the following program: 10 min at 25 °C, 120 min at 37 °C, 5 min at 85 °C, 4 °C on hold.
5. Add 30 μL nuclease-free water.
6. At this stage the cDNA mixture can be stored at −80 °C (*see* **Note 2**).

3.2 Primer Preparation

Primer sequences are indicated in Table 1 (*see* **Notes 3** and **4**).

3.2.1 Primer Preparation for gDNA Template

1. Prepare a 100 μM forward primer mix by pooling each of the 24 IGHV-Leader L2-part primers, all bearing the same barcode, in a microcentrifuge tube. Further dilute this primer mix with nuclease-free water to a final 20 μM concentration.
2. Dilute the reverse IGHJ primer with nuclease-free water to a final 5 μM concentration (*see* **Note 5**).

3.2.2 Primer Preparation for cDNA Template

Primer sequences are indicated in Table 2 (*see* **Notes 3** and **4**).

1. Prepare a 100 μM forward primer mix by pooling each of the 6 IGHV-Leader L1-part primers, all bearing the same barcode, in a microcentrifuge tube.
2. Prepare a 100 μM reverse primer mix by pooling each of the 2 IGHC primers, all bearing the same barcode, in a microcentrifuge tube. Further dilute these primer mixes with nuclease-free water to a final 5 μM concentration (*see* **Note 5**).

3.3 PCR Amplification of IGH Rearrangements

1. Thaw, mix, and briefly centrifuge each component before use.
2. Prepare a PCR master mix by adding the components as shown in Table 3 for gDNA or Table 4 for cDNA.
3. Dispense 40 μL of this mix in each well of the plate.
4. Add 3 μL of forward primer mix.
5. Add 2 μL of reverse primer mix.
6. Add 5 μL of gDNA or cDNA template.
7. Seal the plate with adhesive sheet.
8. Shake briefly and centrifuge (short pulse).
9. Place the plate in a thermocycler. The PCR program is the following: denaturation at 95 °C for 3 min; 35 cycles of 95 °C for 45 s, 63 °C for 45 s, 68 °C for 1 min; final extension at 68 °C for 10 min; 12 °C on hold.

Table 3
PCR mix for gDNA template

Reagents	Volume per reaction	Final concentration
PCR master mix (40 µL per reaction)		
10× High Fidelity PCR Buffer	5 µL	1x
50 mM $MgSO_4$	3.5 µL	3.5 mM
10 mM dNTP mix	1 µL	0.2 mM
Platinum® HighFidelity Taq Polymerase (5 U/µL)	0.2 µL	1 U
PCR-grade water	30.3 µL	–
Primers		
20 µM forward primer mix	3 µL	1.2 µM
5 µM reverse primer	2 µL	0.2 µM
Template		
20 ng/µL gDNA	5 µL	100 ng

Table 4
PCR mix for cDNA template

Reagents	Volume per reaction	Final concentration
PCR master mix (40 µL per reaction)		
10× High Fidelity PCR Buffer	5 µL	1×
50 mM $MgSO_4$	3.5 µL	3.5 mM
10 mM dNTP mix	1 µL	0.2 mM
Platinum® HighFidelity Taq Polymerase (5 U/µL)	0.2 µL	1 U
PCR-grade water	30.3 µL	–
Primers		
5 µM forward primer mix	3 µL	0.3 µM
5 µM reverse primer	2 µL	0.2 µM
Template		
20 ng/µL gDNA	5 µL	100 ng

10. At this stage, the plate can be sealed and stored at −20 °C for later usage.

3.4 PCR Product Purification

1. Preparation.
   (a) Prepare fresh 70% ethanol for optimal results.
   (b) Agencourt AMPure XP bottle should be used at room temperature.

2. Centrifuge briefly the PCR plate.
3. Shake the Agencourt AMPure XP bottle to resuspend the magnetic beads before adding 37.5 μL per well of the PCR plate (*see* **Note 6**).
4. Mix thoroughly by pipetting until the mixture appears homogeneous.
5. Incubate for 5 min at room temperature.
6. Place the reaction plate on the magnetic stand for 5 min.
7. Remove and discard the cleared supernatant (80 μL).
8. Wash the beads by dispensing 200 μL of 70% ethanol (freshly prepared) to each well of the reaction plate, and incubate for 30 s at room temperature; then aspirate and discard the ethanol (200 μL).
9. Repeat for a total of two washes.
10. Dry 5 min at room temperature to ensure all traces of ethanol are removed.
11. To elute purified DNA fragments from beads, remove the reaction plate from the magnetic stand, and then add 35 μL of nuclease-free water to each well of the reaction plate and mix by pipetting until beads are completely resuspended.
12. Incubate for 5 min at room temperature.
13. Place the reaction plate onto the magnetic plate for 2 min to collect the beads.
14. Transfer 25 μL of the eluate to a new microplate.
15. At this stage, the plate can be sealed and stored at −20 °C for later usage.

3.5 Quantification of Purified PCR Products (See Note 7)

1. Dilute Quant-iT™ dsDNA HS reagent 1:200 in Quant-iT™dsDNA HS buffer (sufficient quantity for all PCR samples plus 8 standards and 1 blank).
2. Load 200 μL of the working solution into each microplate well.
3. Add 10 μL of each dsDNA HS standards or 2.5 μL of each PCR sample.
4. Place on the plate shaker at 1200 rpm for 5 min.
5. Briefly spin in a centrifuge.
6. Measure the fluorescence using the Clariostar microplate reader.
7. Use the standard curve to determine the DNA amounts (*see* **Note 8**).

3.6 Library Preparation and Quantification

1. Calculate the concentration in nM according to the formula:

$$[\text{concentration}(\text{ng}/\mu\text{L})/(\text{size amplicon}(\approx 550\,\text{bp}) \times 650] \times 10^6.$$

2. For samples with purified PCR products >10 nM:
   (a) Dilute samples in a new microplate to final 10 nM concentration (5 μL PCR product + H_2O up to 10 nM using the formula: vol. $H_2O = [\text{conc}(\text{nM})/2] - 5$.
   (b) Use 5 μL of this 10 nM dilution for pooling samples in a microfuge tube.
3. For samples with purified PCR products <10 nmm, use 10 μL for library pooling in the same microfuge tube.
4. At this stage, the tube containing the library pool can be sealed and stored at −20 °C for later use.

3.7 Library Denaturation and Illumina MiSeq Sequencing

1. Place the MiSeq Reagent Kit at 4 °C the day before to thaw the reagents overnight.
2. On the day of the run, prepare 1 mL of NaOH 0.2 N in a microcentrifuge tube: 200 μL 1 N NaOH +800 μL H_2O.

   *The following steps (**steps 3–9**) should be performed on ice*:

3. Dilute library at 4 nM in a microcentrifuge tube: add 2 μL library at 10 nM and 3 μL EBT.
4. Denature and dilute library at 2 nM by adding 5 μL of 0.2 N NaOH.
5. Incubate for 5 min at room temperature.
6. Add 990 μL of HT1 buffer resulting in 1 mL of a 20 pM denatured library.
7. Proceed the same way (i.e., **steps 3–6**) to obtain 20 pM denatured PhiX.
8. In a new microcentrifuge tube, add 300 μL of 20 pM library to 300 μL HT1 resulting in 600 μL of 10 pM library; mix by pipetting.
9. In a new microcentrifuge tube, add 540 μL of the 10 pM library and 60 μL of 20 pM denatured PhiX; mix by pipetting.
10. Load 600 μL of the final combined library in the "load sample well" of the MiSeq cartridge.
11. Enter the following parameters in the Local Run Manager (LRM) of the Miseq as depicted in Table 5.
12. Fill the sample table with sample ID, and the associated index well (A01 corresponding to the unique indexes D701 and D501 combination).

Table 5
Parameters to enter into the MiSeq Local Run Manager

Module	GenerateFASTQ
Workflow	GenerateFASTQ
Library Prep Kit	TruSeq DNA-RNA CD Indexes 96 Indexes
Chemistry	Amplicon
Read Type	Paired end
Index Reads	2
Read 1 length	301
Read 2 length	301
Index read 1 length	8
Index read 2 length	8
Adapter Trimming	On
Adapter read 1	AGATCGGAAGAGCACACGTCTGAACTCCAGTCA
Adapter read 2	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

3.8 Bioinformatic Analysis on Vidjil Platform (See Note 9)

1. After completion of the run (56 h), copy the FASTQ files from the MiSeq directory on a hard drive.
2. Connect to your Vidjil server and enter login and password (*see* **Note 10**).
3. Create a run: click on run, and then click on [+ new runs]; enter run ID, run name, date, and other information, and click on [save].
4. Create new patients for each sample in the run: click on [+ new patients] and enter patient ID, first name, last name, and other information; click on [save].
5. Open the created run and click on [+ add samples].
6. Choose pre-process scenario (read merging with Flash2): select [4- M + R2: Merge paired-end reads].
7. Sample1: select R1 (first file) and R2 (second file) FASTQ files by clicking on [Browse...]. Date of sampling and other informations can be added. Importantly, the corresponding patient had to be associated with the sample by clicking its ID in the field [sample information].
8. Repeat for each sample [add other sample].
9. Click on [submit samples].
10. In the created run select process config [IGH] and click on the gear wheel to launch the analysis. The results are available when Completed appears in the status.

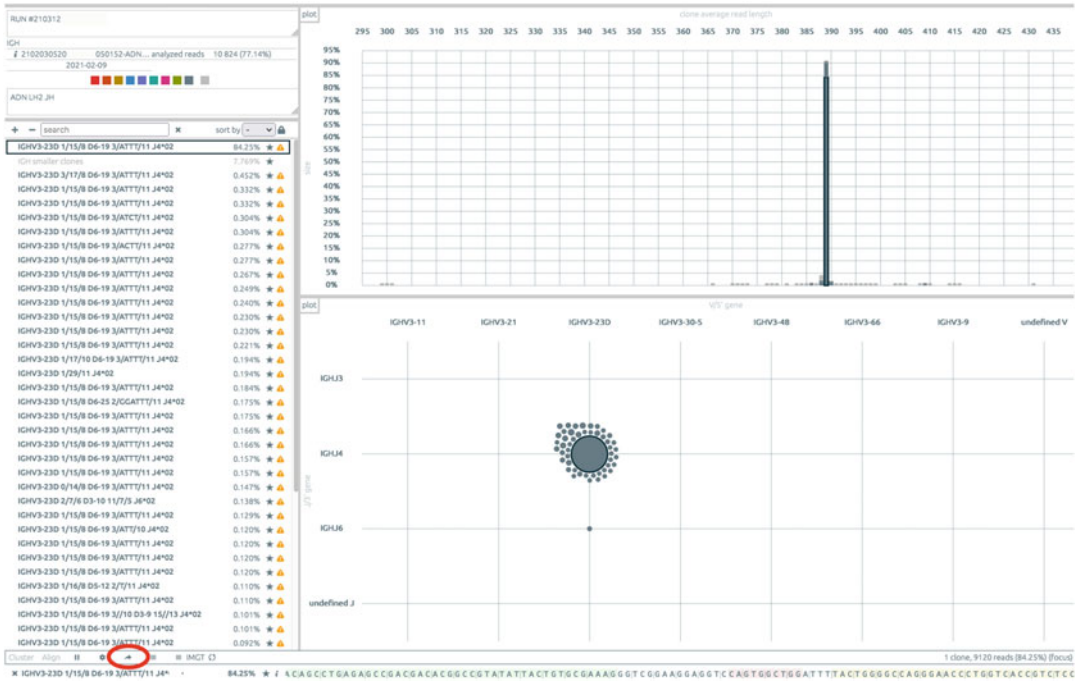


Fig. 1 Screenshot of results displayed Vidjil. The Vidjil platform provides an interactive visualization of antigen receptor repertoire from high-throughput data. The left panel lists the most frequent clonotypes, the most abundant being at the top (squared). By default, the 50 most frequent ones are displayed, the value being adjustable (from 5 to 100). The IGHV, IGHD, and IGHJ genes contributing to each clonotype are indicated as well as number of deleted/inserted nucleotides. Further information is available by clicking on the yellow triangles. At the top of the left panel, a summary of the sample sequencing quality data can be obtained by clicking on the "*i*" symbol. The top-right panel shows the size distribution of the clonotype average read length, simulating the traditional Genescan view of clonality analysis. The bottom-right panel offers a representation of the clonotypes according to their size and IGHV and IGHJ gene composition. Note that, in the vast majority of cases, the CLL dominant IGH clonotype appears surrounded by multiple small variant ones, differing by minor nucleotides changes. Sequence of the selected clonotype appears at the very bottom, the IGHV, IGHD, and IGHJ genes being highlighted. By clicking on the bent arrow above (circled), the sequence is sent automatically to IMGT/V-QUEST, IgBlast, and ARResT/AssignSubsets for further analysis

11. To access results, go to the patient page and click on [IGH].
12. The following information is displayed: (*see* also Fig. 1).
    (a) List of most abundant clonotypes on the left.
    (b) Graphic visualization by abundance and read length (on top).
    (c) Graphic visualization by abundance and V/J usage (on the bottom).
13. Select the most abundant clonotype(s).
14. Click on the curved arrow on the bottom of the page to send the clonotype sequence to IMGT/V-QUEST [18] and ARResT/AssignSubsets [19] (*see* **Note 11**).

4 Notes

1. Peripheral blood is the most common source of material and should be collected (10–20 mL) in EDTA (or citrate)-containing tubes. Tumor material can also be obtained from tissues infiltrated by leukemic cells such as bone marrow or lymph nodes. Frozen biopsies are much preferred over formalin-fixed paraffin-embedded tissue samples due to the need to amplify relatively large PCR products (median size around 400 bp).
2. If necessary, the quality of the cDNA synthesis can be assessed by amplification of a house-keeping gene, although this is not a mandatory step.
3. The sequencing protocol uses dual index PCR primers. Each primer contains, from 5′ to 3′, the following: (1) a set of nucleotides for flow cell binding (P5 or P7), (2) a patient barcode index (forward D501-D508, reverse D701-D712), (3) a set of nucleotides for sequencing initiation, and (4) IGHV-Leader or IGHJ (or IGHC) specific primer sequence. The double barcode indexing allows up to 96 unique combinations in a single run.
4. Primers are ordered according to a standard quality synthesis followed by polyacrylamide gel electrophoresis purification. They are resuspended in nuclease-free water at a 100 μM concentration.
5. When preparing primer mixes, only IGHV-Leader and IGHJ or IGHC primers containing the same barcode can be pooled. Take precautions to avoid cross-contamination between primers with different barcodes.
6. This 0.75:1 (beads/PCR products) ratio allows selective recovery of DNA fragments above 150 bp, thus eliminating primer dimers.
7. Quantification of purified PCR products should be done preferentially by fluorometry. Several types of fluorescence readers can be used, including Qubit® 4 Fluorometer (Thermo Fischer Scientific) or Clariostar (BMG Labtech). Only the latter is described here.
8. In case of concentration above the standards (upper limit 40 ng/μL), dilute the sample and repeat quantification.
9. There are numerous tools to analyze antigen receptor sequences produced by high-throughput sequencing [20]. Here we refer to Vidjil (https://app.vidjil.org/) [21, 22], an easy-to-use platform which does not require specific informatics skills. Note that the current online version is for research only, but an option compliant for clinical use can

be purchased. ArresT/Interrogate developed within the EuroClonality-NGS working group is another well-adapted alternative [23].

10. Several options exist for adding patient data on a Vidjil server, see http://www.vidjil.org/doc/healthcare/
11. More detailed information can be found in the user manual: http://www.vidjil.org/doc/user.

Acknowledgments

Anne Langlois de Septenville and Myriam Boudjoghra contributed equally to this work.

References

1. Hallek M, Shanafelt TD, Eichhorst B (2018) Chronic lymphocytic leukaemia. Lancet 391: 1524–1537
2. Chiorazzi N, Stevenson FK (2020) Celebrating 20 years of IGHV mutation analysis in CLL. Hemasphere 4:e334
3. Sutton LA, Hadzidimitriou A, Baliakas P et al (2017) Immunoglobulin genes in chronic lymphocytic leukemia: key to understanding the disease and improving risk stratification. Haematologica 102:968–971
4. Rossi D, Terzi-di-Bergamo L, De Paoli L et al (2015) Molecular prediction of durable remission after first-line fludarabine-cyclophosphamide-rituximab in chronic lymphocytic leukemia. Blood 126:1921–1924
5. Fischer K, Bahlo J, Fink AM et al (2016) Long-term remissions after FCR chemoimmunotherapy in previously untreated patients with CLL: updated results of the CLL8 trial. Blood 127: 208–215
6. Hallek M, Cheson BD, Catovsky D et al (2018) iwCLL guidelines for diagnosis, indications for treatment, response assessment, and supportive management of CLL. Blood 131: 2745–2760
7. Tonegawa S (1983) Somatic generation of antibody diversity. Nature 302:575–581
8. Rajewsky K (1996) Clonal selection and learning in the antibody system. Nature 381: 751–758
9. Lefranc MP, Giudicelli V, Duroux P et al (2015) IMGT®, the international imMunoGeneTics information system® 25 years on. Nucleic Acids Res 43:D413–D422
10. Damle RN, Wasil T, Fais F et al (1999) Ig V gene mutation status and CD38 expression as novel prognostic indicators in chronic lymphocytic leukemia. Blood 94:1840–1847
11. Hamblin TJ, Davis Z, Gardiner A et al (1999) Unmutated Ig V(H) genes are associated with a more aggressive form of chronic lymphocytic leukemia. Blood 94:1848–1854
12. Stamatopoulos K, Agathangelidis A, Rosenquist R et al (2017) Antigen receptor stereotypy in chronic lymphocytic leukemia. Leukemia 31:282–291
13. Baliakas P, Hadzidimitriou A, Sutton LA et al (2014) Clinical effect of stereotyped B-cell receptor immunoglobulins in chronic lymphocytic leukaemia: a retrospective multicentre study. Lancet Haematol 1:e74–e84
14. Sutton LA, Young E, Baliakas P et al (2016) Different spectra of recurrent gene mutations in subsets of chronic lymphocytic leukemia harboring stereotyped B-cell receptors. Haematologica 101:959–967
15. Rosenquist R, Ghia P, Hadzidimitriou A et al (2017) Immunoglobulin gene sequence analysis in chronic lymphocytic leukemia: updated ERIC recommendations. Leukemia 31: 1477–1481
16. Langerak AW, Brüggemann M, Davi F et al (2017) High-throughput immunogenetics for clinical and research applications in immunohematology: potential and challenges. J Immunol 198:3765–3774
17. Davi F, Langerak AW, de Septenville AL et al (2020) Immunoglobulin gene analysis in chronic lymphocytic leukemia in the era of next generation sequencing. Leukemia 34: 2545–2551
18. Brochet X, Lefranc MP, Giudicelli V (2008) IMGT/V-QUEST: the highly customized

and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. Nucleic Acids Res 36:W503–W508

19. Bystry V, Agathangelidis A, Bikos V et al (2015) ARResT/AssignSubsets: a novel application for robust subclassification of chronic lymphocytic leukemia based on B cell receptor IG stereotypy. Bioinformatics 31:3844–3846
20. Chaudhary N, Wesemann DR (2018) Analyzing immunoglobulin repertoires. Front Immunol 9:462
21. Giraud M, Salson M, Duez M et al (2014) Fast multiclonal clusterization of V(D)J recombinations from high-throughput sequencing. BMC Genomics 15:409
22. Duez M, Giraud M, Herbert R et al (2016) Vidjil: a web platform for analysis of high-throughput repertoire sequencing. PLoS One 11:e0166126. [Erratum in: PLoS One (2017) 12:e0172249]
23. Bystry V, Reigl T, Krejci A et al (2017) ARResT/Interrogate: an interactive immunoprofiler for IG/TR NGS data. Bioinformatics 33:435–437

Chapter 11

NGS-Based B-Cell Receptor Repertoire AnalysisRepertoire analyses in the Context of Inborn Errors of Immunity

Pauline A. van Schouwenburg, Mirjam van der Burg, and Hanna IJspeert

Abstract

Inborn errors of immunity (IEI) are genetic defects that can affect both the innate and the adaptive immune system. Patients with IEI usually present with recurrent infections, but many also suffer from immune dysregulation, autoimmunity, and malignancies.

Inborn errors of the immune system can cause defects in the development and selection of the B-cell receptor (BCR) repertoire. Patients with IEI can have a defect in one of the key processes of immune repertoire formation like V(D)J recombination, somatic hypermutation (SHM), class switch recombination (CSR), or (pre-)BCR signalling and proliferation. However, also other genetic defects can lead to quantitative and qualitative differences in the immune repertoire.

In this chapter, we will give an overview of protocols that can be used to study the immune repertoire in patients with IEI, provide considerations to take into account before setting up experiments, and discuss analysis of the immune repertoire data using Antigen Receptor Galaxy (ARGalaxy).

Key words Next generation sequencing, Primary immunodeficiency, B-cell receptor repertoire, Inborn errors of immunity

1 Introduction

At this moment, more than 450 monogenetic defects have been reported in patients with inborn errors of immunity (IEI) [1]. The most common forms of IEI are patients with a predominant B-cell disorder leading to primary antibody deficiencies. T-cell disorders also have an effect on the development and function of B cells, because they are required for further differentiation of B cells into memory B cells and plasma cells.

IEI can have a direct or indirect effect on the B-cell receptor (BCR) repertoire. Direct effects are found in patients with genetic defects in genes involved in one of the key processes in the formation or shaping of the B-cell repertoire: V(D)J recombination, somatic hypermutation (SHM), class switch recombination

Anton W. Langerak (ed.), *Immunogenetics: Methods and Protocols*, Methods in Molecular Biology, vol. 2453, https://doi.org/10.1007/978-1-0716-2115-8_11,

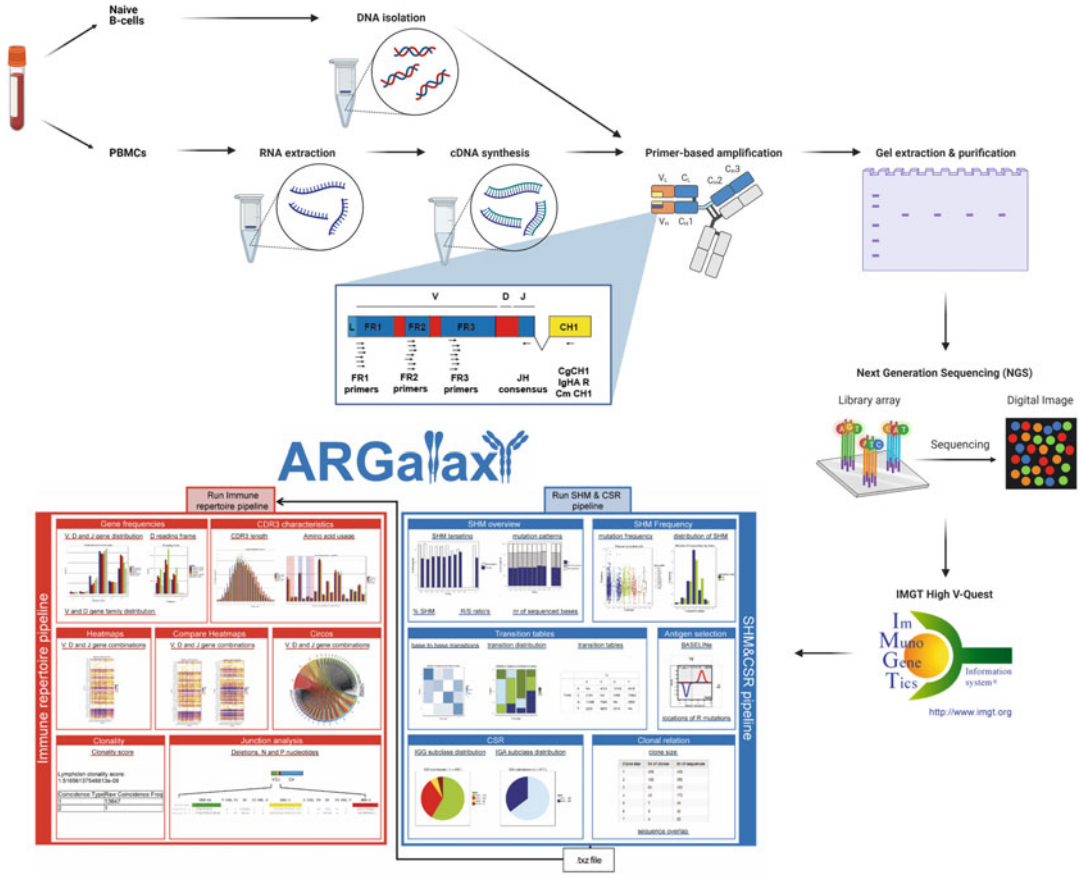


Fig. 1 Schematic overview of the workflow. Summary of the workflow for NGS-based B-cell receptor sequencing using primer-based amplification and analysis using the Antigen Receptor Galaxy (ARGalaxy) pipeline. Created with BioRender.com

(CSR), and (pre-)BCR signalling and proliferation [2–4]. Indirect effects can also be found because recurrent infections and/or autoimmunity can shape the BCR repertoire in IEI patients [5].

The BCR can be studied in several different ways, largely depending on the research question that needs to be answered and on the availability of the material. We will discuss how the BCR repertoire can be studied by amplifying BCR rearrangements from either DNA or cDNA and how to analyze the data using the Antigen Receptor Galaxy (ARGalaxy) analysis tool (Fig. 1). These methods can be applied to every sample, but we will focus on considerations that will affect the setup of the experiments and the data analysis for patients with IEI.

1.1 Selection of the Type of Cell or Tissue

The BCR repertoire can be divided into three classes: the immature BCR repertoire, the naïve BCR repertoire, and the antigen-selected BCR repertoire (Fig. 2). The immature BCR repertoire is derived

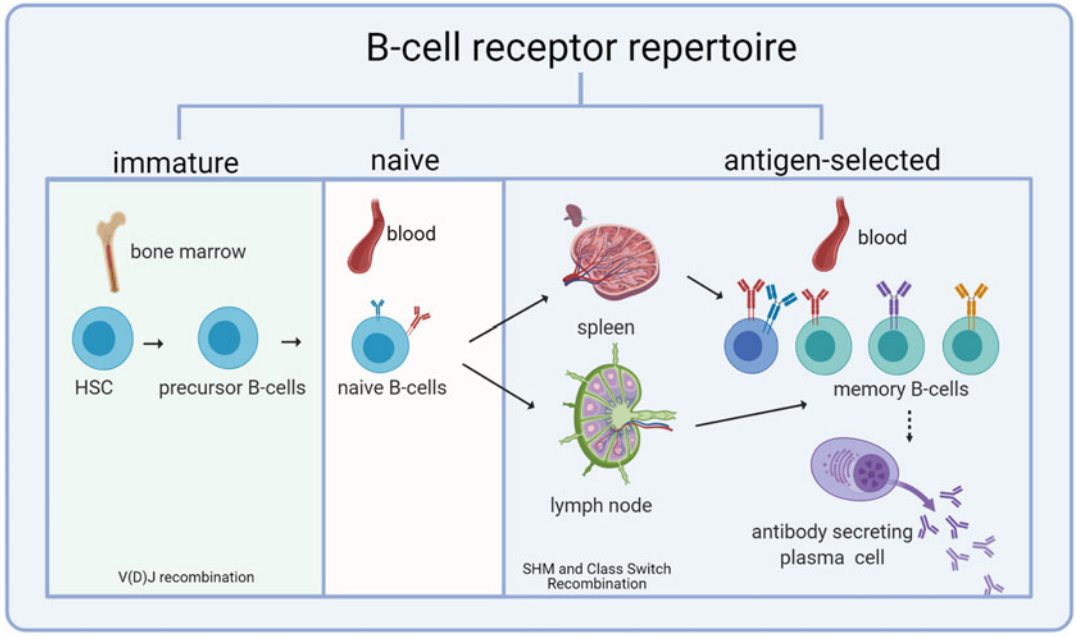


Fig. 2 Overview of the B-cell receptor repertoires. The B-cell receptor (BCR) repertoire can be divided into immature BCR repertoire, naïve BCR repertoire and antigen-selected BCR repertoire. Created with BioRender.com

from precursor B cells that did not undergo selection and/or have not completed their BCR rearrangements. This repertoire is particularly interesting for studying BCR repertoire formation in developing precursor B cells and processes like V(D)J recombination or pre-BCR signalling. Since precursor B-cell development takes place in the bone marrow, the only tissue that can be used to study the immature BCR is bone marrow. The naïve BCR repertoire is derived from naïve B cells that have not been activated. These naïve B cells can be found in peripheral blood. Peripheral blood is the least invasive material to obtain and for most labs easily accessible. However, peripheral blood contains a mixture of B-cell subsets, including naïve, memory, and plasma cells. The antigen-selected repertoire is derived from B cells that have been activated by their antigen. These B cells will differentiate into memory B cells or plasma cells. The antigen-selected B cells can be found in peripheral blood or secondary lymphoid organs, such as spleen or lymph nodes. Because tissues contain a mixture of B-cell subsets, it might be relevant to sort the population of interest before performing immune repertoire analysis of the naïve BCR repertoire or the antigen-selected BCR repertoire.

1.2 DNA Versus RNA

BCR rearrangements can be amplified from either DNA or RNA (cDNA). DNA is more stable than RNA and can be isolated from smaller cell numbers. The advantage of DNA is that is allows to

study unproductive and incomplete (DH-JH) rearrangements, which is not possible with RNA. Furthermore, there is only one DNA copy of a given functional rearrangement per cell, in contrast to RNA where there are many RNA copies per rearrangement per cell. The number of RNA copies is much higher in plasma cells compared to memory B cells. The advantage of RNA is that it allows to only analyze productive rearrangements and to study the constant gene. Furthermore, RNA is also preferred to use unique molecular identifiers (UMI) to identify the single RNA molecules.

1.3 The Number of B Cells that Can Be Studied

The BCR repertoire has been studied for decades by amplifying BCR rearrangements, cloning, and Sanger sequencing. However, since the introduction of next generation sequencing, it is possible to study thousands or even millions of BCR, in a way that is less labor intensive. The challenge of this high-throughput method is to obtain enough B cells to study thousands or millions of BCR rearrangements, especially in patients with a B-cell deficiency. Therefore, in patients with IEI, the starting material is often mononuclear cells obtained from blood or bone marrow. When using mononuclear cells, it is good to determine the frequency of B cells, e.g., using flow cytometry to be able to estimate the number of B-cell rearrangements that can be analyzed.

1.4 Location of the Primers

The IGH locus consist of >100 different variable (V), diversity (D), and joining (J) genes that are recombined to form a BCR. Fortunately, many of the genes have large sequence similarities and can therefore be subdivided in different families, such that primers specific for these gene families can be used in a multiplex PCR to amplify the repertoire. The forward primers can be located in the leader, or the frame work regions (FR) of the VH genes. Preferably, the forward primers should not be located in the complementary determining regions (CDR) regions, because these regions can have a high frequency of somatic hypermutations (SHM) that can decrease the binding efficiency of the primer. The location of the primers is also dependent on the information that is needed from the immune repertoire data. Primers in the leader sequence are least affected by SHM and provide the most accurate information about the hypomorphic alleles, but this results in a long amplicon that might not be suitable for all sequence platforms. In this protocol, we use the 6 IGHV FR1, 7 IGHV FR2, or 7 IGHV FR3 forward primers adapted with the Rd1 adaptor for Illumina sequencing (Fig. 3) (Table 1) [6]. As reverse primer, a single primer in the JH gene is enough to cover all six functional JH genes (Table 1). However, when there is an interest in information about the (sub)-class of the BCR, a primer in the constant (C) gene can be used. These rearrangements can only be amplified using cDNA as starting material. Since the amount of material is often limited in patients

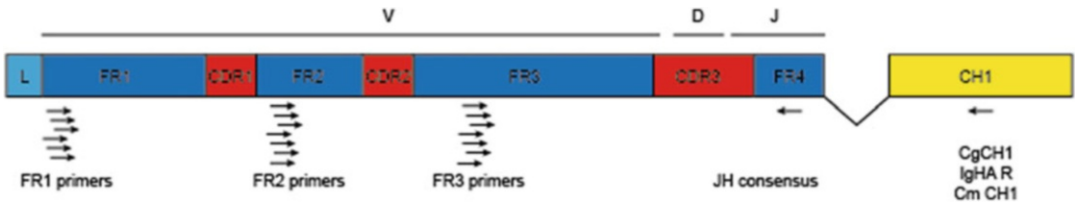


Fig. 3 Overview of IGH locus with primers. The forward primers located in FR1, FR2, or FR2 are indicated. For B-cell receptor rearrangements amplified from DNA the JH consensus can be used. For amplification of the B-cell receptor rearrangements from cDNA, either the JH consensus, the CgCH1, IgHA R, or the Cm CH1 primers can be used

with IEI, using primers in the Cγ or Cα region also allows to select for rearrangements derived from Ig-switched memory B cells without the need of pre-sorting of these cells. Optionally, a reverse primer in the Cμ region can be used. Subsequently, the data can be separated in rearrangements that contain <2% SHM and are likely derived from naïve B cells and rearrangements that have >2% SHM, which are likely derived from memory B cells. The reverse primers should also be adapted by addition of the Rd2 adaptor for Illumina sequencing (Table 1).

1.5 Choosing a Tool to Analyze the Immune Repertoire Data

Next generation sequencing of the BCR repertoire generates thousands of rearrangements and requires bioinformatics tools to analyze. In this last decade, many different analysis tools have been developed. Most tools help to annotate the rearrangements and will aid to visualize the data. The choice of the tool greatly depends on the research question, and it is likely that multiple tools are needed to answer all questions. In this chapter, we will discuss the Antigen Receptor Galaxy (ARGalaxy) tool [7]. This tool is a web-based tool and can be used to analyze many different qualitative measurements. It has two different pipelines, the immune repertoire pipeline which allows the analysis of V, D, and J gene usage, CDR3 characteristics and junction characteristics, and the SHM and CSR pipeline which allows the analysis of SHM, antigen selection, and CSR. Depending on the research question, data can be analyzed with either one or both pipelines.

2 Materials

2.1 Amplification (VH-Cg or VH-Ca from cDNA or VH-JH from DNA)

1. cDNA or 50 ng/μl DNA.
2. PCR cycler.
3. PCR tubes.
4. AmpliTaqGold (Thermo Fisher Scientific) with 10× Buffer Gold.
5. 25 mM $MgCl_2$.

Table 1
Overview of primers sequences

Name primer	Rd1 or Rd2 adaptor illumina	Template-specific sequence	Primer 5′ - 3′
VH1-FR1 Rd1	ACACTCTTTCCCTACACGACGC TCTTCCGATCT	GGCCTCAGTGAAGGTCTCC TGCAAG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGGCC TCAGTGAAGGTCTCCTGCAAG
VH2-FR1 Rd1	ACACTCTTTCCCTACACGACGC TCTTCCGATCT	GTCTGGTCCTACGCTGG TGAAACCC	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGTCTGG TCCTACGCTGGTGAAACCC
VH3-FR1 Rd1	ACACTCTTTCCCTACACGACGC TCTTCCGATCT	CTGGGGGGTCCCTGAGACTC TCCTG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTC TGGGGGGTCCCTGAGACTCTCCTG
VH4-FR1 Rd1	ACACTCTTTCCCTACACGACGC TCTTCCGATCT	CTTCGGAGACCCTGTCCC TCACCTG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTC TTCGGAGACCCTGTCCCTCACCTG
VH5-FR1 Rd1	ACACTCTTTCCCTACACGACGC TCTTCCGATCT	CGGGGAGTCTCTGAAGATC TCCTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATC TCGGGGAGTCTCTGAAGATCTCCTGT
VH6-FR1 Rd1	ACACTCTTTCCCTACACGACGC TCTTCCGATCT	TCGCAGACCCTCTCACTCACC TGTG	ACACTCTTTCCCTACACGACGCTCTTCCGATC TTCGCAGACCCTCTCACTCACCTGTG
VH1-FR2 Rd1	ACACTCTTTCCCTACACGACGC TCTTCCGATCT	CTGGGTGCGACAGGCCCC TGGACAA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTGGG TGCGACAGGCCCCTGGACAA
VH2-FR2 Rd1	ACACTCTTTCCCTACACGACGC TCTTCCGATCT	TGGATCCG TCAGCCCCCAGGGAAGG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGGA TCCGTCAGCCCCCAGGGAAGG
VH3-FR2 Rd1	ACACTCTTTCCCTACACGACGC TCTTCCGATCT	GGTCCGCCAGGC TCCAGGGAA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGG TCCGCCAGGCTCCAGGGAA

VH4-FR2 Rd1	ACACTCTTTCCCTACACGACGC TCTTCCGATCT	TGGATCCGCCAGCCC CCAGGGAAGG	ACACTCTTTCCCTACA CGACGCTCTTCCGATCTTGGA TCCGCCAGCCCCCAGGGAAGG
VH5-FR2 Rd1	ACACTCTTTCCCTACACGACGC TCTTCCGATCT	GGGTGCGCCAGA TGCCCGGGAAAGG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGGG TGCGCCAGATGCCCGGGAAAGG
VH6-FR2 Rd1	ACACTCTTTCCCTACACGACGC TCTTCCGATCT	TGGATCAGGCAGTCCCCA TCGAGAG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGGA TCAGGCAGTCCCCATCGAGAG
VH7-FR2 Rd1	ACACTCTTTCCCTACACGACGC TCTTCCGATCT	TTGGGTGCGACAGGCCCC TGGACAA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTTGGG TGCGACAGGCCCCTGGACAA
VH1-FR2 Rd1	ACACTCTTTCCCTACACGACGC TCTTCCGATCT	TGGAGCTGAGCAGCCTGAGA TCTGA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGGAGC TGAGCAGCCTGAGATCTGA
VH2-FR2 Rd1	ACACTCTTTCCCTACACGACGC TCTTCCGATCT	CAATGACCAACATGGACCCTG TGGA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCAA TGACCAACATGGACCCTGTGGA
VH3-FR2 Rd1	ACACTCTTTCCCTACACGACGC TCTTCCGATCT	TCTGCAAATGAACAGCC TGAGAGCC	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTC TGCAAATGAACAGCCTGAGAGCC
VH4-FR2 Rd1	ACACTCTTTCCCTACACGACGC TCTTCCGATCT	GAGCTCTG TGACCGCCGCCGCGGCACG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGAGCTC TGTGACCGCCGCCGCGGCACG
VH5-FR2 Rd1	ACACTCTTTCCCTACACGACGC TCTTCCGATCT	CAGCACCGCCTACCTGCAG TGGAGC	ACACTCTTTCCCTACACGACGCTCTTCCGATC TCAGCACCGCCTACCTGCAGTGGAGC
VH6-FR2 Rd1	ACACTCTTTCCCTACACGACGC TCTTCCGATCT	GTTCTCCCTGCAGCTGAACTC TGTG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGTTC TCCCTGCAGCTGAACTCTGTG

(continued)

Table 1
(continued)

Name primer	Rd1 or Rd2 adaptor illumina	Template-specific sequence	Primer 5′ - 3′
VH7-FR2 Rd1	ACACTCTTTCCCTACACGACGC TCTTCCGATCT	CAGCACGGCATATCTGCAGA TCAG	ACACTCTTTCCCTACACGACGCTCTTCCGATC TCAGCACGGCATATCTGCAGATCAG
JH cons Rd2	TCGCGAGTTAATGCAACGATCG TCGAAATTCGC	CTTACCTGAGGAGACGG TGACC	TCGCGAGTTAATGCAACGATCGTCGAAATTCGCCTTACC TGAGGAGACGGTGACC
CgCH1 Rd2	TCGCGAGTTAATGCAACGATCG TCGAAATTCGC	GGAAGGTGTGCACGCCGC TGGTC	TCGCGAGTTAATGCAACGATCGTCGAAA TTCGCGGAAGGTGTGCACGCCGCTGGTC
IgHA R Rd2	TCGCGAGTTAATGCAACGATCG TCGAAATTCGC	CTTTCGCTCCAGGTCACAC TGAG	TCGCGAGTTAATGCAACGATCGTCGAAATTCGCC TTTCGCTCCAGGTCACACTGAG
Cm CH1 Rd2	TCGCGAGTTAATGCAACGATCG TCGAAATTCGC	GGGAATTCTCACAGGAGACGA	TCGCGAGTTAATGCAACGATCGTCGAAATTCGCGGGAA TTCTCACAGGAGACGA

6. dNTP solution; prepare a mix with 20 mM of each nucleotide.
7. Bovine serum albumin (BSA) (20 mg/ml).
8. Nuclease-free PCR water.
9. 10 μM (10 pmol/μl) primer: pipet every primer separately.
10. Ethidium bromide (Sigma).
11. Agarose.
12. Tris-Borate-EDTA (TBE) buffer.
13. Loading dye for DNA gels.
14. 100 bp DNA ladder.
15. Gel extraction kit (Qiagen).
16. Scalpels: use 1 scalpel per PCR reaction.

2.2 Nested PCR

1. PCR cycler.
2. PCR tubes.
3. KAPA HiFi Hotstart Ready mix (Roche).
4. TruSeq Custom Amplicon Index Kit (Illumina).

2.3 Merging, Trimming, and Alignment of Reads and Data Analysis

1. https://argalaxy.researchlumc.nl/.
2. PEAR (https://cme.h-its.org/exelixis/web/software/pear/) [8].
3. Cutadapt (http://code.google.com/p/cutadapt) [9].
4. FASTQ to FASTA converter (http://usegalaxy.org/u/dan/p/fastq) [10].
5. IMGT High-V-Quest (http://www.imgt.org/HighV-QUEST/home.action) [11].
6. Immune repertoire tool of ARGalaxy (https://argalaxy.researchlumc.nl/).
7. SHM and CSR tool of ARGalaxy (https://argalaxy.researchlumc.nl/).

3 Methods

3.1 Amplification of VH-Cg, VH-Ca, or VH-Cμ from cDNA

1. Prepare PCR master mix consisting of 28.3 μl water, 5 μl 10× Buffer Gold, 3 μl $MgCl_2$, 0.5 μl dNTPs, 1 μl BSA, and 0.2 μl Taq Gold (*see* **Note 1**).
2. Transfer PCR master mix into PCR reaction tubes (38 μl into each well).
3. Deposit 1 μl of each primer into the corresponding well (*see* **Notes 2**, **3**, and **4**).

4. Add 5 µl cDNA to the corresponding well, and carefully add the lid of the PCR tubes (*see* **Notes 4** and **5**).
5. Run PCR at 95 °C for 7 min; 25–35 cycles at 94 °C for 30 s, 57 °C for 30 s, 72 °C 1 min; 72 °C for 10 min (*see* **Note 6**).
6. Load 50 µl PCR product with 10 µl loading dye onto a 1% agarose gel in TBE buffer containing ethidium bromide and run for 1 h at 180 V.
7. Visualize DNA band under ultraviolet (UV) light (*see* **Note 7**), and cut the PCR band of approximately 500 bp from gel using a scalpel (*see* **Note 8**).
8. Purify the PCR product from gel using the gel extraction kit. Follow the instructions in the manual and eluate with 20 µl elution buffer.
9. Continue with Subheading 3.3, Nested PCR.

3.2 Amplification of VH-JH from DNA

1. Prepare PCR master mix consisting of 31.3 µl water, 5 µl 10× Buffer Gold, 3 µl $MgCl_2$, 0.5 µl dNTPs, 1 µl BSA, and 0.2 µl Taq Gold (*see* **Note 1**).
2. Transfer PCR master mix into PCR reaction tubes (41 µl into each well).
3. Deposit 1 µl of each primer (6 5′ primers and 1 Cg or Ca primer per well) into the corresponding well (*see* **Notes 2** and **4**).
4. Add 2 µl 50 ng/µl DNA to the corresponding well, and carefully add the lid of the PCR tubes (*see* **Notes 4** and **9**).
5. Run PCR at 95 °C for 7 min; 25–35 cycles at 94 °C for 30 s, 57 °C for 30 s, 72 °C 1 min; 72 °C for 10 min (*see* **Note 6**).
6. Load 50 µl PCR product with 10 µl loading dye onto a 1% agarose gel in TBE buffer containing ethidium bromide and run for 1 h at 180 V.
7. Visualize DNA band under ultraviolet (UV) light (*see* **Note 7**), and cut the PCR band of approximately 500 bp from gel using a scalpel (*see* **Note 8**).
8. Purify the PCR product from gel using the gel extraction kit. Follow the instructions in the manual and eluate with 20 µl elution buffer.
9. Continue with Subheading 3.3, Nested PCR.

3.3 Nested PCR and Pooling

1. Add 12.5 µl KAPA HiFi Hotstart Ready mix, 2 µl TruSeq Custom Amplicon Index forward primer, 2 µl TruSeq Custom Amplicon reverse primer, and 8.5 µl purified PCR product from Subheading 3.1 or Subheading 3.2 to a PCR reaction tube.
2. Run PCR at 95 °C for 5 min; 10 cycli at 98 °C for 20 s, 66 °C for 30 s, 72 °C 30 s; 72 °C for 1 min.

3. Measure the concentration of the PCR products (*see* **Note 10**).
4. Mix the PCR product at an equimolar concentration of 50 mM.
5. Purify the pool of PCR products (*see* **Note 11**).
6. The PCR pool can be sequenced using the Illumina platform.

3.4 Merging, Trimming, and Alignment of Reads Using Galaxy

1. Sequencing with the Illumina platform results R1 and R2 reads that need to be merged before they can be aligned to a reference database. This merging can be done with PEAR, which is a pair-end read merger [8], which can be found on "pre-processing" at https://argalaxy.researchlumc.nl/.
2. After the reads are merged, the Illumina Rd1 and Rd2 primer adapters have to be removed from the reads as well as the forward primers. This can be done with the Cutadapt tool [9] (*see* **Note 12**). which can be found on "pre-processing" at https://argalaxy.researchlumc.nl/.
3. Before the reads can be aligned using IMGT/HighV-Quest, the FASTQ files have to be adapted to the FASTA file format. This can be done with the FASTQ to FASTA converter [10], which can be found on "pre-processing" at https://argalaxy.researchlumc.nl/.
4. For alignment and annotation of the BCR rearrangements, the international ImMunoGeneTics system IMGT/HighV-Quest can be used (http://www.imgt.org/HighV-QUEST/analysis.action) [11]. This tool will produce a compressed .txz file that contains 12 text files with alignment information.

3.5 Data Analysis Using the Immune Repertoire Pipeline in Antigen Receptor Galaxy (ARGalaxy) (See Note 13)

1. Open ARGalaxy from https://argalaxy.researchlumc.nl/ [7].
2. Upload the compressed .txz files using: get data → upload file (*see* **Note 12**). The file will appear on the right site of your screen under "History."
3. Select under "Tools" on the left site of the screen "ARGalaxy," and click on the "Immune repertoire pipeline."
4. Select the .txz file you would like to analyze (*see* **Notes 14** and **15**).
5. Enter a name in the "ID" field (*see* **Note 16**).
6. Select the definition of the clonotype (*see* **Note 17**).
7. Select the order in which the V, D, and J genes have to appear in the graphs. The default setting is on alphabetical order and not in the order they appear on the IGH locus.
8. Select "IGH" at the "Locus" field.
9. Choose if you want to visualize the unproductive rearrangements in the graphs (*see* **Note 18**).

Table 2
Example of the overview table of the Immune repertoire pipeline in ARGalaxy showing the number and percentage of (unique) productive and unproductive sequences per donor and per replicate. The definition of unique sequences is based on the clonal type definition filter setting chosen

Donor/replicate	All	Productive	Unique productive	Unproductive	Unique unproductive
TEST	42,488	32,207 (76%)	14,905 (35%)	10,010 (24%)	6067 (14%)
TEST_1	18,911	15,036 (80%)	6452 (34%)	3772 (20%)	2164 (11%)
TEST_2	11,880	9390 (79%)	4852 (41%)	2394 (20%)	1583 (13%)
TEST_3	11,697	7781 (67%)	3601 (31%)	3844 (33%)	2320 (20%)

10. Select if you want to identify overlapping sequences between different replicates within one donor (*see* **Note 19**).
11. Press execute. A new item will be displayed in your history and turn green when the tool is ready with processing the data.
12. Click on the "eye" symbol to open the table that shows an overview of the rearrangements, including the percentage of productive, productive unique, unproductive, and unproductive unique (*see* Table 2 for an example).
13. Press on "Click here for the results" to open the page with the different analysis tabs.
14. The tab "Gene frequencies" shows the percentage of V, D, an J gene usage (*see* **Note 20**). The frequency of the different V, D, and J genes vary slightly between individuals and also between different primer sets that are used to amplify the BCR rearrangements. However, there are some important parameters that can give an indication of changes in the BCR repertoire (*see* Table 3). These changes can be specific for patients with IEI, but are also present between the naïve and antigen-selected BCR repertoire in healthy individuals. For example, the frequency of BCR with the IGHV4–34 and IGHJ6 genes are relatively high in the naïve BCR repertoire, but are significantly lower in antigen-selected B cells. In contrast to IGHJ4 which is less frequently used than IGHJ6 in the naïve BCR repertoire, it is the most frequently used IGHJ gene in the antigen-selected repertoire in healthy individuals (Fig. 4a) [15].
15. The tab "CDR3 characteristics" contains plots that show the distribution of the CDR3 length and the frequency of the different amino acids used in the CDR3. The median CDR3 length is longer in naïve B cells compared to memory B cells, which is likely caused by selection against long CDR3 lengths, because they are more likely to be autoreactive (Fig. 4b) [15].

Table 3
Overview of V, D, and J genes that can be affected in the B-cell receptor repertoire

Gene	Remark
IGHV4-34	B cells expressing this IGHV gene are almost all intrinsically autoreactive [12, 13]. The frequency of VH4-34 is high in naive B cells, but very low in memory B cells. An increased frequency has been observed in several IEI patients
IGHJ4 and IGHJ6	JH4 and JH6 are the most frequently used JH genes in healthy donors; however the distribution might differ between naive and memory B cells. The length of the JH6 gene is significantly longer than the other JH genes. Since memory B cells have shorter CDR3 length, the frequency of JH6 is lower in memory B cells compared to naive B cells in the same healthy donor [4]
IGHD7–27	IGHD7-27 is the smallest D gene and is located immediately adjacent to the IGHJ locus. High frequencies of IGHD7-27 have been observed in fetal B-cell receptor rearrangements [14]

16. In the tab "Heatmaps," the frequency of the different combinations of V-J, V-D, and D-J genes are visualized in heatmaps.
17. In the tab "Compare heatmaps," the heatmaps between different donors can be compared.
18. In the tab "Circos," the frequency of the different combination of V-J, V-D, and D-J genes are visualized using circus plots [16].
19. When the option is chosen to determine the number of sequences that share the same clonal type between replicates or to determine the clonality of the donor, the tab "Shared Clonal Types" or "Clonality" is shown. These tabs include a table with information about the number of BCR rearrangement that is present in multiple replicates of the same donor. When three replicates are present and the option "determine the clonality of the donor" is chosen, the clonality score based on the publication by Boyd et al. is given [17]. In patients with IEI, the diversity of the repertoire is often reduced (Fig. 4c) [2, 7].
20. The tab "Junction analysis" contains a table with the median or mean number of deletions, palindromic (P) nucleotides, or non-templated (N) nucleotides in the productive and unproductive rearrangements. Genetic defects in the non-homologous end joining (NHEJ) pathway have been shown to affect the number of deletions, N-nucleotides and P-nucleotides (*see* Table 4).
21. In the "Download" tab, all data used to create the tables and graphs can be downloaded.

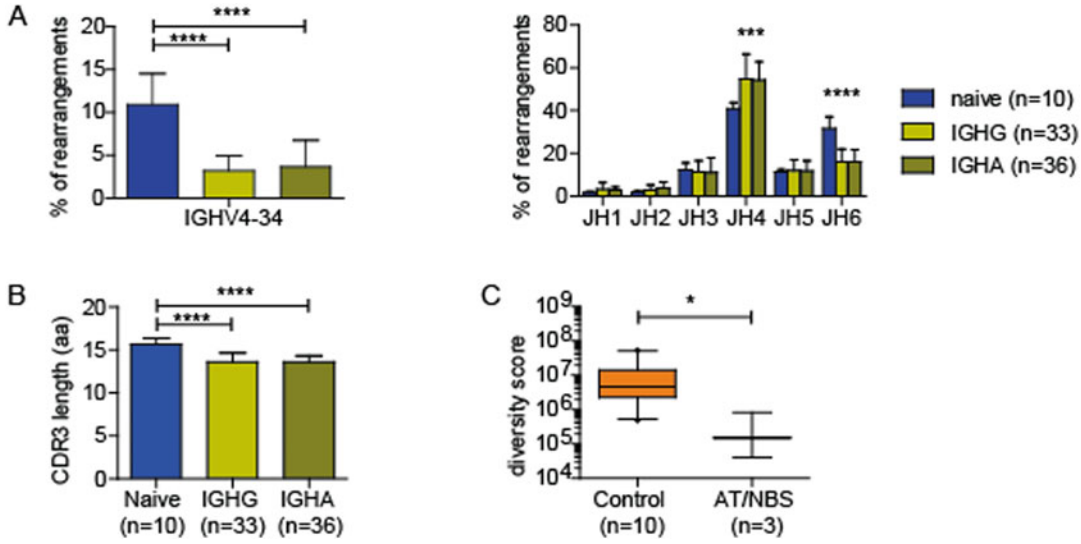


Fig. 4 Examples of analyses with the immune repertoire pipeline. Naïve B cells have a higher frequency of IGHV4–34 and IGHJ6 compared to antigen-selected switched B cells (IGHG, and IGHA) (**a**). The CDR3 length is shorter in antigen-selected switched B cells (IGHG and IGHA) compared to naïve B cells (**b**). Patients with ataxia telangiectasia (AT) or Nijmegen breakage syndrome have a reduced diversity of the naïve BCR repertoire (**c**). The number of samples analyzed is indicated per group. *P*-values <0.001 are indicated by *** and *P*-values <0.0001 are indicated by ****

Table 4
Overview junction characteristics of IEI patients with defects in the non-homologous end joining pathway

Gene	Protein name	Deletions	N-nucleotides	P-nucleotides	References
DCLRE1C	ARTEMIS	↓	↓	↑↑	[18]
PRKDC	DNA-PKcs	↓	↓	↑	[19]
NHEJ1	XLF/Cernunnos	Normal	↓↓↓	Normal	[4]
LIG4	DNA ligase 4	↑↑↑	↓	Normal	[20]
XRCC4	XRCC4	Normal	↓	Normal	[21]

3.6 Data Analysis Using the SHM and CSR Tool in ARGalaxy (See Note 13)

1. Open ARGalaxy from https://argalaxy.researchlumc.nl/ [7].
2. Upload the compressed .txz files using: get data → upload file (*see* **Note 14**). The file will appear on the right site of the screen under "History."
3. Select under "Tools" on the left site of the screen "ARGalaxy" and click on the "SHM and CSR pipeline."
4. Select the .txz file to be analyzed.
5. Select which regions of the BCR rearrangements should be included in the analysis (*see* **Note 21**).

6. Select if only the productive, only the unproductive, or both productive and unproductive sequences should be analyzed.
7. Select if the sequences should be filtered by "remove unique" or "keep unique." The "remove unique filter" removes all sequences that occur only once and the duplicates (based on the nucleotides sequence of the "analyzed region" and the C gene or the sequences that have the same V, J, and amino acid sequence of the CDR3 region). When choosing "remove unique," an additional filter appears that allows to choose the minimal number of duplicates that have to be in a group in order to keep one of the sequences (based on the nucleotides sequence of the "analyzed region" and the C gene or the sequences that have the same V, J, and amino acid sequence of the CDR3 region) . The "keep unique" filter removes all duplicate sequences based on the nucleotides sequence of the "analyzed" region and the C gene.
8. Select if duplicates should be removed based on V, CDR3, and C region (different options possible).
9. The class/subclass filter should only be applied when part of the C region is present. The SHM and CSR pipeline identifies human Cμ, Cα, Cγ, and Cε constant genes by dividing the reference sequences for the subclasses (NG_001019) in eight nucleotide chunks, which overlap by four nucleotides. These overlapping chunks are then individually aligned in the right order to each input sequence. This alignment is used to calculate the chunk hit percentage and the nt hit percentage. The chunk hit percentage is the percentage of the chunks that is aligned. The Nt hit percentage is the percentage of chunks covering the subclass-specific nucleotide match with the different subclasses. The most stringent filter for the subclass is 70% "nt hit percentage" which means that five out of seven subclass-specific nucleotides for Cα or six out of eight subclass specific nucleotides of Cγ should match with the specific subclass. The option ">19% class" can be chosen when only the class (Cα/Cγ/Cμ/Cε) of the sequences is of interest and the length of the sequence is not long enough to assign the subclasses. With the location of the primers used in this protocol, assignment of subclass is not possible and the class can be assigned with the >19% filter.
10. Select if a new IGMT archive output is needed in the history that contains only the sequences based on the filtering options used before (*see* **Note 13**).
11. Select if the generation of new IMGT archives and the analysis of Change-O/Baseline need to be skipped to decrease the time ARGalaxy needs to run the pipeline.

12. Press execute. A new item will be displayed in the history and turns green when the tool is ready with processing the data.
13. Click on the "eye" symbol to open the table that shows the number of rearrangement after each filtering step.
14. Press on "Click here for the results" to open the page with the different analysis tabs.
15. The "SHM overview" tab gives a table with detailed information on the SHM including frequency of SHM, the transversion and transition mutations, replacement and silent mutations, etc. Furthermore, it also contains graphs visualizing the percentage of mutations in AID and pol eta motives, the relative mutation patterns, and the absolute mutation patterns. The frequency of SHM increased during childhood (Fig. 5a) [15], but can be affected in patients with IEI (Fig. 5b). This can be caused by genetic defects in one of the genes involved in the SHM process [22], but can also be the consequence of recurrent infections or immune dysregulation.
16. The "SHM frequency" tab contains graphs that visualize the frequency of SHM per (sub)class.
17. The "transition table" tab contains tables, heatmaps, and bar graph that visualize the SHM per base. This information provides a lot of information about the SHM process and can be used to study the SHM pathway. In patients with genetic defects in genes involved in the DNA repair pathways (*UNG*, *MSH2*, *MSH6*, *PMS2*) crucial for the induction of SHM, the frequency as well as the pattern of SHM is affected (Fig. 5c) [22].
18. The "antigen selection" tab contains bar plots showing the frequency of replacement mutations per amino acid. These graphs can be used to study in which region or amino acids positions replacement mutations are most/least frequent. Furthermore, in this tab, also the plots showing the score for antigen selection based on the BASELINe method are given [23].
19. The "CSR" tab contains circle plots that indicate the subclass distribution of the IGHA or IGHG rearrangements. In patients with IEI, the subclass distribution is often affected (Fig. 5d). This can be caused by defects in CSR, e.g., in patients with ataxia telangiectasia (AT) [2], but is also observed in patients with common variable immunodeficiency [24].
20. The "clonal relation" tab gives a table which indicates the number of clones and the number of sequences within a clone (the definition of the clone is based on the filter settings used) based on the Change-O method [25] (*see* **Note 22**).
21. In the "Download" tab, all data used to create the tables and graphs can be downloaded.

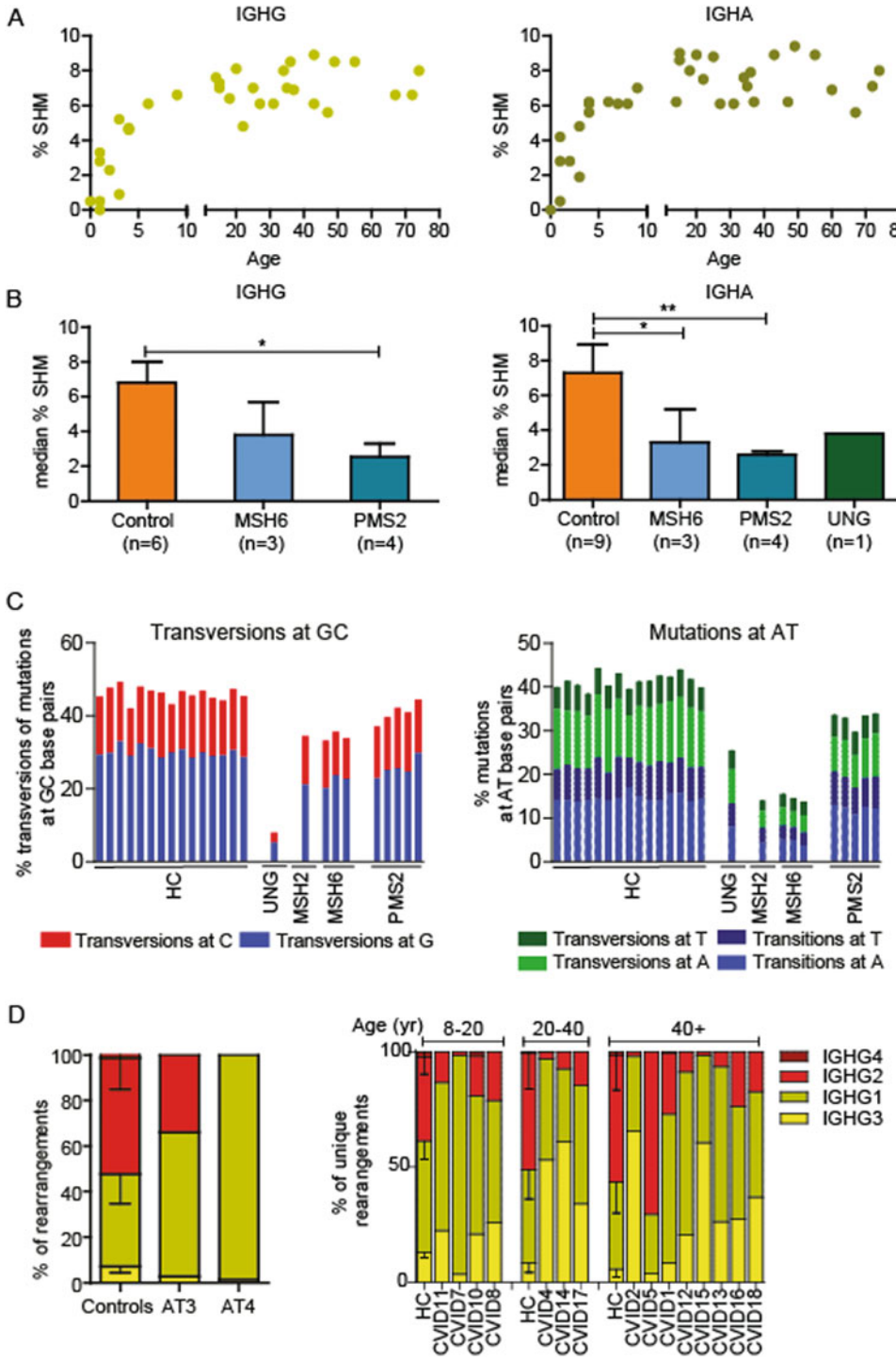


Fig. 5 Examples of analyses with the SHM and CSR pipeline. The median frequency of somatic hypermutations (SHM) increases during childhood in both IGHG and IGHA antigen-selected switched B cells (**a**). The median frequency of SHM is reduced in patients with MSH6, PMS2, or UNG deficiency (**b**). Patients with defects MSH2 and MH6 have a strong reduction in mutation at A and T base pairs compared to controls. Patients with UNG deficiency have a strong reduction in transversion mutations at G and C base pairs (**c**). Patients with ataxia telangiectasia (AT) and common variable immunodeficiency (CVID) have reduced frequency of IGHG2 and IGHG4 switched B cells (**d**). $*P < 0.05$ and $**P < 0$

4 Notes

1. Prepare a master mix for all reactions. Allow a surplus of 10%.
2. When using IGH VH FR1 primers, add six forward primers or seven primers when using IGH VH FR2 or IGH VH FR3.
3. When using reverse primers in the constant region, prepare a separate reaction to amplify the VH-Cα, VH-Cγ, or VH-Cμ rearrangements with the IgHA R Rd2, CgCH1 Rd2, or Cm CH1 Rd2 primers, respectively.
4. In these steps, prevention of cross-contaminations between samples is essential.
5. The amount of cDNA that needs to be added is dependent on the amount of B cells in the samples, and the number of B-cell rearrangement to be analyzed. Furthermore, it has to be taken into account that plasma cells typically have a 1000 times higher copy number of the RNA copies of the B-cell rearrangement compared to other B cells. If less cDNA is needed, nuclease-free PCR water can be added to reach a total volume of 5 μl.
6. The number of PCR cycli should preferably be low enough to remain in the linear amplification stage of the PCR which will reduce amplification bias. However, the number of cycli should be high enough to be able to visualize the PCR product on the agarose gel. The lower the number of B cells in the sample, the higher the number of PCR cycli that should be used.
7. Keep the exposure to UV as short as possible since UV can damage the PCR products.
8. Use a new scalpel for every PCR product to avoid contamination.
9. The amount of DNA that needs to be added is dependent on the amount of B cells in the sample, and the number of B-cell rearrangements to be analyzed. If less volume is needed to add the accurate amount of DNA, nuclease-free PCR water can be added to reach the total volume of 2 μl. The amount of DNA per cells is estimated to be 6 pg. So for DNA isolated from only B cells can be divided by 6 pg. 100 ng DNA corresponds to approximately 16,667 B cells. Since every B cells can have one unproductive and 1 productive rearrangement, 100 ng B-cell DNA can result in maximally 33,334 unique B-cell rearrangements.
10. The amount of PCR product should be measures with a sensitive method for low quantities of double-stranded DNA, e.g., Qubit™ dsDNA BCR Assay Kit.

11. Purification of the PCR library pool can be done with AMPure XP beads (Beckman Coulter) according to the manufacturer's instructions.
12. Removing the adaptor sequences from the reads improves the alignment of the BCR rearrangements. Removal of primer sequences located in the V(D)J region is essential to prevent mismatching of degenerate primers to be classified as SHM.
13. Dependent on your research question, analyze the data using the immune repertoire pipeline, the SHM and CSR pipeline, or both. When interested in using both using the same filtering, one can start with the analysis of the SHM and CSR pipeline, and then select "yes" in the filter "Output new IMGT archives per class into the history." This will provide a new data set in the history with the filtered data (split per class if class is assigned) which then can be analyzed using the immune repertoire pipeline.
14. Select "imgt_archive" by "Type (set all)."
15. In the "Immune repertoire pipeline," multiple .txz files can be analyzed simultaneously. These can be replicates from the same donor or can be derived from multiple donors.
16. Spaces and special characters (except "_") cannot be used in the ID field. Leaving this field empty will result in an error.
17. The data likely contains multiple reads which are identical or nearly identical. These reads can be derived from unique B cells with the same IGH rearrangements, but these can also be technical duplicates. When the BCR rearrangements are amplified from a low number of B-cells and/or many PCR cycles had to be used to obtain a PCR product, the presence of reads with the same clonotype is more likely caused by technical duplicates. Importantly, IGH rearrangements with the same CDR3 sequence at the amino acid level can be derived from unique B cells with a different IGH rearrangement at the nucleotide level. This filter will only include one sequence with the same clonotype definition in the analysis.
18. Unproductive rearrangements are rearrangements that are out-of-frame or contain a stop codon. When IGH rearrangements were amplified from DNA, a large fraction of the rearrangements are non-productive, while in case of amplification of IGH rearrangements from RNA, only a very small fraction of the IGH rearrangements will be unproductive, since unproductive rearrangements are mostly not transcribed.
19. Option 1 "Do not determine overlap (only 1 replicate present)" should be used if only one replicate is analyzed per donor or if there is no interest to determine the presence of overlapping sequences. Option 2 "Determine the number of

sequences that share the same clonal type between the replicate" should be used if the overlap between at least two replicates within the same donor should be determined. Option 3 "Determine the clonality of the donor (minimal 3 replicates) can be used to determine the number of overlapping sequences between at least three replicates within one donor and provides the clonality score described by Boyd et al. [17].

20. The rearrangements used for making the graphs are filtered based on the settings "Clonal type definition" and "Remove the unproductive sequences from graphs." When choosing to filter the data based on clonal type and remove the unproductive sequences from the graph, only the total number of unique productive sequences are included in the graphs.
21. The regions that are/can be included in the analysis depend on the forward primer being used. When using primers in the leader sequence, the complete BCR rearrangement can be used. However, when using primers in the FR regions, these regions have to be excluded because the primers sequences can cause false-positive SHM.
22. To calculate clonal relation, Change-O is used [26]. Transcripts are considered clonally related, if they have maximally three nucleotides difference in their CDR3 sequence and the same first V gene (as assigned by IMGT). Change-O settings used are the nucleotide hamming distance substitution model with a complete distance of maximally three. For clonal assignment, the first genes were used, and the distances were not normalized. In case of asymmetric distances, the minimal distance was used.

References

1. Tangye SG, Al-Herz W, Bousfiha A, Cunningham-Rundles C, Franco JL, Holland SM et al (2021) The ever-increasing array of novel inborn errors of immunity: an interim update by the IUIS committee. J Clin Immunol 41:666–679
2. Driessen GJ, IJspeert H, Weemaes CM, Haraldsson A, Trip M, Warris A et al (2013) Antibody deficiency in patients with ataxia telangiectasia is caused by disturbed B- and T-cell homeostasis and reduced immune repertoire diversity. J Allergy Clin Immunol 131: 1367–1375
3. Kolhatkar NS, Brahmandam A, Thouvenel CD, Becker-Herman S, Jacobs HM, Schwartz MA et al (2015) Altered BCR and TLR signals promote enhanced positive selection of autoreactive transitional B cells in Wiskott-Aldrich syndrome. J Exp Med 212:1663–1677
4. IJspeert H, Rozmus J, Schwarz K, Warren RL, van Zessen D, Holt RA et al (2016) XLF deficiency results in reduced N-nucleotide addition during V(D)J recombination. Blood 128: 650–659
5. Galson JD, Clutterbuck EA, Truck J, Ramasamy MN, Munz M, Fowler A et al (2015) BCR repertoire sequencing: different patterns of B-cell activation after two Meningococcal vaccines. Immunol Cell Biol 93:885–895
6. van Dongen JJ, Langerak AW, Bruggemann M, Evans PA, Hummel M, Lavender FL et al (2003) Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98-3936. Leukemia 17: 2257–2317

7. IJspeert H, Van Schouwenburg P, Van Zessen D, Pico-Knijnenburg I, Stubbs AP, Van der Burg M (2017) Antigen receptor galaxy: a user-friendly web-based tool for analysis and visualization of T and B cell receptor repertoire data. J Immunol 198:4156–4165
8. Zhang J, Kobert K, Flouri T, Stamatakis A (2013) PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics 30: 614–620
9. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal [S.l.], 17(1):10–12. ISSN 2226-6089. https://doi.org/10.14806/ej.17.1.200
10. Blankenberg D, Gordon A, Von Kuster G, Coraor N, Taylor J, Nekrutenko A, Galaxy T (2010) Manipulation of FASTQ data with Galaxy. Bioinformatics 26:1783–1785
11. Alamyar E, Duroux P, Lefranc MP, Giudicelli V (2012) IMGT((R)) tools for the nucleotide analysis of immunoglobulin (IG) and T cell receptor (TR) V-(D)-J repertoires, polymorphisms, and IG mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. Methods Mol Biol 882:569–604
12. Klonowski KD, Primiano LL, Monestier M (1999) Atypical VH-D-JH rearrangements in newborn autoimmune MRL mice. J Immunol 162:1566–1572
13. Wardemann H, Yurasov S, Schaefer A, Young JW, Meffre E, Nussenzweig MC (2003) Predominant autoantibody production by early human B cell precursors. Science 301: 1374–1377
14. Shiokawa S, Mortari F, Lima JO, Nunez C, Bertrand FE 3rd, Kirkham PM et al (1999) IgM heavy chain complementarity-determining region 3 diversity is constrained by genetic and somatic mechanisms until two months after birth. J Immunol 162: 6060–6070
15. IJspeert H, van Schouwenburg PA, van Zessen D, Pico-Knijnenburg I, Driessen GJ, Stubbs AP et al (2016) Evaluation of the antigen-experienced B-cell receptor repertoire in healthy children and adults. Front Immunol 7:410
16. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D et al (2009) Circos: an information aesthetic for comparative genomics. Genome Res 19:1639–1645
17. Boyd SD, Marshall EL, Merker JD, Maniar JM, Zhang LN, Sahaf B et al (2009) Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. Sci Transl Med 1:12–23
18. van der Burg M, Verkaik NS, den Dekker AT, Barendregt BH, Pico-Knijnenburg I, Tezcan I et al (2007) Defective Artemis nuclease is characterized by coding joints with microhomology in long palindromic-nucleotide stretches. Eur J Immunol 37:3522–3528
19. van der Burg M, IJspeert H, Verkaik NS, Turul T, Wiegant WW, Morotomi-Yano K et al (2009) A DNA-PKcs mutation in a radiosensitive T-B- SCID patient inhibits Artemis activation and nonhomologous end-joining. J Clin Invest 119:91–98
20. van der Burg M, van Veelen LR, Verkaik NS, Wiegant WW, Hartwig NG, Barendregt BH et al (2006) A new type of radiosensitive T-B-NK+ severe combined immunodeficiency caused by a LIG4 mutation. J Clin Invest 116:137–145
21. Murray JE, van der Burg M, IJspeert H, Carroll P, Wu Q, Ochi T et al (2015) Mutations in the NHEJ component XRCC4 cause primordial Dwarfism. Am J Hum Genet 96: 412–424
22. IJspeert H, van Schouwenburg PA, Pico-Knijnenburg I, Loeffen J, Brugieres L, Driessen GJ et al (2019) Repertoire sequencing of B cells elucidates the role of UNG and mismatch repair proteins in somatic hypermutation in humans. Front Immunol 10:1913
23. Yaari G, Vander Heiden JA, Uduman M, Gadala-Maria D, Gupta N, Stern JN et al (2013) Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. Front Immunol 4: 358
24. van Schouwenburg PA, IJspeert H, Pico-Knijnenburg I, Dalm V, van Hagen PM, van Zessen D et al (2018) Identification of CVID patients with defects in immune repertoire formation or specification. Front Immunol 9: 2545
25. Yaari G, Uduman M, Kleinstein SH (2012) Quantifying selection in high-throughput immunoglobulin sequencing data sets. Nucleic Acids Res 40:e134

26. Gupta NT, Vander Heiden JA, Uduman M, Gadala-Maria D, Yaari G, Kleinstein SH (2015) Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. Bioinformatics 31 (20):3356–3358

Chapter 12

Generic Multiplex Digital PCR for Accurate Quantification of T Cells in Copy Number Stable and Unstable DNA Samples

Rogier J. Nell, Willem H. Zoutman, Mieke Versluis, and Pieter A. van der Velden

Abstract

An accurate T cell quantification is prognostically and therapeutically relevant in various clinical applications, including oncology care and research. In this chapter, we describe how T cell quantifications can be obtained from bulk DNA samples with a multiplex digital PCR experiment. The experimental setup includes the concurrent quantification of three different DNA targets within one reaction: a unique T cell DNA marker, a regional corrector, and a reference DNA marker. The T cell marker is biallelically absent in T cells due to VDJ rearrangements, while the reference is diploid in all cells. The so-called regional corrector allows to correct for possible copy number alterations at the T cell marker locus in cancer cells. By mathematically integrating the measurements of all three markers, T cells can be accurately quantified in both copy number stable and unstable DNA samples.

Key words T cell quantification, Multiplex digital PCR, DNA markers, Copy number instability, Cancer

1 Introduction

T cells form an essential part of the human adaptive immune system. These cells are able to recognize and bind antigens via unique, antigen-specific cell-surface receptors, referred to as the T cell receptors (TCR). The enormous diversity of TCR molecules is generated by unique genetic mechanisms occurring during early maturation of these cells in the thymus [1, 2]. One of these mechanisms involves the rearrangement of the germline T cell receptor (TR) genes (i.e., TRD, TRG, TRB, and TRA) into a unique TR blueprint. The absolute presence of T cells varies between tissues and body fluids and is influenced by physiological and pathological conditions. For that reason, an accurate quantification of (infiltrated) T cells is relevant in various clinical applications, ranging from autoimmune disorders to infectious disease and cancer [3, 4].

Anton W. Langerak (ed.), *Immunogenetics: Methods and Protocols*, Methods in Molecular Biology, vol. 2453, https://doi.org/10.1007/978-1-0716-2115-8_12,

Traditionally, the presence of immune cells has been assessed by histological or cytological techniques such as immunohistochemistry or flow cytometry, depending on the nature of the input sample. These methods identify cells by making use of antibodies that can bind to T cell-specific epitopes, which should be available and accessible in the samples of interest [5, 6]. For that reason, these methodologies become problematic when the quality or quantity of specimens is limited [7].

More recently, high-resolution technologies (e.g., single-cell RNA sequencing and mass cytometry) have become available to study the presence of immune cells in mixed populations. These approaches, however, have even higher requirements concerning sample quality and quantity than traditional methods and remain financially and technically challenging for common use in research or diagnostics.

Alternatively, the presence of immune cells may be estimated from bulk "omics" data. Based on cell type-specific signature matrices, bulk gene expression or DNA methylation data can be computationally separated into its cellular components, a process called "deconvolution" [8, 9]. These approaches, however, are often less accurate when analyzing mixtures with unknown content or noise (such as cancer cells) and frequently show skewed or nonlinear relationships when compared against ground-truth measurements [8].

As another alternative, the abundance of T cells can be quantified by elaborating the genetic dissimilarities of the TR genes between T cells (i.e., rearranged) and non–T cells (i.e., in germline configuration). While various genomic approaches have been developed, these methods are usually very complex and not entirely quantitative. For example, multiplex PCR-based techniques, like the BIOMED-2 approach, only demonstrate relative differences in V(D)J gene usage and are performed to reveal the clonal expansion of specific T cells, rather than a general quantification of all T cells [2]. High-throughput sequencing can be used to analyze the full repertoire of V(D)J-rearranged TR genes. This approach is, however, relatively vulnerable to preferential amplification, which also limits the possibilities for an absolute quantification. Currently, one of the best solutions is the commercially available ImmunoSEQ™ Assay (Adaptive Biotechnologies). This sequencing-based method makes use of spiked synthetic control DNA, which represents a complete immune repertoire and is co-amplified with the target DNA. Such inline controls allow for the normalization of preferential amplification and offer a more accurate quantification of T cells, as recently demonstrated in melanoma and carcinoma [10, 11]. Nevertheless, it remains a complex, expensive, and time-consuming procedure to obtain a simple T cell quantification.

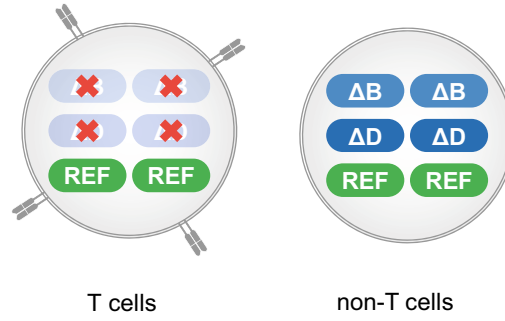


Fig. 1 Schematic overview of the availability of the T cell markers (ΔB in the TRB gene and ΔD in the TRD gene) and stable genomic reference (REF) in T cells and non–T cells. Due to T cell receptor rearrangements, ΔB and ΔD are biallelically absent in T cells specifically. In contrast, REF is present on both alleles in all cells

To overcome these hurdles, we developed a novel, digital PCR-based methodology to measure the abundance of T cells from a bulk DNA sample [12]. Our approach is based on unique, generic markers for rearranged TRB and TRD genes (named ΔB and ΔD, respectively) that facilitate a robust and simple T cell quantification. Due to TR rearrangements, mature T cells have completely lost ΔB and ΔD, whereas the markers are biallelically present in other cells (Fig. 1). By simply comparing the absolute abundance of ΔB or ΔD to a stable genomic reference DNA marker (abbreviated as "REF") that is biallelically present in all cells, the fraction of T cells can be determined based on bulk DNA [12]. Our method can be performed using only 20 ng of DNA and showed a highly accurate and linear relationship when compared to flow cytometry in blood samples from healthy donors and lymphoma patients (Fig. 2) [12]. Moreover, we successfully applied this approach to determine the T cell content of primary uveal melanomas [13]. Recently, it was used as part of assays to quantify the number of infected cells with human T cell leukemia virus type 1 (HTLV-1) and human immunodeficiency virus (HIV) [14, 15]. Furthermore, our methodology has translational applications in validating the purity of isolated or sorted populations of T cells and non–T cells [14].

A drawback of our approach lies in its sensitivity to pathogenic genetic alterations that affect the copy number of the various marker loci. While such variation is unusual in benign samples, copy number alterations (CNAs) are frequently seen in malignancies and premalignant conditions [16]. In such conditions, healthy T cells may be mixed with copy number unstable cancer cells, which can complicate the mathematical interpretation of the obtained marker quantifications. On the one hand, the genomic reference may be lost or gained as part of a chromosomal CNA. This problem

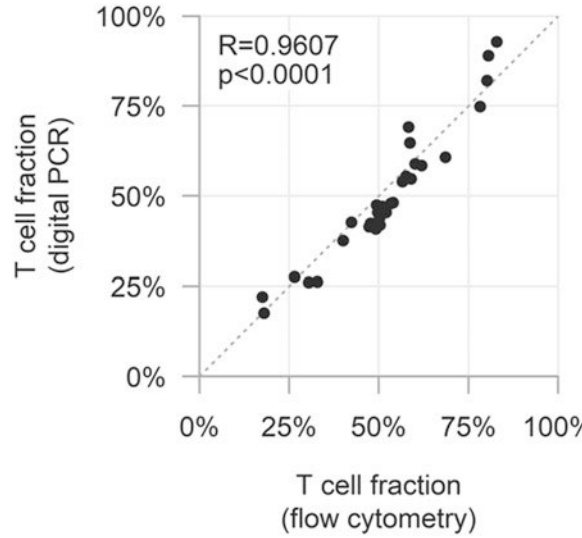


Fig. 2 Comparison of T cell quantifications in 30 peripheral blood mononuclear cell samples from healthy donors and lymphoma (Sézary syndrome) patients obtained by gold standard flow cytometry (measured by CD3+, *x*-axis) and digital PCR (measured by ΔB, *y*-axis) [12]. A strong and linear correlation is observed (Pearson $R = 0.9607$, $p < 0.0001$), demonstrating the high accuracy of our approach

was illustrated earlier and is a common pitfall of various molecular techniques [17]. However, it can be easily resolved by using a target at another chromosome as reference. The identification and selection of such sample-specific stable reference may be supported by tumor-type specific knowledge about common copy number alterations. On the other hand, a CNA in admixed cancer cells may disturb the abundance of our T cell marker (ΔB in the TRB gene on chromosome 7q34 or ΔD in the TRD gene on chromosome 14q11.2). Consequently, this gain or loss of T cell marker DNA may be unjustly attributed to the absence or presence of T cells, leading to under- or overestimated fractions. The strict genomic locations of the T cell markers, however, prevent from freely switching to another chromosome to overcome this problem. Based on copy number profiles of more than 10,000 cases spanning 31 tumor types from the TCGA pan-cancer dataset [16, 18], we previously showed that CNAs involving the ΔB and ΔD marker loci are present in ~24% and ~17% of the tumors [19]. These frequencies indicate that our original methodology (referred to as the *classic model*) gives incorrect T cell fractions in on average one out of four (ΔB) or one out of five (ΔD) of the cancer specimens.

As a robust solution for this problem, we developed an extension (referred to as the *adjusted model*) of our original experimental setup, which enables the recognition and adjustment of copy

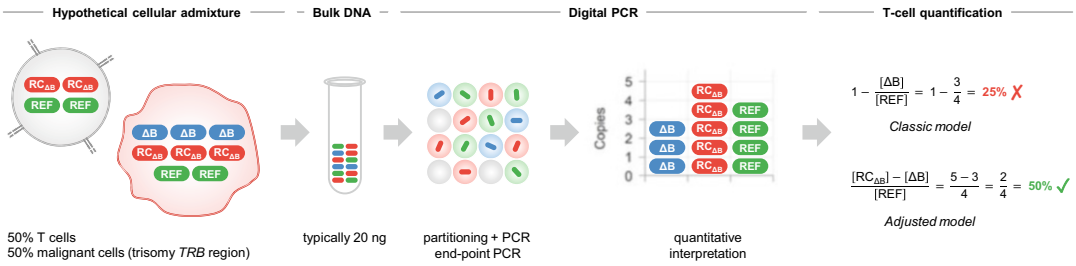


Fig. 3 Workflow to obtain the fraction of T cells from a hypothetical cellular admixture consisting of 50% healthy T cells and 50% copy number unstable non–T cells with a gain of the ΔB T cell marker region at chromosome 7. Typically, 20 ng of isolated DNA is analyzed for T cell marker ΔB, regional corrector ($RC_{\Delta B}$), and reference (REF) using digital PCR, which involves the compartmentalization of the complete PCR reaction (DNA and reagents) into a large number of nanoliter-sized droplets. For each of the DNA markers, the random distribution of the reaction mixture over the droplets results in a certain fraction of the droplets containing this target. PCR amplification only takes place in these droplets and results in a distinctive fluorescence intensity (droplets scored as "positive"). The number of positive droplets of all droplets can then be used to determine the abundances of all DNA targets, by which the T cell fractions can be calculated. Using the *classic model*, the T cell marker locus CNA is not recognized, and an incorrect T cell fraction of 25% is calculated. Following the *adjusted model*, the CNA is detected and adjusted for and a correct T cell fraction of 50% is calculated

number instability involving the T cell marker region. This enhanced approach relies on a so-called regional corrector that measures the copy number of the ΔB or ΔD marker locus. In contrast to the T cell marker, the regional corrector should not be deleted by TR rearrangements and is therefore biallelically present in all T cells. When a CNA in non–T cells involves the T cell marker region, the regional corrector will be affected likewise, allowing for a mathematical correction of the disrupted T cell quantification [19].

The actual quantifications of the T cell marker, regional corrector, and stable reference are obtained via custom-designed PCR assays (consisting of primers and fluorescently labelled probes) using digital PCR, as illustrated in Fig. 3. This technique involves the compartmentalization of a PCR reaction into a large number of small partitions, which are nanoliter-sized droplets when using the Droplet Digital™ PCR System (Bio-Rad Laboratories, Hercules, USA). For each of the DNA markers, the random distribution of the reaction mixture over the droplets results in a certain fraction of the droplets containing this target. PCR amplification only takes place in these droplets and results in a distinctive fluorescence intensity (droplets scored as "positive"). In contrast, droplets without initial presence of the target (but still containing nontarget DNA) remain unaltered and show a low background level of fluorescence (droplets scored as "negative" or "empty"). As usually an end-point PCR is carried out, all positive droplets will have a comparable fluorescence. "Digital" in digital PCR refers to this

dichotomous way of scoring: each droplet can only be positive or negative for a certain target. The number of positive droplets reflects the abundance of the measured DNA target: the more targets are to distribute, the more droplets will be filled and eventually scored as positive. This relationship is, however, not linear, as the random DNA distribution can also lead to droplets containing more than one target molecule. Instead, the relation between positive droplets and number of targets follows a Poisson distribution [20, 21]. For that reason, the final phase of a digital PCR experiment consists of a mathematical interpretation of the experimental outcomes. The statistical uncertainty of the obtained results is usually presented with a 95% confidence interval.

In this chapter, we describe how T cell quantifications can be obtained from bulk DNA samples using multiplex digital PCR. The experimental setup includes the concurrent quantification of three different DNA targets within one reaction: one of the unique T cell DNA markers (ΔB or ΔD), a regional corrector, and an independent reference DNA marker. By mathematically integrating the measurements of all three markers, T cells can be accurately quantified in both copy number stable and unstable DNA samples, as we previously validated [12, 19].

2 Materials

1. QX200™ Droplet Digital™ PCR System with Automated Droplet Generator, Droplet Reader and QuantaSoft™ software (Bio-Rad, *see* **Note 1**).
2. ddPCR™ 96-Well Plates (Bio-Rad).
3. Pierceable adhesive foil seals, for example, Microseal® "F" foil seals (Bio-Rad; *see* **Note 2**).
4. PX1™ PCR Plate Sealer (Bio-Rad) or comparable equipment, with compatible pierceable heat-sealing foil seals (Bio-Rad; *see* **Note 2**).
5. T100™ thermal cycler (Bio-Rad) or comparable programmable PCR thermal cycler with adjustable ramp rates, compatible with the ddPCR™ 96-Well Plates.
6. 2× ddPCR™ Supermix for Probes (No dUTP, Bio-Rad; *see* **Note 3**).
7. ddPCR™ probe assays (Bio-Rad) or 20× TaqMan probe assays (Sigma-Aldrich, Gillingham, UK), consisting of a set of primers and a fluorescently labelled hydrolysis probe (*see* Table 1, **Notes 4** and **5**), for:

Table 1
Overview of PCR assays

ΔB	Forward primer (900 nM): Reverse primer (900 nM): Probe (HEX-labelled, 250 nM):	5′ GCCATGCACTTTCCCTTTCG 3′ 5′ ACAGAGTCCATCCACAGGG 3′ 5′ TGGACCCTCACAGAGGGAGCA 3′
TRBC2	Assay (Sigma-Aldrich, FAM-labelled, premixed)	
	Context sequence: CCCCTGAAACCCTGAAAATGTTCTCTCTTCCACAGGTCAAGAGAAAGGAT TCCAGAGGCTAGCTCCAAAACCATCCCAGGTCATTCTTCATCCTCACCCA GGATTCTCCTGTACCTGCTCCCAATC	
TTC5	Assay (dHsaCP2506733, Bio-Rad, HEX-labelled, premixed)	
	Context sequence: TGGTCGCGATGCCACTGTGGCAACAGCCTGGCTGCTGGATCCCTGAGGC TTCCCATTCACCACTAGCAGGAGGGGCGTCTCCACTCGAACACTGGAAAA GGAATAGTCCTAGAAAAGACAGAC	

(a) A T cell marker, here ΔB (HEX-labelled).

(b) A regional corrector for the chosen T cell marker, here TRBC2 (FAM-labelled).

(c) A stable genomic reference, here TTC5 (HEX-labelled).

8. DNA restriction enzyme HaeIII with 10× CutSmart® buffer (both New England Biolabs, Ipswich, USA; *see* **Note 6**).
9. DNA of good quality and with high molecular weight at a concentration of ideally 20 ng/μL (*see* **Note 7**).
10. DG32™ Automated Droplet Generator Cartridges (Bio-Rad).
11. Automated Droplet Generation Oil for Probes (Bio-Rad).
12. Pipet Tips for the AutoDG™ System (Bio-Rad).
13. ddPCR™ Droplet Reader Oil (Bio-Rad).
14. Software to analyze multiplex digital PCR experiments, such as Roodcom WebAnalysis (https://www.roodcom.nl).
15. General lab equipment (centrifuge, vortex mixer, pipettes).

3 Methods

In Subheadings 3.1 and 3.2, we introduce the multiplex experimental setup to quantify T cells in a sample of interest. In Subheadings 3.4, 3.5, and 3.6, we describe the complete workflow to perform the experiments. Finally, in Subheading 3.7, we discuss the legitimacy of the approach in the analysis of samples with a lymphoproliferative component.

3.1 Choosing an Experimental Setup

The experimental setup to quantify T cells in (possibly) copy number unstable DNA samples involves the measurement of three distinct DNA targets: a T cell marker, its regional corrector, and a stable genomic reference. We previously identified two DNA targets (ΔB and ΔD) that fulfill the role of generic T cell marker: in mature T cells, both markers are biallelically absent [12]. Both assays have been successfully applied to quantify the proportion of T cells [12–15, 17], and for that reason, either ΔB or ΔD can be used in the experimental setup.

The regional corrector is used to quantify the (possibly altered) copy number of the chosen T cell marker locus. Therefore, it should measure a DNA target located in close genomic proximity to the T cell marker, but its abundance should not be altered due to TR rearrangements. For ΔB, we recently validated a regional corrector targeting TRBC2, the secondary constant domain and last region of the TRB gene complex [19]. As TRBC2 is not deleted as part of VDJ rearrangements, it can be considered the closest genomic locus functioning as regional corrector for ΔB. For ΔD, a candidate regional corrector may be found in the constant gene of the TRA gene, as TRD itself is located within TRA and consequently may be lost due to TR rearrangements [22].

The stable genomic reference should measure a copy number invariant DNA target that is biallelically present in all cells. This marker measures the total number of genomes (and thus cells) and is used to normalize the relative loss of the T cell marker. Hereby, the T cell fraction (i.e., fraction of all cells that is a T cell) can be calculated. The selection of a stable reference in cancer specimens may be guided by tumor-type specific information or measurements in individual samples, as we illustrated previously [17].

The experimental setup of this protocol consists of T cell marker ΔB with regional corrector TRBC2 (both on chromosome 7q34) and stable reference TTC5 (chromosome 14q11.2).

3.2 Multiplex Digital PCR

In traditional multiplex (q)PCR reactions, multiple targets of interest can be analyzed simultaneously by using differentially colored fluorescent probes. The QX200™ Droplet Digital™ PCR System is, however, limited to the detection in two optical channels (i.e., FAM and HEX/VIC). Still, it is possible to measure more than two targets in a single digital PCR reaction. By varying the concentration of same-colored probes, distinct probes (and thus distinct targets) may be identified based on different end-point fluorescence intensities [23, 24]. Here, we make use of this strategy to measure the regional corrector (single FAM-labelled assay), the T cell marker (HEX-labelled assay, low concentration), and the genomic reference (HEX-labelled assay, high concentration) in a triplex reaction (*see* Fig. 3). As the signal intensities may differ between assay batches or dilutions, optimization and validation experiments

are needed for each channel with more than one assay. Consequently, in our setup, the mixing of two HEX-labelled assays (ΔB and TTC5) should be optimized.

1. In our triplex experimental setup, three different DNA targets are measured. As droplets (by chance) can contain any combination of these targets, $2^3 = 8$ distinct clusters may be present in the two-dimensional space. With one FAM-labelled assay (TRBC2), $2^1 = 2$ populations of droplets may appear on channel 1: TRBC2− and TRBC2+ clusters. The two HEX-labelled assays give rise to $2^2 = 4$ populations on channel 2: ΔB−/TTC5−, ΔB+/TTC5−, ΔB−/TTC5+, and ΔB+/TTC5+ clusters. To distinguish the two assays on channel 2, one of them should be lowly concentrated (giving a relatively low amplitude) and the other should be highly concentrated (giving a relatively high amplitude). The double-positive cluster (ΔB+/TTC5+) will then have an amplitude that is roughly the sum of the individual single-positive amplitudes. Increasing the input of an assay usually increases the amplitude, and vice versa. To make an educated guess on the relative concentrations of the two HEX-labelled assays, the assay amplitudes (1× input concentration) observed in duplex experiments can be very informative.
2. Perform a multiplex digital PCR experiment with the estimated input volumes of the different assays and evaluate the obtained results. In Fig. 4a, the 2D plot of such multiplex is shown. Although four clusters are entirely separated from each other, the clusters in the middle are still overlapping. The separation of these clusters may be better by slightly decreasing the input of ΔB, while increasing the input of TTC5.
3. Adjust the input of the assays and repeat the multiplex experiment accordingly. Repeat this step until all clusters become visually separated. In Fig. 4b, an optimized multiplex experiment is shown in which all clusters are distinguishable and not overlapping anymore.
4. To verify that correct quantifications are obtained, we recommend to analyze various control DNA samples (*see* **Note 8**) in both duplex and multiplex experiments. The concentrations and ratios obtained with the optimized multiplex setup should be similar to those acquired by the individual duplex experiments (as shown in Fig. 4c).

3.3 Pre-PCR Preparation of Reaction Mixture

1. Bring the ddPCR™ Supermix for Probes (No dUTP) to room temperature and mix thoroughly by pulse-vortexing the tube.
2. Bring the primer/probe mixes of the ΔB, regional corrector, and reference assays to room temperature, and mix thoroughly by pulse-vortexing the tube. Centrifuge briefly to ensure all contents are collected at the bottom of the tube.

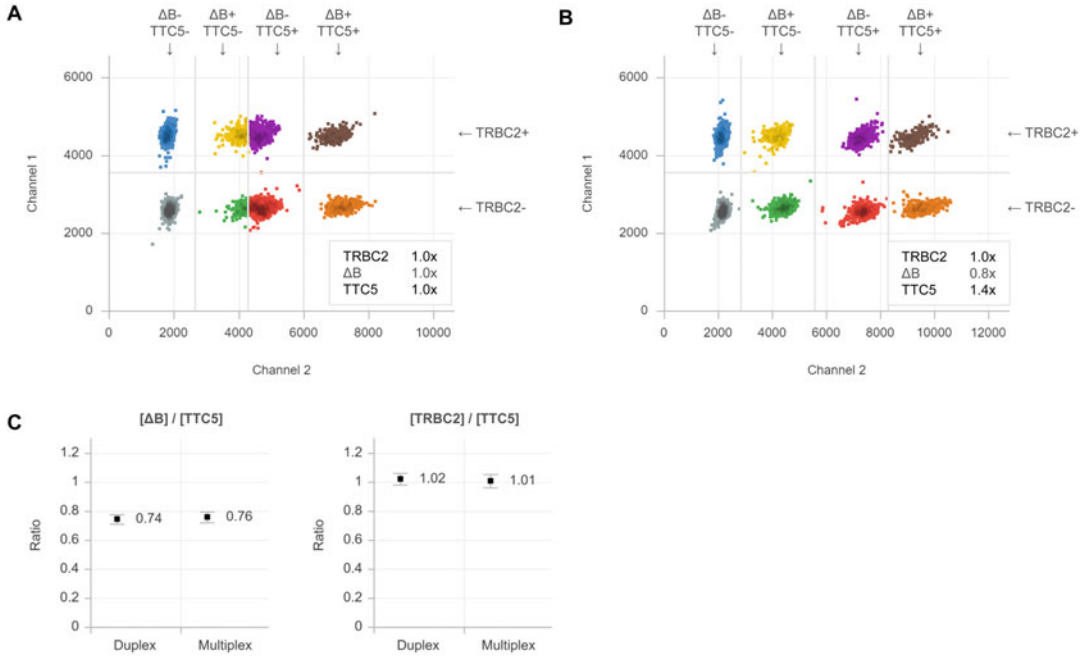


Fig. 4 2D plot of 1 × 2 multiplex digital PCR analyzing regional corrector TRBC2 on channel 1 (FAM) and T cell marker ΔB (assay with lowest fluorescence) and stable reference TTC5 (assay with highest fluorescence) on channel 2 (HEX) in a healthy, copy number stable PBMC sample. Initially (**a**) clusters are overlapping, but after optimization (**b**) all eight clusters are separated. To validate that correct quantifications are obtained with this multiplex, the concentration ratios [ΔB]/[TTC5] and [TRBC2]/[TTC5] are compared with the results obtained in separate duplex experiments (**c**)

3. Prepare a reaction mixture for each DNA sample as follows (*see* **Note 9**):
   (a) 11.0 μL ddPCR™ Supermix for Probes (No dUTP)
   (b) 1.0 μL TRBC2 assay primer/probe mix
   (c) 0.8 μL ΔB assay primer/probe mix (optimized input; *see* Subheading 3.2)
   (d) 1.4 μL TTC5 reference assay primer/probe mix (optimized input; *see* Subheading 3.2)
   (e) 1.0 μL of 2 U/μL HaeIII restriction enzyme, diluted in 1× CutSmart® buffer (*see* **Note 6**)
   (f) 20 ng DNA (*see* **Note 7**)
   (g) Nuclease-free H_2O up to a total volume of 22.0 μL.
4. Mix thoroughly by pulse-vortexing the PCR reaction mix. Centrifuge briefly to ensure all contents are collected at the bottom of the tube, and dispense the reaction mixture in a ddPCR™ 96-Well Plate.

5. Adhere an *adhesive* foil seal to the plate and centrifuge for 1 min at 100 × *g* to ensure all contents are collected at the bottom of each well.
6. Carefully remove the foil (to prevent well-to-well contamination) and gently mix the contents in the plate by pipetting up and down at least 10 times. Make sure not to introduce air bubbles in the content of the wells.
7. Adhere a new *adhesive* foil seal to the plate and centrifuge again for 1 min at 100 × *g*.

3.4 Droplet Generation and PCR Amplification

1. Place the prepared plate into the Automated Droplet Generator, and follow the instructions in the user manual to generate the droplets (*see* **Note 10**). To prevent evaporation, the generated droplets should be collected in a second 96-well plate placed into a properly frozen cooling block.
2. After the droplet generation has finished, remove the plate from the cooling block immediately, and cover it with a *heat-sealed* foil seal, for example, using the PX1™ PCR Plate Sealer. As the generated droplets are fragile in this stage, it is advised to handle the plate with care and to proceed with the next step directly.
3. Place the plate with the generated droplets into a T100™ Thermal Cycler or comparable programmable PCR cycler suitable for the described 96-well plates. The PCR amplification should be carried out with a lid temperature of 105 °C, a ramp rate set to 2 °C/s, and the reaction volume set to 40 μL, using the following protocol:
   (a) 10 min at 95 °C
   (b) 30 s at 94 °C and 1 min at 60 °C, for 40 cycles
   (c) 10 min at 98 °C
   (d) 30 min at 4 °C and (optional) cooling at 12 °C until droplet reading (*see* **Note 11**).

3.5 Droplet Reading

1. After the PCR has been carried out, place the 96-well plate into the plate holder of the QX200™ Droplet Reader and load the holder into the droplet reader.
2. Create a template in the experimental setting section of the QuantaSoft™ software (*see* Fig. 5), and follow further instructions as given in the user manual to start droplet reading (*see* **Note 10**).

3.6 Interpretation of Results

1. After reading the droplets, the end point fluorescence levels of the accepted droplets are available in the QuantaSoft™ software. However, this application has no functionalities to directly analyze multiplex digital PCR experiments. Various

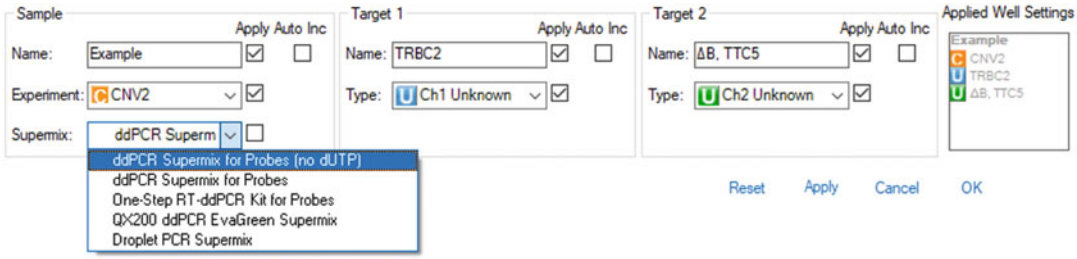


Fig. 5 Example of defining the well template settings for our multiplex experimental setup in the QuantaSoft™ software

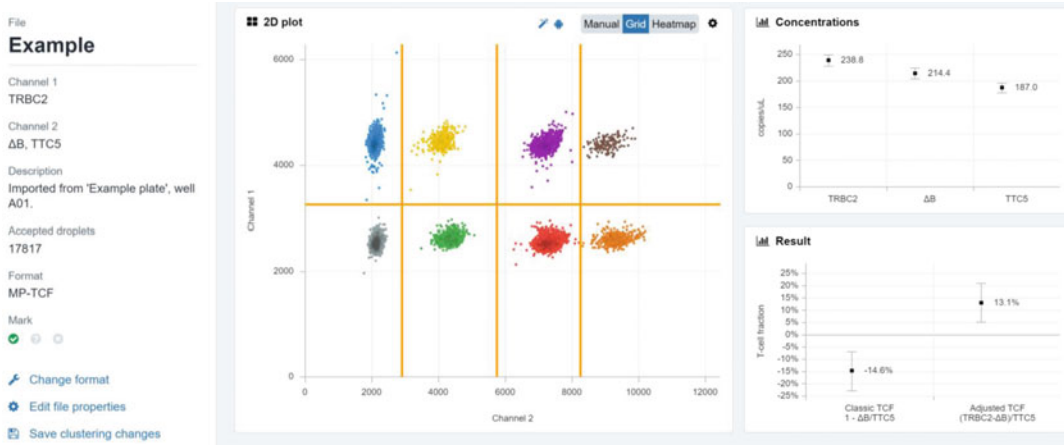


Fig. 6 Example of the analysis of our multiplex experimental setup in Roodcom WebAnalysis. Here, 20 ng DNA from a malignant melanoma is analyzed. In the panel "2D plot," eight distinct clusters are detected that are used to calculate the concentrations of the individual targets (in the panel "Concentrations"). When the experimental format is set to "T cell multiplex" (MP-TCF), the T cell fractions according to the *classic* and *adjusted model* and their associated 95% confidence interval are calculated automatically (in the panel "Results"). In this tumor sample, TTC5 represents the stable genomic reference, but a chromosomal gain of the T cell marker region makes that [ΔB] and [TRBC2] are higher than [TTC5]. Consequently, the T cell fraction according to the *classic model* is negative (−14.6%), which is incorrect and biologically impossible. Using the *adjusted model*, however, the CNA is detected and properly normalized, leading to a positive and correct T cell fraction (13.1%)

(third-party) software applications have been developed and are available for downstream analysis, e.g., QuantaSoft Analysis Pro or QX Manager (both Bio-Rad), ddPCRclust [25] or Roodcom WebAnalysis (https://www.roodcom.nl). Here, we make use of Roodcom WebAnalysis to analyze the data (*see* Fig. 6).

2. As discussed in Subheading 3.2, an optimized triplex reaction will result in $2^3 = 8$ distinct clusters of droplets. Although thresholding may have been carried out automatically by the software, manual examination and, if necessary, adjustment are recommended. To validate that correct quantifications are

Table 2
Formulas to calculate the T-cell fraction (TCF) and its 95% confidence interval [TCF_{low}; TCF_{high}] according to the *classic model*, without correction for CNAs affecting the T cell marker locus, based on the absolute numbers of droplets scored positive for markers ΔB ($n_{\Delta B+}$) and REF (n_{REF+}) and the total number of droplets analyzed (n_{total})

$p_A = \frac{n_{\Delta B+}}{n_{total}}$ $p_{A,\ low} = p_A - 1.96\sqrt{\frac{p_A \cdot (1-p_A)}{n_{total}}}$ $p_{A,high} = p_A + 1.96\sqrt{\frac{p_A \cdot (1-p_A)}{n_{total}}}$
$\lambda_A = -\ln(1-p_A)$ $\lambda_{A,\ low} = -\ln(1-p_{A,\ low})$ $\lambda_{A,\ high} = -\ln(1-p_{A,\ high})$
$p_B = \frac{n_{REF+}}{n_{total}}$ $p_{B,\ low} = p_B - 1.96\sqrt{\frac{p_B \cdot (1-p_B)}{n_{total}}}$ $p_{B,high} = p_B + 1.96\sqrt{\frac{p_B \cdot (1-p_B)}{n_{total}}}$
$\lambda_B = -\ln(1-p_B)$ $\lambda_{B,\ low} = -\ln(1-p_{B,\ low})$ $\lambda_{B,\ high} = -\ln(1-p_{B,\ high})$
$H_{top} = \lambda_{A,\ high} - \lambda_A$ $H_{bottom} = \lambda_A - \lambda_{A,\ low}$
$W_{right} = \lambda_{B,\ high} - \lambda_B$ $W_{left} = \lambda_B - \lambda_{B,\ low}$
$TCF = 1 - \frac{\lambda_A}{\lambda_B}$ $TCF_{low} = 1 - \frac{\lambda_A \cdot \lambda_B + \sqrt{{\lambda_A}^2 \cdot {\lambda_B}^2 - \left({H_{top}}^2 - {\lambda_A}^2\right) \cdot \left({W_{left}}^2 - {\lambda_B}^2\right)}}{{\lambda_B}^2 - {W_{right}}^2}$ $TCF_{high} = 1 - \frac{\lambda_A \cdot \lambda_B - \sqrt{{\lambda_A}^2 \cdot {\lambda_B}^2 - \left({H_{bottom}}^2 - {\lambda_A}^2\right) \cdot \left({W_{right}}^2 - {\lambda_B}^2\right)}}{{\lambda_B}^2 - {W_{right}}^2}$

obtained, we propose various control experiments to be performed next to the samples of interest (*see* **Note 8**). For a general evaluation of the experimental performance, we advise to follow the "MiQE" guidelines described by Huggett et al. [26].

3. Based on the clustering of the droplets, the number of positive droplets per assay and the total number of accepted droplets are used to calculate the concentrations of the individual targets (indicated by square brackets, e.g., [ΔB] is the concentration of T cell marker ΔB; *see* Table 2 for all formulas). In Roodcom WebAnalysis, these results are automatically available in the panel "Concentrations" (*see* Fig. 6).
4. The T cell fraction following the *classic model* (assuming the T cell marker locus is not involved in any copy number alteration) can be calculated as follows:

$$\text{T cell fraction} = 1 - \frac{[\Delta B]}{[TTC5]}$$

Table 3
Formulas to calculate the T cell fraction (TCF) and its 95% confidence interval [TCF_{low}; TCF_{high}] according to the *adjusted model*, with correction for CNAs affecting the T cell marker locus, based on the absolute numbers of droplets scored positive for markers ΔB ($n_{\Delta B+}$), $RC_{\Delta B}$ ($n_{RC_{\Delta B}+}$), and REF (n_{REF+}) and the total number of droplets analyzed (n_{total})

$p_{A1} = 1 - \frac{n_{RC_{\Delta B}+}}{n_{total}}$ $p_{A2} = 1 - \frac{n_{\Delta B+}}{n_{total}}$
$\lambda_A = -\ln\left(\frac{p_{A1}}{p_{A2}}\right)$ $\lambda_{A,low} = \lambda_A - 1.96 * \sqrt{\frac{p_{A1}^{-1} + p_{A2}^{-1} - 2}{n_{total}}}$ $\lambda_{A,high} = \lambda_A + 1.96 * \sqrt{\frac{p_{A1}^{-1} + p_{A2}^{-1} - 2}{n_{total}}}$
$p_B = \frac{n_{REF+}}{n_{total}}$ $p_{B,\ low} = p_B - 1.96\sqrt{\frac{p_B \cdot (1-p_B)}{n_{total}}}$ $p_{B,high} = p_B + 1.96\sqrt{\frac{p_B \cdot (1-p_B)}{n_{total}}}$
$\lambda_B = -\ln(1 - p_B)$ $\lambda_{B,\ low} = -\ln(1 - p_{B,\ low})$ $\lambda_{B,\ high} = -\ln(1 - p_{B,\ high})$
$H_{top} = \lambda_{A,\ high} - \lambda_A$ $H_{bottom} = \lambda_A - \lambda_{A,\ low}$
$W_{right} = \lambda_{B,\ high} - \lambda_B$ $W_{left} = \lambda_B - \lambda_{B,\ low}$
$TCF = \frac{\lambda_A}{\lambda_B}$ $TCF_{low} = \frac{\lambda_A \cdot \lambda_B - \sqrt{\lambda_A^2 \cdot \lambda_B^2 - \left(H_{bottom}^2 - \lambda_A^2\right) \cdot \left(W_{right}^2 - \lambda_B^2\right)}}{\lambda_B^2 - W_{left}^2}$ $TCF_{high} = \frac{\lambda_A \cdot \lambda_B + \sqrt{\lambda_A^2 \cdot \lambda_B^2 - \left(H_{top}^2 - \lambda_A^2\right) \cdot \left(W_{left}^2 - \lambda_B^2\right)}}{\lambda_B^2 - W_{left}^2}$

The T cell fraction following the *adjusted model* (which automatically corrects possible copy number alterations involving the T cell marker locus) can be calculated as follows:

$$\text{T cell fraction} = \frac{[\text{TRBC2}] - [\Delta\text{B}]}{[\text{TTC5}]}$$

When the experimental format is set to "T cell multiplex" (MP-TCF) in Roodcom WebAnalysis, these results are automatically available in the panel "Results" (*see* Fig. 6).

3. We recommend to construct confidence intervals for each obtained T cell fraction (*see* Tables 2 and 3 for all formulas [19, 20]). These intervals, usually with a confidence level of 95%, indicate the precision of the calculated fractions: the wider the interval, the more uncertain the quantification is. The absolute width of such confidence interval depends on several factors, including the amount of DNA input, the total number of accepted droplets, and the copy number of the T cell marker locus. Generally spoken, the width of the interval can be decreased by analyzing more DNA (when available). In Roodcom WebAnalysis, these confidence intervals are automatically presented in the panel "Results" (*see* Fig. 6).

3.7 Analysis of Samples with a Lymphoproliferative Component

This protocol is designed for the quantification of copy number stable T cells, mixed with (potentially) unstable non–T cells. For that reason, particular care should be taken in the analysis of samples from a lymphoproliferative origin, such as T and B cell lymphomas and leukemias. The maturation stage during onset of the malignancy, clonality, and genetic stability of (pre-)T cell malignancies may have different consequences on the availability of our T cell markers. Whereas mature T cell proliferations have undergone VDJ rearrangement and have generally lost the marker on both alleles, the TR genes in immature T cell proliferations might be incompletely rearranged. As a result, our T cell markers may be mono- or biallelically present in (malignant) T cells, not following our mathematical model. Moreover, in lymphoid malignancies, TR gene rearrangements are not restricted to the T cell lineage only. For example, in precursor-B-acute lymphoblastic leukemias and in acute myeloid leukemias, cross-lineage rearrangements of TR genes are found [27, 28]. As a consequence, our T cell markers may be deleted in these admixed leukemic B cells, resulting in an overestimation of the T cell fraction. Therefore, it would be generally recommended to ensure the absence of any of these alterations when analyzing samples with a lymphoproliferative component.

4 Notes

1. This protocol makes use of the QX200™ Droplet Digital™ PCR System with Automated Droplet Generator and Droplet Reader (Bio-Rad). While beyond the scope of this chapter, all experiments can also be carried out with a Manual Droplet Generator (Bio-Rad) or even using another digital PCR system.
2. Two different types of sealing foils are used:
   (a) *Adhesive* foils are used to cover the plate before droplet generation and can be removed easily. These foils should be pierceable by the Automatic Droplet Generator.
   (b) *Heat-sealed* foils are used to cover the plate after droplet generation and during the PCR. These foils should be compatible with the heating steps in the PCR thermal cycler and should be pierceable by the Droplet Reader. Make sure that only a single foil is used and that the plate is sealed completely.
3. The 2× ddPCR™ Supermix for Probes (No dUTP) can be stored in the fridge (short term) or the freezer (long term).
4. To preserve the quality of the fluorescently labelled probes, it is advised to protect them from light as much as possible. Primers

(900 nM) and probes (250 nM) are usually mixed together (the "assay") and can be stored in the fridge (short term) or the freezer (long term).

5. In this chapter, we describe an experimental setup consisting of T cell marker ΔB with regional corrector TRBC2 (both on chromosome 7q37) and stable reference TTC5 (chromosome 14q11.2), similar as previously published [19]. The regional corrector is the single FAM-labelled assay; the T cell marker and the genomic reference are two HEX-labelled assays. Still, it is possible to use other DNA references or fluorescent labels (*see* Subheading 3.1).
6. The addition of a DNA restriction enzyme improves the accessibility of the various DNA targets molecules and may result in better 2D plots. Moreover, it prevents physical linkage between the T cell marker and the regional corrector [19]. It is important that the restriction enzyme does not digest any of the various amplicon sequences. In this protocol, we make use of restriction enzyme HaeIII with its associated 10× CutSmart® buffer (both New England Biolabs), which is compatible with the chemistry used by the QX200™ Droplet Digital™ PCR System.
7. Based on our experience, T cell quantifications can be successfully performed with 20 ng of input DNA (e.g., 1 μL of 20 ng/μL) per reaction, but this can generally be decreased or increased if necessary. The precision of the quantification, however, also depends on the amount of input DNA and can be visualized by the construction of 95% confidence intervals.
8. We previously introduced various controls to validate obtained T cell quantifications [12, 17, 19]. Besides testing these controls, DNA or originating cells of various controls may be mixed to generate standard curves. Examples:
   (a) DNA from 100% non–T cells (e.g., cultured fibroblasts, copy number stable).
   (b) DNA from >95% T cells (e.g., purified/sorted T cells from blood, copy number stable).
   (c) DNA from 100% T cells (e.g., a T cell cell line, copy number stable).
   (d) DNA from 100% non–T cells with a CNA affecting the T cell marker locus (e.g., a pure cancer cell line).
   (e) DNA from samples with a known T cell fraction (e.g., blood samples measured for T cell content using flow cytometry).
9. As the QX200™ Droplet Digital™ PCR System always works with 8 wells (the droplet generation can only be performed per column of a 96-well plate), it is advised to fill up all empty wells

with (control) samples. To reduce the number of pipetting actions and enhance the practical performance, the PCR reaction mixture (without DNA) can usually be prepared for multiple wells at once. As final step, the DNA samples can then be added directly to the various wells.

10. The latest versions of the user manual for the Automated Droplet Generator and Droplet Reader with QuantaSoft™ software are available via the Bio-Rad website (https://www.bio-rad.com).
11. Post-PCR cooling of the plate enhances the performance of experiments carried out with the QX200™ Droplet Digital™ PCR System [29]. For best results, the plate should be placed at 12 °C overnight.

References

1. Davis MM, Bjorkman PJ (1988) T-cell antigen receptor genes and T-cell recognition. Nature 334(6181):395–402
2. van Dongen JJ, Langerak AW, Bruggemann M, Evans PA, Hummel M, Lavender FL et al (2003) Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98-3936. Leukemia 17(12):2257–2317
3. Talmadge JE (2011) Immune cell infiltration of primary and metastatic lesions: mechanisms and clinical impact. Semin Cancer Biol 21(2): 131–138
4. Fridman WH, Galon J, Pages F, Tartour E, Sautes-Fridman C, Kroemer G (2011) Prognostic and predictive impact of intra- and peritumoral immune infiltrates. Cancer Res 71(17):5601–5605
5. Wood B, Jevremovic D, Bene MC, Yan M, Jacobs P, Litwin V (2013) Validation of cell-based fluorescence assays: practice guidelines from the ICSH and ICCS—part V—assay performance criteria. Cytometry B Clin Cytom 84(5):315–323
6. Walker RA (2006) Quantification of immunohistochemistry—issues concerning methods, utility and semiquantitative assessment I. Histopathology 49(4):406–410
7. de Hoog J, Dik WA, Lu L, Heezen KC, Ten Berge JC, Swagemakers SMA et al (2019) Combined cellular and soluble mediator analysis for improved diagnosis of vitreoretinal lymphoma. Acta Ophthalmol 97(6):626–632
8. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y et al (2015) Robust enumeration of cell subsets from tissue expression profiles. Nat Methods 12(5):453–457
9. Chakravarthy A, Furness A, Joshi K, Ghorani E, Ford K, Ward MJ et al (2018) Pan-cancer deconvolution of tumour composition using DNA methylation. Nat Commun 9(1):3220
10. Pruessmann W, Rytlewski J, Wilmott J, Mihm MC Jr, Attrill GH, Dyring-Andersen B et al (2020) Molecular analysis of primary melanoma T cells identifies patients at risk for metastatic recurrence. Nat Cancer 1(2):197–209
11. Van Abel KM, Routman DM, Moore EJ, Ma DJ, Yin LX, Fields PA et al (2020) T cell fraction impacts oncologic outcomes in human papillomavirus associated oropharyngeal squamous cell carcinoma. Oral Oncol 111:104894
12. Zoutman WH, Nell RJ, Versluis M, van Steenderen D, Lalai RN, Out-Luiting JJ et al (2017) Accurate quantification of T cells by measuring loss of germline T-cell receptor loci with generic single duplex droplet digital PCR assays. J Mol Diagn 19(2):236–243
13. de Lange MJ, Nell RJ, Lalai RN, Versluis M, Jordanova ES, Luyten GPM et al (2018) Digital PCR-based t-cell quantification-assisted deconvolution of the microenvironment reveals that activated macrophages drive tumor inflammation in uveal melanoma. Mol Cancer Res 16(12):1902–1911
14. Yurick D, Khoury G, Clemens B, Loh L, Pham H, Kedzierska K et al (2019) Multiplex droplet digital PCR assay for quantification of human T-cell leukemia virus type 1 subtype c

DNA proviral load and T cells from blood and respiratory exudates sampled in a remote setting. J Clin Microbiol 57(2):e01063-18

15. Levy CN, Hughes SM, Roychoudhury P, Reeves DB, Amstuz C, Zhu H et al (2021) A highly multiplexed droplet digital PCR assay to measure the intact HIV-1 proviral reservoir. Cell Rep Med 2(4):100243
16. Taylor AM, Shih J, Ha G, Gao GF, Zhang X, Berger AC et al (2018) Genomic and functional approaches to understanding cancer aneuploidy. Cancer Cell 33(4):676–689 e3
17. Zoutman WH, Nell RJ, van der Velden PA (2019) Usage of droplet digital PCR (ddPCR) assays for T cell quantification in cancer. Methods Mol Biol 1884:1–14
18. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA et al (2013) The cancer genome atlas pan-cancer analysis project. Nat Genet 45(10):1113–1120
19. Nell RJ, Zoutman WH, Calbet-Llopart N, Garcia AP, Menger NV, Versluis M et al (2021) Accurate quantification of T cells in copy number unstable DNA samples using multiplex digital PCR (submitted). J Mol Diagn 24(1):88–100. https://doi.org/10.1016/j.jmoldx.2021.10.007
20. Dube S, Qin J, Ramakrishnan R (2008) Mathematical analysis of copy number variation in a DNA sample using digital PCR on a nanofluidic device. PLoS One 3(8):e2876
21. Hindson BJ, Ness KD, Masquelier DA, Belgrader P, Heredia NJ, Makarewicz AJ et al (2011) High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. Anal Chem 83(22):8604–8610
22. Dik WA, Pike-Overzet K, Weerkamp F, de Ridder D, de Haas EF, Baert MR et al (2005) New insights on human T cell development by quantitative T cell receptor gene rearrangement studies and gene expression profiling. J Exp Med 201(11):1715–1723
23. Zhong Q, Bhattacharya S, Kotsopoulos S, Olson J, Taly V, Griffiths AD et al (2011) Multiplex digital PCR: breaking the one target per color barrier of quantitative PCR. Lab Chip 11(13):2167–2174
24. Whale AS, Huggett JF, Tzonev S (2016) Fundamentals of multiplexing with digital PCR. Biomol Detect Quantif 10:15–23
25. Brink BG, Meskas J, Brinkman RR, Wren J (2018) ddPCRclust: an R package and Shiny app for automated analysis of multiplexed ddPCR data. Bioinformatics 34(15):2687–2689
26. Huggett JF, Foy CA, Benes V, Emslie K, Garson JA, Haynes R et al (2013) The digital MIQE guidelines: minimum information for publication of quantitative digital PCR experiments. Clin Chem 59(6):892–902
27. Szczepanski T, Beishuizen A, Pongers-Willemse MJ, Hahlen K, Van Wering ER, Wijkhuijs AJM et al (1999) Cross-lineage T cell receptor gene rearrangements occur in more than ninety percent of childhood precursor-B acute lymphoblastic leukemias: alternative PCR targets for detection of minimal residual disease. Leukemia 13(2):196–205
28. Przybylski G, Oettle H, Ludwig W, Siegert W, Schmidt C (1994) Molecular characterization of illegitimate TCRδ gene rearrangements in acute myeloid leukaemia. Br J Haematol 87(2):301–307
29. Rowlands V, Rutkowski AJ, Meuser E, Carr TH, Harrington EA, Barrett JC (2019) Optimisation of robust singleplex and multiplex droplet digital PCR assays for high confidence mutation detection in circulating tumour DNA. Sci Rep 9(1):12620

Chapter 13

Gene Engineering T Cells with T-Cell Receptor for Adoptive Therapy

Dian Kortleve, Mandy van Brakel, Rebecca Wijers, Reno Debets, and Dora Hammerl

Abstract

Prior to clinical testing of adoptive T-cell therapy with T-cell receptor (TCR)-engineered T cells, TCRs need to be retrieved, annotated, gene-transferred, and extensively tested in vitro to accurately assess specificity and sensitivity of target recognition. Here, we present a fundamental series of protocols that cover critical preclinical parameters, thereby enabling the selection of candidate TCRs for clinical testing.

Key words T-cell receptor, T-cell engineering, TCR cloning, TCR annotation, Gene transfer, In vitro assays, Specificity, Sensitivity

1 Introduction

Adoptive therapy with T-cell receptor (TCR)-engineered T cells is based on the insertion of genes into the patient's T cells that encode for a TCR directed against a predefined tumor antigen and are re-infused back into the patient. Once transferred to the patient, TCR-engineered T cells specifically migrate toward and kill tumor cells that express this antigen. The promises and challenges of this form of immunotherapy are reviewed elsewhere [1, 2]. Here we provide an overview of steps and details of laboratory protocols necessary to obtain and test TCRs, thereby providing a platform for the identification and selection of those TCRs amenable for further preclinical studies and, when successful, clinical studies.

Epitope-specific T cells and their corresponding TCRs are generally retrieved from tumor-infiltrating lymphocytes (TILs) or peripheral blood mononuclear cells (PBMCs) derived from either patients or healthy donors. In some cases, frequencies of epitope-specific T cells can be amplified in co-culture systems with antigen-presenting cells (*not* part of this chapter, but well be described in Theaker et al. and Wölfl et al. [3, 4]). Epitope-specific T cells can be

Anton W. Langerak (ed.), *Immunogenetics: Methods and Protocols*, Methods in Molecular Biology, vol. 2453, https://doi.org/10.1007/978-1-0716-2115-8_13,

detected and isolated by fluorescent-activated cell sorting (FACS) using peptide-major histocompatibility complexes (pMHC). Then, the RNA of these sorted T cells can be isolated. In Subheadings 2.1 and 3.1, we present materials and protocols to obtain and sequence and identify TCRα and β chains from RNA isolated from pMHC-sorted T cells. TCRα and β sequences are identified with the SMARTer RACE cDNA Amplification Kit (Takara Bio) and Sanger sequencing, after which sequences are annotated using the IMGT database and the HighV-QUEST tool.

Depending on the presence and frequency of T-cell clones, a variable number of TCRα and β chains are identified, and single α and β chains can be co-introduced into T cells to test TCRαβ heterodimers. In Subheadings 2.2 and 3.2, we present materials and protocols to introduce TCRαβ genes into T cells. TCRα and β chains that are molecularly connected with a 2A linker are cloned into an expression vector and retrovirally transduced into T cells. To this end, packaging cells are transfected with the TCRα and β genes as well as retroviral helper constructs, which will enable the secretion of virus particles with RNA encoding the TCR gene construct. PBMCs from healthy donors are activated with stimulatory antibodies and/or cytokines and incubated with the virus particles, leading to a stable integration of TCR genes.

TCR-transduced T cells can be validated in vitro. In Subheadings 2.3 and 3.3, we present materials and protocols to assess TCR surface expression and sensitivity as well as the specificity of T cells expressing an epitope-specific TCR. The surface expression of the TCR transgene is measured using pMHC multimers at the single cell level via flow cytometry. Functional avidity of T cells expressing such TCR transgenes can be determined by measuring IFNγ secretion upon co-culture of these T cells with antigen-presenting cells (APCs) loaded with different concentrations of the cognate epitope. Additionally, the specificity of TCR transgene-expressing T cells is determined by identifying the recognition motif of the TCR, i.e., those amino acids and their positions in the cognate epitope that are critical for recognition by this particular TCR. The more stringent the TCR recognition motif (i.e., the more amino acid residues critically contribute to the epitope's recognition), the lesser the chance that the TCR is cross-reactive. Finally, tumor cell recognition assays can be performed to test if the TCR can recognize epitopes that are the product of endogenous antigen processing and presentation by tumor cells. Extensive in vitro testing of the TCR using sensitivity and specificity assays is crucial to assess its potential clinical value [5]. Collectively, the below protocols provide a stepwise approach to identify TCRαβ sequences, introduce the TCR into T cells, and characterize the TCR in vitro (*see* Fig. 1).

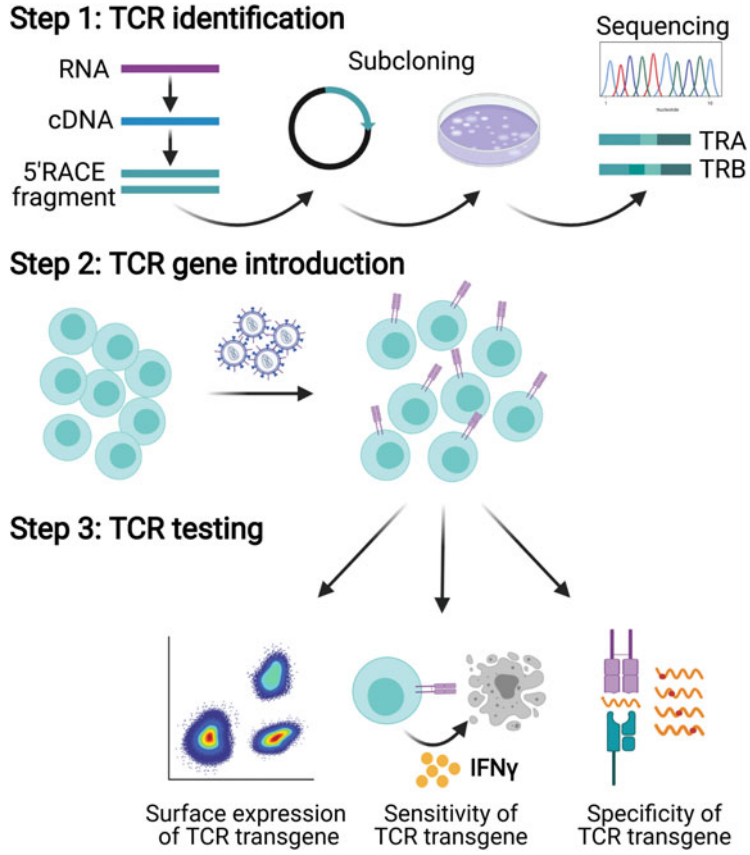


Fig. 1 A stepwise methodological approach to gene-engineered T cells with T-cell receptors for adoptive therapy. The steps are threefold: identification, gene introduction into T cells, and in vitro characterization of TCRαβ sequences (The illustration is created with BioRender.com)

2 Materials

2.1 Identification of TCR from RNA Isolated from pMHC-Positive T Cells

1. SMARTer RACE cDNA Amplification kit (*see* **Notes 1** and **2**).
2. Plasmid isolation kit (*see* **Note 3**).
3. GSP1 primers: prepare a 10 μM stock in sterile dH_2O (*see* **Note 4**). Store at −20 °C.

   GSP1α: GATTACGCCAAGCTTGTTTTGTCTGTGATATACACA.

   GSP1β: GATTACGCCAAGCTTTGCACCTCCTTCCCATTCACCC-ACCAGCTCAGCTC.
4. Nested primers: prepare a 10 μM stock in sterile dH_2O. Store at −20 °C.

NP1α: GATTACGCCAAGCTTGTGACACATTTGTTTGAGAAT.

NP1β: GATTACGCCAAGCTTGGCTCAAACACAGCGACCTC.

5. M13 primers: prepare a 10 μM stock in sterile dH_2O. Store at −20 °C.

   Forward M13: GTAAAACGACGGCCAGT.

   Reverse M13: CAGGAAACAGCTATGAC.

6. 2× Q5 Master Mix (*see* **Note 5**).
7. DreamTaq DNA polymerase and DreamTaq buffer.
8. Deoxynucleotide triphosphates (dNTP) mix (10 mM).
9. 1% agarose gel in modified Tris Acetate-EDTA (TAE) buffer. Dilute TAE buffer in dH_2O to a 1× concentrated solution. Add 1% agarose; heat up the solution in the microwave to dissolve the agarose; and after cooling down to 50–60 °C, pour the agarose in a gel casting tray, with the appropriate well comb in place.
10. 5× DNA loading buffer: dissolve 10 g sucrose in 20 mL H_2O. Add 50 mg Orange G to the solution. Add H_2O up to 50 mL and store at 4 °C.
11. GelRed can be used as a fluorescent nucleic acid dye for visualization of the DNA in the gel (*see* **Note 6**). Add 2.5 μL GelRed to 500 μL 5× DNA loading buffer. Mix 5 μL GelRed/DNA loading buffer, and mix with 20 μL sample before adding onto the gel.
12. LB medium: add 20 g LB to 1 L of dH_2O in a glass bottle. Autoclave the bottle, and let the medium cool down. Store at RT or 4 °C.
13. LB agar plates: add 15 g agar to 1 L LB medium. Autoclave the bottle and let it cool down to approximately 40 °C. Add 100 μg/mL ampicillin, mix gently by shaking, and pour LB/Agar + Amp in 10 cm petri dishes. Let petri dishes cool down at RT until agar is solid before dishes are stored at 4 °C.
14. PCR tubes.
15. 10 cm petri dishes.
16. 1.5 mL and 2 mL Eppendorf tubes.
17. PCR thermocycler.
18. Heating block.
19. Incubator with rotation at 37 °C.
20. Spectrophotometer for nucleic acid quantification.

2.2 Gene Transfer of TCR into T Cells

1. The adherent cell lines 293T and Phoenix-Amp (*see* **Note 7**) should be cultured twice a week using 0.5–1.0 × 10^6 cells per T75 culture flask in 10 mL $DMEM^{++++}$ medium. Additionally, the Phoenix Amp cells need to undergo a 1-week selection procedure with 300 μg/mL Hygromycin B and 1 μg/mL diphtheria toxin, which is repeated after 15 passages of culturing. Cells are cultured for a maximum of 30 passages.
2. Peripheral blood mononuclear cells (PBMCs) (*see* **Note 8**).
3. $DMEM^{++++}$: DMEM medium supplemented with 10% fetal bovine serum (FBS), nonessential amino acids, 200 mM L-glutamine, and 1% penicillin-streptomycin (PS).
4. $RPMI\ Hepes^{HuS++}$: RPMI medium supplemented with 25 mM Hepes, 6% human serum (*see* **Note 9**), 200 mM L-glutamine, and 1% PS.
5. $RPMI\ Hepes^{FBS++}$: RPMI medium supplemented with 25 mM Hepes, 10% FBS, 200 mM L-glutamine, and 1% PS.
6. PBS.
7. PBS/1% FBS: add 5 mL FBS to 500 mL PBS. Store at 4 °C.
8. PBS/0.1% gelatin: add 25 mL 2% gelatin solution to 500 mL PBS.
9. Trypsin/EDTA.
10. Hygromycin B and diphtheria toxin.
11. Promega Calcium Phosphate Transfection Kit.
12. TCR construct in expression vector (e.g., the pMP71 vector).
13. pHIT60 and pColtGalV helper constructs.
14. Ficoll-Paque plus (density: 1.077 g/mL).
15. 10 μg/mL OKT-3 (anti-CD3 MoAb) in PBS stored at −80 °C.
16. Retronectin: 12 μg/mL in dH_2O stored at −20 °C or −80 °C (*see* **Note 10**).
17. 100 IU/mL IL-2 (during transduction) and 360 IU/mL IL-2 (during culture).
18. Trypan blue (TB) for cell counting.
19. Hemacytometer counter and cover slips.
20. Light microscope.
21. T75 culture flasks.
22. 0.45 μm filter.
23. 10 mL syringes.
24. 50 mL tubes.
25. 50 mL Leucosep tubes.
26. Non-tissue culture (NTC) 24-well plate.
27. Parafilm.

2.3 In Vitro Validation of TCR

1. Peptide-MHC-Dextramer with PE label (pMHC-PE) (ProImmune) (*see* **Note 11**).
2. Flow cytometry antibodies (*see* **Note 12**).
   Anti-CD3 FITC (Clone SK7, BD).
   Anti-CD8 APC (Clone SK1 eBioScience): make 1/80 pre-dilution in PBS.
3. FACS buffer: PBS/1% FBS. Add 5 mL FBS to 500 mL PBS. Store at 4 °C.
4. 1% paraformaldehyde in PBS (PFA).
5. RPMI Hepes^{FBS++}: RPMI medium supplemented with 25 mM Hepes, 10% FBS, 200 mM L-glutamine, and 1% PS.
6. Recombinant interferon gamma (IFNγ): dissolve in PBS to a final concentration of 50 ng/mL.
7. Epitope dissolved according to manufacturer (*see* **Note 13**).
8. Epitopes containing individual alanines as replacements at every single position of the cognate epitope (*see* **Note 14**).
9. T2 or BSM cells (*see* **Note 15**).
10. Cell lines expressing target antigen and human leukocyte antigen (HLA) allele of interest (*see* **Notes 16** and **17**).
11. Human IFNγ ELISA Kit (*see* **Note 18**).
12. 5 mL round bottom polystyrene FACS tubes.
13. 96-well tissue culture treated (TCT) round bottom plates.
14. Flow cytometer.

3 Methods

3.1 Identification of TCR from RNA Isolated from pMHC-Positive T Cells

3.1.1 RACE-Ready cDNA, PCR, Cloning, and TCR Sequencing

1. Isolate RNA from epitope-specific T cells according to manufacturer's protocol, and elute the RNA in 10 μL.
2. Measure RNA concentration with a spectrophotometer (*see* **Note 19**).
3. Prepare buffer mix for 5′RACE-ready cDNA synthesis by pipetting 4 μL 5× first strand buffer with 0.5 μL DTT (100 mM) and 1 μL dNTPs (20 mM) in Eppendorf tube 1.
4. Prepare 5′RACE-ready cDNA reaction in Eppendorf tube 2 by mixing 9 μL RNA with 1 μL 5′CDS primer A and 1 μL sterile dH_2O. Incubate the mix 3 min at 72 °C and 2 min at 42 °C, and spin down briefly at 14,000 × *g*.
5. Add 1 μL SMARTer II A oligonucleotide per reaction to Eppendorf tube 2.
6. Add 0.5 μL RNase inhibitor (40 U/μL) and 2.0 μL SMARTScribe reverse transcriptase (100 U) to Eppendorf tube 1. Add

to Eppendorf tube 2, mix by pipetting up and down, and briefly spin down. Incubate Eppendorf tube 2 for 90 min at 42 °C followed by 10 min at 70 °C.

7. Dilute the reaction: add 10 μL of Tricine-EDTA buffer if you started with <200 ng of total RNA, or add 90 μL if you started with >200 ng RNA (*see* **Note 20**).
8. Perform RACE PCR with GSP1 primers. Per condition, two reactions will be performed to separately identify TCRα and β chains. Prepare a master mix containing 15.5 μL PCR-grade H_2O, 25 μL 2× SeqAmp buffer, and 1 μL SeqAmp DNA polymerase.
9. Mix the master mix with 2.5 μL 5′RACE-ready cDNA, 5 μL 10× UPM, and 1 μL GSP1α or β primer, and perform PCR according to the following settings (after putting tubes in PCR thermocycler):

   Five cycles: 94 °C, 30 s; 72 °C, 1.5 min.

   Five cycles: 94 °C, 30 s; 68 °C, 30 s; 72 °C, 1.5 min.

   20 cycles: 94 °C, 30 s; 65 °C, 30 s; 72 °C, 1.5 min
10. Perform nested PCR on the RACE PCR products from **step 9** of Subheading 3.1.1 (*see* **Note 19**). Mix 1 μL RACE PCR product with 1 μL nested universal primer, 1 μL NP1α- or β-primer, 22 μL dH_2O, and 25 μL 2× Q5 master mix in a PCR tube, and perform nested PCR according to the following settings:

    25 cycles: 94 °C, 30 s; 65 °C, 30 s; 72 °C, 1.5 min
11. Load 20 μL of nested PCR product with a fluorescent nucleic acid dye onto a 1% agarose gel.
12. Cut out bands at the correct size of around 800 bp (*see* Fig. 2).
13. Transfer the cut out bands to an Eppendorf tube and add 200 μL NTI buffer. Let the gel dissolve for 5–10 min at 50 °C, while vortexing every 3 min.
14. Place the NucleoSpin column into an Eppendorf tube, and transfer 700 μL of the dissolved sample onto the column. Spin down the Eppendorf tube for 30 s at 11,000 × *g*, and discard the flow-through.
15. Add 700 μL NT3 buffer to the column, and spin down for 30 s at 11,000 × *g*. Discard the flow-through and centrifuge the tube again at 11,000 × *g* for 1 min. Place the column into a new Eppendorf tube. Elute the DNA in 15 μL NE buffer, incubate for 1 min at RT, and centrifuge at 11,000 × *g* for 1 min.
16. Following elution, In-Fusion cloning can be performed according to manufacturer's protocol: transfer 7 μL of eluted

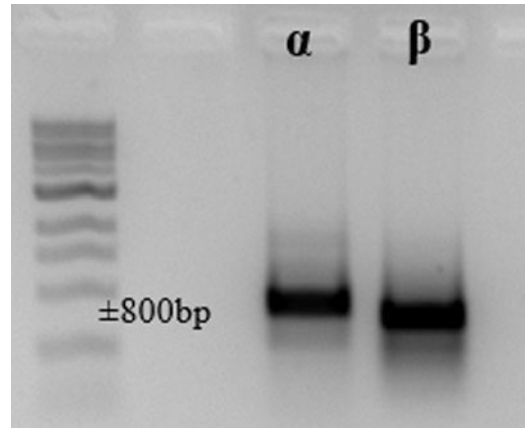


Fig. 2 Correct size of amplified TCRα and β products after nested PCR. Correct size is around 800 bp. Intrinsically the TCRβ chain is larger than the TCRα chain; however, due to the design of RACE primers, the TCRα chain fragment is slightly larger after nested PCR

DNA to an Eppendorf tube. Add 1 μL linearized pRACE vector and 2 μL In-Fusion HD premix to the Eppendorf tube, and mix by vortexing. As a negative control, prepare an empty vector, replacing the eluted DNA with 7 μL dH_2O. The reaction of the positive control provided by the manufacturer consists of 1 μL pUC19 vector, 2 μL 2 kb control insert, 2 μL In-Fusion HD premix, and 5 μL dH_2O. Incubate the reactions for 15 min at 50 °C, and transfer to ice (*see* **Note 21**).

17. Use 2.5 μL cloning reaction from **step 16** of Subheading 3.1.1 for transformation of 25 μL SOC bacteria (*see* **Note 22**). Mix gently by pipetting up and down, after which the reaction should be incubated on ice for 30 min. Perform a heat-shock at 42 °C for 45 s, and put on ice for 1–2 min.
18. Add 225 μL warm super optimal broth (SOC) medium (37 °C) to bacteria, and shake suspension at 200 × *g* for 1 h at 37 °C.
19. Plate out the reaction from **step 18** of Subheading 3.1.1 over 3 LB + Amp plates:
    (a) 1/10: 25 μL of culture +50 μL SOC medium
    (b) 1/100: 2.5 μL of culture +50 μL SOC medium
    (c) Left over.
20. Incubate the plates upside down O/N at 37 °C.
21. Screen for colonies with inserts of expected size. Perform a PCR reaction with M13 primers to amplify DNA inserts of about 20 colonies, and visualize inserts in an agarose gel. The size of the colonies should be around 800–1000 bp for TCRα and β chains. Prepare a PCR premix with following reagents per sample:

(a) 32.5 μL dH_2O

(b) 4 μL 10× DreamTaq buffer

(c) 0.5 μL dNTPs

(d) 1 μL M13 forward primer

(e) 1 μL M13 reverse primer

(f) 0.5 μL DreamTaq DNA polymerase.

22. Add 40 μL premix to each PCR tube (one tube per colony).
23. Pick the colony using a pipette and a small tip, inoculate another LB Agar plate by putting a stripe on the plate with the tip to store the colony, and dip the tip into the PCR mix in the PCR tube. Pipette up and down to mix the DNA with the premix. PCR settings are as followed:

    1 cycle: 95 °C, 5 min

    25 cycles: 95 °C, 30 s; 55 °C, 30 s; 72 °C, 1 min

    1 cycle: 72 °C, 5 min.
24. Load 20 μL sample with a fluorescent nucleic acid dye onto a 1% agarose gel. Select the colonies with the correct size for further analysis.
25. Incubate the LB agar plates at 37 °C overnight.
26. Pick colonies with the correct size the next day with a small pipette tip, and put the tip in a 10–15 mL tube containing 2 mL LB + Amp. Incubate tubes using a 200 × *g* shaker overnight at 37 °C.
27. Transfer bacterial culture to a 2 mL Eppendorf tube, and centrifuge these tubes at 11,000 × *g* for 30 s.
28. Remove supernatant.
29. Continue with plasmid isolation (*see* **Note 3**): add 250 μL buffer A1. Resuspend the pellet by vortexing.
30. Add 250 μL buffer A2. Mix gently by flipping the tube 6–8 times. Incubate for 5 min at RT.
31. Add 300 μL buffer A3. Mix by flipping the tube 6–8 times. Blue color should disappear completely. Centrifuge for 5 min at 11,000 × *g*.
32. Place the NucleoSpin Plasmid Column in a collection tube provided by the kit, and add 750 μL supernatant onto the column. Centrifuge for 1 min at 11,000 × *g*.
33. Discard the flow-through, and place the column back onto the same (now empty) collection tube. Add 500 μL AW buffer, and centrifuge for 1 min at 11,000 × *g*.
34. Discard the flow-through, and place the column back onto the empty collection tube. Add 600 μL buffer A4 and centrifuge for 1 min at 11,000 × *g*.

35. Discard the flow-through, and place the column back onto the empty collection tube. Centrifuge for 2 min at 11,000 × *g* to dry the membrane.
36. Place the NucleoSpin Plasmid Column in a 1.5 mL Eppendorf tube, and add 50 μL buffer AE. Incubate for 1 min at RT and centrifuge for 1 min at 11,000 × *g*.
37. After elution, measure DNA concentration with spectrophotometer.
38. Send samples for Sanger sequencing with M13 primers (*see* **Note 23**).

3.1.2 TCR Sequence Annotation

1. Process TCRα and β chain sequences with alignment software such as Chromas. The software translates the chromatogram file to a sequence.
2. Copy the sequence in plain text or FASTA format.
3. Classify the TCR-V, D, and J genes with the IMGT database and the HighV-QUEST tool (http://www.imgt.org/IMGT_vquest/vquest). Submit the sequence by copy/paste, select *Homo sapiens* in the species section, and select the α (TRA) or β (TRB) sequence in the type of receptor/locus section. The TCR-V, D, and J genes are classified according to the most recent Lefranc nomenclature (*see* **Note 24**).
4. Determine whether the constant region of the β chain is TCRβ constant 1 (Cβ1) or 2 (Cβ2). Align TCR-Cβ of interest with Cβ1 or Cβ2 sequences as reported in https://www.ncbi.nlm.nih.gov/nuccore.
5. Determine the reading frame using the Expasy tool (https://web.expasy.org/translate/). Use Verbose as output format, and determine the in-frame sequence. In the case the sequence has multiple start codons that are in-frame, choose the start codon that is at the exact 5′ end of the leader sequence according to SignalP (http://www.cbs.dtu.dk/services/SignalP/).
6. Design the TCRαβ sequence according to scheme below (*see* **Note 25**).

   *Not*I—GCCACC (Kozak sequence) TCRVβ—Cβ1 or 2 without stop codon—T2A linker—TCRVα—Cα—stop codon—*Eco*RI
7. Order TCRαβ sequences as part of an expression vector (*see* **Note 26**).

3.2 Gene Transfer of TCR into T Cells

3.2.1 Packaging TCR Viruses

1. Precoat T75 flask per condition (*see* **Note 27**) with 5 mL PBS/0.1% gelatin for 10 min at RT.
2. Wash the adherent 293T and Phoenix-Amp cell line with PBS, and loosen the cells with 2 mL trypsin/EDTA at 37 °C.

3. Add 8 mL DMEM^{++++} medium, centrifuge, dissolve in 10 mL fresh DMEM^{++++} medium, and count the cells with TB.
4. Transfer 1.5×10^6 cells of each cell line together in one coated T75 flask in 10 mL DMEM^{++++}.
5. Incubate the cells overnight at 37 °C/5% CO_2.
6. The next day, refresh medium of the packaging cells with 10 mL DMEM^{++++} 3 h prior to transfection.
7. Use the Promega Calcium Phosphate Transfection Kit to transfect the packaging cells. Per T75 flask, prepare in Eppendorf tube 1 the following:
   (a) 10–15 μg TCR construct.
   (b) 5 μg of each helper construct pHIT60 and pColtGalV.
   (c) Add dH_2O to a volume of 500 μL.
   (d) Add 62 μL $CaCl_2$
8. Prepare in Eppendorf tube 2 a 500 μL 2× HBS buffer.
9. Gently vortex the 2× HBS. Slowly add the DNA solution from Eppendorf tube 1 dropwise to the HBS in Eppendorf tube 2 while vortexing. Incubate the mixture at RT for 30 min. Vortex again, and then immediately add the solution to the packaging cells in the T75 flask.
10. Incubate overnight at 37 °C/5% CO_2.
11. The next day, refresh medium of the transfected packaging cells with 10 mL RPMI Hepes^{FBS++}.

3.2.2 Activation of Peripheral Blood Mononuclear Cells (PBMCs)

1. Thaw PBMCs or use freshly collected blood.
2. When using freshly collected blood, isolate PBMCs from healthy donor buffy coats with Leucosep tubes (*see* **Note 8** and steps below).
3. Pipette 15 mL Ficoll Paque in Leucosep tubes.
4. Centrifuge the tube for 10 s at 1000 × *g*.
5. Dilute the buffy coat with PBS/1% FBS in a 1:1 volume ratio.
6. Divide the buffy coat over five Leucosep tubes.
7. Centrifuge the tubes for 10 min at 1000 × *g* with slow deceleration settings.
8. Harvest the cells from the interphase. First, aspirate most of the upper layer (serum), and pour the PBMC (within interphase) in three 50 mL tubes.
9. Add PBS/1%FBS up to 50 mL per tube, and centrifuge for 5 min at 450 × *g*. Aspirate the supernatant, and repeat the washing step three times.
10. Count PBMCs with TB (*see* **Note 28**), and resuspend in RPMI Hepes^{HuS++} at a cell density of 1×10^6/mL.

11. Add 10 ng/mL OKT-3 (*see* **Note 29**).
12. Transfer the PBMCs to a T75 flask, and incubate horizontally for 2 days at 37 °C/5% CO_2.
13. Also coat a non-tissue culture (NTC) 24-well plate with Retronectin. Thaw the Retronectin and add 500 μL to each well; use two wells for each transduction-condition. Seal the plate with parafilm and store overnight at 4 °C.

3.2.3 Transduction of PBMCs

1. Remove Retronectin from the wells (**step 13** of Subheading 3.2.2).
2. Block the wells with 1 mL PBS/2% FBS for 30 min at 37 °C.
3. Harvest virus supernatant from the transfected packaging cells (**step 11** of Subheading 3.2.1), and filter through a 0.45 μM filter using a 10 mL syringe into a 50 mL tube.
4. Add 100 IU/mL IL-2 to the filtered virus supernatant.
5. Add 10 mL fresh RPMI $Hepes^{FBS++}$ to the packaging cells to start a second production round of TCR-encoding virus particles.
6. Aspirate the PBS/2% FBS from the wells of the 24-well plate (**step 2** of Subheading 3.2.3), and add 0.3 mL virus supernatant to each well.
7. Centrifuge for 15 min at 1000 × *g* with slow deceleration settings.
8. Harvest the activated PBMCs (**step 12** of Subheading 3.2.2) by pipetting the cells up and down and using a cell scraper to scrape the cells loose. Transfer the cells to a 50 mL tube.
9. Centrifuge and add RPMI $Hepes^{HuS++}$ to the cells.
10. Count the cells with TB, and use 1×10^6 activated PBMCs per well (*see* **step 11** of Subheading 3.2.3); two wells per condition are used (i.e., use 2×10^6 cells for two wells) (*see* **Note 27**). Transfer the cells to a tube, and centrifuge, aspirate, and resuspend the cells in 0.6 mL virus supernatant.
11. Add 0.3 mL of the PBMC/virus-sup suspension from **step 10** of Subheading 3.2.3 per well (*see* **Note 30**).
12. Seal the plate with parafilm (handle carefully), and centrifuge for 1 h at 1000 × *g* with slow deceleration settings.
13. Remove the parafilm and incubate 5 h (37 °C/5% CO_2).
14. Add 0.8 mL RPMI $Hepes^{HuS++}$ supplemented with 100 IU/ml IL-2.
15. Incubate overnight at 37 °C/5% CO_2.
16. The next day, harvest supernatant from the transfected packaging cells from **step 11** of Subheading 3.2.1, and filter through a 0.45 μM filter using a 10 mL syringe into a 50 mL tube.

17. Add 100 IU/ml IL-2 to the filtered virus supernatant.
18. Carefully remove 1.2 mL from each well, and add 0.6 mL of freshly harvested virus supernatant to each well.
19. Seal the plate with parafilm (handle carefully), and centrifuge for 1 h at 1000 × *g* with slow deceleration settings.
20. Remove the parafilm and incubate 5 h (37 °C/5% CO_2).
21. Harvest the cells, resuspend in RPMI Hepes^{HuS++} supplemented with 360 IU/mL IL-2 at a final concentration of 0.25×10^6/mL. Incubate for 72 h at 37 °C/5% CO_2.
22. TCR- or mock-transduced T cells can be expanded in RPMI Hepes^{HuS++} supplemented with 360 IU/mL IL-2 at a final concentration of 1×10^6/mL for up to 3 weeks (*see* **Note 29**). Culture medium should be refreshed weekly. After these 3 weeks, cells need to be co-cultured with feeder cells (*see* **Note 31**). TCR-engineered T cells can be used for further in vitro validation (*see* **Note 32**).

3.3 In Vitro Validation of TCR

3.3.1 Surface Expression of TCR Transgene

1. Transfer $0.5–1.0 \times 10^6$ TCR- or mock-transduced T cells to FACS tubes.
2. Wash the cells twice with FACS buffer.
3. Aspirate the FACS buffer, and incubate the cells with 5–10 μL pMHC-PE for 10 min at RT in the dark (*see* **Note 11**).
4. Add the following antibodies, and incubate for 20 min at 2–8 °C in the dark.
   (a) 1 μL anti-CD3 FITC
   (b) 2.5 μL 1/80 diluted anti-CD8 APC
   (c) FACS buffer to make a total volume of 10 μL.
5. After incubation, wash the cells twice with FACS buffer.
6. Aspirate and dissolve pellet in 200 μL 1% PFA.
7. Measure surface expression of TCR transgene with flow cytometer (*see* **Note 32**).

3.3.2 Sensitivity of TCR Transgene

1. Harvest TCR- and mock-transduced T cells.
2. Centrifuge and resuspend T cells in RPMI Hepes^{FBS++} at a final concentration of 0.6×10^6 cells/mL.
3. Harvest BSM or T2 cells.
4. Centrifuge and resuspend BSM cells in CTX medium at a final concentration of 0.2×10^6 cells/mL.
5. Transfer 1 mL of BSM cells to each tube.
6. Incubate BSM cells with the cognate epitope for 15 min at 37 °C. Different concentrations per epitope can range from 1 pM to 1 μM (*see* **Note 13**). Also include an irrelevant epitope

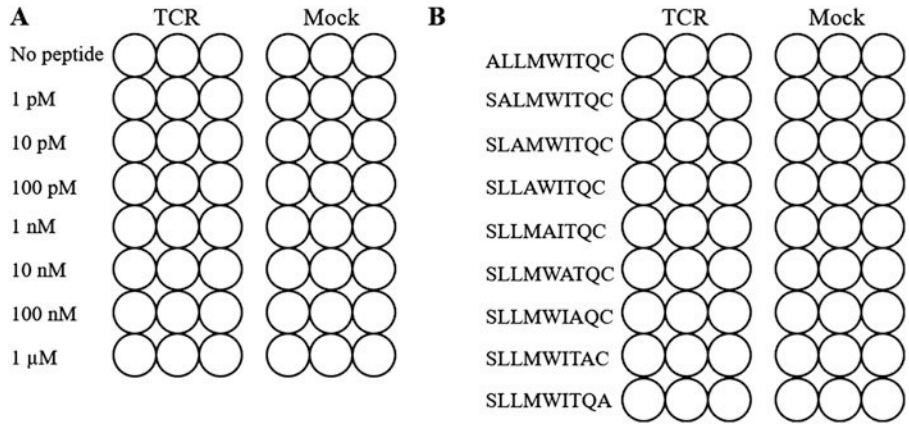


Fig. 3 Pipetting scheme to assess the TCR transgene's sensitivity (**a**) and specificity (**b**) for its cognate epitope. Different concentrations of the cognate epitope (**a**) or different single amino acid mutants of the cognate epitope (illustrated with an example sequence) (**b**) are incubated with BSM or T2 cells prior to co-culture with TCR- or mock-transduced T cells

that should not be recognized by the TCR-transduced T cells as a negative or background control. An example of the layout for the 96-well plate with different epitope concentrations and controls is shown in Fig. 3a.

7. Add 100 μL of T cells from **step 2** of Subheading 3.3.2 to each well (in triplicates) of a 96-well TCT round bottom plate.
8. Add 100 μL of epitope-loaded BSM cells to each well (in triplicates), making a total volume of 200 μL.
9. Centrifuge the plate for 2 min at 220 × *g* with slow deceleration settings.
10. Incubate the plate for 16–24 h at 37 °C/5% CO_2.
11. Centrifuge the plate for 2 min at 220 × *g* with slow deceleration settings.
12. Harvest 150 μL supernatant which can be used to measure IFNγ levels with ELISA-based methods as a readout for sensitivity of TCR transgene-mediated T-cell responses (*see* **Notes 33** and **34**).

3.3.3 Specificity of TCR Transgene

1. Harvest TCR- and mock-transduced T cells.
2. Centrifuge and resuspend T cells in CTX medium at a final concentration of 0.6 × 10^6 cells/mL.
3. Harvest BSM or T2 cells.
4. Centrifuge and resuspend BSM cells in CTX medium at a final concentration of 0.2 × 10^6 cells/mL.
5. Transfer 1 mL of BSM cells to each FACS tube.

6. Incubate BSM cells with different single amino acid mutants of the cognate epitope (10 μM) (*see* **Notes 13** and **14**) for 15 min at 37 °C. Add 1 μL of 10 mM epitope to 1 mL of BSM cells. An example of the layout for the 96-well plate with different epitope mutants is shown in Fig. 3b.
7. Add 100 μL of T cells to each well (in triplicates) of a 96-well TCT round bottom plate.
8. Add 100 μL of epitope-loaded BSM cells to each well (in triplicates), making a total volume of 200 μL.
9. Centrifuge the plate for 2 min at 220 × *g* with slow deceleration settings.
10. Incubate the co-culture for 16–24 h at 37 °C/5% CO_2.
11. Centrifuge the plate for 2 min at 220 × *g* with slow deceleration settings.
12. Harvest supernatant which can be used to measure IFNγ levels with ELISA-based methods as a readout for specificity of TCR transgene-mediated T-cell responses (*see* **Note 35**).

3.3.4 Target Cell Recognition

1. Pre-treat target cells (*see* **Notes 16** and **17**) with 50 pg/mL IFNγ (1:1000 dilution) 48 h prior to co-culture with TCR or mock-transduced T cells.
2. Harvest TCR- and mock-transduced T cells.
3. Centrifuge and resuspend T cells in CTX medium at a final concentration of 0.6 × 10^6 cells/mL.
4. Harvest target cells from **step 1** of Subheading 3.3.4.
5. Centrifuge and resuspend target cells in CTX medium at a final concentration of 0.2 × 10^6 cells/mL.
6. Add 100 μL of T cells to each well (in triplicates) of a 96-well TCT round bottom plate.
7. Add 100 μL of target cells to each well (in triplicates), making a total volume of 200 μL.
8. Centrifuge the plate for 2 min at 220 × *g* with slow deceleration settings.
9. Incubate the plate at for 16–24 h at 37 °C/5% CO_2.
10. Centrifuge the plate for 2 min at 220 × *g* with slow deceleration settings.
11. Harvest supernatant which can be used to measure IFNγ levels with ELISA-based methods as a readout for TCR transgene-mediated recognition of target cells (*see* **Note 36**).

4 Notes

1. The SMARTer RACE cDNA Amplification Kit contains the reagents needed to identify TCRαβ sequences. Components should be stored at different temperatures upon arrival.
2. Other options to identify sequences of the α and β chains of the TCR can be (single cell) RNA sequencing [6].
3. Plasmid isolation can be done with kits from different manufacturers. It is recommended to follow the instructions of the manufacturer.
4. The GSP primers are designed to hybridize within the constant region of the α or β chains of the TCR with a 15 bp PiggyBac vector overhang.
5. 2× Q5 Master Mix can be used as ready-to-use mixture to perform PCR. Also, individual PCR reagents can be used to make a mixture for the PCR reaction.
6. Other fluorescent nucleic acid dyes can also be used to stain DNA or RNA in agarose gels.
7. Use both packaging cell lines for optimal production of virus particles [7, 8].
8. PBMCs can be isolated using Ficoll density gradient centrifugation from healthy donors first described by Böyum in 1968 [9]. They can be freshly used or frozen and thawed on the day of activation of the PBMCs.
9. Human serum used in the culture media is an equal volumetric mixture from five different donors.
10. Retronectin significantly enhances retrovirus-mediated gene transduction into mammalian cells [10, 11].
11. The optimal protocol for pMHC stainings can be different per manufacturer of pMHCs; it is recommended to follow the instructions of the manufacturer.
12. Anti-CD3 and anti-CD8 antibodies with different fluorochromes can be used, as long as spectral overlap between fluorochromes is limited or adequately compensated for.
13. Epitopes (i.e., peptides) should be dissolved in 50–100% DMSO with a final concentration of 10 mM. Dissolvement may very per epitope according to its hydrophobicity profile.
14. To determine the recognition motif of a TCR, an alanine scan is performed. For nine amino acid epitopes, every single position is replaced with an alanine, resulting in nine different epitope mutants.
15. T2 and BSM cells should be cultured twice a week in $RPMI^{+++}$ medium at 0.2×10^6 cells/mL. Depending on the HLA allele

of interest (in these examples, epitopes bound by HLA-A2 are considered), also other cell lines with other HLA alleles can be used.

16. It is important to include positive and negative controls to this assay. Positive controls can include PHA (phytohemagglutinin)/PMA (phorbol myristate acetate), or enterotoxin B (for TCR-independent T-cell stimulation) can be used. Negative controls can include T cells only (without targets), target cells with a different HLA allele, or target cells with the HLA allele under study and no antigen expression. Also blocking antibodies to HLA allele can be used to verify HLA restriction.
17. To test whether the TCR can recognize and kill target cells (i.e., tumor cell lines), it is important to have identified cell or cell lines that express target antigen as well as HLA allele of interest. Expression of target antigen can be determined at RNA (via RT-PCR) and protein level (via immunocytochemistry or Western blot), and expression of HLA allele can be done via flow cytometry.
18. The production of IFNγ by T cells can be used as a readout for TCR-mediated T-cell responses. The enzyme-linked immunosorbent assay (ELISA) is a widely used method to measure levels of IFNγ. The ELISA protocol is generally based on the capturing of the analyte (i.e., IFNγ or other cytokine molecules) by antibodies, which enables the quantification of the analyte present in the sample. Every particular ELISA can have minor adaptations to the standard protocol; therefore, we recommend to follow the instructions of the manufacturer. More detailed information on ELISA-based methods can be found elsewhere [12].
19. Low frequency of antigen-specific T cells in, for instance, PBMC, and difficulties to enrich for these populations can give low RNA quantity and quality. Therefore, a nested PCR on RACE PCR products is recommended for further amplification. If nested PCR products don't show clear bands on agarose gel, the amount of cDNA input can be optimized.
20. cDNA samples can be stored at −80 °C. It is recommended to continue with the RACE and nested PCRs within a week; a decrease in quality can be observed after prolonged storage at −80 °C.
21. The cloning reaction product can be stored at −20 °C. Always take along the positive control in the transformation of bacteria.
22. SOC bacteria are provided as part of the SMARTer RACE cDNA Amplification Kit. Thaw the bacteria on ice and mix by gently pipetting up and down. Store the bacteria at −80 °C and prevent multiple freezing-thawing cycles.

23. Sanger sequencing can be done by several sequencing companies. Generally, DNA and primers have to be provided, but concentrations can differ per company.
24. The IMGT database provides the nucleotide sequences of the TCRα and β chains. It is important to translate this sequence to the amino acid sequence using, for example, the Expasy tool (https://web.expasy.org/translate/), and determine the correct open reading frame.
25. TCRαβ sequences should include the α and β chain sequence connected with a T2A linker [13] and surrounded by restriction sites (*Not*I and *Eco*RI) and a Kozak region prior to the start codon of the variable β-region.
26. TCRαβ sequences can be designed and ordered via an online vector design program. The restriction sites, Kozak region, and 2A linker should be protected, and the rest of the TCRαβ sequences should be codon-optimized for the species *Homo sapiens*.
27. For the transduction of the TCR transgene and further validation of the TCR with in vitro assays, it is important to take mock-transduced T cells along as a negative control. Mock-transduced T cells are created similarly as the TCR-transduced T cells, except that the TCR transgene itself is replaced by an empty vector.
28. Trypan blue (TB) labels death cells with a blue color and is used to facilitate counting of viable cells via a light microscope. Add TB to a cell suspension in a 1:1 ratio (10 μL TB and 10 μL cell suspension, dilution factor is 2) in a 96-well TCT round bottom plate. Mix by pipetting up and down, transfer 10 μL of the mixture to a chamber of a hemacytometer counter, and cover chamber with a slip. Count all cells that are not stained blue in 25 squares, and calculate the concentration of live cells in the cell suspension with the following formula:

$$\text{Concentration/mL} = (\text{counted cells} \times \text{dilution factor} \times 10^3)/(\text{number of squares counted} \times \text{surface area per square in mm}^2 \times \text{depth in mm}).$$

29. Besides OKT-3 and also other antibodies and/or cytokines can be used to activate T cells, such as CD28 mAbs, IL-7, IL-15, and/or IL-21. In fact, the activation of T cells with soluble anti-CD3/CD28 mAbs in the presence of IL-15 and IL-21 resulted in a younger T-cell phenotype and enhanced pMHC binding [14]. The cytokines IL-7, IL-15, and/or IL-21 can also be used to improve T-cell expansion [14–17].
30. TCR gene transfer can also be combined with gene editing using the CRISPR-Cas9 principle [18–21].

31. When necessary to maintain TCR-engineered T cells in culture for longer than 3 weeks, it is required to make use of a feeder cell system as described by Griend et al. This system enables T-cell expansion for up to 2 months while retaining target specificity and cytolytic capacity [22].
32. A TCR transgene is considered to be surface expressed when minimally 5% of $CD3^+$ T cells show an expression of the transgene. Please note that transduction efficiency varies per donor and TCR transgene; therefore, one cannot formally exclude TCR with a surface expression <5% of $CD3^+$ T cells. We therefore recommend to evaluate expression for multiple donors and include other TCR transgenes as controls and use the 5% cutoff as a guideline. In these analyses, mock-transduced T cells can be used as background. Gating strategy for analysis: (1) lymphocytes; (2) $CD3^+$; and (3) $CD8^+$, pMHC multimer$^+$. More detailed information on flow cytometry is well described elsewhere [23]. In the case the TCR genes are only expressed in a low percentage (<10%), cells can be enriched using FACS or magnetic-activated cell sorting (MACS) techniques to increase the percentage of cells expressing the TCR transgene. Again, the need for sorting varies per donor and TCR transgene, for which reason we recommend to use the 10% cutoff as a guideline.
33. Supernatant can be directly used to measure cytokine levels, or the supernatant can be frozen down (−20 °C) and tested at later moments. Avoid repeated freezing-thawing cycles, since this can drastically decrease cytokine levels in supernatants [24].
34. Titrations of cognate epitope and the measurement of IFNγ production by T cells enable the retrieval of half-maximal effective concentrations and thus a measure for functional avidity of the TCR-engineered T cell. A detailed description on how to calculate EC50 can be found elsewhere [25].
35. The recognition motif of a cognate epitope for a particular TCR transgene is the amino acid sequence that is found critical for TCR binding. To determine whether an amino acid is critical for the TCR's recognition, the T-cell response (IFNγ production) to the sequence variant with an alanine replacement is measured. The variant sequence in which a critical amino acid is replaced by an alanine will induce a reduced T-cell response, i.e., at least a twofold drop in IFNγ levels compared to the levels induced by the cognate peptide. The recognition motif, i.e., the sequence of critical amino acids, can be used to search for self-peptides that contain the particular recognition motif (and constitute a potential source for cross-reactive T-cell responses), thereby assessing the risk of off-target toxicity [5].

36. To determine whether a T-cell response against target cell lines is meaningful, IFNγ levels between TCR T cells and mock T cells should be compared. Significant differences in IFNγ levels can be tested with the Mann-Whitney test.

References

1. Debets R, Donnadieu E, Chouaib S, Coukos G (2016) TCR-engineered T cells to treat tumors: seeing but not touching? Semin Immunol 28:10–21. https://doi.org/10.1016/J.SMIM.2016.03.002
2. Lim WA, June CH (2017) The principles of engineering immune cells to treat cancer. Cell 168:724–740
3. Wölfl M, Greenberg PD (2014) Antigen-specific activation and cytokine-facilitated expansion of naive, human CD8+ T cells. Nat Protoc 9:950–966. https://doi.org/10.1038/nprot.2014.064
4. Theaker SM, Rius C, Greenshields-Watson A et al (2016) T-cell libraries allow simple parallel generation of multiple peptide-specific human T-cell clones. J Immunol Methods 430:43–50. https://doi.org/10.1016/j.jim.2016.01.014
5. Kunert A, Obenaus M, Lamers CHJ et al (2017) T-cell receptors for clinical therapy: in vitro assessment of toxicity risk. Clin Cancer Res 23:6012–6020
6. Redmond D, Poran A, Elemento O (2016) Single-cell TCRseq: paired recovery of entire T-cell alpha and beta chain transcripts in T-cell receptors from single-cell RNAseq. Genome Med 8:1. https://doi.org/10.1186/s13073-016-0335-7
7. Lamers CHJ, Willemsen RA, Van Elzakker P et al (2006) Phoenix-ampho outperforms PG13 as retroviral packaging cells to transduce human T cells with tumor-specific receptors: implications for clinical immunogene therapy of cancer. Cancer Gene Ther 13:503–509. https://doi.org/10.1038/sj.cgt.7700916
8. Straetemans T, van Brakel M, van Steenbergen S et al (2012) TCR gene transfer: MAGE-C2/HLA-A2 and MAGE-A3/HLA-DP4 epitopes as Melanoma-specific immune targets. Clin Dev Immunol 2012:1–14. https://doi.org/10.1155/2012/586314
9. Böyum A (1968) Isolation of mononuclear cells and granulocytes from human blood. Isolation of monuclear cells by one centrifugation, and of granulocytes by combining centrifugation and sedimentation at 1 g. Scand J Clin Lab Invest Suppl 97:77–89
10. Hanenberg H, Xiao XL, Dilloo D et al (1996) Colocalization of retrovirus and target cells on specific fibronectin fragments increases genetic transduction of mammalian cells. Nat Med 2:876–882
11. Hanenberg H, Hashino K, Konishi H et al (1997) Optimization of fibronectin-assisted retroviral gene transfer into human CD34+ hematopoietic cells. Hum Gene Ther 8:2193–2206. https://doi.org/10.1089/hum.1997.8.18-2193
12. Hnasko R (2015) ELISA: methods and protocols. Springer, New York
13. Kim JH, Lee SR, Li LH et al (2011) High cleavage efficiency of a 2A peptide derived from porcine teschovirus-1 in human cell lines, zebrafish and mice. PLoS One 6:e18556. https://doi.org/10.1371/journal.pone.0018556
14. Lamers CHJ, van Steenbergen-Langeveld S, van Brakel M et al (2014) T cell receptor-engineered T cells to treat solid tumors: T cell processing toward optimal T cell fitness. Hum Gene Ther Methods 25:345–357. https://doi.org/10.1089/hgtb.2014.051
15. Cieri N, Camisa B, Cocchiarella F et al (2013) IL-7 and IL-15 instruct the generation of human memory stem T cells from naive precursors. Blood 121:573–584. https://doi.org/10.1182/blood-2012-05-431718
16. Grabstein KH, Eisenman J, Shanebeck K et al (1994) Cloning of a T cell growth factor that interacts with the β chain of the interleukin-2 receptor. Science 264:965–968. https://doi.org/10.1126/science.8178155
17. Li Y, Bleakley M, Yee C (2005) IL-21 influences the frequency, phenotype, and affinity of the antigen-specific CD8 T cell response. J Immunol 175:2261–2269. https://doi.org/10.4049/jimmunol.175.4.2261
18. Legut M, Dolton G, Mian AA et al (2018) CRISPR-mediated TCR replacement generates superior anticancer transgenic T cells. Blood 131:311–322. https://doi.org/10.1182/blood-2017-05-787598
19. Schober K, Müller TR, Gökmen F et al (2019) Orthotopic replacement of T-cell receptor α- and β-chains with preservation of near-physiological T-cell function. Nat Biomed Eng 3(12):974–984. https://doi.org/10.1038/s41551-019-0409-0

20. Roth TL, Puig-Saus C, Yu R et al (2018) Reprogramming human T cell function and specificity with non-viral genome targeting. Nature 559:405–409. https://doi.org/10.1038/s41586-018-0326-5
21. Eyquem J, Mansilla-Soto J, Giavridis T et al (2017) Targeting a CAR to the TRAC locus with CRISPR/Cas9 enhances tumour rejection. Nature 543:113–117. https://doi.org/10.1038/nature21405
22. van de Griend RJ, Bolhuis RLH (1984) Rapid expansion of allospecific cytotoxic T cell clones using nonspecific feeder cell lines without further addition of exogenous IL2. Transplantation 38:401–406. https://doi.org/10.1097/00007890-198410000-00017
23. Hawley TS, Hawley RG (2011) Flow cytometry protocols. Humana Press, Totowa, NJ
24. De Jager W, Bourcier K, Rijkers GT et al (2009) Prerequisites for cytokine measurements in clinical trials with multiplex immunoassays. BMC Immunol 10:52. https://doi.org/10.1186/1471-2172-10-52
25. Campillo-Davo D, Flumens D, Lion E (2020) The quest for the best: how TCR affinity, avidity, and functional avidity affect TCR-engineered T-cell antitumor responses. Cell 9:1720

Chapter 14

Combined Analysis of Transcriptome and T-Cell Receptor Alpha and Beta (TRA/TRB) Repertoire in Paucicellular Samples at the Single-Cell Level

Nicolle H. R. Litjens, Anton W. Langerak, Zakia Azmani, Xander den Dekker, Michiel G. H. Betjes, Rutger W. W. Brouwer, and Wilfred F. J. van IJcken

Abstract

With the advent of next-generation sequencing (NGS) methodologies, the total repertoires of B and T cells can be disclosed in much more detail than ever before. Even though many of these strategies do provide in-depth and high-resolution information of the immunoglobulin (IG) and/or T-cell receptor (TR) repertoire, one clear disadvantage is that the IG/TR profiles cannot be connected to individual cells. Single-cell technologies do allow to study the IG/TR repertoire at the individual cell level. This is especially relevant in cell samples in which much heterogeneity of the cell population is expected. By combining the IG/TR repertoire with transcriptome data, the reactivity of the B or T cell can be associated with activation or maturation stages. An additional advantage of such single-cell technologies is that the combination of both IG and both TR chains can be studied on a per cell basis, which better reflects the antigen receptor reactivity of cells. Here we present the ICELL8 single-cell method for the parallel analysis of the TR repertoire and transcriptome, which is especially useful in samples that contain relatively few cells.

Key words T-cell receptor alpha, T-cell receptor beta, Repertoire, Transcriptome, Single cell, Next-generation sequencing

1 Introduction

T cells recognize antigens via unique T-cell receptor (TCR) molecules. Approximately 95% of T cells express a TCRαβ receptor, consisting of a TCRα and a TCRβ chain, whereas the remaining 5% possess a TCRγδ receptor, consisting of a TCRγ and a TCRδ chain. All four TCR chains are highly diverse in their variable domains. Diversity in these variable domains arises from complex recombination processes involving V, D, and J genes in the TCR chain-encoding loci [1]. In this way the V(D)J recombination

Anton W. Langerak (ed.), *Immunogenetics: Methods and Protocols*, Methods in Molecular Biology, vol. 2453, https://doi.org/10.1007/978-1-0716-2115-8_14,

process generates a huge TR repertoire diversity, which is especially apparent in the V(D)J junction. The V(D)J junction is one of the complementarity-determining regions (i.e., CDR3) of the variable domain, which collectively mediate the specific recognition of antigens. Estimates of the number of possible different TCRαβ receptors amount to 10^{12} molecules [2, 3]. Importantly, whereas antigen-inexperienced or naïve T cells have a broad, unselected TCR repertoire [4], antigen-experienced or memory T cells generally contain more narrow TCR repertoires, mostly consisting of particular antigen-selected specificities.

Historically, TCR repertoire diversity assays have mostly focused on TCRβ (TRB) chain profiling. Varying from DNA- [5] or RNA-based [6] TRB bulk sequencing assays to flow cytometry-based single-cell TCRVβ approaches [7], all suffer from drawbacks. A major disadvantage of bulk sequencing approaches is the large number of cells required, whereas flow cytometry-based TCRVβ assays suffer from the limitation that the 24 different TCRVβ antibodies collectively cover only 70% of the normal human TCRVβ repertoire. Moreover, neither of these approaches allows to evaluate the actual composition of the total TCRαβ receptor, as no information on TCRα (TRA) profiles is obtained. Most importantly perhaps, with any of these approaches, it remains difficult to examine changes in TCRαβ repertoire diversity within a heterogeneous pool of T cells or low-abundant population like antigen-specific T cells without purifying them first and/or acquiring large enough numbers of cells.

Over the last 5 years, single-cell transcriptomics has become a popular approach, as it allows to detect the heterogeneity in gene expression among individual cells and the discovery of small subpopulations [8]. The combination of single-cell transcriptomics with TR transcript sequencing provides gene expression and TCR repertoire information at the single-cell level. Several platforms exist for single-cell-combined TCR repertoire and transcriptomics analysis, including 10× Genomics and more recently the ICELL8 single-cell system [9, 10]. Typically, single-cell transcriptomics requires 5–10 K cells [11–13], but little is known about the possibilities of single-cell-based molecular tools for questioning clinically relevant paucicellular samples [9, 10].

Here we describe a method for the combined evaluation of the transcriptome and TRA/TRB repertoire at the single-cell level in clinical samples with low cell numbers. The method covers all the steps from cell dispensation using the ICELL8 single-cell system, double cDNA preparation at the single-cell level, parallel sequencing of transcript and TRA/TRB sequencing libraries, to data evaluation.

2 Materials

2.1 Sample Preparation

1. 15 mL polypropylene tubes.
2. 5 mL polystyrene tube with cell strainer cap.
3. RPMI-1640 medium without L-glutamine.
4. Penicillin/streptomycin (pen/strep) (10^4 U and 10^4 μg/mL stock).
5. DNase I (Sigma-Aldrich; 10 mg/mL stock).
6. Phosphate buffered Saline (1×PBS) without Ca^{2+} and Mg^{2+} pH 7.4 (Invitrogen), sterilized and degassed for at least 1 h using a vacuum system.
7. Fetal bovine serum (FBS) heat-inactivated (HI, 30 min at 56 °C).
8. PAN T cell isolation kit (Miltenyi Biotec).
9. AutoMacs rinsing solution (Miltenyi Biotec).
10. AutoMacs washing solution (Miltenyi Biotec).
11. Bovine serum albumin (BSA, sterile filtered, Sigma-Aldrich; 15% stock solution in AutoMacs rinsing solution).
12. Miltenyi Biotec Automacs Pro Cell Sorter.
13. BD FACSCANTO II flow cytometer (Becton Dickinson).
14. Brilliant Violet (BV)510-labelled antihuman CD3 (Becton Dickinson).
15. Trypan blue (Sigma-Aldrich; 0.4% 0.2 μM filtered before use).
16. Burker counting chamber.

2.2 Cell Dispensation

1. ICELL8 single-cell system (Takara Bio).
2. ICELL8 Collection Kit (Takara Bio).
3. ICELL8 Loading Kit (Takara Bio).
4. Biometra Advanced Thermocycler (Westburg).
5. HulaMixer Sample Mixer (Thermo Fisher Scientific).
6. 2100 Bioanalyzer instrument (Agilent Technologies).
7. Varioskan 3001 microplate reader (Thermo Fisher Scientific).
8. MSND 384-well plates and seals, 20 packs (Takara Bio).
9. Nuclease-free 0.2 mL PCR tubes.
10. Biotix nuclease-free LoBind 1.5 mL microcentrifuge tubes (VWR).
11. 15 mL polypropylene tubes.
12. Magnetic separator/PCR strip (Takara Bio).
13. ICELL8 Human TCR a/b Profiling Reagent Kit (Takara Bio).

14. ICELL8 Human TCR a/b Profiling/Indexing Primer Set (Takara Bio).
15. ICELL8 TCR chip (Takara Bio).
16. Nextera XT DNA Library Preparation Kit (24 samples; Illumina).
17. Nextera XT Index Kit (24 indexes, 96 samples; Illumina).
18. Ethanol absolute (≥99.8%).
19. AccuGENE molecular biology water (Westburg).
20. Helium (≥99.9% purity).
21. Phosphate buffered saline (1×PBS) without Ca^{2+} and Mg^{2+} pH 7.4 (Invitrogen).
22. ReadyProbes Cell Viability Imaging Kit: blue/red contains Hoechst 33342 and propidium iodide (Thermo Fisher Scientific).
23. NucleoSpin® Gel and PCR Clean-up Kit (Macherey-Nagel).
24. Agencourt AMPure XP (Beckman Coulter).
25. Quant-iT™ dsDNA Assay Kit, high sensitivity (Thermo Fisher Scientific).
26. Agilent High-Sensitivity DNA Kit (Agilent Technologies).

2.3 Sequencing

1. Illumina (MiSeq system: HiSeq 2500, 3000, or 4000 or NextSeq 550, 1000, or 2000 system).
2. Paired end 600-Cycle Sequencing Kit (Illumina).
3. HiSeq Rapid SBS Kit v2 (50 cycles) or equivalent (for sequencers other than HiSeq1500, 2500) (Illumina).
4. PhiX (Illumina).

2.4 Analysis

Linux (virtual) machine with at least 16 GB RAM memory and the following software installed:

1. Python3 (version > = 3.6).
2. pyngs (https://github.com/erasmus-center-for-biomics/pyngs).
3. pysc (https://github.com/erasmus-center-for-biomics/pysc).
4. Snakemake (version > = 5.28.0).
5. BCL2Fastq2 (https://emea.support.illumina.com/downloads/bcl2fastq-conversion-software-v2-20.html).
6. AdapterTrimmer (https://github.com/erasmus-center-for-biomics/AdapterTrimmer).
7. IgBLAST (Ye et al., 2013; Reference [14]).
8. Pear (Zhang et al., 2014; Reference [15]).

9. Biomics TCR workflows (https://github.com/erasmus-center-for-biomics/tcr-workflows).
10. R (version $>= 4.0$).
11. tidyverse.
12. Seurat v3 (Stuart et al., 2019; Reference [16]).
13. wesanderson.
14. scales.
15. R studio (version >1).
16. R analysis scripts (https://github.com/erasmus-center-for-biomics/MiMB-R).

3 Methods

3.1 Sample Preparation

1. Prepare DNase medium by adding 5 mL of p/s and 5 mL DNase stock solution to 500 mL of RPMI-1640.
2. Bring DNase medium and FBS-HI to 37 °C.
3. Prepare AutoMacs running buffer by adding 50 mL of 15% BSA to 1450 mL of AutoMacs rinsing solution, and bring buffer to room temperature (*see* **Note 1**).
4. Start up the AutoMacs Pro Cell Sorter according to manufacturer's instruction.
5. Add 5 mL of DNase medium and 1 mL of FBS-HI per vial (1.8 mL) of 10–20 million PBMC to a 15 mL polypropylene tube (use maximal two vials of PBMC per 15 mL tube).
6. Take the required number of vials of peripheral blood mononuclear cells (PBMC) from liquid nitrogen storage or −150 °C freezer (*see* **Note 2**).
7. Thaw PBMC at 37 °C until only a small clump of ice remains.
8. Add the PBMC suspension dropwise to the 15 mL tube containing 5 mL of DNase medium and 1 mL of FBS-HI.
9. Centrifuge at 900 × *g* for 10 min.
10. Discard supernatant, resuspend pellet in DNase medium, add 5 mL of DNase medium.
11. Centrifuge at 900 × *g* for 10 min.
12. Discard supernatant and resuspend pellet in 2 mL AutoMacs running buffer.
13. Take 20 μL of this cell suspension, mix it with 20 μL trypan blue solution, and add 20 μL of this mixture to a Burker counting chamber.
14. Count the number of cells and evaluate their viability (*see* **Note 3**).

15. Fill up the tube using AutoMacs running buffer, and centrifuge at 900 × *g* for 10 min.
16. Discard supernatant, resuspend cell pellet, and label PBMC to enrich for T cells in an untouched manner using the PAN T-cell isolation kit according to manufacturer's instruction (*see* **Note 4**).
17. Following labeling, add AutoMacs running buffer to fill up the tube to 15 mL.
18. Centrifuge cells for 10 min at 900 × *g*.
19. Discard supernatant, and resuspend cells in AutoMacs running buffer to a concentration of 20–25 million of cells/mL (*see* **Note 5**).
20. Filter cells in order to have a single-cell suspension using 5 mL round bottom polystyrene tube with cell strainer snap cap (*see* **Note 6**).
21. Enrich using the DEPLETES protocol according to manufacturer's instruction.
22. Collect enriched sample from Automacs Pro Cell Sorter, and fill up the tube to 15 mL by adding sterile PBS.
23. Take 20 μL to determine cell number within the enriched fraction similar to **step 13**.
24. Take another 20 μL to evaluate purity using flow cytometry (*see* **Note 7**).
25. Discard supernatant, and resuspend in degassed PBS to a concentration of 2–5 × 10^4/mL with a minimum of 270–350 μL/sample.

3.2 Cell Dispensation

3.2.1 Dispense Instrument Pre-checks

1. Before operating the ICELL8 instrument, check the system according to manufacturer's instructions. Check the water level in the pressure reservoir, humidifier, and wash bottle.
2. Check the helium tank pressure. The regulator should be set to a supply input of >435 psi (30 bar). Replace the tank if the pressure is <435 psi (30 bar) (*see* **Note 8**).
3. The regulator should have a supply output of 20–30 psi (1.3–2.0 bar) (*see* **Note 9**).
4. Before starting the cell dispensation, initialize the MSND, preform a daily warm-up, and start the ICELL8 Imaging System according to manufacturer's instructions (*see* **Note 10**).
5. Pre-freeze the empty chip holder at −80 °C.

3.2.2 Staining of Cell Suspension

1. Mix the cell suspension gently by inverting the tube five times, and transfer the required volume from the center of the tube.
2. Stain the transferred cells with Hoechst 33342 and propidium iodide. Add 40 μL of each dye per mL of washed cells.

3. Incubate and mix the cell stain suspension with the HulaMixer at room temperature for 20 min in the dark.
4. Settings for the HulaMixer: orbital rotation at 15 rpm for 15 s, reciprocal rotation at 45° for 10 s, and the vortexing off (*see* **Note 11**).

3.2.3 Dilution and Dispensation of Cells

1. Thaw the second diluent (100×), ICELL8 fiducial mix (1×), and nuclease-Free water on ice.
2. Prepare 100 μL of 5 pg/50 nL positive control Jurkat total RNA by mixing 1 μL second diluent (100×), 1 μL RNase inhibitor (40 U/μL), 97 μL PBS (1× Ca^{2+} and Mg^{2+} free), and 1 μL control Jurkat RNA (10 ng/μL). Keep the dilution on ice.
3. Prepare 100 μL of the negative control mix by mixing 1 μL second diluent (100×), 1 μL RNase inhibitor (40 U/μL), and 98 μL PBS (1× Ca^{2+} and Mg^{2+} free). Keep the dilution on ice.
4. Prepare 200 μL of a 2 cell/50 nl (4×10^3 cells/mL) cell suspension by mixing 1 μL second diluent (100×), 1 μL RNase inhibitor (40 U/μL), 97 μL PBS (1× Ca^{2+} and Mg^{2+} free), and 1 μL control Jurkat RNA (10 ng/μL) (*see* **Note 12**). Keep the dilution on ice.
5. Prepare the 384-well source plate, and start the run to dispense 50 nL of the prepared suspensions into the nanowells according to manufacturer's instructions.

3.2.4 Imaging of Cells

1. In this section, images of all 5184 nanowells of the ICELL8 TCR chip are acquired. *See* manufacturer's instructions (Chapter C) on https://www.takarabio.com/documents/User Manual/ICELL8 Human TCR ab Profiling User Manual/ICELL8 Human TCR ab Profiling User Manual_072219.pdf (*see* **Note 13**).

3.2.5 Analyzing Nanowells (Blank Chip)

1. In this section, CellSelect Software is used to analyze the images of the ICELL8 TCR chip in order to identify the Poisson value of each sample position. *See* manufacturer's instructions (Chapter D) on https://www.takarabio.com/documents/User Manual/ICELL8 Human TCR ab Profiling User Manual/ICELL8 Human TCR ab Profiling User Manual_072219.pdf.
2. Proceed up to "Specify sample names," then return to the Summary tab, and determine the Poisson value for each sample position (*see* **Note 14**).

3.2.6 Analyzing Nanowells (Printed Chip)

1. In this section, CellSelect Software is used to analyze the images of the ICELL8 TCR chip in order to identify nanowells containing viable single cells that are suitable for further

processing and analysis via RT-PCR. See manufacturer's instructions (Chapter D) on https://www.takarabio.com/documents/User Manual/ICELL8 Human TCR ab Profiling User Manual/ICELL8 Human TCR ab Profiling User Manual_072219.pdf (*see* **Note 15**).

2. When instructed to use the Manual triage function, use this function for the following:

   Exclude some wells that were falsely marked as candidate wells.

   Include a lot of wells that were not included by the software:

   (a) Go to Wells tab.

   (b) Sort the wells as follows:

   - "State" (*see* **Note 16**)
   - "HasDeadCells"
   - "Cells1".

   (c) Click on "Manual triage."

   (d) If the well is selected for dispense by the software but needs manual exclusion, click on "Reject – Next Well."

   (e) If the well is not selected by the by the software but needs manual inclusion, click on "Use – Next Well."

3. Save the file using a different filename (<Chip ID>_analysis1.wcd). *See* **Note 17**.

4. The TCR printed chip contains barcodes in triplicates. Exclude wells that have been selected for dispense, so only one unique barcoded well remains.

   (a) Go to the "Wells" tab and select all wells.

   (b) Copy (Ctrl-c) and paste (Ctrl-v) the date into a spreadsheet.

   (c) In the spreadsheet, sort by "For dispense," and delete rows that contain "For dispense- FALSE."

   (d) Highlight wells that have duplicate barcodes by using the conditional format function.

   (e) Select the wells that need to be excluded.

   (f) Switch to the CellSelect software and go to "Wells" tab.

   (g) Manually highlight the wells that needs to be excluded, and exclude these wells.

5. Save the file using a different filename (<Chip ID>_final.wcd).

6. Copy the filter file (<Chip ID>_final.wcd) to the MSND. It is required for dispensing the RT reaction mix.

3.3 First and Second Strand cDNA Synthesis

1. Thaw the required components for the first and second strand cDNA synthesis (except the enzyme) on ice. Vortex and spin down briefly.
2. Thaw the chip (without holder) at room temperature for 10 min. Centrifuge the chip at 3220 × *g* for 3 min at 4 °C.
3. Take the SMARTScribe reverse transcriptase and cDNA amplification polymerase from the freezer just before use. Gently mix, do not vortex, and spin down briefly.
4. Prepare the RT-PCR mix by mixing 56 μL GC Melt (5 M), 24 μL cDNA amplification dNTP mixture (25 mM), 3.2 μL $MgCl_2$, 8.8 μL DTT (100 mM), 30.9 μL SMARTScribe buffer (10×), 33.3 μL cDNA amplification buffer (2×), 5.3 μL Triton X-100 (3%), 3.8 μL Oligo dT Amp primer (100 μM), and 8.8 μL Amp primer (10 μM). Mix well by vortexing until the Triton is dissolved.
5. Add 48 μL SMARTScribe reverse transcriptase (100 U/μL) and 9.6 μL cDNA amplification polymerase (2×) to the RT-PCR mix. Mix by pipetting up and down six times.
6. Prepare the RT-PCR 384-well source plate, and start the run to dispense 50 nL of RT-PCR reaction mix into selected nanowells according to manufacturer's instructions.
7. Run the RT-PCR reaction in a preheated SmartChip cycler with a heated lid. Use the following PCR program: 3 min at 50 °C; 5 min at 4 °C; 90 min at 42 °C; 2 min at 50 °C and 2 min at 42 °C (two cycles); 15 min at 70 °C; 1 min at 95 °C; 10 s at 98 °C, 30 s at 65 °C, and 3 min at 68 °C (24 cycles); and 10 min at 72 °C and 4 °C on hold. The amplified reactions can be stored at 4 °C overnight.
8. Collect the full-length cDNA extraction from the chip to a clean PCR tube with the collection module according to manufacturer's instructions.
9. Measure the volume of the collected cDNA extraction. The measured volume may contain up to 15% less than the expected volume. If more than 15% loss is observed, it gives an indication that cDNA synthesis could have been performed suboptimally. Proceed with steps in Subheadings 3.4 and 3.5 (cDNA cleanup and cDNA QC validation) (*see* **Note 18**).
10. The collected full-length cDNA can be stored at −20 °C.

3.4 Cleanup and Concentration After Full-Length cDNA Extraction

1. Purify and concentrate the amplified full-length cDNA extraction with the Gel and PCR Cleanup Kit according to manufacturer's instructions.
2. Purify the concentrated full-length cDNA eluted from the Gel and PCR Cleanup Kit with the AMPure XP beads.

3. Add 30 μL (0.6×) of AMPure XP beads to the cleaned and concentrated cDNA extraction. Mix by pipetting and spin down briefly.
4. Follow the wash steps of the purification according to manufacturer's instructions.
5. Resuspend the dried beads with 15 μL nuclease-free water. Mix by pipetting and spin down briefly.
6. Incubate the sample at room temperature for 5 min.
7. Place the sample on the magnet and incubate for 2 min or until the solution is clear.
8. Transfer 14 μL supernatant containing purified full-length cDNA to a clean PCR tube. The sample can be stored at −20 °C.

3.5 Validation and Quantification of cDNA

1. Use 1 μL of the purified cDNA product and dilute to 1:3. Then use 1 μL of the diluted cDNA for quantitation using the Quant-iT™ High-Sensitivity dsDNA Assay Kit. Use the manufacturer's instructions on https://assets.thermofisher.com/TFS-Assets/LSG/manuals/Quant_iT_dsDNA_HS_Assay_UG.pdf. *See* **Note 19**.
2. Using the results of the Quant-iT™ High-Sensitivity dsDNA Assay Kit, normalize 1 μL of the purified cDNA product to 2 ng/μL.
3. Use 1 μL of the normalized cDNA (2 ng/μL) to load the Agilent 2100 BioAnalyzer high-sensitivity DNA chip. Use manufacturer's instructions on https://www.agilent.com/cs/library/usermanuals/public/2100_Bioanalyzer_HSDNA_QSG.pdf.
4. Determine cDNA QC: compare the results for your sample with Fig. 1 to verify if the sample is suitable for further processing. Proper cDNA synthesis and purification should yield a broad peak spanning ~400 bp to ~6000 bp (Fig. 1; *see* **Note 20**).

3.6 Preparation of TCR a/b Library by Semi-Nested PCR

1. Thaw the required PCR components for the first PCR reaction (except the enzyme) on ice. Vortex and spin down briefly.
2. Dilute the cDNA to 500 pg/μL. Transfer 1 μL in a 0.2 mL tube. Keep on ice.
3. Prepare the TCRa +TCRb human primer 1 mix by mixing 4 μL of the TCRa human primer 1 and 2 μL of the TCRb human primer 1. Mix well by vortexing and spin down briefly.
4. Take the TCR amplification polymerase from the freezer just before use. Gently mix, do not vortex, and spin down briefly.

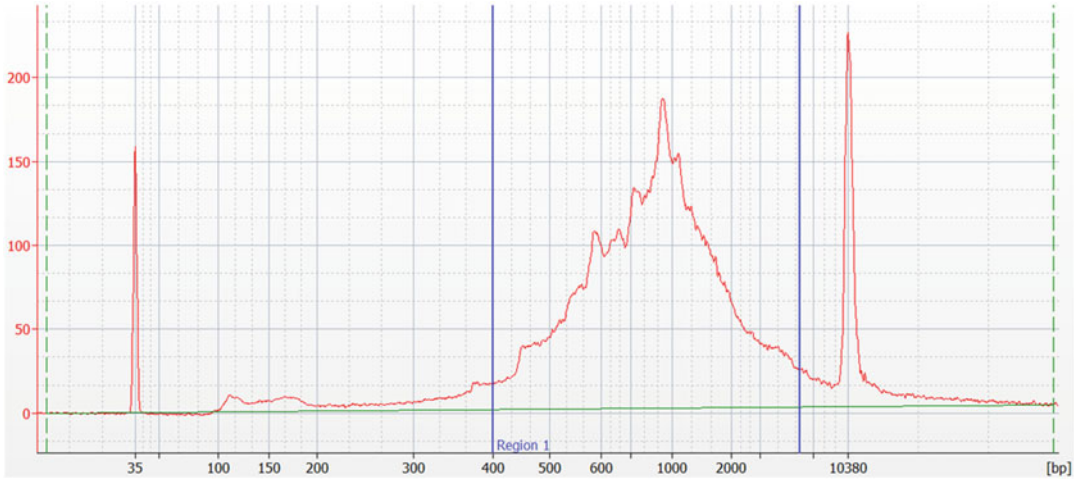


Fig. 1 Typical BioAnalyzer output of full-length cDNA, showing a broad peak spanning ~400 bp to ~6000 bp

5. Prepare 49 µL PCR 1 mastermix in a 0.2 mL tube by combining 10 µL TCR amplification buffer (5×), 4 µL TCR amplification dNTP mixture (2.5 mM each), 1.25 µL primer P5 (5 µM), 3 µL TCRa + TCRb human primer 1 premix, 1 µL TCR amplification polymerase, and 29.75 µL nuclease-free water (*see* **Note 21**).
6. Mix by gently vortexing, and spin down briefly.
7. Add 1 µL of cDNA (500 pg/µL) to the PCR 1 mastermix. Mix by pipetting and spin down briefly.
8. Incubate the reaction in a pre-heated thermal cycler with a heated lid. Use the following PCR program: 1 min at 95 °C and 10 s at 98 °C, 15 s at 60 °C, and 45 s at 68 °C (16 cycles) and 4 °C on hold. The tubes may be stored at 4 °C overnight.
9. Thaw the required PCR components for the second PCR reaction (except the enzyme) on ice. Vortex and spin down briefly.
10. Prepare the ICELL8 TCRa +TCRb human primer 2 mix by mixing 4 µL of the TCRa Human Primer 2 Forward HT Index and 2 µL of the TCRb Human Primer 2 Forward HT Index. Mix well by vortexing and spin down briefly.
11. Take the TCR amplification polymerase from the freezer just before use. Gently mix, do not vortex, and spin down briefly.
12. Prepare 46 µL PCR 2 mix in a 0.2 mL tube by combining 10 µL TCR amplification buffer (5×), 4 µL TCR amplification dNTP mixture (2.5 mM each), 1.25 µL primer P5 (5 µM), 1 µL TCR amplification polymerase, and 29.75 µL nuclease-free water. Mix by gently vortexing and spin down briefly.

13. Add 1 μL of the amplified product from the first PCR reaction and 3 μL of the TCRa + TCRb Human Primer 2 Reverse HT Index Mix (*see* **Note 22**).
14. Mix by pipetting and spin down briefly.
15. Incubate the reaction in a pre-heated thermal cycler with a heated lid. Use the following PCR program: 1 min at 95 °C and 10 s at 98 °C, 15 s at 60 °C, and 45 s at 68 °C (14 cycles) and 4 °C on hold. The amplified reactions may be stored at 4 °C overnight.

3.7 Purification of TCR a/b Library

1. Purify and size-select the TCR library with the AMPure XP beads.
2. Add 22.5 μL (0.45×) of AMPure XP beads to the TCR library to remove fragments larger than ~900 bp. Mix by pipetting and spin down briefly.
3. Incubate the sample at room temperature for 5 min to let the fragments ~900 bp bind to the beads.
4. Place the sample on the magnet, and incubate for 2 min or until the solution is clear.
5. Transfer the supernatant to a clean PCR tube, and add 10 μL of AMPure XP beads. Mix by pipetting and spin down briefly.
6. Incubate the sample at room temperature for 5 min to let the fragments between ~400 and 900 bp bind to the beads.
7. Follow the wash steps of the purification according to manufacturer's instructions.
8. Resuspend the dried beads in 17.5 μL nuclease-free water. Mix by pipetting and spin down briefly.
9. Incubate the sample at room temperature for 5 min.
10. Place the sample on the magnet, and incubate for 2 min or till the solution is clear.
11. Transfer the supernatant containing purified and size-selected TCR library to a clean PCR tube.
12. The purified TCR library can be stored at −20 °C or keep at 4 °C, and proceed directly to Subheading 3.8 (validation and quantification of TCR a/b library).

3.8 Validation and Quantification of TCR a/b Library

1. Use 1 μL of the purified TCR library and dilute to 1:20. Then use 1 μL for quantitation using the Agilent 2100 BioAnalyzer high-sensitivity DNA chip. Use manufacturer's instructions on https://www.agilent.com/cs/library/usermanuals/public/2100_Bioanalyzer_HSDNA_QSG.pdf (*see* **Note 23**). Determine TCR library QC: compare the results for your samples with Fig. 2 to verify if the sample is suitable for further processing. A proper purified TCR library should yield a broad

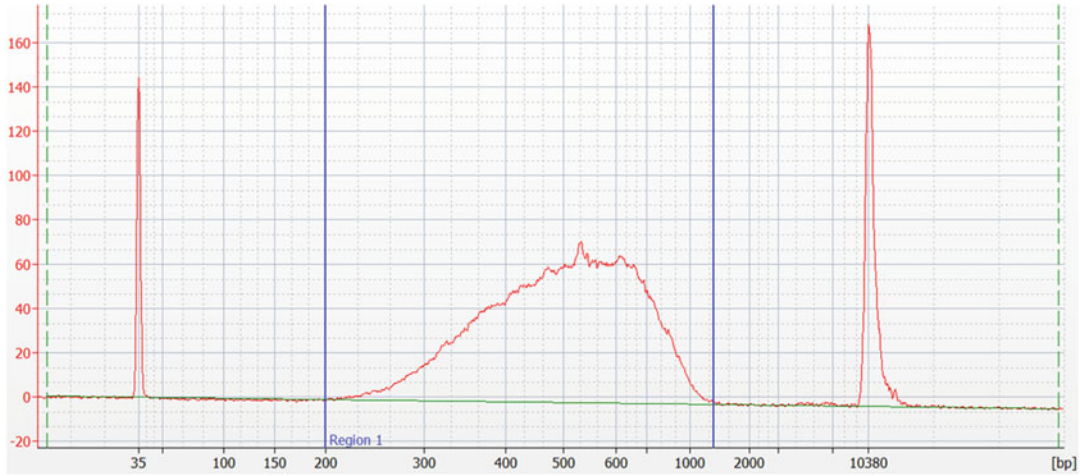


Fig. 2 Typical BioAnalyzer output of purified TCR library, showing a broad peak (550–1200 bp) with a maximum between ~700 bp and ~ 900 bp

peak spanning 550–1200 bp, with a maximum between ~700 bp and ~ 900 bp (Fig. 2).

2. Determine TCR library molarity: set the region table to measure between 550 and 1200 bp, and obtain the molarity in pmol/L.
3. Store the TCR library at −20 °C until sequencing (*see* **Note 24**).

3.9 Preparation of 5′ differential expression (5′ DE) Library

1. Prepare the tagmentation reaction in a 0.2 mL tube (*see* Table 1); keep on ice.
2. Incubate the reaction in a thermal cycler, using the program in Table 2.
3. Add 5 μL of neutralize tagment (NT) buffer to each well.
4. Pipette up and down five times to mix.
5. Incubate for 5 min at room temperature.
6. Prepare the Nextera XT PCR reaction in a 0.2 mL tube (Table 3), and vortex to mix (*see* **Note 25**).
7. Briefly centrifuge the 0.2 mL tube, and incubate in a thermal cycler, using the program in Table 4.

3.10 Purification of 5′ DE Library

1. Purify the amplified 5′ DE library with the AMPure XP beads (use triple purification).
2. First purification: add 1.0× volume (50 μL) of AMPure XP beads to the previous PCR product (~50 μL). Mix by pipetting and spin down briefly.
3. Incubate the sample at room temperature for 5 min.

Table 1
Tagmentation reaction

Component	Volume (μL)
Nextera XT, tagment DNA (TD) buffer	10.0
Sample, purified cDNA (0.2 ng/μL)	5.0
Nextera XT, amplicon tagmentation Mix (ATM)	5.0
Total	*20.0*

Table 2
Thermal cycler program of tagmentation reaction

Step	Number of cycles	Temperature	Time (mm:ss)
1	1	55 °C	5:00
2	1	10 °C	On hold

Table 3
Nextera XT PCR reaction

Component	Volume (μL)
Nextera XT, PCR mastermix (NPM)	15.0
Nextera XT, I7 index primer (orange cap)	5.0
Primer P5 (5 μM) (TCR Kit)	5.0
Neutralized sample	25.0
Total	*50.0*

Table 4
Thermal cycler program for Nextera XT PCR reaction

Step	Number of cycles	Temperature	Time (mm:ss)
1	1	72 °C	03:00
2	1	95 °C	00:30
3	12	95 °C	00:10
4		55 °C	00:30
5		72 °C	00:30
6	1	72 °C	05:00
7	1	10 °C	On hold

4. Place the sample on the magnet, and incubate for 2 min or until the solution is clear.
5. Follow the wash steps of the purification according to manufacturer's instructions.
6. Resuspend the beads with 51 μL nuclease-free water, and transfer eluate (~50 μL) to a clean tube.
7. Second purification: add 0.5× volume (25 μL) of AMPure XP beads to the previous PCR eluate (~50 μL) (*see* **Note 26**). Mix by pipetting and spin down briefly.
8. Incubate the sample at room temperature for 5 min.
9. Place the sample on the magnet, and incubate for 2 min or until the solution is clear.
10. Transfer eluate (~75 μL) to a clean tube.
11. Third purification: add 0.2× volume (15 μL) of AMPure XP beads to the previous PCR product (~75 μL). Mix by pipetting and spin down briefly.
12. Incubate the sample at room temperature for 5 min.
13. Place the sample on the magnet, and incubate for 2 min or until the solution is clear.
14. Follow the wash steps of the purification according to manufacturer's instructions.
15. Resuspend the beads with 13 μL nuclease-free water.
16. The purified 5′ DE library can be stored at −20 °C or keep at 4 °C, and proceed directly to Subheading 3.11 (validation and quantification of 5′ DE library).

3.11 Validation and Quantification of 5′ DE Library

1. Use 1 μL of the purified 5′ DE library and dilute to 1:10. Then use 1 μL for quantitation using the Agilent 2100 BioAnalyzer high-sensitivity DNA chip, following the manufacturer's instructions on https://www.agilent.com/cs/library/usermanuals/public/2100_Bioanalyzer_HSDNA_QSG.pdf.
2. *See* **Note 27**.
3. Determine 5′ DE library QC: Compare the results for your sample with Fig. 3 to verify if the sample is suitable for further processing. A proper purified 5′ DE library should yield a broad peak spanning 200–1400 bp.
4. Determine 5′ DE library molarity: set the region table to measure between 200 and 1400 bp, and obtain the molarity in pmol/L.
5. Store the 5′ DE Library at −20 °C until sequencing (*see* **Note 28**).

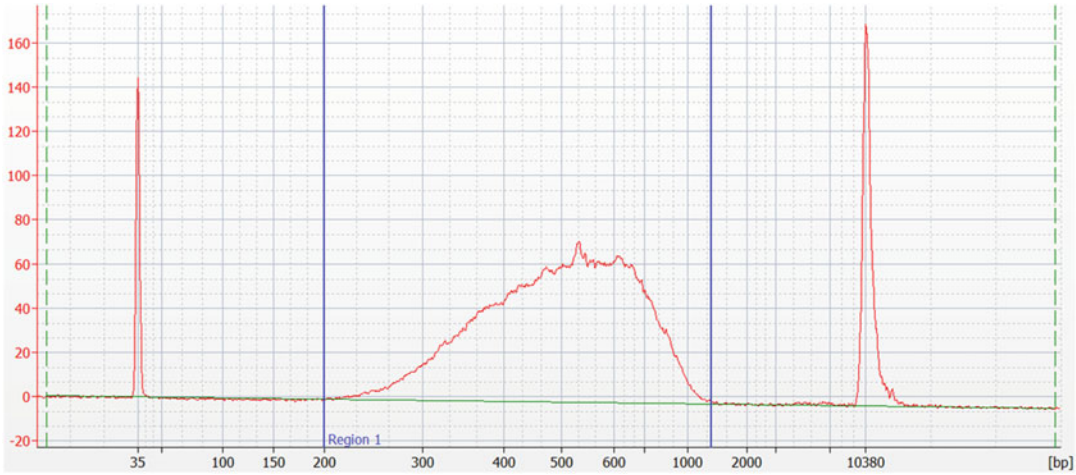


Fig. 3 Typical BioAnalyzer output of purified 5′ DE library, showing a broad peak spanning 200–1400 bp

3.12 Next-Generation Sequencing

1. Thaw the TCR and the 5′ DE sequencing libraries, and prepare them according to the instructions of Illumina (www.illumina.com).
2. In the case you want to sequence multiple samples on one flow cell, check that the libraries have different P7 indices (*see* **Note 29**).
3. Pool the selected samples in equal molarities in a single tube.
4. The TCR library has to be sequenced Paired End 300 (600 cycle kit) on a MiSeq, whereas the 5′ DE library has to be sequenced either single-read 50 or 75 base pairs on a HiSeq or NextSeq sequencer. Sequencing instructions can be found at www.illumina.com.
5. Both single index or dual index sequencing can be applied (*see* **Note 30**).
6. Add the PhIX control sample to both the TCR, and the 5′ DE sequencing runs at 1% of the total output for each flow cell. This enables quantity and quality checks.
7. Proceed with the sequencing procedure as described by the manufacturer. Alternatively, the sequencing procedure can be outsourced to a sequence service provider.
8. Check after sequencing the quality parameters of the PE300 and SR50 sequencing runs. Q30 values should be above 75% for the PE300 reads and above 80% for the single 50 or 75 bp reads.
9. Check after sequencing the yield of the sequencing run. Expected yields vary per sequencer. We advise to obtain at least 5 M clusters for the PE300 run to ensure sufficient data

for subsequent TCR analysis (*see* **Note 31**). For the transcriptome library, a minimal yield of 50 M clusters is advised to ensure sufficient data for transcriptome analysis.

3.13 Data Analysis

In the following sections, code will be typeset in a monospaced font.

3.13.1 Primary Analysis of Single-Cell TCR Data

1. Demultiplex the TCR data using the bcl2fastq2 program from Illumina per the instructions in the BCL2FASTQ2 manual. The I5 indices for these libraries are TCTTTCCC on an Illumina MiSeq sequencer. For other sequencers the reverse complement sequence may be needed. This procedure will result in two files of which the names are constructed as follows: {x}_S{y}_L001_R{z}_001.fastq.gz. In this filename, x denotes the sample ID, y is an arbitrary sample number in the sample sheet, and z is the number of the read (1 for forward, 2 for reverse) (*see* **Note 32**).
2. Make a new working directory in which to perform the TCR assignment, and copy the newly generated FastQ files and the well list obtained during well selection over.
3. Go into the new working directory.
4. Create a directory labeled demultiplexed.
5. Create individual FastQ files per well using pysc (https://github.com/erasmusmc-center-for-biomics/pysc). Both the data start (base 14) and barcode position (read 1 bases 0 to 10) will need to be specified while running this tool as well as a format where to put the output reads. The following command will place the FastQ files per well in the demultiplexed directory:

```
python3/data/Software/python/pysc/bin/pysc demultiplex \
--read_1 TCR_S1_L001_R1_001.fastq.gz \
--read_2 TCR_S1_L001_R2_001.fastq.gz \
--well-list welllist.txt \
 --output-read-1 "demultiplexed/{sample}_{row}_{column}_R1.
fastq" \
 --output-read-2 "demultiplexed/{sample}_{row}_{column}_R2.
fastq" \
--well-barcode-read 1 \
--well-barcode-start 0 \
--well-barcode-end 10 \
--data-start 14
```

6. Run the TCR snakemake workflow (https://github.com/erasmusmc-center-for-biomics/tcr-workflows) which removes sequence adapters introduced during the sample preparation,

merges the forward and reverse reads using PEAR [15], and runs IG-BLAST [14] to align the merged sequences to human V, D, and J genes for each well individually. This workflow yields clonotype reports containing both TRA and TRB structures and their CDR3 sequences per well.

3.13.2 Primary Analysis of RNA-Seq Dat

1. Demultiplex the single-cell 5′ DE RNA-seq data using bcl2fastq2 in a similar manner to the TCR data.
2. Make a new directory in which to perform the scRNA-seq quantification, and copy the well list and scRNA-seq FastQ files over and move into.
3. In the new directory, make a demultiplexed sub-directory.
4. Assign reads to wells using pysc, and place them in a single file per sample. To this end run the following command:

```
python3/data/Software/python/pysc/bin/pysc demultiplex \
--read_1 $file \
--well-list welllist.txt \
--output-read-1 "demultiplexed/{sample}_R1.fastq" \
--well-barcode-start 0 \
--well-barcode-end 10 \
--data-start 14
```

5. Run the expression analysis snakemake workflow (https://github.com/erasmusmc-center-for-biomics/tcr-workflows) which removes the adapters introduced during the sample preparation, aligns the reads with HISAT2 [17], converts mapped regions to BED format, and intersects these with transcripts with bedtools [18]. Finally, the gene expression per gene per well is quantified.

3.13.3 Downstream Analysis in R

1. Make a new directory for the downstream analysis and go into this. It is strongly advised to give this folder a meaningful name to distinguish it from other experiments.
2. Make an R sub-directory and data sub-directory, and in this create a directory named Expression and a directory named TCR.
3. Copy the sample.exon.tsv.gz to the Expression directory and the sample_row_column.report.csv files to the TCR directory.
4. Copy the well list to the data directory as welllist.txt.
5. Start the RStudio (https://www.rstudio.com/), and make a new project from an existing directory. For the directory choose the downstream analysis folder.
6. Copy the files environment.R, seurat_analysis_part1.R, seurat_analysis_part2.R and tcr_analysis.R from https://github.com/erasmusmc-center-for-biomics/MiMB-R to the R directory.

7. Run the environment.R script by typing source("R/environment.R") in the R console. This script will make an R environment with the well list, expression, and TCR data loaded.
8. Run the seurat_analysis_part1.R script by typing source("R/seurat_analysis_part1.R") in the R console. This script will load the data assembled with environment.R and perform normalization, scaling, principal component analysis, and a Jackstraw analysis. It will create a directory named output with figures and tables.
9. Open the pca_variance_explained.png and jackstraw_score.png figures in the output directory (Fig. 4), and determine at which principal component the scores decrease sharply. Up to this component, biologically relevant variation is contained (*see* **Note 33**).
10. Set the ndim variable to the component number previously obtained by issuing the following command ndim <− n, where *n* should be larger than 1.
11. Run the seurat_analysis_part2.R script by typing source("R/seurat_analysis_part2.R") in the R console. This script will run both UMAP and *t*-SNE which will place the cells based on their expression profiles on a two-dimensional plane (*see* **Note 34**). Figures depicting the cell layout are made in the output directory as well as tables with the coordinates per cell obtained for both methods (Fig. 5).
12. Run the TCR analysis script by executing source("R/tcr_analysis.R") in the R console. This script will summarize and filter the alignments to the TCR and yield a receptor to number of cells table (*see* **Note 35**). Two more tables are created in the output directory: one table with the filtered T-cell receptor data (filtered_tcr.txt) and one table with the TCR versus the number of cells (tcr_to_cells.txt) which gives an indication of sample complexity. Furthermore, the TCR complexity is plotted in a figure in the file tcr_to_cells.png (Fig. 6).

3.13.4 Projecting Gene Expression Data on Cell Coordinates

1. Create a variable named gene with identifier of the gene of which the expression will be displayed, for example, gene <− "ENSG00000167286" for CD3d.
2. Make a variable coordinate_type to use either the *t*-SNE (tsne) of UMAP (umap) coordinates.
3. Run the project_expression.R script with source("R/project_expression.R"). A file will be created in the output directory with the gene identifier and the coordinate type. Opening this file will display the *t*-SNE or UMAP projection with the expression depicted over it (Fig. 7). The expression is scaled from 0 to 1 with 0 being the cell(s) with the lowest gene expression and 1 the cell(s) the highest expression.

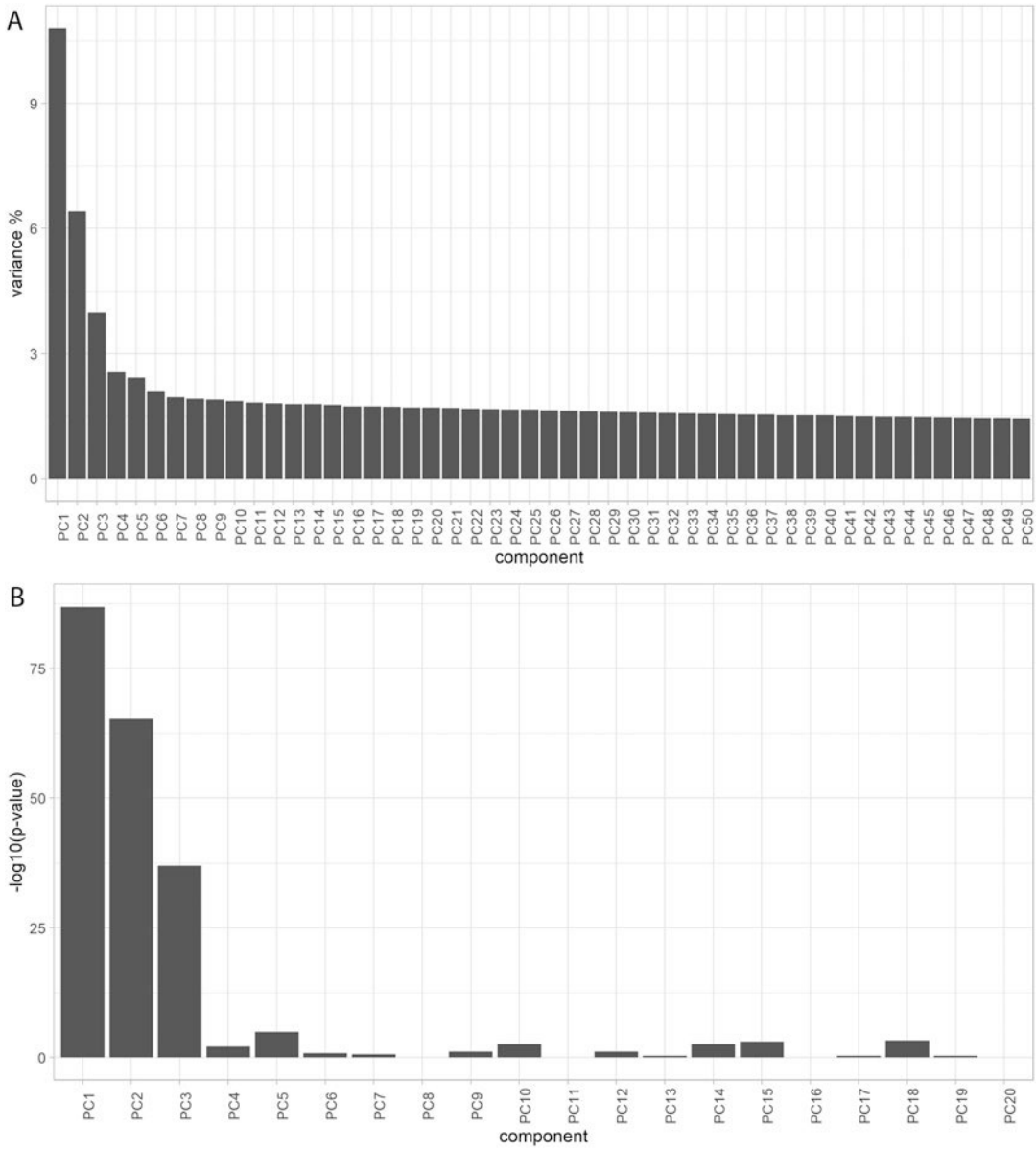


Fig. 4 Results from the PCA (**a**) and Jackstraw (**b**) analyses

3.13.5 Projecting VDJ Usage Data on Cell Coordinates

1. Create a variable named locus with either TRA to display the alpha chains or TRB to display the beta chains.
2. Make a variable coordinate_type to use either the *t*-SNE (tsne) of UMAP (umap) coordinates.
3. Run the project_vdj.R script with source("R/project_vdj.R"). A file will be created in the output directory with the locus and the coordinate type. Opening this file will display the *t*-SNE or UMAP projection with the V(D)J composition projected as the colors (Fig. 8).

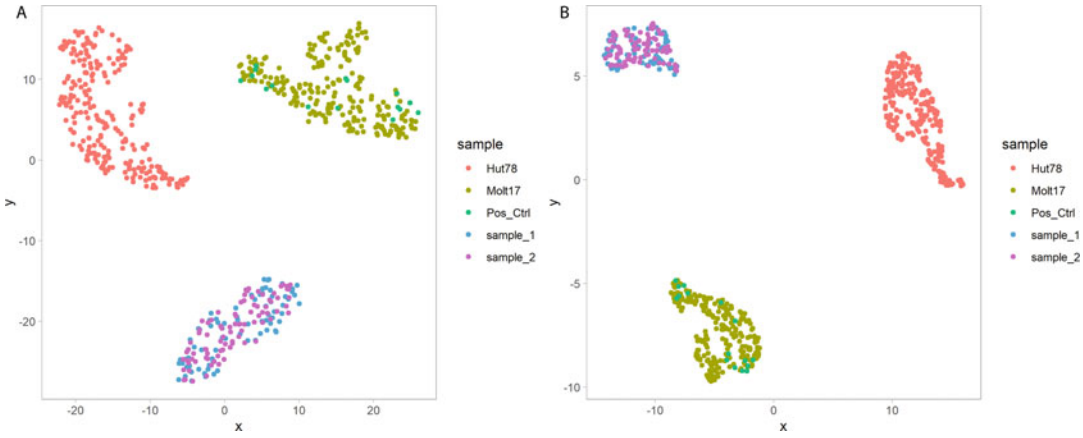


Fig. 5 Projections of single cells on a two-dimensional plane based on their expression profiles using *t*-SNE (**a**) and UMAP (**b**)

4 Notes

1. This buffer needs to be stored at 4 °C and expires 1 month after adding the BSA stock solution (15%).
2. Use freshly isolated PBMC (obtained following Ficoll gradient centrifugation as described before [19] instead of frozen PBMC for the enrichment of T cells. Moreover one can also enrich for activated (antigen-specific) T cells following stimulation with antigens using, for example, CD137, a costimulatory molecule upregulated upon the interaction of a TCR with antigen presented by antigen-presenting cells using flowcytometry-based cell sorting [20].
3. Determine the number of unstained (living) and blue (dead) cells in duplicate by evaluating number of cells in 25 squares. The number of cells multiplied by 2 (dilution factor) and 10^4 return the number of cells/mL. Multiplication by the volume results in total cell number. The number of living cells divided by the total cell count results the fraction of living cells. The number of dead cells divided by the total cell count returns the fraction of dead cells.

   Samples containing a lot of dead cells, for example, 50%, do not perform well in the single-cell RNA sequencing. Typically viability should be more than 80%.
4. Both untouched and positive selection of T cells/cells of interest work for downstream processing/combined analysis of transcript and TCRA/TRB sequencing libraries.

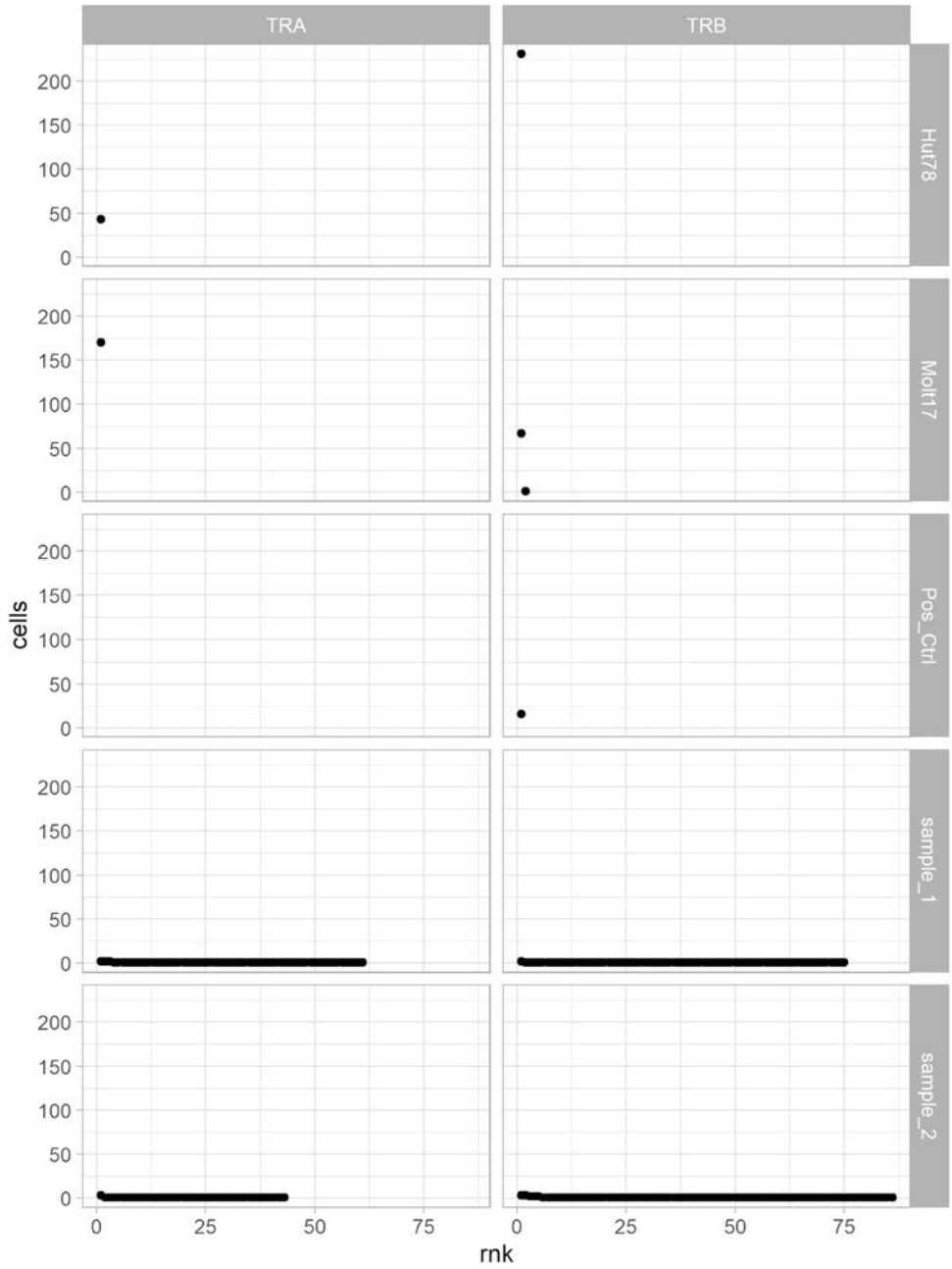


Fig. 6 The number of cells with the same TCR ordered from the most to the least abundant receptor per sample

5. When having more than 25 million of cells, consider to run a second separation tube instead of increasing the volume of the tube. This will result in better purities of enriched fractions. For more than one separation, use a quick rinse in between samples when they are of the same origin, and rinse if you want to separate a sample of a different origin.

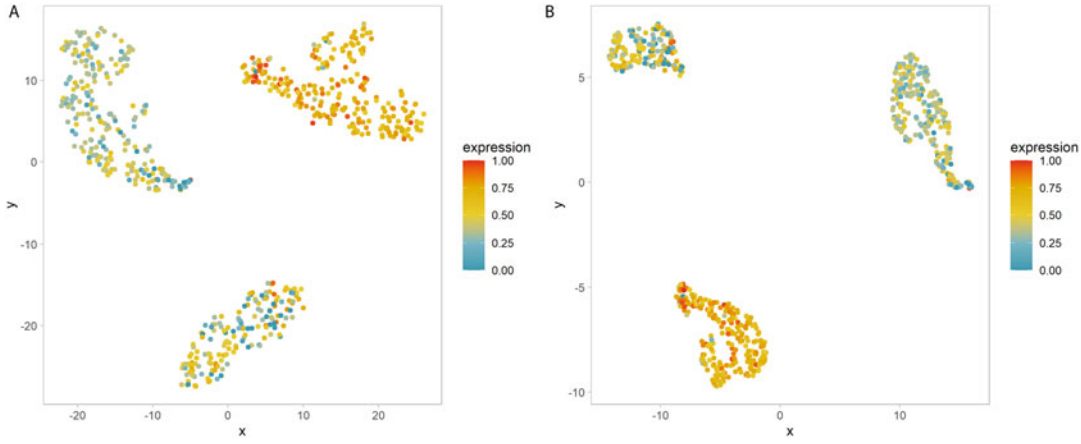


Fig. 7 Expression of CD3d (ENSG00000167286) projected on cells placed with *t*-SNE (**a**) and UMAP (**b**)

6. Only filter if clumps of cells are visible by eye to minimize cell loss due to filtration of cell suspensions. Prewet the filter using AutoMacs running buffer before adding the cell suspension, and rinse the tube afterward. The nylon mesh (pore) size of the cell strainer is 35 μm.
7. Purity evaluation is done by adding 30 μL 1× PBS to 20 μL of cell suspension and staining using 2 μL of BV510-labelled antihuman CD3 for 15 min at room temperature. Following a wash (900 × *g* for 5 min), the supernatant is discarded, pellet resuspended, and the sample measured on the BD FACSCanto II. The proportion of T cells within the live gate represents the purity of the cell sample. Purities over 95% give the best results as most of the cells represent the cells of interest, i.e., T cells.
8. One dispense uses ~73 psi (5 bar). Replace earlier if more experiments are planned.
9. Be aware that 35 psi (2.4 bar) should be the very max. Pressure above 35 psi (2.4 bar) will blow the overpressure valve and cause a Helium leakage.
10. The fluorescence light source requires a warming up period of ~5 min. After switching it on, wait at least 5 min before switching it off.
11. Do not vortex or centrifuge the cells. Only use 200 μL and 1000 μL pipets.
12. The manual advises to make a 1 cell/50 nL solution; however, in practice cell counts are usually overestimated; therefore, we make a 2 cell/50 nL solution. After the blank chip, the solution will be diluted based on the Poisson information. Prepare enough volume for the blank and printed chip; 100 μL is required per source plate well. If there are <8 biological samples for the chip, for example four, prepare at least 300 μL cell

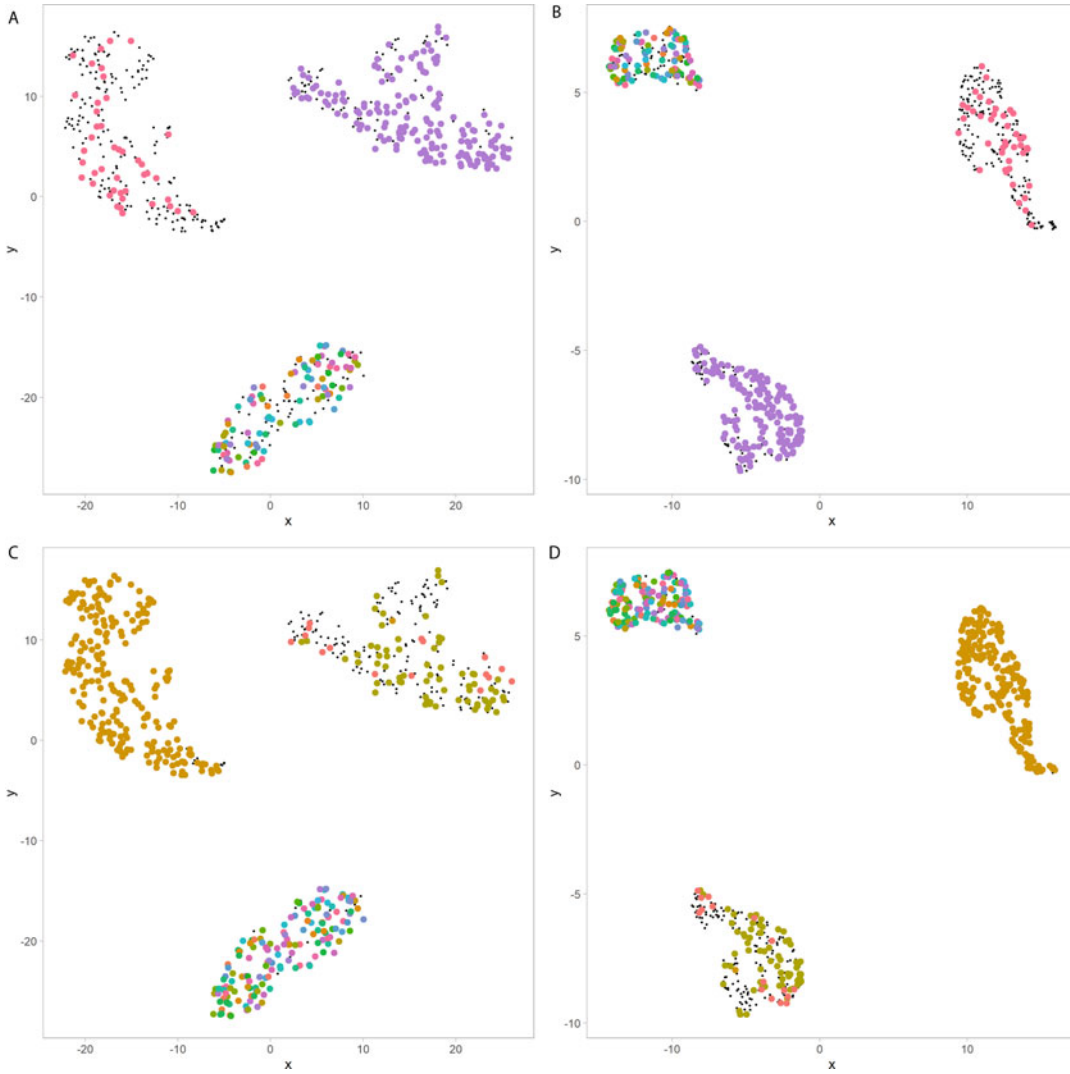


Fig. 8 TCR V(D)J composition projected on cells placed with *t*-SNE (**a**, **c**) and UMAP (**b**, **d**) for TRA (**a**, **b**) and TRB (**c**, **d**)

suspension per sample (100 µL for the blank chip and 200 µL for the barcoded chip). The blank chip can be re-used. Keep the unused wells in the source plate empty. After dispense the wells in the chip that, corresponding to the empty wells from the source plate, will stay clean for the next blank dispense.

13. When prompted for "Run CellSelect with images from: C:\Wafergen\WafergenData For Chip: <Chip ID>?", click NO. After imaging keep the blank chip. It might contain some empty sample positions, or use as balance chip during centrifugation. Store printed chip at −80 °C; this step can also be performed with non-pre-chilled holders.

14. A Poisson value of 0.8–1.0 means that most wells are filled with a single cell. If the Poisson value on the blank chip is within the 0.1–3.0 range, adjust the cell concentration to 0.8 before dispensing on a printed chip. If the Poisson value is outside the 0.1–3.0 range, adjust cell concentration, and perform new blank chip dispense.
15. Before starting the analysis, make a copy of the data folder containing the images:

```
From: C:\Wafergen\WafergenData\ < Chip ID>
To: C:\Users\ICELL8\Desktop\Analyzed Images\ < Chip ID>
```

A maximum of 1728 wells can be selected for further dispense. The downselect function is generally not required.

16. Overview of "States" that might contain wells to include or exclude wells for dispense:
    - Cluster (+LowConfidence): might contain some cells to include
    - Good: might contain some cells to exclude (empty or duplicate)
    - HasDeadCells (+LowConfidence): might contain many cells to include
      (Cells are sometimes incorrectly marked as "dead" due to false detection of well center; dead cells could be included for analysis as well.)
    - Inconclusive: not likely to contain cells to include
    - LowConfidence: might contain some cells to include
    - MultipleCells (+LowConfidence): not likely to contain cells to include
    - NoCells (+LowConfidence): not likely to contain cells to include
    - TooManyCells: not likely to contain cells to include
17. Additional selection (**steps 19–23** in the manufacturer's instructions) is generally not required.
18. An additional volume of 3–10% was observed in our experiments; there are no indications that this affects cDNA quality. Proceed with steps in Subheadings 3.4 and 3.5 (cDNA cleanup and cDNA QC validation).
19. Instead of the Quant-iT™ High-Sensitivity dsDNA Assay Kit, the Denovix instrument (https://www.denovix.com/) is also able to quantify the undiluted purified cDNA product. Use

1 μL to load onto the Denovix instrument. This concentration is overestimated because of optical impurities. To compensate for this effect, the concentration outcome must be divided by 2.

20. If the Denovix was used to normalize the purified cDNA product to 2 ng/μL, then the final concentration must be determined by using the results of the Agilent 2100 BioAnalyzer high-sensitivity DNA chip: set the region table to measure between 400 and 6000 bp, and obtain the yield in pg/μL.
21. If the concentration of the cDNA is <500 pg/μL, add more volume of cDNA and less volume of the nuclease-free water.
22. The index is present in the TCRa and the TCRb human primer 2 reverse primers. When preparing more than one chip, use TCRa and TCRb primers with different indices to enable pooled sequencing.
23. Alternatively, the libraries can be quantified by qPCR using the NGS Library Quantification Kit (Takara Bio, Cat No. 638324). Please refer to the corresponding user manuals for detailed instructions.
24. Following validation, the TCR library is ready for sequencing on the Illumina platforms. It is advised to determine the final molarity of the library by sequencing a small amount of the library first to adjust the sample molarity and perform additional sequencing to the required amount.
25. Do not use the I5 index primer supplied with the Nextera XT Index Kit.
26. In the second purification of the manufacturer's instructions, "50 μL of eluate" is listed. This volume is incomplete as it contains 75 μL (25 μL AMPure XP beads and 50 μL eluate from the first purification). We have used the full eluate volume from the second purification and adjusted the volumes in the third purification: from 10 μL of AMPure XP beads added to ~50 μL eluate to 15 μL of AMPure XP beads added to ~75 μL eluate.
27. Alternatively, the libraries can be quantified by qPCR using the NGS Library Quantification Kit (Takara Bio, Cat No. 638324). Please refer to the corresponding user manuals for detailed instructions.
28. Following validation, the 5′ DE library is ready for sequencing on Illumina platforms. Determine final molarity by sequencing a small amount of the library first. Then normalize the sample molarity and perform further sequencing.
29. Preferred index combinations can be found in the sequencing manuals of your sequencing provider.

30. In the case of dual index sequencing, the second index read will be TCTTTCCC, which is part of the NexteraXT adaptor sequence. This sequence is not an actual index as the TCR and 5′DE sequencing libraries are single-indexed libraries.
31. The advised amount of the TCR reads is based on our previously published experiments [21], where 5 M clusters yielded 71% of the single a TCR signature. Increasing the yield will likely increase the percentage of TCR alpha and beta clonotypes but will also increase cost per sample. The advised amount of 50 M cluster for the 5′ DE sequencing libraries is based on a minimal amount of 50 k clusters per cell multiplied with 1000 single cells, which is a usual number of cells that can be obtained from a Takara TCR and 5′ DE chip.
32. In the unfortunate event that a flow cell yielded an insufficient number of clusters, less than 5 M for a typical experiment, the data files of multiple flow cells can be merged. First, make a new directory to hold the output file, and use the following commands to merge the data over multiple flow cells:

```
zcat \
 {fc_1}/{x}_S{y}_L001_R{z}_001.fastq.gz  \  {fc_n}/{x}_S{y}
_L001_R{z}_001.fastq.gz | \ gzip -c > {dir}/{x}_R{z}.fastq.gz
```

    Please note that neither samples nor reads 1 and 2 should be merged together, and the order of the flow cells remains the same while merging reads 1 and 2. The sample number is variable between flow cells and should not be taken into account while merging.
33. At least two dimensions need to be selected, otherwise the subsequent UMAP will fail. For the visualization, a slightly larger number of dimensions will introduce some noise in the figure, but only in extreme cases will it change the overall topology. Choosing too few dimensions with which to continue will result in a loss of variation which is likely biologically significant. For these reasons, often times a few more dimensions are chosen than the absolute minimum based on the PCA and Jackstraw analyses. If the PCA and Jackstraw analyses are not in accordance, choose the largest number of dimensions.
34. The exact placement of the cells on the plane is arbitrary and uses random number generation. To make the results more reproducible, the seed of the random number generator can be set using the command set.seed. For example, executing set.seed(42) sets the seed of the random number generator to 42.

35. TCR sequences are considered identical, when they are composed of the same V, D, and J genes and have the same CDR3 sequence. As cells should not be counted twice, only the most highly abundant sequence for either TRA or TRB is taken into account.

Acknowledgments

We would like to thank Maaike de Bie, Amy van der List, and Mariska Klepper for technical assistance.

References

1. Garcia KC, Teyton L, Wilson IA (1999) Structural basis of T cell recognition. Annu Rev Immunol 17:369–397
2. Nikolich-Zugich J, Slifka MK, Messaoudi I (2004) The many important facets of T-cell repertoire diversity. Nat Rev Immunol 4: 123–132
3. Miles JJ, Douek DC, Price DA (2011) Bias in the alpha beta T-cell repertoire: implications for disease pathogenesis and vaccination. Immunol Cell Biol 89:375–387
4. Kohler S, Wagner U, Pierer M, Kimmig S, Oppmann B, Mowes B et al (2005) Post-thymic in vivo proliferation of naive CD4+ T cells constrains the TCR repertoire in healthy human adults. Eur J Immunol 35:1987–1994
5. van Dongen JJ, Langerak AW, Bruggemann M, Evans PA, Hummel M, Lavender FL et al (2003) Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 concerted action BMH4-CT98–3936. Leukemia 17: 2257–2317
6. Li B, Li T, Pignon JC, Wang B, Wang J, Shukla SA et al (2016) Landscape of tumor-infiltrating T cell repertoire of human cancers. Nat Genet 48:725–732
7. Langerak AW, van Den Beemd R, Wolvers-Tettero IL, Boor PP, van Lochem EG, Hooijkaas H et al (2001) Molecular and flow cytometric analysis of the V beta repertoire for clonality assessment in mature TCR alpha beta T-cell proliferations. Blood 98:165–173
8. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E et al (2017) Human cell atlas meeting, the human cell atlas. eLife 6: 1–30
9. Valihrach L, Androvic P, Kubista M (2018) Platforms for single-cell collection and analysis. Int J Mol Sci 19:807
10. Papalexi E, Satija R (2018) Single-cell RNA sequencing to explore immune cell heterogeneity. Nat Rev Immunol 18:35–45
11. Aarts M, Georgilis A, Beniazza M, Beolchi P, Banito A, Carroll T et al (2017) Coupling shRNA screens with single-cell RNA-seq identifies a dual role for mTOR in reprogramming-induced senescence. Genes Dev 31: 2085–2098
12. Bergiers I, Andrews T, Vargel Bolukbasi O, Buness A, Janosz E, Lopez-Anguita N et al (2018) Single-cell transcriptomics reveals a new dynamical function of transcription factors during embryonic hematopoiesis. eLife 7: e29312
13. Goldstein LD, Chen YJ, Dunne J, Mir A, Hubschle H, Guillory J et al (2017) Massively parallel nanowell-based single-cell gene expression profiling. BMC Genomics 18:519
14. Ye J, Ma N, Madden TL, Ostell JM (2013) IgBLAST: an immunoglobulin variable domain sequence analysis tool. Nucleic Acids Res 41(W1):W34–W40
15. Zhang JK, Kobert K, Flouri T, Stamatakis A (2014) PEAR: a fast and accurate illumina paired-end ReAd MergeR. Bioinformatics 30(5):614–620
16. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, Satija R (2019) Comprehensive integration of single-cell data. Cell 177(7):1888–1902.e21
17. Daehwan K, Paggi JM, Park C, Bennett C, Salzberg SL (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 37(8): 907–915

18. Quinlan AR (2014) BEDTools: the Swiss-Army tool for genome feature analysis. Curr Protoc Bioinformatics 47:11.12.1–11.1234
19. Litjens NH, Huisman M, Hijdra D, Lambrecht BM, Stittelaar KJ, Betjes MG (2008) IL-2 producing memory CD4+ T lymphocytes are closely associated with the generation of IgG-secreting plasma cells. J Immunol 181(5):3665–3673
20. Litjens NH, de Wit EA, Baan CC, Betjes MG (2013) Activation-induced CD137 is a fast assay for identification and multi-parameter flow cytometric analysis of alloreactive T cells. Clin Exp Immunol 174(1):179–191
21. Litjens NHR, Langerak AW, van der List ACJ, Klepper M, de Bie M, Azmani Z, den Dekker AT, Brouwer RWW, Betjes MGH, Van IJcken WFJ (2020) Validation of a combined transcriptome and T cell receptor alpha/beta (TRA/TRB) repertoire assay at the single cell level for paucicellular samples. Front Immunol 11:1999

Chapter 15

AIRR Community Guide to Planning and Performing AIRR-Seq Experiments

Anne Eugster, Magnolia L. Bostick, Nidhi Gupta, Encarnita Mariotti-Ferrandiz, Gloria Kraus, Wenzhao Meng, Cinque Soto, Johannes Trück, Ulrik Stervbo, and Eline T. Luning Prak and on behalf of the AIRR Community

Abstract

The development of high-throughput sequencing of adaptive immune receptor repertoires (AIRR-seq of IG and TR rearrangements) has provided a new frontier for in-depth analysis of the immune system. The last decade has witnessed an explosion in protocols, experimental methodologies, and computational tools. In this chapter, we discuss the major considerations in planning a successful AIRR-seq experiment together with basic strategies for controlling and evaluating the outcome of the experiment. Members of the AIRR Community have authored several chapters in this edition, which cover step-by-step instructions to successfully conduct, analyze, and share an AIRR-seq project.

Key words AIRR-seq, Immunoglobulin, Antibody, T-cell receptor, Immune repertoire, V(D)J recombination, Next-generation sequencing

1 Introduction

Next-generation sequencing of adaptive immune receptor repertoires (AIRR-seq of immunoglobulin, IG and T-cell receptor, TR rearrangements) has provided a new frontier for in-depth analysis of the immune system. The Adaptive Immune Receptor Repertoire (AIRR) Community was founded with the goal of developing standards for AIRR-seq studies to enable analysis and sharing of AIRR-seq data. In this book, members of the AIRR Community and colleagues have contributed sample methods for immune repertoire profiling studies. These AIRR Community chapters cover experimental (wet lab) and computational (dry lab) methods and encompass all of the many facets of the AIRR Community. While much of our focus in these chapters is on how to adequately control, standardize, annotate, and share data, we found it

Anton W. Langerak (ed.), *Immunogenetics: Methods and Protocols*, Methods in Molecular Biology, vol. 2453, https://doi.org/10.1007/978-1-0716-2115-8_15,

impossible to discuss these attributes of AIRR-seq data without also describing the types of data sets that are generated and then integrating those descriptions with data analysis for commonly encountered use cases. In the companion AIRR Community data analysis chapters, information is provided about study design, data analysis, data use, and the AIRR data commons and how data can be reused and shared. In this chapter we describe how to plan and perform AIRR-seq experiments.

2 Planning the Experiment

Understanding the dynamics, selection, and pathology of immune responses has been aided greatly aided in recent years by next-generation sequencing (NGS)-based approaches to studying the adaptive immune receptor repertoire (AIRR) [1–3]. The AIRR Community is focused on the standardization, sharing, and re-use of these repertoire data [4]. The AIRR is the collection of distinct B-cell and T-cell clones (cells that are derived from a common progenitor cell) that are found in an individual. Each clone is associated with a distinct antigen receptor, which is a B-cell receptor (BCR or IG) or a TR. The DNA sequences that encode IG or TR are very diverse. This diversity is achieved through the recombination of variable (V), diversity (D), and joining (J) gene segments [5, 6]. Moreover, somatic hypermutation (SHM) provides further diversification of IG repertoires through DNA mutation [7, 8]. In addition to facilitating the sampling of diverse and complex immune repertoires, AIRR-seq has opened the door for systematic analysis and comparison of immune responses across different individuals and disease conditions [9–12]. The immune repertoire is dynamic and changes in its composition and diversity with age [13, 14], in different anatomic sites [15] and under diverse conditions such as malignancy, autoimmunity, immunodeficiency, infection, or vaccination [9, 13, 16–21]. In addition to comparing different individuals, AIRR-seq is also a powerful method for studying the evolution of immune responses or tracking specific B- or T-cell populations over time within individuals [22]. For example, clonal expansions can be identified, quantified, and monitored [23]. AIRR-seq studies not only enhance our ability to understand how to diagnose and monitor diseases but also can inform therapeutic approaches [12, 24–31].

When designing a study that leverages AIRR-seq data, there are several considerations including the subjects, sample types, manner in which the samples are processed, timeline and other considerations. The types of samples, their numbers, and budget often drive the types of questions that can be asked and answered using AIRR-seq. Once a suitable question has been defined and appropriate samples have been identified, the next major branch point in the

decision-making process involves the selection of AIRR-seq methods. In this section, we provide a brief overview of the most important considerations when selecting one or more AIRR-seq methods for a research study or clinical evaluation.

2.1 Organisms

This chapter focuses on samples from humans, but of course samples from other vertebrates or synthetic libraries (such as phage display [32]) are possible. If one is planning an experiment with nonhuman or synthetic samples, it is worth considering whether there are established protocols (such as PCR primer sets) and analysis pipelines (to include adequate libraries of validated germline gene sequences for animal species that are not frequently studied) for downstream analysis. With respect to samples derived from humans, there are several considerations [4, 33]. First, are the samples coming from individuals who have been consented for a research study? If not, one should check with the local institutional review board (IRB) or other regulatory body and/or with the investigator who supplies the samples for guidance on whether samples can be studied or if additional regulatory approvals may be required for full analysis and/or sharing of the data. Second, the study design will be impacted by the availability of samples from individuals in different comparison groups or on the availability of samples that are collected over time from the same individuals. Depending on the research question, resources, and time horizon for the project, study participants may be recruited who have a particular disease (in which case the phase of the disease and prior or current therapies may be important). If studying immune responses, longitudinal collections from the same individual at multiple time points and synchronization of those time points across the study cohort may be important to study changes in clonal abundance or, in the case of B cells, the level of SHM within clonal lineages. Demographic characteristics of the individuals in the group under study (including but not limited to age, geographical origin and sex, disease history) and the availability of one or more appropriately matched control groups are additional considerations. For TR-based sequencing, it is also useful to consider the HLA type, as HLA can have a major impact on TRBV gene usage [34]. Finally, if published data are going to be used for comparison, compatibility of the assay platforms and sample types is important.

2.2 Samples and Processing

Studies on humans are often limited by sample availability. The most commonly used sample is peripheral blood, which serves as starting material for a range of different sample types including whole blood (drawn into a tube with an anticoagulant such as EDTA), peripheral blood mononuclear cells (PBMCs, which are typically isolated by centrifugation over a Ficoll gradient), or plasma (the liquid portion of anticoagulated whole blood, which is typically prepared by centrifugation and stored in aliquots frozen for

isolation of cell-free DNA). Samples from other body fluids such as cerebrospinal fluid or bronchoalveolar lavage may also provide important insights if sampled in certain disease states. Tissue samples can be obtained from fine-needle aspirations (where sample quantities may be very limited, particularly if the same samples are being used for both clinical and research purposes) or from biopsies, where larger amounts of tissue can be sampled. In the case of the bone marrow, the aspirate is typically used for the evaluation of clonally expanded populations. In some cases, it is possible to obtain multiple tissues (surveillance biopsies for transplant rejection or bone marrow samples) as well as peripheral blood from the same individual over time. Finally, different tissues can be accessed from the same individual in organ donors or living individuals, as has been described for studies of human tissue-based immunity [35] and in certain disease states, such as type 1 diabetes, lupus, or rheumatoid arthritis [36–43]. From most of these samples, either total cells or isolated cell subsets (obtained after cell sorting using flow cytometry or magnetic bead-based methods) can be analyzed. The sample size and purity of the cell population of interest are important to consider when designing the experiment and interpreting the results.

How samples are processed is a critical consideration for the design of AIRR-seq experiments. Bulk sequencing methods can use samples that are formalin-fixed, lysed, or non-viably cryopreserved. Fixation significantly reduces the quality of the input nucleic acid and may require larger amounts of input DNA or RNA as well as protocols that use shorter amplicons (such as primers that are positioned in FR3 instead of FR1). The longer a sample sits in a fixative or is stored as a formalin-fixed paraffin-embedded (FFPE) tissue section, the poorer the template quality becomes. If it is possible to obtain snap frozen tissues that are not fixed, this is preferable. For certain cell types, such as diffuse large B-cell lymphoma, using tissue sections may provide a higher yield of cells of interest than single-cell suspensions [44]. For single-cell-based methods, viable cells are essential and typically consist of either freshly isolated cells or cryopreserved cells. In the case of cryopreserved cells, one needs to consider whether the method of initial sample preparation has influenced the recovery or phenotype of the cell population of interest.

Cell sorting or enrichment with magnetic beads can be used to selectively recover larger numbers of cells of interest, as, for example, with antigen-specific T cells identified by multimer staining, but these methods can also result in significant loss of sample. Sorting time should be kept to a minimum for plate-based single-cell methods, as cell viability decreases rapidly in the plate; ideally, the time from the addition of a life/dead staining solution to the end of the sort should not exceed 30 min. If longer sorting times are necessary, as is often the case for rare cells, cells can be sorted

into PCR strips instead. For droplet sequencing-based single-cell methods, batches of 1000–20,000 cells are usually collected in PCR tubes that need to be coated to ensure complete recovery of the cells for further processing.

2.3 Bulk vs. Single-Cell Sequencing

There are two complementary approaches to analyze the AIRR by sequencing that are usually driven by the number of cells available and the research question. On the one hand, bulk AIRR-seq methods allow systematic and global analysis of TR and IG repertoires from as few as 1000 cells to hundreds of thousands of cells or more. Bulk methods provide information about the TR (usually alpha + beta) or IG (heavy + light) rearrangements, although the pairing information is lost during the cell lysis step. On the other hand, single-cell AIRR-seq offers the possibility to reconstruct paired chain information for each TR or IG. However, most single-cell methods use lower cell input numbers (usually <20,000 cells, due to constraints in costs associated with kits and sequencing). Hence single-cell approaches, when used on bulk populations, generally tend to be focused on specific cell subsets or antigen-enriched cells to ensure sufficient sampling of the population of interest. In some cases, for example, when multiple samples with different amounts of cell inputs are available from the same individual, it may be preferable to use a tiered approach. For example, one might rely on bulk sequencing to get a view of the overall clonal landscape and then leverage single-cell sequencing to gain detailed insights into the association of specific clones (with paired chain information) and cell phenotypes (either through flow cytometry or by single-cell RNA-seq). The single-cell approach is discussed in detail in the AIRR Community chapter (Chapter 20)

2.4 Template Amplification from DNA vs. RNA

Bulk AIRR-seq can be performed on libraries that have been generated from either genomic DNA (gDNA) or RNA. gDNA-based methods are exclusively based on multiplex PCR approaches, where primers targeting the different V genes (or leader regions) and J genes are combined in the same reaction. Advantages of DNA-based sequencing are the stability of the template and its parsimonious nature (one template per cell), which allows for studies in which large numbers of cells are studied at modest cost. Disadvantages include the potential for primer bias, as PCR primers are usually positioned in the V gene and J gene (due to constraints on sequence length) and the potential loss of amplification in heavily mutated IG sequences. The bulk DNA approach is discussed in the AIRR Community chapter (Chapter 18).

Messenger RNA-based methods can be based on multiplex PCR (with either V and J primer combinations or V and constant region (C) primer combinations), or they can use rapid amplification of cDNA Ends (RACE)-PCR. Advantages of RNA-based sequencing are (1) more “shots on goal” with RNA than DNA

(with individual B/T cells harboring multiple RNA copies vs. only a single DNA copy), allowing for higher yield of amplicons when there are low cell numbers; (2) reduced PCR bias with primers that are in the constant region, (3) the incorporation of unique molecular identifiers (UMI) at the cDNA synthesis step (allowing for the generation of high-fidelity consensus sequences); and (4) the ability to generate data on the constant region usage for isotyping. Disadvantages of RNA-based sequencing methods include greater cost associated with the higher sequencing depths that are required (particularly if UMIs are used) and biases introduced by differences in transcript abundance in different cell types (if mixed rather than sorted populations are used for input). In the AIRR Community chapter (Chapter 19), we focus on the mRNA-based approach to AIRR-seq.

2.5 Commercial Kit vs. Homebrew Bulk Methods

Several commercial kits are now available to generate AIRR-seq data. Currently available commercial kits include gDNA-based methods (e.g., Adaptive Biotechnologies, iRepertoire) as well as mRNA-based methods (e.g., Illumina, Takara Bio, iRepertoire, MiLaboratory). Advantages of commercial-grade AIRR-seq assays are that kit reagents are produced following standards and rigorous quality controls such as qualifying primers, controlling for contamination, and verifying yield and amplification standards. Some vendors obtain certification in meeting rigorous quality standards in their laboratories that manufacture reagents, such as those set forth by the International Organization for Standardization (e.g., ISO 9001). In addition, service providers such as Adaptive Biotechnologies and iRepertoire offer large data sets for comparison and a series of user-friendly data analysis tools. Some disadvantages of commercial methods are that kits are expensive and sometimes these assays are not easily adapted to specific experimental needs. On the other hand, with homebrew assays, there is considerable variation in assay linearity and reproducibility (e.g., *see* ref. 45), and it can take months or even years to set up robust, well-validated assays that are then also not easy to adjust. The use of commercially available kits for in-house experiments can be a compromise to ensure reliability of the reagents and protocol customization.

2.6 Single Cell: Index Sorting and Bead-Based Emulsion Approaches

Single-cell AIRR-seq (scAIRR-seq), as any other single-cell sequencing technology, relies on partitioning each cell. In early protocols, cells were index sorted into plates, and multiplex PCR was used to amplify both chains of immune receptors of a cell concomitantly [46, 47]. The emergence of single-cell RNA-seq (scRNA-seq) has provided another tool for AIRR-seq. Many protocols to recover and sequence mRNA from single cells have been developed and differ in their approaches for cell capture, cDNA synthesis (full-length or tag-based) and amplification (only PCR or PCR following reverse transcription), and library preparation steps

[48]. Probably the most frequently used current commercial protocol for sequencing small cell numbers leverages the scSMARTer technology. With this approach, paired IG/TR information became accessible by combining full-length scRNA-seq amplification approaches with the development of the de novo assembly-based bioinformatics tools (TraCer, scTCR Seq, TRAPes, VDJ Puzzle) [49–52]. Unfortunately, these approaches remain computationally intensive, relatively costly, and are constrained with respect to cell throughput. More recently, bead-based emulsion methods have been developed for higher-throughput single-cell sequencing, allowing access to repertoires of tens of thousands of cells [53]. The formation of droplets in an oil-water emulsion using microfluidics allows single-cell encapsulation, barcoding, and the production of cDNA from each cell and culminates in parallel sequencing of the transcriptomes of thousands of cells [54]. These approaches have been adapted to sequence both TR or IG chains in parallel [55] and are available commercially, via the 10× Genomics platform (Chromium 10×), thereby allowing the processing of samples of 5×10^2 to 1.5×10^4 cells. In addition to paired immune receptor data, it is also possible to obtain scRNA-seq data. Similar approaches are also commercially available including the BD Rhapsody VDJ CDR3 protocol, which relies on cell compartmentation by microwells and allows processing of 1×10^3 to 4×10^4 cells, and the Takara Bio ICELL8 Single-Cell System, which can process ~1×10^3 cells. Recent progress on the throughput of single-cell sorting has been described with CelliGO, which combines cell encapsulation in droplets through microfluidics [56], but sequencing costs are still limiting the widespread adoption of these approaches.

2.7 Cost

Finally, cost may influence the choice of a particular protocol. There are many factors that contribute to the cost of AIRR-seq data generation. For example, the number of samples, the cost of sequencing, the sequencing depth, and the number of cells analyzed per sample are all important considerations. Furthermore, the choice between service providers, commercial kits, and "homebrew" methods will influence costs. In general, gDNA analysis is the most cost-effective method, because it usually requires the lowest-sequencing depth with the largest representation of cells per sample, whereas single-cell analysis is on the opposite end of the spectrum, with bulk cDNA sequencing in the middle [45].

2.8 Overview of Companion AIRR Community Method Chapters

The correct choice of method for a given experimental question is crucial and has to be carefully evaluated. The companion AIRR Community method chapters concern (1) "Bulk gDNA Sequencing of Antibody Heavy-Chain Gene Rearrangements for Detection and Analysis of B-Cell Clone Distribution" (Chapter 18), (2) "Bulk Sequencing from mRNA with UMI for Evaluation of B-Cell

Table 1
Overview of highlighted use cases in associated chapters

Highlighted use case	Major steps
Bulk gDNA Sequencing of Antibody Heavy Chain Gene Rearrangements for Detection and Analysis of B-Cell Clone Distribution; a method by the AIRR Community	
Analysis of the clonal landscape in different samples from the same individual	Bulk gDNA FR1 + JH primers (BIOMED2 adapted) 2 × 300 bp reads Multiple replicates per sample
Bulk Sequencing from mRNA with UMI for Evaluation of B-Cell Isotype and Clonal Evolution; a method by the AIRR Community	
Evaluation of an antibody response to viral infection with clonal evolution	Antigen-enriched cells Bulk RNA (SMARTer Kit) UMI at cDNA synthesis step Amplification of isotypes
Single-Cell Analysis and Tracking of Antigen-Specific T Cells: Integrating Paired-Chain AIRR-Seq and Transcriptome Sequencing; a method by the AIRR Community	
Part A Single-cell-paired TCR chain and/or mRNA sequencing of memory and/or whole CD8+ T cells from COVID-19 patients	Bulk memory CD8+ T cells Feature barcode and sample hashtagging possible Template amplification and size fractionation Paired TCR chain data Single-cell RNA-seq data, feature barcode data
Part B Sequencing of activated and/or SARS-CoV-2 antigen-specific CD8+ T-cell populations to map to whole repertoires	T-cells (sorted from PBMCs) binding to antigen or with activation marker expressionIndex sort possible Single-cell mRNA sequencing (SMART-Seq) Paired TCR chain data Single-cell RNA-seq data
Quality Control: Chain Pairing Precision and Monitoring of Cross-Sample Contamination; a method by the AIRR Community	
Chain pairing reference data and estimation of within sample AIRR-seq reproducibility	Isolation of memory B cells or CD8+ T cells Establishment of replicate stimulation cultures prior to single-cell sequencing

Isotype and Clonal Evolution" (Chapter 19), (3) "Single-Cell Analysis and Tracking of Antigen-Specific T Cells: Integrating Paired-Chain AIRR-Seq and Transcriptome Sequencing" (Chapter 20), and (4) "Quality Control: Chain Pairing Precision and Monitoring of Cross-Sample Contamination" (Chapter 21). These chapters illustrate four basic workflows for AIRR-seq, with a focus on IG for bulk sequencing, TR for single-cell sequencing, and IG and TR replicate analyses for quality control. The four methods are summarized in Table 1 and are discussed further below.

In Chapter 18, we illustrate, using a homebrew method with primer sequences adapted for NGS from the BIOMED2 immunoglobulin heavy-chain (IGH) PCR assays [57], how to evaluate the clonal landscape, including clone size distributions, clonal lineage analysis, and tracking of clones in different samples from the same individual. This method uses multiplex PCR and can be scaled to very high cell inputs as described [15]. The method shown uses long reads that are adequate for robust IGHV gene alignment and SHM evaluation but can also be performed with shorter reads, depending upon the sample type and DNA quality. In Chapter 19, IGH rearrangements are amplified from bulk RNA with UMIs incorporated at the cDNA synthesis step for the generation of high-fidelity consensus sequences using a commercial kit from Takara Bio. This method can be used for low to moderate throughput analysis of antigen-enriched cell populations, for evaluation of SHM, selection, and isotype usage. In Chapter 20, two different but parallel workflows are used to analyze single cells, both for paired TR transcripts as well as for their transcriptome, using two commercial kits, one from Takara Bio and one from 10× Genomics. Single-cell technologies can use a multiplex or RACE-based amplification and can generate long high-quality reads that can be mapped to individual cells but can also be based on AIRR target enrichment. One kit allows for the analysis of small numbers of antigen-enriched, index-sorted cells, useful in the case the cells of interest are present at very low frequencies in the overall sample, while the other kit allows for the analysis of larger cell numbers, providing insights into the overall T-cell repertoire as well as into other immune cell populations, if desired. The combination of paired-chain information and RNA-seq data can provide insights into the nature of the different T-cell populations that are found among expanded clones in various disease settings. Furthermore, through clonal overlap analysis, the data from the antigen-enriched cells can be integrated with the larger data set to further characterize the populations with respect to antigen-binding. In Chapter 21, two workflows are presented. The first is for the isolation of CD27+ memory B cells and their expansion in replicate cultures in vitro, using a cell line that expresses CD40L and a cocktail of cytokines. The second workflow is for the isolation of CD8+ T cells and their expansion using CD3/CD28 and IL-2 stimulation. The generation of these expanded cell cultures provides a larger input of more readily resampled cells that can be used as reference libraries for IG- or TR-paired chain combinations, respectively, as well as providing diverse libraries for the evaluation of within-sample reproducibility.

3 Interpreting the Results

3.1 Overview

Immune repertoire profiling experiments are affected by numerous pre-analytical, experimental, and post-analytical variables. Pre-analytical variables include the quality, quantity, and purity of the target cell population(s) in the sample. Experimental variables include the quality and length of the template for amplification, contamination at the level of the sample or PCR, hybrid PCR products, and PCR jackpots. The sequencing run can be affected by the concentration of the library, which can influence the clustering density; there can be cross-clustering in the flow cell, poor quality or short reads, and issues with controlling for sequencing depth (reads per template). Many technical problems with experiments can be evaluated during data analysis (please *see* the companion AIRR Community commentaries on "TR and IG Gene Annotation" (Chapter 16) and "Repertoire Analysis" (Chapter 17), so here we will limit our comments to basic strategies for controlling and evaluating the adequacy of the experiment on the wet bench side.

3.2 General QC Considerations and Controls

For sample and amplification QC, spectroscopy, agarose gel electrophoresis, or capillary electrophoresis can be used for the evaluation of nucleic acid purity and size distribution. Standardized samples that are put through the same workflow can be used to compare the entire AIRR-seq procedure in one assay run to another run, to help identify and control for batch effects. Bead purification and/or further gel purification can be performed to remove primer dimers, which can swamp sequencing runs and reduce the fraction of informative reads. Capillary electropherograms (e.g., Bioanalyzer) can be used to evaluate library quality, while KAPA quantitation and real-time PCR can be performed to quantify the library. For the sequencing run, the clustering density is important (as described in the individual protocol chapters). Another helpful metric is the fraction of reads that have quality scores of 30 or higher (projected sequencing error rates below 1 per 1000 nucleotides).

3.3 Clonal Recovery

The quality and type of sample have significant effects on the efficiency of amplification and clonal yield. FFPE tissue samples yield ~10-fold fewer clones than the same tissue snap frozen without fixation. Furthermore, the longer a tissue sits in FFPE, the poorer the sample quality becomes. For FFPE samples, using larger amounts of input DNA or RNA into the initial amplification can improve clonal recovery, as can the use of primers that target shorter amplicons (e.g., primers that flank the CDR3 sequence such as FR3 and JH [58]). Another reason for low numbers of clones is if the initial amplification uses primers that do not capture

a high enough fraction of the rearrangements in the sample. With RNA as the starting material, there is bias toward recovering more templates from cells that are activated. Plasma cells, for example, can produce ~100 times as much IG RNA as naive B cells [59]. Primers that amplify DNA are not subject to this problem, but can have other issues, such as the potential for nonuniform amplification of different templates. To correct for PCR bias, some assays use internal calibrators [60, 61]. Amplification of IG rearrangements has an additional challenge if these are highly somatically hypermutated. One hint that this may be occurring is if there is an elevated frequency of nonproductive rearrangements (from a bulk gDNA amplification). Alternative approaches in this situation are to amplify templates that are less prone to SHM such as the leader region in the VH genes or focus on RNA-based sequencing with primers that extend from the constant region [15]. Another approach is to amplify alternative loci (such as light chains, which have about half the level of SHM of heavy chains [62], RS (recombining sequence also known as kappa deleting element) rearrangements [63], or DJ rearrangements [58]).

3.4 PCR Cycle Number

For RNA-based protocols, the gene expression of each IG/TR chains can vary significantly from one cell to another. Therefore, it is challenging to predict how many cycles of PCR will amplify sufficient material for downstream sequencing without overamplification such that there are significant off-target PCR products. One approach is to focus on sorted cell populations to control for the effects of different transcript levels. In addition, one can amplify each chain of interest (e.g., IgH, IgK, IgL, etc.) separately, with different library index combinations for each chain. This can allow for separate optimization of cycling conditions for each chain, as discussed in Chapter 19. It is also possible that the suggested number of cycles will not generate enough material for downstream sequencing. If there is insufficient material for sequencing, we recommend increasing the number of cycles. Conversely, if the library yield is too high, the number of cycles in the library PCR amplification (e.g., PCR2 in Chapter 19) can be decreased.

3.5 Sensitivity

The sensitivity of an AIRR-seq experiment can be determined by titrating spike-ins, such as mixing cells with a known gene rearrangement into a diverse sample at different ratios, as described by Barennes and colleagues [45]. The linearity of the titration also reveals the range of clone concentrations where the method is quantitative or semiquantitative. The threshold of detection of the assay depends upon the biological question being asked, but if rare clonotypes need to be detected (as is the case for detection of minimal residual disease), then it is important to power the analysis on clone sizes. This can be accomplished experimentally by running multiple biological replicates (independent PCR amplifications) on the same sample and determining the fraction of rearrangements

that can be repeatedly sampled in two, three, four, or more replicates, as described previously [15, 64]. Using within-sample clonal overlap as a maximal estimate, one can then evaluate (with greater rigor) the expected overlap between one sample and a different sample [15]. If sensitivity falls below the level required, there are several potential reasons for this including poor-quality sample, too few cells (of the relevant type) in the sample, too small a sample, or a clone size that is too small to be detected. The depth of sequencing can also influence the detection of clones, particularly if one uses rigorous cutoffs for clone size or requires a minimum number of UMIs per clone.

3.6 Amplification Bias

As discussed in the amplification section in Chapter 18, DNA-based amplification methods can exhibit bias in the form of preferential amplification of certain genes over others. RNA-based amplification methods can be biased by transcript abundance, which is higher in certain cell types than others. To evaluate an AIRR-seq experiment for amplification bias, one can use an alternative method, such as flow cytometry with antibodies against known TCR Vβ chains, as a basis for comparison, as described in [45]. In single-cell experiments, one can quantify the recovery of receptors in different cell subsets using RNA-seq profiles to assign cells to different subsets. In addition, spike-in controls and cell mixtures with defined rearrangements can be used during protocol development to quantify bias. Primers with conserved sequence tags can also be used to evaluate bias, as described by Reddy and colleagues [61]. Bias can also occur during the sequencing step. For example, a higher depth of sequencing can result in greater coverage and the detection of smaller clones. However, in samples with few clones, a higher-depth sequencing can also create more sequencing errors which, depending on the bioinformatic pipeline, can result in skewed clone size or SHM profiles. If samples from different sequencing runs are being compared, it is important to consider potential batch effects due to differences in depth of sequencing, clustering density, and sequence quality. To minimize problems associated with batch effects, it is useful to include samples that are being compared to each other in the same run, whenever this is possible. One way to potentially control for (or at least recognize) batch effects is to include an external reference sample (such as pooled spleen or PBMCs) in each run.

3.7 Contamination

During data analysis, one can check for contamination by computing clonal overlap between different samples in the same experiment. Samples from the same individual will exhibit numerous overlapping clones, depending upon the level of sampling, whereas samples between individuals have far fewer overlapping clones. Overlapping clones or identical CDR3 sequences between different individuals cause concern for contamination if they have identical nucleotide sequences and if there are multiple shared sequences

(which is nearly impossible to achieve by chance, particularly for IG sequences, [65]). Spurious clonal overlap between different individuals can arise through mixing of samples prior to nucleic acid amplification, by erroneous assignment of sample barcodes, by PCR contamination, by cross-clustering of samples in the same flow cell, or some combination of these difficulties. Sample mixing can occur during flow cytometry if the instrument is not rigorously flushed between samples. Samples that are assigned the wrong barcode will associate with the "wrong" individual, or if samples come from different species, processing with the wrong pipeline (including the wrong database for reference germline genes) will result in sequences that have very low levels of sequence homology to the (incorrect) germline genes. If this occurs, an IgBLAST [66] search with a few sequences will quickly resolve to which species the genes correspond. With PCR contamination, one may see spurious amplification in the negative control samples (such as water or fibroblast DNA). PCR contamination can also often result in high-copy sequences that are shared by multiple subjects in the same experiment. In contrast, with cross-clustering, there is often a very-high-copy sequence and then a low number of copies of that same sequence in an unrelated individual. There are several process controls that can reduce the risk of contamination. First, there should be physically separate areas for pre- and post-PCR workstations. Second, primers with different barcodes can be used for diagnostic samples (where high-copy clones might be present) vs. MRD samples. Unique dual indices can be used to control for sequencing barcode crosstalk [67]. Third, when in doubt and if more samples are available, repeat the experiment to confirm the results.

3.8 Spurious Amplification Products

Sometimes one obtains unexpected sequences due to technical artifacts. Large clonal expansions can appear with PCR jackpots. In the case of gDNA, independent PCR amplifications of the same sample are sampling different gene rearrangements. If the same expanded clone is present in both biological replicates, it is far more likely to be due to a *bona fide* expansion instead of a PCR jackpot. Another artifact is a hybrid PCR product. With hybrid PCR products, templates with partial sequence homology can cross-amplify [68]. Hybrid products will tend to share sequences at either the 5′ or 3′ end and then exhibit a sharp boundary where the templates crossed over into the other sequence. One way to distinguish hybrid products from gene conversion events or biological variants in V gene sequences or potential convergence (with sharing of CDR3 sequences) is to amplify sequences with TRBV or IGHV gene specific primers and see if the same products can be recreated. In addition, using protocols with fewer PCR cycle numbers can sometimes be helpful in reducing spurious amplification products.

3.9 Data Reporting

The AIRR Community has published a series of data and experimental metadata sharing standards called MiAIRR [33]. The MiAIRR data standards guide the publication, curation, and sharing of AIRR-seq data and metadata and consist of six high-level data sets for study and subject, sample collection, sample processing and sequencing, raw sequences, processing of sequence data, and processed AIRR sequences. All current data fields in the MiAIRR standard can be accessed here: https://docs.airr-community.org/en/stable/miairr/data_elements.html.

More details on how to annotate and report AIRR-seq data and metadata are provided in the AIRR Community companion method chapter "Data sharing and re-use" (Chapter 23).

4 Conclusion

In this chapter, we have given an overview of the considerations needed to plan and execute a successful AIRR-seq experiment. We have also broadly discussed basic strategies for controlling and evaluating the adequacy of the experiment. Each topic touched upon in this chapter is explored in depth in the corresponding AIRR Community companion chapters.

Acknowledgments

This work is supported by NIH research grants awarded to E.L.P. (AI144288, AI106697, P30-AI0450080, P30-CA016520). U.S. was supported by grants from Mercator Stiftung, Germany; German Research Foundation, Germany (DFG, grant 397650460); BMBF e:KID, Germany (01ZX1612A); and BMBF NoChro, Germany (FKZ 13GW0338B). A.E. and G.K. are supported by grants from the Deutsche Forschungsgemeinschaft (BO 3429/3-1 and BO 3429/4-1) and the BMBF (RESET-AID). E.M.F. contributions were funded by iMAP (ANR-16-RHUS-0001), Transimmunom LabEX (ANR-11-IDEX-0004-02), TriPoD ERC Research Advanced Grant (Fp7-IdEAS-ErC-322856), AIR-MI (ANR-18-ECVD-0001), iReceptorPlus (H2020 Research and Innovation Programme 825821), and SirocCo (ANR-21-CO12-0005-01) grants. J.T. was supported by the Swiss National Science Foundation (Ambizione-SCORE: PZ00P3_161147 and PZ00P3_183777).

The authors thank Andrew Farmer for constructive criticism of the manuscript.

E.L.P. is the former Chair of the Adaptive Immune Receptor Repertoire Community, receives research funding from Roche Diagnostics and Janssen Pharmaceuticals for projects unrelated to

the methods presented in this chapter, and is consulting or an advisor for Roche Diagnostics, Enpicom, the Antibody Society, IEDB, and the American Autoimmune Related Diseases Association. J.T. is consulting or an advisor for Enpicom and Merck, Sharp & Dohme (MSD).

References

1. Boudinot P, Marriotti-Ferrandiz ME, Pasquier LD, Benmansour A, Cazenave PA, Six A (2008) New perspectives for large-scale repertoire analysis of immune receptors. Mol Immunol 45(9):2437–2445. https://doi.org/10.1016/j.molimm.2007.12.018
2. Robins HS, Campregher PV, Srivastava SK, Wacher A, Turtle CJ, Kahsai O et al (2009) Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. Blood 114(19):4099–4107. https://doi.org/10.1182/blood-2009-04-217604
3. Georgiou G, Ippolito GC, Beausang J, Busse CE, Wardemann H, Quake SR (2014) The promise and challenge of high-throughput sequencing of the antibody repertoire. Nat Biotechnol 32(2):158–168. https://doi.org/10.1038/nbt.2782
4. Breden F, Luning Prak ET, Peters B, Rubelt F, Schramm CA, Busse CE et al (2017) Reproducibility and reuse of adaptive immune receptor repertoire data. Front Immunol 8:1418. https://doi.org/10.3389/fimmu.2017.01418
5. Tonegawa S (1983) Somatic generation of antibody diversity. Nature 302(5909):575–581. https://doi.org/10.1038/302575a0
6. Sakano H, Kurosawa Y, Weigert M, Tonegawa S (1981) Identification and nucleotide sequence of a diversity DNA segment (D) of immunoglobulin heavy-chain genes. Nature 290(5807):562–565. https://doi.org/10.1038/290562a0
7. Weigert MG, Cesari IM, Yonkovich SJ, Cohn M (1970) Variability in the lambda light chain sequences of mouse antibody. Nature 228(5276):1045–1047. https://doi.org/10.1038/2281045a0
8. Papavasiliou FN, Schatz DG (2002) Somatic hypermutation of immunoglobulin genes: merging mechanisms for genetic diversity. Cell 109(Suppl):S35–S44. https://doi.org/10.1016/s0092-8674(02)00706-7
9. Benichou J, Ben-Hamo R, Louzoun Y, Efroni S (2012) Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. Immunology 135(3):183–191. https://doi.org/10.1111/j.1365-2567.2011.03527.x
10. Six A, Mariotti-Ferrandiz ME, Chaara W, Magadan S, Pham HP, Lefranc MP et al (2013) The past, present, and future of immune repertoire biology—the rise of next-generation repertoire analysis. Front Immunol 4:413. https://doi.org/10.3389/fimmu.2013.00413
11. Langerak AW, Bruggemann M, Davi F, Darzentas N, van Dongen JJM, Gonzalez D et al (2017) High-throughput immunogenetics for clinical and research applications in immunohematology: potential and challenges. J Immunol 198(10):3765–3774. https://doi.org/10.4049/jimmunol.1602050
12. Arnaout RA, Luning Prak ET, Schwab N, Rubelt F, the AIRR Community (2021) The future of blood testing is the immunome. Front Immunol. https://doi.org/10.3389/fimmu.2021.626793
13. Ghraichy M, Galson JD, Kelly DF, Truck J (2018) B-cell receptor repertoire sequencing in patients with primary immunodeficiency: a review. Immunology 153(2):145–160. https://doi.org/10.1111/imm.12865
14. Gibson KL, Wu YC, Barnett Y, Duggan O, Vaughan R, Kondeatis E et al (2009) B-cell diversity decreases in old age and is correlated with poor health status. Aging Cell 8(1):18–25. https://doi.org/10.1111/j.1474-9726.2008.00443.x
15. Meng W, Zhang B, Schwartz GW, Rosenfeld AM, Ren D, Thome JJC et al (2017) An atlas of B-cell clonal distribution in the human body. Nat Biotechnol 35(9):879–884. https://doi.org/10.1038/nbt.3942
16. Kirsch IR, Watanabe R, O'Malley JT, Williamson DW, Scott LL, Elco CP et al (2015) TCR sequencing facilitates diagnosis and identifies mature T cells as the cell of origin in CTCL. Sci Transl Med 7(308):308ra158. https://doi.org/10.1126/scitranslmed.aaa9122
17. Dziubianau M, Hecht J, Kuchenbecker L, Sattler A, Stervbo U, Rodelsperger C et al (2013) TCR repertoire analysis by next generation sequencing allows complex differential diagnosis of T cell-related pathology. Am J

Transplant 13(11):2842–2854. https://doi.org/10.1111/ajt.12431

18. Hou D, Chen C, Seely EJ, Chen S, Song Y (2016) High-throughput sequencing-based immune repertoire study during infectious disease. Front Immunol 7:336. https://doi.org/10.3389/fimmu.2016.00336
19. Galson JD, Kelly DF, Truck J (2015) Identification of antigen-specific B-cell receptor sequences from the total B-cell repertoire. Crit Rev Immunol 35(6):463–478. https://doi.org/10.1615/CritRevImmunol.2016016462
20. Tipton CM, Fucile CF, Darce J, Chida A, Ichikawa T, Gregoretti I et al (2015) Diversity, cellular origin and autoreactivity of antibody-secreting cell population expansions in acute systemic lupus erythematosus. Nat Immunol 16(7):755–765. https://doi.org/10.1038/ni.3175
21. Turner JS, Zhou JQ, Han J, Schmitz AJ, Rizk AA, Alsoussi WB et al (2020) Human germinal centres engage memory and naive B cells after influenza vaccination. Nature 586(7827):127–132. https://doi.org/10.1038/s41586-020-2711-0
22. Davis CW, Jackson KJL, McElroy AK, Halfmann P, Huang J, Chennareddy C et al (2019) Longitudinal analysis of the human B cell response to ebola virus infection. Cell 177(6):1566–1582.e1517. https://doi.org/10.1016/j.cell.2019.04.036
23. Miho E, Yermanos A, Weber CR, Berger CT, Reddy ST, Greiff V (2018) Computational strategies for dissecting the high-dimensional complexity of adaptive immune repertoires. Front Immunol 9:224. https://doi.org/10.3389/fimmu.2018.00224
24. Robinson WH (2015) Sequencing the functional antibody repertoire—diagnostic and therapeutic discovery. Nat Rev Rheumatol 11(3):171–182. https://doi.org/10.1038/nrrheum.2014.220
25. Fink K (2019) Can we improve vaccine efficacy by targeting T and B cell repertoire convergence? Front Immunol 10:110. https://doi.org/10.3389/fimmu.2019.00110
26. Jiang N, Schonnesen AA, Ma KY (2019) Ushering in integrated T cell repertoire profiling in cancer. Trends Cancer 5(2):85–94. https://doi.org/10.1016/j.trecan.2018.11.005
27. Jacobsen LM, Posgai A, Seay HR, Haller MJ, Brusko TM (2017) T cell receptor profiling in type 1 diabetes. Curr Diab Rep 17(11):118. https://doi.org/10.1007/s11892-017-0946-4
28. Theil A, Wilhelm C, Kuhn M, Petzold A, Tuve S, Oelschlagel U et al (2017) T cell receptor repertoires after adoptive transfer of expanded allogeneic regulatory T cells. Clin Exp Immunol 187(2):316–324. https://doi.org/10.1111/cei.12887
29. Sellner L, Bruggemann M, Schlitt M, Knecht H, Herrmann D, Reigl T et al (2017) GvL effects in T-prolymphocytic leukemia: evidence from MRD kinetics and TCR repertoire analyses. Bone Marrow Transplant 52(4):544–551. https://doi.org/10.1038/bmt.2016.305
30. Stervbo U, Nienen M, Hecht J, Viebahn R, Amann K, Westhoff TH et al (2020) Differential diagnosis of interstitial allograft rejection and BKV nephropathy by T-cell receptor sequencing. Transplantation 104(4):e107–e108. https://doi.org/10.1097/TP.0000000000003054
31. Babel N, Stervbo U, Reinke P, Volk HD (2019) The identity card of T cells-clinical utility of T-cell receptor repertoire analysis in transplantation. Transplantation 103(8):1544–1555. https://doi.org/10.1097/TP.0000000000002776
32. Glanville J, Zhai W, Berka J, Telman D, Huerta G, Mehta GR et al (2009) Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. Proc Natl Acad Sci U S A 106(48):20216–20221. https://doi.org/10.1073/pnas.0909775106
33. Rubelt F, Busse CE, Bukhari SAC, Burckert JP, Mariotti-Ferrandiz E, Cowell LG et al (2017) Adaptive Immune Receptor Repertoire Community recommendations for sharing immune-repertoire sequencing data. Nat Immunol 18(12):1274–1278. https://doi.org/10.1038/ni.3873
34. DeWitt WS 3rd, Smith A, Schoch G, Hansen JA, Matsen FA, Bradley P (2018) Human T cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity. eLife 7. https://doi.org/10.7554/eLife.38358
35. Szabo PA, Miron M, Farber DL (2019) Location, location, location: tissue resident memory T cells in mice and humans. Sci Immunol 4(34). https://doi.org/10.1126/sciimmunol.aas9673
36. Kaestner KH, Powers AC, Naji A, Consortium H, Atkinson MA (2019) NIH initiative to improve understanding of the pancreas, islet, and autoimmunity in type 1 diabetes: the human pancreas analysis program (HPAP). Diabetes 68 (7):1394–1402. https://doi.org/10.2337/db19-0058
37. Zhang J, Hu M, Wang B, Gao J, Wang L, Li L et al (2018) Comprehensive assessment of

T-cell repertoire following autologous hematopoietic stem cell transplantation for treatment of type 1 diabetes using high-throughput sequencing. Pediatr Diabetes 19(7): 1229–1237. https://doi.org/10.1111/pedi.12728

38. Seay HR, Yusko E, Rothweiler SJ, Zhang L, Posgai AL, Campbell-Thompson M et al (2016) Tissue distribution and clonal diversity of the T and B cell repertoire in type 1 diabetes. JCI Insight 1(20):e88242. https://doi.org/10.1172/jci.insight.88242
39. Li H, Adamopoulos IE, Moulton VR, Stillman IE, Herbert Z, Moon JJ et al (2020) Systemic lupus erythematosus favors the generation of IL-17 producing double negative T cells. Nat Commun 11(1):2859. https://doi.org/10.1038/s41467-020-16636-4
40. Liu X, Zhang W, Zhao M, Fu L, Liu L, Wu J et al (2019) T cell receptor beta repertoires as novel diagnostic markers for systemic lupus erythematosus and rheumatoid arthritis. Ann Rheum Dis 78(8):1070–1078. https://doi.org/10.1136/annrheumdis-2019-215442
41. Thapa DR, Tonikian R, Sun C, Liu M, Dearth A, Petri M et al (2015) Longitudinal analysis of peripheral blood T cell receptor diversity in patients with systemic lupus erythematosus by next-generation sequencing. Arthritis Res Ther 17:132. https://doi.org/10.1186/s13075-015-0655-9
42. Sui W, Hou X, Zou G, Che W, Yang M, Zheng C et al (2015) Composition and variation analysis of the TCR beta-chain CDR3 repertoire in systemic lupus erythematosus using high-throughput sequencing. Mol Immunol 67(2Pt B):455–464. https://doi.org/10.1016/j.molimm.2015.07.012
43. Bashford-Rogers RJM, Bergamaschi L, McKinney EF, Pombal DC, Mescia F, Lee JC et al (2019) Analysis of the B cell receptor repertoire in six immune-mediated diseases. Nature 574(7776):122–126. https://doi.org/10.1038/s41586-019-1595-3
44. Bertram HC, Check IJ, Milano MA (2001) Immunophenotyping large B-cell lymphomas. Flow cytometric pitfalls and pathologic correlation. Am J Clin Pathol 116(2):191–203. https://doi.org/10.1309/BA3U-RMTU-D7UJ-M8DR
45. Barennes P, Quiniou V, Shugay M, Egorov ES, Davydov AN, Chudakov DM et al (2021) Benchmarking of T cell receptor repertoire profiling methods reveals large systematic biases. Nat Biotechnol 39(2):236–245. https://doi.org/10.1038/s41587-020-0656-3
46. Kim SM, Bhonsle L, Besgen P, Nickel J, Backes A, Held K et al (2012) Analysis of the paired TCR alpha- and beta-chains of single human T cells. PLoS One 7(5):e37338. https://doi.org/10.1371/journal.pone.0037338
47. Eugster A, Lindner A, Catani M, Heninger AK, Dahl A, Klemroth S et al (2015) High diversity in the TCR repertoire of GAD65 autoantigen-specific human CD4+ T cells. J Immunol 194(6):2531–2538. https://doi.org/10.4049/jimmunol.1403031
48. Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M et al (2017) Comparative analysis of single-cell RNA sequencing methods. Mol Cell 65(4): 631–643.e634. https://doi.org/10.1016/j.molcel.2017.01.023
49. Afik S, Yates KB, Bi K, Darko S, Godec J, Gerdemann U et al (2017) Targeted reconstruction of T cell receptor sequence from single cell RNA-seq links CDR3 length to T cell differentiation state. Nucleic Acids Res 45(16):e148. https://doi.org/10.1093/nar/gkx615
50. Stubbington MJT, Lonnberg T, Proserpio V, Clare S, Speak AO, Dougan G et al (2016) T cell fate and clonality inference from single-cell transcriptomes. Nat Methods 13(4):329–332. https://doi.org/10.1038/nmeth.3800
51. Redmond D, Poran A, Elemento O (2016) Single-cell TCRseq: paired recovery of entire T-cell alpha and beta chain transcripts in T-cell receptors from single-cell RNAseq. Genome Med 8(1):80. https://doi.org/10.1186/s13073-016-0335-7
52. Eltahla AA, Rizzetto S, Pirozyan MR, Betz-Stablein BD, Venturi V, Kedzierska K et al (2016) Linking the T cell receptor to the single cell transcriptome in antigen-specific human T cells. Immunol Cell Biol 94(6):604–611. https://doi.org/10.1038/icb.2016.16
53. DeKosky BJ, Kojima T, Rodin A, Charab W, Ippolito GC, Ellington AD et al (2015) In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. Nat Med 21(1):86–91. https://doi.org/10.1038/nm.3743
54. Klein AM, Macosko E (2017) InDrops and Drop-seq technologies for single-cell sequencing. Lab Chip 17(15):2540–2541. https://doi.org/10.1039/c7lc90070h
55. Zemmour D, Zilionis R, Kiner E, Klein AM, Mathis D, Benoist C (2018) Single-cell gene expression reveals a landscape of regulatory T cell phenotypes shaped by the TCR. Nat Immunol 19(3):291–301. https://doi.org/10.1038/s41590-018-0051-0

56. Gerard A, Woolfe A, Mottet G, Reichen M, Castrillon C, Menrath V et al (2020) High-throughput single-cell activity-based screening and sequencing of antibodies using droplet microfluidics. Nat Biotechnol 38(6):715–721. https://doi.org/10.1038/s41587-020-0466-7
57. van Dongen JJ, Langerak AW, Bruggemann M, Evans PA, Hummel M, Lavender FL et al (2003) Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98-3936. Leukemia 17(12):2257–2317. https://doi.org/10.1038/sj.leu.2403202
58. Scheijen B, Meijers RWJ, Rijntjes J, van der Klift MY, Mobs M, Steinhilber J et al (2019) Next-generation sequencing of immunoglobulin gene rearrangements for clonality assessment: a technical feasibility study by EuroClonality-NGS. Leukemia 33(9):2227–2240. https://doi.org/10.1038/s41375-019-0508-7
59. Perry RP, Kelley DE, Coleclough C, Kearney JF (1981) Organization and expression of immunoglobulin genes in fetal liver hybridomas. Proc Natl Acad Sci U S A 78(1):247–251. https://doi.org/10.1073/pnas.78.1.247
60. Carlson CS, Emerson RO, Sherwood AM, Desmarais C, Chung MW, Parsons JM et al (2013) Using synthetic templates to design an unbiased multiplex PCR assay. Nat Commun 4:2680. https://doi.org/10.1038/ncomms3680
61. Friedensohn S, Lindner JM, Cornacchione V, Iazeolla M, Miho E, Zingg A et al (2018) Synthetic standards combined with error and bias correction improve the accuracy and quantitative resolution of antibody repertoire sequencing in human naive and memory B cells. Front Immunol 9:1401. https://doi.org/10.3389/fimmu.2018.01401
62. Brard F, Shannon M, Luning Prak ET, Litwin S, Weigert M (1999) Somatic mutation and light chain rearrangement generate autoimmunity in anti-single-stranded DNA transgenic MRL/lpr mice. J Exp Med 190(5):691–704. https://doi.org/10.1084/jem.190.5.691
63. Hieter PA, Korsmeyer SJ, Waldmann TA, Leder P (1981) Human immunoglobulin kappa light-chain genes are deleted or rearranged in lambda-producing B cells. Nature 290(5805):368–372. https://doi.org/10.1038/290368a0
64. Rosenfeld AM, Meng W, Chen DY, Zhang B, Granot T, Farber DL et al (2018) Computational evaluation of B-cell clone sizes in bulk populations. Front Immunol 9:1472. https://doi.org/10.3389/fimmu.2018.01472
65. Japp AS, Meng W, Rosenfeld AM, Perry DJ, Thirawatananond P, Bacher RL et al (2021) TCR(+)/BCR(+) dual-expressing cells and their associated public BCR clonotype are not enriched in type 1 diabetes. Cell 184(3):827–839.e814. https://doi.org/10.1016/j.cell.2020.11.035
66. Ye J, Ma N, Madden TL, Ostell JM (2013) IgBLAST: an immunoglobulin variable domain sequence analysis tool. Nucleic Acids Res 41(Web server issue):W34–W40. https://doi.org/10.1093/nar/gkt382
67. MacConaill LE, Burns RT, Nag A, Coleman HA, Slevin MK, Giorda K et al (2018) Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. BMC Genomics 19(1):30. https://doi.org/10.1186/s12864-017-4428-5
68. Shuldiner AR, Nirula A, Roth J (1989) Hybrid DNA artifact from PCR of closely related target sequences. Nucleic Acids Res 17(11):4409. https://doi.org/10.1093/nar/17.11.4409

Chapter 16

Adaptive Immune Receptor Repertoire (AIRR) Community Guide to TR and IG Gene Annotation

Lmar Babrak, Susanna Marquez, Christian E. Busse, William D. Lees, Enkelejda Miho, Mats Ohlin, Aaron M. Rosenfeld, Ulrik Stervbo, Corey T. Watson, and Chaim A. Schramm and on behalf of the AIRR Community

Abstract

High-throughput sequencing of adaptive immune receptor repertoires (AIRR, i.e., IG and TR) has revolutionized the ability to carry out large-scale experiments to study the adaptive immune response. Since the method was first introduced in 2009, AIRR sequencing (AIRR-Seq) has been applied to survey the immune state of individuals, identify antigen-specific or immune-state-associated signatures of immune responses, study the development of the antibody immune response, and guide the development of vaccines and antibody therapies. Recent advancements in the technology include sequencing at the single-cell level and in parallel with gene expression, which allows the introduction of multi-omics approaches to understand in detail the adaptive immune response. Analyzing AIRR-seq data can prove challenging even with high-quality sequencing, in part due to the many steps involved and the need to parameterize each step. In this chapter, we outline key factors to consider when preprocessing raw AIRR-Seq data and annotating the genetic origins of the rearranged receptors. We also highlight a number of common difficulties with common AIRR-seq data processing and provide strategies to address them.

Key words AIRR-Seq, B-cell receptor, Germline database, Gene annotation, Preprocessing, Single-cell sequencing, T-cell receptor

1 Introduction

Once an Adaptive Immune Receptor Repertoire sequencing (AIRR-seq, please see the AIRR Community glossary at doi: https://doi.org/10.5281/zenodo.5095381 for definitions of key terms) experiment has been successfully designed and carried out (see discussion in the Chap. 15, attention turns to analyzing the data collected to produce biological insights. Many of the same

Lmar Babrak and Susanna Marquez are shared first authors.

Anton W. Langerak (ed.), *Immunogenetics: Methods and Protocols*, Methods in Molecular Biology, vol. 2453, https://doi.org/10.1007/978-1-0716-2115-8_16,

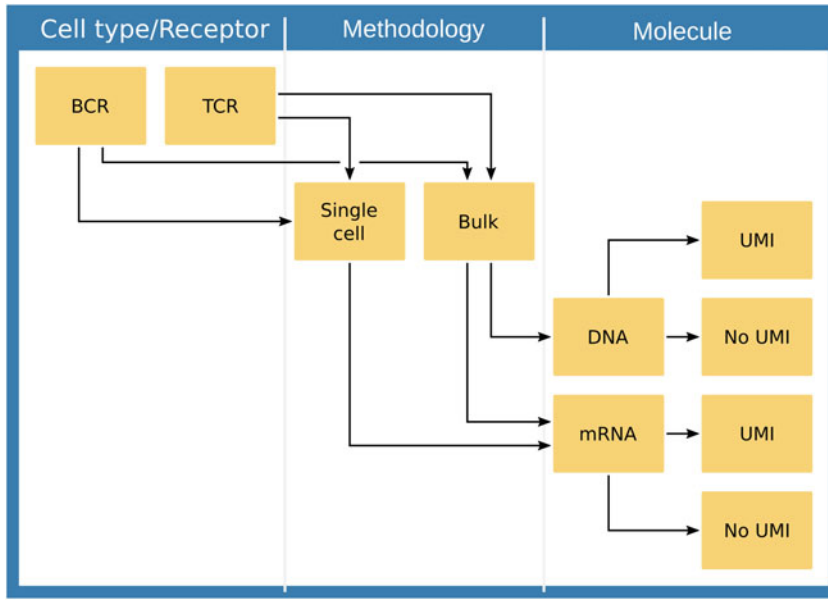


Fig. 1 AIRR-seq decision points. The different ways an AIRR-seq experiment can be constructed. Each choice has implications both for the experimental methodology and for the design of an appropriate analysis strategy

factors that influenced choices in experimental design will be important in planning the computational approach as well. AIRR-seq data to be analyzed may have been generated from genomic DNA or mRNA, with or without unique molecular identifiers (UMIs), and in bulk or single-cell context, as described in the Chap. 15. Each of these alternatives may require (or preclude) the use of certain software tools and influence the interpretation of the analysis. In addition, thought must be given to what computational and storage resources will be necessary given the size of the dataset and the intended analysis.

A clear first decision point in AIRR-seq data analysis is whether IG or TR repertoires are being analyzed (Fig. 1). While many tools such as MiXCR [1], IMGT [2], and others (Table 1) can handle both types of data, some are specific to one or the other. In addition, interest in specialized inquiries like phylogenetic analysis of IGs or calculation of clonal dynamics may require additional specific tools. In such a case, it may be useful to work within a particular ecosystem like Immcantation (http://immcantation.org), VDJServer [18], or SONAR [12], which provide several tools for a thorough analysis from quality control to clonal analysis, to facilitate smooth workflows.

The most critical set of considerations revolve around the origins of the molecules that were actually loaded into the sequencer (*see* Chap. 15). They may have been initially amplified from genomic DNA or from mRNA; the former results in exactly

Table 1
Software tools

Software	Notes/description	URL	
Preprocessing			
Change-O	Data standardization, germline reconstruction, and clonal assignment. Part of the Immcantation suite	https://changeo.readthedocs.io/en/stable/	[3]
pRESTO	Raw data processing. All Immcantation suite tools are certified as compliant with AIRR community software guidelines	https://presto.readthedocs.io/en/stable/	[4]
TraCeR	Extracts and reconstructs rearranged TRs from short read RNA-seq data. Does not support AIRR data representations	https://github.com/Teichlab/tracer/	[5]
VDJPipe	High-performance raw data preprocessing	https://bitbucket.org/vdjserver/vdj_pipe/src/master/	[6]
Gene annotation			
Cell ranger	Proprietary software from 10x genomics for processing AIRR-seq and transcriptomic data generated from the 10× chromium controller	https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger/	
Decombinator	Analysis of TR sequences	https://github.com/innate2adaptive/Decombinator/	[7]
IMGT/high V-QUEST	Free (with registration) access to computational resources to run IMGT/V-quest on up to 1,000,000 sequences at once	http://www.imgt.org/HighV-QUEST/login.action	[8]
IMGT/V-QUEST	Proprietary web tool for annotating IG and TR sequences	http://www.imgt.org/IMGT_vquest/vquest/	[2]
IMSEQ	Error-aware tool for high-throughput AIRR-seq data analysis. Does not support AIRR data representations	http://www.imtools.org	[9]
IgBLAST	BLAST-based identification of IG and TR V genes. Available as both a web interface and a downloadable tool	https://www.ncbi.nlm.nih.gov/igblast/	[10]
MiXCR	Universal tool for annotating and analyzing AIRR-seq data	https://mixcr.readthedocs.io/en/master/	[1]
Partis	Hidden Markov model-based framework for annotating IG and TR sequences	https://github.com/psathyrella/partis/	[11]

(continued)

Table 1
(continued)

Software	Notes/description	URL	
SONAR	BLAST-based with custom wrappers, for IG sequences only. SONAR is certified as compliant with AIRR Community software guidelines	https://github.com/scharch/SONAR/	[12]
Vidjil	Available as both a web interface and a downloadable tool	http://www.vidjil.org	[13, 14]
Gene inference			
TIgGER	Identifies novel alleles based on the intercept of the linear fit. Part of the Immcantation suite	https://tigger.readthedocs.io/en/stable/	[15]
Partis	Identifies novel alleles based on the intercept of the linear fit. Part of the Immcantation suite	https://github.com/psathyrella/partis/	[16]
IgDiscover	Identifies alleles present by iterative clustering	http://docs.igdiscover.se/en/stable/	[17]
Preprocessing, annotation, and analysis environments			
VDJServer	A free, scalable resource for performing immune repertoire analysis and sharing data	https://vdjserver.org	[18]
ImmuneDB	Database and analysis tool for large amounts of AIRR-seq data	https://immunedb.readthedocs.io/en/latest/	[19]

one initial copy of each productive V(D)J rearrangement in a cell, while the latter starts with several or many copies and may vary with cell type and activation state. When amplifying mRNA, the initial molecules may also be labeled with UMIs, which enable the correction of errors introduced by PCR and/or sequencing by identifying reads that are derived from the same original molecule. Of note, while the usage of UMIs enables experimental error correction, their usage necessitates a considerably larger sequencing depth due to consensus read building (for a more nuanced discussion, see, e.g., [20, 21]). UMIs may also be used when sequencing DNA, but that is currently less common in practice. UMIs can also be used to improve quantification, by collapsing apparent expansions due to differential amplification. Some specialized UMI protocols may also require particular matched software tools to fully utilize the advantages of those schemes [22]. Without UMIs, it is advisable to cluster highly similar reads to avoid overcounting, particularly for IG sequences, where errors and somatic hypermutation (SHM) are often indistinguishable.

It is also important to think about how molecules from the full repertoire get included into the pool to be amplified for sequencing. For mRNA-derived libraries, in particular, the efficiency of cDNA generation can be a significant bottleneck and may vary depending on the enzymes and protocol used in the reverse transcription (RT) reaction [23, 24]. The efficiency of the RT reaction can lead to a bias toward abundant species in the repertoire and concomitant dropout of rare ones. In addition, because of the diversity of V and J genes and their surrounding genetic context, many protocols use pools of primers to capture the full repertoire [25]. However, these primers may have different efficiencies in amplifying their respective targets, and some genes might be targeted by more than one primer in a pool. Other protocols circumvent this problem by adding 5′ anchors during reverse transcription [26]. In addition, IGs with high SHM can lose their ability to bind to an intended primer, resulting in the depletion of these sequences from the measured repertoire.

Recently, several high-throughput technologies have become popular for conducting AIRR-seq at single-cell resolution. These provide the most accurate, direct measurements of repertoire statistics and allow more biologically accurate definitions of clones. To do so, however, requires analysis tools that are capable of keeping heavy/light, alpha/beta, or gamma/delta chain sequences linked. The AIRR Community [27] (https://www.antibodysociety.org/the-airr-community/) is developing standardized representations for "receptors" and "cells" to facilitate these analyses and ensure data portability. In addition, single-cell IG and TR data can be easily linked to transcriptomic and other measurements for more comprehensive analyses.

The sequencing technology used must also be taken into account. Illumina paired-end sequencing requires an additional preprocessing step to reassemble the amplicon, and this may result in a bias against longer sequences, with less overlap between the two reads. Meanwhile, more error-prone long-read technologies require extra attention to quality control.

This chapter aims to guide bioinformaticians through the first steps in repertoire analysis, specifically the considerations and preparation of raw data for subsequent repertoire analysis (*see* Chap. 17). Firstly, this chapter provides in-depth information on the materials necessary to conduct the analysis, including computational resources for data preparation, available software tools, and germline database information (Fig. 2). The main portion of the chapter then discusses the considerations on data preprocessing and annotation of raw sequences with a reference germline database.

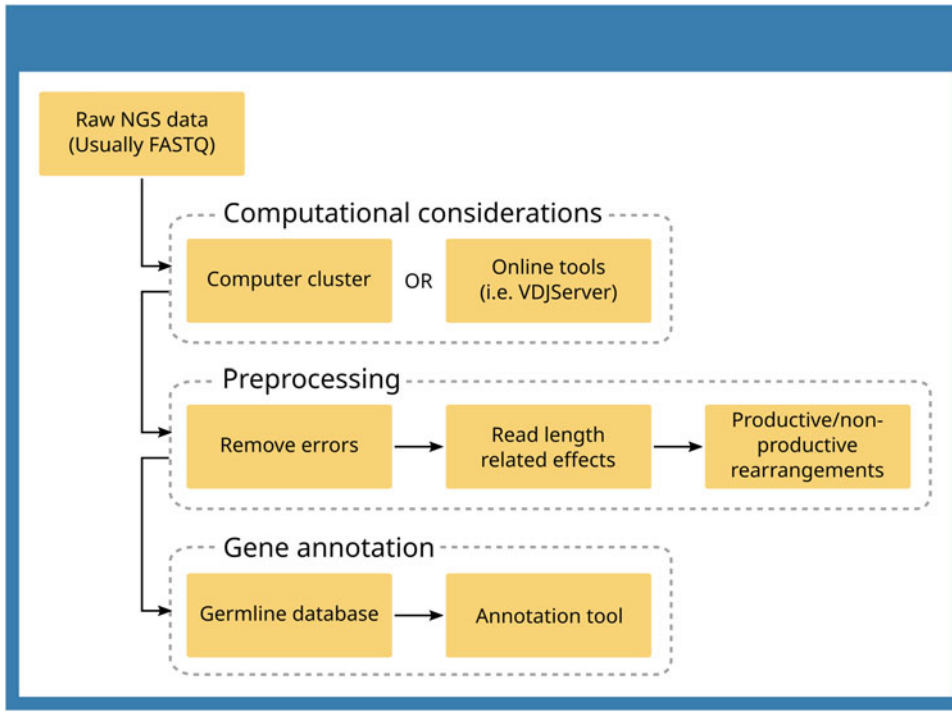


Fig. 2 Process overview. Conceptual steps in designing an AIRR-seq analysis, proceeding from raw inputs to annotated sequences for downstream analysis

2 Materials

2.1 Computing Resources

AIRR-seq data are usually large and require specialized analysis methods and software tools. A typical Illumina MiSeq sequencing run generates 20–30 million 2 × 300 bp paired-end sequence reads which roughly corresponds to 15 GB of sequence data to be processed. Other platforms like NextSeq, which is useful in projects where the full V gene is not needed, creates about 400 million 2 × 150 bp paired-end reads. Because of the size of the datasets, the analysis can be computationally expensive, particularly the early analysis steps like preprocessing and gene annotation that process the majority of the sequence data. A standard desktop PC may take 3–5 days of constant processing for a single MiSeq run, so dedicated high-performance computational resources may be required. The institution may provide a cluster with high-performance computers for running analysis jobs. Commercial services like Amazon Web Services or Google Cloud can provide access to compute resources. However, this may come at added costs and could carry with them privacy concerns. Alternatively, there are free computing resources available. For AIRR-seq data, VDJServer provides free access to high-performance computing at the Texas Advanced Computing Center (TACC) through a graphical user interface

[18]. VDJServer has also parallelized execution for tools such as IgBLAST, so more compute resources are utilized as the size of the input data grows. Analysis that takes days on a desktop PC might take only a few hours on VDJServer. An example workflow is provided in the AIRR Community Chap. 22 with instructions about using VDJServer for immune repertoire analysis.

2.2 Software Tools

Many tools are available for the first steps in AIRR-seq analysis [28–31]. Table 1 highlights several of the more commonly used programs. These are noted particularly because they support standardized AIRR data representations and are mostly free and open source, two key criteria among the AIRR software guidelines (https://docs.airr-community.org/en/stable/swtools/airr_swtools_standard.html). When deciding what are the right software tools to analyze data, besides computational requirements and expertise of the user, we recommend taking into consideration whether these tools use the AIRR Community standards and are AIRR-compliant. Tools that use the standard can easily be incorporated into complex workflows with other tools that share the same data format. Selecting AIRR-compliant software adds an additional layer of transparency to the analysis, because the source code is (1) available for inspection on a publicly available repository, (2) uses a versioning system, (3) has been tested, and (4) is available as a container (Docker, Singularity), among other quality requirements. The use of AIRR standards and of AIRR-compliant software supports the transparency, reproducibility, and rigor of research results.

2.3 Germline Databases

IG and TR germline databases are a requirement for accurate AIRR-seq analyses, regardless of the technique used (e.g., single cell vs. bulk). These databases guide the assignment of sequences to known and novel IG and TR genes/alleles, facilitating downstream sequence annotation and the accurate assessment of various repertoire features (e.g., gene/allele usage, SHM, clonal assignment, etc.; see AIRR Community Chaps. 18–20 for more detail). A germline database should ideally contain the most comprehensive and accurate set of possible IG/TR V, D, and J genes and alleles that best represent the genomic content of an organism. There are various sources of reference germline databases available, and occasionally a tool is limited by which database can be used for a particular analysis. Thus, the use of a particular database, or a combination of databases, may vary depending on the experimental objectives, as well as the particular species in which the AIRR-seq data has been generated. We therefore recommend investing effort in obtaining as accurate a database as possible. Table 2 describes currently available databases, focusing on those that are in active development.

Table 2
Germline reference databases

Database	Description	Website	
Open Germline Receptor Database (OGRDB)	Curated high-quality alleles inferred from AIRR-seq data. Currently only human IG	https://ogrdb.airr-community.org	[32]
IMGT/GENE-DB	IG and TR for a wide range of species	http://www.imgt.org/vquest/refseqh.html	[33]
10X Genomics Germline Reference database	Human and mouse IG and TR, derived from Ensembl	https://support.10xgenomics.com/single-cell-vdj/software/downloads/latest	
MiXCR built-in reference	Human and mouse IG and TR; rat TR only, derived from Genbank	https://github.com/repseqio/library/	[1]
VDJBase	Genotype and haplotype data inferred from human AIRR-seq datasets. Currently IG only, planned expansion to other species and loci in 2021	https://www.vdjbase.org	[34]

IMGT [2] provides the most commonly used reference genome databases, but even for species of substantial research interest, these do not represent species diversity and can contain sequences reported in error [35, 36]. For TR genes and for IG genes from nonhuman species, however, few or no satisfactory alternatives exist. Ongoing initiatives seek to remedy this by continuously improving germline databases across species. Several programs are available to infer personalized databases from AIRR-seq data for each experimental subject (Table 1). VDJbase (https://www.vdjbase.org) is a resource that brings together AIRR-seq and genomic information to study population diversity and identify previously unreported alleles [34]. In 2019, the AIRR Community established the IARC (Inferred Allele Review Committee) to evaluate, document, and name human IGH alleles inferred from AIRR-seq data [37], and it is anticipated that this approach will be extended to other species and loci over time: The IARC's work is supported and published by OGRDB (the Open Germline Receptor Database, https://ogrdb.airr-community.org), which provides full information regarding alleles, metadata on the repertoires from which they originated, and ref. 32.

3 Methods

Preprocessing and gene annotation of AIRR-seq data takes as input the sequencing files and returns a set of high-quality sequences for which V, D, and J allele calls can be made and structural elements can be identified. After further quality control filtering steps, a final set of sequences is selected and can be used to carry out more in-depth analyses (*see* Chap. 17). All steps should be carefully documented to maintain data provenance and allow the analysis to be reproduced; the AIRR Community has defined a set of MiAIRR data processing fields to standardize the representation of analysis steps [38]. Below, we outline the concepts involved in each phase of analysis and then supply detailed protocols, applying them to common use cases. We also provide further information on reporting and sharing AIRR-seq data.

3.1 Preprocessing

While there are several experimental technologies available for AIRR-seq studies from different experimental setups, most approaches typically produce the same raw data file format (.fastq) and share the ultimate goal of obtaining a final set of reads of high quality, particularly in the complementarity-determining region 3 (CDR3) region, representative of each B or T cell in the repertoire. The general steps that need to be performed include (1) filtering reads (e.g., removing PhiX spike-ins, short reads, and reads with a low Phred score or excessive ambiguous base calls), (2) identifying and removing primers and sequencing barcodes (if present), (3) building consensus sequences (using UMI or cell barcodes, if present), (4) merging mate pairs (if using a paired-end protocol), (5) masking low-quality positions, (6) annotating with constant (C) region (if present), and (7) collapsing duplicate sequences. For some of these steps, some considerations and adjustments need to be made depending on whether the data are from genomic DNA or RNA, B cells or T cells; bulk or single cell, paired or unpaired chains, and whether UMIs have been used (Fig. 1).

In the following we describe the important considerations to be made when preprocessing AIRR-seq samples.

3.1.1 Filtering by Sequence or by Clone

Current NGS methods introduce occasional base-call errors which may not be detectable from the associated quality scores. A common approach to avoid incorporating these sequences in downstream analyses is to threshold data based on the frequency of reads. This does not eliminate such errors but can reduce their influence on gross metrics of the underlying immune repertoire. To remove spurious sequences, a common approach taken, e.g., by MiXCR [1] and SONAR [12], is to collapse identical or near-identical sequences and drop those with fewer than a specified number of reads (usually two or three). This approach is preferred where

individual sequences may be of low quality, for instance, if sequencing depth is low. However, this approach to filtering can result in nonuniform loss of data when libraries of different sequencing depths are compared. Alternatively, instead of a preprocessing step, all sequences passing quality control checks can be grouped into clones using the regular workflows described in the AIRR Community method Chaps. 18 and 19, and then clones that include fewer than the specified number of unique sequences are removed prior to downstream analysis. This may be appropriate for high-quality sequences, such as with UMIs and sufficient sequencing depth for robust error correction. Without this correction, errors in the CDR3 can lead to the inference of spurious clones.

3.1.2 Read Length-Related Effects

Long paired-end reads provide useful information for reliable V gene assignment as well as more comprehensive mapping of SHM in the case of IG gene rearrangements [39]. As read length increases, the quality of base calls degrades as sequences are generated, but paired-end sequencing allows for computational alignment of the overlapping regions. After alignment, sequencing errors at the ends of the sequences can be reduced as the higher-quality base call for each position that overlaps can be used. However, for longer sequences such as with RNA libraries capturing the constant region, the read length on the sequencer may need to be increased, reducing the overlapping portion of the 5′ and 3′ reads, resulting in a bias against sequences encoding longer CDR3. Further complicating this issue, a common procedure is to trim the ends of reads of low-quality stretches of base calls, such as with generic tools like fastx-toolkit or pRESTO's `FilterSeq trimqual` [4]. This can in turn reduce the number of full-length high-quality sequences. On the other hand, with RNA-based sequencing, UMIs can be incorporated at the cDNA synthesis step, and, when coupled with very deep sequencing, these can be used for error correction through the construction of consensus sequences that share the same UMI. There is, however, a trade-off between the sequencing depth required for adequate coverage of UMIs and the number of independent sequences that can be sampled.

Long reads covering the entire variable region can also be generated using alternative sequencing platforms, such as those offered by Pacific Biosciences and Ion Torrent [31, 40–43]. These offer the additional advantage of being able to capture large enough parts of the C-region to be able to distinguish between subtypes of IgG. However, lower throughput on these platforms limits the depth of sampling that can be achieved.

Short reads are sometimes used to generate large quantities of data on CDR3 sequences, as sequencing short reads can be done on higher-throughput sequencers at lower cost. This strategy is particularly common for TR rearrangement analysis on gDNA using

commercial platforms such as Adaptive. Short reads may be required if the template is of low quality, as sometimes occurs in formalin-fixed paraffin-embedded samples. Short reads can sometimes compromise TRBV gene assignments but are particularly problematic for IGH gene rearrangements with SHM. Short IGHV gene sequences result in larger numbers of ambiguous V gene assignments which can cause erroneous clustering of unrelated sequences into clones.

gDNA vs. *mRNA templates*. When using genomic DNA as starting material, each cell contributes a fixed number of IG or TR template, providing a parsimonious and cost-effective means of profiling large numbers of cells. gDNA-based sequencing will also capture far more nonproductive gene rearrangements than mRNA-based sequencing. With RNA, nonproductive rearrangements are subjected to nonsense-mediated degradation (although some nonproductive rearrangements can be recovered). gDNA is also more stable than RNA. On the other hand, RNA-based sequencing is more sensitive, with more templates per cell. With mRNA-based sequencing, cells contribute different numbers of templates, based upon cell subset-specific differences in transcript abundance. With mRNA-based libraries, cells can be grouped into subsets using immunophenotyping or single-cell RNA-seq to control for these differences. In the case of IG data where primers can be designed to capture the C-regions, each read can be annotated with its isotype using, for example, pRESTO's `MaskPrimers` routine. Further, unlike gDNA, it is straightforward to incorporate unique molecular identifiers (UMIs) at the RNA to cDNA synthesis step. Each UMI, which should be unique to original individual cDNA templates, can be processed with pRESTO's `BuildConsensus` to generate consensus sequences which can nearly eliminate sequencing error given sufficient sequencing depth [44, 45]. MiXCR, SONAR, and other packages also offer similar tools. The necessary depth might be difficult to achieve, though, for instance, in cases of vastly different expression levels or with samples of large size.

3.1.3 Productive Vs. Nonproductive Rearrangements

For each sample, the *fraction of productive rearrangements* can be an informative metric. On average, it can be expected that approximately 80% of TRB rearrangements and approximately 85% of IGH rearrangements sequenced from mature T or B cells will be productive [46]. Lower frequencies of productive rearrangements can be observed in immature lymphocytes, where selection has not yet been imposed on cells without productive rearrangements [47]. Lower frequencies of productive rearrangements can also be seen in sequencing libraries that are of poor quality. Nonproductive sequences also can be used as a baseline estimator of gene usage frequency in rearrangement [48, 49] and compared to productive

sequences to investigate the effects of tolerance checkpoints on the AIRR [50, 51]. With such comparisons, it may be useful to remove clonal lineages that contain both productive and nonproductive versions of the same rearrangement, as sequencing errors can cause a sequence to appear nonproductive. Nonproductive rearrangements are sometimes also useful for identifying clonal expansions in tumors, particularly if tumors harbor SHM that may interfere with primer binding (the nonproductive rearrangements are usually un-mutated). Nonproductive rearrangements can be found in lymphocytes that have undergone multiple rounds of V(D)J recombination, as can occur with receptor editing; the presence of more than one rearrangement is particularly common with IG light chains [52, 53]. Finally, it is important to computationally filter nonproductive sequences for general analyses, if one is making claims about selected repertoires.

3.2 Gene Annotation

After preprocessing AIRR sequences for good-quality and relevant reads, sequences need to be accurately aligned and annotated to an appropriate reference germline database. This process identifies the V, D, and J genes; CDRs; and framework regions (FWRs) for each sequence in the repertoire. There are numerous annotation tools for IG and TR sequences that are freely available to users, including popular programs such as IgBLAST [10] and IMGT/HighV-QUEST (Table 1) [8]. Depending on the tools, different tool-specific algorithms (e.g., Smith-Waterman) assign the best match among a set of genes in a user-defined reference germline database. Accurate alignment is very important for subsequent analyses such as the identification of SHM for IGs, clustering of clonal groups, and determination of IG/TR diversity. Alignment algorithms have been demonstrated to influence the outcome of V, D, and J gene assignments, even when identical input sequences, tool parameters, and reference germline databases are chosen [31]. Furthermore, differences in the length of alleles of genes in databases may force algorithms to output an incorrect best match in the gene annotation process. To complicate matters, some tools provide alignments to multiple (often highly similar) genes and leave it to users to choose which of the ambiguous calls is most appropriate.

Schemes for IGs and TRs that number amino acid residues facilitate sequence comparisons, protein structure modelling, and engineering [54]. Although many schemes have been proposed and different schemes are employed by different tools, only five schemes are commonly used. Three are specifically for IGs: Kabat [55], Chothia [56], and enhanced Chothia [57]. Two more can be used for both IGs and TRs: IMGT [58] and AHo [59]. Conversion tables and tools like ANARCI [60] can be used to translate between schemes. CDR boundaries can differ substantially between different numbering schemes: care is needed when comparing results

from different studies [54]. In repertoire studies, the IMGT numbering scheme is widely used and supported, and its use is recommended in the absence of other considerations.

One more barrier to direct comparison is the identification in some studies and tools of the "junction" and in others of the CDR3. In IMGT terminology, the junction includes the second conserved cysteine of the V gene and the conserved tryptophan or phenylalanine of the J gene, while the CDR3 omits these residues. The AIRR Community data representation standard uses "junction"; however, it is not universally accepted [31].

Accurate annotation requires an accurate and comprehensive germline database. As noted above, even the currently available human database does not as yet meet this criterion [15, 61], and databases for other species are often partial and based solely on the analysis of a single animal [36, 62–65]. Fortunately, scientific need has resulted in the determination of new germline gene sets [36, 40, 66, 67], but these are not necessarily implemented by public germline gene databases in a timely fashion. The impact of missing or incorrect information in the database will depend upon the nature of the analysis, but one overall point to note is that the databases are updated frequently, and changes in the database can impact results [31]. It is therefore important that an analysis is conducted using a single, consistent, and up-to-date version of the database and that the version (or download date) is recorded for reproducibility. Germline databases are sometimes installed automatically with annotation tools: where that is the case, researchers should check if the installed version meets these requirements, and update it if necessary.

In a repertoire from a single individual, although structural variation and gene duplication give rise to frequent exceptions, we would expect to see a maximum of two alleles of most germline receptor genes: one from the paternal and one from the maternal chromosome. When used with an extensive germline database, annotation tools that are based on sequence similarity tend to call a biologically implausible number of alleles in B-cell repertoires, particularly in repertoires that are highly mutated, and will make a large number of indeterminate calls, where the tool would be unable to determine the likely germline allele unambiguously. Tools are available that will improve allele calls by using probabilistic methods to infer the individual's "personalized" germline set: such tools can also infer the presence of alleles in the individual that were not listed in the annotation tool's germline database [15–17, 68, 69]. While the use of a comprehensive germline database is important in the first instance, the determination of a personalized germline set and re-annotation with just that set is recommended where allele assignment is important: for example, when clonal inference is employed: personalization can also compensate to some extent for deficiencies in the germline database.

The decision of which annotation tool to use is also dependent on the computer skill set of the user. IMGT/HIGHV-QUEST and IgBLAST provide easy-to-use web platforms, suited for researchers that prefer to access a graphic user interface. Other tools, such as the stand-alone version of IgBLAST [10], MiXCR [1], and partis [11], require additional computer expertise, because they need to be installed and are used in the terminal. The advantage of such tools is that they provide more flexibility and can be integrated in automated workflows.

4 Conclusion

In this chapter, we present important considerations involved in the first steps in the preparation of raw data after sequencing and guide bioinformaticians in choosing the appropriate parameters for preprocessing and annotation. These first steps are required for the subsequent repertoire analysis, described in the Chap. 17, as choices made in these first steps have serious implications for the types of data analyses that can be performed and for the accuracy of the results. After the completion of this chapter, the bioinformatician is now ready to begin the in-depth analysis of repertoire features specific to the question at hand.

Acknowledgments

The authors would like to thank Eline T. Luning Prak for the constructive criticism of the manuscript.

References

1. Bolotin DA, Poslavsky S, Mitrophanov I, Shugay M, Mamedov IZ, Putintseva EV et al (2015) MiXCR: software for comprehensive adaptive immunity profiling. Nat Methods 12: 380–381. https://doi.org/10.1038/nmeth.3364
2. Giudicelli V, Brochet X, Lefranc M-P (2011) IMGT/V-QUEST: IMGT standardized analysis of the immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences. Cold Spring Harb Protoc 2011:695–715. https://doi.org/10.1101/pdb.prot5633
3. Gupta NT, Vander Heiden JA, Uduman M, Gadala-Maria D, Yaari G, Kleinstein SH (2015) Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. Bioinformatics 31: 3356–3358. https://doi.org/10.1093/bioinformatics/btv359
4. Vander Heiden JA, Yaari G, Uduman M, Stern JNH, O'Connor KC, Hafler DA et al (2014) pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. Bioinformatics 30: 1930–1932. https://doi.org/10.1093/bioinformatics/btu138
5. Stubbington MJT, Lönnberg T, Proserpio V, Clare S, Speak AO, Dougan G et al (2016) T cell fate and clonality inference from single-cell transcriptomes. Nat Methods 13:329–332. https://doi.org/10.1038/nmeth.3800
6. Christley S, Levin MK, Toby IT, Fonner JM, Monson NL, Rounds WH et al (2017) VDJPipe: a pipelined tool for pre-processing immune repertoire sequencing data. BMC Bioinformatics 18:448. https://doi.org/10.1186/s12859-017-1853-z

7. Peacock T, Heather JM, Ronel T, Chain B (2020) Decombinator V4 - an improved AIRR-compliant software package for T cell receptor sequence annotation. Bioinformatics 37(6):876–878. https://doi.org/10.1093/bioinformatics/btaa758
8. Alamyar E, Duroux P, Lefranc M-P, Giudicelli V (2012) IMGT® tools for the nucleotide analysis of immunoglobulin (IG) and T cell receptor (TR) V-(D)-J repertoires, polymorphisms, and IG mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. In: Christiansen FT, Tait BD (eds) Immunogenetics. Humana Press, Totowa, NJ, pp 569–604. https://doi.org/10.1007/978-1-61779-842-9_32
9. Kuchenbecker L, Nienen M, Hecht J, Neumann AU, Babel N, Reinert K et al (2015) IMSEQ—a fast and error aware approach to immunogenetic sequence analysis. Bioinformatics 31:2963–2971. https://doi.org/10.1093/bioinformatics/btv309
10. Ye J, Ma N, Madden TL, Ostell JM (2013) IgBLAST: an immunoglobulin variable domain sequence analysis tool. Nucleic Acids Res 41:W34–W40. https://doi.org/10.1093/nar/gkt382
11. Ralph DK, Matsen FA (2016) Consistency of VDJ rearrangement and substitution parameters enables accurate B cell receptor sequence annotation. PLoS Comput Biol 12:e1004409. https://doi.org/10.1371/journal.pcbi.1004409
12. Schramm CA, Sheng Z, Zhang Z, Mascola JR, Kwong PD, Shapiro L (2016) SONAR: a high-throughput pipeline for inferring antibody ontogenies from longitudinal sequencing of B cell transcripts. Front Immunol 7:372. https://doi.org/10.3389/fimmu.2016.00372
13. Giraud M, Salson M, Duez M, Villenet C, Quief S, Caillault A et al (2014) Fast multiclonal clusterization of V(D)J recombinations from high-throughput sequencing. BMC Genomics 15:409. https://doi.org/10.1186/1471-2164-15-409
14. Duez M, Giraud M, Herbert R, Rocher T, Salson M, Thonier F (2016) Vidjil: a web platform for analysis of high-throughput repertoire sequencing. PLoS One 11:e0166126. https://doi.org/10.1371/journal.pone.0166126
15. Gadala-Maria D, Yaari G, Uduman M, Kleinstein SH (2015) Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. Proc Natl Acad Sci U S A 112:E862–E870. https://doi.org/10.1073/pnas.1417683112
16. Ralph DK, Matsen FA (2019) Per-sample immunoglobulin germline inference from B cell receptor deep sequencing data. PLoS Comput Biol 15:e1007133. https://doi.org/10.1371/journal.pcbi.1007133
17. Corcoran MM, Phad GE, Bernat NV, Stahl-Hennig C, Sumida N, Persson MAA et al (2016) Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. Nat Commun 7:13642. https://doi.org/10.1038/ncomms13642
18. Christley S, Scarborough W, Salinas E, Rounds WH, Toby IT, Fonner JM et al (2018) VDJServer: a cloud-based analysis portal and data commons for immune repertoire sequences and rearrangements. Front Immunol 9:976. https://doi.org/10.3389/fimmu.2018.00976
19. Rosenfeld AM, Meng W, Luning Prak ET, Hershberg U (2018) ImmuneDB, a novel tool for the analysis, storage, and dissemination of immune repertoire sequencing data. Front Immunol 9:2107. https://doi.org/10.3389/fimmu.2018.02107
20. Barennes P, Quiniou V, Shugay M, Egorov ES, Davydov AN, Chudakov DM et al (2021) Benchmarking of T cell receptor repertoire profiling methods reveals large systematic biases. Nat Biotechnol 39:236–245. https://doi.org/10.1038/s41587-020-0656-3
21. Greiff V, Miho E, Menzel U, Reddy ST (2015) Bioinformatic and statistical analysis of adaptive immune repertoires. Trends Immunol 36:738–749. https://doi.org/10.1016/j.it.2015.09.006
22. Khan TA, Friedensohn S, Gorter de Vries AR, Straszewski J, Ruscheweyh H-J, Reddy ST (2016) Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting. Sci Adv 2:e1501371. https://doi.org/10.1126/sciadv.1501371
23. Schwaber J, Andersen S, Nielsen L (2019) Shedding light: the importance of reverse transcription efficiency standards in data interpretation. Biomol Detect Quantif 17:100077. https://doi.org/10.1016/j.bdq.2018.12.002
24. Zucha D, Androvic P, Kubista M, Valihrach L (2020) Performance comparison of reverse transcriptases for single-cell studies. Clin Chem 66:217–228. https://doi.org/10.1373/clinchem.2019.307835
25. van Dongen JJM, Langerak AW, Brüggemann M, Evans PAS, Hummel M, Lavender FL et al (2003) Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect

lymphoproliferations: report of the BIOMED-2 concerted action BMH4-CT98-3936. Leukemia 17:2257–2317. https://doi.org/10.1038/sj.leu.2403202

26. Douek DC, Betts MR, Brenchley JM, Hill BJ, Ambrozak DR, Ngai K-L et al (2002) A novel approach to the analysis of specificity, clonality, and frequency of HIV-specific T cell responses reveals a potential mechanism for control of viral escape. J Immunol 168:3099–3104. https://doi.org/10.4049/jimmunol.168.6.3099
27. Breden F, Luning Prak ET, Peters B, Rubelt F, Schramm CA, Busse CE et al (2017) Reproducibility and reuse of adaptive immune receptor repertoire data. Front Immunol 8:1418. https://doi.org/10.3389/fimmu.2017.01418
28. Zhang Y, Yang X, Zhang Y, Zhang Y, Wang M, Ou JX et al (2020) Tools for fundamental analysis functions of TCR repertoires: a systematic comparison. Brief Bioinform 21:1706–1716. https://doi.org/10.1093/bib/bbz092
29. López-Santibáñez-Jácome L, Avendaño-Vázquez SE, Flores-Jasso CF (2019) The pipeline repertoire for Ig-seq analysis. Front Immunol 10:899. https://doi.org/10.3389/fimmu.2019.00899
30. Lees WD (2020) Tools for adaptive immune receptor repertoire sequencing. Curr Opin Syst Biol 24:86–92. https://doi.org/10.1016/j.coisb.2020.10.003
31. Smakaj E, Babrak L, Ohlin M, Shugay M, Briney B, Tosoni D et al (2020) Benchmarking immunoinformatic tools for the analysis of antibody repertoire sequences. Bioinformatics 36:1731–1739. https://doi.org/10.1093/bioinformatics/btz845
32. Lees W, Busse CE, Corcoran M, Ohlin M, Scheepers C, Matsen FA et al (2020) OGRDB: a reference database of inferred immune receptor genes. Nucleic Acids Res 48:D964–D970. https://doi.org/10.1093/nar/gkz822
33. Giudicelli V, Chaume D, Lefranc M-P (2005) IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. Nucleic Acids Res 33:D256–D261. https://doi.org/10.1093/nar/gki010
34. Omer A, Shemesh O, Peres A, Polak P, Shepherd AJ, Watson CT et al (2020) VDJbase: an adaptive immune receptor genotype and haplotype database. Nucleic Acids Res 48:D1051–D1056. https://doi.org/10.1093/nar/gkz872
35. Wang Y, Jackson KJL, Sewell WA, Collins AM (2008) Many human immunoglobulin heavy-chain IGHV gene polymorphisms have been reported in error. Immunol Cell Biol 86:111–115. https://doi.org/10.1038/sj.icb.7100144
36. Vázquez Bernat N, Corcoran M, Nowak I, Kaduk M, Castro Dopico X, Narang S et al (2021) Rhesus and cynomolgus macaque immunoglobulin heavy-chain genotyping yields comprehensive databases of germline VDJ alleles. Immunity 54:355–366.e4. https://doi.org/10.1016/j.immuni.2020.12.018
37. Ohlin M, Scheepers C, Corcoran M, Lees WD, Busse CE, Bagnara D et al (2019) Inferred allelic variants of immunoglobulin receptor genes: a system for their evaluation, documentation, and naming. Front Immunol 10:435. https://doi.org/10.3389/fimmu.2019.00435
38. Rubelt F, Busse CE, Bukhari SAC, Bürckert J-P, Mariotti-Ferrandiz E, Cowell LG et al (2017) Adaptive immune receptor repertoire community recommendations for sharing immune-repertoire sequencing data. Nat Immunol 18:1274–1278. https://doi.org/10.1038/ni.3873
39. Zhang B, Meng W, Luning Prak ET, Hershberg U (2015) Discrimination of germline V genes at different sequencing lengths and mutational burdens: a new tool for identifying and evaluating the reliability of V gene assignment. J Immunol Methods 427:105–116. https://doi.org/10.1016/j.jim.2015.10.009
40. Watson CT, Kos JT, Gibson WS, Newman L, Deikus G, Busse CE et al (2019) A comparison of immunoglobulin IGHV, IGHD and IGHJ genes in wild-derived and classical inbred mouse strains. Immunol Cell Biol 97:888–901. https://doi.org/10.1111/imcb.12288
41. Deiss TC, Vadnais M, Wang F, Chen PL, Torkamani A, Mwangi W et al (2019) Immunogenetic factors driving formation of ultralong VH CDR3 in Bos taurus antibodies. Cell Mol Immunol 16:53–64. https://doi.org/10.1038/cmi.2017.117
42. Koning MT, Kiełbasa SM, Boersma V, Buermans HPJ, van der Zeeuw SAJ, van Bergen CAM et al (2017) ARTISAN PCR: rapid identification of full-length immunoglobulin rearrangements without primer binding bias. Br J Haematol 178:983–986. https://doi.org/10.1111/bjh.14180

43. Lay L, Stroup B, Payton JE (2020) Validation and interpretation of IGH and TCR clonality testing by ion torrent S5 NGS for diagnosis and disease monitoring in B and T cell cancers. Pract Lab Med 22:e00191. https://doi.org/10.1016/j.plabm.2020.e00191
44. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B (2011) Detection and quantification of rare mutations with massively parallel sequencing. Proc Natl Acad Sci U S A 108: 9530–9535. https://doi.org/10.1073/pnas.1105422108
45. Shugay M, Britanova OV, Merzlyak EM, Turchaninova MA, Mamedov IZ, Tuganbaev TR et al (2014) Towards error-free profiling of immune repertoires. Nat Methods 11: 653–655. https://doi.org/10.1038/nmeth.2960
46. Rubelt F, Bolen CR, McGuire HM, Heiden JAV, Gadala-Maria D, Levin M et al (2016) Individual heritable differences result in unique cell lymphocyte receptor repertoires of naïve and antigen-experienced cells. Nat Commun 7:11112. https://doi.org/10.1038/ncomms11112
47. Meng W, Yunk L, Wang L-S, Maganty A, Xue E, Cohen PL et al (2011) Selection of individual VH genes occurs at the pro-B to pre-B cell transition. J Immunol 187: 1835–1844. https://doi.org/10.4049/jimmunol.1100207
48. Marcou Q, Mora T, Walczak AM (2018) High-throughput immune repertoire analysis with IGoR. Nat Commun 9:561. https://doi.org/10.1038/s41467-018-02832-w
49. Sethna Z, Elhanati Y, Callan CG, Walczak AM, Mora T (2019) OLGA: fast computation of generation probabilities of B- and T-cell receptor amino acid sequences and motifs. Bioinformatics 35:2974–2981. https://doi.org/10.1093/bioinformatics/btz035
50. Elhanati Y, Sethna Z, Marcou Q, Callan CG, Mora T, Walczak AM (2015) Inferring processes underlying B-cell repertoire diversity. Philos Trans R Soc Lond Ser B Biol Sci 370: 20140243. https://doi.org/10.1098/rstb.2014.0243
51. Sethna Z, Isacchini G, Dupic T, Mora T, Walczak AM, Elhanati Y (2020) Population variability in the generation and selection of T-cell repertoires. PLoS Comput Biol 16:e1008394. https://doi.org/10.1371/journal.pcbi.1008394
52. Langerak AW, van Dongen JJM (2012) Multiple clonal Ig/TCR products: implications for interpretation of clonality findings. J Hematop 5:35–43. https://doi.org/10.1007/s12308-011-0129-1
53. Luning Prak ET, Monestier M, Eisenberg RA (2011) B cell receptor editing in tolerance and autoimmunity. Ann N Y Acad Sci 1217: 96–121. https://doi.org/10.1111/j.1749-6632.2010.05877.x
54. Dondelinger M, Filée P, Sauvage E, Quinting B, Muyldermans S, Galleni M et al (2018) Understanding the significance and implications of antibody numbering and antigen-binding surface/residue definition. Front Immunol 9:2278. https://doi.org/10.3389/fimmu.2018.02278
55. Wu TT, Kabat EA (1970) An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. J Exp Med 132:211–250. https://doi.org/10.1084/jem.132.2.211
56. Al-Lazikani B, Lesk AM, Chothia C (1997) Standard conformations for the canonical structures of immunoglobulins. J Mol Biol 273:927–948. https://doi.org/10.1006/jmbi.1997.1354
57. Abhinandan KR, Martin ACR (2010) Analysis and prediction of VH/VL packing in antibodies. Protein Eng Des Sel 23:689–697. https://doi.org/10.1093/protein/gzq043
58. Lefranc M-P, Pommié C, Ruiz M, Giudicelli V, Foulquier E, Truong L et al (2003) IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. Dev Comp Immunol 27: 55–77. https://doi.org/10.1016/s0145-305x(02)00039-3
59. Honegger A, Plückthun A (2001) Yet another numbering scheme for immunoglobulin variable domains: an automatic modeling and analysis tool. J Mol Biol 309:657–670. https://doi.org/10.1006/jmbi.2001.4662
60. Dunbar J, Deane CM (2016) ANARCI: antigen receptor numbering and receptor classification. Bioinformatics 32:298–300. https://doi.org/10.1093/bioinformatics/btv552
61. Watson CT, Breden F (2012) The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease. Genes Immun 13:363–373. https://doi.org/10.1038/gene.2012.12
62. Ramesh A, Darko S, Hua A, Overman G, Ransier A, Francica JR et al (2017) Structure and diversity of the rhesus macaque immunoglobulin loci through multiple de novo genome assemblies. Front Immunol 8:1407. https://doi.org/10.3389/fimmu.2017.01407
63. Cirelli KM, Carnathan DG, Nogal B, Martin JT, Rodriguez OL, Upadhyay AA et al (2019)

Slow delivery immunization enhances HIV neutralizing antibody and germinal center responses via modulation of immunodominance. Cell 177:1153–1171.e28. https://doi.org/10.1016/j.cell.2019.04.012

64. Retter I, Chevillard C, Scharfe M, Conrad A, Hafner M, Im T-H et al (2007) Sequence and characterization of the Ig heavy chain constant and partial variable region of the mouse strain 129S1. J Immunol 179:2419–2427. https://doi.org/10.4049/jimmunol.179.4.2419
65. Collins AM, Wang Y, Roskin KM, Marquis CP, Jackson KJL (2015) The mouse antibody heavy chain repertoire is germline-focused and highly variable between inbred strains. Philos Trans R Soc Lond Ser B Biol Sci 370:20140236. https://doi.org/10.1098/rstb.2014.0236
66. Magadan S, Krasnov A, Hadi-Saljoqi S, Afanasyev S, Mondot S, Lallias D et al (2019) Standardized IMGT® nomenclature of Salmonidae IGH genes, the paradigm of Atlantic Salmon and rainbow trout: from genomics to repertoires. Front Immunol 10:2541. https://doi.org/10.3389/fimmu.2019.02541
67. Magadan S, Mondot S, Palti Y, Gao G, Lefranc MP, Boudinot P (2021) Genomic analysis of a second rainbow trout line (Arlee) leads to an extended description of the IGH VDJ gene repertoire. Dev Comp Immunol 118:103998. https://doi.org/10.1016/j.dci.2021.103998
68. Zhang W, Wang I-M, Wang C, Lin L, Chai X, Wu J et al (2016) IMPre: an accurate and efficient software for prediction of T- and B-cell receptor germline genes and alleles from rearranged repertoire data. Front Immunol 7:457. https://doi.org/10.3389/fimmu.2016.00457
69. Gadala-Maria D, Gidoni M, Marquez S, Vander Heiden JA, Kos JT, Watson CT et al (2019) Identification of subject-specific immunoglobulin alleles from expressed repertoire sequencing data. Front Immunol 10:129. https://doi.org/10.3389/fimmu.2019.00129

Chapter 17

Adaptive Immune Receptor Repertoire (AIRR) Community Guide to Repertoire Analysis

Susanna Marquez, Lmar Babrak, Victor Greiff, Kenneth B. Hoehn, William D. Lees, Eline T. Luning Prak, Enkelejda Miho, Aaron M. Rosenfeld, Chaim A. Schramm, and Ulrik Stervbo and on behalf of the AIRR Community

Abstract

Adaptive immune receptor repertoires (AIRRs) are rich with information that can be mined for insights into the workings of the immune system. Gene usage, CDR3 properties, clonal lineage structure, and sequence diversity are all capable of revealing the dynamic immune response to perturbation by disease, vaccination, or other interventions. Here we focus on a conceptual introduction to the many aspects of repertoire analysis and orient the reader toward the uses and advantages of each. Along the way, we note some of the many software tools that have been developed for these investigations and link the ideas discussed to chapters on methods provided elsewhere in this volume.

Key words AIRR-seq, B-cell receptor, T-cell receptor, Analysis, Clonal structure

1 Introduction

Once an adaptive immune receptor repertoire (AIRR) experiment has been carried out and the data has been appropriately preprocessed and annotated (see chapter "AIRR Community Guide to TR and IG Gene Annotation"), the next step is to plan a course of analysis to answer the questions posed by the experiment. As AIRRs are complex datasets that can contain thousands or even millions of sequences, it is important to have a working familiarity with the type of information each analysis can provide, as well as the limitations of an analysis. Here we provide an introduction to a variety of widely used techniques and discuss their applicability. In other chapters in this volume, we provide detailed experimental protocols and instructions to perform such analyses for the purpose of addressing specific biological questions. For a definition of terms used

Anton W. Langerak (ed.), *Immunogenetics: Methods and Protocols*, Methods in Molecular Biology, vol. 2453, https://doi.org/10.1007/978-1-0716-2115-8_17,

throughout this chapter, please see the AIRR Community glossary of terms, available at https://zenodo.org/record/5095381.

2 Materials

A breathtaking array of computational tools are available for repertoire analysis. These range from bespoke command line tools written in various programming languages that require facility in a Linux terminal to software with fully developed graphical interfaces and no requirement for programming skills of any kind. Thus, a key factor in choosing which programs to use will be the skill level and comfort of the user. Moreover, most tools have a narrow scope of the types of analysis they can perform, so matching the implementation to the desired goal is also a critical consideration. In addition, thought must be given to the computational resources necessary for repertoire analysis, including both storage and processing.

A comprehensive listing of the available software is out of the scope of this conceptual introduction, but the interested reader is directed to some recent reviews [1–4]. Here we focus on a small selection of commonly used tools, especially those which comply with AIRR Community guidelines for reproducibility and interoperability (https://docs.airr-community.org/en/stable/swtools/airr_swtools_standard.html). These are highlighted in Table 1, and several are discussed in more detail below and in other chapters in this volume, where we demonstrate their application to common analytical tasks.

3 Methods

In this section we introduce some of the most frequently used methods to analyze AIRRs and suggest computational tools that can perform such analysis. Some of the methods are applicable to both IG and TR, and some are specific. In addition, the selection of the method and the interpretation of the results can depend on the specific biological state; for instance, some samples might be expanded from solid tumors, others from antigen-specific cells isolated from peripheral blood or from whole blood from healthy and diseased patients. The theoretical framework presented here can be used to interpret the results of the practical methods detailed in the AIRR Community chapters "Bulk gDNA Sequencing of Antibody Heavy Chain Gene Rearrangements for Detection and Analysis of B-Cell Clone Distribution," "Bulk Sequencing From mRNA With UMI for Evaluation of B-Cell Isotype and Clonal Evolution and Single-Cell Analysis," and "Tracking of Antigen-Specific T Cells: Integrating Paired-Chain AIRR-Seq and Transcriptome Sequencing," all in this volume.

Table 1
Software tools

Software	Notes/description	URL	
Analysis			
Abysis	IG annotation and analysis. Does not support AIRR data representations	http://www.abysis.org/abysis/	[5]
Alakazam	Find clonal lineages and analyze diversity, gene usage, and other repertoire level properties. Alakazam is part of the Immcantation suite	https://alakazam.readthedocs.io/en/stable/	[6]
Change-O	Data standardization, germline reconstruction and clonal assignment. Part of the Immcantation suite	https://changeo.readthedocs.io/en/stable/	[6]
Dowser	Build, visualize, and analyze IG lineage trees. Implements statistical tests for discrete trait analysis of B-cell migration, differentiation, and isotype switching	https://dowser.readthedocs.io/en/stable/	[7]
IGoR	Learn models for the generation of V(D)J rearrangements	https://github.com/qmarcou/IGoR/	[8]
IgPhyML	Build phylogenetic trees and test evolutionary hypotheses regarding B-cell affinity maturation	https://igphyml.readthedocs.io/en/stable/	[9, 10]
Immunarch	Integrated framework for analysis and visualization	https://immunarch.com	[11]
MiXCR	Find and analyze clonal lineages	https://mixcr.readthedocs.io/en/master/	[12]
OLGA	Calculate likelihood of generating a particular V(D)J sequences and simulate repertoires	https://github.com/statbiophys/OLGA/	[13]
Partis	Find clonal lineages, analyze selection, and simulate repertoires	https://github.com/psathyrella/partis/	[14, 15]
RAbHIT	Infer Ig haplotypes from AIRRseq data	https://yaarilab.bitbucket.io/RAbHIT/	[16]
Scirpy	An extension of scanpy for AIRR-seq data. Find and analyze clones, compare repertoires, and integrate transcriptomic data. Scirpy is certified as compliant with AIRR community software guidelines	https://github.com/icbi-lab/scirpy	[17]
SCOPer	Multiple methods for identifying B-cell clones, including spectral clustering. SCOPer is part of the Immcantation suite	https://scoper.readthedocs.io/en/stable/	[18]
SHazaM	Analysis of mutational load, SHM targeting, and selection pressure for IGs. SHazaM is part of the Immcantation suite	https://shazam.readthedocs.io/en/stable/	[6]

(continued)

Table 1
(continued)

Software	Notes/description	URL	
SONAR	Identification and longitudinal phylogenetic analysis of IG sequences clonally related to an antibody of interest. SONAR is certified as compliant with AIRR community software guidelines	https://github.com/scharch/SONAR/	[19]
Sumrep	Assess and compare repertoire properties	https://github.com/matsengrp/sumrep	[20]
TRIgS	Utilities for high-throughput identification and analysis of IG clones	http://cimm.ismb.lon.ac.uk/trigs/	[21]
Vidjil	Find clones and analyze repertoires. Available as both a web interface and a downloadable tool	http://www.vidjil.org	[22, 23]
Preprocessing, annotation, and analysis environments			
VDJServer	A free, scalable resource for performing immune repertoire analysis and sharing data	https://vdjserver.org	[24]
ImmuneDB	Database and analysis tool for large amounts of AIRR-seq data	https://immunedb.readthedocs.io/en/latest/	[25]

3.1 Gene Usage

The V gene is the most diverse gene of the TR and IG loci. This is driven especially by variation in the first and second complementarity-determining regions (CDR1 and CDR2) of the genes, which contribute to the specificity and affinity of the immune receptor. Differences in the distribution of V genes used in the rearranged repertoire can indicate an antigen-specific response or unusual clonal expansions and can be evaluated with the function `compareVGeneDistributions` of the sumrep R package [20] (https://github.com/matsengrp/sumrep). The D and J gene strongly contributes to the CDR3 and can be compared using `compareDGeneDistributions` and `compareJGeneDistributions`. Skewing of the V-J usage can be revealed by plotting the V-J combination as a heatmap (Fig. 1a). The distribution of V-J and V-D-J usage can be compared between two repertoires using the functions `compareVJDistributions` and `compareVDJDistributions` in sumrep.

3.2 Properties of the CDR3

The CDR3 is the most variable part of the rearranged IG/TR, and is a key contributor to the overall specificity of the receptor [26]. Therefore, analyzing the properties of this region is of great interest.

Due to the randomness in addition and deletion of nucleotides during the rearrangement of the receptor, CDR3 lengths will be distributed around a mean value (Fig. 1b). Any changes to this distribution signifies an expansion of cells with a particular immune receptor.

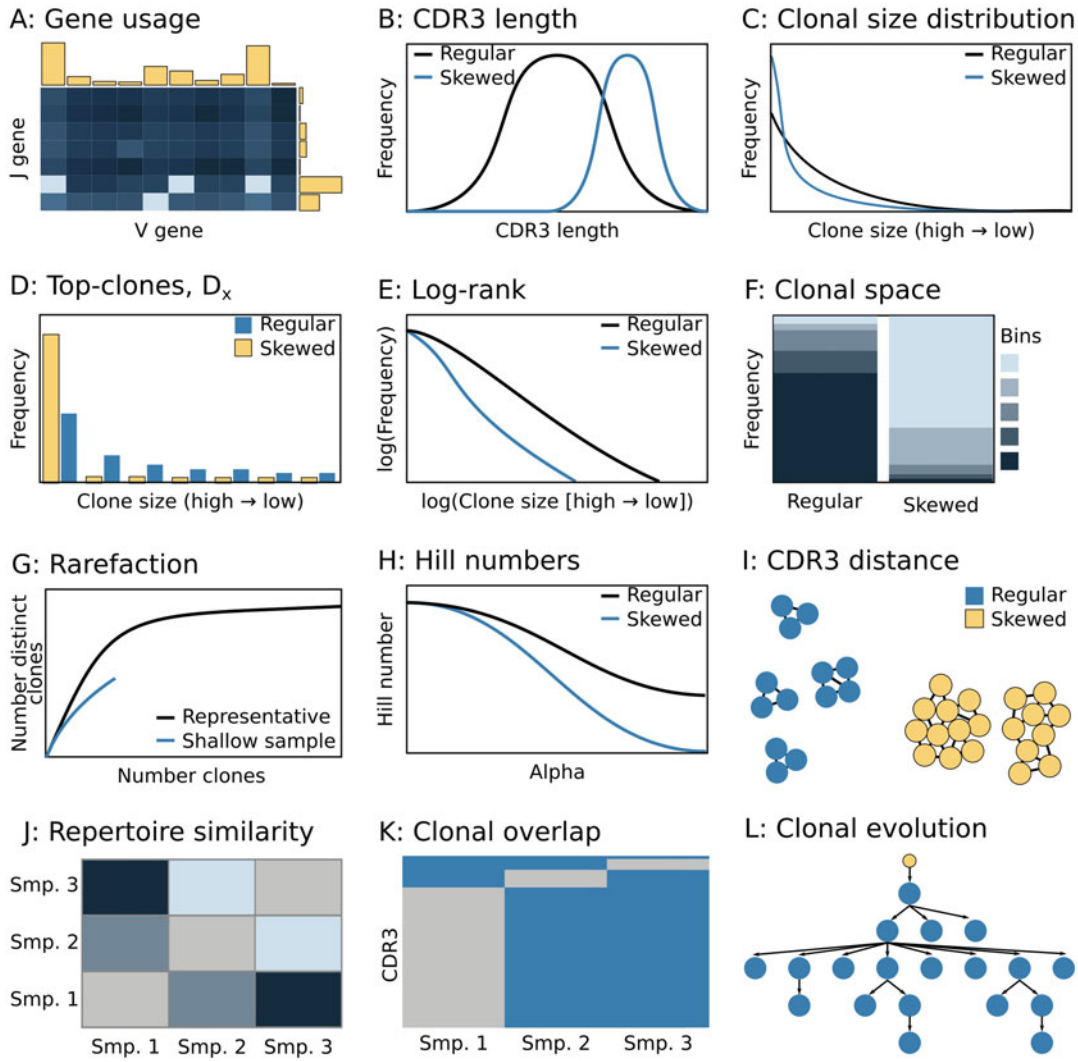


Fig. 1 Data visualizations. Examples of different data visualizations to gain insights into the AIRR. The plot title describes the basic analysis. Smp = sample. For further details, please refer to the main text

Different receptors specific for the same epitope can be expected to share motifs [27, 28]. Such motifs can be a few identical amino acids or amino acids with similar physical properties. Apart from properties like size, charge, and polarity, the properties of amino acids can be described by different factors derived through dimensionality reduction of a larger number of properties. Atchley [29] factors comprise five numerical descriptions, and Kidera [30] factors comprise ten numerical descriptions.

The R package sumrep [20] provides functions to compare the CDR3 properties of two repertoires, such as the CDR3 length and a number of amino acid physicochemical properties [31, 32].

3.3 Clonal Lineages

A clone or a clonal lineage comprises a group of T or B cells descended from the same original naive ancestor. As such, all cells in a clone contain the same set of rearrangements. An important part of AIRR-seq analysis is computationally reconstructing these relationships from the sequences obtained. For TRs, the exercise is relatively straightforward, as only PCR and sequencing error need to be accounted for. With IGs, however, somatic hypermutation can significantly obscure the ancestry of a particular sequence [33], and so more complex strategies are required (*see* Subheading 3.9.2).

When analyzing bulk AIRR-seq data, in which native pairing between heavy and light, alpha and beta, and gamma and delta chains is lost, clones are sometimes defined based on a single chain. This may be sufficient for IGH and TRB rearrangements, which are more diverse and contain most of the information needed to group sequences into clonal lineages [34]. However, care should still be taken in interpreting such data. Many different definitions of clonally related sequences have been offered in the literature (e.g., see the work by Kotouza and co-workers [35]), and methods to infer clones from AIRR-seq data are under active investigation [6, 12, 14, 18, 36].

The distribution of clone sizes in an AIRR can be informative of underlying biology. One visualization is to plot ranks from high to low on the x-axis and associated frequency on the y-axis (Fig. 1c) to reveal clonal expansion. A closer look at the top x (Fig. 1d) helps likewise to identify clonal expansion. When plotting the log of rank and the frequency (Fig. 1e), the slope reveals the distribution of clones, such that the steeper the slope, the less evenly distributed the repertoire. The function `estimateAbundance` in the R package Alakazam [6] can estimate clonal abundance with confidence intervals obtained by bootstrapping.

An alternative visualization makes use of division of clonal frequencies into different groups ("binning") and sums the frequencies in each bin (Fig. 1f). The binning is essentially arbitrary, but binning the clone frequencies into the bins [0.0, 0.001], [0.001, 0.01], [0.01, 0.1], and [0.1, 1] are widely used. Binning by rank is an alternative where the bins [1, 10], [11, 100], [101, 1000], and [1001, inf] are common.

3.4 Diversity

The concept of diversity unites two properties of a repertoire, namely, the number of distinct clones and their distribution. As such, diversity describes the composition and state of a repertoire. For instance, a repertoire derived from a completely naive cell population is much more diverse both in terms of distinct clones and their distribution compared to the repertoire of antigen-specific memory cells.

There are numerous sampling factors that are important to consider when measuring diversity. Perhaps the most important is whether a sample is derived from gDNA or mRNA [37]. As

discussed in Subheading 3.1 in chapter "AIRR Community Guide to TR and IG Gene Annotation," in the case of gDNA, each sampled cell contributes one or two templates, while the number of templates in mRNA data will be skewed by cell subset-specific transcript abundance. In the case of the former, diversity measures will be influenced substantially less by the underlying subset distribution than the latter. For both, one can measure diversity weighted by copy number or by clone number. For DNA data, using copy number-weighted diversity measures can give a sense of how similar *sequences* are in the underlying repertoire while using an unweighted measure will indicate how similar *clones* are. With RNA, using copy number-weighted measures will give a general measure of how similar *large clones* are, and unweighted measures will give a measure of how similar *all clones* are.

Another consideration when analyzing diversity is the depth of sequencing, that is, the proportion of clones that were sequenced compared to how many were actually in the sample. Assessing appropriate sequencing depth is no trivial task, but very important as undersampling can lead to false conclusions. Rarefaction curves [38] can help to evaluate if a repertoire is near full sampling depth. In this visualization, the number of distinct clones are plotted for a given subsample size (Fig. 1g). If the numbers of distinct clones plateau, the repertoire is near full sampling depth. Conversely, the absence of a plateau is an indication that the sampling depth of repertoire is shallow.

Another use of rarefaction is an estimation of the total number of clones from the sample. To achieve this, libraries from the sample of interest must be run in replicates, where more replicates give a more accurate estimate of total clones [39].

There are a large number of diversity metrics. These different metrics are all united in Hill numbers which are calculated over a range of diversities to generate a smooth curve (Fig. 1h) [40–42]. The function `calcDiversity` in the R package Alakazam estimates the Hill numbers for a repertoire. The same function also makes calculation of particular diversity indices straightforward. The function `compareHillNumbers` in the R package sumrep compares one or more Hill numbers of two repertoires. Newer approaches toward diversity metrics specific for AIRR make use of Hill numbers combined with a functional similarity matrix [43].

3.5 Similarity of AIRR Sequences

The similarity of AIRR sequences directly influences antigen recognition breadth: the more dissimilar the receptors are, the larger is the antigen space covered. One major approach to interrogate and measure AIRR sequence similarity is network analysis (Fig. 1i) [44–50]. Networks allow investigation of sequence similarity and thereby add a complementary layer of information to repertoire diversity analysis. Sequence networks are built by defining each nucleotide or amino acid sequence as a node. Two nodes are

connected with an edge if a certain similarity condition is satisfied, which is typically defined as a string distance (e.g., Levenshtein/edit distance). A commonly used distance for both IG and TR is one amino acid difference [44]. For B cells, networks representing amino acid distances of up to 12 amino acids have been reported [47]. Building a sequence similarity network is computationally expensive. This challenge has been approached by at least two methods that allow the construction of large-scale networks from millions of AIRR sequences [47, 51].

Although networks of a few thousand nodes may be visualized using software suites such as igraph, Cytoscape, and Gephi [52, 53], and the visual interpretation of networks becomes indiscernible with a size of $>10^2$ nodes. Furthermore, the visualization of networks does not provide quantitative information regarding the network similarity architecture. To address this problem, graph properties and network analysis have recently been employed to quantify the architecture of large-scale AIRR networks [47]. Architecture analytics may be subdivided into properties that capture the repertoire at the global level (generally one coefficient per network), and those that describe the repertoire at the local level (one coefficient per sequence per repertoire). These network measures may be used to identify enrichment of network clusters (Fig. 1i), potentially originating from an ongoing immune response [46, 47].

To increase precision in isolating immune-associated AIRR sequences and clusters therefore, network analysis may be coupled with AIRR generation probabilities [45]. More generally, it has been observed that sequences that tend to show increased sharing across individuals (discussed in the *see* Subheading 3.7), are also more connected within a repertoire [45, 47, 48] and confer robustness on its architecture with respect to network properties [47].

Recently, sequence similarity and diversity analysis have been combined, providing further insights into AIRR architecture [43].

3.6 Similarity among Repertoires

Similarity indices measure the similarity of two populations by not only considering the number of shared clones but also taking clone count or frequency into account (Fig. 1j). Similarity is sometimes calculated as dissimilarity (for historical reasons), but the index is always in the range of [0, 1]. It is therefore important to indicate the meaning of 0 and 1 to avoid confusion. One of the most popular indices is called Morisita-Horn, implemented in the function `vegdist` in the R package vegan [54]. Numerically, the observed overlaps are usually small, but considering the potential repertoire being sampled, the upfront chance of an overlap is very small. Alternatively, the CDR3s shared between samples can be plotted as a true/false heatmap (Fig. 1k). This is particularly useful when tracking clones over time or assessing the specificity of transplant infiltrating cells [55, 56].

Similarities on other parameters such as different amino acid properties as well as pairwise CDR3 distance and GC content can be compared between repertoires by the function `compareRepertoires` in the R package sumrep.

Other proposed similarity measures make use of feature counting [57], while another B-cell-specific similarity metric focuses on identical CDR3 length together with identical V and J genes considered within and between repertoires [58].

3.7 Public Clones

Though not clones in a true biological sense, the existence of identical TRs and identical or closely similar IGs in multiple individuals due to convergent rearrangement has been noted on several occasions [59–61]. Such rearrangements are termed public clones and can yield insights into common selection patterns, which in turn can elucidate how the immune system responds to disease and if there are commonalities between individuals. The ability to identify public clones in an AIRR depends on the sequencing depth and the number of individuals tested [62, 63]. In addition, the meaning of a public immune receptor must be assessed in the context of the likelihood for it to be generated [8, 13]. Receptors with shorter CDR3s are more likely to be generated by chance and can overlap even between individuals with no exposures in common [60, 64, 65] and do not necessarily indicate a convergent response in multiple individuals to similar antigens. Sequences that share the same (preferably longer) CDR3 amino acid sequence but have different nucleotide sequences are more convincing as candidate public clones, as differences in the nucleotide sequences may indicate independent generation with convergent selection [66].

Functionally identical IG can be identified by allowing some degree of difference in the CDR3. There is no well-defined cutoff to ensure the capture of a majority of receptors with identical specificities without including IGs of unrelated specificity into a particular collection of public IGs. A commonly used cutoff is 10–20% amino acid difference in the CDR3 [67–70]. Although a less restrictive cutoff might detect more divergent public clones [71], care must be taken to avoid identification of spurious public immune receptors [72]. Cross-contamination and index hopping on the sequencer further complicate the identification of public clones [73], and suitable definitions and analysis parameters may be helpful.

3.8 Detection and Monitoring of Cross-Sample Contamination Events

Despite strict quality assurance and control measures, PCR-based sample cross-contamination can occur at any time. Environmental contamination events are expected to arise from the presence of remaining DNA amplicons, which can be re-amplified and incorporated into new, unrelated libraries [74]. PCR contaminations can lead to major losses of reagents, time, and samples, and rapid detection and isolation are critical to the health of an AIRR-

seq research laboratory. There are several experimental precautions that can reduce contamination, including separate work areas and different sample barcodes, as illustrated in the AIRR Community chapter "Quality Control: Chain Pairing Precision and Monitoring of Cross-Sample Contamination."

3.9 B-Cell-Specific Aspects

3.9.1 IG SHM Analysis

SHM is the process driving the affinity maturation of IGs during the adaptive immune response [75]. Mutations are introduced at a rate of ~10^{-3} mutations per base pair per division. These mutations are not randomly distributed along the IG but accumulate more in hotspots and CDRs, whereas coldspots and framework regions are disfavored for mutation. Furthermore, substitution profiles may be germline gene-directed [76–79], possibly as a consequence of specific features of the encoded protein sequence. Understanding SHM biases is key to develop better tools to reconstruct lineages, quantify selection pressure, and generate realistic simulated sequence data [9, 79, 80].

To better understand the distribution of targets for SHM, it is, for instance, possible to use the R package sumrep that provides two functions `getHotspotCountDistribution` and `getColdspotCountDistribution` to the distribution of the hot- and coldspot motifs in the repertoire. In addition, sumrep interfaces with the R package SHazaM [6], which calculates a mutability model for the likelihood for the center base in a 5-mer to be mutated (the function `getMutabilityModel`). The associated function `getSubstitutionModel` provides the relative probabilities that the center base in a 5-mer is mutated into each of the other three nucleotides. SHazaM also provides methods for quantification of selection pressure and whether it has contributed to the nature of the specific IG repertoire during antigenic stimulation [81].

3.9.2 Identification of B-Cell Clones

As noted above, B-cell clones can be inferred from AIRR-seq data by analyzing their CDR3s and/or mutation patterns (Fig. 1l). Repertoires usually consist of hundreds or thousands of clonal lineages. Due to the presence of SHM, members of a B-cell clone cannot be identified solely based on identical CDR3s. There are many methods available to group IGs into clonal lineages (Table 1), but all generally attempt to computationally group sequences which likely share a common progenitor. However, different approaches can drastically change the interpretation of the underlying IG immune repertoire.

Some approaches begin by grouping sequences by their CDR3 independent of their V, D, or J gene usage [22]. Other software first groups sequences by gene (generally just V and J due to the difficulty in D gene annotation) and CDR3 length after which sequences similar in the CDR3 are grouped into clonal lineages [12, 19, 82, 83]. SCOPer does a similar grouping, but then

evaluates the similarity by analyzing shared SHM in the V and J genes [84]. Finally, some pipelines use common mutations in the body of the V gene to group sequences from the same clonal lineage [36, 85]. It is also possible to combine these approaches, but this section focuses on each independently.

Each approach has potential benefits and flaws. Initially grouping sequences by CDR3, either by identity or hierarchical clustering, can result in inflated copy number and sequence counts for common CDR3s (in particular those of short length that incorporate few non-templated bases) which may have arisen independently and utilize different genes. However, this method can be beneficial as some gene calls may be incorrect (in particular when annotation of sequences has not been made using a personalized repertoire as defined above), and similar CDR3 amino-acid sequences, especially those with long lengths, can indicate that sequences are related.

Grouping sequences by both gene annotation and CDR3 length prior to inferring clonal lineages can be beneficial for a number of reasons. Because V gene annotation is generally robust to sequencing error, sequences with similar CDR3s but different V gene assignments are unlikely to derive from the same rearrangement. Binning by gene annotation can therefore prevent erroneous clonal groupings. It also eases the computational burden, as CDR3 identity only needs calculation among smaller sets of sequences. Similar advantages apply to binning by CDR3 length as well, since distance metrics can be calculated more efficiently without the need for alignment. While insertions and deletions can occur as part of SHM, they are relatively rare [86, 87] and can be neglected in many cases.

Once sequences have been binned, hierarchical clustering is a common technique for identifying clonally related sequences [82]. This requires a choice of linkage (e.g., single, average) to define the distance between groups of sequences and a threshold for cutting the hierarchy into discrete groups. A convenient way to set the threshold is to analyze the distribution of distances between nearest neighbors. This distribution is typically bimodal, with the first mode representing sequences in the same clonal lineage, while the second mode represents sequences that do not have any relatives in the data. If the distribution for a particular sample is not bimodal, a set of external sequences from a different subject can be used to establish the threshold [82]. While the threshold for separating the two modes can sometimes be established by visual inspection of the distribution, there are algorithmic methods to determine it more consistently [18].

The last common approach is to group sequences into clones by common mutations in the body of the V gene. This can be done by constructing clonal lineages directly or by inspecting the *k*-mers

of each sequence [36, 88]. Unlike methods that first separate sequences by gene call and junction length, this method takes advantage of infrequent mutations to group sequences into clones. This can be beneficial for a number of reasons in certain circumstances. First, this method does not rely on proper gene calling or sequence alignment, which can be difficult in samples containing highly mutated populations or more generally due to sequencing error. Additionally, it is not sensitive to junction length, allowing sequences that have accumulated insertions and deletions to be grouped into clones [89, 90]. This method necessitates one to define the minimum number of mutations required to group two sequences into the same clone. A fixed value can be used, or the value can be dynamically determined based on the distribution of distances between each pair of sequences.

3.9.3 IG Affinity Maturation

The reconstruction and analysis of IG clonal lineages trees is a powerful method to understand the immune response, affinity maturation, and the generation of broadly neutralizing antibodies (bNAb) [91–93]. Within a B-cell clonal lineage, B cells descended from a shared common ancestor evolve through SHM and antigen-driven selection. While standard algorithms for inferring phylogenetic trees using maximum parsimony and maximum likelihood [94] are often employed, these approaches can be improved [80]. In particular, the unique biology of B cells can present problems for standard phylogenetic approaches and has led to the development of B-cell-specific phylogenetic tools. One cause of the problems is that SHM is enzymatically driven and biased by hotspot and coldspot motifs. This violates the assumption of independent evolution among sites that many likelihood-based phylogenetics methods rely on. To address this challenge, more context-aware phylogenetic methods, such as IgPhyML [9, 10], have been developed. While context-aware models of SHM clearly improve estimates of phylogenetic model parameters used to detect antigen-driven selection [10], it is less clear how much they improve estimates of tree topology and branch lengths [95]. Another problem is that while standard phylogenetic models consider clonal lineages individually, IG repertoires often contain hundreds of independent clones. The use of repertoire-wide models, which allow some parameters to be shared among these multiple clonal lineages, can improve model precision significantly [10]. One important application of B-cell phylogenetics is estimating the series of mutations leading from a clone's unmutated germline ancestor to a sequence of interest, such as a known bnAb sequence. While standard phylogenetic methods can reconstruct intermediate sequences, they are less appropriate for reconstructing the germline ancestral sequence because they do not take into account the biology of V(D)J rearrangement. This has led to the development of tools such as

Clonalyst and linearham [96, 97] that improve the reconstruction of these sequences by combining phylogenetic models with models of V(D)J rearrangement. Another feature of B-cell clonal lineages is that reconstructed intermediate sequences are often identical to observed IG sequences. Some tools, such as IgTree [98] and Alakazam [6], use this fact to simplify the visualization of these lineage trees by collapsing observed and sampled intermediate nodes. Finally, lineage trees containing B cells from multiple tissues, isotypes, and timepoints have the potential to be used to make inferences about how B-cell migration, isotype switching, and evolution over time occur. Multiple analyses have used lineage trees for this purpose [33, 40, 99, 100], and generalized tools for making these inferences from B-cell repertoires, such as Dowser and PopTree, are an area of active development [7].

3.10 T-Cell-Specific Aspects

There is growing evidence that TR repertoire perturbations can serve as a biomarker of immune response toward some solid tumors [101–103] and pathogens such as Epstein-Barr virus (EBV), cytomegalovirus (CMV), Ebola, and SARS-CoV-2 [104–108]. Challenges with studying T-cell repertoires include the dependence of T-cell interactions on the major histocompatibility complex (MHC) [109], changes in TRBV usage based on MHC and significant differences in TRBV usage, and clonality in CD4+ and CD8+ repertoires [110–112].

Antigen-specific TCRs can be isolated either by sorting of MHC-tetramer-positive cells or activated cells after stimulation with overlapping peptide pools. Staining with tetramers requires knowledge of the correct epitope in the right MHC context, and T cells with high affinity tend to be recovered with the highest efficiency. Therefore, tetramer staining sometimes fails to identify some of the relevant TCRs [113]. Stimulation with overlapping peptide pools, on the other hand, can lead to isolation of non-peptide-specific T cells due to bystander activation [114]. The TR of the antigen-enriched cells can be compared to samples from different timepoints to track the frequency of clones of interest [104, 106].

4 Conclusion

In this chapter, we have provided a brief overview of diverse, widely used techniques to uncover biological information in AIRR-seq data. These techniques can be applied to all of the AIRR-seq data created using the methodologies described in this book. They further form the basis for selecting the optimal experimental protocol to address the biological question and choosing the computational methods used in the analysis.

Acknowledgments

The authors would like to thank Mats Ohlin for the constructive criticism of the manuscript. US was supported by grants from Mercator Stiftung, Germany; German Research Foundation, Germany (DFG, grant 397650460); BMBF e:KID, Germany (01ZX1612A); and BMBF NoChro, Germany (FKZ 13GW0338B).

References

1. Zhang Y, Yang X, Zhang Y, Zhang Y, Wang M, Ou JX et al (2020) Tools for fundamental analysis functions of TCR repertoires: a systematic comparison. Brief Bioinform 21: 1706–1716. https://doi.org/10.1093/bib/bbz092
2. López-Santibáñez-Jácome L, Avendaño-Vázquez SE, Flores-Jasso CF (2019) The pipeline repertoire for Ig-seq analysis. Front Immunol 10:899. https://doi.org/10.3389/fimmu.2019.00899
3. Lees WD (2020) Tools for adaptive immune receptor repertoire sequencing. Curr Opin Syst Biol 24:86–92. https://doi.org/10.1016/j.coisb.2020.10.003
4. Smakaj E, Babrak L, Ohlin M, Shugay M, Briney B, Tosoni D et al (2020) Benchmarking immunoinformatic tools for the analysis of antibody repertoire sequences. Bioinformatics 36:1731–1739. https://doi.org/10.1093/bioinformatics/btz845
5. Martin ACR (2010) Protein sequence and structure analysis of antibody variable domains. In: Kontermann R, Dübel S (eds) Antibody engineering. Springer, Berlin, pp 33–51
6. Gupta NT, Vander Heiden JA, Uduman M, Gadala-Maria D, Yaari G, Kleinstein SH (2015) Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. Bioinformatics 31: 3356–3358. https://doi.org/10.1093/bioinformatics/btv359
7. Hoehn KB, Pybus OG, Kleinstein SH (2020) Phylogenetic analysis of migration, differentiation, and class switching in B cells. Immunology
8. Marcou Q, Mora T, Walczak AM (2018) High-throughput immune repertoire analysis with IGoR. Nat Commun 9:561. https://doi.org/10.1038/s41467-018-02832-w
9. Hoehn KB, Lunter G, Pybus OG (2017) A phylogenetic codon substitution model for antibody lineages. Genetics 206:417–427. https://doi.org/10.1534/genetics.116.196303
10. Hoehn KB, Vander Heiden JA, Zhou JQ, Lunter G, Pybus OG, Kleinstein SH (2019) Repertoire-wide phylogenetic models of B cell molecular evolution reveal evolutionary signatures of aging and vaccination. Proc Natl Acad Sci U S A 116:22664–22672. https://doi.org/10.1073/pnas.1906020116
11. ImmunoMind Team (2019) immunarch: an R Package for painless analysis of large-scale immune repertoire data
12. Bolotin DA, Poslavsky S, Mitrophanov I, Shugay M, Mamedov IZ, Putintseva EV et al (2015) MiXCR: software for comprehensive adaptive immunity profiling. Nat Methods 12:380–381. https://doi.org/10.1038/nmeth.3364
13. Sethna Z, Elhanati Y, Callan CG, Walczak AM, Mora T (2019) OLGA: fast computation of generation probabilities of B- and T-cell receptor amino acid sequences and motifs. Bioinformatics 35:2974–2981. https://doi.org/10.1093/bioinformatics/btz035
14. Ralph DK, Matsen FA (2016) Likelihood-based inference of B cell clonal families. PLoS Comput Biol 12:e1005086. https://doi.org/10.1371/journal.pcbi.1005086
15. Ralph DK, Matsen FA (2020) Using B cell receptor lineage structures to predict affinity. PLoS Comput Biol 16:e1008391. https://doi.org/10.1371/journal.pcbi.1008391
16. Gidoni M, Snir O, Peres A, Polak P, Lindeman I, Mikocziova I, IMI test presentation (2019) Mosaic deletion patterns of the human antibody heavy chain gene locus shown by Bayesian haplotyping. Nat Commun 10:628. https://doi.org/10.1038/s41467-019-08489-3
17. Sturm G, Szabo T, Fotakis G, Haider M, Rieder D, Trajanoski Z, IMI test presentation (2020) Scirpy: a Scanpy extension for analyzing single-cell T-cell receptor-sequencing

data. Bioinformatics 36:4817–4818. https://doi.org/10.1093/bioinformatics/btaa611

18. Nouri N, Kleinstein SH (2018) A spectral clustering-based method for identifying clones from high-throughput B cell repertoire sequencing data. Bioinformatics 34: i341–i349. https://doi.org/10.1093/bioinformatics/bty235
19. Schramm CA, Sheng Z, Zhang Z, Mascola JR, Kwong PD, Shapiro L (2016) SONAR: a high-throughput pipeline for inferring antibody ontogenies from longitudinal sequencing of B cell transcripts. Front Immunol 7: 372. https://doi.org/10.3389/fimmu.2016.00372
20. Olson BJ, Moghimi P, Schramm CA, Obraztsova A, Ralph D, Vander Heiden JA et al (2019) Sumrep: a summary statistic framework for immune receptor repertoire comparison and model validation. Front Immunol 10:2533. https://doi.org/10.3389/fimmu.2019.02533
21. Lees WD, Shepherd AJ (2015) Utilities for high-throughput analysis of B-cell clonal lineages. J Immunol Res 2015:1–9. https://doi.org/10.1155/2015/323506
22. Giraud M, Salson M, Duez M, Villenet C, Quief S, Caillault A et al (2014) Fast multiclonal clusterization of V(D)J recombinations from high-throughput sequencing. BMC Genomics 15:409. https://doi.org/10.1186/1471-2164-15-409
23. Duez M, Giraud M, Herbert R, Rocher T, Salson M, Thonier F (2016) Vidjil: a web platform for analysis of high-throughput repertoire sequencing. PLoS One 11:e0166126. https://doi.org/10.1371/journal.pone.0166126
24. Christley S, Scarborough W, Salinas E, Rounds WH, Toby IT, Fonner JM, IMI test presentation (2018) VDJServer: a cloud-based analysis portal and data commons for immune repertoire sequences and rearrangements. Front Immunol 9:976. https://doi.org/10.3389/fimmu.2018.00976
25. Rosenfeld AM, Meng W, Luning Prak ET, Hershberg U (2018) ImmuneDB, a novel tool for the analysis, storage, and dissemination of immune repertoire sequencing data. Front Immunol 9:2107. https://doi.org/10.3389/fimmu.2018.02107
26. Xu JL, Davis MM (2000) Diversity in the CDR3 region of VH is sufficient for most antibody specificities. Immunity 13:37–45. https://doi.org/10.1016/S1074-7613(00)00006-6
27. Glanville J, Huang H, Nau A, Hatton O, Wagar LE, Rubelt F et al (2017) Identifying specificity groups in the T cell receptor repertoire. Nature 547:94–98. https://doi.org/10.1038/nature22976
28. Dash P, Fiore-Gartland AJ, Hertz T, Wang GC, Sharma S, Souquette A et al (2017) Quantifiable predictive features define epitope-specific T cell receptor repertoires. Nature 547:89–93. https://doi.org/10.1038/nature22383
29. Atchley WR, Zhao J, Fernandes AD, Druke T (2005) Solving the protein sequence metric problem. Proc Natl Acad Sci U S A 102: 6395–6400. https://doi.org/10.1073/pnas.0408677102
30. Kidera A, Konishi Y, Oka M, Ooi T, Scheraga HA (1985) Statistical analysis of the physical properties of the 20 naturally occurring amino acids. J Protein Chem 4:23–55. https://doi.org/10.1007/BF01025492
31. Haigh OL, Grant EJ, Nguyen THO, Kedzierska K, Field MA, Miles JJ (2021) Genetic bias, diversity indices, physiochemical properties and CDR3 motifs divide auto-reactive from Allo-reactive T-cell repertoires. Int J Mol Sci 22:1625. https://doi.org/10.3390/ijms22041625
32. Sankar K, Hoi KH, Hötzel I (2020) Dynamics of heavy chain junctional length biases in antibody repertoires. Commun Biol 3:207. https://doi.org/10.1038/s42003-020-0931-3
33. Wu X, Zhang Z, Schramm CA, Joyce MG, Kwon YD, Zhou T et al (2015) Maturation and diversity of the VRC01-antibody lineage over 15 years of chronic HIV-1 infection. Cell 161:470–485. https://doi.org/10.1016/j.cell.2015.03.004
34. Zhou JQ, Kleinstein SH (2019) Cutting edge: Ig H chains are sufficient to determine most B cell clonal relationships. J Immunol 203:1687–1692. https://doi.org/10.4049/jimmunol.1900666
35. Kotouza MT, Gemenetzi K, Galigalidou C, Vlachonikola E, Pechlivanis N, Agathangelidis A et al (2020) TRIP - T cell receptor/immunoglobulin profiler. BMC Bioinformatics 21:422. https://doi.org/10.1186/s12859-020-03669-1
36. Lindenbaum O, Nouri N, Kluger Y, Kleinstein SH (2021) Alignment free identification of clones in B cell receptor repertoires. Nucleic Acids Res 49:e21–e21. https://doi.org/10.1093/nar/gkaa1160

37. Bashford-Rogers RJM, Palser AL, Idris SF, Carter L, Epstein M, Callard RE et al (2014) Capturing needles in haystacks: a comparison of B-cell receptor sequencing methods. BMC Immunol 15:29. https://doi.org/10.1186/s12865-014-0029-0
38. Gotelli NJ, Colwell RK (2001) Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. Ecol Lett 4:379–391. https://doi.org/10.1046/j.1461-0248.2001.00230.x
39. Greiff V, Menzel U, Haessler U, Cook SC, Friedensohn S, Khan TA et al (2014) Quantitative assessment of the robustness of next-generation sequencing of antibody variable gene repertoires from immunized mice. BMC Immunol 15:40. https://doi.org/10.1186/s12865-014-0040-5
40. Stern JNH, Yaari G, Vander Heiden JA, Church G, Donahue WF, Hintzen RQ et al (2014) B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. Sci Transl Med 6:248ra107. https://doi.org/10.1126/scitranslmed.3008879
41. Greiff V, Bhat P, Cook SC, Menzel U, Kang W, Reddy ST (2015) A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. Genome Med 7:49. https://doi.org/10.1186/s13073-015-0169-8
42. Hill MO (1973) Diversity and evenness: a unifying notation and its consequences. Ecology 54:427–432. https://doi.org/10.2307/1934352
43. Arora R, Burke HM, Arnaout R (2018) Immunological diversity with similarity. Immunology
44. Miho E, Yermanos A, Weber CR, Berger CT, Reddy ST, Greiff V (2018) Computational strategies for dissecting the high-dimensional complexity of adaptive immune repertoires. Front Immunol 9:224. https://doi.org/10.3389/fimmu.2018.00224
45. Pogorelyy MV, Minervina AA, Shugay M, Chudakov DM, Lebedev YB, Mora T et al (2019) Detecting T cell receptors involved in immune responses from single repertoire snapshots. PLoS Biol 17:e3000314. https://doi.org/10.1371/journal.pbio.3000314
46. Ben-Hamo R, Efroni S (2011) The whole-organism heavy chain B cell repertoire from zebrafish self-organizes into distinct network features. BMC Syst Biol 5:27. https://doi.org/10.1186/1752-0509-5-27
47. Miho E, Roškar R, Greiff V, Reddy ST (2019) Large-scale network analysis reveals the sequence space architecture of antibody repertoires. Nat Commun 10:1321. https://doi.org/10.1038/s41467-019-09278-8
48. Madi A, Poran A, Shifrut E, Reich-Zeliger S, Greenstein E, Zaretsky I et al (2017) T cell receptor repertoires of mice and humans are clustered in similarity networks around conserved public CDR3 sequences. eLife 6:e22057. https://doi.org/10.7554/eLife.22057
49. Madi A, Shifrut E, Reich-Zeliger S, Gal H, Best K, Ndifon W et al (2014) T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity. Genome Res 24:1603–1612. https://doi.org/10.1101/gr.170753.113
50. Bashford-Rogers RJM, Palser AL, Huntly BJ, Rance R, Vassiliou GS, Follows GA et al (2013) Network properties derived from deep sequencing of human B-cell receptor repertoires delineate B-cell populations. Genome Res 23:1874–1884. https://doi.org/10.1101/gr.154815.113
51. Valkiers S, Van Houcke M, Laukens K, Meysman P (2021) clusTCR: a python interface for rapid clustering of large sets of CDR3 sequences. Bioinformatics
52. Csardi G, Nepusz T (2006) The igraph software package for complex network research. InterJournal Complex Systems 1695
53. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D et al (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–2504. https://doi.org/10.1101/gr.1239303
54. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D et al (2019) Vegan: community ecology package
55. Stervbo U, Nienen M, Hecht J, Viebahn R, Amann K, Westhoff TH et al (2020) Differential diagnosis of interstitial allograft rejection and BKV nephropathy by T-cell receptor sequencing. Transplantation 104:e107–e108. https://doi.org/10.1097/TP.0000000000003054
56. Nienen M, Stervbo U, Mölder F, Kaliszczyk S, Kuchenbecker L, Gayova L et al (2019) The role of pre-existing cross-reactive central memory CD4 T-cells in vaccination with previously unseen influenza strains. Front Immunol 10:593. https://doi.org/10.3389/fimmu.2019.00593
57. Bolen CR, Rubelt F, Vander Heiden JA, Davis MM (2017) The repertoire dissimilarity index as a method to compare lymphocyte receptor repertoires. BMC Bioinformatics 18:155.

https://doi.org/10.1186/s12859-017-1556-5

58. Greiff V, Menzel U, Miho E, Weber C, Riedel R, Cook S et al (2017) Systems analysis reveals high genetic and antigen-driven predetermination of antibody repertoires throughout B cell development. Cell Rep 19:1467–1478. https://doi.org/10.1016/j.celrep.2017.04.054
59. Bradley P, Thomas PG (2019) Using T cell receptor repertoires to understand the principles of adaptive immune recognition. Annu Rev Immunol 37:547–570. https://doi.org/10.1146/annurev-immunol-042718-041757
60. Soto C, Bombardi RG, Branchizio A, Kose N, Matta P, Sevy AM et al (2019) High frequency of shared clonotypes in human B cell receptor repertoires. Nature 566:398–402. https://doi.org/10.1038/s41586-019-0934-8
61. Greiff V, Weber CR, Palme J, Bodenhofer U, Miho E, Menzel U et al (2017) Learning the high-dimensional Immunogenomic features that predict public and private antibody repertoires. J Immunol 199:2985–2997. https://doi.org/10.4049/jimmunol.1700594
62. Elhanati Y, Sethna Z, Callan CG, Mora T, Walczak AM (2018) Predicting the spectrum of TCR repertoire sharing with a data-driven model of recombination. Immunol Rev 284:167–179. https://doi.org/10.1111/imr.12665
63. Greiff V, Miho E, Menzel U, Reddy ST (2015) Bioinformatic and statistical analysis of adaptive immune repertoires. Trends Immunol 36:738–749. https://doi.org/10.1016/j.it.2015.09.006
64. Briney B, Inderbitzin A, Joyce C, Burton DR (2019) Commonality despite exceptional diversity in the baseline human antibody repertoire. Nature 566:393–397. https://doi.org/10.1038/s41586-019-0879-y
65. Soto C, Bombardi RG, Kozhevnikov M, Sinkovits RS, Chen EC, Branchizio A et al (2020) High frequency of shared clonotypes in human T cell receptor repertoires. Cell Rep 32:107882. https://doi.org/10.1016/j.celrep.2020.107882
66. Venturi V, Quigley MF, Greenaway HY, Ng PC, Ende ZS, McIntosh T et al (2011) A mechanism for TCR sharing between T cell subsets and individuals revealed by pyrosequencing. J Immunol 186:4285–4294. https://doi.org/10.4049/jimmunol.1003898
67. Nielsen SCA, Yang F, Hoh RA, Jackson KJL, Roeltgen K, Lee J-Y et al (2020) B cell clonal expansion and convergent antibody responses to SARS-CoV-2. Res Sq
68. Nielsen SCA, Yang F, Jackson KJL, Hoh RA, Röltgen K, Jean GH (2020) Human B cell clonal expansion and convergent antibody responses to SARS-CoV-2. Cell Host Microbe 28:516–525.e5. https://doi.org/10.1016/j.chom.2020.09.002
69. Kim SI, Noh J, Kim S, Choi Y, Yoo DK, Lee Y et al (2021) Stereotypic neutralizing V_H antibodies against SARS-CoV-2 spike protein receptor binding domain in patients with COVID-19 and healthy individuals. Sci Transl Med 13:eabd6990. https://doi.org/10.1126/scitranslmed.abd6990
70. Galson JD, Schaetzle S, Bashford-Rogers RJM, Raybould MIJ, Kovaltsuk A, Kilpatrick GJ et al (2020) Deep sequencing of B cell receptor repertoires from COVID-19 patients reveals strong convergent immune signatures. Front Immunol 11:605170. https://doi.org/10.3389/fimmu.2020.605170
71. Ohlin M (2014) A new look at a poorly immunogenic neutralization epitope on cytomegalovirus glycoprotein B. Is there cause for antigen redesign? Mol Immunol 60:95–102. https://doi.org/10.1016/j.molimm.2014.03.015
72. Japp AS, Meng W, Rosenfeld AM, Perry DJ, Thirawatananond P, Bacher RL et al (2021) TCR+/BCR+ dual-expressing cells and their associated public BCR clonotype are not enriched in type 1 diabetes. Cell 184:827–839.e14. https://doi.org/10.1016/j.cell.2020.11.035
73. Costello M, Fleharty M, Abreu J, Farjoun Y, Ferriera S, Holmes L et al (2018) Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. BMC Genomics 19:332. https://doi.org/10.1186/s12864-018-4703-0
74. Seitz V, Schaper S, Dröge A, Lenze D, Hummel M, Hennig S (2015) A new method to prevent carry-over contaminations in two-step PCR NGS library preparations. Nucleic Acids Res 43(20):e135. https://doi.org/10.1093/nar/gkv694
75. Methot SP, Di Noia JM (2017) Molecular mechanisms of somatic Hypermutation and class switch recombination. Adv Immunol 133:37–87
76. Sheng Z, Schramm CA, Kong R, Comparative Sequencing Program NISC, Mullikin JC,

Mascola JR et al (2017) Gene-specific substitution profiles describe the types and frequencies of amino acid changes during antibody somatic Hypermutation. Front Immunol 8: 537. https://doi.org/10.3389/fimmu.2017.00537

77. Schramm CA, Douek DC (2018) Beyond hot spots: biases in antibody somatic hypermutation and implications for vaccine design. Front Immunol 9:1876. https://doi.org/10.3389/fimmu.2018.01876
78. Kirik U, Persson H, Levander F, Greiff L, Ohlin M (2017) Antibody heavy chain variable domains of different germline gene origins diversify through different paths. Front Immunol 8:1433. https://doi.org/10.3389/fimmu.2017.01433
79. Zhou JQ, Kleinstein SH (2020) Position-dependent differential targeting of somatic Hypermutation. J Immunol 205: 3468–3479. https://doi.org/10.4049/jimmunol.2000496
80. Yermanos A, Greiff V, Krautler NJ, Menzel U, Dounas A et al (2017) Comparison of methods for phylogenetic B-cell lineage inference using time-resolved antibody repertoire simulations (AbSim). Bioinformatics 33: 3938–3946. https://doi.org/10.1093/bioinformatics/btx533
81. Yaari G, Uduman M, Kleinstein SH (2012) Quantifying selection in high-throughput immunoglobulin sequencing data sets. Nucleic Acids Res 40:e134–e134. https://doi.org/10.1093/nar/gks457
82. Gupta NT, Adams KD, Briggs AW, Timberlake SC, Vigneault F, Kleinstein SH (2017) Hierarchical clustering can identify B cell clones with high confidence in Ig repertoire sequencing data. J Immunol 198:2489–2499. https://doi.org/10.4049/jimmunol.1601850
83. Aouinti S, Malouche D, Giudicelli V, Kossida S, Lefranc M-P (2015) IMGT/HighV-QUEST statistical significance of IMGT Clonotype (AA) diversity per gene for standardized comparisons of next generation sequencing immunoprofiles of immunoglobulins and T cell receptors. PLoS One 10: e0142353. https://doi.org/10.1371/journal.pone.0142353
84. Nouri N, Kleinstein SH (2020) Somatic hypermutation analysis for improved identification of B cell clonal families from next-generation sequencing data. PLoS Comput Biol 16:e1007977. https://doi.org/10.1371/journal.pcbi.1007977
85. Briney B, Le K, Zhu J, Burton DR (2016) Clonify: unseeded antibody lineage assignment from next-generation sequencing data. Sci Rep 6:23901. https://doi.org/10.1038/srep23901
86. Briney BS, Willis JR, Crowe JE (2012) Location and length distribution of somatic hypermutation-associated DNA insertions and deletions reveals regions of antibody structural plasticity. Genes Immun 13: 523–529. https://doi.org/10.1038/gene.2012.28
87. Briney BS, Willis JR, Crowe JE (2012) Human peripheral blood antibodies with long HCDR3s are established primarily at original recombination using a limited subset of germline genes. PLoS One 7:e36750. https://doi.org/10.1371/journal.pone.0036750
88. Yaari G, Benichou JIC, Vander Heiden JA, Kleinstein SH, Louzoun Y (2015) The mutation patterns in B-cell immunoglobulin receptors reflect the influence of selection acting at multiple time-scales. Philos Trans R Soc B Biol Sci 370:20140242. https://doi.org/10.1098/rstb.2014.0242
89. Wilson PC, de Bouteiller O, Liu Y-J, Potter K, Banchereau J, Capra JD et al (1998) Somatic Hypermutation introduces insertions and deletions into immunoglobulin V genes. J Exp Med 187:59–70. https://doi.org/10.1084/jem.187.1.59
90. Ohlin M, Borrebaeck CAK (1998) Insertions and deletions in hypervariable loops of antibody heavy chains contribute to molecular diversity. Mol Immunol 35:233–238. https://doi.org/10.1016/S0161-5890(98)00030-3
91. Shlomchik MJ, Marshak-Rothstein A, Wolfowicz CB, Rothstein TL, Weigert MG (1987) The role of clonal selection and somatic mutation in autoimmunity. Nature 328:805–811. https://doi.org/10.1038/328805a0
92. Haynes BF, Kelsoe G, Harrison SC, Kepler TB (2012) B-cell-lineage immunogen design in vaccine development with HIV-1 as a case study. Nat Biotechnol 30:423–433. https://doi.org/10.1038/nbt.2197
93. Liao H-X, Lynch R, Zhou T, Gao F, Alam SM, Boyd SD et al (2013) Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. Nature 496:469–476. https://doi.org/10.1038/nature12053
94. Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood

approach. J Mol Evol 17:368–376. https://doi.org/10.1007/BF01734359

95. Davidsen K, Matsen FA (2018) Benchmarking tree and ancestral sequence inference for B cell receptor sequences. Front Immunol 9:2451. https://doi.org/10.3389/fimmu.2018.02451
96. Kepler TB (2013) Reconstructing a B-cell clonal lineage. I. Statistical inference of unobserved ancestors. F1000Res 2:103. https://doi.org/10.12688/f1000research.2-103.v1
97. Dhar A, Ralph DK, Minin VN, Matsen FA (2020) A Bayesian phylogenetic hidden Markov model for B cell receptor sequence analysis. PLoS Comput Biol 16:e1008030. https://doi.org/10.1371/journal.pcbi.1008030
98. Barak M, Zuckerman NS, Edelman H, Unger R, Mehr R (2008) IgTree: creating immunoglobulin variable region gene lineage trees. J Immunol Methods 338:67–74. https://doi.org/10.1016/j.jim.2008.06.006
99. Horns F, Vollmers C, Croote D, Mackey SF, Swan GE, Dekker CL et al (2016) Lineage tracing of human B cells reveals the in vivo landscape of human antibody class switching. eLife 5:e16578. https://doi.org/10.7554/eLife.16578
100. Vieira MC, Zinder D, Cobey S (2018) Selection and neutral mutations drive pervasive mutability losses in long-lived anti-HIV B-cell lineages. Mol Biol Evol 35:1135–1146. https://doi.org/10.1093/molbev/msy024
101. Cui J-H, Lin K-R, Yuan S-H, Jin Y-B, Chen X-P, Su X-K et al (2018) TCR repertoire as a novel indicator for immune monitoring and prognosis assessment of patients with cervical cancer. Front Immunol 9:2729. https://doi.org/10.3389/fimmu.2018.02729
102. Vollmer T, Schlickeiser S, Amini L, Schulenberg S, Wendering DJ, Banday V et al (2021) The intratumoral CXCR3 chemokine system is predictive of chemotherapy response in human bladder cancer. Sci Transl Med 13:eabb3735. https://doi.org/10.1126/scitranslmed.abb3735
103. Li N, Yuan J, Tian W, Meng L, Liu Y (2020) T-cell receptor repertoire analysis for the diagnosis and treatment of solid tumor: a methodology and clinical applications. Cancer Commun (Lond) 40:473–483. https://doi.org/10.1002/cac2.12074
104. Dziubianau M, Hecht J, Kuchenbecker L, Sattler A, Stervbo U, Rödelsperger C et al (2013) TCR repertoire analysis by next generation sequencing allows complex differential diagnosis of T cell-related pathology: NGS allows complex differential diagnosis. Am J Transplant 13:2842–2854. https://doi.org/10.1111/ajt.12431
105. Wolf K, Hether T, Gilchuk P, Kumar A, Rajeh A, Schiebout C et al (2018) Identifying and tracking low-frequency virus-specific TCR Clonotypes using high-throughput sequencing. Cell Rep 25:2369–2378.e4. https://doi.org/10.1016/j.celrep.2018.11.009
106. Pogorelyy MV, Minervina AA, Touzel MP, Sycheva AL, Komech EA, Kovalenko EI et al (2018) Precise tracking of vaccine-responding T cell clones reveals convergent and personalized response in identical twins. Proc Natl Acad Sci U S A 115:12704–12709. https://doi.org/10.1073/pnas.1809642115
107. Schober K, Buchholz VR, Busch DH (2018) TCR repertoire evolution during maintenance of CMV-specific T-cell populations. Immunol Rev 283:113–128. https://doi.org/10.1111/imr.12654
108. Gittelman RM, Lavezzo E, Snyder TM, Zahid HJ, Elyanow R, Dalai S, IMI test presentation (2020) Diagnosis and tracking of SARS-CoV-2 infection By T-cell receptor sequencing. Infectious diseases (except HIV/AIDS)
109. Klein L, Kyewski B, Allen PM, Hogquist KA (2014) Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see). Nat Rev Immunol 14:377–391. https://doi.org/10.1038/nri3667
110. Logunova NN, Kriukova VV, Shelyakin PV, Egorov ES, Pereverzeva A, Bozhanova NG, IMI test presentation (2020) MHC-II alleles shape the CDR3 repertoires of conventional and regulatory naïve CD4+ T cells. Proc Natl Acad Sci U S A 117:13659–13669. https://doi.org/10.1073/pnas.2003170117
111. Lu J, Van Laethem F, Bhattacharya A, Craveiro M, Saba I, Chu J, IMI test presentation (2019) Molecular constraints on CDR3 for thymic selection of MHC-restricted TCRs from a random pre-selection repertoire. Nat Commun 10:1019. https://doi.org/10.1038/s41467-019-08906-7
112. Migalska M, Sebastian A, Radwan J (2019) Major histocompatibility complex class I diversity limits the repertoire of T cell receptors. Proc Natl Acad Sci U S A 116:5021–5026. https://doi.org/10.1073/pnas.1807864116
113. Rius C, Attaf M, Tungatt K, Bianchi V, Legut M, Bovay A, IMI test presentation

(2018) Peptide-MHC class I tetramers can fail to detect relevant functional T cell clonotypes and underestimate antigen-reactive T cell populations. J Immunol 200: 2263–2279. https://doi.org/10.4049/jimmunol.1700242

114. Martin MD, Jensen IJ, Ishizuka AS, Lefebvre M, Shan Q, Xue H-H, IMI test presentation (2019) Bystander responses impact accurate detection of murine and human antigen-specific CD8 T cells. J Clin Invest 129:3894–3908. https://doi.org/10.1172/JCI124443

Chapter 18

Bulk gDNA Sequencing of Antibody Heavy-Chain Gene Rearrangements for Detection and Analysis of B-Cell Clone Distribution: A Method by the AIRR Community

Aaron M. Rosenfeld, Wenzhao Meng, Kalisse I. Horne, Elaine C. Chen, Davide Bagnara, Ulrik Stervbo, and Eline T. Luning Prak and on behalf of the AIRR Community

Abstract

In this method we illustrate how to amplify, sequence, and analyze antibody/immunoglobulin (IG) heavy-chain gene rearrangements from genomic DNA that is derived from bulk populations of cells by next-generation sequencing (NGS). We focus on human source material and illustrate how bulk gDNA-based sequencing can be used to examine clonal architecture and networks in different samples that are sequenced from the same individual. Although bulk gDNA-based sequencing can be performed on both IG heavy (IGH) or kappa/lambda light (IGK/IGL) chains, we focus here on IGH gene rearrangements because IG heavy chains are more diverse, tend to harbor higher levels of somatic hypermutations (SHM), and are more reliable for clone identification and tracking. We also provide a procedure, including code, and detailed instructions for processing and annotation of the NGS data. From these data we show how to identify expanded clones, visualize the overall clonal landscape, and track clonal lineages in different samples from the same individual. This method has a broad range of applications, including the identification and monitoring of expanded clones, the analysis of blood and tissue-based clonal networks, and the study of immune responses including clonal evolution.

Key words Antibody, Clone, Lineage, Immune repertoire profiling, Immunoglobulin, V(D)J recombination, Next-generation sequencing

1 Introduction

Antibodies or immunoglobulins (IGs) on B cells are generated through somatic recombination of variable (V), diversity (D), and joining (J) genes [1, 2] and further diversified through somatic hypermutation (SHM) [3, 4]. The collection of different B cells in an individual, also known as the immune repertoire, is complex,

Aaron M. Rosenfeld and Wenzhao Meng are shared first authors.

Anton W. Langerak (ed.), *Immunogenetics: Methods and Protocols*, Methods in Molecular Biology, vol. 2453, https://doi.org/10.1007/978-1-0716-2115-8_18,

containing many different B cells with different antibodies. B cells that derive from the same progenitor are clonally related and harbor gene rearrangements that are identical or have very similar nucleotide sequences (differing only by SHM or sequencing errors). The grouping of antibody gene rearrangement sequences into clones provides a means of characterizing the immune repertoire with respect to the distribution, size, complexity, and dynamics of clones in different cell types and tissues [5–7].

Here we describe a homebrew method, with primer sequences adapted for NGS from the BIOMED2 IG heavy-chain (IGH) PCR assays [8], to evaluate samples for evidence of B-cell clonal expansion and track clones in bulk gDNA samples. Similar methods exist as commercial services (e.g., Adaptive Biotechnologies, iRepertoire), and there are also similar homebrew methods for the analysis of T-cell AIRR-seq data (e.g., [9]). This homebrew method for IGH rearrangements uses multiplex PCR and can be scaled to very high cell inputs as described in [10]. DNA is more robust than RNA and has a simpler relationship to cell numbers (one template per cell) than RNA. For these reasons, bulk gDNA-based sequencing is typically the method of choice for clinical-grade assays to evaluate malignant clonal expansions [11], as well as the in-depth study of clones in different tissues to study clonal networks in the body [10]. The method shown uses long reads that are adequate for robust IGHV gene alignment and evaluation of SHM, but this method can also be performed with shorter reads, depending upon the sample type and DNA quality.

In this chapter, we also illustrate how to use pRESTO [12] and ImmuneDB [13] to analyze sequencing data generated following the wet bench protocol. In this dry bench analysis, we describe how to filter the raw read data, group highly similar rearrangements into clones using both the IGHV gene and CDR3 sequences, estimate clone size distributions, and track clones of interest in other samples.

2 Materials

2.1 Primers

All IG gene amplification primers are synthesized by Integrated DNA Technologies, and HPLC purification is recommended for sequences that are longer than 60 bp and any sequence that contains one or more "Ns" (random nucleotides). Dual indices are provided to distinguish clone identification from tracking primers (*see* **Note 1**).

1. Human (Hu) IGH amplification primers for clone identification:

   NexteraR2-Hu-VH1-FW1:GTCTCGTGGGCTCGGAGAT GTGTATAAGAGACAGGGCCT

CAGTGAAGGTCTCCTGCAAG

NexteraR2-Hu-VH2-FW1:GTCTCGTGGGCTCGGAGAT TGTATAAGAGACAGGTCTG

GTCCTACGCTGGTGAAACCC

NexteraR2-Hu-VH3-FW1:GTCTCGTGGGCTCGGAGAT GTGTATAAGAGACAGCTGG

GGGGTCCCTGAGACTCTCCTG

NexteraR2-Hu-VH4-FW1:GTCTCGTGGGCTCGGAGATG TGTATAAGAGACAGCTTC

GGAGACCCTGTCCCTCACCTG

NexteraR2-Hu-VH5-FW1:GTCTCGTGGGCTCGGAGATG TGTATAAGAGACAGCGGG

GAGTCTCTGAAGATCTCCTGT

NexteraR2-Hu-VH6-FW1:GTCTCGTGGGCTCGGAGATG TGTATAAGAGACAGTCGC

AGACCCTCTCACTCACCTGTG

NexteraR1-Hu-JHmix1:TCGTCGGCAGCGTCAGATGTG TATAAGAGACAGTACGTNC

TTACCTGAGGAGACGGTGACC

NexteraR1-Hu-JHmix2:TCGTCGGCAGCGTCAGATGTG TATAAGAGACAGCTGCNCT

TACCTGAGGAGACGGTGACC

NexteraR1-Hu-JHmix3:TCGTCGGCAGCGTCAGATGTG TATAAGAGACAGAGNCTTA

CCTGAGGAGACGGTGACC

2. Hu IGH amplification for clone tracking:
   These primers use dual ID barcodes to distinguish them from the identification sample amplicons (the bold font indicates the barcode sequences).

NexteraR2-Barcoded-Hu-VH1-FW1: GTCTCGTGGGCTCGGAGATGTGTATAAGAGAC

AG**AGGCTATA**GGCCTCAGTGAAGGTCTCCTGCAAG

NexteraR2-Barcoded-Hu-VH2-FW1: GTCTCGTGGGCTCGGAGATGTGTATAAGAGAC

AG**GCCTCTAT**GTCTGGTCCTACGCTGGTGAAACCC

NexteraR2-Barcoded-Hu-VH3-FW1: GTCTCGTGGGCTCGGAGATGTGTATAAGAGAC

AG**AGGATAGG**CTGGGGGGTCCCTGAGACTCTCCTG

NexteraR2-Barcoded-Hu-VH4-FW1: GTCTCGTGGGCTCGGAGATGTGTATAAGAGAC

AG**TCAGAGCC**CTTCGGAGACCCTGTCCCTCACCTG

NexteraR2-Barcoded-Hu-VH5-FW1:
GTCTCGTGGGCTCGGAGATGTGTATAAGAGAC
AG**CTTCGCCT**CGGGGAGTCTCTGAAGATCTCCTGT
NexteraR2-Barcoded-Hu-VH6-FW1:
GTCTCGTGGGCTCGGAGATGTGTATAAGAGAC
AG**TAAGATTA**TCGCAGACCCTCTCACTCACCTGTG
NexteraR1-Barcoded-Hu-JHmix4:
TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG
ATTACTCGTACGTNCTTACCTGAGGAGACGGTGACC
NexteraR1-Barcoded-Hu-JHmix5:
TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG
TCCGGAGACTGCNCTTACCTGAGGAGACGGTGACC
NexteraR1-Barcoded-Hu-JHmix6:
TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG
CGCTCATTAGNCTTACCTGAGGAGACGGTGACC

3. NexteraXT index primers S5XX and N7XX. These primers are synthesized with HPLC purification to create sets A, B, C, and D for different barcode combinations (available from Illumina).

2.2 DNA Extraction

Use molecular biology-grade reagents.

1. Isopropanol, DNase/RNase free.
2. 200 proof ethanol.
3. 10 mM Tris and 0.1 mM EDTA, pH 8.0.
4. DNase/RNase-free water.
5. Glycogen.
6. RBC lysis solution (Qiagen).
7. Cell lysis solution (Qiagen).
8. Protein precipitation solution (Qiagen).
9. RNase A solution (Qiagen).
10. DNA-off (Thermo Fisher Scientific).

2.3 Library Preparation

1. Multiplex PCR Kit (Qiagen).
2. Ultrapure agarose (Thermo Fisher Scientific).
3. 100 bp DNA Ladder (New England Biolabs).
4. 50× Tris-acetate electrophoresis buffer (Quality Biological)
5. Agencourt AMPure XP beads (Beckman Coulter).
6. Gel Extraction Kit (Qiagen).
7. 3 M sodium acetate, pH 5.5 (Sigma).
8. 10 mg/ml ethidium bromide aqueous solution (Sigma-Aldrich).

2.4 Library QC and Sequencing

Use molecular biology-grade solutions.

1. 10 M Sodium hydroxide solution, BioUltra (Sigma-Aldrich).
2. 600-Cycle MiSeq Reagent Kit v3 (Illumina).
3. Qubit dsDNA High-Sensitivity Kit (EMSCO/Thermo Fisher Scientific).
4. KAPA Library Quantification Kit (EMSCO/Thermo Fisher Scientific).
5. PhiX Control V3 Kit (Illumina).
6. Tween 20 (Sigma-Aldrich).

2.5 Disposables

1. 96-well ABI-style PCR plate (Thomas Scientific).
2. Plate-sealing film, aluminum, cold storage, and sterile (Thomas Scientific).
3. Microseal "B" adhesive seals for thermo cycling (Bio-Rad Laboratories).
4. 1.5 ml Posi-Click tubes (Thomas Scientific).
5. Reagent reservoir, 25 ml, sterile, and individually wrapped (Thomas Scientific).
6. DNA LoBind tubes, 1.5 ml (Eppendorf).
7. PCR plate 96 LoBind, semi-skirted (Eppendorf).
8. Qubit assay tubes (Life Technologies).
9. Polypropylene conical tube 15 ml bulk wrap sterile (Thomas Scientific).
10. P2/P10 extra-long filter pipet tips (Thomas Scientific).
11. P-20 filter pipet tips (Thomas Scientific, *see* **Note 2**).
12. P-200 filter pipet tips (Thomas Scientific).
13. P-1000E filter pipet tips (Thomas Scientific).

2.6 Equipment

1. Veriti 96-well thermal cycler.
2. NanoDrop 1000 spectrophotometer.
3. Agilent 2100 bioanalyzer.
4. Qubit 4 fluorometer.
5. Illumina MiSeq.
6. PCR workstation (C.B.S. Scientific).
7. 96S super ring magnet plate (Thomas Scientific).
8. Labnet mini plate spinner (Thomas Scientific).
9. Gel Doc XR Imaging System with Universal Hood II (Bio-Rad).
10. Owl™ EasyCast™ B2 mini gel electrophoresis system.

3 Methods

The major steps of the wet bench procedure are outlined in Fig. 1.

3.1 Lab Setup

The lab should have separate areas for pre-PCR and post-PCR work, to prevent contamination. Two separate rooms are recommended with all DNA extraction: the PCR setup is performed in the pre-PCR room, and all of the gel running, AMPure bead purification, and sequencing are performed in the post-PCR room. Use DNA-off to wipe down the workstation and UV treat pipets before and after each experiment.

3.2 DNA Purification

This protocol starts from high-purity genomic DNA (gDNA) that has been isolated from a population of cells, such as peripheral blood mononuclear cells (PBMCs), or cells from tissues or sorted cells (Fig. 1a and *see* **Note 3**). To prepare the sorted cells for sequencing, sort cells directly into 300 μl of cell lysis solution if

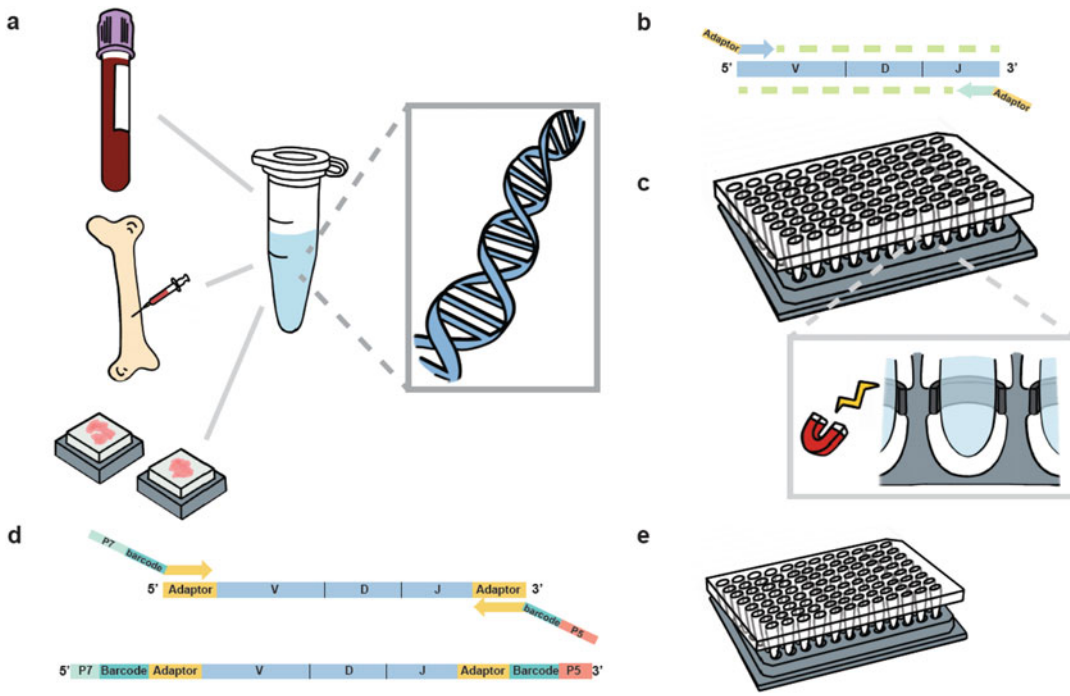


Fig. 1 Workflow for IGH sequencing from bulk gDNA. (**a**) Starting from PBMCs, bone marrow aspirate, or formalin-fixed paraffin-embedded samples, gDNA is extracted from bulk populations. (**b**) Next, IGH gene rearrangements are amplified from gDNA using primer cocktails in FR1 and JH along with Illumina adapters. V = variable, D = diversity, and J = joining genes. (**c**) Amplicons from this first round of PCR are purified using AMPure beads and (**d**) subjected to second-round amplification using primers that include sample barcodes (see primers in Subheading 2.1 for DNA sequence information). (**e**) Sequencing libraries are subjected to further purification, size selection, quality control, and pooling prior to loading onto the sequencer

the expected lymphocyte yield is less than 50,000 cells and use a DNA LoBind tube. If the expected yield is more than 50,000 cells per population, sort the cells into sorting buffer, centrifuge the cells, remove the supernatant, and resuspend the cell pellet in a cell lysis buffer (add 300 µl cell lysis buffer for up to two million cells). DNA is extracted from whole blood, bone marrow, or sorted cells using protocols from Gentra Puregene (Qiagen) handbook using the manufacturer's recommendations. Outlined below is the protocol (with notes) for 3 ml of whole blood.

1. Mix 3 ml whole blood with 9 ml of RBC lysis solution in a 15 ml conical tube, and gently invert ten times. Incubate at room temperature for 5 min (minutes), and invert the mixture at least once during the incubation (*see* **Note 4**).
2. Centrifuge at 2000 × *g* for 2 min to pellet the white blood cells. Carefully discard the supernatant by pipetting or pouring the supernatant to a waste tank containing water with 10% bleach. Leave behind ~200 µl of the residual liquid, and vortex to resuspend the pellet in the residual liquid (*see* **Note 5**).
3. Add 3 ml of cell lysis solution, and vortex for 10 s (seconds). Stopping point: Once samples are fully suspended in cell lysis solution, the DNA will be stable for 2 years (*see* **Notes 6** and 7).
4. Add 1 ml protein precipitation solution, and vortex for 20 s at high speed. Centrifuge at 2000 × *g* for 5 min (*see* **Note 8**).
5. Transfer the supernatant to a new 15 ml conical tube, and add 3 ml isopropanol. Mix by inverting 30 times until the DNA is visible as threads or a clump (*see* **Note 9**). Stopping point: DNA can be precipitated overnight.
6. Centrifuge at 2000 × *g* for 3 min. After centrifugation, the DNA may be visible as a small white pellet. Carefully pour off the supernatant to a waste isopropanol/ethanol container. Drain the residual liquid in the tube by inverting on a clean piece of absorbent paper.
7. Add 3 ml of 70% ethanol (prepared with molecular biology grade ethanol and DNase/RNase-free water), and invert several times to wash the DNA pellet. Centrifuge at 2000 × *g* for 1 min, and carefully pour the supernatant to a waste isopropanol/ethanol container. As the DNA pellet may be loose at this step, decant the liquid from the tube carefully.
8. Perform a quick spin at 2000 × *g* for 30 s to bring down the residual ethanol to the bottom of the tube, and use a P-200 µl filter tip to remove the residual ethanol. Allow the DNA pellet to air dry for 10 min or until no ethanol can be observed. Note: Avoid overdrying the DNA pellet, as then the DNA will be difficult to dissolve.

9. Resuspend DNA in 100 ul of TE (low EDTA) buffer (10 mM Tris and 0.1 mM EDTA), and check the DNA quality using NanoDrop. The OD_{260}/OD_{280} ratio should be close to 1.8. If the DNA concentration is <100 ng/μl using the NanoDrop instrument, repeat the DNA concentration measurement using Qubit HS DNA Kit for a more accurate measurement.

3.3 Template Amplification and Initial Quality Control

Before beginning, make sure that all of the workstations are clean, and perform all template amplification procedures in a separate pre-PCR area. Aliquot all primers (equimolar mixture of primers for both VH and JH primer mixes), PCR-grade water, and PCR master mix buffers before use. The PCR product that is amplified from gDNA is shown in Fig. 1b.

1. Use water and PCR master mix from Qiagen Multiplex PCR Kit, and prepare the PCR mix (*see* **Notes 10–13**):

Reagent	Volume (μl)
DNA	4
2× master mix PCR buffer	12.5
5′ VHFW1 mix (5 μM)	3
3′ JH mix (5 μM)	3
Nuclease-free water	2.5
Total volume	**25**

2. Thermal cycling. If using plates, use microseal B adhesive seal. Perform a quick spin of the plate before loading onto the thermal cycler, and run the following program:

   First PCR program.

Temperature and time	Number of cycles
95 °C 7 min	1
95 °C 45 s, 60 °C 45 s, 72 °C 90 s	35
72 °C 10 min	1

   Stopping point: Amplified samples can be stored at 4 °C for up to 48 h.

3. Agarose gel electrophoresis of PCR products (*see* **Note 14**). Gel electrophoresis is performed to ensure that the first-round PCR has generated a sufficient quantity of amplicons of the correct length and that there is no evidence of contamination in the negative controls.

(a) Prepare a 2% agarose gel (ultrapure agarose) in 1× TAE buffer. Ethidium bromide can be mixed into the gel, or the gel can be stained afterward to visualize the DNA.

(b) Load 5 μl of the first-round PCR amplicons per lane, and check the amplicon size under UV light by comparing the products to the molecular weight ladder (100 bp ladder).

(c) The amplicon on the gel should be the same as the positive control band, and the size is ~440 bp. If water and fibroblast DNA controls show contamination, the whole experiment needs to be rerun.

4. AMPure bead purification (Fig. 1c) is performed on the remaining 20 μl of the sample from the first-round PCR to enrich for PCR products of the appropriate length and remove primers and primer dimers following the manufacturer's protocol (Beckman Coulter). Aliquot beads before use. Equilibrate beads to room temperature, and prepare fresh 85% ethanol each time before use. Use filter tips.

(a) Mix an equal volume of AMPure beads with amplicons (in this case, 20 μl of beads, *see* **Note 15**).

(b) Mix the beads and amplicons together by pipetting up and down 20 times. Incubate the mixture at room temperature for 1 min. Set the plate on the magnet for 5 min until the mixture is clear.

(c) Keep the PCR plate on the magnet, remove the supernatant, and discard.

(d) Wash the beads by adding 180 μl of fresh 85% ethanol, *do not mix*, incubate at room temperature for 30 s, and remove and discard the supernatant.

(e) Use P2/P10 extra-long tips to remove the residual ethanol from each well, and air dry at room temperature for up to 5 min. Note: Do not allow beads to air dry for more than 5 min.

(f) Remove the PCR plate from the magnet, add 40 μl of TE (low EDTA) buffer into each sample well. Mix by pipetting up and down ten times to resuspend the beads. Incubate at room temperature for 2 min.

(g) Return the PCR plate to the magnet, and incubate at room temperature for 5 min. With the plate on the magnet, transfer 38 μl of the eluates to a new 96-well PCR plate. Stopping point: At this step, the new plate with the purified first PCR amplicons can be sealed and stored at −20 °C for later use.

3.4 Second-Round PCR and Product Purification

In this section of the protocol, the bead-purified amplicons from the first step are amplified using primers that are tagged with Illumina barcodes. A schematic illustration of the PCR product is shown in Fig. 1d. All procedures for preparing the PCR mix are performed in the pre-PCR room, *except* for the addition of the first-round PCR amplicons, which is performed in a PCR hood in the post-PCR room. Aliquot all primers (Nextera XT index primers), PCR-grade water, and PCR master mix buffers before use.

1. Use water and PCR master mix from the Qiagen Multiplex PCR Kit, and prepare the PCR mix:

Reagent	Volume (μl)
Purified first-round PCR amplicons	4
2× master mix PCR buffer	12.5
NexteraXT index primer S5XX	2.5
NexteraXT index primer N7XX	2.5
Nuclease-free water	3.5
Total volume	**25**

2. Run the second-round PCR program:

Temperature and time	Cycles
95 °C 10 min	1
95°C 30 s, 60°C 30 s, 72°C 45 s	8
72°C 10 min	1

Stopping point: Amplified samples can be stored at 4 °C for up to 48 h.

3. Sample pooling and analysis of pooled second-round PCR products (*see* **Note 16**).
   (a) Add equal volumes (typically ~5 μl) of the individual sample amplicons (replicates) together into a "pooled library" for sequencing. Samples can be pooled together at this stage, because the amplicons have sample-specific barcodes.
   (b) Prepare a 2% agarose gel, and add 5 μl of the second-round PCR amplification mixture. The expected amplicon size on the gel is ~510 bp and should be present in the positive control sample. If water or fibroblast have amplification products, the second-round PCR experiment needs to be rerun. Stopping point: The second-round PCR samples can be stored at −20 °C in a post-PCR freezer for later use (*see* **Note 17**).

4. Optional gel extraction step. If primer dimers are observed at the size of ~200 bp, a gel purification step using QIAquick Gel Extraction Kit is recommended to enrich for products of the right length for sequencing. Gel extraction is preferred over AMPure beads for this step, because the beads do not remove this amplicon size well.
   (a) Run the pooled samples on a 2% agarose gel with a low-voltage setting (~60 V) to allow the amplicons to migrate slowly on the gel.
   (b) After 3 h of gel running, cut out the expected size (510 bp) band under long wavelength UV light to minimize DNA damage. Weigh the gel slice in a 1.5 ml Eppendorf tube.
   (c) Add 3 volumes of buffer QG to 1 volume of gel (100 mg gel corresponds to ~100 μl of liquid volume). The maximum amount of gel per spin column is 400 mg. Incubate at 50 °C for 10 min (invert the tube to help dissolve gel) or until the gel slice has dissolved completely.
   (d) If the color of the mixture is orange or violet, add 10 μl of 3 M sodium acetate until the color turns yellow. Add 1 gel volume of isopropanol to the sample, and mix by inverting the tube ten times.
   (e) Apply 750 μl of the gel-isopropanol mixture to a QIAquick spin column in the provided 2 ml collection tube, and centrifuge at 17,900 × *g* for 1 min.
   (f) Discard the flow-through, and place the QIAquick column back into the same tube.
   (g) Apply the rest of the mixture (if any is remaining) to the same column, and repeat steps 4e and 4 f.
   (h) Add 750 μl buffer PE to the QIAquick column, and centrifuge at 17,900 × *g* for 1 min to wash the column. Discard flow-through, and place the QIAquick column back into the same collection tube.
   (i) Centrifuge the QIAquick column for 1 min to remove the residual wash buffer, and place the QIAquick column into a clean 1.5 ml Eppendorf tube.
   (j) Add 50 μl buffer EB to the center of the QIAquick membrane, let the column stand for 2–3 min, and then centrifuge for 1 min. Stopping point: Gel-purified product (the eluate in the clean 1.5 ml Eppendorf tube) can be stored at −20 °C in the post-PCR freezer for later use.

3.5 Library Pooling, Purification, and Quantification

1. Gather up all of the pooled libraries and gel-purified pooled libraries, if any, that are going to be included in the sequencing run (*see* **Note 18**).

2. Starting from the pooled libraries, perform two rounds of AMPure bead purification as described previously (Fig. 1e).
3. The final purified libraries can be eluted in ½ or ¼ of the original pooled sample volume to concentrate the library, if needed.
4. Run 1 μl of each pooled library with a Bioanalyzer high-sensitivity DNA assay to verify the size and purity. Check the concentration of each pooled library on Qubit using a dsDNA High-Sensitivity Kit with 2 μl of each pool.
5. Once the molarity is calculated for each pooled library (*see* **Note 19**), normalize the library inputs in the sequencing run. The goal is to have the number of molecules per sample be equal across the different libraries. For example, suppose that one pooled library (library A) has 34 samples with an overall molarity of 50 nM and a second pooled library (library B) has 46 samples with a molarity of 35 nM. If 10 μl of library B is used for the final pooled library, then the volume of library A is calculated by solving for A in the following expression:

$$(\text{A } \mu\text{l} \times 50 \text{ nM})/34 \text{ samples} = (10 \ \mu\text{l} \times 35 \text{ nM})/46 \text{ samples}.$$

$$\text{A} = 5.17 \ \mu\text{l}.$$

The concentration of the final pooled library is determined by Qubit and calculated as molarity (*see* **Note 19**).

3.6 Sequencing

1. Prepare a fresh dilution of 0.2 N NaOH. Dilute the original 10 N NaOH to 1 N, and discard the aliquot after 3 months. Mix 80 μl of water and 20 μl of 1 N NaOH for a total of 100 μl of 0.2 N NaOH.
2. Prepare 4 nM of the final pooled sequencing library by diluting the concentrated one with TE (low EDTA).
3. Mix 5 μl of 0.2 N NaOH and 5 ul of 4 nM library by pipetting up and down for 20 times in a 1.5 DNA LoBind tube. Denature at room temperature for 5 min.
4. Add 990 μl prechilled HT1 (from the MiSeq Kit), and incubate on ice immediately. The final concentration for the denatured library is 20 pM.
5. Prepare 20 pM of PhiX. Mix 2 μl of PhiX control with 3 μl of TE (low EDTA) in a 1.5 ml DNA LoBind tube by pipetting. Add 5 μl of freshly diluted 0.2 N NaOH, mix by pipetting up and down 20 times, and incubate at room temperature for 5 min. Next, add 990 μl prechilled HT1 (from the MiSeq Kit), and incubate on ice immediately (*see* **Note 20**).
6. To spike in 10% PhiX into the final sequencing library, take 100 μl of the 20 pM denatured library out and discard, and add in 100 μl of 20 pM denatured PhiX. This will yield 20 pM of the final sequencing library with 10% PhiX (*see* **Note 21**). Load

600 μl of this library to the pre-thawed MiSeq cartridge MiSeq® Reagent Kit v3 (2X300 cycles). The run takes 2.5 days to complete.

7. General sequencing run QC. For the MiSeq (2X300 cycle) V3 Kit, the optimal raw cluster density is 1200–1400 K/mm^2 (Illumina provides additional details on clustering density online). The percentage of reads for the entire run that have Q scores above 30 (Q30, 1 in 1000 base calls may be incorrect) should be at least 70%. Finally, the percentage of clusters passing filter (PF%) should be $> = 80\%$. If a run does not pass all three of these thresholds, the sequencing should be repeated. Under passing conditions, each replicate has on average 100,000 to 300,000 valid reads (using pRESTO processing with Q30 filtering, please see following sections for data analysis).

3.7 Software Installation

Before processing raw sequencing data, analysis software must be installed as follows:

1. Install pRESTO. pRESTO [12] is used for quality control prior to running the rest of the pipeline. It can be installed with `pip3 install presto`.
2. Install the ImmuneDB Docker image. The ImmuneDB [13] Docker image should be installed for gene identification (via pre-installed IgBLAST), clonal inference, database-backed storage, exporting, and a web interface. For illustrative purposes, we will use version 0.29.10, which can be pulled with `docker pull arosenfeld/immunedb:v0.29.10`.

3.8 Raw Data Processing

Raw data from NGS platforms are generally output in a format providing base calls for each read along with a quality score for each base. Depending on the sequencing method, there are a number of different steps to transform and filter these data into a format that is readily available for further analyses. In general, if reads are paired, the matching 5′ and 3′ reads must be aligned to form full-length sequences. Specifically, each pair of reads is iteratively compared until the maximal number of overlapping nucleotides is found. Nucleotides in the overlapping segment that do not match are assigned the base from whichever read has a higher-quality score.

Following this, short and low-quality sequences should be removed as they do not provide sufficient information to make accurate gene calls. Then, primer sequences which were incorporated into the DNA/RNA templates should be masked as not to skew later mutation analyses. Individual base calls with low confidence (generally either a Phred score < 20 or < 30) should be masked to reduce their influence on downstream analyses. Finally, genes should be annotated with IgBLAST for downstream processing. The commands for this entire process, assuming paired input

files from an Illumina-based sequencing platform and applying a Phred quality score filter of 30, are as follows:

1. Locate the sequencing FASTQ files. First, change the working directory to where the sequencing is located. For example, if the data are in `$HOME/seq_data`, run `cd $HOME/seq_data`.
2. Run pRESTO:

```
PairSeq.py -1 *R1*.fastq -2 *R2*.fastq
AssemblePairs.py align -1 *R1*_pair-pass.fastq \
 -2 *R2*_pair-pass.fastq \
 --coord illumina
FilterSeq.py quality -s *assemble-pass.fastq
FilterSeq.py trimqual -s *quality-pass.fastq -q 30 --win 20
FilterSeq.py length -s *trimqual-pass.fastq -n 100
FilterSeq.py maskqual -s *length-pass.fastq -q 30
FilterSeq.py missing -s *maskqual-pass.fastq -n 10
```

3. Move the quality-controlled data into a new directory. The remaining steps of this method only use the final resulting files which will end in `missing-pass.fastq`. These files should now be moved to a location to mount into the ImmuneDB Docker container.

```
mkdir $HOME/immunedb_share/input
mv *missing-pass.fastq $HOME/immunedb_share/input
```

4. Annotate raw sequences which have passed general quality control filters with gene information. For IGH sequences V, D, and J genes should be associated with each sequence. IgBLAST is the preferred annotation tool which provides AIRR-compliant output for gene calls in addition to other alignment information [14]. For ease, IgBLAST and a helper script are included in the ImmuneDB Docker image. To begin annotation, perform the following:
   (a) Run the docker container.
   To begin an interactive session, run the following:

```
docker run -v $HOME/immunedb_share:/share \
-p 8080:8080 \
-it arosenfeld/immunedb:v0.29.10
```

One should see output similar to the following, after which a terminal prompt will be shown:

```
Moving MySQL to Volume
* Starting MariaDB database server mysqld [ OK ]
Setting up database
Starting webserver
```

(b) Run IgBLAST on the QC'd FASTQ files. In the Docker container, a helper script run_igblast.sh can be used to annotate sequences. Reference genes are provided for humans and mice for IGH, IGL, IGK, TRA, and TRB. In this protocol, we will focus on human IGH. Run the following:

```
run_igblast.sh human IGH /share/input /share/input
mkdir -p /share/sequences
mv /share/input/*.fast[aq] /share/sequences
```

After this step, TSV files annotated in AIRR format [15] will be located in the Docker container at `/share/input` (which is also accessible at `$HOME/immunedb_share/input` on the host).

3.9 Importing Metadata and Sequence Data into ImmuneDB

1. Specifying sample metadata. Prior to importing these annotated data into ImmuneDB for clonal inference and downstream analyses, metadata must be specified for each sample file.

   (a) Create a template metadata file. Although a metadata file is simply a TSV which could be created manually, ImmuneDB provides a helper script to create a template as follows:

```
cd /share/input
immunedb_metadata --use-filenames
```

   (b) Add relevant metadata**.** With the command above, a metadata file with one row per file will be generated, and the sample name for each file will be set to the filename stripped of its extension.

   On the **host**, open the metadata file in a spreadsheet editor. The headers included by default are required; file_name and sample_name will already be filled in from the previous step, but the `study_name`, `subject` must be filled in (*see* **Note 22**).

2. Importing sequences into ImmuneDB. The data are now ready for importing and further processing before clonal inference. To do so, in the Docker container, run the following steps.

   (a) Create a database for the project. The first step is to create a database into which the AIRR-compliant sequencing data annotated by IgBLAST will be stored. For this method we will call the database `my_db`, but it can be any valid name for a MySQL database (*see* **Note 23**).

```
immunedb_admin create my_db /share/configs
```

(b) Import the annotated data and trace duplicate sequences. The next commands import all the annotated sequences into the previously created database and annotate (collapses) duplicate reads within and between samples. Counting duplicates is useful for downstream filtering and clone size estimation.

```
immunedb_import /share/configs/my_db.json airr \
/root/germlines/igblast/human/IGHV.gapped.fasta \
/root/germlines/igblast/human/IGHJ.gapped.fasta \
/share/input \
--trim-to 80
immunedb_collapse /share/configs/my_db.json
```

One important parameter in the previous commands is `--trim-to`. This masks the bases on the 5′ end of each read with the ambiguity character *N*. This avoids the primer sequences, which are incorporated into the resulting reads, from being incorporated into downstream mutational analyses. The value of 80 was chosen for this chapter due to the use of framework 1 (FWR1) primers. If different primers are used, the IMGT position of the 3′ end of the primer sequence should be used instead.

3.10 Clonal Inference from Sequencing Data and General Statistics

1. Once the data are imported and collapsed, sequences likely originating from a common progenitor cell can be grouped into clones.

```
immunedb_clones /share/configs/my_db.json cluster
```

The default parameters used by immuneDB to specify clonally related sequences are the use of the same IGHV and IGHJ genes, the same CDR3 length, and at least 85% amino acid sequence similarity in the CDR3 (*see* **Note 24**).

2. Calculating statistics. To make downstream analyses more efficient, ImmuneDB pre-calculates a number of statistics about clones and samples (*see* **Note 25**).

```
immunedb_clone_stats /share/configs/my_db.json
immunedb_sample_stats /share/configs/my_db.json
```

3. Create lineage trees for each clone. Optionally, lineage trees can be constructed for each clone. Like clonal inference, this process has many parameters, and the following is for general use and may need to be tweaked depending on sequencing depth, error rates, and the underlying biological samples:

```
immunedb_clone_trees /share/configs/my_db.json --min-seq-co-
pies 2
```

More details on clonal lineages can be found in Subheading 3.3 of the chapter "AIRR Community Guide to Repertoire Analysis."

3.11 Analysis of Clone Numbers and Size Distributions

1. Sample clone count (*see* **Note 26**). One can do a quick "back-of-the-envelope" calculation to estimate the maximal number of expected unique IGH rearrangements in a bulk gDNA sequencing using the equation below [16] if the nanogram input is known:

$$\text{Max.\#of rearrangements} = (\text{ng input})\,(1000\ \text{pg/ng}) \times (1.4\ \text{rearrangements/cell})/6.7\ \text{pg/cell}.$$

   Or, equivalently, about 150 cells per nanogram of input DNA. These equations assume that 100% of the cells in the samples are the B or T cells of interest that there is quantitative recovery of all possible rearrangements and that each cell has an average of 1.4 rearrangements (due to some cells having more than one IGH or TRB rearrangement [17], *see* **Note 27**). Obtaining fewer or more clones than expected can reveal potential technical or analytical problems with the experiment or data analysis pipeline, respectively (*see* **Note 28**).

2. Clone size distribution. There are at least two size metrics which can be applied to each clone to generate distributions of estimated clone sizes. First, one can define the size of each clone as its number of sequence copies. This metric is particularly useful when looking at malignancies where nearly all copies reside in the same single clone (or a small number of clones). However, this approach can also be affected by PCR jackpots, sample-specific subset differences, or other inter-sample copy number variability (*see* **Note 29**). A second approach is to compute the number of unique sequences in each clone. This metric can be influenced by the level of SHM in the clone. It can also be affected by sequencing error, with larger numbers of unique sequences per clone arising in more deeply sequenced samples.

   (a) Histogram of top-ranked clones. As shown in Fig. 2a, one can plot the copy number fraction of the 20 clones in a sample that have the highest copy numbers. Investigating the top copy number clones in datasets can highlight expanded clones as compared to the overall repertoire, giving insight into a range of different biological processes (*see* **Note 30**). In healthy individuals, expanded B-cell clones in the peripheral blood generally have copy numbers within the same order of magnitude of non-expanded clones (*see* **Note 31**).

   (b) D_x index. One can compute the fraction of sequence copies that are occupied by the top x percent of clones in a sequencing library. D_x is the fraction of total copies

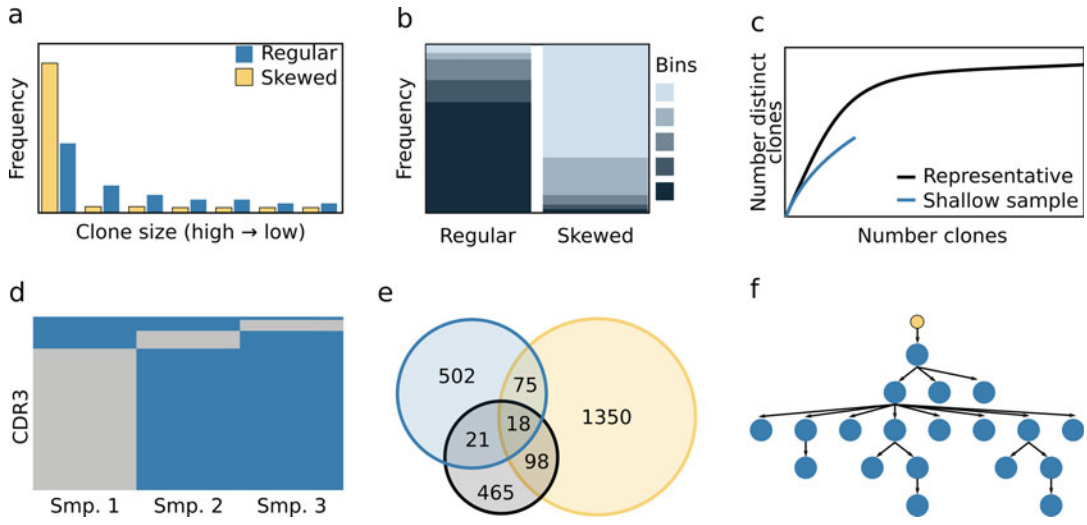


Fig. 2 Clone visualization scheme. All plots are illustrative. (**a**) Top clone plot. An example plot showing the size of the top clones as measured by copy number in two samples, one shown in blue and one in yellow. Each set of columns represents the clone of a given rank, and the y-axis shows the copy number frequency as a fraction of the entire sample. (**b**) Clone rank plot. An example of a clone rank plot for two samples. Each bar represents a sample; each color represents the copy number fraction for a bin of clones of a given range of ranks (sizes) with lighter blue indicating higher-ranked (larger) clones and darker blue representing lower-ranked (smaller) clones. A generally darker sample indicates that the majority of clones are not expanded, and a lighter sample indicates a more oligoclonal repertoire. (**c**) Rarefaction curves. Illustrative rarefaction curves for two hypothetical samples showing sufficient and insufficient sampling. The x-axis indicates number of clones, and the y-axis indicates the measured number of total (unique) clones. Curves in which the number of distinct clones continues to increase as the number of sampled clones increases indicate potential under-sampling (blue), whereas curves that begin to plateau (black) indicate the sampled clones are becoming more representative of the true underlying clonal population. (**d**) Clonal string plots visualizing the degree of clonal overlap between three samples. Each row represents a clone and each column a sample (smp). The presence of a clone in a given sample is indicated by blue and its absence by gray. Only clones that overlap in two or more samples are shown. (**e**) Venn diagram. Three different hypothetical samples (demarcated by the blue, yellow, and black circles) from the same individual. Numbers indicate clone counts that are found uniquely in one, two, or three of the samples. (**f**) Clonal lineage. An inferred hypothetical lineage of clonally related sequences. Each blue node represents a unique sequence, and the yellow node represents the nearest germline reference sequence. The edge length between two nodes indicates the total number of accumulated mutations from the parent sequence to the child sequence

occupied by the top x clones. A common value of x is 20 [10] which, when looking at copy number distribution, reveals if there are one or more dominating clones.

(c) Clone rank plot. Unlike the top-ranked clone plot and D_x index, clone rank plots provide a snapshot of the clone size distribution in the entire repertoire. Clone rank plots achieve this by segregating clones by rank as shown in Fig. 2b. In such plots, each column represents a sample, or a pool of samples, and the height of each bar represents the proportion of copies in the given clonal range bracket.

For example, in this example, the red bars show the proportion of sequence copies in the top ten ranked clones. In oligoclonal repertoires, both the D_x index and the clone rank plot, the top copy clones contain the majority of copies. In contrast, for polyclonal repertoires, range plots can provide a nuanced view of clonal abundance by stratifying clones into categories based on their copy number distributions.

3.12 Clonal Overlap Analysis

Determining how many samples or replicates are necessary to sufficiently reveal the clonal landscape of the underlying immune repertoire is challenging. Undersampling a repertoire can lead to underpowered analyses and false biological conclusions (e.g., claiming lack of overlap), whereas oversampling can be expensive and time-consuming.

1. Rarefaction analysis can provide insight into the level of sampling, providing a means of powering the clonal overlap analysis. Rarefaction stems from ecology where one wants to estimate the total number of species in a region by repeated sampling of organisms [18]. Species which occur in multiple independent samples are likely more abundant than those which only occur in a few samples. One can apply the same principles to the analysis of clones (*see* **Note 32**). This analysis is generally plotted as shown in Fig. 2c where the x-axis is the sample size (e.g., the number of clones sampled) and the y-axis is the diversity (e.g., the estimated total number of clones). Curves that plateau indicate that sampling more clones will likely not affect the total estimated number of clones in the underlying population. On the other hand, curves which do not flatten show that a larger sample size is required to reveal additional unknown clones.
2. Visualizing and quantifying clonal overlap. Clones which are found in multiple samples are of particular interest. The sizes of clones which are present at baseline and after treatment can reveal insights into the efficacy of treatment and a patient's response to therapy. Clonal overlap can also be applied to samples which were acquired from different tissues, cell subsets, and other biologically relevant populations.
   (a) Clone definitions for the evaluation of clonal overlap. Most frequently used are clonal annotations or shared CDR3 amino acid sequences. In ImmuneDB, for example, clones are annotated with a unique clone ID that can be scanned across all of the samples in a given subject, allowing for the construction of clonal networks across all of the different samples in an individual. Alternatively, one can trace the consensus CDR3 amino acid sequence of each clone through samples to determine overlap.

(b) The Jaccard index [19] is the cardinality of the intersection of two samples divided by the cardinality of the union of the same samples. Specifically, for two (potentially overlapping) sets of clones A and B, the Jaccard index J is calculated with

$$J = \frac{A \cap B}{A \cap B}$$

(c) Cosine similarity. The cosine similarly also gives an indication of overlap between samples. However, unlike the Jaccard index, it takes into account clone size rather than only presence or absence in samples. For each of the two samples to compare, a one-dimensional vector is constructed, the values of which indicate the size of each clone in copies. The order of clone sizes must be the same for both samples. Specifically, given two vectors of clone sizes from two samples, A and B, the cosine similarity S is defined as

$$S = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i} \sqrt{\sum_{i=1}^{n} B_i}}$$

(d) Overlapping clones can be visualized in line plots (Fig. 2d) in which each column is a sample and each row is a clone (line). The lines can be heat mapped to indicate the abundance (e.g., copy number fraction) of a clone in a given sample. Line plots only show clones that overlap in two or more of the analyzed samples. To gain insight into the fraction of overlapping clones in each sample and their distribution, Venn diagrams (Fig. 2e) can be used. Venn diagrams show the numbers of overlapping and nonoverlapping clones but become difficult to visualize when four or more samples are being compared.

(e) Clones can be further visualized in multidimensional datasets. The temporal relationship of sequences derived from the same clone can be further visualized as lineages which are rooted, directed graphs, showing the progression of mutations (Fig. 2f). This can be useful to analyze the changes within each clone across tissues, subsets, timepoints, or other metadata of interest. There are multiple ways to infer lineages from a collection of clonally related sequences. Two common approaches are neighbor joining and maximum parsimony. Neighbor joining begins with every sequence being its own node and iteratively adds parent nodes between those which are most similar [20]. Maximum parsimony [21] takes as input the same

sequences but instead attempts to construct a tree which requires the minimum number of total mutations. Both have positives and negatives. For example, neighbor joining can create trees which are not optimal (e.g., mutations occurring multiple times or incorrectly grouping clades), but it is computationally more efficient than maximum parsimony. Maximum parsimony, however, guarantees some properties of the tree such as minimizing its height, but is computational intractable to calculate for large clonal lineages.

4 Notes

1. Sample barcodes can become associated with the wrong sample in the flow cell during sequencing (i.e., failure to accurately demultiplex the samples in a sequencing run), a phenomenon known as cross-clustering. For example, a very large clone in one sample can sometimes be found at very low copy number in an unrelated individual. Cross-clustering can also occur when the cluster density is too high. With dual indices, there is a second barcode that links a sample with an individual, providing a means of computationally resolving issues with cross-clustering [22].
2. All micropipet tips used in this protocol are SHARP® Precision Barrier Tips (available from Thomas Scientific) that use low retention polymer technology. Some of the aerosol-resistant tips from other vendors trap liquids.
3. The recommended DNA inputs for different samples are up to 1 μg per reaction for formalin-fixed paraffin-embedded tissue (particularly if lymphopenic by histology), up to 400 ng/reaction for unsorted cells in which the cells of interest make up 5% or more of total nucleated cells or up to 100 ng/reaction for sorted lymphocytes per reaction. Lower input amounts can be used if less sample is available.
4. For fresh blood (collected within 1 h before starting the protocol), increase the incubation time to 8 min to ensure complete red blood cell lysis.
5. If the pellet does not break apart well, flick the bottom of the tube with your fingers.
6. If cell clumps are visible, incubate at 37 °C until the solution is homogeneous.
7. If RNA-free DNA is required, add 15 μl RNase A solution, and mix by inverting 25 times. Incubate for 15 min at 37 °C. Then incubate for 3 min on ice to quickly cool the sample for storage.

8. If the protein precipitation step does not form a tight pellet, incubate on ice for 5 min, and repeat the centrifugation at 4 °C.
9. If no DNA threads or clump is observed, add 5 μl of glycogen (20 mg/ml), and incubate the mix at −20 °C for 1 h.
10. The volumes of DNA and water can be adjusted based on DNA concentration and the experimental input. The volume of input DNA is recommended to be a minimum of 4 μl for pipetting accuracy.
11. When assembling the PCR mix, the DNA should be added last and pipetted up and down ten times using a separate filter tip for each reaction.
12. Include at least two negative controls (such as water, human fibroblast DNA, 50–200 ng) and one positive control (such as pooled human DNA from the spleen or PBMCs from plasmapheresis donors, 50–200 ng).
13. Include up to 48 replicates in 1 96-well plate. Place the samples in every other well in rows and columns to reduce the risk of cross contamination. If fewer than 48 samples are used, spread the samples out on the plate as far apart as possible.
14. Ethidium bromide is a carcinogen. Wear appropriate personal protective equipment (lab coat, gloves, and eye protection), and discard the gel and disposables in the appropriate hazardous waste containers.
15. The beads need to be mixed well by vortexing briefly; then do a quick spin to bring down the leftover beads from the cap, and pipet up and down ten times right before mixing them with the amplicons.
16. All of these steps are performed in the post-PCR room.
17. The pooled library can also be stored at −20 °C until the day of the MiSeq run.
18. For survey-level sequencing, run at least two replicates (independent amplifications starting from gDNA). For deeper sequencing, run three or more replicates.
19. The reading of sample concentration from Qubit is in ng/μl and needs to be converted to nM using this formula: [Concentration by Qubit (ng/μl) × 10^6]/(660 × size of the amplicon in base pairs). The size of the IgH FW1 library amplicon is 510 bp.
20. 20 pM PhiX can be used for up to 3 weeks when aliquoted into LoBind tubes and stored at −20°C.
21. If one is using this method for the first time, a bioanalyzer analysis is highly recommended to evaluate the purity of the final library, and a KAPA quantification is recommended to compare with the Qubit concentration measurement. The

method presented in this chapter uses Qubit for concentration measurement and uses 20 pM of the final library based on the Qubit calculation. Bioanalyzer and KAPA quantification may give different concentrations, and the optimal input library concentrations calculated based on these methods may differ.

22. Additional custom columns may be added for relevant metadata such as tissue, timepoint, etc., following the nomenclature conventions proposed by the AIRR Community [23]. For the updated AIRR-C nomenclature, please visit https://docs.airr-community.org/en/stable/datarep/metadata.html#repertoire-fields
23. The only technical limitations for database names are those documented in the MySQL requirements (https://dev.mysql.com/doc/refman/8.0/en/identifiers.html). In addition, we recommend that the names consist exclusively of lowercase Latin characters, integers, and underscores.
24. There are multiple built-in clonal inference methods including by edit distance and hierarchical clustering, both of which are highly customizable [24–27]. The `–help` command can be used to show additional clustering options.
25. After sequence alignment and clonal inference, it is useful to calculate high-level measures for each replicate (individual sequencing library), sample (pooled replicates), and subject. There are a number of different metrics that can indicate the quality of sequencing and also highlight potentially interesting biological phenomena, such as the number of total reads, valid reads (sequence copies), unique sequences, and clones. Copies, unique sequences, and clones can be heavily influenced by the input DNA amount and cell count. A low number of unique sequences as compared to copies can indicate PCR jackpots or oligoclonality of the underlying repertoire. A number of unique sequences that is similar to the total copies may indicate insufficient sequencing depth.
26. The clone count is influenced by the degree of clonal expansion, diversity of clones, and the amount (and purity) of the cell population of interest. The clone count is also influenced by how clonally related sequences are defined.
27. Typically cells with more than one IGH rearrangement have one productive and one nonproductive rearrangement as cells with two productive IGH [28, 29] or TRB rearrangements [30] are very infrequent. In contrast cells with two productive IGK [31] or TRA rearrangements [32] are more common.
28. If one obtains far fewer clones than expected, possible reasons include clonal expansion, low fraction of T cells or B cells in the sample, poor-quality template (such as an old FFPE sample [33]), technical problem with template amplification or

sequencing such that only a few of the available rearrangements are being amplified, or a filtering procedure that results in an unacceptably large fraction of the data being removed or a clone collapsing procedure that groups unrelated sequences together into the same clones, under-calling the number of different clones. If, on the other hand, one obtains more clones than the predicted maximum number, there may be an issue with the computational pipeline in terms of how clones are defined. For example, if a very high level of sequence similarity is used on a sample enriched for memory B cells with high levels of SHM, clonally related sequences may be grouped falsely into separate clones.

29. If sampling a modest number of cells, in addition to spurious oligoclonality, the experiment may be more susceptible to artifacts such as PCR jackpots, in which one or a few templates "take over" the reaction, leading to misleading evidence of clonal expansion or dominance. The analysis of independently amplified biological replicates can provide important insights into clonal expansions (as discussed in greater detail in the method chapter "Quality Control: Chain Pairing Precision and Monitoring of Cross-Sample Contamination"). Clones which have a high copy number in one replicate but not others suggest a PCR jackpot or a highly oligoclonal sample with very few templates. Conversely, clones which reproducibly have high copy numbers may indeed be expanded.

30. Significant clonal expansions can be encountered in the setting of malignant or nonmalignant lymphoproliferative disorders or during acute immune responses. The degree of clonal expansion is influenced by the cell type under study (e.g., more expanded clones are encountered among memory populations than naive cells), the tissue (e.g., B cells make up a much larger fraction of total cells in the spleen than in the GI tract), and the level of sampling (sequencing libraries that contain fewer clones will have higher average numbers of copies per clone).

31. If one is surveying B cells in the peripheral blood, one can use the copy number fraction and fold-change over the next most frequent nondominant clone to report a significant clonal expansion. For example, one might use a cutoff of 5% of total sequence copies for an IGH rearrangement that is also at least threefold more frequent than the next most frequent nondominant rearrangement. In most individuals the top 20 most frequent clones in the blood typically occupy $<2\%$ of total sequence copies. The additional requirement of being threefold more frequent than the next most frequent nondominant rearrangement helps to limit false calls of significant expansions with oligoclonal samples (which could be due to poor sample quality, a low number of input cells due to B-cell lymphopenia,

or other factors). The term nondominant is used in case there are expanded clones with more than one amplifiable IGH rearrangement, for example, one productive and one nonproductive IGH gene rearrangement in the same cell.

32. If multiple replicates are not available for the dataset of interest, one can also computationally resample the dataset, mimicking the effect of multiple replicates [34].

Acknowledgments

This work is supported by NIH research grants awarded to ELP (AI144288, AI106697, P30-AI0450080, P30-CA016520). US is supported by grants from Mercator Stiftung, the German Research Foundation (DFG 397650460), BMBF e:KID (01ZX1612A), and BMBF NoChro (FKZ 13GW0338B). The authors thank members of the AIRR Community Biological Resources Working Group and Diagnostics Working Group for helpful discussions and feedback on the manuscript.

ELP is the director of the Human Immunology Core facility at the University of Pennsylvania, which uses this protocol. She is also the former Chair of the AIRR Community, receives research funding from Roche Diagnostics and Janssen Pharmaceuticals for projects unrelated to the method presented in this chapter, and is consulting or an advisor for Roche Diagnostics, Enpicom, the Antibody Society, IEDB, and the American Autoimmune Related Diseases Association.

References

1. Tonegawa S (1983) Somatic generation of antibody diversity. Nature 302(5909): 575–581. https://doi.org/10.1038/302575a0
2. Sakano H, Kurosawa Y, Weigert M, Tonegawa S (1981) Identification and nucleotide sequence of a diversity DNA segment (D) of immunoglobulin heavy-chain genes. Nature 290(5807):562–565. https://doi.org/10.1038/290562a0
3. Weigert MG, Cesari IM, Yonkovich SJ, Cohn M (1970) Variability in the lambda light chain sequences of mouse antibody. Nature 228(5276):1045–1047. https://doi.org/10.1038/2281045a0
4. Papavasiliou FN, Schatz DG (2002) Somatic hypermutation of immunoglobulin genes: merging mechanisms for genetic diversity. Cell 109(Suppl):S35–S44. https://doi.org/10.1016/s0092-8674(02)00706-7
5. Georgiou G, Ippolito GC, Beausang J, Busse CE, Wardemann H, Quake SR (2014) The promise and challenge of high-throughput sequencing of the antibody repertoire. Nat Biotechnol 32(2):158–168. https://doi.org/10.1038/nbt.2782
6. Benichou J, Ben-Hamo R, Louzoun Y, Efroni S (2012) Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. Immunology 135(3):183–191. https://doi.org/10.1111/j.1365-2567.2011.03527.x
7. Six A, Mariotti-Ferrandiz ME, Chaara W, Magadan S, Pham HP, Lefranc MP et al (2013) The past, present, and future of immune repertoire biology - the rise of next-generation repertoire analysis. Front Immunol 4:413. https://doi.org/10.3389/fimmu.2013.00413

8. van Dongen JJ, Langerak AW, Bruggemann M, Evans PA, Hummel M, Lavender FL et al (2003) Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 concerted action BMH4-CT98-3936. Leukemia 17(12): 2257–2317. https://doi.org/10.1038/sj.leu.2403202
9. Ritz C, Meng W, Stanley NL, Baroja ML, Xu C, Yan P et al (2020) Postvaccination graft dysfunction/aplastic anemia relapse with massive clonal expansion of autologous $CD8^+$ lymphocytes. Blood Adv 4(7):1378–1382. https://doi.org/10.1182/bloodadvances.2019000853
10. Meng W, Zhang B, Schwartz GW, Rosenfeld AM, Ren D, Thome JJC et al (2017) An atlas of B-cell clonal distribution in the human body. Nat Biotechnol 35(9):879–884. https://doi.org/10.1038/nbt.3942
11. Langerak AW, Bruggemann M, Davi F, Darzentas N, van Dongen JJM, Gonzalez D et al (2017) High-throughput Immunogenetics for clinical and research applications in immunohematology: potential and challenges. J Immunol 198(10):3765–3774. https://doi.org/10.4049/jimmunol.1602050
12. Vander Heiden JA, Yaari G, Uduman M, Stern JN, O'Connor KC, Hafler DA et al (2014) pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. Bioinformatics 30(13):1930–1932. https://doi.org/10.1093/bioinformatics/btu138
13. Rosenfeld AM, Meng W, Luning Prak ET, Hershberg U (2018) ImmuneDB, a novel tool for the analysis, storage, and dissemination of immune repertoire sequencing data. Front Immunol 9:2107. https://doi.org/10.3389/fimmu.2018.02107
14. Ye J, Ma N, Madden TL, Ostell JM (2013) IgBLAST: an immunoglobulin variable domain sequence analysis tool. Nucleic Acids Res 41(Web Server issue):W34–W40. https://doi.org/10.1093/nar/gkt382
15. Vander Heiden JA, Marquez S, Marthandan N, Bukhari SAC, Busse CE, Corrie B et al (2018) AIRR community standardized representations for annotated immune repertoires. Front Immunol 9:2206. https://doi.org/10.3389/fimmu.2018.02206
16. Rosenfeld AM, Meng W, Chen DY, Zhang B, Granot T, Farber DL et al (2018) Computational evaluation of B-cell clone sizes in bulk populations. Front Immunol 9:1472. https://doi.org/10.3389/fimmu.2018.01472
17. Alt FW, Yancopoulos GD, Blackwell TK, Wood C, Thomas E, Boss M et al (1984) Ordered rearrangement of immunoglobulin heavy chain variable region segments. EMBO J 3(6):1209–1219
18. Gotelli NJ, Colwell RK (2001) Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. Ecol Lett 4:379–391. https://doi.org/10.1046/j.1461-0248.2001.00230.x
19. Jaccard P (1912) The distribution of the flora in the alpine zone. New Phytol 11(2):37–50
20. Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4(4):406–425. https://doi.org/10.1093/oxfordjournals.molbev.a040454
21. Farris JS (1970) Methods for computing Wagner trees. Syst Zool 19(1):83–92. https://doi.org/10.2307/2412028
22. MacConaill LE, Burns RT, Nag A, Coleman HA, Slevin MK, Giorda K et al (2018) Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. BMC Genomics 19(1):30. https://doi.org/10.1186/s12864-017-4428-5
23. Rubelt F, Busse CE, Bukhari SAC, Burckert JP, Mariotti-Ferrandiz E, Cowell LG et al (2017) Adaptive immune receptor repertoire community recommendations for sharing immune-repertoire sequencing data. Nat Immunol 18(12):1274–1278. https://doi.org/10.1038/ni.3873
24. Lindenbaum O, Nouri N, Kluger Y, Kleinstein SH (2021) Alignment free identification of clones in B cell receptor repertoires. Nucleic Acids Res 49(4):e21. https://doi.org/10.1093/nar/gkaa1160
25. Gupta NT, Adams KD, Briggs AW, Timberlake SC, Vigneault F, Kleinstein SH (2017) Hierarchical clustering can identify B cell clones with high confidence in Ig repertoire sequencing data. J Immunol 198(6):2489–2499. https://doi.org/10.4049/jimmunol.1601850
26. Kepler TB (2013) Reconstructing a B-cell clonal lineage. I. Statistical inference of unobserved ancestors. F1000Res 2:103. https://doi.org/10.12688/f1000research.2-103.v1
27. Ralph DK, Matsen FA IV (2016) Likelihood-based inference of B cell clonal families. PLoS Comput Biol 12(10):e1005086. https://doi.org/10.1371/journal.pcbi.1005086
28. Pernis B, Chiappino G, Kelus AS, Gell PG (1965) Cellular localization of immunoglobulins with different allotypic specificities in

rabbit lymphoid tissues. J Exp Med 122(5): 853–876. https://doi.org/10.1084/jem.122.5.853

29. Barreto V, Cumano A (2000) Frequency and characterization of phenotypic Ig heavy chain allelically included IgM-expressing B cells in mice. J Immunol 164(2):893–899. https://doi.org/10.4049/jimmunol.164.2.893
30. Balomenos D, Balderas RS, Mulvany KP, Kaye J, Kono DH, Theofilopoulos AN (1995) Incomplete T cell receptor V beta allelic exclusion and dual V beta-expressing cells. J Immunol 155(7):3308–3312
31. Casellas R, Zhang Q, Zheng NY, Mathias MD, Smith K, Wilson PC (2007) Igkappa allelic inclusion is a consequence of receptor editing. J Exp Med 204(1):153–160. https://doi.org/10.1084/jem.20061918
32. Petrie HT, Livak F, Schatz DG, Strasser A, Crispe IN, Shortman K (1993) Multiple rearrangements in T cell receptor alpha chain genes maximize the production of useful thymocytes. J Exp Med 178(2):615–622. https://doi.org/10.1084/jem.178.2.615
33. Mathieson W, Thomas GA (2020) Why formalin-fixed, paraffin-embedded biospecimens must be used in genomic medicine: an evidence-based review and conclusion. J Histochem Cytochem 68(8):543–552. https://doi.org/10.1369/0022155420945050
34. Colwell RK, Chao A, Gotelli NJ, Lin S-Y, Mao CX, Chazdon RL et al (2012) Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. J Plant Ecol 5(1): 3–21. https://doi.org/10.1093/jpe/rtr044

Chapter 19

Bulk Sequencing from mRNA with UMI for Evaluation of B-Cell Isotype and Clonal Evolution: A Method by the AIRR Community

Nidhi Gupta, Susanna Marquez, Cinque Soto, Elaine C. Chen, Magnolia L. Bostick, Ulrik Stervbo, and Andrew Farmer

Abstract

During the course of an immune response to a virus such as influenza, B cells undergo activation, clonal expansion, isotype switching, and somatic hypermutation (SHM). Members of an antigen-experienced B-cell clone can have different sequence features including SHM in the immunoglobulin heavy-chain V (IGHV) gene and can use the same IGVH gene in combination with different constant regions or isotypes (e.g., IgM, IgG, IgA). To study these features of expanded clones in an immune response by AIRR-seq, we provide a bulk RNA-based sequencing experimental procedure with unique molecular identifiers (UMIs) and the accompanying bioinformatics analytical workflow.

Key words BCR, B cells, Repertoire, Bulk RNA, Sequencing, AIRR, Immunoglobulin, Bulk RNA sequencing, UMI, Heavy and light chain

1 Introduction

This protocol enables users to generate indexed libraries with full-length transcripts that are ready for sequencing on Illumina platforms (Fig. 1). It allows for the analysis of both immunoglobulin heavy (IGH) and kappa/lambda light-chain (IGK/IGL) gene rearrangements and has a sample input range from 10 ng to 1 μg of total RNA from peripheral blood mononuclear cells (PBMCs) or 1 to 100 ng of total RNA from purified B cells.

The protocol leverages SMART technology (**s**witching **m**echanism **a**t 5′ end of **R**NA **t**emplate) and employs a 5' RACE-like approach to capture complete V(D)J variable regions of BCR/IG transcripts. It also incorporates unique molecular identifiers (UMIs). First-strand cDNA synthesis is oligo-dT primed and

Nidhi Gupta and Susanna Marquez are shared first authors.

Anton W. Langerak (ed.), *Immunogenetics: Methods and Protocols*, Methods in Molecular Biology, vol. 2453, https://doi.org/10.1007/978-1-0716-2115-8_19,

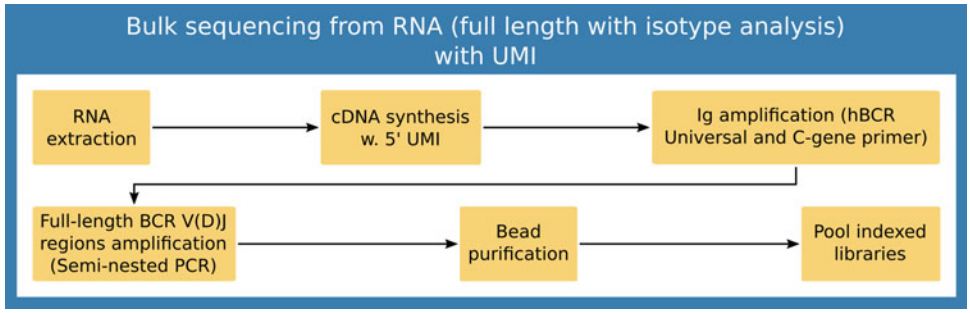


Fig. 1 Overview of the SMARTer human BCR procedure. cDNA is synthesized from RNA isolated from PBMCs or B cells, followed by two rounds of PCR and finally purified and pooled to prepare the libraries for sequencing

catalyzed by SMARTScribe™ reverse transcriptase (RT), which adds non-templated nucleotides at the 5′ end of each mRNA template. The SMART UMI Oligo anneals to these non-templated nucleotides, serves as a template for incorporation of a PCR handle into the first-strand cDNA, and uniquely tags each cDNA molecule with a UMI (UMIs allow for the generation of consensus sequences during data analysis, thereby minimizing PCR and sequencing errors). Following reverse transcription, two rounds of PCR are performed to amplify cDNAs. To capture the entire V(D)J region, primers in these PCRs anneal to sequence added by the SMART UMI Oligo at the 5′ end and the IG constant region(s) at the 3′ end. The second PCR takes the product from the first PCR as a template and uses semi-nested primers to amplify the entire IG variable region and a small portion of the constant region (Fig. 2).

We also provide a computational workflow to analyze the sequencing data with the Immcantation framework (immcantation.org). The workflow covers preprocessing, isotype assignment, quality control and filtering, gene annotation, gene usage, population structure determination, and lineage reconstruction.

2 Materials

2.1 General Reagents

All components are available in SMARTer Human BCR IgG IgM / IgK/IgL Profiling Kit (Takara Bio, *see* **Note 1**).

1. Control RNA (human spleen total RNA, 1 μg/μL).
2. SMART UMI Oligo.
3. dT Primer.
4. 5× first-strand buffer.
5. 100 U/μL SMARTScribe reverse transcriptase.
6. Nuclease-free water.
7. 40 U/μL RNase inhibitor.

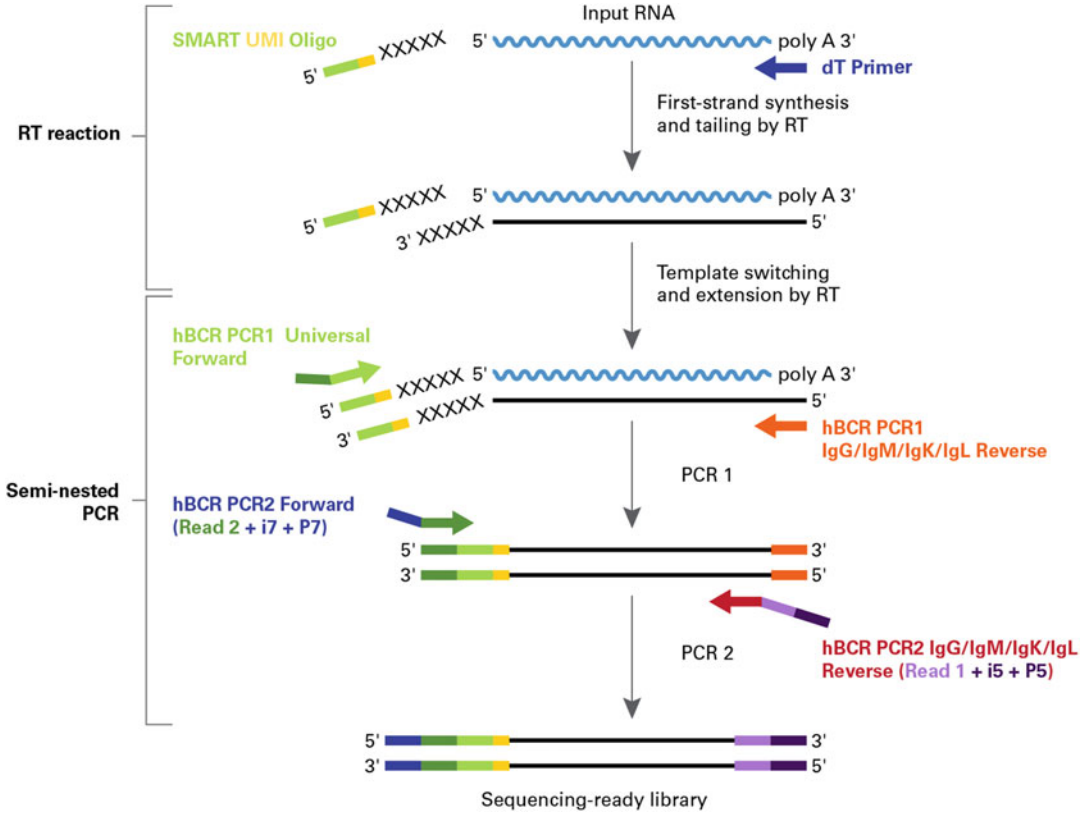


Fig. 2 A schematic of dT-primed first-strand cDNA synthesis followed by two rounds of successive PCR for amplification of cDNA sequences. After post-PCR purification, size selection, and quality analysis, the library is ready for sequencing

8. BCR enhancer.
9. 10 mM Tris–HCl elution buffer (pH 8.5).
10. hBCR PCR1 Universal Forward.
11. hBCR PCR1 IgG reverse.
12. hBCR PCR1 IgM reverse.
13. hBCR PCR1 IgK reverse.
14. hBCR PCR1 IgL reverse.
15. 1.25 U/μL PrimeSTAR GXL DNA polymerase.
16. 5× PrimeSTAR GXL buffer.
17. 2.5 mM dNTP mixture.
18. hBCR PCR2 IgG reverse 1–4.
19. hBCR PCR2 IgM reverse 1–4.
20. hBCR PCR2 IgK reverse 1–4.
21. hBCR PCR2 IgL reverse 1–4.
22. hBCR PCR2 Forward 1–12.

2.2 Primers

Illumina indexes are incorporated into human BCR profiling libraries through both forward and reverse PCR primers. The corresponding Illumina indexes are listed below.

2.2.1 Human BCR Indexing Primer Set HT for Illumina Sequences

2.2.2 BCR PCR2 Forward Primer i7 HT Index

Primers are listed with the name, Illumina ID, and index sequence.

1. hBCR PCR2 Universal Forward 1, D701, ATTACTCG.
2. hBCR PCR2 Universal Forward 2, D702, TCCGGAGA.
3. hBCR PCR2 Universal Forward 3, D703, CGCTCATT.
4. hBCR PCR2 Universal Forward 4, D704, GAGATTCC.
5. hBCR PCR2 Universal Forward 5, D705, ATTCAGAA.
6. hBCR PCR2 Universal Forward 6, D706, GAATTCGT.
7. hBCR PCR2 Universal Forward 7, D707, CTGAAGCT.
8. hBCR PCR2 Universal Forward 8, D708, TAATGCGC.
9. hBCR PCR2 Universal Forward 9, D709, CGGCTATG.
10. hBCR PCR2 Universal Forward 10, D710, TCCGCGAA.
11. hBCR PCR2 Universal Forward 11, D711, TCTCGCGC.
12. hBCR PCR2 Universal Forward 12, D712, AGCGATAG.

2.2.3 BCR Indexing Reverse Primer Set HT for Illumina Index Sequences

Primers are listed with name, Illumina ID, and index sequences, as read on a MiSeq instrument.

1. hBCR PCR2 IgG Reverse 1, D501, TATAGCCT.
2. hBCR PCR2 IgM Reverse 1, D501, TATAGCCT.
3. hBCR PCR2 IgK Reverse 1, D501, TATAGCCT.
4. hBCR PCR2 IgL Reverse 1, D501, TATAGCCT.
5. hBCR PCR2 IgG Reverse 2, D502, ATAGAGGC.
6. hBCR PCR2 IgM Reverse 2, D502, ATAGAGGC.
7. hBCR PCR2 IgK Reverse 2, D502, ATAGAGGC.
8. hBCR PCR2 IgL Reverse 2, D502, ATAGAGGC.
9. hBCR PCR2 IgG Reverse 3, D503, CCTATCCT.
10. hBCR PCR2 IgM Reverse 3, D503, CCTATCCT.
11. hBCR PCR2 IgK Reverse 3, D503, CCTATCCT..
12. hBCR PCR2 IgL Reverse 3, D503, CCTATCCT.
13. hBCR PCR2 IgG Reverse 4, D504, GGCTCTGA.
14. hBCR PCR2 IgM Reverse 4, D504, GGCTCTGA.
15. hBCR PCR2 IgK Reverse 4, D504, GGCTCTGA.
16. hBCR PCR2 IgL Reverse 4, D504, GGCTCTGA.

2.3 Equipment

1. Pipettes: 10 μL, 20 μL, and 200 μL.
2. Filter pipette tips: 2 μL, 20 μL, and 200 μL.
3. Microcentrifuge tubes: 1.5 mL.
4. Minicentrifuge 0.2 mL tubes or strips.
5. NucleoSpin RNA Plus, mini kit for RNA purification with DNA removal column (Macherey-Nagel).
6. Thermal cyclers, separate dedicated instruments for first-strand cDNA synthesis, and PCR amplification.
7. Agilent 2100 Bioanalyzer – DNA 1000 kit; for validation, alternatively use the TapeStation.
8. TapeStation, (Agilent), for validation, alternatively use the bioanalyzer.
9. Qubit dsDNA HS Kit (Thermo Fisher Scientific).
10. Nuclease-free thin wall PCR tubes, 96-well plates, or strips (USA Scientific).
11. Nuclease-free low-adhesion 1.5 mL tubes.
12. NucleoMag NGS clean-up and size select, Takara Bio 5 mL size for bead purification.
13. 100% ethanol, molecular biology grade.
14. SMARTer-Seq™ Magnetic Separator, PCR Strip, Takara Bio 8-tube strips or Thermo Fisher Scientific 96-well plates.
15. Low-speed benchtop centrifuge for 96-well plate.

2.4 Software

Immcantation suite Docker container. *See* **step 3** ("obtain the software") in Subheading 3.10.

3 Methods

3.1 Overview of Wet Bench Protocol

The major steps of the procedure are outlined in Fig. 3. This sequencing protocol has been optimized for 10 ng of total RNA per sequencing library, which corresponds to ~1000 cells. But the extraction of RNA is far more efficient and yields higher purity and higher-quality product at much higher numbers of input cells, on the scale of tens of thousands to millions. RNA extracted from PBMCs yields approximately 1 μg per mL or ~one million nucleated cells, of which only a small fraction (1–10%) are B cells.

3.2 RNA Extraction

The following is an illustration of RNA isolation from approximately 1×10^7 cultured cells using the NucleoSpin RNA Plus protocol.

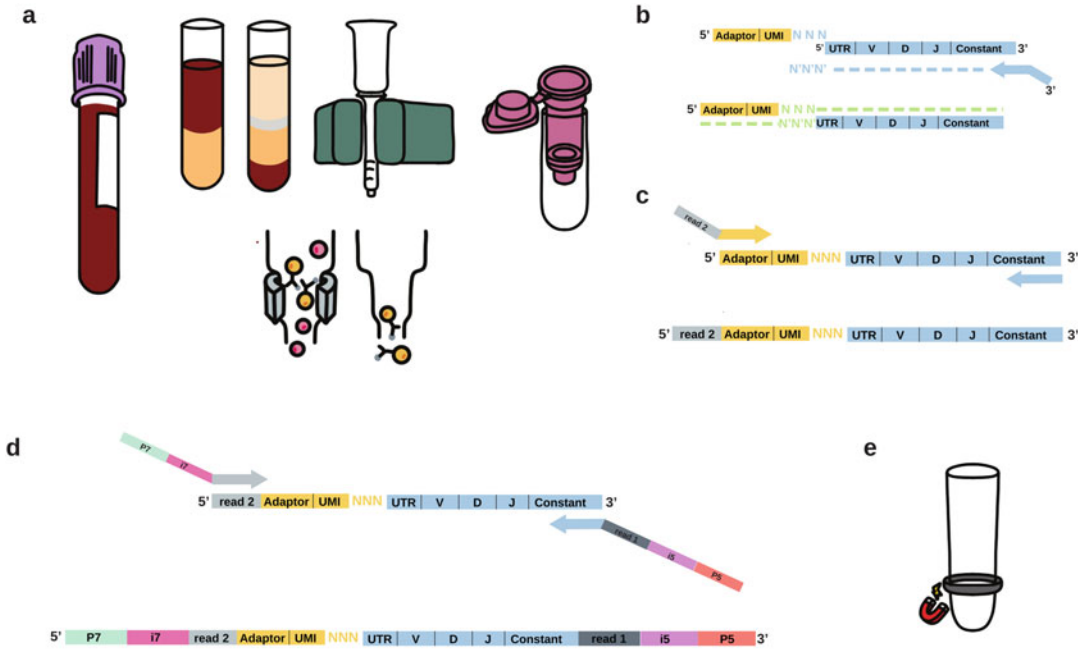


Fig. 3 Bulk RNA sequencing with unique molecular identifiers (UMIs). (**a**) This protocol begins with a single-cell suspension that can be isolated from whole blood, peripheral blood mononuclear cells, or by purification using either magnetic beads or flow cytometry. Total RNA is extracted from the cell population(s) of interest. (**b**) cDNA is reverse transcribed from RNA, and unique molecular identifiers (UMIs) are included in the SMART Oligo, tagging each parental cDNA molecule. (**c**) In PCR1, IgG, IgM, IG kappa, and IG lambda chains (including both the variable and part of the constant region) are separately amplified (only the IGH are shown). (**d**) In PCR2, Illumina indices are added to generate the sequencing libraries. (**e**) After size selection, QC, and normalization of library input, libraries are sequenced using the Illumina platform

1. Cell collection: Transfer cells to an appropriate tube, and pellet by centrifugation for 5 min at 300 × *g*. Remove supernatant.
2. Homogenize and lyse sample by adding 350 µL buffer LBP to the cell pellet. Mixing the cells with a lysis buffer is usually sufficient for complete lysis.
3. Remove gDNA and filtrate lysate: Place the NucleoSpin® gDNA Removal Column (yellow ring) in a 2 mL collection tube, transfer the homogenized lysate to the NucleoSpin® gDNA Removal Column, and centrifuge for 30 s at 11,000 × *g*. Discard the column and continue with the flowthrough.
4. Adjust RNA binding conditions: Add 100 µL binding solution BS to the flowthrough, and mix well by moderate vortexing or by pipetting up and down several times. After addition of binding solution BS, a stringy precipitate may become visible which will not affect the RNA isolation. Be sure to disaggregate any precipitate by mixing and load all of the precipitate on the

column as described in the following step. Do not centrifuge the lysate after addition of binding solution before loading it onto the column in order to avoid pelleting the precipitate.

5. Bind RNA: Transfer the whole lysate (~450 μL) to the NucleoSpin® RNA Plus Column (light blue ring) preassembled with a collection tube. Centrifuge for 15 s at 11,000 × *g*.
6. Wash and dry silica membrane:
   (a) First wash: Add 200 μL buffer WB1 to the NucleoSpin® RNA Plus Column. Centrifuge for 15 s at 11,000 × *g*. Discard the flowthrough with the collection tube, and place the column into a new 2 mL collection tube.
   (b) Second wash: Add 600 μL buffer WB2 to the NucleoSpin® RNA Plus Column. Centrifuge for 15 s at 11,000 × *g*. Discard flowthrough, and place the column back into the collection tube.
   (c) Third wash: Add 250 μL buffer WB2 to the NucleoSpin® RNA Plus Column. Centrifuge for 2 min at 11,000 × *g* to dry the membrane completely. Place the column into a nuclease-free 1.5 mL collection tube.
7. Elute RNA: Add 30 μL RNase-free H_2O and centrifuge at 11,000 × *g* for 1 min. Add an additional 30 μL RNase-free H_2O to the column, and centrifuge again at 11,000 × *g* for 1 min.
8. After RNA extraction, if sample size is not limiting, we recommend evaluating total RNA quality using the Agilent RNA 6000 Pico Kit or an equivalent platform. Refer to the manufacturer for instructions (*see* **Note 2**).

3.3 First-Strand cDNA Synthesis

First-strand cDNA synthesis (from RNA) is primed by the dT Primer. Here we illustrate cDNA synthesis using the SMART UMI Oligo for template switching at the 5′ end of the transcript.

1. Thaw the First-Strand Buffer at room temperature. Thaw BCR enhancer, SMART UMI Oligo, and dT Primer on ice. Gently vortex each reagent to mix and centrifuge briefly. Store all but the First-Strand Buffer on ice. Remove the SMARTScribe reverse transcriptase and RNase inhibitor from the freezer immediately before use, centrifuge briefly, and store on ice (*see* **Note 3**).
2. Preheat the thermal cycler to 72 °C.
3. On ice, prepare samples and controls in nuclease-free thin-wall PCR tubes, plates, or strips by adding the reagents in the order shown below.

Component	Volume (μL)
Sample or control[a]	1–9.5
Nuclease-free water	Up to 8.5
BCR enhancer	1
dT primer	2
Total volume	**12.5**

[a]Control RNA is supplied at a concentration of 1 μg/μL. It should be thawed on ice and diluted serially in nuclease-free water

4. Mix by gently vortexing and then centrifuge briefly.
5. Incubate the tubes at 72 °C in the preheated, heated-lid thermal cycler for 3 min. During this incubation, prepare the RT Master Mix.
6. At room temperature, prepare RT Master Mix by combining the following in the order shown. Wait to add the SMARTScribe reverse transcriptase to the master mix until just prior to use in **step 10** of Subheading 3.3.

Component	Volume (μL)
First-Strand buffer[a]	4
SMART UMI oligo	1
RNase inhibitor	0.5
SMARTScribe reverse transcriptase	2
Total volume	**7.5**

[a]Ensure the First-Strand Buffer is completely in solution. Vortex gently to remove any cloudiness before use

7. Mix the RT Master Mix well by gently pipetting up and down, and then centrifuge briefly.
8. Immediately after the 3-min incubation at 72 °C (**step 5** of Subheading 3.3), place the samples on ice for 2 min.
9. Reduce the temperature of the thermal cycler to 42 °C.
10. Add 7.5 μL of the RT Master Mix (**step 6** of Subheading 3.3) to each reaction tube. Mix the contents of each tube by pipetting gently and centrifuge briefly.
11. Place the tubes in a thermal cycler with a heated lid, preheated to 42 °C. Run the following program: 42 °C, 90 min; 70 °C, 10 s; 4 °C hold.

    Stopping point: The tubes can be stored at 4 °C overnight.

Table 1
Cycling guidelines based on amount of starting material

RNA source	Input amount	Number of PCR 1 cycles	Number of PCR 2 cycles[a]
PBMC	10 ng	18	21
PBMC	100 ng	18	18
PBMC	1 μg	18	16
B cell	1 ng	18	21
B cell	10 ng	18	18
B cell	100 ng	18	16
Whole blood	100 ng	18	25
Spleen	10 ng	18	20
Bone marrow	10 ng	18	20
Control RNA	10 ng	18	20
Control RNA	100 ng	18	18
Control RNA	1 μg	18	16

[a]If the number of cycles generates an insufficient library for sequencing, repeat PCR2 with more cycles

3.4 First-Round Amplification

Semi-nested PCR amplifies the entire V(D)J region and a portion of the constant region of IG cDNA(s) and incorporates adapters and barcodes for Illumina sequencing platforms. Expression of different IG chains can vary significantly among B-cell populations. Thus, we recommend separately amplifying each chain of interest. Table 1 provides PCR cycling recommendations, but optimal parameters may vary for different sample types, input amounts, and thermal cyclers. We recommend trying a range of cycle numbers to determine the minimum number necessary to obtain the desired yield.

In the first round of PCR amplification, also referred to as PCR1, one performs separate IgG/IgM/IgK/IgL amplification. This PCR selectively amplifies full-length BCR V(D)J regions from first-strand cDNA. A portion of the first-strand cDNA is used for each amplification reaction. The hBCR PCR1 Universal Forward primer anneals to the 5′ end of transcripts via the SMART UMI Oligo sequence. The hBCR PCR1 IgG/IgM/IgK/IgL reverse primers anneal to sequences in the constant regions of IG heavy- and light-chain cDNAs.

1. Thaw 5× PrimeSTAR GXL buffer, dNTP mix, primers, and nuclease-free water on ice. Gently vortex each reagent to mix and centrifuge briefly. Store on ice. Remove the PrimeSTAR GXL DNA polymerase from the freezer immediately before use, gently pipet to mix, centrifuge briefly, and store on ice.

2. Prepare a PCR1 Master Mix for each IgG/IgM/IgK/IgL chain of interest, by combining the following in the order shown, on ice. Gently vortex to mix and centrifuge briefly (*see* **Note 4**).

Component	Volume (μL)
Nuclease-free water	29
5× PrimeSTAR GXL PCR buffer	10
dNTP mixture	4
hBCR PCR1 universal forward	1
hBCR PCR1 IgG reverse OR	1
hBCR PCR1 IgM reverse OR	
hBCR PCR1 IgK reverse OR	
hBCR PCR1 IgL reverse	
PrimeSTAR GXL polymerase	1
Total volume	**46**

3. Add 46 μL of the appropriate IgG/IgM/IgK/IgL PCR1 Master Mix to nuclease-free, thin-wall 0.2-mL PCR plate/tube(s).
4. Add 4 μL of first-strand cDNA from Subheading 3.3 to the corresponding tube(s) containing PCR1 Master Mix. Gently vortex to mix, and centrifuge briefly.
5. Place the plate/tube(s) in a preheated thermal cycler with a heated lid, and run the following program (lid temperature: 105 °C): 95°C 1 min; 98°C 10 s, 60°C 15 s, and 68°C 45 s (18 cycles); 4°C hold. *Consult Table 1 for PCR cycle number guidelines. Stopping point: The tubes may be stored at 4 °C overnight.

3.5 Second-Round PCR Amplification

In the second round of PCR amplification, termed PCR2, sequencing libraries are generated. PCR2 further amplifies the full-length IG V(D)J regions and adds Illumina indexes using a semi-nested approach. The hBCR PCR2 Universal Forward 1–12 primers add P7/i7 index sequences. The hBCR PCR2 IgG/IgM/IgK/IgL reverse 1–4 primers anneal to the constant region of the IG sequence and add P5/i5 index sequences (*see* **Note 5**).

1. Thaw 5X PrimeSTAR GXL buffer, dNTP Mix, primers, and nuclease-free water on ice. Gently vortex each reagent to mix and centrifuge briefly. Store on ice. Remove the PrimeSTAR GXL DNA polymerase from the freezer immediately before use, gently pipet mix, centrifuge briefly, and store on ice.
2. For each IgG/IgM/IgK/IgL chain of interest, prepare a PCR2 Master Mix by combining the following in the order

shown, on ice. Gently vortex to mix and centrifuge briefly (*see* **Note 6**).

Component	Volume (μL)
Nuclease-free water	32
5X PrimeSTAR GXL PCR buffer	10
dNTP mixture	4
hBCR PCR2 IgG reverse 1–4 OR	1
hBCR PCR2 IgM reverse 1–4 OR	
hBCR PCR2 IgK reverse 1–4 OR	
hBCR PCR2 IgL reverse 1–4	
PrimeSTAR GXL polymerase	1
Total volume	**48**

3. For each reaction, add 48 μL of PCR2 Master Mix to nuclease-free, thin-wall, 0.2-mL PCR plate/tube(s).
4. Add 1 μL of appropriate PCR1 product to each corresponding PCR 2 tube.
5. Add 1 μL of the appropriate hBCR PCR2 Universal Forward 1–12 primer to each sample. Gently vortex to mix and centrifuge briefly.
6. Place the plate/tube(s) in a preheated thermal cycler with a heated lid, and run the following program (lid temperature: 105 °C): 95°C 1 min; 98°C 10 s, 60°C 15 s, 68°C 45 s (X* cycles); 4°C Hold. *Consult Table 1 for PCR cycle number guidelines.

Stopping point: The tubes may be stored at 4 °C overnight.

3.6 Purification of Amplified Libraries

Here we illustrate amplified library purification using NucleoMag NGS clean-up and size select beads (*see* **Note 7**).

1. Vortex NucleoMag beads until evenly mixed, and then add 25 μL of the NucleoMag beads to each sample.
2. Mix thoroughly by gently pipetting the entire volume up and down at least ten times (*see* **Note 8**).
3. Incubate at room temperature for 8 min to let the DNA bind to the beads.
4. Briefly spin the samples to collect the liquid from the side of the tube or sample well. Place the samples on the magnetic separation device for ~5 min or longer until the liquid appears completely clear, and there are no beads left in the supernatant. The time required for the solution to clear will depend on the strength of the magnet (*see* **Note 9**).

5. While the reaction tubes are sitting on the magnetic separation device, use a pipette to transfer the supernatant (which contains your library) to clean PCR tubes.
6. Remove the tubes containing the beads from the magnetic separation device, and discard them.
7. Add 10 μL of NucleoMag beads to each tube containing supernatant (*see* **Note 10**).
8. Mix thoroughly by gently pipetting the entire volume up and down at least ten times.
9. Incubate at room temperature for 8 min to let the DNA bind to the beads.
10. Place the tubes on the magnetic separation device for ~10 min or until the solution is completely clear.
11. While the tubes are sitting on the magnetic separation device, remove the supernatant with a pipette and discard it (the library is now bound to the beads).
12. Keep the tubes on the magnetic separation device. Add 200 μL of freshly made 80% ethanol to each sample, without disturbing the beads, to wash away contaminants. Wait for 30 s, and use a pipette to carefully remove the supernatant containing contaminants. The library will remain bound to the beads during the washing process.
13. Repeat the ethanol wash (**step 12** of this section) once more.
14. Briefly spin the tubes (~2000 × *g*) to collect the remaining liquid at the bottom of each tube. Place the tubes on the magnetic separation device for 30 s, and then remove all remaining liquid with a pipette.
15. Let the sample tubes rest open on the magnetic separation device at room temperature for ~2–2.5 min until the pellet appears dry and is no longer shiny. You may see a tiny crack in the pellet (*see* **Note 11**).
16. Once the bead pellet has dried, remove the tubes from the magnetic separation device, and add 17 μL of elution buffer to cover the pellet. Mix thoroughly by pipetting up and down to ensure complete bead dispersion (*see* **Note 12**).
17. Incubate at room temperature for at least 5 min to rehydrate.
18. Briefly spin the samples to collect the liquid from the side of the tube or sample well. Place the samples back on the magnetic separation device for 2 min or longer until the solution is completely clear (*see* **Note 13**).
19. Transfer clear supernatant containing purified BCR/IG library from each tube to a nuclease-free, low-adhesion tube. Label each tube with sample information and store at −20 °C.

    Stopping point: The tubes may be stored at 4 °C overnight.

3.7 Library Validation

To assess the success of library preparation, purification, and size selection, we recommend quantifying the libraries with a Qubit dsDNA HS Kit and evaluating the libraries' size distributions with an Agilent 2100 Bioanalyzer and the DNA 1000 Kit.

1. Compare the results for your samples with Fig. 4 to verify whether each sample is suitable for further processing. High-quality libraries should yield no product for negative control reactions and a broad peak spanning 500 bp to 1200 bp, with a maximum between ~600 bp and ~900 bp for positive controls and samples containing PCR-amplified IG libraries. The position and shape of electropherogram peaks will vary depending on which chain sequences are included in the library, the nature of the originally included RNA sample, and the analysis method. In general, electropherogram peaks obtained with the Fragment Analyzer tend to be sharper than those obtained with the bioanalyzer.
2. Following validation, libraries are ready for sequencing on Illumina platforms.

3.8 Pooling of Samples to Generate Libraries for Sequencing

Following library validation by Qubit and bioanalyzer, the desired library pools should be prepared for the sequencing run. Prior to pooling, libraries must be carefully quantified. By combining the quantification obtained with the Qubit with the average library size determined by the bioanalyzer, the concentration in ng/μL can be converted to nM. The following web tool is convenient for the conversion: http://www.molbiol.edu.ru/eng/scripts/01_07.html. Alternatively, libraries can be quantified by qPCR using the Library Quantification Kit from Takara Bio.

Most Illumina sequencing library preparation protocols require libraries with a final concentration of 4 nM, including the MiSeq instrument that we recommend for this protocol.

Prepare a pool of 4 nM as follows:

1. Dilute each library to 4 nM in nuclease-free water. To avoid pipetting errors, use at least 2 μL of each original library for dilution.
2. Pool the diluted libraries by combining an equal amount of each library in a low-bind 1.5-mL tube. Mix by vortexing at low speed or by pipetting up and down. Use at least 2 μL of each diluted library to avoid pipetting error.
3. Use a 5 μL aliquot of the 4-nM-concentration-pooled libraries. Follow the library denaturation protocol according to the latest edition of your Illumina sequencing instrument's user guide.

You should also plan to include a 10% PhiX control spike-in (PhiX Control v3, Illumina). The addition of the PhiX control is essential to increase the nucleotide diversity and achieve high-quality data generation (*see* **Note 14**).

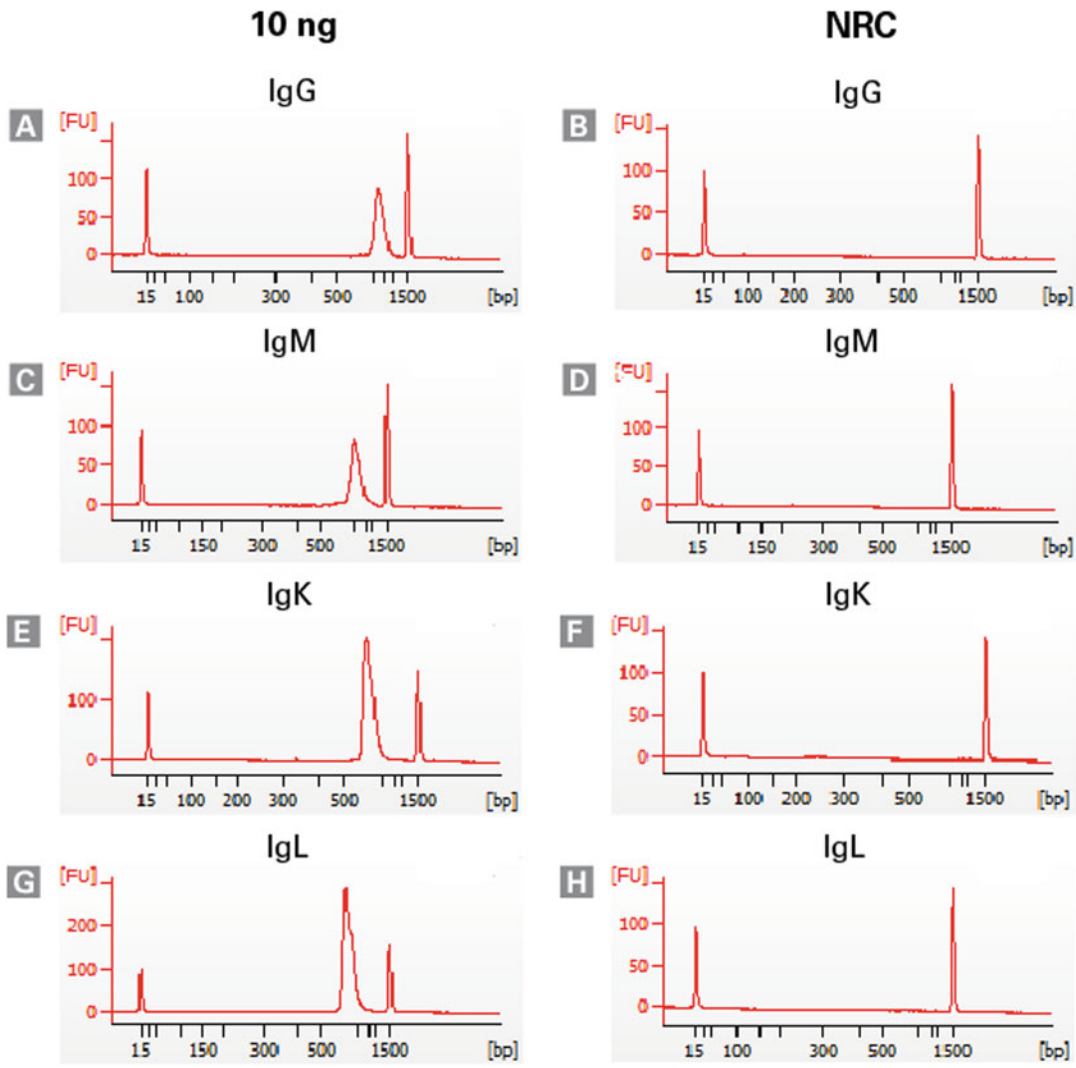


Fig. 4 Validation of IG heavy- and IG light (kappa or lambda)-chain libraries from human spleen that were generated using the SMARTer Human BCR Profiling Kit. Purified and size-selected libraries were analyzed on an Agilent 2100 Bioanalyzer (Panels A–H). Panels A, C, E, and G show broad peaks between ~500 and 1200 bp and maximal peaks in the range of ~600–900 bp (typical results for a library generated from spleen RNA). RNA control (NRC) samples (Panels B, D, F, and H) show no library produced and a flat Bioanalyzer profile within the predicted amplicon range of 500–1200 bp

Sequencing should be performed on an Illumina MiSeq sequencer using the 600-cycle MiSeq Reagent Kit v3 with paired-end, 2 × 300 base pair reads. When relying on Qubit quantification, we recommend diluting the pooled denatured libraries to a final concentration of 12.5 pM to achieve optimal cluster density. If using qPCR for quantification, one may need to use a lower final concentration.

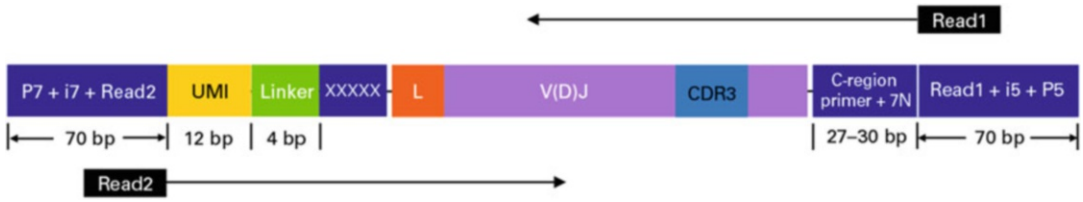


Fig. 5 SMARTer human BCR IgG IgM H/K/L profiling library structure. First 19 nt from Read2 can be trimmed off if UMI analysis is not performed

The complexity of the human IG repertoire varies from person to person. We generally recommend a minimum of 200,000 reads for IG heavy-chain libraries (IgG and IgM) from an input of 10 ng PBMC RNA (or 1 ng B-cell RNA). and a minimum of 500,000 reads for IG light (IGK and IGL) chains from an input of 10 ng PBMC RNA (or 1 ng B-cell RNA). For libraries generated from >10 ng PBMC RNA, higher sequencing depth is recommended. However, the optimal conditions may vary for different samples types, sample masses, sample complexities, and desired outcomes. We recommend trying a higher sequencing depth, then down sampling to determine the optimal sequencing depth.

As shown in Fig. 5, a human BCR profiling library contains a 12-nucleotide UMI that can be used to create consensus reads for sequences that share the same UMI, allowing correction for sequencing error correction.

Upon completion of a sequencing run, data can be analyzed with Takara Bio Cogent NGS Immune Profiler Software or other software. In the following sections, we provide a workflow to analyze data with Immcantation, a suite that provides tools to perform preprocessing, population structure determination, and repertoire analysis. Immcantation is certified as compliant with AIRR Community software guidelines.

3.9 Data Analysis Overview

In this workflow, we show how to use Immcantation (immcantation.org) to analyze sequencing data generated following the experimental protocol described in Subheading 3.8. An overview is given in Fig. 6. In the "AIRR Community Guide to TR and IG Gene Annotation" and "AIRR Community Guide to Repertoire Analysis" chapters the goals of multiple common AIRR-seq analysis techniques are described in detail, which can be useful to interpret the analysis performed in this section.

3.10 Raw Data Processing

The command line tool pRESTO [1] provides utilities to execute all stages of sequencing data processing prior to germline gene assignment. The tools are modular and can be combined to build highly customizable workflows. pRESTO includes features for quality control, primer masking, annotation of reads with sequence-embedded barcodes, generation of unique molecular

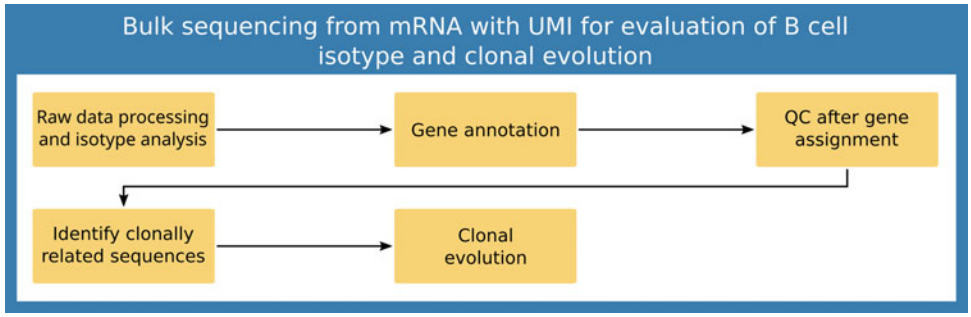


Fig. 6 Overview of data processing and analysis steps

identifier (UMI) consensus sequences, assembly of paired-end reads, and identification of duplicate sequences.

1. Remove PhiX

   If spike-in PhiX was not removed by the sequencing facility, it is recommended [2] to filter out these reads.

2. Understand the Read Layout

   It is important to have a good understanding of the read layout and know what region each read covers, where the primers and barcodes are located, and how long they are. In this example, R1 starts in the constant region of the rearranged sequence and R2 upstream the V region. *See* Fig. 5 for details on the read layout.

   Primers from the vendor are not available. To identify isotypes, it is possible to use as primers the consensus sequences of the constant region available online from the `protocols/Universal` directory in the Immcantation repository (https://bitbucket.org/kleinstein/immcantation). These sequences have been created after analyzing the first 30 nucleotides of the human constant region sequences available from IMGT.

3. Obtain the Software

   Immcantation, with its dependencies, accessory scripts, and IgBLAST [4] and IMGT [3] reference germlines, is available as a Docker container on docker hub under `immcantation/suite:x.x.x` where x.x.x stands for a release number. This protocol is using the container release 4.3.0.

   To start an interactive session inside the container and share local files in the current working directory with the `/data` folder in the container, use.

```
docker run -it -v $(pwd):/data:z --workdir /data immcantation/suite:4.3.0 bash
```

Once inside the container, you can use the commands `versions report` and `builds report` to know the versions of the software installed.

If you type `pwd`, you should get the result `/data`, as expected after starting the container with `--workdir /data`. If you type `ls`, you should see the files that you have in the local directory from which you launched the container. Being inside the container session, create the output directories presto and logs, and verify that the folder also becomes available locally in your computer:

```
mkdir presto
mkdir logs
```

4. Remove Low-Quality Sequences

   To remove reads with a mean quality lower than 20, use the command.

```
FilterSeq.py quality -s data/S5_R1.fastq -q 20 --nproc 8 \
 --outname CRR --outdir presto --log "logs/quality-crr.log"

FilterSeq.py quality -s data/S5_R2.fastq -q 20 --nproc 8 \
 --outname VRR --outdir presto --log "logs/quality-vrr.log"
```

   Output data files for the constant region reads will use the prefix CRR, and data files for the V region reads, will use the prefix VRR.

5. Identify Primers and UMI

   The next step is to remove or mask primers and extract UMI barcodes from the sequence but keeping this information as annotations in the FASTQ file headers. We recommend to mask or remove primers so that sequencing errors in the primers do not affect downstream analyses. Here we remove barcodes and primers. We know that the kit used to generate the data has a 12-nucleotide-long UMI (`--start 12`), followed by a linker sequence and a template switch (`--len 7`). With this command, pRESTO will extract the first 12 bp and annotate the fastq file header with the field `BARCODE`.

```
MaskPrimers.py extract -s presto/VRR_quality-pass.fastq \
 --start 12 --len 7 --barcode --bf BARCODE --mode cut \
 --log "logs/primers-vrr.log" \
 --outname VRR --outdir presto
```

   An example output FASTQ header is as follows:

```
@M03355:144:000000000-CH2WP:1:1104:17528:20342 2:N:0:CGCTCATT
+TATAGCCT|PRIMER=GTACGGG|BARCODE=TTGAAGTTATTC
```

6. Annotate R1 with Internal C-Region

   Use the following command to annotate the CRR FASTQ file with a constant region call. This step requires a reference FASTA file containing the reverse-complement of short sequences from the front of CH-1. The C-region sequences (`-p`) are available in the container. For each sequence, MaskPrimers.py align will look for good matches (maximum error of `--maxerror 0.3`) to the reference sequences in the first 100 nucleotides (`--maxlen 100`). The matching and preceding region will be cut out from the sequence. The matching sequence name will be added as an annotation into the FASTQ header, under the field `C_CALL`.

```
MaskPrimers.py align -s presto/CRR_quality-pass.fastq \
 -p /usr/local/share/protocols/Universal/Human_IG_CRegion_RC.
fasta \
 --maxlen 100 --maxerror 0.3 \
 --mode cut --skiprc --pf C_CALL \
 --log "logs/cregion.log" --outname "CRR" --nproc 8
```

An example output FASTQ header is as follows:

```
@M03355:144:000000000-CH2WP:1:2116:18550:17244 1:N:0:CGCTCATT
+TATAGCCT|SEQORIENT=F|C_CALL=IGHM
```

pRESTO tools save logs that can be converted into tabulated files with ParseLog.py. It is useful to use these files to generate diagnostic plots. To extract the information to make figures, inspect the C_CALLs made, and identify the starting position of the match, use the command below. It will create a tabulated file with the fields ID, PRIMER, ERROR, and PRSTART, which can be used to create such plots.

```
ParseLog.py -l "logs/cregion.log" -f ID PRIMER ERROR PRSTART
--outdir logs
```

Once the log has been converted to a tabulated file, it can be easily loaded into R, to count the different isotypes that have been identified:

```
cregion_table <- read.delim("logs/cregion_table.tab")
table(cregion_table$PRIMER)
```

Example output:

```
 IGHA IGHD IGHE IGHG IGHM IGKC IGLC1 IGLC3
 4 32 510 242981 220640 244015 250521 40647
```

These results match the expectations for this experimental protocol, because it uses a kit designed for IgM, IgG, IgK, and IgL. The isotype count can also be visualized (Fig. 7a):

```
# Create a color palette
color_palette <- c(
"IGHA"="882255",
"IGHD"="#AA4499",
"IGHE"="#88CCEE",
"IGHG"="#CC6677",
"IGHM"="#6699CC",
"IGKC"="#44AA99",
"IGLC1"="#888888",
"IGLC3"="#DDCC77"
)
isotypes <- sort(unique(cregion_table$PRIMER))
names(color_palette) <- isotypes
```

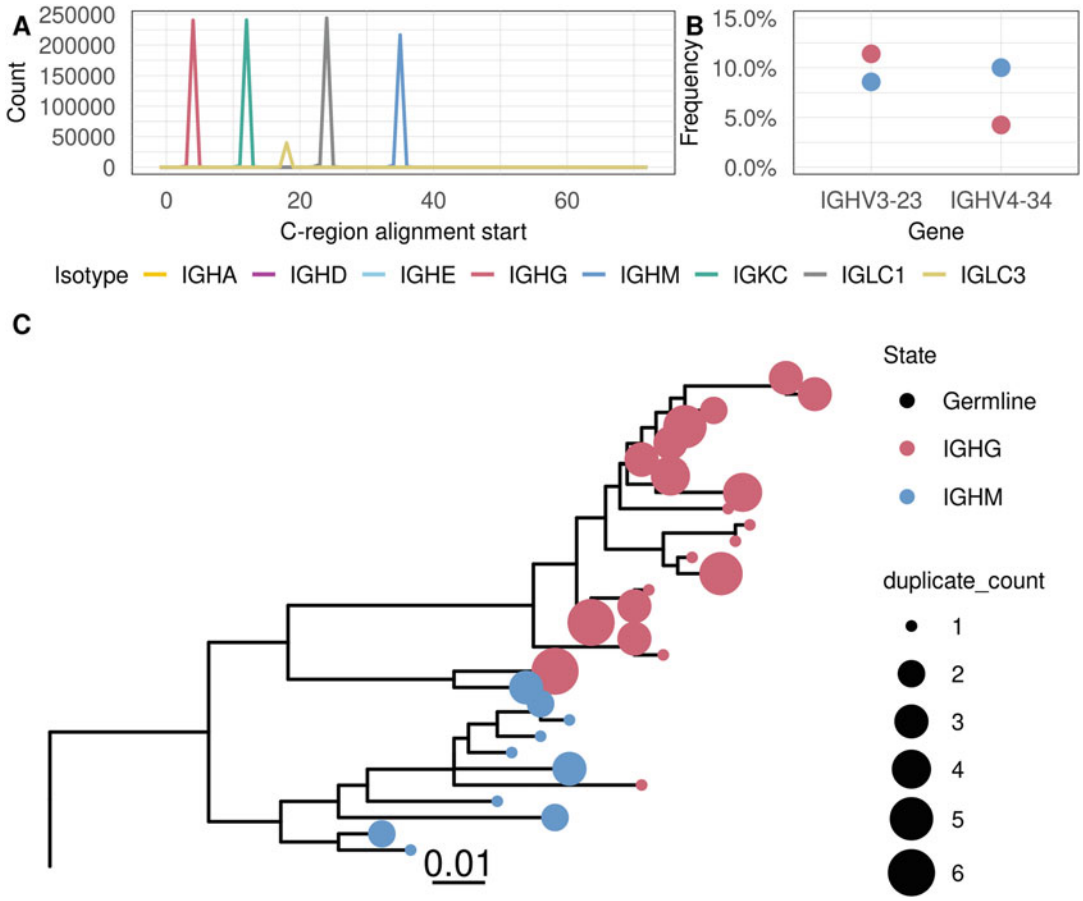


Fig. 7 Evaluation of B-cell isotype and clonal evolution. (**a**) Count and position of isotype primers. (**b**) Gene usage by isotype. (**c**) Reconstructed lineage tree

```
# Plot isotype primer position
cprimer_plot <- ggplot(cregion_table, aes(x=PRSTART, color=-
PRIMER)) +
 geom_freqpoly(size = 0.5,binwidth=1) +
 scale_color_manual(values = color_palette) +
 theme_minimal() +
 labs(x = "C-region alignment start", y = "Count", colour =
"Primer") +
  theme(legend.key.height = unit(0.1, "lines"), legend.key.
width = unit(0.5, "lines"))
cprimer_plot
```

7. Copy Annotations Between Reads

   Propagation of annotations between mate pairs is accomplished with PairSeq.py, which also removes unpaired reads and sorts mate pairs in both files. In this example, the UMI barcode is part of read VRR, and C_CALL is part of read VCC. We need to transfer this information to be able to build consensus sequences for groups of reads sharing the same UMI and C_CALL.

```
PairSeq.py -1 presto/VRR_primers-pass.fastq \
 -2 presto/CRR_primers-pass.fastq \
 --1f BARCODE --2f C_CALL --coord illumina
```

8. Generation of UMI Consensus Sequences

   If UMIs are available, it is possible to correct sequencing errors maintaining true mutations introduced by SHM. Reads sharing a UMI barcode are reads that originated from the same RNA molecule. Ideally, if the primers used are different enough, and the UMIs have enough diversity, each UMI will represent one mRNA molecule, and each mRNA molecule will be represented by one UMI. BuildConsensus.py can then be used to generate a consensus sequence for a set of aligned reads sharing the same UMI. Finding more than one primer in a UMI group suggests sequences may not be aligned, as we expect reads originating from the same mRNA molecule should be amplified with the same primer. If the multiplex pool contains similar primers, they could be incorporated into the same UMI group during amplification, and the reads will have variations in the start positions. This situation can be mitigated by first aligning the reads.

   (a) Multiple Align UID Read Groups

   If the reads are not aligned, a correction strategy is to use MUSCLE [5] and AlignSets.py to perform a multiple alignment of each UMI read group, before generating the consensus sequence in the next step.

```
AlignSets.py muscle -s "presto/VRR_primers-pass_pair-pass.
fastq" --exec /usr/local/bin/muscle --nproc 8 --log "logs/
align-vrr.log" --outname "VRR"
AlignSets.py muscle -s "presto/CRR_primers-pass_pair-pass.
fastq" --exec /usr/local/bin/muscle --nproc 8 --log "logs/
align-crr.log" --outname "CRR"
```

(b) Build the Consensus Sequence

BuildConsensus.py will group sequences sharing the same barcode to build a consensus sequence. If a UMI group has a number of average mismatches larger than 0.1 (`--maxerror 0.1`), it will be dismissed. Sequences with the same barcode have originated from the same original mRNA molecule, and they should also have the same isotype. `--pf C_CALL` and `--prcons 0.6` are used to require that 60% of the UMI group have the same C_CALL.

```
BuildConsensus.py -s presto/CRR_align-pass.fastq \
 --bf BARCODE --pf C_CALL --prcons 0.6 \
 -n 1 -q 0 --maxerror 0.1 --maxgap 0.5 \
 --nproc 8 --log "logs/consensus-crr.log" \
 --outdir presto --outname "CRR"
```

Example output:

```
@TGTTGGTTGGGT|CONSCOUNT=5|PRCONS=IGHM|
PRFREQ=0.8333333333333334
```

CONSCOUNT shows the number of sequences that contributed to build the consensus. In the example above, the consensus isotype (PRCONS) is IGHM, with a frequency of 0.83. In the starting UMI group, there were six sequences, and one of them was an IGKC. This sequence was not used to build the consensus.

The same process needs to be repeated for the other reads:

```
BuildConsensus.py -s presto/VRR_align-pass.fastq \
 --bf BARCODE --pf C_CALL --prcons 0.6 \
 -n 1 -q 0 --maxerror 0.1 --maxgap 0.5 \
 --nproc 8 --log "logs/consensus-vrr.log" \
 --outdir presto --outname "VRR"
```

9. Synchronize Reads

This step puts pairs of reads in the same order.

```
PairSeq.py -1 "presto/VRR_consensus-pass.fastq" -2 \
 "presto/CRR_consensus-pass.fastq" \
 --coord presto
```

10. Assemble Pairs

    Consensus sequences are paired in two steps, starting with joining overlapping mate pairs. For read pairs failing this step, the tool proceeds to perform a reference guided alignment, using ungapped V-segment reference sequences to properly space nonoverlapping reads.

```
AssemblePairs.py sequential -1 "presto/VRR_consensus-pass_pair-pass.fastq" \
 -2 "presto/CRR_consensus-pass_pair-pass.fastq" \
 -r /usr/local/share/igblast/fasta/imgt_human_ig_v.fasta \
 --coord presto --rc tail --1f CONSCOUNT --2f PRCONS CONSCOUNT \
 --minlen 8 --maxerror 0.3 --alpha 1e-5 --scanrev \
 --minident 0.5 --evalue 1e-5 --maxhits 100 --aligner blastn \
 --nproc 8 --log "logs/assemble.log" \
 --outname "S5"
```

    Example output: `@ACTAGGGTTCAT|CONSCOUNT = 4,4|PRCONS=IGHM`.

    `PRCONS` is the consensus `C_CALL` from the CRR file.

11. Mask Low-Quality positions

    Positions with a low consensus quality can be masked with Ns.

```
FilterSeq.py maskqual -s presto/S5_assemble-pass.fastq -q 30 --nproc 8 \
 --outname "S5-MQ" --log "logs/maskqual.log"
```

12. Track the Number of Sequences that Contributed to the Consensus

    It is important to know the number of unique sequences that contributed to build the consensus, as this information will be used in a later step.

```
ParseHeaders.py collapse -s presto/S5-MQ_maskqual-pass.fastq -f CONSCOUNT --act min \
 --outname "S5-final"
mv "presto/S5-final_reheader.fastq" "presto/S5-final_total.fastq"
```

13. Collapse Duplicates

    The goal is to remove duplicated sequences to retain in the repertoire one representative sequence per cell. The argument

"`-n 0 --inner`" will determine how to handle N and gap characters. In this example, we allow 0 ambiguous characters, ignoring any continuous N or gap characters that occur at any end of the sequence. "`---uf`" specifies fields that should be used to define groups of unique sequences. "`--cf CONSCOUNT`" requests to copy the field CONSCOUNT and then perform the action "`--act sum`," to obtain a final unique sequence with CONSCOUNT equal to the sum of the CONSCOUNTS of the collapsed sequences.

```
CollapseSeq.py -s "presto/S5-final_total.fastq" -n 0 \
 --uf PRCONS --cf CONSCOUNT --act sum --inner \
 --keepmiss --outname "S5-final"
```

14. Subset to Sequences Seen at Least Twice

    We recommend filtering the data to focus the analysis on sequences with at least two contributing reads. Sequences with CONSCOUNT of 1 are generated with only one sequence contributing to the UMI group, and this suggests the existence of sequencing error.

```
SplitSeq.py group -s presto/S5-final_collapse-unique.fastq -f
CONSCOUNT \
 --num 2
```

15. Explore the Logs

    All pRESTO tools provide the option to generate detailed logs that can be used to generate diagnostic plots. The log files can be converted to tabulated text files with `ParseLog.py`. The tabulated text files can be loaded into R or python to generate plots.

    (a) Obtain Tabulated Data

        The output files are parsed to generate tables of data for the repertoire.

```
ParseHeaders.py table -s "presto/S5-final_total.fastq" \
 -f ID PRCONS CONSCOUNT --outname "final-total" \
 --outdir logs
ParseHeaders.py table -s "presto/S5-final_collapse-unique.
fastq" \
 -f ID PRCONS CONSCOUNT DUPCOUNT --outname "final-unique" \
 --outdir logs
ParseHeaders.py table -s "presto/S5-final_collapse-unique_a-
tleast-2.fastq" \
  -f ID PRCONS CONSCOUNT DUPCOUNT --outname "final-unique-
atleast2" \
 --outdir logs
```

To see a summary of the final isotype assignments:

```
log <- read.delim("logs/final-unique-atleast2_headers.tab")
table(log$PRCONS)

 IGHE IGHG IGHM IGKC IGLC1 IGLC3
 1 48485 50955 37578 46879 7609
```

(b) Process the Log Files Generated at Each Step

Log files are also parsed into tabulated files.

```
ParseLog.py -l "logs/primers-vrr.log" -f ID BARCODE ERROR \
 --outdir logs
ParseLog.py -l "logs/consensus-vrr.log" "logs/consensus-crr.
log" \
 -f BARCODE SEQCOUNT CONSCOUNT PRIMER PRCONS PRCOUNT PRFREQ
ERROR \
 --outdir logs
ParseLog.py -l "logs/assemble.log" \
 -f ID REFID LENGTH OVERLAP GAP ERROR PVALUE EVALUE1 EVALUE2
IDENTITY FIELDS1 FIELDS2 \
 --outdir logs
ParseLog.py -l "logs/maskqual.log" -f ID MASKED \
 --outdir logs
```

3.11 Gene Annotation

Raw sequences which have passed general quality control filters should and then be annotated with gene information: for IGH sequences V, D, and J genes and for IGK/IGL only V and J genes. The IgBLAST executable and the reference database are available in the Immcantation container.

1. Convert FASTQ to FASTA

   IgBLAST takes as input FASTA file. The FASTQ files obtained at the end of the raw data processing section need to be converted to FASTA format. Simultaneously, rename `PRCONS` to `C_CALL`.

```
ParseHeaders.py rename -s presto/S5-final_collapse-unique_a-
tleast-2.fastq --fasta -f PRCONS -k C_CALL
```

2. Run IgBLAST

   The wrapper tool AssignGenes.py, from Change-O [6], uses IgBLAST, and a reference database created with germlines from IMGT, to make V(D)J allele calls.

```
mkdir changeo
AssignGenes.py igblast -s presto/S5-final_collapse-unique_a-
tleast-2_reheader.fasta \
 --organism human --loci ig \
 -b /usr/local/share/igblast --format blast --nproc 8 \
 --outdir changeo --outname "S5"
```

3. Data Standardization

   IgBLAST's results need to be converted into an AIRR-formatted file (https://immcantation.readthedocs.io/en/stable/datastandards.html) suitable for downstream analysis.

```
MakeDb.py igblast -s presto/S5-final_collapse-unique_atleast-
2_reheader.fasta \ -i changeo/S5_igblast.fmt7 \
--extended --failed --format airr \
-r /usr/local/share/germlines/imgt/human/vdj/ --outname S5
```

   Some sequences don't pass this MakeDb step with these settings. This could be because a junction could not be identified, there are Ns in the junction, there is a stop codon, or the reads are partial, among other possible reasons.

3.12 Quality Control After Gene Assignment

Once sequences have been annotated with allele calls, and the aligned rearranged sequence is available, further collapsing of duplicates and removal of low-quality sequences is possible. Here we demonstrate how to perform some common additional QC steps using R and Immcantation tools (alakazam [6]). The goal is to keep sequences with at least 200 informative positions, with coherent gene and locus calls, and with a limited number of ambiguous nucleotides. It is also common to focus the analysis in productive sequences. Here we will keep productive sequences and will remove sequences with junction length that is not a multiple of three. Finally, chimeric reads will be identified and removed.

1. Identify Short Sequences.

```
library(airr)
library(alakazam)
library(stringi)
library(dplyr)
airr <- read_rearrangement("changeo/S5_db-pass.tsv")
# Min length 200 nt
long_seq <- stri_count(airr[['sequence_alignment']],re-
gex="[^-.N]") >= 200
```

2. Identify Reads with Coherent Gene, Primer, and Isotype Calls

   The goal is to remove sequences with incoherent gene and isotype calls. For example, a sequence that has a V gene

assigned, but an IG light-chain-constant region will be removed.

```
# Keep reads with coherent gene, primer and isotype calls
same_locus <- getLocus(airr[['v_call']]) == airr[['locus']] &
 getLocus(airr[[c_call]]) == airr[['locus']]
```

3. Identify Reads with an Acceptable Number of Ambiguous Nucleotides.

```
# Max 10% N
num_n <- stri_count(airr[['sequence_alignment']],fixed="N")
len <- stri_count(airr[['sequence_alignment']],regex="[^-.]")
low_n <- num_n/len <= 0.10
```

4. Identify Productive Sequences.

```
prod <- airr[['productive']]
```

5. Identify Sequences with Junction Length Multiple of Three.

```
m3 <- airr[['junction_length']] %% 3 == 0
```

6. Filter and Save.

```
filter_pass <- long_seq &
 same_locus &
 low_n &
 prod &
 m3
write_rearrangement(airr[filter_pass,], file="changeo/S5_fil-
ter-pass.tsv")
```

7. Reconstruct Germline Sequences

   Identify the V(D)J germline sequences from which each of the sequences is derived. These germlines will be used to analyze mutation patterns in a sliding window to identify chimeric sequences.

```
CreateGermlines.py -d changeo/S5_filter-pass.tsv \
-r /usr/local/share/germlines/imgt/human/vdj \
 -g dmask --format airr
```

8. Identify and Remove Chimeric Sequences

   Chimeric sequences can be identified by analyzing their mutation frequencies. The function `slideWindowDb`, from shazam [6], identifies which sequences in the repertoire

contain excessive mutations in a given length of consecutive nucleotides (a "window") when compared to their respective germline sequence.

```
library(airr)
library(shazam)

airr <- read_rearrangement("changeo/S5_filter-pass_germ-pass.
tsv")
is_chimeric <- slideWindowDb(
 airr,
 sequenceColumn = "sequence_alignment",
 germlineColumn = "germline_alignment_d_mask",
 mutThresh=6,
 windowSize=10
)
table(is_chimeric)
airr <- airr[!is_chimeric,]
```

9. Collapse Duplicates

   Once the sequences in the repertoire are aligned following the IMGT scheme, further collapsing of duplicate sequences can be done with the function collapseDuplicates.

```
library(dplyr)
num_fields <- c("consensus_count", "duplicate_count")

# Data comes one sample, so no need to add
# sample identifier groups
collapse_groups <- c("v_gene",
 "j_gene",
 "junction_length",
 "c_call",
 "productive")

airr <- airr %>%
 mutate(v_gene=getGene(v_call),
 j_gene=getGene(j_call)) %>%
 group_by(.dots=collapse_groups) %>%
 do(collapseDuplicates(.,
 id = "sequence_id",
 seq = "sequence_alignment",
 text_fields = NULL,
 num_fields = num_fields,
 seq_fields = NULL,
 add_count = TRUE,
 ignore = c("N", "-", ".", "?"),
 sep = ",",
```

```
 dry = FALSE,
 verbose = FALSE
 )) %>%
 ungroup() %>%
 select(-v_gene, -j_gene)
```

3.13 Identify Clonally Related Sequences

The goal is to partition sequences into clonal lineages. Each clonal lineage is a group of sequences derived from the same original cell. There are several methods to identify clonal lineages (*see* Subheading 3.9.2: Identification of B-Cell Clones in the chapter "AIRR Community Guide to Repertoire Analysis"). Here, we first group by V gene, J gene, and junction length. Then we compare the junctions and apply a threshold to separate sequences into clonal lineages.

1. Calculate the Distance to the Nearest Distribution

   Hierarchical clustering requires a measure of distance between pairs of sequences and a choice of linkage to define the distance between groups of sequences. The result is a hierarchy, and a threshold is needed to cut the tree into clonal groups.

```
# Subset to heavy chain sequences
airr_heavy <- airr %>%
 filter(locus == "IGH")

# Group by V gene, J gene and junction length, and calculate
the distance
# to the nearest sequence in the group
airr_heavy <- distToNearest(airr_heavy, sequenceColumn="junc-
tion",
 vCallColumn="v_call", jCallColumn="j_call",
 model="ham", first=FALSE, normalize="len",
 nproc=params$nproc)
write_rearrangement(airr_heavy,  file="changeo/IB7_heavy_col-
lapse-pass.tsv")
```

2. Find a Threshold

   It is possible to determine a threshold by analyzing the distribution of the distances. The distribution is usually bimodal. The first mode represents sequences that have a close relative. The second mode is representative of sequences without clonal relatives. The goal is to select a threshold that separates the two modes.

```
threshold <- findThreshold(airr_heavy[['dist_nearest']],
method="density")
plot(threshold, binwidth=0.02, silent=FALSE)
clone_threshold <- round(threshold@threshold,)
clone_threshold
```

3. Identify Clonally Related Sequences

   Once a threshold is selected, it is applied to identify groups of related sequences:

```
DefineClones.py -d changeo/S5_heavy_collapse-pass.tsv --model
ham \
 --dist 0.09 --mode gene --act set --nproc 8 \
  --outname S5 --outdir changeo --format airr --log "logs/
clone.log"
```

4. Reconstruct Clonal Germline

   The next step is to identify the V(D)J germline sequences from which each of the observed sequences is derived. These germlines are used as the reference to analyze mutations.

```
CreateGermlines.py -d changeo/S5_clone-pass.tsv \
-r /usr/local/share/germlines/imgt/human/vdj \
 -g dmask --format airr --cloned --outname S5-airr
```

3.14 Gene Usage by Isotype

When isotype information is available, it is possible to investigate biases in gene usage at the isotype level.

```
library(alakazam)
library(airr)
library(dplyr)

airr <- read_rearrangement("changeo/S5-airr_germ-pass.tsv")

# Gene usage by Isotype for one sample with only heavy chain
data
v_usage_isotype <- countGenes(airr, "v_call",group="c_call",
fill=T)

most_used_v <- v_usage_isotype %>%
 filter(c_call != "IGHE") %>%
 group_by(c_call) %>%
 slice_max(.,seq_freq,n=1)

# Plot the most used V gene(s)
library(scales)
```

```
gene_usage_plot <- ggplot(v_usage_isotype %>%
 filter(c_call != "IGHE" & gene %in% most_used_v[['gene']]),
 aes(x=gene,y=seq_freq, color=c_call)) +
 scale_color_manual(values=color_palette) +
 scale_y_continuous(labels=percent) +
 geom_point(size=2) + theme_minimal() +
 xlab("Gene") + ylab("Frequency") +
 guides(color=guide_legend(title="Isotype"))
gene_usage_plot
```

The gene usage by isotype can be visualized in Fig. 7b.

3.15 Clonal Lineage Tree Analysis

Dowser [7] provides tools for building and visualizing IG lineage trees using multiple methods and implements statistical tests for discrete trait analysis of B-cell migration, differentiation, and isotype switching.

1. Format

   First, data must be formatted into a data table of AIRR clone objects. The `formatClones` function will change non-nucleotide characters to N characters, collapse sequences that are either identical or differ only by ambiguous characters, and remove uninformative sequence sites in which all sequences have N characters.

```
library(dowser)
clones <- formatClones(airr, traits=c("c_call"),
 num_fields=c("duplicate_count"), columns=c("d_call"),
 minseq=10)
```

2. Build the Trees

   There are several lineage reconstruction methods implemented in dowser. Maximum parsimony trees (topologies that minimize the number of mutations needed along the tree) can be built with the `getTrees` function.

```
# build maximum parsimony trees
clones <- getTrees(clones)
```

3. Visualize

   The function `plotTrees` makes plotting lineages easy. Branch lengths by default represent the number of mutations per site between nodes. It is also possible to show numerical or categorical information associated with the tree tips, such as the duplicate count or the isotype.

```
# Plot the trees. Save them in a list of plots.
# Use tip metadata: c_call and duplicate_count
```

```
tree_plots <- plotTrees(clones, tips="c_call",
 tipsize="duplicate_count",
 tip_palette=c(color_palette, "Germline"="#000000"))
```

Example output is given in Fig. 7c.

4 Notes

1. Other reagents may be substituted but require additional optimization to ensure adequate performance of the protocol. Store BCR enhancer at −20 °C. Once thawed, the buffer can be stored at 4 °C. Store nuclease-free water at −20 °C. Once thawed, the water can be stored at 4 °C. Store elution buffer at −20 °C. Once thawed, the buffer can be stored at room temperature.
2. The success of the experiment depends on the quality of the input RNA. The RNA should be of high integrity (RIN > 7) to enable oligo(dT) priming. Prior to cDNA synthesis, ensure that the RNA is in nuclease-free water, is intact, and is free of contaminants. Input RNA should also be free from poly (A) carrier RNA that will interfere with oligo(dT)-primed cDNA synthesis.
3. The First-Strand Buffer may form precipitates. Thaw this buffer at room temperature, and vortex before using to ensure all components are completely in solution.
4. Each reverse PCR primer is used in a separate PCR Master Mix. Alternatively, plan to add 1 μL of each primer individually instead of including them in the PCR1 Master Mix, particularly if the number of samples is low.
5. Different combinations of hBCR PCR2 Universal Forward 1–12 and hBCR PCR2 IgG/IgM/IgK/IgL reverse 1–4 indices must be used for each sample if samples are to be pooled and loaded on a single flow cell.
6. Each PCR primer is used in a separate PCR Master Mix. Alternatively, plan to add 1 μL of each primer individually instead of including them in the PCR2 Master Mix, particularly if the number of samples is low.
7. Aliquot NucleoMag beads into 1.5-mL tubes upon receipt in the laboratory. Before each use, bring bead aliquots to room temperature for at least 30 min, and mix well to disperse. Prepare fresh 80% ethanol for each experiment. You will need 400 μL per sample. You will need a magnetic separation device for 0.2-mL tubes, strip tubes, or a 96-well plate.

8. The beads are viscous; pipette the entire volume and push it out slowly. Do not vortex. Vortexing will generate bubbles, making subsequent handling of the beads difficult.
9. Ensure that the solution is completely clear, as any bead carry-over will decrease the efficiency of size selection. There is no disadvantage to separating the samples for longer than 5 min.
10. Ensure that the beads are fully resuspended before use. If the beads appear to have settled at the bottom of the tube, vortex to ensure that they are completely mixed.
11. Be sure to dry the pellet only until it is just dry. The pellet will look matte with no shine.
12. Be sure that the beads are completely resuspended. The beads can sometimes stick to the sides of the tube.
13. Gently pipette any remaining beads that are in suspension toward the magnet where the rest of the beads have already pelleted. Continue the incubation until there are no beads left in the supernatant.
14. Follow Illumina guidelines on how to denature, dilute, and combine a PhiX control library with your own pool of libraries. Make sure to use a fresh and reliable stock of the PhiX control library.

Acknowledgments

The authors would like to thank Eline Luning Prak for assistance with manuscript content and formatting of text and Chaim Schramm, Johannes Truck, and Wenwen Xiang for constructive criticism of the manuscript. Conflict of interest: NG and AF are employees at Takara Bio, Inc., San Jose, CA, USA, that produces the kit described in this protocol.

References

1. Vander Heiden JA, Yaari G, Uduman M, Stern JNH, O'Connor KC, Hafler DA et al (2014) pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. Bioinformatics 30: 1930–1932. https://doi.org/10.1093/bioinformatics/btu138
2. Mukherjee S, Huntemann M, Ivanova N, Kyrpides NC, Pati A (2015) Large-scale contamination of microbial isolate genomes by Illumina PhiX control. Stand Genomic Sci 10:18. https://doi.org/10.1186/1944-3277-10-18
3. Giudicelli V, Chaume D, Lefranc M-P (2005) IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. Nucleic Acids Res 33: D256–D261. https://doi.org/10.1093/nar/gki010
4. Ye J, Ma N, Madden TL, Ostell JM (2013) IgBLAST: an immunoglobulin variable domain sequence analysis tool. Nucleic Acids Res 41: W34–W40. https://doi.org/10.1093/nar/gkt382
5. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797. https://doi.org/10.1093/nar/gkh340

6. Gupta NT, Vander Heiden JA, Uduman M, Gadala-Maria D, Yaari G, Kleinstein SH (2015) Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. Bioinformatics 31:3356–3358. https://doi.org/10.1093/bioinformatics/btv359

7. Hoehn KB, Pybus OG, Kleinstein SH (2020) Phylogenetic analysis of migration, differentiation, and class switching in B cells. Immunology. https://doi.org/10.1101/2020.05.30.124446

Chapter 20

Single-Cell Analysis and Tracking of Antigen-Specific T Cells: Integrating Paired Chain AIRR-Seq and Transcriptome Sequencing: A Method by the AIRR Community

Nidhi Gupta, Ida Lindeman, Susanne Reinhardt, Encarnita Mariotti-Ferrandiz, Kevin Mujangi-Ebeka, Kristen Martins-Taylor, and Anne Eugster

Abstract

Single-cell adaptive immune receptor repertoire sequencing (scAIRR-seq) offers the possibility to access the nucleotide sequences of paired receptor chains from T-cell receptors (TCR) or B-cell receptors (BCR). Here we describe two protocols and the downstream bioinformatic approaches that facilitate the integrated analysis of paired T-cell receptor (TR) alpha/beta (TRA/TRB) AIRR-seq, RNA sequencing (RNAseq), immunophenotyping, and antigen-binding information. To illustrate the methodologies with a use case, we describe how to identify, characterize, and track SARS-CoV-2-specific T cells over multiple time points following infection with the virus. The first method allows the analysis of pools of memory $CD8^+$ cells, identifying expansions and contractions of clones of interest. The second method allows the study of rare or antigen-specific cells and allows studying their changes over time.

Key words Single-cell sequencing, TR gene, IG gene, Rearrangement, Transcriptome, 10x Genomics, SMART-seq, Multi-omic analysis

1 Introduction

Single-cell adaptive immune receptor repertoire sequencing (scAIRR-seq) aims at describing the sequences of T-cell receptor (TR) or immunoglobulin (IG) rearrangements at the single-cell level. scAIRR-seq has been used since the mid-1990s [1, 2] and has seen rapidly increasing adoption by the scientific community over the last few years [3]. This has been facilitated by a plethora of protocols, commercial kits, and platforms as well as by the associated software tools developed in the last decade as discussed in detail in the two AIRR Community commentary chapters in this

Nidhi Gupta, Ida Lindeman, and Susanne Reinhard are shared first authors.

Anton W. Langerak (ed.), *Immunogenetics: Methods and Protocols*, Methods in Molecular Biology, vol. 2453, https://doi.org/10.1007/978-1-0716-2115-8_20,

volume. Workflows for scAIRR-seq distinguish themselves from bulk methods by several features: First, they preserve the chain-pairing information of the complete IG/TR, which is critical for the experimental reconstruction and measurement of receptor reactivities. Second, they allow an unbiased view of clonal expansion, as observed sequences can be attributed to and normalized by the individual cell from which they originate. Third, using the individual cell as a common reference point, they allow the integration with other single-cell resolution data, such as transcriptome, cell-surface phenotype as well as antigen-specificities. All of this, however, comes at a reduced throughput and lower sensitivity for the detection of rare clones when compared to bulk sequencing.

The vast majority of currently used scAIRR-seq methods, including the two presented here (Fig. 1), maintains the compartmentalization of the cells throughout the process, either physically or via barcoding [4]. The 10x Genomics Chromium is a microfluidic-based platform, which allows the encapsulation and barcoding of up to 3000 to 10,000 cells at a time. The Chromium Next GEM Single Cell V(D)J Reagent Kits described here allow the generation of three libraries (Fig. 2): (1) full-length, paired AIRR sequences, (2) the cell transcriptome (derived from all polyadenylated transcripts), and (3) feature barcodes linked to the surface protein expression and antigen specificity (e.g., CITE-seq) as well as barcoding of libraries for multiplexing (e.g., hash-tagging). Single-cell SMART-seq is a method to collect single-cell AIRR-seq and gene expression data from cells which are sorted into 96-well plates (Fig. 3 and Fig. 4), allowing the analysis of rare cells. Single-cell SMART-seq is based on the SMART® (switching mechanism at 5′ end of RNA template) technology. The SMART-Seq Single Cell Kit used to generate mRNA-seq libraries is particularly useful for the analysis of cells with very low RNA content, such as PBMCs.

In both methods AIRR-seq as well as transcriptome sequences are obtained from RNA, making use of the template-switching activity of reverse transcriptase to enrich for full-length cDNAs and to add defined PCR adapters directly to both ends of the first-strand cDNA. This ensures that the final cDNA libraries contain the 5′ end of the mRNA and maintain a true representation of the original mRNA transcripts. These factors are critical for AIRR-seq and transcriptome sequencing.

A use case for these two methods is to identify, characterize, and track SARS-CoV-2- or other virus-specific T cells over multiple time points following infection, remittance, or vaccination. The Chromium Next GEM Single Cell V(D)J method can be used to broadly analyze thousands of memory $CD8^+$ cells from several time points of an individual, before and after the immunizing event, in a multiplexed manner (through hash-tagging). This allows identifying expansions and contractions of clones of interest (through AIRR-seq), phenotyping them (through transcriptome analysis

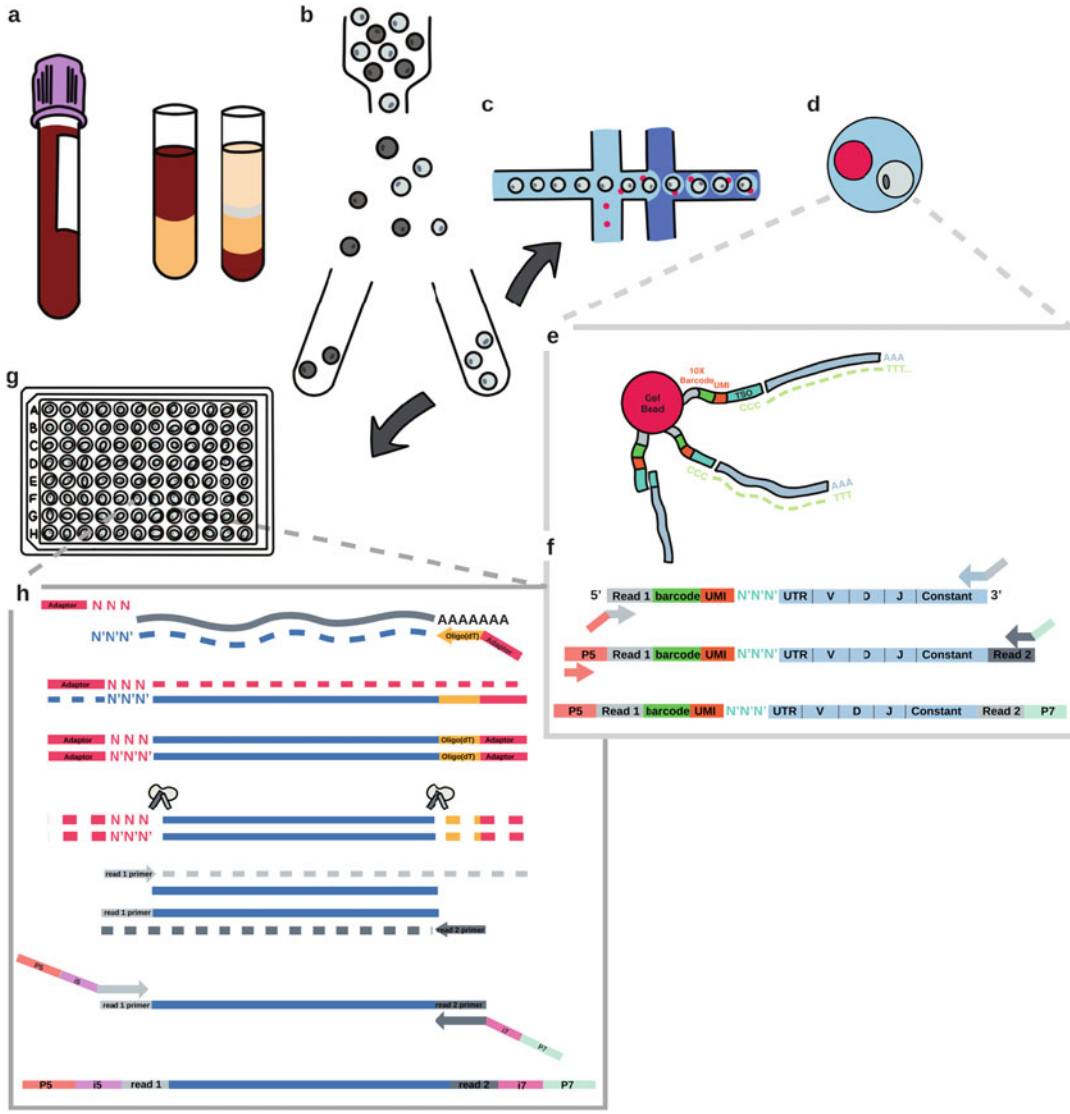


Fig. 1 Schematic illustrating the main characteristics of single cell paired chain AIRR-seq. (**a**) A blood sample is processed by Ficoll gradient centrifugation to obtain PBMC. (**b**) PBMCs are stained and cells of interest sorted by FACS. As described in Subheading 3.1, using the 10x Genomics fluidics system (panel **c**), cells are processed for transcript barcoding. After encapsulation in a droplet, a "GEM" is created (panel **d**). Gel bead primers containing a 10× barcode, a UMI, and a template switch oligo (TSO) bind to the transcripts after cell lysis (panel **e**). Gel bead primers also capture the cell surface feature barcodes. Barcoded transcripts and feature barcodes are then reverse transcribed, and through size selection and enrichment, a library containing amplified AIRR (panel **f**), a library containing whole cell transcriptome, and one containing feature barcodes are prepared through the sequential additions of primers containing the P5 and P7 sites required for sequencing. As described in Subheading 3.1, cells are deposited into a plate (panel **g**). Here reverse transcription, cDNA amplification, and preparation of a library containing amplified AIRR and of a library containing whole cell transcriptome take place through the sequential additions of primers that include the P5 and P7 sites required for sequencing by the SMARTseq method (panel **h**)

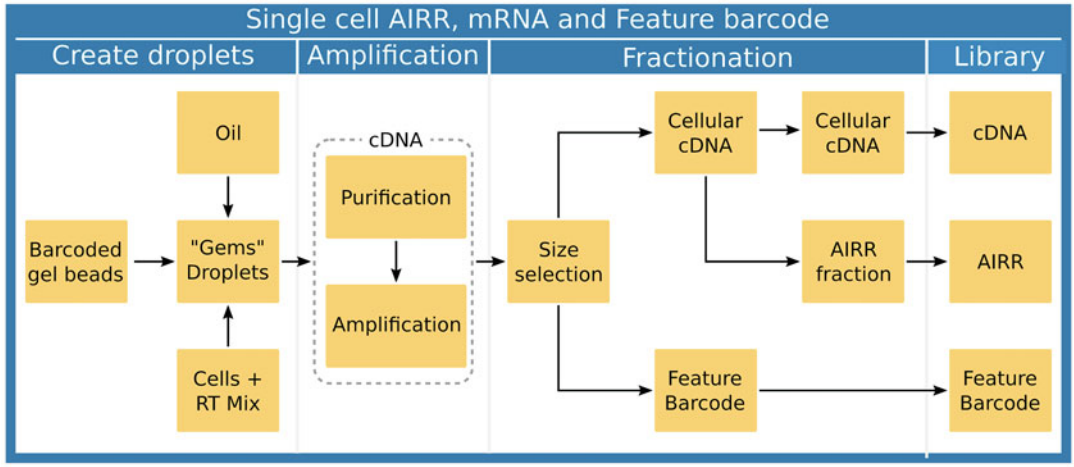


Fig. 2 Overview of the main steps of the Chromium Next GEM Single-Cell procedure: The creation of droplets (left), in which the RNA is captured and barcoded is followed by breaking the GEMs, the amplification of cDNA, the fractionation that allows separation of cDNA from feature barcodes from cellular cDNA, and finally by the preparation of the three libraries

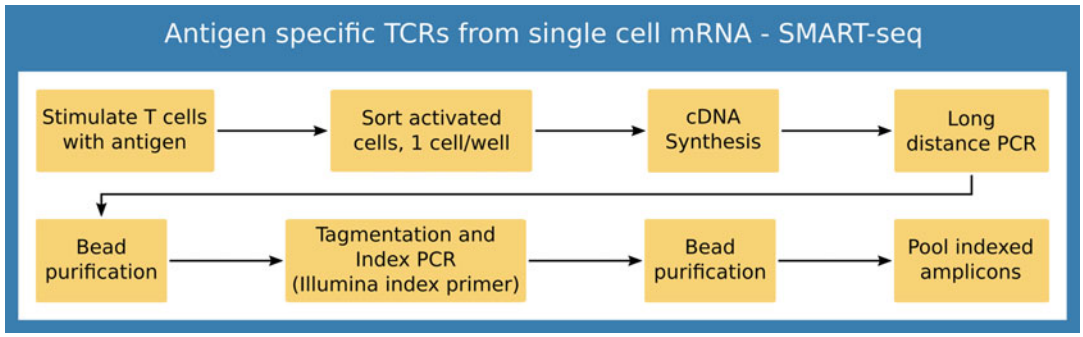


Fig. 3 Overview of main steps for SMART-seq scTCR procedure: Stimulated cells are sorted into PCR plates, followed by cDNA synthesis, two rounds of PCR and purification, and are finally pooled to prepare the library for sequencing

and feature barcode analysis), and correlating them to disease status. Clones or cells of interest can be defined through their activation state or their antigen specificity and are isolated by flow cytometry after surface marker or multimer staining, respectively. Clones or cells of interest can be activated $CD8^+CD25^+CD137^+$ T cells [5] from several time points during and after a viral infection, or cells stimulated with an antigen of interest. Single-cell SMART-seq of these often rare clones after isolation gives access to their AIRR data that can then be matched to the data obtained from Chromium Next GEM Single Cell V(D)J. Here we provide protocols and detailed information for the generation, processing, and analysis of scAIRR-seq- and associated data produced with the two platforms described (Fig. 5).

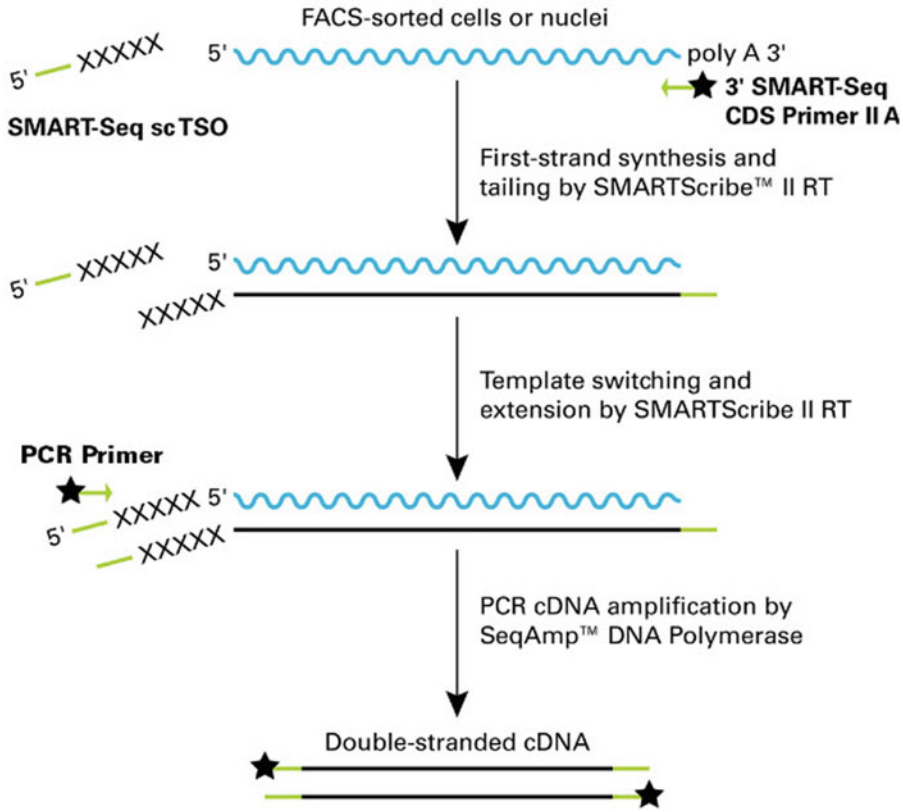


Fig. 4 Schematic of the technology in the SMART-Seq Single-Cell Kit. Non-templated nucleotides (indicated by Xs) added by the SMARTScribe II reverse transcriptase (RT) hybridize to the SMART-Seq single-cell template-switching oligonucleotide (SMART-Seq sc TSO), which provides a new template for the RT. The SMART adapters used for amplification during PCR added by the oligo(dT) primer (3′ SMART-Seq CDS Primer II A) and TSO are indicated in green. Chemical modifications to block ligation (if using a ligation-based library preparation method) are present on some primers (indicated by black stars)

2 Materials

2.1 10x Genomics Chromium Next GEM Single-Cell V(D)J Kit

1. Single-channel pipette: 10 μl, 20 μl, 200 μl, and 1000 μl.
2. 8-channel or 12-channel pipette (recommended): 20 μl and 200 μl.
3. Filter pipette tips: 2 μl, 20 μl, 200 μl, and 1000 μl.
4. Minicentrifuge for 1.5-ml tubes.
5. Minicentrifuge for 0.2-ml tubes or strip.
6. Bioanalyzer or TapeStation for library validation.
7. All the related cell-sorting equipment.

cDNA Synthesis and Amplification

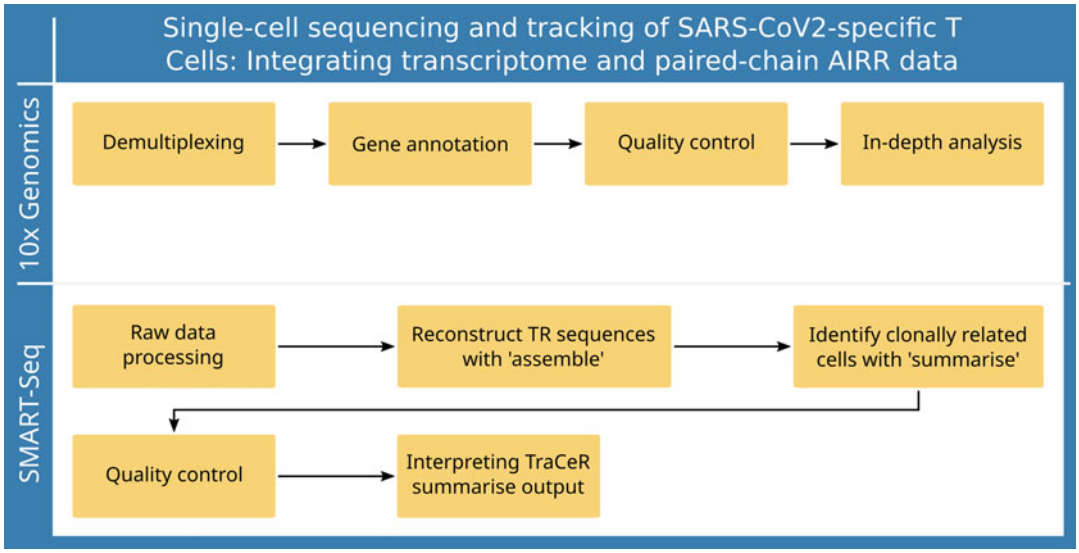


Fig. 5 Overview of the main steps of the analysis of single-cell AIRR-seq data. Libraries created with the 10x Genomics technology (upper panel) are processed using the CellRanger software. In brief, sequencing libraries are demultiplexed before TR sequences are extracted and annotated. The quality of each library is assessed, and TR sequences may be combined with transcriptional and feature libraries for an in-depth integrated analysis. For libraries created by plate-based sequencing technologies such as SMART-seq (lower panel), TR sequences are computationally reconstructed with TraCeR. Low-quality cells or potential duplets may be filtered out, before clonally related cells are identified and visualized in clonal networks

8. Two thermal cyclers with heated lids (*see* **Note 1**): one thermal cycler used only for first-strand cDNA synthesis; one thermal cycler used only for double-stranded cDNA amplification by PCR.
9. 96-well semi-skirted plates (Thermo Fisher Scientific) or 8-tube strips (Thermo Fisher Scientific).
10. Thermo Scientific Adhesive PCR Plate Seals (Thermo Fisher Scientific) for 96-well plates or flat cap strips (Thermo Fisher Scientific) for 96-well plates or 8-tube strips.

10x Genomics Kits and Reagents (10X Genomics, Unless Mentioned)

10. Chromium Next GEM Single Cell 5′ Library and Gel Bead Kit v1.1, 16 rxns.
11. DynaBeads® MyOne™ Silane Beads (Thermo Fisher Scientific).
12. Chromium Single Cell 5’ Library Construction Kit, 16 rxns.
13. Chromium Single Cell 5’ Feature Barcode Library Kit, 16 rxn.

14. Chromium™ Single Cell V(D)J Enrichment Kit, human T cell/mouse T cell.
15. Chromium™ Single Cell V(D)J Enrichment Kit, human B cell/mouse B cell.
16. Chromium Next GEM Chip G Single Cell Kit, six chips.
17. Single Index Kit N Set A, Single Index Kit T Set A.

Other Supplies

18. Nuclease-free water.
19. Low TE buffer: 10 mM Tris–HCl pH 8.0, 0.1 mM EDTA.
20. Ethanol, pure (200 Proof, anhydrous).
21. Tween 20.
22. Glycerin (glycerol), 50% (v/v) aqueous solution.
23. Elution buffer (EB).
24. NGS HS Fragment Analysis Kit (Agilent) or comparable chemistry to run QC.
25. Bovine serum albumin (BSA).
26. SPRIselect Reagent Kit (Beckman Coulter).

Bead Purifications

27. NucleoMag NGS cleanup and size select (*see* **Note 2**) Takara Bio) or the AMPure XP PCR Purification Kit (Beckman Coulter).

For cDNA and Illumina Library Quantification and Preparation

28. High-Sensitivity DNA Kit (Agilent) for bioanalyzer or equivalent high-sensitivity electrophoresis method (may be used in Sections V.D and VI.D).
29. Quant-iT PicoGreen dsDNA Assay Kit (Thermo Fisher Scientific) or Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific).
30. Library Quantification Kit (Takara Bio).
31. Nuclease-free, PCR-grade, thin-wall PCR strips or 96-well plates.
32. Nuclease-free, low-adhesion 1.5-ml tubes.
33. Benchtop cooler, such as VWR CryoCoolers.

Cell Preparation

34. Benzonase (10 U/ml).

2.2 Single-Cell SMART-Seq

The same equipment and supplies are used as for 10× Chromium, except for the following:

General Lab Equipment

1. 96-well PCR chiller rack, such as IsoFreeze PCR Rack, or 96-well aluminum block.

Sample Preparation

2. Nuclease-free, PCR-grade 8-tube strips secured in a PCR rack, or 96-well plates that have been validated to work with your FACS instrument.
3. Microplate film for sealing tube strips/plates before sorting.
4. Aluminum single-tab foil seal or cap strips for sealing tubes/plates after sorting.
5. Low-speed benchtop centrifuge for 96-well plates or tube strips.
6. Dry ice in a suitable container for flash freezing cell.

Bead Purifications

7. 80% ethanol: freshly made for each experiment from molecular-biology-grade 100% ethanol.
8. Strong magnetic separation device for plates that accommodates 96 samples in 96-well V-bottom plates (500 μl; VWR).
9. Low-speed benchtop centrifuge for a 96-well plate.
10. Strong magnetic separation device for tubes.
11. Nuclease-free, PCR-grade 8-tube strips secured in a PCR rack, or 96-well plates.

Sequencing Library Generation

12. Nextera XT DNA Library Preparation Kit (Illumina).
13. Nextera XT Index Kit (Illumina) or other Nextera-compatible indexes.
14. 96-well PCR plate.
15. Microseal "B" adhesive seals.
16. Freshly prepared 80% ethanol (EtOH), as above.
17. 96-well 0.8-ml polypropylene deep-well storage plate (midi plate).
18. Microseal "F" foil seals.
19. Nuclease-free water.

cDNA Synthesis (Takara Bio Unless Otherwise Specified)

20. 1 μg/μl control total RNA.
21. 10× lysis buffer.
22. 40 U/μl RNase inhibitor.
23. Nuclease-free water.
24. SMART-Seq sc TSO.
25. 3′ SMART-Seq CDS Primer II A.

26. SMARTScribe II reverse transcriptase.
27. SMART-Seq sc First-Strand Buffer (5×).
28. SeqAmp DNA polymerase.
29. 2× SeqAmp CB PCR buffer.
30. PCR primer.
31. 10 mM Tris-Cl elution buffer (pH 8.5).

Nextera Library Preparation (Illumina Unless Otherwise Mentioned)

32. Amplicon Tagment Mix (ATM).
33. Tagment DNA Buffer (TD).
34. Neutralize Tagment Buffer (NT).
35. Nextera PCR Master Mix (NPM).
36. Resuspension Buffer (RSB).

Nextera Indices

37. Index 1 (i7) Adapters (N7xx—Nextera XT Index Kit v2, Nextera Index Kit): N701; TCGCCTTA, N702; CTAGTACG, N703; TTCTGCCT, N704; GCTCAGGA, N705; AGGAGTCC, N706; CATGCCTA, N707; GTAGAGAG, N710; CAGCCTCG, N711; TGCCTCTT, N712; TCCTCTAC, N714; TCATGAGC, N715; and CCTGAGAT (i7 index name; bases in adapter).
38. Index 2 (i5) adapters (S5xx—Nextera XT Index Kit v2): CTCTCTAT, S503; TATCCTCT, S505; GTAAGGAG, S506; ACTGCATA, S507; AAGGAGTA, S508; CTAAGCCT, S510; CGTCTAAT, S511; and TCTCTCCG (i5 index name; bases in adapter S502).

2.3 10x Genomics Data Processing and Analysis

1. Cell Ranger software, provided by 10x Genomics free of charge, required for raw data handling and various secondary analysis.
2. Linux workstation (minimum 8 cores, 64 GB RAM, 1 TB storage) running a recent version of CentOS/RHEL or Ubuntu, to run Cell Ranger.
3. Loupe Brower and Loupe VDJ Browser, also available via the 10x Genomics website, which provide a complementary set of analysis tools.
4. Windows or Macintosh operating systems to run Loupe Browser.

2.4 Single-Cell SMART-Seq Data Processing and Analysis

1. TraCeR, installed as a Docker container, with all of its dependencies installed and properly configured. Alternatively, TraCeR can be installed from GitHub.
2. Dependencies, installed and configured by the user (see https://github.com/Teichlab/tracer), (if installing TraCer from GitHub).

3 Methods

3.1 10x Genomics Chromium Next GEM Single-Cell V(D)J Kit

Before starting, please refer to considerations regarding the kits used (*see* **Note 3**), sample multiplexing (*see* **Note 4**), and surface protein detection (*see* **Note 5**).

3.1.1 Coat Tubes for Cell Sort and Count Cells

1. Fill PCR tubes with 1% BSA in PBS and incubate them overnight. Remove the BSA completely by shortly centrifuging the tubes after the first removal and removing the collected liquid. Prepare one tube for each sample and pre-lye 1 μl PBS.
2. After staining and washing (*see* **Note 6**), sort cells into the tube (*see* **Note 7**), and use 1–2 μl from the sample to verify the cell quality and number (*see* **Note 8**) under a light microscope. Proceed to loading the chip taking into account the time required for the sort (*see* **Note 9**).

3.1.2 Load Next GEM Chip G

To avoid contamination, this section should be carried out on a separate bench dedicated to RNA/cell work.

1. Equilibrate Single-Cell VDJ 5' Gel Beads v1.1 (−80 °C), RT reagent B, additive A, poly-dT-RT primer (−20 °C), 50% glycerol solution (max. 1.4 ml per chip) to room temperature for 30 min; place RT Enzyme Mix B and cell suspension on ice.
2. Prepare RT mix, by mixing 18.8 μl of RT Reagent B (blue lid), 6.4 μl poly-dT RT primer (blue lid), 2 μl additive A (blue-green lid), and 10 μl RT Enzyme Mix B (white lid), resulting in a total volume of 37.2 μl. Mix 15 times, centrifuge, and keep on ice.
3. Place a Next GEM G Chip in a 10× chip holder, and add 50% glycerol in the following order: (1) 70 μl in row 1, (2) 50 μl in row 2, and (3) 45 μl in row 3.
4. Dispense 34.8 μl RT Mix into an eight-stripe tube per sample, and add the amount of water needed to dilute to the final cell concentration and to obtain 35.5 μl cell suspension; resuspend the cells very slowly and add them to the RT mix. Each tube should contain a total volume of 70.3 μl.
5. With a pipette set to 70 μl slowly mix the cell-RT mix five times, and finally transfer them to the wells in row 1 in the chip. Do not introduce bubbles.
6. Vigorously vortex gel beads for 30 s, and flick sharply to collect beads at the bottom (meanwhile the cell RT mix is priming on the chip). Slowly load 50 μl of the viscous gel beads with a multichannel pipette into row 2 in the chip (avoid bubbles).
7. Pipette 45 μl partitioning oil into row 3 of the chip (*see* **Note 10**).

8. Attach the 10× gasket. The notched cut is at the left top corner. Check that the holes are aligned with the wells. Avoid touching the smooth gasket side. Do not press on top of the gasket! Avoid wetting. Keep assembly horizontal while loading the chip. Tap the touchscreen to eject and insert tray and start the run. Proceed immediately after the completion of the run (~18 min).
9. Remove the chip from the Chromium controller, discard the gasket, and open the chip holder in a 45-degree angle position (*see* **Note 11**). Check for volume uniformity in gel bead and sample wells.
10. With a multichannel pipette transfer very slowly 100 μl uniformly opaque GEMs from row 1 (backwards triangle) into a precooled eight-tube PCR strip. Avoid air bubbles as they may compromise the GEMs. The presence of excess *clear* partitioning oil indicates a potential clog. Discard the used Next GEM chip G.
11. Perform the reverse transcription as follows (lid temperature: 53 °C, reaction volume 125 μl): 20°C, hold; 53°C, 45 min; 85°C, 5 min, 4°C, hold.

    Stopping point: At this point the samples can be stored at 4 °C for up to 72 h or at −20 °C up to a week. If samples were frozen, keep them at room temperature for 10 min before continuing. The aqueous phase will look translucent (rather than clear).

3.1.3 Post GEM Cleanup and cDNA Amplification

To avoid contamination, this section should be carried out on a separate bench dedicated to RNA work.

1. Equilibrate DynaBeads MyOne silane beads (4 °C, white lid), additive A (−20 °C), amplification master mix (−20 °C), and SC5′ feature cDNA primers (yellow lid) from 5' FBC Kit to room temperature for 30 min.
2. Thaw buffer sample cleanup 1 (−20 °C) for 10 min at 65 °C, and vigorously mix until no precipitate is visible anymore, and cool down to room temperature.
3. Apply 130 μl recovery agent dropwise on top of the post incubation GEMs, and wait 2 min until you see a clear aqueous upper phase. If biphasic separation is incomplete, mix by inverting the capped tube strip 5×, and centrifuge briefly.
4. Slowly remove 130 μl of the pink lower phase and discard. Be careful not to aspirate the clear aqueous phase containing your cDNA. A small volume of recovery reagent will remain.
5. Vortex DynaBeads MyOne silane beads, and prepare cleanup mix by mixing 5 μl nuclease-free water, 182 μl buffer sample cleanup 1 (green), 8 μl DynaBeads MyOne silane (white), and 5 μl additive A (blue-green) to a total of 200 μl.

6. Add 200 µl cleanup mix to each sample, and pipette five times. Incubate 10 min at room temperature.
7. Prepare 2.5 ml fresh 80% ethanol per sample.
8. Prepare elution solution I by mixing 98 µl EB buffer (Qiagen), 1 µl 10% Tween 20, and 1 µl additive A (blue-green lid) to a total of 100 µl for each reaction, and mix thoroughly and centrifuge.
9. After a 10 min incubation, place the tube strip into the 10× magnet (HIGH position) until the supernatant is clear (a white interface may appear between the phases). Carefully remove and discard the supernatant.
10. Add 300 µl 80% ethanol to the pellet, wait 30 s, and remove 300 µl volume.
11. Add 200 µl 80% ethanol to the pellet, wait 30 s, and remove 200 µl volume.
12. Pulse-spin tube, and place it on the 10× magnet (low position). Remove residual ethanol. Air-dry beads for 2 min.
13. Elute sample in 35 µl elution solution I. DynaBeads might be difficult to resuspend. Let beads rehydrate for 1 min.
14. Place the strip in a 10× magnet (low position). Transfer 35 µl of GEM-RT product into a new tube.

3.1.4 cDNA and Feature Barcode Amplification

1. Prepare cDNA (*see* **Note 12**) and feature barcode amplification reaction mix on ice by mixing 50 µl amplification master mix (blue lid) and 15 µl SC5′ feature cDNA primers to a total of 65 µl.
2. Add 65 µl of amplification mix to each tube containing 35 µl GEM-RT product. Mix and centrifuge.
3. Perform amplification as follows (lid temperature: 105 °C, reaction volume 100 µl).

   98°C, 45 s; 14 cycles of (98°C, 20 s; 68°C, 30 s; 72°C,1 min); 72°C, 1 min; 4°C, hold. The number of cycles depends on cell size and the number of cells recovered.

   Stopping point: At this point the samples can be stored at 4 °C for up to 72 h.

3.1.5 Feature Barcode and cDNA Fractionation by Size Selection

In this section the amplified feature barcode fraction is separated from the amplified cDNA by size selection, so that both fractions can be further processed separately. These steps should be carried out on a bench dedicated to cDNA work.

1. Add 60 µl (0.6×) of resuspended SPRIselect beads to the amplification tube. Mix well, pulse-spin the tube, and incubate for 5 min at room temperature. Place the tube on a 10× magnet (high position) to separate beads from supernatant.

2. Transfer 80 µl of the supernatant containing the feature barcode fraction into a new clean tube, and keep the remaining supernatant on the beads with the cDNA (incubate for further 5 min).
3. Add 70 µl (2×) SPRIselect beads to the feature barcode supernatant. Mix well. Pulse-spin the tube. Incubate for 5 min at room temperature.
4. Place the cDNA and feature barcode tubes onto a 10× magnet (high position). Remove supernatant (without disturbing the beads).
5. Wash the samples *twice* with 200 µl of freshly prepared 80% EtOH while the tube remains on the 10× magnet. Incubate for at least 30 s, and then remove and discard all of the supernatant.
6. After the final ethanol wash, pulse-spin the beads, and place them on the magnet (low position). Remove any residual ethanol. Air-dry beads. Do not exceed 2 min as this will lead to decreased elution efficiency.
7. Resuspend beads with 45 µl EB. Quickly spin the tube, and incubate for 2 min at room temperature.
8. Place the sample in the 10× magnet (low position), and transfer 45 µl supernatant to a fresh PCR tube.
9. Quantify the cDNA using the NGS Standard Sensitivity Kit for the Fragment Analyzer (Agilent) in a region of 400–5500 bp. Alternatively, an Agilent Bioanalyzer High-Sensitivity chip can be used.

   Stopping point: At this point the samples can be stored at 4 °C for up to 72 h or at −20 °C up to 4 weeks.

3.1.6 Library Construction

The following steps describe the preparation of three types of libraries: a feature barcode library (that will yield information on cell surface proteins (features) or hash-tags, made from the purified feature barcode fraction) (a), AIRR libraries (TR and/or IG) that require target enrichment (made from the purified cDNA fraction), and 5′ gene expression libraries (made from the purified cDNA fraction) (b, c).

Feature Barcode Library Construction by Index-PCR and Purification

1. Prepare the index-PCR by mixing 10 µl amplification master mix (blue lid), 2.5 µl SI Primer, and 2.5 µl Single-Index Kit N Set A to a total of 15 µl, for each sample. Record the assignment of the used indices.
2. Add 5 µl feature barcode sample fraction, mix, pulse-spin, and start the following PCR program (lid temperature: 105 °C): 20°C, Hold; 98°C, 45 s; 98°C, 20 s; 54°C, 30 s; 72°C, 20 s (eight cycles); 72°C, 1 min; 4°C, hold.

The following steps should be carried out on a separate, post-PCR-dedicated bench.

3. Vortex SPRIselect beads and add 24 μl (1.2×) beads to each sample. Incubate for 5 min at room temperature. Place the tube on a 10× magnet (high position). Remove supernatant.
4. Wash samples *twice* with 200 μl freshly prepared 80% EtOH, while the tube remains on the 10× magnet. Incubate for at least 30 s, and then remove and discard all of the supernatant. After the final ethanol wash, pulse-spin the beads, and place them on the 10× magnet (low position). Remove any residual EtOH. Air-dry the beads. Do not exceed 1 min to ensure maximum elution efficiency.
5. Resuspend beads with 15 μl nuclease-free water. Quickly spin the tube and incubate for 2 min at room temperature.
6. Place the sample in the 10× magnet (low position). Transfer 15 μl supernatant to a new PCR tube.
7. Measure concentration with Qubit and dilute library appropriately with TE. Check the size and concentration with the FA NGS Standard Sensitivity (region of 190–210 bp).

   Stopping point: At this point the samples can be stored at 4 °C for up to 72 h or at −20 °C for long-term storage.

Target Enrichment for AIRR Libraries

1. Prepare the enrichment reaction mix 1 on ice by mixing 5 μl amplified cDNA, 35 μl nuclease-free water, 50 μl amplification master mix (blue lid), 5 μl cDNA additive (pink lid), and 5 μl T-cell mix 1 (red lid) **or** B-cell mix 1 (blue lid). Pipet 95 μl of enrichment mix to each tube, and add 5 μl amplified cDNAμ to a total of 100 μl. Mix by pipetting and centrifuge.
2. Perform amplification as follows (lid temperature: 105 °C, reaction volume 100 μl): 98°C, 45 s; x***** cycles of (98°C, 20 s; 67°C, 30 s; 72°C, 1 min); 72°C, 1 min; 4°C, hold. *****Six cycles for IG and ten cycles for TR.

   Stopping point: At this point the samples can be stored at 4 °C for up to 72 h.

   The following steps should be carried out on a bench dedicated to amplified cDNA.
3. Purification of amplified cDNA: Add 80 μl (0.8×) resuspended SPRIselect beads to the cDNA. Mix well. Pulse-spin the tube. Incubate for 5 min at room temperature. Place the tube on a 10× magnet (high). Discard supernatant without disturbing the beads.
4. Wash the samples *twice* with 200 μl of freshly prepared 80% EtOH, while the tube remains in the magnetic rack. Incubate for at least 30 s, and then remove and discard all of the supernatant.

5. After the final ethanol wash, pulse-spin the beads, place them back on the rack (low position), and remove any residual ethanol. Air-dry the beads. Do not exceed 2 min as this will lead to decreased elution efficiency.
6. Resuspend beads with 35 μl EB. Quickly spin the tube, and incubate for 2 min at room temperature.
7. Place the sample in the 10× magnet (low). Transfer supernatant to a clean nuclease-free PCR tube.
8. Prepare enrichment reaction mix 2 on ice by mixing 5 μl nuclease-free water, 50 μl amplification master mix (blue lid), 5 μl cDNA additive (pink lid), and 5 μl T-cell mix 2 (red lid) **or** B-cell mix 2 (blue lid) to a total of 65 μl.
9. Pipet 65 μl of enrichment mix to each tube and add 35 μl amplified cDNA. Mix by pipetting and centrifuge.
10. Perform amplification as follows (lid temperature: 105 °C, reaction volume 100 μl):

    20°C, hold; 98°C, 45 s; 98°C, 20 s and 67°C, 30 s and 72°C, 1 min (six cycles); 72°C, 1 min; 4 °C, hold.

    Stopping point: At this point the samples can be stored at 4 °C for up to 72 h.
11. Perform size selection on AIRR-enriched cDNA: Vortex SPRIselect beads, and add 50 μl (0.5×) beads to each sample and incubate for 5 min at room temperature. Place the tube on a 10× magnet (high position). Keep supernatant.
12. Transfer 145 μl supernatant to a new tube strip, and add 30 μl (0.8×) SPRIselect reagent to each sample. Mix and incubate for 5 min at room temperature.
13. Place the tube on a 10× magnet (high). Discard supernatant without disturbing the beads.
14. Wash the samples *twice* with 200 μl of freshly prepared 80% EtOH, while the tube remains in the magnetic rack. Incubate for at least 30 s, and then remove and discard all of the supernatant.
15. After the final EtOH wash, pulse-spin the beads, and place them on the 10× magnet (low). Remove any residual ethanol. Air-dry beads. Do not exceed 1 min to ensure maximum elution efficiency.
16. Resuspend beads with 22 μl nuclease-free water (NFW). Quickly spin the tube and incubate for 2 min at room temperature.
17. Place the sample in the 10× magnet (low). Transfer 22 μl supernatant to a new PCR tube.

18. Quantify samples using the FA NGS Standard Sensitivity in a region of 200–5500 bp. Alternatively, an Agilent Bioanalyzer High-Sensitivity chip can be used.

    Stopping point: At this point the samples can be stored at −20 °C for up to 1 week (or 4 °C for up to 72 h).

5′ Gene Expression and AIRR Library Construction: Fragmentation, Adaptor Ligation, and Library Amplification (*See* **Note 13**)

1. Equilibrate the fragmentation buffer to room temperature for 30 min, and verify that there is no precipitate; otherwise vortex (the same should be done later for the ligation buffer, adaptor mix, and SI-PCR primer as well). Keep fragmentation enzyme blend (and later DNA ligase, amplification master mix (−20 °C)) on ice.
2. Prepare your cDNA samples: You will need 50 ng (maximally) in 20 µl in PCR tubes. Keep on ice. Prepare PCR tubes with 20 µl AIRR fraction.
3. Vortex the fragmentation buffer (check if precipitate is still visible), and prepare fragmentation mix (the volume required for both cDNA and AIRR fraction fragmentation) by mixing 20 µl nuclease-free water, 5 µl fragmentation buffer (white lid), and 10 µl fragmentation enzyme blend (purple lid) to a total of 30 µl.
4. Start cDNA fragmentation: Dispense 30 µl of the fragmentation mix into the PCR tubes containing the 50 ng cDNA.
5. Incubate the samples under the following conditions (lid temperature: 65 °C, reaction volume 50 µl): pre-cooling block (4 °C, Hold); fragmentation (32 °C, 5 min); end Repair and A-tailing (65 °C, 30 min); 4 °C, hold.
6. Once cDNA fragmentation has started, add enriched AIRR fraction to the corresponding fragmentation mix tubes. Immediately after cDNA fragmentation, proceed to size selection!
7. cDNA size selection: Vortex SPRIselect beads, and add 30 µl beads to each sample (0.6×). Incubate for 5 min at room temperature. Place the tube on a magnetic rack (high position). Keep supernatant!
8. Meanwhile, add enriched AIRR fraction to the corresponding fragmentation mix tubes, and start AIRR fragmentation: Incubate the samples under the following conditions (lid temperature: 65 °C, reaction volume 50 µl): pre-cooling block (4 °C, hold); fragmentation (32 °C, 2 min); end repair and A-tailing (65 °C, 30 min); 4 °C, hold.
9. Continue cDNA purification: Place the tube on a magnetic rack (high position). Keep supernatant.
10. Transfer 75 µl cDNA supernatant to a new tube. Add 10 µl SPRIselect beads (0.8×) to each sample and mix. Incubate for 5 min at room temperature. Place the tube on a 10× magnet (high). Discard 80 µl supernatant.

11. Wash the beads *twice* with 150 µl of freshly prepared 80% EtOH, while the tube remains in the magnetic rack. Incubate for at least 30 s, and then remove and discard all of the supernatant.
12. After the final EtOH wash, pulse-spin the beads, place them on the 10× magnet (low). Remove any residual ethanol. Air-dry beads. Do not exceed 1 min to ensure maximum elution efficiency.
13. Resuspend beads with 50.5 µl EB. Quickly spin the tube, and incubate for 2 min at room temperature.
14. Place the sample on the 10× magnet (low). Transfer 50 µl supernatant to a new PCR tube.
15. Prepare adapter ligation mix for cDNA and AIRR fraction by mixing 20 µl ligation buffer (green lid), 17.5 µl nuclease-free water, 10 µl DNA ligase (yellow lid), and 2.5 µl adaptor mix (blue-green lid) to a total of 50 µl (*see* **Note 14**).
16. Add 50 µl ligation mix to each sample, mix, and briefly spin down the sample. Incubate as follows (lid temperature: 30 °C, reaction volume 100 µl): 20°C, 15 min; 4°C, hold.
17. Proceed immediately to the cleanup: Add 80 µl (0.8×) of resuspended SPRIselect beads to the sample. Mix well. Pulse-spin the tube. Incubate for 5 min at room temperature. Place the tube on a 10× magnet (high). Discard supernatant.
18. Wash the samples *twice* with 200 µl of freshly prepared 80% EtOH, while the tube remains in the magnetic rack. Incubate for at least 30 s, and then remove and discard all of the supernatant.
19. After the final ethanol wash, pulse-spin the beads, and place them on the 10× magnet (low). Remove any residual ethanol. Air-dry beads. Do not exceed 2 min as this will lead to decreased elution efficiency.
20. Resuspend beads with 30.5 µl EB. Quickly spin the tube and incubate for 2 min at room temperature.
21. Place the sample on the 10× magnet (low). Transfer 30 µl supernatant to a new PCR tube.
22. Prepare the index-PCR by mixing 50 µl amplification master mix (blue lid), 8 µl nuclease-free water, and 2 µl SI-PCR primer (purple lid) to a total of 60 µl.
23. Add 60 µl PCR mix and 10 µl of Single-Index Kit T Set A (PN 1000213) primer (12.5 µM) to each 30 µl sample, and record the assignment of the used indices.
24. Mix, briefly spin down the sample, and start the following PCR program (lid temperature: 105 °C): 20°C, hold; 98°C, 45 s; 98°C, 20 s and 54°C, 30 s and 72°C, 20 s (x* cycles); 72°C,

1 min; 4°C, hold. *Cycles: cDNA: adjust to your input: 1–25 ng cDNA: 14–16 cycles, 26–50 ng cDNA: 10–14 cycles; AIRR: 8 cycles.

Stopping point: At this point the samples can be stored at 4 °C for up to 72 h.

25. Perform size selection on final cDNA libraries and AIRR libraries, followed by a 0.8 x purification using SPRIselect beads: Vortex SPRIselect beads, add 60 μl (0.6×) beads to each cDNA sample, and incubate for 5 min at room temperature. Place the tube on a 10× magnet (high position). Keep supernatant.
26. Vortex SPRIselect beads, add 55 μl (0.55×) beads to each BCR/TCR sample, and incubate for 5 min at room temperature. Place the tube on a 10× magnet (high position). Keep supernatant.
27. Transfer 150 μl of the cDNA supernatant to a new tube, and add 20 μl (0.8×) SPRIselect reagent to each sample. Mix and incubate for 5 min at room temperature.
28. Transfer 150 μl of the IG/TR supernatant to a new tube, and add 20 μl (0.75×) SPRIselect reagent to each sample. Mix and incubate for 5 min at room temperature.
29. Place all tubes (cDNA and IG/TR) on a 10×magnet (high). Discard 165 μl supernatant without disturbing the beads.
30. Wash the samples *twice* with 200 μl of freshly prepared 80% EtOH, while the tube remains in the magnetic rack. Incubate for at least 30 s, and then remove and discard all of the supernatant.
31. After the final ethanol wash, pulse-spin the beads, and place them back on the rack (low). Remove any residual ethanol. Air-dry beads not longer than 1 min to ensure maximum elution efficiency.
32. Resuspend beads with 50 μl NFW. Quickly spin the tube, and incubate for 2 min at room temperature.
33. Vortex SPRIselect beads. Add 40 μl (0.8×) beads to each sample. Incubate for 5 min at room temperature. Place the tube on a 10× magnet (high). Discard 85 μl supernatant (without beads).
34. Wash the samples *twice* with 200 μl of freshly prepared 80% EtOH, while the tube remains in the magnetic rack. Incubate for at least 30 s, and then remove and discard all of the supernatant.
35. After the final ethanol wash, pulse-spin the beads, and place them back on the rack (low). Remove any residual ethanol. Air-dry beads not longer than 1 min to ensure maximum elution efficiency.

36. Resuspend beads with 15 µl NFW. Quickly spin the tube and incubate for 2 min at room temperature.
37. Place the sample in the 10× magnet (low). Transfer 15 µl of supernatant to a clean PCR tube.
38. Quantify libraries with Qubit and with the correct dilution load the Fragment Analyzer using the NGS High-Sensitivity Kit. Calculate concentration in a region of 200–700 bp (Fig. 6).

3.1.7 Sequencing

Prepare libraries for Illumina sequencing (*see* **Note 15**).

3.2 Single-Cell SMART-Seq

Due to the sensitivity of these protocols, cells should be collected under clean-room conditions to avoid contamination. The whole process of cDNA synthesis should be carried out in a PCR clean workstation under clean-room conditions.

3.2.1 Cell Sorting and cDNA Synthesis

Buffer Preparations

1. At room temperature, thaw the SMART-Seq sc First-Strand Buffer. On ice, thaw all the remaining reagents (except the enzyme) needed for first-strand cDNA synthesis: 10× lysis buffer, RNase inhibitor, nuclease-free water, SMART-Seq sc TSO (*see* **Note 16**), and 3′ SMART-Seq CDS Primer II A. Gently vortex each reagent to mix and spin down briefly. Store all reagents on ice except the SMART-Seq sc First-Strand Buffer (*see* **Note 17**).
2. Assemble the plain sorting solution (PSS; without 3′ SMART-Seq CDS Primer II A) by mixing 104.5 µl 10× lysis buffer (*see* **Note 18**), 5.5 µl RNase inhibitor, and 1155 µl nuclease-free water to a total of 1265 µl for 96 wells, volume includes ~10% extra for overage (*see* **Note 19**).
3. Mix briefly, and then spin down.
4. Aliquot 11.5 µl of PSS into the appropriate number of wells of PCR tube strips or a 96-well plate (*see* **Note 20**).
5. Seal the plate/tube strips, and briefly spin to ensure the PSS collects at the bottom of the wells (*see* **Note 21**).
6. Store the plate/tube strips at −20 °C for 10 min at a minimum and for up to 24 h. As the volume of PSS is small, the tubes/plate should be kept at −20 °C until just before sorting.

Cell Sorting

1. When ready to sort, unseal the prepared plate/tube strips, and sort cells into the sorting solution according to the FACS system manual and desired parameters.
2. Seal the plate/tube strips with an aluminum foil seal or PCR strip caps. Ensure the plate/tube strips are sealed firmly to minimize any evaporation.
3. Immediately after sorting the cells and sealing the plate, spin briefly to collect the cells at the bottom of each well in the PSS.
4. Place the plate on dry ice to flash-freeze the sorted cells (*see* **Note 22**).

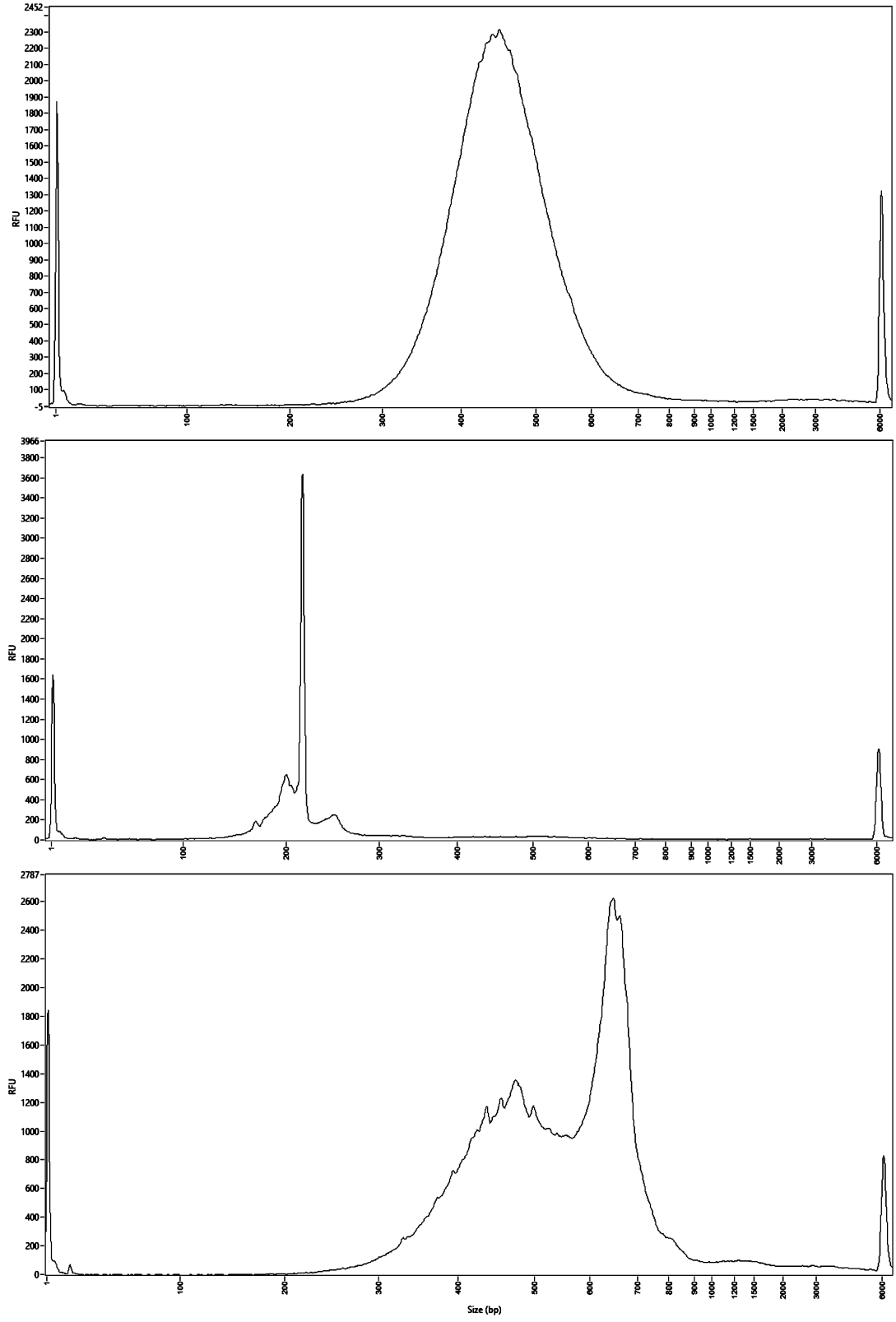


Fig. 6 Fragment analysis of the final libraries. Top: example of the distribution of a cDNA library. Middle: example of the distribution of a feature barcode library. Bottom: example of the distribution of an AIRR library

Preparing Controls

See below for guidelines on setting up your positive and negative controls alongside your cell samples.

1. Prepare each control in 8-well strips or a 96-well plate by mixing 9.5 μl nuclease-free water, 2 μl control sample (cells or RNA [*see* **Note 23**]), and 1 μl 3′ SMART-Seq CDS Primer II A to a total volume of 12.5 μl.

cDNA Synthesis

1. When ready to start, remove the samples (plate or PCR strips containing the sorted cells) from the freezer and briefly spin to collect the contents at the bottom of the tubes.
2. Place the samples on ice, and add any necessary remaining reagents, including 1 μl of 3′ SMART-Seq CDS Primer II A (*see* **Note 24**). Mix well by gently vortexing, and then spin the tube(s) briefly to collect the contents at the bottom of the tube (*see* **Note 25**).
3. Immediately incubate the tubes at 72 °C in a preheated, hot-lid thermal cycler for 3 min.
4. Prepare RT master mix while the samples are incubating. Prepare enough for all the reactions, plus 10% of the total reaction mix volume, by mixing at room temperature 422.4 μl SMART-Seq sc First-Strand Buffer, 105.6 μl SMART-Seq sc TSO, 52.8 μl RNase inhibitor (40 U/μl), and 211.2 μl SMART-Scribe II reverse transcriptase to a total volume of 792 μl for 96 wells (Includes 10% overage). Add the SMARTScribe II reverse transcriptase just prior to use (in **step** 7 part (b) of this section).
5. Immediately after the 3-min incubation at 72 °C, place the samples on ice for at least 2 min (but no more than 10 min).
6. Preheat the thermal cycler to 42 °C.
7. Add the SMARTScribe II reverse transcriptase to the RT master mix. Mix well by gently vortexing, and then spin the tube briefly in a mini-centrifuge to collect the contents at the bottom of the tube.
8. Add 7.5 μl of the RT master mix to each sample. Mix the contents of the tubes by gently vortexing, and spin briefly to collect the contents at the bottom of the tubes.
9. Place the tubes in a thermal cycler with a heated lid, preheated to 42 °C. Run the following program: 42°C, 180 min; 70°C, 10 s; 4°C, hold.

   Stopping point: The tubes can be stored at 4°C overnight.

3.2.2 cDNA Amplification by LD PCR

The PCR primers amplify the cDNA by priming to the sequences introduced by the 3′ SMART-Seq CDS Primer II A and the SMART-Seq sc TSO.

1. Thaw SeqAmp CB PCR buffer and PCR primer on ice. Do not thaw SeqAmp DNA polymerase. Gently vortex each reagent tube to mix and spin down briefly. Store on ice.
2. Prepare enough PCR master mix for all the reactions by mixing 2640 µl SeqAmp CB PCR buffer (2×), 105.6 µl PCR primer, 105.6 µl SeqAmp DNA polymerase, and 316.8 µl nuclease-free water to a total volume of 3168 µl for 96 wells (including 10% overage).

   Remove the SeqAmp DNA polymerase from the freezer, gently mix the tube without vortexing, and add to the master mix just before use. Mix the master mix well by vortexing gently, and spin the tube briefly to collect the contents at the bottom of the tube.
3. Add 30 µl of PCR master mix to each tube containing 20 µl of the first-strand cDNA product. Mix well by gently vortexing, and briefly spin to collect the contents at the bottom of the tube. Transfer the samples from the PCR clean work station to the general lab. All downstream processes should be performed in the general lab.
4. Place the tubes in a preheated thermal cycler with a heated lid, and run with the following conditions: 95 °C, 1 min; 98 °C,10 s; 65 °C, 30 s and 68 °C, 3 min (X***** cycles); 72 °C, 10 min; 4 °C, hold. PCR cycle number guidelines depend on the cell type (*****_see_ **Note 26**).

   Stopping point: The tubes may be stored at 4 °C overnight.

3.2.3 Purification of Amplified cDNA

1. If purification is performed directly in the PCR tubes or strips using the Takara Bio SMARTer-Seq Magnetic Separator-PCR Strip, add 40 µl of beads (_see_ **Note 27**) to each sample (_see_ **Note 28**). Mix thoroughly by vortexing for 3–5 s or pipetting the entire volume up and down at least ten times. Proceed to **step 3** of this section.
2. If you are performing purification with the Thermo Fisher Magnetic Stand-96 (recommended if processing 48–96 samples), cDNA samples need to be transferred to a 96-well V-bottom plate. Distribute 40 µl of beads (_see_ **Note 28**) to each well of the 96-well V-bottom plate, and then use a multichannel pipette to transfer the cDNA. Pipette the entire volume up and down at least ten times to mix thoroughly. Proceed to **step 3** of this section.
3. Incubate the beads-cDNA mixture at room temperature for 8 min to let the cDNA bind to the beads.
4. Briefly spin the samples to collect the liquid from the side of the tubes or plate wells (centrifugation is generally not necessary if using a 96-well V-bottom plate).

5. Place the samples on the magnetic separation device for ~5 min or longer, until the liquid appears completely clear and there are no beads left in the supernatant.
6. While the samples are on the magnetic separation device, remove and discard the supernatant. Take care not to disturb the beads.
7. Keep the samples on the magnetic separation device. Add 200 μl of freshly made 80% ethanol to each sample without disturbing the beads (*see* **Note 27**). Incubate for 30 s. Then, carefully remove and discard the supernatant, taking care not to disturb the beads. The cDNA remains bound to the beads during the washing process.
8. Repeat the ethanol wash (**step 7** of this section) once more.
9. Briefly centrifuge the samples to collect the liquid from the side of the tubes or plate wells. Place the samples on the magnetic separation device for 30 s, and then remove any residual ethanol with a pipette.
10. Incubate the samples at room temperature for ~2–2.5 min, until the pellet is no longer shiny, but before cracks appear (*see* **Note 29**).
11. Once the beads are dry, remove the samples from the magnetic separation device, and add 17 μl of elution buffer to cover the bead pellet. Mix thoroughly by pipetting or gently vortexing to resuspend the beads.
12. Incubate at room temperature for at least 2 min to rehydrate.
13. Briefly spin the samples to collect the liquid from the side of the tubes or plate wells. Place the samples back on the magnetic separation device for 1 min or longer, until the solution is completely clear.
14. Transfer clear supernatant (~15 μl) containing purified cDNA from each tube/well to a new tube/plate. Do not pool samples at this point. Take care not to carry over any beads with your sample.
15. Proceed to validation immediately or store at −20 °C.

3.2.4 Validation Using the Agilent 2100 Bioanalyzer

1. Aliquot 1 μl of the amplified cDNA for validation using the Agilent 2100 Bioanalyzer and Agilent's High-Sensitivity DNA Kit. See the Agilent High-Sensitivity DNA Kit User Manual for instructions.
2. Compare the results for your samples and controls to verify whether the sample is suitable for further processing. Successful cDNA synthesis and amplification should yield no product in the negative control, and a distinct peak spanning 400 bp to 10,000 bp, peaked at ~2500 bp for the positive control RNA sample, yielding approximately 3.4–17 ng of cDNA (depending on the input type and amount).

3.2.5 Library Preparation for Next-Generation Sequencing

The following sections describe a modified Illumina Nextera XT DNA library preparation protocol that has been fully validated to work with the SMART-Seq Single-Cell Kit. The reaction size has been reduced to a quarter volume of what is recommended by Illumina.

Dilute and Prepare cDNA for Tagmentation

1. Dilute each cDNA to 100 pg/μl with nuclease-free water in a plate or PCR strips (*see* **Note 30**). Do not pool at this step. Vortex at medium speed for 20 s and centrifuge at 350 × *g* for 1 min.
2. Warm Tagment DNA Buffer and NT Buffer to room temperature. Visually inspect NT Buffer to ensure that there is no precipitate. If there is a precipitate, vortex the buffer until all particles are resuspended.
3. After thawing, gently invert the tubes 3–5 times, followed by centrifuging the tubes briefly, to ensure all reagents are adequately mixed.
4. Label a new 96-well PCR plate "Library Prep."
5. In a 1.5-ml PCR tube, prepare tagmentation premix by mixing 300 μl Tagment DNA Buffer and 150 μl amplification tagment mix to a total volume of 450 μl (calculated based on a 25% excess). Vortex gently for 20 s and centrifuge the tube briefly.
6. Distribute 3.75 μl of the tagmentation premix into each well of the "Library Prep" plate.
7. Transfer 1.25 μl of each diluted cDNA sample to the "Library Prep" plate.
8. Seal the plate and vortex at medium speed for 20 s. Centrifuge at 2000 × *g* for 5 min to remove bubbles.
9. Place the "Library Prep" plate in a thermal cycler with a heated lid, and run the following program: 55 °C, 10 min; 10 °C, hold.
10. Once the thermal cycler reaches 10 °C, pipette 1.25 μl of NT Buffer into each of the tagmented samples to neutralize the samples (*see* **Note 31**).
11. Seal the plate and vortex at medium speed, and then centrifuge at 2000 × *g* for 1 min.
12. Incubate at room temperature for 5 min.

Amplify the Tagmented cDNA

1. Pipette 3.75 μl of Nextera PCR Master Mix (NPM) into each well of the "Library Prep" plate using an eight-channel pipette (*see* **Note 32**).

   Select appropriate Index 1 (N7xx) and Index 2 (S5xx) primers for the number of samples in your experiment (*see* **Note 33**).

2. Pipette 1.25 μl of Index 1 Primers (N7xx) into the corresponding wells of each row of the "Library Prep" plate. As a result, each of the 12 wells in row "A" will contain different Index 1 Primers.
3. Pipette 1.25 μl of Index 2 Primers (S5xx) into the corresponding wells of each column of the "Library Prep" plate. As a result, each of the 8 wells in column "1" will contain different Index 2 Primers.
4. Seal the plate with adhesive film and vortex at medium speed for 20 s. Centrifuge at 2000 × *g* for 2 min. Place the "Library Prep" plate into a thermal cycler, and perform PCR amplification using the following program: 72°C, 3 min; 95°C, 30 s; 95°C, 10 s; 55°C, 30 s; and 72°C, 60 s (12 cycles); 72°C, 5 min; 10°C, hold.

   Samples can be left overnight in the thermal cycler at 4 °C. If not processed within the next day, freeze the PCR products at −20 °C.

Pooling and Purification of Amplified Libraries

1. Determine the number of libraries to be pooled based on the desired sequencing depth and sequencer throughput. If preferred, clean up libraries individually before pooling (*see* **Note 34**).
2. Pool the libraries by pipetting a fixed volume from each sample into a 1.5-ml tube or PCR tube. Volumes between 2 μl and 8 μl are appropriate. Do not use less than 2 μl per sample to ensure greater accuracy (e.g., to pool 96 libraries, add 2 μl of each library (total 192 μl) and 154 μl of bead volume to a 1.5 ml tube. The bead volume is approximately 80% of the total pool volume).
3. Add a volume of beads representing 80% of the volume of the pooled libraries. If cleaning up libraries individually, add 40 μl of beads to each 50-μl sample.
4. Mix well by vortexing or pipetting the entire mixture up and down ten times (*see* **Note 35**).
5. Incubate at room temperature for 5 min to let the cDNA libraries bind to the beads.
6. Briefly spin the sample to collect the liquid from the side of the tube. Place the tube on a magnetic stand for ~2 min or until the liquid appears completely clear, and there are no beads left in the supernatant.
7. While the samples are on the magnetic separation device, remove and discard the supernatant. Take care not to disturb the beads.
8. Keep the samples on the magnetic separation device. Add 200 μl of fresh 80% ethanol to each sample without disturbing

the beads. Incubate for 30 s, and then remove and discard the supernatant, taking care not to disturb the beads. The cDNA remains bound to the beads during washing.

9. Repeat the ethanol wash (**step 8** of part c) of this section) once more.
10. Briefly centrifuge the samples to collect the liquid from the side of the tube or plate well. Place the samples on the magnetic separation device for 30 s, and then remove any residual ethanol with a pipette.
11. Incubate the samples at room temperature for ~5–15 min, until the pellet is no longer shiny, but before cracks appear (*see* **Note 29**). The pooled samples requiring higher bead volumes take longer to dry.
12. Once the beads are dry, elute the pooled, purified libraries by adding the required volume of nuclease-free water (provided), based on the number of samples pooled (*see* **Note 36**).
13. Remove from the magnetic separation device, and vortex the tube for 3 s to mix thoroughly. Incubate at room temperature for ~5 min to rehydrate the beads.
14. Briefly spin to collect the liquid from the side of the tube. Place the tube back on the magnetic separation device for ~2 min or longer until the solution is completely clear.
15. Transfer the clear supernatant containing purified libraries to a nuclease-free, low-adhesion tube. Label each tube with sample information. The purified libraries can be stored at −20 °C.

3.2.6 Sequencing

Sequence the SMART-Seq single-cell library with Illumina sequencing (*see* **Note 37**).

3.3 10x Genomics Chromium Next GEM Single-Cell V(D)J Kit Data Processing and Analysis

10X Genomics data is analyzed using the Cell Ranger software, provided by 10x Genomics free of charge. Cell Ranger allows (1) sequencing raw data demultiplexing, (2) quality control of the raw data obtained, (3) raw data alignment to the reference genome, and (4) data matrix preparation for further in-depth analyses using dimensionality reduction methods. Three Cell Ranger pipelines are now available: `cellranger count` (for transcriptome and feature data), `cellranger vdj` (for AIRR data) and `cellranger multi` (which does the integrative analysis of transcriptome, AIRR, and feature data). Loupe Brower and Loupe VDJ Browser, which are also available via the 10x Genomics website, provide a complementary set of analysis tools. Note that the Loupe Browsers are only available for Windows or macOS environments.

3.3.1 Setup

Cell Ranger can be installed in a folder named "cellranger" in the home directory. Before running Cell Ranger, ensure that this folder is included in the PATH environment variable: `export PATH = $PATH:$HOME/cellranger.`

3.3.2 Demultiplexing

Libraries can be demultiplexed using the following command (*see* **Note 38**): `cellranger mkfastq --id=LBA-01 --run=data/BCL --samplesheet=cellranger-tiny-bcl-sample-sheet-1.2.0.csv --output-dir=data/VDJ.`

Arguments:

--id: The folder name that will be created by cellranger mkfastq (here: LBA-01).

--run: Path of Illumina BCL Sample run folder.

--output: Path to where your folder will be located (here: data/VDJ) (optional).

--samplesheet: Path to an Illumina Experiment Manager (IEM) sample sheet format which contains all the information needed for describing samples.

--csv: Path to a csv file which contains some information for describing samples. It is an alternative if your Illumina Experiment Manager is not provided.

3.3.3 Alignment

V(D)J genes, gene transcripts, and feature barcodes can be analyzed by running either "cellranger vdj" (a) or "cellranger count" (b) using the following commands (*see* **Note 39**):

1. `cellranger vdj --id LBA-01-VDJ --sample LBA-01-Sample --fastqs data/VDJ --reference data/refdata/refdata-cellranger-vdj-GRCh38-alts-ensembl-5.0.0.`

   Arguments (*see* **Note 40**):

   (a) --id: The folder that will contain the output of the pipeline (here: LBA-01-VDJ)

   (b) --sample: Sample name as specified in the FASTQ file (here: LBA-01-Sample)

   (c) --fastqs: Path to where your FASTQ file is located (here: data/VDJ)

   (d) --reference: Path to a Cell Ranger compatible VDJ reference (*see* **Notes 41** and **42**).

2. `cellranger count --id LBA-01-GEX --transcriptome data/refdata/refdata-gex-GRCh38-2020-A --fastqs data/GEX --sample LBA-01-GEX-Sample`

   Arguments:

   (a) --id: The folder that will contain the output of the pipeline (here: LBA-01-GEX)

   (b) --sample: Sample name as specified in the FASTQ file (here: LBA-01-GEX-Sample)

(c) --fastqs: Path to where your FASTQ file is located (here: data/GEX)

(d) --transcriptome: Path to a Cell Ranger compatible GEX reference (*see* **Notes 41** and **42**).

3.3.4 Quality Control

Visualize the HTML report file generated by Cell Ranger on your computer. Ensure that there are no warnings regarding the RNA content, the number of gene/VDJ count, the number of cells detected, or the average number of genes per cell. When feature barcoding is used, ensure the absence of aggregates, of unbound antibodies as this may dramatically reduce the cells to be analyzed. In case you encounter major warnings, please reconsider your protocol (antibody dilution, washing steps, etc.). This step is critical to ensure the quality of the data that will be further analyzed.

3.3.5 Exploratory Analysis

Loupe Browser (version 5.0.1) software can be employed for a first analysis using the vloupe.vloupe file to explore the aligned results using classical single-cell approaches and determine whether expected variables have been detected (and therefore sequenced) (*see* **Note 43**).

3.3.6 In-Depth Analysis

Further analysis can be performed using additional tools (*see* **Note 44**).

3.4 Single-Cell SMART-Seq Data Processing and Analysis

The steps below outline how to use TraCeR [6] as a Docker container using the test data in the TraCeR GitHub repository. The TraCeR pipeline consists of two main steps, which are run separately: assemble and summarise (*see* **Note 45**).

3.4.1 Raw Data Processing

1. Consider whether your data corresponds to the demands for TraCeR (*see* **Note 46**).
2. Check raw data files (*see* **Note 47**).
3. Perform standard quality control, and trim sequencing reads for bad quality reads/nucleotides and adapter sequences by using tools such as Trim Galore (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore) and Cutadapt [7].

3.4.2 Obtain the Software

1. Download or clone the TraCeR GitHub repository by running: `git clone` https://github.com/Teichlab/tracer.
2. Pull the Docker container from DockerHub by running: `docker pull teichlab/tracer`.
3. Increase the Docker memory limit to 6–8 GB to avoid TraCeR running out of memory during the *assemble* step. Details on how to increase the Docker memory limit are found at https://docs.docker.com/docker-for-windows/#advanced for Windows and https://docs.docker.com/docker-for-mac/#advanced for Mac.

4. Run a TraCeR test to verify that the installation of TraCeR and all its dependencies runs smoothly and gives the expected outputs. To do this, enter the main directory of the downloaded GitHub TraCeR repository, and run `docker run -it --rm -v $PWD:/scratch -w /scratch teichlab/tracer test -o test_data`.
5. Inspect output of the TraCeR test by comparing the output in *test_data/results/filtered_TCR_summary* with the expected results in *test_data/expected_summary*. The test output should consist of three cells, of which Cell 1 and Cell 2 are clonally related. Each cell should have a productive TRA, a nonproductive TRA, a productive TRB, and a nonproductive TRB rearrangement.

3.4.3 Reconstruct TR Sequences with the Assemble Mode

1. Run TraCeR following commands followed by any appropriate arguments, from the directory containing the input data, to ensure that TraCeR runs on Docker (*see* **Note 48**):

```
docker run -it --rm -v $PWD:/scratch -w /scratch teichlab/tracer assemble [options] <file_1> [<file_2>] <cell_name> <output_directory>
```

The main arguments expected by TraCeR are as follows:

<*file_1*>: FASTQ file containing #1 mates from paired-end sequencing or all reads from single-end sequencing. If paired-end sequencing is used, provide #2 mates after the #1 mates FASTQ file.

<*cell_name*>: Name of the cell chosen by the user. This name will be used in file names and labels.

<*output_directory*>: Directory for output. The cell-specific output from the *assemble* mode will be found in <output_directory>/<cell_name>.

TraCeR also accepts several options, which are detailed at https://github.com/Teichlab/tracer and in [8].

2. Reconstruct TRA, TRB, TRD, and TRG rearrangements from paired-end data using one processor core by running the following command (here for a hypothetical example dataset consisting of T cells from humans):

```
docker run -it --rm -v $PWD:/scratch -w /scratch teichlab/tracer assemble cell_1_R1.fq.gz cell_1_R2.fq.gz cell_1 Exp_1 -c my_config_file -s Hsap --loci A B G D -m assembly
```

3.4.4 Identify Clonally Related Cells with the Summarise Mode

1. Remove low-quality cells, defined by standard quality control pipelines for single-cell RNA-seq data, to obtain the most accurate results, before running the TraCeR *summarise* mode.
2. Define cell populations to analyze. The output of the *assemble* mode of TraCeR is a directory for each cell. Before running the *summarise* mode, create a new directory for cells you want to analyze together, and move the relevant TraCeR result directories for these cells into the new directory. If cells from multiple donors are present in the dataset, run the summarise mode separately for each donor in order to define true clonally related cells.
3. Run TraCeR in *summarise* mode on Docker with the following command, where <*input_dir*> is the path to the directory containing the output of TraCeR *assemble* mode for all the cells to be summarized together:

```
docker run -it --rm -v $PWD:/scratch -w /scratch teichlab/
tracer summarise [options] <input_dir>
```

For a hypothetical example dataset consisting of T cells from humans, the following command could be run using one processor core:

```
docker run -it --rm -v $PWD:/scratch -w /scratch teichlab/
tracer summarise Exp_1 -c my_config_file -s Hsap --loci A B G D
-g svg -u
```

3.4.5 Quality Control

1. Create a new directory for filtering out the *assemble* result directories for suspicious cells.
2. Remove likely cell doublets/multiplets affecting clonal assignments with more or less strict criteria depending on the dataset and biological questions (*see* **Note 49**) by visually inspecting the clonal graph outputs created by TraCeR *summarise* run with the `-u` flag. Likely doublets/multiplets can be seen as cells with two or more sets of rearranged TRA and TRB chains (or TRD and TRG), connecting smaller clone groups that otherwise do not share rearranged sequences with each other. If such likely doublet/multiplets exist in the data, we recommend to remove the result directories from *assemble* for these cells and rerun TraCeR *summarise* mode without the `-u` flag.
3. Remove likely cell doublets/multiplets/contaminations by opening `TCR_summary.txt` in the unfiltered summary folder and looking at the section named `#Cells with more than two recombinants for a locus`. Move the *assemble* result folder for any cell with more than three reconstructed TR

rearrangements for any locus to the directory containing cells to be filtered out.

4. Remove potential cell doublets/contaminations by opening `<cell_name>/unfiltered_TCR_seqs/unfiltered_TCRs.txt` for each cell with three reconstructed TR rearrangements for a locus. Discard a cell if all reconstructed sequences for the locus in question are substantially different from each other and have nonzero expression values. If two or more rearrangements use the same gene segments, they have probably not been collapsed to one sequence due to PCR errors or misassemblies, and the sequence with the highest expression value is likely the true rearranged sequence. In such cases, the cells need not be filtered out.
5. Consider removing other possible cell doublets: While the expression of two different TRA rearrangements is not uncommon in T cells, it is less common for TRB rearrangements [9]. Thus, cells expressing two TRA and two TRB rearrangements may be cell doublets and can be filtered out depending on the desired balance of discarding false cell doublets versus keeping potential cell doublets.

3.4.6 Interpreting TraCeR Summarise Output

The output of the *summarise* step is written to `<input_dir>/filtered_TCR_summary` or `<input_dir>/unfiltered_TCR_summary` depending on whether the `-u` flag was used. The most useful output files of `summarise` are as follows:

`TCR_summary.txt`: Summary statistics for TR reconstruction and a list of clonally related cells.

`recombinants.txt`: List of TR identifiers, lengths, productivities, and CDR3 sequences reconstructed for each cell.

`clonotype_sizes.pdf` and `clonotype_sizes.txt`: Distribution of clone sizes as bar graph and text file.

`clonotype_network_[with|without]_identifiers.<graph_format>`: Graphical representation of TR rearrangements with full identifiers or just lines indicating presence/absence of rearrangements of the different loci (Fig. 7). The graph without identifiers gives an overview of the degree of clonality in the cell population (A). The graph with identifiers only shows clonally expanded cells, with details on the shared TR rearrangements within each clone (B). Edges between nodes represent cells that share one or more reconstructed TR rearrangements, colored by the TR locus (*see* **Note 50**).

3.4.7 In-Depth Analysis

Further analysis can be performed using additional tools (*see* **Note 44**).

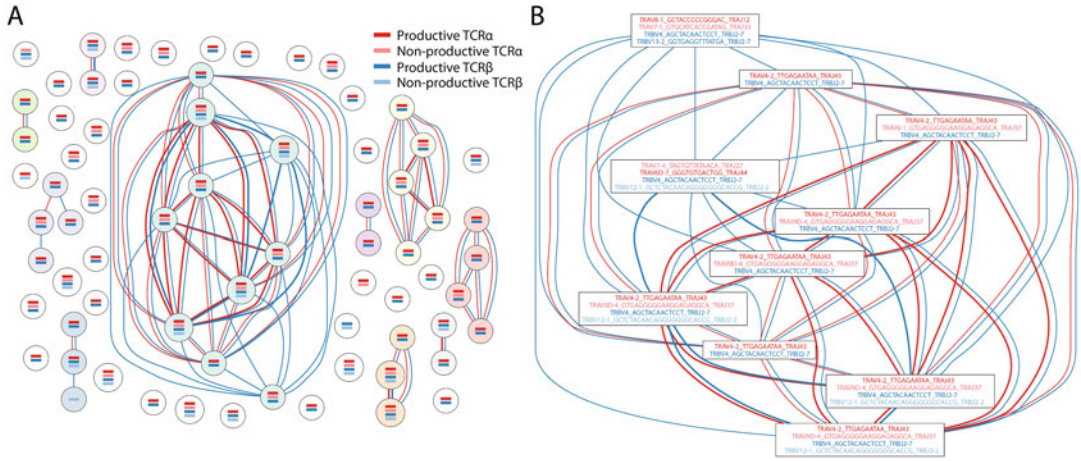


Fig. 7 Overview of clonality of a T-cell population based on TR sequences reconstructed by TraCeR. Each cell is represented by a node, while each reconstructed TR sequence is represented by a horizontal line (**a**) or a sequence identifier (**b**; only showing the largest clone group), colored according to chain type. Cells sharing identical TR sequences are connected with edges colored by chain type

4 Notes

1. The thermal cycler should always be used with the heated lid option turned on. If prompted to input a specific temperature, use 105 °C. Most thermal cyclers with heated lids will automatically adjust the lid temperature just above the highest block temperature within a cycling program. However, if the thermal cycler does make this automatic adjustment, one may want to follow the manufacturer's instructions to choose a lower lid temperature for the reverse transcription step.
2. The kit has been specifically validated with the beads listed above. Any substitutions may lead to unexpected results. Beads need to come to room temperature before the container is opened. We strongly recommend aliquoting the beads into 1.5-ml tubes upon receipt and then refrigerating the aliquots. Individual tubes can be removed for each experiment, allowing them to come to room temperature more quickly (~30 min). Aliquoting is also instrumental in decreasing the chances of bead contamination. Immediately before use, vortex the beads until they are well dispersed. The color of the liquid should appear homogeneous. Confirm that there is no remaining pellet of beads at the bottom of the tube. Mix well to disperse before adding the beads to your reactions. The beads are viscous, so pipette them slowly.
3. Please consult the Chromium Next GEM Single-Cell V(D)J Kit manual (v1.1 or 2) from the 10x Genomics website (https://support.10xgenomics.com/single-cell-vdj/overview/doc/user-guide-chromium-single-cell-vdj-reagent-kits-

v11-chemistry) carefully, and follow all instructions regarding general reagent handling, Chromium Next GEM Chip handling, assembly, loading, and all other technical instructions. The evolving kit versions (v1., v1.1, v2) differ with respect to certain volumes and concentrations. The following protocol is based on v1.1.

4. Multiplexing samples (hash-tagging): multiple small cell samples can be multiplexed into one large sample. This is accomplished by labelling an ubiquitously expressed cell surface protein with an antibody conjugated to a feature barcode oligonucleotide, followed by the direct capture of the feature barcode by the gel bead primer and amplification of the feature barcode (such "hashtag" antibodies are provided by BioLegend (https://www.biolegend.com/totalseq)). A lipid anchored hash-tagging strategy will be available from 10x Genomics in the future (Cellplex).
5. Measuring the expression of cell surface proteins or bound antigenic peptides. This is accomplished by labeling cell surface proteins with antibodies conjugated to a feature barcode oligonucleotide, followed by the direct capture of the feature barcode by the gel bead primer. Cells can also be labeled using a feature barcode oligonucleotide conjugated to a MHC dextramer, such as a dCODE dextramer (detailed information can be found here from 10x Genomics:
   https://support.10xgenomics.com/single-cell-gene-expression/sample-prep/doc/demonstrated-protocol-cell-surface-protein-labeling-for-single-cell-rna-sequencing-protocols
   https://support.10xgenomics.com/single-cell-vdj/sample-prep/doc/demonstrated-protocol-cell-labeling-with-dextramers-for-single-cell-rna-sequencing-protocols).
6. This protocol starts from a cell suspension, usually isolated by flow cytometry. If thawing cells before staining and sorting, 10 μl benzonase per 10 ml cell suspension must be used. Incubation with FACS antibodies and/or TotalSeq-C antibodies is done on ice for 30 min. FACS sorting buffer washing steps should be repeated three times after staining.
7. Sorting for example 20,000 cells using the 70 μm nozzle and four-way purity mode yields 26.5 μl of sorted cell suspension. If the total sample is loaded, the volume should be topped to a final volume of 35.5 μl required to be loaded onto the chip. The final number of cells that are recovered depends on the cell concentration (*see* Table 1).
8. Counting the obtained cell suspension (after FACS sorting) under a microscope (or with a cell counter such as a Vi-Cell instrument) is crucial. A small excess of cells should be sorted, to allow for the removal of a small aliquot for cell counting (2 μl will be sufficient).

Table 1
Cell recovery as a function of cell concentration

	Cell recovery (in thousands)								
Cells/μl	2	3	4	5	6	7	8	9	10
800	**4.1** 33.7	**6.2** 31.6	**8.3** 29.5	**10.3** 27.5	**12.4** 25.4	**14.4** 23.4	**16.5** 21.3	**18.6** 19.2	**20.6** 17.2
900	**3.7** 34.1	**5.5** 32.3	**7.3** 30.5	**9.2** 28.6	**11.0** 26.8	**12.8** 25.0	**14.7** 23.1	**16.5** 21.3	**18.3** 19.5
1000	**3.3** 34.5	**5.0** 32.8	**6.6** 31.2	**8.3** 29.5	**9.9** 27.9	**11.6** 26.3	**13.2** 24.6	**14.9** 23.0	**16.5** 21.3
1100	**3.0** 34.8	**4.5** 33.3	**6.0** 31.8	**7.5** 30.3	**9.0** 28.8	**10.5** 27.3	**12.0** 25.8	**13.5** 24.3	**15.0** 22.8
1200	**2.8** 35.1	**4.1** 33.7	**5.5** 32.3	**6.9** 30.9	**8.3** 29.5	**9.6** 28.2	**11.0** 26.8	**12.4** 25.4	**13.8** 24.0

Top rows (bold): μl cell suspension; bottom rows: μl PBS or nuclease-free water (NFW)

9. If the sorting process of three samples takes <20 min for each, the samples can be loaded onto the same chip, whereas if sorting of each sample takes >30 min, the samples should be run on separate chips.
10. Failure to add partitioning oil can damage the Chromium controller!
11. Avoid storing GEMs longer than 1 h on ice! Do not centrifuge!
12. The cDNA contains AIRR transcripts and all the remaining cellular transcripts.
13. The time required for fragmentation of the AIRR library is shorter, and no size selection is performed between fragmentation and adapter ligation. Thus, you can start with the cDNA fragmentation, and while purifying the fragments, the IG/TR fraction can be fragmented. The master mix can be prepared for both libraries together.
14. The viscosity of the ligation mix is higher than for other mixes. Please ensure thorough mixing of all components, otherwise the ligation efficiency could be reduced.
15. AIRR, gene expression, and feature barcode libraries (BCL) are standard Illumina paired-end constructs which begin with P5 and end with P7. Libraries are prepared for Illumina sequencing with the following components: using version v1.1 with Single-Index Kit N Set A (for feature barcode libraries); Single-Index Kit T Set A (for gene expression and BCR/TCR libraries) and using v2 with dual indices (Dual Index Kit TN Set A (for feature barcode libraries); and Dual Index Kit TT Set A (for gene expression and IG/TR libraries).

These libraries include a P5 part that binds to the flow cell, the primer binding site for read 1 which contains a 16 bp 10× barcode to identify the cell assignment, followed by a 10 mer UMI for counting the transcripts, the TSO, and the poly-A-stretch. The transcript insert follows and is sequenced in read 2, followed by a region for the sequencing primer, the i7 index, and the P7 part that binds to the flow cell. The minimum sequencing lengths are 26 bp for read 1 and 91 bp for read 2. Sequencing these libraries produces standard Illumina BCL data. The optimal sequencing depth is 25–30 K reads/cell for cDNA libraries and 10 K reads/cell for AIRR and feature barcode libraries. For library loading we recommend the following: MiSeq (2× 150 bp reads): 15 pM; NovaSeq in XP mode: cDNA lib 250 pM; feature barcode library: 190–200 pM; AIRR library: 300 pM; NovaSeq in standard mode: cDNA lib 450 pM; and feature barcode library 300 pM: AIRR library 500 pM.

16. First-strand cDNA synthesis is primed by the 3′ SMART-Seq CDS Primer II A and uses the SMART-Seq sc TSO for template switching at the 5′ end of the transcript.
17. The SMART-Seq sc First-Strand Buffer forms precipitates. Keep this buffer at room temperature until you use it. Vortex before using it to ensure all components to be dissolved.
18. The lysis buffer contains a detergent, avoiding bubbles when mixing.
19. Due to small pipetting volumes, prepare no less than 250 μl of plain sorting solution (PSS) (enough for 18 wells). Scale up as needed. Be sure to count any negative control reactions you wish to include. If you need to sort large numbers of cells compared to the number of cDNA reactions you plan to prepare, you have the option to purchase the 10× lysis buffer (Takara Bio) separately.
20. To minimize bubble formation, set single- or multichannel pipettes to 11.6 μl, and pipette only to the first stop when aliquoting. Changing tips often also minimizes bubble formation.
21. If using PCR strips, strip caps can be used instead of aluminum foil, but are not practical when sorting a large number of samples.
22. If using PCR strips, leave them secured on the PCR rack for freezing. Store sorted samples at −80 °C until ready to proceed with cDNA synthesis. To use PCR strips sealed with an aluminum foil seal, use a clean razor blade to separate the individual strips, and then push up slightly on the tubes from under the PCR rack to loosen them before taking out the desired number of strips. Long-term storage at −80 °C may impact the

efficiency of cDNA synthesis; however, it is safe to store the cells for several weeks prior to cDNA synthesis.

23. Control cells should be in PSS described above. PSS does not contain the 3′ SMART-Seq CDS Primer II A, so it must be added here.

    The Control Total RNA is supplied at a concentration of 1 μg/μl. It should be diluted to match the concentration of your test sample using serial dilutions. For positive and negative controls, replace the cell sample with 2 μl of the diluted control RNA and water, respectively.

24. Since the PSS does not include the 3′ SMART-Seq CDS Primer II A, one needs to add it when thawing the samples.

25. If more than one reagent needs to be added (i.e., the 10× reaction buffer, the 3′ SMART-Seq CDS Primer II A and extra nuclease-free water), they can be pooled in a master mix and incorporated into each sample as a single addition.

26. Cycle guideline: When using good-quality control RNA, such as the Control Total RNA from the kit (mouse brain total RNA), 2 pg will generate 500–1500 pg/μl if using 19 cycles. It is best to choose a number of cycles that will yield <1500 pg/μl.

    Cycling guidelines based on cell type and pg RNA per cell: PBMCs (1–5 pg), 20 cycles; Jurkat cells (5 pg), 17 cycles; and lymphoblastoid cells (2–15 pg), 17–19 cycles.

27. Aliquot beads into 1.5-ml tubes upon receipt in the laboratory. Before each use, bring bead aliquots to room temperature for at least 30 min, and mix well by vortexing. Here the bead: sample ratio is 0.8:1. Prepare fresh 80% ethanol for each experiment. One will need 400 μl per sample. Use a magnetic separation device for 0.2-ml tubes, strip tubes, or a 96-well plate.

28. Do not pool the samples at the cDNA purification step. If pooling is desired, it can only be performed after library preparation.

29. Be sure to dry the pellet only until it is just dry. The pellet will look matte with no shine. If the pellet is under-dried, ethanol will remain in the sample wells. The ethanol might reduce the amplified cDNA recovery rate and ultimately the cDNA yield. Allow the plate to sit at room temperature until the pellet is dry. If the pellet is over-dried, there will be cracks in the pellet. It will take longer than 2 min to rehydrate (in the next step) and may reduce amplified cDNA recovery and yield.

30. The optimal cDNA input for Nextera XT library preparation is 100–300 pg. A larger amount of cDNA will generate libraries that are too large for sequencing on an Illumina instrument. The protocol below uses 125 pg of cDNA (in a volume of

1.25 μl), but any input between 100 and 300 pg will work. If all samples are correctly quantified and normalized to a uniform input amount before Nextera XT library preparation, sequencing libraries can be pooled before cleanup, and a relatively uniform amount of sequencing reads will be obtained. However, sample-to-sample read coverage varies, and one may observe some underrepresented or overrepresented samples with the pooling option. Always use a minimum of 2 μl of cDNA to make dilutions. Samples containing less than 100 pg/μl can still be used without dilution, but one may get fewer reads than for other samples if pooled for cleanup. If negative controls are going to be sequenced, they should be used without dilution.

31. If processing a large volume of samples, aliquot equal amounts of Tagmentation Premix into each tube of an eight-tube strip and then use an eight-channel pipette to distribute the Tagmentation Premix.
32. If processing a large number of samples, aliquot equal amounts of NPM into each tube of an eight-tube strip, and then use an eight-channel pipette to distribute the NPM.
33. Consult Illumina literature (Index Adapters Pooling Guide 1000000041074, available at https://support.illumina.com/downloads/index-adapters-pooling-guide-1000000041074.html) for proper index primer selection before proceeding to PCR amplification of the tagmented cDNA.
34. PCR-amplified libraries can be purified individually, or, optionally, the libraries can be pooled if the input cDNA was quantified and normalized to a uniform input amount before library preparation. The libraries are then purified by immobilization on NucleoMag NGS cleanup and size select (available from Takara Bio) beads. The beads are then washed with 80% ethanol, and then the cDNA is eluted with nuclease-free water.
35. The beads are viscous; pipette the entire volume and push it out slowly.
36. If 96 libraries were pooled, elute using 96 μl of nuclease-free water. Nuclease-free water volume is 50% of the original pool volume. If libraries were cleaned up individually, elute in 25 μl of nuclease-free water.
37. The SMART-Seq Single-Cell library is generated using Nextera XT and can be treated as a regular Nextera XT library. The libraries contain p5-i5-read 1 and p7-i7-read 2 sequences on 5′ and 3′ ends, respectively. The recommended read length is 75 bp. The sequencing depth depends on the number of pooled libraries and sample complexity. General considerations regarding sequencing depth and QC are discussed in AIRR Community method chapters “Bulk gDNA Sequencing of

Antibody Heavy-Chain Gene Rearrangements for the Detection and Analysis of B-Cell Clone Distribution" and "Bulk Sequencing from mRNA with UMI for Evaluation of B-Cell Isotype and Clonal Evolution," in **Note 15** and on the Illumina website.

38. Several libraries are usually sequenced together, and each library is individually barcoded. The raw sequencing data obtained from Illumina sequencers are in BCL file format, which contains the raw base calls for all the libraries of the sequencing run. The first step of the demultiplexing consists of splitting the whole BCL file into individual FASTQ files. Each of the resulting FASTQ files will contain the sequence reads linked to a quality score for each base. As sequencing is performed in a paired-end manner and on two lanes of the sequencer, the command will generate four FASTQ files for each library (two lanes x two read directions). To increase reproducibility, it is recommended to start the processing from the original BCL files, even if demultiplexed FASTQ files are provided by the sequencing provider. It is recommended to initially perform the analysis strategy separately for each sequencing lane, until quality control steps. If no issues have been found on the different lanes, the analysis can be repeated on the complete dataset. To extract the reads from one or more specific lanes, argument `--lanes=1` (for lane 1) and `--lanes=2` (for lane 2) can be used.
39. The commands cellranger vdj and cellranger count will generate several outputs, including an HTML report for each type of library, a vloupe.vloupe file for an exploratory analysis, a detailed description of each identified VDJ rearrangements in the files airr_rearrangement.tsv (annotated contigs and consensus sequences of VDJ rearrangements in the AIRR TSV format) and clonotype.csv (high-level descriptions of each clonotype), and a features.tsv.gz file containing the gene expression counts. The last two files will be used for in-depth analysis.
40. For a complete reference, see the 10x Genomics support website: https://support.10xgenomics.com/single-cell-vdj/software/pipelines/latest/using/vdj.
41. The example mentioned here is available on the 10x Genomics website. The command will provide the VDJ and UMI counts for each cell within the library. The main pipeline output files will be in a subfolder named "outs" with the folder indicated by `"--id"`. Likewise, the quantification of gene expression can be performed using cellranger count, which will provide the gene and UMI counts per gene for each cell within the library.

42. The choice of reference libraries and cell numbers should be carefully considered when processing 10× sequencing data. Reference libraries: The read alignment procedures require appropriate reference sequences for VDJ and gene expression profiling. For human transcriptome data, "refdata-gex-GRCh38–2020-A" should be used, and for human AIRR-seq data, "refdata-cellranger-vdj-GRCh38-alts-ensembl-5.0.0" is recommended; both libraries are available from 10x Genomics. Cell numbers: When performing a single-cell experiment, a given number of cells have been loaded for the library preparation. Cell Ranger automatically detects unique cells, assuming a high RNA content per cell. However, in heterogeneous populations, cells with low RNA content could accidentally be discarded. To circumvent this, Cell Ranger allows to indicate the expected cell number using the --expect-cells parameter (default: 3000 cells).
43. The Loupe Browser (version 5.0.1) software is a complementary tool provided by 10×. This tool uses the vloupe.vloupe file and can run on MacOS or Windows.
44. In-depth analysis can be done using various tools, mainly R packages. The Seurat package [10] can be used to analyze the gene expression data obtained. The TR sequences can be analyzed using the scRepertoire package [11]. The input for the scRepertoire R package is the *filtered_contig_annotations.csv* file from Cell Ranger. The individual files can be combined into the list element, and the list is then used as an input for the *combineTCR/BCR* function, which combines individual chains into clones for each cell. One can also exclude single-chain clones, or filter the chains based on their UMI count in the case of multiple chains of the same type in a single cell by setting different parameters of the function. There are several definitions of a clone based on the user's needs, which one can choose with the *cloneCall* parameter of the various functions. The package provides analysis for a number of modalities, including calling clones, clonal space/homeostasis, clonal diversity, and repertoire overlap between samples. Moreover, the package allows the investigation of the clones with regard to the cellular gene expression by combining the clones with Seurat objects by using *combineExpression* function. One needs to pay attention to consistent naming of the barcodes between the Seurat and scRepertoire objects. Scirpy is another recently developed Python toolkit that allows simple and straightforward analysis of and visualization of scAIRR data and also allows integration with transcriptome data [12]. VDJView is also a recently published tool providing similarly multi-omic integration [13].

45. The two main steps in the TraCeR pipeline are assemble and summarise. *Assemble*. Sequencing reads derived from TRs are extracted from single-cell RNA-seq data by aligning the reads from each cell against synthetic TR rearrangements made up of all possible combinations of V and J genes with a masked junctional region. TR-derived reads are then assembled into paired full-length TR sequences and characterized by the V and J gene usage and junctional sequence. TraCeR reports both the full nucleotide sequence of reconstructed TR rearrangements, as well as a shorter sequence identifier (such as TRBV31_AG TCTTGACACAAGA _TRBJ2–5) for each rearrangement. *Summarize*. Sequence identifiers for each cell are used to identify groups of clonally related T cells. Both productive and nonproductive TR rearrangements are considered clonally related if they have the same sequence identifier (thus the same V gene, J gene, and junctional nucleotide sequence). Clonal relationships are visualized as clonal networks, where each cell is represented by a node in the graph. Reconstructed TR rearrangements are represented within nodes by horizontal lines or sequence identifiers colored according to locus and productivity. Edges between the nodes represent shared TR rearrangements and are also colored according to locus.
46. Data usable in TraCeR needs to fulfil the following demands. TraCeR requires data from a sequencing library protocol that generates sequencing reads from full-length mRNA transcripts, for example, SMART-Seq. Sequencing approaches that only provide partial coverage of transcript ends are not suitable for TR reconstruction by TraCeR. TraCeR works with both paired-end and single-end reads, although paired-end reads give the highest accuracy and reconstruction rate. The library must be sequenced with a minimum read length of 50 bases. The recommended read depth is 250,000–500,000 paired-end reads per cell (in total 0.5–1 million reads per cell), depending on the cell activation state and how much information about the rest of the transcriptome is needed.
47. TraCeR expects raw reads to be demultiplexed according to cell of origin. This means that two files (one for R1 and one for R2) for each cell are expected for paired-end output and one file per cell for single-end output. TraCeR accepts FASTQ (fastq or fastq.gz) files as input.
48. TraCeR reconstructs TR sequences from raw reads when run in *assemble* mode. The reconstruction is performed separately for

each cell but may be performed for multiple cells in parallel if running TraCeR on a computational cluster.

49. It is important to remove cell duplets or multiplets from single-cell data. TraCeR should only be applied to single-cell data, as the tool assumes all input reads provided to the *assemble* step to derive from one cell. A single T cell should have no more than two reconstructed TR rearrangements per TR locus, and the number of reconstructed TR rearrangements for a given locus may thus be used to identify cell duplets/multiplets or potential contaminations.
50. Whether sharing of rearranged TR sequences between cells is defined as evidence of clonality depends on the type of TR chain and how strict the definition of clonality should be. All TR rearrangements present in a given cell may not always be reconstructed as the detection sensitivity is never 100% and depends on experimental as well as biological factors. Thus, clone groups can be inferred based on various levels of evidence. We therefore recommend that clone groups should be interpreted by the user depending on the specific biological questions of interest. Some typical patterns are as follows: (1) Clone groups consisting of cells sharing the same TR rearrangements, including a productive TRA and TRB (or TRD and TRG), are clonally related with a high confidence. Sharing of additionally rearranged sequences when these are detected in cells (such as a nonproductive rearrangement) further increases the confidence. (2) Sometimes clone groups consisting of subclones may exist due to the possibility of a TRB rearrangement to be found in combination with different TRA rearrangements due to the TRB rearrangement preceding cell proliferation and TRA rearrangement. Such patterns can be seen as cells all sharing the same TRB rearrangement, with clusters of cells within the clone sharing TRA rearrangements. Such patterns may indicate that the TRB is particularly important for antigen specificity of the clone. It is up to the user whether such composite clones should be divided into subclones sharing the same TRA or not. (3) Because of unsuccessful reconstruction of all TR rearrangements in some cells, it may occur that the only shared TR rearrangement between a cell and a clone group is nonproductive or a TRA. In such cases it cannot be certain that the cell belongs to the clone group, but there is also no evidence to the contrary. Such cells may be excluded manually by the user if strict definitions of clones are desired.

Acknowledgments

We thank Magnolia Bostick, Christian Busse, Eline T. Luning Prak, Gloria Kraus, Chaim Schramm, Nicolas Tchitchek, Ulrik Stervbo, and Johannes Trück for helpful discussions and inspiration around the manuscript and for editing and proofreading and Elaine Chen for contributing figures. EMF and KME were funded by iMAP (ANR-16-RHUS-0001), Transimmunom LabEX (ANR-11-IDEX-0004-02), TriPoD ERC Research Advanced Grant (Fp7-IdEAS-ErC-322856), AIR-MI (ANR-18-ECVD-0001), iReceptorPlus (H2020 Research and Innovation Programme 825821), and SirocCo (ANR-21-CO12-0005-01) grants. AE is supported by grants from the Deutsche Forschungsgemeinschaft (BO 3429/3-1 and BO 3429/4-1) and the BMBF (RESET-AID). IL is funded by KG Jebsen (project SKGJ-MED-017). Conflict of interest: AE, EMF, KME, IL, NG, and SR declare no conflict of interest. KM is an employee at 10x Genomics, Pleasanton, CA, USA, and NG is an employee at Takara Bio, Mountain View, CA, USA. Both companies produce a kit described in this protocol.

References

1. Kantor AB, Merrill CE, Herzenberg LA, Hillson JL (1997) An unbiased analysis of V(H)-D-J(H) sequences from B-1a, B-1b, and conventional B cells. J Immunol 158:1175–1186
2. Brezinschek HP, Brezinschek RI, Lipsky PE (1995) Analysis of the heavy chain repertoire of human peripheral B cells using single-cell polymerase chain reaction. J Immunol 155: 190–202
3. de Simone M, Rossetti G, Pagani M (2018) Single cell T cell receptor sequencing: techniques and future challenges. Front Immunol 9: 1638
4. Wardemann H, Busse CE (2017) Novel approaches to analyze immunoglobulin repertoires. Trends Immunol 38:471–482
5. Fuchs YF, Sharma V, Eugster A, Kraus G, Morgenstern R, Dahl A, Reinhardt S, Petzold A, Lindner A, Löbel D, Bonifacio E (2019) Gene expression-based identification of antigen-responsive CD8+ T Cells on a single-cell level. Frontiers in Immunology. https://doi.org/10.3389/fimmu.2019.02568
6. Stubbington MJT, Lönnberg T, Proserpio V, Clare S, Speak AO, Dougan G et al (2016) T cell fate and clonality inference from single-cell transcriptomes. Nat Methods 13:329–332
7. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J 17:10
8. Lindeman I, Stubbington MJT (2019) Antigen receptor sequence reconstruction and clonality inference from scRNA-seq data. Methods Mol Biol 1935:223–249
9. Schuldt NJ, Binstadt BA (2019) Dual TCR T Cells: Identity Crisis or Multitaskers? J Immunol 202:637–644
10. Satija R, Farrell JA, Gennert D, Schier AF, Regev A (2015) Spatial reconstruction of single-cell gene expression data. Nat Biotechnol 33:495–502

11. Borcherding N, Bormann NL, Kraus G (2020) scRepertoire: an R-based toolkit for single-cell immune receptor analysis. F1000Res 9:47
12. Sturm G, Szabo T, Fotakis G, Haider M, Rieder D, Trajanoski Z (2020) Scirpy: a Scanpy extension for analyzing single-cell T-cell receptor-sequencing data. Bioinformatics 36: 4817–4818
13. Samir J, Rizzetto S, Gupta M, Luciani F (2020) Exploring and analysing single cell multi-omics data with VDJView. BMC Med Genomics 13:29

Chapter 21

Quality Control: Chain Pairing Precision and Monitoring of Cross-Sample Contamination: A Method by the AIRR Community

Cheng-Yu Chung, Matías Gutiérrez-González, Sheila N. López Acevedo, Ahmed S. Fahad, and Brandon J. DeKosky and on behalf of the AIRR Community

Abstract

New approaches in high-throughput analysis of immune receptor repertoires are enabling major advances in immunology and for the discovery of precision immunotherapeutics. Commensurate with growth of the field, there has been an increased need for the establishment of techniques for quality control of immune receptor data. Our laboratory has standardized the use of multiple quality control techniques in immunoglobulin (IG) and T-cell receptor (TR) sequencing experiments to ensure quality control throughout diverse experimental conditions. These quality control methods can also validate the development of new technological approaches and accelerate the training of laboratory personnel. This chapter describes multiple quality control techniques, including split-replicate cell preparations that enable repeat analyses and bioinformatic methods to quantify and ensure high sample quality. We hope that these quality control approaches can accelerate the technical adoption and validated use of unpaired and natively paired immune receptor data.

Key words B-cell receptor, T-cell receptor, Next-generation sequencing, PCR, Single-cell analysis, Replicate analysis, Quality control

1 Introduction

Recent developments in single-cell technologies have made it possible to capture the sequences of both heavy and light chains of immunoglobulins (IG) at high throughput [1–3]. These high-throughput methods require single-cell isolation approaches to enable the identification of the IG heavy and light chain cognate pairs. Clonal spikes of immortalized B-cell lines can provide some measure of IG heavy and light chain pairing quality control in single cell assays, but the expression level of clonal spikes is often very different from that of native B cells in a given sample. One method

Anton W. Langerak (ed.), *Immunogenetics: Methods and Protocols*, Methods in Molecular Biology, vol. 2453, https://doi.org/10.1007/978-1-0716-2115-8_21,

to accurately analyze the quality of single-cell analyses relies on the determination of B-cell IG heavy and light chain pairs in split-replicate samples. By processing two or more replicates of the same sample, the experimental or technical performance and reproducibility can be analyzed via these replicates for methods development and high-quality data and to perform sample-specific quality control. For highest experimental accuracy and statistical validity, it is important that both replicates are treated in exactly the same way throughout all stages of the experiment and subsequent data analysis.

B-cell replicate analyses can be performed both with two aliquots of the same patient samples (e.g., split aliquots of the same blood PBMCs), which is often the simplest method. Some information can also be obtained using different tissue samples from the same individual or animal (e.g., comparing spleen and bone marrow compartments from the same mouse), although with some statistical compromises when the cell sources are different, because the samples are therefore not true replicates. A robust experimental analysis of split replicates can be performed after in vitro B-cell expansion is providing a large pool of expanded B cells, which will be distributed across the two replicates. This approach enables precise determination of IG heavy and light chain pairing accuracy using a given single-cell technique. The first experimental section of this methods article describes an effective way to generate expanded B cell populations that are ready for split-replicate analyses and associated statistical determinations of the pairing precision accuracy.

In addition to IG sequence analysis, the high-throughput sequencing of T-cell receptor (TR) gene rearrangements provides insights into dynamic cellular adaptive immune responses. TR screening can also be useful for discovery of therapeutic T cells or evaluations of T cell–based therapies [4, 5]. Single-cell TR sequencing approaches further accelerate progress by yielding paired alpha and beta TCR chain information. Methods development and quality control of single T-cell sequencing are facilitated by split-replicate studies for high-quality statistical determination of technical single-cell accuracy. The second experimental protocol describes the retrieval of frozen cell samples, the purification of a T-cell subset of interest (here we use CD8+ T cells as a demonstration), and T-cell expansion in vitro prior to TR sequencing.

One important approach for quality control of single-cell sequencing experiments is the use of technical split-replicate samples, which provide a powerful tool to measure the reproducibility of a technique or assay. Split replicates can also provide major advantages for experimental training of new individuals and for the sample quality control of critical samples. While not strictly necessary, we find that in vitro expansion of B or T cells prior to analysis can provide a major boost to the number of overlapping

clones in a given split-replicate analysis, thereby enhancing statistical accuracy of the pairing precision analysis, although the in vitro expansion can alter the original distribution of clonal frequency within a dataset. To determine the single-cell chain pairing precision, we assume that an IG heavy-light chain pair found across multiple replicates is a true positive, while an IG heavy chain paired with a different IG light chain found across replicates is a false positive in at least one replicate [2, 6]. We base our pairing precision analyses on the CDR3 nucleotide sequence for highest accuracy, as many IG heavy or light chain CDR3s can be encoded by similar amino acid sequences across individuals but still derive from unique V(D)J rearrangement events. In the third experimental protocol here, we describe the evaluation of single-cell pairing precision for both IG and TR sequences using a common bioinformatic approach.

A major source of potential experimental error is PCR contamination, which can occasionally appear at any research group and must quickly and effectively be eliminated to avoid continued spread. In particular, it is important to track experimental samples for the presence of potential PCR contamination across an entire laboratory and group of researchers as a means of ensuring high sample quality in an ongoing and up-to-date basis. Our final protocol describes the construction and analysis of a database of samples previously analyzed in a laboratory or research group to monitor for cross-sample contamination in new samples. The database can be as simple as a collected set of files containing the information needed, and ongoing additions to the database permits facile monitoring and analysis for any potential PCR contamination events to ensure high quality control.

The four protocols provided here describe both experimental and bioinformatic methods that help ensure robust and rigorous data from next-generation sequencing technologies. We believe that these key quality control methods can be useful for other laboratories and can accelerate the growth of sequence data and associated information derived from single-cell adaptive immune receptor sequencing techniques.

2 Materials

2.1 B-Cell Stimulation to Generate Split-Replicate Cell Samples

1. EasySep Human B Cell Enrichment Kit II w/o CD43 depletion (Stemcell Technologies).
2. EasySep Magnet (Stemcell Technologies).
3. RoboSep Buffer (Stemcell Technologies).
4. Cell strainer (35 μm, BD Falcon).
5. Human CD27+ MicroBeads (Miltenyi Biotec).

6. Biosafety Cabinet.
7. Hemocytometer.
8. Microscope.
9. MiniMACS MS Separation columns (Miltenyi Biotec).
10. MiniMACS Magnetic Separator and Stand (Miltenyi Biotec).
11. Microcentrifuge Tubes, 1.5 mL (NEST Scientific USA).
12. Refrigerated Centrifuge (Thermo Fisher Scientific; Sorvall; Legend XTR).
13. MACS Buffer: DPBS (Gibco) + 0.5% BSA (VWR) + 2 mM EDTA (Life Technologies). Sterile filter and store at 4 °C.
14. IMDM Media: Iscove's Modified Dulbecco's Medium (Thermo Fisher Scientific) supplemented with 10% FBS (Fisher Scientific), 1% Penicillin/Streptomycin (Thermo Fisher Scientific), and 1% nonessential amino acids (Fisher Scientific).
15. Human Interleukin-2 (IL-2, Sigma-Aldrich).
16. Human Interleukin-21 (IL-21, PeproTech).
17. 3T3-CD40L cells (available from Boston Cellron, 3T3-msCD40L).
18. 48 Well Cell Culture Plate (NEST Scientific USA).
19. Autoclaved water.
20. Cell incubator.

2.2 T-Cell Stimulation to Generate Split-Replicate Cell Samples

1. Complete RPMI 1640 Medium (cRPMI): RPMI 1640 (ATCC modification) (Thermo Fisher Scientific) with 10% heat inactivated FBS (Thermo Fisher Scientific) and Penicillin–Streptomycin (final concentration 100 units/mL of penicillin and 100 μg/mL of streptomycin).
2. IL-2 from Biopharmaceutical Development Program (BDP) at the Frederick National Laboratory for Cancer Research (FNLCR) (1×10^6 U per vial, stock concentration 1000 U/μL).
3. ImmunoCult™ Human CD3/CD28 T Cell Activator (Stemcell Technologies) for T-cell expansion when T-cell sources is from PBMC.
4. EasySep™ Human CD8+ T Cell Isolation Kit (Stemcell Technologies).
5. EasySep™ Magnet: Magnet for column-free immunomagnetic separation for Human CD8+ T Cell Isolation Kit (Stemcell Technologies).
6. Dynabeads™ Human T-Activator CD3/CD28/CD137 (Thermo Fisher Scientific) for T cell expansion when T-cell sources are from splenocytes.

7. Falcon™ Round-Bottom Polystyrene Test Tubes with Cell Strainer Snap Cap, 5 mL.
8. 2-Mercaptoethanol (MilliporeSigma).

2.3 Technical Precision Analysis of Paired IG Heavy/Light or Paired TR Alpha/Beta Sequencing

1. The bash script `precision_calculator.sh` is available from GitHub at https://github.com/dekoskylab/quality_control.
2. Paired sequence data for each replicate should be organized into separate tab-separated files, with the observed paired read counts in the first column, IG heavy or TR beta chain CDR3 nucleotide junction sequences in the second column, and IG light or TR alpha chain CDR3 nucleotide junction sequence in the third column (*see* **Note 1**).

2.4 Laboratory-Scale Global Detection and Monitoring of Cross-Sample Contamination Events

1. The python script `PCR_QC_analysis.py` is available from GitHub at https://github.com/dekoskylab/quality_control. The script compares data output and rapidly searches for the presence of PCR contamination analysis in IG and/or TR data.
2. `PCR_QC_analysis.py` requires the software dependencies Python ≥v3.6 and pandas ≥0.25. The script has not been tested in older versions of python or pandas.

3 Methods

3.1 B-Cell Stimulation for the Generation of Split-Replicate Samples as Repeated Analyses

Here, we describe a protocol for CD27+ antigen-experienced B-cell retrieval via magnetic-activated cell sorting (MACS) that uses magnetic beads that are coated with antibodies or enzymes associated with surface markers of our targeted cells. Alternatively, flow cytometry can be used to isolate high-purity B-cell populations of interest. Next, B-cell activation by cells presenting CD40 ligand (CD40L), along with other cytokines, is performed to induce B-cell proliferation in vitro [2, 6–8]. This robust selection and stimulation protocol yields a substantial B-cell population, normally expanded two- or threefold after a 5-day culture period, ready for subsequent split-replicate analyses, and single B-cell quality control studies. In this section, **steps 1–13** describe human B-cell enrichment without CD43 depletion, while **steps 14–29** describe CD27+ memory B cell selection, and finally, **steps 30–38** describe the procedures for cell culture for in vitro expansion.

1. Resuspend cells such at 50 million cells/mL, while staying within the range of 0.25–2.0 mL total volume (*see* **Note 2**).
2. Pass the cell solution over a cell strainer into a 5 mL polystyrene tube.
3. Add 50 μL/mL of cocktail enhancer to the sample and mix.

4. Add 50 μL/mL of enrichment cocktail to the cells. Flick to mix and incubate for 5 min at room temperature.
5. Vortex rapid spheres (until evenly dispersed).
6. Add 35 μL/mL of rapid spheres to the sample (not through the filter) and flick to mix.
7. Calculate the current volume based on amounts added in **steps 1–6**, then top up to 2.5 mL with RoboSep and mix.
8. Place in magnet, and let sit at room temperature for 3 min.
9. Collect the mix into a new cell strainer tube.
10. Leave the magnet and tube inverted for 2–3 s, then return upright. Do not shake or blot off any drops that may remain hanging from the mouth of the tube.
11. Place this new tube with the cells in the magnet and incubate at room temperature for 1 min.
12. Collect the mix into a new 15 mL conical tube.
13. Proceed to count the recovered B cells (cell count required for next part of the protocol).
14. Once the cell number is determined (**step 13**), spin cells in a centrifuge at 300 × *g* for 10 min at 4 °C.
15. Transfer the supernatant completely into a separate tube, and set the supernatant aside.
16. Flick the pellet and resuspend in 100 μL MACS buffer.
17. Add 25 μL of CD27 MicroBeads (per 10^7 cells for a 100 μL cell suspension).
18. Mix well by flicking the solution several times. Incubate this mixture for 15 min at 4 °C.
19. After incubation, transfer the bead cell mixture (~125 μL) to a 1.5 mL microtube, and wash the 15 mL tube with 1 mL MACS buffer to collect any remaining microbeads and cells.
20. Add the 1 mL wash to the 1.5 mL microtube. Centrifuge the 1.5 mL microtube by placing it inside a 15 mL conical tube at 300 × *g* for 10 min at 4 °C.
21. Label three new different 1.5 mL microtubes as "pre-wash," "flow-through," and "elute," respectively.
22. Remove the supernatant completely, and add the supernatant to the microtube labeled as "pre-wash." Flick and resuspend in 500 μL MACS buffer.
23. Place the magnetic separator and the stand along with a magnetic column inside the biosafety hood cabinet over the "pre-wash" tube.
24. Prepare the column by rinsing it with 500 μL of plain MACS Buffer. Collect the rinsed solution into the "pre-wash" tube (*see* **Note 3**).

25. Take the 1.5 mL tube labeled "flow-through," and place it under the magnetic column. Apply the cell suspension onto the magnetic column placed on the magnetic separator. Collect the flow through containing unlabeled cells.
26. Wash the column with 500 μL of MACS Buffer. This step requires two washes, each of 500 μL. Collect the flow through (*see* **Note 4**).
27. Once the flow through is collected, remove the magnetic column from the separator. Take the magnetic column far away from the magnet and immediately place it on a suitable new 1.5 mL microtube ("elute" tube).
28. Pipet 1 mL MACS Buffer onto the column and quickly plunge out the magnetically labeled cells by pushing the plunger into the column and collecting the eluted fluid. Label this elute tube as CD27+ B cells (*see* **Note 5**).
29. Proceed to count the recovered CD27+ B cells using a hemocytometer.
30. Using the recovered cells from **step 29**, determine the volume of stimulation components and cells to add as detailed below (*see* **Note 6**).

Component	Desired final concentration
IL-21	50 ng/mL
IL-2	100 units/mL
Irradiated 3T3-CD40L cells	100,000 cells/mL
CD27+ B cells	150,000 cells/mL
IMDM media	to final total volume of 1 mL

31. Spin the isolated CD27+ B cells in a centrifuge at 300 × *g* for 10 min.
32. Remove the supernatant without disrupting cell pellet. Flick the pellet, then suspend in a volume of culture medium determined in **step 30**.
33. Determine the number of wells for culture: 150,000 CD27+ B cells per well in 1 mL final volume (for a 48-well plate) (*see* **Note 7**).
34. Add treatment components or control components to their respective mixes and culture the cells on a 48-well plate. Add culture medium from **step 30** to reach the recommended total volume.
35. Add sterile water to all remaining empty wells (autoclaved water) to prevent evaporation.

36. Incubate for 5 days in a humidified 37 °C incubator in 5% carbon dioxide.
37. Visualize under a microscope on day 1 and 5, take pictures of the stimulation progress as desired (*see* **Note 8**).
38. When cells are appropriately expanded, continue with single-cell sequencing and analysis using the desired techniques. The expanded cell samples may be split into two (or more) replicates as a cell source for parallel analyses.

3.2 T-Cell Stimulation for the Generation of Split-Replicate Samples as Repeated Analyses

T-cell populations can be divided as replicates prior to single-cell analyses for robust statistical determination of technical performance. To obtain primary T cell populations, MACS or immunophenotyping can be used. MACS is fast, easy to scale, and may have higher viability post-purification than FACS. Alternatively, FACS-based selection permits more complex cell subset isolations, including fluorochrome-labeled multimeric peptide-MHC screening [9], peptide-pulsed antigen-presenting cells [10], and peptide mega-pools [11].

Direct TR sequencing analysis after T-cell isolation using MACS for split-replicate TR analysis is feasible, although the overlap will not be as high as for in vitro expanded T-cell populations. Larger cell numbers following stimulation can allow for more complete coverage of the TR repertoire when multiple assays are designed (e.g., staining cells for multiple peptide/MHC targets). Split-replicate samples can be used for their TRα/TRβ chain pairings to test the statistical accuracy of single-cell TR sequencing. The following protocol describes the collection and in vitro expansion of T cell populations to generate split-replicate populations for single-cell technical analysis studies. In this section, **steps 1–7** detail how to thaw frozen PBMC or splenocyte samples that contain the desired T-cell populations, while **steps 8–15** describe how to purify CD8+ T cells from PBMC samples, and finally **steps 16–18** describe how to expand the T cells in vitro.

1. Pre-warm cRPMI in a 37 °C water bath. Retrieve the EasySep™ buffer from the refrigerator, and allow it to warm up to room temperature (15–25 °C). Aliquot 9 mL of pre-warmed cRPMI into a 15 mL centrifuge tube.
2. Retrieve the cryovial(s) containing the cells from the liquid nitrogen tank and immediately place into the 37 °C water bath for thawing. Continuously swirl the cryovial until a small piece of ice is left (*see* **Note 9**).
3. Wipe down the cryovial with 70% ethanol, and bring the vial into the biosafety cabinet.
4. Transfer the cells from cryovial dropwise into the pre-warmed cRPMI in the centrifuge tube. Gently invert the tube three times to mix the cells with cRPMI.

5. Centrifuge the cells for 300 × *g* for 10 min at 22 °C.
6. For PBMC samples, please proceed to step 2-CD8 T-cell purification from PBMC samples. For splenocyte samples, measure the cell density and resuspend the cells at a density of 1 x 10^6 cell/mL using cRPMI supplied with 200 IU/mL of IL-2 and 0.05 mM of 2-mercaptoethanol.
7. Add prewashed Dynabeads™ Human T-Activator CD3/CD28/CD137 at a bead-to-cell ratio of 1:5. Add IL-2 at a final concentration of 200 IU/mL. Proceed to the following steps in T-cell culture expansion.
8. Remove the supernatant and resuspend the cells in 0.5 mL EasySep buffer. Pass the cell mixture through the cell strainer. Measure the cell density and adjust the cell density to 0.5 × 10^7 cell/mL. The reaction volume for CD8 T cell purification should be between 0.25 and 2 mL. If you do not have enough cells, we suggest proceeding to the next step with a reaction volume of 0.25 mL regardless of the cell density. If you have more than 1 × 10^7 cells, please split the cells into two reactions.
9. Retrieve the EasySep™ Human CD8+ T cell isolation kit from the refrigerator. Add 50 μL Isolation Cocktail per milliliter of reaction volume. Mix the reaction by pipetting up and down gently three times. Incubate the reaction for 5 min.
10. Vortex the tube containing rapid spheres from the kit for 30 s. Ensure that the rapid spheres are evenly dispersed.
11. Add 50 μL of rapid spheres per milliliter of reaction volume. Then top off with EasySep buffer to a final total volume of 2.5 mL.
12. Mix the reaction by pipetting up and down gently three times. Remove the snap cap and place the tube into the EasySep™ Magnet. Incubate the reaction for 3 min. Do not move the magnet during the incubation.
13. In one continuous movement, invert the magnet along with the tube containing the cells and pour the cell suspension into a 50 mL centrifuge tube.
14. Count the cells and centrifuge the cells at 300 × *g* for 10 min.
15. Remove the supernatant, and resuspend the cell at a density of 1 × 10^6 cell/mL. Use 25 μL of ImmunoCult™ Human CD3/CD28 T Cell Activator per milliliter of the T cell culture (*see* **Note 10**). Add IL-2 into the culture medium at a final concentration of 200 IU/mL.
16. Feed the PBMC-derived T cells regularly to maintain the cell density between 0.5 × 10^6 cell/mL and 2 × 10^6 cell/mL using cRPMI with 200 IU/mL IL-2 during the expansion.

17. Feed the splenocyte-derived T cells regularly to maintain the cell density between 1×10^6 cell/mL and 2×10^6 cell/mL using cRPMI with 200 IU/mL IL-2 and 50 mM 2-mercaptoethanol during the expansion.
18. When cells are appropriately expanded, continue with single cell sequencing and analysis using the desired techniques (*see* **Note 11**). The expanded cell samples may be split into two (or more) replicates as a cell source for parallel analyses (*see* **Note 12**).

3.3 Monitoring the IG Heavy/Light and TR Alpha/Beta Technical Pairing Precision Using Split-Replicate Single-Cell Sequencing Samples

Here, we describe how to run the software for automated pairing precision analysis.

1. Execute the command:

```
bash precision_calculator.sh <REPLICATE_FILE1> <REPLI
CATE_FILE2>
```

2. In an automated fashion, the script will carry out the following steps:
   (a) Extract and count repeated CDR-H3/CDR-β3 sequences. Only CDR-H3/CDR-β3 sequences that overlap between replicates are counted.
   (b) Count true positives (TP). CDR-H3/CDR-β3 paired with exact match CDR-L3/CDR-α3 sequences. An exact match is equal length and identical sequence.
   (c) Count false positives (FP). CDR-H3/CDR-β3 paired with different CDR-L3/CDR-α3 sequences, defined by different lengths, in at least one replicate.
   (d) Count TP and FP among CDR-H3/CDR-β3 with mismatched CDR-L3/CDR-α3.

   Clonally expanded variants can have mutated CDR-L3/CDR-α3 sequences either due to somatic hypermutation and sequencing error (IG), or sequencing error alone (TR), but still clearly derive from the same V-J rearrangement and represent different variants of the same BCR/TCR cell clone. CDR-L3/CDR-α3 with equal length and not more than 20% mismatches can generally be considered TP, while equal length CDR-L3/CDR-α3 with more than 20% mismatches can be considered FP.

   The Hamming distance is then used to calculate the degree of mismatch:

$$\%\ \text{CDRL3 difference} = \frac{(\text{Number of mismatches})}{(\text{Sequence length})}$$

(e) The script will then calculate and report the chain pairing-precision in the following fashion (*see* **Notes 13** and **14**):

The pairing precision (P) is calculated from the number of TP and FP [12]:

$$P = \frac{(\mathrm{TP})}{(\mathrm{TP} + \mathrm{FP})}$$

Therefore, the collective precision of two independently technical replicates (R1 and R2) mentioned above as follows [2, 6]:

$$P_{\mathrm{R1\ and\ R2}} = \frac{(\mathrm{TP}_{\mathrm{R1\ and\ R2}})}{(\mathrm{TP}_{\mathrm{R1\ and\ R2}} + \mathrm{FP}_{\mathrm{R1\ and\ R2}})}$$

The probability of independent events is equal to the product of the independent event probabilities. Moreover, the P of technical replicates 1 and 2 (PR1 and PR2) are considered to be equal, a property of technical replicates.

$$P_{1^2} = P_{\mathrm{R1}} \times P_{\mathrm{R2}} = P^2$$

Solve for the above two equations and estimate P for a single analysis:

$$P_{\mathrm{R1\ and\ R2}} = P^2 \frac{(\mathrm{TP}_{\mathrm{R1\ and\ R2}})}{(\mathrm{TP}_{\mathrm{R1\ and\ R2}} + \mathrm{FP}_{\mathrm{R1\ and\ R2}})}$$

3.4 Laboratory-Scale Global Detection and Monitoring of Cross-Sample Contamination Events

In this section, **steps 1** and **2** detail how to prepare the data files, while **steps 3** and **4** describe how to perform the analysis for large-scale contamination monitoring.

1. Prepare tabulated files containing annotated IG and/or TR sequencing results. The files should include the following named columns in a comma separated file: sequence identifier (column: `sequence_id`), CDR3 sequences or CDR3 junction nucleotide sequences (column: `cdr3`), and constant region gene assignment (column: `c_call`). Files should have a commonly arranged naming system that can be used as a pattern for quick capture using character expansion (such as EXP_NAME_final_file.txt).
2. Prepare a metadata file containing experimental information for each sample. This file must have a column named `file` that contains the file names, which is used to match the contamination analysis output. Although the script does not have specific metadata requirements, we recommend the metadata file that include all MiAIRR information and use AIRR standard names whenever possible.

3. Execute the command:

```
python PCR_QC_analysis.py
<sequencing_files_pattern> <metadata_file>
```

4. The script will carry out the following steps:
   (a) Search for all desired files using a pattern expansion strategy. For example, using the term <PBMC> as <sequencing_files_pattern> variable will match all files with the phrase "PBMC" in the current directory. Then, the script will create all pairwise combinations of matched files (*see* **Note 15**).
   (b) CDR-H3 nucleotide sequences are matched across files. In this example the number of matches is divided by the total number of clones. The protocol does not discriminate between potential convergent or public responses and actual cross-contamination events. Presence of shared CDR-H3s should be assessed on a case-by-case basis, considering the nature of each experiment and the extent of shared sequences. Unrelated samples from literature or from different laboratories can be used to set a minimum threshold for convergent responses within your database. Cross-contamination events will be readily detectable and will exceed background levels established for convergent responses by a substantial margin.
   (c) For IG sequences, clones are also binned by antibody isotype to allow a closer analysis of potential contaminations.
   (d) Output overlap fractions are processed as pairwise comparisons and annotated from an external database containing important file metadata. As another control, the provided script also reports the fraction of shared CDRs for a single file. The sum result of the constituent fractions for a complete repertoire should always be 1.
5. Expected results:

   For a pair of unrelated samples, the level of shared CDR-H3 sequences should be close to zero. However, a low level of shared sequences can be expected, and a low threshold can be considered (e.g., <1 in 10^4 clones) [13] and should be addressed on a case-by-case basis. For the CDR-L3, a lower diversity translates into a higher number of public sequences [14], which should be taken into consideration when interpreting results. In the case of paired VH/VL paired sequences, the considered threshold for CDR-H3 must be lower than the used for CDR-L3, since the chances of shared CDR-H3 and is so much lower than that of a CDR-L3. Some samples may

show low-level overlapping CDR-H3s simply from being analyzed on the same sequencing run, for example, as a result of index hopping during Illumina sequencing. If contamination across samples from different species is observed (e.g., human/rhesus), a V(D)J gene annotation tool (e.g., IgBLAST) with a search database that includes both species can accurately reflect the level of cross-contamination and help with robust identification of the source of contamination. Similar approaches could be used with cell barcodes or index barcodes, provided that sufficient barcode space is available and rare or no reuse of index barcodes for a substantial period of time to enable robust tracking and sequence overlap analysis.

4 Notes

1. Technical replicate analyses could also include TR gamma/delta data if desired.
2. For <50 million cells/mL, use a volume of 0.25 mL.
3. Avoid generating bubbles that can clog the column.
4. The flow through contains the CD27- B cells.
5. While plunging, be careful with the collection tube. If excess pressure is applied, the solution in the collection tube can splash or spill out.
6. It is recommended to prepare a treatment condition and a negative control. Here, the treatment will include the CD27+ cells (from **step 28** of Subheading 3.1), 3T3-CD40L cells, and IL-2 and IL-21. The negative control will include all the mentioned components, except the CD27+ cells.
7. CD27+ B cells will be cocultured with irradiated 3T3-CD40L fibroblast cells that secrete CD40L to aid B-cell expansion.
8. Stimulated cells will assemble into visible clumps or grape-like clusters, which can be dissociated gently via mixing with a pipet prior to single-cell experimental analyses.
9. Please wear proper PPE when removing the cryovials from the liquid nitrogen freezer (safety eye goggles, cryogenic gloves, lab coat, and closed toe shoes), and wear appropriate PPE for all manipulations with human source material (eye protection, lab coat, and gloves).
10. Do not add human T-Activator CD3/CD28/CD137 Dynabeads™ again during the feeding unless the restimulation is needed.
11. We typically observe the initial sign of T cell expansion around day 5 or 6 when cells exhibit an increase in size and irregular shape. Expanded cells typically begin to show slower growth around day 8–10.

12. We usually obtain >10 million T cells for split replicate TCR sequencing analysis, with an initial pre-expansion cell number of around 0.1 million T cells.
13. For a hypothetical pairing precision analysis using the values identified in **step 2e** of Subheading 3.3:
    (a) Total CDR-H3 sequences overlapping in both replicates: 1000.
    (b) CDR-H3 observed to be exactly matched with the same CDR-L3s in R1 and R2 (TP): 950.
    (c) CDR-H3 observed to be paired with CDR-L3's of different lengths in R1 or R2 (FP): 30.
    (d) CDR-H3 with matched length, but an inexact CDR-L3 nt sequence match in R1 or R2: 20.

    Analyzing the 20 mismatched light chain sequences using a Hamming distance formula:

    (a) 15 CDRL-L3 were equal or less than 20% different in R1 and R2 by nt Hamming distance (i.e., assume TP).
    (b) 5 CDRL-L3 were more than 20% different in R1 or R2 by nt Hamming distance (i.e., assume FP).

    Total number of TP: 950 + 15 = 965.
    Total number of FP: 30 + 5 = 35.

$$P = \sqrt{\frac{965}{(965 + 35)}} = 98.2\%$$

14. If single cells are prepared using a flow focusing technology, pairing precision for heavy and light chains is normally around 98%, and between 96 and 99% can be expected [2, 6, 15].
15. In a high-performance computing platform, the analysis of ~1 million sequences requires around 10 min.

Acknowledgments

We thank Susanna Marquez, Eline T. Luning Prak, Chaim Schramm, and Ulrik Stervbo for assistance with the manuscript and David Price and Daniel Douek for scientific guidance. This work was supported by the University of Kansas Departments of Pharmaceutical Chemistry and Chemical Engineering, the KU Cancer Center, the US Department of Defense W81XWH1810296, and by NIH grants DP5OD023118, P20GM103418, R21AI143407, and R21AI144408.

References

1. Setliff I, Shiakolas AR, Pilewski KA, Murji AA, Mapengo RE, Janowska K et al (2019) High-throughput mapping of B cell receptor sequences to antigen specificity. Cell 179: 1636–1646 e15
2. DeKosky BJ, Kojima T, Rodin A, Charab W, Ippolito GC, Ellington AD et al (2015) In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. Nat Med 21:86–91
3. Kwong PD, DeKosky BJ, Ulmer JB (2020) Antibody-guided structure-based vaccines. Semin Immunol 50:101428
4. Robbins PF, Morgan RA, Feldman SA, Yang JC, Sherry RM, Dudley ME et al (2011) Tumor regression in patients with metastatic synovial cell sarcoma and melanoma using genetically engineered lymphocytes reactive with NY-ESO-1. J Clin Oncol 29:917–924
5. Eggenhuizen PJ, Ng BH, Ooi JD (2020) Treg enhancing therapies to treat autoimmune diseases. Int J Mol Sci 21(19):7015
6. Lagerman CE, López Acevedo SN, Fahad AS, Hailemariam AT, Madan B, DeKosky BJ (2019) Ultrasonically-guided flow focusing generates precise emulsion droplets for high-throughput single cell analyses. J BiosciBioengineer 128:226–233
7. Recher M, Berglund LJ, Avery DT, Cowan MJ, Gennery AR, Smart J et al (2011) IL-21 is the primary common gamma chain-binding cytokine required for human B-cell differentiation in vivo. Blood 118:6824–6835
8. Doria-Rose NA, Schramm CA, Gorman J, Moore PL, Bhiman JN, DeKosky BJ et al (2014) Developmental pathway for potent V1V2-directed HIV-neutralizing antibodies. Nature 509:55–62
9. Spindler MJ, Nelson AL, Wagner EK, Oppermans N, Bridgeman JS, Heather JM et al (2020) Massively parallel interrogation and mining of natively paired human TCRalphabeta repertoires. Nat Biotechnol 38: 609–619
10. Mayassi T, Ladell K, Gudjonson H, McLaren JE, Shaw DG, Tran MT et al (2019) Chronic inflammation permanently reshapes tissue-resident immunity in celiac disease. Cell 176: 967–981 e19
11. Grifoni A, Weiskopf D, Ramirez SI, Mateus J, Dan JM, Moderbacher CR et al (2020) Targets of T cell responses to SARS-CoV-2 coronavirus in humans with COVID-19 disease and unexposed individuals. Cell 181:1489–1501 e15
12. Saha S, Raghava GP (2006) AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. Nucleic Acids Res 34:W202–W209
13. Briney B, Inderbitzin A, Joyce C, Burton DR (2019) Commonality despite exceptional diversity in the baseline human antibody repertoire. Nature 566:393–397
14. Collins AM, Watson CT (2018) Immunoglobulin light chain gene rearrangements, receptor editing and the development of a self-tolerant antibody repertoire. Front Immunol 9:2249
15. McDaniel JR, DeKosky BJ, Tanno H, Ellington AD, Georgiou G (2016) Ultra-high-throughput sequencing of the immune receptor repertoire from millions of lymphocytes. Nat Protoc 11:429–442

Chapter 22

Immune Repertoire Analysis on High-Performance Computing Using VDJServer V1: A Method by the AIRR Community

Scott Christley, Ulrik Stervbo, and Lindsay G. Cowell and on behalf of the AIRR Community

Abstract

AIRR-seq data sets are usually large and require specialized analysis methods and software tools. A typical Illumina MiSeq sequencing run generates 20–30 million 2 × 300 bp paired-end sequence reads, which roughly corresponds to 15 GB of sequence data to be processed. Other platforms like NextSeq, which is useful in projects where the full V gene is not needed, create about 400 million 2 × 150 bp paired-end reads. Because of the size of the data sets, the analysis can be computationally expensive, particularly the early analysis steps like preprocessing and gene annotation that process the majority of the sequence data. A standard desktop PC may take 3–5 days of constant processing for a single MiSeq run, so dedicated high-performance computational resources may be required.

VDJServer provides free access to high-performance computing (HPC) at the Texas Advanced Computing Center (TACC) through a graphical user interface (Christley et al. Front Immunol 9:976, 2018). VDJServer is a cloud-based analysis portal for immune repertoire sequence data that provides access to a suite of tools for a complete analysis workflow, including modules for preprocessing and quality control of sequence reads, V(D)J gene assignment, repertoire characterization, and repertoire comparison. Furthermore, VDJServer has parallelized execution for tools such as IgBLAST, so more compute resources are utilized as the size of the input data grows. Analysis that takes days on a desktop PC might take only a few hours on VDJServer. VDJServer is a free, publicly available, and open-source licensed resource. Here, we describe the workflow for performing immune repertoire analysis on VDJServer's high-performance computing.

Key words AIRR-Seq, B-cell receptor, T-cell receptor, High-performance computing, Cloud computing

1 Introduction

Immune repertoire sequencing produces large, highly complex data sets that require specialized analysis methods and software tools. We developed VDJServer to address critical barriers in broader adoption of immune repertoire sequencing, namely, the

Anton W. Langerak (ed.), *Immunogenetics: Methods and Protocols*, Methods in Molecular Biology, vol. 2453, https://doi.org/10.1007/978-1-0716-2115-8_22,

lack of a complete, start-to-finish analysis pipeline, the lack of a data management infrastructure, and limited access for many researchers to high-performance computing (HPC) resources. VDJServer fills these gaps, specifically providing (1) an open suite of interoperable repertoire analysis tools that allows users to upload a set of sequences and pass them through a seamless workflow that executes all steps in an analysis, (2) access to sophisticated analysis tools running in an HPC environment, (3) interactive visualization capabilities for exploratory analysis, (4) a data management infrastructure, and (5) a graphical user interface to facilitate use by experimental and clinical research groups that lack extensive bioinformatics expertise.

Here, we describe the workflow for performing immune repertoire analysis on VDJServer's high-performance computing. The major steps of the workflow include creating a project to hold sequencing data and analysis results, uploading and preparing immune repertoire sequencing files, preprocessing the raw sequence data, performing V(D)J assignment and annotation of the processed sequences, defining study metadata and analysis comparison groups, performing repertoire characterization and comparison, and visualizing and downloading analysis results.

2 Materials

VDJServer requires a user account with a valid email address to access the system. Creating an account is free, as well as using the VDJServer resources. Accounts are used to insure data and results are private and secure. Create an account at https://vdjserver.org to get started. Contact VDJServer with any questions or concerns by using the *Feedback* option on the website or send email to vjdserver@utsouthwestern.edu.

3 Methods

For researchers without access to high-performance computing (HPC), VDJServer provides free access to the Texas Advanced Computing Center (TACC) through a standard web browser via a graphical user interface [1]. A suite of tools for a complete analysis workflow are provided, including modules for preprocessing and quality control of sequence reads, V(D)J gene assignment, repertoire characterization, and repertoire comparison (Fig. 1). VDJServer incorporates analysis software from the Immcantation suite [2, 3], VDJPipe [4], and other interoperability tools [5, 6]. Germline gene sets for human and mouse are derived from IMGT [7], and a draft germline set for Indian origin rhesus macaque IG is also provided [8]. VDJServer provides the Community Data Portal for

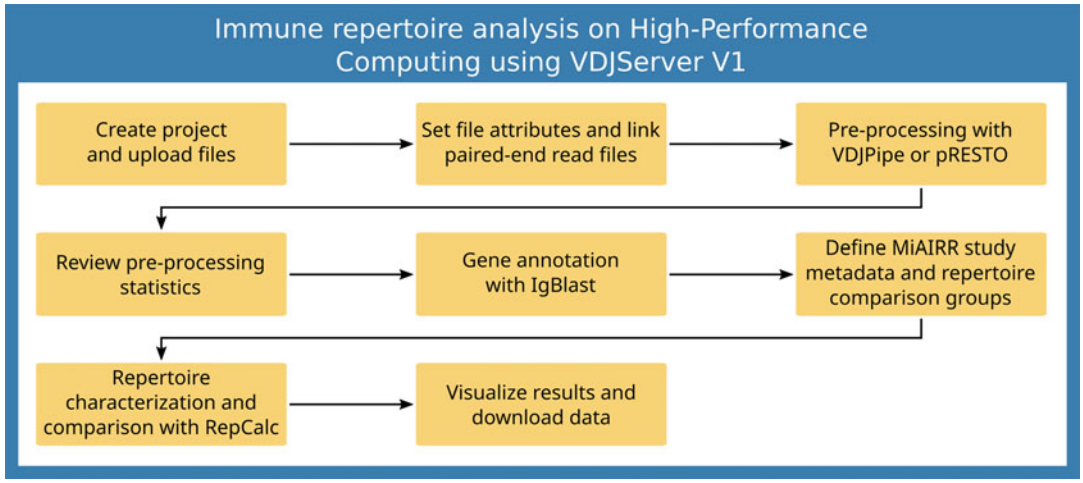


Fig. 1 Workflow immune repertoire analysis on high-performance computing using VDJServer V1

publicly sharing data and analysis results, and studies can be published to the AIRR Data Commons [9], which is not covered in this workflow. To publicly share data, please see the AIRR Community method chapter entitled, "Data Sharing and Re-Use." Here, we discuss the different steps in the immune repertoire analysis workflow using VDJServer.

3.1 Create Project

After login with a user account, click on the *Add Project* button to create a new project and give the project a name. Each project is a logical container for files, jobs, analysis results, and visualizations, and any number of projects may be created. All data in a project is private to the user account but can be shared with other VDJServer users by adding them on the *Project Settings* page (*see* **Note 1**).

3.2 Upload Files into Project

From the *Upload and Browse Project Data* page, click on the *Upload* button and select files from the local computer, from Dropbox, or from a URL (ftp/http), to be uploaded; multiple files can be selected (*see* **Note 2**). Click the *Start* button to start uploading. Upload FASTQ sequence read files (compress with gzip for faster upload), FASTA files with barcode and primer sequences, TSV containing metadata, or any file to be associated with the project.

3.3 Set File Attributes and Link Paired-End Read Files

VDJServer attempts to detect the file type (FASTQ, FASTA, AIRR TSV, etc.) from the file extension, but this can be changed with the *File Type* setting. Use *Barcode* or *Primer* for files containing those sequences. For paired-end read sequencing files, set the *Read Direction* on each file to either the Forward or Reverse orientation, and then link the two files together on the *Link Paired Read Files* page. The forward orientation refers to the V gene end of the template, and reverse orientation refers to the J gene or constant

region end of the template. Correct orientation is necessary for proper matching of barcodes and primers. Linked files will show together as a pair on the *Upload and Browse Project Data* page.

3.4 Preprocessing with VDJPipe or pRESTO

From the *Upload and Browse Project Data* page, select sequence read files for preprocessing by clicking on the checkbox next to each file. Click on the *Run Job* button and select either VDJPipe or pRESTO; a job submission screen will be displayed. VDJPipe and pRESTO have similar capabilities; pRESTO should be used for UMI; otherwise, VDJPipe is significantly faster (up to 20×) on larger data sets. A single workflow is available for pRESTO, while VDJPipe offers a number of customized workflows. VDJPipe's single function workflows perform individual preprocessing steps, while the complete workflow performs all steps. If unsure about filtering parameters to use for preprocessing, such as length or quality settings, it is useful to run VDJPipe's *Sequence Statistics* workflow. This will visualize length, quality, and nucleotide distributions of the read data. The job submission screen will provide parameters, with default values, that can be changed for the individual preprocessing steps (*see* **Note 3**). Finally, click the *Launch Job* button to submit the preprocessing job to the TACC supercomputer. The user will receive an email when the job is finished.

3.5 Review Preprocessing Statistics

When the preprocessing job is complete, the job on the *View Analyses and Results* page will change from an *In Progress* label to a *View Output* button. Click the button to show the *View Output* page, which has three main sections: job output files, analysis charts, and log files. Job output files provides a list of output files generated from the preprocessing job. Analysis Charts provides visualizations for pre- and post-filtering statistics, and log files are job error logs and workflow provenance metadata. The provided visualizations include:

1. Nucleotide composition for each read position.
2. GC% histogram.
3. Sequence length histogram.
4. Mean quality score histogram.
5. Quality score distribution for each read position.

Use the Analysis Charts to review the preprocessing results; they show the pre- and post-filtering statistics to understand how preprocessing has affected the data. If preprocessing removed too many reads, or alternatively has not removed enough, then a new preprocessing job should be run with looser or more stringent parameters. Among the job, log files are summary logs that will give information about the number of reads processed during each preprocessing step.

3.6 Make Job Output Files Available in Project Data Area

Once satisfied with the preprocessing results, the appropriate job output files need to be made available in the project data area so that they can be selected as input for additional analysis jobs. This can be done in two ways. The first is on the *View Analyses and Results* page. Click the *Job Actions* button for the job and select *Include Job Output*; this will make all output files available. Conversely, select *Exclude Job Output* from the *Job Actions* button, which will remove all output files for the project data area. Alternatively, the second way, the user can make individual job output files available from the *View Output* page for the job by clicking the *Make Available in Project Data Area* button next to each file. Clicking that button again will remove the file from the project data area. Job output files available in the project data area will show in their own section on the *Upload and Browse Project Data* page, grouped together by the job with the job name as a title.

3.7 Gene Annotation with IgBLAST

Select files for IgBLAST processing, either job output files or uploaded FASTA files, on the *Upload and Browse Project Data* page by clicking on the checkbox next to each file (*see* **Note 4**). Click on the *Run Job* button and select IgBLAST; a job submission screen will be displayed. Select the organism species (human, mouse, or rhesus macaque), the strain (if appropriate), and the sequence type (IG or TR). VDJServer maintains separate germline databases, so processing multiple sequence types and/or organism species requires running multiple IgBLAST jobs. Finally, click the *Launch Job* button to submit the preprocessing job to the TACC supercomputer. The user will receive an email when the job is finished.

As with all analysis jobs on VDJServer, job status is shown on the *View Analyses and Results* page, and the job output is available with the *View Output* button. Multiple output formats are provided including VDJServer's custom RepSum TSV, VDML, Change-O TSV, and AIRR TSV. Individual files can be downloaded by clicking on the filename, or all output files can be downloaded by clicking on *Archive of Output Files* in the log file section. It is recommended that AIRR TSV files are used for any custom analysis as they contain the most comprehensive annotations, and they are interoperable with many AIRR-seq tools.

3.8 Define MiAIRR Study Metadata and Repertoire Comparison Groups

By this point in the workflow, raw sequence data has been preprocessed, and sequences have been annotated. However, to achieve the greatest utility of repertoire analysis, it is recommended that metadata is entered and comparison groups are defined, though it is not strictly necessary as individual files can be analyzed in isolation. Entering metadata also has the benefit of providing MiAIRR compliance when it's time to publish the study. Metadata is entered on the *Metadata Entry* page and consists of the six MiAIRR components: study, subject, diagnosis, sample, cell processing, and nucleic

acid processing. VDJServer adds a seventh component with sample groups for doing group comparisons. Metadata can be manually entered on the page, but it is typically more efficient to prepare the metadata in a separate spreadsheet file, then import that spreadsheet into VDJServer (*see* **Note 5**). To do this, go to the appropriate section on the *Metadata Entry* page, click on the *Metadata Actions* button, and select *Export to File*. Open the spreadsheet file in Excel or another program, use one row for each entry, and fill in the values for each column. Save the file as Tab-delimited Text and upload the file into the project. Finally, on the *Metadata Entry* page, click on the *Metadata Actions* button and select *Import From File*. A panel will be shown where the user can pick the file to import and choose to either replace or append the current metadata.

Sample groups are a specialized feature of VDJServer that allows sample repertoires to be grouped together for performing intragroup and intergroup comparisons. Sample groups are defined by using one or more grouping operations. These grouping operations include:

1. Grouping by the values of a study metadata field. The *Group By* option provides a popup list of all the possible fields. VDJServer will determine all of the values for that field among the study metadata and create a separate group for each value.
2. Grouping by a logical operation. The *Logical* option defines a simple Boolean expression. Sample repertoires where that expression evaluates as true will be included in the sample group, while those for which the expression evaluates as false will be excluded. Pick the study metadata field from the popup list of available fields, pick the comparison operator, and provide a value.
3. Individual samples can be picked. The *Repertoires* option provides a popup list of sample repertoires. By default, when no sample repertoires are selected, then all sample repertoires are included in the sample group. Click on specific sample repertoires in the list to include them in the sample group.
4. The three grouping operations can be combined together. A sample repertoire needs to satisfy all applicable grouping operations to be included in the sample group.

3.9 Repertoire Characterization and Comparison with RepCalc

RepCalc performs a wide variety of analysis functions including clonal assignment, gene usage, gene combination usage, CDR3 length distribution and amino acid properties, CDR3 and clonal sharing and uniqueness, clonal abundance, diversity profile, and B cell–specific mutation analysis and clonal lineage. RepCalc uses a combination of tools to perform the analyses including VDJServer's custom repertoire summarization and Change-O, Alakazam, and SHazaM from the Immcantation suite.

To run RepCalc, no files need to be selected on the *Upload and Browse Project Data* page; instead, RepCalc will directly access the appropriate output files from a previous IgBLAST job. Click the *Run Job* button and select RepCalc; a job submission screen will be displayed. Pick the IgBLAST job to use as input. If study metadata was defined, the screen will indicate its availability and automatically perform group comparison; otherwise, RepCalc will only perform analysis on individual files. Change the default values to include or exclude specific analysis functions. Finally, click the *Launch Job* button to submit the job to the TACC supercomputer. The user will receive an email when the job is finished.

3.10 Visualize Analysis Results and Download Data

When the RepCalc job has completed successfully, click the *View Output* button on the *View Analyses and Results* page to display the analysis results. For RepCalc jobs, the *View Output* page has three main sections: job output files, analysis charts, and log files. Job output files provides a list of clonal assignment output files, Analysis charts provide analysis visualizations, and log files are job error logs and workflow provenance metadata. RepCalc produces a set of interactive analysis charts:

1. Absolute and relative gene usage.
2. Nucleotide and amino acid CDR3 length distribution.
3. Clonal abundance and cumulative abundance.
4. Diversity profile curves.
5. Selection pressure quantification.

Each chart provides three pop-up lists for selecting files, sample repertoires, or sample groups to be displayed on the chart. Chart figures can be downloaded by clicking on the *Download Chart* button, which will generate a figure identical to the chart being displayed in the browser, and the data for the chart can be downloaded by clicking on the *Download Data* button. Not all analysis output has an associated visualization but can be downloaded by clicking on *Archive of Output Files* in the log file section, with the data provided in TSV format for easy import into Excel and other tools.

4 Notes

1. Processing on shared facilities external to the user's institution raises privacy concerns. We recommend that data is fully anonymized before analysis begins.
2. Uploading a large number (hundreds) of files at one time is susceptible to network errors and timeouts that may prevent a file or two from not being uploaded. Be sure to check the total file count to insure it matches. If files are missing, use the *Search*

field to narrow the list and verify. Another technique is to upload the files in batches, e.g., 20 files at a time, and check after each batch that all the files got uploaded.

3. Steps that will not be used, as indicated by a red warning box, should be removed by clicking on the red X button.
4. With many files, use the *Search* field to restrict the files shown to just the desired files, and then click *Select All* to select them all at once.
5. While VDJServer V1 collects MiAIRR study metadata, it does not yet utilize the AIRR Standards Repertoire metadata format for import/export interoperability. However, VDJServer V2 will directly utilize the AIRR Data Model and AIRR Standards data formats. Conversion scripts exist for converting VDJServer V1 into VDJServer V2 metadata so no data will be lost or require reentry.

References

1. Christley S, Scarborough W, Salinas E, Rounds WH, Toby IT, Fonner JM et al (2018) VDJServer: a cloud-based analysis portal and data commons for immune repertoire sequences and rearrangements. Front Immunol 9:976
2. Gupta NT, Vander Heiden JA, Uduman M, Gadala-Maria D, Yaari G, Kleinstein SH (2015) Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. Bioinformatics 31:3356–3358
3. Vander Heiden JA, Yaari G, Uduman M, Stern JN, O'Connor KC, Hafler DA et al (2014) pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. Bioinformatics 30: 1930–1932
4. Christley S, Levin MK, Toby IT, Fonner JM, Monson NL, Rounds WH et al (2017) VDJPipe: a pipelined tool for pre-processing immune repertoire sequencing data. BMC Bioinformatics 18:448
5. Toby IT, Levin MK, Salinas EA, Christley S, Bhattacharya S, Breden F et al (2016) VDJML: a file format with tools for capturing the results of inferring immune receptor rearrangements. BMC Bioinformatics 17:333
6. Vander Heiden JA, Marquez S, Marthandan N, Bukhari SAC, Busse CE, Corrie B et al (2018) AIRR community standardized representations for annotated immune repertoires. Front Immunol 9:2206
7. Giudicelli V, Brochet X, Lefranc MP (2011) IMGT/V-QUEST: IMGT standardized analysis of the immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences. Cold Spring Harb Protoc 2011:695–715
8. Cottrell CA, van Schooten J, Bowman CA, Yuan M, Oyen D, Shin M et al (2020) Mapping the immunogenic landscape of near-native HIV-1 envelope trimers in non-human primates. PLoS Pathog 16:e1008753
9. Christley S, Aguiar A, Blanck G, Breden F, Bukhari SAC, Busse CE et al (2020) The ADC API: a web API for the programmatic query of the AIRR data commons. Frontiers Big Data 3:22

Chapter 23

Data Sharing and Reuse: A Method by the AIRR Community

Brian D. Corrie, Scott Christley, Christian E. Busse, Lindsay G. Cowell, Kira C. M. Neller, Florian Rubelt, and Nicholas Schwab and on behalf of the AIRR Community

Abstract

High-throughput sequencing of adaptive immune receptor repertoires (AIRR, i.e., IG and TR) has revolutionized the ability to study the adaptive immune response via large-scale experiments. Since 2009, AIRR sequencing (AIRR-seq) has been widely applied to survey the immune state of individuals (see "The AIRR Community Guide to Repertoire Analysis" chapter for details). One of the goals of the AIRR Community is to make the resulting AIRR-seq data FAIR (Findable, Accessible, Interoperable, and Reusable) (Wilkinson et al. Sci Data 3:1–9, 2016), with a primary goal of making it easy for the research community to reuse AIRR-seq data (Breden et al. Front Immunol 8:1418, 2017; Scott and Breden. Curr Opin Syst Biol 24:71–77, 2020). The basis for this is the MiAIRR data standard (Rubelt et al. Nat Immunol 18:1274–1278, 2017). For long-term preservation, it is recommended that researchers store their sequence read data in an INSDC repository. At the same time, the AIRR Community has established the AIRR Data Commons (Christley et al. Front Big Data 3:22, 2020), a distributed set of AIRR-compliant repositories that store the critically important annotated AIRR-seq data based on the MiAIRR standard, making the data findable, interoperable, and, because the data are annotated, more valuable in its reuse. Here, we build on the other AIRR Community chapters and illustrate how these principles and standards can be incorporated into AIRR-seq data analysis workflows. We discuss the importance of careful curation of metadata to ensure reproducibility and facilitate data sharing and reuse, and we illustrate how data can be shared via the AIRR Data Commons.

Key words AIRR-seq, B-cell receptor, Immunoglobulin, T-cell receptor, FAIR data, Data sharing, Data reuse

1 Introduction

Once an adaptive immune receptor repertoire sequencing (AIRR-seq, *see* Table 1, of the "AIRR Community Guide to TR and IG Gene Annotation" chapter for a glossary of terms) experiment has been successfully designed and carried out (*see* the "AIRR

Brian D. Corrie and Scott Christley shared first author.

Anton W. Langerak (ed.), *Immunogenetics: Methods and Protocols*, Methods in Molecular Biology, vol. 2453, https://doi.org/10.1007/978-1-0716-2115-8_23, © The Author(s) 2022

Table 1
AIRR-seq-related data repository resources

Software	Notes/description	URL	
AIRR-compliant resources			
iReceptor	A web based user interface for the discovery, analysis and download of data from the AIRR data commons	https://gateway.ireceptor.org	[7]
iReceptor turnkey	An open source, easy to install software stack for operating an ADC compliant repository	https://github.com/sfu-ireceptor/turnkey-service-php	[7]
VDJServer	A free, scalable resource for performing immune repertoire analysis and sharing data	https://vdjserver.org	[8]
Other AIRR-seq-related resources			
OAS	OAS: A public database for AIRR-seq data	http://opig.stats.ox.ac.uk/webapps/oas/	[21]
IEDB	Catalog of experimental data on antibody and T-cell epitopes across multiple species and conditions	https://www.iedb.org	[26]
ImmuneDB	Database and analysis tool for large amounts of AIRR-seq data	https://immunedb.readthedocs.io/en/latest/	[9]
McPAS-TCR	A manually curated catalogue of TR sequences associated with pathology	http://friedmanlab.weizmann.ac.il/McPAS-TCR/	[27]
PIRD	A collection of raw and processed AIRR-seq data from human and other vertebrate species with different phenotypes	https://db.cngb.org/pird/home/	[22]
SystimsDB	Preprocessed and annotated IG and TR sequences from the scientific literature	https://www.systimsdb.ethz.ch/index.html	
TBAdb	Antigen-specific TRs from PIRD	https://db.cngb.org/pird/tbadb/	[22]
TCRdb	Curated TR repertoires from various conditions	http://bioinfo.life.hust.edu.cn/TCRdb/	[23]
VDJdb	A curated database of TR sequences with known antigen specificities	https://vdjdb.cdr3.net	[28]

Community Guide to Planning and Performing AIRR-seq Experiments"chapter) and the data have been processed and analyzed for the experimental purpose of the study (see the "AIRR Community Guide to Repertoire Analysis" chapter), it is necessary to consider how to report on and share the AIRR-seq data from that study according to FAIR data principles. The FAIR data principles, which state that data should be Findable, Accessible, Interoperable, and Reusable[1], provide a number of benefits to the research

community and to individual researchers. The principles ensure that the generated data can easily be reused within the laboratory that generated the data, thus maximizing the potential of the data for that lab. Externally, the principles can increase the visibility and recognition of the work and thereby attract new partnerships within the research community and with policy makers (*see* **Note 1**). In the scientific community at large, the FAIR data principles support scientific transparency and reproducibility, increasing the rigor of scientific results. Additionally, they facilitate data reuse for the exploration of new questions, particularly for questions that benefit from the ability to integrate data originally generated in different studies.

1.1 Experimental Reporting: Minimal Information Standards

The primary purpose of the AIRR Community-endorsed MiAIRR Standard [4] is to establish a community-based standard for the recording and reporting of experimental results involving AIRR-seq data. The MiAIRR paper states that such a standard is considered necessary "... for the interpretation and comparison of AIRR-seq experiments conducted by different groups" and is a critical component of making AIRR-seq "interoperable" and "reusable" (the "I" and "R" in FAIR) (*see* **Note 2**). It is also critical for scientific transparency and reproducibility. The MiAIRR standard covers six high-level sets of data, including recommendations on how to capture data and metadata at the following levels: (1) study/subject/diagnosis, (2) sample collection, (3) sample processing and sequencing, (4) raw sequences, (5) data processing, and (6) sequence annotations. At each level, there are a set of metadata fields that are recommended for consideration when designing AIRR-seq studies and curating the metadata during the performance and reporting of such a study.

1.2 Data Sharing: Data Formats

Although minimal information standards are necessary, they are not sufficient to completely enable data sharing, interoperability, and reuse. Subsequent to the establishment of the MiAIRR Standards, the AIRR Community established a set of computable specifications and accompanying file formats that facilitate sharing of data and support analysis tool interoperability. This includes a file format for AIRR-seq rearrangement data [6] and a file format for study and repertoire metadata [5]. In addition, the AIRR Community has established a software certification process that provides a "badge"-based system for tool developers to certify that their software supports the AIRR Standards as specified on the AIRR Software Compliance web page (https://docs.airr-community.org/en/stable/swtools/airr_swtools_standard.html).

1.3 Data Sharing: AIRR Data Commons

Critical to the concept of FAIR is making AIRR-seq data "Findable" and "Accessible." The AIRR Community has established the AIRR Data Commons (ADC) [5], a network of geographically

distributed AIRR-compliant repositories that adhere to the AIRR Standards. Of particular importance in creating the ADC is the establishment of the AIRR Data Commons web API (ADC API) [5] for finding, querying, and exploring data in AIRR-compliant repositories. The ADC API is a web-based query API that makes AIRR-seq studies and their associated annotated sequence data findable and accessible (the "F" and "A" in FAIR) (*see* **Note 3** in regard to challenges in data discovery). Because the ADC API utilizes the MiAIRR data standard and AIRR file formats, the ADC also promotes and facilitates interoperability and data reuse (the "I" and "R" in FAIR), thereby supporting reproducibility, data integration, and meta-analysis. The ADC has grown an order of magnitude since 2018, from just under 400 million annotated adaptive immune receptor rearrangements in late 2018 to its current size of five distributed repositories with over 60 studies, 6000 repertoires, and 4 billion rearrangements available for data exploration and download. Of the five distributed repositories, there are two community repositories in Canada (the iReceptor Public Archive (IPA) and iReceptor COVID-19 repositories managed by iReceptor [7]), one community repository in the United States (managed by VDJServer [8]), and two research group-specific repositories: the i3 AIRR repository at Sorbonne University in France and the VDJBase AIRR-seq repository at Bar Ilan University in Israel. We expect two new repositories to be added to the ADC in the near future, with ImmuneDB [9] at the University of Haifa, Israel, and sciReptor [10] at DKFZ in Germany working on implementations of the ADC API for their respective repositories. The ADC can be searched using CURL at the command line or interactively using a web user interface through the iReceptor Gateway [7]. In the near future, interactive search through the VDJServer web interface will also be possible [8].

As a result of the COVID-19 pandemic, the AIRR Community made a call for open data [11]. In response to this call, the iReceptor and VDJServer groups have collaborated with a number of researchers to curate publicly available COVID-19 AIRR-seq data in the ADC [3]. As of the first quarter of 2021, there are 13 studies, over 3500 repertoires, and over 1 billion rearrangements from COVID-19 studies available in the ADC.

2 Materials

2.1 Study Design and Reproducibility

A diverse range of clinically important conditions—including infections, vaccinations, autoimmune diseases, transplants, transfusion reactions, aging, and cancers—can lead to unique, measurable AIRR responses [12]. Therefore, AIRR-seq not only can be used to advance our understanding of disease and how the immune system responds but also can provide a unique opportunity for

diagnostic and prognostic approaches. Analysis of AIRR-seq data provides the opportunity for advancing personalized medicine in the form of a highly multiplexed diagnostic tool, with the potential of a near-universal blood test. Sample processing, sequencing methodology, and bioinformatic analysis are all critical to generating reliable and meaningful data. However, bioinformatics is the key component when using AIRR-seq to approach clinical questions. While powerful enough for the current era of personalized immuno-medicine, data science, and data-driven patient care, there is a particularly high need for standardization and reproducibility. Even though the first AIRR-based applications are already in clinical use in, e.g., leukemia and COVID-19 [13–15], many challenges remain before AIRR-seq-based blood testing can become a useful component in daily clinical practice. Many aspects in addition to the sample processing (please refer to the "AIRR Community Guide to TR and IG Gene Annotation," "AIRR Community Guide to Planning and Performing AIRR-Seq Experiments," and "AIRR Community Guide to Repertoire Analysis" chapters for details) need to be considered.

First, immune responses (particularly in investigations of infections or vaccine responses) follow very time-sensitive kinetics, with initial responses detectable within days and durations of weeks up to life-long. Similarly, sample choice including both tissue sources (e.g., peripheral blood or tumor tissue) and cell subset (e.g., memory B cells, effector T cells) will heavily influence results. In all cases, every aspect of any AIRR-seq based approach for diagnostic purposes will have to be rigorously validated, as potential regulatory approval will be contingent upon successful validation. Finally, obtaining meaningful data at scale requires a sample set with detailed clinical annotations, while still maintaining adherence to patient privacy directives and related legislation. Designing an AIRR-seq-based study with these key points in mind will ultimately determine the range of conclusions which can be obtained from it. The power of bioinformatic analysis depends not only on high quality sample processing and sequencing methods (see the "AIRR Community Guide to Planning and Performing AIRR-Seq Experiments" chapter) but also equally on careful study design.

In addition to validation for clinical uses, ensuring that bioinformatic analyses are reproducible is important for all research. This can be challenging, as analysis of AIRR-seq data is lengthy and requires the use of specialized software, with settings that can be project-specific and reference germlines that change in time, as new alleles are discovered [2]. To guarantee the reproducibility of results, we recommend to always record the versions of software and germlines used, as well as the arguments that were decisive to choose one setting or another in the metadata of an analysis such as an AIRR-compliant Repertoire file (https://docs.airr-community.org/en/stable/datarep/metadata.html). Analysis environments

such as VDJServer often capture analysis metadata automatically. For custom scripts, we recommend documenting the code and avoid creating new names for fields that are already described in the AIRR Community standards.

2.2 Software Tools

Many tools are available for AIRR-seq analysis [16–19]. Table 1 highlights several of the more commonly used programs that are free and open source and support standardized AIRR data representations, which facilitates data sharing via the ADC.

3 Methods

3.1 How to Share AIRR-Seq Data: General Information

The AIRR Community provides substantial documentation around the processes of sharing and finding AIRR-seq data. We summarize these processes below. For more detailed descriptions, please refer to the AIRR Community documentation web site (https://docs.airr-community.org/en/stable/standards/data_submission.html). Refer to **Notes 1** and **3** for the costs and benefits of sharing AIRR-seq data.

3.1.1 Curating Using MiAIRR

One of the most critical steps in sharing data is to ensure that studies are curated with appropriate terms as specified in the MiAIRR Standard. The AIRR Community has published the list of MiAIRR fields (https://docs.airr-community.org/en/stable/miairr/data_elements.html) with field definitions, types, and examples. In addition, the Center for Expanded Data Annotation and Retrieval (CEDAR, https://metadatacenter.org) project has created a MiAIRR-compliant web user interface (Cedar for AIRR or CAIRR [20] (https://docs.airr-community.org/en/stable/cairr/overview.html) for capturing and entering AIRR-seq study metadata. Projects such as iReceptor [7] and VDJServer [8] provide tools for MiAIRR metadata curation as part of their platforms, ranging from template metadata spreadsheets to web interfaces for capturing MiAIRR study metadata.

3.1.2 Storing AIRR-Compliant Data in INSDC Repositories

In order to promote Open Science, the AIRR Community recommends that the source sequence data from studies be stored in a sustainable repository such as those maintained by the International Nucleotide Database Collaboration (INSDC) (e.g., NCBI SRA or EBI's ENA). The AIRR Community has collaborated with NCBI to create a protocol for storing MiAIRR compliant study metadata in the NCBI resources [4]. This protocol maps MiAIRR field names to metadata in NCBI entities such as BioProject and BioSample. In addition, the CAIRR pipeline [20] supports and facilitates publication of MiAIRR compliant metadata to the NCBI.

3.1.3 Publishing Your Data in the AIRR Data Commons

The primary difference between data in the INSDC repositories and the data in the ADC is that the ADC repositories store data that has gone through quality control and annotation pipelines as described in the "AIRR Community Guide to Repertoire Analysis" and "AIRR Community Guide to TR and IG Gene Annotation" chapters (*see* **Note 5**). The annotation process compares the expressed sequences to a reference database (see the "AIRR Community Guide to TR and IG Gene Annotation" chapter) and identifies the most likely V, D, and J genes that contributed to the expressed genes. By sharing this processed data, other researchers avoid having to rerun these computationally complex and sometimes expensive annotation pipelines. Additionally, this facilitates querying of AIRR-seq data based on the annotations, such as querying for sequences that use a particular V gene or contain a particular CDR3 sequence.

Although there are a number of large repositories in the ADC that curate data from a broad range of research groups, there are also multiple mechanisms by which data generators can themselves publicly share their AIRR-seq data into the ADC. Researchers can (1) collaborate with one of the existing ADC repositories to publish their data, (2) self-publish data into those ADC repositories that provide such a service, (3) install and manage their own ADC repository, or (4) implement the ADC API against an existing repository making their own repository ADC-compliant. These options are described in more detail below.

Collaborating with an Existing ADC Repository Provider

A number of the large repository providers in the ADC (e.g., iReceptor, VDJServer) are community repositories that curate and store data on behalf of the community. Although it does not scale for these groups to curate and store data from the hundreds of AIRR-seq studies that are currently published each year, these large repository providers often collaborate with researchers to help them publish their AIRR-seq data. For example, in answer to the AIRR Community's call for sharing COVID-19 data, both the iReceptor and VDJServer repositories collaborated with a number of research groups to curate and share their COVID-19 studies [3]. Additionally, users who have analyzed their data using the VDJServer analysis portal can work with the VDJServer team to directly share their project data into the ADC.

Installing and Running Your Own ADC-Compliant Repository

Developing, installing, and managing an AIRR-seq repository to facilitate data sharing can be challenging, but for some groups, this may be the best option. For example, large research groups that manage and process their own data and have the bioinformatics and technical expertise to manage database platforms may want to manage their own repository. In addition, groups that want to

more closely manage the stewardship of their data (due to ethics requirements) may also want to operate their own AIRR-compliant repository.

To enable this, the iReceptor Project has developed a software stack called the iReceptor Turnkey [7] that is designed to make the download, installation, and management of an AIRR-compliant repository as straightforward as possible. The software is open source and uses container-based software management (Docker) to implement an AIRR-compliant database, a data curation service, and an ADC API web service for querying the database. As a result, a research group can download and install the software, curate their data, and easily have an AIRR-compliant repository that can participate as a member in the ADC. Such a repository would then automatically be searched by tools that search the ADC, such as the iReceptor Gateway. Currently, the iReceptor COVID-19, the i3 AIRR Sorbonne University repository, and the VDJBase Bar Ilan repository are all using the iReceptor Turnkey software as the platform for their ADC repositories. For more information on installing an iReceptor Turnkey, *see* Subheading 3.3.4.

Implementing the ADC API in an Existing Repository

Several groups have preexisting repositories that already contain AIRR-seq data, having been developed prior to or in parallel with the ADC [9, 10, 21–24]. In order to interoperate with repositories in the ADC, it is necessary to perform a data transformation to bring the data and metadata into compatible formats. Alternatively, a repository can implement the ADC API, thereby avoiding the need to transform data. Although implementing such an API is not a trivial task, if a research group has a significant investment in an existing repository technology and wants to add their data to the ADC, this is one practical option. The AIRR Community has developed a reference implementation for the ADC API (https://github.com/airr-community/adc-api). This implementation provides a JavaScript-based ADC API web implementation that performs simple queries against an AIRR-compliant MongoDB repository. This provides a framework for implementing the ADC API against an existing repository. In addition, the AIRR Community provides an extensive suite of test queries (https://github.com/airr-community/adc-api-tests) to help implementers ensure that their ADC API implementation is compliant with the ADC API specification. Although nontrivial, we know through iReceptor and VDJServer that this approach works, and indeed a number of the repositories described above are currently working on implementations of the ADC API (e.g., ImmuneDB [9] and sciReptor [10]). We expect them to be searchable as part of the ADC in the near future.

3.1.4 Sharing Data Through a Non-ADC but AIRR-Compliant Repository

All ADC repositories are AIRR-compliant, but it is possible to be compliant with the AIRR formats and not be part of the ADC. The key difference between AIRR Standards-compliant repositories and those that are part of the ADC is that ADC repositories implement the ADC API for queries, while AIRR-compliant repositories do not. AIRR-compliant repositories store valuable and useful annotated AIRR-seq data in an AIRR-compliant format and use the AIRR Standard file formats for data exchange, but cannot be directly queried by external clients using the ADC API. There are a number of repositories of this type. For example, the Observed Antibody Space (OAS) [21] repository has over 1 billion annotated sequences from a number of AIRR-seq studies that can be queried and downloaded. Data from OAS is interoperable with data from repositories in the ADC, but ADC queries do not work against the OAS repository.

In addition, it is possible to download, install, and curate your own data into an AIRR-compliant repository. ImmuneDB (http://immunedb.com/) [9] allows the curation of AIRR-seq study data and the sharing of that data through a web-interface or password-protected mysql repository. Installation through the Docker image is simple, and once installed, ImmuneDB allows users to annotate raw data, load previously annotated data, and share that data using its web interface. ImmuneDB uses AIRR-compliant file formats for data exchange and is currently working on implementing the ADC API for queries. We expect these services to be provided as part of the ADC in the near future.

3.2 Finding Data in the AIRR Data Commons

There are two primary mechanisms for finding, downloading, analyzing, and re-using data in the ADC, that is using the ADC API or using a web-based user interface.

3.2.1 Using the ADC API

The ADC API [5] is the primary mechanism to search the ADC and is required for a repository to be part of the ADC. The ADC API specifies a rich query language that allows researchers to pose complex queries across all keywords defined in the MiAIRR data standard. The same query will work on all ADC-compliant repositories, providing a consistent mechanism to identify data sets of interest. Queries can be made against repertoire metadata at the study, subject, and sample level, including how the sample was obtained, prepared for sequencing, and processed after sequencing, via the ADC web API repertoire query end point. Queries can be made at the sequence annotation or rearrangement level, such as for V, D, or J gene annotations, via the ADC web API rearrangement query end point. Once a set of repertoires of interest are identified (e.g., all repertoires generated using primers that target IGH genes), it is possible to filter the rearrangements from those repertoires based on sequence annotation fields, such as for specific V gene calls or CDR3 amino acid sequences. Finally, once a data set

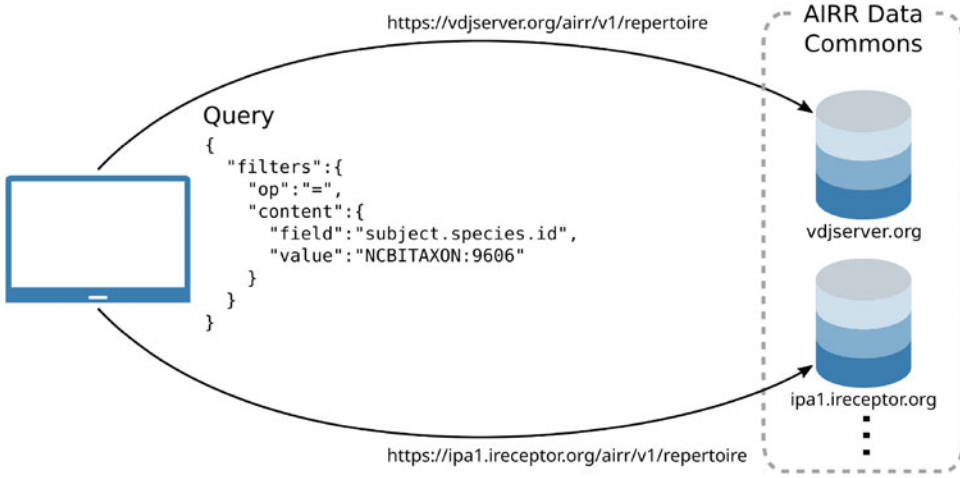


Fig. 1 Using the AIRR Data Commons API. Searching two repositories in the AIRR Data Commons for all human repertoires

is identified, it is possible to download that data in an AIRR-compliant file format.

An example query, which searches two of the repositories in the ADC for all human repertoires, is given in Fig. 1. The AIRR Community provides documentation for using the ADC API (https://docs.airr-community.org/en/stable/api/adc_api.html), including a number of additional example queries.

There are many repositories in the ADC, and when using the ADC API, it is necessary to query each one independently and federate the query results. As such, use of the ADC API is targeted at users who are comfortable writing code that uses web API queries. Subheading 3.4 contains examples on how to query VDJServer and iReceptor from the command line, python, and R.

3.2.2 Using a Web-Based User Interface

Web-based user interfaces (UIs) for the ADC are targeted at the more general AIRR-seq data user. User interfaces such as the iReceptor Gateway [7] are designed to hide the complexity of the fact that the user is querying multiple, international repositories. Web-based UIs typically implement a specific workflow, through web-based forms and menus, providing the user with the ability to issue complex queries across the entire ADC. These queries are usually targeted at a specific scientific use case and workflow, allowing the web-based UI to optimize the queries performed and to provide a simple UI for that specific use case.

For example, the iReceptor Gateway currently implements two data exploration workflows:

1. One that allows the user to generate complex queries that span the rich study, subject, sample, and processing metadata of the MiAIRR standard to find specific repertoires of interest across thousands of repertoires.

2. One that allows users to generate simple sequence annotation queries across gene and other sequence annotation fields to find specific sequences of interest across billions of sequences.

Queries are sent out by the iReceptor Gateway to each repository in the ADC, then results are federated and presented in a manner that helps the user find data of interest.

Typical workflows might be to iteratively search for data of interest from the entire ADC. For example, a researcher might start by limiting the data to all subjects that were diagnosed with COVID-19, then search for IG data sets (IGH, IGK, or IGL data), and finally refine that search to those data sets that only have paired IG heavy and light chain data (*see* **Note 6** for more discussion on finding data of interest). For this example, each step is accomplished by a few UI interactions (choosing menu items, typing in keywords), drilling down from over 6000 repertoires, 60 studies, and over 4 billion annotated sequences to the two such studies currently included in the ADC, which together comprise approximately 245,000 annotated sequences from 87 repertoires found in two international repositories. Once data of interest are found, the user can visualize a number of statistics on each repertoire (such as V, D, or J gene usage and CDR3 length distribution), search for a sequence annotation feature (such as search for a specific V, D, or J gene or CDR3), or request that the iReceptor Gateway federate and download the annotated sequence data and the repertoire metadata for further analysis and reuse (*see* **Note 7** for considerations when combining data from different studies). A screenshot of such a query, as implemented in the iReceptor Gateway, is given in Fig. 2.

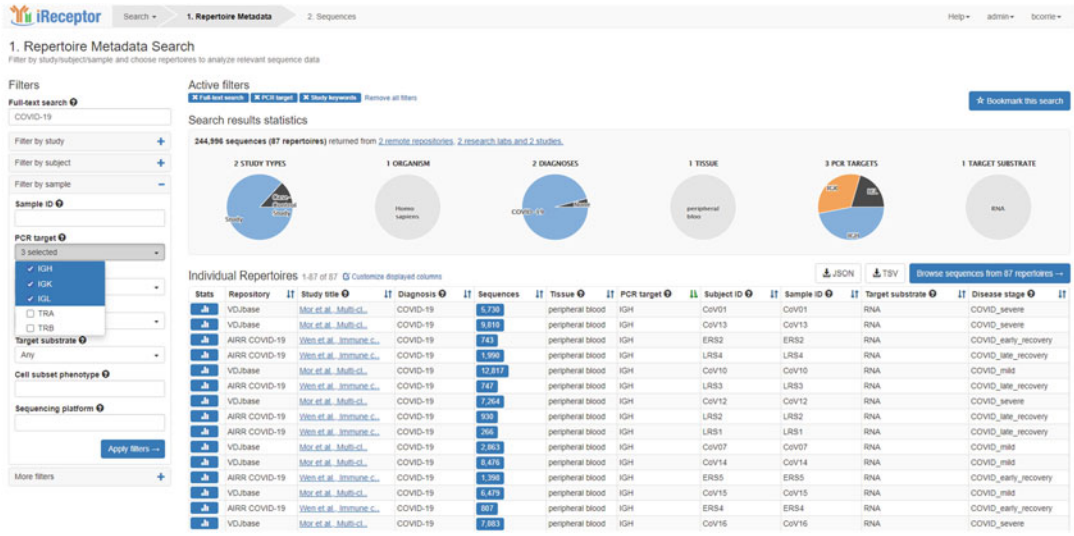


Fig. 2 iReceptor Gateway

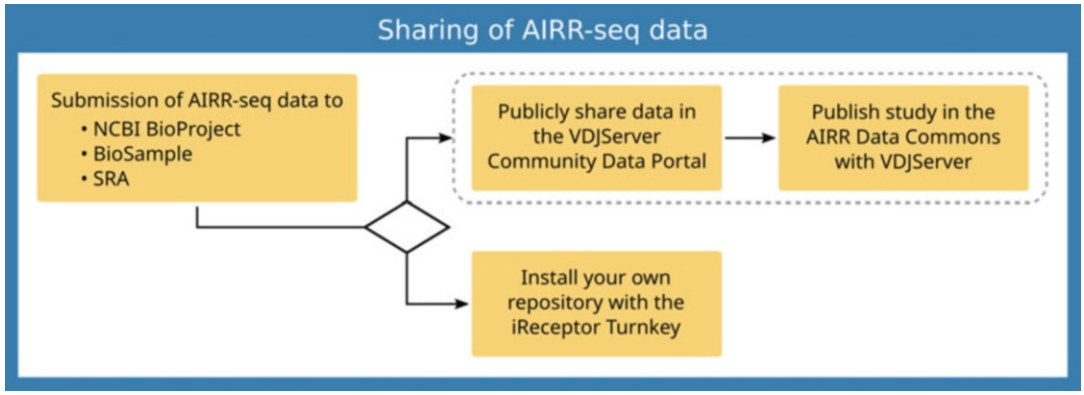


Fig. 3 Workflow sharing of AIRR-seq data

3.3 Methods for Sharing of AIRR-Seq Data

The AIRR Community provides processes for sharing AIRR-seq data. The following subsections provide detailed instructions for some of the common processes. An overview is given in Fig. 3.

3.3.1 Submission of AIRR-Seq Data to NCBI BioProject, BioSample, and SRA

At a minimum, raw sequence read data should be submitted to an INSDC repository such as NCBI for long-term archival along with study and sample metadata. The normal NCBI submission form is used, but, in place of the standard NCBI spreadsheets, the AIRR Community XLS spreadsheets, which contain MiAIRR 1.0-compliant data elements, are used:

1. AIRR BioSample XLS (AIRR_BioSample_v1.0.xls).
2. AIRR SRA XLS (AIRR_SRA_v1.0.xls).

Download these spreadsheets from the AIRR Community GitHub repository (https://github.com/airr-community/airr-standards/tree/master/NCBI_implementation/templates_XLS), fill them out with sample and sequencing run metadata, and then upload them as part of the NCBI submission. Here are the detailed steps:

1. Go to https://submit.ncbi.nlm.nih.gov/subs/sra and login with an NCBI account.
2. Click the *New Submission* button. This will start a multistep SRA submission process.
3. Fill out the forms. The process will allow one to attach the sequencing data to an existing BioProject and BioSample records, or it will create new records.
4. When requested to provide BioSample Attributes, select the upload XLS file option and provide the completed AIRR BioSample XLS file.

5. When requested to provide SRA Metadata, select the upload XLS file option and provide the completed AIRR SRA XLS file.
6. Upload the sequencing files and finish the submission.

When your submission is published, the BioSample and SRA records will contain all of the MiAIRR metadata provided in the spreadsheets.

Even with raw sequencing data available in an INSDC repository, it is not immediately useful as the data needs to be preprocessed and annotated before it can be used for immune repertoire analysis. It is strongly encouraged to make postprocessed data (annotated sequences, clonal analysis, clonal lineage) available. One method is to publicly share the data so that it can be downloaded; Subheading 3.3.2 describes how to do that with the VDJServer Community Data Portal. The other method is to publish your AIRR-seq data in the ADC; Subheading 3.3.3 describes publishing in VDJServer's repository, and Subheading 3.3.4 describes running your own repository using the iReceptor Turnkey.

3.3.2 Publicly Share Data in the VDJServer Community Data Portal

If VDJServer is used for immune repertoire analysis as described above in the AIRR Community "Immune Repertoire Analysis on High-Performance Computing Using VDJServer V1" chapter, the data and analysis files are already available within VDJServer; in this case, skip the first steps and go directly to **step 3** below to publicly share the project. It is not necessary to have performed the analysis on VDJServer in order to publicly share the data. Create an account at https://vdjserver.org to get started.

Create Project

Click on the *Add Project* button to create a new project and give the project a name. To help users identify the project, use a descriptive name such as the title of the study publication. After the project is created, go to the *Metadata Entry* page and fill out the Project/Study Metadata with a long study description (e.g., abstract of paper), PI and contact information, grant information, publication identifiers (e.g., Pubmed ID), and the BioProject ID. Click the *Save Project Metadata* button to save changes. It is not necessary to enter metadata for the other sections.

Upload Files into Project

From the *Upload and Browse Project Data* page, click on the *Upload* button and select files from the local computer, from Dropbox, or from a URL (ftp/http), to be uploaded; multiple files can be selected. Click the *Start* button to start uploading.

Publish Project

On the *Project Settings* page, copy/paste the VDJServer UUID. This is a long identifier with numbers, letters, and dashes that uniquely identifies the project. Provide this UUID in the Data Availability section of the publication so users can directly search

and find the project. Finally, click the *Project Actions* button and select Publish Project. This will initiate publishing, and one will receive an email when the project is publicly available. Changes cannot be directly made to a published project, but as project owner one can unpublish the project at any time to correct information or files. On the *Community Data* page, find the project and go to the *Project Settings* page. Click the Project Actions button and select Unpublish Project; one will receive an email when the project has been unpublished. Make the necessary corrections to the project then publish it to make it publicly available again.

3.3.3 Publish the AIRR-Seq Study in the ADC with VDJServer

Publishing the AIRR-seq study in the ADC with VDJServer is not a completely automated process; there are a number of manual validation steps that need to be performed. Furthermore, loading the data into the repository database can take hours, days, or even a week depending upon the size of the data; therefore, the load process is initiated by a VDJServer administrator. The basic requirements include:

1. Study metadata in AIRR Repertoire format. Validation scripts are run to verify the metadata is valid and complete. If the study metadata has been provided in VDJServer's Metadata Entry page, that metadata can be automatically converted into the AIRR Repertoire format.
2. Rearrangement (annotated sequence) data in AIRR TSV format. If the rearrangement data is not in the AIRR TSV format, it may need to be converted or run through the IgBLAST tool on VDJServer. Validation scripts are run to verify that the annotations are valid and complete.
3. VDJServer administrator loads the study into VDJServer's repository. Contact VDJServer (vdjserver@utsouthwestern.edu) to initiate publishing the study.

3.3.4 Install a Local Repository with the iReceptor Turnkey

The iReceptor Turnkey [7] is a self-contained AIRR compliant database platform that makes it easy for a research group to curate and share their own data. The iReceptor Turnkey software (https://github.com/sfu-ireceptor/turnkey-service-php/blob/master/README.md) is open source and is available for download via Github. The software uses Docker containers to manage the installation and includes a container for the repository itself (MongoDB), a container for the web service that implements the ADC API to query the repository, and a container to load data into the repository. It assumes that one is installing the software on a Unix platform and have appropriate privileges to install software.

Installation is a simple four-part process. More detailed instructions for installing and managing an iReceptor Turnkey repository are available on the iReceptor Turnkey repository github site.

Download the Software

Downloading the software is straightforward with the following command:

```
git clone --branch production-v3 https://github.com/sfu-ireceptor/turnkey-service-php.git
```

This downloads the v3.0 production release (June 2020), which includes the Docker configuration files and the basic commands one that use to control and manage the repository.

Install the Software

Installing the software is simple through a provided installation script. Note that this installation script installs Docker, docker_compose, and downloads multiple Docker images from DockerHub. Total time estimate: 5–10 min.

```
cd turnkey-service-php
scripts/install_turnkey.sh
```

After this step, it should notify one that the system is installed and running. If the software is running correctly, one should be able to query the repository using typical command line URL software such as curl. Because no data has yet been loaded, the query below will return an empty data set.

```
curl --data "{}" "http://localhost/airr/v1/repertoire"
```

Loading AIRR-Seq Data

This is typically a two-step process; first it is necessary to load repertoire metadata that describes the study, subject, and samples that are in the study. The input file can either be an AIRR Repertoire JSON file or a simple comma separated file with each column header mapping to an AIRR Standard field name. We have provided some simple test data to test out the installation. To load the test repertoire metadata, simply issue the following command:

```
scripts/load_metadata.sh ireceptor test_data/PRJNA330606_Wang_1_sample_metadata.csv
```

One can check that it worked with the following command once again:

```
curl --data "{}" "http://localhost/airr/v1/repertoire"
```

One should now see a single repertoire returned as a JSON object.

Next, load a set of sequence annotation files. There is typically one set of AIRR-seq annotation files loaded for each row in the metadata file loaded above. Again, an example sequence annotation

file with 1000 rearrangements generated using the MiXCR annotation tool is provided and can be loaded using the following command:

```
scripts/load_rearrangements.sh mixcr test_data/
SRR4084215_aa_mixcr_annotation_1000_lines.txt
```

Finally, check that the rearrangements were loaded correctly.

```
curl --data "{}" "http://localhost/airr/v1/rearrangement"
```

One now has a running AIRR compliant repository, containing one repertoire with 1000 sequence annotations loaded.

Domain Name, Security, and Public Access to Your Repository

The first three steps provide an AIRR-compliant repository for local access. If the repository machine has a publicly accessible IP address, it can be added to the iReceptor Gateway to test and verify that the AIRR-seq data can be queried. Contact the iReceptor team (support@ireceptor.org) to enable the repository in the iReceptor Gateway.

To make the repository publicly accessible on the Internet, it should be given a domain name and an SSL certificate for https access. A domain name (e.g., repository.example.org) can be acquired through any number of companies, but one should contact their own institution regarding domain name policies as the institution may be able to provide and manage a domain name. Acquiring an SSL certificate is highly recommended because many modern browsers have restrictive policies that will prevent users from accessing the repository through a nonsecure connection. Similar to the domain name, an SSL certificate can be acquired through any number of companies, but one should contact their own institution as they may be able acquire and manage SSL certificates.

Finally, the repository can be added to the list of repositories in the ADC on the AIRR Community documentation website (https://docs.airr-community.org/en/stable/api/adc.html). File an issue at the AIRR Standards Github (https://github.com/airr-community/airr-standards) with information about the repository to get it added to the list.

3.4 Methods to Query AIRR-Seq Data in the ADC

The ADC API provides a programmatic method for accessing data from the ADC. This is possible through any programming language or tool that supports web queries. Examples are given below for performing queries from the Unix command line, python, and R. Note that the ADC consists of a large number of repositories, and when using the ADC API, it is necessary to send the query of interest to all repositories in the ADC if one wants all data in the ADC that meets the query constraints. The examples

below use only a single repository. For a complete list of ADC repositories please refer to the AIRR Community ADC web page (https://docs.airr-community.org/en/stable/resources/adc_support.html) or the AIRR Data Commons repository list on Fairsharing.org (https://fairsharing.org/biodbcore/?q=AIRR).

Queries are sent to ADC API repositories using the ADC API query language (in JSON format) and the responses from the API are provided in JSON as well. For more information on using the ADC API, please refer to the AIRR Community ADC API web page (https://docs.airr-community.org/en/stable/api/adc_api.html).

3.4.1 Using the UNIX Curl Command

The Unix curl command can be used to send a query to any ADC compliant repository. An example searching for all repertoires that are from subjects with species equal to the *Homo sapiens* ontology ID to the iReceptor repository http://covid19-1.ireceptor.org would be issued as follows:

```
curl -s --data \
  '{"filters": { "op":"=", "content":
  { "field":"subject.species.id", "value":"NCBITAXON:9606"}}}'
\
  http://covid19-1.ireceptor.org/airr/v1/repertoire
```

3.4.2 Query ADC API with R

A more complex example, using the R programming language and querying the same repository but looking for subjects with a COVID-19 disease diagnosis and specifically IG data, is given below. In addition, this example also takes the results of the initial repertoire query to retrieve a small number of rearrangements from a single repertoire that was returned in the first query.

```
# Load required libraries
library(yaml)
library(httr)
library(dplyr)
library(jsonlite)

# Find Covid-19 repertoires
repertoire_api <-
  'http://covid19-1.ireceptor.org/airr/v1/repertoire'

# Set up a query for IG sequences from humans with COVID-19
# See https://docs.airr-community.org/en/stable/datarep/meta-
data.htm
# for further information on the fields and values
query_repertoires <- '{
   "filters":{
```

```
        "op":"and",
        "content": [
          {
              "op":"=",
              "content": {
                   "field":"subject.organism.id",
                   "value":"9606"
              }
          },
              {
              "op":"in",
              "content": {
          "field":"sample.pcr_target.pcr_target_locus",
                   "value": ["IGH",
                          "IGK",
                          "IGL"]
              }
              },
              {
              "op":"=",
              "content": {
          "field":"subject.diagnosis.disease_diagnosis",
                   "value":"DOID:0080600"
              }
              }
        ]
    }
}'

repertoires_response <-
  POST(url = repertoire_api, body = query_repertoires)

repertoires <-
  jsonlite::fromJSON(
      httr::content(
      repertoires_response,
      as = "text",
      encoding = "UTF-8"),
      simplifyDataFrame = TRUE)

selected_repertoires_id <-
  unique(repertoires$Repertoire$repertoire_id)
```

Once the repertoire_id is known, it is possible to use a loop to retrieve the sequence data. This example shows how to retrieve three sequences from the first Covid-19 repertoire returned from the previous query:

```
# API's Rearrangement endpoint url
rearrangements_api <-
  "http://covid19-1.ireceptor.org/airr/v1/rearrangement"

# Prepare the query
rearrangement_query <-
  paste0(
        '{"filters": {"op":"=","content": {"field":"repertoir-
e_id","value":"',
       selected_repertoires_id[1],
       '"}},"size":3}'
  )

# Submit the query
rearrangement_response <- POST(rearrangements_api,
                                  body = rearrangement_query)

# Parse the response
rearrangement <-
  jsonlite::fromJSON(
       httr::content(
       rearrangement_response,
       as = "text",
       encoding = "UTF-8"),
       simplifyDataFrame = TRUE
  )

# Explore the response:
# General information
rearrangement$Info

# Inspect the first 3 rows and columns of the Rearrangement
rearrangement$Rearrangement[1:3, 1:3]
```

3.4.3 Query ADC API with Python

A slightly different query using the python programming language is provided below. This queries the VDJServer repository (https://vdjserver.org) for all repertoires that contain TRB data from a specific study (with Study ID PRJNA300878) and then writes that data to a file. A second query in this example downloads 1000 productive rearrangements from a single repertoire from the same repository.

```
import airr
import requests

# This study is stored at VDJServer data repository
host_url = 'https://vdjserver.org/airr/v1'
```

```
#
# Query the repertoire endpoint
#

# POST data is sent with the query. Here we construct an object
for
# the query ((study_id == "PRJNA300878") AND (locus == "TRB"))

query = {
    "filters": {
        "op": "and",
        "content": [
            {
                "op": "=",
                "content": {
                    "field": "study.study_id",
                    "value": "PRJNA300878"
                }
            },
            {
                "op": "=",
                "content": {
                                "field": "sample.pcr_target.
pcr_target_locus",
                    "value": "TRB"
                }
            }
        ]
    }
}

# Send the query
resp = requests.post(host_url + '/repertoire', json=query)

# The data is returned as JSON, use AIRR library to write out
data
data = resp.json()
airr.write_repertoire('repertoires.airr.json',
                      data['Repertoire'], info=data['Info'])

# Construct a query to retrieve the 1000 productive sequences
from the
#  repertoire  with  repertoire_id  ==  2603354229190496746-
242ac113-0001-012
query = {
    "filters": {
        "op": "and",
```

```
        "content": [
            {
                "op": "=",
                "content": {
                    "field": "repertoire_id",
                  "value": "2603354229190496746-242ac113-0001-
012"
                }
            },
            {
                "op": "=",
                "content": {
                    "field": "productive",
                    "value": True
                }
            }
        ]
    },
    "size": 1000,
    "from": 0
}

# Send the query
resp = requests.post(host_url + '/rearrangement', json=query)
data = resp.json()
rearrangements = data['Rearrangement']
```

4 Notes

1. The "research value" of data: Even though the ADC contains over 4 billion annotated sequences from over 6000 repertoires and 60 studies, this is a small fraction of the AIRR-seq data that has been produced. As part of its AIRR-seq data curation effort, the AIRR Community has been attempting to document publicly available AIRR-seq data sets (data that are available in SRA/ENA repositories) on the B-T.cr web site (https://b-t.cr/t/publicly-available-airr-seq-data-sets-b-cells/470). Currently, there are 110 B-cell and 109 T-cell studies listed with known publicly available AIRR-seq data. Only 60 of these 219 studies are currently available in the ADC. As a result, it is likely that a researcher looking for a very specific type of AIRR-seq data (e.g., data from a rare disease with certain subject characteristics) may not be able to find it. This limit is primarily driven by the fact that not enough data has been shared in an easily usable form.

It is no surprise that the data available in the ADC reflects the priorities of researchers who took the time to curate this data. For example, the iReceptor Public Archive (IPA) repositories are currently cancer focused, as the iReceptor project chose cancer AIRR-seq studies as an important and broadly valuable resource for the general community. These studies were primarily curated in 2018 and 2019. Similarly, in response to the COVID-19 pandemic, the AIRR Community made a significant effort to curate COVID-19 AIRR-seq studies, resulting in 13 studies, 3500 repertoires, and over 1 billion sequence annotations being made available in 2020 from COVID-19 patients.

This is unfortunately a dilemma, with the costs of curating data for sharing being balanced against the rewards and value of that data, in turn driving the amount and type of data that is currently available for reuse. For example, the importance and value of COVID-19 data and its reuse in terms of fighting the COVID-19 pandemic has compelled researchers to share their data, even in some cases in the preprint stage. This, combined with standards for sharing being in place (the AIRR Standards) and readily available resources in which to store and share this data (e.g., iReceptor and VDJServer), has made COVID-19 data the single largest available disease diagnosis in the ADC [3]. This illustrates that by working toward a common goal of data accessibility, it is possible to resolve the conflicting costs of data sharing against the value of that shared data, leading to an increase in data reuse.

2. Interoperability and reusability: Although the AIRR Standards provide mechanisms to enable reuse of AIRR-seq data, challenges remain. These standards have varying levels of rigor applied to the AIRR fields. For example, many fields are defined as ontology terms, coming from well-known and widely used ontologies such as the NCBI Taxonomy (https://www.ncbi.nlm.nih.gov/taxonomy) and the Disease Ontology (https://disease-ontology.org). Other terms come from AIRR defined controlled vocabularies (e.g., pcr_target_locus, as used above in the ADC query API, comes from a controlled vocabulary of IGH, IGI, IGK, IGL, TRA, TRB, TRD, TRG). Such rigor makes it possible to be very precise about sharing, exploring, and reusing data across studies. In particular, it is possible to compare such metadata computationally, taking away the need for an expert to interpret such fields.

   At the same time, it is not possible to be so precise about all fields. The AIRR-seq world is evolving rapidly, and agreement around rigor on all terms in the standard is not feasible and in many cases is not desirable. Researchers need flexibility to describe parts of their study in ways that we are not currently

able to capture precisely. As the domain matures, more rigor will be applied to more fields in the standard, but in some cases, it is still challenging to compare two repertoires interoperable way. The most prominent fields where this is prevalent is in the fields involving data processing. Given the complexities discussed around preprocessing, annotation, and analysis of AIRR-seq data, standardization around fields and their contents are yet to be determined. Although the MiAIRR data standard has fields to capture this information, the community is working toward a concise specification for describing these processes in detail.

3. Knowing where to look for data: In working from the ADC API level to find and access data, one challenge when searching for data across the ADC is knowing which repositories are actually part of the ADC. Currently, there is no central registry where one can get a list of the repositories in the ADC; at present, the main central resource providing this information is the AIRR Community ADC documentation page (https://docs.airr-community.org/en/stable/resources/adc_support.html) with the AIRR Community also maintaining a registry of repositories on Faisharing.org (https://fairsharing.org/biodbcore/?q=AIRR). This page provides links to the existing repository providers. If searching for data in the ADC using a programmatic interface (Python, R), then it is necessary for the end user to manage the list of repositories and send queries to the appropriate repositories as required.

   Web-based user interfaces, such as the iReceptor Gateway, typically hide the fact that there are multiple repositories being queried, and these platforms maintain an internal registry so the end user does not need to worry about where the data resides.

   The AIRR Community is working on the specification and establishment of a central registry that will provide programmatic access to a list of repositories known to support the ADC API.

4. The time and cost of data sharing: As evidenced by the nuances discussed in both this and the "AIRR-Community Guide: Planning and Performing AIRR-Seq Experiments" chapter, the process of defining and performing a study involving AIRR-seq data is incredibly complex. The MiAIRR and other AIRR Standards are designed to guide researchers in capturing study/data processing features important for data reuse. Because of this completeness, the standards necessarily contain a lot of detail. Although the AIRR Community encourages researchers to be as complete as possible in documenting their study design and process, the more critical aspect of data sharing and re-use is that when one does document a part of a

study, it is done in a standard's compliant way. Although the MiAIRR Standard has a large number of fields, all of them considered "important," only a small subset of these fields are "essential/required" for a study to be AIRR-compliant.

There is a balance to be struck between the time it takes to capture study metadata at an appropriate level, both for internal requirements to perform the study and external requirements for data sharing and reuse. The MiAIRR Standard was created to help on both of these fronts, providing researchers with a comprehensive list of metadata fields that they might consider when designing a study and fields that enable comparison of study methods and guide decisions on data reuse. One of the main reasons that it is currently costly in terms of time and effort to reuse data (see below) is because the data has not been curated in a manner that facilitates this reuse.

5. The challenges of starting with FASTA/FASTQ: Although the INSDC sequence archive repositories (SRA, ENA, etc.) are a critical resource for long term storage of raw AIRR-seq data, it can be challenging for some researchers to reuse this data. This challenge comes from the need to transform data as stored in SRA/ENA to data that can be reused in analysis. As discussed in the "AIRR Community Guide: TR and IG Gene Annotation" chapter, the transformation of raw sequence data to annotated data, which is in turn the basis for data reuse in analysis, can be complex. If all data reuse requires starting from FASTA/FASTQ sequence files, each case of reuse requires the reannotation of the data, including the expertise and time to run the annotation pipelines. Even for experienced bioinformaticians, lack of data preprocessing information (e.g. primer/barcode sequences or thresholds for trimming/filtering) adds significant time to data reannotation and can potentially impact reproducibility of results. In some cases, researchers may need to contact the data generator directly to obtain critical details such as barcodes for sample demultiplexing; depending on the level of collaboration, this could pose a barrier to data reuse. In addition, unless the AIRR-to-NCBI pipeline (Subheading 3.1.2) was used to process study metadata, it is unlikely that the study, subject, and sample metadata will be stored in a standard-based and reusable format. Mapping study metadata and rerunning annotation pipelines is error prone, and this process needs to be redone by each researcher wanting to reuse a data set obtained from INSDC sequence archives.

   The AIRR Community has attempted to minimize this cost, firstly by providing a process to store the critical study, subject, and sample metadata in the INSDC repositories and secondly by providing a mechanism—the AIRR Standards and

the AIRR Data Commons—by which researchers can curate and share annotated AIRR-seq data. Through storing AIRR-seq data in a standard-based curated and annotated form, each AIRR-seq data set can be annotated and stored once, and then reused by many others without the costly overheads of reannotation. Because the AIRR Standard supports the ability to curate the annotation process, it is possible to have a data set from a single study, annotated multiple times using different annotation tools and have the user differentiate and choose the annotation that they think is the most appropriate for their data reuse.

6. Using the right data set: Making data FAIR is critical to the advancement of science not only by supporting transparency and reproducibility of published results but also by making data available for reuse in support of new research questions. In order to reuse data in this manner, it is critical to evaluate the data sets to ensure that they are comparable. The criteria required to make this decision would vary dramatically based on the research question being asked, ranging from questions around subjects (gender and age), disease diagnosis, sample preparation, and data processing. These are complex questions that span many fields across the MiAIRR data standard, and because some of these fields still lack rigor, determining the comparability of two data sets for such reuse requires manual intervention and often interpretation by the researcher. Although still challenging, through the use of the existing MiAIRR terms and the tools that make it possible to compare such data sets (e.g., the ADC API and the iReceptor Gateway), it is currently possible to find, compare, and make appropriate decisions about reuse of AIRR-seq data from the ADC [25].
7. Considerations when reusing data from different studies: After selecting studies for joint analysis based on study metadata as discussed above, it is important to consider how variations in experimental, data preprocessing, and rearrangement annotation protocols across the studies may impact the validity of the results. Data reuse for joint analysis can involve directly combining data sets for reanalysis or combining the results of earlier studies for meta-analysis. Clearly, in the case of meta-analysis, any results that will be combined or compared should have been derived using the same method. For example, meta-analysis involving diversity should ensure that the same diversity metric or set of metrics was used across the included studies. Perhaps less obvious, however, is that differences in the experimental and data preprocessing and annotation protocols used in the different studies can lead to invalid conclusions. This is equally true for reanalysis of combined data sets and for meta-analysis.

Regarding differences in experimental protocols, each of the decision points outlined in the "AIRR Community Guide to TR and IG Gene Annotation" chapter, Fig. 1 (single cell versus bulk sequencing, gDNA versus mRNA as the starting molecule, and whether UMIs were used), as well as differences in sequencing depth, sequencing error rates, read length, and the primers used can impact the data in ways that can make certain conclusions invalid. For example, differences in hybridization and amplification efficiencies between primers in a primer set can influence gene usage estimates. Thus, combining or comparing gene usage estimates between studies that used different primer sets can result in gene usage differences that are attributable to experimental artifacts and do not reflect true repertoire differences. The potential impact of specific experimental protocol choices on analysis conclusions are briefly outlined here and in Table 2. All of these experimental protocol differences can impact the number of unique receptor sequences observed in a sample, particularly whether rare sequences are observed. This can in turn result in artifacts when computing any of the common diversity metrics, such as clonality, repertoire overlap, and when constructing sequence networks. Similarly, protocol differences that affect relative abundances of starting templates, including whether gDNA or mRNA was used, primer differences, and whether UMIs were used, can result in artifacts when computing diversity and clonality measures and when analyzing gene usage. Finally, differences in the use of UMIs, sequencing platform error rates, and primers can result in artifacts for somatic hypermutation analyses.

Differences in preprocessing and sequence annotation protocols can similarly result in artifacts during re- or meta-analysis (Table 2). Differences in sequence filtering and deduplication can impact the number of unique AIRR sequences observed in a sample, as well as estimates of starting template relative abundances, in turn impacting diversity, clonality, and overlap measures, as well as sequence networks and gene usage estimates. The germline gene database and alignment algorithm used for germline gene annotations can impact germline gene assignments, thereby impacting gene usage estimates and somatic hypermutation analyses. These can also impact diversity, clonality, and other analyses when they are conducted at the annotation rather than sequence level (i.e., defining unique rearrangements according to their V gene, J gene, and CDR3 sequence rather than their full-length sequence). Finally, assigning clonal membership can depend on the clonal assignment algorithm used and could thereby impact the results of IG affinity maturation.

Table 2
Considerations in re- and meta-analysis

Protocol differences	# of unique sequences	Estimates of template relative abundance	Germline gene assignments	Clone assignment
Experimental				
Single-cell vs. bulk sequencing	X	X		
DNA vs mRNA	X	X		
Use of UMIs	X	X	X	X
Sequencing depth	X	X		
Sequencing error rate and bias	X	X	X	X
Read length	X	X	X	X
Primer hybridization and amplification bias	X	X		
Primer hybridization location			X	X
Preprocessing				
Filtering and deduplication stringency	X	X		
Rearrangement annotation				
Germline gene database			X	X
Germline gene Annotation algorithm			X	X
Clone assignment algorithm				X
Analyses				
Richness	X		O	X
Diversity/clonality	X	X	O	X
Repertoire overlap	X		O	
Average sequence similarity	X		O	
Repertoire similarity network	X		O	
Gene usage		X	X	

(continued)

Table 2
(continued)

Protocol differences	# of unique sequences	Estimates of template relative abundance	Germline gene assignments	Clone assignment
Somatic hypermutation			X	
IG affinity maturation				X

Upper panel—Experimental, preprocessing, and rearrangement annotation protocol differences are shown in rows, and the impacted AIRR-seq data set features are shown in columns. X indicates when a data set feature is expected to be impacted by a particular protocol difference. **Lower panel**—Analysis types are shown in rows. X indicates when an analysis type is expected to be impacted by the data set feature in the corresponding column. O indicates that an analysis type is expected to be impacted when the analysis is conducted at the annotation- rather than sequence-level (e.g., V gene, J gene, and CDR3 sequence rather than full-length sequence)

The above and Table 2 are not meant to be comprehensive, but instead serve as a guide when designing analyses that combine data sets or results across different studies. When selecting studies for such an analysis and formulating the research questions that can be reliably addressed, it is important to identify differences in experimental, preprocessing and annotation protocols and understand how these protocol differences can affect relevant data set features and analyses. It is recommended to choose research questions and analysis methods that are independent of any protocol differences where possible. Furthermore, common experimental design best practices are encouraged, such as ensuring that protocol differences do not partition with treatment groups and incorporating methods to control for and/or estimate their effects, as one would batch effects.

Acknowledgments

We would like to thank our colleagues from the AIRR Community, who have dedicated many hours to the development of the community and the standards and initiatives on which this chapter is based. In particular, we would like to thank the authors of the other AIRR Community chapters in this volume, with a special thanks to Susanna Marquez, William Lees, and Ulrik Stervbo who assisted with content and editing of this chapter.

References

1. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A et al (2016) The FAIR guiding principles for scientific data management and stewardship. Sci Data 3:1–9. https://doi.org/10.1038/sdata.2016.18

2. Breden F, Luning Prak ET, Peters B, Rubelt F, Schramm CA, Busse CE et al (2017) Reproducibility and reuse of adaptive immune receptor repertoire data. Front Immunol 8:1418.

https://doi.org/10.3389/fimmu.2017.01418

3. Scott JK, Breden F (2020) The adaptive immune receptor repertoire community as a model for FAIR stewardship of big immunology data. Curr Opin Syst Biol 24:71–77. https://doi.org/10.1016/j.coisb.2020.10.001
4. Rubelt F, Busse CE, Bukhari SAC, Bürckert J-P, Mariotti-Ferrandiz E, Cowell LG et al (2017) Adaptive immune receptor repertoire community recommendations for sharing immune-repertoire sequencing data. Nat Immunol 18:1274–1278. https://doi.org/10.1038/ni.3873
5. Christley S, Aguiar A, Blanck G, Breden F, Bukhari SAC, Busse CE et al (2020) The ADC API: a web API for the programmatic query of the AIRR data commons. Front Big Data 3:22. https://doi.org/10.3389/fdata.2020.00022
6. Vander Heiden JA, Marquez S, Marthandan N, Bukhari SAC, Busse CE, Corrie B et al (2018) AIRR community standardized representations for annotated immune repertoires. Front Immunol 9:2206. https://doi.org/10.3389/fimmu.2018.02206
7. Corrie BD, Marthandan N, Zimonja B, Jaglale J, Zhou Y, Barr E et al (2018) iReceptor: a platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories. Immunol Rev 284:24–41. https://doi.org/10.1111/imr.12666
8. Christley S, Scarborough W, Salinas E, Rounds WH, Toby IT, Fonner JM et al (2018) VDJServer: a cloud-based analysis portal and data commons for immune repertoire sequences and rearrangements. Front Immunol 9:976. https://doi.org/10.3389/fimmu.2018.00976
9. Rosenfeld AM, Meng W, Luning Prak ET, Hershberg U (2018) ImmuneDB, a novel tool for the analysis, storage, and dissemination of immune repertoire sequencing data. Front Immunol 9:2107. https://doi.org/10.3389/fimmu.2018.02107
10. Imkeller K, Arndt PF, Wardemann H, Busse CE (2016) sciReptor: analysis of single-cell level immunoglobulin repertoires. BMC Bioinformatics 17:67. https://doi.org/10.1186/s12859-016-0920-1
11. Borghardt P (2020) COVID-19 Demands Increased Public Sharing of Biomedical Research Data. https://perma.cc/UC5Q-X4J2. Accessed 5 Mar 2021
12. Arnaout RA, Prak ETL, Schwab N, Rubelt F, Arora R, Bashford-Rogers R et al (2021) The future of blood testing is the Immunome. Front Immunol 12:228. https://doi.org/10.3389/fimmu.2021.626793
13. Brüggemann M, Kotrová M, Knecht H, Bartram J, Boudjogrha M, Bystry V et al (2019) Standardized next-generation sequencing of immunoglobulin and T-cell receptor gene recombinations for MRD marker identification in acute lymphoblastic leukaemia; a EuroClonality-NGS validation study. Leukemia 33:2241–2253. https://doi.org/10.1038/s41375-019-0496-7
14. Gittelman RM, Lavezzo E, Snyder TM, Zahid HJ, Elyanow R, Dalai S et al (2020) Diagnosis and tracking of SARS-CoV-2 infection by T-cell receptor sequencing. Preprint, infectious diseases (except HIV/AIDS). MedRXiv preprint, downloaded 2022–01–15. https://doi.org/10.1101/2020.11.09.20228023
15. Commissioner O of the (2021) Coronavirus (COVID-19) update: FDA authorizes adaptive biotechnologies T-detect COVID test. In: FDA https://www.fda.gov/news-events/press-announcements/coronavirus-covid-19-update-fda-authorizes-adaptive-biotechnologies-t-detect-covid-test. Accessed 9 Mar 2021
16. Zhang Y, Yang X, Zhang Y, Zhang Y, Wang M, Ou JX et al (2020) Tools for fundamental analysis functions of TCR repertoires: a systematic comparison. Brief Bioinform 21:1706–1716. https://doi.org/10.1093/bib/bbz092
17. López-Santibáñez-Jácome L, Avendaño-Vázquez SE, Flores-Jasso CF (2019) The pipeline repertoire for Ig-Seq analysis. Front Immunol 10:899. https://doi.org/10.3389/fimmu.2019.00899
18. Lees WD (2020) Tools for adaptive immune receptor repertoire sequencing. Curr Opin Syst Biol 24:86–92. https://doi.org/10.1016/j.coisb.2020.10.003
19. Smakaj E, Babrak L, Ohlin M, Shugay M, Briney B, Tosoni D et al (2020) Benchmarking immunoinformatic tools for the analysis of antibody repertoire sequences. Bioinformatics 36:1731–1739. https://doi.org/10.1093/bioinformatics/btz845
20. Bukhari SAC, O'Connor MJ, Martínez-Romero M, Egyedi AL, Willrett D, Graybeal J et al (2018) The CAIRR pipeline for submitting standards-compliant B and T cell receptor repertoire sequencing studies to the National Center for biotechnology information

repositories. Front Immunol 9:1877. https://doi.org/10.3389/fimmu.2018.01877

21. Kovaltsuk A, Leem J, Kelm S, Snowden J, Deane CM, Krawczyk K (2018) Observed antibody space: a resource for data mining next-generation sequencing of antibody repertoires. J Immunol 201:2502–2509. https://doi.org/10.4049/jimmunol.1800708
22. Zhang W, Wang L, Liu K, Wei X, Yang K, Du W et al (2019) PIRD: pan immune repertoire database. Bioinformatics 36(3):897–903. https://doi.org/10.1093/bioinformatics/btz614
23. Chen S-Y, Yue T, Lei Q, Guo A-Y (2021) TCRdb: a comprehensive database for T-cell receptor sequences with powerful search function. Nucleic Acids Res 49:D468–D474. https://doi.org/10.1093/nar/gkaa796
24. Adaptive Biotechnologies immuneACCESS Data. https://clients.adaptivebiotech.com/immuneaccess. Accessed 3 Mar 2021
25. Heming M, Li X, Räuber S, Mausberg AK, Börsch A-L, Hartlehnert M et al (2021) Neurological manifestations of COVID-19 feature T cell exhaustion and dedifferentiated monocytes in cerebrospinal fluid. Immunity 54: 164–175.e6. https://doi.org/10.1016/j.immuni.2020.12.011
26. Randi, Vita Swapnil, Mahajan James A, Overton Sandeep Kumar, Dhanda Sheridan, Martini Jason R, Cantrell Daniel K, Wheeler Alessandro, Sette Bjoern, Peters (2019) (2018) The Immune Epitope Database (IEDB): 2018 update. Nucleic Acids Research 47(D1) D339–D343. https://doi.org/10.1093/nar/gky1006
27. Nili, Tickotsky Tal, Sagiv Jaime, Prilusky Eric, Shifrut Nir, Friedman Jonathan, Wren (2017) McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. Bioinformatics 33(18):2924–2929. https://doi.org/10.1093/bioinformatics/btx286
28. Mikhail, Shugay Dmitriy V, Bagaev Ivan V, Zvyagin Renske M, Vroomans Jeremy Chase, Crawford Garry, Dolton Ekaterina A, Komech Anastasiya L, Sycheva Anna E, Koneva Evgeniy S, Egorov Alexey V, Eliseev Ewald, Van Dyk Pradyot, Dash Meriem, Attaf Cristina, Rius Kristin, Ladell James E, McLaren Katherine K, Matthews E Bridie, Clemens Daniel C, Douek Fabio, Luciani Debbie, van Baarle Katherine, Kedzierska Can, Kesmir Paul G, Thomas David A, Price Andrew K, Sewell Dmitriy M, Chudakov (2018) (2017) VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. Nucleic Acids Research 46(D1):D419–D427. https://doi.org/10.1093/nar/gkx760

Chapter 24

IMGT® Immunoinformatics Tools for Standardized V-DOMAIN Analysis

Véronique Giudicelli, Patrice Duroux, Maël Rollin, Safa Aouinti, Géraldine Folch, Joumana Jabado-Michaloud, Marie-Paule Lefranc, and Sofia Kossida

Abstract

The variable domains (V-DOMAIN) of the antigen receptors, immunoglobulins (IG) or antibodies and T cell receptors (TR), which specifically recognize the antigens show a huge diversity in their sequences. This diversity results from the complex mechanisms involved in the synthesis of these domains at the DNA level (rearrangements of the variable (V), diversity (D), and joining (J) genes; N-diversity; and, for the IG, somatic hypermutations). The recognition of V, D, and J as "genes" and their entry in databases mark the creation of IMGT by Marie-Paule Lefranc, and the origin of immunoinformatics in 1989. For 30 years, IMGT®, the international ImMunoGeneTics information system® http://www.imgt.org, has implemented databases and developed tools for IG and TR immunoinformatics, based on the IMGT Scientific chart rules and IMGT-ONTOLOGY concepts and axioms, and more particularly, the princeps ones: IMGT genes and alleles (CLASSIFICATION axiom) and the IMGT unique numbering and IMGT Collier de Perles (NUMEROTATION axiom). This chapter describes the online tools for the characterization and annotation of the expressed V-DOMAIN sequences: (a) IMGT/V-QUEST analyzes in detail IG and TR rearranged nucleotide sequences, (b) IMGT/HighV-QUEST is its high throughput version, which includes a module for the identification of IMGT clonotypes and generates immunoprofiles of expressed V, D, and J genes and alleles, (c) IMGT/StatClonotype performs the pairwise comparison of IMGT/HighV-QUEST immunoprofiles, (d) IMGT/DomainGapAlign analyzes amino acid sequences and is frequently used in antibody engineering and humanization, and (e) IMGT/Collier-de-Perles provides two-dimensional (2D) graphical representations of V-DOMAIN, bridging the gap between sequences and 3D structures. These IMGT® tools are widely used in repertoire analyses of the adaptive immune responses in normal and pathological situations and in the design of engineered IG and TR for therapeutic applications.

Key words IMGT, Immunogenetics, Immunoinformatics, Immunoglobulin, Antibody, T cell receptor, V-DOMAIN sequence analysis, Adaptive immune repertoire, IMGT Collier de Perles, IMGT-ONTOLOGY

Anton W. Langerak (ed.), *Immunogenetics: Methods and Protocols*, Methods in Molecular Biology, vol. 2453, https://doi.org/10.1007/978-1-0716-2115-8_24,

1 Introduction

Immunoglobulins (IG) or antibodies [1, 2] and T cell receptors (TR) [3] are antigen receptors of the adaptive immune responses in vertebrates with jaws (gnathostomata) [4]. The huge diversity of the variable domains (V-DOMAIN) of the IG and TR chains of the immune repertoires results from several mechanisms that occur during their synthesis [1–4]. In particular, the combinatorial diversity depends on the number of variable (V), diversity (D), and joining (J) genes found in the IG and TR loci, which potentially can rearrange to form V-DOMAIN encoded by V-(D)-J regions [1–4]. It is the recognition of the V, D, and J as "genes" and their entry in databases that mark the creation of IMGT in 1989 by Marie-Paule Lefranc (Université de Montpellier, CNRS) at Human Gene Mapping 10 (HGM10) and is at the origin of immunoinformatics, a new science at the interface between immunogenetics and bioinformatics [4].

Other mechanisms of diversity comprise the junctional diversity with exonuclease trimming at the ends of the V, D, and J genes and the random addition of nontemplated nucleotides, preferably "g" and "c," by the terminal deoxynucleotidyl transferase (TdT) encoded by the DNA nucleotidylexotransferase (DNTT) gene, creating the N-regions [1–3], and for IG, somatic hypermutations [1, 2]. For 30 years, IMGT®, the international ImMunoGeneTics information system® http://www.imgt.org, has implemented databases and developed tools for IG and TR immunoinformatics [5], based on the IMGT Scientific chart rules (*see* Subheading 2) and IMGT-ONTOLOGY concepts and axioms [6, 7], and more particularly, the princeps ones: IMGT genes and alleles (CLASSIFICATION axiom) [8–12] and the IMGT unique numbering [13–17] and IMGT Collier de Perles [18–21] (NUMEROTATION axiom). This chapter describes the online analysis tools for the characterization and annotation of the expressed V-DOMAIN nucleotide (nt) and amino acid (AA) sequences, available from "IMGT tools" section of the IMGT® Home page. Protocols for their use and the description of main results are presented in this chapter. These concern the following: (a) IMGT/V-QUEST [22, 23] is the IMGT® online tool for the analysis of IG and TR nucleotide rearranged sequences (*see* Subheading 3); (b) IMGT/HighV-QUEST [24–27], the high throughput version of IMGT/V-QUEST, can analyze sets of up to one million sequences. It includes a module for the identification of IMGT clonotypes (AA) and the generation of IG and TR gene profiles for the diversity and expression of IMGT clonotypes (AA) (*see* Subheading 4); (c) IMGT/StatClonotype [28, 29] is a standalone package that performs statistical pairwise comparisons of IMGT clonotype (AA) diversity or expression between two IMGT/HighV-QUEST result sets (*see* Subheading

5); (d) IMGT/DomainGapAlign [30, 31] analyses domain AA sequences and two dimensional (2D) structures, and its results are used in antibody engineering and humanization [32, 33] (*see* Subheading 6); and (e) IMGT/Collier-de-Perles tool [21] generates IMGT Colliers de Perles graphical 2D representations for AA domain sequences [18–20] (*see* Subheading 7), it is available from the IMGT Home page and is also automatically launched by IMGT/V-QUEST and IMGT/DomainGapAlign.

2 IMGT Scientific Chart Rules for the Analysis of the V-DOMAIN

2.1 IMGT Gene and Allele Nomenclature and IMGT Reference Directory Sets

The IMGT gene names of the IG and TR V, D, J, and C genes [1–4] were approved by the Human Genome Organization (HUGO) Nomenclature Committee (HGNC) in 1999 [8, 9, 12] and were endorsed by the WHO-IUIS Nomenclature Subcommittee for IG and TR [10, 11]. IMGT gene and allele names are based on the concepts of classification of IMGT-ONTOLOGY "Group," "Subgroup," "Gene," and "Allele" [1–4, 10–12]. Alleles are the polymorphic variants of a gene: they are identified by their IMGT reference sequence, which corresponds to the coding V-REGION, D-REGION, J-REGION, and C-REGION sequence at the nucleotide level of V, D, J, and C gene alleles, respectively. IMGT reference directory sets include the allele IMGT reference sequences from functional (F) genes and alleles, open reading frame (ORF), and pseudogenes (P) [5]. IMGT germline V, D, and J genes and alleles, with their characteristics, their reference sequence and other sequences from the literature are managed in IMGT/GENE-DB [34] and in IMGT Repertoire (IG and TR) Gene tables and Alignments of alleles Web resources [1–4]. The tools for V-DOMAIN analysis compare user sequences with IMGT reference directory sets for the identification of V, D, and J genes and alleles and the evaluation of mutations and AA changes.

2.2 IMGT Unique Numbering for the IG and TR V Domains

An IG or TR V-DOMAIN comprises about 100 amino acids and is made of nine antiparallel beta strands (A, B, C, C′, C″, D, E, F, and G) linked by beta turns (AB, CC′, C″D, DE, and EF) or loops (BC, C′C″, and FG) [35]. At the structural level, they form a sandwich of two sheets closely packed against each other through hydrophobic interactions and joined together by a disulfide bridge between 1st-CYS at position 23 in B-STRAND (in the first sheet) and 2nd-CYS at position 104 in F-STRAND (in the second sheet) [13].

The IMGT unique numbering for IG and TR V-DOMAIN [13] delimits (1) the four framework regions: FR1-IMGT (A and B strands, from positions 1 to 26), FR2-IMGT (C and C′ strands, from positions 39 to 55), FR3-IMGT (C″, D, E and F strands, from positions 66 to 104), FR4-IMGT (G strand, from positions 118 to 128), and (2) the three hypervariable or complementarity

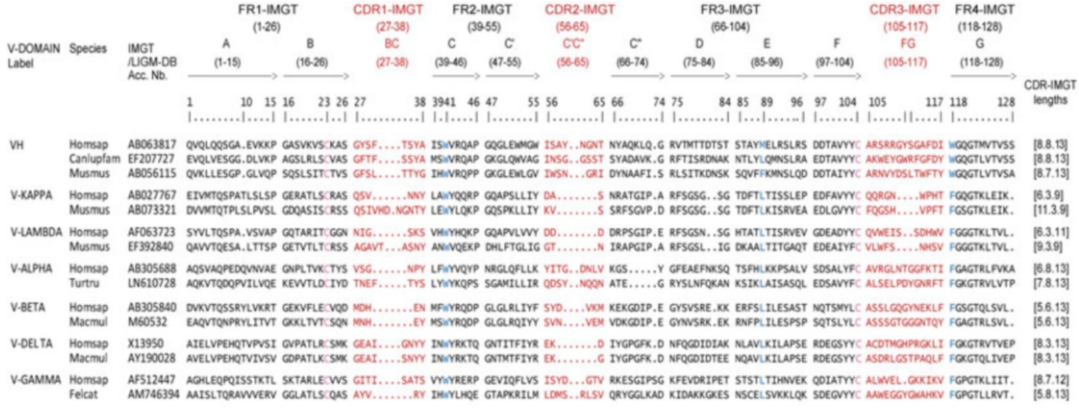


Fig. 1 Protein displays of IG and TR V-DOMAIN based on the IMGT unique numbering for V-DOMAIN [13]. The V-DOMAIN translations were obtained from the analysis by IMGT/V-QUEST [22, 23] (*see* Subheading 3) of the nucleotide sequences of shown accession numbers in IMGT/LIGM-DB [36]. The identification of FR-IMGT and CDR-IMGT and of beta strands and loops was performed by IMGT/DomainGapAlign [30, 31] (*see* Subheading 6), which provides a standardized delimitation whatever the species, the receptor type, and the chain type. CDR-IMGT lengths are indicated between brackets, separated by dots (column on the right). 1st-CYS 23 and 2nd-CYS 104 are in pink, and W 41, hydrophobic AA 89, and W or F 118 are in blue. Taxons are in the IMGT 6- or 9-letter abbreviation: Homsap for *Homo sapiens*, Canlupfam for *Canis lupus familiaris* (dog), Musmus for *Mus musculus* (mouse), Turtru for *Tursiops truncatus* (dolphin), Macmul for *Macaca mulatta* (Rhesus monkey), and Felcat for *Felis catus* (cat)

determining regions involved in the ligand recognition: CDR1-IMGT (BC loop, positions 27 to 38), CDR2-IMGT (C′C″ loop, positions 56 to 65), and CDR3-IMGT (FG loop, positions 105 to 117, with additional positions 112.1, 111.1, 112.2 etc., if longer than 13 codons (or AA)). FR-IMGT positions, which delimit the three CDR-IMGT, are designated as anchors: they are 26 and 39, 55 and 66, and 104 and 118, respectively (Fig. 1), and shown as squares in IMGT Colliers de Perles [18–21]. According to the IMGT unique numbering [13], a V-DOMAIN is characterized by five highly conserved AA: 1st-CYS 23, tryptophan 41 (CONSERVED-TRP), hydrophobic amino acid 89, 2nd-CYS 104, and J-PHE or J-TRP 118 of the J-MOTIF (F/W-G-X-G, 118–121, where F is phenylalanine, W tryptophan, G glycine, X, any AA). The three CDR-IMGT lengths characterize a V-DOMAIN. By convention, they are indicated between brackets, separated by dots (for example [8.8.13]). The CDR1-IMGT and CDR2-IMGT are encoded by the V-REGION, whereas the CDR3-IMGT results from the V-(D)-J rearrangement. The IMGT Collier de Perles [18–20] can be generated by the IMGT/Collier-de-Perles tool [21] (*see* Subheading 7).

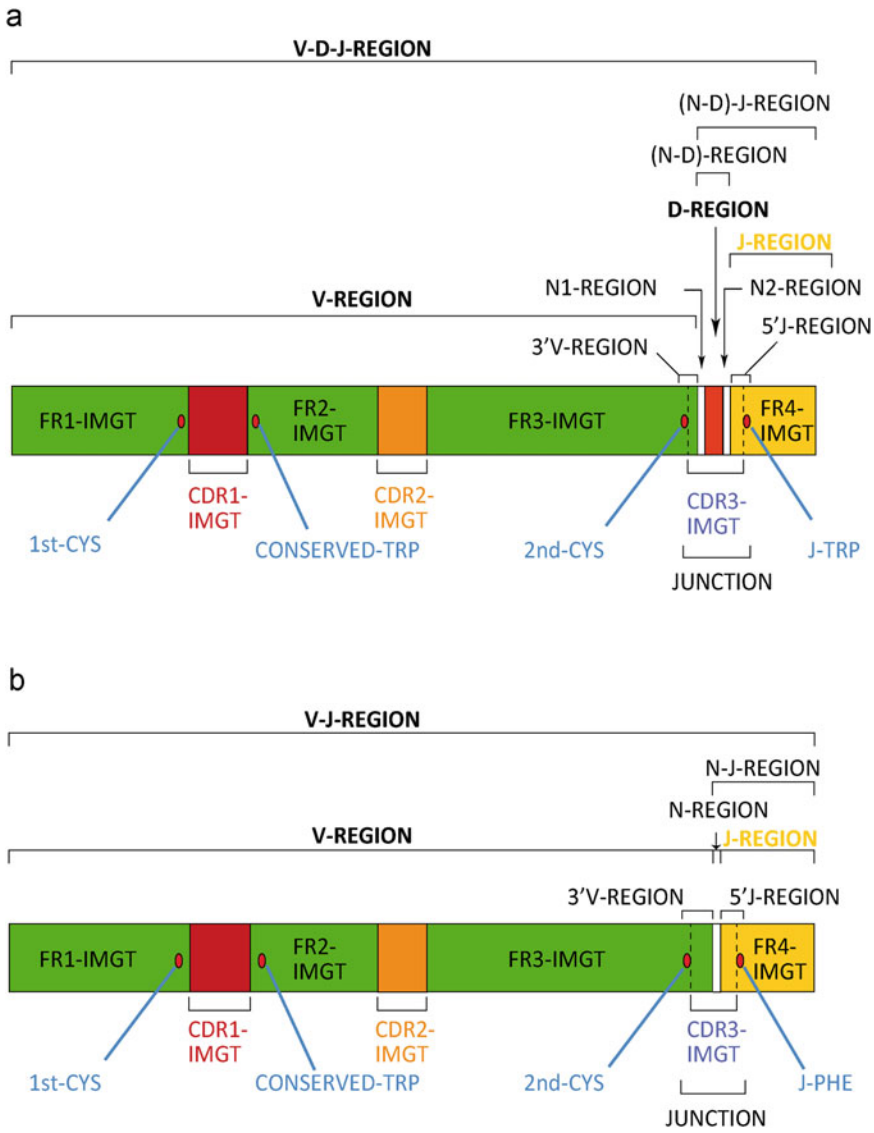


Fig. 2 Graphical representation or prototypes of IG and TR V-DOMAIN with IMGT labels at the nucleotide level. (**a**) V-D-J-REGION. (**b**) V-J-REGION [1–4]. The JUNCTION encompasses 2nd-CYS 104, CDR3-IMGT, and J-TRP or J-PHE 118, and its length is therefore two AA longer than CDR3-IMGT. Potential palindromic nucleotides ("P") identified in case of untrimmed V, D, and/or J regions during the DNA rearrangement are not shown (With permission from M-P. Lefranc and G. Lefranc, LIGM, Founders and Authors of IMGT®, the international ImMunoGeneTics information system®, http:/www.imgt.org)

2.3 IMGT Standardized Labels and Sequence Description

The IMGT tools, which perform the analysis of sequences, provide the description of the V-DOMAIN with IMGT standardized labels (written in capital letters). The V-DOMAIN corresponds either to a V-D-J-REGION (in IG heavy (IGH)), TR beta (TRB), and TR delta (TRD) chains) (Fig. 2a) or to a V-J-REGION (in IG light

lambda (IGL) and IG kappa (IGK)), TR alpha (TRA) and TR gamma (TRG) chains) (Fig. 2b), encoded by V-D-J or V-J rearrangements, respectively.

The V-DOMAIN labels according to the chain type or locus are: VH, V-KAPPA, V-LAMBDA for the IGH, IGK, and IGL, respectively, and V-ALPHA, V-BETA, V-DELTA, V-GAMMA for the TRA, TRB, TRD, and TRG, respectively [1–4].

2.4 IMGT Functionality of IG and TR Genes and Alleles and of Rearranged Sequences

The "Functionality" concept identifies the functionality based on the configuration of the IG and TR genes. The functionality of the germline (V, D and J) and undefined (C) IG and TR genes and alleles, defined on the same criteria as conventional genes and alleles, is either functional (F), open reading frame (ORF), or pseudogene (P). The functionality of the IG and TR V-(D)-J rearranged sequences is either "productive" (no stop codon and in-frame JUNCTION (2nd-CYS 104 and J-TRP/J-PHE 118 in the same reading frame)) or "unproductive" (stop codons and/or out-of-frame JUNCTION) [2].

3 IMGT/V-QUEST

IMGT/V-QUEST [22, 23] identifies the V, D, and J genes and alleles in IG and TR V domains. It characterizes the nucleotide (nt) mutations and amino acid (AA) changes resulting from somatic hypermutations in IG V-REGION. It provides a detailed characterization of the V-D-J or V-J junctions by the integrated IMGT/JunctionAnalysis tool [37, 38] and the full annotation of the V-DOMAIN with IMGT labels by IMGT/Automat [39, 40].

3.1 IMGT/V-QUEST Sequence Submission

The top of the IMGT/V-QUEST Welcome page (Fig. 3) provides two links: the first one gives access to the list of the IMGT/V-QUEST reference directory sets to which the users' own sequences can be compared (*see* **Note 1**), and the second one provides examples of human rearranged sequences to test the tool.

The page then includes five sections to configure the analysis:

3.1.1 Your Selection

1. Select the species or taxon first and then the receptor type or locus (*see* **Note 2**) in the lists.

3.1.2 Sequence Submission

IMGT/V-QUEST analyses up to 50 FASTA formatted IG or TR rearranged nucleotide sequences per run, from genomic DNA or cDNA indifferently.

1. Enter the sequences in the text area "Type (or copy/paste) your nucleotide sequence(s) in FASTA format".

Fig. 3 IMGT/V-QUEST Welcome page with the five sections: "Your selection," "Sequence submission," "Display results," "Advanced parameters," and "Advanced functionalities" [22, 23]

2. Alternatively, upload the sequences as a text file by selecting the option "Or give the path access to a local file containing your sequence(s) in FASTA format" (*see* **Note 3**).

3.1.3 Display Results

Three choices of display for the results are available [22, 23]. "A. Detailed view" and "B. Synthesis view" are displayed online in HTML (by default) or text format. Both include sequence alignments for which the user can define the number of nucleotides per line (60 by default). The third type of display, "C. Excel file," is dedicated for the download of the results.

1. Select "A. Detailed view" to get the results for each sequence individually. Results consist in a "Result summary" with the main results of the analysis and 14 detailed result sections that can be selectively checked or unchecked by the user (*see* **Note 4**). In sequence alignments, the number of IMGT reference sequences aligned with the user sequence (five by default) can be modified from 1 to 20.
2. Select "B. Synthesis view" to display the sequences that express the same V gene and allele aligned together. Results include a "Summary table" with the main results of the analysis which can be ordered by "V-GENE and allele name" (default) or by the sequence "input" order. There are eight detailed result sections that can be checked or unchecked (*see* **Note 5**).
3. Select "C. Excel file" to download the results, either in a spreadsheet (default) or as a zip archive. The results may include 11 sheets (or text files in the zip archive (*see* **Note 6**)), which can be checked or unchecked. The 12th sheet (or text file) is available if the option "Analysis of single chain Fragment variable (scFv)" is selected in "Advanced functionalities" (*see* Subheading 3.1.5) [41]. An alternative is to display the content of one given sheet in your browser ("Display 1 CSV file in your browser") or to "Download AIRR formatted results" as a zip archive (*see* **Note 7**) [42, 43].

3.1.4 Advanced Parameters

The default values of the advanced parameters are used by IMGT/V-QUEST for classical analyses [22, 23]. They may be modified for specific studies and/or unusual sequences. The user may:

1. Select the relevant set to be compared with the submitted sequences in "Selection of IMGT reference directory set": 'F+ORF', 'F+ORF+in frame P' (by default), 'F+ORF including orphons', or 'F+ORF+in frame P including orphons' (*see* **Note 8**), "With all alleles" of genes or "With allele *01 only" in order to restrict the IMGT reference directory to one representative sequence per gene only.
2. Choose to "Search for insertions and deletions in V-REGION" or not. By default, IMGT/V-QUEST does not search for

insertions and/or deletions. Selecting "Yes" allows to identify the somatic hypermutations by nucleotide insertions and deletions in the V-REGION that may occur in normal and malignant cells [44] and/or potential sequencing errors.

3. Set the values for "Parameters for IMGT/JunctionAnalysis" that include:
   (a) "Nb of accepted D-GENE" (number of D genes searched by the tool in IGH, TRB or TRD junctions).
   (b) "Nb of accepted mutations" in 3′V-REGION, D-REGION, and 5′J-REGION: by default, 2, 4 and 2 mutations are accepted in the 3′V-REGION, D-REGION, and 5′J-REGION, respectively, for IGH, 7 in the 3′V-REGION and 5′J-REGION for IGK and IGL junctions (*see* **Note 9**). By default, no mutation is accepted for the TR junctions.
4. Set "Parameters for Detailed view":
   (a) "Nb of nucleotides to exclude in 5′ of the V-REGION for the evaluation of the number of mutations" (useful in case of primer specific nucleotides).
   (b) "Nb of nucleotides to add (or exclude) in 3′ of the V-REGION for the evaluation of the alignment score" (useful in case of low (or high) exonuclease activity).

3.1.5 Advanced Functionalities

"Advanced functionalities" [22, 23] corresponds to specific analyses, with additional dedicated results, for engineered/artificial sequences, and for the search of specific sequences for clinical applications. The user may:

1. Select "Analysis of single chain Fragment variable (scFv)" if the submitted set contains engineered single chains with two V-DOMAIN connected by a linker. IMGT/V-QUEST will search for the two V-DOMAIN in the submitted sequence (*see* **Note 10**) [41]. This functionality is generic for IG and TR.
2. Select "Clinical application: search for CLL subsets #2 and #8" for sequences from patients with chronic lymphocytic leukemia (CLL). The analysis of IGH sequences includes the search of specific rearrangements and stereotyped patterns associated to the two CLL subset #2 and subset #8 (*see* **Note 11**).

3.2 IMGT/V-QUEST Results for A. Detailed View

The page "A. Detailed results for the IMGT/V-QUEST analyzed sequences" [22, 23] indicates at the top the number of analyzed sequences and the list of sequences identifiers with links allowing to browse directly the corresponding individual results. Individual results include the FASTA submitted sequence and the "Result summary" of the analysis, followed by the detailed result sections selected in the Welcome page. Importantly, the result sections allow

to explore in depth the results of the analysis regarding the identification of V, (D), J genes and alleles, the description of the V-DOMAIN with the delimitation of FR-IMGT and CDR-IMGT, and the characterization of the mutations.

3.2.1 Sequence and Result Summary

The numbers of 5′ trimmed-n and 3′ trimmed-n from the submitted sequence before the analysis if any (*see* **Note 3**), the sequence length, the sequence analysis category (*see* **Note 12**), and the IMGT reference directory set with which the sequence was compared (e.g., *Homo sapiens* (human) IG set) are indicated above the submitted sequence provided in FASTA format [22, 23] (Fig. 4). The part of the sequence corresponding to the V-DOMAIN is underlined in green. If a sequence was submitted in antisense orientation, it is complementary reversed and displayed, as well as the results, in the V gene sense orientation.

The "Result summary" provides the main characteristics of the analyzed sequence [22, 23]:

1. The evaluation of the sequence functionality: "Productive" or "Unproductive." Only productive sequences are expressed in antigen receptors.
2. The identification of the closest V, (D), and J genes and alleles: the names of the closest "V-GENE and allele" and "J-GENE and allele" are provided with their alignment score (*see* **Note 13**), the percentage of identity and the ratio of the number of identical nucleotides (nt)/number of aligned nt. The name of the closest "D-GENE and allele" determined by IMGT/JunctionAnalysis [37, 38] is indicated with the D-REGION reading frame.
3. The length of the four FR-IMGT and of the three CDR-IMGT between square brackets and separated by dots and the amino acid (AA) JUNCTION sequence.
4. The JUNCTION length (in nt) and the JUNCTION decryption [45], which describe the length (in nt) of the IMGT labels that compose the JUNCTION (*see* **Note 14**).

IMGT/V-QUEST provides warnings (not shown) that appear as notes in red to alert the user, if potential insertions or deletions are suspected in the V-REGION (*see* **Note 15**), or if other possibilities for the J-GENE and allele names are identified. Users are encouraged to check alignments in related detailed result sections.

Below the "Result summary," notes in black (not shown) may appear to indicate:

1. The number of missing nt in the 5′ part of the V-REGION and/or the number of missing nt in the 3′ part of the J-REGION in case of a partial V-DOMAIN.
2. The number of V-REGION uncertain nt number("n") within the analyzed sequence if any.

Species	Homo sapiens (human)
Receptor type or locus	IG
IMGT directory reference set	F+ORF+ in-frame P
Search for insertions and deletions	no

A. Detailed results for the IMGT/V-QUEST analysed sequences

Number of analysed sequences: 1

1. seq_1

This release of IMGT/V-QUEST uses IMGT/JunctionAnalysis for the analysis of the JUNCTION

Hyphens (-) show nucleotide identity, dots (.) represent gaps

Sequence: 1 seq_1

Analysed sequence length: 417.

Sequence analysis category: 1 (no indel search).

Sequence compared with the *Homo sapiens* (human) IG set from the IMGT reference directory (set: F+ORF+ in-frame P)

```
>seq_1
atggagtttgggctgagctggctttttcttgtggctattttaaaaggtgtccagtgtgaa
gtgcagctggtggagtctgagggaggcttggtacagcctggcaggtcccagagactctcc
tgtgcagcctctggattcacctttgatgattatgccatgcactgggtccggcaagctcca
gggaagggcctggagtgggtctcaggtattagttggaatagtggtagcataggctatgca
gactctgtgaagggccgattcaccatctccagagacaacgccaagaactccctgtatctg
caaatgaacagtctgagagctgaggacacggccttgtattactgtgcaaaggggatttt
ggagtggttaacccccttgactactggggccagggaaccctggtcaccgtctcctca
```

Result summary: seq_1	**Productive IGH rearranged sequence** (no stop codon and in-frame junction)		
V-GENE and allele	Homsap IGHV3-9*01 F	score = 1413	identity = **98.96%** (285/288 nt)
J-GENE and allele	Homsap IGHJ4*02 F	score = 208	identity = 93.62% (44/47 nt)
D-GENE and allele by IMGT/JunctionAnalysis	Homsap IGHD3-3*01 F	D-REGION is in reading frame 3	
FR-IMGT lengths, CDR-IMGT lengths and AA JUNCTION	[25.17.38.11]	[8.8.13]	CAKGIFGVVNPLDYW
JUNCTION length (in nt) and decryption	45 nt = (8)-5{3}-7(17)-7{6}-6(11)	(3'V)3'{N1}5'(D)3'{N2}5'(5'J)	

Fig. 4 IMGT/V-QUEST "Detailed results" [22, 23]. The parameters of the analysis are recalled on the top of the page. The first part the "Detailed results" for "seq_1" (IMGT/LIGM-DB [36] accession number X81732) includes the sequence in FASTA format (the first 57 nt not underlined in green are not part of the V-DOMAIN) and the "Result summary." Seq_1 functionality is "productive." This human IGHV sequence expresses the IGHV3-9*01, IGHD3-3*01, and IGHJ4*02 genes and alleles. The lengths of the four FR-IMGT are 25, 17, 38, and 11. The lengths of the three CDR-IMGT are 8, 8, and 13. The JUNCTION length is 45 nt, and the decryption [45] shows that it is composed of 8 nt for the 3′V-REGION (5 nt were trimmed from the germline V during DNA rearrangement), 3 nt for N1-REGION, and 17 nt for the D-REGION (7 nt in 5′ and 7 in 3′ were trimmed from the germline D, 6 nt for the N2-REGION, and 11 for the 5′J-REGION (6 nt were trimmed from the germline J))

3.2.2 Detailed Result Sections

If selected in the Welcome page, the 14 detailed result sections are displayed [22, 23]. They allow to verify, detail, and complete the "Result summary."

1. Detailed result sections for V, D, and J genes and alleles identification: in sections 1–3, the alignments for V, D, and J genes and alleles display the alignments of the user sequence with the five (default value in option "Nb of aligned reference sequences") closest germline V, D, and J gene alleles, respectively, with their alignment score and their identity percentage. All V or J genes and alleles with an identical highest identity percentage in alignments are solutions and are provided in the "Result summary" table (*see* **Note 16**). The alignment for D-GENE and allele should be considered with caution since it may show discrepancies with the results obtained by IMGT/JunctionAnalysis [37, 38] (*see* **Note 17**).
2. Detailed analysis of the JUNCTION: the section 4 provides the Results of IMGT/JunctionAnalysis [37, 38], which include:
   (a) The "Analysis of the JUNCTION" (Fig. 5) [22, 23] shows the details of the junction at the nucleotide level with delimitation of the IMGT labels (Fig. 2 in Subheading 2.3). Dots indicate the number of nucleotides trimmed at the germline V, D, and J gene ends. Vmut, Dmut, and Jmut indicate the number of mutations in the 3′V-REGION, D-REGION, and 5′J-REGION, respectively, and the corresponding mutated nucleotides are underlined in the sequence. "Ngc" corresponds to the ratio of the number of g+c nucleotides to the total number of N nucleotides. The JUNCTION decryption is also provided [45] (*see* **Note 14**). If selected "Eligible D genes" (not shown), all D genes, which match the junction with their corresponding score, are displayed below.
   (b) The "Translation of the JUNCTION" displays the AA JUNCTION with AA colored according to the eleven IMGT physicochemical classes [46] (*see* **Note 18**), the JUNCTION frame ('+' for in-frame, and '-' for out-of-frame), the CDR3-IMGT length, the molecular mass, the isoelectric point (pI), and a link to detailed physicochemical descriptor (not shown). Gaps (represented by dots) are inserted in "out-of-frame" JUNCTION to maintain the J-REGION frame, and the corresponding codon, which cannot be translated, is represented by "#" in AA translation (not shown).
3. The section "5. Sequence of the JUNCTION ("nt" and "AA")" provides the JUNCTION in nt and AA with IMGT unique numbering for in-frame JUNCTION, and in the

4.Results of IMGT/JunctionAnalysis

Maximum number of accepted mutations in: 3'V-REGION = 2, D-REGION = 4, 5'J-REGION = 2
Maximum number of accepted D-GENE: 1

Analysis of the JUNCTION

D-REGION is in reading frame 2

Click on mutated (underlined) nucleotide to see the original one:

Input	V name	3'V-REGION	N1	D-REGION	N2	5'J-REGION
seq_2	Homsap IGHV3-11*05 F	tgtgtgaga..	ataatc	.ggaatagcagcagctgg...	ccc	....ctttgaccagtgg

J name	D name	Vmut	Dmut	Jmut	Ngc	JUNCTION decryption
Homsap IGHJ4*02 F	Homsap IGHD6-13*01 F	1	1	2	4/9	(9)-2{6}-1(17)-3{3}-4(13)

Translation of the JUNCTION

Click on mutated (underlined) amino acid to see the original one:

	104	105	106	107	108	109	110	111	112.1	112	113	114	115	116	117	118	Frame	CDR3-IMGT length	Molecular mass	pI
	C	V	R	I	I	G	I	A	A	A	G	P	F	D	Q	W				
seq_2	tgt	gtg	aga	ata	atc	gga	ata	gca	gca	gct	ggc	ccc	ttt	gac	cag	tgg	+	14	1,717.03	6.15

Fig. 5 IMGT/V-QUEST "Detailed results" [22, 23]. Results of IMGT/JunctionAnalysis [37, 38] for AB063867 IMGT/LIGM-DB accession number. This human IGH sequence results from the rearrangement of IGHV3-11*05 F, IGHD6-13*01 F, and IGHJ4*02 F. The JUNCTION is in-frame. The length of the CDR3-IMGT is 14 AA (42 nt). The JUNCTION length is of 48 nt, and the decryption [45] shows that it is composed of 9 nt for the 3′V-REGION (2 nt were trimmed from the germline V), 6 nt for N1-REGION, and 17 nt for the D-REGION (1 nt in 5′ and 3 in 3′ were trimmed from the germline D, 3 nt for the N2-REGION, and 13 for the 5′J-REGION (4 nt was trimmed from the germline J)

FASTA format with the formatted header required as input by IMGT/JunctionAnalysis online [37, 38]. These results are provided even if IMGT/JunctionAnalysis gives no results.

4. Delimitation of the FR-IMGT and CDR-IMGT in V-REGION: the sections 6, 7 and 8 provide three displays of the V-REGION [22, 23]:
   (a) "6. V-REGION alignment according to the IMGT unique numbering" for the nt sequences with the FR-IMGT and CDR-IMGT delimitations according to the IMGT unique numbering [13].
   (b) "7. V-REGION translation" for the nt sequence and its AA translation, aligned with the closest germline V-REGION.

(c) "8. V-REGION protein display" for the AA translation of the input sequence, aligned with the V-REGION translation of the closest germline V-GENE, and with, on the third line of the alignment and shown in bold, the AA of the input sequence which are different from the closest germline V-REGION.

5. Analysis of the mutations: the sections 9, 10 and 11 are dedicated to the analysis of the nt mutations and AA changes observed in the V-REGION by comparison with the closest germline V gene and allele [22, 23]:

   (a) "9. V-REGION mutation and AA change table" lists the nt mutations and, if nonsilent, the corresponding AA changes. They are described for each FR-IMGT and CDR-IMGT with their nt and codon positions according to the IMGT unique numbering [13]. In parentheses, the "AA class Change Type" indicates if, between germline AA and replaced AA, the hydropathy, volume, and physicochemical properties have been conserved (+) or not (−) according to the IMGT physicochemical classes [46].

   (b) "10. V-REGION mutation and AA change statistics" comprises two tables for the detailed and complete characterization of nt mutations and AA changes: "Nucleotide (nt) mutations" table quantifies nt positions with or without gaps, the identical nt, the total number of mutations, and the silent and nonsilent ones for the V-REGION and per FR-IMGT and CDR-IMGT. It then details the same evaluation for the four types of transitions and of the eight types of transversions. "Amino acid (AA) changes" table quantifies the codons or amino acid positions, with or without gaps, the unchanged AA, and AA changes for the V-REGION and per FR-IMGT and CDR-IMGT (*see* **Note 19**). It then evaluates the number of changes in 4 "AA class Similarity Degree": "Very similar" (the three properties hydropathy, volume, and physicochemical properties are conserved), "Similar" (one of the three properties is changed), "Dissimilar" (two of the three properties are changed), and "Very dissimilar" (the three properties are changed).

   (c) "11. V-REGION mutation hot spots" shows the localization of the hot spot patterns (a/t)a (or wa) and (a/g) g(c/t)(a/t) (or rgyw) and their complementary reverse motifs t(a/t) (or tw) and (a/t)(a/g)c(c/t) (or wrcy) in the closest germline V gene and allele. Finally, this section includes a table for the "Correlation between V-REGION mutations, AA changes, codons changes, and hotspot motifs." It provides a synthesis for each mutation: the position in nt, the AA change and its position according

to the AA numbering [13, 16, 17], the AA class Change Type, the germline and mutated codon, and the corresponding hotspot if any. An illustration is provided in Fig. 6.

6. Sequence annotation with IMGT labels:
   (a) "12. V-REGION and V-(D)-J-REGION" provides nt and AA FASTA sequences with gaps according to the IMGT unique numbering [13, 16, 17] of the V-REGION (nt sequence with access to the IMGT/PhyloGene tool [47]) and of V-J or V-D-J-REGION. In case of out-of-frame junctions V-J or V-D-J-REGION, a note is added, and the V-J or V-D-J-REGION is shown in red.
   (b) "13. Annotation by IMGT/Automat" provides a full automatic annotation for the V-J-REGION or V-D-J-REGION by IMGT/Automat [39, 40] with IMGT labels (*see* Subheading 2.3).
7. Graphical representation of the V-DOMAIN [22, 23]: "14. IMGT Collier de Perles" allows to display the IMGT Collier de Perles for analyzed V-DOMAIN either through a "link to IMGT/Collier-de-Perles tool" [21] (*see* Subheading 7 IMGT/Collier-de-Perles) or as a direct representation integrated in IMGT/V-QUEST results depending on the user selection.

3.2.3 Sequence and Result Summary with the Search for Insertions and Deletions in V-REGION

The insertions and/or deletions that are detected by using the "Advanced parameters" and "Search for insertions" are described in the "Result summary" row [22, 23] (Fig. 7) with their localization in FR-IMGT or CDR-IMGT, the number of inserted or deleted nt, and, for insertions, the inserted nucleotides, the presence or absence of frameshift, the V-REGION codon from which the insertion or deletion starts, and the nt position in the user sequence.

3.2.4 Top of Detailed Results for the Analysis of single chain Fragment variable (scFv)

The Advanced functionality "Analysis of single chain Fragment variable (scFv)" [41] allows the analysis of scFv sequences from phage-display combinatorial libraries [48, 49]. IMGT/V-QUEST [22, 23] identifies, localizes, and characterizes the two V-DOMAIN of a scFv (Fig. 8). At the top of the result page, the number of analyzed sequences and the number of identified V-DOMAIN are indicated. V-DOMAIN identifiers are automatically generated by adding to the sequence identifier a suffix composed of an underscore plus a letter for the locus (H, K, L for IGH, IGK, IGL or A, B, D, G for TRA, TRB, TRD, TRG, respectively). Below the list of V-DOMAIN identifiers is a table that indicates the positions of each V-DOMAIN and of the linker in the identified scFv. The detailed analysis of each individual V-DOMAIN is then provided classically.

a

7. V-REGION translation

```
                                                                  FR1 - IMGT
                  <------------------------------------------
                  1                   5                  10                  15
                   E   V   Q   L   V   Q   S   G   A       E   V   K   K   P
seq_3             gag gtg cag ctg gtg cag tct gga gca ... gag gtg aaa aag ccc

M99686 Homsap IGHV5-51*01 F --- --- --- --- --- --- --- --- --- ... --- --- --- --- ---


                  ------------------------------------------>_________________
                                      20                  25                  30
                   G   E   S   L   K   I   S   C   K   A   S   G   Y   S   S
seq_3             gga gag tcc ctg aag atc tcc tgt aag gct tct gga tac agt tct
                                                      G                       F
M99686 Homsap IGHV5-51*01 F --g --- --t --- --- --- --- --- --- -g- --- --- --- --c -t-
```

b

11. V-REGION mutation hotspots

Hotspots motifs and localizations in germline V-REGION

(a/t)a
wa

Motif	Localization
aa	37-38 (FR1)
aa	38-39 (FR1)
aa	39-40 (FR1)
aa	40-41 (FR1)
aa	58-59 (FR1)
ta	69-70 (FR1)
aa	70-71 (FR1)

t(a/t)
tw

Motif	Localization
ta	69-70 (FR1)
tt	75-76 (FR1)

(a/g)g(c/t)(a/t)
rgyw

Motif	Localization
agct	8-11 (FR1)
agca	24-27 (FR1)
ggtt	73-76 (FR1)

(a/t)(a/g)c(c/t)
wrcy

Motif	Localization
agct	8-11 (FR1)
agcc	41-44 (FR1)

Correlation between V-REGION mutations, AA changes, codons changes and hotspots motifs

FR1-IMGT
g48>a, G16 ; G16 ggg 46-48>G gga
t54>c, S18 ; S18 tct 52-54>S tcc
g74>c, G25>A (- + -) ; G25 ggt 73-75[ggtt 73-76]>A gct

Fig. 6 IMGT/V-QUEST "Detailed results" [22, 23]. Correlation between V-REGION mutations, AA changes, codons changes, and hotspots motifs in FR1-IMGT of seq_3 (accession number AJ006165 of IMGT/LIGM DB [36]). (**a**) "7. V-REGION translation" and (**b**) "11. V-REGION mutation hotspots." Only the FR1-IMGT parts of the results are displayed: the two silent mutations g48>a (G16) and t54>c (S18) are shown in green and light blue rectangles, respectively. The nonsilent mutation g74>c (shown in the red rectangle) leads to the AA dissimilar change G25>A (hydropathy and physicochemical properties are not conserved). The codon "ggt" in position 73-75 is changed in "gct." The nt mutation occurs in the hotspot "ggtt" in position 73–76

Species	Homo sapiens (human)
Receptor type or locus	IG
IMGT directory reference set	F+ORF+ in-frame P
Search for insertions and deletions	yes

A. Detailed results for the IMGT/V-QUEST analysed sequences

Number of analysed sequences: 1

1. seq_4

This release of IMGT/V-QUEST uses IMGT/JunctionAnalysis for the analysis of the JUNCTION

Hyphens (-) show nucleotide identity, dots (.) represent gaps

Sequence: 1 seq_4

Analysed sequence length: 379.
Sequence analysis category: 2 (indel search & correction).
Sequence compared with the *Homo sapiens* (human) IG set from the IMGT reference directory (set: F+ORF+ in-frame P)

>seq_4
caggtgcagctacagcagtggggcgcaggactgttgaagccttcggagaccctgtccctc
acctgcgctgtctatggtgggtccttcagtggttactactggagctggatccgccagtcc
ccagagacggggctggagtggcttggcgaaTTCGATCTTGGTGGAAGCatcactcatagt
agaggcaccaactacaacccgtcgctcaagagtcgagtcaccatctcaggagacacgtcc
aagaaccagttctccctgaaactgacctctgtgaccgccgcggacacggctgtctattac
tgtgcgagaggtttagcaatgggtggaactaaggagtttgactcctggggccagggaacc
ctggtcaccgtctcctcag

Result summary: seq_4

Nucleotide insertions have been detected and automatically removed for this analysis: they are displayed as capital letters in the user submitted sequence above.

localization in V-REGION	nb of inserted nt	inserted nt	causing frameshift	from V-REGION codon	from nt position in user submitted sequence
CDR2-IMGT	18	TTCGATCTTGGTGGAAGC	no	56	151

IMGT/V-QUEST results after removal of the insertion(s)
Potentially productive IGH rearranged sequence (no stop codon and in-frame junction)
(Check also your sequence with BLAST against IMGT/GENE-DB reference sequences to eventually identify out-of-frame pseudogenes)

V-GENE and allele	Homsap IGHV4-34*01 F, or Homsap IGHV4-34*12 F	score = 1299	identity = 95.09% (271/285 nt) [94.74% (270/285 nt)]
J-GENE and allele	Homsap IGHJ4*02 F (a)	score = 204	identity = 91.67% (44/48 nt)
D-GENE and allele by IMGT/JunctionAnalysis	Homsap IGHD6-19*01 F	D-REGION is in reading frame 2	
FR-IMGT lengths, CDR-IMGT lengths and AA JUNCTION	[25.17.38.11]	[8.7.14]	CARGLAMGGTKEFDSW
JUNCTION length (in nt) and decryption	48 nt = (11)0{2}-5(16)0{7}-5(12)	(3'V)3'{N1}5'(D)3'{N2}5'(5'J)	

(a) Other possibilities: Homsap_IGHJ5*02 (highest number of consecutive identical nucleotides)

Fig. 7 IMGT/V-QUEST "Detailed results" [22, 23]. Sequence and Result summary with "Search for insertions and deletions." An insertion of 18 nt is identified in seq_4 (IMGT/LIGM-DB [36] accession number MG950400) from CDR2-IMGT position 56 (from nt 151 in the submitted sequence). The insertion is shown in capital letters

3.3 IMGT/V-QUEST Results for B. Synthesis View

At the top of the page, the parameters used for the analysis are recalled, and the number of analyzed sequences is indicated. The results include a summary table and potentially eight detailed result sections if selected by the user.

3.3.1 Summary Table

The "Summary table" (Fig. 9) displays one row for each input sequence with the corresponding results, including 22 columns [22, 23]: (1) the sequence order in the submission; (2) the sequence identifier (Sequence ID); (3) the name of the closest V-GENE and allele; (4) the functionality of the sequence (when found, the presence of stop codons is indicated); (5) the V-REGION score; (6) the V-REGION percentage of identity with, between parentheses, the ratio of number of identical nucleotides (nt)/number of aligned nt; (7) the name of the closest J-GENE and allele; (8) the J-REGION score; (9) the J-REGION percentage of identity and the ratio of number of identical nucleotides (nt)/number of aligned nt; and provided according to the IMGT/JunctionAnalysis results [37, 38] (10) the D-GENE and allele name; (11) the D reading frame; (12) the CDR-IMGT lengths; (13) the AA JUNCTION; and (14) the JUNCTION frame (in the absence of results of IMGT/JunctionAnalysis, only the AA JUNCTION defined by IMGT/V-QUEST [22, 23] is displayed); (15) the JUNCTION nt length and decryption [45]; (16) the number of missing in 5′ partial V-REGION; (17) the number of uncertain nt; (18) the number of missing nt in 3′ partial J-REGION; (19) and (20) the numbers of 5′ and 3′ trimmed 'n' nucleotides; (21) the length of the sequence; and (22) the sequence analysis category (*see* **Note 12**). Clicking on the sequence ID provides the corresponding Detailed View in a separate tab (depending on your browser). Warnings in red may be indicated to highlight specific features of the sequence (Fig. 9).

3.3.2 Detailed Analysis of the JUNCTION

A link to access the IMGT/JunctionAnalysis [37, 38] results is provided for sequences of the same locus. AA translations are aligned on the longest CDR3 length according to the IMGT unique numbering [13] (Fig. 10).

3.3.3 Detailed Result Sections for Alignment of Sequences Expressing the Same V Gene and Allele

In "Alignment with the closest alleles" below the summary table, the V genes and alleles are listed with the number of assigned sequences in parentheses [22, 23]. Click on the associated link to reach the corresponding detailed result sections. They provide six different displays (if all were selected) of alignment of sequences

Fig. 7 (continued) in the FASTA sequence. IMGT/V-QUEST then performs a classical analysis (for gene and allele identification, analysis of the JUNCTION, evaluation of nt mutation, and AA changes) after removal of the insertion(s) and addition of gaps to replace the deletions. The evaluation of the identity percentage added in square brackets includes each insertion or deletion as an additional mutation

Species	Homo sapiens (human)
Receptor type or locus	IG
IMGT directory reference set	F+ORF+ in-frame P
Search for insertions and deletions	no
Analysis of scFv	yes

A. Detailed results for the IMGT/V-QUEST analysed sequences

Number of analysed sequences: 2
Number of analysed V-DOMAIN: 4

1. scFv_1_H, 2. scFv_1_K, 3. scFv_2_H, 4. scFv_2_K

Identified scFv:

Sequence ID	5'V-DOMAIN ID	5'V-DOMAIN positions	5'V-DOMAIN length	linker positions	linker length	3'V-DOMAIN ID	3'V-DOMAIN positions	3'V-DOMAIN length
scFv_1	1_scFv_1_H	1..349	349	350..390	41	2_scFv_1_K	391..714	324
scFv_2	3_scFv_2_H	1..361	361	362..405	44	4_scFv_2_K	406..727	322

This release of IMGT/V-QUEST uses IMGT/JunctionAnalysis for the analysis of the JUNCTION
Hyphens (-) show nucleotide identity, dots (.) represent gaps

V-DOMAIN: 1 scFv_1_H (associated V-DOMAIN: 2 scFv_1_K)

Analysed sequence length: 714
Sequence analysis category: 1 (no indel search).
Sequence compared with the *Homo sapiens* (human) IG set from the IMGT reference directory (set: F+ORF+ in-frame P)

>scFv_1_H
gaggtgcagctgttggagtctgggggaggcttggtacagcctgggggtccctgagactc
tcctgtgcagcctctggattcacctttagcagctatgccatgagctgggtccgccaggct
ccagggaaggggctggagtgggtctcagctattagtggtagtggtggtagcacatactac
gcagactccgtgaagggccggttcaccatctccagagacaattccaagaacacgctgtat
ctgcaaatgaacagcctgagagccgaggacacggccgtatattactgtgcgaaatctctt
cttctttttgactactggggccagggaaccctggtcaccgtctcgagtggcgatgggtcc
agtggcggtagcgggggcgcgtcgactggcgaaattgtgttgacgcagtctccaggcacc
ctgtctttgtctccaggggaaagagccaccctctcctgcagggccagtcagagtgttagc
agcagctacttagcctggtaccagcagaaacctggccaggctcccaggctcctcatctat
ggtgcatccagcagggccactggcatcccagacaggttcagtggcagtgggtctgggaca
gacttcactctcaccatcagcagactggagcctgaagattttgcagtgtattactgtcag
cagtggggtgagaagcccttgacgttcggccaagggaccaaggtggaaatcaaa

Result summary: scFv_1_H	**Productive IGH rearranged sequence** (no stop codon and in-frame junction)		
V-GENE and allele	Homsap IGHV3-23*01 F, or Homsap IGHV3-23D*01 F	score = 1440	identity = **100.00%** (288/288 nt)
J-GENE and allele	Homsap IGHJ4*02 F	score = 159	identity = 81.25% (39/48 nt)
D-GENE and allele by IMGT/JunctionAnalysis	Homsap IGHD2-15*01 F	D-REGION is in reading frame 1	
FR-IMGT lengths, CDR-IMGT lengths and AA JUNCTION	[25.17.38.11]	[8.8.9]	CAKSLLLFDYW
JUNCTION length (in nt) and decryption	33 nt = (9)-2{3}-21(8)-2{1}-5(12)	(3'V)3'{N1}5'(D)3'{N2}5'(5'J)	

Fig. 8 IMGT/V-QUEST "Detailed results" [22, 23]. Top of Detailed results for the "Advanced functionality" "Analysis of single chain fragment variable (scFv)". scFv_1 and scFv_2 sequences correspond to the accession numbers AJ006120 and AF117956 in the IMGT/LIGM-DB database [36]. The 5′V-DOMAIN are VH and 3′ V-DOMAIN are V-KAPPA for both scFv. The detailed results are then provided for each domain

B. Synthesis for the IMGT/V-QUEST analysed sequences

Number of analysed sequences: 4

Sequence compared with the *Homo sapiens* (human) IG set from the IMGT reference directory (set: F+ORF+ in-frame P)

Summary table:

a

Sequence Number	Sequence ID	V-GENE and allele	V-DOMAIN Functionality	V-REGION score	V-REGION identity % (nt)	J-GENE and allele	J-REGION score	J-REGION identity % (nt)	D-GENE and allele	D-REGION reading frame	CDR-IMGT lengths
1	AB021529	Homsap IGHV3-9*01 F	productive	1359	96.88% (279/288 nt)	Homsap IGHJ4*02 F	129	78.57% (33/42 nt)	Homsap IGHD3-16*01 F	2	[8.8.14]
2	X81732	Homsap IGHV3-9*01 F	productive	1413	98.96% (285/288 nt)	Homsap IGHJ4*02 F	208	93.62% (44/47 nt)	Homsap IGHD3-3*01 F	3	[8.8.13]
3	AB245095	Homsap IGHV5-51*01 F	productive	1251	92.71% (267/288 nt)	Homsap IGHJ4*02 F, or Homsap IGHJ4*03 F (a)	141	77.08% (37/48 nt)	Homsap IGHD5-12*01 F	2	[8.8.13]
4	AJ006165	Homsap IGHV5-51*01 F	productive	1287	94.10% (271/288 nt)	Homsap IGHJ6*02 F	238	87.10% (54/62 nt)	Homsap IGHD3-10*01 F	2	[8.8.15]

b

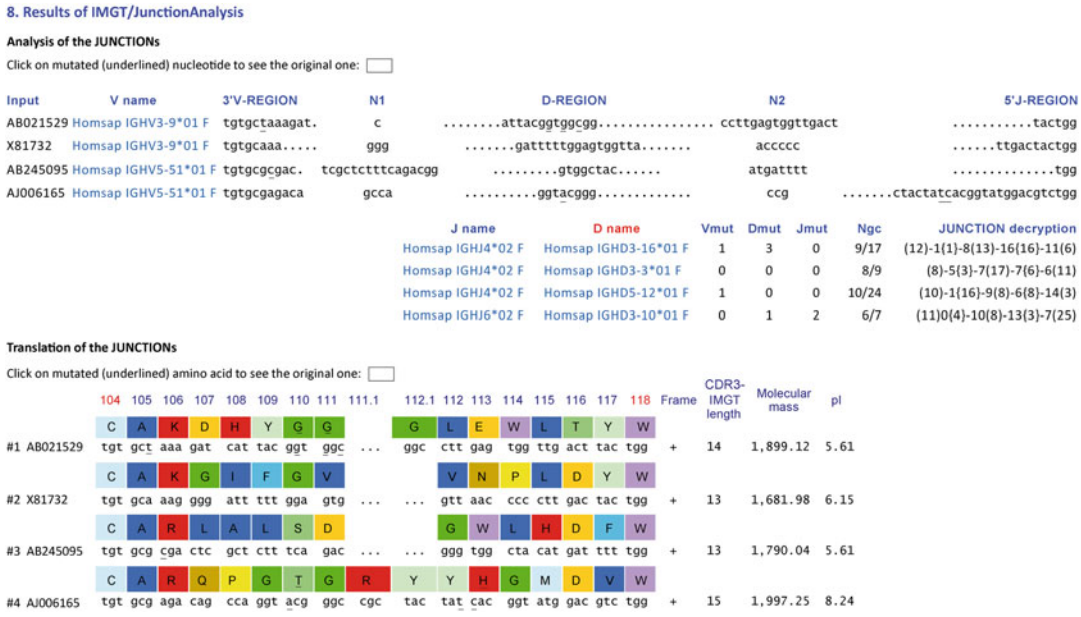
	AA JUNCTION	JUNCTION frame	JUNCTION length (in nt) and decryption	V-REGION partial 5prime missing nt nb	V-REGION uncertain nt nb	J-REGION partial 3prime missing nt nb	5prime trimmed-n nb	3prime trimmed-n nb	Analysed sequence length	Sequence analysis category
1	CAKDHYGGGLEWLTYW	in-frame	48 nt = (12)-1(1)-8(13)-16(16)-11(6)	0	0	5	-	-	358	1 (noindelsearch)
2	CAKGIFGVVNPLDYW	in-frame	45 nt = (8)-5(3)-7(17)-7(6)-6(11)	0	0	0	-	-	417	1 (noindelsearch)
3	CARLALSDGWLHDFW	in-frame	45 nt = (10)-1(16)-9(8)-6(8)-14(3)	0	0	0	-	-	684	1 (noindelsearch)
4	CARQPGTGRYYHGMDVW	in-frame	51 nt = (11)0(4)-10(8)-13(3)-7(25)	0	0	0	-	-	412	1 (noindelsearch)

Fig. 9 IMGT/V-QUEST Synthesis view [22, 23]. (**a**) The 12 first columns of the "Summary table": the four analyzed sequences are shown in the "V-GENE and allele name" order in the Summary table. The sequences ID are accession numbers of IMGT/LIGM-DB [36]. The hyperlinks allow to get the corresponding results in "A Detailed view." The two sequences assigned to IGHV3–9 and the two sequences assigned to IGHV5-51 will be, respectively, aligned together in the detailed result sections 1 to 6. In the column "J-GENE and allele," a warning "(**a**)" in red indicates that other IGHJ genes and alleles may be solutions for the sequence 3 (not shown). (**b**) The last 12 columns of the "Summary table": the V-DOMAIN of sequence 1 is partial: 5 nt are missing in the 3′ part of the J-REGION

that express the same V gene and alleles: "1. Alignment for V-GENE," "2. V-REGION alignment according to the IMGT unique numbering" [13], "3. V-REGION translation," and three different formats for the "V-REGION protein display." Section "7. V-REGION most frequently occurring AA per position and per FR-IMGT and CDR-IMGT" shows the most frequent AA in sequences expressing the same V genes and alleles per FR-IMGT and CDR-IMGT and per position according to the IMGT unique numbering [13].

3.4 IMGT/V-QUEST Output for Excel File

"Excel file" allows the users to open and save a spreadsheet including the results of the IMGT/V-QUEST analysis [22, 23]. The file contains 11 sheets or 12 for the Advanced Functionality "Analysis of single chain Fragment variable (scFv)" [41] (*see* Subheading 4.3 for the detail of their content in IMGT/HighV-QUEST sequence analysis results).



8. Results of IMGT/JunctionAnalysis

Analysis of the JUNCTIONs

Click on mutated (underlined) nucleotide to see the original one:

Input	V name	3'V-REGION	N1	D-REGION	N2	5'J-REGION
AB021529	Homsap IGHV3-9*01 F	tgtgctaaagat.	c	.....attacggtggcgg.....	ccttgagtggttgact	..........tactgg
X81732	Homsap IGHV3-9*01 F	tgtgcaaa.....	ggg	.....gatttttggagtggtta.....	accccc	.....ttgactactgg
AB245095	Homsap IGHV5-51*01 F	tgtgcgcgac.	tcgctctttcagacgg	.....gtggctac.....	atgatttt	.....tgg
AJ006165	Homsap IGHV5-51*01 F	tgtgcgagaca	gcca	.....ggtacggg.....	ccg	.....ctactatcacggtatggacgtctgg

J name	D name	Vmut	Dmut	Jmut	Ngc	JUNCTION decryption
Homsap IGHJ4*02 F	Homsap IGHD3-16*01 F	1	3	0	9/17	(12)-1(1)-8(13)-16(16)-11(6)
Homsap IGHJ4*02 F	Homsap IGHD3-3*01 F	0	0	0	8/9	(8)-5(3)-7(17)-7(6)-6(11)
Homsap IGHJ4*02 F	Homsap IGHD5-12*01 F	1	0	0	10/24	(10)-1(16)-9(8)-6(8)-14(3)
Homsap IGHJ6*02 F	Homsap IGHD3-10*01 F	0	1	2	6/7	(11)0(4)-10(8)-13(3)-7(25)

Translation of the JUNCTIONs

Click on mutated (underlined) amino acid to see the original one:

	104	105	106	107	108	109	110	111	111.1		112.1	112	113	114	115	116	117	118	Frame	CDR3-IMGT length	Molecular mass	pI
	C	A	K	D	H	Y	G	G			G	L	E	W	L	T	Y	W				
#1 AB021529	tgt	gct	aaa	gat	cat	tac	ggt	ggc	...		ggc	ctt	gag	tgg	ttg	act	tac	tgg	+	14	1,899.12	5.61
	C	A	K	G	I	F	G	V				V	N	P	L	D	Y	W				
#2 X81732	tgt	gca	aag	ggg	att	ttt	gga	gtg	...		...	gtt	aac	ccc	ctt	gac	tac	tgg	+	13	1,681.98	6.15
	C	A	R	L	A	L	S	D				G	W	L	H	D	F	W				
#3 AB245095	tgt	gcg	cga	ctc	gct	ctt	tca	gac	...		...	ggg	tgg	cta	cat	gat	ttt	tgg	+	13	1,790.04	5.61
	C	A	R	Q	P	G	T	G	R		Y	Y	H	G	M	D	V	W				
#4 AJ006165	tgt	gcg	aga	cag	cca	ggt	acg	ggc	cgc		tac	tat	cac	ggt	atg	gac	gtc	tgg	+	15	1,997.25	8.24

Fig. 10 IMGT/V-QUEST Synthesis view [22, 23]. Results of IMGT/JunctionAnalysis for four IGH junctions [37, 38]: "Analysis of the JUNCTIONs" displays for the four junctions, the sequences of the IMGT labels 3′V-REGION, N1, D-REGION, N2, and 5′J-REGION. The mutated nt are underlined. "Translation of the JUNCTIONs" displays the four junctions aligned per position according to the IMGT unique numbering [13]

4 IMGT/HighV-QUEST

IMGT/HighV-QUEST [24–27] is the high throughput version of IMGT/V-QUEST [22, 23]. It is freely available for academics, but it requires the user's registration. This allows the tool to automatically notify the users on the availability of the results. A link to the "New user" form is provided in the IMGT/HighV-QUEST welcome page. When the user logs in, the tool uses reCAPTCHA (https://developers.google.com/recaptcha) to protect the site from spam and abuse.

IMGT/HighV-QUEST provides two main functionalities [24–27]:

1. The high throughput analysis of IG and TR rearranged sequences based on the IMGT/V-QUEST algorithm (use the "IMGT/HighV-QUEST Search page" for sequence submission (*see* Subheading 4.1) and use the "Analysis history" page for the download of sequence analysis results (*see* Subheading 4.2)).
2. The evaluation of the diversity and of the expression of the IMGT clonotypes (AA) in analyzed sequence sets (use the "Launch statistics" page for IMGT clonotypes evaluation (*see*

Subheading 4.4), and use "Statistics history" page for the download of IMGT clonotypes results (*see* Subheading 4.5)).

Links to the four pages are displayed on the top of the IMGT/HighV-QUEST web interface.

4.1 IMGT/HighV-QUEST Sequence Set Submission

IMGT/HighV-QUEST Search page (Fig. 11) is provided when the user logs in. It includes the four following sections [24–27]:

4.1.1 The Sequence Submission Form

1. Provide an analysis title, select the species (*see* **Note 1**), and the receptor type or locus (*see* **Note 2**) as for IMGT/V-QUEST.
2. Upload a simple text-formatted file containing your FASTA sequences (up to 1,000,000 of IG or TR rearranged sequences can be submitted in a single run).
3. When an analysis is launched ("Start" button), it is firstly dispatched and queued on the IMGT servers and is then performed depending on the available resources. Choose to be notified by e-mail "when analysis is queued" and/or "when analysis is completed" (selected by default).

4.1.2 Display Results

1. Select Result format: the default "CSV" result format includes 11 (or 12 with the Advanced Functionality "Analysis of single chain Fragment variable (scFv)") CSV files equivalent to those provided by the "Excel file" of IMGT/V-QUEST. Result format "AIRR" [42, 43] (*see* **Note 7**) or "Both formats" can be selected.
2. The individual result files (equivalent to IMGT/V-QUEST "Detailed view" in text format) can be included in the results for submissions of maximally 200,000 sequences only.

4.1.3 Advanced Parameters

The analysis can be customized with exactly the same advanced parameters as proposed by IMGT/V-QUEST (*see* Subheading 3.1.4).

4.1.4 Advanced Functionalities

"Analysis of single-chain Fragment variable (scFv)" [41] can be included in the analysis (default is "no") (*see* Subheading 3.1.5).

4.2 IMGT/HighV-QUEST Analysis History Page: Follow-Up and Download of Results

The "Analysis history" page allows the user to check the status of the submitted analyses [24–27]. A table displays for each of them its title, its status (queued, running, or completed), the submission date, the number of submitted sequences, the species and the receptor type or locus (as selected by the user), and the actions that can be performed. When the analysis is completed, the user can download the results as a single archive file in TXZ format (commonly supported by archive tools for windows and other operating systems). The availability of the results is guaranteed for two weeks

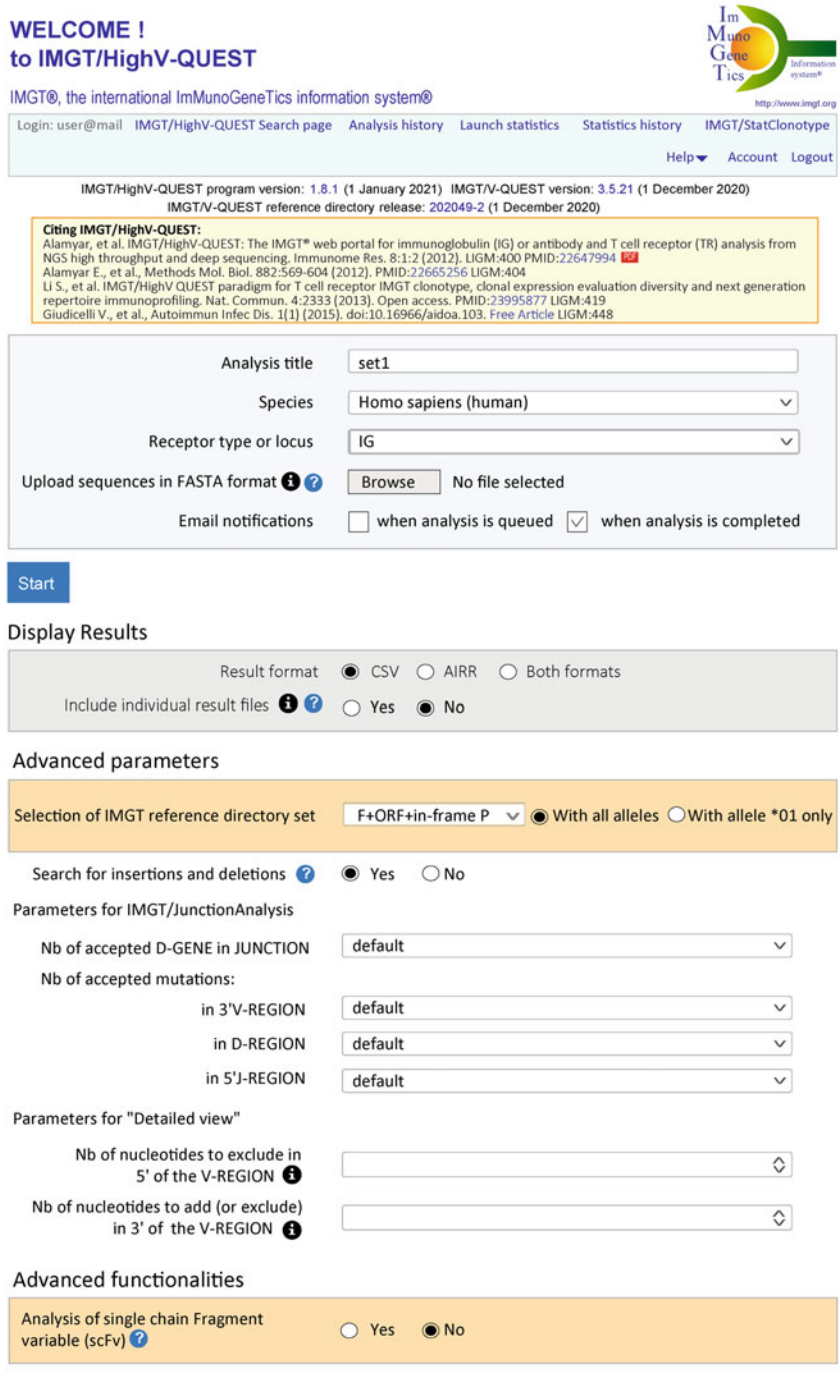


Fig. 11 The IMGT/HighV-QUEST Search page [24–27]

after the analysis is completed. After that, the files can be removed by the system. In that case, "File removed" is indicated in red instead of the archive logo.

A user may delete an analysis at any time except if it is used by the second module "Statistics" of IMGT/HighV-QUEST. In such cases, "Used by Statistics" is indicated in place of the "delete" button.

4.3 IMGT/HighV-QUEST Sequence Analysis Results

The content of the TXZ file depends on the selected "Result format" ("CSV," "AIRR," or "Both formats") [24–27]:

1. "CSV" format contains a tar folder (which needs to be extracted by an archive tool) with 11 (or 12) files (equivalent to the results of the excel file provided by the classical IMGT/V-QUEST) in CSV format, and, if selected in the IMGT/HighV-QUEST Search page, one subfolder with individual result files, in text format for each sequence (equivalent to the classical IMGT/V-QUEST "Detailed view" results (*see* Subheading 3.2)).

   The content of each CSV file is indicated in Table 1 [27].
2. "AIRR" format contains a "vquest_airr.tsv" file, generated by the tool in AIRR format [42, 43] (described in the IMGT/V-QUEST [22, 23] Documentation http://www.imgt.org/IMGT_vquest/vquest_airr) and the "11_Parameters.txt" for the parameters used for the analysis.
3. "Both formats" selection includes "CSV" and "AIRR" format results.

4.4 IMGT/HighV-QUEST Launch Statistics Page for the Evaluation of IMGT Clonotypes

An IMGT clonotype (AA) is defined by a unique V-(D)-J-rearrangement (V and J genes and alleles), with a unique CDR3-IMGT amino acid sequence and the presence of the conserved anchors C 104 and W/F 118 [26]. An IMGT clonotype (AA) is linked to one or more IMGT clonotype (nt): they are defined by a unique V-(D)-J-rearrangement with a unique CDR3-IMGT nucleotide sequence, whose translation corresponds to the CDR3-IMGT of the IMGT clonotype (AA) [26]. When IMGT/HighV-QUEST "Statistics" is launched, the tool evaluates IMGT clonotypes in batches of analyzed sequence sets per locus and provides immunoprofiles for IMGT clonotypes (AA) diversity (number of different IMGT clonotypes per V, D, and J gene and allele) and expression (number of sequences assigned to IMGT clonotypes (AA) per V, D, and J gene and allele) [26]. Moreover, IMGT/HighV-QUEST can perform the comparison of multiple batches and provide the list of the IMGT clonotype (AA), which are common to two or more batches [26].

Table 1
List of the IMGT/HighV-QUEST CSV files with the number of columns and result content [27]

File number	Result type	File name	Number of columns	Results content (*see* Note 20)
#1	Gene and allele identification, summary of the analysis, characterization of insertions and deletions	"Summary"	33 (or 29)	Result overview: · Sequence order and sequence ID · V-DOMAIN Functionality · Identification of V, D, and J genes and alleles · Identity percentage with the closest V and J genes and alleles and alignment scores · FR-IMGT and CDR-IMGT lengths · Amino acid (AA) JUNCTION · Description of insertions and deletions (indels) in V-REGION if any · User sequence in the direct orientation · Sequence orientation at the submission, the number of trimmed "n" before analysis if any, sequence length, sequence analysis category This file may include notes regarding the evaluation of the functionality, the V, D, and J gene and allele identification, potential indels in V-REGION in three dedicated columns (V-DOMAIN Functionality comment, V-REGION potential ins/del, J-GENE, and allele comment)
#2	Sequence description and annotation	"IMGT-gapped-nt-sequences"	18	Sequences for main labels of the gapped nt V-(D)-J-REGION: · Nucleotide (nt) sequences gapped according to the IMGT unique numbering for V-D-J-REGION, V-J-REGION, V-REGION, FR1-IMGT, CDR1-IMGT, FR2-IMGT, CDR2-IMGT, FR3-IMGT · nt sequences of CDR3-IMGT, JUNCTION, J-REGION and FR4-IMGT

(continued)

Table 1
(continued)

File number	Result type	File name	Number of columns	Results content (*see* Note 20)
#3		"Nt-sequences"	118	Full annotation of V-(D)-J-REGION nucleotide sequence with IMGT labels: · nt sequences of all labels that can be automatically annotated by IMGT/Automat · Start and end positions of annotated labels The four last columns evaluate the V-REGION reading frame, the number of missing 5′ and 3′ nt for partial V-(D)-J-REGION, and the number of uncertain nt in V-REGION
#4		"IMGT-gapped-AA-sequences"	18	Sequences for main IMGT labels of the gapped AA V-(D)-J-REGION: · AA sequences gapped according to the IMGT unique numbering for the labels V-D-J-REGION, V-J-REGION, V-REGION, FR1-IMGT, CDR1-IMGT, FR2-IMGT, CDR2-IMGT, FR3-IMGT · AA sequences of CDR3-IMGT, JUNCTION, J-REGION, and FR4-IMGT
#5		"AA-sequences"	18	Sequences for main IMGT labels of the non-gapped AA V-(D)-J-REGION: · Same columns as "IMGT-gapped-AA-sequences" (#4), but sequences of IMGT labels are without IMGT gaps
#6		"Junction"	85	Results of IMGT/JunctionAnalysis [37, 38]: 37 columns for IGL, IGK, TRA and TRG sequences, 51 (if one D), 63 (if two D) or 78 (if 3 D) columns for IGH, TRB, and TRD sequences

#7	Analysis of mutations	"V-REGION-mutation-and-AA-change table"	11	Correlation between V-REGION mutations, AA changes [46], codons changes and hotspots motifs Description of the mutations for V-REGION, FR1-IMGT, CDR1-IMGT, FR2-IMGT, CDR2-IMGT, FR3-IMGT and germline CDR3-IMGT, each of them characterized by: the nt mutation, the AA changes and the 3 AA class identity (+) or change (−), the codon change, and the corresponding hotspot with their localization
#8		"V-REGION-nt-mutation-statistics"	130	Characteristics and number of nt mutations: · Number of nt positions including IMGT gaps, number of nt, number of identical nt, total number of mutations, number of silent mutations, and number of nonsilent mutations · Number of transitions (a>g, g>a, c>t, t>c) and number of transversions (a>c, c>a, a>t, t>a, g>c, c>g, g>t, t>g) for V-REGION, FR1-IMGT, CDR1-IMGT, FR2-IMGT, CDR2-IMGT, FR3-IMGT, and germline CDR3-IMGT
#9		"V-REGION-AA-change-statistics"	109	Number of AA positions including IMGT gaps, number of AA, number of identical AA, total number of AA changes, number of AA changes according to the AA class Change Type (+++, ++−, +−+, +−−, −+−, −−+, −−), and number of AA class changes according to AA class Similarity Degree (Very similar, Similar, Dissimilar, and Very dissimilar) for V-REGION, FR1-IMGT, CDR1-IMGT, FR2-IMGT, CDR2-IMGT, FR3-IMGT, and germline CDR3-IMGT [46]

(continued)

Table 1
(continued)

File number	Result type	File name	Number of columns	Results content (*see* Note 20)
#10		"V-REGION-mutation-hotspots"	8	Hotspot motifs (a/t)a, t(a/t), (a/g)g(c/t)(a/t), and (a/t)(a/g)c(c/t) detected in the closest germline V-REGION with their localization in FR-IMGT and CDR-IMGT
#11	"Parameters"	"Parameters"		· Date of the analysis · IMGT/V-QUEST program version, IMGT/V-QUEST reference directory release · Parameters used for the analysis: species, receptor type or locus, IMGT reference directory set, advanced parameters, advanced functionalities
#12	Sequence description	scFv	40	Available only for the advanced functionality "Analysis of single chain Fragment variable (scFv)," one line per scFv: Positions and length of the V-(D)-J-REGION, CDR_lengths, JUNCTION for the 2 V-DOMAIN of the scFv, positions and length of the linker [41]

For launching statistics, the following steps should be followed:

1. Provide a statistical analysis title.
2. Indicate if an email notification should be sent when the statistical analysis is completed.
3. Provide comments on the analysis (optional).
4. Choose if the "Multiple batch comparison" will be performed (yes is selected by default).
5. Define a list of batches: in order to define a batch, click on the "show" button in the form "Define a batch" of the page. It allows to list the available sequences analyses already performed by IMGT/HighV-QUEST. They are displayed in a table with their title, the user name, the status of the analysis, the number of submitted sequences, the species and receptor type of locus, and the main information for analysis parameters (IMGT reference directory set and Search for insertions/deletions) (*see* **Note 21**). Provide a short title for the batch (six characters or less) before adding it to the list. Up to 15 batches can be defined.
6. Click on the start button to launch the run.

4.5 IMGT/HighV-QUEST Statistics History Page: Follow-Up and Download of Statistics

It allows to follow the status of the submitted statistics analysis and to download the results once completed [24–27]. The IMGT/HighV-QUEST statistical output is provided as a zip file.

1. Extract the archive.
2. Open the file "open_to_start.html" localized in the main folder with a web browser.

4.5.1 Results Sections to be Displayed in the User Web Browser

The IMGT/HighV-QUEST statistics output [26] is organized in the sections listed in Table 2 (see also http://www.imgt.org/HighV-QUEST/doc.action#statistical-outputs-results).

The illustration of the content of file 4.2.1 is shown in Fig. 12: it shows the first seven most expressed IMGT clonotypes (AA) of a list of 27,080.

4.5.2 "Data" Directory

Importantly, the archive includes a "data" directory: it contains text files named 'stats_xxx' where 'xxx' is composed of 'batch name'_'-locus'. They include the list of all the IMGT clonotypes (AA) (that are displayed through html sections of Table 2) and their characteristics separated by tabulations. These files include the fields needed by the external IMGT/StatClonotype [28, 29] tool (*see* Subheading 5). Their content is described in the IMGT/HighV-QUEST Documentation at http://www.imgt.org/HighV-QUEST/doc.action#datastatsxxx.

Table 2
Documents included in IMGT/HighV-QUEST statistics output [26]

Documents	File type	Content description
1. "Selected parameters" and "batch list table"	html	Species, Receptor type or locus, IMGT reference directory set, Search for insertions/deletions (yes or no), the total number of sequences, Batch IMGT clonotype comparison (yes or no), and then the list of batches with the titles of the sequence analyses, the number of sequences, the species, Receptor type (or locus), the program version of IMGT/HighV-QUEST and IMGT/V-QUEST, and the release of the IMGT reference directory
2. Result summary for batches	html	List of batches including their ID, nb of sequences, of "1 copy," "1 copy with indels," "More than 1," "More than 1 with indels" (*see* **Note 22**), the number of sequences with no J-GENE, No junction, Warnings, Unknown functionality, and with No results
3. Result summary for IMGT clonotypes (AA)	html	Number of IMGT clonotypes (AA), of assigned sequences, number of in-frame sequences not assigned to IMGT clonotypes (AA), number of productive sequences, in-frame unproductive sequences, out-of-frame sequences, sequences "1 copy" + "More than 1," "single gene," "several genes," and of submitted sequences per batch and per locus
4. Detailed results per batch	html	
4.1 "Results categories" and V, D, and J genes and alleles for genotype analysis ("1 copy" "single gene" for V and J)	zip	Includes five pdf reports including the list of filtered out sequences, the number of "1 copy single gene" and of "1 copy several genes," and a folder of graphics
4.2 Detailed IMGT clonotype (AA and nt) results per locus	html	
4.2.1 IMGT clonotypes (AA) per Nb	html	List of IMGT clonotypes (AA) ordered by decreasing number of assigned sequences with the IMGT clonotype (AA) definition, the IMGT clonotype (AA) representative sequence, and access of corresponding FASTA "1 copy" sequences
4.2.2 IMGT clonotypes (AA) per number with detailed clonotypes (nt)	html	Same as 4.2.1 with associated IMGT clonotype (nt)
4.2.3 IMGT clonotypes (AA) per V gene	html	Identical to 4.2.1 ordered by V gene

(continued)

Table 2
(continued)

Documents	File type	Content description
4.2.4 IMGT clonotypes (AA) per V gene with detailed clonotypes (nt)	html	Identical to 4.2.2 ordered by V gene
4.2.5 IMGT clonotypes (AA) per CDR3-IMGT length (AA)	html	Identical to 4.2.1 ordered by CDR3 length
4.2.6 IMGT clonotypes (AA) per CDR3-IMGT length (AA) with detailed clonotypes (nt)	html	Identical to 4.2.2 ordered by CDR3 length
4.2.7 IMGT clonotypes (AA) with identical CDR3-IMGT (AA) with detailed clonotypes (nt) per CDR3-IMGT length (AA)	html	IMGT clonotypes (AA) grouped by CDR3-IMGT (AA) with detailed clonotypes (nt) per CDR3-IMGT length (AA)
4.2.8 IMGT clonotype (AA) diversity and expression histograms: per V, (D), J-GENE and per CDR3-IMGT length	html	· IMGT clonotype (AA) expression histograms: number of sequences assigned to an IMGT clonotype (AA) per V-GENE (green color), D-GENE (for IGH, TRB, TRD) (red color) and J-GENE (yellow color) and per CDR3-IMGT length · IMGT clonotype (AA) diversity histograms: number of different IMGT clonotypes (AA) per V-GENE, D-GENE (for IGH, TRB, TRD) and J-GENE (pink color) and per CDR3-IMGT length
4.2.9 IMGT clonotype (AA) diversity and expression tables: per V, (D), J-GENE and per CDR3-IMGT length	html	Tables for the number of sequences assigned to an IMGT clonotype (AA) and the number of IMGT clonotypes (AA) per V-GENE, D-GENE (for IGH, TRB, TRD) and J-GENE, and per CDR3-IMGT length
4.2.10 V gene and allele table: Rearrangements, number of sequences and number IMGT clonotypes (AA) per V-GENE and allele	html	Number of sequences assigned to an IMGT clonotype (AA), number of different IMGT clonotypes (AA), number of out-of-frame sequences, and number of sequences of other categories per V-GENE and allele
5. IMGT clonotype (AA) results comparison	html	
5.1 IMGT clonotype (AA) comparison: Full results	html	Lists of IMGT clonotypes (AA) unique for each batch and lists for common IMGT clonotypes (AA) in two or more batches
5.2 IMGT clonotype (AA) comparison: Synthesis table	html	Number of IMGT clonotypes (AA) (diversity) and the number of sequences assigned to IMGT clonotypes (AA) (expression) only present ('exclusive') in a single batch or common to two or more batches
	html	"Number of IMGT clonotypes (AA)," "Number of in-frame sequences assigned to

(continued)

Table 2
(continued)

Documents	File type	Content description
5.3 IMGT clonotype (AA) comparison: Result summary table per V-GENE, D-GENE (for IGH, TRB, TRD), J-GENE		IMGT clonotypes (AA)," per gene, and for each batch

a

ID		Nb			IMGT clonotype (AA) definition							IMGT clonotype (AA) representative sequence			IMGT clonotypes (nt)
#	Exp. ID	Total nb of '1 copy'	Total nb of 'More than 1'	Total	V gene and allele	D gene and allele	J gene and allele	CDR3-IMGT length (AA)	CDR3-IMGT sequence (AA)	Anchors 104, 118	V %	Sequence length	Functio -nality	Sequence ID	Sequences file ('1 copy')
1	21129-S3	224	11	235	Homsap IGHV4-4*07 F	Homsap IGHD1-1*01 F	Homsap IGHJ6*03 F	13 AA	ARGTTFYYYYMDV	C,W	100	497	productive	SRR1168790.43 G9YJURD01DTO length=497_NA	Sequences file
2	15294-S3	148	0	148	Homsap IGHV4-4*07 F	Homsap IGHD3-16*01 F	Homsap IGHJ4*02 F	15 AA	ARDPLGGNSALTFDY	C,W	98.6	557	productive	SRR1168790.29 G9YJURD01AKHJ length=557_NA	Sequences file
3	13555-S3	125	0	125	Homsap IGHV4-39*01 F	Homsap IGHD5-12*01 F	Homsap IGHJ4*02 F	16 AA	ARLAPQSKSHVSAPDY	C,W	99.66	515	productive	SRR1168790.63 G9YJURD01AZDA length=515_NA	Sequences file
4	5172-S3	121	1	122	Homsap IGHV4-59*01 F	Homsap IGHD6-6*01 F	Homsap IGHJ6*03 F	20 AA	ARTPIGHYSSSSKRYYYMDV	C,W	100	521	productive	SRR1168790.60 G9YJURD01B57Z length=521_NA	Sequences file
5	23185-S3	106	1	107	Homsap IGHV4-61*02 F	Homsap IGHD6-25*01 F	Homsap IGHJ3*01 F	12 AA	ARGSGIAPVMDV	C,W	90.34	499	productive	SRR1168790.349 G9YJURD01B3TE length=499_NA	Sequences file
6	5153-S3	105	0	105	Homsap IGHV4-34*01 F	Homsap IGHD2-21*01 F	Homsap IGHJ4*02 F	20 AA	ARSWGYCGSIDCQTPVGLGY	C,W	97.19	509	productive	SRR1168790.59 G9YJURD01ARW length=509_NA	Sequences file
7	17162-S3	96	2	98	Homsap IGHV4-31*03 F	Homsap IGHD2-15*01 F, or Homsap IGHD2-21*01 F	Homsap IGHJ4*02 F	14 AA	ACDVQTSQYVAFDY	C,W	87.29	517	productive	SRR1168790.42 G9YJURD01BH3 length=517_NA	Sequences file

b

#	CDR3-IMGT length (nt)	Nb diff CDR3-IMGT (nt)	CDR3-IMGT sequence (nt)	Nb diff nt	V gene and allele	D gene and allele	J gene and allele	Anchors 104, 118	V % mean	V-REGION length mean	J % mean	J-REGION length mean	Sequence length mean	Total nb of '1 copy'	Total nb of 'More than 1'	Total
7	42	3	gcgtgcgacgtccagacgtcacaatatgtagcttttgactac	1	Homsap IGHV4-31*03 F	Homsap IGHD2-15*01 F	Homsap IGHJ4*02 F	C,W	86.55	298	60.42	43	510	1	0	1
			gcgtgtgacgtccagacgtcacaatatgtagccttgactac	1	Homsap IGHV4-31*03 F	Homsap IGHD2-21*01 F	Homsap IGHJ4*02 F	C,W	87.24	298	81.25	44	506	1	0	1
			gcgtgtgacgtccagacgtcacaatatgtagcttttgactac	0	Homsap IGHV4-31*03 F	Homsap IGHD2-15*01 F	Homsap IGHJ4*02 F	C,W	86.4	298	80.59	43	514	94	2	96

Fig. 12 Top of the file 4.2.1 'IMGT clonotypes (AA) per Nb' [26]. (**a**) List of IMGT clonotypes (AA) ordered by decreasing number of assigned sequences. The first seven of IMGT clonotypes (AA) are shown. The table provides the Exp. ID (IMGT clonotype (AA) identifier in the set), the numbers of "1 copy," "More than one," and the total. The IMGT clonotype (AA) definition includes the names of the V, D, and J genes and alleles, the CDR3-IMGT length, the AA CDR3-IMGT, and the anchors. The IMGT clonotype (AA) representative sequence is characterized by the identity percentage with the closest V gene and allele, the sequence length, the sequence functionality, and a link to the FASTA sequence. An additional link allows to display all "1 copy" assigned to the IMGT clonotype (AA). The batch S3 results from the analysis of the run SRR1168790 available on Sequence Read Archive (SRA) (https:/www.ncbi.nlm.nih.gov/sra). (**b**) Example of IMGT clonotypes (nt) linked to the IMGT clonotypes (AA) #7 (extracted from file 4.2.2) to which 96 "1 copy" sequences were assigned. Ninety-four of them are assigned to the same IMGT clonotype (nt) with a CDR3-IMGT of 42 nucleotide "gcgtgtgacgtccagacgtcacaatatgtagcttttgactac". Two other IMGT clonotypes (nt) (with one sequence each) are also linked to #7. One shows a mutation (t>c) on the nt 6 of the CDR3 and the second a mutation (t>c) on the nt 33 of the CDR3 (shown in red in the figure)

5 IMGT/StatClonotype

IMGT/StatClonotype [28, 29] provides statistical pairwise comparison of the diversity and of the expression of the IMGT clonotypes (AA) between two IMGT/HighV-QUEST statistics output results [24, 26]. The tool evaluates the statistical significance of the differences in proportions per variable (V), diversity (D), and joining (J) gene and allele of a given IG or TR locus according to seven multiple testing procedures for the adjustment of the *p*-value: this allows the user to choose the stringency, which is the most relevant for the aim of a given study [28, 29]. IMGT/StatClonotype includes the characterization of the CDR-IMGT with the analysis of the distribution of CDR-IMGT length (for IMGT clonotype (AA) diversity or expression) and, for a given length, the distribution per position of the amino acids according to IMGT AA physicochemical classes [46] and variability indexes. Results for the evaluation of V-(D)-J associations are provided through heatmaps.

5.1 IMGT/StatClonotype Installation and Launch

IMGT/StatClonotype [28, 29] is a standalone IMGT® tool that needs to be installed locally on the user's computer. Running IMGT/StatClonotype requires the prior installation of the R program (*see* **Note 23**):

1. Install R program and IMGTStatClonotype R package following the steps described in "IMGTStatClonotype R package installation" (http://www.imgt.org/StatClonotype/IMGTStatClonotypeDoc.html#pack).
2. In the console of R program, following the prompt ">," enter the two command lines:

   library(IMGTStatClonotype)

   launch()

   The IMGT/StatClonotype web interface is launched on your default web browser.

5.2 IMGT/StatClonotype Uploading of Input Sets of IMGT Clonotypes (AA)

1. In the left panel of IMGT/StatClonotype welcome page, choose the IMGT/HighV-QUEST set 1 and IMGT/HighV-QUEST set 2 to be compared (Fig. 13). The IMGT/StatClonotype input sets must be selected from IMGT/HighV-QUEST statistical analysis output folders, which are already stored on your computer (*see* Subheading 4.5, IMGT/HighV-QUEST "Statistics history" page: follow-up and download of statistics, Subheading 4.5.2 "data" directory).
2. Select CDR3-IMGT length range of IMGT clonotypes (AA) in order to eliminate outliers from the statistical procedures (default range for CDR3-IMGT lengths is ≥4 and ≤45).

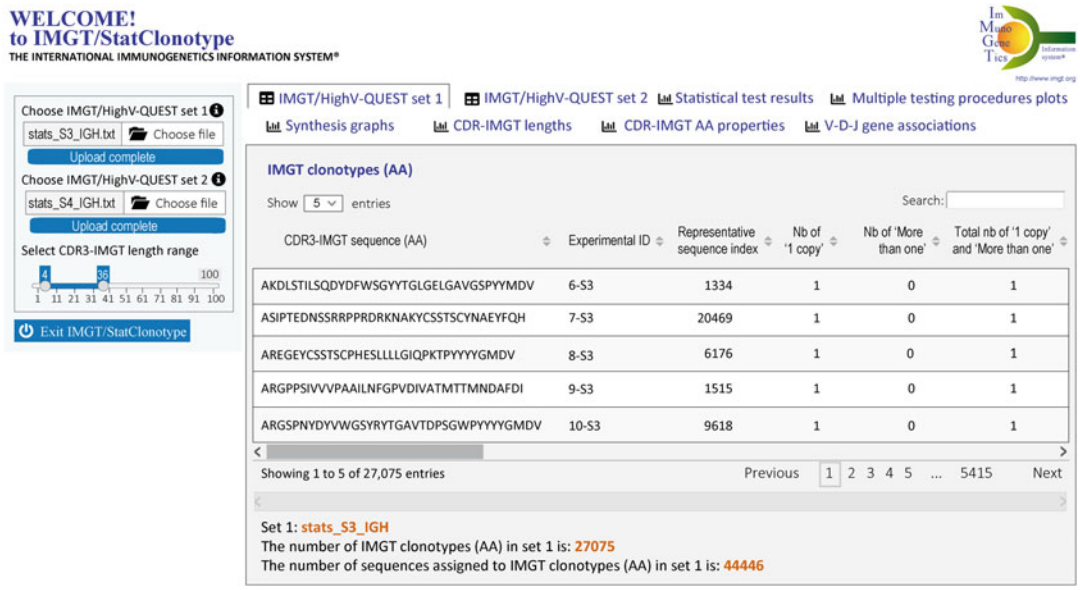


Fig. 13 IMGT/StatClonotype Welcome page [28, 29]. The files stats_S3_IGH.txt (IMGT/HighV-QUEST set 1) and stats_S4_IGH.txt (IMGT/HighV-QUEST set 2) were uploaded from the "data" directory of IMGT/HighV-QUEST statistical output (obtained from the runs SRR1168790 and SRR1168789, respectively, available on Sequence Read Archive (SRA) (https:/www.ncbi.nlm.nih.gov/sra)). The range for CDR3-IMGT lengths is $\geq$4 and $\leq$36. IMGT/HighV-QUEST set 1 includes 27,075 IMGT clonotypes (AA) to which 44,446 sequences were assigned

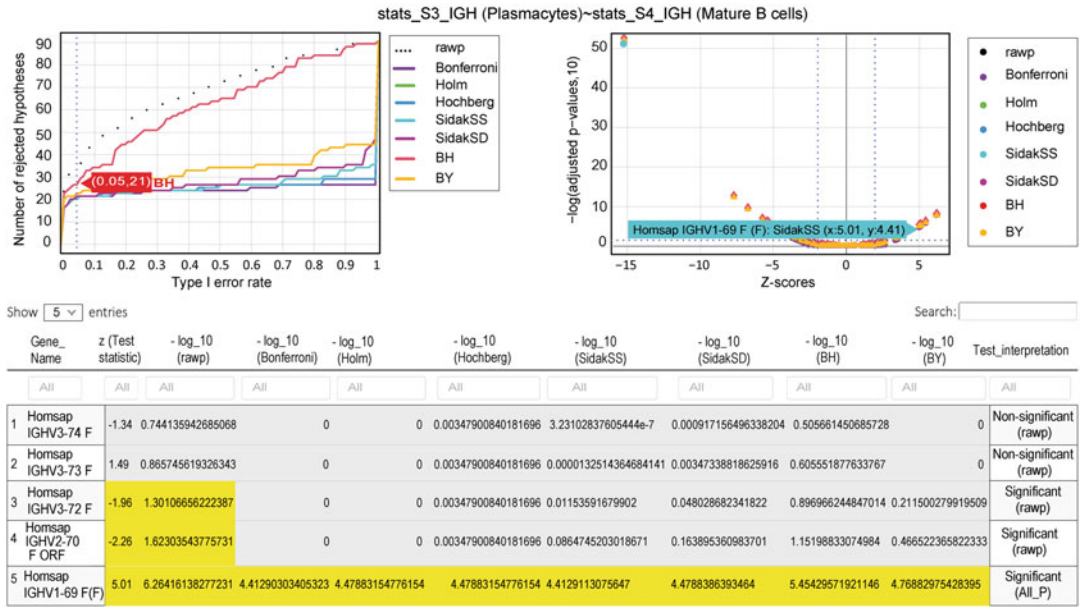
5.3 IMGT/StatClonotype Results

IMGT/StatClonotype [28, 29] results are displayed in the major right panel in eight distinct tabs:

1. The "IMGT/HighV-QUEST set 1" and "IMGT/HighV-QUEST set 2" tabs display, for set 1 and set 2, respectively, a table of the IMGT clonotypes (AA) with a CDR3-IMGT length in the range selected in the left panel (Fig. 13). The table includes [28, 29]: CDR3-IMGT sequence (AA), experimental ID, representative sequence index, Nb of "1 copy," Nb of "More than one" (*see* **Note 22**), total number of "1 copy" and "More than one," "1 copy" indexes, V gene, V allele, D gene, D allele, J gene, J allele, CDR1-IMGT, CDR2-IMGT, CDR1-IMGT-gapped sequence (AA), CDR2-IMGT-gapped sequence (AA), V-REGION %identity, sequence length, C104, F/W118, anchors (true or false), sequence ID, functionality, sequence file number, and sequence clonotype number.

   The total number of the selected IMGT clonotypes (AA) and the number of sequences assigned are indicated below the table. At the bottom of the page, a second table lists the IMGT clonotypes (AA) corresponding to CDR3-IMGT length outliers that are not taken into account for statistical procedures (not shown).

2. The "statistical test results" tab [28, 29] displays the statistical test results of differences in proportions for the IMGT clonotypes (AA) in both sets, for genes (top of the page) and for alleles (bottom of the page; *see* **Note 24**), without adjusted *p*-values or with adjusted *p*-values according to the seven multiple testing procedures (Bonferroni, Holm, Hochberg, ŠidákSS, ŠidákSD, Benjamini & Hochberg, Benjamini & Yekutieli) [28]. The results are displayed in a table of 21 columns (*see* **Note 25**). Columns 1–12 provide gene (or allele) name, gene (or allele) type, "Nb of IMGT clonotypes (AA)," "Proportion" and "normalized proportion" for set 1 and set 2 (*see* **Note 26**), "Difference in proportions," *z*-scores values, and lower and upper bound confidence interval (CI) of the difference in proportions. Unadjusted *p*-values (rawp) and adjusted *p*-values from multiple testing are given from column 13 to column 20 of the table. The column 21 provides a test interpretation for the significance of the difference in proportion. The "Download" button allows to save the tables as CSV files. Below the main table is the "Show/Hide Table" button that displays (or not) the list of genes (or alleles) with null or small occurrences. Use the left panel to customize the display: (1) select the results for IMGT clonotype (AA) diversity or for IMGT clonotype (AA) expression, (2) include results for several genes (or alleles) or for single genes (or alleles) only, (3) display or not the null or smallest gene (or allele) occurrences, (4) select or unselect one or more columns of the "Statistical test results for genes" and "Statistical test results for alleles" tables to be shown.
3. "Multiple testing procedures plots" tab [28, 29] displays, in the major right panel, interactive line graphs, and scatter plots for genes (on the top) and for alleles (at the bottom), for the comparison of the differences in proportions for IMGT clonotypes (AA) between sets 1 and 2 (Fig. 14). On the left, the line graphs display the number of rejected null hypotheses (therefore the number of significant differences in proportions) for a chosen Type I error for the seven procedures. On the right, the scatter plots show negative decimal logarithms (−log10) of unadjusted p-values (black symbols) and adjusted *p*-values obtained by each multiple testing procedure (colored symbols): it highlights the V, D, and J genes of a locus with the most significant differences (positive or negative) in proportions. Numerical values and z-scores are reported in a table below, the plots with a yellow background for significant positive or negative differences in proportions.
4. "Synthesis Graphs" tab [28, 29] displays a synthesis graph that combines a normalized bar graph of proportions (*see* **Note 26**) and the differences in proportions with significance and



	Gene_ Name	z (Test statistic)	- log_10 (rawp)	- log_10 (Bonferroni)	- log_10 (Holm)	- log_10 (Hochberg)	- log_10 (SidakSS)	- log_10 (SidakSD)	- log_10 (BH)	- log_10 (BY)	Test_interpretation
1	Homsap IGHV3-74 F	-1.34	0.744135942685068	0	0	0.00347900840181696	3.23102837605444e-7	0.000917156496338204	0.505661450685728	0	Non-significant (rawp)
2	Homsap IGHV3-73 F	1.49	0.865745619326343	0	0	0.00347900840181696	0.000013251436468414 1	0.00347338818625916	0.605551877633767	0	Non-significant (rawp)
3	Homsap IGHV3-72 F	-1.96	1.30106656222387	0	0	0.00347900840181696	0.01153591679902	0.048028682341822	0.896966244847014	0.211500279919509	Significant (rawp)
4	Homsap IGHV2-70 F ORF	-2.26	1.62303543775731	0	0	0.00347900840181696	0.0864745203018671	0.163895360983701	1.15198833074984	0.466522365822333	Significant (rawp)
5	Homsap IGHV1-69 F(F)	5.01	6.26416138277231	4.41290303405323	4.47883154776154	4.47883154776154	4.4129113075647	4.4788386393464	5.45429571921146	4.76882975428395	Significant (All_P)

Fig. 14 IMGT/StatClonotype Multiple testing procedures plots for genes [28, 29]. In the left panel, "IMGT clonotype (AA) diversity," "Single gene," and "Hide null or smallest gene occurrences" were selected (not shown). "Multiple testing procedures plots" displays an interactive line graph on the left and a scatter plot on the right for genes. Hovering the mouse on the interactive the line graph on the left allows the display of the exact number of significant differences in proportions, that is 21 for a Type I error $\alpha = 0.05$ with the multiple testing procedure BH. On the right, the scatter plot shows the coordinates of z-score and −log_10 (SidakSS) for IGHV1-69, for which the difference in proportion is significant whatever the multitesting procedure as indicated in the table. The graphs can be saved in PNG, JPG, or PDF and the tables in CSV format

confidence intervals (CI), for genes (on the top) (Fig. 15) and for alleles (at the bottom) (*see* **Note 27**). In synthesis graphs for genes, IMGT gene names are ordered by their positions in the locus with their known functionalities. Below the normalized bar graph are listed the not ordered genes (not shown). They are grouped and shown at the bottom of the gene list in the synthesis graph. The values for the normalized proportions of genes (or alleles) in set 1 and set 2, the differences in proportions, the lower and upper bound of the confidence indices for differences in proportions, and the Test interpretation are recorded in "Statistical test results" tab Tables.

5. "CDR-IMGT lengths" tab [28, 29] displays, in the right panel, interactive bar graphs for set 1 and set 2 showing the distribution of the number of IMGT clonotypes (AA) (for IMGT clonotype (AA) diversity) or of the number of sequences assigned to IMGT clonotypes (AA) (for IMGT clonotype (AA) expression), per CDR-IMGT length (*see* **Note 28**). The left panel allows to choose the CDR-IMGT (CDR1-IMGT, CDR2-IMGT, or CDR3-IMGT) and to select the length of

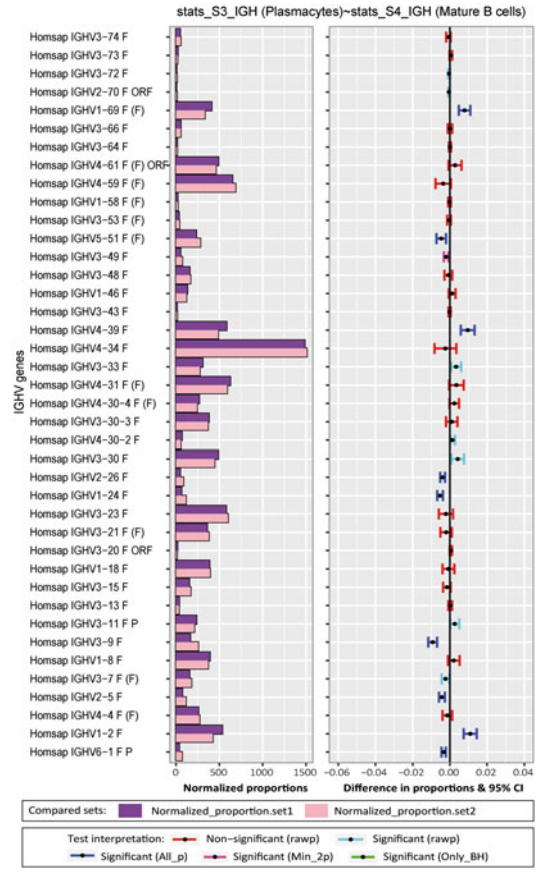


Fig. 15 IMGT/StatClonotype synthesis graph for IMGT clonotype (AA) diversity per V gene [28, 29]. It displays visual comparison of the normalized proportions of IMGT clonotype (AA) diversity of the IGHV genes between sets 1 and 2. For example, the diversity of IMGT Clonotypes (AA) expressing the IGHV1-2, IGHV4-39, and IGHV1-69 genes is significantly higher in set 1 than in set 2 whatever the multiple testing procedure. In the left panel, "Single gene" and "Hide null or smallest gene occurrences" were selected. Synthesis graphs are downloadable in PNG, JPG, or PDF format

the CDR-IMGT for the "List of IMGT clonotypes (AA) with selected CDR3-IMGT length" displayed below the bar graphs for sets 1 and 2.

6. "CDR-IMGT AA Properties" tab displays the distribution of the IMGT classes [46] of the 20 amino acids at CDR-IMGT (CDR1-IMGT, CDR2-IMGT, or CDR3-IMGT) positions in sets 1 and 2 for a given CDR-IMGT length. The left panel allows (1) to select the IMGT classes to be displayed for the amino acids: "20 amino acids," "Physicochemical" (Fig. 16), "Hydropathy," "Volume," "Chemical," "Charge," "Hydrogen donor or acceptor atoms," and "Polarity"; (2) to show

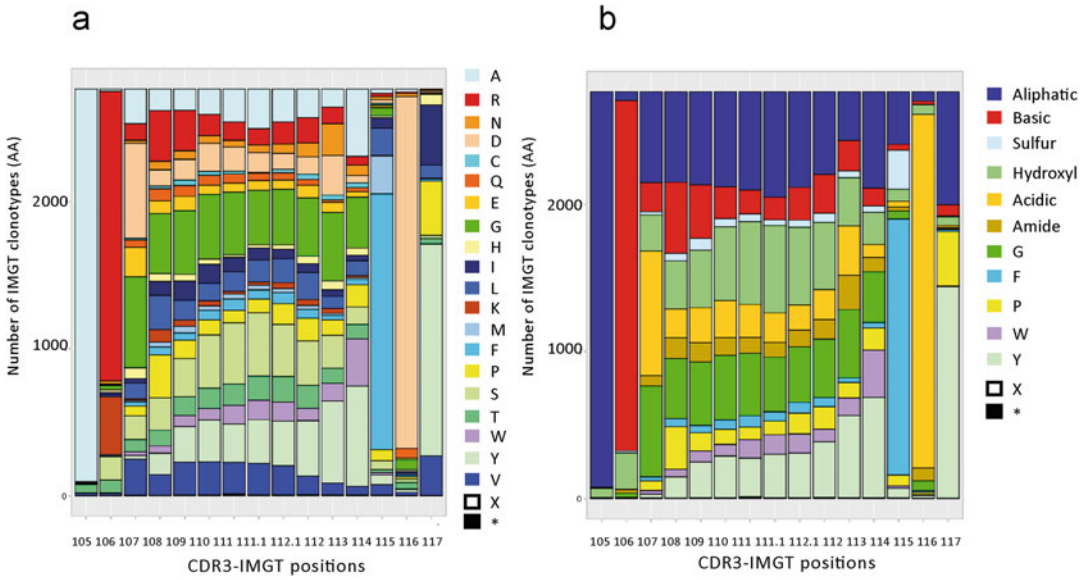


Fig. 16 IMGT/StatClonotype CDR-IMGT AA properties distribution [28, 29]. Examples for the CDR3-IMGT of length 15 in set 1: (**a**) "20 amino acids" and (**b**) "Physicochemical"

results by absolutes values (number of occurrences of an amino acid (or IMGT amino acid class) at a given position, for a given CDR length) or percentages; and (3) to modify the length and width of the graphs. The major right panel includes, for each set, a table with numbers (or percentages) of the amino acids (or IMGT amino acid classes) (in rows), at a given position (in columns). The table includes a row for undefined amino acids ("X") and for stop codons. The tables can be downloaded as CSV files. The corresponding graphical representation is shown as an interactive bar graph to visualize the amino acid distribution per position. At the bottom of the page, the variability plots based on the indexes according to "Shannon entropy," 'Wu-Kabat variability," or "Simpson index" with tables for numerical values are displayed. Comparisons of two sets are useful in detecting the characteristics of amino acids at positions important for the V domain antibody diversity or, by contrast, for maintaining its structure.

7. "V-D-J gene associations" tab [28, 29] displays interactive heat maps to represent V-J, V-D, or D-J gene associations in set 1 and set 2. The left panel allows (1) to display the Dendrogram for V-J, V-D, or D-J gene association, (2) to get the results with clustering or not, (3) to get the results in normalized values, and (4) to select the color palettes. The major right panel includes interactive heat maps to represent V-J, V-D, or D-J gene associations in set 1 and set 2. If the "Results with clustering" is selected, a double Ward hierarchical clustering with Euclidean distance is performed (this classification

operates simultaneously on the lines and columns of a matrix intersecting two different types of genes), otherwise heat maps are shown without dendrograms and ordering. Such an analysis permits to detect genes with similar diversity or expression profiles, which can be further explored for given and/or related specificities in immune repertoire comparative analysis. Under heat maps, tables crossing the V-J, V-D, or D-J gene occurrences in set 1 and set 2 are given.

6 IMGT/DomainGapAlign

IMGT/DomainGapAlign [30, 31] analyzes the amino acid sequences of the IG and TR V-DOMAIN (*see* **Note 29**). IMGT/DomainGapAlign identifies the closest V and genes and alleles of the user's amino acid domain sequences by comparison with the IMGT reference directory sets composed of the translations of the germline V and J regions of the genes managed in IMGT/GENE-DB [34]. The reference amino acid sequences are available by querying IMGT/DomainDisplay (IMGT® Home page, http://www.imgt.org). Importantly, IMGT/DomainGapAlign can analyze V-DOMAIN from different species and different locus in a single run. The tool gaps the sequences, numbers the AA of each V-DOMAIN, and provides the delimitations of the FR-IMGT and CDR-IMGT and those of the beta strands and loops by applying the IMGT unique numbering [13]. It also characterizes the amino acid changes (*see* **Note 30**).

6.1 IMGT/DomainGapAlign Query and Customization of the Analysis

6.1.1 Standard Parameters and Sequence (s)

1. Paste your FASTA amino acid sequences in the text area or upload them from a text file.
2. By default, the analysis is performed on the V-DOMAIN (Domain type "V′) [30, 31] (Fig. 17).
3. Select the species (by Latin name or English name) of the reference directory sets with which the sequences will be compared or let the tool (with default "any") detecting the best alignments whatever the species.
4. Set, with the option "Smith-Waterman score above," the threshold of the Smith-Waterman score above, which the alignments will be displayed in the results (*see* **Note 31**) (default is 0).

   Select the number of alignments displayed for each V-DOMAIN in the results (default is 5).
5. Check "IMGT Colliers de Perles" [21] to include the IMGT Collier de Perles [18–20] in the results (*see* Subheading 7)

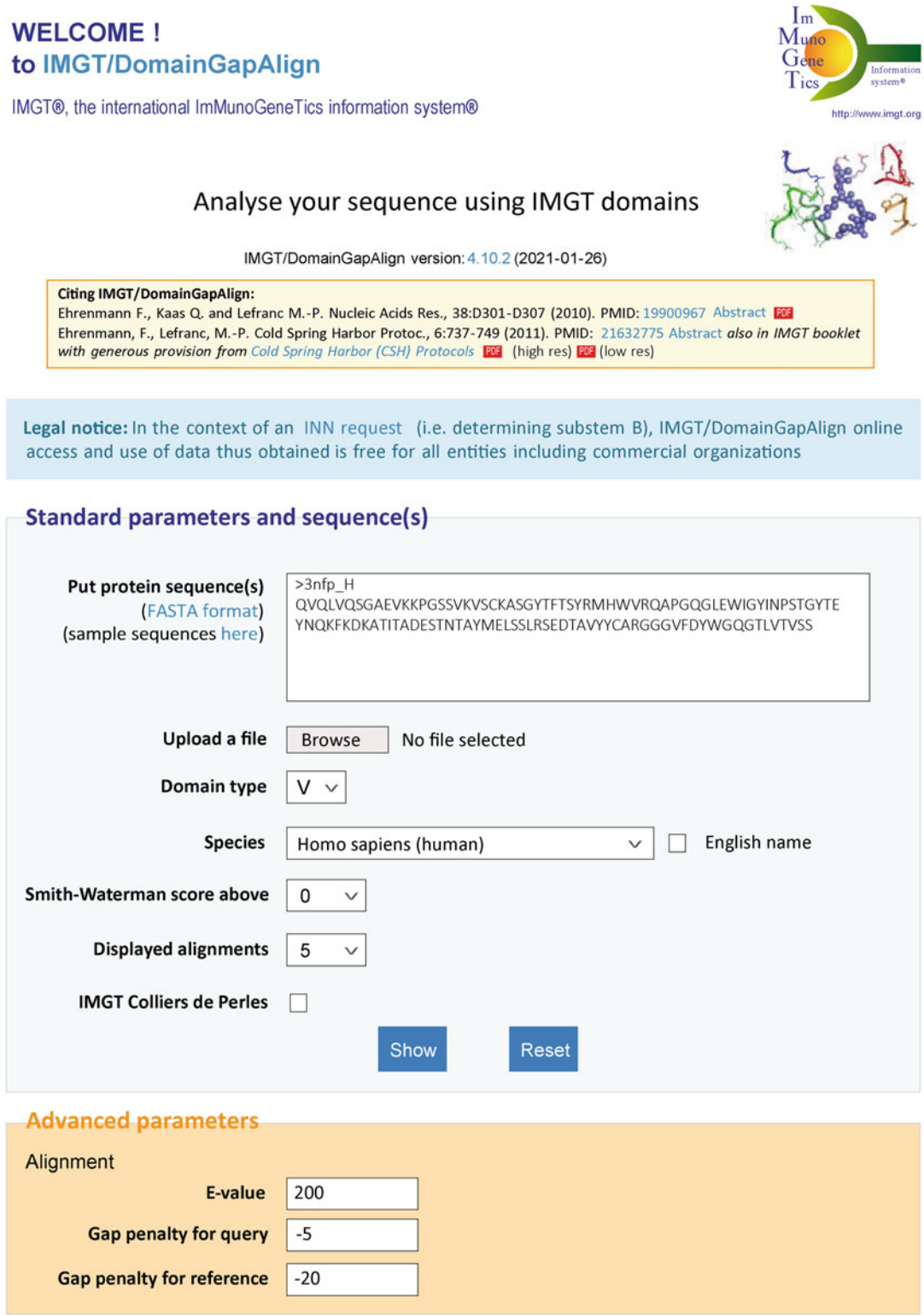


Fig. 17 IMGT/DomainGapAlign Welcome page [30, 31]

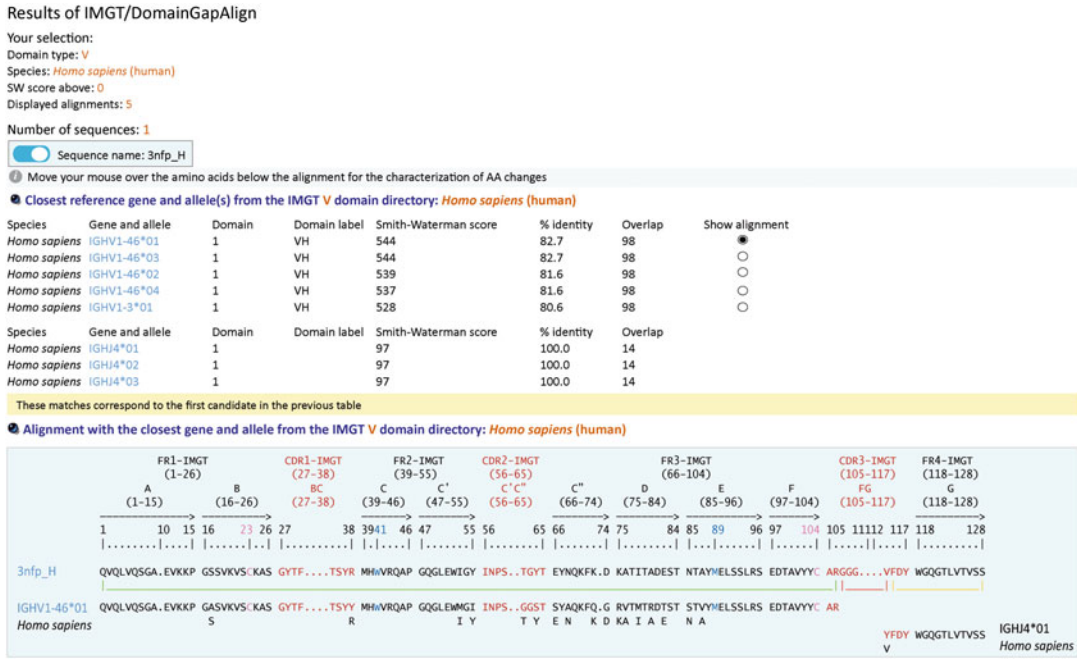


Fig. 18 IMGT/DomainGapAlign Results [30, 31]. Top of the result page: the VH domain of daclizumab 3nfp_H chain (PDB code 3nfp of IMGT/3Dstructure-DB [30, 50, 51]) is compared with the *Homo sapiens* reference directory. It is aligned with the human IGHV1-46*01 and IGHJ4*01 genes alleles

6.1.2 Advanced Parameters

Modify if necessary the "E-value," the "Gap penalty," and "Gap penalty for reference" used for Smith-Waterman alignments.

6.2 IMGT/DomainGapAlign Results

"Your selection," on the top of the IMGT/DomainGapAlign results page [30, 31] (Fig. 18), recalls the parameters and values selected for the AA sequence submission. Following the "Number of sequences", a switch button allows to display (or not) the results of the corresponding sequence.

1. "Closest reference gene and allele(s) from the IMGT V domain directory" (Fig. 18) shows the name of the species of which the AA references sequences are compared with the user sequence ("All species" is indicated if "any" was selected).

   A table summarizes the five (selected by default) best aligned V genes and alleles including the species, the IMGT V gene and allele name, the number of the domain, the Domain label, and the Smith-Waterman alignment score, with the identity percentage and the overlap (number of aligned amino acids assigned to the V-REGION) between the user and the IMGT AA reference sequence, and a radio button for the alignment to display (if the Smith-Waterman score is equal or higher than the threshold selected for the submission).

A second table is displayed for J genes and alleles with the species, the IMGT J gene and allele name, the number of the domain, the Smith-Waterman alignment score, the identity percentage, and the overlap.

2. "Alignment with the closest gene and allele from the IMGT V domain directory" [30, 31] (Fig. 18): the header of the alignment indicates the length and delimitation of the 4 FR-IMGT and 3 CDR-IMGT and of the 9 beta strands (A, B, C, C′, C″, D, E, F, and G) and 3 loops (BC, C′C″, and FG) according to the IMGT unique numbering for V domain [13]. The submitted AA-gapped sequence is aligned with the V region of the germline gene and allele. Between both sequences, a green line delimits the V-REGION, a red line delimits the (N-D)-REGION, and a yellow one delimits the J-REGION. Below the alignment, the AA changes compared with the germline are shown. The AA J-REGION is aligned with the closest J gene and allele.
3. "Region(s) and domain(s) identified in your sequence (by comparison with the closest genes and alleles" (Fig. 19) allows to download the V-DOMAIN amino acid sequence with or without IMGT gaps.
4. "Results summary (by comparison with the closest genes and alleles)" (Fig. 19) provides the first table, which includes the percentage identity with the V-REGION, the CDR-IMGT lengths, the total number of different AA in CDR1-IMGT and CDR2-IMGT, the FR-IMGT lengths, the number of different AA in FR-IMGT, and the total number of amino acid changes.

   Below are displayed two additional parallel tables: on the left the "AA changes in strands and loops" and on the right the "AA changes in FR-IMGT and CDR-IMGT" with the number of different AA, the description of the AA change with the "AA class Change Type'" (+) or not (−) (for hydropathy, volume and physicochemical characteristics [46] according to the AA IMGT classes), and "AA class Similarity Degree" (very similar, similar, dissimilar, and very dissimilar).
5. IMGT Colliers de Perles [18–20] (*See* Subheading 7) are shown, if selected, on one or two layers, without or with AA change positions shown in pink circles (or squares for CDR-IMGT anchors).

7 IMGT/Collier-de-Perles

The IMGT/Collier-de-Perles tool [21] generates "'IMGT Colliers de Perles" [18–20]. For V-DOMAIN, IMGT Colliers de Perles are obtained on one or two layers, provided that the V-DOMAIN

Region(s) and domain(s) identified in your sequence (by comparison with the closest genes and alleles: *Homo sapiens* IGHV1-46*01 and IGHJ4*01)

QVQLVQSGAEVKKPGSSVKVSCKASGYTFTSYRMHWVRQAPGQGLEWIGY
INPSTGYTEYNQKFKDKATITADESTNTAYMELSSLRSEDTAVYYCARGG
GVFDYWGQGTLVTVSS

Sequence without gaps Sequence with gaps

Results summary (by comparison with the closest genes and alleles *Homo sapiens* IGHV1-46*01 and IGHJ4*01)

Sequence name	V-REGION identity percentage	CDR-IMGT lengths	Number of different AA in CDR1- and CDR2-IMGT	FR-IMGT lengths	Number of different AA in FR-IMGT	Total number of AA changes in V-DOMAIN
3nfp_H	82.7%	[8.8.9]	3	[25.17.38.11] = 91 AA	14	17

AA changes in strands and loops

Strands	Number of different AA	AA changes
A (1-15)	0	-
B (16-26)	1	A17>S (- + -) dissimilar
C (39-46)	0	-
C' (47-55)	2	M53>I (+ + -) similar I55>Y (- - -) very dissimilar
C'' (66-74)	4	S66>E (- - -) very dissimilar A68>N (- - -) very dissimilar Q72>K (+ - -) dissimilar G74>D (- - -) very dissimilar
D (75-84)	5	R75>K (+ + +) very similar V76>A (+ - +) similar M78>I (+ + -) similar R80>A (- - -) very dissimilar T82>E (- - -) very dissimilar
E (85-96)	2	S85>N (- - -) very dissimilar V87>A (+ - +) similar
F (97-104)	0	-
G (118-128)	0	-

Loops	Number of different AA	AA changes
BC (27-38)	1	Y38>R (- - -) very dissimilar
C'C'' (56-65)	2	G62>T (+ - -) dissimilar S64>Y (+ - -) dissimilar
FG (105-117)	0	-

AA changes in FR-IMGT and CDR-IMGT

FR-IMGT	Number of different AA	AA changes
FR1-IMGT (1-26)	1	A17>S (- + -) dissimilar
FR2-IMGT (39-55)	2	M53>I (+ + -) similar I55>Y (- - -) very dissimilar
FR3-IMGT (66-104)	11	S66>E (- - -) very dissimilar A68>N (- - -) very dissimilar Q72>K (+ - -) dissimilar G74>D (- - -) very dissimilar R75>K (+ + +) very similar V76>A (+ - +) similar M78>I (+ + -) similar R80>A (- - -) very dissimilar T82>E (- - -) very dissimilar S85>N (- - -) very dissimilar V87>A (+ - +) similar
FR4-IMGT (118-129)	0	-

CDR-IMGT	Number of different AA	AA changes
CDR1-IMGT (27-38)	1	Y38>R (- - -) very dissimilar
CDR2-IMGT (56-65)	2	G62>T (+ - -) dissimilar S64>Y (+ - -) dissimilar
CDR3-IMGT (105-117)	0	-

Fig. 19 IMGT/DomainGapAlign Results [30, 31]. Bottom of the result page for VH domain of daclizumab 3nfp_H chain (PDB code 3nfp of IMGT/3Dstructure-DB [30, 50, 51]): the CDR-IMGT lengths are [8.8.9] with a total of three AA changes. The FR-IMGT lengths are [25.17.38.11] with a total of 14 AA changes

(AA) sequence is gapped according to the IMGT unique numbering [13] (*see* **Note 32**). Resulting IMGT Colliers de Perles show the standardized delimitation of FR-IMGT and CDR-IMGT, and of beta strands with their orientation in the IG and TR V-DOMAIN, allowing the visualization of the amino acids, which are important for a 3D structural configuration and bridging the gap between sequences and structures.

7.1 IMGT/Collier-de-Perles Launched from IMGT Sequence Analysis Tools

1. In order to generate the IMGT Colliers de Perles from a V-DOMAIN nucleotide sequence, use IMGT/V-QUEST [22, 23] (*see* Subheading 3) and select A. Detailed view and result section 14.
2. Starting from a V-DOMAIN amino acid sequence, use IMGT/DomainGapAlign [30, 31] (*see* Subheading 6) to generate the IMGT Colliers de Perles and select "IMGT Colliers de Perles" in the submission form (*see* **Note 33**).

7.2 IMGT/Collier-de-Perles Submission Interface

Alternatively, using the IMGT/Collier-de-Perles interface [21] (Fig. 20) offers complete display options. The submitted V domain (AA) sequence must be gapped according to the IMGT unique numbering for V-DOMAIN [13] and the CDR3-IMGT length must be 13 or longer. The user may:

1. Select the "Domain type" ("Variable (V)"), the number of layers for the IMGT Collier de Perles representation (1 or 2) (*see* **Note 34**).
2. Select the "CDR-IMGT color type" [46] according to the locus of the sequence (1 for IGH, TRB, or TRD sequences and 2 for IGK, IGL, TRA or TRG sequences) and the "Background color," which will be applied to the FR-IMGT positions (*see* **Note 35**).
3. Enter the CDR3-IMGT length.
4. Enter the gapped AA sequence without any header.
5. In case of detected amino acid insertions compared with the IMGT unique numbering for V domain [13], provide in "Amino acid insertions" the position that precedes the insertion, its length in AA, and the numbering label for each inserted position.
6. A title for the resulting IMGT Collier de Perles can be optionally provided.
7. Click on "Show" to launch the tool.

7.3 IMGT/Collier-de-Perles Results

The IMGT Collier de Perles for a V-DOMAIN [18–21] displays the graphical representation of a V-DOMAIN with one position (1 AA) per bead (circle or square). Numbers allow an easy delimitation of the FR-IMGT, of the CDR-IMGT, and of the beta strands

Fig. 20 The IMGT/Collier-de-Perles Welcome page [21]

and the localization of the conserved amino acids. The anchor positions of CDR-IMGT are in square (*see* Subheading 2.2). The hatched positions represent gaps according to the IMGT unique numbering for V domain [13]. AA written in red letters indicate the five conserved positions in V-DOMAIN (1st-CYS 23, CONSERVED-TRP 41, hydrophobic 89, 2nd-CYS 104 and J-TRP or J-PHE 118). CDR-IMGT are colored according to "IMGT CDR-IMGT color type" [46] of the corresponding locus and

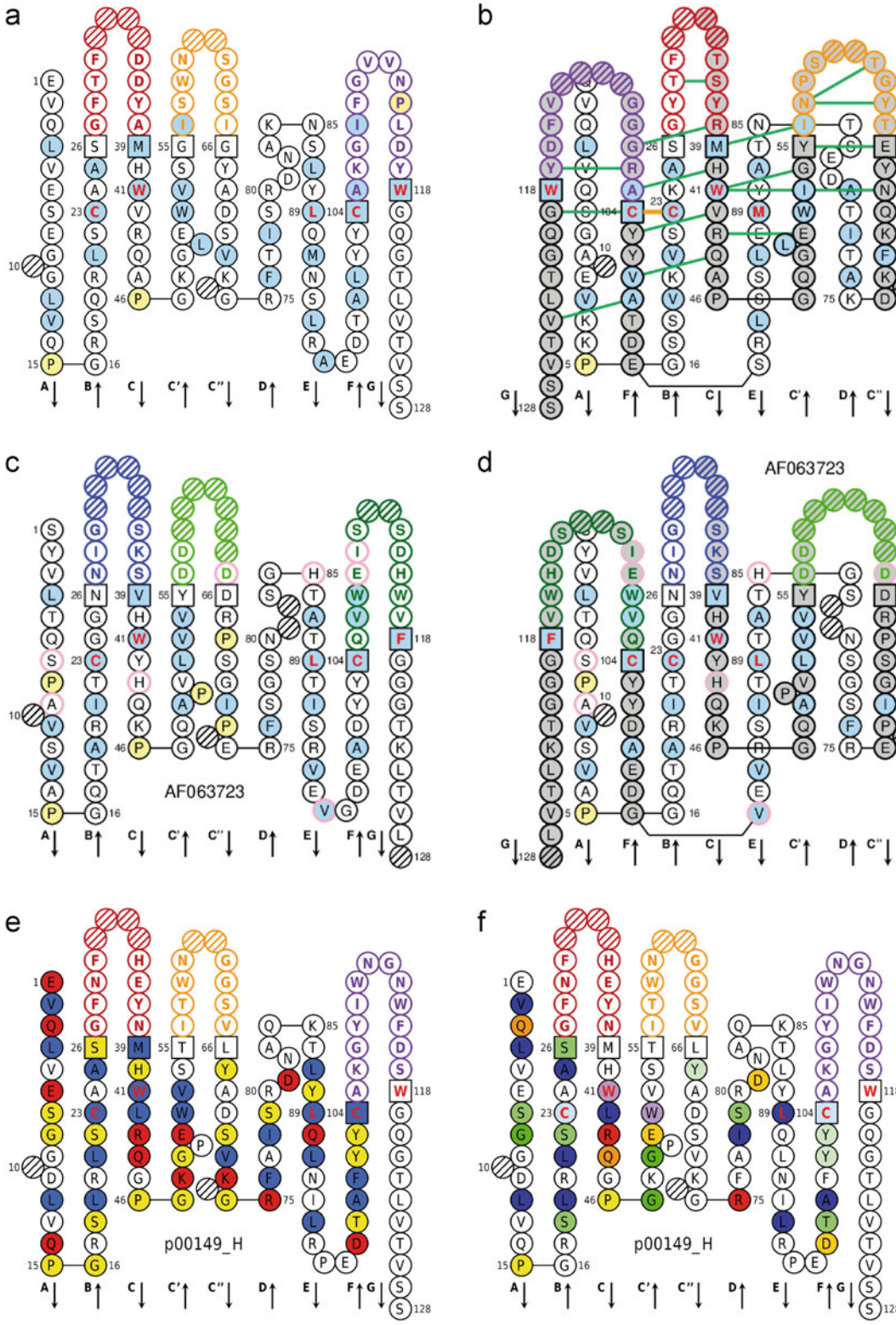


Fig. 21 IMGT Colliers de Perles for V-DOMAIN [18–21]. (**a**–**d**) Background color is "50% Hydrophobic positions," and Proline (P) is in yellow [46]. (**a**) IMGT Collier de Perles on one layer generated from the

FR-IMGT according to the "Background color" (*see* **Note 35**) selected by the user. The orientations of the nine beta-strands are indicated at the bottom of the IMGT/Collier-de-Perles. Illustrations of IMGT/Collier-de-Perles output are shown in Fig. 21.

8 Notes

1. The IMGT/V-QUEST reference directory sets [22, 23] are defined for species, which have been extensively studied, such as human, mouse, and dog, as well as for the species of which germline IG or TR repertoires are not fully available. The IMGT/V-QUEST results should be interpreted according to the available IMGT reference directories for a given species or taxon and a given locus. IMGT/V-QUEST reference directories have been also set for groups of taxons (e.g., Teleostei or Chondrichthyes), which contain pooled reference sequences from several species: these latter are not available in IMGT/HighV-QUEST.
2. Selecting "IG" or "TR" allows to submit sequences of different locus ("IGH," "IGK," and "IGL" for "IG" and "TRA," "TRB," "TRD," and "TRG" for "TR") in a same run (the locus will be automatically determined for each sequence), while the selection of a given locus forces the tool to compare the user sequences to the IMGT/V-QUEST reference directory set of the selected locus only.
3. The nucleotides "n" at 5′ and/or 3′ end of the submitted sequences are automatically trimmed before the analysis. The numbers of 5′ trimmed-n and 3′ trimmed-n are indicated in the results if any (*see* "Trimming of nucleotides "n" at 5' and/or 3' end of the submitted sequences" in the IMGT/V-QUEST [22, 23] Documentation).
4. The 14 sections are linked to the corresponding IMGT/V-QUEST [22, 23] Documentation in order to help the user in choosing of checking or unchecking them (see also Subheading 3.2.2). When "Uncheck all" is selected, only the "Result summary" is displayed.

Fig. 21 (continued) IMGT/V-QUEST [22, 23] analysis of a VH (nt) (accession number X81732 of IMGT/LIGM-DB [36]). (**b**) IMGT Collier de Perles on two layers of a VH with hydrogen bonds between the amino acids of the C, C′, C″, F, and G strands and those of the CDR-IMGT (daclizumab 3nfp_H, PDB code 3nfp of IMGT/3Dstructure-DB [30, 50, 51]). (**c**, **d**) IMGT Colliers de Perles with AA changes of a V-LAMBDA domain generated from the IMGT/DomainGapAlign [30, 31] analysis (translation of the AF063723 IMGT/LIGM-DB accession number), (**c**) on one layer, and (**d**) on two layers. (**e**, **f**) IMGT/Collier-de-Perles [21] results on one layer for the entry code p00149 from IMGT/2Dstructure-DB (**e**) with background color "IGH 80% hydrophopathy classes [46] and (**f**) with background color "IGH 80% physicochemical classes" [46]

5. The eight sections are linked to the corresponding IMGT/V-QUEST [22, 23] Documentation in order to help the user in choosing of checking or unchecking them. When "Uncheck all" is selected, only the "Summary table" is displayed.
6. The contents of the 11 or 12 text files are identical to those of the results provided by IMGT/HighV-QUEST [24–27], the high throughput version of IMGT/V-QUEST [22, 23].
7. The "AIRR formatted results" archive includes two text files: vquest_airr.tsv and Parameters.txt (IMGT/V-QUEST [22, 23] parameters used for the analysis). The vquest_airr.tsv contains the fields of the "Rearrangement Schema" provided by Adaptive Immune Receptor Repertoire (AIRR) Community [42, 43] plus additional IMGT fields (see http://www.imgt.org/IMGT_vquest/vquest_airr).
8. Including orphons [1–3] in the IMGT reference sets is relevant for genomic studies only.
9. In case of unmutated IG V-REGION (no mutations in FR1-IMGT, CDR1-IMGT, FR2-IMGT, CDR2-IMGT and FR3-IMGT), the number of accepted mutations is adjusted to 0 in 3'V-REGION and 5'J-REGION, and 2 in D-REGION for IGH, and 2 in 3'V-REGION and 5'J-REGION of IGK and IGL, in order to reflect the low probability of somatic hypermutations.
10. Both V-DOMAIN of a scFv must be in the same orientation [41]. In addition to the results for each V-DOMAIN individually, "Detailed view" includes a table for the identified scFv that links and localizes the two V-DOMAIN. In "Synthesis view," the two V-DOMAIN of a scFv are always displayed in consecutive rows of the "Summary table." In "Excel file," an additional 12th sheet provides one row per scFv with main characteristics and positions of the two V-DOMAIN.
11. Stereotyped sequences of Chronic Lymphocytic Leukemia (CLL) of subset #2 are characterized by a IGHV3-21/IGHJ6 rearrangement, a CDR3-IMGT of 9 AA with pattern "XX[D/E]XXXMDV" (X is for any AA, [D/E] means D or E). Sequences of subset #8 are characterized by an IGHV4-39/IGHJ5 rearrangement, an IGHV identity % is >98% and a CDR3-IMGT of 19 AA with a pattern "AXXXXXSSXWXXXXXWFDP". CLL patients whose malignant B clone carries a B-cell receptor with a heavy chain of subset #2 or subset #8 a re clinically associated with a poor prognosis [52, 53].
12. Four categories for sequence analysis are defined: (1) analysis without "Search for insertions and deletions in V-REGION," (2) analysis with "Search for insertions and deletions in V-REGION" and corrections if any, (3) analysis on complementary reverse sequence without "Search for insertions and

deletions in V-REGION," and (4) analysis on complementary reverse sequence with "Search for insertions and deletions in V-REGION" and corrections if any.

13. The score of the alignment for two sequences is calculated by counting +5 for each identical nt at a given position (match) and −4 for position with different nt (mismatch) [22, 23].
14. The JUNCTION decryption [45] for sequences with 1 D gene and allele provides lengths (in nt) of 3′V-REGION (3′V), D-REGION (D), and 5′J-REGION (5′J) (numbers between parentheses) of N1-REGION {N1} and N2-REGION {N2} (numbers between braces), and numbers between these regions indicate at the 3′ of the end of V, at 5′ or 3′ of D, and at the 5′ of the end of J, either trimmed nt (negative (−) values) or palindromic P nucleotides (positive (+) values) (trimmed or P nt are mutually exclusive) [45]. See IMGT/JunctionAnalysis [37, 38] Documentation (http://www.imgt.org/IMGT_jcta/decryption) [45] for sequences with 2 or 3 D genes and alleles.
15. Potential insertions or deletions are suspected by IMGT/V-QUEST [22, 23] when the V-REGION score is very low (less than 200), and/or the percentage of identity is less than 85%, and/or when the input sequence has different CDR1-IMGT and/or CDR2-IMGT lengths, compared with those of the closest germline V.
16. Several V or J genes and alleles with same highest identity percentage can be found generally: (1) if the sequence is partial in 5′ (for V) and/or in 3′ (for J), (2) if the numbers of mutations are identical (whatever their positions), (3) if reference sequences are identical (in case of duplicated genes or alleles), and (4) in case of polymorphism between different alleles in the germline CDR3-IMGT.
17. The algorithms for D gene and allele identification differ between IMGT/V-QUEST [22, 23] and IMGT/JunctionAnalysis [37, 38] and may provide different solutions. The results of IMGT/JunctionAnalysis are the most precise and are those reported in the "Summary of results." IMGT/V-QUEST results may be helpful to solve ambiguous cases and when IMGT/JunctionAnalysis does not provide results.
18. The 20 amino acids have been classified in 11 "IMGT physicochemical classes," which are based on "Hydrophathy," "Volume," and "Chemical" characteristics of the AA (http://www.imgt.org/, section 'Amino acids' in IMGT Education > Aide-mémoire) [46].
19. In case of differences due to the 5′ primer in V-REGION, it is possible to exclude a given number of nucleotides (IMGT/V-QUEST [22, 23] Search page, "Advanced parameters," Parameters for "Detailed view," and "Nb of nucleotides to exclude

in 5' of the V-REGION for the evaluation of the number of mutations") before launching the analysis.

20. Files from #2 to #6 also include six additional columns: the order of the sequence in file, the Sequence identifier, the V-DOMAIN Functionality, and the names of the V, D, and J genes and alleles. Files from #7 to #10 include four additional columns: the order of the sequence in file, the sequence identifier, the V-DOMAIN functionality, and the name of the V gene and allele.
21. The selection of several "completed" sequence analyses in the same time will combine them as a given batch. Only pertinent combinations or comparisons are allowed. For example, the selection of the sequence analyses from different species or receptor types is forbidden. The selected analyses must include the result format "CSV."
22. "1 copy" are unique sequences from which is built the list of IMGT clonotypes (AA) or (nt). "More than 1" are sequences which are fully identical to one of the "1 copy" set: they are taken into account for the evaluation of the number of sequences assigned to a given IMGT clonotype (AA) or (nt).
23. R is a language and environment for statistical computing and graphics available as free software and downloadable at the CRAN (Comprehensive R Archive Network) website (http://cran.r-project.org/) for Windows, Linux, or Macintosh operating systems. If R is already installed on your computer, please check that the R version is equal or higher to the one indicated in the IMGT/StatClonotype [28, 29] Documentation (http://www.imgt.org/StatClonotype/IMGTStatClonotypeDoc.html#pack).
24. Only alleles of genes having significant differences in proportions validated by all multiple testing procedures are analyzed [28, 29]. By displaying statistical test results per allele, in the case of individuals heterozygous for a given gene, it becomes possible to detect if significant differences in gene proportions, validated by all multiple testing procedures, depend on one allele or not.
25. Above the table, on the left, the number of displayed rows, five by default, can be modified [28, 29]. On the right "Search" allows to enter value in order to filter rows with one or several fields containing it. Clicking on the column title allows to sort the values (alphabetical or number order depending on the column type). Below the name of each column, a filter allows to select values for text fields (e.g., a gene name in column "Gene_Name") or a range for numerical values.

26. Normalized proportions for set 1 and for set 2 represent the numbers of IMGT clonotypes (AA) for a given gene obtained from the IMGT/HighV-QUEST [24–27] outputs normalized for 10,000 IMGT clonotypes (AA) (for clonotype diversity) or for 10,000 sequences assigned to IMGT clonotypes (AA) (for clonotype expression).
27. In addition to other parameters, the left panel allows: (1) the selection of the gene type (V, D, or J); (2) the addition of the locus type to graph axis title: IGH, IGK, IGL, TRA, TRB, TRG, or TRD; (3) the change of the bar colors for Normalized_proportion.set1 and Normalized_proportion.set2; (4) the addition of a title to the graphs for genes and alleles; and (5) the selection of the height and width of the graphs for genes and alleles [28, 29].
28. The CDR-IMGT lengths in the x-axis are not necessarily consecutive values: only CDR-IMGT lengths found in one or both of compared sets are displayed in the graphs.
29. IMGT/DomainGapAlign [30, 31] can analyze also amino acid sequences of C-DOMAIN of IG and TR [14], of V-LIKE-DOMAIN and C-LIKE-DOMAIN of the IgSF other than IG or TR [13, 14, 16, 17], of G-DOMAIN of major histocompatibility (MH) [15], and of G-LIKE-DOMAIN of MhSF other than MH [15–17].
30. In the context of humanization, IMGT/DomainGapAlign [30, 31] allows to precisely define the CDR1-IMGT, CDR2-IMGT, and CDR3-IMGT to be grafted and to select the most appropriate human FR-IMGT by alignment of V-DOMAIN amino acid sequence of the original species (mouse or other species) with the *Homo sapiens* V-REGION and J-REGION reference sets [32, 33].
31. The Smith-Waterman algorithm is used for local sequence alignments of the user AA sequences with the AA IMGT reference directories for V-REGION and J-REGION. The highest alignment scores correspond to the highest sequence similarities.
32. IMGT/Collier-de-Perles [21] provides also 2D graphical representations of C-DOMAIN of IG and TR [14], of V-LIKE-DOMAIN and C-LIKE-DOMAIN of the IgSF other than IG or TR [13, 14, 16, 17], of G-DOMAIN of major histocompatibility (MH), and of G-LIKE-DOMAIN of MhSF other than MH [15–17].
33. "IMGT Colliers de Perles" [18–21] are also provided in IMGT/3Dstructure-DB and IMGT/2Dstructure-DB database entries [30, 50, 51]: the hydrogen bonds within a V-DOMAIN, determined from experimental structural data, are shown as green lines in generated IMGT Collier de Perles on two layers.

34. The number of layers "2" allows to display the two sheets of beta strands of a V-DOMAIN.
35. The background color by default "50% Hydrophobic positions" displays in blue, the positions that have an hydrophobic amino acid (hydropathy index with positive value) or a tryptophan (W) in 50% or more of analyzed V domains [46]. Other background colors have been set for each IGH, IGK, and IGL AA sequences showing the positions, which belong to the same hydropathy classes, volume classes, or physicochemical classes in 80% or more of the analyzed V-DOMAIN.

Acknowledgements

We are very grateful to Gérard Lefranc, founder of the Laboratoire d'ImmunoGénétique Moléculaire LIGM (Université de Montpellier and CNRS), for his unique contribution in the creation of IMGT® in 1989 and his unwavering support for these 30 years. We thank all members of the IMGT® team for their expertise and constant motivation. IMGT® was funded in part by the BIOMED1 (BIOCT930038), Biotechnology BIOTECH2 (BIO4CT960037), 5th PCRDT Quality of Life and Management of Living Resources (QLG2-2000-01287), and 6th PCRDT Information Science and Technology (ImmunoGrid, FP6 IST-028069) programs of the European Union (EU). IMGT® received financial support from the GIS IBiSA, the Agence Nationale de la Recherche (ANR) Labex MabImprove (ANR-10-LABX-53-01), the Région Occitanie Languedoc-Roussillon (Grand Plateau Technique pour la Recherche (GPTR), and BioCampus Montpellier. IMGT® is currently supported by the Centre National de la Recherche Scientifique (CNRS), the Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation (MESRI), the University of Montpellier, and the French Infrastructure Institut Français de Bioinformatique (IFB) ANR-11-INBS-0013. IMGT® is a registered trademark of CNRS. IMGT® is member of the International Medical Informatics Association (IMIA) and a member of the Global Alliance for Genomics and Health (GA4GH). This work was granted access to the High Performance Computing (HPC) resources of Meso@LR and of Centre Informatique National de l'Enseignement Supérieur (CINES), to Très Grand Centre de Calcul (TGCC) of the Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA) and to Institut du développement et des ressources en informatique scientifique (IDRIS) under the allocation 036029 (2010-2022) made by GENCI (Grand Equipement National de Calcul Intensif).

References

1. Lefranc M-P, Lefranc G (2001) The Immunoglobulin FactsBook. Academic Press, London, UK
2. Lefranc M-P, Lefranc G (2020) Immunoglobulins or antibodies: IMGT® bridging genes, structures and functions. Biomedicines 8(9): 319 https://doi.org/10.3390/biomedicines8090319
3. Lefranc M-P, Lefranc G (2001) The T cell receptor FactsBook. Academic Press, London, UK
4. Lefranc M-P (2014) Immunoglobulin and T cell receptor genes: IMGT(®) and the birth and rise of immunoinformatics. Front Immunol 5:22. https://doi.org/10.3389/fimmu.2014.00022
5. Lefranc M-P, Giudicelli V, Duroux P et al (2015) IMGT®, the international ImMunoGeneTics information system® 25 years on. Nucleic Acids Res 43:D413–D422. https://doi.org/10.1093/nar/gku1056
6. Giudicelli V, Lefranc M-P (2012) IMGT-ONTOLOGY 2012. Front Genet 3:79. https://doi.org/10.3389/fgene.2012.00079
7. Duroux P, Kaas Q, Brochet X et al (2008) IMGT-Kaleidoscope, the formal IMGT-ONTOLOGY paradigm. Biochimie 90: 570–583. https://doi.org/10.1016/j.biochi.2007.09.003
8. Lefranc M-P (2000) Nomenclature of the human immunoglobulin genes. In: Coligan JE, Bierer BE, Margulies DE, Shevach EM, Strober W (eds) Current Protocols in Immunology. John Wiley and Sons, Hoboken N.J, pp A.1P.1–A.1P.37
9. Lefranc M-P (2000) Nomenclature of the human T cell receptor genes. In: Coligan JE, Bierer BE, Margulies DE, Shevach EM, Strober W (eds) Current Protocols in Immunology. John Wiley and Sons, Hoboken N.J, pp A.1O.1–A.1O.23
10. Lefranc M-P (2007) WHO-IUIS Nomenclature Subcommittee for immunoglobulins and T cell receptors report. Immunogenetics 59: 899–902. https://doi.org/10.1007/s00251-007-0260-4
11. Lefranc M-P (2008) WHO-IUIS Nomenclature Subcommittee for immunoglobulins and T cell receptors report August 2007, 13th International Congress of Immunology, Rio de Janeiro, Brazil. Dev Comp Immunol 32: 461–463. https://doi.org/10.1016/j.dci.2007.09.008
12. Lefranc M-P (2011) From IMGT-ONTOLOGY CLASSIFICATION Axiom to IMGT standardized gene and allele nomenclature: for immunoglobulins (IG) and T cell receptors (TR). Cold Spring Harb Protoc 2011:627–632. https://doi.org/10.1101/pdb.ip84
13. Lefranc M-P, Pommié C, Ruiz M et al (2003) IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. Dev Comp Immunol 27:55–77
14. Lefranc M-P, Pommié C, Kaas Q et al (2005) IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains. Dev Comp Immunol 29:185–203. https://doi.org/10.1016/j.dci.2004.07.003
15. Lefranc M-P, Duprat E, Kaas Q, Tranne M, Thiriot A, Lefranc G (2005) IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhcSF) G-LIKE-DOMAIN. Dev Comp Immunol 29: 917–938. https://doi.org/10.1016/j.dci.2005.03.003
16. Lefranc M-P (2011) IMGT unique numbering for the variable (V), constant (C), and groove (G) domains of IG, TR, MH, IgSF, and MhSF. Cold Spring Harb Protoc 2011:633–642. https://doi.org/10.1101/pdb.ip85
17. Lefranc M-P (2014) Immunoinformatics of the V, C, and G domains: IMGT® definitive system for IG, TR and IgSF, MH, and MhSF. Methods Mol Biol 1184:59–107. https://doi.org/10.1007/978-1-4939-1115-8_4
18. Ruiz M, Lefranc M-P (2002) IMGT gene identification and colliers de Perles of human immunoglobulins with known 3D structures. Immunogenetics 53:857–883. https://doi.org/10.1007/s00251-001-0408-6
19. Kaas Q, Lefranc M-P (2007) IMGT Colliers de Perles: standardized sequence-structure representations of the IgSF and MhcSF superfamily domains. Curr Bioinforma 2:21–30
20. Kaas Q, Ehrenmann F, Lefranc M-P (2007) IG, TR and IgSF, MHC and MhcSF: what do we learn from the IMGT Colliers de Perles? Brief Funct Genomic Proteomic 6:253–264. https://doi.org/10.1093/bfgp/elm032
21. Ehrenmann F, Giudicelli V, Duroux P, Lefranc M-P (2011) IMGT/collier de Perles: IMGT standardized representation of domains (IG, TR, and IgSF variable and constant domains, MH and MhSF groove domains). Cold Spring

Harb Protoc 2011:726–736. https://doi.org/10.1101/pdb.prot5635

22. Brochet X, Lefranc M-P, Giudicelli V (2008) IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. Nucleic Acids Res 36:W503–W508. https://doi.org/10.1093/nar/gkn316
23. Giudicelli V, Brochet X, Lefranc M-P (2011) IMGT/V-QUEST: IMGT standardized analysis of the immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences. Cold Spring Harb Protoc 2011:695–715. https://doi.org/10.1101/pdb.prot5633
24. Alamyar E, Giudicelli V, Shuo L et al (2012) IMGT/HighV-QUEST: the IMGT® web portal for immunoglobulin (IG) or antibody and T cell receptor (TR) analysis from NGS high throughput and deep sequencing. Immun Res 8(1):3
25. Alamyar E, Duroux P, Lefranc M-P, Giudicelli V (2012) IMGT(®) tools for the nucleotide analysis of immunoglobulin (IG) and T cell receptor (TR) V-(D)-J repertoires, polymorphisms, and IG mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. Methods Mol Biol 882:569–604. https://doi.org/10.1007/978-1-61779-842-9_32
26. Li S, Lefranc M-P, Miles JJ et al (2013) IMGT/HighV QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. Nat Commun 4:2333. https://doi.org/10.1038/ncomms3333
27. Giudicelli V, Duroux P, Lavoie A, Aouinti S, Lefranc M-P, Kossida S (2015) From IMGT-ONTOLOGY to IMGT/HighVQUEST for NGS immunoglobulin (IG) and T cell receptor (TR) repertoires in autoimmune and infectious diseases. Autoimmune Infect Dis 1(1). https://doi.org/10.16966/2470-1025.103
28. Aouinti S, Malouche D, Giudicelli V, Kossida S, Lefranc MP (2015) IMGT/HighV-QUEST statistical significance of IMGT clonotype (AA) diversity per gene for standardized comparisons of next generation sequencing immunoprofiles of immunoglobulins and T cell receptors. PLoS One 10(11): e0142353. https://doi.org/10.1371/journal.pone.0142353
29. Aouinti S, Giudicelli V, Duroux P, Malouche D, Kossida S, Lefranc MP (2016) IMGT/StatClonotype for pairwise evaluation and visualization of NGS IG and TR IMGT clonotype (AA) diversity or expression from IMGT/HighV-QUEST. Front Immunol 7: 339. https://doi.org/10.3389/fimmu.2016.00339
30. Ehrenmann F, Kaas Q, Lefranc M-P (2010) IMGT/3Dstructure-DB and IMGT/DomainGapAlign: a database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhcSF. Nucleic Acids Res 38: D301–D307. https://doi.org/10.1093/nar/gkp946
31. Ehrenmann F, Lefranc M-P (2011) IMGT/DomainGapAlign: IMGT standardized analysis of amino acid sequences of variable, constant, and groove domains (IG, TR, MH, IgSF, MhSF). Cold Spring Harb Protoc 2011: 737–749. https://doi.org/10.1101/pdb.prot5636
32. Lefranc M-P, Ehrenmann F, Ginestoux C, Giudicelli V, Duroux P (2012) Use of IMGT (®) databases and tools for antibody engineering and humanization. Methods Mol Biol 907: 3–37. https://doi.org/10.1007/978-1-61779-974-7_1
33. Lefranc M-P (2014) IMGT® immunoglobulin repertoire analysis and antibody humanization. In: Alt F, Honjo T, Radbruch A, Reth M (eds) Molecular Biology of B Cells, vol 27, 2nd edn. Elsevier Ltd., London, UK, pp 481–514
34. Giudicelli V, Chaume D, Lefranc M-P (2005) IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. Nucleic Acids Res 33: D256–D261. https://doi.org/10.1093/nar/gki010
35. Lefranc M-P, Lefranc G (2019) IMGT® and 30 years of immunoinformatics insight in antibody V and C domain structure and function. Antibodies 8:29. https://doi.org/10.3390/antib8020029
36. Giudicelli V, Duroux P, Ginestoux C et al (2006) IMGT/LIGM-DB, the IMGT comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. Nucleic Acids Res 34:D781–D784. https://doi.org/10.1093/nar/gkj088
37. Yousfi Monod M, Giudicelli V, Chaume D, Lefranc M-P (2004) IMGT/JunctionAnalysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J JUNCTIONs. Bioinformatics 20(Suppl 1):i379–i385. https://doi.org/10.1093/bioinformatics/bth945
38. Giudicelli V, Lefranc M-P (2011) IMGT/JunctionAnalysis: IMGT standardized analysis of the V-J and V-D-J junctions of the rearranged immunoglobulins (IG) and T cell receptors (TR). Cold Spring Harb Protoc 2011:716–725. https://doi.org/10.1101/pdb.prot5634

39. Giudicelli V, Protat C, Lefranc M-P (2003) The IMGT strategy for the automatic annotation of IG and TR cDNA sequences: IMGT/Automat. In: Proceedings of the European Conference on Computational Biology (ECCB 2003). INRIA (DISC/Spid), Paris, DKB-31, pp 103–104
40. Giudicelli V, Chaume D, Jabado-Michaloud J, Lefranc M-P (2005) Immunogenetics sequence annotation: the strategy of IMGT based on IMGT-ONTOLOGY. Stud Health Technol Inform 116:3–8
41. Giudicelli V, Duroux P, Kossida S, Lefranc M-P (2017) IG and TR single chain fragment variable (scFv) sequence analysis: a new advanced functionality of IMGT/V-QUEST and IMGT/HighV-QUEST. BMC Immunol 18:35. https://doi.org/10.1186/s12865-017-0218-8
42. Rubelt F, Busse CE, Bukhari SAC et al (2017) Adaptive Immune Receptor Repertoire Community recommendations for sharing immune-repertoire sequencing data. Nat Immunol 18:1274–1278. https://doi.org/10.1038/ni.3873
43. Vander Heiden JA, Marquez S, Marthandan N et al (2018) AIRR Community Standardized Representations for Annotated Immune Repertoires. Front Immunol 9:2206. https://doi.org/10.3389/fimmu.2018.02206
44. Belessi CJ, Davi FB, Stamatopoulos KE et al (2006) IGHV gene insertions and deletions in chronic lymphocytic leukemia: "CLL-biased" deletions in a subset of cases with stereotyped receptors. Eur J Immunol 36:1963–1974. https://doi.org/10.1002/eji.200535751
45. Rollin M, Giudicelli V, Lefranc M-P IMGT/JunctionAnalysis: IMGT JUNCTION decryption values for (3′V)3′{N}[5′(D)3′{N}]5′(5′J). http://www.imgt.org/IMGT_jcta/decryption. Accessed 14 Jan 2022
46. Pommié C, Levadoux S, Sabatier R, Lefranc M-P (2004) IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. J Mol Recognit 17:17–32. https://doi.org/10.1002/jmr.647
47. Elemento O, Lefranc M-P (2003) IMGT/PhyloGene: an on-line tool for comparative analysis of immunoglobulin and T cell receptor genes. Dev Comp Immunol 27:763–779
48. Hemadou A, Giudicelli V, Smith ML et al (2017) Pacific Biosciences Sequencing and IMGT/HighV-QUEST analysis of full-length single chain Fragment variable from an in vivo selected phage-display combinatorial library. Front Immunol 8:1796. https://doi.org/10.3389/fimmu.2017.01796
49. Han SY, Antoine A, Howard D et al (2018) Coupling of single molecule, long read sequencing with IMGT/HighV-QUEST analysis expedites identification of SIV gp140-specific antibodies from scFv phage display libraries. Front Immunol 9:329. https://doi.org/10.3389/fimmu.2018.00329
50. Kaas Q, Ruiz M, Lefranc M-P (2004) IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. Nucleic Acids Res 32:D208–D210. https://doi.org/10.1093/nar/gkh042
51. Ehrenmann F, Lefranc M-P (2011) IMGT/3Dstructure-DB: querying the IMGT database for 3D structures in immunology and immunoinformatics (IG or antibodies, TR, MH, RPI, and FPIA). Cold Spring Harb Protoc 2011:750–761. https://doi.org/10.1101/pdb.prot5637
52. Agathangelidis A, Darzentas N, Hadzidimitriou A et al (2012) Stereotyped B-cell receptors in one-third of chronic lymphocytic leukemia: a molecular classification with implications for targeted therapies. Blood 119:4467–4475. https://doi.org/10.1182/blood-2011-11-393694
53. Agathangelidis A, Chatzidimitriou A, Gemenetzi K et al (2020) Higher-order connections between stereotyped subsets: implications for improved patient classification in CLL. Blood 137(10):1365–1376. https://doi.org/10.1182/blood.2020007039

Chapter 25

IMGT/3Dstructure-DB: T-Cell Receptor TR Paratope and Peptide/Major Histocompatibility pMH Contact Sites and Epitope

Marie-Paule Lefranc and Gérard Lefranc

Abstract

T-cell receptors (TR), the antigen receptors of T cells, specifically recognize peptides presented by the major histocompatibility (MH) proteins, as peptide/MH (pMH), on the cell surface. The structure characterization of the trimolecular TR/pMH complexes is crucial to the fields of immunology, vaccination, and immunotherapy. IMGT/3Dstructure-DB is the three-dimensional (3-D) structure database of IMGT®, the international ImMunoGenetics information system®. By its creation, IMGT® marks the advent of immunoinformatics, which emerged at the interface between immunogenetics and bioinformatics. The IMGT® immunoglobulin (IG) and TR gene and allele nomenclature (CLASSIFICATION axiom) and the IMGT unique numbering and IMGT/Collier-de-Perles (NUMEROTATION axiom) are the two founding breakthroughs of immunoinformatics. IMGT-ONTOLOGY concepts and IMGT Scientific chart rules generated from these axioms allowed IMGT® bridging genes, structures, and functions. IMGT/3Dstructure-DB contains 3-D structures of IG or antibodies, TR and MH proteins of the adaptive immune responses of jawed vertebrates (*gnathostomata*), IG or TR complexes with antigens (IG/Ag, TR/pMH), related proteins of the immune system of any species belonging to the IG and MH superfamilies, and fusion proteins for immune applications. The focus of this chapter is on the TR V domains and MH G domains and the contact analysis comparison in TR/pMH interactions. Standardized molecular characterization includes "IMGT pMH contact sites" for peptide and MH groove interactions and "IMGT paratopes and epitopes" for TR/pMH complexes. Data are available in the IMGT/3Dstructure database, at the IMGT Home page http://www.imgt.org.

Key words IMGT, T-cell receptor, CDR-IMGT, Major histocompatibility, Paratope, Epitope, TR/pMH, IMGT-ONTOLOGY, Immunoinformatics, IMGT/3Dstructure-DB

1 Introduction

The adaptive immune responses were acquired by jawed vertebrates (or *gnathostomata*) more than 450 million years ago and are found in all extant jawed vertebrate species from fishes to humans [1]. The adaptive immune responses are characterized by a remarkable specificity and memory, which are the properties of the B and T cells owing to an extreme diversity of their antigen receptors [1]. The

Anton W. Langerak (ed.), *Immunogenetics: Methods and Protocols*, Methods in Molecular Biology, vol. 2453, https://doi.org/10.1007/978-1-0716-2115-8_25, © The Author(s) 2022

specific antigen receptors comprise the immunoglobulins (IG) or antibodies of the B cells and plasma cells [2–5] and the T-cell receptors (TR) [6]. Whereas the IG recognize antigens in their native (unprocessed) form, the TR recognize processed antigens, which are presented as peptides by the highly polymorphic major histocompatibility (MH) proteins (in humans HLA for human leukocyte antigens, encoded by genes in the MHC locus) (Fig. 1). T cells are involved in cell-mediated immune response, against a stress of viral, bacterial, fungal, or tumoral origin and identify antigenic peptides presented by the MH proteins as

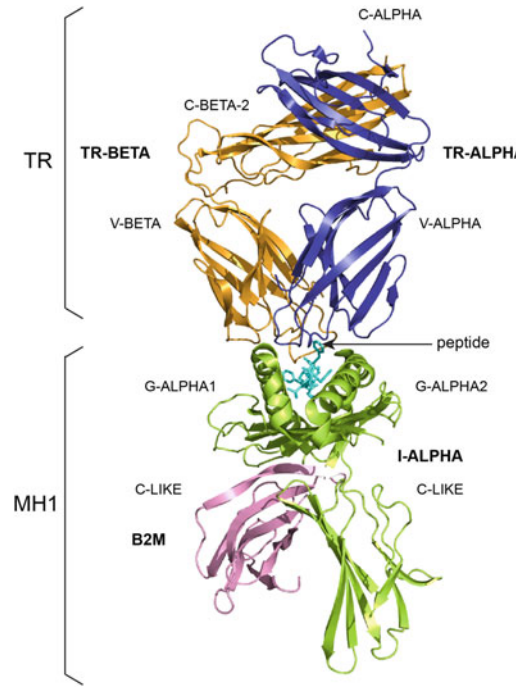


Fig. 1 A T-cell receptor (TR)/peptide-major histocompatibility 1 (pMH1) complex. A TR (here, TR-alpha_beta) is shown (on top, upside down) in complex with an MH (here, MH1) presenting a peptide in its groove [1]. In vivo, a TR is anchored in the membrane of a T cell as part of the signaling T-cell receptor (TcR = TR +CD3). A TR is made of two chains, each comprising a variable domain (V-DOMAIN) at the N-terminal end and a constant domain (C-DOMAIN) at the C-terminal end. The domains are V-ALPHA and C-ALPHA for the TR-ALPHA chain and V-BETA and C-BETA for the TR-BETA chain. An MH1 is made of the I-ALPHA chain with two G-DOMAIN (G-ALPHA1 and G-ALPHA2) and a C-LIKE-DOMAIN (C-LIKE), noncovalently associated with the B2M (a C-LIKE-DOMAIN). In this representation (with G-ALPHA1 on the left, G-ALPHA2 and B2M on the right), the peptide is oriented in the groove from front of the figure to back. The TR/pMH1 complex structure is 3qfj from IMGT/3Dstructure-DB (http:/www.imgt.org). (With permission from M-P. Lefranc and G. Lefranc, LIGM, Founders and Authors of IMGT®, the international ImMunoGeneTics information system®, http:/www.imgt.org)

peptide/MH (pMH) on cell surface [1]. The recognition and signal transduction are carried out by the multiprotein bifunctional T-cell receptor (TcR) assembly that comprises the TR responsible of the specific pMH recognition plus the associated transmembrane signaling CD3 proteins [6]. The TcR is itself associated, in the immunological synapse, with the CD4 or CD8 coreceptors, to the activating CD28 and inhibitory CTLA4 costimulatory proteins, to the CD2 adhesion molecule and to intracellular kinases. The CD8 expressed on most cytotoxic T cells binds the MH class I (MH1) that is expressed ubiquitously on cells of the organism [7]. The CD4 expressed on most helper T cells binds the MH class II (MH2) that is expressed by professional antigen presenting cells (dendritic cells, macrophages, monocytes, and B cells) [7].

IMGT®, the international ImMunoGeneTics information system®, http://www.imgt.org, is the global reference in immunogenetics and immunoinformatics [1], founded in 1989 by Marie-Paule Lefranc at Montpellier (Université de Montpellier and CNRS). It is a high quality integrated knowledge resource comprising 7 databases, 17 online tools, and more than 25,000 pages of web resources [8–11]. IMGT® is specialized in the sequences, structures, and genetic data of the IG, TR, and MH of human and other vertebrate species, in the immunoglobulin superfamily (IgSF) and the MH superfamily (MhSF) of vertebrates and invertebrates, and in related proteins of the immune system (RPI), fusion protein for immune applications, and composite proteins for clinical applications [1, 8–11]. IMGT/3Dstructure-DB [12–14] is the three-dimensional (3-D) structure database of IMGT®. This database provides the standardized IMGT annotation and analysis of the 3-D structures of the TR, pMH, and TR/pMH complexes and comprises detailed molecular characterization and description of their interactions [15–17]. The standardized analysis is based on the concepts of IMGT-ONTOLOGY, the first ontology in immunogenetics and immunoinformatics [18–24]. The IMGT-ONTOLOGY concepts are generated from seven axioms [25–31], of which the CLASSIFICATION axiom (IG and TR gene and allele nomenclature) (*see* **Note 1**) at the birth of IMGT® and immunoinformatics [1, 25] and the NUMEROTATION axiom (IMGT unique numbering [7, 26–27, 32–35] and IMGT Colliers de Perles [28, 29, 36–39]) allow bridging sequences, structures, and functions. The IMGT unique numbering for variable (V) domain includes the IG and TR V-DOMAIN and the V-like domains of IgSF other than IG and TR [32–34]. The IMGT unique numbering for constant (C) domain includes the IG and TR C-DOMAIN and the C-like domains of IgSF other than IG and TR [35]. The IMGT unique numbering for G domain includes the groove (G) domains of the MH G-DOMAIN and the G-like domains of MhSF other than MH (or RPI-MH1Like) [7]. The IMGT/DomainGapAlign tool [13, 40, 41] analyzes the amino acid

sequences of the V, C, and G domains using the IMGT unique numbering [7, 34, 35] and provides a direct link to the IMGT/Collier-de-Perles tool [39]. The IMGT Scientific chart rules provide a standardized description of the contact analysis [15–17] and comparison of TR/pMH complexes and their interactions, irrespective of the TR chains and domains, the MH class (MH1 or MH2), or the species (*Homo sapiens*, *Mus musculus,* etc.). Eleven "IMGT pMH contact sites" were defined for the comparison of pMH interactions, regardless of the peptide lengths [15–17]. The "IMGT pMH contact sites" visualize the interactions between the amino acids (AA) (*see* **Note 2**) of the peptide and those of the MH groove based on the contact analysis. They are a useful asset in peptide vaccine design and epitope prediction, and they precisely identify and visualize AA of the peptide located in the MH2 groove. The standardized "IMGT paratope and epitope" for TR/pMH complexes comprises the TR paratope and the pMH epitope, determined from contact analysis, in IMGT/3Dstructure-DB, at the IMGT Home page http://www.imgt.org.

2 TR and MH Standardized Description in IMGT/3Dstructure-DB

2.1 TR and MH Chains and Domains

2.1.1 TR Chains and Domains

The TR is made of two chains, an alpha chain (TR-ALPHA) and a beta chain (TR-BETA) for the TR-ALPHA_BETA receptor and a gamma chain (TR-GAMMA) and a delta chain (TR-DELTA) for the TR-GAMMA_DELTA receptor [6] (Table 1). Each complete TR chain comprises an extracellular region made up of a V-DOMAIN) (for instance, V-ALPHA for the alpha chain) and a C-DOMAIN) (for instance, C-ALPHA for the alpha chain), a

Table 1
IMGT standardized labels for the DESCRIPTION of the T-cell receptors (TR) and of their chains and domains. IMGT® labels (concepts of description) are written in capital letters [1]

IMGT receptor description	IMGT chain description	IMGT domain description	IMGT region labels
TR-ALPHA_BETA	TR-ALPHA	V-ALPHA	V-J-REGION
		C-ALPHA	Part of C-REGION[a]
	TR-BETA	V-BETA	V-D-J-REGION
		C-BETA	Part of C-REGION[a]
TR-GAMMA_DELTA	TR-GAMMA	V-GAMMA	V-J-REGION
		C-GAMMA	Part of C-REGION[a]
	TR-DELTA	V-DELTA	V-D-J-REGION
		C-DELTA	Part of C-REGION[a]

[a]The TR chain C-REGION also includes the CONNECTING-REGION (CO), the TRANSMEMBRANE-REGION (TM), and the CYTOPLASMIC-REGION (CY), which are not present in the 3-D structures (IMGT® http://www.imgt.org, IMGT Scientific chart >1. Sequence and 3D structure identification and description > Correspondence between labels for IG and TR domains in IMGT/3Dstructure-DB and IMGT/LIGM-DB)

connecting region (CONNECTING-REGION (CO)), a transmembrane region (TRANSMEMBRANE-REGION (TM)), and a short cytoplasmic region (CYTOPLASMIC-REGION (CY)) [6, 7] (Fig. 2, Table 1). The TR V domains that are directly involved in the TR/pMH interactions are described in Subheading 2.2.

2.1.2 MH Chains and Domains

The MH1 is formed by the association of a heavy chain (I-ALPHA) and a light chain (beta-2-microglobulin or B2M). The MH2 is an heterodimer formed by the association of an alpha chain (II-ALPHA) and a beta chain (II-BETA) [7] (Table 2) The I-ALPHA chain of the MH1 and the II-ALPHA and II-BETA chains of the MH2 comprise an extracellular region, made of three domains for the MH1 chains and of two domains for the MH2 chains, and CO, TM, and CY regions [7] (Fig. 2, Table 2). The I-ALPHA chain comprises two groove domains (G-DOMAIN), G-ALPHA1 [D1] and G-ALPHA2 [D2], and one C-LIKE domain [D3] [7]. The B2M corresponds to a single C-LIKE domain. The II-ALPHA chain and the II-BETA chain each comprises two domains, G-ALPHA [D1] and C-LIKE [D2], and G-BETA [D1] and C-LIKE [D2] [7] (Fig. 2). Only the extracellular region that corresponds to these domains has been crystallized. The MH G domains that are directly involved in the TR/pMH interactions are described in Subheading 2.3.

2.2 TR V Domains

2.2.1 Definition

A V domain [32–34] comprises about 100 AA and is made of nine antiparallel beta strands (A, B, C, C′, C″, D, E, F, and G) linked by beta turns (AB, CC′, C″D, DE, and EF) or loops (BC, C′C″, and FG) and forming a sandwich of two sheets (Table 3). The sheets are closely packed against each other through hydrophobic interactions giving a hydrophobic core and joined together by a disulfide bridge between first-CYS at position 23 in the B-STRAND in the first sheet and the second-CYS 104 in the F-STRAND in the second sheet [34]. The V domain type includes the V-DOMAIN of the TR (and IG), which corresponds to the V-J-REGION or V-D-J-REGION encoded by V-(D)-J rearrangements [1–6, 36], and the V-LIKE-DOMAIN of the IgSF other than IG and TR [37–44]. In a V-DOMAIN, the three hypervariable loops BC, C′C″, and FG involved in the ligand (antigen for IG or pMH for TR) recognition are designated as complementarity determining regions (CDR-IMGT) [1–6].

2.2.2 IMGT Unique Numbering for V Domain

The V domain strands and loops and their delimitations and lengths are based on the IMGT unique numbering for V domain (V-DOMAIN and V-LIKE-DOMAIN) [33, 34] (Table 3). In the IG and TR V-DOMAIN, the G-STRAND is the C-terminal part of the J-REGION, with J-PHE or J-TRP 118 and the canonical motif F/W-G-X-G at positions 118–121 [1]. The loop length (number of AA (or codons), which is the number of occupied positions, is a

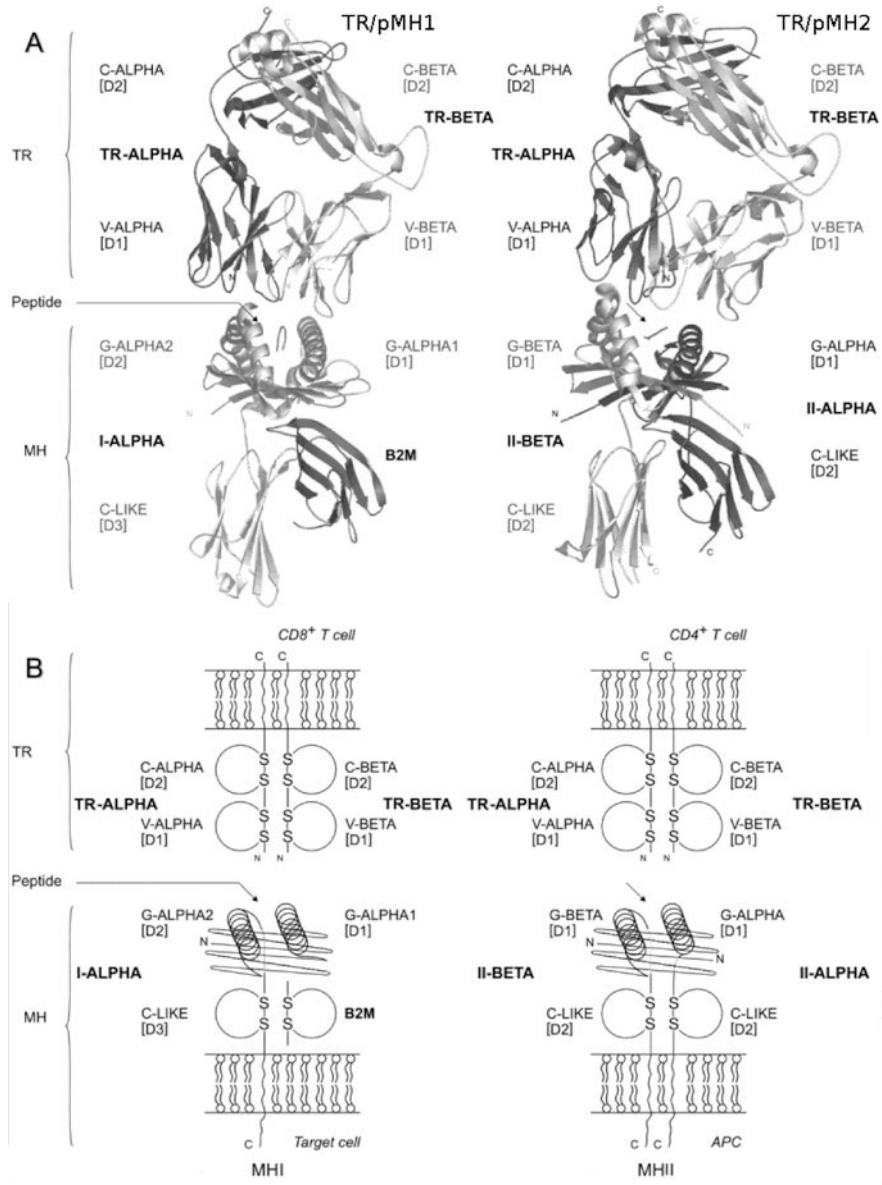


Fig. 2 T-cell receptor/peptide/MH complexes with MH class I (TR/pMH1) and MH class II (TR/pMH2). (**a**) 3-D structures of TR/pMH1 and TR/pMH2. (**b**) Schematic representation of TR/pMH1 and TR/pMH2. The TR (TR-ALPHA and TR-BETA chains), the MH1 (I-ALPHA and B2M chains), and the MH2 (II-ALPHA and II-BETA chains) are shown with the extracellular domains (V-ALPHA and C-ALPHA for the TR-ALPHA chain; V-BETA and C-BETA for the TR-BETA chain; G-ALPHA1, G-ALPHA2, and C-LIKE for the I-ALPHA chain; C-LIKE for B2M; G-ALPHA and C-LIKE for the II-ALPHA chain; II-BETA and C-LIKE for the II-BETA chain), and the connecting, transmembrane, and cytoplasmic regions. [D1], [D2], and [D3] indicate the domains. Arrows indicate the peptide localization in the MH groove made of two G-DOMAIN [7]. In these representations (with G-ALPHA1 on the right, G-ALPHA2 and B2M on the left), the peptide is oriented in the groove from back of the figures to front. (With permission from M-P. Lefranc and G. Lefranc, LIGM, Founders and Authors of IMGT®, the international ImMunoGeneTics information system®, http:/www.imgt.org)

Table 2
IMGT-standardized labels for the DESCRIPTION of the major histocompatibility (MH) and of their chains and domains. IMGT® labels (concepts of description) are written in capital letters [1]

IMGT receptor description	IMGT chain description	IMGT domain description	IMGT domain number
MHC-I-ALPHA_B2M	I-ALPHA	G-ALPHA1	[D1]
		G-ALPHA2	[D2]
		C-LIKE-DOMAIN	[D3][a]
	B2M	C-LIKE-DOMAIN	[D]
MHC-II-ALPHA_BETA	II-ALPHA	G-ALPHA	[D1]
		C-LIKE-DOMAIN	[D2][a]
	II-BETA	G-BETA	[D1]
		C-LIKE-DOMAIN	[D2][a]

[a]The I-ALPHA, II-ALPHA and II-BETA chains includes at the C-terminal end of the C-LIKE-DOMAIN, the CONNECTING-REGION (CO), the TRANSMEMBRANE-REGION (TM), and the CYTOPLASMIC-REGION (CY), which are not present in the 3-D structures

Table 3
V domain strands and loops, IMGT positions and lengths, based on the IMGT unique numbering for V domains (V-DOMAIN and V-LIKE-DOMAIN) [33, 34]

V domain strands and loops[a]	IMGT positions	Lengths[b]	Characteristic Residue@Position[c]	V-DOMAIN FR-IMGT and CDR-IMGT
A-STRAND	1–15	15 (14 if gap at 10)		FR1-IMGT
B-STRAND	16–26	11	1st-CYS 23	
BC-LOOP	27–38	12 (or less)		CDR1-IMGT
C-STRAND	39–46	8	CONSERVED-TRP 41	FR2-IMGT
C′-STRAND	47–55	9		
C′C″-LOOP	56–65	10 (or less)		CDR2-IMGT
C″-STRAND	66–74	9 (or 8 if gap at 73)		FR3-IMGT
D-STRAND	75–84	10 (or 8 if gaps at 81, 82)		
E-STRAND	85–96	12	Hydrophobic 89	
F-STRAND	97–104	8	2nd-CYS 104	
FG-LOOP	105–117	13 (or less, or more)		CDR3-IMGT
G-STRAND	118–128	11 (or 10)	V-DOMAIN J-PHE 118 or J-TRP 118[d]	FR4-IMGT

[a]IMGT® labels (concepts of description) are written in capital letters
[b]In number of AA (or codons)
[c]*See* Subheading 2.4
[d]In the IG and TR V-DOMAIN, the G-STRAND (or FR4-IMGT) is the C-terminal part of the J-REGION, with J-PHE or J-TRP 118 and the canonical motif F/W-G-X-G at positions 118–121. The JUNCTION refers to the CDR3-IMGT plus the two anchors second-CYS 104 and J-PHE or J-TRP 118 [1]

crucial and original concept of IMGT-ONTOLOGY. The lengths of the loops BC (or CDR1-IMGT), C′C″, (or CDR2-IMGT) and FG (or CDR3-IMGT) characterize the V-DOMAIN (Table 3). They are delimited by anchor positions (*see* **Note 3**). The BC loop (or CDR1-IMGT) comprises positions 27–38, the C′C″ (or CDR2-IMGT) positions 56–65, and the FG (or CDR3-IMGT) positions 105–117. In a V-DOMAIN, the CDR3-IMGT that encompasses the V-(D)-J junction resulting from V-J or V-D-J rearrangements [1] is more variable in sequence and length than the CDR1-IMGT and CDR2-IMGT that are encoded by the V-REGION only. The lengths of the three loops BC, C′C″, and FG are shown in number of AA (or codons), into brackets and separated by dots. For example, [9.6.9] means that the BC, C′C″, and FG loops (or CDR1-IMGT, CDR2-IMGT, and CDR3-IMGT for a V-DOMAIN) have a length of 9, 6, and 9 AA (or codons), respectively.

2.2.3 IMGT Colliers de Perles for V Domain

The V domain nine strands are indicated, with their orientation, in the IMGT Colliers de Perles [28, 29, 32–34, 36–39], which are IMGT 2D graphical representations based on the IMGT unique numbering. IMGT Colliers de Perles of the TR V-ALPHA and V-BETA domains from 1ao7 (*see* **Note 4**) a TR/pMH1 3-D structure complex are shown as examples (Fig. 3). The V-ALPHA and V-BETA domains share the main conserved characteristics of the V-DOMAIN, which are the disulfide bridge between cysteine 23 (first-CYS) and cysteine 104 (second-CYS), and the three other hydrophobic core residues tryptophan 41 (CONSERVED-TRP), leucine (or hydrophobic) 89, and phenylalanine 118 (J-PHE) (*see* **Note 5**). In Fig. 3, the V-ALPHA (1ao7_D chain; [6.6.11]) has a CDR1-IMGT and a.

CDR2-IMGT of 6 AA and a CDR3-IMGT of 11 AA, whereas the V-BETA (1ao7_E chain [5.6.14]) has a CDR1-IMGT, CDR2-IMGT, and CDR3-IMGT of 5, 6, and 14 AA, respectively (Subheading 2.2.2) (*see* **Note 6**). In IMGT/3Dstructure-DB, the IMGT genes and alleles that contribute to the V-DOMAIN are determined automatically by IMGT/DomainGapAlign [13, 40, 41], based on the standardized IMGT nomenclature [1, 2, 6] and IMGT unique numbering [34]. Thus, the V-ALPHA of 1ao7_D corresponds to *Homo sapiens* TRAV12-2*02-TRAJ24*02 and the V-BETA of 1ao7_E corresponds to *Homo sapiens* TRBV6-5*01-(TRBD2)-TRBJ2-7*01 [16, 17].

2.3 MH G Domains

2.3.1 Definition

A G domain [7] comprises about 90 AA and is made of a sheet of four antiparallel beta strands linked by turns and of a helix (Table 4); the helix sits on the beta strands, its axis forming an angle of about 40 degrees with the strands [16, 17]. Two G domains are needed to form the MhSF groove made of a “floor” and two “walls” [7]. Each G domain contributes by its four strands

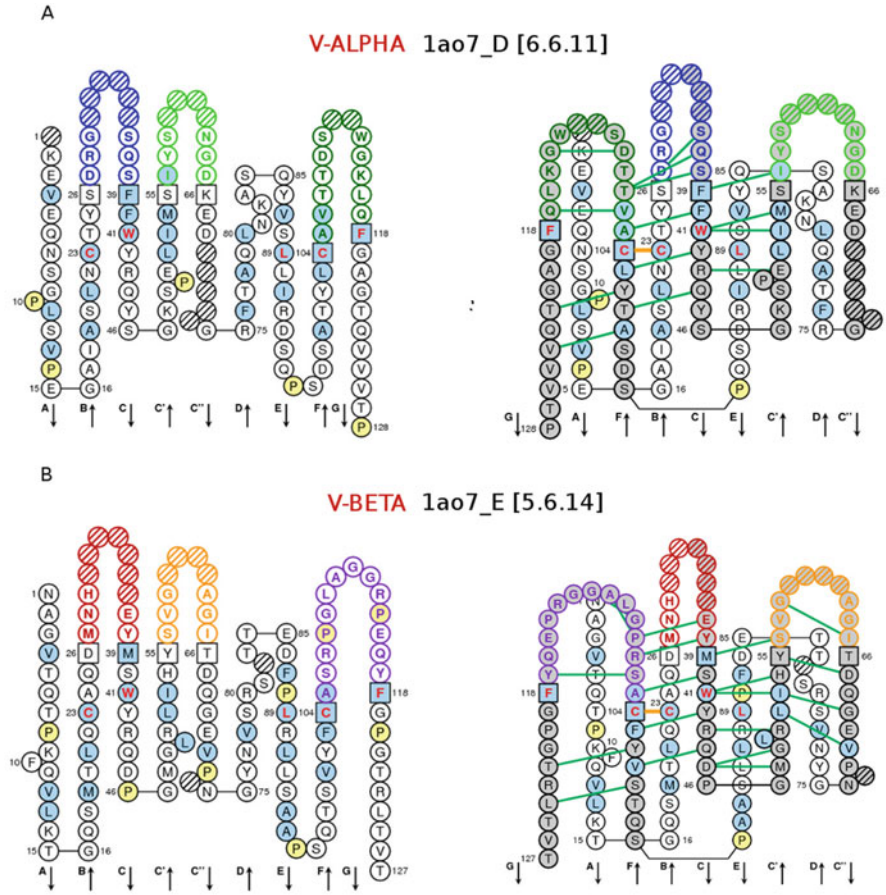


Fig. 3 IMGT/Collier-de-Perles for TR V domain (V-DOMAIN). (**a**) IMGT/Collier-de-Perles for TR V-ALPHA (chain 1ao7_D). The CDR-IMGT lengths are [6.6.11]. (**b**) IMGT/Collier-de-Perles for TR V-BETA (chain 1ao7_E). The CDR-IMGT lengths are [5.6.14]. AA ais shown in the one-letter abbreviation (*see* **Note 2**). Position at which hydrophobic AA (hydropathy index with positive value: I, V, L, F, C, M, A) and tryptophan (W) are found in more than 50% of analyzed sequences are shown in blue, online. All proline (P) are shown in yellow, online. Anchor positions are shown in squares (*see* **Note 3**). Arrows indicate the direction of the beta strands [28, 29]. Hatched circles correspond to missing positions according to the IMGT unique numbering for V domain [33, 34]. IMGT color menu for CDR1-IMGT, CDR2-IMGT, and CDR3-IMGT is blue, green, and greenblue, for V-ALPHA, and red, orange and purple, for V-BETA (*see* **Note 6**). IMGT/Collier-de-Perles are shown on one layer (on the left hand side) and two layers (on the right hand side). The IMGT Colliers de Perles on two layers show, in the forefront, the GFCC′C″ strands and, in the back, the ABED strands. Hydrogen bonds (from the IMGT/3Dstructure-DB entry) are show in green, online. Only those between the AA of the C, C′, C″, F, and G strands (in the forefront) and those of the CDR-IMGT are shown here. IMGT/Collier-de-Perles are from IMGT/3Dstructure-DB, http:/www.imgt.org. (With permission from M-P. Lefranc and G. Lefranc, LIGM, Founders and Authors of IMGT®, the international ImMunoGeneTics information system®, http:/www.imgt.org)

and turns to half of the groove floor and by its helix to one wall of the groove [7, 16, 17]. The G domain type includes the G-DOMAIN of the MH [7] and the G-LIKE-DOMAIN of the MhSF other than MH or RPI-MH1Like [7, 45, 46] (*see* **Note 7**).

Table 4
G domain strands, turns, and helix, IMGT positions and lengths, based on the IMGT unique numbering for G domains (G-DOMAIN and G-LIKE-DOMAIN) [7]

G domain strands, turns and helix[a]	IMGT positions	Lengths[b]	Characteristic Residue@Position[c] and additional positions[d]
A-STRAND	1–14	14	7A, CYS-11
AB-TURN	15–17	3 (or 2 or 0)	
B-STRAND	18–28	11 (or 10[e])	
BC-TURN	29–30	2	
C-STRAND	31–38	8	
CD-TURN	39–41	3 (or 1[f])	
D-STRAND	42–49	8	49.1 to 49.5
HELIX	50–92	43 (or less or more)	54A, 61A, 61B, 72A, CYS-74, 92A

[a]IMGT® labels (concepts of description) are written in capital letters
[b]In number of AA (or codons)
[c]*See* Subheading 2.4
[d]For details on the characteristic Residue@Position and additional positions, *see* Ref. [7]
[e]Or 9 in some G-BETA
[f]Or 0 in some G-ALPHA2-LIKE [7]

2.3.2 IMGT Unique Numbering for G Domain

The G domain strands, turns, and helix and their delimitations and lengths are detailed in Table 4, based on the IMGT unique numbering for G domain (G-DOMAIN and G-LIKE-DOMAIN) [7].

2.3.3 IMGT Colliers de Perles for G Domain

The MH groove in which the peptide binds is made of 2 G-DOMAIN, belonging to the same chain (I-ALPHA) for MH1 or to two different chains (II-ALPHA and II-BETA) for MH2 [7, 37–41]. For the RPI-MH1Like (*see* **Note 7**), the 2 G-LIKE-DOMAIN also belong, as for the MH1, to the same chain (I-ALPHA-LIKE) [7, 37–41]. The IMGT Colliers de Perles (Fig. 4) show, in the upper part of the groove representation, G-ALPHA1 ([D1] of I-ALPHA chain), G-ALPHA ([D1] of II-ALPHA chain) or G-ALPHA1-LIKE ([D1] of I-ALPHA-LIKE chain), and, respectively, in the lower part of the groove representation, G-ALPHA2 ([D2] of I-ALPHA chain), G-BETA ([D1] of II-BETA chain), or G-ALPHA2-LIKE ([D2] of I-ALPHA-LIKE chain). IMGT Colliers de Perles for the MH1 G-ALPHA1 and G-ALPHA2 (1ao7_A chain) are represented in Fig. 4a. IMGT Colliers de Perles for the MH2 G-ALPHA and G-BETA (1j8h_A and 1j8h_B chains, respectively) are represented in Fig. 4b. In IMGT/3Dstructure-DB, the IMGT genes and alleles that encode the G-DOMAIN are determined automatically by IMGT/

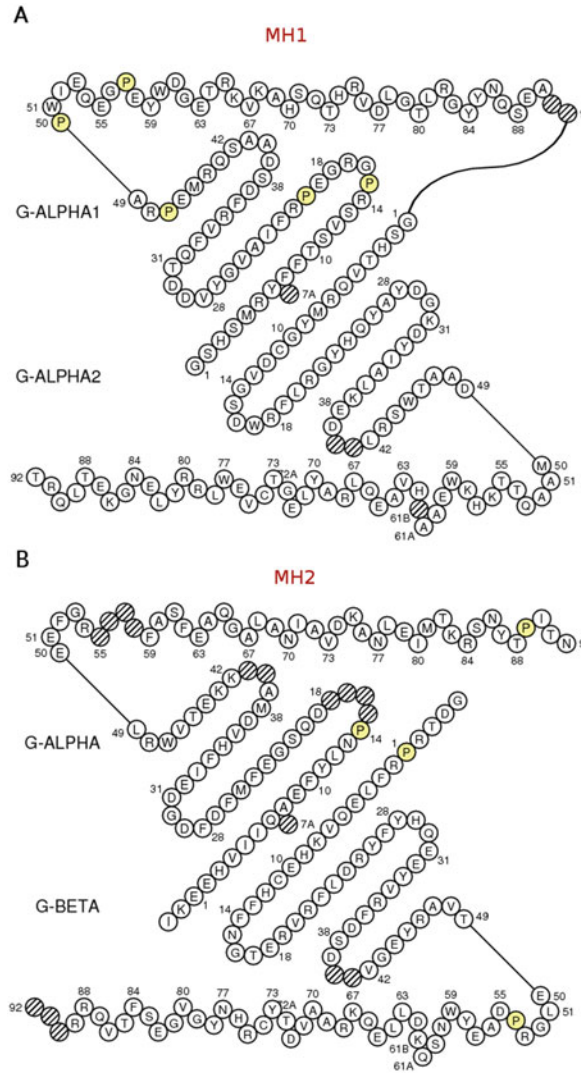


Fig. 4 IMGT/Collier-de-Perles of MH G domains (G-DOMAIN). (**a**) MH1 G-ALPHA1 and G-ALPHA2 domains from 1ao7 (I-ALPHA chain 1ao7_A). (**b**) MH2 G-ALPHA and G-BETA domains from 1j8h (II-ALPHA chain 1j8h_A and II-BETA 1j8h_B, respectively). AA positions and gaps (hatched positions) are according to the IMGT unique numbering for G domain [7]. Positions 61A, 61B, and 72A are characteristic of the G-ALPHA2 and G-BETA domains (and are not reported in the G-ALPHA1 and G-ALPHA IMGT/Collier-de-Perles) [7]. IMGT/Collier-de-Perles are from IMGT/3Dstructure-DB, http:/www.imgt.org. G-domain terminal hatched positions (MH1 G-ALPHA1 91 and 92 and MH2 G-BETA 90, 91 and 92) are not reported in online IMGT/Collier-de-Perles. The IMGT Colliers de Perles can also be obtained, with the sequences gapped by IMGT/DomainGapAlign [40, 41], using the IMGT/Collier-de-Perles tool [39]. (With permission from M-P. Lefranc and G. Lefranc, LIGM, Founders and Authors of IMGT®, the international ImMunoGeneTics information system®, http:/www.imgt.org)

DomainGapAlign [13, 40, 41], based on the standardized IMGT nomenclature and numbering [1, 7]. Thus the G-ALPHA1 and G-ALPHA2 of 1ao7_A are encoded by HLA-A*0201 [15–17].

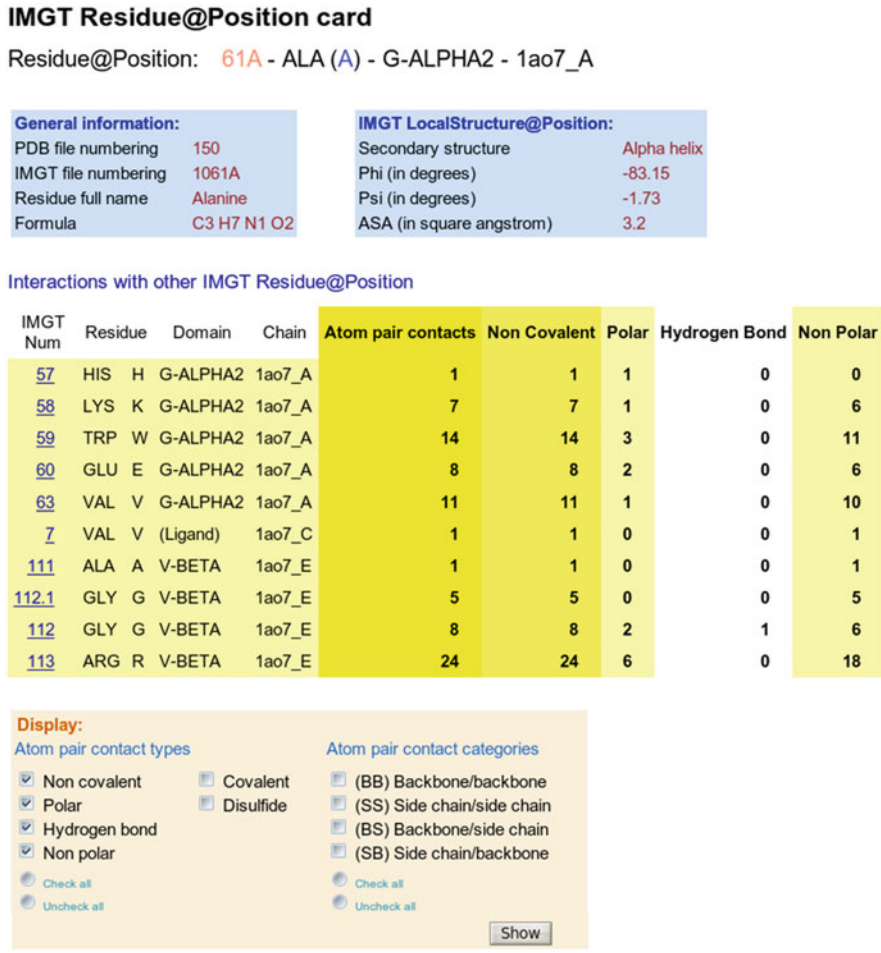
2.4 Residue@Position and Atom Pair Contacts

"Residue@Position" is an IMGT® concept of numerotation that numbers the position of a given residue (or by extension that of a conserved property AA class [47]), based on the IMGT unique numbering (*see* **Note 8**). A "Residue@Position" (R@P) is defined by the position numbering according to the IMGT unique numbering [7, 34, 35], the residue name (in the 3-letter abbreviation and/or in the one-letter abbreviation) (*see* **Note 2**), the IMGT domain label (Tables 1 and 2), and either the gene and allele name for AA sequences (*see* **Note 1**), or the "IMGT chain ID" for 3-D structures. In IMGT/3Dstructure-DB, a 'Residue@Position is described in a "Residue@Position card" (Fig. 5) that provides information on its characteristics (*see* **Note 9**) and the list of the other R@P with which it interacts [16, 17]. Each interaction is characterized by the total number of "atom pair contacts" (*see* **Note 10**) and, as selected by the user for display, the number of atom pair contacts per type ("noncovalent," "polar," "hydrogen bond," "non polar," "covalent," or "disulfide") and/or per category ("(BB) Backbone/backbone," "(SS) Side chain/side chain," "(BS) Backbone/side chain," and "(SB) Side chain/backbone") [16, 17].

3 IMGT pMH Contact Analysis

3.1 IMGT pMH Contact Sites Definition and Determination

"IMGT pMH contact sites" [15–17] highlight the contacts between the amino acids of a presented peptide and those of the floor and helix walls of the MH groove, in 3-D structures of pMH and TR/pMH complexes [12–14]. The "IMGT pMH contact sites" are visualized in IMGT Colliers de Perles for G-DOMAIN [7]. The "IMGT pMH contact sites" provide a standardized comparison of the interactions between a presented peptide and the MH, regardless of the MH class (MH1 or MH2), the G domain (G-ALPHA1, G-ALPHA2, G-ALPHA, and G-BETA), and the peptide length. The "IMGT pMH contact sites" also allow one to precisely identify the AA that is effectively bound in the MH groove. This is particularly informative for the peptides bound to MH2 as these peptides can be much longer than the actual groove length with the N-terminal and C-terminal ends extending outside the groove [7]. In order to deal with different peptide lengths in the groove, 11 standard "IMGT pMH contact sites" were defined (C1–C11) [15–17] (Fig. 6). They correspond to a theoretical maximum length of 11 AA in the groove. This means that, in 3-D structures, some (usually two or three) "IMGT pMH contact sites" are absent as peptides are shorter than 11 AA (usually nine or eight AA long).



IMGT Residue@Position card

Residue@Position: 61A - ALA (A) - G-ALPHA2 - 1ao7_A

General information:	
PDB file numbering	150
IMGT file numbering	1061A
Residue full name	Alanine
Formula	C3 H7 N1 O2

IMGT LocalStructure@Position:	
Secondary structure	Alpha helix
Phi (in degrees)	-83.15
Psi (in degrees)	-1.73
ASA (in square angstrom)	3.2

Interactions with other IMGT Residue@Position

IMGT Num	Residue		Domain	Chain	Atom pair contacts	Non Covalent	Polar	Hydrogen Bond	Non Polar
57	HIS	H	G-ALPHA2	1ao7_A	1	1	1	0	0
58	LYS	K	G-ALPHA2	1ao7_A	7	7	1	0	6
59	TRP	W	G-ALPHA2	1ao7_A	14	14	3	0	11
60	GLU	E	G-ALPHA2	1ao7_A	8	8	2	0	6
63	VAL	V	G-ALPHA2	1ao7_A	11	11	1	0	10
7	VAL	V	(Ligand)	1ao7_C	1	1	0	0	1
111	ALA	A	V-BETA	1ao7_E	1	1	0	0	1
112.1	GLY	G	V-BETA	1ao7_E	5	5	0	0	5
112	GLY	G	V-BETA	1ao7_E	8	8	2	1	6
113	ARG	R	V-BETA	1ao7_E	24	24	6	0	18

Display:

Atom pair contact types: Non covalent, Polar, Hydrogen bond, Non polar, Covalent, Disulfide; Check all, Uncheck all

Atom pair contact categories: (BB) Backbone/backbone, (SS) Side chain/side chain, (BS) Backbone/side chain, (SB) Side chain/backbone; Check all, Uncheck all

Show

Fig. 5 IMGT Residue@Position card. The "Residue@Position: 61A—ALA (A)—G-ALPHA2—1ao7_A" is defined by the position numbering ("61A") according to the IMGT unique numbering for G domain [7], the residue name in the three-letter abbreviation and in the one-letter abbreviation for AA ("ALA (A)") (*see* **Note 2**), the IMGT domain label (G-ALPHA2) (Table 2) and the IMGT chain ID (1ao7_A) (*see* **Note 4**). The list of atom pair contacts shows that this R@P interacts with 5 R@P of the same domain (G-ALPHA2) and, of interest for the TR/pMH interactions, with 4 R@P of the V-BETA and one of the peptide (Ligand). The "Residue@Position" card is from IMGT/3Dstructure-DB, http:/www.imgt.org. (With permission from M-P. Lefranc and G. Lefranc, LIGM, Founders and Authors of IMGT®, the international ImMunoGeneTics information system®, http:/www.imgt.org)

The peptide binding mode to MH1 is characterized by the N-terminal and C-terminal peptide ends docked deeply with the C1 and C11 contact sites (red and pink, respectively, in the IMGT/Collier-de-Perles) and by the peptide length that mechanically constrains the peptide conformation in the groove. Thus, for a peptide of 10 AA, one "IMGT pMH contact sites" is absent (C2), and for a peptide of 9 AA, two "IMGT pMH contact sites" are absent (C2 and C7), whereas for a peptide of 8 AA, three pMH contact sites are absent (C2, C7, and C8) [15–17] (*see* **Note 11**).

Standard 'IMGT pMH contact sites'

	MH1 bound peptides			MH2 bound peptides
	8-AA peptides	9-AA peptides	10-AA peptides	9 AA in the groove
C1	1	1	1	1
C2	-	-	-	2
C3	2	2	2	3
C4	3	3	3	4
C5	4	4	4	5
C6	5	5	5	6
C7	-	-	6	-
C8	-	6	7	-
C9	6	7	8	7
C10	7	8	9	8
C11	8	9	10	9

Fig. 6 Standard "IMGT pMH contact sites'. Eleven standard 'IMGT pMH contact sites' (C1 to C11) were defined for the standardized analysis and comparison of pMH interactions [16, 17]. (With permission from M-P. Lefranc and G. Lefranc, LIGM, Founders and Authors of IMGT®, the international ImMunoGeneTics information system®, http:/www.imgt.org)

The peptide binding mode to MH2 is different with the peptide lying in the groove. Thus, for nine amino acids lying in an MH2 groove, C2 is present but there are no C7 and C8. For a given 3-D structure in IMGT/3Dstructure-DB, the determination of the "IMGT pMH contact sites" combines contact analysis between the peptide and the MH (in a pMH or in a TR/pMH complex), with an interaction scoring function (*see* **Note 12**). The MH AA automatically selects the highest score that is listed and displayed in "IMGT/Collier-de-Perles with pMH contact sites." The characterization of the "IMGT pMH contact sites" based on contact analysis has superseded the previous identification of "pockets" in the MH groove (*see* **Note 13**).

3.2 Access to IMGT pMH Contact Sites

1. In the IMGT® Home page at http://www.imgt.org, click the link "IMGT/3Dstructure-DB and IMGT/2Dstructure-DB" to access the IMGT/3Dstructure-DB Welcome page [12–14].
2. In "IMGT complex type," select "pMH1" or "pMH2" (*see* **Note 14**), or "TR/pMH1" or "TR/pMH2" (*see* **Note 15**), to retrieve the corresponding IMGT/3Dstructure-DB entries.
3. Click on the "IMGT entry ID" to access an individual IMGT/3Dstructure-DB card [13, 14].
4. Click the "Contact analysis" section (in "Chain details" of the MH chain(s)) to access the "IMGT pMH contact sites."

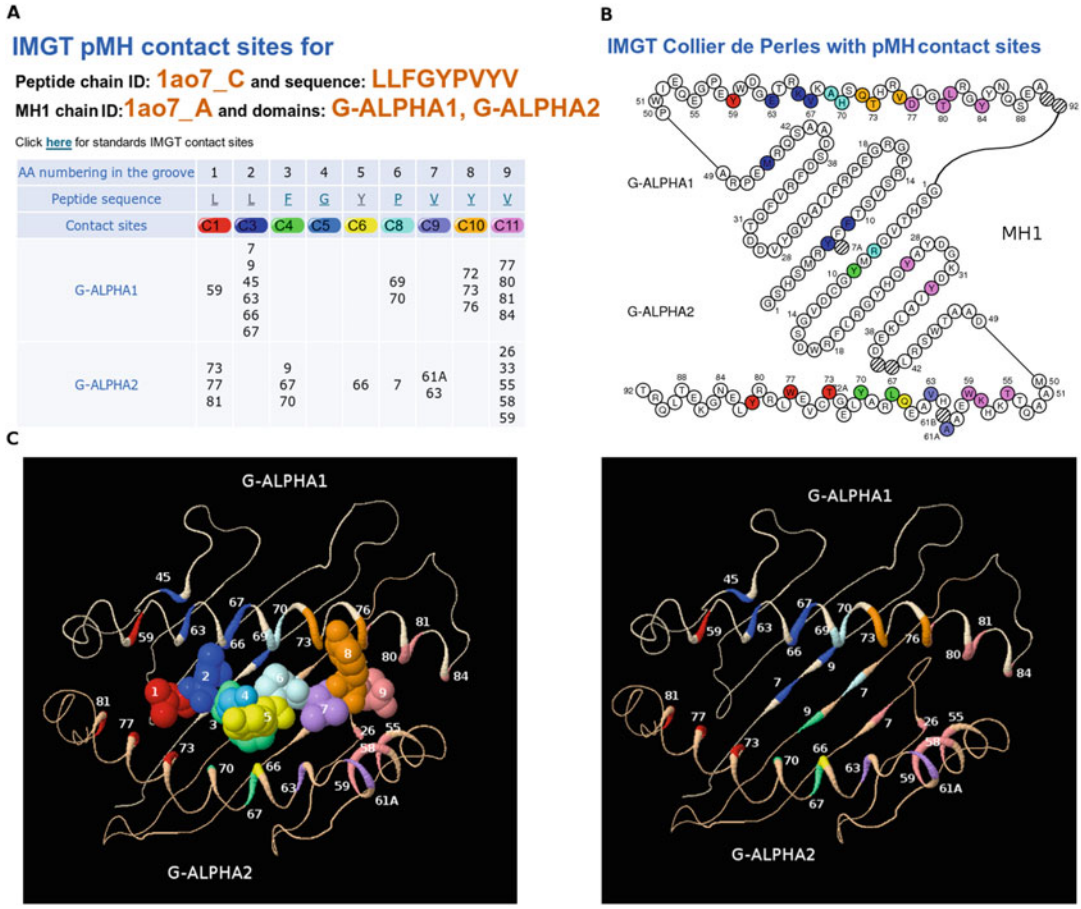


Fig. 7 "IMGT pMH contact sites" between MH1 and a 9-AA peptide. (**a**) "IMGT pMH contact sites" for MH1 (human HLA-A*0201, 1ao7_A) and peptide 1ao7_C. The numbers 1–9 refer to the peptide AA numbering (LLFGYPVYV). C1–C11 refer to the "IMGT pMH contact sites" (there are no C2 and C7 in agreement with MH1 binding a 9-AA peptide). In that 3-D structure, there is no C5 because the glycine G4 score is too low. The G-ALPHA1 and G-ALPHA2 AA positions assigned automatically to the "IMGT pMH contact sites" are listed. (**b**) "IMGT/Collier-de-Perles with pMH contact sites." View is from above the cleft, with G-ALPHA1 on top and G-ALPHA2 on bottom. (**c**) Groove 3-D structure. The groove is shown with and without the peptide (on the left and right hand side, respectively). The IMGT Color menu for "IMGT pMH contact sites" is used in (**a**), (**b**), and (**c**). (**a**) and (**b**) are from IMGT/3Dstructure-DB, http:/www.imgt.org. (With permission from M-P. Lefranc and G. Lefranc, LIGM, Founders and Authors of IMGT®, the international ImMunoGeneTics information system®, http:/www.imgt.org)

3.2.1 IMGT pMH Contact Sites for pMH1

An example of "IMGT pMH contact sites" for pMH1 is shown in Fig. 7. In that 3-D structure of a TR/pMH1 complex (1ao7), the groove made by the G-ALPHA1 and G-ALPHA2 of the I-ALPHA chain (1ao7_A) binds a 9-AA peptide (1ao7_C). "IMGT pMH contact sites" results provide first a table, which shows the positions 1–9 of the peptide AA (each AA is clickable, giving access to its Residue@Position card). Nine of the 11 C1–C11 contact sites are displayed, C2 and C7 being absent, in agreement with a 9-AA

peptide bound in a MH1 groove (*see* Subheading 3.1). The G-ALPHA1 and G-ALPHA2 AA positions that contribute to each "IMGT pMH contact site" are listed. For example, G-ALPHA1 59 and G-ALPHA2 73, 77, and 81 contribute to the "IMGT pMH contact site" C1 that predominantly interacts with leucine (L) 1 of the peptide (N-terminal end) (Fig. 7a). The "IMGT pMH contact sites" are displayed in "IMGT/Collier-de-Perles with pMH contact sites" (Fig. 7b). Clicking on one residue in the IMGT/Collier-de-Perles gives access to its "IMGT Residue@Position card" (*see* Subheading 2.4). The 3-D structure, with or without peptide, is shown in Fig. 7c.

3.2.2 IMGT pMH Contact Sites for pMH2

An example of "IMGT pMH contact sites" for pMH2 is shown in Fig. 8. In that 3-D structure of a TR/pMH2 complex (1j8h), the groove made by the G-ALPHA (of the II-ALPHA chain) (1j8h_A) and the G-BETA (of the II-BETA chain) binds a 13-AA peptide (1j8h_C). "IMGT pMH contact sites" results provide first a table, which shows the AA 1–9 in the groove (each AA is clickable, giving access to its Residue@Position card). However, in contrast to MH1 (*see* Subheading 3.2.1), the nine AA shown in Fig. 8a only correspond to the central part of the peptide. Indeed, the peptide bound to MH2 is longer than the length of the groove and extends outside its N-terminal and C-terminal ends, as the MH2 groove is "open" at both ends [7]. One major breakthrough of the "IMGT pMH contact sites" is the identification of the AA that is located in the MH2 groove [15–17]. Whereas the peptide (1j8h_C) is 13 AA long (PKYVKQNTLKLAT), the "IMGT pMH contact sites" results allow one to determine that the 9 AA in the MH2 groove are YVKQNTLKL (Fig. 8a). Nine of the 11 C1–C11 contact sites are displayed, C7 and C8 being absent, in agreement with 9 AA inside a MH2 groove (*see* Subheading 3.1). The G-ALPHA and G-BETA AA positions that contribute to each 'IMGT pMH contact sites' are listed. They are visualized in the "IMGT Collier de Perles with pMH contact sites" (Fig. 8b). Clicking on one residue in the IMGT Colliers de Perles gives access to its "IMGT Residue@Position card" (*see* Subheading 2.4). The 3-D structure, with or without peptide, is shown in Fig. 8c.

4 IMGT/3Dstructure-DB Domain Pair Contacts

4.1 IMGT/3DStructure-DB Domain Pair Contacts (Overview)

"IMGT/3Dstructure-DB Domain pair contacts (overview)" (Fig. 9) is accessed by clicking on "Domain contacts (overview)" of "Contact analysis" in an IMGT/3Dstructure-DB card. The example shown in Fig. 9 is that of the TR/pMH1 structure 1ao7. Eight "Domain pair contacts" are of interest for TR/pMH interactions, two for pMH1 (*see* Subheading 4.1.1) and six for TR/pMH1 (*see* Subheading 4.1.2). Similar results are obtained for the

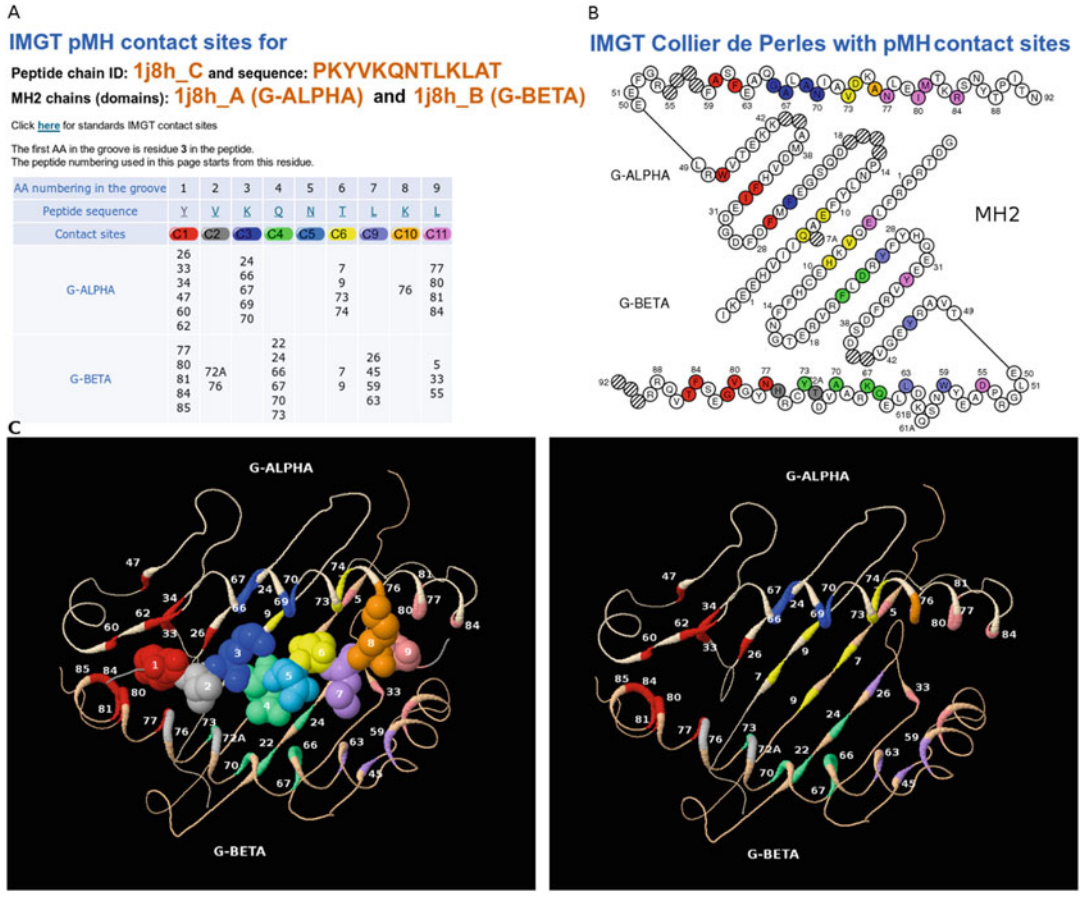


Fig. 8 "IMGT pMH contact sites" between MH2 and 9 AA in the groove. (**a**) "IMGT pMH contact sites" for MH2 (human HLA-DRA*0101_HLA-DRB1*0401) (1j8h_A-1j8h_B) and a 13 AA long peptide (1j8_C). The numbers 1–9 refer to the AA numbering in the groove (YVKQNTLKL) as determined by the "IMGT pMH contact sites." C1–C11 refer to the "IMGT pMH contact sites" (there are no C7 and C8 in agreement with MH2 binding 9 AA in the groove). In that 3-D structure, there is no C5 because the asparagine N5 score is too low. The G-ALPHA and G-BETA AA positions assigned automatically to the "IMGT pMH contact sites" are listed. (**b**) "IMGT/Collier-de-Perles with pMH contact sites." View is from above the cleft, with G-ALPHA on top and G-BETA on bottom. (**c**) Groove 3-D structure. The groove is shown with and without the peptide (on the left and right hand side, respectively). The IMGT Color menu for "IMGT pMH contact sites" is used in (**a**), (**b**), and (**c**). (**a**) and (**b**) are from IMGT/3Dstructure-DB, http:/www.imgt.org. (With permission from M-P. Lefranc and G. Lefranc, LIGM, Founders and Authors of IMGT®, the international ImMunoGeneTics information system®, http:/www.imgt.org)

TR/pMH2 structure, e.g., 1j8h (the only difference being the names of the G-DOMAIN, G-ALPHA, and G-BETA, instead of G-ALPHA1 and G-ALPHA2) (not further detailed here).

4.1.1 Domain Pair Contacts for pMH1 Interactions

The two domain pair contacts for pMH1 interactions are "(Ligand)/G-ALPHA1" and "(Ligand)/G-ALPHA2" (Fig. 9). Thus, for the pMH1 interactions in 1ao7, the domain pair "(Ligand)/G-ALPHA1" shows that 26 residues are involved,

IMGT/3Dstructure-DB Domain pair contacts (overview) of 1ao7

	Unit 1 Domain	Unit 1 Chain	Unit 2 Domain	Unit 2 Chain	Residue pair contacts	Number of residues Total	From 1	From 2	Atom pair contact types Total	Noncovalent	Polar	Hydrogen	Covalent	Disulfide
DomPair	V-ALPHA	1ao7_D	G-ALPHA1	1ao7_A	15	16	9	7	126	126	22	3	0	0
DomPair			G-ALPHA2	1ao7_A	12	15	7	8	105	105	17	2	0	0
DomPair			(Ligand)	1ao7_C	15	13	7	6	109	109	20	3	0	0
DomPair			C-ALPHA	1ao7_D	1	2	1	1	7	7	1	0	0	0
DomPair			V-BETA	1ao7_E	57	42	20	22	401	401	46	7	0	0
DomPair			C-BETA-2	1ao7_E	1	2	1	1	9	9	2	0	0	0
DomPair	C-ALPHA	1ao7_D	V-ALPHA	1ao7_D	1	2	1	1	7	7	1	0	0	0
DomPair	V-BETA	1ao7_E	G-ALPHA1	1ao7_A	3	4	1	3	23	23	0	0	0	0
DomPair			G-ALPHA2	1ao7_A	11	10	5	5	82	82	17	3	0	0
DomPair			(Ligand)	1ao7_C	14	13	9	4	119	119	9	2	0	0
DomPair			V-ALPHA	1ao7_D	57	42	22	20	401	401	46	7	0	0
DomPair			C-BETA-2	1ao7_E	32	27	12	15	236	236	30	1	0	0
DomPair	C-BETA-2	1ao7_E	V-ALPHA	1ao7_D	1	2	1	1	9	9	2	0	0	0
DomPair			V-BETA	1ao7_E	32	27	15	12	236	236	30	1	0	0
DomPair	G-ALPHA1	1ao7_A	G-ALPHA2	1ao7_A	119	77	36	41	961	961	137	22	0	0
DomPair			C-LIKE	1ao7_A	7	8	3	5	62	62	9	2	0	0
DomPair			C-LIKE	1ao7_B	18	18	11	7	153	153	18	6	0	0
DomPair			(Ligand)	1ao7_C	29	26	18	8	305	305	31	5	0	0
DomPair			V-ALPHA	1ao7_D	15	16	7	9	126	126	22	3	0	0
DomPair			V-BETA	1ao7_E	3	4	3	1	23	23	0	0	0	0
DomPair	G-ALPHA2	1ao7_A	G-ALPHA1	1ao7_A	119	77	41	36	961	961	137	22	0	0
DomPair			C-LIKE	1ao7_A	13	13	4	9	98	98	19	1	0	0
DomPair			C-LIKE	1ao7_B	25	20	11	9	246	246	20	3	0	0
DomPair			(Ligand)	1ao7_C	26	24	16	8	281	281	20	5	0	0
DomPair			V-ALPHA	1ao7_D	12	15	8	7	105	105	17	2	0	0
DomPair			V-BETA	1ao7_E	11	10	5	5	82	82	17	3	0	0
DomPair	C-LIKE	1ao7_A	G-ALPHA1	1ao7_A	7	8	5	3	62	62	9	2	0	0
DomPair			G-ALPHA2	1ao7_A	13	13	9	4	98	98	19	1	0	0
DomPair			C-LIKE	1ao7_B	31	25	12	13	310	310	43	8	0	0
DomPair	C-LIKE	1ao7_B		1ao7_1	6	7	6	1	64	64	0	0	0	0
DomPair				1ao7_2	2	3	2	1	6	6	0	0	0	0
DomPair			G-ALPHA1	1ao7_A	18	18	7	11	153	153	18	6	0	0
DomPair			G-ALPHA2	1ao7_A	25	20	9	11	246	246	20	3	0	0
DomPair			C-LIKE	1ao7_A	31	25	13	12	310	310	43	8	0	0

Fig. 9 IMGT/3Dstructure-DB Domain pair contacts (overview). The IMGT/3Dstructure-DB entry is the TR/pMH1 3-D structure 1ao7. The domain partners considered are designated as "Unit 1" and "Unit 2." The number of residue pair contacts, the number of residues involved (total, from Unit 1 and from Unit 2), the number of total atom pair contacts, and, as selected by the user for the display, the number of contacts per type and/or by category are provided. "(Ligand)" refers to the peptide. Two red frames highlight the domain pair contacts for pMH interactions. Two blue rectangles highlight the domain pair contacts for TR/pMH interactions, three for V-ALPHA and three for V-BETA. (With permission from M-P. Lefranc and G. Lefranc, LIGM, Founders and Authors of IMGT®, the international ImMunoGeneTics information system®, http:/www.imgt.org)

8 AA of the peptide (Ligand) interacting with 18 AA of G-ALPHA1 (creating 29 residue pair contacts with a total of 305 atom pair contacts). Similarly, the domain pair "(Ligand)/G-ALPHA2" shows that 24 residues are involved, 8 AA of the peptide interacting with 16 AA of G-ALPHA2 (creating 26 residue pair contacts with a total of 281 atom pair contacts).

4.1.2 Domain Pair Contacts for TR/pMH1 Interactions

The six domain pair contacts for TR/pMH1 interactions include three domain pairs involving V-ALPHA and three domain pairs involving V-BETA (Fig. 9). The TR/pMH1 interactions in 1ao7 are the following:

1. "V-ALPHA/G-ALPHA1": 9 AA of V-ALPHA interact with 7 AA of G-ALPHA1 (15 residue pair contacts with a total of 126 atom pair contacts).
2. "V-ALPHA/G-ALPHA2": 7 AA of V-ALPHA interact with 8 AA of G-ALPHA2 (12 residue pair contacts with a total of 105 atom pair contacts).
3. "V-ALPHA/(Ligand)": 7 AA of V-ALPHA interact with 6 AA of the peptide (15 residue pair contacts with a total of 109 atom pair contacts).
4. "V-BETA/G-ALPHA1": 1 AA of V-BETA interacts with 3 AA of G-ALPHA1 (3 residue pair contacts with a total of 23 atom pair contacts).
5. "V-BETA/G-ALPHA2": 5 AA of V-BETA interact with 5 AA of G-ALPHA2 (11 residue pair contacts with a total of 82 atom pair contacts).
6. "V-BETA/(Ligand)": 9 AA of V-BETA interact with 4 AA of the peptide (14 residue pair contacts with a total of 119 atom pair contacts).

4.2 IMGT/3Dstructure-DB Domain Pair Contacts (Per Pair)

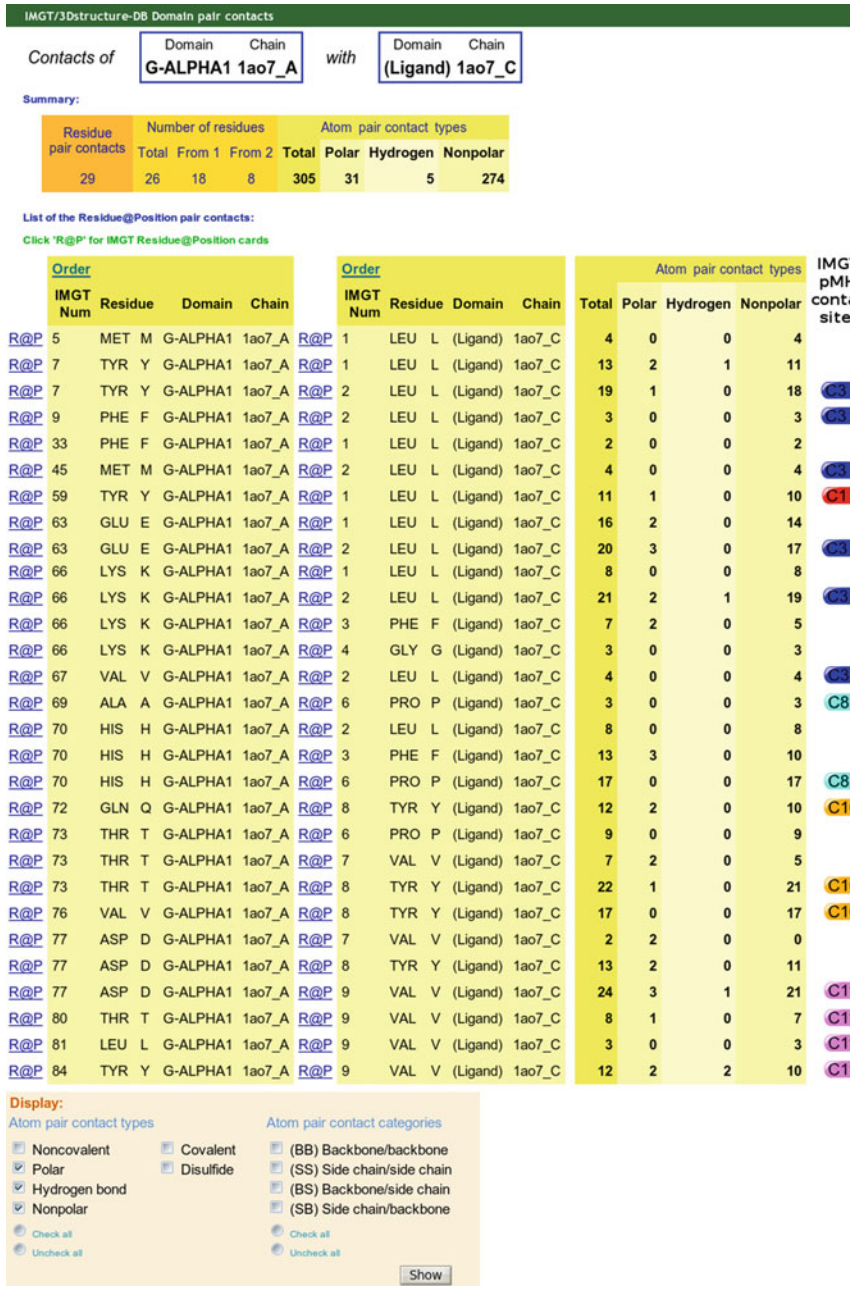
The corresponding detailed description of the "Domain pair contacts" (per pair) that characterize the interactions pMH and TR/pMH is accessed by clicking on "DomPair" (Fig. 9) [12–17].

4.2.1 pMH1 Interactions

The pMH1 interactions "G-ALPHA1/(Ligand)" (Fig. 10a) and "G-ALPHA2/(Ligand)" (Fig. 10b) provide the details of the residue pair contacts with the number of atom pair contact types (total, polar, hydrogen, and nonpolar) (identical results are obtained in the reciprocal queries "(Ligand)/G-ALPHA1" and "(Ligand)/G-ALPHA2"). In these tables, all the peptide—MH1 AA interactions—are listed, in contrast to the "IMGT pMH contact sites," that only visualize those with the highest score (Fig. 7) (*see* Subheading 3).

4.2.2 TR/pMH1 Interactions

The interactions "V-ALPHA/G-ALPHA1," "V-ALPHA/G-ALPHA2," and "V-ALPHA/(Ligand)" are shown in Fig. 11 ((A), (B), and (C), respectively). The interactions "V-BETA/G-ALPHA1," "V-BETA/G-ALPHA2," and "V-BETA/(Ligand)" are shown in Fig. 12((A), (B) and (C), respectively). Positions that belong to the CDR-IMGT are highlighted according to the IMGT color menu: blue (CDR1-IMGT), green (CDR2-IMGT), and greenblue (CDR3-IMGT) for V-ALPHA (Fig. 11) and red (CDR1-IMGT) and purple (CDR3-IMGT) for V-BETA (Fig. 12) (there is no contact with the CDR2-IMGT in 1ao7). Positions can be localized in the IMGT Colliers de Perles (Fig. 3, for V-ALPHA and B-BETA, and Fig. 4 for G-ALPHA1 and G-ALPHA2).



IMGT/3Dstructure-DB Domain pair contacts

Contacts of — Domain: G-ALPHA1, Chain: 1ao7_A — with — Domain: (Ligand), Chain: 1ao7_C

Summary:

Residue pair contacts	Number of residues			Atom pair contact types			
	Total	From 1	From 2	Total	Polar	Hydrogen	Nonpolar
29	26	18	8	305	31	5	274

List of the Residue@Position pair contacts:

Click 'R@P' for IMGT Residue@Position cards

	Order					Order				Atom pair contact types				IMGT pMH contact sites
	IMGT Num	Residue	Domain	Chain		IMGT Num	Residue	Domain	Chain	Total	Polar	Hydrogen	Nonpolar	
R@P	5	MET M	G-ALPHA1	1ao7_A	R@P	1	LEU L	(Ligand)	1ao7_C	4	0	0	4	
R@P	7	TYR Y	G-ALPHA1	1ao7_A	R@P	1	LEU L	(Ligand)	1ao7_C	13	2	1	11	
R@P	7	TYR Y	G-ALPHA1	1ao7_A	R@P	2	LEU L	(Ligand)	1ao7_C	19	1	0	18	C3
R@P	9	PHE F	G-ALPHA1	1ao7_A	R@P	2	LEU L	(Ligand)	1ao7_C	3	0	0	3	C3
R@P	33	PHE F	G-ALPHA1	1ao7_A	R@P	1	LEU L	(Ligand)	1ao7_C	2	0	0	2	
R@P	45	MET M	G-ALPHA1	1ao7_A	R@P	2	LEU L	(Ligand)	1ao7_C	4	0	0	4	C3
R@P	59	TYR Y	G-ALPHA1	1ao7_A	R@P	1	LEU L	(Ligand)	1ao7_C	11	1	0	10	C1
R@P	63	GLU E	G-ALPHA1	1ao7_A	R@P	1	LEU L	(Ligand)	1ao7_C	16	2	0	14	
R@P	63	GLU E	G-ALPHA1	1ao7_A	R@P	2	LEU L	(Ligand)	1ao7_C	20	3	0	17	C3
R@P	66	LYS K	G-ALPHA1	1ao7_A	R@P	1	LEU L	(Ligand)	1ao7_C	8	0	0	8	
R@P	66	LYS K	G-ALPHA1	1ao7_A	R@P	2	LEU L	(Ligand)	1ao7_C	21	2	1	19	C3
R@P	66	LYS K	G-ALPHA1	1ao7_A	R@P	3	PHE F	(Ligand)	1ao7_C	7	2	0	5	
R@P	66	LYS K	G-ALPHA1	1ao7_A	R@P	4	GLY G	(Ligand)	1ao7_C	3	0	0	3	
R@P	67	VAL V	G-ALPHA1	1ao7_A	R@P	2	LEU L	(Ligand)	1ao7_C	4	0	0	4	C3
R@P	69	ALA A	G-ALPHA1	1ao7_A	R@P	6	PRO P	(Ligand)	1ao7_C	3	0	0	3	C8
R@P	70	HIS H	G-ALPHA1	1ao7_A	R@P	2	LEU L	(Ligand)	1ao7_C	8	0	0	8	
R@P	70	HIS H	G-ALPHA1	1ao7_A	R@P	3	PHE F	(Ligand)	1ao7_C	13	3	0	10	
R@P	70	HIS H	G-ALPHA1	1ao7_A	R@P	6	PRO P	(Ligand)	1ao7_C	17	0	0	17	C8
R@P	72	GLN Q	G-ALPHA1	1ao7_A	R@P	8	TYR Y	(Ligand)	1ao7_C	12	2	0	10	C10
R@P	73	THR T	G-ALPHA1	1ao7_A	R@P	6	PRO P	(Ligand)	1ao7_C	9	0	0	9	
R@P	73	THR T	G-ALPHA1	1ao7_A	R@P	7	VAL V	(Ligand)	1ao7_C	7	2	0	5	
R@P	73	THR T	G-ALPHA1	1ao7_A	R@P	8	TYR Y	(Ligand)	1ao7_C	22	1	0	21	C10
R@P	76	VAL V	G-ALPHA1	1ao7_A	R@P	8	TYR Y	(Ligand)	1ao7_C	17	0	0	17	C10
R@P	77	ASP D	G-ALPHA1	1ao7_A	R@P	7	VAL V	(Ligand)	1ao7_C	2	2	0	0	
R@P	77	ASP D	G-ALPHA1	1ao7_A	R@P	8	TYR Y	(Ligand)	1ao7_C	13	2	0	11	
R@P	77	ASP D	G-ALPHA1	1ao7_A	R@P	9	VAL V	(Ligand)	1ao7_C	24	3	1	21	C11
R@P	80	THR T	G-ALPHA1	1ao7_A	R@P	9	VAL V	(Ligand)	1ao7_C	8	1	0	7	C11
R@P	81	LEU L	G-ALPHA1	1ao7_A	R@P	9	VAL V	(Ligand)	1ao7_C	3	0	0	3	C11
R@P	84	TYR Y	G-ALPHA1	1ao7_A	R@P	9	VAL V	(Ligand)	1ao7_C	12	2	2	10	C11

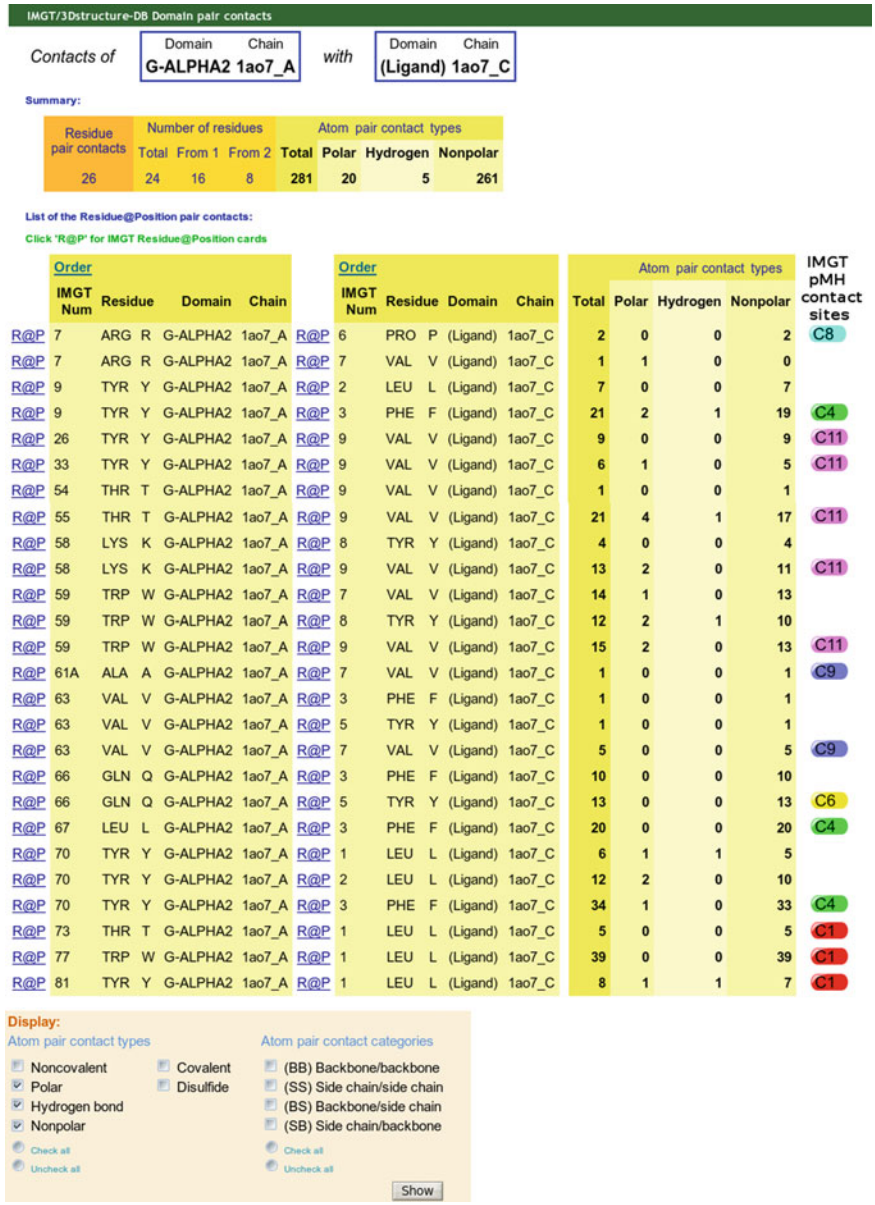
Display:

Atom pair contact types: Noncovalent, Polar, Hydrogen bond, Nonpolar; Covalent, Disulfide; Check all, Uncheck all

Atom pair contact categories: (BB) Backbone/backbone, (SS) Side chain/side chain, (BS) Backbone/side chain, (SB) Side chain/backbone; Check all, Uncheck all

Show

Fig. 10 pMH1 interactions. (**a**) Interactions "G-ALPHA1/(Ligand)" of 1ao7. (**b**) Interactions "G-ALPHA2/(Ligand)" of 1ao7. Clicking on a R@P link gives access to the corresponding IMGT Residue@Position card. "(Ligand)" refers to the peptide. The contact analysis of the TR/pMH 3-D structure 1ao7 is from IMGT/3Dstructure-DB, http:/www.imgt.org. The "IMGT pMH contact sites" for G-ALPHA1 (**a**) and G-ALPHA2 (**b**) were added on the right hand side of the figure, for a comparison with Fig. 7. (With permission from M-P. Lefranc and G. Lefranc, LIGM, Founders and Authors of IMGT®, the international ImMunoGeneTics information system®, http:/www.imgt.org)



IMGT/3Dstructure-DB Domain pair contacts

Contacts of — Domain: G-ALPHA2, Chain: 1ao7_A — with — Domain: (Ligand), Chain: 1ao7_C

Summary:

Residue pair contacts	Number of residues			Atom pair contact types			
	Total	From 1	From 2	Total	Polar	Hydrogen	Nonpolar
26	24	16	8	281	20	5	261

List of the Residue@Position pair contacts:

Click 'R@P' for IMGT Residue@Position cards

	Order						Order					Atom pair contact types				IMGT pMH contact sites
	IMGT Num	Residue		Domain	Chain		IMGT Num	Residue		Domain	Chain	Total	Polar	Hydrogen	Nonpolar	
R@P	7	ARG	R	G-ALPHA2	1ao7_A	R@P	6	PRO	P	(Ligand)	1ao7_C	2	0	0	2	C8
R@P	7	ARG	R	G-ALPHA2	1ao7_A	R@P	7	VAL	V	(Ligand)	1ao7_C	1	1	0	0	
R@P	9	TYR	Y	G-ALPHA2	1ao7_A	R@P	2	LEU	L	(Ligand)	1ao7_C	7	0	0	7	
R@P	9	TYR	Y	G-ALPHA2	1ao7_A	R@P	3	PHE	F	(Ligand)	1ao7_C	21	2	1	19	C4
R@P	26	TYR	Y	G-ALPHA2	1ao7_A	R@P	9	VAL	V	(Ligand)	1ao7_C	9	0	0	9	C11
R@P	33	TYR	Y	G-ALPHA2	1ao7_A	R@P	9	VAL	V	(Ligand)	1ao7_C	6	1	0	5	C11
R@P	54	THR	T	G-ALPHA2	1ao7_A	R@P	9	VAL	V	(Ligand)	1ao7_C	1	0	0	1	
R@P	55	THR	T	G-ALPHA2	1ao7_A	R@P	9	VAL	V	(Ligand)	1ao7_C	21	4	1	17	C11
R@P	58	LYS	K	G-ALPHA2	1ao7_A	R@P	8	TYR	Y	(Ligand)	1ao7_C	4	0	0	4	
R@P	58	LYS	K	G-ALPHA2	1ao7_A	R@P	9	VAL	V	(Ligand)	1ao7_C	13	2	0	11	C11
R@P	59	TRP	W	G-ALPHA2	1ao7_A	R@P	7	VAL	V	(Ligand)	1ao7_C	14	1	0	13	
R@P	59	TRP	W	G-ALPHA2	1ao7_A	R@P	8	TYR	Y	(Ligand)	1ao7_C	12	2	1	10	
R@P	59	TRP	W	G-ALPHA2	1ao7_A	R@P	9	VAL	V	(Ligand)	1ao7_C	15	2	0	13	C11
R@P	61A	ALA	A	G-ALPHA2	1ao7_A	R@P	7	VAL	V	(Ligand)	1ao7_C	1	0	0	1	C9
R@P	63	VAL	V	G-ALPHA2	1ao7_A	R@P	3	PHE	F	(Ligand)	1ao7_C	1	0	0	1	
R@P	63	VAL	V	G-ALPHA2	1ao7_A	R@P	5	TYR	Y	(Ligand)	1ao7_C	1	0	0	1	
R@P	63	VAL	V	G-ALPHA2	1ao7_A	R@P	7	VAL	V	(Ligand)	1ao7_C	5	0	0	5	C9
R@P	66	GLN	Q	G-ALPHA2	1ao7_A	R@P	3	PHE	F	(Ligand)	1ao7_C	10	0	0	10	
R@P	66	GLN	Q	G-ALPHA2	1ao7_A	R@P	5	TYR	Y	(Ligand)	1ao7_C	13	0	0	13	C6
R@P	67	LEU	L	G-ALPHA2	1ao7_A	R@P	3	PHE	F	(Ligand)	1ao7_C	20	0	0	20	C4
R@P	70	TYR	Y	G-ALPHA2	1ao7_A	R@P	1	LEU	L	(Ligand)	1ao7_C	6	1	1	5	
R@P	70	TYR	Y	G-ALPHA2	1ao7_A	R@P	2	LEU	L	(Ligand)	1ao7_C	12	2	0	10	
R@P	70	TYR	Y	G-ALPHA2	1ao7_A	R@P	3	PHE	F	(Ligand)	1ao7_C	34	1	0	33	C4
R@P	73	THR	T	G-ALPHA2	1ao7_A	R@P	1	LEU	L	(Ligand)	1ao7_C	5	0	0	5	C1
R@P	77	TRP	W	G-ALPHA2	1ao7_A	R@P	1	LEU	L	(Ligand)	1ao7_C	39	0	0	39	C1
R@P	81	TYR	Y	G-ALPHA2	1ao7_A	R@P	1	LEU	L	(Ligand)	1ao7_C	8	1	1	7	C1

Display:

Atom pair contact types: Noncovalent, Polar, Hydrogen bond, Nonpolar; Covalent, Disulfide; Check all, Uncheck all

Atom pair contact categories: (BB) Backbone/backbone, (SS) Side chain/side chain, (BS) Backbone/side chain, (SB) Side chain/backbone; Check all, Uncheck all

Show

Fig. 10 (continued)

4.3 IMGT Paratope and Epitope

IMGT paratope and epitope are concepts of the "SpecificityType" in IMGT-ONTOLOGY [48–50]. *Paratope*, or "antigen-binding site," identifies the part of the V-DOMAIN of an IG or antibody ("IG paratope") or of a TR ("TR paratope") that, respectively, recognizes (binds to) the antigen (Ag) or the peptide/major histocompatibility (pMH) ("epitope" or "antigenic determinant") [49]. *Epitope*, or "antigenic determinant," identifies the part of the antigen (Ag) or of the peptide/major histocompatibility

a

IMGT/3Dstructure-DB Domain pair contacts

Contacts of | Domain: V-ALPHA [D1] | Chain: 1ao7_D | with | Domain: G-ALPHA1 [D1] | Chain: 1ao7_A

Summary:

Residue pair contacts	Number of residues			Atom pair contact types			
	Total	From 1	From 2	Total	Polar	Hydrogen	Nonpolar
15	16	9	7	126	22	3	104

List of the Residue@Position pair contacts:

Click 'R@P' for IMGT Residue@Position cards

Order						Order						Atom pair contact types			
	IMGT Num	Residue		Domain	Chain		IMGT Num	Residue		Domain	Chain	Total	Polar	Hydrogen	Nonpolar
R@P	2	LYS	K	V-ALPHA [D1]	1ao7_D	R@P	58	GLU	E	G-ALPHA1 [D1]	1ao7_A	7	1	0	6
R@P	26	SER	S	V-ALPHA [D1]	1ao7_D	R@P	58	GLU	E	G-ALPHA1 [D1]	1ao7_A	3	2	0	1
R@P	27	ASP	D	V-ALPHA [D1]	1ao7_D	R@P	58	GLU	E	G-ALPHA1 [D1]	1ao7_A	24	6	1	18
R@P	28	ARG	R	V-ALPHA [D1]	1ao7_D	R@P	58	GLU	E	G-ALPHA1 [D1]	1ao7_A	1	1	0	0
R@P	37	GLN	Q	V-ALPHA [D1]	1ao7_D	R@P	66	LYS	K	G-ALPHA1 [D1]	1ao7_A	4	1	0	3
R@P	108	THR	T	V-ALPHA [D1]	1ao7_D	R@P	65	ARG	R	G-ALPHA1 [D1]	1ao7_A	5	2	1	3
R@P	108	THR	T	V-ALPHA [D1]	1ao7_D	R@P	66	LYS	K	G-ALPHA1 [D1]	1ao7_A	1	0	0	1
R@P	109	ASP	D	V-ALPHA [D1]	1ao7_D	R@P	62	GLY	G	G-ALPHA1 [D1]	1ao7_A	1	1	0	0
R@P	109	ASP	D	V-ALPHA [D1]	1ao7_D	R@P	65	ARG	R	G-ALPHA1 [D1]	1ao7_A	19	5	1	14
R@P	109	ASP	D	V-ALPHA [D1]	1ao7_D	R@P	66	LYS	K	G-ALPHA1 [D1]	1ao7_A	14	1	0	13
R@P	113	TRP	W	V-ALPHA [D1]	1ao7_D	R@P	65	ARG	R	G-ALPHA1 [D1]	1ao7_A	12	1	0	11
R@P	113	TRP	W	V-ALPHA [D1]	1ao7_D	R@P	68	LYS	K	G-ALPHA1 [D1]	1ao7_A	8	0	0	8
R@P	113	TRP	W	V-ALPHA [D1]	1ao7_D	R@P	69	ALA	A	G-ALPHA1 [D1]	1ao7_A	16	0	0	16
R@P	113	TRP	W	V-ALPHA [D1]	1ao7_D	R@P	72	GLN	Q	G-ALPHA1 [D1]	1ao7_A	4	0	0	4
R@P	114	GLY	G	V-ALPHA [D1]	1ao7_D	R@P	65	ARG	R	G-ALPHA1 [D1]	1ao7_A	7	1	0	6

b

IMGT/3Dstructure-DB Domain pair contacts

Contacts of | Domain: V-ALPHA [D1] | Chain: 1ao7_D | with | Domain: G-ALPHA2 [D2] | Chain: 1ao7_A

Summary:

Residue pair contacts	Number of residues			Atom pair contact types			
	Total	From 1	From 2	Total	Polar	Hydrogen	Nonpolar
12	15	7	8	105	17	2	88

List of the Residue@Position pair contacts:

Click 'R@P' for IMGT Residue@Position cards

Order						Order						Atom pair contact types			
	IMGT Num	Residue		Domain	Chain		IMGT Num	Residue		Domain	Chain	Total	Polar	Hydrogen	Nonpolar
R@P	28	ARG	R	V-ALPHA [D1]	1ao7_D	R@P	77	TRP	W	G-ALPHA2 [D2]	1ao7_A	14	1	0	13
R@P	28	ARG	R	V-ALPHA [D1]	1ao7_D	R@P	80	ARG	R	G-ALPHA2 [D2]	1ao7_A	13	5	0	8
R@P	29	GLY	G	V-ALPHA [D1]	1ao7_D	R@P	77	TRP	W	G-ALPHA2 [D2]	1ao7_A	6	0	0	6
R@P	37	GLN	Q	V-ALPHA [D1]	1ao7_D	R@P	70	TYR	Y	G-ALPHA2 [D2]	1ao7_A	8	0	0	8
R@P	37	GLN	Q	V-ALPHA [D1]	1ao7_D	R@P	73	THR	T	G-ALPHA2 [D2]	1ao7_A	10	2	0	8
R@P	57	TYR	Y	V-ALPHA [D1]	1ao7_D	R@P	65	GLU	E	G-ALPHA2 [D2]	1ao7_A	4	0	0	4
R@P	57	TYR	Y	V-ALPHA [D1]	1ao7_D	R@P	66	GLN	Q	G-ALPHA2 [D2]	1ao7_A	16	1	0	15
R@P	57	TYR	Y	V-ALPHA [D1]	1ao7_D	R@P	69	ALA	A	G-ALPHA2 [D2]	1ao7_A	7	1	0	6
R@P	58	SER	S	V-ALPHA [D1]	1ao7_D	R@P	69	ALA	A	G-ALPHA2 [D2]	1ao7_A	3	1	0	2
R@P	63	ASN	N	V-ALPHA [D1]	1ao7_D	R@P	76	GLU	E	G-ALPHA2 [D2]	1ao7_A	4	2	0	2
R@P	82	LYS	K	V-ALPHA [D1]	1ao7_D	R@P	73	THR	T	G-ALPHA2 [D2]	1ao7_A	5	2	1	3
R@P	82	LYS	K	V-ALPHA [D1]	1ao7_D	R@P	76	GLU	E	G-ALPHA2 [D2]	1ao7_A	15	2	1	13

Fig. 11 TR V-ALPHA/pMH1 interactions. (**a**) Interactions "V-ALPHA/G-ALPHA1" of 1ao7. (**b**) Interactions "V-ALPHA/G-ALPHA2" of 1ao7. (**c**) Interactions "V-ALPHA/(Ligand)" of 1ao7. Clicking on a R@P link gives access to the corresponding IMGT Residue@Position card. "(Ligand)" refers to the peptide. The contact analysis of the TR/pMH 3-D structure 1ao7 is from IMGT/3Dstructure-DB, http:/www.imgt.org. The IMGT color menu is blue, green, and greenblue for CDR1-IMGT, CDR2-IMGT, and CDR3-IMGT, respectively (*see* **Note 6**). (With permission from M-P. Lefranc and G. Lefranc, LIGM, Founders and Authors of IMGT®, the international ImMunoGeneTics information system®, http:/www.imgt.org)

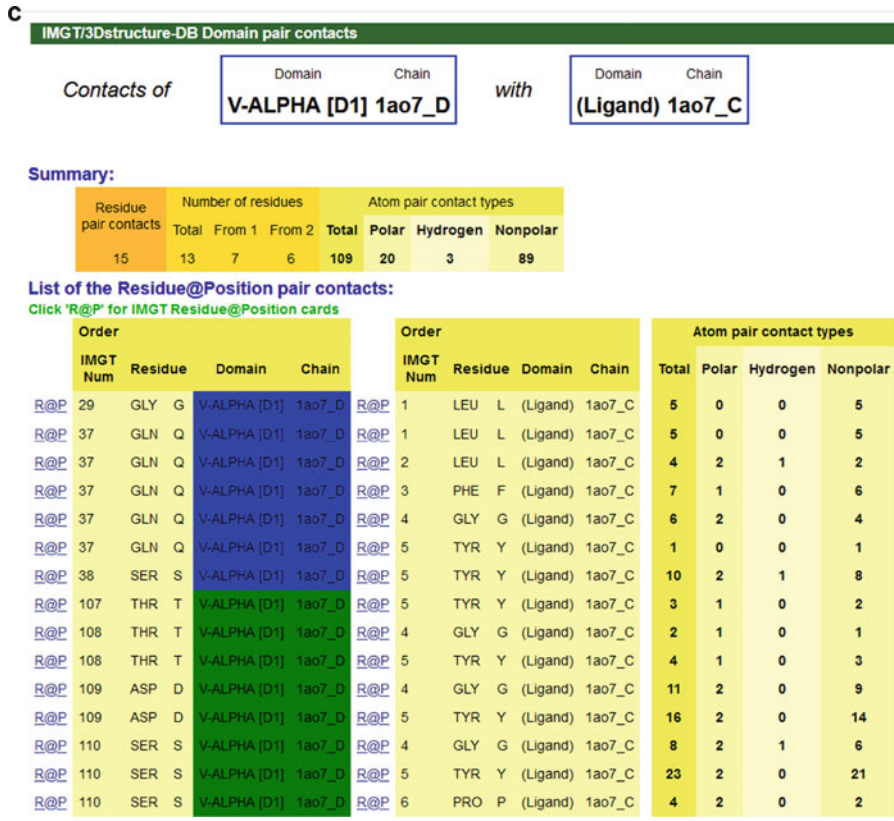


Fig. 11 (continued)

(pMH) that is recognized by the paratope of the V-DOMAIN of an IG or antibody or of a TR, respectively [50].

The amino acids that constitute the TR paratope belong to the paired V domains of a TR (V-alpha and V-beta for a TR-alpha_beta, V-gamma, and V-delta for a TR-gamma_delta), and more precisely to the CDR-IMGT [49]. Among the CDR-IMGT, the CDR3-IMGT that results from the V-J and V-D-J junction play the major role in TR/pMH interactions [15–17]. T-cell epitopes are usually identified as "linear" when referring to the processed peptide (p) presented in the groove of the MH proteins. However, in IMGT-ONTOLOGY, the "T-cell epitope" concept is identified as "discontinuous" as it comprises amino acids of the MH that bind to the TR V domains [50]. Thus, in a TR/pMH complex, the AA in contact at the interface between the TR and the pMH constitute the paratope on the TR surface and the epitope on the pMH surface (Fig. 13). In IMGT/3Dstructure-DB, the "IMGT paratope and epitope" for TR/pMH complexes are determined by combining contact analysis (Table 5) with an interaction scoring function, which roughly complies with the true mean energy ratio [15–17]. A standardized description of the "IMGT paratope and

a

IMGT/3Dstructure-DB Domain pair contacts

Contacts of | Domain: V-BETA [D1] | Chain: 1ao7_E | with | Domain: G-ALPHA1 [D1] | Chain: 1ao7_A

Lorem Ipsum

Summary:

Residue pair contacts	Number of residues			Atom pair contact types			
	Total	From 1	From 2	Total	Polar	Hydrogen	Nonpolar
3	4	1	3	23	0	0	23

List of the Residue@Position pair contacts:

Click 'R@P' for IMGT Residue@Position cards

Order						Order					Atom pair contact types			
	IMGT Num	Residue	Domain	Chain		IMGT Num	Residue	Domain	Chain	Total	Polar	Hydrogen	Nonpolar
R@P	110	LEU L	V-BETA [D1]	1ao7_E	R@P	69	ALA A	G-ALPHA1 [D1]	1ao7_A	9	0	0	9
R@P	110	LEU L	V-BETA [D1]	1ao7_E	R@P	72	GLN Q	G-ALPHA1 [D1]	1ao7_A	4	0	0	4
R@P	110	LEU L	V-BETA [D1]	1ao7_E	R@P	73	THR T	G-ALPHA1 [D1]	1ao7_A	10	0	0	10

Display:

Atom pair contact types: ☐ Noncovalent ☐ Covalent ☑ Polar ☑ Hydrogen bond ☑ Nonpolar ☐ Disulfide

Atom pair contact categories: ☐ (BB) Backbone/backbone ☐ (SS) Side chain/side chain ☐ (BS) Backbone/side chain ☐ (SB) Side chain/backbone

○ Check all ○ Uncheck all (for each group)

Show

b

IMGT/3Dstructure-DB Domain pair contacts

Contacts of | Domain: V-BETA [D1] | Chain: 1ao7_E | with | Domain: G-ALPHA2 [D2] | Chain: 1ao7_A

Summary:

Residue pair contacts	Number of residues			Atom pair contact types			
	Total	From 1	From 2	Total	Polar	Hydrogen	Nonpolar
11	10	5	5	82	17	3	65

List of the Residue@Position pair contacts:

Click 'R@P' for IMGT Residue@Position cards

Order						Order					Atom pair contact types			
	IMGT Num	Residue	Domain	Chain		IMGT Num	Residue	Domain	Chain	Total	Polar	Hydrogen	Nonpolar
R@P	111	ALA A	V-BETA [D1]	1ao7_E	R@P	61A	ALA A	G-ALPHA2 [D2]	1ao7_A	1	0	0	1
R@P	112.1	GLY G	V-BETA [D1]	1ao7_E	R@P	61A	ALA A	G-ALPHA2 [D2]	1ao7_A	5	0	0	5
R@P	112	GLY G	V-BETA [D1]	1ao7_E	R@P	61A	ALA A	G-ALPHA2 [D2]	1ao7_A	8	2	1	6
R@P	112	GLY G	V-BETA [D1]	1ao7_E	R@P	62	HIS H	G-ALPHA2 [D2]	1ao7_A	4	1	0	3
R@P	112	GLY G	V-BETA [D1]	1ao7_E	R@P	63	VAL V	G-ALPHA2 [D2]	1ao7_A	4	0	0	4
R@P	112	GLY G	V-BETA [D1]	1ao7_E	R@P	66	GLN Q	G-ALPHA2 [D2]	1ao7_A	10	2	1	8
R@P	113	ARG R	V-BETA [D1]	1ao7_E	R@P	61	ALA A	G-ALPHA2 [D2]	1ao7_A	5	2	1	3
R@P	113	ARG R	V-BETA [D1]	1ao7_E	R@P	61A	ALA A	G-ALPHA2 [D2]	1ao7_A	24	6	0	18
R@P	113	ARG R	V-BETA [D1]	1ao7_E	R@P	62	HIS H	G-ALPHA2 [D2]	1ao7_A	12	2	0	10
R@P	113	ARG R	V-BETA [D1]	1ao7_E	R@P	66	GLN Q	G-ALPHA2 [D2]	1ao7_A	2	1	0	1
R@P	114	PRO P	V-BETA [D1]	1ao7_E	R@P	66	GLN Q	G-ALPHA2 [D2]	1ao7_A	7	1	0	6

Fig. 12 TR V-BETA/pMH1 interactions. (**a**) Interactions "V-BETA/G-ALPHA1" of 1ao7. (**b**) Interactions "V-BETA/G-ALPHA2" of 1ao7. (**c**) Interactions "V-BETA/(Ligand)." Clicking on a R@P link gives access to the corresponding IMGT Residue@Position card. "(Ligand)" refers to the peptide. The contact analysis of the TR/pMH 3-D structure (1ao7) is from IMGT/3Dstructure-DB, http:/www.imgt.org. R@P belonging to CDR1-IMGT is in red and those belonging to the CDR3-IMGT are in purple according to the IMGT color menu (*see* **Note 6**). (With permission from M-P. Lefranc and G. Lefranc, LIGM, Founders and Authors of IMGT®, the international ImMunoGeneTics information system®, http:/www.imgt.org)

c

IMGT/3Dstructure-DB Domain pair contacts

Contacts of | Domain: V-BETA [D1] | Chain: 1ao7_E | with | Domain: (Ligand) | Chain: 1ao7_C

Summary:

Residue pair contacts	Number of residues			Atom pair contact types			
	Total	From 1	From 2	Total	Polar	Hydrogen	Nonpolar
14	13	9	4	119	9	2	110

List of the Residue@Position pair contacts:

Click 'R@P' for IMGT Residue@Position cards

	Order						Order					Atom pair contact types			
	IMGT Num	Residue		Domain	Chain		IMGT Num	Residue		Domain	Chain	Total	Polar	Hydrogen	Nonpolar
R@P	37	GLU	E	V-BETA [D1]	1ao7_E	R@P	8	TYR	Y	(Ligand)	1ao7_C	10	2	1	8
R@P	107	ARG	R	V-BETA [D1]	1ao7_E	R@P	5	TYR	Y	(Ligand)	1ao7_C	5	2	0	3
R@P	109	GLY	G	V-BETA [D1]	1ao7_E	R@P	6	PRO	P	(Ligand)	1ao7_C	1	1	0	0
R@P	110	LEU	L	V-BETA [D1]	1ao7_E	R@P	6	PRO	P	(Ligand)	1ao7_C	12	1	0	11
R@P	110	LEU	L	V-BETA [D1]	1ao7_E	R@P	7	VAL	V	(Ligand)	1ao7_C	10	1	0	9
R@P	110	LEU	L	V-BETA [D1]	1ao7_E	R@P	8	TYR	Y	(Ligand)	1ao7_C	32	1	1	31
R@P	111	ALA	A	V-BETA [D1]	1ao7_E	R@P	7	VAL	V	(Ligand)	1ao7_C	4	0	0	4
R@P	111	ALA	A	V-BETA [D1]	1ao7_E	R@P	8	TYR	Y	(Ligand)	1ao7_C	5	0	0	5
R@P	112.1	GLY	G	V-BETA [D1]	1ao7_E	R@P	5	TYR	Y	(Ligand)	1ao7_C	1	0	0	1
R@P	112.1	GLY	G	V-BETA [D1]	1ao7_E	R@P	7	VAL	V	(Ligand)	1ao7_C	8	0	0	8
R@P	112	GLY	G	V-BETA [D1]	1ao7_E	R@P	5	TYR	Y	(Ligand)	1ao7_C	9	0	0	9
R@P	112	GLY	G	V-BETA [D1]	1ao7_E	R@P	7	VAL	V	(Ligand)	1ao7_C	2	0	0	2
R@P	113	ARG	R	V-BETA [D1]	1ao7_E	R@P	5	TYR	Y	(Ligand)	1ao7_C	1	0	0	1
R@P	114	PRO	P	V-BETA [D1]	1ao7_E	R@P	5	TYR	Y	(Ligand)	1ao7_C	19	1	0	18

Fig. 12 (continued)

epitope" is provided. Thus, the pMH1 epitope of 1ao7 (Fig. 13) comprises AA of G-ALPHA1 and G-ALPHA2 (1ao7_A) (HLA-A*0201) and of the peptide (1ao7_C, Tax peptide 11–19). Twenty-four AA form the pMH1 epitope: sixteen from the MH1 (six from G-ALPHA1 and ten from G-ALPHA2) and eight from the peptide. Each AA that belongs to the epitope is characterized by its position according to the IMGT unique numbering for G domain [7] and by its position in the peptide.

The TR paratope of 1ao7 (T-cell receptor A6) (Fig. 13) comprises AA of V-ALPHA (1ao7_D chain) and of V-BETA (1ao7_E chain). Sixteen AA of the TR (11 from V-ALPHA and 5 from V-BETA) form the paratope. The IMGT/Collier-de-Perles (Fig. 3) show that nine out of the 11 AA of the V-ALPHA paratope belong to the CDR-IMGT (D27, R28 and G29, Q37 to the CDR1-IMGT, Y57 to the CDR2-IMGT, T108, D109, and W113 and G114 to the CDR3-IMGT) and that five AA of the V-BETA paratope belong to the CDR3-IMGT and are localized at the top of loop. Clicking on "Epitope IMGT Residue@Position cards" and "Paratope IMGT Residue@Position cards" (Fig. 13) provide detailed contacts for each AA belonging to the epitope and paratope, respectively. IMGT paratope and epitope are

Fig. 13 "IMGT paratope and epitope" of an IMGT TR/pMH complex. Each AA that belongs to the pMH epitope is characterized by its position in the peptide or in the G domains according to the IMGT unique numbering [7]. For examples, "E (58G1_A)" means that the glutamate (E) is at position 58 of the G-ALPHA1 domain (1ao7_A), "AAH (61G2-62G2_A)" means that the alanine (A), alanine (A), and histidine (H) are at positions 61, 61A, and 62 of the G-ALPHA2 domain (1ao7_A) (*see* also Fig. 4a). Each AA that belongs to the TR paratope is characterized by its position in the V domains according to the IMGT unique numbering [32–34]. Thus, "DRG (27 V1-29V1_D1)" means that the aspartate (D), arginine (R), and glycine (G) are at positions 27, 28, and 29 of the V domain 1 of 1ao7_D (V-ALPHA) (*see* also Fig. 3a). In the same way, "GGRP (112.1V1-114V1_E)" means that the glycine (G), glycine (G), arginine (R), and proline (P) are at positions 112.1, 112, 113, and 114 of the V domain 1 of 1ao7_E (V-BETA) (*see* also Fig. 3b). The "IMGT paratope and epitope" analysis of the TR/pMH1 3-D structure (1ao7) is from IMGT/3Dstructure-DB, http:/www.imgt.org. (With permission from M-P. Lefranc and G. Lefranc, LIGM, Founders and Authors of IMGT®, the international ImMunoGeneTics information system®, http:/www.imgt.org)

determined automatically for the TR/pMH 3D structures in IMGT/3Dstructure-DB (*see* **Note 15**). Clicking on the "References and links" tag in the IMGT/3Dstructure-DB card gives access to external links. Links to the Immune Epitope Database (IEDB) [51, 52] are provided. Clicking on the "IMGT numbering comparison" displays, per chain, a table providing the correspondence between the IMGT unique numbering per domain and the PDB numbering of the chain entry.

4.4 Bridging IMGT Clonotype (AA), TR-Mimic Antibody and Paratope

4.4.1 IMGT Clonotype (AA) Repertoire and TR Paratope

Next-generation sequencing (NGS) data, analyzed by IMGT/HighV-QUEST, provides a standardized characterization of the TR repertoire diversity and expression in normal (e.g., before and after vaccination) and pathological situations. The results include the IMGT variable (V), diversity (D), and joining (J) gene and allele names (identified at the nucleotide level) and the identification of the JUNCTION lengths and amino sequences, which, together, characterize the IMGT clonotypes (AA) [53]. TR V domain analysis, using IMGT nomenclature [1, 6] and IMGT unique numbering [34] for both NGS and 3-D structures of TR/pMH complexes

Table 5

Amino acids of the TR paratope (V-ALPHA, V-BETA) and of the pMH1 epitope (G-ALPHA1, peptide, and G-ALPHA2) of 1ao7 based on Contact analysis from IMGT/3Dstructure-DB, http://www.imgt.org [12–14]. (A) TR V-ALPHA paratope – pMH1 epitope. (B) TR V-BETA paratope—pMH1 epitope. Amino acid positions of the TR V-ALPHA and V-BETA are according to the IMGT unique numbering for V-DOMAIN [33, 34]. Amino acid positions of the TR V-ALPHA and V-BETA are according to the IMGT unique numbering for G-DOMAIN [7]. The list of contact analysis below is complete. Differences observed in visual displays are due to filters, based on contact types or scores

(A) TR V-ALPHA paratope – pMH1 epitope of 1ao7				
PARATOPE		**EPITOPE**		
TR V-ALPHA [6.6.11][a]		**G-ALPHA1**	**(Ligand) Peptide**	**G-ALPHA2**
[15] 16 (9/7)[b] G-ALPHA1 [15] 13 (7/6)[b] peptide [12] 15 (7/8)[b] G-ALPHA2	1ao7_D	1ao7_A 7 amino acids (126:22,3,104|604) 58 E, 62 G, 65 R, 66 K, 68 K, 69 A, 72 Q	1ao7_C 6 amino acids (109:20,3,89|549) 1 L, 2 L, 3 F, 4 G, 5 Y, 6 P	1ao7_A 8 amino acids (105:17,2,88|468) 65 E, 66 Q, 69 A, 70 Y, 73 T, 76 E,77 W, 80 R
FR1-IMGT	2 K	58 E (7:1,0,6|26)		
	26 S	58 E (3:2,0,1|41)		
CDR1-IMGT	27 D	58 E (24:6,1,18|158)		
	28 R	58 E (1:1,0,0|20)		77 W (14:1,0,13|33) 80 R (13:5,0,8|108)
	29 G		1 L (5:0,0,5|5)	77 W (6:0,0,6|6)
	37 Q	66 K (4:1,0,3|23)	1 L (5:0,0,5|5) 2 L (4:2,1,2|62) 3 F (7:1,0,6|26) 4 G (6:2,0,4|44) 5 Y (1:0,0,1|1)	70 Y (8:0,0,8|8) 73 T (10:2;0,8|48)
	38 S		5 Y (10:2,1,8|68)	
CDR2-IMGT	57 Y			65 E (4:0,0,4|4) 66 Q (16:1,0,15|35) 69 A (7:1,0,6|26)

(continued)

Table 5
(continued)

(A) TR V-ALPHA paratope – pMH1 epitope of 1ao7				
PARATOPE		**EPITOPE**		
TR V-ALPHA [6.6.11][a]		**G-ALPHA1**	**(Ligand) Peptide**	**G-ALPHA2**
	58 S			69 A (3:1,0,2|22)
	63 N			76 E (4:2:0,2|42)
FR3-IMGT	82 K			73 T (5:2,1,3|63) 76 E (15:2,1,13|73)
CDR3-IMGT	107 T		5 Y (3:1,0,2|22)	
	108 T	65 R (5:2,1,3|63) 66 K (1:0,0,1|1)	4 G (2:1,0,1|21) 5 Y (4:1,0,3|23)	
	109 D	62 G (1:1,0,0|20) 65 R (19:5,1,14|134) 66 K (14:1,0,13|33)	4 G (11:2,0,9|49) 5 Y (16:2,0,14|54)	
	110 S		4 G (8:2,1,6|66) 5 Y (23:2,0,21|61) 6 P (4:2,0,2|42)	
	113 W	65 R (12:1,0,11|31) 68 K (8:0,0,8|8) 69 A (16:0,0,16|16) 72 Q (4:0,0,4|4)		
	114G	65 R (7:1,0,6|26)		

(B) TR V-BETA paratope – pMH1 epitope of 1ao7				
PARATOPE			**EPITOPE**	
TR V-BETA [5.6.14][a]		G-ALPHA1	(Ligand) Peptide	G-ALPHA2
[3] 4 (1/3)[b] G-ALPHA1 [14] 13 (9/4)[b] peptide [11] 10 (5/5)[b] G-ALPHA2	1ao7_E	1ao7_A 3 amino acids (23:0,0,23|23) 69 A, 72 Q, 73 T	1ao7_C 4 amino acids (119:9,2110|330) 5 Y, 6 P, 7 V, 8 Y	1ao7_A 6 amino acids (97:19,4,78|538) 61 A, 61A A, 62 H, 63 V, 66 Q, 76 E
CDR1-IMGT	37 E		8 Y (10:2,1,8|68)	
CDR3-IMGT	107 R		5 Y (5:2,0,3|43)	76 E (15:2,1,13|73)
	109 G		6 P (1:1,0,0|20)	
	110 L	69 A (9:0,0,9|9) 72 Q (4:0,0,4|4) 73 T (10:0,0,10|10)	6 P (12:1,0,11|31) 7 V (10:1,0,9|29) 8 Y (32:1,1,31|71)	
	111 A		7 V (4:0,0,4|4) 8 Y (5:0,0,5|5)	61A A (1:0,0,1|1)
	112.1 G		5 Y (1:0,0,1|1) 7 V (8:0,0,8|8)	61A A (5:0,0,5|5)
	112 G		5 Y (9:0,0,9|9) 7 V (2:0,0,2|2)	61A A (8:2,1,6|66) 62 H (4:1,0,3|23) 63 V (4:0,0,4|4) 66 (155) Q (10:2,1,8|68)
	113 R		5 Y (1:0,0,1|1)	61 (149) A (5:2,1,3|63) 61A (150) A (24:6,0,18|138) 62 (151) H (12:2,0,10|50) 66 (155) Q (2:1,0,1|21)
	114 P		5 Y (19:1,0,18|38)	66 (155) Q (7:1,0,6|26)

[a]CDR-IMGT lengths
[b][Residue pair contacts] Number of residues Total (From paratope/From epitope)
[c]Atom pair contact types (Total: polar, hydrogen, nonpolar|score)

[7, 12–14], provides a paradigm for bridging IMGT clonotypes (AA) of NGS repertoires and TR paratope CDR-IMGT (particularly CDR3-IMGT) delimitations.

4.4.2 TR-Mimic Antibody Paratope

The 3-D structures of an engineered TR-mimic antibody and that of a TR targeting peptide-HLA were recently compared**:** the IG Fab 3M4E5 and the TR 1G4_a58b61 are receptors targeting the NY-ESO-1 peptide presented by HLA-A*02:01 [54]. The pMH contacts of the NY-ESO-1 peptide SLLMWITQC with the MH1 HLA-A*02:01 groove are similar in the two peptide-HLA complexes, as expected [55]**.** The paratope of the IG Fab (TR-mimic antibody) includes amino acids of VH [8.8.12] and V-LAMBDA [9.3.9], whereas the paratope of the TR, classically, includes amino acids of V-BETA [5.6.12] and V-ALPHA [6.7.13]. The IMGT unique numbering for V-DOMAIN [34] was used for the four domains in the description of the paratopes. Similarly, in both 3-D structures, the IMGT unique numbering for G-DOMAIN [7] was used for the description of the pMH epitope, which comprises the G-ALPHA1 helix, the peptide, and the G-ALPHA2 helix [55].

5 Availability and Citation

Authors who use IMGT® databases and tools are encouraged to cite this article and to quote the IMGT® Home page, http://www.imgt.org. Online access to IMGT® databases and tools is freely available for academics and under licenses and contracts for companies.

6 Notes

1. Since the creation of IMGT® in 1989, at New Haven during the tenth Human Genome Mapping Workshop (HGM10), the standardized classification and nomenclature of the IG and TR of human and other vertebrate species have been under the responsibility of the IMGT Nomenclature Committee (IMGT-NC). In 1995, following the first demonstration online of the nucleotide database IMGT/LIGM-DB at the ninth International Congress of Immunology in San Francisco, IMGT-NC has become the World Health Organization-International Union of Immunological Societies (WHO-IUIS)/IMGT Nomenclature SubCommittee for IG and TR. IMGT® gene and allele names are based on the concepts of classification of

"Group," "Subgroup," "Gene," and "Allele," generated from the IMGT-ONTOLOGY CLASSIFICATION axiom. The IMGT® gene nomenclature for IG and TR genes was approved at the international level by the Human Genome Organisation (HUGO) Nomenclature Committee (HGNC) in 1999 and by the WHO-IUIS [56, 57]. The IMGT® IG and TR gene names [2, 6, 58, 59] are the official reference for the vertebrate genome projects and, as such, have been entered in IMGT/GENE-DB, the IMGT® gene database [60], in National Center for Biotechnology Information (NCBI) Gene [61], in European Bioinformatics Institute (EBI) Ensembl, and in the Vega Genome Browser (Wellcome Trust Sanger Institute).

2. AA one-letter and three-letter abbreviations: A (Ala), alanine; C (Cys), cysteine; D (Asp), aspartic acid; E (Glu), glutamic acid; F (Phe), phenylalanine; G (Gly), glycine; H (His), histidine; I (Ileu), isoleucine; K (Lys), lysine; L (Leu), leucine; M (Met), methionine; N (Asn), asparagine; P (Pro), proline; Q (Gln), glutamine; R (Arg), arginine; S (Ser), serine; T (Thr), threonine; V (Val), valine; W (Trp), tryptophan; and Y (Tyr), tyrosine. In Residue@Position (Subheading 2.4), the AA three-letter abbreviation is in capital letters. AA physicochemical properties [47] are described in IMGT Aide-mémoire, in the section "Amino acids," http://www.imgt.org.
3. Anchor positions, first defined for V domains, belong to the strands (or FR-IMGT in V-DOMAIN) and represent "anchors" supporting the three BC, C'C", and FG loops (CDR1-IMGT, CDR2-IMGT, and CDR3-IMGT, respectively, in V-DOMAIN). Anchor positions for V domains (V-DOMAIN and V-LIKE-DOMAIN) are positions 26 and 39, 55 and 66, and 104 and 118 [34]. By analogy, anchor positions were defined in C domains at positions 26 and 39, 45 and 77 (delimiting the transverse CD strand), and 104 and 118 [35]. Anchor positions are shown in squares in the IMGT Colliers de Perles.
4. "1ao7" is the code of a 3-D structure entry in the *Research Collaboratory for Structural Bioinformatics* (RCSB) Protein Data Bank (PDB) [62], or "PDB code" (comprising four letters and/or numbers). IMGT® uses the "PDB code" as "IMGT entry ID" for the 3-D structures in IMGT/3Dstructure-DB, http://www.imgt.org [12–14]. An additional letter separated by a "_" identifies the different chains in a 3-D structure. For example, the 1ao7 entry (a TR/pMH1 3D structure) comprises the following chains: 1ao7_D (TR-ALPHA) and 1ao7_E (TR-BETA-1) for the TR, 1ao7_C for the peptide, and 1ao7_A (I-ALPHA) and 1ao7_B (B2M) for the MH1.

5. The other characteristic AA at position 118 of V-DOMAIN is tryptophan (J-TRP) (Table 3) that is found in the IG heavy (IGH) joining (J) regions [3] (and is also found in one human T-cell receptor alpha TRA J region [6]).
6. IMGT color menu for the CDR-IMGT of a V-DOMAIN indicates the type of rearrangement, V-J or V-D-J [1]. Thus, the IMGT color menu for CDR1-IMGT, CDR2-IMGT, and CDR3-IMGT is blue, green, and greenblue for V-ALPHA (encoded by a V-J-REGION resulting from a V-J rearrangement) and red, orange, and purple for V-BETA (encoded by a V-D-J-REGION resulting from a V-D-J rearrangement). The color menu red, orange, and purple is also used for the V-LIKE-DOMAIN BC, C′C″, and FG loops, respectively. The assignment is done automatically by IMGT/DomainGapAlign [13, 40, 41].
7. MhSF proteins other than MH only include RPI-MH1Like proteins (there is no "RPI-MH2Like" identified so far) [45, 46]. The RPI-MH1Like in humans comprise: AZGP1 (that regulates fat degradation in adipocytes), CD1A to CD1E proteins (that display phospholipid antigens to T cells and participate in immune defense against microbian pathogens), FCGRT (that transports maternal immunoglobulins through placenta and governs neonatal immunity), HFE (that interacts with transferring receptor and takes part in iron homeostasis by regulating iron transport through cellular membranes), MICA and MICB (that are induced by stress and involved in tumor cell detection), MR1 (that may regulate mucosal immunity), PROCR, previously EPCR, (that interacts with activated C protein and is involved in the blood coagulation pathway), and RAET1E, RAETG, and RAET1L (that are inducible by retinoic acid and stimulate cytokine/chemokine production and cytotoxic activity of NK cells) [45, 46].
8. The princeps references for the IMGT unique numbering, used to define Residue@Position, are available as pdf on the IMGT® web site (http://www.imgt.org) in the IMGT Scientific chart section: IMGT unique numbering for V domain (V-DOMAIN of IG and TR and V-LIKE-DOMAIN of IgSF other than IG and TR) [32–34], IMGT unique numbering for C domain (C-DOMAIN of IG and TR and C-LIKE-DOMAIN of IgSF other than IG and TR) [35], and IMGT unique numbering for G domain (G-DOMAIN of MH and G-LIKE-DOMAIN of MhSF other than MH) [7].
9. "Residue@Position" characteristics include general information (PDB file numbering, IMGT file numbering, residue full name and formula) and structural information "IMGT Local-Structure@Position" (secondary structure, Phi and Psi angles (in degrees), and accessible surface area (ASA) (in square angstrom)).

10. Atom pair contacts identify interactions between atoms of two "R@P." They are obtained in IMGT/3Dstructure-DB by a local program in which atoms are considered to be in contact when no water molecule can take place between them [16, 17].
11. The "absent" "IMGT pMH contact sites" are correlated to MH class and to peptide length in the groove (Fig. 6). However, it is worthwhile to note that the "IMGT pMH contact sites" were initially defined from statistical analysis using experimental data from IMGT/3Dstructure-DB, without a priori of the MH class and of the peptide lengths [16, 17].
12. For the determination of the "IMGT pMH contact sites" in IMGT/3Dstructure-DB, all direct contacts (defined with a cutoff equal to the sum of the atom van der Waals radii and of the diameter of a water molecule) and water-mediated hydrogen bonds are taken into account [16, 17]. Then, MH AA involved in the pMH binding interface are filtered and classified in "IMGT pMH contact sites" by combining contact analysis with an interaction scoring function. The score assigned to each contact is a constant value, independent on the distance between atoms [40 for direct hydrogen bond, 20 for water mediated hydrogen bond, 20 for polar interaction, and 1 for nonpolar interaction, which roughly complies with the true mean energy ratio] [16, 17].
13. The "IMGT pMH contact sites" C1 and C11 correspond approximatively to the MH1 "pockets" A and F, respectively. A correspondence between the "IMGT pMH contact sites" and the other "pockets" is much more approximative. Thus, for MH1 with a 8-AA peptide, the "IMGT pMH contact sites" C3, C4, C6, and C9 correspond roughly to the B, D, C, and E pockets, and for MH1 with a 9-AA peptide, "IMGT pMH contact sites" C3, C4, and C9 correspond roughly to the B, D, and E pockets. For MH2, the correspondence is not possible because the pockets are poorly defined.
14. In January 2021, IMGT/3Dstructure-DB [12–14] contained 847 pMH (of which 754 are pMH1 and 93 are pMH2). Two hundred thirty-nine of the pMH belong to trimolecular TR/pMH complexes) [15–17] (*see* **Note 15**).

	Number of pMH in IMGT/3Dstructure-DB			
Complex type	***Homo sapiens***	***Mus musculus***	**Other species**	**Total**
pMH1	576	170	8	754
pMH2	84	8	1	93
Total	660	178	9	847

15. In January 2021, IMGT/3Dstructure-DB [12–14] contained 239 TR/pMH complexes (of which 186 are TR/pMH1 and 53 are TR/pMH2) [15–17].

	Number of TR/pMH complexes in IMGT/3Dstructure-DB			
Complex type	***Homo sapiens***	***Mus musculus***	**Other species**	**Total**
TR/pMH1	135	50	1	186
TR/pMH2	31	21	1	53
Total	166	71	2	239

Acknowledgements

We thank Patrice Duroux for the IMGT/3Dstructure-DB database and associated tools computing management, Anjana Kushwaha for the IMGT/3Dstructure-DB entries biocuration, and François Ehrenmann for help with the 3-D structure figures. We are grateful to the IMGT® team for its expertise and constant motivation. We thank Cold Spring Harbor Protocol Press for the pdf of the IMGT Booklet available in IMGT references. IMGT® is a registered trademark of CNRS. IMGT® is member of the International Medical Informatics Association (IMIA) and a member of the Global Alliance for Genomics and Health (GA4GH). All figures are used with permission from M-P. Lefranc and G. Lefranc, LIGM, Founders and Authors of IMGT®, the international ImMunoGeneTics information system®, http://www.imgt.org).

Funding: IMGT® was funded in part by the BIOMED1 (BIOCT930038), Biotechnology BIOTECH2 (BIO4CT960037), fifth PCRDT Quality of Life and Management of Living Resources (QLG2-2000-01287), and sixth PCRDT Information Science and Technology (ImmunoGrid, FP6 IST-028069) programs of the European Union (EU). IMGT® received financial support from the GIS IBiSA, the Agence Nationale de la Recherche (ANR) Labex MabImprove (ANR-10-LABX-53-01), the Région Occitanie Languedoc-Roussillon (Grand Plateau Technique pour la Recherche (GPTR), and BioCampus Montpellier. IMGT® is currently supported by the Centre National de la Recherche Scientifique (CNRS), the Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation (MESRI), and the University of Montpellier.

References

1. Lefranc M-P (2014) Immunoglobulin (IG) and T cell receptor genes (TR): IMGT® and the birth and rise of immunoinformatics. Front Immunol 5:22. https://doi.org/10.3389/fimmu.2014.00022
2. Lefranc M-P, Lefranc G (2001) The immunoglobulin FactsBook. Academic Press, London, UK, pp 1–458
3. Lefranc M-P, Lefranc G (2020) Immunoglobulins or antibodies: IMGT® bridging genes, structures and functions. Biomedicines 8(9): E319. https://doi.org/10.3390/biomedicines8090319
4. Lefranc LM-P, G. (2019) IMGT® and 30 years of Immunoinformatics Insight in Antibody V and C Domain Structure and Function. In Jefferis R; Strohl W. R., Kato K. Antibodies (Basel) 8(2):29. https://doi.org/10.3390/antib8020029
5. Lefranc M-P (2014) Immunoglobulins: 25 years of Immunoinformatics and IMGT-ONTOLOGY. Biomol Ther 4(4):1102–1139. https://doi.org/10.3390/biom4041102
6. Lefranc M-P, Lefranc G (2001) The T cell receptor FactsBook. Academic Press, London, UK, pp 1–398
7. Lefranc M-P, Duprat E, Kaas Q, Tranne M, Thiriot A, Lefranc G (2005) IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhcSF) G-LIKE-DOMAIN. Dev Comp Immunol 29:917–938
8. Lefranc M-P, Clément O, Kaas Q, Duprat E, Chastellan P, Coelho I, Combres K, Ginestoux C, Giudicelli V, Chaume D, Lefranc G (2005) IMGT-choreography for immunogenetics and immunoinformatics. In Silico Biol 5(1):45–60
9. Lefranc M-P (2011) IMGT, the International ImMunoGeneTics Information System. Cold Spring Harb Protoc 2011(6):595–603. https://doi.org/10.1101/pdb.top115
10. Lefranc M-P (2013) IMGT® information system. In: Dubitzky W, Wolkenhauer O, Cho K-H, Yokota H (eds) Encyclopedia of systems biology. Springer Science+Business Media, LLC, New York, pp 959–964
11. Lefranc M-P, Giudicelli V, Duroux P, Jabado-Michaloud J, Folch G, Aouinti S, Carillon E, Duvergey H, Houles A, Paysan-Lafosse T, Hadi-Saljoqi S, Sasorith S, Lefranc G, Kossida S (2015) IMGT®, the international ImMunoGeneTics information system® 25 years on. Nucleic Acids Res 43:D413–D422. . Epub 2014 Nov 5. https://doi.org/10.1093/nar/gku1056
12. Kaas Q, Ruiz M, Lefranc M-P (2004) IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. Nucleic Acids Res 32:D208–D210. https://doi.org/10.1093/nar/gkh042
13. Ehrenmann F, Kaas Q, Lefranc M-P (2010) IMGT/3Dstructure-DB and IMGT/DomainGapAlign: a database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhcSF. Nucleic Acids Res 38: D301–D307. https://doi.org/10.1093/nar/gkp946. Epub 2009 Nov 9
14. Ehrenmann F, Lefranc M-P (2011) IMGT/3Dstructure-DB: querying the IMGT database for 3D structures in immunology and Immunoinformatics (IG or antibodies, TR, MH, RPI, and FPIA). Cold Spring Harb Protoc 2011(6):750–761. https://doi.org/10.1101/pdb.prot5637
15. Lefranc M-P. (2007) MHC, what do we learn from IMGT colliers de Perles? ImmunoRIO2007, 13th international congress of immunology. Main symposia "Immunogenetics of MHC". A tribute to Jean Dausset. Rio de Janeiro, Brazil, 21 August 2007
16. Kaas Q, Lefranc M-P (2005) T cell receptor/peptide/MHC molecular characterization and standardized pMHC contact sites in IMGT/3Dstructure-DB. In Silico Biol 5:505–528
17. Kaas Q, Duprat E, Tourneur G, Lefranc M-P (2008) IMGT standardization for molecular characterization of the T cell receptor/peptide/MHC complexes. In: Schoenbach C, Ranganathan S, Brusic V (eds) Immunoinformatics. Immunomics reviews, series of springer science and business media LLC, Springer, New York, USA, pp 19–49
18. Giudicelli V, Lefranc M-P (1999) Ontology for Immunogenetics: IMGT-ONTOLOGY. Bioinformatics 15:1047–1054
19. Lefranc M-P, Giudicelli V, Ginestoux C, Bosc N, Folch G, Guiraudou D, Jabado-Michaloud J, Magris S, Scaviner D, Thouvenin V, Combres K, Girod D, Jeanjean S, Protat C, Yousfi Monod M, Duprat E, Kaas Q, Pommié C, Chaume D, Lefranc G (2004) IMGT-ONTOLOGY for Immunogenetics and Immunoinformatics, http://www.imgt.org. In Silico Biol 4(1):17-29
20. Lefranc M-P (2004) IMGT-ONTOLOGY and IMGT databases, tools and web resources for immunogenetics and immunoinformatics. Mol Immunol 40:647–660

21. Lefranc M-P, Giudicelli V, Regnier L, Duroux P (2008) IMGT, a system and an ontology that bridge biological and computational spheres in bioinformatics. Brief Bioinform 9(4):263–275. https://doi.org/10.1093/bib/bbn014. Epub 2008 Apr 19
22. Duroux P, Kaas Q, Brochet X, Lane J, Ginestoux C, Lefranc M-P, Giudicelli V (2008) IMGT-kaleidoscope, the formal IMGT-ONTOLOGY paradigm. Biochimie 90:570–583. https://doi.org/10.1016/j.biochi.2007.09.003. Epub 2007 Sep11
23. Giudicelli V, Lefranc M-P (2012) IMGT-ONTOLOGY 2012. Frontiers in bioinformatics and computational biology. Front Genet 3: 79. Epub 2012 May 23
24. Giudicelli V, Lefranc M-P (2013) IMGT-ONTOLOGY. In: Dubitzky W, Wolkenhauer O, Cho K-H, Yokota H (eds) Encyclopedia of systems biology. Springer Science+Business Media, LLC, New York, pp 964–972
25. Lefranc M-P (2011) From IMGT-ONTOLOGY CLASSIFICATION axiom to IMGT standardized gene and allele nomenclature: for immunoglobulins (IG) and T cell receptors (TR). Cold Spring Harb Protoc 6: 627–632. https://doi.org/10.1101/pdb.ip84
26. Lefranc M-P (2011) IMGT unique numbering for the Variable (V), Constant (C), and Groove (G) domains of IG, TR, MH, IgSF, and MhSF. Cold Spring Harb Protoc 2011(6):633–642. https://doi.org/10.1101/pdb.ip85
27. Lefranc M-P (2013) IMGT unique numbering. In: Dubitzky W, Wolkenhauer O, Cho K-H, Yokota H (eds) Encyclopedia of systems biology. Springer Science+Business Media, LLC, New York, pp 952–959
28. Lefranc M-P (2011) IMGT collier de Perles for the variable (V), constant (C), and groove (G) domains of IG, TR, MH, IgSF, and MhSF. Cold Spring Harb Protoc 6:643–651. https://doi.org/10.1101/pdb.ip86
29. Lefranc M-P (2013) IMGT Collier de Perles. In: Dubitzky W, Wolkenhauer O, Cho K-H, Yokota H (eds) Encyclopedia of systems biology. Springer Science+Business Media, LLC, New York, pp 944–952
30. Lefranc M-P (2011) (2011) from IMGT-ONTOLOGY DESCRIPTION axiom to IMGT standardized labels: for immunoglobulin (IG) and T cell receptor (TR) sequences and structures. Cold Spring Harb Protoc 6: 614–626. https://doi.org/10.1101/pdb.ip83
31. Lefranc M-P (2011) (2011) from IMGT-ONTOLOGY IDENTIFICATION axiom to IMGT standardized keywords: for immunoglobulins (IG), T cell receptors (TR), and conventional genes. Cold Spring Harb Protoc 6:604–613. https://doi.org/10.1101/pdb.ip82
32. Lefranc M-P (1997) Unique database numbering system for immunogenetic analysis. Immunol Today 18:509
33. Lefranc M-P (1999) The IMGT unique numbering for immunoglobulins, T cell receptors and Ig-like domains. Immunologist 7: 132–136
34. Lefranc M-P, Pommié C, Ruiz M, Giudicelli V, Foulquier E, Truong L, Thouvenin-Contet V, Lefranc G (2003) IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. Dev Comp Immunol 27:55–77
35. Lefranc M-P, Pommié C, Kaas Q, Duprat E, Bosc N, Guiraudou D, Jean C, Ruiz M, Da Piedade I, Rouard M, Foulquier E, Thouvenin V, Lefranc G (2005) IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains. Dev Comp Immunol 29: 185–203
36. Ruiz M, Lefranc M-P (2002) IMGT gene identification and colliers de Perles of human immunoglobulins with known 3D structures. Immunogenetics 53:857–883
37. Kaas Q, Lefranc M-P (2007) IMGT colliers de Perles: standardized sequence-structure representations of the IgSF and MhcSF superfamily domains. Curr Bioinforma 2:21–30
38. Kaas Q, Ehrenmann F, Lefranc M-P (2007) IG, TR and IgSf, MHC and MhcSF: what do we learn from the IMGT colliers de Perles? Brief Funct Genomic Proteomic 6(4): 253–264. https://doi.org/10.1093/bfgp/elm032. Epub 2008 Jan 21
39. Ehrenmann F, Giudicelli., V, Duroux, P., Lefranc, M.-P. (2011) IMGT/collier de Perles: IMGT standardized representation of domains (IG, TR, and IgSF variable and constant domains, MH and MhSF groove domains). Cold Spring Harb Protoc 2011(6):726–736
40. Ehrenmann F, Lefranc M-P (2011) IMGT/DomainGapAlign: IMGT standardized analysis of amino acid sequences of variable, constant, and groove domains (IG, TR, MH, IgSF, MhSF). Cold Spring Harb Protoc 2011(6): 737–749. https://doi.org/10.1101/pdb.prot5636
41. Ehrenmann F, Lefranc M-P (2012) IMGT/DomainGapAlign: the IMGT® tool for the analysis of IG, TR, MH, IgSF, and MhSF domain amino acid polymorphism. Methods

Mol Biol 882:605–633. https://doi.org/10.1007/978-1-61779-842-9_33

42. Duprat E, Kaas Q, Garelle V, Lefranc G, Lefranc M-P (2004) IMGT standardization for alleles and mutations of the V-LIKE-DOMAINs and C-LIKE-DOMAINs of the immunoglobulin superfamily. In: Recent Research Developments in Human Genetics, vol 2. Research Signpost, Trivandrum, Kerla, pp 111–136
43. Bernard D, Hansen JD, du Pasquier L, Lefranc M-P, Benmansour A, Boudinot P (2005) Costimulatory receptors in jawed vertebrates: conserved CD28, odd CTLA4 and multiple BTLAs. Dev Comp Immunol 31:255–271
44. Garapati VP, Lefranc M-P (2007) IMGT colliers de Perles and IgSF domain standardization for T cell costimulatory activatory (CD28, ICOS) and inhibitory (CTLA4, PDCD1 and BTLA) receptors. Dev Comp Immunol 31: 1050–1072
45. Frigoul A, Lefranc M-P (2005) MICA: standardized IMGT allele nomenclature, polymorphisms and diseases. In: Pandalai SG (ed) Recent research developments in human genetics, vol 3. Research Signpost, Trivandrum, Kerala, pp 95–145
46. Duprat E, Lefranc M-P, Gascuel O (2006) A simple method to predict protein binding from aligned sequences - application to MHC superfamily and beta2-microglobulin. Bioinformatics 22:453–459
47. Pommié C, Levadoux S, Sabatier R, Lefranc M-P (2004) IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION AA properties. J Mol Recognit 17:17–32
48. Lefranc M-P (2013) IMGT-ONTOLOGY, SpecificityType. In: Dubitzky W, Wolkenhauer O, Cho KH, Yokota H (eds) Encyclopedia of systems biology. Springer, New York, NY. https://doi.org/10.1007/978-1-4419-9863-7_677
49. Lefranc M-P (2013) Paratope. In: Dubitzky W, Wolkenhauer O, Cho KH, Yokota H (eds) Encyclopedia of systems biology. Springer, New York, NY. https://doi.org/10.1007/978-1-4419-9863-7_673
50. Lefranc M-P (2013) Epitope. In: Dubitzky W, Wolkenhauer O, Cho KH, Yokota H (eds) Encyclopedia of systems biology. Springer, New York, NY. https://doi.org/10.1007/978-1-4419-9863-7_663
51. Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, Wheeler DK, Sette A, Peters B (2019) The immune Epitope database (IEDB): 2018 update. Nucleic Acids Res 47: D339–D343. https://doi.org/10.1093/nar/gky1006
52. Mahajan S, Vita R, Shackelford D, Lane J, Schulten V, Zarebski L, Jespersen MC, Marcatili P, Nielsen M, Sette A, Peters B (2018) Epitope specific antibodies and T cell receptors in the immune Epitope database. Front Immunol 9:2688. https://doi.org/10.3389/fimmu.2018.02688
53. Li S, Lefranc M-P, Miles JJ, Alamyar E, Giudicelli V, Duroux P, Freeman JD, Corbin VDA, Scheerlinck J-P, Frohman MA, Cameron PU, Plebanski M, Loveland B, Burrows SR, Papenfuss AT, Gowans EJ (2013) IMGT/HighV-QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. Nat Commun 4: 2333. https://doi.org/10.1038/ncomms3333
54. Holland CJ, Crean RM, Pentier JM, de Wet B, Lloyd A, Srikannathasan V, Lissin N, Lloyd KA, Blicher TH, Conroy PJ, Hock M, Pengelly RJ, Spinner TE, Cameron B, Potter EA, Jeyanthan A, Molloy PE, Sami M, Aleksic M, Liddy N, Robinson RA, Harper S, Lepore M, Pudney CR, van der Kamp MW, Rizkallah PJ, Jakobsen BK, Vuidepot A, Cole DK (2020) Specificity of bispecific T cell receptors and antibodies targeting peptide-HLA. J Clin Invest 130(5):2673–2688. https://doi.org/10.1172/JCI130562
55. Lefranc M-P, Lefranc G. (2021) Antibody sequence and structure analyses using IMGT®: 30 years of immunoinformatics. Computer aided antibody design: methods and protocols.In: Tsumoto K, Kuroda D, eds. Methods Mol biol, in press
56. Lefranc M-P (2007) WHO-IUIS nomenclature subcommittee for immunoglobulins and T cell receptors report. Immunogenetics 59: 899–902
57. Lefranc M-P (2008) WHO-IUIS nomenclature subcommittee for immunoglobulins and T cell receptors report august 2007, 13th international congress of immunology, Rio de Janeiro, brazil. Dev Comp Immunol 32: 461–463
58. Lefranc M-P (2000) Nomenclature of the human immunoglobulin genes. In: Coligan JE, Bierer BE, Margulies DE, Shevach EM, Strober W (eds) Current protocols in immunology. John Wiley and Sons, Hoboken N.J, pp A.1P.1–A.1P.37
59. Lefranc M-P (2000) Nomenclature of the human T cell receptor genes. In: Coligan JE, Bierer BE, Margulies DE, Shevach EM, Strober W (eds) Current protocols in immunology.

John Wiley and Sons, Hoboken N.J, pp A.1O.1–A.1O.23
60. Giudicelli V, Chaume D, Lefranc M-P (2005) IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. Nucleic Acids Res 33: D256–D261
61. Maglott D, Ostell J, Pruitt KD, Tatusova T (2007) Entrez gene: gene-centered information at NCBI. Nucleic Acids Res 35:D26–D31
62. Rose PW, Beran B, Bi C, Bluhm WF, Dimitropoulos D, Goodsell DS, Prlic A, Quesada M, Quinn GB, Westbrook JD, Young J, Yukich B, Zardecki C, Berman HM, Bourne PE (2011) The RCSB protein data Bank: redesigned web site and web services. Nucleic Acids Res 39:D392–D401

Chapter 26

ARResT/Interrogate Immunoprofiling Platform: Concepts, Workflows, and Insights

Nikos Darzentas

Abstract

ARResT/Interrogate was built within the EuroClonality-NGS working group to meet the challenge of developing and applying assays for the high-throughput sequence-based profiling of immunoglobulin (IG) and T-cell receptor (TR) repertoires. We herein present basic concepts, outline the main workflow, delve into EuroClonality-NGS-specific aspects, and share insights from our experiences with the platform.

Key words Immunoglobulin, Antigen receptors, IG, TR, Bioinformatics, Pipeline, Sequence analysis

1 Introduction

Immunoglobulins (IG) and T-cell receptors (TR) are highly adaptive molecular receptors involved in antigen recognition and enormously variable immunological responses. The advent of sequence-based profiling of IG and TR repertoires has been instrumental for understanding such responses, both normal and pathologic, the latter encompassing a wide range of diseases with an underlying immune cause. This unprecedented capability has also brought along novel and unique challenges [1]; this chapter will cover the bioinformatic one, from the perspective of the ARResT/Interrogate immunoprofiling platform.

ARResT (abbreviation of Antigen Receptors Research Tool, http://bat.infspire.org/arrest) comprises a handful of tools developed over the years within focused groups. It originated in the days of Sanger sequence analysis toward delineating subsets of stereotyped antigen receptor sequences in chronic lymphocytic leukemia (CLL) [2, 3].

ARResT/Interrogate [http://arrest.tools/interrogate] was built from the grounds up within the EuroClonality-NGS working group [http://euroclonality.org] to initially support the development of the group's NGS assays and eventually to apply them in

Anton W. Langerak (ed.), *Immunogenetics: Methods and Protocols*, Methods in Molecular Biology, vol. 2453, https://doi.org/10.1007/978-1-0716-2115-8_26,

research and clinical applications [4, 5]. ARResT/Interrogate is able to: automatically paired-end-join and concatenate input files; use spreadsheet sample sheets to make data and metadata available to itself and the user; identify, tag, trim, and report on primer sequences (and primer dimers); annotate and identify all rearrangement types (or 'junction classes') of all IG/TR loci; offer powerful interactive tools to the user for mining results; identify, filter, and use the EuroClonality-NGS central in-tube quality/quantification control (cIT-QC, or spike-ins) for abundance normalization; generally support EuroClonality-NGS assays, also with bespoke analytical and visual functionalities; and provide detailed logs and feedback to the user.

ARResT/Interrogate will be continuously updated, and therefore bioinformatic and user interface details included herein may not stay the same over time. We advise readers/users to seek the latest information on the ARResT/Interrogate browser [http://arrest.tools/interrogate] and on the EuroClonality-NGS website [http://euroclonality.org]. For the same reason, we chose not to focus on application-specific methods, also because they are covered in other chapters in this book. Still, the general concepts and workflows included in this chapter should be considered safe, as should the Notes below from our years of experience both developing and using ARResT/Interrogate.

1.1 Design

ARResT/Interrogate consists of the pipeline and the browser (its user interface). The browser features four main "panels" for logically organized and ordered steps (Fig. 1):

1. Access to pipeline ("processing").
2. Access to pipeline results ("file").
3. Analysis of immunogenetic features ("questions").
4. Retrieval and analysis of sequences ("forensics") (*see* **Note 1**).

There is also the "HQ" panel that offers introductory text and specific notes and advice (in separate "tabs"). There are more panels to serve special applications, e.g., clonality assessment, but they are by default hidden and may be accessed by switching "user modes" with the widget on the top left (set at "Interrogate.simple" in Fig. 1).

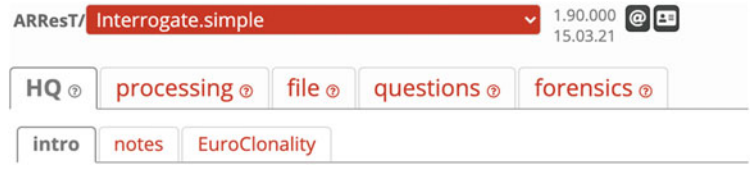


Fig. 1 ARResT/Interrogate browser (user interface), with "panels," "tabs," and the "user mode" selection widget

1.2 Primers

ARResT/Interrogate is able to identify, tag, trim, and report on primer sequences (and primer dimers), including making the results available for a fully interactive analysis. The trimming allows for less artificial sequence data to be processed more accurately and more efficiently, while the reporting allows for the primer-based results to be directly used for quality control and development. *See* also **Notes 2** and **3** about trimming.

1.3 Rearrangements and Junctions

ARResT/Interrogate is able to annotate and identify all rearrangement types of all IG/TR loci. We call these rearrangement types "junction classes." They include "complete," e.g., IG's VJ:Vh-(Dh)-Jh; "incomplete," e.g., TR's DJ:Db-Jb; and "other," e.g. IG's Vk-Kde or intron-Kde (Table 1). For junction classes with no biologically relevant junctional anchors (i.e., residues that define the CDR3 region, as per IMGT), we decided to introduce virtual ones—this enables consistent and informative results across all junction classes, assisting the user to focus on the most variable part of the rearrangement. For the D genes in DJ, VD, and DD incomplete junction classes, we use recombination signal sequence (RSS) heptamers: the last triplet of the heptamer in 5′ and the first triplet of the heptamer in 3′. For the intron RSS in the IGK locus, we use a CCC triplet between the EuroClonality-NGS primer and the RSS heptamer, while for Kde in the IGK locus the final triplet after the RSS heptamer and before the EuroClonality-NGS primer is used. In the majority of cases, these anchors are far enough from the junctional point to allow for nucleotide trimming without affecting their presence, but ARResT/Interrogate is anyway able to report rearrangements even with the anchors trimmed or mutated. This is also true for normal anchors in complete rearrangements.

Anchors overview:

5′ side of junction.

V genes: C aa = TG[CT] nt.

D genes: V aa = GT[any] nt, the last triplet of the 5′ heptamer.

intron: P aa = CCC nt, a triplet between primer and heptamer.

3′ side of junction.

J genes: W aa = TGG nt or F aa = TT[CT] nt.

D genes: H aa = CA[CT] nt, the first triplet of the 3′ heptamer.

Kde: R aa = CGA nt, final triplet after heptamer and before primer.

Table 1
Junction classes supported by ARResT/Interrogate and the EuroClonality-NGS amplicon and capture assays

Junction class	Receptor	Chain	Type	EuroClonality-NGS amplicon primer set	EuroClonality-NGS DNA capture (ECNDC)
VJ:Vh-(Dh)-Jh	IG	H	Complete	IGH-VJ	Yes
DJ:Dh-Jh	IG	H	Incomplete	IGH-DJ	Yes
VJ:Vk-Jk	IG	K	Complete	IGK-VJ-Kde (IGK)	Yes
Vk-Kde	IG	K	Other	IGK-VJ-Kde (IGK)	Yes
intron-Kde	IG	K	Other	intron-Kde (IGK)	Yes
VJ:Vl-Jl	IG	L	Complete		Yes
VJ:Va-Ja	TR	A	Complete	TRD	Yes
VJ:Va-(Dd)-Jd	TR	D	Complete	TRD	Yes
VJ:Vd-(Dd)-Ja	TR	D	Complete	TRD	Yes
DJ:Dd-Ja	TR	D	Incomplete	TRD	Yes
VD:Va-Dd3	TR	D	Incomplete		Yes
VJ:Vb-(Db)-Jb	TR	B	Complete	TRB-VJ	Yes
DJ:Db-Jb	TR	B	Incomplete	TRB-DJ	Yes
VD:Vb-Db	TR	B	Incomplete		Yes
DD:Db-Db	TR	B	Other		Yes
VJ:Vd-(Dd)-Jd	TR	D	Complete	TRD	Yes
DJ:Dd2-Jd	TR	D	Incomplete	TRD	Yes
VD:Vd-Dd3	TR	D	Incomplete	TRD	Yes
DJ:Dd3-Jd	TR	D	Incomplete		Yes
VD:Vd-Dd2	TR	D	Incomplete		Yes
DD:Dd2-Dd3	TR	D	Other	TRD	Yes
DD:Dd3-Dd2	TR	D	Other		Yes
VJ:Vg-Jg	TR	G	Complete	TRG	Yes

2 Materials

2.1 Sequence Input

1. All sequences are uploaded through the "processing" panel.
2. Sample sequences should be uploaded in FASTQ (preferably) or FASTA format, preferably also compressed in "gunzip" format (extension ".gz"). Also *see* **Note 4**.

3. Primer or tracer sequences should be uploaded in uncompressed FASTA format.
4. Filenames should not contain spaces or any special characters; underscores and hyphens are allowed (in fact, encouraged for clarity). This is—generally speaking—a good advice for any files to be used with bioinformatic tools.
5. ARResT/Interrogate can automatically recognize forward/reverse and multilane sequence files, the former being paired-end-joined and the latter concatenated. Since our code is based on Illumina, such files should contain "_L001_R1_" incremented accordingly. The user will be alerted if we believe there are issues with this logic, e.g., if we only have forward (R1) and not reverse (R2).
6. There are more checks on files, including for bad format, zero size, etc.; the user should watch out for relevant pipeline feedback.

2.2 Availability, Requirements, Contact

ARResT/Interrogate is currently available online at arrest.tools/interrogate (*see* **Note 5**); therefore, compute and storage requirements on the user side are limited. We nevertheless urge the use of a modern computer and web browser. In case of trouble using ARResT/Interrogate, please email contact@arrest.tools with as many details as possible, on what was done and if the issue persisted after a fresh start. Screenshots are invaluable, even if the browser has crashed and is grayed out.

2.3 Sample Sheet

1. We will mention sample sheets a number of times below. A sample sheet, a spreadsheet in Microsoft Excel format, is uniquely useful to provide the pipeline and the browser (and the user) with data and metadata that can run different pipeline options for different samples, provide cell counts (deduced from amount of DNA) for spike-in-based normalization, and help select/filter/order/rename/identify user's samples in the browser.
2. The ARResT/Interrogate sample sheet offers a number of predefined columns (i.e., ARResT/Interrogate expects these column names for the information to be used properly) and the possibility to add many others with flexible column names.
3. The most important predefined columns (again, do not change the column names or use them for other purposes) are:
   (a) Sample: required—unique for every sample and part or whole of the sample's sequence filenames.
   (b) Cells: number of cells, based on amount of DNA of, e.g., patient, to be used for quantification.

sample	cells	primer set	primers	rearrangements	scenario	tracers	patient	entity	timepoint
pt1_IGH-VJ-FR1_S1_L001_R1_001	15000	IGH-VJ-FR1	EC-NGS.set	all	EC-NGS.marker_identification		pt1	B-ALL	diagnosis
pt1_IGH-DJ_S2_L001_R1_001	15000	IGH-DJ	EC-NGS.set	all	EC-NGS.marker_identification		pt1	B-ALL	diagnosis
pt1_IGK-VJ-Kde_S3_L001_R1_001	15000	IGK-VJ-Kde	EC-NGS.set	all	EC-NGS.marker_identification		pt1	B-ALL	diagnosis
pt1_intron-Kde_S4_L001_R1_001	15000	intron-Kde	EC-NGS.set	all	EC-NGS.marker_identification		pt1	B-ALL	diagnosis
pt1_TRB-VJ_S5_L001_R1_001	15000	TRB-VJ	EC-NGS.set	all	EC-NGS.marker_identification		pt1	B-ALL	diagnosis
pt1_TRB-DJ_S6_L001_R1_001	15000	TRB-DJ	EC-NGS.set	all	EC-NGS.marker_identification		pt1	B-ALL	diagnosis
pt1_TRD_S7_L001_R1_001	15000	TRD	EC-NGS.set	all	EC-NGS.marker_identification		pt1	B-ALL	diagnosis
pt1_TRG_S8_L001_R1_001	15000	TRG	EC-NGS.set	all	EC-NGS.marker_identification		pt1	B-ALL	diagnosis
pt2_IGH-VJ-FR1_S9_L001_R1_001	100000	IGH-VJ-FR1	IGH-leader.set	IGH	routine.follow_up	marker.fasta	pt2	CLL	post-therapy

Fig. 2 Sample sheet example. Red are predefined columns, and blue are flexible columns

(c) Primer set: please use one of IGH-VJ, IGH-DJ, IGK-VJ-Kde, intron-Kde, TRB-VJ, TRB-DJ, TRD, and TRG.

(d) Primers: name(s) of file(s) of primers.

(e) Scenario: if one wants to run different pipeline scenarios for different samples.

(f) Rearrangements: rearrangement type(s) to be identified for each sample.

(g) Tracers: name(s) of file(s) of tracers (i.e., rearrangements of interest, including spike-ins or artifacts).

(h) Select: which samples should be analyzed, also in batches (i.e., could be 'x' or a batch number).

4. Check the example sample sheet in Fig. 2, in which red are predefined columns and blue are flexible columns.

3 Methods

3.1 A Basic Workflow

1. Visit http://arrest.tools/interrogate (*see* **Note 5**) and log in; this requires an account, which can be requested by emailing contact@arrest.tools.
2. Switch to the "processing" panel.
   (a) Create a new analysis or select an existing one, otherwise the "default" will be used, which is OK. Also *see* **Note 6**.
   (b) Upload sample sequences in compressed FASTQ/A format (*see* Subheading 2).
   (c) The default scenario ("ARResT.profile") should work fine in any case. One may select a different user mode or pipeline scenario, especially when deploying EuroClonality-NGS assays (*see* Subheading 3.2).
   (d) One may use own primer sequences by uploading them in uncompressed FASTA format and selecting them under "scenario options" (there are instructions on the user interface). In general, please study primers (e.g., *see* **Notes 3**, **7**, and **8** as to why). Also *see* **Note 9**.

(e) Click on the blue "test it" button when ready; if the test goes well (otherwise please follow the advice in the "process output" tab), click on the green "process" button to start the actual run.

(f) There is no need to wait, one may even close the browser; either log in later or, better, make sure to provide an email address to receive email notifications. Also *see* **Note 10**.

(g) If the run was not successful, the email notification's subject will include "(SOME SAMPLES) FAILED", pay attention to the pipeline feedback as to why, or email contact@arrest.tools.

3. Switch to the "file" panel.

(a) Select results in the drop-down widget, select filtering level (*see* **Note 11**), click "load results".

(b) One may browse the run and sample reports, paying attention to quality control (QC) information, alarms (and our hints and tips for possible causes and solutions), basic numbers like percentage of reads with junction (*see* **Note 7**) that are also color-coded to provide visual feedback. Alarms include:

- Low number of reads "5′ primed in R1" or "3′ primed in R2", indicating wrong or missing primers, noisy reads, i.e., compromised primer alignment, etc.
- Low number of reads "3′ primed in R1" or "5′ primed in R2", indicating long or trimmed amplicons (with FR1 or FR2 primers for example) not covered by the sequenced read length, or wrong or missing primers.
- High number of reads "short"—sequence artifacts are generally an explanation, and if primers were used with the pipeline, one may also see an alarm about primer dimers.

4. Switch to the "questions" panel.

(a) The main series of widgets are split into "select" on the left and "filter" on the right (Fig. 3).

(b) Note that if samples are "QC-failed" (*see* **Note 12**), they will not be available here by default; uncheck appropriate widget in "samples options" to include them back in.

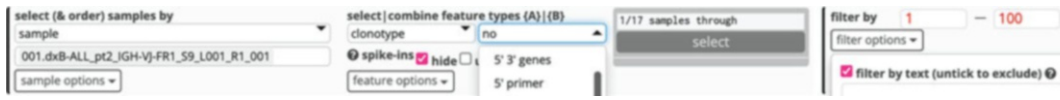


Fig. 3 "Select" and "filter" widgets on the "questions" panel

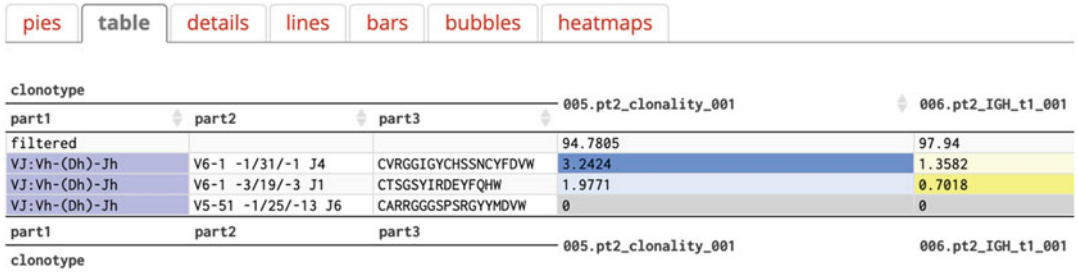


Fig. 4 "Table" visualization of the "questions" panel

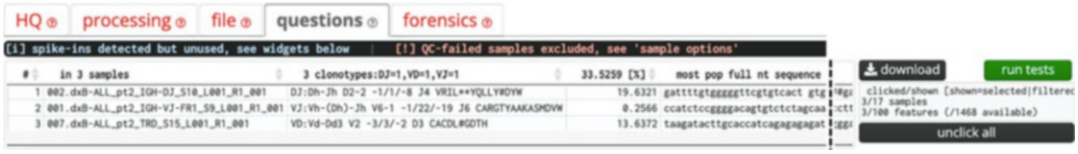


Fig. 5 The "minitable" for tabulation and downloading of selected features and their most popular sequences

(c) By default, "clonotypes" will be selected to be shown across samples. One may select one or two feature types among many.

(d) Make sure to click on "select" or "filter" after changing options, in which case these buttons are black (vs. gray).

(e) The panel provides access to multiple tabs for different visualizations, with an interactive table being the default (Fig. 4).

(f) One may click on specific clonotypes, in which case the corresponding information is shown in what is called a "minitable" near the top of the page; importantly, here also the most popular sequence of the clonotypes is provided (Fig. 5). Apart from being able to download this table for reporting, one may also "run tests" on the sequences (see below).

5. Switch to the "forensics" panel, after having clicked on at least one clonotype in "questions."

(a) One may retrieve and download all stored sequences of the clonotype in the "sequences" tab. The sequence variation that will undoubtedly appear in the retrieved sequences could be biological variability (e.g., somatic hypermutation) or technical noise including PCR or sequencing errors or amplification by different primers (*see* **Note 3**). Also, the retrieved sequences are not necessarily all possible sequences from the original sample, as we mainly avoid storing sequences supported by a single read unless they are the only representative of a combination of features.

(b) The "tests" tab (also accessible via "runs tests" in the "questions" panel) offers the possibility to annotate the sequences in different ways. When checking the "Interrogate" option, one will get more color-coded information than what is available in "questions," including D genes and more detailed segmentation. "AssignSubsets" provides access to ARResT/AssignSubsets for assignment of IGH rearrangements to major stereotyped subsets of chronic lymphocytic leukemia (CLL) [http://bat.infspire.org/arrest/assignsubsets] [3].

3.2 EuroClonality-NGS Assays

We will now provide more information on EuroClonality-NGS-specific aspects.

1. EuroClonality-NGS primer sets.
   (a) The EuroClonality-NGS amplicon assay uses eight tubes for the eight EuroClonality-NGS primer sets: IGH-VJ-FR [1–3], IGH-DJ, IGK-VJ-Kde, intron-Kde, TRB-VJ, TRB-DJ, TRD, and TRG.
   (b) It is useful for ARResT/Interrogate to know the primer set of the sample, and therefore, we try to auto-detect it, otherwise the sample is considered "pooled." If the sample is not pooled, these primer set names should be used bookended by _, e.g., sample1_IGH-VJ-FR1_[...] or sample2_IGK-VJ-Kde_[...]. If still detected wrongly, the name should be edited accordingly as to affect the process, either way. Another way is with a sample sheet and its "primer set" column (*see* Subheading 2.3).
   (c) Starting from version 1.90, ARResT/Interrogate specifically tags rearrangements that do not match the primer set (e.g., an VJ:Vg-Jg in an IGK tube) as contamination—one of the advantages of the EuroClonality-NGS assays using one primer set per tube.
2. EuroClonality-NGS central in-tube quality/quantification control (cIT-QC), or spike-ins.
   (a) If one uses spike-ins and wants to access normalized values (i.e., number of cells instead of number of reads), it is also necessary to provide the number of cells (derived from the DNA amount) in the sample, e.g., ~15,000 cells from 100 ng of DNA; this will be used as the denominator for the ratio calculation. There is a widget in the "processing" panel and the same in the "questions" panel (Fig. 6), which sets the value for all samples; if different values need to be set for different samples, this needs to be done with a sample sheet and its "cells" column (*see* Subheading 2.3). Do not include spike-in cells in those numbers.

Fig. 6 Messages and widgets related to cIT-QC (spike-ins)

(b) One should be able to see extra relevant widgets and messages in "questions" (and remember to hover over the "?" tooltip anchors); to see normalized abundances, check the "use" box (Fig. 6).

3. Pipeline scenarios and browser functionalities for EuroClonality-NGS assays.

   (a) It is important to select the appropriate user modes to properly analyze data from EuroClonality-NGS assays. One of the automations is the preset of appropriate pipeline scenarios in the "processing" panel with the aforementioned primers and spike-ins.

   (b) Switch to the "Interrogate.EC-NGS marker identification" user mode for the assays described in [6]. These assays involve one primer set per tube, plus spike-ins in each tube.

   (c) Switch to the "Interrogate.EC-NGS clonality assessment" user mode for the assays described in [7]. These assays pool the primer set tubes after PCR but before sequencing; therefore, ARResT/Interrogate needs to computationally separate them before calculating abundances. There are currently no spike-ins included. This user mode also enables a bespoke panel, "reporting," in which ARResT/Interrogate separates the different primer sets from the pooled data sample creating one view for each—see the VJ:Vh-(Dh)-Jh and DJ:Dh-Jh views (the latter partially and with a faint red background because of the low number of reads—121, in dark red background—included in it) in Fig. 7.

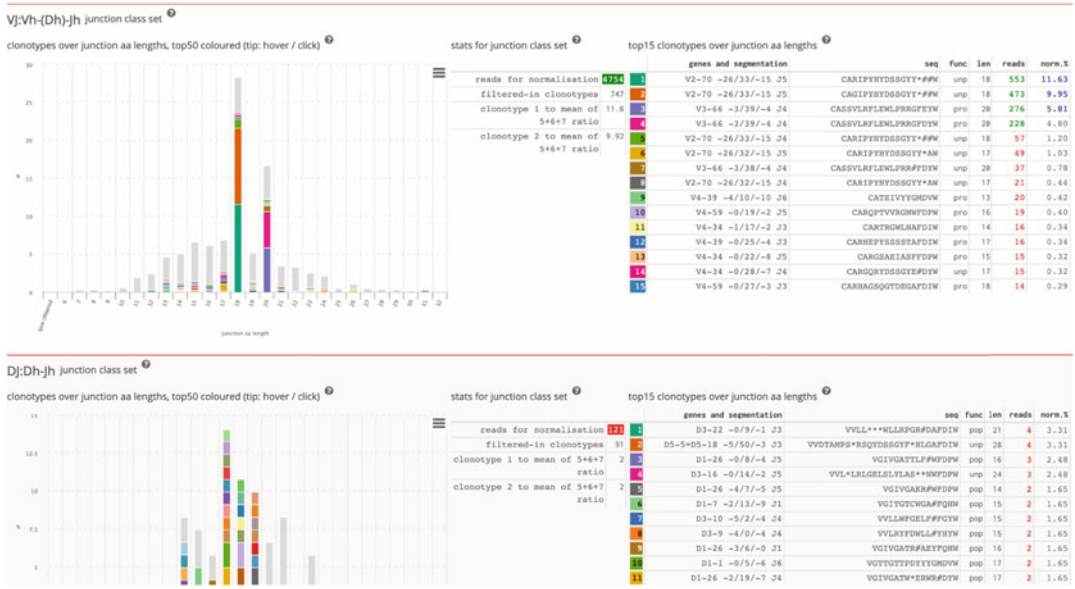


Fig. 7 Views from the bespoke "reporting" panel of the "Interrogate.EC-NGS clonality assessment" user mode, with two of the pooled primer sets separated, normalized, and presented to the user

4 Notes

1. We strongly advise to spend some time clicking around and hovering over the tooltip markers ("?"), especially after switching to the "Interrogate.advanced" user mode to enable more widgets and options; one should at least be generally aware of what is possible.
2. Although the primer is artificial and may compromise downstream analyses, it sometimes is necessary to keep it on in order to have enough sequence to annotate and thus identify a rearrangement (currently, for EuroClonality-NGS primers, IGH-D, IGK-DE, TRG-J primers are kept on the sequence).
3. Depending on the primer sequences used, one may face potentially confounding situations, especially since ARResT/Interrogate by default trims away the primer sequences.

   Amplification by different primers annealing on the same template may result in slightly different sequences of, e.g., the same clonotype—keep that in mind when looking at combinations of primers and clonotypes, or retrieve sequences of such a clonotype.

   Amplification by different primers annealing on the same template that result in the same sequence and length means that to fully study primers one needs to disable primer trimming so that the sequences remain separate; otherwise, only one primer is remembered per unique sequence that

might not represent the full picture. To do this, enable the "primer_ext" pipeline option, or use the "ARResT.profile.primer_ext" scenario as a template, or email contact@arrest.tools.

4. If the data are too big and/or one is facing upload issues, please email contact@arrest.tools to ask for access to our FTP service (it is planned to make it available by default via the ARResT/Interrogate user interface).
5. If the server (or "station" as we call them) is busy or too slow, please revisit arrest.tools/interrogate to be redirected to a different station—please do not bookmark any final link with specific station numbers.
6. To best analyze and report on markers across diagnostic and follow-up samples (the latter usually coming from a separate, later NGS run), eventually upload all files in the same "analysis" and process together with the same ARResT/Interrogate version.
7. When facing a low percentage of reads with a junction, check the "postmortem" section of the sample report. The first example in Fig. 8 (and below) without a junction has 1181 high-quality and 12 low-quality forward reads and was

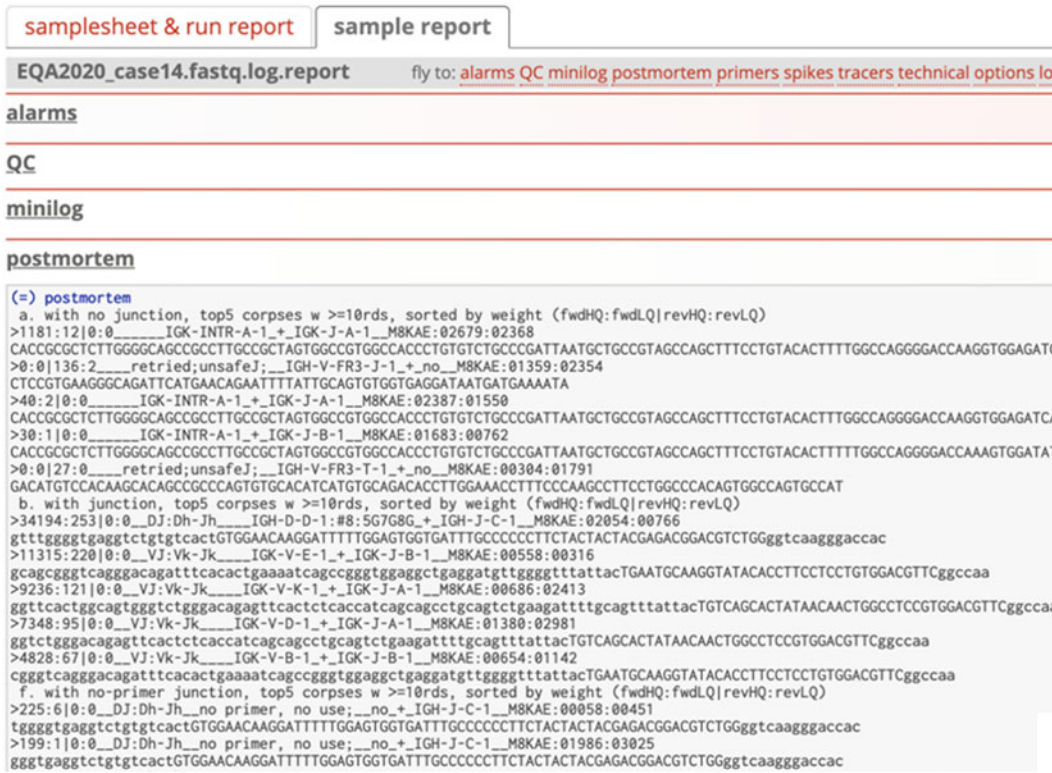


Fig. 8 View of the sample report with the "postmortem" section expanded—most abundant examples of sequences with and without junction are shown

bookended by the IGK-INTR-A-1 and IGK-J-A-1 EuroClonality-NGS primers, which actually do not make sense as a pair (IGK intron and IGK J). The second example has reverse reads; it had to go through a more sensitive workflow ("retried") and ended up with an unsafe IGHJ gene assignment ("unsafeJ"), and only had the 5′ IGHV primer on the sequence.

(a) *With no junction, top5 corpses w >=10rds, sorted by weight (fwdHQ:fwdLQ|revHQ:revLQ).*
>1181:12|0:0_____IGK-INTR-A-1_+_IGK-J-A-1__M8KAE:02679:02368.
CACCGCGCTCTTGGGGCAGCCGCCTTGCCGCTAGTGGCCGTGG[...]
>0:0|136:2____retried;unsafeJ;__IGH-V-FR3-J-1_+_no__M8KAE:01359:02354.
CTCCGTGAAGGGCAGATTCATGAACAGAATTTTATTGCAGTGTG[...]

8. Primers are useful to safeguard amplicon completeness and thus junction safety. Demand that both primers are present on the sequence of interest if, for example, the lab work is known to produce incomplete amplicons.
9. Do not mix IGH-VJ-FR* primer sets in pipeline options, e.g., FR3 primers will wrongly trim FR1/FR2 reads.
10. Running times may vary heavily, depending on sample number, depth and clonality, read length, and sequence noise.
11. Regarding "results filtering" in the "file" panel, keep in mind to switch to "not pre-filtered" when looking for very low abundance clonotypes.
12. As a "fail" "QC status" in the run report does not necessarily mean that the sample is unusable, one may reinsert it back into the analysis in the "questions" panel. Final decision is with the user, based on context (kind of sample, DNA quality, and purpose); QC status is just meant to attract the user's attention to potential issues.

References

1. Langerak AW, Brüggemann M, Davi F, Darzentas N, van Dongen JJM, Gonzalez D et al (2017) High-throughput Immunogenetics for clinical and research applications in immunohematology: potential and challenges. J Immunol 198:3765–3774
2. Darzentas N, Hadzidimitriou A, Murray F, Hatzi K, Josefsson P, Laoutaris N et al (2010) A different ontogenesis for chronic lymphocytic leukemia cases carrying stereotyped antigen receptors: molecular and computational evidence. Leukemia 24:125–132
3. Bystry V, Agathangelidis A, Bikos V, Sutton LA, Baliakas P, Hadzidimitriou A et al (2015) ARResT/AssignSubsets: a novel application for robust subclassification of chronic lymphocytic leukemia based on B cell receptor IG stereotypy. Bioinformatics 31:3844–3846
4. Bystry V, Reigl T, Krejci A, Demko M, Hanakova B, Grioni A et al (2017) ARResT/

interrogate: an interactive immunoprofiler for IG/TR NGS data. Bioinformatics 33:435–437

5. Knecht H, Reigl T, Kotrová M, Appelt F, Stewart P, Bystry V et al (2019) Quality control and quantification in IG/TR next-generation sequencing marker identification: protocols and bioinformatic functionalities by EuroClonality-NGS. Leukemia 33:2254–2265
6. Brüggemann M, Kotrová M, Knecht H, Bartram J, Boudjogrha M, Bystry V et al (2019) Standardized next-generation sequencing of immunoglobulin and T-cell receptor gene recombinations for MRD marker identification in acute lymphoblastic leukaemia; a EuroClonality-NGS validation study. Leukemia 33:2241–2253
7. Scheijen B, RWJ M, Rijntjes J, van der Klift MY, Möbs M, Steinhilber J et al (2019) Next-generation sequencing of immunoglobulin gene rearrangements for clonality assessment: a technical feasibility study by EuroClonality-NGS. Leukemia 33:2227–2240

Chapter 27

Purpose-Built Immunoinformatics for BcR IG/TR Repertoire Data Analysis

Chrysi Galigalidou, Laura Zaragoza-Infante, Anastasia Chatzidimitriou, Kostas Stamatopoulos, Fotis Psomopoulos, and Andreas Agathangelidis

Abstract

The study of antigen receptor gene repertoires using next-generation sequencing (NGS) technologies has disclosed an unprecedented depth of complexity, requiring novel computational and analytical solutions. Several bioinformatics workflows have been developed to this end, including the T-cell receptor/immunoglobulin profiler (TRIP), a web application implemented in R shiny, specifically designed for the purposes of comprehensive repertoire analysis, which is the focus of this chapter. TRIP has the potential to perform robust immunoprofiling analysis through the extraction and processing of the IMGT/HighV-Quest output, via a series of functions, ensuring the analysis of high-quality, biologically relevant data through a multilevel process of data filtering. Subsequently, it provides in-depth analysis of antigen receptor gene rearrangements, including (a) clonality assessment; (b) extraction of variable (V), diversity (D), and joining (J) gene repertoires; (c) CDR3 characterization at both the nucleotide and amino acid level; and (d) somatic hypermutation analysis, in the case of immunoglobulin gene rearrangements. Relevant to mention, TRIP enables a high level of customization through the integration of various options in key aspects of the analysis, such as clonotype definition and computation, hence allowing for flexibility without compromising on accuracy.

Key words Antigen receptor, B-cell receptor, Immunoglobulin, T-cell receptor, Immunoinformatics, Clonality, Immune repertoire, Somatic hypermutation

1 Introduction

Profiling the B-cell receptor immunoglobulin (BcR IG) and T-cell receptor (TR) gene repertoires using next-generation sequencing (NGS) technologies advanced our understanding of various clinical conditions and biological processes, extending from infections, vaccination, autoimmunity, to malignancy. NGS immunogenetics has applications in both diagnostics (e.g., assessment of clonality in

Chrysi Galigalidou and Laura Zaragoza-Infante are equal first authors.

Fotis Psomopoulos and Andreas Agathangelidis are equal senior authors.

Anton W. Langerak (ed.), *Immunogenetics: Methods and Protocols*, Methods in Molecular Biology, vol. 2453, https://doi.org/10.1007/978-1-0716-2115-8_27, © The Author(s) 2022

samples investigated for a possible lymphoproliferation or detection of minimal residual disease in patients with lymphoid malignancies) and research [1, 2]. To date, several pipelines that perform both BcR IG/TR sequence annotation and meta-data analysis have been made publicly available [3–6]: in that regard, notable examples include the "IMGT/StatClonotype" tool [7, 8], the MiXCR software [9], the Vidjil platform [10], and the ARReST|Interrogate application [11], among others.

Our contribution in this field concerns the T-cell receptor/immunoglobulin profiler (TRIP) software [12], which was designed in order to enable the comprehensive characterization of BcR IG and TR gene repertoires based on an integrated, robust, and user-friendly interface. TRIP has been utilized in projects on hematological malignancies, such as chronic lymphocytic leukemia (CLL) and multiple myeloma (MM) [13–16], as well as other contexts, e.g., infections [17, 18], providing valuable insight into the selection forces that shape the architecture of the respective immune repertoires.

This chapter will focus on the features of TRIP, particularly aiming to highlight how the functionalities offered by this software address the challenges of repertoire analysis in both diagnostic and, particularly, research settings.

2 Data Processing

2.1 Preprocessing of the Raw Data

The raw sequencing data is transferred from the sequencer server to the dedicated workspace.

The initial processing actions depend on the selected sequencing strategy; e.g., there might be need for some steps, such as demultiplexing, adapter masking, and format conversion to the FASTQ data type.

2.1.1 Demultiplexing

Demultiplexing concerns the separation of sequencing reads to the respective samples in cases of multiplexed sequencing, i.e., the simultaneous sequencing of multiple samples in a single run. During sample preparation, unique index sequences are attached to the sequences of each individual sample that will be used as identifiers during the demultiplexing process.

2.1.2 Adaptor Masking

Adapter sequence masking leads to the identification of adapter sequences and removes them from consideration in the downstream analysis steps. This process is essential in order to avoid artificial mismatches and alignment issues during sequence annotation.

2.1.3 Format Conversion

Relevant to mention, the output file(s) of this process are of the FASTQ type, since this is the format required for the downstream steps. The FASTQ file format is very commonly used in bioinformatics in order to process raw sequencing data, since it contains information regarding the sequence reads and their quality. FASTQ files contain four lines of information for each individual read:

1. The first line begins with the "@" symbol followed by a read identifier, which is given during the sequencing process.
2. The second line contains the nucleotide sequence of the read.
3. The third line has a "+" symbol, used as a line separator.
4. The fourth line has information about the quality of each base of the sequence, represented as Phred quality score. The value of these quality scores can be retrieved from ASCII charts.

Additional information about FASTQ files is provided at the following link https://www.ebi.ac.uk/ega/submission/sequence#fastq_format.

In the case of sequencing on an Illumina platform, the bcl2fastq2 software is the one most commonly used for demultiplexing sequencing data and for the masking of the adaptor sequences and/or UMIs (unique molecular identifiers), if present. Moreover, bcl2fastq2 transforms base call files (BCL), which is the default format of raw data when obtained from the Illumina sequencer platform, into FASTQ files. Some sequencing platforms, such as the MiniSeq or MiSeq, provide the option to automatically transform BCL files to the FASTQ format.

In case another sequencing platform is used, it is necessary to follow the instructions specified for each scenario, check the data format of the sequencer server output files, and transform them to FASTQ.

2.2 Filtering of the Raw Data

As a first step, quality filters should be applied to all reads in the FASTQ file(s) in order to ensure that only high-quality data will be subjected to further analysis. A set of filtering parameters can be selected according to the type of data and the design of the experiment. The reads that do not fulfill all the requirements will be filtered out. The most common parameters are related to the read length, the quality score of each individual nucleotide, and the overall quality score of each read.

The level of strictness of the parameters is chosen according to the overall quality of the NGS run and the minimum quality threshold that would allow the extraction of biologically meaningful results depending on the project design.

Indicative examples of parameters for the analysis of BcR IG/TR data include: minimum length of the raw reads, 150 nucleotides; quality threshold for each nucleotide, 14; accepted minimum mean sequence quality for each read, 20; maximum percentage of

low-quality nucleotides, 0.2 (20%); and minimum percentage of accepted unidentified nucleotides, 0.01 (1%).

2.3 Synthesis of Paired-End Reads

Given the extreme intrinsic variability of BcR IG/TR rearrangement sequences, paired-end sequencing protocols are usually applied. In this scenario, two individual reads, namely, R1 and R2, are obtained from each sequence ensuring the high quality of the sequences and the accuracy of the immunogenetic annotation.

1. After checking the quality of each individual read (*see* Subheading 2.2), perform the synthesis of full-length reads by merging the individual R1 and R2 reads corresponding to each sequence, through the identification of an overlapping region.
2. Apply quality filters to the synthesized, full-length reads. Examples of these filters are: minimum length of the overlap between R1 and R2 reads, 20 nucleotides; mismatch ratio of the overlapping area, 0.25 (25%); threshold for the continuous match of the overlapping area, 10 nucleotides; quality thresholds for the classification of individual nucleotides either as of "bad quality" (and be replaced by "N") or of "high quality", 14 and 35, respectively; quality mean score of the synthesized reads, 25; minimum length of the synthesized reads, 280 nucleotides; percentage of nucleotides that can have low quality in the synthesized reads, 0.15 (15%); quality threshold of individual nucleotides in the synthesized reads, 20; percentage of "bad quality" nucleotides in the synthesized reads excluding the CDR3, 0.005 (0.5%); and CDR3 quality threshold, 25. This set of filtering criteria was designed in order to take into account the intrinsic properties of the BcR IG/TR rearrangement sequences, i.e., the extreme variability of the CDR3.
3. Compare the number of synthesized reads that have passed all filters with the number of raw reads, and check the percentage of reads that is discarded due to each filter. If the percentage of the synthesized, high quality reads is low (e.g., below 50–60%), a revision of relevant filter(s) should be considered.

The final synthesized reads that have successfully passed all filters from each sample are deposited in a FASTA file. This file consists of two lines of information per sequence: the first line begins with a ">" symbol followed by the read identifier, and the second line contains the nucleotide sequence.

3 Sequence Annotation with IMGT/HighV-QUEST

IMGT (the international ImMunoGeneTics information system) is the worldwide reference in immunogenetics and

immunoinformatics [19]. IMGT has incorporated the most extensive and updated reference datasets for human BcR IG/TR genes. IMGT/HighV-QUEST is the web portal for BcR IG/TR data analysis from NGS high-throughput and deep sequencing [20].

In the IMGT/HighV-QUEST home page (http://www.imgt.org/HighV-QUEST/home.action), the user can customize the analysis through a series of options, including a job title, the species, the antigen receptor type (BcR IG or TR), and the specific locus (for instance, BcR IGH or IGL). The data has to be uploaded in FASTA format, and the submission limit is 1,000,000 sequences. Once the analysis is finished, the results can be downloaded from the "Analysis history" tab.

The output for each sample is a folder with ten files in text (.txt) format, with each of them containing different types of immunogenetic information. More specifically, the output files are the following: "1_Summary.txt," containing a summary table of basic immunogenetic information, such as the rearranged V(D)J genes, the % of identity with the germline, the presence of indels etc.; "2_IMGT-gapped-nt-sequences.txt"; "3_Nt-sequences.txt"; 4_IMGT_gapped_AA_sequences.txt"; "5_AA-sequences.txt; "6_Junction.txt"; "7_V-REGION-mutation-and-AA-change-table.txt"; "8_V-REGION-nt-mutation-statistics.txt"; "9_V-REGION-AA-change-statistics.txt"; "10_V-REGION-mutation-hotspots.txt"; "11_Parameters", with the set of parameters applied in the analysis; and "README.txt", with technical information about the analysis.

4 IMGT/HighV-QUEST Meta-Data Analysis with TRIP

The T-cell receptor/immunoglobulin profiler (TRIP) tool [12] is a web application that provides an in-depth meta-data analysis based on the processing of the IMGT/HighV-QUEST output files, through a number of interoperable modules. The TRIP tool can be downloaded from the following link: https://bio.tools/TRIP_-_T-cell_Receptor_Immunoglobulin_Profiler.

1. Since IMGT/HighV-QUEST has a submission threshold of 1,000,000 sequences, if a sample contains a larger number of sequences, the user must split them into different batches of sequences before analyzing them with IMGT/HighV-QUEST. Thus, multiple output folders will be generated by the tool for the same sample. In this case, the folders should be named using the same identifier with a different extension, following a numerical order starting from 0, i.e., "-0", "-1", "-2", etc. With this approach, TRIP can trace the origin of these files to the same sample and will combine the respective data for the analysis.

2. The first step of the analysis with TRIP concerns the selection of the directory containing the IMGT/HighV-QUEST output data. At this step, it is also possible to restore previous sessions.
3. The next step concerns data selection. TRIP allows for the simultaneous analysis of several datasets. In that case, the analysis is performed both individually for each dataset and for all datasets together ("All Data" output files). Moreover, if more than one dataset is selected, there will be additional available steps in the downstream analysis (such as the shared clonotype computation or the repertoire comparison, see below).
4. The relevant output files from IMGT/HighV-QUEST are selected, depending on the type of downstream analysis. For most types of analysis, the necessary files are "1_Summary.txt," "2_IMGT-gapped-nt-sequences.txt," "4_IMGT-gapped-AA-sequences.txt," and "6_Junction.txt."

   After loading the files (option "Load Data"), TRIP scans the data and gives a notification in the case of data headers with a different or an unknown value. In that case, data headers should be replaced with the appropriate ones.
5. The antigen receptor type is selected: BcR IG or TR.
6. The type of data to be analyzed is selected: high throughput (NGS data) or low throughout (Sanger sequencing data). Henceforth, this chapter will focus on the analysis of high-throughput data.

A summary of all aforementioned analytical steps is depicted in Fig. 1.

5 High-Throughput Data Analysis with TRIP

5.1 Preselection (Data Curation)

All the preselection filters should be applied:

1. The *Only take into account functional V-gene* filter ensures the exclusive analysis of sequences utilizing a functional V gene; sequences with pseudogenes (P) or open reading frame (ORF) genes will be excluded from downstream analysis.
2. The *Only take into account CDR3 with no special characters* filter removes from the analytical process sequences with characters others than those of the 20 amino acids.
3. The *Only take into account productive sequences* filter limits the analysis to productive BcR IG/TR rearrangement sequences; sequences with stop codons or frameshifts will be filtered out.
4. The filter entitled *Only take into account CDR3 with valid start/end landmarks* ensures that only sequences with well-annotated CDR3 will be subjected to further analysis. The CDR3 is delimited by a cysteine at IMGT position

Fig. 1 A summary of all major steps in the analytical workflow starting from the NGS BcR IG/TR raw data up to the extraction of biologically meaningful results

104 (second-CYS 104) and a tryptophan or a phenylalanine (for BcR IG and TR sequences, respectively) at IMGT position 118 (i.e., J-PHE or J-TRP 118). If necessary, it is possible to add more than one landmark in the analysis by separating them with the "|" symbol.

Filters are applied consecutively and, as soon as one of the criteria is not passed, the sequence is filtered out; it is important to keep in mind that only the first non-passed criterion is reported. Only sequences that were of high quality according to all aforementioned standards will be further analyzed.

The results of the preselection process are summarized in four different tables and can be found at the "Pre-selection" tab:

1. A summary of the included and excluded sequences based on the application of each criterion ("Summary").
2. The entire set of data ("All data table").
3. The set of filtered-in sequences that will be analysed in the following steps ("Clean table").
4. The set of filtered-out sequences ("Clean out table").

The last data column in the "Clean out" table indicates the criterion that was not passed for each individual sequence. All tables can be downloaded in text (.txt) format.

5.2 Selection (Filtering)

Sequences, meeting all the pre-selection criteria, are further filtered during the Selection step.

The range of the V-region identity % should be selected. Sequences with a V region identity % that does not fall into the selected range are excluded from the analysis. The selection % of identity depends largely on:

1. The type of antigen receptor gene sequence data, e.g., the SHM mechanism is operational exclusively in B cells.
2. The expected error rate induced by the amplification protocol or the sequencing process.

In more detail, in the case of BcR IG sequences, a typical range of the V-region identity would be 85–100%, whereas in TR, the range would be narrower (95–100%).

The rest of the available filters enable the selection of sequences with specific immunogenetic features, namely, V, D, and J genes, CDR3 length and the presence of particular CDR3 amino acid sequence motifs. These filters allow for a high level of customization of the analytical procedure.

Again, four output files are produced, which are located at the "Selection" tab:

1. A summary table with all filtered-in and -out sequences for each individual parameter ("Summary).
2. The entire set of sequences that passed all the preselection criteria ("All Data table").
3. All sequences that passed through the Selection filters ("Filter in table").
4. The sequences that did not meet the selection criteria and were, thus, excluded from further analysis ("Filter out table").

The last column of the "Filter out" table indicates the criterion that was relevant for the exclusion of each individual sequence. These tables can be downloaded in text (.txt) format.

The Pre-selection and Selection steps were developed in order to ensure that only relevant, high-quality BcR IG/TR sequence data will be included in the downstream Analytical Pipeline of TRIP.

5.3 TRIP Analytical Pipeline

Once the NGS data has been curated and filtered, it is subjected to the TRIP Analytical Pipeline (located at the "Pipeline" tab). The workflow of the analysis can be customized according to the biological context of the project.

5.3.1 Clonotype Computation

The first step of the pipeline refers to the clonotype computation. It concerns the grouping of the analyzed sequences in clonotypes, based on a set of shared immunogenetic properties.

The clustering process depends on the definition of the clonotype. TRIP provides ten different options for defining the clonotype, in order to facilitate the selection of the most relevant immunogenetic properties. If, for example, "IGHV gene and CDR3 aa sequence" is chosen as definition, all the reads expressing the same IGHV gene and identical CDR3 at the aa sequence level will be grouped together into a single clonotype.

There is also the option "Load clonotypes," which allows to directly upload precomputed clonotypes from analyzed datasets.

The output is located at the tab "Clonotypes" and can be downloaded in text (.txt) format. The output contains a series of information regarding each individual clonotype:

1. Utilized V gene and the CDR3 amino acid sequence.
2. Absolute number of clustered sequences ("N").
3. Relative frequency.
4. Analysis of convergent evolution referring to the number of different nucleotide sequences that encode for the CDR3 aa sequence of each given clonotype.

Each clonotype is also a link leading to a table with the immunogenetic information of all the assigned reads. At this step, each clonotype is given a unique cluster identifier (cluster ID).

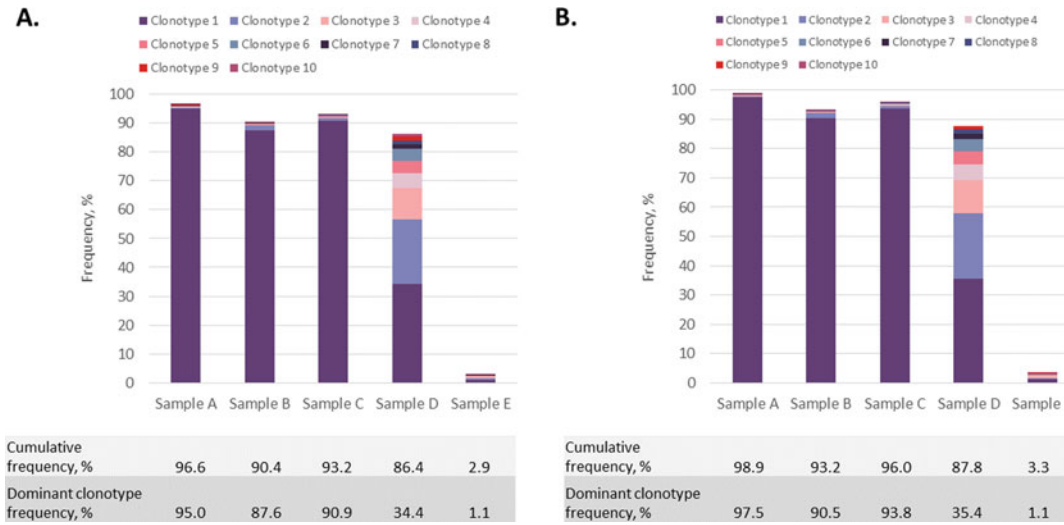
Clonotype computation can provide important biological information mostly in regard to the BcR IG/TR clonality levels in a given setting. Some examples of different approaches supported by TRIP are the following:

1. Frequency of the most expanded, dominant clonotype (the clonotype with the highest frequency).
2. Average cumulative frequency of the "top 10" clonotypes (the ten clonotypes with the highest frequency).
3. Average frequency of the abundant clonotypes (those with a frequency above a specific frequency threshold; this threshold may vary according to the aims of the project).

An example of clonality assessment using the top 10 clonotypes is illustrated in Fig. 2a.

5.3.2 Computation of Highly Similar Clonotypes

Following the previous approach on clonotype definition, namely "V gene and CDR3 aa sequence," two or more clonotypes would be considered as highly similar, if displaying the same CDR3 amino acid length and a low number of amino acid mismatches. TRIP allows for the grouping of highly similar clonotypes obtained at the "Clonotypes computation" step (Subheading 5.3.1). The number



Cumulative frequency, %	96.6	90.4	93.2	86.4	2.9
Dominant clonotype frequency, %	95.0	87.6	90.9	34.4	1.1

Cumulative frequency, %	98.9	93.2	96.0	87.8	3.3
Dominant clonotype frequency, %	97.5	90.5	93.8	35.4	1.1

Fig. 2 Clonality assessment through the analysis of the top 100 clonotypes for five samples, using the either the "Clonotype computation" (**a**) or the "Highly similar clonotypes computation" option (**b**). The first three samples (namely Samples A, B, and C) display a monoclonal profile, characterized by predominance of a single clonotype with a frequency of 95%, 87.6%, and 90.9%, respectively. Sample D is oligoclonal, with multiple clonotypes exhibiting high frequency; the dominant clonotype accounts for 34.4% of the repertoire, whereas the cumulative frequency of the top 10 clonotypes is 86.4%. Finally, Sample E is polyclonal with the top 10clonotypes accounting for a very small fraction of the repertoire (2.9%). The option of merging together the "Highly similar clonotypes" resulted in an increase in the cumulative frequency of the top 10clonotypes in all samples (range 0.4–2.8%) indicating the presence of minor clonotypes exhibiting strong immunogenetic relations with the top 10 clonotypes

of allowed CDR3 aa mismatches can be either chosen manually for each individual CDR3 length or through the application of a percentage (%) threshold.

One of the most typical approaches is based on the CDR3 length and allows for a low number of aa mismatches, thus ensuring a strong connection between highly similar clonotypes:

1. One aa mismatch for BcR IG/TR sequences with CDR3 lengths of up to 13 aa.
2. Two aa mismatches for BcR IG/TR sequences with CDR3 lengths between 14 and 24 aa.
3. Three aa mismatches for BcR IG/TR sequences with CDR3 lengths of 25 or more aa.

This process is implemented by considering the most frequent clonotype for each given CDR3 length as the reference for all the remaining clonotypes with the same CDR3 length. After merging the highly similar clonotypes, their relative frequencies are calculated accordingly.

Another parameter given by TRIP for the computation of highly similar clonotypes concerns the rearranged V gene. The

application of this parameter enables the consideration of the whole variable domain of the BcR IG/TR into the clonotype grouping process, yet depends on the context of the given project.

The output files from this process are given as text (.txt) files and contain the following information:

1. Cluster identifiers of the merged clonotypes (consisting of the V gene and amino acid CDR3 sequence).
2. Number of sequences belonging to each merged clonotype.
3. Relative frequency of each merged clonotype.
4. Cluster identifiers of the clonotypes computed at the previous step ("Clonotype computation"), which formed each merged clonotype.
5. Detailed information about the clonotype merging process for each individual CDR3 length.

The output files from this step can be found under the tab entitled "Highly Similar Clonotypes."

The effect of the grouping of highly similar clonotypes on clonality assessment is given in Fig. 2b.

5.3.3 Repertoire Extraction

The next step of the analysis enables the extraction of the V, D, and J repertoires either at the gene or at the gene allele level. The V, D, or J gene repertoires are extracted from the output file of the previous step (Subheading 5.3.2) that includes all the clonotypes of the dataset (Fig. 3). Here, it is important to keep in mind that the relative frequency of each V, D, or J gene is calculated at the clonotype level rather than at the sequence level. The output of this process is provided as a text (.txt) file and contains information on the gene names, and the absolute number and relative frequency of clonotypes utilizing each specific V, D, and J gene.

At the end of this section, TRIP allows the user to choose whether the repertoire extraction will be based on the computation before or after the grouping of highly similar clonotypes. The output of this part of the pipeline can be found under the tab "Repertoires."

5.3.4 CDR3 Length Distribution

The distribution of the CDR3 length is calculated based on the number of clonotypes corresponding to each individual length. In case the user would like to perform the analysis after the grouping of highly similar clonotypes, the results will be modified accordingly. The output is provided in the form of a table and a graph and can be found under the tab "Visualization." Characteristic examples of CDR3 length distribution are given in Fig. 4.

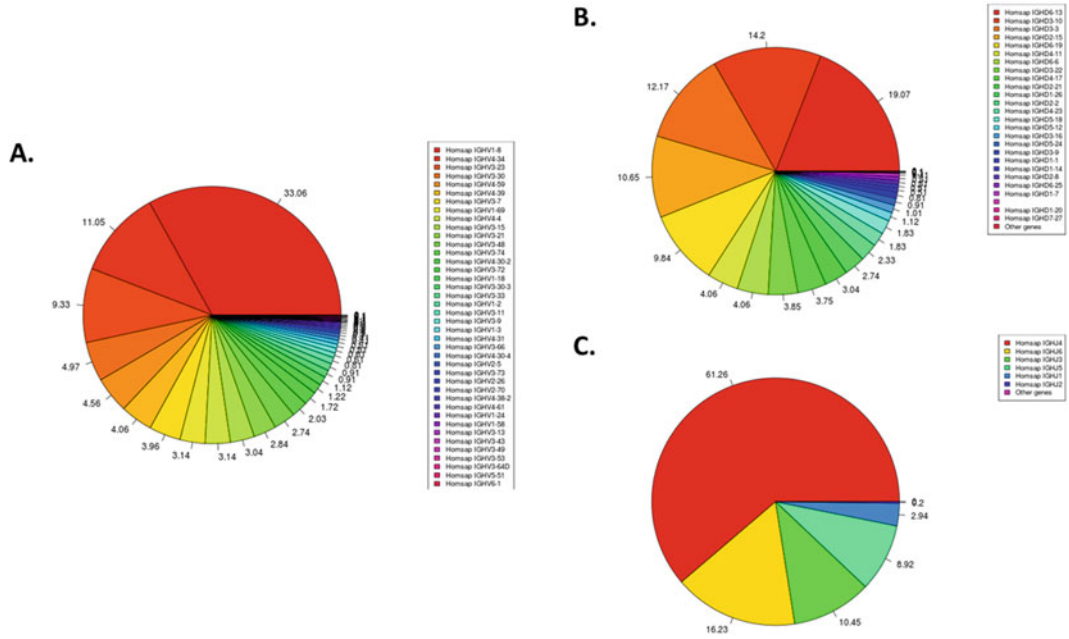


Fig. 3 The TRIP output for "Repertoire extraction." IGHV (**a**), IGHD (**b**), and IGHJ (**c**) gene repertoires at the clonotype level for Sample A. Strong biases were identified in all cases, characterized by predominance of the IGHV1–8, IGHD6–13, and IGHJ4 genes

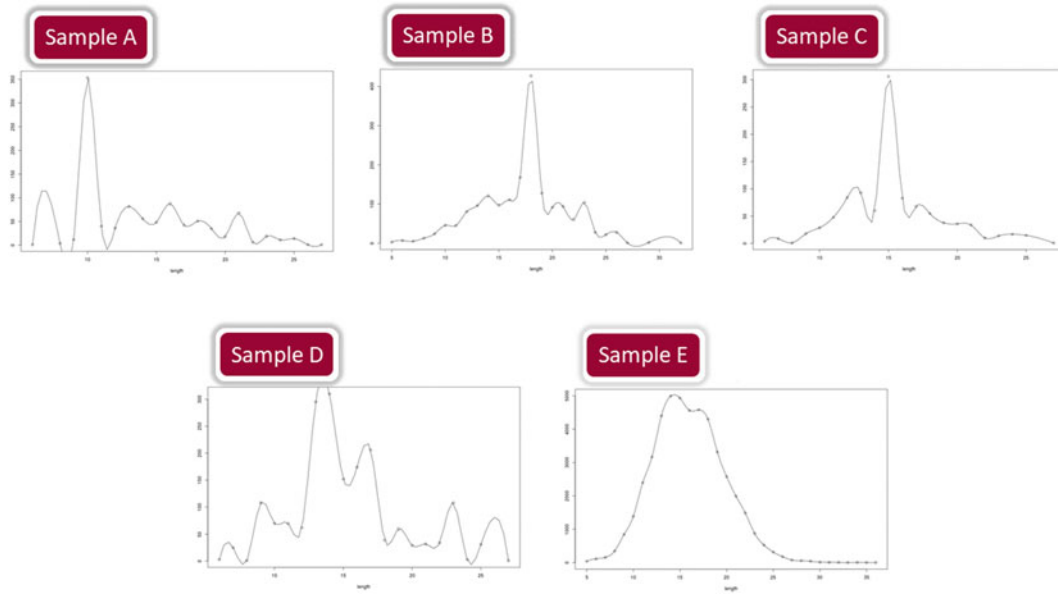


Fig. 4 The distribution of the CDR3 length in samples A–E. The x axis refers to the CDR3 length, and the y axis to the respective number of clonotypes. Samples A–C exhibited strong restrictions, in line with their monoclonal profiles. In contrast, restrictions were less prevalent in Sample D, in line with its oligoclonal clonotype repertoire. Finally, in the case of sample E, an almost Gaussian distribution of the CDR3 length is evident

5.3.5 pI Distribution

Next, the isoelectric point (pI, pH(I), and IEP) values of the CDR3 of each clonotype is extracted from the corresponding IMGT/HighV-QUEST output file, which is the pH at which the respective CDR3 carries no electrical charge or is electrically neutral. The pI of a given CDR3 is largely dependent on its amino acid composition. TRIP provides the distribution of the pI in a given dataset, based on the selection of either all or the merged clonotypes from the previous steps. A graph referring to the pI distribution can be found at the "Visualization" tab (Fig. 5).

5.3.6 Multiple Value Comparison

Different pairs of immunogenetic variables can be selected at this part of the pipeline. TRIP uses the output file from the computation of either all clonotypes or the merged clonotypes and performs comparisons between any given set of variables. The output file contains the values for each of the two selected variables and the number and relative frequency of clonotypes for each possible combination of values.

Eleven different variables that can be selected at this step include:

1. V gene.
2. V gene and allele.
3. J gene.
4. J gene and allele.
5. D gene.
6. D gene and allele.
7. CDR3 length.
8. D region reading frame.
9. Molecular mass.
10. pI.
11. V region identity %.

Figure 6 illustrates two examples of comparisons when using the V-gene and J-gene variables. The output files for the selected comparisons can be found at the "Multiple value comparison" tab.

5.3.7 Computation of Shared Clonotypes

In this section, TRIP scans different samples for the presence of identical clonotypes. The output file is provided in text (.txt) format with each row corresponding to a unique clonotype and each column to a different sample. Results include the absolute number of reads and the relative frequency of each clonotype in each sample (Column A: Sample id 1_Reads/Total, Column B: Sample id 1_Freq, Column C: Sample id 2_Reads/Total, Column D: Sample id 2_Freq). This type of analysis is based on the selection of either all clonotypes or just the merged clonotypes.

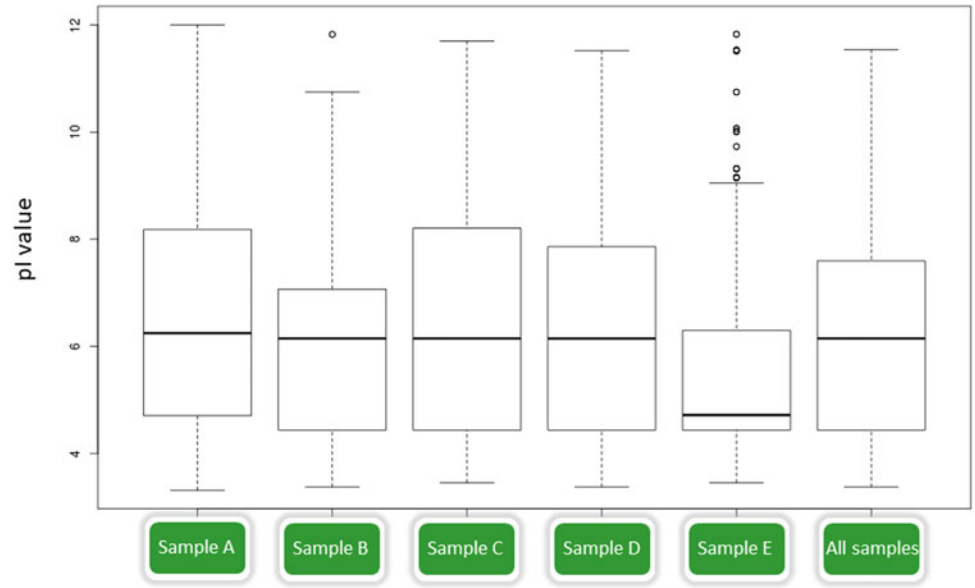


Fig. 5 The pI distribution for samples A–E individually and all samples together, using a boxplot. Clonotypes from Samples A–D displayed a similar pI distribution, whereas the clonotypes from sample E exhibited lower pI values

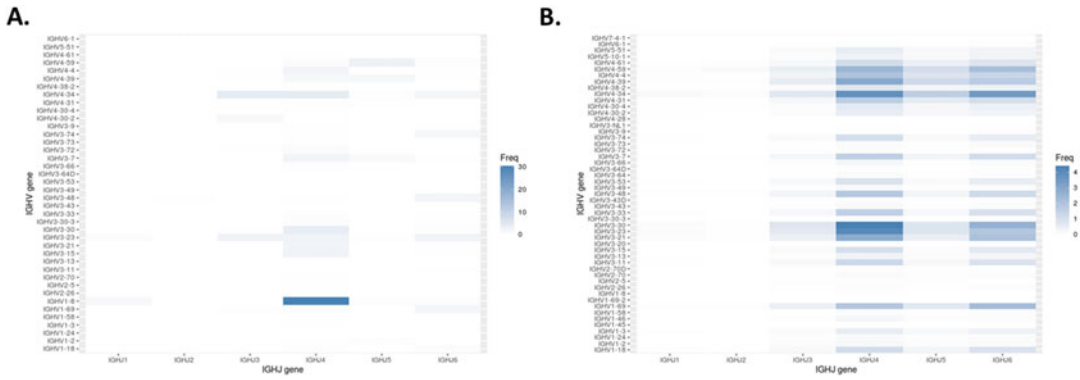


Fig. 6 Comparisons between IGHV and IGHJ gene utilization in monoclonal (Sample A) (**a**) versus polyclonal cases (Sample E) (**b**) using heatmaps. (**a**) A strong association between the IGHV1–8 and IGHJ4 genes is evident in Sample A, corresponding to the dominant clonotype. (**b**) Several associations are evident in Sample E, reflecting the polyclonal profile of this sample

5.3.8 Repertoire Comparison

Similar to the comparison of clonotypes, TRIP allows the comparison of gene or gene allele repertoires (*see* Subheading 5.3.3), between two or more samples/datasets. The output consists of a table where each row represents a unique gene and each column a sample. Results include the absolute number and relative frequency of the clonotypes expressing each gene in every individual sample (Column A: Sample id 1_N/Total, Column B: Sample id 1_Freq, Column C: Sample id 2_N/Total, Column D: Sample id 2_Freq).

Again, this type of analysis can be performed on either all clonotypes or just the merged clonotypes.

5.3.9 Clustering of CDR3 Sequences with Maximum Length Difference of One Amino Acid

As in the previous section concerning the merging of highly similar clonotypes (Subheading 5.3.2), at this point, TRIP allows for the merging of clonotypes differing by one amino acid in CDR3 length that are identical over the same length. In this case, TRIP adds one amino acid at a specified position of the shorter CDR3 resulting in the formation of two identical CDR3s. The output graph can be found at the "Visualization" tab.

5.3.10 Alignment

TRIP provides the option to align all clonotypes using the IMGT germline reference of the VDJ or VJ region at both the nucleotide and amino acid levels. An alignment table and a grouped alignment table based on the corresponding region are computed, and they are both available at the "Alignment" tab. Relevant gene alleles or a different reference sequence can be provided by the user.

5.3.11 Insert Identity Groups

At this point, TRIP enables the customization of the SHM analysis that can be applied at the next step (*see* Subheading 5.3.12). In detail, the user can specify the number of clonotype groups and the respective germline identity % thresholds that will be used for the SHM analysis. In certain clinical contexts, especially chronic lymphocytic leukemia (CLL), mutational categories defined by specific identity % thresholds have distinct clinical course, including responses to different treatments [21]. In that case, TRIP allows defining three distinct groups through the application of the 85–98% (*see* Subheading 5.2 on Selection for the application of the 85% cutoff), 98%–100%, and 100% cutoffs. The first group corresponds to "IG-mutated CLL" (M-CLL), the second to "IG-unmutated CLL" (U-CLL), and the third to "truly IG-unmutated" CLL cases. In terms of clonotype selection, TRIP gives the user the option to perform this part of the analysis on either all or just the merged clonotypes. Figure 7 depicts the application of these identity % thresholds in a series of cases.

This part of the analysis ("Insert identity groups") along with the next one ("Somatic hypermutation") apply to BcR IG datasets only, since the SHM mechanism does not operate in cells other than B cells.

5.3.12 Somatic Hypermutation Analysis

For SHM analysis, TRIP uses as reference the alignment tables produced at the Alignment step. This type of analysis can be applied to the entire dataset or only to clonotypes exhibiting either high frequency or specific immunogenetic properties.

As an output, TRIP offers information on:

1. The type of nucleotide mutations and relevant amino acid changes.

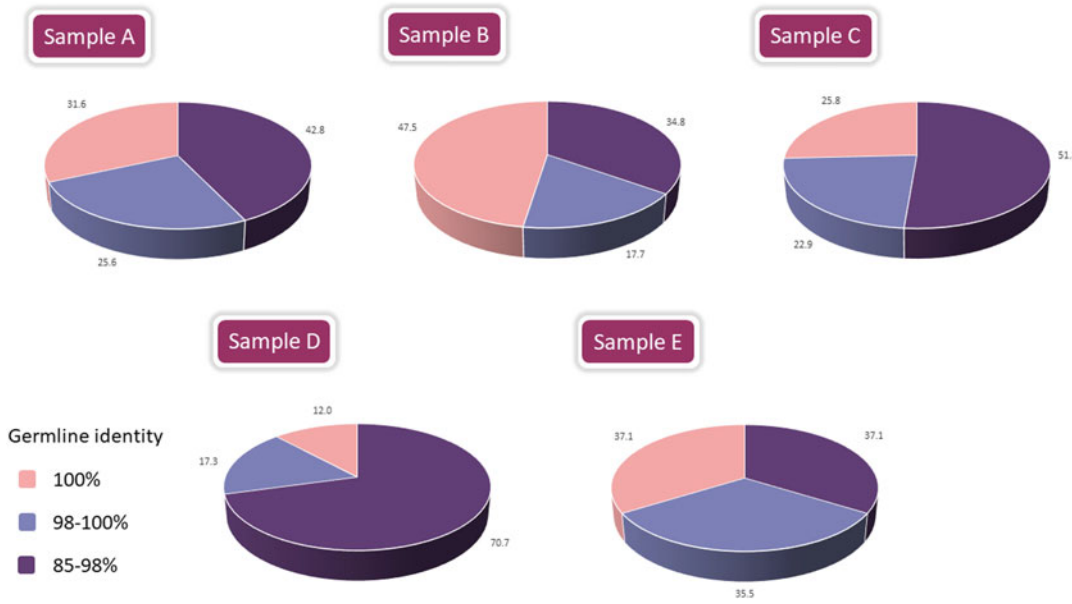


Fig. 7 Relative frequency of the three IG mutational subgroups, in Samples A–E. The mutated subgroup (germline identity, GI 85–98%) was dominant in Samples A, C, and D. Truly unmutated clonotypes (GI 100%) accounted for the largest fraction of the repertoire in Sample B, indicating a different biological context. Finally, polyclonal Sample E was characterized by similar frequency levels for all mutational subgroups, perhaps due to the lack of strong selection mechanisms

2. The exact position of each change, at both the nucleotide and amino acid levels.
3. The topology of each change, i.e., the region of the BcR IG V domain (FR-CDR).
4. The total number of clonotypes carrying each mutation.
5. The frequency of each change at the gene level.

5.3.13 Logo Creation

The final step of the analysis concerns the creation of a table containing information about the frequency of each aa at each specific position of the sequence, for sequences of the same length. The region of focus is set by the user and can be either the CDR3 or the entire VDJ or VJ region. The user can also choose if TRIP will provide a frequency table for all clonotypes or just for the top N clonotypes (based on their relative frequency). Subsequently, the data in the frequency table is used for the creation of a sequence Logo. The color code used in the Logo is based on the IMGT guidelines. The output of this step can be found at the "Logo" tab.

5.3.14 Visualization

The tab "Visualization" on the TRIP interface includes all different graph types that were produced during the course of the analytical pipeline. The first graph is a bar plot of either all clonotypes or the merged clonotypes, with the option of a frequency threshold. The

next graphs are pie charts of the selected V, D, and J gene repertoires. The option of applying a frequency threshold is also given here. The visualization of convergent evolution is next, with different options including a 3-D plot. Next on this tab is a pie chart and corresponding table concerning the selected identity groups for the SHM analysis along with the absolute number and frequency of clonotypes assigned to each group. The last graphs of this section are a candlestick chart for the depiction of the pI distribution and the line graph for the CDR3 distribution, below which the corresponding table is presented.

5.3.15 Overview

At this last tab of the TRIP interface, all main steps of the analysis are given, including the Preselection, Selection and Pipeline sections (Subheadings 5.1–5.3). The overview can be downloaded in pdf format. Furthermore, TRIP provides the user with the option to download all the output tables from every step of the analysis. Each table, though, can be downloaded separately, too, from its corresponding tab.

5.3.16 Dependencies

In the TRIP tool pipeline, different steps can be run independently. However, there are some dependencies:

1. It is necessary to select the option "Clonotype computation" in order to apply the following types of analysis:
   (a) "Highly similar Clonotypes computation."
   (b) "Repertoires Extraction". In the case that the "Highly Similar Clonotypes Computation" has been selected, the repertoires will be extracted for both the total clonotypes and the merged clonotypes.
   (c) "Alignment" using the option "Select top N clonotypes."
   (d) "Mutations" using the options "Select top N clonotypes" or "Select clonotypes separately."
   (e) "Logo" using the "Select top N clonotypes" option.
2. The "Somatic hypermutation status" is applied using the groups that have been selected using the "Insert identity groups" option.
3. If both "Alignment" and "Clonotypes computation" have been selected, the cluster ID in the alignment table is the same as the one in the Clonotype table. Otherwise, all elements in the "cluster_ID" column of the alignment table will be set to 0.
4. To apply "Mutations," "Alignment" should have run previously, using the "AA or Nt" option. The Mutation table is computed based on the grouped alignment table.

Acknowledgments

This work was supported in part by the Framework of the Hellenic Republic: Siemens Settlement Agreement, through the Hellenic Precision Medicine Network on Oncology project; the ERA-NET on Translational Cancer Research (TRANSCAN-2) acronym NOVEL project code (MIS) 5041673; and the Hellenic Foundation for Research and Innovation (HFRI) and the General Secretariat for Research and Technology (GSRT), under grant agreement No 336 (Project CLLon); the project ODYSSEAS (Intelligent and Automated Systems for enabling the Design, Simulation, and Development of Integrated Processes and Products) implemented under the "Action for the Strategic Development on the Research and Technological Sector," funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014–2020) and @co-financed by Greece and the European Union, with grant agreement no: MIS 5002462; and the EuroClonality-NGS working group.

Disclosures: Kostas Stamatopoulos has received honoraria and research support from Abbvie, Janssen, Astra-Zeneca, and Gilead.

References

1. Rawstron AC, Fazi C, Agathangelidis A et al (2016) A complementary role of multiparameter flow cytometry and high-throughput sequencing for minimal residual disease detection in chronic lymphocytic leukemia: an European research initiative on CLL study. Leukemia 30(4):929–936. https://doi.org/10.1038/leu.2015.313
2. Rodriguez-Vicente AE, Bikos V, Hernandez-Sanchez M et al (2017) Next-generation sequencing in chronic lymphocytic leukemia: recent findings and new horizons. Oncotarget 8(41):71234–71248. https://doi.org/10.18632/oncotarget.19525
3. Bolotin DA, Shugay M, Mamedov IZ et al (2013) MiTCR: software for T-cell receptor sequencing data analysis. Nat Methods 10(9):813–814. https://doi.org/10.1038/nmeth.2555
4. Kuchenbecker L, Nienen M, Hecht J et al (2015) IMSEQ--a fast and error aware approach to immunogenetic sequence analysis. Bioinformatics 31(18):2963–2971. https://doi.org/10.1093/bioinformatics/btv309
5. Thomas N, Heather J, Ndifon W et al (2013) Decombinator: a tool for fast, efficient gene assignment in T-cell receptor sequences using a finite state machine. Bioinformatics 29(5):542–550. https://doi.org/10.1093/bioinformatics/btt004
6. Yang X, Liu D, Lv N et al (2015) TCRklass: a new K-string-based algorithm for human and mouse TCR repertoire characterization. J Immunol 194(1):446–454. https://doi.org/10.4049/jimmunol.1400711
7. Aouinti S, Giudicelli V, Duroux P et al (2016) IMGT/StatClonotype for pairwise evaluation and visualization of NGS IG and TR IMGT Clonotype (AA) diversity or expression from IMGT/HighV-QUEST. Front Immunol 7:339. https://doi.org/10.3389/fimmu.2016.00339
8. Aouinti S, Malouche D, Giudicelli V et al (2015) IMGT/HighV-QUEST statistical significance of IMGT Clonotype (AA) diversity per gene for standardized comparisons of next generation sequencing Immunoprofiles of immunoglobulins and T cell receptors. PLoS One 10(11):e0142353. https://doi.org/10.1371/journal.pone.0142353
9. Bolotin DA, Poslavsky S, Mitrophanov I et al (2015) MiXCR: software for comprehensive adaptive immunity profiling. Nat Methods 12(5):380–381. https://doi.org/10.1038/nmeth.3364
10. Duez M, Giraud M, Herbert R et al (2016) Vidjil: a web platform for analysis of high-throughput repertoire sequencing. PLoS One 11(11):e0166126. https://doi.org/10.1371/journal.pone.0166126

11. Bystry V, Reigl T, Krejci A et al (2017) ARResT/interrogate: an interactive immunoprofiler for IG/TR NGS data. Bioinformatics 33(3):435–437. https://doi.org/10.1093/bioinformatics/btw634
12. Kotouza MT, Gemenetzi K, Galigalidou C et al (2020) TRIP - T cell receptor/immunoglobulin profiler. BMC Bioinformatics 21(1):422. https://doi.org/10.1186/s12859-020-03669-1
13. Gemenetzi K, Agathangelidis A, Sutton L-A et al (2018) Remarkable functional constraints on the antigen receptors of CLL stereotyped subset #2: high-throughput Immunogenetic evidence. Blood 132(Supplement 1):1839. https://doi.org/10.1182/blood-2018-99-119125
14. Vardi A, Vlachonikola E, Mourati S et al (2019) High-throughput B-cell immunoprofiling at diagnosis and relapse offers further evidence of functional selection throughout the natural history of chronic lymphocytic leukemia. HemaSphere 3:512. https://doi.org/10.1097/01.hs9.0000562808.48237.52
15. Vardi A, Vlachonikola E, Papazoglou D et al (2020) T-cell dynamics in chronic lymphocytic leukemia under different treatment modalities. Clin Cancer Res 26(18):4958–4969. https://doi.org/10.1158/1078-0432.CCR-19-3827
16. Vlachonikola E, Vardi A, Kastritis E et al (2018) Longitudinal T cell Immunoprofiling of patients with relapsed and/or refractory myeloma who receive Daratumumab monotherapy: a subanalysis of a phase 2 study (the REBUILD study). Blood 134(Supplement 13167):3167. https://doi.org/10.1182/blood-2019-124655
17. Galigalidou C, Papadopoulou A, Stalika E et al (2018) High-throughput T cell receptor (TR) repertoire analysis of virus-specific T cells: implications for T cell immunotherapy and viral infection risk stratification. Blood 132(Supplement 1):2057. https://doi.org/10.1182/blood-2018-99-118851
18. Gemenetzi K, Stalika E, Agathangelidis A et al (2018) Evidence for epitope-specific T cell responses in HIV-associated non neoplastic lymphadenopathy: High-Throughput Immunogenetic Evidence. Blood 132(Supplement 1):1117. https://doi.org/10.1182/blood-2018-99-118975
19. Lefranc MP, Giudicelli V, Duroux P et al (2015) IMGT(R), the international ImMunoGeneTics information system(R) 25 years on. Nucleic Acids Res 43(Database issue): D413–D422. https://doi.org/10.1093/nar/gku1056
20. Li S, Lefranc MP, Miles JJ et al (2013) IMGT/HighV QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. Nat Commun 4: 2333. https://doi.org/10.1038/ncomms3333
21. Chiorazzi N, Stevenson FK (2020) Celebrating 20 years of IGHV mutation analysis in CLL. HemaSphere 4(1):e334. https://doi.org/10.1097/HS9.0000000000000334

INDEX

Anton W. Langerak (ed.), *Immunogenetics: Methods and Protocols*, Methods in Molecular Biology, vol. 2453, https://doi.org/10.1007/978-1-0716-2115-8,

L

M

N

P

Q

R

S

T

U

V

W