Compendium of Plant Genomes
Series Editor: Chittaranjan Kole

Rudi Appels
Kellye Eversole
Catherine Feuillet
Dusti Gallagher *Editors*

The Wheat Genome

OPEN ACCESS

Springer

Compendium of Plant Genomes

Series Editor

Chittaranjan Kole, President, International Climate Resilient Crop Genomics Consortium (ICRCGC), President, International Phytomedomics and Nutriomics Consortium (IPNC) and President, Genome India International (GII), Kolkata, India

Whole-genome sequencing is at the cutting edge of life sciences in the new millennium. Since the first genome sequencing of the model plant *Arabidopsis thaliana* in 2000, whole genomes of about 100 plant species have been sequenced and genome sequences of several other plants are in the pipeline. Research publications on these genome initiatives are scattered on dedicated web sites and in journals with all too brief descriptions. The individual volumes elucidate the background history of the national and international genome initiatives; public and private partners involved; strategies and genomic resources and tools utilized; enumeration on the sequences and their assembly; repetitive sequences; gene annotation and genome duplication. In addition, synteny with other sequences, comparison of gene families and most importantly potential of the genome sequence information for gene pool characterization and genetic improvement of crop plants are described.

Rudi Appels · Kellye Eversole ·
Catherine Feuillet · Dusti Gallagher
Editors

The Wheat Genome

Springer

Editors
Rudi Appels
University of Melbourne and AgriBio
La Trobe University
Melbourne, VIC, Australia

Kellye Eversole
International Wheat Genome
Sequencing Consortium
Eau Claire, WI, USA

Catherine Feuillet
Inari Agriculture (United States)
Cambridge, MD, USA

Dusti Gallagher
Fritz Consulting
Wamego, KS, USA

ISSN 2199-4781 ISSN 2199-479X (electronic)
Compendium of Plant Genomes
ISBN 978-3-031-38292-5 ISBN 978-3-031-38294-9 (eBook)
https://doi.org/10.1007/978-3-031-38294-9

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Paper in this product is recyclable.

This book series is dedicated to my wife Phullara and our children Sourav and Devleena

Chittaranjan Kole

Preface to the Series

Genome sequencing has emerged as the leading discipline in the plant sciences coinciding with the start of the new century. For much of the twentieth century, plant geneticists were only successful in delineating putative chromosomal location, function, and changes in genes indirectly through the use of a number of "markers" physically linked to them. These included visible or morphological, cytological, protein, and molecular or DNA markers. Among them, the first DNA marker, the RFLPs, introduced a revolutionary change in plant genetics and breeding in the mid-1980s, mainly because of their infinite number and thus potential to cover maximum chromosomal regions, phenotypic neutrality, absence of epistasis, and codominant nature. An array of other hybridization-based markers, PCR-based markers, and markers based on both facilitated construction of genetic linkage maps, mapping of genes controlling simply inherited traits, and even gene clusters (QTLs) controlling polygenic traits in a large number of model and crop plants. During this period, a number of new mapping populations beyond F_2 were utilized and a number of computer programs were developed for map construction, mapping of genes, and for mapping of polygenic clusters or QTLs. Molecular markers were also used in the studies of evolution and phylogenetic relationship, genetic diversity, DNA fingerprinting, and map-based cloning. Markers tightly linked to the genes were used in crop improvement employing the so-called marker-assisted selection. These strategies of molecular genetic mapping and molecular breeding made a spectacular impact during the last one and a half decades of the twentieth century. But still they remained "indirect" approaches for elucidation and utilization of plant genomes since much of the chromosomes remained unknown and the complete chemical depiction of them was yet to be unraveled.

Physical mapping of genomes was the obvious consequence that facilitated the development of the "genomic resources" including BAC and YAC libraries to develop physical maps in some plant genomes. Subsequently, integrated genetic–physical maps were also developed in many plants. This led to the concept of structural genomics. Later on, emphasis was laid on EST and transcriptome analysis to decipher the function of the active gene sequences leading to another concept defined as functional genomics. The advent of techniques of bacteriophage gene and DNA sequencing in the

1970s was extended to facilitate sequencing of these genomic resources in the last decade of the twentieth century.

As expected, sequencing of chromosomal regions would have led to too much data to store, characterize, and utilize with the-then available computer software. But the development of information technology made the life of biologists easier by leading to a swift and sweet marriage of biology and informatics, and a new subject was born—bioinformatics.

Thus, the evolution of the concepts, strategies, and tools of sequencing and bioinformatics reinforced the subject of genomics—structural and functional. Today, genome sequencing has traveled much beyond biology and involves biophysics, biochemistry, and bioinformatics!

Thanks to the efforts of both public and private agencies, genome sequencing strategies are evolving very fast, leading to cheaper, quicker, and automated techniques right from clone-by-clone and whole-genome shotgun approaches to a succession of second-generation sequencing methods. The development of software of different generations facilitated this genome sequencing. At the same time, newer concepts and strategies were emerging to handle sequencing of the complex genomes, particularly the polyploids.

It became a reality to chemically—and so directly—define plant genomes, popularly called whole-genome sequencing or simply genome sequencing.

The history of plant genome sequencing will always cite the sequencing of the genome of the model plant *Arabidopsis thaliana* in 2000 that was followed by sequencing the genome of the crop and model plant rice in 2002. Since then, the number of sequenced genomes of higher plants has been increasing exponentially, mainly due to the development of cheaper and quicker genomic techniques and, most importantly, the development of collaborative platforms such as national and international consortia involving partners from public and/or private agencies.

As I write this preface for the first volume of the new series "Compendium of Plant Genomes," a net search tells me that complete or nearly complete whole-genome sequencing of 45 crop plants, eight crop and model plants, eight model plants, 15 crop progenitors and relatives, and three basal plants is accomplished, the majority of which are in the public domain. This means that we nowadays know many of our model and crop plants chemically, i.e., directly, and we may depict them and utilize them precisely better than ever. Genome sequencing has covered all groups of crop plants. Hence, information on the precise depiction of plant genomes and the scope of their utilization are growing rapidly every day. However, the information is scattered in research articles and review papers in journals and dedicated Web pages of the consortia and databases. There is no compilation of plant genomes and the opportunity of using the information in sequence-assisted breeding or further genomic studies. This is the underlying rationale for starting this book series, with each volume dedicated to a particular plant.

Plant genome science has emerged as an important subject in academia, and the present compendium of plant genomes will be highly useful to both students and teaching faculties. Most importantly, research scientists involved in genomics research will have access to systematic deliberations on the plant genomes of their interest. Elucidation of plant genomes is of interest not only for the geneticists and breeders, but also for practitioners of an array of plant science disciplines, such as taxonomy, evolution, cytology, physiology, pathology, entomology, nematology, crop production, biochemistry, and obviously bioinformatics. It must be mentioned that information regarding each plant genome is ever-growing. The contents of the volumes of this compendium are, therefore, focusing on the basic aspects of the genomes and their utility. They include information on the academic and/or economic importance of the plants, description of their genomes from a molecular genetic and cytogenetic point of view, and the genomic resources developed. Detailed deliberations focus on the background history of the national and international genome initiatives, public and private partners involved, strategies and genomic resources and tools utilized, enumeration on the sequences and their assembly, repetitive sequences, gene annotation, and genome duplication. In addition, synteny with other sequences, comparison of gene families, and, most importantly, the potential of the genome sequence information for gene pool characterization through genotyping by sequencing (GBS) and genetic improvement of crop plants have been described. As expected, there is a lot of variation of these topics in the volumes based on the information available on the crop, model, or reference plants.

I must confess that as the series editor, it has been a daunting task for me to work on such a huge and broad knowledge base that spans so many diverse plant species. However, pioneering scientists with lifetime experience and expertise on the particular crops did excellent jobs editing the respective volumes. I myself have been a small science worker on plant genomes since the mid-1980s and that provided me the opportunity to personally know several stalwarts of plant genomics from all over the globe. Most, if not all, of the volume editors are my longtime friends and colleagues. It has been highly comfortable and enriching for me to work with them on this book series. To be honest, while working on this series I have been and will remain a student first, a science worker second, and a series editor last. And, I must express my gratitude to the volume editors and the chapter authors for providing me the opportunity to work with them on this compendium.

I also wish to mention here my thanks and gratitude to Springer staff, particularly Dr. Christina Eckey and Dr. Jutta Lindenborn, for the earlier set of volumes and presently Ing. Zuzana Bernhart for all their timely help and support.

I always had to set aside additional hours to edit books beside my professional and personal commitments—hours I could and should have given to my wife, Phullara, and our kids, Sourav and Devleena. I must mention that

they not only allowed me the freedom to take away those hours from them but also offered their support in the editing job itself. I am really not sure whether my dedication of this compendium to them will suffice to do justice to their sacrifices for the interest of science and the science community.

New Delhi, India Chittaranjan Kole

Preface

In 2005, we embarked on the journey to develop a high-quality, reference sequence for the bread wheat genome. The vision was to produce a sequence comparable in quality to the rice genome. Many told us that it would be impossible, and the best strategy would be a low coverage sequence. Yet, we persisted. We persisted because we listened to the breeders and future users of the genome sequence. They asked for a high-quality sequence of the hexaploid bread wheat because it contributes more to the human diet than any other crop species and is the most widely grown crop in the world. With the publication of the IWGSC RefSeq v1.0 and its accompanying annotation in 2018, we marked the attainment of this vision. This journey was accomplished with contributions from scientists all over the globe and would not have been possible without the public–private partnerships that ensued and the determination of many who never lost sight of the vision.

IWGSC RefSeq v1.0 and its annotation created a paradigm shift that ushered in a new world for wheat breeders and scientists, and the advancements have been rapid and simply amazing. Thanks to technological advancements, genome sequencing of large polyploid genomes with highly repetitive content like wheat has now become routine and platinum quality sequences of wheat can be delivered in weeks at a reasonable cost. New resources for the molecular genetic study of wheat and its application for wheat improvement are arriving at a record pace.

The production of the *Wheat Genome* book is essential and highlights the groundbreaking research ongoing for this critical crop. This volume includes papers describing the development of the reference sequence, new assemblies of commercial varieties, genome-wide studies, and the accelerated cloning of agronomically important genes and provides valuable resources and literature for fundamental and applied research, crop improvement and teaching. It illustrates the value and impact of having high-quality reference genomes for overall crop improvements that address the dual challenges of producing a reliable, safe, and sustainable supply of wheat while facing a rapidly changing climate.

We are indebted to our colleagues from the global wheat community who contributed to this book and for their continuous commitment to provide resources and knowledge to scientists and breeders around the world.

Due to the significance of this material and the desire to ensure accessibility to everyone, a group joined together to cover the costs associated with publishing an open access book. We are grateful to the following entities and individuals who contributed to the costs:

- Rudi Appels, Kellye Eversole, and Catherine Feuillet
- The National Research Institute for Agriculture, Food and Environment (INRAE)
- The International Maize and Wheat Improvement Center (CIMMYT)
- Syngenta

Melbourne, Australia
Eau Claire, USA
Cambridge, USA
Wamego, USA,
February 2023

Rudi Appels
Kellye Eversole
Catherine Feuillet
Dusti Gallagher

Contents

1 The Bread Wheat Reference Genome Sequence

Jane Rogers

Abstract

In 2018, the International Wheat Genome Sequencing Consortium published a reference genome sequence for bread wheat (*Triticum aestivum* L.). The landmark achievement was the culmination of a thirteen-year international effort focused on the production of a genome sequence linked to genotypic and phenotypic maps to advance understanding of traits and accelerate improvements in wheat breeding. In this chapter, we describe the challenges of the project, the strategies employed, how the project adapted over time to incorporate technological improvements in genome sequencing and the project outcomes.

Keywords

IWGSC · Bread wheat · Genome sequence · Trait improvement

International Wheat Genome Sequencing ConsortiumJ. Rogers (✉) · International Wheat Genome Sequencing Consortium
International Wheat Genome Sequencing Consortium, Eau Claire, WI, USA
e-mail: janerogersh@gmail.com

1.1 Introduction

In 2018, the International Wheat Genome Sequencing Consortium published a reference genome sequence for bread wheat (*Triticum aestivum* L.). The landmark achievement was the culmination of a thirteen-year international effort focused on the production of a genome sequence linked to genotypic/phenotypic maps to advance understanding of traits and accelerate improvements in wheat breeding. In this monograph, we bring together contributions from colleagues to highlight the advances and document the resources now available for wheat research and its relatives.

This first chapter describes the challenges of developing the bread wheat reference genome sequence project, the strategies employed, how the project adapted over time to incorporate technological improvements in genome sequencing and the project outcomes. The following chapters include Chap. 2 for a comprehensive documentation of available data repositories; Chap. 3 using chromosomes as a focus underpinning the establishment of a high-quality assembly; Chap. 4 on the challenge of the structural and functional annotation of the genome; Chap. 5 the wheat transcriptome and functional gene networks; Chap. 6 covering the genome-level diversity within cultivated wheats; Chap. 7 highlights the advances in sequencing ancient wheat DNA; Chap. 8 examines the

R. Appels et al. (eds.), *The Wheat Genome*, Compendium of Plant Genomes,
https://doi.org/10.1007/978-3-031-38294-9_1

impact of the durum wheat genome in identifying new germplasm for breeding; Chap. 9 demonstrates the use of the genome sequence to identify genes underpinning agronomic traits; Chap. 10 examines new and faster approaches to cloning disease resistance; Chap. 11 documents the genome views of the CIMMYT breeding programme; Chap. 12 reviews the gene pools contributing to wheat genetic variation; Chap. 13 provides an overview of approaches to integrating genomics into breeding strategies; Chap. 14 explores pan-genomes for capturing new functionalities and refining wheat genomics; Chap. 15 provides insights into the extensive germplasm resources established within the wheat community.

1.2 Origins of the Wheat Genome Project

Since the early 1990s, there has been a growing realization across the world that to feed a rapidly growing human population grain production needs to increase by an annual rate of 2% on an area of land equivalent to that already under cultivation. Wheat was one of the first domesticated food crops and continues to be the most important food grain source for humans today. Wheat is grown on a greater area than any other crop (approx. 255 m ha, Bonjean et al. 2016; https://www.fao.org/faostat/en/#data) and is best adapted to temperate regions of the world.

By 2003, demand for wheat already regularly outstripped annual global production, and, faced with an estimated 25% annual loss due to biotic (pests) and abiotic stresses (heat, frost, drought and salinity), it was clear that a paradigm shift was needed in wheat breeding and understanding of wheat biology to attain a sustainable food supply. At the time, other areas of biology were benefitting from access to genome data generated through high throughput DNA sequencing projects. The largest genome sequence available was the human genome sequence (3 Gb), for which draft and finished versions were published in 2001 (Lander et al. 2001; Venter et al. 2001) and 2004 (International Human Genome Consortium 2004), respectively. The sequence rapidly yielded new information about the structure, organisation, genes, genetic traits and genome variation to make an immediate impact on human biology and medicine. The *Arabidopsis thaliana* genome sequence (ca.100 Mb) published in 2000 (The Arabidopsis Genome Initiative 2000) was similarly impacting understanding of genes and genetic traits in plants, and genome sequencing projects for rice (450 Mb) (Eckhardt 2000; International Rice Genome Sequencing Project and Sasaki 2005) and maize (ca 1 Gb) (Chandler and Brender 2002) were underway.

In November 2003, a USDA-NSF workshop was convened to consider the feasibility and requirements of a wheat genome sequence Gill et al. 2004). The development of genomic resources for wheat lagged behind the other major crops due to the genome posing three major challenges. First, the wheat genome is very large. The genome size estimated from DNA-Feulgen studies of root tip nuclei was ca. 17 Gb, over five times the size of the human genome. Second, early cytogenetic studies established that several *Triticeae* species, including bread wheat, are polyploid and originated from spontaneous hybridisation of diploid genomes (Kihara 1944; McFadden and Sears 1946). The genome of bread wheat is allohexaploid, comprising 21 pairs of homologous chromosomes originating from three homeologous sets of seven chromosomes, referred to as the A, B and D sub-genomes. The hexaploid wheat genome arose from two hybridisation events, estimated to have taken place between 0.8 and 0.5 million years ago and 8–10,000 years ago, respectively. The first hybridisation event occurred between a species related to *Triticum urartu* ($2n=2x=14$; A^uA^u) and one or more species from the Sitopsis section related most closely to *Aegilops speltoides* ($2n=2x=14$; SS), believed to be the closest living relative to the B genome progenitor. The resulting

fertile tetraploid ($2n = 4x = 28$; AABB)) was domesticated over 10,000 years ago and developed into emmer wheat (*Triticum turgidum*). The hybridisation of emmer wheat in a region south of the Caspian Sea some 8–10,000 years ago with *Aegilops tauschii* ($2n = 2x = 14$), a wild diploid with a D genome, led to the fertile hexaploid with an AABBDD genome, the ancestral bread wheat (Zohary et al. 2012). This has subsequently undergone a number of structural and functional rearrangements, including slight reductions (2–10%) in the size of the homoeologous genomes compared to the diploid ancestors, to produce the stable genome of bread wheat of today (Feldman and Levy 2009). Because these events have taken place over a short evolutionary timescale, the three sub-genomes exhibit high levels homology, with similar gene contents and high levels of synteny with other grass species and diploid wheat relatives. These high levels of similarity have hampered genome sequence assembly and the assignment of genes or other tag sequences to specific sub-genomes to distinguish between specific variants that may have phenotypic importance.

The additional challenge for sequencing the wheat genome is its very high repetitive sequence content. Early studies suggested that approximately 83% of the genome comprises transposable elements (TE) that arose from massive amplifications of inserted elements in the ancestral *Triticeae* genome. These have subsequently evolved independently in individual sub-genomes to give rise to characteristic quantitative and qualitative variations in the A, B and D genomes of modern bread wheat. Repeat elements have proved challenging for all sequence assembly algorithms, and the extent to which qualitative and quantitative differences in types of repeats and their distribution across the homoeologous chromosomes of hexaploid wheat could be or needed to be resolved to understand genomic function was an important consideration (see also Chap. 4).

The USDA-NSF workshop participants recognised that a high-quality reference genome sequence for wheat would underpin future wheat improvement by providing access to a complete gene catalogue, an unlimited number of molecular markers to enable genome-based selection of new varieties and a framework for the efficient exploitation of natural and induced genetic diversity. It would also provide insights into the functioning of a polyploid genome. It was agreed that a wheat genome project should focus on the hexaploid wheat variety CHINESE SPRING, for which resources that had been developed previously included large genetic stocks of aneuploid lines (Sears 1954, 1966) and sets of tag sequences, used to evaluate the gene content. In recognition of the complexity of the genome, several pilot projects were proposed to inform the development of a sequencing strategy. These included (i) construction of an accurate, sequence-ready physical map based on ordered BAC contigs; (ii) assessment of the feasibility of a chromosome-based approach for mapping and sequencing; and (iii) exploration of different strategies for gene enrichment. The outcomes of these projects were evaluated under the umbrella of the International Wheat Genome Consortium (IWGSC) which was established in 2005. The aims of the Consortium focus on advancing agricultural research for wheat production and utilisation by developing DNA-based tools and resources resulting from the complete sequence of the hexaploid wheat genome.

1.3 Wheat Genome Strategy Development

The size and complexity of the bread wheat genome initially caused many to believe that determining a genome sequence would be impossible within a reasonable time frame and budget. Several projects were initiated that aimed to reduce the complexity by focusing on diploid relatives of wheat A and D genomes (*T. urartu*, Ling et al. 2013; Ling et al. 2018; *A. tauschii*, Jia et al. 2013) or by focusing only on the assembly of genic regions from the hexaploid wheat genome (see Chap. 4). Bread wheat

breeders and researchers, however, realised that to provide the tools and resources for bread wheat research would ultimately require the genome of the hexaploid (Feuillet et al. 2016).

The determination of the DNA sequence of whole genomes is achieved by piecing together shorter lengths of DNA sequence in the order and orientation in which they occur in the organism from which the DNA was extracted. By 2005, two main approaches to genome sequencing had been established and were being applied to different genomes.

1.3.1 The Hierarchical Shotgun Strategy

This strategy is based on a two-step approach entailing initial construction of a physical map of the target genome followed by sequencing and assembly of short DNA fragments (typically 500 bp–1 kb) generated from sets of overlapping clones that represent a minimal tiling path (MTP) across the genomic DNA. Sequences representing typically at least tenfold coverage of each clone in paired sequence reads are assembled into longer pieces (contigs) using an assembly algorithm that identifies and joins matching sequences. The number of contigs into which each clone is assembled depends on a variety of factors, including clone representation in sequence fragments, sequence depth and quality and the repeat content of the DNA. Once an initial assembly has been made further, directed sequencing can be undertaken to improve the sequence quality, close gaps and resolve ambiguities. Finally, sequence overlaps between clones are identified after removal of cloning and sequencing vector sequences, and the clone sequences are linked to produce a pseudomolecule representing chromosomal DNA. The hierarchical shotgun approach was used to produce the first reference sequence for the human genome (Lander et al. 2001) and to produce the first reference genome sequences for plants, *A. thaliana* (The Arabidopsis Genome Initiative 2000) and rice (International Rice Genome sequencing Project and Sasaki 2005). It has subsequently been used in the production of reference sequences for the legume *Medicago truncatula* (Young et al. 2011) and to manage the complexity of the highly repetitive 3.5 Gb maize genome (Schnable et al. 2009). By requiring prior generation of a physical map, the hierarchical approach to genome sequencing increased the timespan and cost of genome projects. Some of the advantages, however, were that it enabled targeted sequencing of regions and targeted resolution of problems, and it facilitated project and cost sharing by enabling distribution of mapping and sequencing among multiple groups. It also generated clone resources that have been used to sequence specific genes or regions of interest ahead of the genome sequence becoming available. Until the very recent introduction of improved algorithms for short read sequence assembly (Clavijo et al. 2017; Avni et al. 2017), accurate sequencing reads in excess of 15–20 kb (De Coster et al. 2021) and the development of alternatives to physical maps for long-range structural organisation, such as optical maps (Keeble-Gagnère et al. 2018) and chromosome conformation capture sequencing (Hi-C, Burton et al. 2013), the hierarchical shotgun approach produced the most complete and accurate reference genome sequences, supporting detailed annotation and downstream applications in functional genomics.

1.3.2 Whole Genome Sequencing (WGS) Strategy

The WGS strategy is based on the random fragmentation (shotgun fragmentation) of whole genome DNA, sequencing the ends of the fragments and assembly of the overlapping sequences to build up longer lengths of DNA. Typically, fragments of different sizes are used and pairs of sequences from the ends of sized fragments representing at least 30-fold coverage of the genome are assembled. In 1977, Sanger et al. (1977) reported the use of whole genome shotgun sequencing to assemble the genome of the bacteriophage ϕX174 (5386 bp). Subsequently, the approach has been used to

sequence genomes of increasing complexity, including a wide variety of plants. It was championed in the late 1990s by C. Venter to sequence the genomes of *Haemophilus influenzae* (Fleischmann et al. 1995), *Drosophila melanogaster* (Adams et al. 2000) and the human genome (Venter et al. 2001). As sequencing costs have fallen with the introduction of second-generation sequencing technologies, whole genome shotgun approaches were considered a more tractable way to access large genomes, particularly those of plants (Feuillet et al. 2011; Jackson et al. 2011).

Factors affecting the quality of the assembly that can be achieved with this approach include the completeness and depth of coverage of the genome in sequence fragments, the level of bias in the fragmentation, cloning and sequencing processes caused by specific sequence motifs or repetitive elements, the sequence depth (number of times each individual piece of DNA is sequenced) and the power of the assembly algorithm. Highly repetitive genomes are particularly challenging where sequence read lengths are shorter than the length of repeats and reads cannot be positioned uniquely. As a result, they are often not assembled in the genome, leaving gaps.

Although the hierarchical and whole genome sequencing strategies have often been regarded as strategic competitors, they can be used to complement each other to achieve a more complete result. Methods to integrate whole genome sequence data into a BAC-based genome and integration of BAC sequences into a whole genome shotgun have been developed resulting in many of the higher-quality genome sequences being hybrid assemblies (e.g. mouse (Mouse Genome Sequencing Consortium 2002), zebrafish (Howe, et al. 2013), Drosophila (Celniker and Rubin 2003), Medicago (Young et al. 2011), maize (Schnable et al. 2009), rice (International Rice Genome Sequencing Project and Sasaki 2005) and tomato (The Tomato Genome Sequencing Consortium 2012)). Such assemblies achieve more complete coverage of the genome, enabling more accurate annotation, whilst still delivering resources for targeted improvement, gene cloning, etc.

1.4 IWGSC Strategic Roadmap

The IWGSC published its first roadmap for the bread wheat genome in 2006. The strategy proposed was based on reducing the complexity of the genome by generating physical maps and sequences for individual chromosome arms. This had the advantage of reducing the size of the assembly challenge to between 200 and 800 Mb, comparable to the sizes of other plant genomes (Doležel et al. 2007). It also largely eliminated problems of mis-assembling similar regions or sequences originating from homoeologous chromosomes. This chromosome-based approach was dependent upon the technological advances in flow cytometric chromosome sorting developed by the group of J. Doležel (Institute of Experimental Botany, Czech Republic) (see Chap. 3.). Between 2004 and 2013, the group flow sorted and produced BAC libraries representing 37 bread wheat chromosome/chromosome arms. These comprised a single library for chromosome 3B (Šafář et al. 2004), a composite library for chromosomes 1D, 4D and 6D (Janda et al. 2004) and individual libraries for each arm of the remaining 17 chromosomes. The complete set of BAC libraries contains 2,713,728 clones (Šafář et al. 2010). In 2008, Paux et al. (2008) reported the construction of the first physical map of a wheat chromosome, 3B. The map covered approximately 82% of the estimated size of the chromosome and provided a minimal tile path of physically mapped clones for sequencing. It also provided a 'proof of principle' for the hierarchical chromosome-based strategy to map and sequence the hexaploid wheat genome. Following the generation of the first physical maps, the IWGSC continued its focus on the production of physical maps for the whole genome, recruiting groups throughout the world to join the enterprise. In total, 17 groups from 14 countries contributed and the physical maps for all chromosomes were complete by January 2014.

Throughout the course of the wheat genome project, the strategy and roadmap evolved to take account of technological advances. In 2010, the roadmap was updated to incorporate

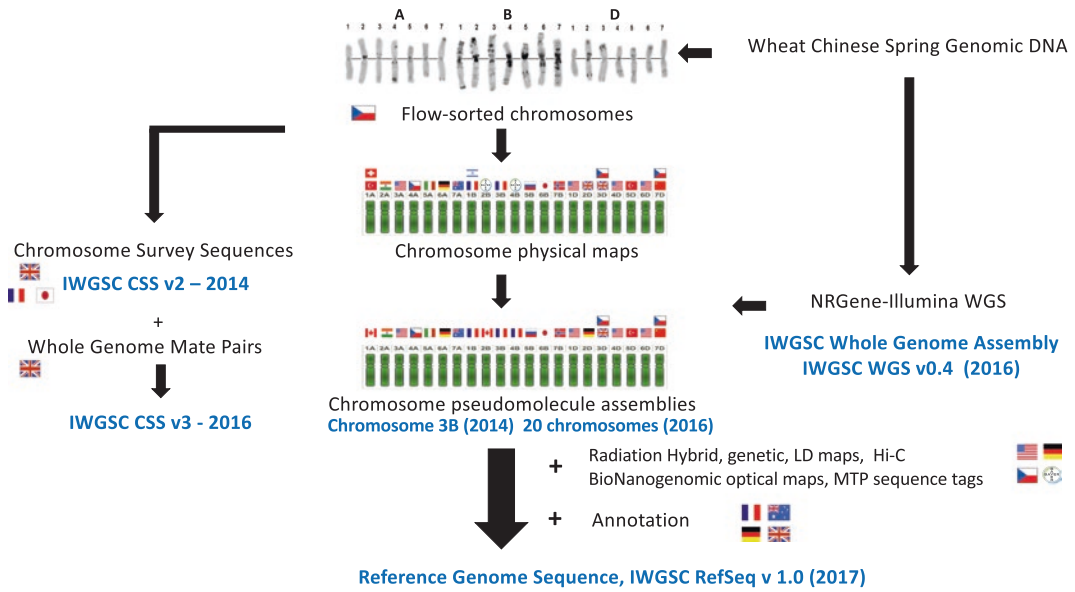


Fig. 1.1 Overview of the global community contributing to the sequencing of the wheat genome. National flags indicate the country-of-origin of the research groups contributing to the establishment of the high-quality *Triticum aestivum* cv. CHINESE SPRING reference genome assembly (IWGSC RefSeq v1.0) including involvement in the flow sorting, chromosome shotgun, generation of additional resources and annotation. The times for the data set releases are indicated in blue

the generation of chromosome-based short read sequence data into the strategy. The data provided the first genome-wide information about the distribution of genic sequences across the 21 chromosomes and provided an intermediate gene catalogue for wheat research (International Wheat Genome Sequencing Consortium 2014). Two further strategic modifications were made in 2014 and 2016, respectively. The first enabled the integration of the physical maps with genome-wide sequence data by generating short sequence tag data from minimal tile paths of BACs for chromosomes mapped using the SNaPShot approach (see International Wheat Genome Sequencing Consortium 2018). The final update to the IWGSC wheat genome roadmap reflected the breakthrough in sequence assembly software developed by NRGene (www.nrgene.com) and others (Clavijo et al. 2017) which made it possible to assemble a whole genome sequence of bread wheat. By integrating a whole genome shotgun assembly with data derived from chromosomal maps and genetic maps, the first reference genome sequence for hexaploid bread wheat was produced (Fig. 1.1).

1.5 Impact of Sequencing Technology Improvement on IWGSC Strategy

At the time of the USDA-NSF workshop, high throughput DNA sequencing was in a state of transition. Previously, the predominant sequencing platforms had been based on fluorescent dideoxy nucleotide sequencing (so-called Sanger sequencing) which delivered of the order of 350–1000 bases per sequence using automated gel-based or capillary separation systems. Driven by the human genome project and other large genome projects, between 1994 and 2004 the sequence accuracy and output rose to around 1 million bases per day per instrument, but the cost of sequencing remained relatively high at ca. 0.3 USD per sequence read (500 USD per raw Mb). The high cost and relatively slow pace of sequencing meant that even medium-sized

genomes (500 Mb–1 Gb) required large, multi-year projects to produce even draft versions of genomes with wildly differing quality, depending on the size and composition of repeat sequences.

Around 2004, the first second-generation sequencing instruments began to emerge. The first was the 454 Life Sciences pyrosequencer (later acquired by Roche Diagnostics) that measured sequential DNA polymerase catalysed sequencing reactions in picotiter plate arrays (Ronaghi et al. 1998; Margulies, et al. 2005). Early instruments generated around 100 million bases per day from ca. 0.5 million sequences of up to 100 nucleotides. The output improved with further development to approximately 400 million bases from sequences up to 400 nucleotides long in a 10-h run at a cost of around 15 USD per raw Mb by 2009. Whilst the 454 brought speed and cost benefits to high throughput sequencing, the accuracy was lower than 'Sanger sequencing', largely due to problems with accurate determination of bases in homopolymers (Metzker 2010; Mardis 2011). This could be accommodated and corrected to some extent by sequence analysis and assembly software, but it still caused some problems for some genome sequences.

The emergence of the highly parallelised pyrosequencing instrumentation of 454 Life Sciences led the way for more 'second-generation' platforms offering massively parallel sequencing. The most successful of these was developed by Solexa and subsequently commercialised by Illumina™. The platform uses 'sequencing by synthesis' to measure the incorporation of fluorescent nucleotides into millions of growing chains of DNA anchored to a glass surface which are scanned using a confocal microscope (Bennett et al. 2005). Initially, sequence read lengths were limited to around 30 bases, but as the technology matured improvements in chemistry, imaging technology and software have reduced the sequence ascertainment bias and enabled routine collection of paired sequence reads up to 300 bases long from sized DNA fragments. As a result, rates of data collection rose from 300 Mb to over 100 Gb per day with high levels of sequence accuracy (Schatz 2015) and reduced the costs compared to Sanger sequencing by 4–5 orders of magnitude. By assembling overlapping sequences from paired reads derived from small fragments (300–400 bp), longer sequences can be built up that help to overcome some of the problems encountered in using Illumina technology to sequence large or repetitive genomes. There has also been significant investment in developing data management and sequence assembly pipelines in both the public and private domains to meet the challenges of documenting and assembling very large volumes of short read sequence data (see Chap. 2). These benefits have resulted in the Illumina technology becoming the most widely used second-generation technology with a broad range of applications including de novo genome sequencing, comparative genomics, gene expression, transcriptomics, DNA–protein interactions and methylation profiling.

The earliest wheat genome-wide sequencing projects focused on genic sequences with the sequencing of expressed sequence tags (ESTs) and cDNAs. A set of 1,073,845 EST sequences derived from polyA-tailed transcripts were released by the Triticeae EST Cooperative in 1998 and used to produce a set of 40,000 Unigenes (http://www.ncbi.nlm.nih.gov/dbEST/dbESTsummary.html). In 2008, a Japanese initiative released 15,871 annotated cDNA sequences (http://trifldb.psc.riken.jp). Subsequently, relatively small studies of sequences from plasmids, from the 3B BAC library and from a gene-enriched methyl filtration library, were used to develop estimates of the gene and repeat contents of the genome based on 'Sanger' sequencing. Low sample sizes and sampling bias, however, produced widely ranging estimates of between 36,000 and 300,000 for gene number and a repeat content ranging from 68 to 86%.

The introduction of higher throughput new sequencing technologies facilitated the production of more extensive genome-wide data sets. In 2012, Brenchley et al. published the results of analysis of 85 Gb of sequence generated on the Roche 454 GS FLX Titanium and

GS FLX+platforms. Around 5 million scaffolds were assembled from 20 million sequence reads representing approximately fivefold coverage of the CHINESE SPRING wheat genome. Although the data were highly fragmented, they provided 132,000 SNPs for use in genotyping studies and estimates of the gene numbers at between 94,000 and 96,000 per sub-genome, with a repeat content of 79%.

In 2014, the IWGSC published the results of Illumina™ short read survey sequencing of chromosome 3B and the chromosome arms of the other 20 chromosomes of the wheat genome (IWGSC 2014). Based on between 30-fold and 240-fold depth of sequence reads, sequences with contig L50s ranging from 1.8 to 8.9 kb were assembled after removal of repetitive sequences that could not be assembled uniquely to give an estimated coverage of between 0.5 and 0.8 of each chromosome. From the sequence analysis, 124,000 gene models were allocated across the chromosome arms and ca. 75,000 were ordered using SNP genotyping and/or synteny with other grass genomes. Whilst most of the genes were incomplete and the data provided little or no information about gene duplications and pseudogenisation, nor the structural relationships between genes and repeat sequences, these analyses still provided the first genome-wide view of the distribution of wheat genes across homoeologous chromosomes. They also provided sets of chromosome-specific markers for gene selection and future genome-wide analyses.

In addition to genome surveys, the new sequencing technologies were used for high-quality sequencing. 454 sequencing technology was used to produce the first reference quality sequence of a wheat chromosome, 3B (Choulet et al. 2014). Sequences generated from 8452 MTP BAC clones in pools of ten BACs using 8 kb paired-end barcoded libraries were incorporated into an assembly of 833 Mb with a N50 for the sequence scaffolds of 892 kb (i.e. half of the chromosome sequence is represented by scaffolds greater than 892 kb). Using 2594 anchored SNP markers, 1358 sequence scaffolds comprising 774.4 Mb with a scaffold N50 of 949 kb were used to construct a pseudomolecule representing the 3B chromosome. Annotation of the chromosome with the automated Triannot pipeline (Leroy et al. 2012) identified and positioned 5326 functional genes and 1938 pseudogenes. It was also possible for the first time to annotate transposable elements and obtain a view of their distribution along the chromosome (Choulet et al. 2014).

Having established the principle of chromosomal MTP BAC sequencing for wheat, the sequencing of 3B was swiftly followed by projects for other chromosomes. By January 2015, MTP sequencing of 1A, 1B, 2B, 3A, 3D, 4A, 5B, 6B, 7A, 7B and 7D was underway in 11 countries, using predominantly Illumina™ sequencing to take advantage of higher throughput and lower costs relative to other sequencing platforms. A variety of strategies were employed to increase the contiguity of BAC sequences, which assembled into between 1 and 200 contigs per BAC, depending on the nature of the sequence, the quality and depth of the sequence data and the assembly software employed (see Chap. 3). Additional targeted efforts included combining sequence data from different fragment sizes (e.g. data from 500 bp to 1 kb fragments with paired-end sequences (mate pairs) from fragments between 1 and 10 kb), incorporation of long read sequence data generated on new platforms and comparison with BioNano Optical maps generated for individual BACs from flow-sorted chromosomes (see Chap. 3). Many of these efforts were ultimately superceded by the whole genome assembly, but much of the data has contributed to the refinement of the whole genome sequence to produce the first high-quality reference genome sequence for bread wheat.

1.6 Building the Reference Genome Sequence of Bread Wheat

One of the greatest challenges for genome sequencing is being confident that the sequence accurately represents the genome in coverage

and in organisation along the chromosomes. Chromosome 3B was the first wheat chromosome to achieve reference sequence quality and set a high standard for the rest of the genome. Representing more than 90% of the chromosome, the BAC sequence contigs and scaffolds were organised along the chromosome using additional information derived from integrating chromosomal Illumina shotgun data, BAC end sequences and information from the physical map and high density genetic maps.

As the second-generation short read sequencing technologies became established, the throughput and data quality improved and the overall cost of data generation declined. In other spheres, population genetics studies were beginning to be based on whole genome comparisons, prompting the development of new methods for the rapid assembly and comparative analysis of increasingly large and complex genomes. Whole genome assemblies of hexaploid bread wheat based on defined sets of paired sequences generated from the ends of sized DNA fragments were generated by Chapman et al. (2015) and Clavijo et al. (2017). These assemblies were greatly improved over previous assemblies covering 8.2 Gb and 13.4 Gb, with reported N50 contig sizes of 24.8 kb and 88.8 kb, respectively. The organisation of the assembled sequence contigs and scaffolds relied, as in the case of chromosome 3B on alignment to orthogonal genetic linkage maps. These were generated for wheat using the POPSEQ method enabled by high throughput sequencing and demonstrated initially in barley (Mascher et al. 2013; Chapman et al. 2015).

In 2016, the IWGSC released a whole genome assembly of Illumina short read sequence data assembled with DeNovoMAGIC2™, software developed by NRGene that assembles Illumina™ short reads into highly accurate long, phase sequences, even when the data are derived from highly repetitive genomes. The assembled sequences totalled 14.5 Gb and were assigned to chromosomal locations using POPSEQ data (Chapman et al. 2015) and a chromosome conformation capture (Hi-C) map constructed from Illumina sequence data produced from four independent Hi-C libraries. The assembly was released as IWGSC WGAv0.4. It represented over 90% of the genome and contains over 97% of known genes. Additional work was undertaken to integrate IWGSCv0.4 with chromosome-based physical maps, Whole Genome Profiling Tags generated from chromosomal BAC MTPs (van Oeveren et al. 2011), sequenced BACs and optical maps (available at the time for the Group 7 chromosomes). This resulted in the IWGSC Reference Sequence v1.0 released in January 2017 together with gene annotation based on extensive RNASeq data, annotations of transposable elements, duplicated regions and integration of molecular markers (IWGSC 2018).

The goal of the IWGSC wheat genome project was to produce an annotated reference genome sequence for wheat and make it available in the public domain to underpin wheat research and improvement. The release of IWGSC RefSeq v1 and the first analyses published in 2018 marked the culmination of the project and the beginning of the next chapter of wheat research. Throughout the genome project, verified sequence data sets were released through the IWGSC repository hosted at INRA, France, GrainGenes and the major public sequence data repositories hosted at EBI, NCBI and DDBJ (see Chap. 2). New insights have emerged about the structure of the genome and the distribution of features, including genes, repeat sequences and regulatory factors, together with information about temporal and spatial tissue-specific gene expression and regulation. The genome sequence has prompted the development of new tools for population studies to identify genomic features associated with specific traits. For example, genome-wide SNP assays and computational platforms for analysis are being developed together with tools for the assembly and comparative analyses of multiple genome sequences (Chap. 6; Walkowiak et al. 2020). The high quality of the sequence is also enabling targeted genetic manipulation work (see Chap. 10).

Whilst IWGSC RefSeq1 represented a highly contiguous genome sequence covering approximately 94% of the genome with contig,

scaffold and super-scaffold N50s of 52 kb, 7 Mb and 22.8 Mb, respectively, gaps remained. As new data becomes available, the sequence will be updated and improved. The first updated sequence, IWGSC Reference sequence v2.1 (Zhu et al. 2021) was based on alignments to optical maps, refined the reference genome to correct the orientation of some scaffolds as well as filling gaps in the genome sequence. With the improvement in so-called third-generation long read sequencing technologies, further updates to the reference genome sequence can be expected. In 2020, Alonge et al. used data from IWGSC RefSeq v1 to improve and annotate a sequence assembly generated from PacBio long read sequence data (Alonge et al. 2020). PacBio long read sequence data were also used to assemble the sequence of the bread wheat *Triticum aestivum* cultivar KARIEGA (Athiyannan et al. 2022), and Oxford Nanopore long read sequence data were used to assemble *Triticum aestivum* cultivar RENAN (Aury et al. 2021) to enable functional studies of these varieties.

The goal of the IWGSC was to produce a reference genome sequence for bread wheat that would enable wheat research and breeding improvements. IWGSC RefSeq v 1 has provided an excellent foundation that is shared by the international wheat community for future developments.

References

Adams MD, Celnicke SE, Holt RA, Evans CA et al (2000) The genome sequence of *Drosophila melanogaster*. Science 287:2185–2195. https://doi.org/10.1126/science.287.5461.2185

Alonge M, Shumata A, Pulu D, Zimin AV, Salzberg SL (2020) Chromosome-scale assembly of the bread wheat genome reveals thousands of additional gene copies. Genetics 216:599–608. https://doi.org/10.1534/genetics.120.303501

Athiyannan N, Abrouk M, Boshoff WHP, Cauet S, Rodde N, Kudrna D, Mohammed N, Bettgenhaeuser J, Botha K, Derman SS, Wing RA, Prins R, Krattinger SG (2022) Long-read genome sequencing of bread wheat facilitates disease resistance gene cloning. Nat Genet 54:227–231. https://doi.org/10.1038/241588-022-0102201

Aury J-M, Engelen S, Istace B, Monat C, Lasserre-Zuber P, Belser C, Cruaud C, Rimbert H, Leroy P, Arribat S, Dufau I, Bellec A, Grimbichler D, Papon N, Paux E, Ranoux M, Alberti A, Wincker P, Choulet F (2021) Long-read and chromosome-scale assembly of the hexaploidy wheat genome achieves higher resolution for research and breeding. bioRxiv preprint. https://doi.org/10.1101/2021.08.24.457458

Avni R, Nave M, Barad O, Baruch K, Twardziok SO, Gundlach H, Hale I, Mascher M, Spannagl M, Wiebe K, Jordan KW, Golan G, Deek J, Ben-Zvi B, Ben-Zvi G, Himmelbach A, MacLachlan RP, Sharpe AG, Fritz A, Ben-David R, Budak H, Fahima T, Korol A, Faris JD, Hernandez A, Mikel MA, Levy AA, Steffenson B, Maccaferri M, Tuberosa R, Cattivelli L, Faccioli P, Ceriotti A, Kashkush K, Pourkheirandish M, Komatsuda T, Eilam T, Sela H, Sharon A, Ohad N, Chamovitz DA, Mayer KFX, Stein N, Ronen G, Peleg Z, Pozniak CJ, Akhunov ED, Distelfeld A (2017) Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. Science 357(6346):93–97. https://doi.org/10.1126/science.aan0032. PMID: 28684525

Bennett ST, Barnes C, Cox A, Davies L, Brown C (2005) Toward the $1000 human genome. Pharmacogenomics 6:373–382

Bonjean A (2016) The saga of wheat—the successful story of wheat and human interaction. In: Bonjean A et al (eds) The world wheat book: a history of wheat breeding, vol 3. Lavoisier, Paris, pp 1–90

Brenchley R, Spannagl M, Pfeifer M, Barker GLA, D'Amore R, Allen AM, McKenzie N, Kramer M, Kerhornou A, Bolser D et al (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. Nature 491:705–710

Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J (2013) Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat Biotechnol 31(12):1119–1125. https://doi.org/10.1038/nbt.2727

Celniker SE, Rubin GM (2003) The *Drosophila melanogaster* genome. Annu Rev Genomics Hum Genet 4(1):89–117

Chandler VA, Brender V (2002) The maize genome sequencing project. Plant Physiol 130(4):1594–1597. https://doi.org/10.1104/pp.015594

Chapman JA, Mascher M, Buluç A, Barry K, Georganas E, Session A, Stradova V, Jenkins J, Sehgal S, Oliker L, Scmutz J, Yelick K, Scholz U, Waugh R, Poland J, Muehlbauer G, Stein N, Rokhsar D (2015) A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome. Genome Biol 16:26. https://doi.org/10.1186/s13059-015-0582-8

Choulet F, Alberti A, Theil S, Glover N, Barbe V, Daron J, Pingault L, Sourdille P, Couloux A, Paux E et al (2014). Structural and functional partitioning of bread wheat chromosome 3B. Science 345

Clavijo BJ, Venturini L, Schudoma C, Accinelli GG, Kaithakottil G, Wright J, Borrill P, Kettleborough G, Heavens D, Chapman H, Lipscombe J, Barker T,

Lu FH, McKenzie N, Raats D, Ramirez-Gonzalez RH, Coince A, Peel N, Percival-Alwyn L, Duncan O, Trösch J, Yu G, Bolser DM, Namaati G, Kerhornou A, Spannagl M, Gundlach H, Haberer G, Davey RP, Fosker C, Palma FD, Phillips AL, Millar AH, Kersey PJ, Uauy C, Krasileva KV, Swarbreck D, Bevan MW, Clark MD (2017) An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. Genome Res 27(5):885–896. https://doi.org/10.1101/gr.217117.116.PMID:28420692;PMCID:PMC5411782

De Coster W, Weissensteiner MH, Sedlazeck FJ (2021) Towards population-scale long-read sequencing. Nat Rev Genet 22:572–587. https://doi.org/10.1038/s41576-021-00367-3

Doležel J, Kubaláková M, Paux E, Bartoš J, Feuillet C (2007) Chromosome-based genomics in cereals. Chromosome Res 15

Eckhardt NA (2000) Sequencing the rice genome. Plant Cell 12(11):2011–2018. https://doi.org/10.1105/tpc.12,11.2011

Feldman M, Levy AA (2009) Genome evolution in allopolyploid wheat—a revolutionary reprogramming followed by gradual changes. J Genet Genomics 36:511–518

Feuillet C, Leach JE, Rogers J, Schnable PS, Eversole K (2011) Crop genome sequencing: lessons and rationales. Trends Plant Sci 16:77–88

Feuillet C, Rogers J, Eversole K (2016) Progress towards achieving a reference genome sequence to accelerate the selection of improved wheat varieties. In: Bonjean A et al (eds) The world wheat book: a history of wheat breeding, vol 3. Lavoisier, Paris, pp 965–999

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM et al (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269(5223):496–512. https://doi.org/10.1126/science.7542800

Gill BS, Appels R, Botha-Oberholster A-M et al (2004) A workshop report on wheat genome sequencing: international genome research on wheat consortium. Genetics 168:1087–1096. https://doi.org/10.1534/genetics.104.034769

Howe K, Clark M, Torroja C et al (2013) The zebrafish reference genome sequence and its relationship to the human genome. Nature 496:498–503. https://doi.org/10.1038/nature12111

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431:931–945

International Rice Genome Sequencing Project and Sasaki (2005) The map-based sequence of the rice genome. Nature 436:793–800. https://doi.org/10.1038/nature03895

International Wheat Genome Sequencing Consortium (IWGSC) (2014) A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. Science 345(6194):1251788. https://doi.org/10.1126/science.1251788. PMID: 25035500

International Wheat Genome Sequencing Consortium (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science 361(6403). https://doi.org/10.1126/science.aar7191

Jackson SA, Iwata A, Lee S-H, Schmutz J, Shoemaker R (2011) Sequencing crop genomes: approaches and applications. New Phytol 191:915–925

Janda J, Bartoš J, Šafář J, Kubaláková M, Valárik M, Číhalíková J, Šimková H, Caboche M, Sourdille P, Bernard M et al (2004) Construction of a subgenomic BAC library specific for chromosomes 1D, 4D and 6D of hexaploid wheat. Theor Appl Genet 109

Jia J, Zhao S, Kong X, Li Y, Zhao G, He W, Appels R, Pfeifer M, Tao Y, Zhang X et al (2013) *Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation. Nature 496:91–95

Keeble-Gagnère G, Rigault P, Tibbits J, Pasam R, Hayden M, Forrest K, Frenkel Z, Korol A, Huang BE, Cavanagh C, Taylor J, Abrouk M, Sharpe A, Konkin D, Sourdille P, Darrier B, Choulet F, Bernard A, Rochfort S, Dimech A, Watson-Haigh N, Baumann U, Eckermann P, Fleury D, Juhasz A, Boisvert S, Nolin MA, Doležel J, Šimková H, Toegelová H, Šafář J, Luo MC, Câmara F, Pfeifer M, Isdale D, Nyström-Persson J, Iwgsc, Koo DH, Tinning M, Cui D, Ru Z, Appels R (2018) Optical and physical mapping with local finishing enables megabase-scale resolution of agronomically important regions in the wheat genome. Genome Biol 19(1):112. https://doi.org/10.1186/s13059-018-1475-4

Kihara H (1944) Discovery of the DD analyser, one of the ancestors of *T. vulgare*. Agric Hortic 19:889–890

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

Leroy P, Guilhot N, Sakai H, Bernard A, Choulet F, Theil S, Reboux S, Amano N, Flutre T, Pelegrin C et al (2012) TriAnnot: a versatile and high performance pipeline for the automated annotation of plant genomes. Front Plant Sci 3

Ling H-Q, Zhao S, Liu D, Wang J, Sun H, Zhang C, Fan H, Li D, Dong L, Tao Y et al (2013) Draft genome of the wheat A-genome progenitor *Triticum urartu*. Nature 496:87–90

Ling HQ, Ma B, Shi X et al (2018) Genome sequence of the progenitor of wheat A subgenome *Triticum urartu*. Nature 557:424–428. https://doi.org/10.1038/s41586-018-0108-0

Mardis ER (2011) A decade's perspective on DNA sequencing technology. Nature 470:198–203

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380

Mascher M, Muehlbauer GJ, Rokhsar DS, Chapman J, Schmutz J, Barry K, Muñoz-Amatriaín M, Close TJ, Wise RP, Schulman AH et al (2013) Anchoring and ordering NGS contig assemblies by population sequencing (POPSEQ). Plant J 76:718–727

McFadden ES, Sears ER (1946) The origin of *Triticum spelta* and its free-threshing hexaploid relatives: hybrids of synthetic *T. spelta* with cultivated hexaploids. J Hered 37:107–116

Metzker ML (2010) Sequencing technologies—the next generation. Nat Rev Genet 11:31–46

Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562. https://doi.org/10.1038/nature01262

Paux E, Sourdille P, Salse J, Saintenac C, Choulet F, Leroy P, Korol A, Michalak M, Kianian S, Spielmeyer W et al (2008) A physical map of the 1-gigabase bread wheat chromosome 3B. Science 322:101–104

Ronaghi M, Uhlén M, Nyrén P (1998) A sequencing method based on real-time pyrophosphate. Science 281:363–365

Šafář J, Bartoš J, Janda J, Bellec A, Kubaláková M, Valárik M, Pateyron S, Weiserová J, Tušková R, Číhalíková J et al (2004) Dissecting large and complex genomes: flow sorting and BAC cloning of individual chromosomes from bread wheat. Plant J 39:960–968

Šafář J,·Šimková H,·Kubaláková M, Číhalíková J, Suchánková P, Bartoš J, Doležel J (2010) Development of chromosome-specific BAC resources for genomics of bread wheat. Cytogenet Genome Res 129:211–223.https://doi.org/10.1159/000313072

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74:5463–5467

Schatz MC (2015) Biological data sciences in genome research. Genome Res 25:1417–1422

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA et al (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112–1115

Sears ER (1954) The aneuploids of common wheat. Mo Agr Exp Sta Res Bulletin 572:1–58

Sears ER (1966) Chromosome mapping with the aid of telocentrics. Hereditas 2(Supplement):370–381

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815

The Tomato Genome Consortium (2012) The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485:635–641. https://doi.org/10.1038/nature11119

van Oeveren J, de Ruiter M, Jesse T, van der Poel H, Tang J, Yalcin F, Janssen A, Volpin H, Stormo KE, Bogden R, van Eijk MJ, Prins M (2011) Sequence-based physical mapping of complex genomes by whole genome profiling. Genome Res:618–625. https://doi.org/10.1101/gr.112094.110

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA et al (2001) The sequence of the human genome. Science 291:1304–1351

Walkowiak S, Gao L, Monat C et al (2020) Multiple wheat genomes reveal global variation in modern breeding. Nature 588:277–283. https://doi.org/10.1038/s41586-020-2961-x

Young N, Debellé F, Oldroyd G et al (2011) (2011) The *Medicago* genome provides insight into the evolution of rhizobial symbioses. Nature 480:520–524. https://doi.org/10.1038/nature10625

Zhu T, Wang l, Rimbert H, Rodriguez J, Deal K, De Oliveira R, Choulet F, Keeble-Gagnère G, Tibbitts J, Rogers J, Eversole K, Appels R, Gu Y, Mascher N, Dvorak J, Luo, MC (2021) Optical maps refine the bread wheat *Triticum aestivum* cv. CHINESE SPRING genome assembly. Plant J 107:303–314.https://doi.org/10.1111/tpj.15289

Zohary D, Weiss E, Hopf M (2012) Domestication of plants in the old world. OUP, Oxford

2 Wheat Data Integration and FAIRification: IWGSC, GrainGenes, Ensembl and Other Data Repositories

Michael Alaux, Sarah Dyer and Taner Z. Sen

Abstract

Wheat data integration and FAIRification are key to tackling the challenge of wheat improvement. The data repositories presented in this chapter play a central role in generating knowledge and allow data exchange and reuse. These repositories rely on international initiatives such as (i) the International Wheat Genome Sequencing Consortium (IWGSC), which delivers common genomics resources such as reference sequences, communal Web-based seminars and (ii) the Wheat Information System (WheatIS) of the Wheat Initiative (http://www.wheatis.org), which improves the interoperability and findability of the wheat data across the repositories.

Keywords

Wheat data · Data repositories · IWGSC · GrainGenes · Ensembl · FAIR

M. Alaux (✉)
Université Paris-Saclay, INRAE, URGI, 78026 Versailles, France
e-mail: michael.alaux@inrae.fr

Université Paris-Saclay, INRAE, BioinfOmics, Plant Bioinformatics Facility, 78026 Versailles, France

S. Dyer
European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK
e-mail: sdyer@ebi.ac.uk

T. Z. Sen
Western Regional Research Center, Crop Improvement and Genetics Research Unit, United States Department of Agriculture-Agricultural Research Service, Albany, CA, USA
e-mail: taner.sen@usda.gov

University of California, Department of Bioengineering, Berkeley, CA, USA

2.1 Introduction

According to the Food and Agriculture Organisation (FAO), wheat is the most widely cultivated crop on Earth, contributing about a fifth of the total calories consumed by humans (https://www.fao.org/faostat/en/#data). To meet the challenge of delivering safe, high-quality and health-promoting food and feed in an environmentally sensitive, economical and sustainable manner, it is generally considered that wheat improvement needs molecular breeding to complement more standard approaches. Furthermore, the efforts of breeding happen in a context of climate change but are still limited by insufficient knowledge and understanding of the molecular basis of central agronomic traits. In order to address the scientific questions related to this challenge, the wheat research community generates large and

R. Appels et al. (eds.), *The Wheat Genome*, Compendium of Plant Genomes,
https://doi.org/10.1007/978-3-031-38294-9_2

heterogeneous datasets. The greatest value of these data lies in their integration to generate new knowledge as a result of effective sharing to allow transparency and openness.

The wheat data landscape relies on repositories centred on (i) one or multiple data types (such as genomics, genetics or phenomics) that are highly curated and integrated with a common reference genome (e.g. the accession CHINESE SPRING developed by the IWGSC, 2018), (ii) projects or community of users with dedicated tools to mine the data. To improve the FAIRness (Findable, Interoperable, Accessible, and Reusable, Wilkinson et al. 2016a) of the wheat datasets and databases, the WheatIS expert working group of the Wheat Initiative recommended standards and developed a data discovery tool dedicated to improve the findability of wheat data across repositories (Dzale Yeumo et al. 2017; Sen et al. 2020).

In this chapter, we describe major wheat data repositories and tools, and how they integrate different types of wheat data following the FAIR principles.

2.2 IWGSC Data Repository

The International Wheat Genome Sequencing Consortium (IWGSC) has developed a variety of resources for bread wheat (*Triticum aestivum* L.) through its long-term efforts to achieve a high quality and functionally annotated reference wheat genome sequence (accession CHINESE SPRING). These data are available in a dedicated IWGSC data repository (https://wheat-urgi.versailles.inrae.fr/Seq-Repository, Alaux et al. 2018) categorised by data type as shown in Fig. 2.1.

2.2.1 Sequence Assemblies and Annotations

IWGSC wheat genome sequence assemblies are available for download, BLAST (Altschul et al. 1990), and display in genome browsers. The assembly dataset includes the draft and the reference sequences, along with their annotations.

The draft survey sequence assembly (IWGSC Chromosome Survey Sequence (CSS) v1, IWGSC 2014) and the chromosome 3B reference sequence (the first reference quality chromosome sequence obtained by the consortium, Choulet et al. 2014) were released in 2014, followed by two improved versions of the CSS (v2 and v3). The virtual gene order map generated for the CSS, the POPSEQ data were used to order sequence contigs on chromosomes (Mascher et al. 2013), and mapped marker sets were associated with these assemblies.

The reference sequence of the bread wheat genome released in 2018 (IWGSC RefSeq v1.0, 2018) included the whole genome, pseudomolecules of individual chromosomes or chromosome arms, scaffolds with the structural and functional annotation of genes, transposable elements (TEs) and non-coding RNAs. In addition, mapped markers as well as annotations supported with alignments of nucleic acid and protein evidence were made available. Manual annotations for specific gene families or regions of specific chromosomes (ca. 3685 genes) were included in the IWGSC RefSeq v1.1 annotations. This v1.1 annotation set was updated to v1.2 by integrating a set of 117 novel genes and 81 microRNAs manually curated by the wheat community following guidelines provided by IWGSC.

The improved version IWGSC RefSeq v2.1 assembly was released in 2021 (Zhu et al. 2021), which relied on whole-genome optical maps and contigs assembled from whole-genome-shotgun Pacific Biosciences (PacBio) reads (Zimin et al. 2017). Optical maps were used to detect and resolve chimeric scaffolds, anchor unassigned scaffolds, correct ambiguities in positions and orientations of scaffolds, create super-scaffolds and estimate gap sizes more accurately. PacBio contigs were used for gap closing, and pseudomolecules of the 21 CHINESE SPRING chromosomes were re-constructed to develop this new reference sequence. The corresponding IWGSC v2.1 annotation accompanying the IWGSC RefSeq v2.1 assembly was also completed. The transposable elements (TEs) in the resulting assembly IWGSC RefSeq v2.1 were

Fig. 2.1 Homepage of the IWGSC data repository hosted by the Wheat@URGI portal [Retrieved in August 2023]

reannotated, and gene annotations were updated by transferring the known gene models from previous annotations using a fine-tuned, dedicated strategy implemented in the Marker-Assisted Gene Annotation Transfer for *Triticeae* pipeline (https://forgemia.inra.fr/umr-gdec/magatt). The released IWGSC Annotation v2.1 contains 266,753 genes comprising 106,913 high-confidence genes and 159,840 low-confidence genes (Zhu et al. 2021).

In addition to the bread wheat reference sequence, the IWGSC also sequenced the genome of the Turkish bread wheat elite cultivar SONMEZ (Nelson et al. 2005) along with seven diploid and tetraploid species: *Triticum durum* cv. CAPPELLI, *Triticum durum* cv. STRONGFIELD, *Triticum durum* cv. SVEVO, *Triticum monococcum*, *Triticum urartu*, *Aegilops speltoides* and *Aegilops sharonensis* (IWGSC 2014). Download and BLAST services are available for these datasets at https://wheat-urgi.versailles.inrae.fr/Seq-Repository/Assemblies.

More broadly, the IWGSC is responsible for organising workshops and seminars and making genomics tools available to the community (https://www.wheatgenome.org/) as shown in Fig. 2.2. For example, the Apollo portal from national Australian Research Data Commons (https://apollo-portal.genome.edu.au/) has been set up to allow the curation of the IWGSC v2.1 annotation.

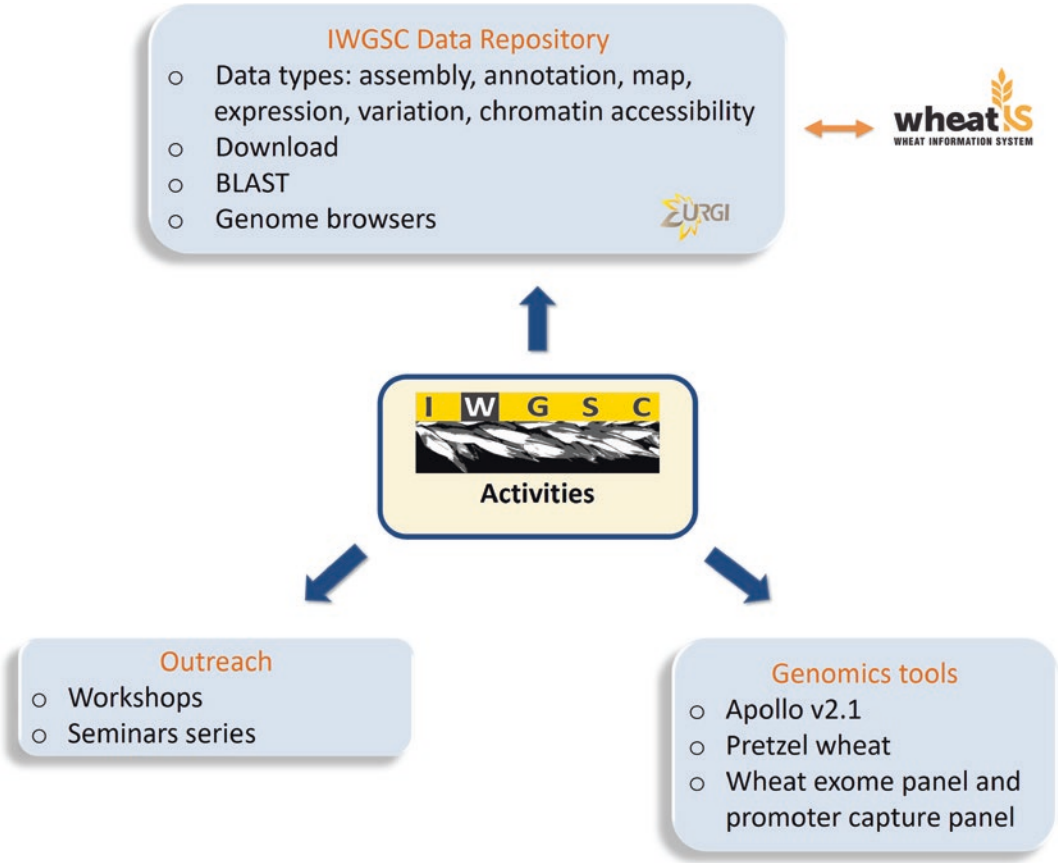


Fig. 2.2 Summary of IWGSC activities

2.2.2 Physical Maps and BAC Libraries

Physical maps of the 21 bread wheat chromosomes, based on high information content fluorescence fingerprinting (Nelson et al. 2005) or whole-genome profiling (Philippe et al. 2012) of flow-sorted chromosome or chromosome-arm specific BAC libraries, are stored and displayed in a dedicated browser. The BAC clone assemblies were produced by IWGSC members using fingerprinted contigs (Soderlund et al. 2000) or LTC (Frenkel et al. 2010) software. The positions of individual BAC clones, markers and deletion bins were mapped onto physical contigs. The wheat physical map browser also provides a link to request the BAC clones from the French plant genomic resource centre.

2.2.3 Expression Data

RNA-Seq expression data are available as read counts and transcripts per kilobase million mapped reads for the IWGSC RefSeq v1.1 annotation. A transcriptome atlas developed from 850 RNA-Seq datasets representing 32 tissues at different growth stages and stresses were mapped to the IWGSC RefSeq annotations v1.0 and v1.1 (Ramírez-González et al. 2018).

2.2.4 Variation Data

These datasets consist of the 1000 wheat exome project (He et al. 2019), whole exome capture and genotyping-by-sequencing approaches of 62 diverse wheat lines (Jordan et al. 2015)

and varietal and intervarietal SNPs (Rimbert et al. 2018). VCF data files are downloadable, and the variant calls can be displayed in the browser (https://wheat-urgi.versailles.inrae.fr/Seq-Repository/Variations).

2.2.5 Chromatin Accessibility

Using a differential nuclease sensitivity assay, the chromatin states were investigated in the coding and TE-rich repetitive regions of the allopolyploid wheat genome. Micrococcal nuclease (MNase) scores in BigWig format for IWGSC RefSeq v1.0 assembly are available to download (Jordan et al. 2020).

2.3 Wheat@URGI

The Wheat@URGI portal, developed by INRAE (French National Research Institute for Agriculture, Food and Environment) URGI unit, hosts the IWGSC data repository and GnpIS, a dedicated information system following the Findable Accessible Interoperable Reusable (FAIR) principles: https://wheat-urgi.versailles.inrae.fr/ (Alaux et al. 2018; Pommier et al. 2019).

GnpIS encompasses a set of integrated databases to manage genomic data using well-known tools such as BLAST, JBrowse, GBrowse and InterMine. An in-house database called GnpIS-coreDB developed by URGI to manage genetic and phenomic plant data, especially wheat, has been produced from French, European and international projects since 2000. A significant amount of this data is available as open access, and some project-restricted data can be obtained through a material transfer agreement.

Data managed by GnpIS-coreDB include: genetic information (markers, quantitative trait loci (QTLs), germplasm, genome-wide association studies (GWAS), genomic information (SNP discovery experiments, genotyping and synteny) and phenomic data. The phenomic data are available as whole trials including phenotypic and environmental observations on well-identified plant material provided by reference sources such as European genebanks. Detailed descriptions of these datasets are available in Alaux et al. (2018) and Pommier et al. (2019), and Table 2.1 presents a data summary.

The genetic and phenomic data have been produced from large collaborative projects such as BreedWheat (Paux et al. 2022) and Whealbi (Pont et al. 2019).

These different types of data are linked within the GnpIS information system. This integration is organised around key data, also called "pivot data" as they are pivotal objects which allow integration between data types. The key objects used to link genomic resources to genetic data are markers and traits. Markers are mapped to the genome sequences and provide information on neighbour genes and their function. They also have links to GnpIS-coreDB genetic maps, QTLs, genotyping and GWAS data. Traits link the genetic data to the phenomic data in GnpIS-coreDB and to synteny data displayed by the PlantSyntenyViewer tool (Flores et al. 2023; Pont et al. 2013).

The FAIRness of these data (including metadata) can be summarised as follows:

- Findability: (i) the data are searchable using our data discovery tools (WheatIS data discovery and FAIDARE, see below), Web interfaces (genome browsers), analysis tool (BLAST), data mining tool (WheatMine); (ii) digital object identifiers (DOIs) were generated for each accession.
- Accessibility: phenotyping data are accessible through Breeding API (BrAPI) Web services (Selby et al. 2019) and file downloads.
- Interoperability: the data are in standard formats (gff3, VCF, MCPD, MIAPPE, Papoutsoglou et al. 2020), and phenotyping data follow an ontology developed within the BreedWheat project and merged with the international wheat crop ontology (CO_321, Shrestha et al. 2012).
- Reusability: (i) all the GnpIS tools have general terms of use and licence. Open access data including code are in CC BY 4.0; (ii) the data are sufficiently described to allow their reuse in new analysis.

Table 2.1 Genetic and phenomic wheat data summary hosted in the GnpIS-coreDB database of the Wheat@URGI portal in August 2023

Data type		Total number of data points	Open access	Restricted access
Germplasm	Taxon	56	56	0
	Accession	15,031	10,448	4583
Genetic map	Map	30	29	1
	Marker	716,745	314,390	402,355
	QTL	749	465	284
Genotyping	Experiment	23	1	22
	Sample	9556	42	9543
	Marker	680,463	0	680,463
	SNP discovery	724,132	280,321	443,811
Phenotyping	Trial	895	833	62
	Seed lot	8461	5037	3653
	Variable	405	107	301
	Observation	1,488,199	602,553	885,646
GWAS	Analysis	2013	43	1970
	Sample	3096	2361	735
	Variable	313	37	279
	Marker	160,774	4109	156,665
	Association	1,014,694	48,596	966,098

2.4 GrainGenes

The GrainGenes repository (https://wheat.pw.usda.gov, Yao et al. 2022, Fig. 2.3) is a digital platform and a community service provider that has been continuously supported by U.S. congressional funds since 1992 through the U.S. Department of Agriculture. Its stakeholders are primarily global small grain research communities who work on wheat, barley, rye and oat (Blake et al. 2022). Unlike many other small grain repositories, GrainGenes has decades-worth of genetic data: GrainGenes contains rich, peer-reviewed, curated data content (Odell et al. 2017), ranging from genetic to genomic, phenotypic to traits, and people to publications, with a myriad of search and visualisation tools to enhance data findability and information discovery. GrainGenes also provides services, such as the GrainGenes email list and a Twitter feed, for small grain communities through communicating community announcements, open positions, upcoming conference information and grant opportunities.

The range of genome browsers at GrainGenes for wheat-related species attest the data growth as a result of increasingly accessible sequencing platforms, advanced assembly algorithms and annotation pipelines (https://wheat.pw.usda.gov/GG3/genome_browser). GrainGenes, in addition to IWGSC's CHINESE SPRING v1 and v2 assemblies, houses assemblies and annotations for *Aegilops longissima*, *A. speltoides*, *A. sharonensis*, five *Aegilops tauschii* accessions, wild emmer (ZAVITAN) and durum wheat SVEVO, as well as *Triticum aestivum* genomes from the 10+Wheat Genome project and the hexaploid wheat pangenome. The genome browsers at GrainGenes are shared with the Triticeae Toolbox (T3) database for the benefit of small grain researchers.

In its IWGSC CHINESE SPRING v1 genome browser, GrainGenes has many tracks overlapped with the IWGSC's data depository at Wheat@URGI and Ensembl Plants. In addition to those tracks, T3 created several tracks for variants, genome-wide association studies (GWAS), primers and quantitative trait loci (QTLs). The GrainGenes team created the guanine-quadruplex (G4) track, for this newly emergent transcription regulation element class (Cagirici and Sen 2020).

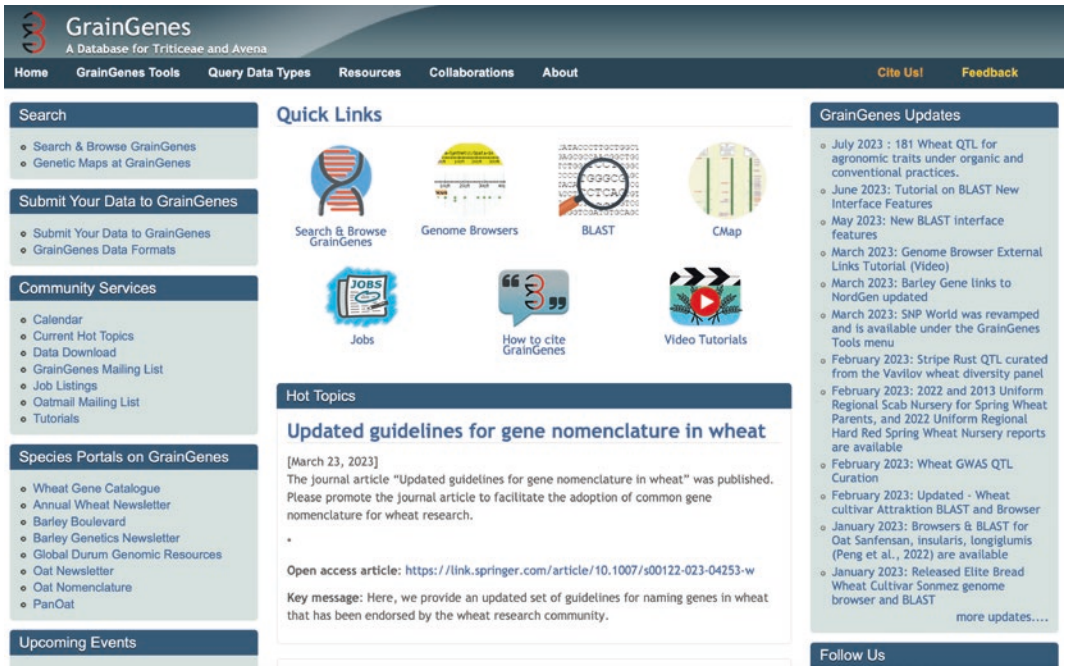


Fig. 2.3 Homepage of GrainGenes (https://wheat.pw.usda.gov) [Retrieved in July 2022]

Some of GrainGenes' genome browsers overlap with the genome browsers at other repositories such as Wheat@URGI or Ensembl Plants. This duplication of displays is not in excess, but ultimately serve the interest of small grains researchers, because having the same datasets at multiple repositories allows users to harness different tools built on top of these datasets, for example, BLAST services at GrainGenes (https://wheat.pw.usda.gov/blast/) or the Ensembl Variant Effect Predictor at Ensembl Plants.

One of the added values of using genome browsers at GrainGenes is their integration with the BLAST service at GrainGenes. When users run their nucleotide/protein sequences at GrainGenes, the results are linked to hit regions in the browsers, which allow users to go to those regions with a single mouse click. GrainGenes also uniquely allows rubber banding selection of a genome region on its JBrowse-based browsers for automatic copy pasting of underlying sequence data for subsequent BLASTing.

Those who are not familiar with genome browser operations and their relationship to other pages at GrainGenes can benefit from the several YouTube tutorial videos that were created by the GrainGenes team. This is especially useful for those who would like to learn how to jump from genomic to reach genetic data, and vice versa in GrainGenes. The videos are linked at https://wheat.pw.usda.gov/GG3/tutorials.

2.5 Ensembl Plants

The Ensembl Plants platform (https://plants.ensembl.org) provides a Web browser, databases, tools and programmatic access to integrated public genomic data for a breadth of plant species (Cunningham et al. 2022, Fig. 2.4). Ensembl Plants imports genomes and community gene annotations into the platform, annotates genomic repeat regions, imports variation data and identifies homologues via Ensembl's comparative genomics analysis pipeline. Users can access bioinformatics tools such as BLAST (Altschul et al. 1990) for sequence similarity searching or the Ensembl Variant Effect

Predictor (VEP, McLaren et al. 2016) to predict the functional consequences of variants.

The first version of the IWGSC Chromosome Survey Sequence (CSS) and gene annotation for the cultivar CHINESE SPRING was made available in Ensembl Plants in 2014. At that time there were three other triticeae genomes also included: *A. tauschii*, *Hordeum vulgare* and *T. urartu*. The TGACv1 whole-genome assembly (Clavijo et al. 2017) which used the CSS reads to assign scaffolds to chromosome arms became available via Ensembl Plants in 2015 and was subsequently replaced by the release of IWGSC RefSeq v1.0 in 2018, although all assemblies can still be accessed via Ensembl's archive sites. As of April 2023, Ensembl Plants contains an additional 17 bread wheat cultivar genomes from the 10+ project (Walkowiak et al. 2020, https://www.wheatinitiative.org/10-wheat-genome-project), making 26 triticeae genomes in total. Each of the bread wheat cultivars displays the annotation from IWGSC RefSeq v1.1 projected onto the cultivar assembly. In addition, de novo genes predicted by the Plant Genome and Systems Biology Group (PGSB) at Helmholtz, Munich and the Earlham Institute (EI) for the nine chromosome-level assemblies are also displayed.

In addition to genome annotations, Ensembl Plants also displays variation data, primarily from the 35 K and 820 K Axiom SNP breeders array, as provided by CerealsDB (Wilkinson et al. 2016b) and also EMS mutations mapped from the EMS TILLing populations (Krasileva et al. 2017) maintained by JIC's SeedStor (https://www.seedstor.ac.uk) for CADENZA (hexaploid bread wheat) and KRONOS (tetraploid durum wheat). This allows users to visualise where variants are located with respect to the IWGSC genome, and where those variants occur in the proximity of gene models the Ensembl Variant Effect Predictor will provide estimates of the likely impact of those variants on predicted gene and protein sequences. This helps users to identify those variants most likely to cause disruption to genes, and Ensembl Plants also provides a route to connect to SeedStor to order materials from the EMS populations which have those variants.

Ensembl's comparative genomics pipelines (Cunningham et al. 2019) provide gene/protein trees based on sequence homology and whole-genome alignments (WGA) between the majority of species within the platform. The IWGSC v1.0 assembly has gene trees and WGA available which allow users to explore gene family loss and expansions, identifying orthologues and paralogues and regions of synteny between genomes in Ensembl Plants. The 10+ wheat cultivars have wheat-specific gene trees available which provide a mechanism for users to explore gene conservation within the current bread wheat pan-genome (Fig. 2.5).

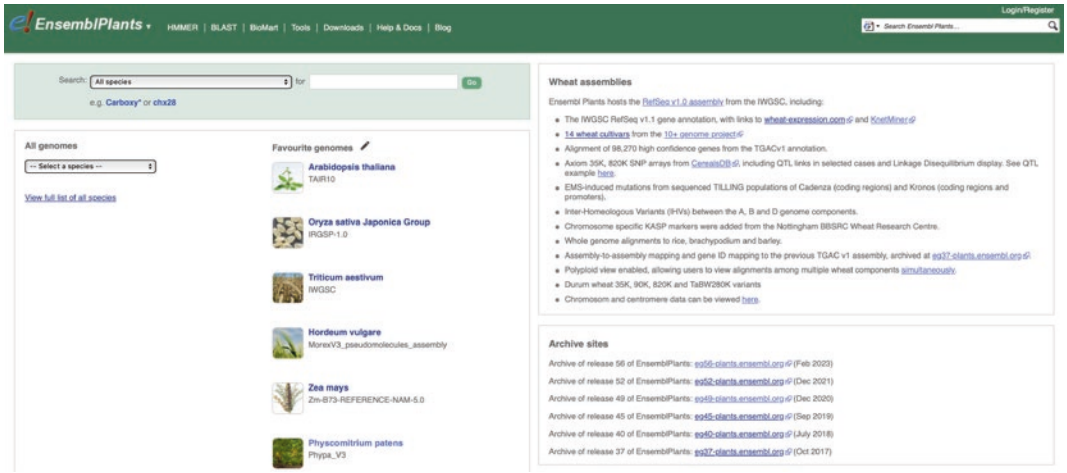


Fig. 2.4 Homepage of Ensembl Plants (https://plants.ensembl.org) [Retrieved in August 2023]

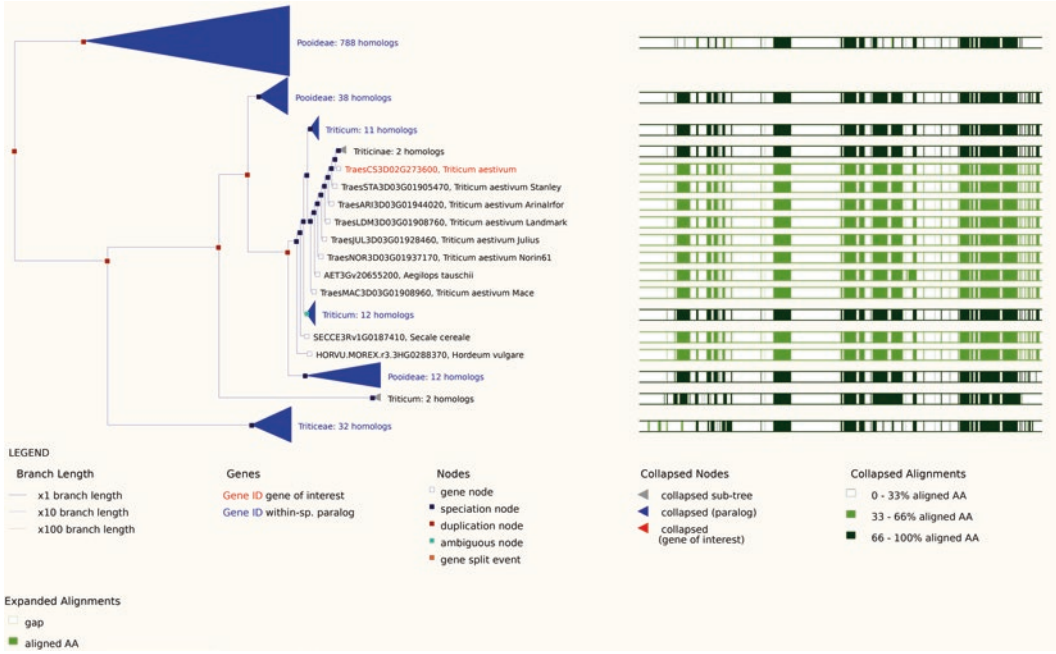


Fig. 2.5 Cultivar comparative gene tree for gene *TraesCS3D02G273600*, a heat shock protein located on chromosome 3D in IWGSC CHINESE SPRING v1.0, shown in red [Retrieved in September 2022]

Ensembl (Cunningham et al. 2022) provides user access via Web-based searches through the Ensembl browser or BioMart (which allows structured user querying to select subsets of data), FTP download access to complete sets of sequence data, annotations, gene trees and databases and programmatic access via Ensembl's APIs. All Ensembl data and tools are open access and freely available, and extensive documentation, training materials (https://training.ensembl.org) and a helpdesk are available to support user access. Ensembl Plants can also be accessed through the Gramene resource (https://www.gramene.org, Tello-Ruiz et al. 2022).

2.6 Some Other Repositories

It is beyond the purview of this chapter to provide all available wheat repositories worldwide, but the following are extremely valuable sites that we will mention briefly. Reading the publications for these repositories will be useful to learn more about their data content and features.

The Triticeae Toolbox (T3) (https://wheat.triticeaetoolbox.org, Blake et al. 2016). T3's mission is to create tools for researchers that work on genotypic–phenotypic relationships. As such, T3 played a centralised role in past projects with a strong breeding focus, such as Triticeae Coordinated Agricultural Project (TCAP) in the past, and, currently, in the Wheat Coordinated Agricultural Project (WheatCAP), both funded by the U.S. Department of Agriculture, National Institute of Food and Agriculture.

T3 houses many Web-based tools for breeders. It has capabilities that allow users to (1) upload their raw genome-wide association (GWAS) and genotype-by-sequencing datasets onto the Website, (2) perform computations such as principal component analyses and (3) visualise histograms for phenotypic observations, screen-plots of principal component eigenvalues, Q–Q plots displaying observed and expected -$\log_{10}$ *p*-values and Manhattan plots. In addition, T3 provides Web-based tools to generate selection indexes for multiple

traits simultaneously, which is a useful method for breeding programs to select and advance germplasms. As mentioned in the previous section, T3 has a very close collaboration with GrainGenes. Both databases maintain and share the same genome browsers, which enable users to go back and forth between two databases seamlessly.

Gramene (https://www.gramene.org, Tello-Ruiz et al. 2021). Gramene offers a rich data content and a wide range of tools for comparative functional genomics for 118 reference genomes and 124,010 gene family trees (Release #65, May 2022). These genomes encompass a wide range of species, including various accessions of wheat (similar to other databases discussed in this chapter). Gramene is also the home of the Plant Reactome portal (Gupta et al. 2022), which contains pathways information and gene expression displays for 106 species. Gramene has a close partnership with Ensembl Plants and displays genomes, gene models, variations and annotations collaboratively. In addition to multiple visualisation and analysis tools, such as Ensembl genome browsers, BLAST and FTP download, it also houses the Ensembl-Compara-based GeneTrees visualiser tool for sequence-based protein family classification (Vilella et al. 2009).

2.7 WheatIS Data Discovery

An expert working group of the international Wheat Initiative has built an international wheat information system, called WheatIS, with the aim of providing Web-based one-stop access to all available wheat data resources, bioinformatics tools and recommended standards (http://wheatis.org/, Dzale Yeumo et al. 2017; Sen et al. 2020). The data repositories described in this chapter are major data providers of the WheatIS federation that facilitate the availability of genomic, genetic and phenomic data to the community using a data discovery tool. This tool developed by INRAE-URGI is a search engine that indexes the metadata of each database of the federation and provides links back to the source repositories. Long-term sustainability has been achieved through a close collaboration with the ELIXIR European infrastructure for Life Science to develop a common data discovery tool usable both for WheatIS and for ELIXIR (FAIDARE, FAIR Data-finder for Agricultural REsearch, https://urgi.versailles.inrae.fr/faidare/) extended to all plants data.

Figure 2.6 and Table 2.2 present the wheat resources queried by the WheatIS data discovery tool in August 2023: https://urgi.versailles.inrae.fr/wheatis/.

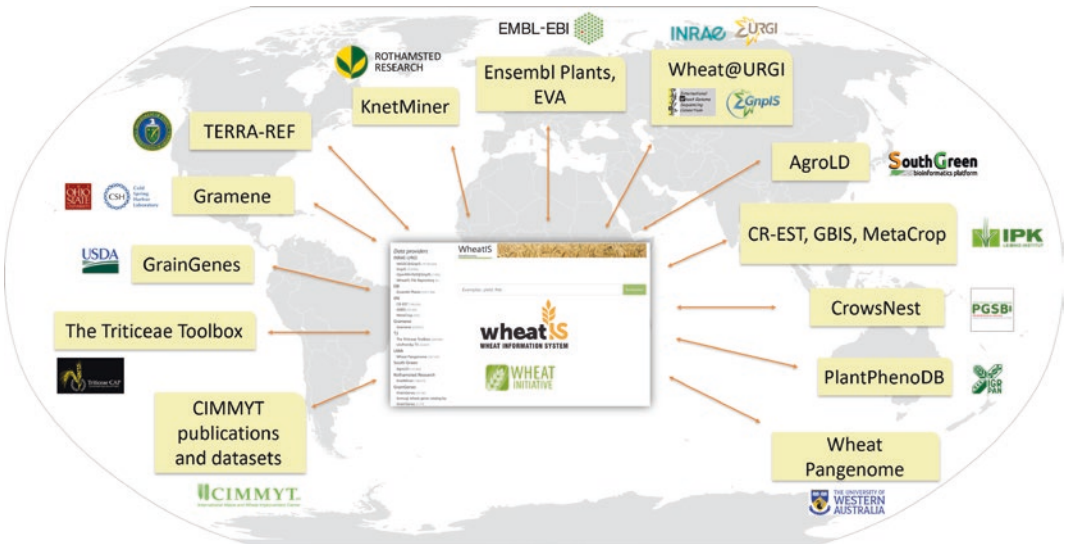


Fig. 2.6 Wheat resources queried by the WheatIS data discovery tool

Table 2.2 Number of data per wheat resource indexed by the WheatIS data discovery

Resource	Institution	Number of indexed data
TERRA-REF	U.S. Department of Energy	284
Wheat@URGI (including IWGSC data repository and GnpIS)	INRAE-URGI	19,844,409
GrainGenes (including Wheat Gene Catalogue at Komugi)	USDA-ARS	23,309
Ensembl Plants (including EVA)	EMBL-EBI	3,071,899
The Triticeae Toolbox (including UniProt)	Triticeae CAP	223,013
Gramene	CSH, OSU	229,851
AgroLD	SouthGreen	137,060
CIMMYT publications and datasets	CIMMYT	1,788
CR-EST, GBIS and MetaCrop	IPK	250,877
CrowsNest	PGSB	13,324
KnetMiner	Rothamsted Research	108,474
PlantPhenoDB	IPG PAS	6
Wheat pangenome	UWA	167,167

ARS Agricultural Research Service; *EMBL* European Molecular Biology Laboratory; *EBI* European Bioinformatics Institute; *INRAE* French National Research Institute for Agriculture, Food and Environment; *URGI* Research Unit in Genomics and Bioinformatics; *USDA* U.S. Department of Agriculture, *Triticeae CAP* Triticeae Coordinated Agricultural Product; *CSH* Cold Spring Harbor Laboratory; *OSU* Ohio State University; *CIMMYT* International Maize and Wheat Improvement Center; *IPK* Leibniz Institute of Plant Genetics and Crop Plant Research; *PGSB* Plant Genome and Systems Biology; *IPG PAS* Institute of Plant Genetics of the Polish Academy of Sciences; *UWA* University of Western Australia

2.8 Conclusion

In a context of increasingly dispersed and numerous wheat data production, the data integration and FAIRification are fundamental. The resources detailed in this chapter contribute to facilitating data discovery by helping researchers and breeders to use genetic and genomic information to improve wheat varieties. The involvement of the wheat bioinformatics community in global initiatives, such as AgBioData, ELIXIR or Research Data Alliance for an open science through standardisation, requires a long-term commitment in order to continue to contribute to research and plant breeding worldwide.

Acknowledgements The authors would like to thank the following people from Ensembl Plants: Guy Naamati, Shradha Saraf and former members Bruno Contreras-Moreira, Dan Bolser, Arnaud Kerhornou and Paul Kersey; from INRAE-URGI: Anne-Françoise Adam-Blondon, Cyril Pommier, Célia Michotey, Raphaël Flores, Nicolas Francillonne, Erik Kimmel; the GrainGenes team members.

Thanks to the International Wheat Genome Sequencing Consortium and its sponsors, the Wheat Initiative and especially the WheatIS expert working group, the Plant Bioinformatics Facility (https://doi.org/10.15454/1.5572414581735654E12), the following projects: BreedWheat (ANR-10-BTBR-03, France Agrimer, FSOV), Whealbi (EU FP7–613556), AGENT (European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 862613), Wheat Genomics for Sustainable Agriculture (BB/J00328X/1), Designing Future Wheat (BB/P016855/1), Elixir: the research infrastructure for life-science data and the European Molecular Biology Laboratory.

References

Alaux M et al (2018) Linking the International Wheat Genome Sequencing Consortium bread wheat reference genome sequence to wheat genetic and phenomic data. Genome Biol 19:111

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Blake VC et al (2016) The Triticeae toolbox: combining phenotype and genotype data to advance small-grains breeding. Plant Genome 9

Blake VC, Wight CP, Yao E, Sen TZ (2022) GrainGenes: tools and content to assist breeders improving oat quality. Foods 11:914

Cagirici HB, Sen TZ (2020) Genome-wide discovery of G-quadruplexes in wheat: distribution and putative functional roles. G3(Bethesda) 10:2021–2032

Choulet F et al (2014) Structural and functional partitioning of bread wheat chromosome 3B. Science 345:1249721

Clavijo BJ et al (2017) An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. Genome Res 27:885–896

Cunningham F et al (2019) Ensembl 2019. Nucleic Acids Res 47:D745–D751

Cunningham F et al (2022) Ensembl 2022. Nucleic Acids Res 50:D988–D995

Dzale Yeumo E et al (2017) Developing data interoperability using standards: a wheat community use case. F1000Res 6:1843

Flores et al (2023) SyntenyViewer: a comparative genomics-driven translational research tool. Database 2023:baad027

Frenkel Z, Paux E, Mester D, Feuillet C, Korol A (2010) LTC: a novel algorithm to improve the efficiency of contig assembly for physical mapping in complex genomes. BMC Bioinformatics 11:584

Gupta P et al (2022) Plant reactome and PubChem: the plant pathway and (Bio)chemical entity knowledgebases. Methods Mol Biol 2443:511–525

He F et al (2019) Exome sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome. Nat Genet 51:896–904

International Wheat Genome Sequencing Consortium (IWGSC) (2014) A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. Science 345:1251788

International Wheat Genome Sequencing Consortium (IWGSC) (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science 361:eaar7191

Jordan KW et al (2015) A haplotype map of allohexaploid wheat reveals distinct patterns of selection on homoeologous genomes. Genome Biol 16:48

Jordan KW, He F, de Soto MF, Akhunova A, Akhunov E (2020) Differential chromatin accessibility landscape reveals structural and functional features of the allopolyploid wheat chromosomes. Genome Biol 21:176

Krasileva KV et al (2017) Uncovering hidden variation in polyploid wheat. Proc Natl Acad Sci USA 114:E913–E921

Mascher M et al (2013) Anchoring and ordering NGS contig assemblies by population sequencing (POPSEQ). Plant J 76:718–727

McLaren W et al (2016) The ensembl variant effect predictor. Genome Biol 17:122

Nelson WM et al (2005) Whole-genome validation of high-information-content fingerprinting. Plant Physiol 139:27–38

Odell SG, Lazo GR, Woodhouse MR, Hane DL, Sen TZ (2017) The art of curation at a biological database: principles and application. Curr Plant Biol 11–12:2–11

Papoutsoglou EA et al (2020) Enabling reusability of plant phenomic datasets with MIAPPE 1.1. New Phytol 227:260–273

Paux E et al (2022) Breeding for economically and environmentally sustainable wheat varieties: an integrated approach from genomics to selection. Biology (Basel) 11:149

Philippe R et al (2012) Whole genome profiling provides a robust framework for physical mapping and sequencing in the highly complex and repetitive wheat genome. BMC Genomics 13:47

Pommier C et al (2019) Applying FAIR principles to plant phenotypic data management in GnpIS. Plant Phenomics 2019:1671403

Pont C et al (2013) Wheat syntenome unveils new evidences of contrasted evolutionary plasticity between paleo- and neoduplicated subgenomes. Plant J 76:1030–1044

Pont C et al (2019) Tracing the ancestry of modern bread wheats. Nat Genet 51:905–911

Ramírez-González RH et al (2018) The transcriptional landscape of polyploid wheat. Science 361:eaar6089

Rimbert H et al (2018) High throughput SNP discovery and genotyping in hexaploid wheat. PLoS ONE 13:e0186329

Selby P et al (2019) BrAPI-an application programming interface for plant breeding applications. Bioinformatics 35:4147–4155

Sen TZ, Caccamo M, Edwards D, Quesneville H (2020) Building a successful international research community through data sharing: the case of the wheat information system (WheatIS). F1000Res 9:536

Shrestha R et al (2012) Bridging the phenotypic and genetic data useful for integrated breeding through a data annotation using the crop Ontology developed by the crop communities of practice. Front Physiol 3:326

Soderlund C, Humphray S, Dunham A, French L (2000) Contigs built with fingerprints, markers, and FPC V4.7. Genome Res 10:1772–1787

Tello-Ruiz MK et al (2021) Gramene 2021: harnessing the power of comparative genomics and pathways for plant research. Nucleic Acids Res 49:D1452–D1463

Tello-Ruiz MK, Jaiswal P, Ware D (2022) Gramene: a resource for comparative analysis of plants genomes and pathways. Methods Mol Biol 2443:101–131

Vilella AJ et al (2009) EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. Genome Res 19:327–335

Walkowiak S et al (2020) Multiple wheat genomes reveal global variation in modern breeding. Nature 588:277–283

Wilkinson MD et al (2016a) The FAIR guiding principles for scientific data management and stewardship. Sci Data 3:160018

Wilkinson PA et al (2016b) CerealsDB 3.0: expansion of resources and data integration. BMC Bioinform 17:256

Yao E et al (2022) GrainGenes: a data-rich repository for small grains genetics and genomics. Database (Oxford) 2022:baac034

Zhu T et al (2021) Optical maps refine the bread wheat *Triticum aestivum* cv. Chinese Spring genome assembly. Plant J 107:303–314

Zimin AV et al (2017) The first near-complete assembly of the hexaploid bread wheat genome, *Triticum aestivum*. Gigascience 6:1–7

Wheat Chromosomal Resources and Their Role in Wheat Research

3

Hana Šimková, Petr Cápal and Jaroslav Doležel

Abstract

Bread wheat (*Triticum aestivum* L.) is grown on more area of land than any other crop, and its global significance is challenged only by rice. Despite the socioeconomic importance, the wheat genome research was lagging behind other crops for a long time. It was mainly a high complexity of the genome, polyploidy and a high content of repetitive elements that were laying obstacles to a thorough genome analysis, gene cloning and genome sequencing. Solution to these problems came in the beginning of the new millennium with the emergence of chromosome genomics—a new approach to studying complex genomes after dissecting them into smaller parts—single chromosomes or their arms. This lossless complexity reduction, enabled by flow-cytometric chromosome sorting, reduced the time and cost of the experiment and simplified downstream analyses. Since the approach overcomes difficulties due to sequence redundancy and the presence of homoeologous subgenomes, the chromosomal genomics was adopted by the International Wheat Genome Sequencing Consortium (IWGSC) as the major strategy to sequence bread wheat genome. The dissection of the wheat genome into single chromosomes enabled the generation of chromosome survey sequences and stimulated international collaboration on producing a reference-quality assembly by the clone-by-clone approach. In parallel, the chromosomal resources were used for marker development, targeted mapping and gene cloning. The most comprehensive approaches to gene cloning, such as MutChromSeq and assembly via long-range linkage, found their use even in the post-sequencing era. The chapter provides a two-decade retrospective of chromosome genomics applied in bread wheat and its relatives and reports on the chromosomal resources generated and their applications.

Keywords

Wheat chromosomes · Wheat genomics

H. Šimková (✉) · P. Cápal · J. Doležel
Centre of Plant Structural and Functional Genomics, Institute of Experimental Botany of the Czech Academy of Sciences, Olomouc, Czech Republic
e-mail: simkovah@ueb.cas.cz

P. Cápal
e-mail: capal@ueb.cas.cz

J. Doležel
e-mail: dolezel@ueb.cas.cz

R. Appels et al. (eds.), *The Wheat Genome*, Compendium of Plant Genomes,
https://doi.org/10.1007/978-3-031-38294-9_3

Abbreviations

BAC	Bacterial artificial chromosome
CSS	Chromosome survey sequence
CS	cv. Chinese Spring
FISH	Fluorescence in situ hybridization
FISHIS	Fluorescence in situ hybridization in suspension
HICF	High information content fingerprinting
HMW DNA	High molecular weight DNA
IWGSC	International Wheat Genome Sequencing Consortium
MDA	Multiple-displacement amplification
MTP	Minimal tiling path
MutChromSeq	Mutant Chromosome Sequencing
OM	Optical map
TE	Transposable element
TACCA	TArgeted Chromosome-based Cloning via long-range Assembly
WGP	Whole genome profiling

3.1 Development of Wheat Chromosome Genomics

The development of DNA sequencing technique by Sanger et al. (1977) marked the beginning of genomics with a prospect of obtaining complete genome sequences and studying entire genomes. The progress in DNA sequencing and genome assembly technologies, which followed the pioneering projects on small bacterial genomes (Fleischmann et al. 1995; Fraser et al. 1995), made it possible to deliver the first genome of a plant—*Arabidopsis thaliana* (Arabidopsis Genome Initiative 2000), followed by *Oryza sativa* (International Rice Genome Sequencing Project 2005). Together with the progress in human genome sequencing (Lander et al. 2001) these achievements stimulated the interest to produce genome sequence of hexaploid bread wheat (*Triticum aestivum*, $2n = 6x = 42$), one of the three most important crops worldwide. This was a daunting task at that time given its genome size exceeding 15 Gb (IWGSC 2018), presence of three homoeologous genomes and high repeat content.

Despite the difficulties foreseen, participants of the workshop on wheat genome sequencing held in Washington DC in 2003 agreed on a need for a bread wheat genome sequence (Gill et al. 2004). Among available strategies, it was decided to explore the use of DNA libraries prepared from individual chromosomes and chromosome arms for the assembly of a global physical map and chromosome sequencing. As individual chromosomes and chromosome arms represent only about 4–6% and 1–3% of the bread wheat genome, respectively, dissecting the genome to chromosomes or even chromosome arms offered a dramatic and lossless reduction in DNA sample complexity to facilitate targeted development of DNA markers, gene mapping and cloning as well as genome sequencing. The chromosome-based approach avoided problems due to the presence of homoeologous DNA sequences and enabled a division of labor so that different groups could work on physical mapping and sequencing different chromosomes simultaneously (Gill et al. 2004). A principal condition for the application of this approach was the ability to purify particular chromosomes and chromosome arms in sufficient numbers (~10^3–10^6) so that enough DNA may be obtained. Until today, the only method suitable for this task is flow-cytometric sorting.

3.1.1 Flow Cytogenetics

Unlike microscopy, flow cytometry analyzes condensed mitotic metaphase chromosomes during their movement, one after another, in a narrow liquid stream. To distinguish this approach from microscopic analysis, the term flow cytogenetics has been coined. Prior to flow cytometry, chromosomes are stained by a DNA fluorochrome so that they can be classified

according to relative DNA content. The analysis can be performed at rates of ~10^3 s so that large numbers of chromosomes can be interrogated to obtain statistically accurate data and potentially discriminate individual chromosomes. A histogram of DNA content thus obtained is termed flow karyotype, and ideally, each chromosome is represented by a well-discriminated peak. In fact, the extent to which the chromosome peak is discriminated from peaks of other chromosomes determined the purity in the sorted fraction, or the frequency of contaminating chromosomes in flow-sorted fraction. Not all flow cytometers are equipped by a sorting module, and only some are designed to physically separate (sort) microscopical particles with particular optical parameters. Gray et al. (1975a, b), Stubblefield et al. (1975) and Carrano et al. (1976) were the first to confirm that flow cytometry can be used not only to classify mammalian chromosomes according to DNA content, but also to sort them. These experiments paved the way to the use of flow-sorted chromosomes during the initial phases of human genome sequencing (Van Dilla and Deaven 1990).

The samples for flow cytometry must have a form of a concentrated suspension of intact chromosomes. In contrast to animals and human, their preparation in plants is hampered by low frequency of dividing mitotic cells and by the presence of a rigid cell wall. A successful approach has been to artificially induce cell cycle synchrony in root tips of hydroponically grown seedlings, accumulate dividing cells at mitotic metaphase and release intact chromosomes from formaldehyde-fixed root tips by mechanical homogenization. This high-yielding procedure was developed for faba bean (Doležel et al. 1992), and by optimizing it for wheat, Vrána et al. (2000) set a foundation for using flow-sorted chromosomes in wheat genomics (Figs. 3.1 and 3.2).

3.1.2 Chromosome Sorting in Wheat

The study of Vrána and co-workers (Vrána et al. 2000) revealed that out of the 21 chromosomes of bread wheat, only chromosome 3B could be discriminated from other chromosomes and sorted at high purity (Fig. 3.3a). The remaining chromosomes formed three composite peaks on a flow karyotype, each of them representing three to ten chromosomes, which could be only sorted as groups. In order to determine chromosome content in the flow-sorted fractions, samples of ~10^3 chromosomes were sorted onto a microscopic slide and microscopically identified after fluorescence in situ hybridization with probes giving chromosome-specific

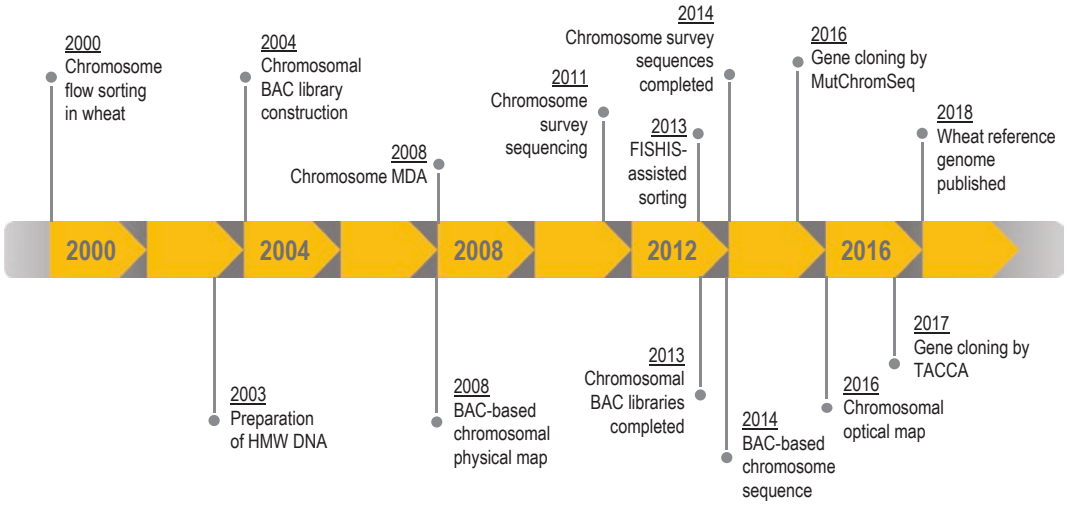


Fig. 3.1 Major developments in wheat chromosomal genomics

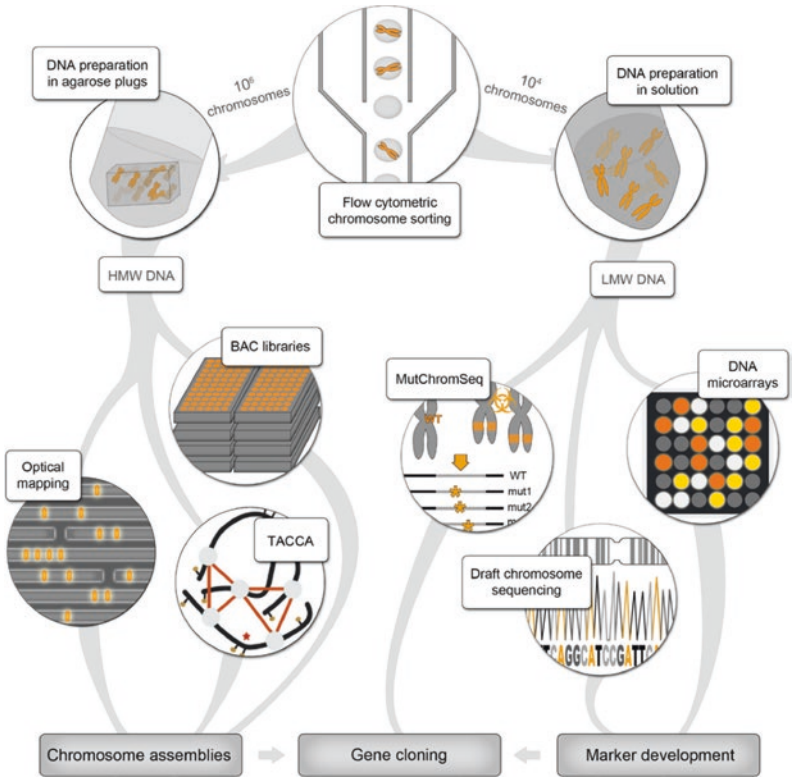


Fig. 3.2 Applications of wheat chromosomal resources. Depending on downstream application, flow-sorted chromosomes can be processed by two distinct approaches. For applications with high demand on DNA amount and contiguity, i.e., BAC libraries, optical mapping and TArgeted Chromosome-based Cloning via long-range Assembly (TACCA), high molecular weight (HMW) DNA is prepared by purifying chromosomes embedded in agarose plugs. Low molecular weight (LMW) DNA, to be used for short-read sequencing or DArT marker development (DNA microarrays), is obtained after treating chromosomal DNA in solution

labeling patterns (Fig. 3.3e; Kubaláková et al. 2002). The study of Vrána et al. (2000) indicated the suitability of chromosomal stocks with altered chromosome sizes for purification of other chromosomes than 3B. In two cultivars of wheat, the authors identified and sorted translocation chromosome 5BL.7BL, which is larger than chromosome 3B (Fig. 3.3c). A subsequent study of Kubaláková et al. (2002) confirmed the potential of cytogenetic stocks. The most important observation concerned the ability to sort any single chromosome arm, either in the form of a telosome or isochromosome. As almost all telosomic lines were developed in the background of cv. CHINESE SPRING (Sears and Sears 1978), their use offered a possibility to analyze the wheat genome chromosome-by-chromosome. In 13 double-ditelosomic lines, both chromosome arms could be discriminated and sorted simultaneously (Fig. 3.3b), saving time to collect DNA from both arms (Doležel et al. 2012).

While this advance made chromosome flow sorting technology ready to support various genomics analyses in bread wheat (Fig. 3.2), including genome sequencing, its dependence on cytogenetic stocks limited its potential for marker development and gene cloning in other wheat genotypes. To overcome this obstacle, Giorgi et al. (2013) developed a protocol for fluorescent labeling repetitive DNA of chromosomes using fluorescence in situ hybridization in suspension (FISHIS). Chromosome classification based on two fluorescence parameters:

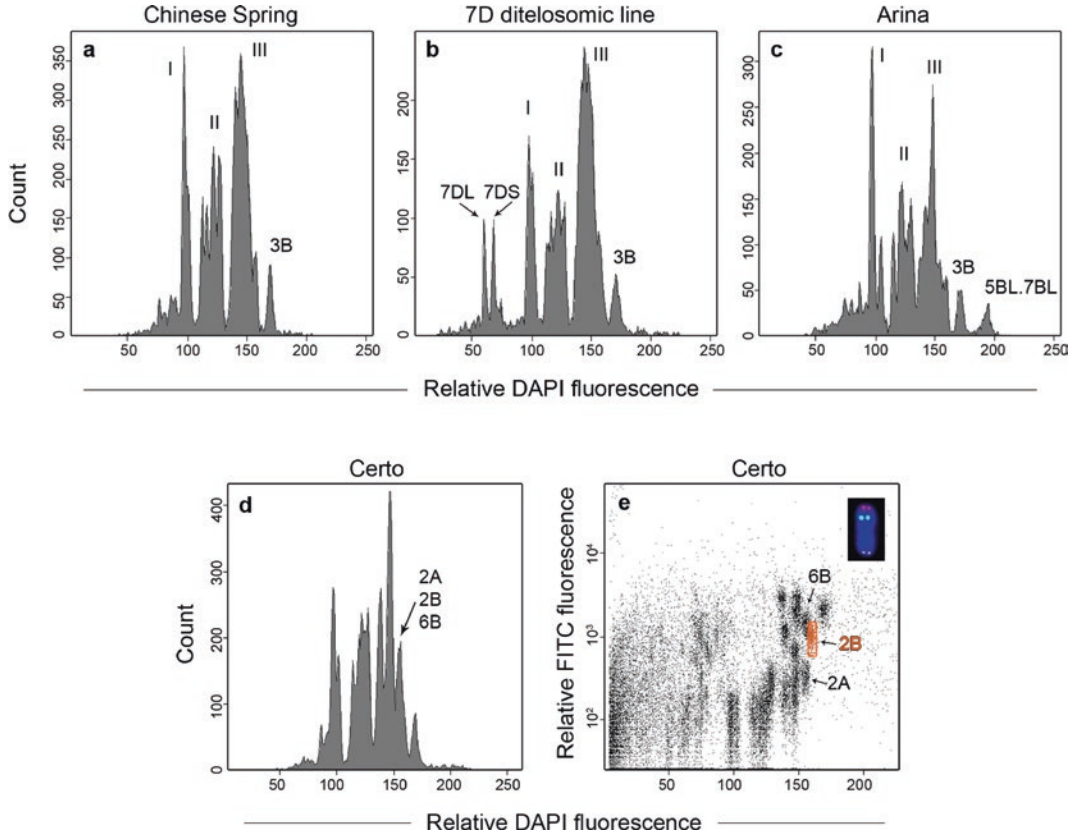


Fig. 3.3 Flow karyotyping of bread wheat. Histograms of relative DAPI fluorescence intensities representing chromosomes of varying sizes are termed flow karyotypes. **a** Flow karyotype of cv. CHINESE SPRING consists of three composite peaks, harboring 3, 7 and 10 chromosomes, respectively, and a standalone peak representing the largest wheat chromosome 3B. **b** Flow karyotype of 7D double ditelosomic line, where both the long and the short arm of chromosome 7D are discriminated and can be sorted simultaneously. **c** The translocated chromosome 5BL.7BL, present in cv. ARINA and some other cultivars, is the largest one in the karyotype and can be sorted with a high purity. **d** Standard monoparametric flow karyotype of cultivar CERTO, where three chromosomes from composite peak III—2A, 2B and 6B—form a defined but still unresolvable sub-population. **e** Bivariate flow karyotype of the same cultivar, where the difference in relative abundance of GAA repeat motif allows further discrimination of these chromosomes and results in well-defined populations containing a single chromosome type each. The chromosome 2B, shown in the inset, can be sorted with purity exceeding 85%. For the purity check, FISH was done with probes for GAA (green) and Afa repeats (red)

DNA (after staining by a DNA fluorochrome) and fluorescence of regions containing DNA repeats (typically GAA microsatellites) labeled by FITC enabled discrimination of chromosomes with the same or very similar DNA content from each other. Depending on genotype, bivariate flow karyotyping after FISHIS typically allows discrimination of ~13 out of 21 wheat chromosomes (Fig. 3.3d, e) and provides to date the most powerful approach to dissect the wheat genome to single chromosomes.

If the FISHIS procedure of Giorgi et al. (2013) is not compatible with a downstream application of sorted chromosomes and, at the same time, appropriate cytogenetic stocks are not available, the option is to partition composite peaks as observed on monovariate flow karyotypes (Fig. 3.3a) (Vrána et al. 2015). Although this approach does not allow discrimination and sorting of single chromosomes, it is suitable for obtaining sub-genomic fractions comprising only a few chromosomes, with one of them

being more abundant. Vrána et al. (2015) calculated a so-called enrichment factor defined as the relative proportion of chromosomal DNA in the wheat genome to the proportion of chromosomal DNA in a sorted fraction and found that a fivefold enrichment was obtained for 17 out of 21 wheat chromosomes. Importantly, subgenomic fractions for 15 out of the 21 chromosomes were not contaminated by homoeologs.

3.1.3 Sorting Chromosomes of Wild Wheat Relatives

The method for flow-cytometric chromosome analysis and sorting, originally developed for hexaploid bread wheat and subsequently modified for tetraploid durum wheat *Triticum turgidum* Desf. var. *durum*, $2n = 4x = 28$ (Kubaláková et al. 2005) was also found to be suitable to sort chromosomes from their wild relatives. In fact, two options were explored. One involved sorting chromosomes from alien chromosome introgression lines of wheat. The samples are prepared from synchronized wheat root tips and, if the alien chromosome can be discriminated on a flow karyotype, it may be sorted (Molnár et al. 2011, 2015; Zwyrtková et al. 2022). In a similar manner, wheat chromosomes carrying introgressions from wild relatives can be purified (Tiwari et al. 2014; Janáková et al. 2019; Bansal et al. 2020). Second and straightforward option is to sort chromosomes directly from wild relatives. Thus, the protocol of Vrána et al. (2000) for wheat has been optimized for a variety of species from *Aegilops*, *Agropyron* and *Haynaldia* (*Dasypyrum*) genera (summarized in Doležel et al. 2021). While in some of them (like *Aegilops comosa*), all chromosomes may be discriminated and sorted (Said et al. 2021), in majority of species (including *Aegilops geniculata, Aegilops biuncialis, Aegilops cylindrica, Haynaldia villosa, Agropyron cristatum* and others) their chromosomes can only be sorted in groups of two to five (Molnár et al. 2011, 2015; Grosso et al. 2012; Said et al. 2019). As in case of wheat, fluorescent labeling of chromosomes by FISHIS prior to flow cytometry increased the number of chromosomes that could be discriminated and sorted. Availability of separated chromosomes of the relatives enabled comparative studies with the bread wheat genome (Molnár et al. 2014, 2016) and have been applied to support cloning of genes from the tertiary gene pool (see Sect. 3.5.1).

3.2 Toward Bread Wheat Reference Genome

Need for a quality bread wheat genome that would provide access to the complete gene catalogue, an unlimited amount of molecular markers to support genome-based selection of new varieties and a framework for the efficient exploitation of natural and induced genetic diversity (Choulet et al. 2014a) stimulated the establishment of the International Wheat Genome Sequencing Consortium, a collaborative platform launched in 2005 (https://www.wheatgenome.org). By that time, a proven strategy to obtaining high-quality reference sequences of large genomes was the clone-by-clone approach, i.e., sequencing clones from large-insert DNA libraries ordered in physical maps. These constituted a technology-neutral resource for accessing complex genomes, enabling possible resequencing of the ordered clones by more advanced technologies. Considering the ability to dissect the wheat genome to individual chromosomes or chromosome arms (Vrána et al. 2000; Kubaláková et al. 2002), and after confirming the feasibility of constructing large-insert DNA libraries from the flow-sorted chromosomes (Šafář et al. 2004; Janda et al. 2004), the Consortium settled on coupling the chromosome purification with the clone-by-clone strategy and producing clone-based physical maps of individual wheat chromosomes that would allow the engagement of multiple teams in the challenging sequencing effort.

3.2.1 Generation of Chromosomal BAC Resources

The prerequisite of the proposed strategy was the ability to separate by flow sorting each of bread wheat chromosomes or chromosome arms. This was only possible in cultivar CHINESE SPRING (CS), for which a complete set of telosomic lines, essential to sort the chromosome arms, was available (Sears and Sears 1978), predestining the cultivar to become the reference genome of bread wheat. The primary resource needed to construct a clone-based physical map is a large-insert genomic DNA library, commonly cloned in the bacterial artificial chromosome (BAC) vector, typically bearing inserts of 100–200 kb. To generate a library of these parameters, several micrograms of high molecular weight (HMW) DNA are needed. Achieving this from the flow-sorted material involved the elaboration of a customized protocol (Šimková et al. 2003) including DNA preparation in agarose plugs (Fig. 3.2), which enabled cumulating samples from multiple sorting days. Based on this advance, Šafář et al. (2004) constructed the first-ever chromosome-specific BAC library in a eukaryotic organism. The library, prepared from two million 3B chromosomes flow-sorted over 18 working days, comprised 67,968 clones with 103 kb average insert size, representing 6.2 equivalents of the chromosome 3B, whose molecular size is close to one gigabase. Further improvements in the procedure permitted the construction of BAC libraries with chromosome coverage up to 18 × and average insert size exceeding 120 Kb (https://olomouc.ueb.cas.cz/en/resources/dna-libraries (Šafář et al. 2010; Table 3.1 and references therein). The effort toward preparing the full set of CS libraries for the chromosomal physical maps lasted over ten years and was completed in the end of 2013 (Fig. 3.1). Individual clones and BAC libraries used to construct chromosome-specific physical maps are publicly available and can be obtained at https://cnrgv.toulouse.inrae.fr/en/Library/Wheat. Besides the 'CHINESE SPRING' BAC libraries generated for the reference genome project, several customized chromosomal libraries from other cultivars were created for the purpose of gene cloning projects, including 3B-specific library from cv. HOPE (Mago et al. 2014) and a BAC library from 4AL arm of cv. TÄHTI, bearing an introgressed segment of *Triticum militinae* (Janáková et al. 2019) (Table 3.1).

Upon their construction, the CS libraries were distributed among national teams engaged in the IWGSC effort who embarked on constructing physical maps. In a proof-of-concept experiment, Paux and co-workers (2008) generated the first chromosomal physical map from chromosome 3B, employing SNaPShot-based High Information Content Fingerprinting (HICF) technology (Luo et al. 2003) to generate fingerprints and FingerPrinted Contig (FPC) software to assemble the physical map and select minimal tiling path (MTP) for sequencing. This achievement validated the feasibility of constructing sequence-ready physical maps of hexaploid wheat by the chromosome-by-chromosome approach and the strategy was subsequently followed for other chromosome arms (Table 3.1; IWGSC 2018). As alternative procedures, Whole Genome Profiling (WGP, van Oeveren et al. 2011) was applied for BAC fingerprinting in several projects and Linear Topological Contig (LTC, Frenkel et al. 2010) software was developed and utilized for map assembly and validation. Procedures applied for individual chromosomes/arms are summarized in IWGSC 2018. The resulting chromosomal physical maps are available at https://urgi.versailles.inra.fr/download/iwgsc/Physical_maps/ and displayable at https://urgi.versailles.inra.fr/gb2/gbrowse/wheat_phys_pub/. In addition to the construction of physical maps for several chromosomes, the WGP technology was utilized to profile MTP clones identified from chromosome physical maps constructed previously by the HICF procedure. Thus generated WGP tags of all 21 wheat chromosomes were used to support the assembly of the IWGSC RefSeq v1.0 genome and are available for download from IWGSC-BayerCropScience WGP™ tags https://urgi.versailles.inra.fr/download/iwgsc/IWGSC_BayerCropScience_WGPTM_tags.

Table 3.1 Wheat chromosomal BAC resources

Library name	Cultivar	Chromosome/arm	Number of clones	Insert size (kb)	Coverage	BAC assembly	References*
TaaCsp146eA	CHINESE SPRING	1D, 4D, 6D	26,112	110	1.3×		
TaaCsp146hA	CHINESE SPRING	1D, 4D, 6D	87,168	85	3.4×		Janda et al. (2004)[1]
TaaCsp146hB	CHINESE SPRING	1D, 4D, 6D	148,224	102	6.9×		
TaaCsp146hC	CHINESE SPRING	1D, 4D, 6D	138,240	116	7.4×		
TaaCsp1ALhA	CHINESE SPRING	1AL	49,536	103	8.0×		Lucas et al. (2013)[1,2]
TaaCsp1ALhB	CHINESE SPRING	1AL	43,008	109	7.7×		
TaaCsp1AShA	CHINESE SPRING	1AS	31,104	111	11.8×		Breen et al. (2013)[1,2]
TaaCsp1BLhA	CHINESE SPRING	1BL	92,160	114	15.4×	Yes	Philippe et al. (2013)[1,2]
TaaCsp1BShA	CHINESE SPRING	1BS	55,296	113	15.7×	Yes	Raats et al. (2013)[1,2]
TaaCsp2ALhA	CHINESE SPRING	2AL	76,800	120	15.8×		
TaaCsp2AShA	CHINESE SPRING	2AS	56,832	123	15.4×		
TaaCsp2BLhA	CHINESE SPRING	2BL	70,656	120	15.1×		
TaaCsp2BShA	CHINESE SPRING	2BS	67,968	116	15.6×		
TaaCsp2DLhA	CHINESE SPRING	2DL	58,368	124	15.3×		
TaaCsp2DShA	CHINESE SPRING	2DS	43,008	132	15.6×		
TaaCsp3ALhA	CHINESE SPRING	3AL	55,296	106	10.2×		
TaaCsp3ALhB	CHINESE SPRING	3AL	24,576	114	5.2×		
TaaCsp3AShA	CHINESE SPRING	3AS	55,296	80	10.9×		Luo et al. (2010)[1]
TaaCsp3AShB	CHINESE SPRING	3AS	55,296	115	15.9×		
TaaCsp3BFeA	CHINESE SPRING	3B	21,120	107	1.9×		
TaaCsp3BFhA	CHINESE SPRING	3B	67,968	103	6.2×	Yes	Šafář et al. (2004)[1], Paux et al. (2008)[2], Choulet et al. (2014b)[3]
TaaCsp3BFhB	CHINESE SPRING	3B	82,176	126	9.1×		
TaaCsp3DLhA	CHINESE SPRING	3DL	64,512	105	12.2×	Yes	
TaaCsp3DShA	CHINESE SPRING	3DS	36,864	110	11.0×	Yes	Luo et al. (2010)[1], Holušová et al. (2017)[2]
TaaCsp4ALhA	CHINESE SPRING	4AL	92,160	126	17.3×	Yes	Shorinola et al. (2017)[2]
TaaCsp4AShA	CHINESE SPRING	4AS	49,152	131	16.6×		
TaaCsp4BLhA	CHINESE SPRING	4BL	63,744	118	15.0×		
TaaCsp4BShA	CHINESE SPRING	4BS	58,368	123	15.4×		
TaaCsp5ALhA	CHINESE SPRING	5AL	90,240	123	18.3×		Barabaschi et al. (2015)[1,2]
TaaCsp5AShA	CHINESE SPRING	5AS	46,080	120	16.5×		

(continued)

Table 3.1 (continued)

Library name	Cultivar	Chromosome/arm	Number of clones	Insert size (kb)	Coverage	BAC assembly	References*
TaaCsp5BLhA	CHINESE SPRING	5BL	76,800	126	15.7×		
TaaCsp5BShA	CHINESE SPRING	5BS	43,776	122	15.8×	Yes	Salina et al. (2018)[1,2]
TaaCsp5DLhA	CHINESE SPRING	5DL	72,960	128	16.0×		
TaaCsp5DShA	CHINESE SPRING	5DS	36,864	137	17.0×		Akpinar et al. (2015b)[1,2]
TaaCsp6ALhA	CHINESE SPRING	6AL	55,296	123	15.7×		Poursarebani et al. (2014)[1,2]
TaaCsp6AShA	CHINESE SPRING	6AS	49,152	125	16.0×		
TaaCsp6BLhA	CHINESE SPRING	6BL	76,032	130	18.0×	Yes	Kobayashi et al. (2015)[1,2]
TaaCsp6BShA	CHINESE SPRING	6BS	57,600	132	15.3×	Yes	
TaaCsp7ALhA	CHINESE SPRING	7AL	61,056	124	15.3×	Yes	Keeble-Gagnère et al. (2018)[2,3]
TaaCsp7AShA	CHINESE SPRING	7AS	58,368	134	15.4×	Yes	
TaaCsp7BLhA	CHINESE SPRING	7BL	72,960	136	15.1×	Yes	
TaaCsp7BShA	CHINESE SPRING	7BS	27,648	182	12.5×	Yes	
TaaCsp7BShB	CHINESE SPRING	7BS	21,504	112	5.8×	Yes	
TaaCsp7DLhA	CHINESE SPRING	7DL	50,304	115	14.8×	Yes	Šimková et al. (2011)[1], Feng et al. (2020)[2,3]
TaaCsp7DShA	CHINESE SPRING	7DS	49,152	114	12.2×	Yes	Šimková et al. (2011)[1], Tulpová et al. (2019a)[2,3]
TaaHop3BFhA	HOPE	3B	92,160	78	6.0×		Mago et al. (2014)[1]
TaaHop3BFhB	HOPE	3B	43,776	160	6.3×		
TaaPav1BShA	PAVON	1BS	65,280	82	14.5×		Janda et al. (2006)[1]
TaaPav1BLhA	PAVON	1BL	41,088	130	8×		
TaaPmt4ALhA	TÄHTI—*T. militinae*introgression	4AL	43,008	113	6.8×		Janáková et al. (2019)[1]

Additional information about the libraries is provided at https://olomouc.ueb.cas.cz/en/resources/dna-librariesand in Šafář et al. (2010)

*All CHINESE SPRING libraries and related physical maps are summarised in IWGSC (2018). The provided references specifically refer to a particular BAC library[1], related physical map[2], and BAC assembly[3]

3.2.2 BAC Clone Sequencing

Availability of BAC clones ordered in chromosomal physical maps opened avenue to systematic analyses of bread wheat genome and its selected parts. The early studies, based on sequencing ends of BAC clones by Sanger technology, provided first insights into gene and repeat content of particular chromosomes, enabled comparative analyses of homoeologous chromosomes and delivered information for targeted marker development (Paux et al. 2006; Sehgal et al. 2012; Lucas et al. 2012).

Later studies, employing next-generation sequencing of whole BAC contigs, provided more comprehensive information about organization of genes and transposable elements (TEs). Choulet et al. (2010) sequenced and annotated 13 BAC contigs, totaling 18 Mb sequence, selected from different regions of the 3B chromosome and revealed that genes were present along the entire chromosome and clustered mainly into numerous small islands of 3–4 genes separated by large blocks of repetitive elements. They observed that wheat genome expansion had occurred homogeneously along the chromosome through specific bursts of TEs. Bartoš et al. (2012), after sequencing a megabase-sized region from wheat arm 3DS and comparing it with the homoeologous region on wheat chromosome 3B, revealed similar rates of non-collinear gene insertion in wheat B and D subgenomes with a majority of gene duplications occurring before their divergence. Li et al. (2013) provided valuable information about the structure of wheat centromeres. Analyzing 1.1-Mb region from the centromere of chromosome 3B, they revealed that 96% of the DNA consisted of TEs. The youngest elements, *CRW* and *Quinta*, were targeted by the centromere-specific histone H3 variant CENH3—the marker of the functional centromere. In contrast to the TEs, long arrays of satellite repeats found in the region were not associated with CENH3. Several other studies employing sequencing of BAC contigs focused on analysis of narrow regions comprising their genes of interest (Breen et al. 2010; Mago et al. 2014; Janáková et al. 2019; Tulpová et al. 2019b).

Although these studies markedly advanced the knowledge on bread wheat genome, the major breakthrough came only with the generation of chromosome-scale sequence assemblies. Choulet and co-workers (2014b) produced a BAC-based reference sequence of the largest bread wheat chromosome—3B. After sequencing 8452 BAC clones, representing the 3B MTP, the authors assembled a sequence of 833 Mb split in 2808 scaffolds, 1358 of which, containing 774 Mb sequence, had known position on the chromosome. The assembly comprised 5326 protein-coding genes, 1938 pseudogenes and 85% of transposable elements. Most interestingly, the distribution of structural and functional features along the chromosome revealed partitioning correlated with meiotic recombination. Comparative analyses with other grass genomes indicated high wheat-specific inter- and intrachromosomal gene duplication activities that were postulated to be sources of variability for adaption. As a contribution to the IWGSC sequencing effort, sequence assemblies of BAC clones representing complete or partial MTPs of seven chromosomes and two chromosome arms were produced (Table 3.1 and references therein; IWGSC 2018) and are publicly available at https://urgi.versailles.inrae.fr/download/iwgsc/BAC_Assemblies/. These assemblies, complemented by information from chromosomal physical maps, and—for group 7 chromosomes—also chromosomal optical maps, were applied to support the assembly of the bread wheat reference genome, IWGSC RefSeq v1.0 (IWGSC 2018), as described in Chap. 2.

It is clear nowadays that the whole-genome-shotgun became the predominant approach to sequencing, even for large polyploid genomes. Still, the generated wheat chromosomal physical maps and BAC clones integrated therein remain a valuable genomic resource for bread wheat, enabling a fast access to and a detailed analysis of a region of interest. The availability of BAC clones with a known genomic position facilitated a focused and affordable resequencing of

a region of interest with long-read technologies, revealing discrepancies and missing segments in the previously generated bread wheat assemblies (Kapustová et al. 2019; Tulpová et al. 2019b).

3.3 Chromosome Survey Sequencing

While the generation of the full set of chromosomal libraries, physical maps and BAC clone sequences proved to be a long-distance run, the requirement for homoeolog-resolved wheat genome information was increasing over time. Apparently, this demand could be met by low-pass chromosome sequencing, which would provide approximate information about the genic component of individual chromosomes. The separation of each bread wheat chromosome or chromosome arm was, in principle, feasible but the yield of flow-sorted chromosomes, typically 1–2 × 10^5 per sorting day, did not meet the demands of the early sequencing technologies on the DNA input, which was in the microgram range. Coupling of chromosome flow sorting with multiple-displacement amplification (MDA) of the chromosomal DNA, originally developed for physical mapping on DNA microarrays (Šimková et al. 2008), opened the door to shotgun sequencing of cereal chromosomes one-by-one. Wheat genome researchers adopted the strategy of chromosome survey sequencing (CSS) developed for barley (Mayer et al. 2009, 2011). In barley, low-coverage (1–3×) chromosomal data, obtained by 454 sequencing, were compared with reference genomes of rice, sorghum and Brachypodium, and EST or full-length-cDNA datasets, which led to the estimation of gene content for each of the barley chromosomes. Moreover, an integration of the shotgun sequence information with the collinear gene order of orthologous rice, sorghum and Brachypodium genes allowed proposing virtual gene order maps of individual chromosomes. The syntenic integration, known as genome zipper, resolved gene order in regions with limited genetic resolution, such as genetic centromeres, which were intractable to genetic mapping.

The first experiments with the CSS in bread wheat were done to compare chromosome arms of homoeologous group 1 (Wicker et al. 2011), and it methodologically followed the barley model, employing the low-pass 454 sequencing. The study revealed that all three wheat subgenomes had similar sets of genes that were syntenic with the model grass genomes but the number of genic sequences in non-syntenic positions outnumbered that of the syntenic ones. Further analysis indicated that a large proportion of the genes that were found in only one of the three homoeologous wheat chromosomes were most probably pseudogenes resulting from transposon activity and double-strand break repair. These findings were supported by a study of Akhunov et al. (2013) who, working with CSSs of both arms of chromosome 3A, found that ~35% of genes had experienced structural rearrangements leading to a variety of mis-sense and non-sense mutations—a finding concordant with other studies indicating ongoing pseudogenization of the bread wheat genome. Another focus of the CSS studies was the evolutionary rearrangement of wheat chromosomes. Hernandez et al. (2012) analyzed bread wheat chromosome 4A, which has undergone a major series of evolutionary rearrangements. Using the genome zipper approach, the authors produced an ordered gene map of chromosome 4A, embracing ~85% of its total gene content, which enabled precise localization of the various translocation and inversion breakpoints on chromosome 4A that differentiate it from its progenitor chromosome in the A-subgenome diploid donor.

In contrast to the above studies, Berkman and co-workers, aiming to shotgun sequence wheat 7DS arm, favored the use of the more cost-efficient Illumina technology and compensated its short reads (75–100 bp) by higher sequencing coverage (34×), which allowed a partial assembly of the reads and capture of ~40% of the sequence content of the chromosome arm (Berkman et al. 2011). Using the same technology, the team proceeded with sequencing the 7BS arm (Berkman et al. 2012) and supplemented the 4A study by delimiting the 7BS segment that was involved in the reciprocal

translocation that gave rise to the modern 4A chromosome. After extending the sequencing effort to all group7 homoeologs (Berkman et al. 2013), the team compared the sequences and concluded that there had been more gene loss in 7A and 7B than in 7D chromosome. Chromosome survey sequences of additional chromosomes/arms followed and were mostly utilized in estimating gene and repeat content of particular chromosomes (Vitulo et al. 2011; Tanaka et al. 2014; Sergeeva et al. 2014; Helguera et al. 2015; Garbus et al. 2015; Kaur et al. 2019), synteny-based ordering of arising clone-based physical maps (Lucas et al. 2013), identifying miRNA-coding sequences (Vitulo et al. 2011; Kantar et al. 2012; Deng et al. 2014; Tanaka et al. 2014) and delimiting lineage-specific translocations (Lucas et al. 2014). Utilization of the chromosome sequencing for gene mapping and cloning is described further in Sect. 3.5.1.

The chromosome survey sequencing in bread wheat has been crowned by a joint effort coordinated by the IWGSC, which exploited the existing Illumina-based CSSs and complemented them by newly produced Illumina data for the remaining chromosomes. The sequences were applied to generate draft assemblies and genome zippers for all wheat chromosomes (IWGSC 2014). As a result, a total of 124,201 gene loci were annotated and more than 75,000 genes were positioned along chromosomes. The IWGSC team anchored more than 3.6 million marker loci to chromosome sequences, uncovered the molecular organization of the three subgenomes and described patterns in gene expression across the subgenomes. The study also provided new insights into the phylogeny of hexaploid bread wheat, which was elaborated in detail in an accompanying study of Marcussen et al. (2014). Moreover, this new wheat genome information was used as a reference to analyze the cell type-specific expression of homoeologous genes in the developing wheat grain (Pfeifer et al. 2014).

The technique of chromosome survey sequencing soon expanded beyond the cultivated crop and was successfully applied to explore individual chromosomes or whole genomes of close wheat relatives, such as *Aegilops tauschii* (Akpinar et al. 2015a) and *Triticum dicoccoides* (Akpinar et al. 2015c; 2018), and even species from the tertiary gene pool, including *Ae. geniculata* (Tiwari et al. 2015), *H. villosa* (Xiao et al. 2017), *Ae. comosa*, *Aegilops umbellulata* (Said et al. 2021) and *A. cristatum* (Zwyrtková et al. 2022). These studies informed about the chromosome gene content and organization, enabling comparative studies important for gene transfer from the wild species to the crop as well as identifying the sequences enabling marker development for tracing introgressions in wheat. Specific examples are provided in Sect. 3.5.1 and Table 3.2.

3.4 Optical Mapping

Extensive experience with preparing quality HMW DNA from flow-sorted chromosomes paved the way to establish a new branch of wheat chromosomal genomics—chromosome optical mapping (OM). The OM technology, commercialized by Bionano Genomics and therefore also known as Bionano genome mapping, is a physical mapping technique based on labeling and imaging short sequence motives along 150 kb to 1 Mb long DNA molecules (Lam et al. 2012). Resulting restriction maps, assembled from high-coverage single-molecule data, are composed of contigs up to > 100 Mb in size, which are instrumental in finishing steps of genome assemblies by enabling contig scaffolding, gap sizing and assembly validation. The optical maps also provided a high-resolution and cost-effective tool for comparative structural genomics.

Staňková et al. (2016) demonstrated the feasibility of generating optical maps from DNA of flow-sorted chromosomes and constructed

Table 3.2 Leveraging wheat chromosomal resources in gene mapping and cloning

Phenotype	Locus	Sorted chrom./arm	Applied approach	References
Stem rust resistance	*Sr2*	3B	BAC	McNeil et al. (2008) Mago et al. (2014)
	SuSr-D1	7D	MutChromSeq	Hiebert et al. (2020)
Green bug resistance	*Gb3*	7DL	BAC	Šimková et al. (2011)
Powdery mildew resistance	*QPm-tut-4A*	4AL	SynSNP	Jakobson et al. (2012)
			CSS, ChromSeq, RICh BAC, CSS	Abrouk et al. (2017) Janáková et al. (2019)
	Pm2	5D	MutChromSeq	Sánchez-Martín et al. (2016)
	Pm21	6V	TACCA	Xing et al. (2018)
	Pm4	2A	MutChromSeq	Sánchez-Martín et al. (2021)
	Pm1a	7A	ChromSeq	Hewitt et al. (2021)
Species cytoplasm specific	*scs*	1D	SynSNP	Michalak de Jimenez et al. (2013)
Leaf rust resistance	*Lr14a*	7BL	SynSNP	Terracciano et al. (2013)
	Lr57	$5M^g$	ChromSeq	Tiwari et al. (2014)
	Lr22	2D	TACCA	Thind et al. (2017)
	Lr49	4B	ChromSeq	Nsabiyera et al. (2020)
	Lr76	5D/5U	ChromSeq	Bansal et al. (2020)
	Lr14a	7BL.5BL	MutChromSeq	Kolodziej et al. (2021)
Glume blotch resistance	*QSng.sfr-3BS*	3B	ChromSeq	Shatalina et al. (2013, 2014)
Stripe rust resistance	*Yr40*	$5M^g$	ChromSeq	Tiwari et al. (2014)
	YrAW1	4AL	ChromSeq	Randhawa et al. (2014)
	Yr70	5D/5U	ChromSeq	Bansal et al. (2020)
Russian wheat aphid resistance	*Dn2401*	7DS	CSS, SynSNP BAC, OM	Staňková et al. (2015) Tulpová et al. (2019b)
Pre-harvest sprouting resistance	*Phs-A1*	4AL	BAC	Shorinola et al. (2017)
Semi-dwarfism	*Rht18*	6A	MutChromSeq	Ford et al. (2018)
Yellow Early Senescence	*YES-1*	3A	ChromSeq	Harrington et al. (2019)
Fusarium head blight resistance	*Fhb*	7EL	ChromSeq	Konkin et al. (2022)

BAC BAC-based physical map/BAC sequencing
CSS Chromosome survey sequence
ChromSeq chromosome sequencing
MutChromSeq Mutant chromosome sequencing
OM Optical map
RICh Rearrangement identification and characterization
SynSNP Synteny-based SNP marker development
TACCA TArgeted chromosome-based cloning via long-range assembly

the first-ever optical map for the bread wheat genome. Using 1.6 million flow-sorted 7DS chromosome arms and the first-generation platform of Bionano Genomics, the authors prepared a map consisting of 371 contigs with N50 of 1.3 Mb, which supported a physical-map and a BAC-based sequence assembly of the chromosome arm (Tulpová et al. 2019a). Applied in a gene cloning project, the OM posed a targeted tool for sequence validation and analysis of structural variability in a region of interest (Tulpová et al. 2019b). Similar maps have been constructed for other group-7 chromosome arms and were used in the process of assembling the wheat reference genome (IWGSC 2018), as well as a complementary BAC-based assembly of chromosome 7A (Keeble-Gagnère et al. 2018).

Another set of chromosomal optical maps was prepared from chromosome arms 1AS, 1BS, 6BS and 5DS, the last being generated on the second-generation platform of Bionano Genomics, with the aim to position and characterize 45S rDNA loci located on those arms. The chromosome-based approach applied in the rDNA project enabled analyzing the loci one-by-one and provided more comprehensive information about individual loci than achieved in long-read bread wheat assemblies (Tulpová et al. 2022).

3.5 Gene Mapping and Cloning

In parallel with the chromosome sequencing efforts, the wheat community started exploiting flow-sorted chromosomes for targeted marker development, aiming to generate a high-density map in a region of interest and, possibly, clone a gene by a map-based approach. This conventional strategy was later complemented by new methods of 'rapid gene cloning' (reviewed in Bettgenhaeuser and Krattinger, 2019). Some of these still capitalize on the complexity reduction by chromosome flow sorting but they avoid the lengthy step of marker development and map saturation while employing mutation genetics and comprehensive sequencing techniques to assemble a highly contiguous sequence for the chromosome of interest.

3.5.1 Marker Development and Map-Based Gene Cloning

The first effort toward massive marker development from a selected chromosome or chromosome arm was bound with the microarray platform of Diversity Array Technologies, able to identify and utilize polymorphic DNA markers without knowledge of the underlying sequence (Jaccoud et al. 2001). Wenzl et al. (2010) demonstrated that a chromosome-enriched DArT array could be developed from only a few nanograms of chromosomal DNA. Of 711 polymorphic markers derived from non-amplified DNA of bread wheat chromosome 3B, 553 (78%) mapped to the chromosome, and even higher efficiency (87%) was observed for the short arm of bread wheat chromosome 1B (1BS).

Before the availability of wheat chromosomal survey sequences, researchers aiming to develop new markers for their locus of interest mined data from sequenced genomes of model grasses, mainly rice, Brachypodium and sorghum. Efficiency of this synteny-based approach was compromised by limitations in designing gene-derived primers with sufficient specificity to distinguish homoeologous genes in polyploid wheat. Amplified DNA from individual wheat chromosome arms used as a template for locus-specific PCR and subsequent amplicon sequencing, significantly increased the efficiency of the procedure and the facilitated targeted generation of gene-associated SNP markers in a time- and cost-effective manner (Jakobson et al. 2012; Michalak de Jimenez et al. 2013; Terracciano et al. 2013; Staňková et al. 2015). Additionally, particular chromosomal arms used as a PCR template were applied to validate specificity of the newly designed markers (Staňková et al. 2015; Janáková et al. 2019).

Advancement in marker development came along with the release of 'CHINESE SPRING'

CSSs and genome zippers that informed about putative gene content and order in the region of interest in the reference genome. Nevertheless, studies comparing shotgun sequences of CS chromosomes with those of other wheat accessions revealed extensive intra- and interchromosomal rearrangements in CS (Ma et al. 2014, 2015; Liu et al. 2016), implying limitations in the transferability of data from the wheat reference to other genomes. Moreover, it became obvious that agronomically important traits were frequently controlled by rare, genotype-specific alleles or had even been introgressed to wheat from its relatives. Under such scenario, genetic maps had to be created from a mapping population derived from a donor of the trait and sequence information from the donor was essential for marker development. As a proof-of-concept experiment, Shatalina et al. (2013) generated tenfold coverage of Illumina data from chromosome 3B isolated from wheat cultivars ARINA and FORNO—the parents of their mapping population. Relying on a synteny with the Brachypodium genome, they identified sequences close to coding regions and used them to develop 70 SNP markers, which were found dispersed over the entire 3B chromosome and contributed to fourfold increase in the number of available markers. The new markers were utilized for mapping a QTL conferring resistance to *Stagonospora nodorum* glume blotch located on 3BS (Shatalina et al. 2014). Chromosome sequencing was then applied by other groups to fine-map *Yellow Early Senescence 1* (Harrington et al. 2019), leaf rust resistance gene *Lr49* (Nsabiyera et al. 2020) and powdery mildew resistance gene *Pm*1 (Hewitt et al. 2021).

The procedure was also adopted to develop markers in species from wheat tertiary gene pool, such as *Ae. geniculata* (Tiwari et al. 2014) and *H. villosa* (Wang et al. 2017; Zhang et al. 2021), with the aim to trace the alien chromatin in the wheat background. For this purpose, the method was refined by Abrouk et al. (2017) who developed an in silico pipeline termed Rearrangement Identification and Characterization (RICh). To delimit a segment transferred from *T. militinae* to the long arm of chromosome 4A of bread wheat cv. TÄHTI, the authors generated a virtual gene order of 'TÄHTI' chromosome 4A. Comparison of homoeologous gene density between 4AL arm of CS and the arm with the introgression, which harbored powdery mildew resistance locus *QPm.tut-4A*, identified alien chromatin with 169 putative genes originating from *T. militinae*. A similar approach was used by Bansal et al. (2020) to fine-map leaf rust and stripe rust resistance genes *Lr76* and *Yr70* introduced from *Ae. umbellulata*. The authors sequenced flow-sorted chromosomes 5U from *Ae. umbellulata*, 5D from a bread wheat-*Ae. umbellulata* introgression line and 5D from the recurrent parent. Sequencing reads were explored with the aim to identify introgression-specific SNP markers whose projection on the IWGSC RefSeq v1.0 sequence (IWGSC 2018) delimited the introgression to a 9.47 Mb region, in which candidates for *Lr76* and *Yr70* genes were identified. Konkin et al. (2022), streaming to identify genes for resistances to several fungal pathogenes, including fusarium head blight, sequenced 7EL telosome, originated from *Thinopyrum elongatum* and existing as addition in CS wheat. They thus built a reference for comparative transcriptome analysis between CS and CS-7EL addition line, which resulted in a list of candidate genes for the resistance.

Alongside the wheat chromosomal survey sequences, emerging BAC assemblies from individual chromosomes of 'CHINESE SPRING,' just as customized chromosomal BAC libraries from other cultivars showed instrumental in gene cloning projects. Šimková et al. (2011) demonstrated that BAC libraries constructed from chromosome arms 7DS and 7DL, consisting of tens of thousands BAC clones, were highly representative and easy to screen, which facilitated fast chromosome walking in a region of green bug resistance gene *Gb3* in 7DL. The 7DS BAC library was screened for markers tightly linked to a Russian wheat aphid resistance locus *Dn2401* (Staňková et al. 2015) and a BAC contig spanning the locus was identified in a 7DS physical map (Tulpová et al. 2019a).

BAC clones from 0.83 cM interval, delimited by *Dn2401*-flanking markers, were sequenced by combination of short Illumina and long nanopore reads and the resulting sequence assembly, validated by optical mapping of the 7DS arm (Staňková et al. 2016), revealed six high-confidence genes. Comparison of 7DS-specific optical maps prepared from susceptible cv. CHINESE SPRING and resistant line CI2401 revealed structural variation in proximity of *Epoxide hydrolase 2*, which gave support to the gene as the most likely *Dn2401* candidate (Tulpová et al. 2019b). Similarly, a BAC library and physical map of CS 4A chromosome were used to approach and analyse pre-harvest sprouting resistance locus *Phs-A1*, which revealed a causal role of *TaMKK3-A* for the trait (Shorinola et al. 2017). Customized BAC libraries constructed from 3B chromosome of cv. HOPE and 4AL telosome bearing introgressed segment of *T. militinae* were utilized to clone stem rust resistance gene *Sr2* (Mago et al. 2014) and to approach powdery mildew resistance locus *Qpm.tut-4A* (Janáková et al. 2019), respectively.

3.5.2 Contemporary Approaches

The completion and release of the 'CHINESE SPRING' reference genome (IWGSC 2018) in hand with rapid technological advancements, allowing resequencing and large-scale pan-genome projects even in a crop with a complex polyploid genome, revolutionized strategies of gene cloning in bread wheat. Whole-genome long-read sequencing, resulting in high-quality sequence with resolved gene duplications, became realistic for wheat but challenges of producing, handling and analyzing the big data still appear too high for the majority of wheat gene cloning projects. Apart from the WGS and pan-genome efforts, several approaches to rapid gene cloning have been developed (Bettgenhaeuser and Krattinger 2019, and Chap. 10 of this book), including several utilizing the complexity reduction by chromosome flow sorting. Among them, Mutant Chromosome Sequencing (MutChromSeq; Sánchez-Martín et al. 2016) and TArgeted Chromosome-based Cloning via long-range Assembly (TACCA; Thind et al. 2017) have been used most widely. As indicated by the acronym, the former method couples chromosome flow sorting and sequencing with reference-free forward genetics. A chromosome bearing the gene of interest is Illumina-sequenced from both wild type and several independent ethyl methanesulfonate (EMS) mutants and the sequences are compared. A candidate gene is identified based on overlapping mutations in a genic region. The feasibility and efficiency of the method were first demonstrated by re-cloning barley *Eceriferum-q* gene and by de novo cloning wheat powdery mildew resistance gene *Pm2* (Sánchez-Martín et al. 2016). This speedy, cost-efficient approach to gene cloning generated a lot of interest in both wheat and barley community (reviewed in Steuernagel et al. 2017). It was successfully applied to identify the semi-dwarfism locus *Rht18* in *T. durum* (Ford et al. 2018) and the *SuSr-D1* gene that suppresses resistance to stem rust in bread wheat (Hiebert et al. 2020). Moreover, it contributed to cloning the race-specific leaf rust resistance gene *Lr14a* (Kolodziej et al. 2021) and the powdery mildew resistance gene *Pm4* (Sánchez-Martín et al. 2021) from hexaploid wheat.

MutChromSeq is a method of choice for traits with a strong phenotype, for which the production of independent mutants is feasible. As an alternative, suitable for any phenotype, Thind et al. (2017) proposed a procedure based on producing a high-quality de novo assembly of the gene-bearing chromosome and named it TACCA. The procedure utilized the so-called Chicago mapping technique (Putnam et al. 2016) developed by Dovetail Genomics. To clone leaf rust resistance gene *Lr22a*, the authors flow-sorted and Illumina-sequenced wheat chromosome 2D from resistant line CH CAMPALA *Lr22*a. The resulting sequences were scaffolded with Chicago long-range linkage. The assembly comprised 10,344 scaffolds with an N50 of 9.76 Mb and with the longest scaffold of 36.4 Mb. The high contiguity of the chromosomal assembly significantly reduced the number of markers needed to delimit the

gene in a narrow interval and, complemented by information from EMS mutants, allowed rapid cloning of this broad-spectrum resistance gene. The TACCA approach was also applied by Xing et al. (2018) to clone powdery mildew resistance gene *Pm21*, introduced to bread wheat from *H. villosa* chromosome 6V. Besides, the quality chromosomal assemblies generated by long-range linkage were used for comparative analyses with chromosomes of the wheat reference genome (Thind et al. 2018; Xing et al. 2021).

3.6 Conclusions and Perspectives

Since its establishment in 2000, flow-cytometric chromosome sorting contributed to major achievements in bread wheat genomics, including the generation of the wheat reference genome. Due to the rapid advancements in next-generation sequencing technologies, the reduction of genome complexity is no more essential in the context of whole-genome sequencing, but remains beneficial in gene cloning projects that call for a high-quality sequence from a narrow region of the genome. This demand was met in coupling chromosome sorting with the long-range linkage method, which resulted in contiguous chromosome assemblies. Since Dovetail Genomics discontinued the Chicago method, other approaches need to be developed to satisfy the demand of the wheat community. Long-read sequencing technologies, such as PacBio or nanopore sequencing, appear to be the logical tools for achieving the goal but to make them compatible with the flow-sorted material, challenges relating to inherent features of the flow sorting technique—formaldehyde fixation and a high laboriousness of producing large DNA amounts—still need to be resolved. Low-input protocols, being developed by the sequencing companies, go toward this demand.

Acknowledgements We wish to acknowledge the support by the European Regional Development Fund project 'Plants as a tool for sustainable global development' (No. CZ.02.1.01/0.0/0.0/16_019/0000827).

References

Abrouk M, Balcárková B, Šimková H, Komínkova E, Martis MM, Jakobson I, Timofejeva L, Rey E, Vrána J, Kilian A, Järve K, Doležel J, Valárik M (2017) The *in silico* identification and characterization of a bread wheat/*Triticum militinae* introgression line. Plant Biotechnol J 15:249–256

Akhunov ED, Sehgal S, Liang H, Wang S, Akhunova AR, Kaur G, Li W, Forrest KL, See D, Šimková H, Ma Y, Hayden MJ, Luo M, Faris JD, Doležel J, Gill BS (2013) Comparative analysis of syntenic genes in grass genomes reveals accelerated rates of gene structure and coding sequence evolution in polyploid wheat. Plant Physiol 161:252–265

Akpinar BA, Lucas SJ, Vrána J, Doležel J, Budak H (2015a) Sequencing chromosome 5D of *Aegilops tauschii* and comparison with its allopolyploid descendant bread wheat (*Triticum aestivum*) Plant Biotechnol J 13:740–752

Akpinar BA, Magni F, Yuce M, Lucas SJ, Šimková H, Šafář J, Vautrin S, Bergès H, Cattonaro F, Doležel J, Budak H (2015b) The physical map of wheat chromosome 5DS revealed gene duplications and small rearrangements. BMC Genomics 16:453

Akpinar BA, Yuce M, Lucas S, Vrána J, Burešová V, Doležel J, Budak H (2015c) Molecular organization and comparative analysis of chromosome 5B of the wild wheat ancestor *Triticum dicoccoides*. Sci Rep 5:10763

Akpinar BA, Biyiklioglu S, Alptekin B, Havránková M, Vrána J, Doležel J, Distelfeld A, Hernandez P, The IWGSC, Budak H (2018) Chromosome-based survey sequencing reveals the genome organization of wild wheat progenitor *Triticum dicoccoides*. Plant Biotechnol J 16:2077–2087

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815

Bansal M, Adamski NM, Toor PI, Kaur S, Molnár I, Holušová K, Vrána J, Doležel J, Valárik M, Uauy C, Chhuneja P (2020) *Aegilops umbellulata* introgression carrying leaf rust and stripe rust resistance genes *Lr76* and *Yr70* located to 9.47 Mb region on 5DS telomeric end through a combination of chromosome sorting and sequencing. Theor Appl Genet 133:903–915

Barabaschi D, Magni F, Volante A, Gadaleta A, Šimková H, Scalabrin S, Prazzoli ML, Bagnaresi P, Lacrima K, Michelotti V, Desiderio F, Orru L, Mazzamurro V, Fricano A, Mastrangelo AM, Tononi P, Vitulo N, Jurman I, Frenkel Z, Cattonaro F, Morgante M, Blanco A, Doležel J, Delledonne M, Stanca AM, Cattivelli L, Vale G (2015) Physical mapping of bread wheat chromosome 5A: an integrated approach. Plant Genome 8:1–24

Bartoš J, Vlček Č, Choulet F, Džunková M, Cviková K, Šafář J, Šimková H, Pačes J, Strnad H, Sourdille P, Bergés H, Cattonaro F, Feuillet C, Doležel J (2012) Intraspecific sequence comparisons reveal similar rates of non-collinear gene insertion in the B and D genomes of bread wheat. BMC Plant Biol 12:155

Berkman PJ, Skarshewski A, Lorenc MT, Lai K, Duran C, Ling EYS, Stiller J, Smits L, Imelfort M, Manoli S, McKenzie M, Kubaláková M, Šimková H, Batley J, Fleury D, Doležel J, Edwards D (2011) Sequencing and assembly of low copy and genic regions of isolated *Triticum aestivum* chromosome arm 7DS. Plant Biotechnol J 9:768–775

Berkman PJ, Skarshewski A, Manoli S, Lorenc MT, Stiller J, Smits L, Lai K, Campbell E, Kubaláková M, Šimková H, Batley J, Doležel J, Hernandez P, Edwards D (2012) Sequencing wheat chromosome arm 7BS delimits the 7BS/4AL translocation and reveals homoeologous gene conservation. Theor Appl Genet 124:423–432

Berkman PJ, Visendi P, Lee HC, Stiller J, Manoli S, Lorenc MT, Lai K, Batley J, Fleury D, Šimková H, Kubaláková M, Weining S, Doležel J, Edwards D (2013) Dispersion and domestication shaped the genome of bread wheat. Plant Biotechnol J 11:564–571

Bettgenhaeuser J, Krattinger SG (2019) Rapid gene cloning in cereals. Theor Appl Genet 132:699–711

Breen J, Li D, Dunn DS, Békés F, Kong X, Zhang J, Jia J, Wicker T, Mago R, Ma W, Bellgard M, Appels R (2010) Wheat beta-expansin (EXPB11) genes: identification of the expressed gene on chromosome 3BS carrying a pollen allergen domain. BMC Plant Biol 10:99

Breen J, Wicker T, Shatalina M, Frenkel Z, Bertin I, Philippe R, Spielmeyer W, Šimková H, Šafář J, Cattonaro F, Scalabrin S, Magni F, Vautrin S, Berges H, International Wheat Genome Sequencing Consortium, Paux E, Fahima T, Doležel J, Korol A, Feuillet C, Keller B (2013) A physical map of the short arm of wheat chromosome 1A. PLoS ONE 8:e80272

Carrano AV, Gray JW, Moore DH, Minkler JL, Mayall BH, Van Dilla MA, Mendelsohn ML (1976) Purification of the chromosomes of the Indian muntjac by flow sorting. J Histochem Cytochem 24:348–354

Choulet F, Wicker T, Rustenholz C, Paux E, Salse J, Leroy P, Schlub S, Le Paslier MC, Magdelenat G, Gonthier C, Couloux A, Budak H, Breen J, Pumphrey M, Liu S, Kong X, Jia J, Gut M, Brunel D, Anderson JA, Gill BS, Appels R, Keller B, Feuillet C (2010) Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. Plant Cell 22:1686–1701

Choulet F, Caccamo M, Wright J, Alaux M, Šimková H, Šafář J, Leroy P, Doležel J, Rogers J, Eversole K, Feuillet C (2014a) The Wheat Black Jack: advances towards sequencing the 21 chromosomes of bread wheat. In: Tuberosa R, Graner A, Frison E (eds) Genomics of plant genetic resources. Springer, Dordrecht, pp 405–438

Choulet F, Alberti A, Theil S, Glover N, Barbe V, Daron J, Pingault L, Sourdille P, Couloux A, Paux E, Leroy P, Mangenot S, Guilhot N, Le Gouis J, Balfourier F, Alaux M, Jamilloux V, Poulain J, Durand C, Bellec A, Gaspin C, Šafář J, Doležel J, Rogers J, Vandepoele K, Aury JM, Mayer K, Berges H, Quesneville H, Wincker P, Feuillet C (2014b) Structural and functional partitioning of bread wheat chromosome 3B. Science 345:1249721

Deng P, Nie X, Wang L, Cui L, Liu P, Tong W, Biradar SS, Edwards D, Berkman P, Šimková H, Doležel J, Luo M, You F, Batley J, Fleury D, Appels R, Weining S (2014) Computational identification and comparative analysis of miRNAs in wheat group 7 chromosomes. Plant Mol Biol Rep 32:487–500

Doležel J, Číhalíková J, Lucretti S (1992) A high-yield procedure for isolation of metaphase chromosomes from root tips of *Vicia faba* L. Planta 188:93–98

Doležel J, Vrána J, Šafář J, Bartoš J, Kubaláková M, Šimková H (2012) Chromosomes in the flow to simplify genome analysis. Funct Integr Genom 12:397–416

Doležel J, Lucretti S, Molnár I, Cápal P, Giorgi D (2021) Chromosome analysis and sorting. Cytometry 99:328–342

Feng K, Cui L, Wang L, Shan D, Tong W, Deng P, Yan Z, Wang M, Zhan H, Wu X, He W, Zhou X, Ji J, Zhang G, Mao L, Karafiátová M, Šimková H, Doležel J, Du X, Zhao S, Luo MC, Han D, Zhang C, Kang Z, Appels R, Edwards D, Nie X, Weining S (2020) The improved assembly of 7DL chromosome provides insight into the structure and evolution of bread wheat. Plant Biotechnol J 18:732–742

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM et al (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496–512

Ford BA, Foo E, Sharwood R, Karafiátová M, Vrána J, MacMillan C, Nichols DS, Steuernagel B, Uauy C, Doležel J, Chandler PM, Spielmeyer W (2018) *Rht18* semidwarfism in wheat is due to increased *GA 2-oxidaseA9* expression and reduced GA content. Plant Physiol 177:168–180

Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM, Fritchman RD, Weidman JF, Small KV, Sandusky M, Fuhrmann J, Nguyen D, Utterback TR, Saudek DM, Phillips CA, Merrick JM, Tomb JF, Dougherty BA, Bott KF, Hu PC, Lucier TS, Peterson SN, Smith HO, Hutchison CA 3rd, Venter JC (1995) The minimal gene complement of *Mycoplasma genitalium*. Science 270:397–403

Frenkel Z, Paux E, Mester D, Feuillet C, Korol A (2010) LTC a novel algorithm to improve the efficiency of contig assembly for physical mapping in complex genomes. BMC Bioinf 11:584–601

Garbus I, Romero JR, Valárik M, Vanžurová H, Karafiátová M, Cáccamo M, Doležel J, Tranquilli G, Helguera M, Echenique V (2015) Characterization of repetitive DNA landscape in wheat homeologous group 4 chromosomes. BMC Genomics 16:375

Gill BS, Appels R, Botha-Oberholster A-M, Buell CR, Bennetzen JL, Chalhoub B, Chumley F, Dvorák J, Iwanaga M, Keller B, Li W, McCombie WR, Ogihara Y, Quetier F, Sasaki T (2004) A workshop report on wheat genome sequencing: international genome research on Wheat Consortium. Genetics 168:1087–1096

Giorgi D, Farina A, Grosso V, Gennaro A, Ceoloni C, Lucretti S (2013) FISHIS: fluorescence *in situ* hybridization in suspension and chromosome flow sorting made easy. PLoS ONE 8:e5799

Gray JW, Carrano AV, Steinmetz LL, Van Dilla MA, Moore HH, Mayall BH, Mendelsohn ML (1975a) Chromosome measurement and sorting by flow systems. Proc Natl Acad Sci USA 72:1231–1234

Gray JW, Carrano AV, Moore HH, Steinmetz LL, Minkler J, Mayall BH, Mendelsohn ML, Van Dilla MA (1975b) High-speed quantitative karyotyping by flow microfluorometry. Clin Chem 21:1258–1262

Grosso V, Farina A, Gennaro A, Giorgi D, Lucretti S (2012) Flow sorting and molecular cytogenetic identification of individual chromosomes of *Dasypyrum villosum* L. (*H. villosa*) by a single DNA probe. PLoS ONE 7:e50151

Harrington SA, Cobo N, Karafiátová M, Doležel J, Borrill P, Uauy C (2019) Identification of a dominant chlorosis phenotype through a forward screen of the *Triticum turgidum* cv. Kronos TILLING Population. Front Plant Sci 10:963

Helguera M, Rivarola M, Clavijo B, Martis MM, Vanzetti LS, González S, Garbus I, Leroy P, Šimková H, Valárik M, Caccamo M, Doležel J, Mayer KFX, Feuillet C, Tranquilli G, Paniego N, Echenique V (2015) New insights into the wheat chromosome 4D structure and virtual gene order, revealed by survey pyrosequencing. Plant Sci 233:200–221

Hernandez P, Martis M, Dorado G, Pfeifer M, Gálvez S, Schaaf S, Jouve N, Šimková H, Valárik M, Doležel J, Mayer KFX (2012) Next-generation sequencing and syntenic integration of flow-sorted arms of wheat chromosome 4A exposes the chromosome structure and gene content. Plant J 69:377–386

Hewitt T, Müller MC, Molnár I, Mascher M, Holušová K, Šimková H, Kunz L, Zhang J, Li J, Bhatt D, Sharma R, Schudel S, Yu G, Steuernagel B, Periyannan S, Wulff B, Ayliffe M, McIntosh R, Keller B, Lagudah E, Zhang P (2021) A highly differentiated region of wheat chromosome 7AL encodes a Pm1a immune receptor that recognizes its corresponding AvrPm1a effector from *Blumeria graminis*. New Phytol 229:2812–2826

Hiebert CW, Moscou MJ, Hewitt T, Steuernagel B, Hernández-Pinzón I, Green P, Pujol V, Zhang P, Rouse MN, Jin Y, McIntosh RA, Upadhyaya N, Zhang J, Bhavani S, Vrána J, Karafiátová M, Huang L, Fetch T, Doležel J, Wulff BBH, Lagudah E, Spielmeyer W (2020) Stem rust resistance in wheat is suppressed by a subunit of the mediator complex. Nat Commun 11:1123

Holušová K, Vrána J, Šafář J, Šimková H, Balcárková B, Frenkel Z, Darrier B, Paux E, Cattonaro F, Berges H, Letellier T, Alaux M, Doležel J, Bartoš J (2017) Physical map of the short arm of bread wheat chromosome 3D. Plant Genome 10:1–11

International Wheat Genome Sequencing Consortium (2014) A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. Science 345:1251788

International Wheat Genome Sequencing Consortium (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science 361:7191

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature 436:793–800

Jaccoud D, Peng K, Feinstein D, Kilian A (2001) Diversity arrays: a solid state technology for sequence information independent genotyping. Nucl Acids Res 29:e25

Jakobson I, Reis D, Tiidema A, Peusha H, Timofejeva L, Valárik M, Kladivová M, Šimková H, Doležel J, Järve K (2012) Fine mapping, phenotypic characterization and validation of non-race-specific resistance to powdery mildew in a wheat-*Triticum militinae* introgression line. Theor Appl Genet 125:609–623

Janáková, E, Jakobson, I, Peusha, H, Abrouk, M, Škopová, M Šimková, H, Šafář, J, Vrána, J, Doležel, J Järve, K, Valárik, M (2019) Divergence between bread wheat and *Triticum militinae* in the powdery mildew resistance *QPmtut 4A* locus and its implications for cloning of the resistance gene. Theor Appl Genet 132:1061-1072

Janda J, Bartoš J, Šafář J, Kubaláková M, Valárik M, Číhalíková J, Šimková H, Caboche M, Sourdille P, Bernard M, Chalhoub B, Doležel J (2004) Construction of a subgenomic BAC library specific for chromosomes 1D, 4D, and 6D of hexaploid wheat. Theor Appl Genet 109:1337–1345

Janda J, Šafář J, Kubaláková M, Bartoš J, Kovářová P, Suchánková P, Pateyron S, Číhalíková J, Sourdille P, Šimková H, Faivre-Rampant P, Hřibová E, Bernard M, Lukaszewski A, Doležel J, Chalhoub B (2006) Advanced resources for plant genomics: BAC library specific for the short arm of wheat chromosome 1B. Plant J 47:977–986

Kantar M, Akpınar BA, Valárik M, Lucas SJ, Doležel J, Hernández P, Budak H, International Wheat Genome Sequencing Consortium (2012) Subgenomic analysis of microRNAs in polyploid wheat. Funct Integr Genomics 12:465–479

Kapustová V, Tulpová Z, Toegelová H, Novák P, Macas J, Karafiátová M, Hřibová E, Doležel J, Šimková H (2019) The dark matter of large cereal genomes: long tandem repeats. Int J Mol Sci 20:2483

Kaur P, Yadav IS, Yadav B, Mahato A, Gupta OP, Doležel J, Singh NK, Khurana JP, Singh K (2019) In silico

annotation of 458 genes identified from comparative analysis of Full length cDNAs and NextGen Sequence of chromosome 2A of hexaploid wheat. J Plant Biochem Biotechnol 28:25–34

Keeble-Gagnère G, Rigault P, Tibbits J, Pasam R, Hayden M, Forrest K, Frenkel Z, Korol A, Huang BE, Cavanagh C, Taylor J, Abrouk M, Sharpe A, Konkin D, Sourdille P, Darrier B, Choulet F, Bernard A, Rochfort S, Dimech A, Watson-Haigh N, Baumann U, Eckermann P, Fleury D, Juhasz A, Boisvert S, Nolin MA, Doležel J, Šimková H, Toegelová H, Šafář J, Luo MC, Câmara F, Pfeifer M, Isdale D, Nyström-Persson J, IWGSC, Koo DH, Tinning M, Cui D, Ru Z, Appels R (2018) Optical and physical mapping with local finishing enables megabase-scale resolution of agronomically important regions in the wheat genome. Genome Biol 19:112

Kobayashi F, Wu J, Kanamori H, Tanaka T, Katagiri S, Karasawa W, Kaneko S, Watanabe S, Sakaguchi T, Hanawa Y, Fujisawa H, Kurita K, Abe C, Iehisa JCM, Ohno R, Šafář J, Šimková H, Mukai Y, Hamada M, Saito M, Ishikawa G, Katayose Y, Endo TR, Takumi S, Nakamura T, Sato K, Ogihara Y, Hayakawa K, Doležel J, Nasuda S, Matsumoto T, Handa H (2015) A high-resolution physical map integrating an anchored chromosome with the BAC physical maps of wheat chromosome 6B. BMC Genomics 16:595

Kolodziej MC, Singla J, Sánchez-Martín J, Zbinden H, Šimková H, Karafiátová M, Doležel J, Gronnier J, Poretti M, Glauser G, Zhu W, Köster P, Zipfel C, Wicker T, Krattinger SG, Keller B (2021) A membrane-bound ankyrin repeat protein confers race-specific leaf rust disease resistance in wheat. Nature Comm 12:956

Konkin D, Hsueh Y-C, Kirzinger M, Kubaláková M, Haldar A, Balcerzak M, Han F, Fedak G, Doležel J, Sharpe A, Ouellet T (2022) Genomic sequencing of *Thinopyrum elongatum* chromosome arm 7EL, carrying fusarium head blight resistance, and characterization of its impact on the transcriptome of the introgressed line CS-7EL. BMC Genomics 23:228

Kubaláková M, Vrána J, Čihalíková J, Šimková H, Doležel J (2002) Flow karyotyping and chromosome sorting in bread wheat (*Triticum aestivum* L.). Theor Appl Genet 104:1362–1372

Kubaláková M, Kovářová P, Suchánková P, Číhalíková J, Bartoš J, Lucretti S, Watanabe N, Kianian SF, Doležel J (2005) Chromosome sorting in tetraploid wheat and its potential for genome analysis. Genetics 170:823–829

Lam ET, Hastie A, Lin C, Ehrlich D, Das SK, Austin MD, Deshpande P, Cao H, Nagarajan N, Xiao M, Kwok PY (2012) Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nat Biotechnol 30:771–776

Lander ES, Linton LM, Birren B et al (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

Li B, Choulet F, Heng Y, Hao W, Paux E, Liu Z, Yue W, Jin W, Feuillet C, Zhang X (2013) Wheat centromeric retrotransposons: the new ones take a major role in centromeric structure. Plant J 73:952–965

Liu M, Stiller J, Holušová K, Vrána J, Liu D, Doležel J, Liu C (2016) Chromosome-specific sequencing reveals an extensive dispensable genome component in wheat. Sci Rep 6:36398

Lucas SJ, Šimková H, Šafář J, Jurman I, Cattonaro F, Vautrin S, Bellec A, Berges H, Doležel J, Budak H (2012) Functional features of a single chromosome arm in wheat (1AL) determined from its structure. Funct Integr Genom 12:173–182

Lucas SJ, Akpınar BA, Kantar M, Weinstein Z, Aydınoglu F, Šafář J, Šimková H, Frenkel Z, Korol A, Magni F, Cattonaro F, Vautrin S, Bellec A, Berges H, Doležel J, Budak H (2013) Physical mapping integrated with syntenic analysis to characterize the gene space of the long arm of wheat chromosome 1A. PLoS ONE 8:e59542

Lucas SJ, Akpınar BA, Šimková H, Kubaláková M, Doležel J, Budak H (2014) Next-generation sequencing of flow-sorted wheat chromosome 5D reveals lineage-specific translocations and widespread gene duplications. BMC Genomics 15:1080

Luo M-C, Thomas C, You FM, Hsiao J, Ouyang S, Buel CR, Malandro M, McGuire PE, Anderson OD, Dvorak J (2003) High-throughput fingerprinting of bacterial artificial chromosomes using the SNaPshot labelling kit and sizing of restriction fragments by capillary electrophoresis. Genomics 82:378–389

Luo M-C, Ma Y, You FM, Anderson OD, Kopecký D, Šimková H, Šafář J, Doležel J, Gill B, McGuire P, Dvořák J (2010) Feasibility of physical map construction from fingerprinted bacterial artificial chromosome libraries of polyploid plant species. BMC Genomics 11:122

Ma J, Stiller J, Wei Y, Zheng YL, Devos KM, Doležel J, Liu C (2014) Extensive pericentric rearrangements in the bread wheat (*Triticum aestivum* L.) genotype "Chinese Spring" revealed from chromosome shotgun sequence data. Genome Biol Evol 6:3039–3048

Ma J, Stiller J, Zheng Z, Wei Y, Zheng Y-L, Yan G, Doležel J, Liu C (2015) Putative interchromosomal rearrangements in the hexaploid wheat (*Triticum aestivum* L.) genotype 'Chinese Spring' revealed by gene locations on homoeologous chromosomes. BMC Evol Biol 15:37

Mago R, Tabe L, Vautrin S, Šimková H, Kubaláková M, Upadhyaya N, Berges H, Kong X, Breen J, Doležel J, Appels R, Ellis JG, Spielmeyer W (2014) Major haplotype divergence including multiple germin-like protein genes, at the wheat *Sr2* adult plant stem rust resistance locus. BMC Plant Biol 14:379

Marcussen T, Sandve SR, Heier L, Spannagl M, Pfeifer M, Rogers J, Doležel J, Pozniak C, Eversole K, Feuillet C, Gill B, Friebe B, Lukaszewski AJ, Sourdille P, Endo TR, Kubaláková M, Číhalíková J,

Dubská Z, Vrána J, Šperková R, Šimková H, Febrer M, Clissold L, McLay K, Singh K, Chhuneja P, Singh NK, Khurana J, Akhunov E, Choulet F, Alberti A, Barbe V, Wincker P, Kanamori H, Kobayashi F, Itoh T, Matsumoto T, Sakai H, Tanaka T, Wu J, Ogihara Y, Handa H, Maclachlan PR, Sharpe A, Klassen D, Edwards D, Batley J, Olsen OA, Sandve SR, Lien S, Steuernagel B, Wulff B, Caccamo M, Ayling S, Ramirez-Gonzalez RH, Clavijo BJ, Wright J, Pfeifer M, Spannagl M, Martis MM, Mascher M, Chapman J, Poland JA, Scholz U, Barry K, Waugh R, Rokhsar DS, Muehlbauer GJ, Stein N, Gundlach H, Zytnicki M, Jamilloux V, Quesneville H, Wicker T, Faccioli P, Colaiacovo M, Stanca AM, Budak H, Cattivelli L, Glover N, Pingault L, Paux E, Sharma S, Appels R, Bellgard M, Chapman B, Nussbaumer T, Bader KC, Rimbert H, Wang S, Knox R, Kilian A, Alaux M, Alfam F, Couderc L, Guilhot N, Viseux C, Loaec M, Keller B, Praud S, Jakobsen KS, Wulff BB, Steuernagel B, Mayer KF, Olsen OA (2014) Ancient hybridizations among the ancestral genomes of bread wheat. Science 345:1250092

Mayer KFX, Taudien S, Martis M, Šimková H, Suchánková P, Gundlach H, Wicker T, Petzold A, Felder M, Steuernagel B, Scholz U, Graner A, Platzer M, Doležel J, Stein N (2009) Gene content and virtual gene order of barley chromosome 1H. Plant Physiol 151:496–505

Mayer KFX, Martis M, Hedley PE, Šimková H, Liu H, Morris JA, Steuernagel B, Taudien S, Roessner S, Gundlach H, Kubaláková M, Suchánková P, Murat F, Felder M, Nussbaumer T, Graner A, Salse J, Endo TR, Sakai H, Tanaka T, Itoh T, Sato K, Platzer M, Matsumoto T, Scholz U, Doležel J, Waugh R, Stein N (2011) Unlocking the barley genome by chromosomal and comparative genomics. Plant Cell 23:1249–1263

McNeil M, Kota R, Paux E, Dunn D, McLean R, Feuillet C, Li D, Kong X, Lagudah E, Zhang JC, Jia JZ, Spielmeyer W, Bellgard M, Appels R (2008) BAC-derived markers for assaying the stem rust resistance gene, *Sr2*, in wheat breeding programs. Mol Breeding 22:15–24

Michalak de Jimenez MK, Bassi FM, Ghavami F, Simons K, Dizon R, Seetan RI, Alnemer LM, Denton AM, Doğramacı M, Šimková H, Doležel J, Seth K, Luo M-C, Dvorak J, Gu YQ, Kianian SF (2013) A radiation hybrid map of chromosome 1D reveals synteny conservation at a wheat speciation locus. Funct Integr Genom 13:19–32

Molnár I, Kubaláková M, Šimková H, Cseh A, Molnár-Láng M, Doležel J (2011) Chromosome isolation by flow sorting in *Aegilops umbellulata* and *Ae. comosa* and their allotetraploid hybrids *Ae. biuncialis* and *Ae. geniculata.* PLoS ONE 6:e27708

Molnár I, Kubaláková M, Šimková H, Farkas A, Cseh A, Megyeri M, Vrána J, Molnár-Láng M, Doležel J (2014) Flow cytometric chromosome sorting from diploid progenitors of bread wheat, *T. urartu*, *Ae. speltoides* and *Ae. tauschii*. Theor Appl Genet 127:1091–1104

Molnár I, Vrána J, Farkas A, Kubaláková M, Sceh A, Molnár-Láng M, Doležel J (2015) Flow sorting of C-genome chromosomes from wild relatives of wheat *Aegilops markgrafii*, *Ae. triuncialis* and *Ae. cylindrica*, and their molecular organization. Annals Bot 116:189–200

Molnár I, Vrána J, Burešová V, Cápal P, Farkas A, Darkó É, Cseh A, Kubaláková M, Molnár-Láng M, Doležel J (2016) Dissecting the U, M, S and C genomes of wild relatives of bread wheat (*Aegilops* spp.) into chromosomes and exploring their synteny with wheat. Plant J 88:452–467

Nsabiyera V, Baranwal D, Qureshi N, Kay P, Forrest K, Valárik M, Doležel J, Hayden MJ, Bariana HS, Bansal UK (2020) Fine mapping of *Lr49* using 90K SNP Chip Array and flow-sorted chromosome sequencing in wheat. Front Plant Sci 10:1787

Paux E, Roger D, Badaeva E, Gay G, Bernard M, Sourdille P, Feuillet C (2006) Characterizing the composition and evolution of homoeologous genomes in hexaploid wheat through BAC-end sequencing on chromosome 3B. Plant J 48:463–474

Paux E, Sourdille P, Salse J, Saintenac C, Choulet F, Leroy P, Korol A, Michalak M, Kianian S, Spielmeyer W, Lagudah E, Somers D, Kilian A, Alaux M, Vautrin S, Berges H, Eversole K, Appels R, Šafář J, Šimková H, Doležel J, Bernard M, Feuillet C (2008) A physical map of the 1-gigabase bread wheat chromosome 3B. Science 322:101–104

Pfeifer M, Kugler KG, Sandve SR, Zhan B, Rudi H, Hvidsten TR, Rogers J, Doležel J, Pozniak C, Eversole K, Feuillet C, Gill B, Friebe B, Lukaszewski AJ, Sourdille P, Endo TR, Kubaláková M, Číhalíková J, Dubská Z, Vrána J, Šperková R, Šimková H, Febrer M, Clissold L, McLay K, Singh K, Chhuneja P, Singh NK, Khurana J, Akhunov E, Choulet F, Alberti A, Barbe V, Wincker P, Kanamori H, Kobayashi F, Itoh T, Matsumoto T, Sakai H, Tanaka T, Wu J, Ogihara Y, Handa H, Maclachlan PR, Sharpe A, Klassen D, Edwards D, Batley J, Olsen OA, Sandve SR, Lien S, Steuernagel B, Wulff B, Caccamo M, Ayling S, Ramirez-Gonzalez RH, Clavijo BJ, Wright J, Pfeifer M, Spannagl M, Martis MM, Mascher M, Chapman J, Poland JA, Scholz U, Barry K, Waugh R, Rokhsar DS, Muehlbauer GJ, Stein N, Gundlach H, Zytnicki M, Jamilloux V, Quesneville H, Wicker T, Faccioli P, Colaiacovo M, Stanca AM, Budak H, Cattivelli L, Glover N, Pingault L, Paux E, Sharma S, Appels R, Bellgard M, Chapman B, Nussbaumer T, Bader KC, Rimbert H, Wang S, Knox R, Kilian A, Alaux M, Alfam F, Couderc L, Guilhot N, Viseux C, Loaec M, Keller B, Praud S, Mayer KF, Olsen OA (2014) Genome interplay in the grain transcriptome of hexaploid bread wheat. Science 345:1250091

Philippe R, Paux E, Bertin I, Sourdille P, Choulet F, Laugier C, Šimková H, Šafář J, Bellec A, Vautrin S,

Frenkel Z, Cattonaro F, Magni F, Scalabrin S, Martis MM, Mayer KFX, Korol A, Bergès H, Doležel J, Feuillet C (2013) A high density physical map of chromosome 1BL supports evolutionary studies, map-based cloning and sequencing in wheat. Genome Biol 14:R64

Poursarebani N, Nussbaumer T, Šimková H, Šafář J, Witsenboer H, van Oeveren J, Doležel J, Mayer KFX, Stein N, Schnurbusch T (2014) Whole-genome profiling and shotgun sequencing delivers an anchored, gene-decorated, physical map assembly of bread wheat chromosome 6A. Plant J 79:334–347

Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, Troll CJ, Fields A, Hartley PD, Sugnet CW, Haussler D, Rokhsar DS, Green RE (2016) Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome Res 26:342–350

Raats D, Frenkel Z, Krugman T, Dodek I, Sela H, Šimková H, Magni F, Cattonaro F, Vautrin S, Bergès H, Wicker T, Keller B, Leroy P, Philippe R, Paux E, Doležel J, Feuillet C, Korol A, Fahima T (2013) The physical map of wheat chromosome 1BS provides insights into its gene space organization and evolution. Genome Biol 14:R138

Randhawa M, Bansal U, Valárik M, Klocová B, Doležel J, Bariana H (2014) Molecular mapping of stripe rust resistance gene *Yr51* in chromosome 4AL of wheat. Theor Appl Genet 127:317–324

Šafář J, Bartoš J, Janda J, Bellec A, Kubaláková M, Valárik M, Pateyron S, Weiserová J, Tušková R, Číhalíková J, Vrána J, Šimková H, Faivre-Rampant P, Sourdille P, Caboche M, Bernard M, Doležel J, Chalhoub B (2004) Dissecting large and complex genomes: flow sorting and BAC cloning of individual chromosomes from bread wheat. Plant J 39:960–968

Šafář J, Šimková H, Kubaláková M, Číhalíková J, Suchánková P, Bartoš J, Doležel J (2010) Development of chromosome-specific BAC resources for genomics of bread wheat. Cytogenet Genome Res 129:211–223

Said M, Kubaláková M, Karafiátová M, Molnár I, Doležel J Vrána J (2019) Dissecting the complex genome of crested wheatgrass by chromosome flow sorting. Plant Genome 12:180096

Said M, Holušová K, Farkas A, Ivanizs L, Gaál E, Cápal P, Abrouk M, Martis-Thiele MM, Kalapos B, Bartoš J, Friebe B, Doležel J, Molnár I (2021) Development of DNA markers from physically mapped loci in *Aegilops comosa* and *Aegilops umbellulata* using single-gene FISH and chromosome sequences. Front Plant Sci 12:689031

Salina EA, Nesterov MA, Frenkel Z, Kiseleva AA, Timonova EM, Magni F, Vrána J, Šafář J, Šimková H, Doležel J, Korol A, Sergeeva EM (2018) Features of the organization of bread wheat chromosome 5BS based on physical mapping. BMC Genomics 19:80

Sánchez-Martín J, Steuernagel B, Ghosh S, Herren G, Hurni S, Adamski N, Vrána J, Kubaláková M, Krattinger SG, Wicker T, Doležel J, Keller B, Wulff BBH (2016) Rapid gene isolation in barley and wheat by mutant chromosome sequencing. Genome Biol 17:221

Sánchez-Martín J, Widrig V, Herren G, Wicker T, Zbinden H, Gronnier J, Spörri L, Praz CR, Heuberger M, Kolodziej MC, Isaksson J, Steuernagel B, Karafiátová M, Doležel J, Zipfel C, Keller B (2021) Wheat Pm4 resistance to powdery mildew is controlled by alternative splice variants encoding chimeric proteins. Nature Plants 7:327–341

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74:5463–5467

Sears ER, Sears LMS (1978) The telocentric chromosomes of common wheat. In: Ramanujams S (ed) Proceedings of 5th international wheat genetics symposium, pp 389–407. Indian Agricultural Research Institute, New Delhi

Sehgal SK, Li W, Rabinowicz PD, Chan A, Šimková H, Doležel J, Gill BS (2012) Chromosome arm-specific BAC end sequences permit comparative analysis of homoeologous chromosomes and genomes of polyploid wheat. BMC Plant Biol 12:64

Sergeeva EM, Afonnikov DA, Koltunova MK, Gusev VD, Miroshnichenko LA, Vrána J, Kubaláková M, Poncet C, Sourdille P, Feuillet C, Doležel J, Salina EA (2014) Common wheat chromosome 5B composition analysis using low-coverage 454 sequencing. Plant Genome 7:1–16

Shatalina M, Wicker T, Buchmann JP, Oberhaensli S, Šimková H, Doležel J, Keller B (2013) Genotype-specific SNP map based on whole chromosome 3B sequence information from wheat cultivars Arina and Forno. Plant Biotechnol J 11:23–32

Shatalina M, Messmer M, Feuillet C, Mascher F, Paux E, Choulet F, Wicker T, Keller B (2014) High-resolution analysis of a QTL for resistance to *Stagonospora nodorum* glume blotch in wheat reveals presence of two distinct resistance loci in the target interval. Theor Appl Genet 127:573–586

Shorinola O, Balcárková B, Hyles J, Tibbits JFG, Hayden MJ, Holušová K, Valárik M, Distelfeld A, Torada A, Barrero JM, Uauy C (2017) Haplotype analysis of the pre-harvest sprouting resistance locus *Phs-A1* reveals a causal role of TaMKK3-A in global germplasm. Front Plant Sci 8:1555

Šimková H, Číhalíková J, Vrána J, Lysák MA, Doležel J (2003) Preparation of HMW DNA from plant nuclei and chromosomes isolated from root tips. Biol Plant 46:369–373

Šimková H, Svensson JT, Condamine P, Hřibová E, Suchánková P, Bhat PR, Bartoš J, Šafář J, Close TJ, Doležel J (2008) Coupling amplified DNA from flow-sorted chromosomes to high-density SNP mapping in barley. BMC Genomics 9:294

Šimková H, Šafář J, Kubaláková M, Suchánková P, Číhalíková J, Robert-Quatre H, Azhaguvel P, Weng Y, Peng J, Lapitan NLV, Ma Y, You FM, Luo M-Ch,

Bartoš J, Doležel J (2011) BAC Libraries from wheat chromosome 7D: efficient tool for positional cloning of aphid resistance genes. J Biomed Biotechnol:302543

Staňková H, Valárik M, Lapitan NLV, Berkman PJ, Batley J, Edwards D, Luo M-C, Tulpová Z, Kubaláková M, Stein N, Doležel J, Šimková H (2015) Chromosomal genomics facilitates fine mapping of a Russian wheat aphid resistance gene. Theor Appl Genet 128:1373–1383

Staňková H, Hastie AR, Chan S, Vrána J, Tulpová Z, Kubaláková M, Visendi P, Hayashi S, Luo M, Batley J, Edwards D, Doležel J, Šimková H (2016) BioNano genome mapping of individual chromosomes supports physical mapping and sequence assembly in complex plant genomes. Plant Biotechnol J 14:1523–1531

Steuernagel B, Vrána J, Karafiátová M, Wulff BBH, Doležel J (2017) Rapid gene isolation using MutChromSeq. In: Periyannan S (ed) Wheat rust diseases: methods and protocols. Springer, Dordrecht, pp 231–243

Stubblefield E, Cram S, Deaven L (1975) Flow microfluorometric analysis of isolated Chinese-hamster chromosomes. Exp Cell Res 94:464–468

Tanaka T, Kobayashi F, Joshi GP, Onuki R, Sakai H, Kanamori H, Wu J, Šimková H, Nasuda S, Endo TR, Hayakawa K, Doležel J, Ogihara Y, Itoh T, Matsumoto T, Handa H (2014) Next-generation survey sequencing and the molecular organization of wheat chromosome 6B. DNA Res 21:103–114

Terracciano I, Maccaferri M, Bassi F, Mantovani P, Sanguineti MC, Salvi S, Šimková H, Doležel J, Massi A, Ammar K, Kolmer J, Tuberosa R (2013) Development of COS-SNP and HRM markers for high-throughput and reliable haplotype-based detection of *Lr14a* in durum wheat (*Triticum durum* Desf). Theor Appl Genet 126:1077–1101

Thind AK, Wicker T, Šimková H, Fossati D, Moullet O, Brabant C, Vrána J, Doležel J, Krattinger SG (2017) Rapid cloning of genes in hexaploid wheat using cultivar-specific long-range chromosome assembly. Nat Biotechnol 35:793–796

Thind AK, Wicker T, Müller T, Ackermann PM, Steuernagel B, Wulff BBH, Spannagl M, Twardziok SO, Felder M, Lux T, Mayer KFX International Wheat Genome Sequencing Consortium, Keller B, Krattinger SG (2018) Chromosome-scale comparative sequence analysis unravels molecular mechanisms of genome dynamics between two wheat cultivars. Genome Biol 19:104

Tiwari VK, Wang S, Sehgal S, Vrána J, Friebe B, Kubaláková M, Chhuneja P, Doležel J, Akhunov E, Kalia B, Sabir J, Gill BS (2014) SNP discovery for mapping alien introgressions in wheat. BMC Genomics 15:273

Tiwari VK, Wang S, Danilova T, Koo DH, Vrána J, Kubaláková M, Hřibová E, Rawat N, Kalia B, Singh N, Friebe B, Doležel J, Akhunov E, Poland J, Sabir JSM, Gill BS (2015) Exploring the tertiary gene pool of bread wheat: sequence assembly and analysis of chromosome 5Mg of *Aegilops geniculata*. Plant J 84:733–746

Tulpová Z, Luo MC, Toegelová H, Visendi P, Hayashi S, Vojta P, Paux E, Kilian A, Abrouk M, Bartoš J, Hajdúch M, Batley J, Edwards D, Doležel D, Šimková H (2019a) Integrated physical map of bread wheat chromosome arm 7DS to facilitate gene cloning and comparative studies. New Biotechnol 48:12–19

Tulpová Z, Toegelová H, Lapitan NLV, Peairs FB, Macas J, Novák P, Lukaszewski AJ, Kopecký D, Mazáčová M, Vrána J, Holušová K, Leroy P, Doležel J, Šimková H (2019b) Accessing a Russian wheat aphid resistance gene in bread wheat by long-read technologies. Plant Genome 12:180065

Tulpová Z, Kovařík K, Toegelová H, Navrátilová P, Kapustová V, Hřibová E, Vrána J, Macas J, Doležel J, Šimková H (2022) Fine structure and transcription dynamics of bread wheat ribosomal DNA loci deciphered by a multi-omics approach. Plant Genome 15:e20191

Van Dilla MA, Deaven LL (1990) Construction of gene libraries for each human chromosome. Cytometry 11:208–218

van Oeveren J, de Ruiter M, Jesse T, van der Poel H, Tang J, Yalcin F, Janseen A, Volpin H, Stormo KE, Bogden R, van Eijk MJ, Prins M (2011) Sequence-based physical mapping of complex genomes by whole genome profiling. Genome Res 21:618–625

Vitulo N, Albiero A, Forcato C, Campagna D, Dal Pero F, Bagnaresi P, Colaiacovo M, Faccioli P, Lamontanara A, Šimková H, Kubaláková M, Perrotta G, Facella P, Lopez L, Pietrella M, Gianese G, Doležel J, Giuliano G, Cattivelli L, Valle G, Stanca AM (2011) First survey of the wheat chromosome 5A composition through a next generation sequencing approach. PLoS ONE 6:e26421

Vrána J, Kubaláková M, Šimková H, Číhalíková J, Lysák MA, Doležel J (2000) Flow-sorting of mitotic chromosomes in common wheat (*Triticum aestivum* L.). Genetics 156:2033–2041

Vrána J, Kubaláková M, Číhalíková J, Valárik M, Doležel J (2015) Preparation of sub-genomic fractions enriched for particular chromosomes in polyploid wheat. Biol Plant 59:445–455

Wang H, Dai K, Xiao J, Yuan C, Zhao R, Doležel J, Wu Y, Cao A, Chen P, Zhang S, Wang X (2017) Development of intron targeting (IT) markers specific for chromosome arm 4VS of *Haynaldia villosa* by chromosome sorting and next-generation sequencing. BMC Genomics 18:167

Wenzl P, Suchánková P, Carling J, Šimková H, Huttner E, Kubaláková M, Sourdille P, Paul E, Feuillet C, Kilian A, Doležel J (2010) Isolated chromosomes as a new and efficient source of DArT markers for the saturation of genetic maps. Theor Appl Genet 121:465–474

Wicker T, Mayer KFX, Gundlach H, Martis M, Steuernagel B, Scholz U, Šimková H, Kubaláková M, Choulet F, Taudien S, Platzer M, Feuillet C, Fahima

T, Budak H, Doležel J, Keller B, Stein N (2011) Frequent gene movement and pseudogene evolution is common to the large and complex genomes of wheat, barley, and their relatives. Plant Cell 23:1706–1718

Xiao J, Dai K, Fu L, Vrána J, Kubaláková M, Wan W, Sun H, Zhao J, Yu C, Wu Y, Abrouk M, Wang H, Doležel J, Wang X (2017) Sequencing flow-sorted short arm of *Haynaldia villosa* chromosome 4V provides insights into its molecular structure and virtual gene order. BMC Genomics 18:791

Xing L, Hu P, Liu J, Witek K, Zhou S, Xu J, Zhou W, Gao L, Huang Z, Zhang R, Wang X, Chen P, Wang H, Jones JDG, Karafiátová M, Vrána J, Bartoš J, Doležel J, Tian Y, Wu Y, Cao A (2018) *Pm21* from *Haynaldia villosa* encodes a CC-NBS-LRR protein conferring powdery mildew resistance in wheat. Mol Plant 11:874–878

Xing L, Yuan L, Lv Z, Wang Q, Yin C, Huang Z, Liu J, Cao S, Zhang R, Chen P, Karafiátová M, Vrána J, Bartoš J, Doležel J, Cao A (2021) Long-range assembly of sequences helps to unravel the genome structure and small variation of the wheat-*Haynaldia villosa* translocated chromosome 6VS.6AL. Plant Biotechnol J 19:1567–1578

Zhang X, Wan W, Li M, Yu Z, Liu J, Holušová K, Vrána J, Doležel J, Wu Y, Wang H, Xiao J, Wang X (2021) Targeted sequencing of the short arm of chromosome 6V of a wheat relative *Haynaldia villosa* for marker development and gene mining. Agronomy 11:1695

Zwyrtková J, Blavet N, Doležalová A, Cápal P, Said M, Molnár I, Vrána J, Doležel J, Hřibová E (2022) Draft sequencing a chromosomes identified evolutionary structural changes and genes and facilitated the development of SSR markers. Int J Mol Sci 23:3191

4 Structural and Functional Annotation of the Wheat Genome

Frédéric Choulet, Xi Wang, Manuel Spannagl, David Swarbreck, Hélène Rimbert, Philippe Leroy, Pauline Lasserre-Zuber and Nathan Papon

Abstract

Wheat genome sequencing has passed through major steps in a decade, starting from the sequencing of large contiguous sequences obtained from chromosome-specific BAC libraries, to reach high-quality genome assemblies of a dozen of bread wheat varieties and wild relatives. While access to an assembled genome sequence is crucial for research, the resource that is mainly used by the community is not the sequence itself, but rather the annotated features, i.e., genes and transposable elements. In this chapter, we describe the work performed to predict the repertoire of 107 k high-confidence genes and 4 million TE copies in the hexaploid wheat genome (cultivar CHINESE SPRING; IWGSC RefSeq) and the procedures established to transfer the annotation through the different releases of genome assembly. Limitations and implications for building a wheat pangenome are discussed, as well as the possibilities for future improvements of structural annotation, and opportunities offered by novel approaches for functional annotation.

Keywords

Wheat genome · Annotation · Gene function · Transposable elements

F. Choulet (✉) · H. Rimbert · P. Leroy · P. Lasserre-Zuber · N. Papon
UCA, INRAE, GDEC, Clermont-Ferrand, France
e-mail: frederic.choulet@inrae.fr

H. Rimbert
e-mail: helene.rimbert@inrae.fr

P. Leroy
e-mail: philippe.leroy.2@inrae.fr

P. Lasserre-Zuber
e-mail: pauline.lasserre-zuber@inrae.fr

N. Papon
e-mail: nathan.papon@inrae.fr

X. Wang
BASF Belgium Coordination Center CommV, Trait Research, Gent Zwijnaarde, Belgium
e-mail: xi.wang@basf.com

M. Spannagl
PGSB Plant Genome and Systems Biology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany
e-mail: manuel.spannagl@helmholtz-muenchen.de

D. Swarbreck
Earlham Institute, Norwich Research Park, Norwich, Norfolk, UK
e-mail: david.swarbreck@earlham.ac.uk

4.1 Introduction

The International Wheat Genome Sequencing Consortium (IWGSC; http://www.wheat-genome.org) was launched in 2005 with the aim of accelerating research in wheat by delivering molecular markers and genomic resources with

R. Appels et al. (eds.), *The Wheat Genome*, Compendium of Plant Genomes,
https://doi.org/10.1007/978-3-031-38294-9_4

the long-term goal of getting a high-quality reference genome sequence for the hexaploid wheat (Feuillet and Eversole 2007). It represents more than a decade of coordinated efforts from the completion of the first chromosome-specific BAC library construction (Paux et al. 2008) to the assembly of the 21 chromosome sequences of cultivar CHINESE SPRING (IWGSC 2018). Since the first release in 2018, the IWGSC integrated additional information coming from optical mapping and long reads in order to improve the quality of the assembly by correcting misordered scaffolds and filling gaps. This led to release RefSeq v2.0 and v2.1 in 2021 (Zhu et al. 2021).

Besides the methodological challenge of assembling this genome, the work performed to deliver an annotation is not well known and often poorly considered. Annotation consists of the identification of sequence features providing biological information, and it represents one of the most difficult tasks in genome sequencing projects. It is far from being obvious. However, annotation is the data mostly accessed by users, contrary to the genome sequence. Achieving a robust structural and functional genome sequence annotation is, thus, essential to provide the foundation for further relevant biological studies (Yandell and Ence 2012). Annotation of the RefSeq v1.0 required the coordinated effort of the IWGSC Annotation Group, bringing together researchers from three different Institutes: GDEC (France), PGSB (Germany), and Earlham Institute (UK). In addition, after the first release of the annotation, additional work has been performed in order to incorporate manual curation, and especially to update the annotation following changes to the genome assembly. This was achieved by developing fine-tuned bioinformatics approaches.

In this chapter, we present an overview of the processes that were established in order to release the first version of the annotation of RefSeq v1.0 and the updates since the first version. Besides the description of the work performed, this chapter is also a current opinion to consider the degree of approximation, the limits of the resources available and used for downstream analyses, and thus, a critical view of the quality of the data. The chapter also includes the plans for future versions not only for the structural annotation, but also for functional annotation.

4.2 Methods, Strategies, Resources for Structural Annotation of Genomes and Their Implications in Wheat Pangenomics

4.2.1 General Aspects of Structural Annotation

Depending on the sequence features targeted for study, and depending on the organism, genome annotation can be either trivial or complicated. This is why there may be a confusion for non-experts who may believe annotation is routine in genome sequencing projects. This is not the case for many species, and especially, this was not the case for wheat. For instance, in compact bacterial genomes, coding genes are intronless and represent the very wide majority of the genome so that predicting the presence of coding open reading frames is obvious and does not even require human curation. For species already widely studied, like in human for instance, with several genomes already assembled and annotated, annotation may be routine since it is based purely on similarity with available highly conserved genomes. The difficulty of annotation increases with the size of the genome, the repeat content and active transposable element (TE) expression, the ploidy, the fragmentation of coding genes into small exons, and with the phylogenetic distance to an already well-characterized genome. The difficulty also increases with the level of conservation of the predicted features. A protein-coding gene highly conserved among distant species will be easily predicted with high confidence, while predicting poorly conserved features with a high level of accuracy is more complicated.

Annotation relies on the combination of approaches: (i) the homology-based method using alignment/mapping algorithms searching for sequence similarity either with proteins, showing that a sequence is conserved across evolution, and/or transcriptomic data, showing that a sequence is expressed; (ii) the ab initio methods, i.e., predictions using statistical models such as hidden Markov models (HMMs); (iii) structural feature-based method through the identification of intrinsic information like motifs at the borders of transposons. It thus relies on a combination of software, algorithms, and adapted reference libraries. Annotation needs to be automated, i.e., performed through a pipeline that combines all different programs and minimizes the subsequent long and laborious step of manual curation.

4.2.2 Sequence Features Usually Annotated and Common Ambiguities

In the plant genomics area, publications usually report on genes and repeats. Both terms are, however, confusing and the shortcut widely accepted by the community to distinguish genes and repeats is ambiguous. First, for convenience, the term "gene" is used as a shortcut for protein-coding gene. It will be the case in this chapter too. When a "number of genes" is given, it nearly always refers to a number of protein-coding genes. However, genomes also carry non-coding RNA (ncRNA) genes which are biologically important. In the annotation area, we distinguish two types of non-coding RNA genes: (i) highly conserved ncRNAs involved in essential cellular processes (splicing, translation) which are ribosomal RNAs, transfer RNAs, small nuclear and nucleolar RNAs, and (ii) less evolutionary conserved ncRNAs like micro-RNAs, long-non-coding RNAs, and others involved in specific regulation processes. Annotating conserved and non-conserved ncRNAs follows two completely different approaches. rRNA, tRNAs, snoRNAs, snRNAs are easily identified by a simple similarity-search approach; however, they tend not to be annotated. The reason for that is probably that they are of interest only for research groups working specifically on them and that are able to identify them with specific tools. In contrast, annotation is much more complicated for the species-specific ncRNAs. It requires the availability of small RNASeq reads that could be mapped to identify transcribed regions as a first clue before concluding to the presence of an ncRNA gene. Second, genes are repeats. In bread wheat, the majority of the "genes" are repeated with only 17% (30,948/181,036) of single-copy genes (IWGSC 2018) so reference to genes versus repeats brings confusion particularly when some repeats carry genes. "Repeats" is a general term encompassing simple repeats as satellite DNA, telomeric repeated motifs, but also transposable elements (TEs), and their mobilizable or inactive derivatives. Usually in plant genome annotation, the term TE is used to describe all elements whatever their status, autonomous, non-autonomous, transposable, mobilizable, or inactive. TEs can carry genes and/or pseudogenes that encode proteins involved in transposition. In species like wheat, where the genome is massively comprised of TEs, it is essential to identify them to avoid calling genes that are in fact derived from TEs and, thus, are/were involved in transposition rather than a function related to a phenotype and under selection pressure.

The problems described above limit our ability to determine if a sequence is a functional protein-coding gene, a pseudogene, or part of a TE, with high confidence. In addition, the lack of evidence sometimes limits our ability to precisely determine the structure of a gene. Positions of the start codon and borders between coding exons and introns can remain doubtful in many cases. Transcriptomic data like RNASeq are extremely useful to determine exon/intron borders, the existence of alternative transcripts, and the extent of untranslated regions (UTRs of the mRNA upstream the start and downstream the stop codons). Fixing the start codon position,

however, often requires protein sequence homology. Usually in whole-genome annotation projects, for each gene, the most important is to predict the coordinates of the CDS features (i.e., the coding exons). With RNASeq, it became a routine to also annotate the positions of UTRs and all alternatively spliced mRNAs, while defining one representative mRNA/CDS per gene (usually the longest or the most conserved with other species, numbered "1" by convention). For low or non-expressed genes, UTR and mRNA coordinates may not be predicted because of a lack of information. In that case, the gene coordinates are limited to the CDS, which remains the basic essential annotation for a protein-coding gene. For wheat, our main goal was to predict CDS first and, if possible, to add the layer of UTRs and transcripts, these later ones being highly dependent on the RNASeq samples available and methods used.

Wheat gene models have been assigned a confidence category, namely high versus low confidence (HC, LC). This could be misleading since confidence may rely either on the existence of a gene or rather on its exon/intron structure. For instance, one can be highly confident that a sequence encodes a gene while weakly confident on its exact exon coordinates. Both are related. Doubt of the existence of a gene at a given locus is associated with lack of homology evidence. In RefSeq v1, the HC/LC categories classified genes based on their level of similarity (complete or partial) with proteins from other plants. The consequence is that HC genes are likely functional and conserved among *Poaceae* even if some might be predicted with a doubtful structure. LC genes share partial similarity with known proteins and can be well-defined functional genes but the qualitative judgment is of low confidence.

Refinement of automated annotation pipelines to deal with the LC "challenge" is expected to engage manual curation by experts. Manual curation is required to improve the overall quality of the automated annotation. However, manual curation may be mistakenly considered as a validation. Both computer and human algorithms take a decision based on a priori knowledge on the structure of genes and on homology information. When the decision is obvious, typically for genes widely conserved, homology with known proteins and mapped transcripts, if consistent, human curation is not needed. When homology is weak or partial, with a lack of transcription evidence, manual curation does not allow to achieve high confidence neither on the existence of a gene nor on its structure. Curation has a positive impact only in particular cases: missing genes (with evidence slightly under default thresholds), chimeric tandem duplicated genes, start codon mis-assignment, and correction of gene models that are in fact pseudogenes because truncated or with frameshift mutations. These are all particular cases where the situation deviates from standard and is too complex for algorithms.

For TEs, especially in large genomes, manual curation has a much stronger impact than it has for genes. Automated TE modeling is extremely complicated in genomes like wheat where TEs cover 85% of the genome. The history of nested insertions of young elements into old ones has shaped a mosaic of TEs highly fragmented. For instance, manual curation led to identify blocks of nested TEs in which the two extremities of the older element are separated by > 200 kb (Choulet et al. 2010). Such reconstruction is a computational challenge, and manual curation still has a major impact on the quality of the TE annotation. However, with around 4 million TEs in the wheat genome, manual curation was limited to small regions for the moment.

4.2.3 TEs *Versus* Genes: The Crucial Point of Having a Manually Curated TE Library

Providing the complete (protein-coding) gene catalog of a sequenced genome is the priority of annotation. The impact of our knowledge about TEs on our ability to determine if an ORF is part of a functional gene, or if it is a TE-related ORF, is illustrated in rice, where the first releases in 2002 over-predicted around 50,000 genes (Goff et al. 2002; Yu et al. 2002;

Bennetzen et al. 2004) because of unknown TEs. In the wheat context, in the first release (RefSeq v1.1), the predicted CDSs represented 143 Mb [i.e., 107,891 HC genes; (IWGSC 2018)] which is not even 1% of the genome versus 85% for TEs. Considering the possibility that if even only 5% of the TEs are not correctly identified, the amount of "TE-related ORFs" considered as potential functional genes would exceed the total number of predicted genes. Consistent with such a high degree of uncertainty was the initial number of 908,149 candidate loci (after filtering out TE-matching loci) that matched either transcripts and/or homologous proteins in the wheat draft genome annotated in 2014 (IWGSC 2014). RNASeq analysis highlighted 976,962 potentially expressed loci in this study (generating polyA-tailed transcripts), a number considered to be well in excess compared to what was expected based on studies in model grasses. Releasing an annotation that is a good representation of the biological reality is therefore a challenge, and the availability of a curated TE library is of major importance since it could filter out thousands of mis-called genes.

In the development of a representative wheat genome sequence, the long-standing effort to build a high-quality curated TE library has provided a sound foundation. From the beginning of BAC sequencing in wheat, barley, and related *Triticeae*, which all share common TE families, several groups around the world have contributed to manually annotate TEs while defining their exact borders (by searching for terminal repeated motifs). These TEs were organized, classified, and distributed through the Triticeae Repeat (TREP) library maintained by Thomas Wicker at Zurich University, a resource extremely useful for masking TEs, a common task in genome annotation meaning that nucleotides assigned to TEs are converted to Ns (or to lowercases). In 2010, the first large contiguous wheat sequences (obtained from BAC-contigs) were published, representing 18 Mb (Choulet et al. 2010). Although it accounted only 0.1% of the genome, it doubled the amount of wheat sequences available at that time. Even though our knowledge of the wheat genome was still extremely partial, similarity-searches against TREP already identified 75% of the sequence as TEs. This early work demonstrated that manual annotation of a small fraction of the genome allowed the identification of all the abundant TE families, highly repeated, that comprised most of the genome. It also revealed that CACTAs were underrepresented in the library, contrary to LTR-Retrotransposons (LTR-RTs) Gypsy/Copia. The main reason being that the level of variability/diversity of LTR-RTs is low compared to CACTAs. This impacts TE annotation/masking because similarity-search (at low stringency) allows cross-matching between LTR-RT families, meaning that it is not necessary to have identified all families to mask the unknown ones. In contrast, for CACTA families, similarity between families is often limited to the extremities of the element while the internal part is much more variable. This is why a special effort was made, in 2010, to manually curate 3222 elements, especially 330 CACTAs, in order to enrich the wheat TE library (Choulet et al. 2010). This led to the proportion of predicted TEs increasing from 75 to 85% of the genome. In 2014, these ca. 3200 new elements were combined to TREP and classified de novo and a more exhaustive library called ClariTeRep was established (Daron et al. 2014). ClariTeRep is mostly enriched in CACTAs compared to the original TREP library and has a clear impact on TE annotation of *Triticeae* genomes. Several *Triticeae* sequencing projects concluded that CACTAs represent 5–6% of the genome (Jia et al. 2013; Ling et al. 2013), while their proportion is around 15% based on ClariTeRep.

4.2.4 Ab Initio, Homology-Based Predictions, and the RNASeq Revolution for Gene Calling in Complex Genomes

Pipelines for automated structural annotation usually require to combine information from ab initio predictors and evidence of similarity with known proteins in other species or transcriptome sequences (ESTs, full-length cDNAs,

RNASeq [short reads], IsoSeq [long reads] data). For large genomes like in wheat, the problem of ab initio predictors is the very high number of false positives. Indeed, since TEs are estimated to cover at least 85% of the genome, while genes would cover 1–2%, the remaining 13–14% of unannotated DNA account for approximately 2 Gb where gene finders predict gene models because of the presence of ORFs that look likely coding. The reason is that the unannotated part is shaped by low-copy TE-derived sequences, old TE relics, not identified with default TE identification approaches, that carry ORFs that are/were coding (e.g., fragment of transposase) and thus are mistakenly recognized by gene predictors.

Because of the TE-derived ambiguity, biological evidence of homology with related species has always been the criteria of choice to accurately predict genes in wheat. The bad point for wheat was that the number of related species with a sequenced genome was limited, among the *Poaceae*, to *Oryza sativa*, *Zea mays*, *Sorghum bicolor*, and *Brachypodium distachyon*. Outside the *Poaceae* (common ancestor 60 MYA), sequence similarity is too weak to ensure accurate homology-based predictions. This raised a serious problem: wheat genes conserved among the *Poaceae* were well-predicted but our ability to predict less conserved genes was very limited at the early stages of annotation before 2010, especially for species-specific genes.

Transcriptome sequencing considerably enhanced our ability to determine which regions of the genome carry genes because it showed evidence of transcription. Transcriptome sequencing started with a massive effort to sequence millions of ESTs and full-length cDNAs (Ogihara et al. 2004; Zhang et al. 2004) and was followed by the emergence of RNASeq technical capacity which provided unprecedented power to drive structural annotation. First use of an RNASeq expression atlas for wheat gene annotation at the chromosome scale was published in 2014 (Choulet et al. 2014; Pingault et al. 2015). In brief, 7264 gene models were predicted but only 5185 (71%) showed transcription evidence in an RNASeq atlas covering five plant organs at three developmental stages each. In addition, 3692 transcribed regions were detected in the unannotated sequences showing that 42% of the loci likely expressed did not correspond to predicted protein-coding genes. This indicated a high level of uncertainty in describing biological reality when annotating the wheat genome. In this chapter, we propose a critical view of automated gene annotation pipelines, namely that bioinformatics can predict but not demonstrate that a sequence is a gene and that a gene is not a pseudogene. Although RNASeq became a primary resource for structural annotation, the correspondence between RNASeq-read mapping loci and the final filtered gene set was far from perfect, with 29% of chr3B gene models showing no transcription evidence and 42% of transcribed regions not looking like protein-coding genes. Homology with related species remains an important benchmark.

4.2.5 Single-Gene Duplications Raise More Problems Than Polyploidy for Structural Annotation

Given the weight of similarity-search with transcripts and proteins in structural annotation, intrinsic features of the genome significantly impact the difficulty to identify the correct gene structure since sequence alignments underpin all the studies. A first important intrinsic feature to impact annotation is the fragmentation level, i.e., the number of exons per gene. As a CDS is fragmented into several exons, the difficulty to predict the correct intron/exon structure increases. In wheat, considering RefSeq Annotation v2.1, the average number of exons per CDS is only 4. Sixty percent of the CDSs are split into a maximum of 3 exons. Actually, only 10% of the gene set corresponds to CDSs split into ten exons or more. Thus, the fragmentation problem is limited in wheat.

Other important criteria are the lengths of exons and introns. Small exons might be missed by sequence alignments because under the

default thresholds of automated pipelines. Large introns also raise problems for spliced-alignments. In the current wheat annotation release, the average exon length is 498 bps and the average intron length is 280 bps (considering only one representative transcript per gene). Thus, exons are, on average, large enough for high-scoring alignments, and introns are small enough for the efficiency of spliced-alignments. So, although it is commented that the wheat genome is complex, some intrinsic features are rather less complex than in many other eukaryotes.

Does polyploidy impact our ability to call genes? The main problem with alignment-based methods for gene calling is obviously multiple mapping, i.e., the fact that a transcript/protein matches at multiple loci along the genome. But it does not mean at all that single-copy genes are easier to predict than duplicated genes. In contrast, the fact that a gene is repeated on, e.g., chromosomes 1A, 1B, and 1D, because of polyploidy is rather in favor of accurate structural annotation. Since each copy is carried by a different chromosome, it is annotated independently and this does not generate problems due to multiple mapping. The three subgenomes A-B-D could be annotated as if they were three genomes of three different species. If a gene copy is silenced and thus does not generate an RNASeq signal, reads coming from the copies that are transcribed can be used to predict the structure of all copies. So, again, to our opinion polyploidy is an advantage here for structural annotation. To go further, we can even consider that we did not fully exploit the advantage provided by this intrinsic redundancy of the genome for structural annotation of the IWGSC RefSeq. We will present this in more detail in the paragraph below describing future plans for improvements.

Large chromosomes such as found in wheat are usually fragmented into "chunks" that are annotated independently in parallel. The problems with multiple mapping arise when repeated copies of a gene are carried by the same chunk. This is typically the case for tandemly duplicated genes. This is why automated structural annotation of tandem duplicates is the most complicated task. Single-gene duplications are much more problematic than whole-genome duplication (i.e., polyploidy). This is true for every genome to be annotated mainly via the homology-based approach. However, for wheat, this problem has strong implications because we demonstrated that single-gene duplications intensively affected the gene repertoire during its recent evolution (Glover et al. 2015). In the IWGSC RefSeq v1.1, we found that 27% of genes were present as tandem duplicates (IWGSC 2018). Multiple mapping of homologous proteins and transcripts on tandem duplicates may lead to artificially link exons from the two copies and, thus, to predict chimeric genes. This is especially the case for highly identical copies that are separated by a small intergenic region, compatible with a classical intron length. Some highly repeated gene families such as the kinase genes and disease resistance genes are well known to fall into this category. Unfortunately, these genes are often the favorite candidates to control phenotypes of interest, and in that case, manual curation is a required step to improve significantly the accuracy of automated annotation.

4.3 RefSeq V1.0 Structural Annotation

4.3.1 The Impact of Annotation Procedure on Gene Predictions Is Very Strong

Sequencing the wheat genome has a long story. Different initiatives have been launched following the advances of sequencing technologies to tackle the hexaploid genome and also the genome of the diploid and tetraploid relative species. For CHINESE SPRING itself, before completing RefSeq v1, a draft genome assembly (named CSSs for chromosome survey sequences) was released in 2014 (IWGSC 2014) together with a chromosome-scale assembly of the entire chromosome 3B using a BAC-by-BAC approach, hereafter named "3B-BAC-2014"

(Choulet et al. 2014). In addition, another version of the CHINESE SPRING genome was produced and annotated in 2017 named TGACv1 (Clavijo et al. 2017). Hence, when the annotation of RefSeq v1 started, chromosome 3B has already been annotated three times independently: 3B-BAC-2014 with the TriAnnot pipeline at GDEC Institute (Clermont-Ferrand, France), CSS-3B-v2.2 at PGSB Institute (Munich, Germany), and TGACv1 at Earlham Institute (EI, Norwich, UK) with homemade pipelines. Here, we compared these three gene catalogs to have a flavor of the impact of the methods on the results released: among the 7264 CDSs predicted on 3B-BAC-2014, only 26% (1884) and 12% (867) were strictly identical in TGACv1 and CSSv2.2 (sharing strictly identical protein sequences). These percentages appear extremely low if one considers these are three independent initiatives to sequence/annotate the same genotype. It demonstrates the impact of the annotation procedure on the released gene catalog as well as the possible impact of the sequencing strategy and assembly quality.

4.3.2 Gene Annotation Through a Federated Approach

Given the strong differences observed when comparing results obtained by different groups, the IWGSC established an Annotation Working Group in order to coordinate the efforts and establish an integrated approach to annotate RefSeq v1. Genes were predicted independently by two groups using two different pipelines and two different strategies: GDEC and PGSB. Both were then integrated at EI to end up with a single annotation. This led to v1.0 which was quickly updated into v1.1 after integrating ~4000 manually curated genes (see below for details on curation).

In v1.1, 107,891 high-confidence (HC) protein-coding loci were identified, with a relatively equal distribution across the A, B, and D subgenomes (35,345, 35,643, and 34,212, respectively). In addition, 161,537 other protein-coding genes were classified as low-confidence (LC) genes, representing partially supported gene models, gene fragments, and orphans. On ChrUn (unplaced scaffolds), 2691 HC and 675 LC gene models were identified. Evidence for transcription was found for 85% (94,114) of the HC genes versus 49% of the LC genes. In addition, 303,818 pseudogenes were also annotated. The quality of RefSeq Annotation v1.1 was estimated with BUSCO v3 (24). It revealed that 99% (1436/1440) of the BUSCO v3 genes were present in at least one complete copy and 90% (1292/1440) in three complete copies.

4.3.2.1 Gene Modeling Using TriAnnot

The TriAnnot pipeline was developed and updated over a period of more than 10 years to enable automated robust structural and functional annotation of protein-coding genes, transposable elements, and conserved non-coding RNA genes in *Triticeae* genomes (Leroy et al. 2012). It was dedicated to large-scale annotation projects and is executable through the command line on high-performance computing infrastructures for parallelization with task dependencies. TriAnnot was initially used for the annotation of BACs (Choulet et al. 2010) and then for the entire chromosome 3B (Choulet et al. 2014). Thus, it was intensively trained and customized specifically for wheat before we assembled RefSeq v1.

The specificities of the annotation strategy implemented in TriAnnot included: (i) mask TEs first in order to restrict the gene modeling to the non-TE space; (ii) use both evidence-based and ab initio approaches before selecting the best gene model at each locus. It was launched individually on each scaffold (or chunks for large ones) of RefSeq v1.0 in parallel while positions of features were subsequently calculated on pseudomolecules. The different steps and tools launched by the pipeline are described below:

- Step 1: TE annotation and sequence masking. TEs were identified by similarity-search using CLARITE and ClariTeRep (Daron et al. 2014). CLARITE used RepeatMasker with cross_match as search engine for optimized accuracy (Smit et al. 1996–2004).

Nucleotides assigned to TEs were then masked so that the following steps, i.e., ab initio predictions and similarity-searches, were all performed on the masked genome sequence.

- Step 2: Gene modeling. Ab initio gene models were predicted using two gene finders previously trained with a wheat gene dataset: FGeneSH (http://linux1.softberry.com/berry.phtml) and AUGUSTUS (Stanke et al. 2006). Evidence-driven gene predictions were also computed following three different strategies giving different weights to protein and transcript similarities. The first approach was based on homology with proteomes of related species. Similarity-search was performed using BLAST (Zhang et al. 2000) and significant hits, filtered with fine-tuned thresholds, were then used for spliced-alignment using EXONERATE (Slater and Birney 2005). The query proteins were those predicted in main *Poaceae* species for which a genome sequence was available: *O. sativa* (International Rice Genome Sequencing Project 2005), *B. distachyon* (The International Brachypodium Initiative 2010), *S. bicolor* (Paterson et al. 2009), *Z. mays* (Schnable et al. 2009), and *Hordeum vulgare* (International Barley Genome Sequencing Consortium et al. 2012). This approach is well suited to precisely determine the obvious structure of a large fraction of the protein-coding genes by taking advantage of their evolutionary conserved nature. However, the main limit here was the lack of similarity at the protein extremities which may lead to incomplete alignment that prevents from finding the start and/or stop codons. Thus, TriAnnot utilized an iterative extension in order to identify in-frame start and stop codons for gene modeling. Models with partial structure were flagged pseudogenes.

  The second evidence-driven approach (SIMsearch module) was based on transcripts first, rather than proteins. SIMsearch module is a gene modeling program based on FPGP (Amano et al. 2010) and adapted specifically for wheat to address problems generated by tandem repeated genes. SIMsearch identified the loci that are transcribed by spliced-alignment using est2genome (Mott 1997) of a series of wheat transcript libraries. The CDS coordinates were predicted afterward through similarity with *Poaceae* proteomes. SIMsearch was launched twice using two databanks of wheat transcripts: (1) predicted transcripts derived from a large RNASeq experiment that targeted five plant organs at three development stages each in two replicates (Pingault et al. 2015); (2) all available wheat full-length cDNAs available at EBI-ENA and from Ogihara et al. (2004). Thus, TriAnnot did not use RNASeq reads directly as an input. Read mapping and transcript calling were computed prior to gene annotation, and the predicted transcripts were provided as FASTA input for spliced-alignment during the process of gene modeling.
- Step 3: Selection of the best gene model at every locus. In summary, TriAnnot predicts gene models through five approaches: two ab initio and three evidence-based (one derived from spliced-alignment of homologous proteins + two derived from transcript evidence). One gene may obviously be predicted through different ways. Thus, the final step is the selection of the best gene model at each locus. Indeed, at that step, there was no combination of different overlapping models to create a new one.

A scoring process was applied in order to validate the existence of a gene and to retain its most probable structure. For scoring, TriAnnot used BLASTP to search for similarity of each model with proteomes of related *Poaceae*, including *Aegilops tauschii* and *Triticum urartu*, and calculate a score while considering metrics of the best hit alignment (percentage of identity and coverage, presence of canonical splicing sites, presence of start and stop codons).

Gene models not supported by homology with *Poaceae* proteins or by transcription evidence were simply discarded (i.e., ab initio only). Models sharing similarity with known

proteins and for which splicing sites were supported by transcript evidence were classified as high confidence. Low-confidence genes also share similarity with known proteins and transcripts but lack support for some splicing sites and/or position of start/stop codons. Finally, genes sharing similarity with known proteins but over less than 70% of the length of its best BLAST hit were classified as pseudogenes. Thus, TriAnnot predicted 107,226 gene models: 65,884 HC and 41,342 LC genes, plus an additional 73,044 pseudogenes on the IWGSC RefSeq v1.

4.3.2.2 PGSB Gene Prediction Pipeline

The procedure implemented in the PGSB annotation pipeline differs in many aspects from that of TriAnnot. It is based on mapping all available evidence on unmasked genome sequence and filtering out TE-related predictions afterward. It was all evidence-driven, not using any ab initio gene finder.

- Step 1: Mapping. The PGSB annotation pipeline combined spliced-alignments of reference proteins, IsoSeq reads and full-length cDNAs (flcDNAs), and RNASeq transcript predictions. In addition to the RNASeq atlas from Pingault et al. (2015) also used in TriAnnot, additional samples were added here. There were Illumina reads produced on grain-specific samples (Pfeifer et al. 2014), whole transcriptome PacBio sequenced samples (PRJEB15048), and disease resistance gene enriched transcriptome samples (PRJEB23081). The latter were all from CHINESE SPRING but there were also transcriptomic data generated from other accessions cultivated under drought and heat stresses (SRP045409) and under infection by *Fusarium graminearum* (E-MTAB-1729). Mapping outputs were all combined, and mapped reads were assembled into transcripts with StringTie (Pertea et al. 2015).
  Protein sequences from the five species *Arabidopsis thaliana*, *B. distachyon*, *O. sativa*, *S. bicolor,* and *Setaria italica*, and complete proteins from *Triticeae* in UniProt (UniProt Consortium 2018) were aligned with GenomeThreader independently on each chromosome. flcDNAs from wheat and barley (Mochida et al. 2009), together with wheat IsoSeq reads (Clavijo et al. 2017) were mapped with Gmap (Wu and Watanabe 2005) and included in the prediction pipeline.
- Step 2: Prediction and selection of open-reading frames. Predictions originating from protein alignments, full-length transcript alignments, and RNASeq were combined while removing redundancy (using Cuffcompare and StringTie). Then, TransDecoder (https://github.com/TransDecoder/TransDecoder/) was used to predict the coding frame for each transcript while considering the most upstream start codon by default. These predictions were then aligned against a set of reference proteins from angiosperms in UniProt, and protein domains were also searched for. These data were given to TransDecoder for selecting the most probable CDS for each model.

Since TEs were not masked prior to mapping evidence, PGSB predictions were filtered out afterward based on similarity-search with TE-related proteins from the PTREP library (https://botserv2.uzh.ch/kelldata/trep-db).

4.3.2.3 Integration of TriAnnot and PGSB Gene Models with Mikado

Selection of the best representative model at each locus was applied through a rule-based approach that combined supporting evidence and intrinsic gene features. PacBio transcripts, RNASeq reads, and homologous protein alignments over the genome were used to measure the accuracy of predictions and a set of high-confidence splicing sites was established from RNASeq mapped reads. Mikado (Venturini et al. 2018) was used to cluster genes from the two pipelines into loci, to calculate an overall score to each gene model, and to select the highest-scoring gene model. The score reflected the congruence between a model and its supporting

evidence, calculated with an average F1-score (reflecting precision and recall) and metrics of gene feature, e.g., a penalty was applied to introns larger than 10 kb. After selecting the representative model, Mikado was used to identify additional high-quality alternatively spliced transcripts, only those that met a series of stringent requirements. The most important were: a CDS overlapping at least 60% of the representative CDS, without any retained intron, and with only verified exon/intron junctions. Eventually, to enrich the annotation, coordinates of UTRs were added based on comparing models and aligned transcripts with PASA (Haas et al. 2008).

4.3.2.4 Gene Confidence Assignment: HC Versus LC

Despite the sophisticated combination of both TriAnnot and PGSB predictions, the final number of models was very high: 269,428, representing approximately 90,000 protein-coding genes per (haploid) subgenome. As previously observed in wheat, regions showing traces of expression or homology with known proteins are much more abundant than expected, given that the number of protein-coding genes is a quite stable parameter in plant genomes with ~ 30,000 genes per haploid genome. It suggested that many gene models were in fact pseudogenes or doubtful non-coding transcribed regions for instance. However, both included filtering steps to discard models matching wheat transposons, before gene modeling for TriAnnot, after for PGSB. Thus, a confidence category was assigned to each gene model: high confidence versus low confidence. The idea was to provide a single filtered dataset of HC genes to people only interested in large-scale whole-genome analyses while keeping information of LC genes to people interested in the characterization of a particular region.

First classification parameter was the completeness of the model, i.e., the presence of both a start and a stop codon. HC genes were complete with significant homology with plant (*Magnoliophyta*) proteins retrieved from Swiss-Prot and TrEMBL. LC genes were, either complete but without significant homology with plant proteins or, incomplete with or without significant homology. The 269,428 gene models were split into 107,891 HC (40%) and 161,537 LC (60%) protein-coding genes. The number of HC genes was much closer to the expected value for plants (~ 35,000 genes per haploid genome), and this became the reference dataset used by the community.

However, within all the limits explained here, we encourage users to always keep in mind the level of uncertainty behind the annotation space. To the question "how many protein-coding genes are there in wheat?" we should answer: We do not know because the proportion of doubtful predictions is just too high.

4.3.2.5 What Should Be Known About the LC Genes and Pseudogenes

The consequence of confidence assignment is that the LC category gathered genes that were non-conserved, i.e., might be species-specific, for which we did not have enough evidence to conclude it is functional, together with (highly) conserved genes that are either pseudogenes or just partially assembled or mis-predicted. One must consider that a part of the LC genes is conserved but exhibits a structure likely incomplete. This has strong implications for researchers interested in a particular gene family or a particular locus.

In addition, a specific search for pseudogenes was launched at the whole-genome level, based on finding DNA fragments sharing similarity with HC genes but only partially or with frameshifts and/or internal stop codons. In total, 288,939 pseudogenes were discovered with 10,440 corresponded to LC genes. Thus, the coding landscape is even more complicated than often believed, with 108 k HC, 162 k LC, and 279 k gene fragments and so if a gene is considered to be absent based on HC genes only, it is important to consider the pool of LC genes.

4.3.3 Comparing Genes Between A, B, and D Subgenomes

4.3.3.1 Finding Homeologous Groups Based on HC Genes Only Can Lead to False Conclusions and Highlights the Requirement of Considering LC Genes

Considering the conclusion of the latter paragraph, it implies that comparing the A-B-D gene repertoires was strongly impacted by the input gene dataset. Homeologous groups were inferred from gene trees. Initially, trees were built with the complete set of HC and LC genes which revealed that considering HC genes only led to considerably overestimate the level of variability between A-B-D subgenomes, because many LC genes were, in fact, orthologous to HC genes (i.e., homeologous in the hexaploid) even though functional annotation revealed that some LC genes represented mis-predicted TE-genes (e.g., transposase-like genes). The solution adopted was to work on a filtered gene dataset: 181,036 genes (103,757 HC and 77,279 LC genes; instead of 269 k initially) that do not correspond to either TE-related functions or to pseudogenes. This led to determine a total of 39,238 homeologous groups (i.e., clades of A-B-D orthologous deduced from gene trees) and 33% of them include LC genes. In total, 28,829 LC genes have homeologous partners and were thus valid for biological analyses.

The main conclusion of the A-B-D comparison was that the gene repertoire of the three subgenomes is much more different than previously thought. The default hypothesis is often that a gene is present in three pairs of homeologous copies in bread wheat because it is a hexaploid. The reality is that only 55% of the homeologous groups are triads, i.e., single-gene copy per subgenome (configuration 1:1:1). Thus, 45% of the groups represent cases where gene loss and/or duplications occurred after A-B-D divergence. Gene loss after A-B-D divergence represents the same proportion for A, B, and D: ~ 10% of the homeologous groups. Regarding gene duplications, they also occurred in the same proportions in A, B, and D. This analysis suggested that the three lineages leading to A-B-D genomes have independently accumulated differences (gene loss and gene duplications) at similar rates.

4.3.3.2 No Evidence of Any Biased Gene Fractionation and Importance of Gene Duplications

Regarding gene presence/absence, no evidence for biased partitioning was observed (IWGSC 2018). In contrast, comparisons support gradual loss/duplications that have occurred after A-B-D divergence in the diploid, tetraploid ancestors, and after hexaploidization event in modern bread wheat. Before gene loss, a gene may lose function because of silencing or change in expression, so that the first evidence of diploidization might be observed at the expression level. Hence, RNASeq data analyses showed that there was an equal contribution of the three homeologous genomes to the overall gene expression, demonstrating the absence of global subgenome dominance (IWGSC 2014).

4.3.4 TE Modeling

Given the amount of TEs shaping the wheat genome, predicting the presence of TE copies along assembled sequences has always been a prerequisite to avoid false predictions of coding genes that are in fact coding parts of TEs. Efforts to manually annotate TEs with their precise borders were made since the beginning of wheat BAC sequencing and a high-quality reference databank of wheat TE sequences was initiated in 2002 with TREP (Wicker et al. 2002) and completed in 2014 with the ClariTeRep library (Daron et al. 2014) (which includes TREP). ClariTeRep originated from manual curation of ~ 3200 TEs along the first large (Mb-sized) contiguous sequences produced on chromosome 3B (Choulet et al. 2010). This implies that the wheat TE library used for similarity-search might be biased toward elements from the B-subgenome, and depleted for A and

D subgenomes. However, it was shown that TE families that shaped the three subgenomes are the same, although subfamilies (variants) have differentially invaded the A-B-D genomes in the diploid ancestors (Wicker et al. 2018).

Thus, TE modeling in RefSeq v1.0 was performed only via a similarity-search approach against ClariTeRep. There was no de novo repeat-based discovery of new TEs. This led to the prediction of 3,968,974 copies, classified among 505 TE families, and representing 86%, 85%, and 83% of the A, B, and D genomes, respectively. Such proportions imply that TEs shape large clusters with recently inserted TEs into older ones, a mosaic of nested insertions which is a computational challenge to reconstruct. This step was dealt with CLARITE (Daron et al. 2014) for RefSeq v1.0. CLARITE uses RepeatMasker (Smit et al. 1996–2004) with the cross-match engine for the first step of similarity-search between the genome of the TE library. The main problems with using RepeatMasker in TE-rich genomes are as follows (i) the over-fragmentation: one copy is often not predicted into a single feature but rather split into adjacent fragments; (ii) the overlap of predictions, i.e., a locus could match with several reference; and (iii) scattered pieces of a TE that has been fragmented by subsequent TE insertions (nested pattern) are not joint. The CLARITE pipeline has been developed specifically for wheat, based on ClariTeRep, in order to overcome these three limitations. It uses classification information: all TEs in ClariTeRep were classified into families and subfamilies by sequence clustering. It also uses positions of LTRs in LTR-retrotransposons, which correspond to long terminal repeats (ca. hundreds of bps) that are largely involved in the fragmentation observed after RepeatMasker because both 5′ and 3′ LTRs cross-match since they are almost identical subsequences. Family classification and LTR positions are the two main points implemented in CLARITE. They allowed accurate defragmentation, while preventing chimeric merging of adjacent features, and accurate reconstruction of nested TEs.

4.4 RefSeq V1.0 Functional Annotation

Gene ontology terms, PFAM, and InterPro domains were assigned to gene models. A function was assigned to 82% (90,919) of HC genes in RefSeq Annotation v1.0. RNASeq-based transcription evidence was found for 85% and 49% of HC and LC genes, respectively. In addition, naming of gene function for each gene was performed by using the AHRD tool (Automated Assignment of Human Readable Descriptions, https://github.com/groupschoof/AHRD, version 3.3.3). This program generates informative functional annotations from BLAST outputs while avoiding retrieving too many "unknown" or "uncharacterized" functions. BLAST outputs against the following databases were parsed by AHRD: Swiss-Prot, *Arabidopsis* Araprot 11, and a subset of TrEMBL for *Viridiplantae*. A filter was then applied in order to discard genes with functions related to TEs. Genes were thus tagged as G (canonical gene), TE (obvious transposon), TE? (potential transposon), or U for unknown. Based on this, 3294 HC genes with a TE tag were moved subsequently to the LC category in RefSeq annotation v1.1.

4.5 RefSeq Annotation V1.1: Integration of Manually Curated Genes

Once Annotation v1.0 was released to the community, researchers who are experts of some specific gene families brought corrections to the automated predictions: Sometimes gene copies were missing, sometimes the predicted exon/intron structure needed to be curated. Feedback was made from the experts to the IWGSC Annotation Group in order to release an updated version 1.1. This concerns gene families CBFs, NLRs, PPRs, Prolamins, WAKs, and amino-acid transporters. A semiautomated process was developed in order to integrate manually curated gene models. It relies on a Python script using common tools like GenomeTools (Gremme et al.

2013), GFFCompare (Pertea and Pertea 2020), pyBEDTools (Dale et al. 2011). GffCompare was used to check that the curated genes did not overlap each other (different teams may have curated the same gene) and also to identify the RefSeq Annotation v1.0 models that required to be updated. Five types of correction were considered: (i) addition of a new gene model that was absent from v1.0; (ii) merging of two gene models; (iii) splitting of a gene model into two genes; (iv) correction of exon positions of a gene model; (v) complex cases which combined splitting and merging. RefSeq Annotation v1.1 includes updates of 3685 manually curated genes, of which 528 were not predicted by the automated annotation process and 354 corresponded to LC gene models. The final v1.1 HC gene set contained 107,891 genes.

4.6 RefSeq Annotation V2: The Challenge of Transferring Gene Annotation Through the Different Versions of Genome Assembly

In 2021, an update of the CHINESE SPRING IWGSC RefSeq Assembly was published (Zhu et al. 2021). Corrections were brought to the initial release by using new resources: Bionano and PacBio contigs. Inconsistencies between pseudomolecules and Bionano maps were reconciled, and 279 unplaced scaffolds were positioned into pseudomolecules. PacBio contigs publicly available (Zimin et al. 2017) were used to fill gaps. Contrary to scaffold reordering, the gap-filling step led to complete changes in the positions of gene models predicted along pseudomolecules, so that it was not possible to calculate new gene position from v1 to v2 with a simple conversion of coordinates. This raised two possibilities: compute de novo gene prediction or transferring the knowledge of the previous annotation release. Since annotation v1.1 was the outcome of an extensive effort to combine different annotation pipelines, the choice was made to try to transfer as many models as possible while trying to optimize the traceability and to minimize the differences between Annotations v1 and v2.

However, finding the new position of a gene required sequence alignment, which raised many problems in hexaploid wheat. For example, we used GMAP to map 298,775 HC and LC genes onto Assembly v2 and observed that 32,152 (11%) could not be transferred accurately because of spurious alignments. Such high error rate was not acceptable and it was decided to develop a transfer-strategy dedicated to this task for wheat. It was implemented in the MAGATT pipeline (https://forgemia.inra.fr/umr-gdec/magatt). The strategy relies on reducing the alignment space to the shortest region predicted to carry the gene to be mapped. In wheat, genes are always flanked by TEs. Although TEs are repeats, each copy is inserted into a different site. Thus, the junction between a TE extremity and its insertion site is unique at the genome level. We derived all such tags from the TE annotation. They represent one tag every 3 kb (compared to one gene every 130 kb on average) that can be uniquely mapped from one assembly version to the other. We used these TE tags as anchors to define the smallest target interval before mapping a gene. The average size of an interval was 9.6 kb, which reduced the alignment space and avoided most problems due to multiple mapping of repeated genes. Even for clusters of tandemly repeated genes in which copies could share 100% identity, this strategy enabled the assignment of the correct interval for each copy and lead to the transfer of annotation of all copies without any cross-matching. MAGGAT succeeded to transfer 90% of HC/LC genes without any difference between v1 and v2 assemblies either in the introns or the exons, and 8% with mismatches due to nucleotide differences incorporated at the gap-filling step (in gap-flanking sequences). Indels were observed for 1% of the genes, and the remaining 1% corresponded to genes for which the sequence was discarded when assembling v2 (Zhu et al.

2021). This step gave rise to the IWGSC RefSeq Annotation v2.1.

Defining the target interval prior to mapping has a major consequence: It avoided the computation of a spliced-alignment of a query transcript/CDS. Indeed, by default MAGATT starts by mapping the entire gene feature (exons + introns + UTRs) with BLAT (Kent 2002) against the short, kb-sized, target sequence. In the majority of the cases (90%), it identified a full perfect match which enabled the repositioning of all sub-features (i.e., exons and UTRs of all alternative spliced mRNAs) from a previous to a new assembly that shared strict identity. This was of major importance because spliced-alignments could have led to errors, especially when exons are very small. When only mismatches (no Indels) were observed between the two assemblies for a given gene (3% of genes), automated repositioning was also possible. Spliced-alignments of mRNAs were computed only when BLAT returned Indels and/or partial match between a query gene and its target.

MAGATT was developed with the objective of transferring a gene annotation to a new assembly release for a given genotype. However, the strategy applies very well to the problem of annotating genes in the genome assemblies of other genotypes and is, thus, significant in the context of post-reference genome sequencing and pangenomics. Pangenomics aims at identifying conserved *versus* non-conserved genes in a series of assembled genomes. The main limit in this area is the quality of the gene predictions. It is therefore possible that presence–absence of a gene may simply be the consequence of annotation artifacts. Thus, MAGATT needs to be considered for delivering an annotation of gene models in new assemblies that mimics as much as possible the reference gene calls and avoid "polluting" the apparent dispensable gene set with differences in gene predictions.

4.7 Plans for Future Improvements

4.7.1 Improving Gene Structural Annotation

The repertoire of 107,891 genes delivered in 2018 for CHINESE SPRING is definitely a reference widely used by the community. However, the methodological limits mentioned above make us consider there are improvement levers. First of all, we must remind here that what we call genes here, by default, correspond to protein-coding genes. Non-coding RNA genes remains largely unexplored in this complex genome although we have no doubt their prediction along the genome sequence represents one of the most challenging tasks but also one of the most impacting novel information to increase our understanding of the functional sequences.

Regarding protein-coding genes, when we discuss the improvement of structural annotation, we distinguish two different things: (i) existence of the gene and (ii) structure of the gene. In other words, improvements concern, on one side, genes that are missing in the annotation and gene models that do actually not correspond to real genes. On the other side, improvements concern the exact structure of a gene and its transcripts.

A key question that impacts on both aspects is the presence of pseudogenes. Pseudogenes are sequences derived from functional genes but that have accumulated mutations (frameshift, in-frame stop codon, truncation) which switched its function off. Pseudogenes are hard to model automatically because gene modeling usually uses structural features (coding frame, start and stop codons) to call a gene while in case of pseudogenes, these features are disturbed. Manual curation of genes remains the best way to classify a sequence as a pseudogene. Although community annotation (jamboree) event was not organized in the framework of the

IWGSC, the IWGSC did establish a procedure in order to integrate curation made by different expert groups at the international level. This led to several updates: annotation releases v1.1, v1.2, and v2.1. Manual curation by experts represents 2–3% of the gene content in v2.1.

The current status with respect to wheat gene models is: 108 k HC genes, 162 k LC genes plus an additional 279 k gene fragments found by scanning for fragments of coding DNA in the unannotated part of the genome. It is clear that, with such a complicated landscape, manual curation is an endless task. However, lots could be done through bioinformatic approaches combined with manual curation in order to increase annotation quality. But even curators need information for taking decision on the most probable gene structure to consider and an open question is "which information/resources are lacking and which strategies could be useful for helping with increasing the quality of gene model predictions?".

4.7.1.1 Transcription Evidence, Gene Finders, and Homology with Related Species: Comparing A-B-D is the Most Highly Valuable Option to Improve the Quality of Structural Annotation

Finding a gene is based on three pieces of evidence: (i) a sequence is transcribed (RNASeq); (ii) a sequence shares similarity with proteins already predicted in divergent genomes; (iii) a sequence has a high probability to be protein-coding (based on hidden Markov models).

Do we miss transcript data? As early as in 2014, up to one million loci matching RNASeq data (short reads) were highlighted but even then, there were still 15% of the HC genes for which no transcription evidence was found (IWGSC 2018).

What about gene finders? The wheat genome is made of ca. 12 Gb of transposon-derived sequences while gene models represent 0.13–0.23 Gb (depending on whether or not LC genes are considered). The wheat genome is full of coding-like DNA but the very wide majority is related to TEs (transposase, reverse transcriptase, integrase, etc.). The consequence is that the unannotated part of the genome, representing ca. 10–15% (1.5–2.0 Gb), i.e., 10 times more than the gene space, often corresponds to unidentified degenerated TEs. This means that ORFs derived from degenerated TEs are an extremely abundant source of false positive predictions for gene finders.

Detecting sequence homology with related genomes appears to us an underestimated lever of improvement. This evidence relies the evolutionary definition of a gene: an entity submitted to selection pressure. If a sequence is conserved across millions of years of evolution, we can be confident it is a gene. Predicted proteomes of *Poaceae* have been used in wheat gene modeling. However, improvements seem here obvious since there were not that many genomes available. Among the *Poaceae*, knowledge from the sequenced and annotated genomes of *O. sativa*, *Z. mays*, *S. bicolor*, *B. distachyon*, and *S. italica* were used for wheat gene modeling. They share a common ancestor with wheat between 30 and 60 MYA. Outside the *Poaceae*, fewer genes are conserved and sequence identity, even at the protein level, is low (around 55% with *Arabidopsis* for instance) which would not be of great interest to improve the annotation. Indeed, widely conserved genes are the easiest to annotate. In contrast, the challenge of annotation relies on finding genes that are specific to the *Triticeae* tribe, the *Triticum/Aegilops* genera, or even to the *T. aestivum* species. So, the most helpful resource to ensure efficient gene modeling in wheat is the *Triticeae* species, where genomes diverged 3–13 MYA, and which share high level of synteny and high level of gene sequence conservation. For instance, 88% of the predicted wheat genes (IWGSC v2.1) share on average 84% protein identity with barley predicted proteins (based on first BLAST hit alignment with thresholds 50% query overlap, 35% identity) (Mascher et al. 2017). But even TEs share sequence similarity between *Triticeae* genomes, meaning that conservation is not synonymous of selection pressure when aligning barley and wheat genomes. However, we could take advantage of the near-complete TE turnover (Wicker

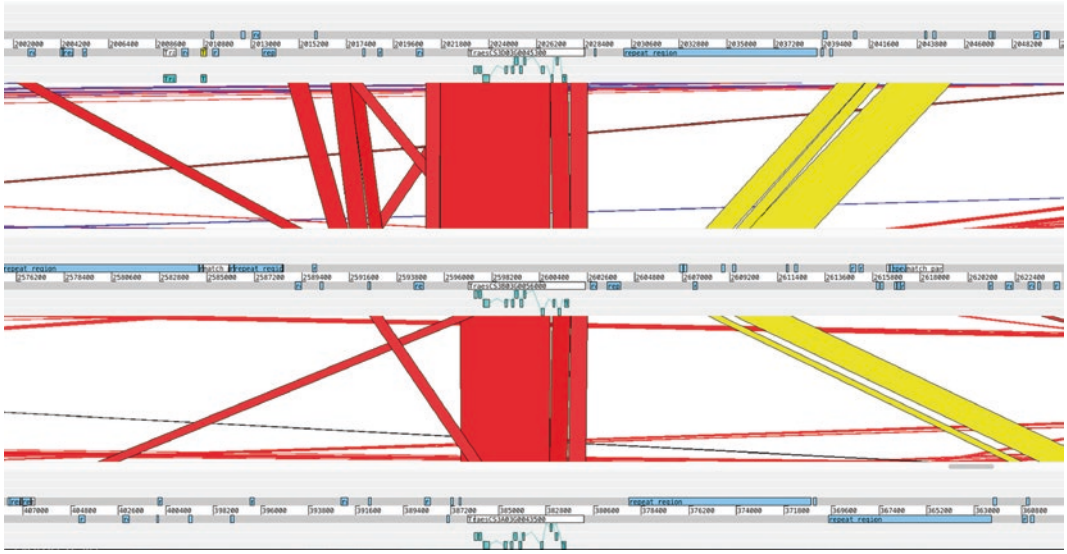


Fig. 4.1 Sequence alignments visualized with ACT (Carver et al. 2005) of three wheat homeologous regions of chromosomes 3A, 3B, and 3D. CDSs are represented in light blue, genes in white, and TEs in blue, across the six coding frames. Red blocks represent sequence conservation (> 85% identity) between A-B-D regions carrying homeologous genes and surrounding regions while TEs are not conserved between homeologous loci. Yellow blocks indicate the presence of a highly conserved unannotated sequence (neither gene nor TE) between A-B-D which strongly suggests the presence of a functional sequence subject to selection pressure that may correspond to a yet uncharacterized gene

et al. 2018) that led to erase ancestral TEs so that there are (almost) no syntenic/orthologous TEs between A, B, D (*Triticum* and *Aegilops*), H (barley; *Hordeum*), and R (rye; *Secale*) genomes. All these genomes diverged between 3 and 13 MYA, a timeframe consistent with (1) a complete TE turnover (2) within a conserved gene backbone. This is the ideal situation to identify new genes based on aligning syntenic regions. Each segment of conserved sequence between A-B-D-H-R genomes (and others) at a micro-syntenic location is evidence for selection pressure and, thus, for the presence of a gene (protein-coding or not) or a sequence involved in regulation processes called conserved non-coding sequence (CNS) as shown in Fig. 4.1.

4.7.1.2 To What Extent Sequencing More Wheat Genomes Help Improving the Reference Wheat Gene Catalog?

As explained above, the divergence window 3–13 MYA of *Triticum*, *Aegilops*, *Hordeum*, *Secale*, and others combines the advantages of a high level of gene conservation with the (almost) absence of orthologous TEs. Sequencing more *T. aestivum* genomes will not be useful in that regard. Indeed, divergence is too low so that sequence conversation is not evidence for selection pressure. Most TEs are conserved (orthologous) even between divergent accessions from the Asian and European pools, as highlighted by the Renan *versus* Chinese Spring comparison (Aury et al. 2022). However, sequencing more wheat genomes will be exploited for building the wheat pangenome.

4.7.2 De Novo Annotation *Versus* Annotation Transfer

With the advances made in sequencing technologies, assembling reference-quality wheat genome sequences is not a limit anymore (Guo et al. 2020; Walkowiak et al. 2020; Sato et al. 2021; Athiyannan et al. 2022; Aury et al. 2022). Building a wheat pangenome is thus a crucial objective in order to distinguish core *versus*

dispensable genes, especially since dispensable genes are the best candidates for adaptation to the environment, like response to specific pathogens. In contrast, core genes are enriched in essential genes, somehow not the privileged targets to search for genetic diversity controlling contrasted phenotypes.

Presence/absence (and copy number) variations of genes between two genotypes are limited to a few percent (De Oliveira et al. 2020). Using resequencing data of chromosome 3B from 20 *T. aestivum* accessions, it was shown that variable genes represent between 2 and 6% of pairwise comparisons with CHINESE SPRING. This weak percentage implies that approximations due to incomplete genome assembly and differences in gene predictions will strongly impact our capabilities to determine if a gene a really absent. Thus, an underestimated limit that prevents from accurate pangenome construction is the annotation step. Automated gene modeling is strongly dependent on the methods, tools, thresholds, used so that two annotations of the same genome are systematically different. Additionally, these differences are not only background noise. For instance, when the IWGSC RefSeq Annotation v1.0 was produced by combining independent predictions from two pipelines (TriAnnot and PGSB), 20% of each gene set did not overlap any prediction from the other one. Moreover, only 67 and 48% of TriAnnot and PGSB gene models were predicted with highly similar structures. These differences exceed largely the real presence/absence variations. The consequence is that pangenomic analyses are dependent on accurate mapping of a reference gene annotation to another assembly. This is why we believe annotation transfer tools like MAGATT (see paragraph RefSeq Annotation v2) are highly valuable in the pangenomic area as well as for maintaining improvements performed through manual curation. Eventually, in future wheat genome assemblies, genes will be transferred/projected from a reference pangenome and de novo annotation should be restricted to specific (non-conserved) regions. Indeed, gene projection was already applied for the annotation of chromosome pseudomolecule assemblies of 15 wheat accessions with the objective of building a wheat pangenome (Walkowiak et al. 2020). Besides the methodological challenge, issues of multiple identifiers (IDs) for a gene will become more and more problematic, as exemplified in the review of Adamski et al. (2020). Authors have highlighted the fact that one gene is already represented by many IDs, sometimes following different nomenclatures, due to the existence of multiple assemblies of the CHINESE SPRING genome sequence itself plus the release of gene models from wild wheat relatives and other cultivated genotypes. There is, thus, a strong need for integrating these data.

4.7.3 Functional Annotation: Opportunities

Automated functional annotation workflow based on sequence similarity and domain search has been established by IWGSC to assign gene ontology (GO) and function descriptions to the wheat reference gene set (IWGSC 2018). Although approaches based on local alignment search such as BLAST are straightforward and work well for certain species and gene families, the drawbacks are clear. It suffers from low sensitivity or specificity, depending on threshold choice and evolutionary distance of query gene set to species in the annotation source (Sasson et al. 2006). In addition, error or lack of robust annotation evidence in the source databases hinder or bias the large-scale functional annotation analysis, especially in non-model crop species.

To overcome these limitations, integrating various omics datasets from high-throughput experiments in combination with novel computational approaches has been considered for complementation to local sequence alignment methods, facilitating annotation of unknown genes or transferring functional knowledge from one gene to another. For example, generation and analysis of large-scale biomolecule interaction networks is a useful approach that utilizes omics data beyond gene/protein sequences. The basic idea is "guilt by association," where

a gene can be assigned a particular function if it is co-expressed with one or several genes of same known function, as the chance that they are co-regulated and needed for the same process or pathway is high (Tohge and Fernie 2012; Aoki et al. 2016). In addition to co-expression, gene–gene relationships such as protein-DNA binding and protein–protein interactions can be used to assign and transfer function from one gene to the other (Cho et al. 2016). Such interactome data can now be generated with advanced high-throughput experimental techniques such as single/bulk RNAseq, Yeast 2-Hybrid, and DNA affinity purification sequencing (DAPseq). Each type of interactome networks can be analyzed separately or in a combined manner to build multi-omics integrated network, followed by computational interpretation, from naive method of evidence aggregation to probabilistic modeling (Yu et al. 2015). The ranking or scoring reflecting proximity or connectivity of genes in the network is then used to link and transfer function from one gene to the other. Beyond the classic "single-gene" approach, integrated network-based approaches provide a more holistic view of gene function and gene–gene relationships, enabling functional annotation of unknown genes that are not related on sequence level but functionally interacted with studied genes (see also Chap. 11).

Choice of cutoff for sequence similarity-search and network mining is crucial but highly arbitrary, which can create bias or error in functional annotation process. In addition, link between various protein features (structure, text description, and interaction) and annotation label that can be utilized for functional annotation are sometimes beyond human knowledge and difficult to be revealed. In contrast, machine learning tools are suitable to identify these hidden features and assess their contribution to functions by analyzing a training set where a group of genes with these features are functionally characterized (Mahood et al. 2020). Quantitative contribution of different features learnt by computer is then exploited to predict the most possible function of unknown genes possessing same feature types. Several tools have been developed to learn relationship between GO and heterogeneous data (text and sequence information, protein structure) and propose a predictor for annotating unknown genes (Törönen et al. 2018; You et al. 2018, 2019).

Although highly advantageous compared to classical approaches, conventional machine learning is achieved using handcrafted features. Deep learning using neural networks, on the other hand, can extract abstracted and high-level features from raw data directly and build a predictor, without human inference. The availability of omics data and computational resources allows to develop sophisticated deep learning algorithms for large-scale functional annotation. Various deep learning architectures have been built using, e.g., deep, convolutional and recurrent neural network, which have specific strength in learning different features (Cao et al. 2017; Sureyya Rifaioglu et al. 2019; Du et al. 2020). Tools built on these architectures predict GO terms either by learning protein sequence (Kulmanov and Hoehndorf 2020; Cao and Shen 2021), protein structure (Tavanaei et al. 2016; Jumper et al. 2021) or heterogenous data and networks (Cai et al. 2020; Peng et al. 2021). Several factors limit the application of deep machine learning approach for functional annotation in large-scale and unbiased manner. Firstly, although various omics and structure data are useful, only primary sequence is available for majority of unknown genes. Secondly, imbalance and incompleteness of GO database with respect to species and function categories can bias the learning step, and GO prediction task itself is a complex multi-label problem. Lastly, the quality of transferring gene model information between species that are evolutionarily distant needs to be assessed carefully. Nevertheless, despite these challenges, deep machine learning-based functional annotation and GO assignment have been successfully applied and will continue in many studies, with the support of the continuing expansion of high-quality omics and experimental datasets.

References

Adamski NM, Borrill P, Brinton J, Harrington SA, Marchal C, Bentley AR, Bovill WD, Cattivelli L, Cockram J, Contreras-Moreira B, Ford B, Ghosh S, Harwood W, Hassani-Pak K, Hayta S, Hickey LT, Kanyuka K, King J, Maccaferrri M, Naamati G, Pozniak CJ, Ramirez-Gonzalez RH, Sansaloni C, Trevaskis B, Wingen LU, Wulff BB, Uauy C (2020) A roadmap for gene functional characterisation in crops with large genomes: lessons from polyploid wheat. Elife 9

Amano N, Tanaka T, Numa H, Sakai H, Itoh T (2010) Efficient plant gene identification based on inter-species mapping of full-length cDNAs. DNA Res 17:271–279

Aoki Y, Okamura Y, Tadaka S, Kinoshita K, Obayashi T (2016) ATTED-II in 2016: a plant coexpression database towards lineage-specific coexpression. Plant Cell Physiol 57:e5

Athiyannan N, Abrouk M, Boshoff WHP, Cauet S, Rodde N, Kudrna D, Mohammed N, Bettgenhaeuser J, Botha KS, Derman SS, Wing RA, Prins R, Krattinger SG (2022) Long-read genome sequencing of bread wheat facilitates disease resistance gene cloning. Nat Genet 54:227–231

Aury JM, Engelen S, Istace B, Monat C, Lasserre-Zuber P, Belser C, Cruaud C, Rimbert H, Leroy P, Arribat S, Dufau I, Bellec A, Grimbichler D, Papon N, Paux E, Ranoux M, Alberti A, Wincker P, Choulet F (2022) Long-read and chromosome-scale assembly of the hexaploid wheat genome achieves high resolution for research and breeding. GigaScience 11

Bennetzen JL, Coleman C, Liu R, Ma J, Ramakrishna W (2004) Consistent over-estimation of gene number in complex plant genomes. Curr Opin Plant Biol 7:732–736

Cai Y, Wang J, Deng L (2020) SDN2GO: an integrated deep learning model for protein function prediction. Front Bioeng Biotechnol 8:391

Cao Y, Shen Y (2021) TALE: transformer-based protein function annotation with joint sequence-label embedding. Bioinformatics 37:2825–2833. https://doi.org/10.1093/bioinformatics/btab198

Cao R, Freitas C, Chan L, Sun M, Jiang H, Chen Z (2017) ProLanGO: protein function prediction using neural machine translation based on a recurrent neural network. Molecules 22

Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J (2005) ACT: the Artemis comparison tool. Bioinformatics 21:3422–3423

Cho H, Berger B, Peng J (2016) Compact integration of multi-network topology for functional analysis of genes. Cell Syst 3:540-548.e545

Choulet F, Wicker T, Rustenholz C, Paux E, Salse J, Leroy P, Schlub S, Le Paslier MC, Magdelenat G, Gonthier C, Couloux A, Budak H, Breen J, Pumphrey M, Liu S, Kong X, Jia J, Gut M, Brunel D, Anderson JA, Gill BS, Appels R, Keller B, Feuillet C (2010) Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. Plant Cell 22:1686–1701

Choulet F, Alberti A, Theil S, Glover N, Barbe V, Daron J, Pingault L, Sourdille P, Couloux A, Paux E, Leroy P, Mangenot S, Guilhot N, Le Gouis J, Balfourier F, Alaux M, Jamilloux V, Poulain J, Durand C, Bellec A, Gaspin C, Safar J, Dolezel J, Rogers J, Vandepoele K, Aury JM, Mayer K, Berges H, Quesneville H, Wincker P, Feuillet C (2014) Structural and functional partitioning of bread wheat chromosome 3B. Science 345:1249721

Clavijo BJ, Venturini L, Schudoma C, Accinelli GG, Kaithakottil G, Wright J, Borrill P, Kettleborough G, Heavens D, Chapman H, Lipscombe J, Barker T, Lu FH, McKenzie N, Raats D, Ramirez-Gonzalez RH, Coince A, Peel N, Percival-Alwyn L, Duncan O, Trösch J, Yu G, Bolser DM, Namaati G, Kerhornou A, Spannagl M, Gundlach H, Haberer G, Davey RP, Fosker C, Palma FD, Phillips AL, Millar AH, Kersey PJ, Uauy C, Krasileva KV, Swarbreck D, Bevan MW, Clark MD (2017) An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. Genome Res 27:885–896

Dale RK, Pedersen BS, Quinlan AR (2011) Pybedtools: a flexible python library for manipulating genomic datasets and annotations. Bioinformatics 27:3423–3424

Daron J, Glover N, Pingault L, Theil S, Jamilloux V, Paux E, Barbe V, Mangenot S, Alberti A, Wincker P, Quesneville H, Feuillet C, Choulet F (2014) Organization and evolution of transposable elements along the bread wheat chromosome 3B. Genome Biol 15:546

De Oliveira R, Rimbert H, Balfourier F, Kitt J, Dynomant E, Vrána J, Doležel J, Cattonaro F, Paux E, Choulet F (2020) Structural variations affecting genes and transposable elements of chromosome 3B in wheats. Front Genet 11:891

Du Z, He Y, Li J, Uversky VN (2020) DeepAdd: protein function prediction from k-mer embedding and additional features. Comput Biol Chem 89:107379

Feuillet C, Eversole K (2007) Physical mapping of the wheat genome: a coordinated effort to lay the foundation for genome sequencing and develop tools for breeders. Israel J Plant Sci 55:307–313

Glover NM, Daron J, Pingault L, Vandepoele K, Paux E, Feuillet C, Choulet F (2015) Small-scale gene duplications played a major role in the recent evolution of wheat chromosome 3B. Genome Biol 16:188

Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, Hadley D, Hutchison D, Martin C, Katagiri F, Lange BM, Moughamer T, Xia Y, Budworth P, Zhong J, Miguel T, Paszkowski U, Zhang S, Colbert M, Sun WL, Chen L, Cooper B, Park S, Wood TC, Mao L, Quail P, Wing R, Dean R, Yu Y, Zharkikh A, Shen R, Sahasrabudhe S, Thomas A, Cannings R, Gutin A,

Pruss D, Reid J, Tavtigian S, Mitchell J, Eldredge G, Scholl T, Miller RM, Bhatnagar S, Adey N, Rubano T, Tusneem N, Robinson R, Feldhaus J, Macalma T, Oliphant A, Briggs S (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science 296:92–100

Gremme G, Steinbiss S, Kurtz S (2013) GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. IEEE/ACM Trans Comput Biol Bioinform 10:645–656

Guo W, Xin M, Wang Z, Yao Y, Hu Z, Song W, Yu K, Chen Y, Wang X, Guan P, Appels R, Peng H, Ni Z, Sun Q (2020) Origin and adaptation to high altitude of Tibetan semi-wild wheat. Nat Commun 11:5085

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. Genome Biol 9:R7

International Barley Genome Sequencing Consortium, Mayer KF, Waugh R, Brown JW, Schulman A, Langridge P, Platzer M, Fincher GB, Muehlbauer GJ, Sato K, Close TJ, Wise RP, Stein N (2012) A physical, genetic and functional sequence assembly of the barley genome. Nature 491:711–716

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature 436:793–800

IWGSC (2014) A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. Science 345:1251788

IWGSC (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science 361

Jia J, Zhao S, Kong X, Li Y, Zhao G, He W, Appels R, Pfeifer M, Tao Y, Zhang X, Jing R, Zhang C, Ma Y, Gao L, Gao C, Spannagl M, Mayer KF, Li D, Pan S, Zheng F, Hu Q, Xia X, Li J, Liang Q, Chen J, Wicker T, Gou C, Kuang H, He G, Luo Y, Keller B, Xia Q, Lu P, Wang J, Zou H, Zhang R, Xu J, Gao J, Middleton C, Quan Z, Liu G, Yang H, Liu X, He Z, Mao L, Consortium IWGS (2013) *Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation. Nature 496:91–95

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596:583–589

Kent WJ (2002) BLAT–the BLAST-like alignment tool. Genome Res 12:656–664

Kulmanov M, Hoehndorf R (2020) DeepGOPlus: improved protein function prediction from sequence. Bioinformatics 36:422–429

Leroy P, Guilhot N, Sakai H, Bernard A, Choulet F, Theil S, Reboux S, Amano N, Flutre T, Pelegrin C, Ohyanagi H, Seidel M, Giacomoni F, Reichstadt M, Alaux M, Gicquello E, Legeai F, Cerutti L, Numa H, Tanaka T, Mayer K, Itoh T, Quesneville H, Feuillet C (2012) TriAnnot: a versatile and high performance pipeline for the automated annotation of plant genomes. Front Plant Sci 3:5

Ling HQ, Zhao S, Liu D, Wang J, Sun H, Zhang C, Fan H, Li D, Dong L, Tao Y, Gao C, Wu H, Li Y, Cui Y, Guo X, Zheng S, Wang B, Yu K, Liang Q, Yang W, Lou X, Chen J, Feng M, Jian J, Zhang X, Luo G, Jiang Y, Liu J, Wang Z, Sha Y, Zhang B, Tang D, Shen Q, Xue P, Zou S, Wang X, Liu X, Wang F, Yang Y, An X, Dong Z, Zhang K, Luo MC, Dvorak J, Tong Y, Yang H, Li Z, Wang D, Zhang A (2013) Draft genome of the wheat A-genome progenitor *Triticum urartu*. Nature 496:87–90

Mahood EH, Kruse LH, Moghe GD (2020) Machine learning: a powerful tool for gene function prediction in plants. Appl Plant Sci 8:e11376

Mascher M, Gundlach H, Himmelbach A, Beier S, Twardziok SO, Wicker T, Radchuk V, Dockter C, Hedley PE, Russell J, Bayer M, Ramsay L, Liu H, Haberer G, Zhang XQ, Zhang Q, Barrero RA, Li L, Taudien S, Groth M, Felder M, Hastie A, Šimková H, Staňková H, Vrána J, Chan S, Muñoz-Amatriaín M, Ounit R, Wanamaker S, Bolser D, Colmsee C, Schmutzer T, Aliyeva-Schnorr L, Grasso S, Tanskanen J, Chailyan A, Sampath D, Heavens D, Clissold L, Cao S, Chapman B, Dai F, Han Y, Li H, Li X, Lin C, McCooke JK, Tan C, Wang P, Wang S, Yin S, Zhou G, Poland JA, Bellgard MI, Borisjuk L, Houben A, Doležel J, Ayling S, Lonardi S, Kersey P, Langridge P, Muehlbauer GJ, Clark MD, Caccamo M, Schulman AH, Mayer KFX, Platzer M, Close TJ, Scholz U, Hansson M, Zhang G, Braumann I, Spannagl M, Li C, Waugh R, Stein N (2017) A chromosome conformation capture ordered sequence of the barley genome. Nature 544:427–433

Mochida K, Yoshida T, Sakurai T, Ogihara Y, Shinozaki K (2009) TriFLDB: a database of clustered full-length coding sequences from Triticeae with applications to comparative grass genomics. Plant Physiol 150:1135–1146

Mott R (1997) EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. Comput Appl Biosci 13:477–478

Ogihara Y, Mochida K, Kawaura K, Murai K, Seki M, Kamiya A, Shinozaki K, Carninci P, Hayashizaki Y, Shin IT, Kohara Y, Yamazaki Y (2004) Construction of a full-length cDNA library from young spikelets of hexaploid wheat and its characterization by large-scale sequencing of expressed sequence tags. Genes Genet Syst 79:227–232

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Feltus

FA, Gowik U, Grigoriev IV, Lyons E, Maher CA, Martis M, Narechania A, Otillar RP, Penning BW, Salamov AA, Wang Y, Zhang L, Carpita NC, Freeling M, Gingle AR, Hash CT, Keller B, Klein P, Kresovich S, McCann MC, Ming R, Peterson DG, Mehboobur R, Ware D, Westhoff P, Mayer KF, Messing J, Rokhsar DS (2009) The *Sorghum bicolor* genome and the diversification of grasses. Nature 457:551–556

Paux E, Sourdille P, Salse J, Saintenac C, Choulet F, Leroy P, Korol A, Michalak M, Kianian S, Spielmeyer W, Lagudah E, Somers D, Kilian A, Alaux M, Vautrin S, Berges H, Eversole K, Appels R, Safar J, Simkova H, Dolezel J, Bernard M, Feuillet C (2008) A physical map of the 1-gigabase bread wheat chromosome 3B. Science 322:101–104

Peng J, Xue H, Wei Z, Tuncali I, Hao J, Shang X (2021) Integrating multi-network topology for gene function prediction using deep neural networks. Brief Bioinform 22:2096–2105

Pertea G, Pertea M (2020) GFF utilities: GffRead and GffCompare. F1000Res 9

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol 33:290–295

Pfeifer M, Kugler KG, Sandve SR, Zhan B, Rudi H, Hvidsten TR, International Wheat Genome Sequencing C, Mayer KF, Olsen OA (2014) Genome interplay in the grain transcriptome of hexaploid bread wheat. Science 345:1250091

Pingault L, Choulet F, Alberti A, Glover N, Wincker P, Feuillet C, Paux E (2015) Deep transcriptome sequencing provides new insights into the structural and functional organization of the wheat genome. Genome Biol 16:29

Sasson O, Kaplan N, Linial M (2006) Functional annotation prediction: all for one and one for all. Protein Sci 15:1557–1562

Sato K, Abe F, Mascher M, Haberer G, Gundlach H, Spannagl M, Shirasawa K, Isobe S (2021) Chromosome-scale genome assembly of the transformation-amenable common wheat cultivar 'Fielder'. DNA Res 28

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, Chen W, Yan L, Higginbotham J, Cardenas M, Waligorski J, Applebaum E, Phelps L, Falcone J, Kanchi K, Thane T, Scimone A, Thane N, Henke J, Wang T, Ruppert J, Shah N, Rotter K, Hodges J, Ingenthron E, Cordes M, Kohlberg S, Sgro J, Delgado B, Mead K, Chinwalla A, Leonard S, Crouse K, Collura K, Kudrna D, Currie J, He R, Angelova A, Rajasekar S, Mueller T, Lomeli R, Scara G, Ko A, Delaney K, Wissotski M, Lopez G, Campos D, Braidotti M, Ashley E, Golser W, Kim H, Lee S, Lin J, Dujmic Z, Kim W, Talag J, Zuccolo A, Fan C, Sebastian A, Kramer M, Spiegel L, Nascimento L, Zutavern T, Miller B, Ambroise C, Muller S, Spooner W, Narechania A, Ren L, Wei S, Kumari S, Faga B, Levy MJ, McMahan L, Van Buren P, Vaughn MW, Ying K, Yeh CT, Emrich SJ, Jia Y, Kalyanaraman A, Hsia AP, Barbazuk WB, Baucom RS, Brutnell TP, Carpita NC, Chaparro C, Chia JM, Deragon JM, Estill JC, Fu Y, Jeddeloh JA, Han Y, Lee H, Li P, Lisch DR, Liu S, Liu Z, Nagel DH, McCann MC, SanMiguel P, Myers AM, Nettleton D, Nguyen J, Penning BW, Ponnala L, Schneider KL, Schwartz DC, Sharma A, Soderlund C, Springer NM, Sun Q, Wang H, Waterman M, Westerman R, Wolfgruber TK, Yang L, Yu Y, Zhang L, Zhou S, Zhu Q, Bennetzen JL, Dawe RK, Jiang J, Jiang N, Presting GG, Wessler SR, Aluru S, Martienssen RA, Clifton SW, McCombie WR, Wing RA, Wilson RK (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112–1115

Slater GS, Birney E (2005) Automated generation of heuristics for biological sequence comparison. BMC Bioinform 6:31

Smit AFA, Hubley R, Green P (1996–2004) RepeatMasker Open-3.0. http://www.repeatmasker.org

Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B (2006) AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res 34:W435-439

Sureyya Rifaioglu A, Doğan T, Jesus Martin M, Cetin-Atalay R, Atalay V (2019) DEEPred: automated protein function prediction with multi-task feed-forward deep neural networks. Sci Rep 9:7344

Tavanaei A, Maia AS, Kaniymattam A, Loganantharaj (2016) Towards recognition of protein function based on its structure using deep convolutional networks. In: IEEE international conference on bioinformatics and biomedicine (BIBM). https://doi.org/10.1109/BIBM.2016.7822509

The International Brachypodium Initiative (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature 463:763–768

Tohge T, Fernie AR (2012) Annotation of plant gene function via combined genomics, metabolomics and informatics. J Vis Exp e3487

Törönen P, Medlar A, Holm L (2018) PANNZER2: a rapid functional annotation web server. Nucleic Acids Res 46:W84–W88

UniProt Consortium (2018) UniProt: the universal protein knowledgebase. Nucleic Acids Res 46:2699

Venturini L, Caim S, Kaithakottil GG, Mapleson DL, Swarbreck D (2018) Leveraging multiple transcriptome assembly methods for improved gene structure annotation. Gigascience 7. https://doi.org/10.1093/gigascience/giy093

Walkowiak S, Gao L, Monat C, Haberer G, Kassa MT, Brinton J, Ramirez-Gonzalez RH, Kolodziej MC, Delorean E, Thambugala D, Klymiuk V, Byrns B, Gundlach H, Bandi V, Siri JN, Nilsen K, Aquino C,

Himmelbach A, Copetti D, Ban T, Venturini L, Bevan M, Clavijo B, Koo DH, Ens J, Wiebe K, N'Diaye A, Fritz AK, Gutwin C, Fiebig A, Fosker C, Fu BX, Accinelli GG, Gardner KA, Fradgley N, Gutierrez-Gonzalez J, Halstead-Nussloch G, Hatakeyama M, Koh CS, Deek J, Costamagna AC, Fobert P, Heavens D, Kanamori H, Kawaura K, Kobayashi F, Krasileva K, Kuo T, McKenzie N, Murata K, Nabeka Y, Paape T, Padmarasu S, Percival-Alwyn L, Kagale S, Scholz U, Sese J, Juliana P, Singh R, Shimizu-Inatsugi R, Swarbreck D, Cockram J, Budak H, Tameshige T, Tanaka T, Tsuji H, Wright J, Wu J, Steuernagel B, Small I, Cloutier S, Keeble-Gagnère G, Muehlbauer G, Tibbets J, Nasuda S, Melonek J, Hucl PJ, Sharpe AG, Clark M, Legg E, Bharti A, Langridge P, Hall A, Uauy C, Mascher M, Krattinger SG, Handa H, Shimizu KK, Distelfeld A, Chalmers K, Keller B, Mayer KFX, Poland J, Stein N, McCartney CA, Spannagl M, Wicker T, Pozniak CJ (2020) Multiple wheat genomes reveal global variation in modern breeding. Nature 588:277–283

Wicker T, Matthews DE, Keller B (2002) TREP: a database for Triticeae repetitive elements. Trends Plant Sci 7:561–562

Wicker T, Gundlach H, Spannagl M, Uauy C, Borrill P, Ramirez-Gonzalez RH, De Oliveira R, International Wheat Genome Sequencing C, Mayer KFX, Paux E, Choulet F (2018) Impact of transposable elements on genome structure and evolution in bread wheat. Genome Biol 19:103

Wu TD, Watanabe CK (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21:1859–1875

Yandell M, Ence D (2012) A beginner's guide to eukaryotic genome annotation. Nat Rev Genet 13:329–342

You R, Huang X, Zhu S (2018) DeepText2GO: improving large-scale protein function prediction with deep semantic text representation. Methods 145:82–90

You R, Yao S, Xiong Y, Huang X, Sun F, Mamitsuka H, Zhu S (2019) NetGO: improving large-scale protein function prediction with massive network information. Nucleic Acids Res 47:W379–W387. https://doi.org/10.1093/nar/gkz388

Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, Cao M, Liu J, Sun J, Tang J, Chen Y, Huang X, Lin W, Ye C, Tong W, Cong L, Geng J, Han Y, Li L, Li W, Hu G, Li J, Liu Z, Qi Q, Li T, Wang X, Lu H, Wu T, Zhu M, Ni P, Han H, Dong W, Ren X, Feng X, Cui P, Li X, Wang H, Xu X, Zhai W, Xu Z, Zhang J, He S, Xu J, Zhang K, Zheng X, Dong J, Zeng W, Tao L, Ye J, Tan J, Chen X, He J, Liu D, Tian W, Tian C, Xia H, Bao Q, Li G, Gao H, Cao T, Zhao W, Li P, Chen W, Zhang Y, Hu J, Liu S, Yang J, Zhang G, Xiong Y, Li Z, Mao L, Zhou C, Zhu Z, Chen R, Hao B, Zheng W, Chen S, Guo W, Tao M, Zhu L, Yuan L, Yang H (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science 296:79–92

Yu G, Rangwala H, Domeniconi C, Zhang G, Zhang Z (2015) Predicting protein function using multiple kernels. IEEE/ACM Trans Comput Biol Bioinform 12:219–233

Zhang Z, Schwartz S, Wagner L, Miller W (2000) A greedy algorithm for aligning DNA sequences. J Comput Biol 7:203–214

Zhang D, Choi DW, Wanamaker S, Fenton RD, Chin A, Malatrasi M, Turuspekov Y, Walia H, Akhunov ED, Kianian P, Otto C, Simons K, Deal KR, Echenique V, Stamova B, Ross K, Butler GE, Strader L, Verhey SD, Johnson R, Altenbach S, Kothari K, Tanaka C, Shah MM, Laudencia-Chingcuanco D, Han P, Miller RE, Crossman CC, Chao S, Lazo GR, Klueva N, Gustafson JP, Kianian SF, Dubcovsky J, Walker-Simmons MK, Gill KS, Dvorák J, Anderson OD, Sorrells ME, McGuire PE, Qualset CO, Nguyen HT, Close TJ (2004) Construction and evaluation of cDNA libraries for large-scale expressed sequence tag sequencing in wheat (*Triticum aestivum* L.). Genetics 168:595–608

Zhu T, Wang L, Rimbert H, Rodriguez JC, Deal KR, De Oliveira R, Choulet F, Keeble-Gagnère G, Tibbits J, Rogers J, Eversole K, Appels R, Gu YQ, Mascher M, Dvorak J, Luo MC (2021) Optical maps refine the bread wheat *Triticum aestivum* cv Chinese spring genome assembly. Plant J 107:303–314

Zimin AV, Puiu D, Hall R, Kingan S, Clavijo BJ, Salzberg SL (2017) The first near-complete assembly of the hexaploid bread wheat genome, *Triticum aestivum*. Gigascience 6:1–7

5 The Wheat Transcriptome and Discovery of Functional Gene Networks

Tayyaba Andleeb, James Milson and Philippa Borrill

Abstract

Gene expression patterns have been a widely applied source of information to start understanding gene function in multiple plant species. In wheat, the advent of increasingly accurate and complete gene annotations now enables transcriptomic studies to be carried out on a routine basis and studies by groups around the world have compared gene expression changes under an array of environmental and developmental stages. However, associating data from differentially expressed genes to understanding the biological role of these genes and their applications for breeding is a major challenge. Recently, the first steps to apply network-based approaches to characterise gene expression have been taken in wheat and these networks have enabled the prediction of gene functions in wheat but only for a handful of traits. Combining advanced analysis methods with better sequencing technology will increase our capacity to place gene expression in wheat in the context of functions of genes that influence agronomically important traits.

Keywords

Wheat transcriptome · Gene networks · Response to environment · Development

5.1 Gene Function Through Gene Expression

In order to understand gene function, one of the first things researchers would like to do is measure gene expression—when, where and how much of a gene's transcript is present? Measuring the expression level of a single gene through quantitative PCR can reveal insight into a specific gene and its potential biological role. However, to explore the integrated nature of gene expression and how entire biological processes work at the transcriptional level, it is desirable to measure the expression level of multiple genes simultaneously using transcriptomics. In model species, transcriptomics has shed insight into the regulation of developmental processes, responses to the environment and

Tayyaba Andleeb and James Milson have contributed equally to this work.

T. Andleeb
Department of Plant Sciences, Faculty of Biological Sciences, Quaid-i-Azam University, Islamabad, Pakistan

T. Andleeb · J. Milson
School of Biological Sciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK

J. Milson · P. Borrill (✉)
Department of Crop Genetics, John Innes Centre, Norwich Research, Norwich NR4 7UH, UK
e-mail: philippa.borrill@jic.ac.uk

R. Appels et al. (eds.), *The Wheat Genome*, Compendium of Plant Genomes,
https://doi.org/10.1007/978-3-031-38294-9_5

genotype-specific responses, all of which would be highly advantageous to understand for wheat improvement. Therefore, transcriptomics has been widely applied in wheat biology.

Initially, transcriptomics largely relied upon microarray approaches. These were useful in determining gene expression patterns, but microarrays in wheat were limited because of the incomplete gene model annotations available when microarrays were designed, therefore many genes were missing from the arrays. The advent of RNA-seq to measure gene expression enabled more accurate measurement of the wheat transcriptome. Transcriptomics could be applied even before high-quality genome assemblies were available because de novo transcriptome assemblies could be generated to answer specific biological questions using individual datasets. However, to get the highest quality and most comprehensive results in a transcriptomic experiment, having a reference transcriptome is valuable and also removes the requirement to carry out a de novo assembly for each new project. Furthermore, the availability of a reference transcriptome facilitates the identification of homoeolog-specific transcripts and therefore allows gene expression to be quantified in a homoeolog-specific manner.

5.2 Measuring Homoeolog-Specific Gene Expression

As consequence of the polyploid nature of wheat, >50% of genes in the wheat genome are present as triads of related homoeologous genes on the A, B and D subgenomes (IWGSC et al. 2018). Studies on a gene-by-gene basis have revealed that each homoeolog in wheat can have different expression levels. For example, the calcium-dependent protein kinase *TaCPK2* has differential responses to stress between homeologs with the A homeolog upregulated in response to powdery mildew infection and the D homoeolog upregulated in response to cold stress (Geng et al. 2013). However, to analyse homoeolog-specific expression using qPCR is labour-intensive and requires the design of homoeolog-specific primers for each gene of interest. The use of transcriptomics allows quicker and easier homoeolog-specific gene expression measurements. Several different ways to quantify homoeolog-specific gene expression in allopolyploids have been implemented including alignment to the individual subgenomes and read classification according to mismatches or inter-homoeolog SNPs (Kuo et al. 2020), alignment to the whole genome sequence using a standard aligner and selecting only uniquely mapping reads (e.g. He et al. 2022) or pseudoalignment to the transcriptome using kallisto which has been demonstrated to assign reads to appropriate homoeolog using nullitetrasomic lines (Borrill et al. 2016; Ramírez-González et al. 2018). Homoeolog-specific gene analysis has been used to study multiple biological questions and has for example revealed homoeolog-specific gene expression responses to stress conditions (e.g. Clavijo et al. 2017) and developmental stage and tissue-specific homoeolog expression (Ramírez-González et al. 2018). In order to maximise information gained from applying transcriptomic approaches, it is necessary to define which genes are present within the genome and have accurate gene annotations to capture the complexities of gene expression in this polyploid species.

5.3 Building Transcriptome Annotations in Wheat

5.3.1 Expressed Sequence Tags and Full-Length cDNAs

The large size of the wheat genome made sequencing the entire wheat genome and the genes within it a difficult prospect in the 1990s and 2000s due to the high cost and sequencing technology limitations (see also Chap. 1). However, the importance and usefulness of having gene sequence information was clear. An alternative way to obtain gene sequence focussed on expressed sequence tags (ESTs), which provided a quicker way to determine

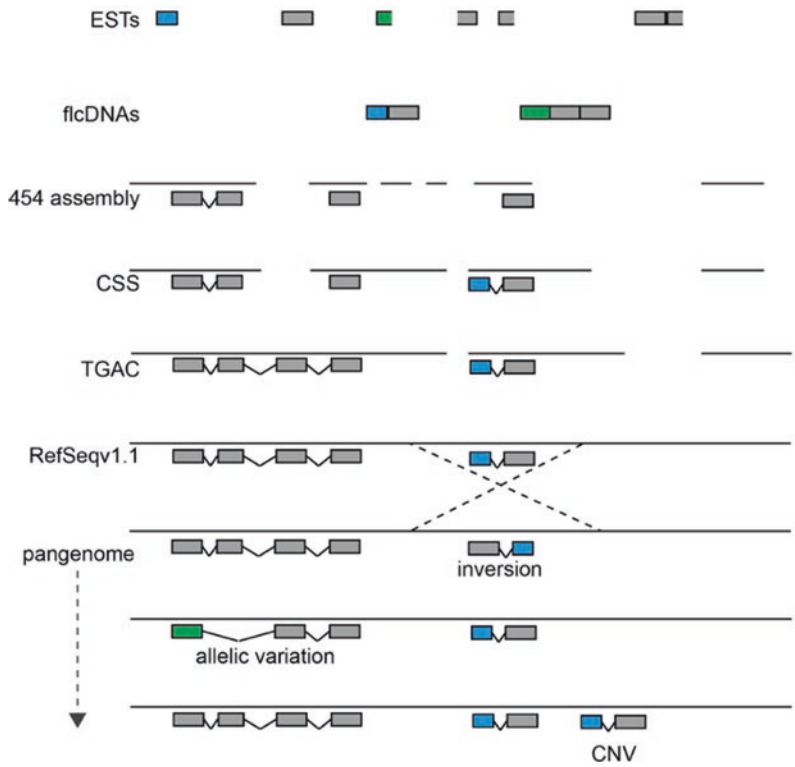


Fig. 5.1 Improvements in transcriptome assemblies in the last 20 years. Transcriptome sequences have progressed from expressed sequence tags (EST) which had unknown chromosomal positions and were often partial sequences, through full-length cDNAs (flcDNAs) to the initial genome assemblies (454 assembly) which often lacked annotation, through to fragmented assemblies with gene model predictions such as the CHINESE SPRING Survey (CSS) and The Genome Analysis Centre (TGAC) assembly, to highly complete transcriptome assemblies on contiguous chromosome-scale scaffolds (RefSeqv1.1). Sequencing and assembly of transcriptomes for multiple wheat cultivars will reveal the pan-transcriptome and variation therein including copy number variation (CNV)

gene sequences and expression information (Fig. 5.1). ESTs were generated by extracting RNA from a tissue or tissues of interest and building a cDNA library in *E. coli*. Plasmids from the *E. coli* library were extracted and sequenced through Sanger sequencing before bioinformatic analysis to group sequences into contigs containing related sequences. ESTs were generated from multiple wheat tissues (Ogihara et al. 2003; Manickavelu et al. 2012) and samples grown under stress conditions (Chao et al. 2006; Mochida et al. 2006) resulting in the identification of over 1 million EST sequences grouped into tens of thousands of contigs. By filtering these contigs for sequences containing both start and stop codons, it was possible to identify full-length cDNA representing entire coding sequences, although the numbers were significantly lower than the number of ESTs. For example, the 1 million EST generated by Manickavelu et al. (2012) were classified into 37,138 contigs of which ~ 7000 were full length. Significant efforts were made to obtain a good representation of full-length cDNAs, and the resulting sequences (~ 20,000 full-length cDNAs) were gathered into databases (Kawaura et al. 2009; Mochida et al. 2009).

5.3.2 Integrating Gene Annotation into Genome Assemblies

In parallel with the development of flcDNA libraries, many groups embarked upon

projects to sequence the wheat genome. The first sequence of a wheat genome with associated gene annotations was published in 2012 using the cultivar CHINESE SPRING (Brenchley et al. 2012). The low sequencing coverage (5x) using 454 technology meant that the assembly was highly fragmented (over 5 million scaffolds), yet it was extremely useful to researchers offering the first extensive set of genomic sequences. Approximately 95,000 genes were annotated using orthologs to flcDNAs from rice, sorghum, Brachypodium and barley. Two-thirds of these genes were assigned to the A, B or D subgenome but it was not possible to assign genes to individual chromosomes. This data provided larger number of gene annotations than were available from flcDNAs, although not all flcDNAs were represented and many of the gene models were fragmented (Fig. 5.1). Nonetheless, this assembly illustrated that whole genome sequencing of wheat was possible and could make major contributions to generating a complete set of gene models.

The next major improvement in gene models was achieved by applying flow-sorting technology to separate individual chromosome arms prior to sequencing (see Chap. 3). This allowed gene models to be assigned to individual chromosome arms, identifying homoeologous genes with confidence, and positional information was added through the use of synteny and genetic mapping approaches. In total 124,201 genes were annotated and assigned to individual chromosomes, and 75,183 had positional information. These genes were located across a total 10.2 Gb assembly of CHINESE SPRING (the CHINESE SPRING Survey; CSS; Fig. 5.1; IWGSC et al. 2014). However, the fragmented nature of this assembly with only 70% of the assembly in contigs longer than 1 kb, meant that although the number of genes identified was high, many genes were not full length for example due to a gene model being truncated at the end of a contig (Brinton et al. 2018).

Improvements to assembling complete gene models came largely through improved contiguity in genome assemblies. The use of varying sized mate-pair libraries and a new assembly algorithm produced a new CHINESE SPRING assembly (Clavijo et al. 2017) with a longer contig size with over 80% of the assembly having contigs larger than 32 kb. In total 104,091 gene models were annotated, which is ~ 20,000 genes fewer than in the CSS assembly (IWGSC et al. 2014), but these new gene models were generally more complete because the higher assembly contiguity meant it was much less likely that a gene model was truncated at the end of a contig (Fig. 5.1). An additional CHINESE SPRING assembly (Triticum3.1) achieved much-increased contiguity by combining Illumina short reads with PacBio long reads, with over 50% of the assembly having contigs larger than 232 kb (Zimin et al. 2017), but this assembly lacked gene annotations.

The next step change came with the publication of the RefSeqv1.0 CHINESE SPRING genome assembly (IWGSC et al. 2018). This pseudomolecule-level 14.5 Gb assembly used a de novo assembly approach, an improved assembly method and additional layers of genetic, physical and sequencing data to generate a long-range ordered assembly with accurate assignment of homoeologs. In total 107,891 high-confidence genes were annotated by combining the outputs of two prediction pipelines. These gene models represented a higher proportion of conserved BUSCO single-copy genes than previous assemblies with 90% of BUSCO genes present as three complete copies in the RefSeq assembly, compared to 70% in the TGAC assembly and 25% in the CSS assembly. Approximately, 2,000 gene models were manually refined, resulting in the RefSeqv1.1 gene model set (Fig. 5.1).

Although highly complete, further improvements have been made to these gene models. By combining the long-read-based Triticum_aestivum_3.1 genome assembly with information from the RefSeqv1.0 assembly to improve scaffolding and annotation, a more complete (15.1 GB) annotated CHINESE SPRING assembly was obtained: *Triticum_aestivum*_4.0 (Alonge et al. 2020). The use of long reads enabled many repeat regions to be expanded in this assembly, including regions containing

thousands of additional gene copies. This gave a total of 108,639 genes localised to individual chromosomes. In parallel, further refinements were made to the RefSeqv1.0 by incorporating optical maps and PacBio long reads to generate RefSeqv2.1 (Zhu et al. 2021). Although the total assembly size did not change much (14.6 GB in RefSeqv2.1 vs. 14.5 GB in RefSeqv1.0), positions and orientations of scaffolds were corrected for 10% of the genome and gaps were filled. In total 106,913 high-confidence genes were annotated by aligning gene annotations from the RefSeqv1.1 and community annotations.

5.3.3 Remaining Challenges to Improve the Accuracy and Completeness of the Gene Model Set

Discrepancies remain between the *Triticum_aestivum*_4.0 and RefSeqv2.1 assemblies in some regions, and integration of new data types will be required to resolve localised gaps or errors, and to assign all scaffolds to accurate positions. Gene annotations may also be inaccurate in a minority of regions due to remaining gaps or inaccuracies. Both these assemblies rely on the transfer of gene models from RefSeqv1.1, so there may be value in re-annotating these genomes from de novo predictions and RNA-seq data to take advantage of these more accurate sequences. A final consequence of relying largely on the RefSeqv1.1 gene models is that alternative spliced isoforms may not be fully represented with only 15.7% of high-confidence genes having alternative isoforms (IWGSC et al. 2018), due to conservative parameters used during the transcriptome assembly.

Although technical challenges remain to perfect the CHINESE SPRING gene models, a more pressing challenge will be to identify variation between gene models in different wheat cultivars. Work by Montenegro et al. (2017) showed that gene content was variable between 18 wheat cultivars, with ~ 81,000 genes shared between all cultivars and an additional 60,000 genes detected in at least one cultivar. The large average number of genes detected in each cultivar in this study (128,656) may be an artefact of basing gene model discovery on the fragmented CSS assembly; nonetheless, the variation in gene models is likely to have significant consequences to understanding wheat biology (see Chap. 4). More recently whole genome sequencing of 15 cultivars in additional to CHINESE SPRING revealed extensive structural and haplotype divergence between wheat cultivars (Fig. 5.1; Walkowiak et al. 2020). Significant differences were found in gene content between cultivars with ~ 12% of genes showing presence–absence variation, although this was based on projecting gene annotations from CHINESE SPRING, rather than de novo genome annotation tailored to each cultivar. Individual genome annotations for each of these high-quality genome sequences will be a valuable resource for biologists and breeders alike and is likely to identify genes absent from CHINESE SPRING.

Beyond increasing the number of cultivars, it will also be important to increase the accuracy of gene models beyond the coding region, which is so far the most accurate portion of wheat gene models. The 5′ and 3′ untranslated regions are annotated in many genes, but their accuracy is not known and specialised next-generation sequencing approaches could be used, such as CAGE-seq to identify transcription start sites and PolyA-seq to identify transcription end sites, as has been done in cotton to generate accurate untranslated region annotations (Wang et al. 2019). The use of PacBio Iso-seq long reads in conjunction with Illumina short reads and stringent filtering can also increase the accuracy of transcript start and end sites, as well as providing information about splice junctions. This has been achieved in wheat's close relative barley (Coulter et al. 2021). This approach identified that 73% of multi-exonic barley genes had two or more transcript isoforms, suggesting that the current wheat annotations may be missing transcript isoforms in many multi-exonic genes.

5.4 Methods of Measuring Gene Expression at the Genome-Wide Level

The availability of high-quality gene models now facilitates the accurate measurements of gene expression using RNA-seq. The most common type of RNA-seq is the enrichment and subsequent sequencing of polyadenylated RNA to study mRNA levels. Reduced representation sequencing can also be applied to reduce costs. For example, 3′ end sequencing can be used for investigating the expression profile of genes at a lower cost due to reduced sequencing requirements and targeted RNA-seq can be used to sequence-specific targets, primarily those with low expression profiles. More recently, low input RNA-seq methods from small tissues to single-cell approaches have been developed. These enable the measurement of gene expression in different cell types and determine co-expression and gene regulation in single cells, although their application in wheat remains limited.

5.5 Diverse Biological Questions Can Be Answered with Transcriptomics

Transcriptomics approaches have been applied in many different types of studies in wheat. These include observing changes in the transcriptome over a developmental time course, studying gene expression responses to different stresses or investigating the effect of a specific gene on downstream molecular pathways (Fig. 5.2).

5.6 Elucidating Genetic Control of Developmental Processes

Transcriptomic approaches can help build understanding of developmental processes by studying gene expression throughout a time course or by focussing on the transcriptional changes induced by manipulating a gene regulating development, for example through mutants or overexpression. Here we will discuss typical approaches which use RNA-seq to understand developmental processes in wheat.

5.6.1 Studying Gene Expression During Time Courses

Grain development is an important process which influences final yield and quality in all cereal crops and has therefore been examined at the transcriptomic level by several groups. For example, using the CHINESE SPRING Survey (CSS) sequence annotation, Pfeifer et al. (2014) identified cell-type and homoeolog-specific gene expression during grain development at three timepoints. Building upon this work Chi et al. (2019) studied gene expression across four timepoints in grain development, although they did not dissect grains into individual cell types. Differentially expressed genes were clustered into groups based on developmental stages and assigned putative functions based on gene ontology (GO) and Kyoto Encyclopaedia of Genes and Genomes (KEGG) enrichment analyses. Many more differentially expressed genes were identified than was possible using previous microarray-based approaches and the more accurate and complete gene models facilitated the analysis (Yu et al. 2016). A similar approach was used to investigate wheat spike development at four different stages (Feng et al. 2017). Clustering analysis of genes differentially expressed over the time course identified dynamically expressed transcription factors which the authors hypothesise may regulate spikelet initiation and floral organ patterning, inferred from their times of expression and orthologs in model plants. The putative functions of the differentially expressed genes found in this study were assigned using GO enrichment analysis, giving an insight into the functions of individual genes as well as temporal dynamics of expression (Feng et al. 2017).

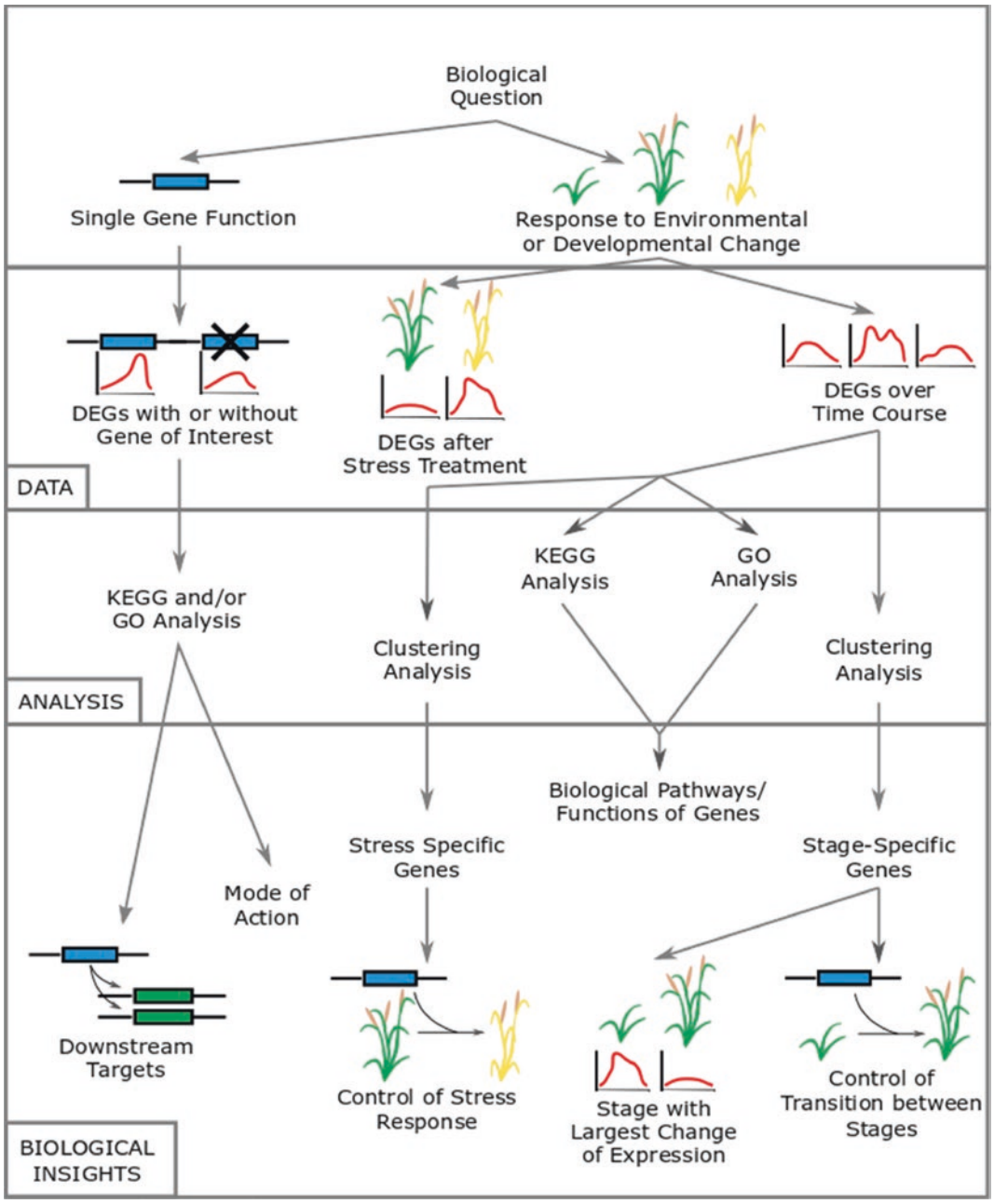


Fig. 5.2 RNA-seq is frequently used to assess the effects of altering a single gene or environmental/developmental change on gene expression. The data collected is used to identify differentially expressed genes (DEGs) which can then be analysed through methods including Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway or gene ontology (GO) analysis, or by clustering gene expression profiles. Specific exploring of differentially expressed genes, pathway and clustering information can uncover the biological pathways and mechanisms through which a gene or environmental/developmental response operates

5.6.2 Understanding the Influence of Individual Genetic Components on a Developmental Process

Understanding general expression changes during development is important, but many geneticists aim to characterise the precise effects of individual genes and RNA-seq can contribute to this goal. Flowering time is one of the best-characterised processes in wheat with many important genes identified. Transcriptomic approaches have deepened our understanding of flowering time pathways by comparing the expression profiles of wild type and plants mutated in or overexpressing key floral regulators (see also Chap. 11). For example, Pearce et al. (2016) studied the phytochrome light receptors using RNA-seq-based methods to better understand how they regulate the developmental transitions controlled by changes in light levels. Under long-day conditions, *PHYB* was found to regulate approximately six times more genes than *PHYC* and only a small number of genes were under transcriptional control of both phytochrome genes. Similarly, under short-day conditions *PHYB* influenced the transcription of approximately five times more genes than *PHYC* (Kippes et al. 2020). Surprisingly in *phyB* and *phyC* mutants flowering was accelerated under short-day conditions, which is unexpected in a long-day plant like wheat. Transcriptomic analysis revealed this may be mediated through flowering promoting genes *VRN-A1* and *PPD-B1*. This work shows that these RNA-seq transcriptome methods can uncover the functions of genes in a developmental process as well as identify downstream targets of these genes.

5.6.3 Atlases of Gene Expression

Beyond individual studies of gene expression, collating gene expression data for future analysis via gene expression atlases allows researchers to address a range of biological questions without the need to carry out more RNA-sequencing. Several different atlases have been built for wheat including the expVIP gene expression atlas which contains RNA-seq data from > 1,000 RNA-seq samples, including diverse tissue types, developmental stages, cultivars and environmental conditions (Borrill et al. 2016; Ramírez-González et al. 2018). A pictorial representation of gene expression across 70 different tissue-developmental stages is also available through the wheat eFP browser which provides a powerful tool for intuitive gene expression exploration (Winter et al. 2007; Ramírez-González et al. 2018).

5.7 Response to Environmental Stress

Transcriptome analyses are also a powerful tool to understand how wheat responds to different environmental stresses, including both abiotic and biotic stresses. Genome-wide scale changes in the transcriptome can be investigated by examining the transcriptome changes after the application of the stress or differences between plants with susceptible or resistant genotypes. The effect of single genes on the response can be investigated by comparing lines with precise genetic differences such as near-isogenic lines, overexpression or mutant lines.

5.7.1 Genome-Wide Transcriptional Responses to Stress Conditions

RNA-seq has been used to characterise gene expression changes in response to a wide range of environmental stresses from pathogen infection (e.g. Zhang et al. 2014; Dobon et al. 2016) through to abiotic stresses including drought, heat, salinity and cold (e.g. Liu et al. 2015; Xiong et al. 2017; Li et al. 2018; Gálvez et al. 2019). The effects of yellow rust infection on gene expression is one of the best studied pathogen infections in wheat, at the transcriptional level. Here we will explore insights that have been gained using RNA-seq to study rust infection, which may be widely applicable to other pathosystems and to other environmental interactions.

Early studies using RNA-seq examined temporal changes in gene expression in wheat (Zhang et al. 2014), or in both wheat and the fungal pathogen itself revealing temporal interactions between host and pathogen (Dobon et al. 2016). Comparisons between susceptible and resistant lines have also proved fruitful. Infection with a mixture of powdery mildew and leaf rust revealed that a specific set of genes were downregulated only in the susceptible line. These genes had functions related to programmed cell death and response to cellular damage, indicating that the two fungal pathogens evade the wheat defense system by inducing transcriptional level changes (Poretti et al. 2021). This agrees with earlier results which examined a time course of RNA-seq in wheat plants infected with yellow rust. Immune response regulators were rapidly upregulated after yellow rust infection, but this upregulation was suppressed in subsequent timepoints. Only in resistant interactions was this suppression alleviated, while in susceptible reactions the immune response regulators continued to be suppressed (Dobon et al. 2016). This parallels the findings of Poretti et al. (2021) that specific suppression is required in susceptible wheat lines for successful infection.

Transcriptomics studies are also now leading to the identification and functional characterisation of genes involved in pathogen resistance and susceptibility. Corredor-Moreno et al. (2021) used data from 68 pathogen-infected wheat varieties to investigate genes which influence wheat rust susceptibility. Since samples were collected from different varieties, growth conditions and developmental stages, the authors clustered gene expression profiles to identify genes linked to yellow rust susceptibility. This reduced the amount of background differentially expressed genes which are not involved in the infection response, but instead are linked to variety, growth condition or developmental stage. By focussing on clusters which showed strong expression differences between the most and least susceptible cultivars, susceptibility-associated genes were identified. These susceptibility-associated genes were enriched for branched-chain amino acid (BCAA) biosynthetic genes. Comparison with publicly available data highlighted the gene *branched-chain aminotransferase 1* (*TaBCAT1*) as a candidate gene, which was ultimately validated as a susceptibility gene using mutant lines. This study highlights a new way of identifying genes with roles in infection response and shows the potential genetic variation we can find beyond the pairwise comparisons of lines with different susceptibilities, which is the more routine approach.

5.7.2 Elucidating Biological Mechanisms of Stress-Associated Genes Using Transcriptomics

It is becoming increasingly routine to characterise lines with phenotypic alternations in stress responses using RNA-seq. This can provide insight into the molecular pathways through which a gene involved in stress responses operates and identify future breeding targets downstream in the process.

Taking drought stress as an example, several studies have recently associated NAC transcription factors with drought tolerance and studied the pathways through which they act. The first NAC gene (*TaSNAC8-6A*) improved seedling stage drought tolerance (Mao et al. 2020). RNA-seq analysis in roots showed that even under well-watered conditions, genes with GO terms associated with drought, auxin and ABA responses were upregulated in lines overexpressing this gene. Under drought conditions, more genes associated with drought, auxin and ABA response were upregulated, in the overexpression line than in well-watered conditions. The authors hypothesise that these changes enhance root development and increase water use efficiency, leading to increased drought tolerance. The second NAC (*TaNAC071-A*) increased yield under drought conditions by increasing water use efficiency (Mao et al. 2022). RNA-seq in leaves revealed that stress-responsive pathways such as response to abscisic acid and response to osmotic stress were upregulated in

lines overexpressing this NAC. Furthermore, orthologs of well-established drought-inducible genes were upregulated in the overexpression lines including genes involved in stomatal closure, suggesting that *TaNAC071-A* may increase drought tolerance by more quickly closing the stomata and reducing the transpiration rate. Interestingly, a separate study revealed through RNA-seq that increasing stomatal closure under drought is a common mechanism controlled by NAC transcription factors in wheat Ma et al. (2022).

5.8 Limitations of Current Transcriptomic Studies

A common limitation in many species is that RNA-seq has generally been carried out on pooled tissue which results in the loss of a large amount of potential information from single cells or individual tissue types. For example, by sampling a whole leaf and grinding it up prior to RNA extraction, the generated expression profiles are an average across many cell types. Therefore, any spatial differences expression within a tissue cannot be observed. Until recently, large quantities of RNA were needed for RNA-seq; therefore in order to study specific cell/tissue types, labour-intensive methods had to be used to gather large quantities of material such as aleurone and endosperm from developing grain (Pfeifer et al. 2014) and developing meiocytes (Martín et al. 2018). However, the development of low input RNA-seq methods now allows gene expression studies with much reduced sample collection requirements and enables studies on very small tissue samples which were not feasible before. Low input methods were used by Backhaus et al. (2022) to investigate the gene expression patterns in different regions of the developing spike. The developing spike was dissected at double ridge and glume primordia stage into three sections (apical, central, basal) for sequencing, without any pooling of different samples required. Surprisingly Backhaus et al. (2022) found that the largest differences in the transcriptome were between the basal and apical sections, rather than between different consecutive timepoints of development. The discovery that position has a stronger effect than the developmental time point could not have been made by doing bulk-RNA-seq of the whole spike, as has been done by previous studies (e.g. Feng et al. 2017), uncovering the unique and powerful information available using this low input approach.

While the ability to sequence small samples is a major step forwards, resolution at the single-cell level is now being applied in other plant species such as *Arabidopsis* (Thibivilliers et al. 2020). However, single-cell RNA-seq (scRNA-seq) still has limitations including the complexity of the method itself, mainly the capture of single cells (Chen et al. 2019) and the risk of overamplification based on the small amount of RNA provided from a single or small number of cells (Hrdlickova et al. 2017). However, the main issue for scRNA-seq in plant transcriptomics is the need to degrade the cell wall, with the different compositions and types meaning different protocols are required (Thibivilliers et al. 2020). The application of scRNA-seq will present new opportunities for wheat research, and success in applying this method to monocots such as rice and maize (e.g. Xu et al. 2021; Zhang et al. 2021) lay the groundwork for future studies.

A second key limitation of many studies to date has been the use of glasshouse and controlled environment conditions, to minimise variations in transcriptome changes due to factors other than what is being experimentally manipulated. However, this is not necessarily indicative of gene expression during development or responses to stress in the field environment. It is becoming increasingly important to understand gene expression in real-world fluctuating environments, and field-based studies are becoming more common (e.g. Quijano et al. 2015; Li et al. 2018; Corredor-Moreno et al. 2021). Field-based studies can develop increased insight into biological pathways and provide important information for breeding. For example, a field-based experiment revealed that multiple interactive pathways that influence cold tolerance to

prepare for over-winter stress, and these complex interactions may have been missed in controlled environment conditions where changes are often abrupt (Li et al. 2018). However, variability in gene expression caused by environmental influence can be strong and make analysing changes due to a single gene difficult, as was found for the powdery mildew resistance allele *Pm3b* (Quijano et al. 2015). Therefore, researchers will need to assess the relative benefits of the realistic nature of gene expression under field conditions against the potential pitfalls for each experiment.

5.9 Constructing Gene Networks for Hypothesis Generation and Candidate Gene Identification

Although comparisons of gene expression between samples at different timepoints or in different environmental conditions can be informative, applying network approaches to understand gene interactions and pathway-level responses to environmental and developmental changes is a complementary and powerful approach. Networks can integrate a wide range of information from gene expression and co-expression through to protein-level interactions and scientific literature links (Hassani-Pak et al. 2016), but here we will focus on gene networks built mainly from gene expression measurements.

5.9.1 Co-expression Networks

Co-expression networks can be built from thousands of genes using the similarity in their expression patterns across multiple conditions to determine which genes are grouped (Fig. 5.3a). Based on "guilt-by-association" genes that belong to the same co-expression group are often considered to be co-regulated, for example by shared transcription factors, and to be part of the same biological process.

An important application of gene co-expression networks is the functional annotation of uncharacterised genes (Serin et al. 2016). The development of a high-quality reference sequence for wheat enabled the generation of detailed co-expression networks focussing on specific wheat tissues (leaf, grain, root and spike) and stress conditions (abiotic and biotic) (Ramírez-González et al. 2018). A comparison of the four tissue-specific networks revealed modules of genes which were uniquely co-expressed in the root including several genes whose orthologs regulate root development in Arabidopsis. The other genes present in these root-specific modules represent novel genes that according to "guilt-by-association" may play roles in root development. Additional studies have used co-expression networks to identify candidate genes involved in meiosis, grain development and flowering time pathways (IWGSC et al. 2018; Alabdullah et al. 2019; Chi et al. 2019).

While these studies showed the potential of co-expression networks to identify candidate genes associated with a biological process of interest, functional validation of newly identified genes was lacking. The value of these predictions has been illustrated in wheat using the disease-related network generated by Ramírez-González et al. (2018). Polturak et al. (2022) revealed that the top pathogen-induced modules contained multiple clusters of physically adjacent genes that correspond to six pathogen-induced biosynthetic pathways. Heterologous expression of these co-expressed genes in *Nicotiana benthamiana* produced flavonoids and terpenes that may play a role in defence signalling or as phytoalexins. This study shows the power of co-expression to assign functions to previously uncharacterised genes.

Several online tools have been developed which allow wheat researchers to identify genes that are co-expressed. WheatOmics allows users to search for genes co-expressed with a gene of interest in either grain or multi-tissue co-expression networks (Ma et al. 2021) and KnetMiner integrates information about co-expression

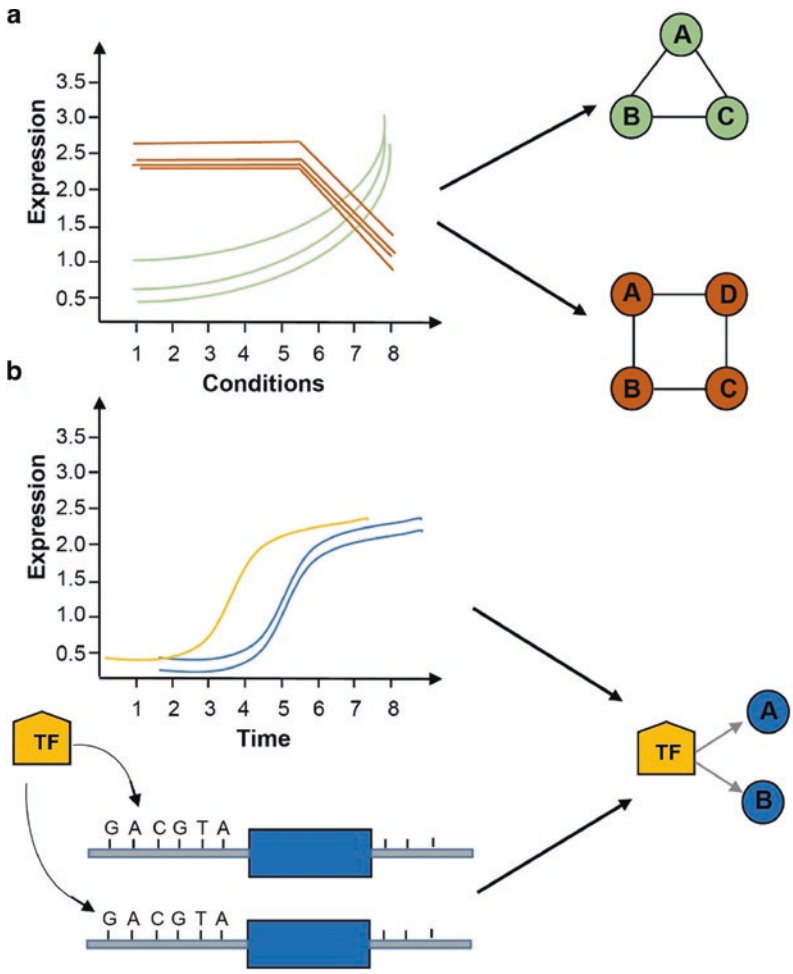


Fig. 5.3 Graphical representation of gene networks. **a** Gene co-expression networks group genes with similar expression patterns across multiple conditions. Interactions between genes (circles) can be direct or indirect. **b** Gene regulatory networks represent direct interactions between genes with directionality. In the example here, a transcription factor (TF; yellow pentagon) is expressed earlier in time and binds to the promoter sites of two downstream genes (blue); the regulatory network on the right shows the directionality of these interactions (arrowheads)

from a network built using 850 wheat RNA-seq samples with a meiosis-specific co-expression, network (IWGSC et al. 2018; Alabdullah et al. 2019; Hassani-Pak et al. 2021). Online tools are also available to construct co-expression networks using custom datasets, such as unpublished RNA-seq data including CoExpNetViz (Tzfadia et al. 2016) and Gene Network Construction Tool Kit (GeNeCK) (Zhang et al. 2019).

5.9.2 Gene Regulatory Networks

In contrast to co-expression networks, the links within gene regulatory networks (GRNs) represent direct gene interactions rather than the association of expression patterns (Fig. 5.3b). GRNs can be built using transcriptome data alone, or they can incorporate additional data types for transcription factor-DNA interactions which inform the network structure (reviewed in

Ko and Brandizzi 2020). GRNs typically have a scale-free network architecture with a few hub genes with multiple connections to other genes and many poorly connected nodes (Barabasi and Oltvai 2004). The hub genes act as master regulators of a GRN and play important roles in biological systems and therefore identifying and manipulating hub genes may enable the manipulation of a biological process of interest.

GRNs in wheat have been used to generate hypotheses about gene function and to identify hub genes which have a strong influence on a biological process. A large GRN was built using 850 RNA-seq samples to predict transcription factor-target interactions using the machine learning-based GENIE3 algorithm (Huynh-Thu et al. 2010). To test the validity of the transcription factor targets identified by GENIE3, Harrington et al. (2020) compared the target genes of the senescence-regulating transcription factor *NAM-A1* to genes differentially expressed in *nam-a1* mutant lines compared to wild type. The *NAM-A1* target genes predicted by GENIE3 overlapped considerably with the differentially expressed genes in lines with reduced *NAM-A1* expression, indicating that GENIE3 can provide biologically relevant predictions. Furthermore, additional senescence-associated transcription factors were identified by combining GENIE3 target information with independent senescence-related expression data. Similarly, combining the GENIE3 network with co-expression networks enabled the identification of candidate genes involved in root development and stress responses (Ramírez-González et al. 2018).

While the GENIE3 approach relies upon diverse RNA-seq samples from different tissues and conditions, GRNs have also proved valuable to understand developmental timeseries in wheat. A ten-timepoint time course of flag leaf senescence was sampled and the resulting RNA-seq data was used to construct a GRN using the time-aware causal structure inference algorithm (Penfold and Wild 2011; Borrill et al. 2019). Filtering the GRN for highly connected and central hub genes identified known senescence regulator *NAM-A1* amongst the 36 top-ranked genes, indicating that this approach identified biologically relevant genes. Functional validation of *NAM-A2*, another top-ranked gene and an uncharacterised paralog of *NAM-A1*, showed the power of this approach to identify genes regulating senescence.

5.9.3 Limitations of Gene Networks

The first attempts to use gene networks in wheat have focussed on hypothesis generation and identifying candidate genes involved in a biological process of interest. While useful insights have been gained, there is still more work to be done to fully leverage the power of gene networks. To date, most gene networks in wheat have been built using gene expression data, although some other types of information are incorporated into tools such as Knetminer and inetbio (Lee et al. 2017; Hassani-Pak et al. 2021). In other species, the accuracy of networks has been improved by incorporating additional data sources such as transcription factor binding sites, open chromatin regions and protein–protein interactions (reviewed in Haque et al. 2019; Ko and Brandizzi 2020). In wheat, these types of data are becoming available, for example with the publication of accessible chromatin regions identified by ATAC-seq (Concia et al. 2020) and this information could be incorporated into future networks to improve the predictive ability.

A second challenge is the validation of gene networks in wheat. In model systems comparison to "gold standard" networks allows the accuracy of different network construction methods to be determined (Marbach et al. 2012). However, in wheat, we know little about the true topology of gene networks so validation using this approach is not possible. Instead, network predictions can be validated on an individual gene basis by examining mutant or gene-edited lines for predicted phenotypic effects (Borrill et al. 2019). Alternatively, gene interactions in the network could be tested using molecular biology approaches. Another promising

approach is to integrate several different network construction approaches which can boost the breadth and accuracy of gene interactions in biological networks (Marbach et al. 2012).

A final issue which affects wheat gene networks is that having a large polyploid genome with > 110,000 genes presents practical challenges for some GRN construction techniques. Although co-expression can be carried out on thousands of genes simultaneously (e.g. IWGSC et al. 2018; Ramírez-González et al. 2018), some widely used GRN approaches only permit tens to hundreds of genes due to computational constraints. One method to circumvent this limitation is to filter genes likely to be of interest before entering them into the GRN to reduce the number of genes (e.g. Borrill et al. 2019). Alternatively, some algorithms such as GENIE3 can use tens of thousands of genes as input, although the computational steps take several weeks on a high-performance computing cluster, therefore this approach will not be accessible to all.

5.10 Conclusions and Future Outlook

The use of transcriptomics has greatly increased in wheat over the past few years, benefitting from a high-quality genome annotation and decreasing sequencing costs. Accurate gene models now simplify the analysis of transcriptomic data and increase the value of the biological information gained. While traditional studies have focussed on understanding changes in gene expression in response to environmental stresses or developmental changes, there are an increasingly varied applications of RNA-seq from identifying candidate genes by surveying genetically diverse populations through to building gene regulatory networks for hypothesis generation. Rapid developments in technologies for transcriptomics will enable us to deepen our understanding of wheat biology for example uncovering high-resolution gene expression patterns.

References

Alabdullah AK, Borrill P, Martin AC, Ramirez-Gonzalez RH, Hassani-Pak K, Uauy C, Shaw P, Moore G (2019) A co-expression network in hexaploid wheat reveals mostly balanced expression and lack of significant gene loss of homeologous meiotic genes upon polyploidization. Front Plant Sci 1325

Alonge M, Shumate A, Puiu D, Zimin AV, Salzberg SL (2020) Chromosome-scale assembly of the bread wheat genome reveals thousands of additional gene copies. Genetics 216:599–608

Backhaus AE, Lister A, Tomkins M, Adamski NM, Simmonds J, Macaulay I, Morris RJ, Haerty W, Uauy C (2022) High expression of the MADS-box gene *VRT2* increases the number of rudimentary basal spikelets in wheat. Plant Physiol 189: 1536–1552

Barabasi A-L, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5:101–113

Borrill P, Ramirez-Gonzalez R, Uauy C (2016) expVIP: a customizable RNA-seq data analysis and visualization platform. Plant Physiol 170:2172–2186

Borrill P, Harrington SA, Simmonds J, Uauy C (2019) Identification of transcription factors regulating senescence in wheat through gene regulatory network modelling. Plant Physiol 180:1740–1755

Brenchley R, Spannagl M, Pfeifer M, Barker GLA, D'Amore R, Allen AM, McKenzie N, Kramer M, Kerhornou A, Bolser D, Kay S, Waite D, Trick M, Bancroft I, Gu Y, Huo N, Luo M-C, Sehgal S, Gill B, Kianian S, Anderson O, Kersey P, Dvorak J, McCombie WR, Hall A, Mayer KFX, Edwards KJ, Bevan MW, Hall N (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. Nature 491:705–710

Brinton J, Simmonds J, Uauy C (2018) Ubiquitin-related genes are differentially expressed in isogenic lines contrasting for pericarp cell size and grain weight in hexaploid wheat. BMC Plant Biol 18:22–22

Chao S, Lazo GR, You F, Crossman CC, Hummel DD, Lui N, Laudencia-Chingcuanco D, Anderson JA, Close TJ, Dubcovsky J, Gill BS, Gill KS, Gustafson JP, Kianian SF, Lapitan NLV, Nguyen HT, Sorrells ME, McGuire PE, Qualset CO, Anderson OD (2006) Use of a large-scale Triticeae expressed sequence tag resource to reveal gene expression profiles in hexaploid wheat (*Triticum aestivum* L.). Genome 49:531–544

Chen G, Ning B, Shi T (2019) Single-cell RNA-Seq technologies and related computational data analysis. Front Genet 10

Chi Q, Guo L, Ma M, Zhang L, Mao H, Wu B, Liu X, Ramirez-Gonzalez RH, Uauy C, Appels R, Zhao H (2019) Global transcriptome analysis uncovers the gene co-expression regulation network and key genes involved in grain development of wheat (*Triticum aestivum* L.). Funct Integr Genomics 19:853–866

Clavijo BJ, Venturini L, Schudoma C, Accinelli GG, Kaithakottil G, Wright J, Borrill P, Kettleborough G, Heavens D, Chapman H (2017) An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. Genome Res 27:885–896

Concia L, Veluchamy A, Ramirez-Prado JS, Martin-Ramirez A, Huang Y, Perez M, Domenichini S, Rodriguez Granados NY, Kim S, Blein T, Duncan S, Pichot C, Manza-Mianza D, Juery C, Paux E, Moore G, Hirt H, Bergounioux C, Crespi M, Mahfouz MM, Bendahmane A, Liu C, Hall A, Raynaud C, Latrasse D, Benhamed M (2020) Wheat chromatin architecture is organized in genome territories and transcription factories. Genome Biol 21:104

Corredor-Moreno P, Minter F, Davey PE, Wegel E, Kular B, Brett P, Lewis CM, Morgan YML, Macías Pérez LA, Korolev AV, Hill L, Saunders DGO (2021) The branched-chain amino acid aminotransferase TaBCAT1 modulates amino acid metabolism and positively regulates wheat rust susceptibility. Plant Cell 33:1728–1747

Coulter M, Entizne JC, Guo W, Bayer M, Wonneberger R, Milne L, Schreiber M, Haaning A, Muehlbauer G, McCallum N, Fuller J, Simpson C, Stein N, Brown JWS, Waugh R, Zhang R (2021) BaRTv2: a highly resolved barley reference transcriptome for accurate transcript-specific RNA-seq quantification. bioRxiv: 2021.2009.2010.459729

Dobon A, Bunting DCE, Cabrera-Quio LE, Uauy C, Saunders DGO (2016) The host-pathogen interaction between wheat and yellow rust induces temporally coordinated waves of gene expression. BMC Genomics 17:380

Feng N, Song G, Guan J, Chen K, Jia M, Huang D, Wu J, Zhang L, Kong X, Geng S, Liu J, Li A, Mao L (2017) Transcriptome profiling of wheat inflorescence development from spikelet initiation to floral patterning identified stage-specific regulatory genes. Plant Physiol 174:1779–1794

Gálvez S, Mérida-García R, Camino C, Borrill P, Abrouk M, Ramírez-González RH, Biyiklioglu S, Amil-Ruiz F, Dorado G, Budak H, Gonzalez-Dugo V, Zarco-Tejada PJ, Appels R, Uauy C, Hernandez P, The I (2019) Hotspots in the genomic architecture of field drought responses in wheat as breeding targets. Funct Integr Genomics 19:295–309

Geng S, Li A, Tang L, Yin L, Wu L, Lei C, Guo X, Zhang X, Jiang G, Zhai W, Wei Y, Zheng Y, Lan X, Mao L (2013) TaCPK2-A, a calcium-dependent protein kinase gene that is required for wheat powdery mildew resistance enhances bacterial blight resistance in transgenic rice. J Exp Bot 64:3125–3136

Haque S, Ahmad JS, Clark NM, Williams CM, Sozzani R (2019) Computational prediction of gene regulatory networks in plant growth and development. Curr Opin Plant Biol 47:96–105

Harrington SA, Backhaus AE, Singh A, Hassani-Pak K, Uauy C (2020) The wheat GENIE3 network provides biologically-relevant information in polyploid wheat. G3: Genes Genomes, Genetics 10:3675–3686

Hassani-Pak K, Castellote M, Esch M, Hindle M, Lysenko A, Taubert J, Rawlings C (2016) Developing integrated crop knowledge networks to advance candidate gene discovery. Appl Transl Genomics 11:18–26

Hassani-Pak K, Singh A, Brandizi M, Hearnshaw J, Parsons JD, Amberkar S, Phillips AL, Doonan JH, Rawlings C (2021) KnetMiner: a comprehensive approach for supporting evidence-based gene discovery and complex trait analysis across species. Plant Biotechnol J 19:1670–1678

He F, Wang W, Rutter WB, Jordan KW, Ren J, Taagen E, DeWitt N, Sehgal D, Sukumaran S, Dreisigacker S, Reynolds M, Halder J, Sehgal SK, Liu S, Chen J, Fritz A, Cook J, Brown-Guedira G, Pumphrey M, Carter A, Sorrells M, Dubcovsky J, Hayden MJ, Akhunova A, Morrell PL, Szabo L, Rouse M, Akhunov E (2022) Genomic variants affecting homoeologous gene expression dosage contribute to agronomic trait variation in allopolyploid wheat. Nat Commun 13:826

Hrdlickova R, Toloue M, Tian B (2017) RNA-Seq methods for transcriptome analysis. Wires RNA 8:e1364

Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P (2010) Inferring regulatory networks from expression data using tree-based methods. PLoS ONE 5:e12776

IWGSC TIWGSC, Mayer Klaus FX, Rogers J, Doležel J, Pozniak C, Eversole K, Feuillet C, Gill B, Friebe B, Lukaszewski Adam J, Sourdille P, Endo Takashi R, Kubaláková M, Číhalíková J, Dubská Z, Vrána J, Šperková R, Šimková H, Febrer M, Clissold L, McLay K, Singh K, Chhuneja P, Singh Nagendra K, Khurana J, Akhunov E, Choulet F, Alberti A, Barbe V, Wincker P, Kanamori H, Kobayashi F, Itoh T, Matsumoto T, Sakai H, Tanaka T, Wu J, Ogihara Y, Handa H, Maclachlan PR, Sharpe A, Klassen D, Edwards D, Batley J, Olsen O-A, Sandve Simen R, Lien S, Steuernagel B, Wulff B, Caccamo M, Ayling S, Ramirez-Gonzalez Ricardo H, Clavijo Bernardo J, Wright J, Pfeifer M, Spannagl M, Martis Mihaela M, Mascher M, Chapman J, Poland Jesse A, Scholz U, Barry K, Waugh R, Rokhsar Daniel S, Muehlbauer Gary J, Stein N, Gundlach H, Zytnicki M, Jamilloux V, Quesneville H, Wicker T, Faccioli P, Colaiacovo M, Stanca Antonio M, Budak H, Cattivelli L, Glover N, Pingault L, Paux E, Sharma S, Appels R, Bellgard M, Chapman B, Nussbaumer T, Bader Kai C, Rimbert H, Wang S, Knox R, Kilian A, Alaux M, Alfama F, Couderc L, Guilhot N, Viseux C, Loaec M, Keller B, Praud S (2014) A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. Science 345:1251788

IWGSC, Appels R, Eversole K, Stein N, Feuillet C, Keller B, Rogers J, Pozniak CJ, Choulet F,

Distelfeld A, Poland J, Ronen G, Sharpe AG, Barad O, Baruch K, Keeble-Gagnère G, Mascher M, Ben-Zvi G, Josselin A-A, Himmelbach A, Balfourier F, Gutierrez-Gonzalez J, Hayden M, Koh C, Muehlbauer G, Pasam RK, Paux E, Rigault P, Tibbits J, Tiwari V, Spannagl M, Lang D, Gundlach H, Haberer G, Mayer KFX, Ormanbekova D, Prade V, Šimková H, Wicker T, Swarbreck D, Rimbert H, Felder M, Guilhot N, Kaithakottil G, Keilwagen J, Leroy P, Lux T, Twardziok S, Venturini L, Juhász A, Abrouk M, Fischer I, Uauy C, Borrill P, Ramirez-Gonzalez RH, Arnaud D, Chalabi S, Chalhoub B, Cory A, Datla R, Davey MW, Jacobs J, Robinson SJ, Steuernagel B, Ex Fv, Wulff BBH, Benhamed M, Bendahmane A, Concia L, Latrasse D, Bartoš J, Bellec A, Berges H, Doležel J, Frenkel Z, Gill B, Korol A, Letellier T, Olsen O-A, Singh K, Valárik M, Vossen Evd, Vautrin S, Weining S, Fahima T, Glikson V, Raats D, Číhalíková J, Toegelová H, Vrána J, Sourdille P, Darrier B, Barabaschi D, Cattivelli L, Hernandez P, Galvez S, Budak H, Jones JDG, Witek K, Yu G, Small I, Melonek J, Zhou R, Belova T, Kanyuka K, King R, Nilsen K, Walkowiak S, Cuthbert R, Knox R, Wiebe K, Xiang D, Rohde A, Golds T, Čížková J, Akpinar BA, Biyiklioglu S, Gao L, N'Daiye A, Kubaláková M, Šafář J, Alfama F, Adam-Blondon A-F, Flores R, Guerche C, Loaec M, Quesneville H, Condie J, Ens J, Maclachlan R, Tan Y, Alberti A, Aury J-M, Barbe V, Couloux A, Cruaud C, Labadie K, Mangenot S, Wincker P, Kaur G, Luo M, Sehgal S, Chhuneja P, Gupta OP, Jindal S, Kaur P, Malik P, Sharma P, Yadav B, Singh NK, Khurana JP, Chaudhary C, Khurana P, Kumar V, Mahato A, Mathur S, Sevanthi A, Sharma N, Tomar RS, Holušová K, Plíhal O, Clark MD, Heavens D, Kettleborough G, Wright J, Balcárková B, Hu Y, Salina E, Ravin N, Skryabin K, Beletsky A, Kadnikov V, Mardanov A, Nesterov M, Rakitin A, Sergeeva E, Handa H, Kanamori H, Katagiri S, Kobayashi F, Nasuda S, Tanaka T, Wu J, Cattonaro F, Jiumeng M, Kugler K, Pfeifer M, Sandve S, Xun X, Zhan B, Batley J, Bayer PE, Edwards D, Hayashi S, Tulpová Z, Visendi P, Cui L, Du X, Feng K, Nie X, Tong W, Wang L (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science 361:eaar7191. https://doi.org/10.1126/science.aar7191

Kawaura K, Mochida K, Enju A, Totoki Y, Toyoda A, Sakaki Y, Kai C, Kawai J, Hayashizaki Y, Seki M, Shinozaki K, Ogihara Y (2009) Assessment of adaptive evolution between wheat and rice as deduced from full-length common wheat cDNA sequence data and expression patterns. BMC Genomics 10:271–271

Kippes N, VanGessel C, Hamilton J, Akpinar A, Budak H, Dubcovsky J, Pearce S (2020) Effect of phyB and phyC loss-of-function mutations on the wheat transcriptome under short and long day photoperiods. BMC Plant Biol 20:297

Ko DK, Brandizzi F (2020) Network-based approaches for understanding gene regulation and function in plants. Plant J 104:302–317

Kuo TCY, Hatakeyama M, Tameshige T, Shimizu KK, Sese J (2020) Homeolog expression quantification methods for allopolyploids. Brief Bioinform 21:395–407

Lee T, Hwang S, Kim CY, Shim H, Kim H, Ronald PC, Marcotte EM, Lee I (2017) WheatNet: a genome-scale functional network for hexaploid bread wheat, *Triticum aestivum*. Mol Plant 10:1133–1136

Li Q, Byrns B, Badawi MA, Diallo AB, Danyluk J, Sarhan F, Laudencia-Chingcuanco D, Zou J, Fowler DB (2018) Transcriptomic insights into phenological development and cold tolerance of wheat grown in the field. Plant Physiol 176:2376–2394

Liu Z, Xin M, Qin J, Peng H, Ni Z, Yao Y, Sun Q (2015) Temporal transcriptome profiling reveals expression partitioning of homeologous genes contributing to heat and drought acclimation in wheat (*Triticum aestivum* L.). BMC Plant Biol 15:1–20

Ma S, Wang M, Wu J, Guo W, Chen Y, Li G, Wang Y, Shi W, Xia G, Fu D (2021) WheatOmics: a platform combining multiple omics data to accelerate functional genomics studies in wheat. Mol Plant 14:1965–1968

Ma J, Tang X, Sun B, Wei J, Ma L, Yuan M, Zhang D, Shao Y, Li C, Chen K-M, Jiang L (2022) A NAC transcription factor, TaNAC5D-2, acts as a positive regulator of drought tolerance through regulating water loss in wheat (*Triticum aestivum* L.). Environ Exp Bot 196:104805

Manickavelu A, Kawaura K, Oishi K, Shin-I T, Kohara Y, Yahiaoui N, Keller B, Abe R, Suzuki A, Nagayama T, Yano K, Ogihara Y (2012) Comprehensive functional analyses of expressed sequence tags in common wheat (*Triticum aestivum*). DNA Res 19:165–177

Mao H, Li S, Wang Z, Cheng X, Li F, Mei F, Chen N, Kang Z (2020) Regulatory changes in TaSNAC8-6A are associated with drought tolerance in wheat seedlings. Plant Biotechnol J 18:1078–1092

Mao H, Li S, Chen B, Jian C, Mei F, Zhang Y, Li F, Chen N, Li T, Du L, Ding L, Wang Z, Cheng X, Wang X, Kang Z (2022) Variation in cis-regulation of a NAC transcription factor contributes to drought tolerance in wheat. Mol Plant 15:276–292

Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, Aderhold A, Allison KR, Bonneau R, Camacho DM, Chen Y, Collins JJ, Cordero F, Costello JC, Crane M, Dondelinger F, Drton M, Esposito R, Foygel R, de la Fuente A, Gertheiss J, Geurts P, Greenfield A, Grzegorczyk M, Haury A-C, Holmes B, Hothorn T, Husmeier D, Huynh-Thu VA, Irrthum A, Kellis M, Karlebach G, Küffner R, Lèbre S, De Leo V, Madar A, Mani S, Marbach D, Mordelet F, Ostrer H, Ouyang Z, Pandya R, Petri T, Pinna A, Poultney CS, Prill RJ, Rezny S, Ruskin HJ, Saeys Y, Shamir R, Sîrbu A, Song M,

Soranzo N, Statnikov A, Stolovitzky G, Vega N, Vera-Licona P, Vert J-P, Visconti A, Wang H, Wehenkel L, Windhager L, Zhang Y, Zimmer R, Kellis M, Collins JJ, Stolovitzky G, The DC (2012) Wisdom of crowds for robust gene network inference. Nat Methods 9:796–804

Martín AC, Borrill P, Higgins J, Alabdullah A, Ramírez-González RH, Swarbreck D, Uauy C, Shaw P, Moore G (2018) Genome-wide transcription during early wheat meiosis is independent of synapsis, ploidy level, and the Ph1 locus. Front Plant Sci 9

Mochida K, Kawaura K, Shimosaka E, Kawakami N, Shin IT, Kohara Y, Yamazaki Y, Ogihara Y (2006) Tissue expression map of a large number of expressed sequence tags and its application to in silico screening of stress response genes in common wheat. Mol Genet Genomics 276:304–312

Mochida K, Yoshida T, Sakurai T, Ogihara Y, Shinozaki K (2009) TriFLDB: a database of clustered full-length coding sequences from Triticeae with applications to comparative grass genomics. Plant Physiol 150:1135–1146

Montenegro JD, Golicz AA, Bayer PE, Hurgobin B, Lee H, Chan C-KK, Visendi P, Lai K, Doležel J, Batley J, Edwards D (2017) The pangenome of hexaploid bread wheat. Plant J 90:1007–1013

Ogihara Y, Mochida K, Nemoto Y, Murai K, Yamazaki Y, Shin-I T, Kohara Y (2003) Correlated clustering and virtual display of gene expression patterns in the wheat life cycle by large-scale statistical analyses of expressed sequence tags. Plant J 33:1001–1011

Pearce S, Kippes N, Chen A, Debernardi JM, Dubcovsky J (2016) RNA-seq studies using wheat PHYTOCHROME B and PHYTOCHROME C mutants reveal shared and specific functions in the regulation of flowering and shade-avoidance pathways. BMC Plant Biol 16:141

Penfold CA, Wild DL (2011) How to infer gene networks from expression profiles, revisited. Interface Focus 1:857–870

Pfeifer M, Kugler Karl G, Sandve Simen R, Zhan B, Rudi H, Hvidsten Torgeir R, Mayer Klaus FX, Olsen O-A (2014) Genome interplay in the grain transcriptome of hexaploid bread wheat. Science 345:1250091

Polturak G, Dippe M, Stephenson Michael J, Chandra Misra R, Owen C, Ramirez-Gonzalez Ricardo H, Haidoulis John F, Schoonbeek H-J, Chartrain L, Borrill P, Nelson David R, Brown James KM, Nicholson P, Uauy C, Osbourn A (2022) Pathogen-induced biosynthetic pathways encode defense-related molecules in bread wheat. Proc Natl Acad Sci 119:e2123299119

Poretti M, Sotiropoulos AG, Graf J, Jung E, Bourras S, Krattinger SG, Wicker T (2021) Comparative transcriptome analysis of wheat lines in the field reveals multiple essential biochemical pathways suppressed by obligate pathogens. Front Plant Sci 12

Quijano CD, Brunner S, Keller B, Gruissem W, Sautter C (2015) The environment exerts a greater influence than the transgene on the transcriptome of field-grown wheat expressing the Pm3b allele. Transgenic Res 24:87–97

Ramírez-González RH, Borrill P, Lang D, Harrington SA, Brinton J, Venturini L, Davey M, Jacobs J, van Ex F, Pasha A, Khedikar Y, Robinson SJ, Cory AT, Florio T, Concia L, Juery C, Schoonbeek H, Steuernagel B, Xiang D, Ridout CJ, Chalhoub B, Mayer KFX, Benhamed M, Latrasse D, Bendahmane A, Wulff BBH, Appels R, Tiwari V, Datla R, Choulet F, Pozniak CJ, Provart NJ, Sharpe AG, Paux E, Spannagl M, Bräutigam A, Uauy C, Korol A, Sharpe Andrew G, Juhász A, Rohde A, Bellec A, Distelfeld A, Akpinar Bala A, Keller B, Darrier B, Gill B, Chalhoub B, Steuernagel B, Feuillet C, Chaudhary C, Uauy C, Pozniak C, Ormanbekova D, Xiang D, Latrasse D, Swarbreck D, Barabaschi D, Raats D, Sergeeva E, Salina E, Paux E, Cattonaro F, Choulet F, Kobayashi F, Keeble-Gagnere G, Kaur G, Muehlbauer G, Kettleborough G, Yu G, Šimková H, Gundlach H, Berges H, Rimbert H, Budak H, Handa H, Small I, Bartoš J, Rogers J, Doležel J, Keilwagen J, Poland J, Melonek J, Jacobs J, Wright J, Jones Jonathan DG, Gutierrez-Gonzalez J, Eversole K, Nilsen K, Mayer Klaus FX, Kanyuka K, Singh K, Gao L, Concia L, Venturini L, Cattivelli L, Spannagl M, Mascher M, Hayden M, Abrouk M, Alaux M, Luo M, Valárik M, Benhamed M, Singh Nagendra K, Sharma N, Guilhot N, Ravin N, Stein N, Olsen O-A, Gupta Om P, Khurana P, Chhuneja P, Bayer Philipp E, Borrill P, Leroy P, Rigault P, Sourdille P, Hernandez P, Flores R, Ramirez-Gonzalez Ricardo H, King R, Knox R, Appels R, Zhou R, Walkowiak S, Galvez S, Biyiklioglu S, Nasuda S, Sandve S, Chalabi S, Weining S, Sehgal S, Jindal S, Belova T, Letellier T, Wicker T, Tanaka T, Fahima T, Barbe V, Tiwari V, Kumar V, Tan Y (2018) The transcriptional landscape of polyploid wheat. Science 361:eaar6089. https://doi.org/10.1126/science.aar6089

Serin EA, Nijveen H, Hilhorst HW, Ligterink W (2016) Learning from co-expression networks: possibilities and challenges. Front Plant Sci 7:444

Thibivilliers S, Anderson D, Libault M (2020) Isolation of plant root nuclei for single cell RNA sequencing. Curr Protoc Plant Biol 5:e20120

Tzfadia O, Diels T, De Meyer S, Vandepoele K, Aharoni A, Van de Peer Y (2016) CoExpNetViz: comparative co-expression networks construction and visualization tool. Front Plant Sci 6:1194

Walkowiak S, Gao L, Monat C, Haberer G, Kassa MT, Brinton J, Ramirez-Gonzalez RH, Kolodziej MC, Delorean E, Thambugala D, Klymiuk V, Byrns B, Gundlach H, Bandi V, Siri JN, Nilsen K, Aquino C, Himmelbach A, Copetti D, Ban T, Venturini L, Bevan M, Clavijo B, Koo D-H, Ens J, Wiebe K, N'Diaye A, Fritz AK, Gutwin C, Fiebig A, Fosker C, Fu BX, Accinelli GG, Gardner KA, Fradgley N, Gutierrez-Gonzalez J, Halstead-Nussloch G, Hatakeyama M, Koh CS, Deek J, Costamagna AC, Fobert P, Heavens

D, Kanamori H, Kawaura K, Kobayashi F, Krasileva K, Kuo T, McKenzie N, Murata K, Nabeka Y, Paape T, Padmarasu S, Percival-Alwyn L, Kagale S, Scholz U, Sese J, Juliana P, Singh R, Shimizu-Inatsugi R, Swarbreck D, Cockram J, Budak H, Tameshige T, Tanaka T, Tsuji H, Wright J, Wu J, Steuernagel B, Small I, Cloutier S, Keeble-Gagnère G, Muehlbauer G, Tibbets J, Nasuda S, Melonek J, Hucl PJ, Sharpe AG, Clark M, Legg E, Bharti A, Langridge P, Hall A, Uauy C, Mascher M, Krattinger SG, Handa H, Shimizu KK, Distelfeld A, Chalmers K, Keller B, Mayer KFX, Poland J, Stein N, McCartney CA, Spannagl M, Wicker T, Pozniak CJ (2020) Multiple wheat genomes reveal global variation in modern breeding. Nature 588:277–283

Wang K, Wang D, Zheng X, Qin A, Zhou J, Guo B, Chen Y, Wen X, Ye W, Zhou Y, Zhu Y (2019) Multi-strategic RNA-seq analysis reveals a high-resolution transcriptional landscape in cotton. Nat Commun 10:4714

Winter D, Vinegar B, Nahal H, Ammar R, Wilson GV, Provart NJ (2007) An "electronic fluorescent pictograph" browser for exploring and analyzing large-scale biological data sets. PLoS ONE 2:e718

Xiong H, Guo H, Xie Y, Zhao L, Gu J, Zhao S, Li J, Liu L (2017) RNAseq analysis reveals pathways and candidate genes associated with salinity tolerance in a spaceflight-induced wheat mutant. Sci Rep 7:2731–2731

Xu X, Crow M, Rice BR, Li F, Harris B, Liu L, Demesa-Arevalo E, Lu Z, Wang L, Fox N, Wang X, Drenkow J, Luo A, Char SN, Yang B, Sylvester AW, Gingeras TR, Schmitz RJ, Ware D, Lipka AE, Gillis J, Jackson D (2021) Single-cell RNA sequencing of developing maize ears facilitates functional analysis and trait candidate gene discovery. Dev Cell 56:557-568.e556

Yu Y, Zhu D, Ma C, Cao H, Wang Y, Xu Y, Zhang W, Yan Y (2016) Transcriptome analysis reveals key differentially expressed genes involved in wheat grain development. Crop J 4:92–106

Zhang H, Yang Y, Wang C, Liu M, Li H, Fu Y, Wang Y, Nie Y, Liu X, Ji W (2014) Large-scale transcriptome comparison reveals distinct gene activations in wheat responding to stripe rust and powdery mildew. BMC Genomics 15:1–14

Zhang M, Li Q, Yu D, Yao B, Guo W, Xie Y, Xiao G (2019) GeNeCK: a web server for gene network construction and visualization. BMC Bioinform 20:1–7

Zhang T-Q, Chen Y, Liu Y, Lin W-H, Wang J-W (2021) Single-cell transcriptome atlas and chromatin accessibility landscape reveal differentiation trajectories in the rice root. Nat Commun 12:2053

Zhu T, Wang L, Rimbert H, Rodriguez JC, Deal KR, De Oliveira R, Choulet F, Keeble-Gagnère G, Tibbits J, Rogers J, Eversole K, Appels R, Gu YQ, Mascher M, Dvorak J, Luo M-C (2021) Optical maps refine the bread wheat *Triticum aestivum* cv. Chinese Spring genome assembly. Plant J 107:303–314

Zimin AV, Puiu D, Hall R, Kingan S, Clavijo BJ, Salzberg SL (2017) The first near-complete assembly of the hexaploid bread wheat genome, *Triticum aestivum*. GigaScience 6:1–7

6 Genome Sequence-Based Features of Wheat Genetic Diversity

Xueyong Zhang and Rudi Appels

Abstract

Common wheat is a hexaploid species crop that is widely recognized as an important staple food crop. The establishment of a gold standard reference genome sequences of the well-studied CHINESE SPRING, and its progenitors (including *Triticum turgidum* ssp. *dicoccoides* accession *Zavitan*, *Triticum durum* accession Svevo, *Triticum urartu*, *Aegilops tauschii*), in the last 5 years has dramatically promoted our understanding of wheat genome diversity and evolution through the resequencing of collections of wheat and its progenitors. In this chapter, we review progress in the analysis and interpretation of genome-based studies of wheat focusing on geographic genome differentiation, interspecies gene flow, haplotype blocks, and gene diversity in breeding. We also consider approaches for efficiently discovering and integrating the genes and genome variations, hidden in Genebank collections, into wheat breeding programs.

X. Zhang (✉)
Center for Crop Genomics and Molecular Design, Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing, China
e-mail: zhangxueyong@caas.cn

R. Appels
University of Melbourne, Food and Nutrition, Parkville, and AgriBio (Latrobe University), Bundoora, Melbourne, Australia

Keywords

Bread wheat · Diversity · Geographic differentiation · Introgression · Human selection · Haplotype blocks

6.1 Wheat Origin and Spread in the World

Common wheat (*Triticum aestivum* L.) provides approximately 20% of the total calories for human intake globally. The origins of the common hexaploid wheat species were through natural crosses between cultivated emmer (*Triticum turgidum*, AABB) and *Aegilops tauschii* (DD) and is considered to be the first domesticated crop in the "hilly flanks of the Fertile Crescent" in southwestern Asia between 10,000 and 7000 BC (Feldman and Levy 2012). Key advances for the domestication process included the absence of head brittleness and free-threshing grains. The dispersal of wheat selections prior to the 5th millennium BC was extensive as several *Triticum* taxa spread from the Fertile Crescent westwards across central Europe and along the northern coastal line of the Mediterranean (Fig. 6.1). To the East, wheat is documented in archaeological records to be present in Turkmenistan and Pakistan before 5000 BC. It was introduced into west China in 2000 BC and into central and east China in approximately

R. Appels et al. (eds.), *The Wheat Genome*, Compendium of Plant Genomes,
https://doi.org/10.1007/978-3-031-38294-9_6

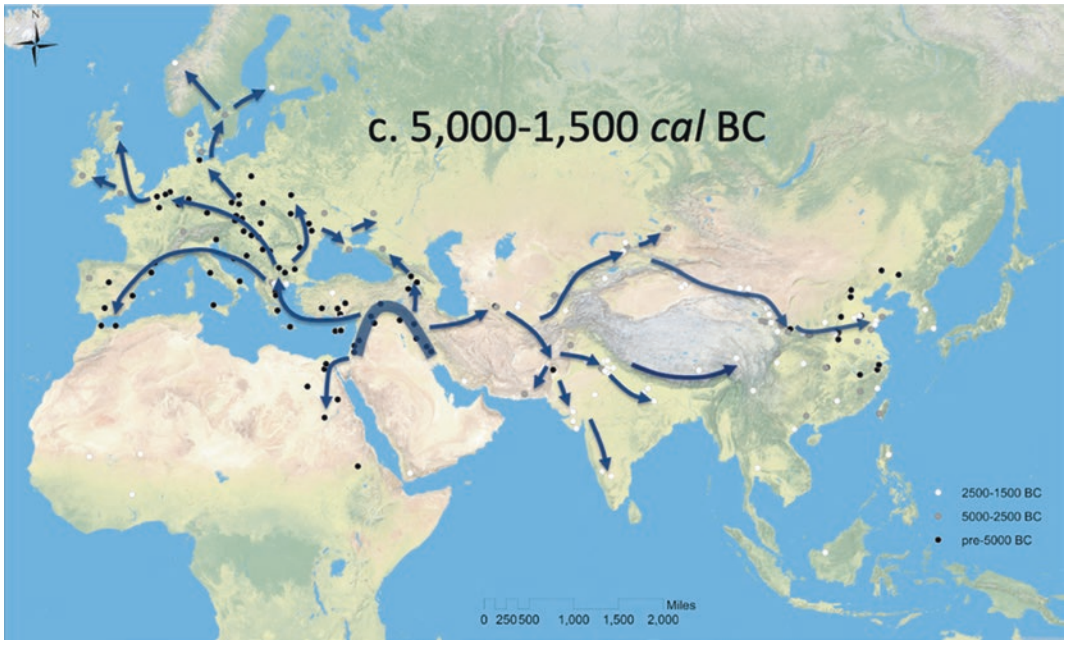


Fig. 6.1 Map showing hypothesized dispersals of domesticated *Triticum* and *Hordeum* taxa (i.e. wheat and barley) originated in southwestern Asia across the Old World dating between 5000 and 1500 cal BC. Black circles: sites older than 5000 BC; gray circles: sites dated between 5000 and 2500 BC; white circles: sites dated between 2500 and 1500 BC; solid line: parsimonious inference from botanical evidence from dated archaeological context (the density of which varies greatly across Eurasia). Map is originally presented in Liu et al. (2019a, b), modified with permission

1500 BC (Liu et al. 2016, 2019a, b), based on archaeological discoveries.

Colonization of wheat in the very new and distinct environments eventually replaced native crops as the staple crop and resulted in field-level selections of traits with very strong geographic characters to meet the local cultivation and consumption of variant human populations. These genetic changes were basically retained in the genome variation between cultivars, especially the landraces. Establishment of the so-called gold standard wheat genome sequence, taken together with the assemblies of the reference genomes of its progenitor species as well as other hexaploid varieties (Avni et al. 2017; Luo et al. 2017; Zhao et al. 2017; Ling et al. 2018; IWGSC 2018; Zhu and Luo 2021), has provided the basis for high-density SNP-chips and resequencing analyses. Advances such as the production of SNP-chips with 90, 285, 660 K SNPs have provided the means for the wide use to elucidate genome diversity. Germplasm exchange and the development of genomics now provide a new opportunity to re-evaluate and reconsider the evolution and dispersal process of wheat from this new point of view. These works also pave the way for associating the allelic variation with phenotypes for physical mapping of variation in the genome (Varshney et al. 2021).

6.2 Global Distribution of Wheat Genome Diversity and the Leading Role of 3B in Geographic Differentiation

The genotyping of 632 world wheat landraces using the 285 K SNP array-markers on chromosome 3B, allowed Paux et al. (2008) to

Fig. 6.2 Haplotypes in landraces on 3B and their global distribution (provided by Dr. Etienne Paux, INRAE). The red, pink, blue, and green dots refer to different haplotypes (see Paux et al. 2008) and the clustering of the different colored haplotypes across the landscape from Europe to China is evident

define the very strong geographic differentiation (Fig. 6.2). In a follow-up diversity analysis of Chinese wheat landraces using a 660 K SNP array, we found they could be basically classified into two sub-groups, the north-China sub-group and middle-south China sub-group (Wang et al. 2021). Among the 21 chromosomes, 3B and 7A were particularly prominent in being associated with the stratified domestication in China, based on the standard *Fst* values for SNP allele frequencies that differentiate populations in two groups, namely *Triticum aestivum*-L1 and *T. aestivum*-L2 (Fig. 6.3).

When the differentiation of populations *T. aestivum*-L1 and *T. aestivum*-L2 were narrowed down to the analysis of the crucial regions of 280–375 Mb on 3B and 211.7–272.9 Mb on 7A in the CS reference 1.0 (IWGSC 2018), the *Fst* reached 0.84 and 0.66, respectively (Fig. 6.3a; quantified in Fig. 6.3b), and were associated with grain size and length in multi-environment BLUP phenotype data (Wang et al. 2021). Accessions in *T. aestivum*-L1 were mainly distributed in northwestern China, whereas those in *T. aestivum*-L2 were mainly from central to eastern China (Fig. 6.4). The most distinct agronomic trait was grain size (TKW), i.e., the *T. aestivum*-L2 accessions usually had smaller grain size than the *T. aestivum*-L1 accessions, which was achieved by reduction in grain length (Wang et al. 2021).

Haplotype analysis in genotyped collections including wild emmer, domesticated emmer, common wheat landraces, and Chinese modern cultivars based on the 660 K SNP genotyped data clearly revealed wild emmer (WE) was the donor for the hap-block in L1 (see Haps 1, 2, 4, and 12 in Fig. 6.4). This is consistent with the suggested intercross and genome introgression between common wheat and wild emmer (He et al. 2019; Cheng et al. 2019).

GWAS based on the multi-year agronomic trait phenotypes revealed strong association of the crucial region on 3B (280–375 Mb) with spike length ($-\log_{10}(p) \geq 5.0$). In Chinese landraces, the northwest haplotype-group (L1) usually has longer spike and larger grains than the southeast haplotype-group (L2). Breeding selection in the seven decades from 1950 to 2020 favored the L1-haplotypes from the wild emmer (Fig. 6.4; Hao et al. 2020; Wang et al. 2021). Therefore, we estimate that this genomic region might also

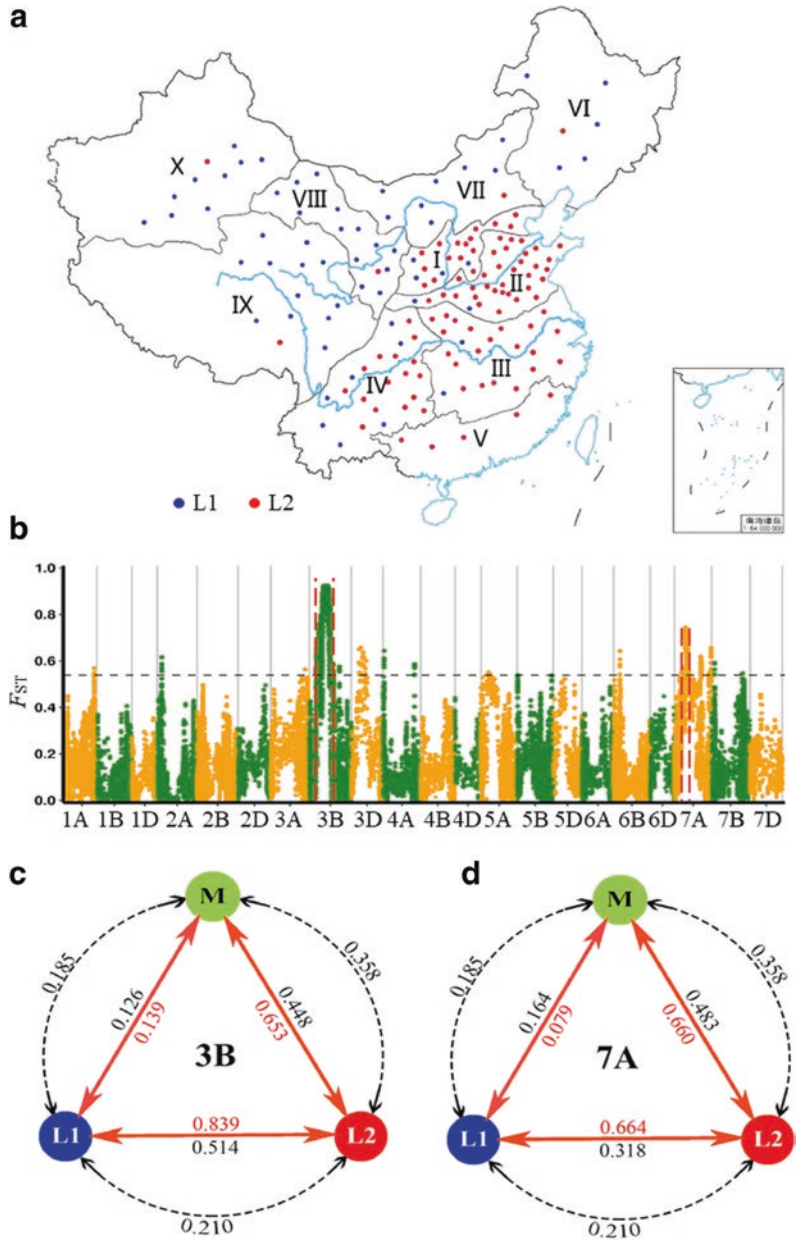


Fig. 6.3 Very strong geographic and genetic differentiation happened in Chinese wheat landraces, forming two subsets, L1 (blue) and L2 (red). The 3B and 7A lead the differentiation among the 21 chromosomes. **a** Quite distinct distribution of collections in L1 and L2. **b** The F_{ST} value between L1and L2 along the 21 chromosomes, which was estimated based on the 660 K SNP markers using CS R 1.0 as reference. **c** and **d**. The triangles/arrows indicate F_{ST} values between L1 (blue) and L2 (red), L1 and modern cultivars (M, green), and L2 and M on 3B and 7A. The red lettering along the arrows focuses on the crucial genomic regions (3B: 280–375 Mb) and (7A: 211.7–272.9 Mb). The data along the dashed circles were F_{ST} values between subsets in whole genome of wheat (adapted from Wang et al. 2021)

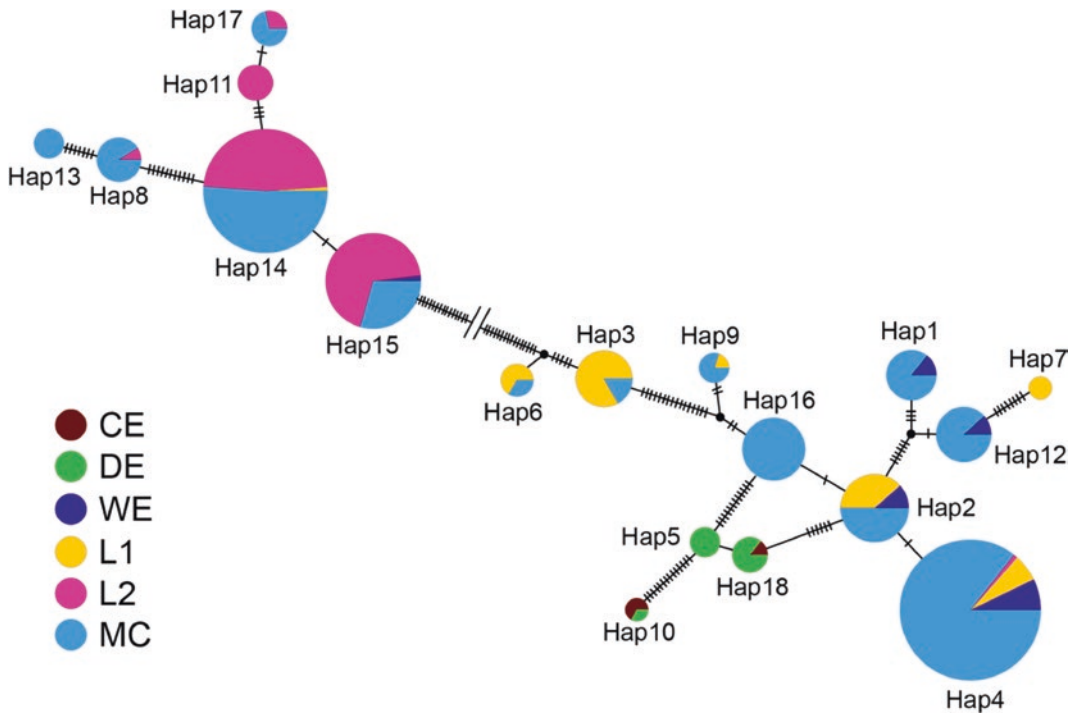


Fig. 6.4 Haplotype network based on SNPs on 3B in cultivated emmer (CE), domesticated emmer (DE), Northeast landrace group 1 (L1), Southeast landrace group 2 (L2), modern Chinese cultivar (MC), and wild emmer (WE). Circle size is proportional to the number of accessions for a given haplotype. The short lines between two nodes indicate the number of mutations

relate to NUE or WUE of cultivars. The great increase of L1-haplotypes (including Haps 1, 2, 4, and 12) in the modern Chinese cultivars (MC) at this genome location also correlated with the cooking style from full grain in history to wheat flour products today, because small grain was favored in full grain cooking, but larger grain was favored in flour-product consumption because of higher yield (Liu et al. 2014, 2016). Based on the analysis in the 10+pangenome, we found large structure variation (SV) existing in this region across the 3B centromere (Fig. 6.5).

6.3 Frequent Gene Flow Between Species and Its Effects on Diversity

In Israel, wild emmer wheat with intermediate phenotypes grew at the boundaries of cultivated areas. These wild plants may have originated from hybridization of wild emmer with *T. turgidum* cultivars. They are indicative of gene flow between wild and domesticated populations (Matsuoka 2011). Dvorak et al. (2006) provided initial molecular evidence for existence of introgressions from wild emmer (*Triticum dicoccoides*) into common wheat, which was indirectly supported by the fact that wild emmer usually existed as an accompanying weed of durum and common wheat in wheat origin/domestication regions. Hexaploid and tetraploid wheats were also cultivated as a mixture in the field in these regions (Matsuoka 2011). The overall consequence was that the mixed cropping provided opportunities for gene flow between species through natural hybridization.

The identity score (IS) is widely used to reveal the parent's genetic contribution to their derived cultivars in breeding. The IS is defined with reference to similar nucleotide sequences present in two, or more than two, individuals

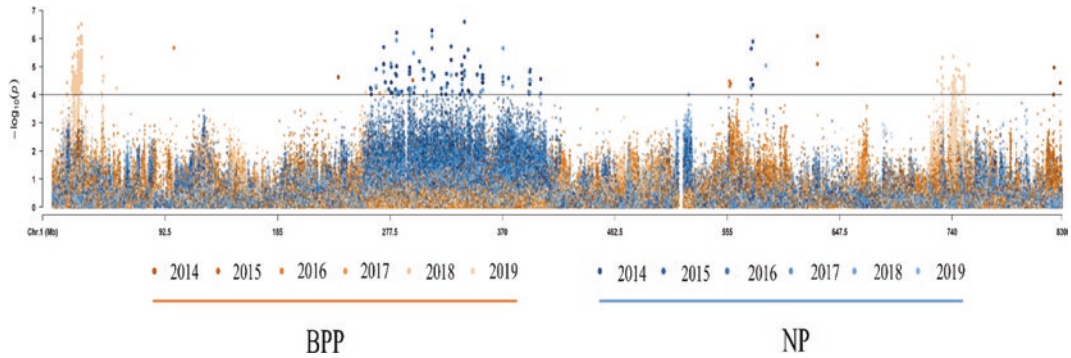


Fig. 6.5 GWAS based on 660 K SNP array with multi-year phenotyping data of landraces (NP) and biparental population (BPP) indicated that high genome differentiation at 280–375 Mb was associated with spike length on 3B chromosome as indicated by the scores for *Fst* exceeding the significance cutoff, across the 280–375 Mb region

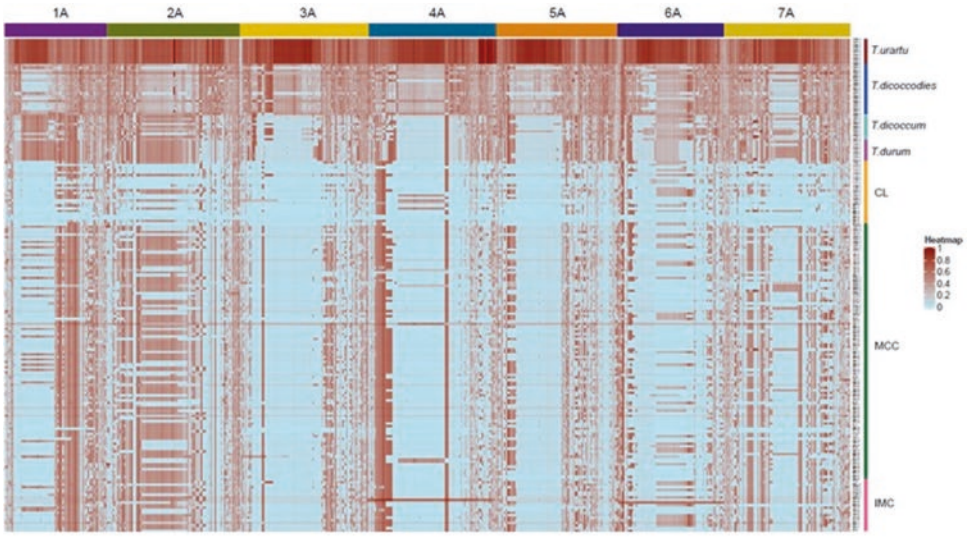
through replication of the same ancestral copy of respective sequences. Our IS graph file analysis based on resequencing analysis in common wheat (landraces and modern cultivars), wild emmer, domesticated emmer and durum wheat, revealed frequent genomic introgressions between common wheat and wild emmer, as well as cultivated emmer, where CS 1.0 was used as reference (light blue, IWGSC 2018). Of course, more frequent introgressions between the tetraploid species were detected as expected, because they shared the common genome AABB (Fig. 6.6a). Independent research by two other groups also revealed the wild existence of introgressions from wild emmer into common wheat (Fig. 6.6b) (Cheng et al. 2019; He et al. 2019). In addition, global wheat diversity research was also strongly promoted by the establishment golden standard reference genomes of common wheat and *T. dicoccoides* and *Triticum durum* (IWGSC 2018; Avni et al. 2017; Maccaferri et al. 2019; Pont et al. 2019; Sansaloni et al. 2020), all of which were sequentially perfected with the integration of more assembles based on 3rd generation sequence reads (Zhu et al. 2021).

Alien introgression usually reduces the recombination frequency and leading to strong LD in natural or breeding population, which results in decline of diversity in the respective genome regions. However, the SNP density was usually increased because suppression of recombination has retained the intact of the introgression fragments, which retain regions rich SNPs in their comparison with CS reference genome (Fig. 6.7). We found that the evenness of recombination rate along the D sub-genome chromosomes is much better than either A- or B- sub-genome chromosomes. This might be caused by the introgressions from wild emmer, which mainly existed within the A- and B-sub-genomes of common wheat (Fig. 6.8). Sufficient genome differentiation should happen between the hexaploid and tetraploid AB genome, which prevents occurrence of recombination between the "introgressions" and original homology fragments. The great difference on recombination rate across the centromeres between 7A and 7B in Chinese modern cultivars directly supported our hypothesis of introgression suppression to recombination, because there is a large introgression detected on 7A across centromeric region (Fig. 6.8, ca 230–430 Mb, Cheng et al. 2019; Hao et al. 2020).

6.4 LD and Haplotype Blocks in Wheat Evolution and Breeding

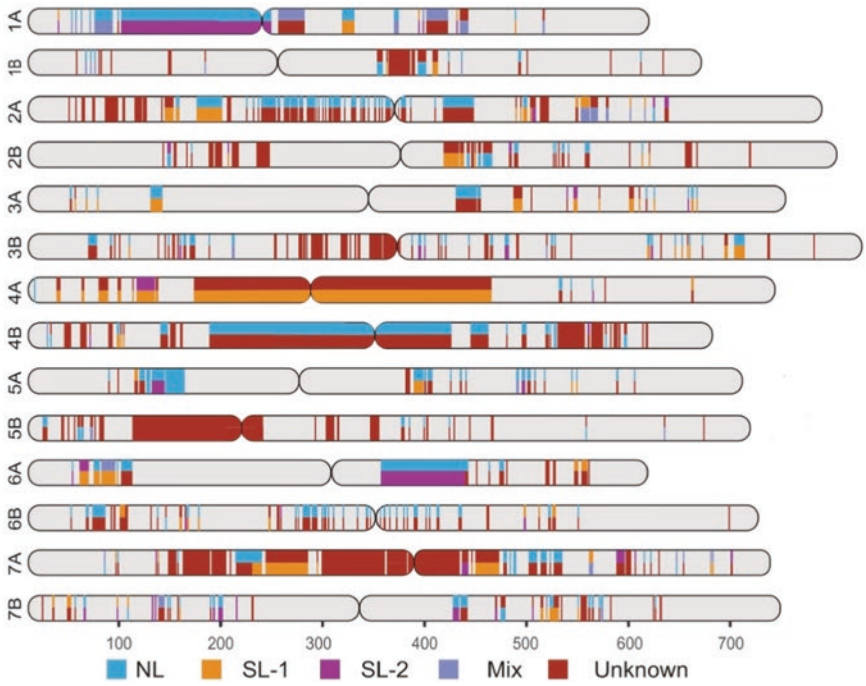
Linkage disequilibrium (LD) is a common phenomenon in the population genetics analysis of crops. For a long time, it was believed that strong positive selection for a gene usually led

a



CL: Chinese landrace, IMC: introduced modern cultivars. MCC: Modern Chinese Cultivars

b



NL: Northern Levant), SL: Southern Levant, Mix: other regions.

Fig. 6.6 Frequent genome introgressions between species in *Triticum* genus revealed by 1-IBD within the A sub-genome chromosomes, where CS 1.0 was used as reference and expressed in light blue. **a** Graph based on (1-IBD) indicated frequent introgressions from wild emmer to domesticated emmer and common wheat. It also revealed reverse introgression from common wheat to domesticated emmer and wild emmer. **b** Genomic introgressions detected in global common wheat on the 14 chromosomes of A and B sub-genomes from four wild emmer populations into common wheat (adapted from Cheng et al. 2019)

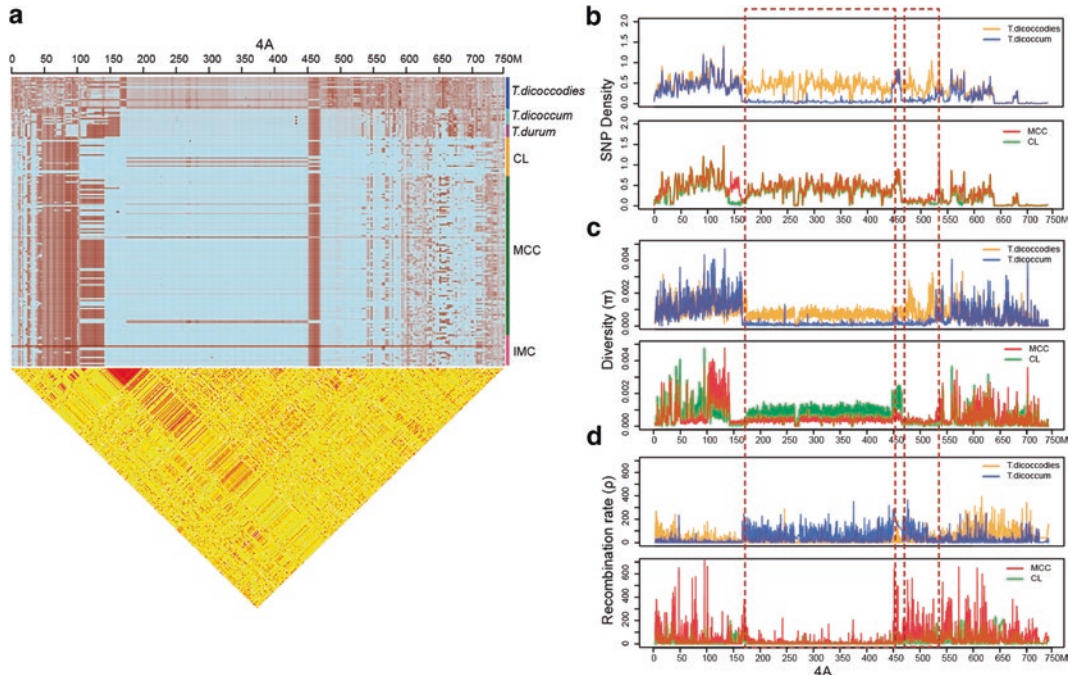


Fig. 6.7 A wild emmer introgression (~ 172–448 Mb) and their effects on SNP density, recombination ratio (ρ), and genome diversity (π) in comparison with the neighbor region without introgression (468–530 Mb) on chromosome 4A. CL: Chinese landrace, IMC: introduced modern cultivars

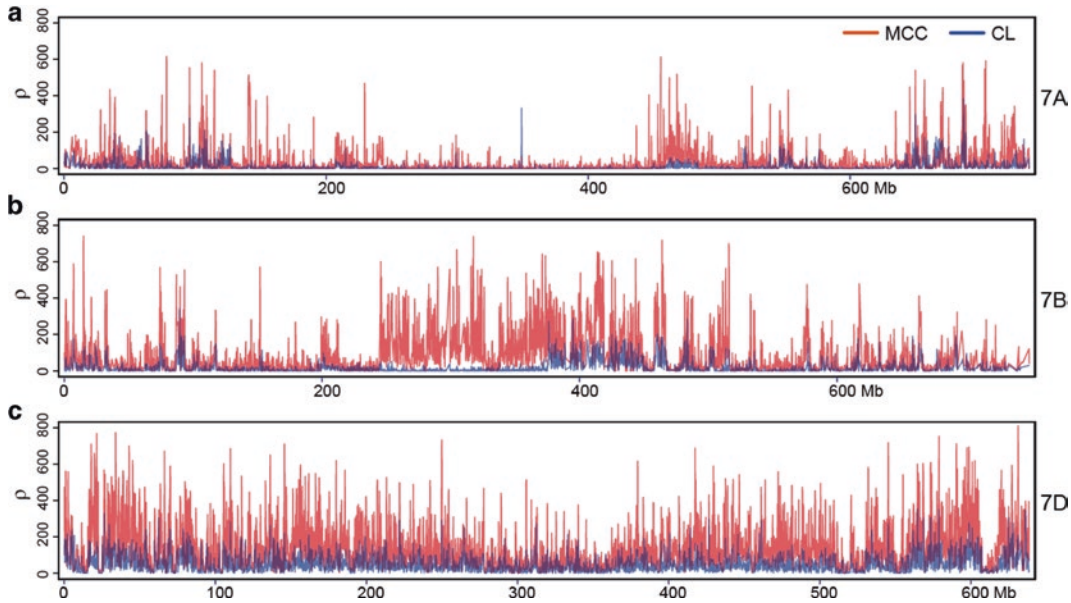


Fig. 6.8 Recombination ratio (ρ) along the three chromosomes of homologous group 7 in Chinese wheat landraces (blue) and modern cultivars (red). There is more recombination disequilibrium on 7A and 7B than on 7D chromosome

to strong LD around the loci because of hitchhiking effect. LD was usually affected by population diversity, selection pressure at the crucial loci, as well as recombination rate. We found that chromosomes in A and B sub-genomes have larger and stronger LD blocks in wheat (Hao et al. 2020). This might be caused by two factors (1) more QTLs controlling agronomic traits on the A and B sub-genomes (Peng et al. 2003, 2011). (2) Partial suppression of recombination along A and B sub-genome chromosomes caused by the introgressions from wild *T. turgidum* species (Figs. 6.6, 6.7 and 6.8).

6.4.1 Haplotype Block Size Along a Chromosome in Wheat

On each of the 21 chromosomes, five chromosomal regions were defined by the IWGSC, based on the overall recombination pattern observed in wheat (IWGSC 2018). There are fewer but very large (> 10 kb, Fig. 6.9a example for chromosome 2B) haplotype blocks at regions across the centromeres, and smaller haplotype blocks at the R1 and R2 regions (Fig. 6.9). The identification of the R1–R3 blocks of chromosome regions in the wheat chromosomes is based on the recombination rate characteristics, gene density, and tissue-specific vs household expression variation across each of the 21 wheat chromosomes. The R1 and R3 designate the distal ends of the short and long chromosomal arms, respectively; R2a and R2b designate the interstitial regions of the short and long arms and the C region and identify the pericentromeric regions (IWGSC 2018).

The box plots in Fig. 6.9 provide a statistical assessment of the R1, R2a, R2b, and R3 designations across the wheat genome based on recombination frequency and indicated that the difference between C and the terminal blocks was significant. Consistent with this significant difference in recombination frequency, Jordan et al. (2020) found that DNS scores assessing DNA accessibility to Micrococcal Nuclease

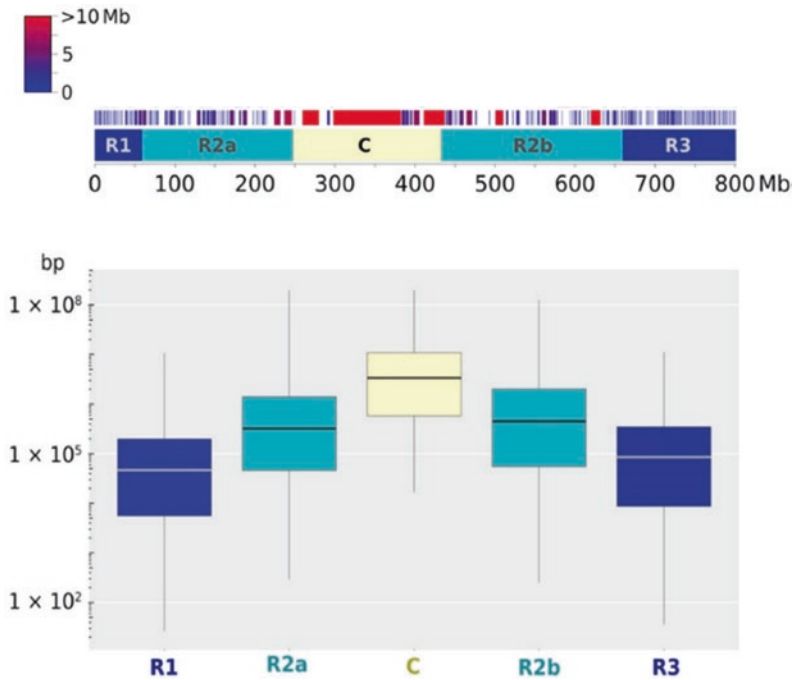


Fig. 6.9 Size difference of haplotypic blocks along wheat chromosomes using 2B as an example (adapted from Balfourier et al. 2019). The designations for the different chromosome regions in the upper panel derive from the overall recombination patterns observed in wheat (IWGSC 2018)

(MNase), and thus the more open or compacted state of the chromatin, were significantly higher (= more open chromatin) for the genome space in the R1 and R3 regions.

6.5 Large Haplotype Blocks and Their Role in Breeding

Identification of haplotype blocks and big PAVs and tracking their variants in evolution and breeding are notable aspects in self-pollination crops in the current genomics era. The investigation of gene-network contributions to the well-studied thousand grain weight (TGW) phenotype that contributes to yield in wheat, for example, provided an unexpected influence of structural variation for the presence/absence of the 5AS chromosome arm (Taagen et al. 2021). A combination of transcriptome data and high-resolution marker maps for the TGW QTL initially thought to be on 5AL, in fact indicated that the QTL resulted from linkage to the presence/absence of the 5AS arm. On a larger scale, the resequencing of 145 land marker cultivars in China, it was found that there were more long-range haplotypes on A and B sub-genomes rather than on D sub-genome in common wheat (Jordan et al. 2015; Hao et al. 2020). The first reason was that the gene flow occurred from the wild tetraploid *T. dicoccoides* during early co-cultivation of tetraploid and hexaploid wheat, where wild emmer was also present as a weed in wheat fields. This was therefore expected to result in a substantial increase in polymorphism on the A and B sub-genomes relative to the D sub-genome in modern bread wheat. This also partially, negatively, affected the homologous recombination occurrence within A and B sub-genomes at crucial genomic regions because of differentiation on the intergenic repeats among wild emmer, cultivated emmer, and common wheat, leading to a very uneven distribution of recombination ratio and SNPs along chromosomes (Fig. 6.8). The second reason is asymmetric distribution of agronomic traits among the three sub-genomes. There are more QTLs or genes controlling domestication and yield traits mapped on the A sub-genome than either on B or D sub-genomes, leading to stronger selection on the A sub-genomes (Peng et al. 2011; Jordan et al. 2015).

Haplotype-based breeding (HBB) can now be proposed following the genome resequencing of larger number of cultivars, because it represents a promising breeding approach for dealing with and identifying, superior haplotypes and their deployment in breeding programs (Varshney et al. 2021). We propose that for self-pollinated crops with a long breeding history, it will be possible to take advantage of hap-block identification to select ideal parent materials to achieve new high-performing cultivars via HBB (Figs. 6.9 and 6.10).

We dissected diversity features along chromosomes 6A (Fig. 6.10a) and 1A (Fig. 6.10b) in cultivar subsets released pre- and post-development of the two hallmark Chinese cultivars AIMENGNIU (AMN) and XIAOYAN 6 (XY6) based on their pedigrees. Fixation of big haplotype blocks from 224 to 442 Mb, on 6A in post-XY6 cultivars were detected but relatively higher diversity was retained in AMN-post cultivars. From 100 to 300 Mb, the haplotype block was fixed in post-AMN cultivars but not in post-XY6 on chromosome 1A. This indicated the haplotype block carried by XY6 on 6A and that carried by AMN on 1A provided sufficiently high-quality attributes for breeders to then retain them. An interesting but less pronounced trend was also found from 178 to 472 Mb on chromosome 2A, but with both XY6- and AMN-derived new cultivars, this genomic region maintained a higher diversity. This indicates haplotypes carried by either XY6 or AMN are not sufficiently high-quality enough for breeders to retain them. The very large sizes of the haplotype blocks also highlight the feasibility of HBB in wheat.

6.6 Human Selection and Gene Diversity

Cloning the gene responsible of a trait or QTL and analyzing its natural variation to find valuable new alleles is one major task for scientists

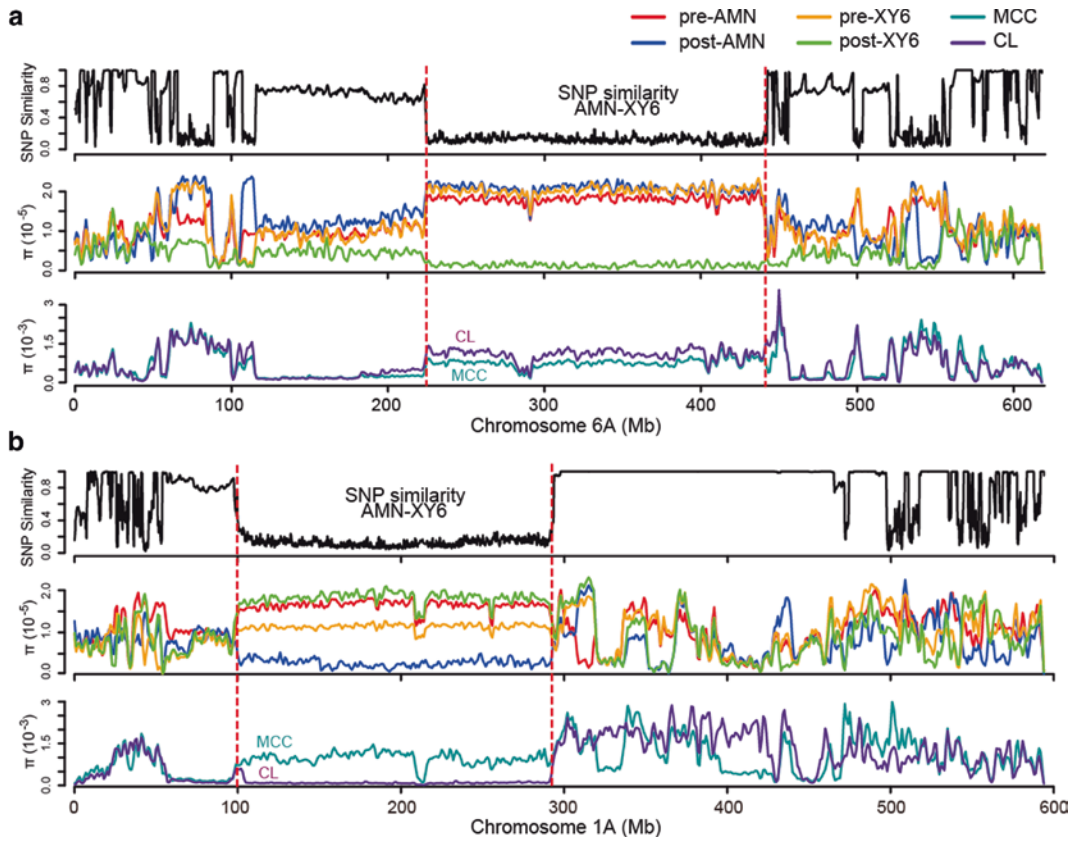


Fig. 6.10 Diversity features at key haplotype blocks considering cultivars subsets released pre- and post-development of the two hallmark Chinese cultivars AIMENGNIU (AMN) and XIAOYAN 6 (XY6) as well as within Chinese landraces (CL) and modern Chinese cultivars (MCC). SNP similarity between AMN and XY6, population diversity for each sub-set (pre-AMN vs. post-AMN, pre-XY6 vs. post-XY6, MCC vs. CL) on chromosomes 6A (**a**), 1A (**b**). Adapted from Hao et al. (2020)

working in crop genetic resources. Fine mapping of QTL through advanced backcross QTL analysis was regarded as the best reliable method for a long time (Tanksley and Nelson 1996). However, QTL mapping-based cloning of genes in wheat is very time-consuming because of the complexity of genome and polyploid nature. The gold reference genome sequence promotes gene cloning via GWAS under the assistance of gene editing and transformation in wheat. The successful mapping of *Rht24* through GWAS in large collections using the CHINESE SPRING genome as reference for 6A is a landmark indicator for gene mapping strategy that complements the QTL fine mapping in biparental recombination population to GWAS-fine mapping in natural population. Through anchoring the flanking markers on the RefSec v1.0, the candidate gene of *Rht24* was narrowed down to 50 Mb region between 400 and 450 Mb on 6A chromosome, which was actively selected in breeding since 1990s (Würschum et al. 2017).

It is very hard to precisely map genes at pericentromeric region through recombination in biparental populations. But through GWAS, we can use long historic recombinations to carry out mapping and dissection of the crucial region. For example, we found a grain thickness-associated locus on the long arm of the chromosome 5A marked by the peak SNP chr5A_430246395 ($-\log_{10}(p)=4.17$) because the region was

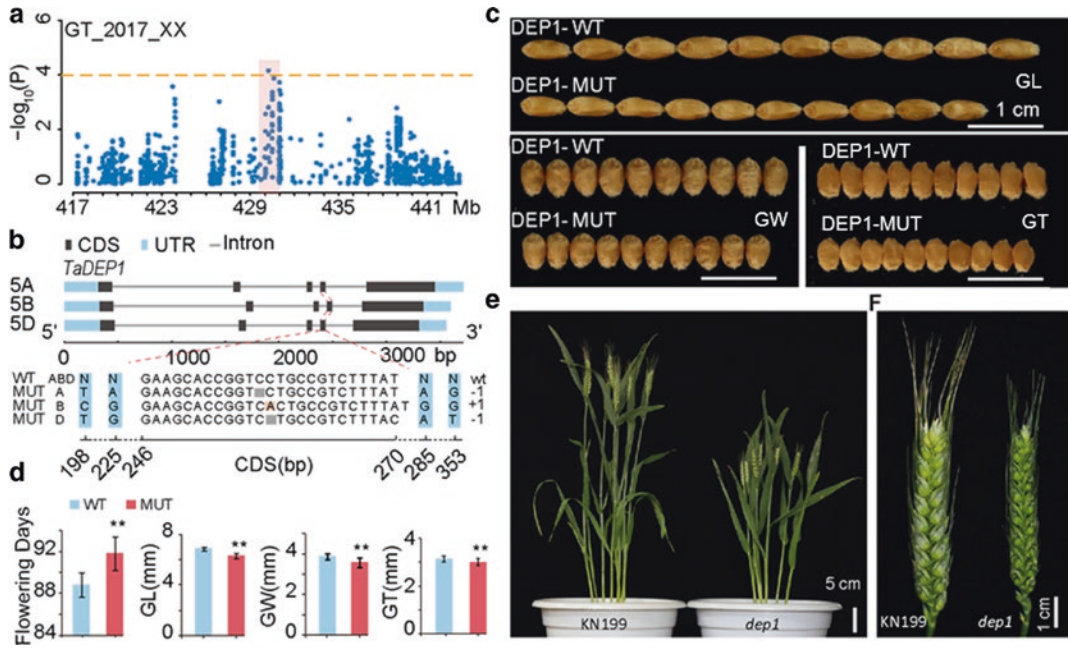


Fig. 6.11 GWAS make it possible to precisely map and verify genes in the low recombination region using CS golden reference and historic recombinations under the assistance of genome re-sequence and gene editing. A grain thickness locus on chromosome 5A carrying the rice DENSE AND ERECT PANICLE ortholog *TaDEP1*. **a** GWAS signals at 430.24 Mb on 5A. **b** Three homologous of DEP1 and their mutated sites by CRISPR-Cas9 in KELONG 199. **c** Seed size difference between the triplet mutant (*DEP1-MUT*) and wild type (WT). **d** Statistics difference between the *DEP1-MUT* and WT on flowering time, grain length (GL), grain wide (GW), and grain thickness (GT). **e** and **f** phenotype difference on plant morphology and spike. Adapted from Li et al. (2022)

overlapping with selection sweeps and contained the wheat homolog of the rice *DEP1* gene (Fig. 6.11) that has been shown in rice to enhance grain yield by promoting nitrogen utilization efficiency (Huang et al. 2009; Xu et al. 2019). The LD block was ~ 1.3 Mb and contained 15 genes. A total of 33 SNPs were present in the region. Haplotype analysis of these SNPs showed that the grain thickness of accessions with haplotype 1 (Hap1) was significantly thicker ($P<0.001$) than that of other accessions (Fig. 6.11), and these two sets of accessions also had significant increases in average thousand grain weight, but reduced plant height. The locus in fact shows pleiotropic effects on multiple agronomic traits and RNA-seq data showed that *TaDEP1* expressed significantly different ($P<0.001$) between accessions of thin-grain and thick-grain. We then used CRISPR/Cas9 editing to introduce deletion mutations into all three *TaDEP1* homoeologs in cv. KENONG199. The edited plants displayed significant ($P<0.01$) reductions in grain size, the edited mutants also showed short stem, more tillers, and compact spike (Fig. 6.11), confirming that *TaDEP1* is a gene with pleiotropic effects and functionally essential for wheat grain size development (Li et al. 2022).

There are more PAV and other SV in common wheat than other crops. Therefore, if the agronomic target is located in the SV region, there are likely to be difficulties to map QTL precisely using biparent population, even in the natural population by GWAS using a single reference genome. A graph of the pangenome for functional genomics and HBB in wheat would be a major advance.

6.7 Yield Genes and Their Diversity

For yield genes, because of their conserved characters among cereals, much work was carried out based on the synteny and collinearity among cereal genomes, especially the good collinearity between wheat and rice. Three very interesting points were found. The first is dominance of the three homologous genes in the hexaploid species. The second is that most of the natural variations occurred within the promotor regions of the crucial genes among cultivars. The third is strong correlation of haplotypes with the water and fertility of the soil as well as sunlight and temperature resources in growing season. For example, the *GS5* was recognized as one gene strongly influencing grain size in cereals (Li et al. 2011), and in wheat, it is preferentially expressing in young spikes and developing grains, and positively regulating grain size. Among the three homoelogous genes, *GS5-3D* has the dominant expression, *GS5-3B* is almost silenced with very low expression level, while the *GS5-3A* is seen to have medium expression levels (Fig. 6.12a). Only one SNP (T/G) was identified at 2334 bp downstream of the ATG start codon at the *TaGS5-3A*. Two alleles were detected on GS5-3A in world modern cultivars, with average 6–7 g difference on thousand grain weight. The global distribution of frequency of larger grain size allele *TaGS5-3A-T* exhibited very good correlation with water resources during wheat growing season (Fig. 6.12b). No diversity was detected at either 3B or 3D loci (Ma et al. 2016).

6.8 Adaptation of Cultivars to Environments

Based on the whether or not a cold temperature vernalization is required to promote flowering, wheat cultivars are classified into winter and spring types. This vernalization requirement prevents temperate plants from flowering under freezing winter conditions. Wheat cultivars grown in different environments need diverse vernalization characteristics to ensure flowering and reproductive development at appropriate time to meet the need for higher yield and mature on time.

In wheat, flowering time is controlled by both vernalization system and photoperiod reaction system together. For the vernalization, there are four genes, *TaVrn1*, *TaVrn2*, *TaVrn3,* and *TaVrn4*, that have been positionally cloned; *TaVrn4* was identified as a duplication of *TaVrn1* (Yan et al. 2003, 2004, 2006; Kippes et al. 2015). The expression level of *TaVrn3* (FT) is the key element determining flowering or not flowering. However, expression of *TaVrn3* is strongly, positively, regulated by *TaVrn1* and *TaVrn4* and *PPD1*, but negatively regulated by *TaVrn2*. Any function mutants in *TaVrn1*, *TaVrn2*, *TaVrn4,* and *PPD1* influence expression of *TaVrn3*, and subsequently the flowering time. This provides wheat with an extensive range of variation to adapt particular combinations of variants to grow environments through combining different alleles at the four loci. Mutations at promoter region of *VRN3* that result in a loss of binding site for *VRN2* lead to complete loss of suppression of *VRN3* by *VRN2* and result in a full spring type wheat (Yan et al. 2003, 2004, 2006; Kippes et al. 2015). Furthermore, it was found *TaVrn1* had significant epistatic effects on flowering time (Xie et al. 2021). Copy number variation (CNV) was also detected at *VRN1* loci, which negatively influences expression level of itself (Diaz et al. 2012). In addition, TFs binding with *cis* at promoter regions of *VRN1*, *VRN2*, and *VRN3* often affecting wheat heading and flowering time (Liu et al., JIBP 2019a). Furthermore, genes in the pathway of auxin were also involved in the regulation of leaf senescence and re-mobility of nutrients from leave and stems to grains in wheat (Li et al. 2023). A detailed summary of the vernalization system and photoperiod reaction networks in wheat is provided by Sehgal et al. (see Chap. 11 in this volume).

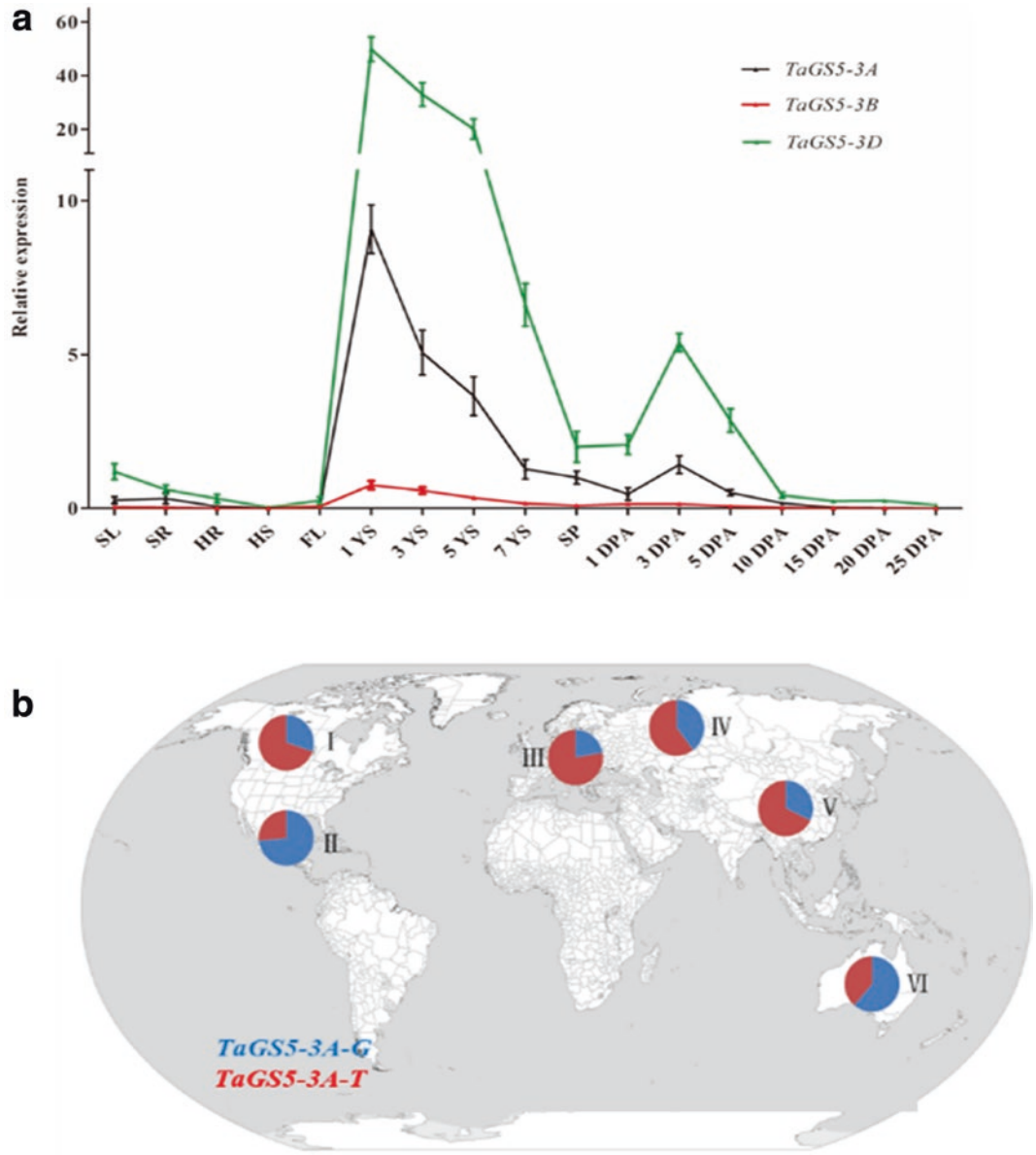


Fig. 6.12 Dominance among the three *GS5* homeology genes in wheat (**a**) and global distribution of alleles in modern cultivars (**b**). **a** Temporal and spatial expression of *TaGS5* homoeologues. SL, seedling leaf; SR, seedling root; HR, root at the heading stage; HS, stem at the heading stage; FL, flag leaf; 1 YS, 3 YS, 5 YS, and 7 YS, young spikes of 1, 3, 5, and 7 cm in length; SP, spike at heading stage; various stages of grain development, including 1 DPA, 3 DPA, 5 DPA, 10 DPA, 15 DPA, 20 DPA, and 25 DPA. The expression of *TaGS5-3A* in the spike at heading stage was assumed to be 1. **b** Distributions of T*aGS5-3A*-T and *TaGS5-3A*-G alleles in wheat cultivars from different ecological regions. North America (I), CIMMYT (II), Europe (III); former USSR (IV); China (V); and Australia (VI)

6.9 Disease-Resistant Genes and Their Diversity

In one life cycle, wheat is threatened by many diseases and pests. Some pathogens' races, such as rusts, powdery mildew change very quickly from year to year. Therefore, wheat disease resistance breeding is a constant evolutionary arms race with their pathogens. Therefore, there must be enough diversity in *R* genes for this race. Fortunately, disease-resistant genes usually mapped at the high recombination regions (R1 and R3, Fig. 6.9) on chromosomes. Frequent recombinations often create new variation and PAV and CNV, which bring great opportunity to create new genes for resistance. Therefore, cloning R genes has been a high priority in wheat molecular biology in the past 10 years and is expected to continue to be a high priority.

Until now, there are three major types of disease resistance genes cloned in wheat (see also Chap. 10). Resistance genes with typical CC-NBS-LRR domains, such as powdery mildew resistance genes *Pm1*, *Pm2,* and *Pm3*, leaf rust resistance genes *Lr1* and *Lr13*, stem rust resistance genes *Sr33* and *Sr35*. 2) *R* genes containing Kinase-MCTP structure, such as the *Yr36*, has a START-Kinase structure. The *Pm4* has a Kinase-MCTP structure. The *Yr15* (*WTK1*), *Sr60* (*WTK2*), and *Pm24* (*WTK3*) have a tandem kinase structure. (3) Disease-resistant genes cording proteins with transmembrane transport functions, such as durable resistance genes *Lr34*/*Yr18*/*Sr57*/*Pm38* and *Lr67*/*Yr46*/*Sr55*/*Pm46*. There are rare natural mutants in landraces carrying good-resistant genes, such as the famous *Fhb1*, *Pm5e* which encode an amino acid mutation in the NLR protein; the deletion of two amino acids in the powdery mildew resistance gene *Pm24* (*WTK3*) confers broad-scope resistance to powdery mildew. Besides common wheat collections, the ancestral species of wheat usually contain abundant disease resistance genes, such as *Pm60* from *T. urartu*, the *Yr15* (*WTK1*), *Yr36,* and *Pm41* from the *T dicoccoides*, the *Lr21* and the powdery mildew resistance gene *WTK4* derived from *A. tauschii*. In addition, distantly related wild species were also good resources to transfer resistance genes to wheat, such as the *Pm21* from *Haynaldia villosa*, conferring durable and broad-spectrum resistance to wheat powdery mildew (Xing et al. 2018). The *Fhb7* from *Thinopyrum elongatum* has good resistance for fusarium head blight spreading in wheat (Wang et al. 2020).

6.10 Prospects

The value of germplasm resources is in the genes hidden within them. The value of a gene is determined by its activity per se as well as the genetic background in which it is recovered. Only by transferring them from un-adapted germplasm into a good genetic background, assaying their value, and integrating them into breeding, we can truly activate them and realize their value for human life.

Genome segment introgression is a major source of genetic variation in wheat. Genomic regions of introgression have provided the hot spots for structural variation that contains many dispensable genes such as tolerant genes to bio-stress and abio-stress. Wheat pangenomes will enable genome-wide high-resolution admixture mapping across species and help figure out causal genetic mutations underlying specific traits (Lei et al. 2021). Furthermore, the pangenome-based research of hallmark cultivars will break through the limitation of having a single reference genome, for revealing the important contributions of chromosomal structural variations (translocation, inversion, duplication/deletion, PAV) in the formation of variety traits. Therefore, a pangenome within and across *Triticum* species will be of interest for wheat genomics in the next 5–10 years for interpreting and utilizing variation at the genome level for breeding and evolution (Khan et al. 2020).

Crucial founder genotypes should be sequenced by the third-generation technology and carefully annotated. Using the founder genotype genome sequences as reference, a set of genetics relative cultivars can be sequenced by cheaper second-generation sequence technology

to reveal the haplotype blocks contributed by the founder genotypes in their genomes. The tracking markers could then be developed for haplotype-based breeding. Using the newest breeding founder genotypes as the recurrent parents crosses with core collections selectively backcrossing the hybrid for 2–3 generations can then establish AB-NAM populations. It is envisaged that through intercrossing between good lines from different AB-NAMs would be efficient strategy to realize the integration of breeding-beneficial alleles with desired haplotype blocks for create super-lines for breeding (Tanksley and Nelson 1996; Hao et al. 2020; Varshney et al. 2021). In addition, tetraploid wheats (see Chap. 8) would be an important and good gene resource for the improvement of common wheat. In view of the fact that the chromosome segments from wild tetraploid wheat have suppression effect on recombination, it is recommended that in the construction process of AB-NAM population, priority should be given to founder genotypes containing more introgressions from wild emmer to increase the recombination ratio, to create more variation, and efficiently to exclude the genetic drag from the wild species.

Acknowledgements We appreciate Dr. Xinyi Liu, Washington University, St Louis, for his valuable help on the archaeological and transmission and permission for use Fig. 6.1. We also appreciate Dr. Zhiyong Liu, Institute of Genetics and Development Biology, CAS for his help on the disease-resistant gene diversity. Thanks should be given to Drs Paux, E and Balfourier, F, GDEC-INREA for providing Figs. 6.2 and 6.9 and fruitful discussion under umbrella of CAAS-INREA joint LAB. We also appreciate Mr Jiao CZ, Drs. Hao CY, Li T in Zhang group, ICS-CAAS, for their professional work in preparing most of the genome diversity figures. Zhang X was supported by the Key Research and Development Program of China (2016YFD01003) and CAAS Innovation Program. Appels R was supported by the University of Melbourne as an Honorary Professor.

References

Avni R, Nave M, Barad O et al (2017) Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. Science 357:93–97. https://doi.org/10.1126/science.aan0032

Balfourier F, Bouchet S, Robert S, De Oliveira R, Rimbert H, Kitt J, Choulet F, International Wheat Genome Sequencing C, BreedWheat C, Paux E (2019) Worldwide phylogeography and history of wheat genetic diversity. Sci Adv 5:eaav0536. https://doi.org/10.1126/sciadv.aav0536

Cheng H, Liu J, Wen J, Nie X, Xu L, Chen N, Li Z, Wang Q, Zheng Z, Li M, Cui L, Liu Z, Bian J, Wang Z, Xu S, Yang Q, Appels R, Han D, Song W, Sun Q, Jiang Y (2019) Frequent intra- and inter-species introgression shapes the landscape of genetic variation in bread wheat. Genome Biol 20:136. https://doi.org/10.1186/s13059-019-1744-x

Díaz A, Zikhali M, Turner AS, Isaac P, Laurie DA (2012) Copy number variation affecting the *Photoperiod-B1* and *Vernalization-A1* genes is associated with altered flowering time in wheat (*Triticum aestivum*). PLoS ONE 7:e33234. https://doi.org/10.1371/journal.pone.0033234

Dvorak J, Akhunov ED, Akhunov AR, Deal KR, Luo MC (2006) Molecular characterization of a diagnostic DNA marker for domesticated tetraploid wheat provides evidence for gene flow from wild tetraploid wheat to hexaploid wheat. Mol Biol Evol 23:1386–1396. https://doi.org/10.1093/molbev/msl004

Feldman M, Levy AA (2012) Genome evolution due to allopolyploidization in wheat. Genetics 192:763–774. https://doi.org/10.1534/genetics.112.146316

Hao C, Jiao C, Hou J, Li T, Liu H, Wang Y, Zheng J, Liu H, Bi Z, Xu F, Zhao J, Ma L, Wang Y, Majeed U, Liu X, Appels R, Maccaferri M, Tuberosa R, Lu H, Zhang X (2020) Resequencing of 145 landmark cultivars reveals asymmetric sub-genome selection and strong founder genotype effects on wheat breeding in China. Mol Plant 13:1733–1751. https://doi.org/10.1016/j.molp.2020.09.001

He F, Pasam R, Shi F, Kant S, Keeble-Gagnere G, Kay P, Forrest K, Fritz A, Hucl P, Wiebe K, Knox R, Cuthbert R, Pozniak C, Akhunova A, Morrell PL, Davies JP, Webb SR, Spangenberg G, Hayes B, Daetwyler H, Tibbits J, Hayden M, Akhunov E (2019) Exome sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome. Nat Genet 51:896–904. https://doi.org/10.1038/s41588-019-0382-2

Huang X, Qian Q, Liu Z, Sun H, He S, Luo D, Xia G, Chu C, Li J, Fu X (2009) Natural variation at the DEP1 locus enhances grain yield in rice. Nat Genet 41:494–497. https://doi.org/10.1038/ng.352

IWGSC (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science 361:eaar7191. https://doi.org/10.1126/science.aar7191

Jordan KW, Wang S, Lun Y, Gardiner LJ, MacLachlan R, Hucl P, Wiebe K, Wong D, Forrest KL, Consortium I, Sharpe AG, Sidebottom CH, Hall N, Toomajian C, Close T, Dubcovsky J, Akhunova A, Talbert L, Bansal UK, Bariana HS, Hayden MJ, Pozniak C, Jeddeloh JA, Hall A, Akhunov E (2015) A haplotype map of

allohexaploid wheat reveals distinct patterns of selection on homoeologous genomes. Genome Biol 16:48. https://doi.org/10.1186/s13059-015-0606-4

Jordan KW, He F, de Soto MF, Akhunova A, Akhunov E (2020) Differential chromatin accessibility landscape reveals structural and functional features of the allopolyploid wheat chromosomes. Genome Biol 21:176. https://doi.org/10.1186/s13059-020-02093-1

Khan AW, Garg V, Roorkiwal M, Golicz AA, Edwards D, Varshney RK (2020) Super-pangenome by integrating the wild side of a species for accelerated crop improvement. Trends Plant Sci 25:148–158. https://doi.org/10.1016/j.tplants.2019.10.012

Kippes N, Debernardi JM, Vasquez-Gross HA, Akpinar BA, Budak H, Kato K, Chao S, Akhunov E, Dubcovsky J (2015) Identification of the *VERNALIZATION 4* gene reveals the origin of spring growth habit in ancient wheats from South Asia. Proc Natl Acad Sci USA 112:e5401-5410. https://doi.org/10.1073/pnas.1514883112

Lei L, Goltsman E, Goodstein D, Wu GA, Rokhsar DS, Vogel JP (2021) Plant pan-genomics comes of age. Annu Rev Plant Biol 72:411–435. https://doi.org/10.1146/annurev-arplant-080720-105454

Li Y, Fan C, Xing Y, Jiang Y, Luo L, Sun L, Shao D, Xu C, Li X, Xiao J, He Y, Zhang Q (2011) Natural variation in *GS5* plays an important role in regulating grain size and yield in rice. Nat Genet 43:1266–1269. https://doi.org/10.1038/ng.977

Li AL, Hao CY, Wang ZY, Geng SF, Jia ML, Wang F, Han X, Kong XC, Yin LJ, Tao S, Deng ZY, Liao RY, Sun GL, Wang K, Ye XG, Jiao CZ, Lu HF, Zhou Y, Fu XD, Zhang XY, Mao L (2022) Wheat breeding history reveals synergistic selection of pleiotropic genomic sites for plant architecture and grain yield. Mol Plant. https://doi.org/10.1016/j.molp.2022.01.004

Ling HQ, Ma B, Shi X, Liu H, Dong L, Sun H, Cao Y, Gao Q, Zheng S, Li Y, Yu Y, Du H, Qi M, Li Y, Lu H, Yu H, Cui Y, Wang N, Chen C, Wu H, Zhao Y, Zhang J, Li Y, Zhou W, Zhang B, Hu W, van Eijk MJT, Tang J, Witsenboer HMA, Zhao S, Li Z, Zhang A, Wang D, Liang C (2018) Genome sequence of the progenitor of wheat A subgenome *Triticum urartu*. Nature 557:424–428. https://doi.org/10.1038/s41586-018-0108-0

Li HF, Liu H, Hao CY, Li T, Wang XL, Yang YX, Zheng J, Zhang XY (2023) The auxin response factor TaARF15-A1 negatively regulates senescence in common wheat (Tritcum aestivum L.). Plant Physiol, 191:1254–1271

Liu X, Lightfoot E, O'Connell TC, Wang H, Li S, Zhou L, Hu Y, Motuzaite-Matuzeviciute G, Jones MK (2014) From necessity to choice: dietary revolutions in west China in the second millennium BC. World Archaeol 46:661–680. https://doi.org/10.1080/00438243.2014.953706

Liu X, Lister DL, Zhao Z, Staff RA, Jones PJ, Zhou L, Pokharia AK, Petrie CA, Pathak A, Lu H, Matuzeviciute GM, Bates J, Pilgram TK, Jones MK (2016) The virtues of small grain size: potential pathways to a distinguishing feature of Asian wheats. Quat Int 426:107–119. https://doi.org/10.1016/j.quaint.2016.02.059

Liu H, Li T, Wang YM, Zheng J, Li HF, Hao CY, Zhang XY (2019a) *TaZIM-A1* negatively regulates flowering time in common wheat (*Triticum aestivum* L.). J Integr Plant Biol 61(3):359/376. https://doi.org/10.1111/jipb.12720

Liu X, Jones PJ, Motuzaite Matuzeviciute G, Hunt HV, Lister DL, An T, Przelomska N, Kneale CJ, Zhao Z, Jones MK (2019b) From ecological opportunism to multi-cropping: mapping food globalisation in prehistory. Quat Sci Rev 206:21–28. https://doi.org/10.1016/j.quascirev.2018.12.017

Luo MC, Gu YQ, Puiu D, Wang H, Twardziok SO, Deal KR, Huo N, Zhu T, Wang L, Wang Y, McGuire PE, Liu S, Long H, Ramasamy RK, Rodriguez JC, Van SL, Yuan L, Wang Z, Xia Z, Xiao L, Anderson OD, Ouyang S, Liang Y, Zimin AV, Pertea G, Qi P, Bennetzen JL, Dai X, Dawson MW, Muller HG, Kugler K, Rivarola-Duarte L, Spannagl M, Mayer KFX, Lu FH, Bevan MW, Leroy P, Li P, You FM, Sun Q, Liu Z, Lyons E, Wicker T, Salzberg SL, Devos KM, Dvorak J (2017) Genome sequence of the progenitor of the wheat D genome *Aegilops tauschii*. Nature 551:498–502. https://doi.org/10.1038/nature24486

Ma L, Li T, Hao C, Wang Y, Chen X, Zhang X (2016) *TaGS5-3A*, a grain size gene selected during wheat improvement for larger kernel and yield. Plant Biotechnol J 14:1269–1280. https://doi.org/10.1111/pbi.12492

Maccaferri M, Harris NS, Twardziok SO, Pasam RK, Gundlach H, Spannagl M, Ormanbekova D, Lux T, Prade VM, Milner SG, Himmelbach A, Mascher M, Bagnaresi P, Faccioli P, Cozzi P, Lauria M, Lazzari B, Stella A, Manconi A, Gnocchi M, Moscatelli M, Avni R, Deek J, Biyiklioglu S, Frascaroli E, Corneti S, Salvi S, Sonnante G, Desiderio F, Mare C, Crosatti C, Mica E, Ozkan H, Kilian B, De Vita P, Marone D, Joukhadar R, Mazzucotelli E, Nigro D, Gadaleta A, Chao S, Faris JD, Melo ATO, Pumphrey M, Pecchioni N, Milanesi L, Wiebe K, Ens J, MacLachlan RP, Clarke JM, Sharpe AG, Koh CS, Liang KYH, Taylor GJ, Knox R, Budak H, Mastrangelo AM, Xu SS, Stein N, Hale I, Distelfeld A, Hayden MJ, Tuberosa R, Walkowiak S, Mayer KFX, Ceriotti A, Pozniak CJ, Cattivelli L (2019) Durum wheat genome highlights past domestication signatures and future improvement targets. Nat Genet 51:885–895. https://doi.org/10.1038/s41588-019-0381-3

Matsuoka Y (2011) Evolution of polyploid triticum wheats under cultivation: the role of domestication, natural hybridization and allopolyploid speciation in their diversification. Plant Cell Physiol 52:750–764. https://doi.org/10.1093/pcp/pcr018

Paux E, Sourdille P, Salse J, Saintenac C, Choulet F, Leroy P, Korol A, Michalak M, Kianian S, Spielmeyer W, Lagudah E, Somers D, Kilian A, Alaux M, Vautrin S, Berges H, Eversole K, Appels R, Safar J, Simkova H, Dolezel J, Bernard M, Feuillet C (2008) A physical map of the 1-gigabase bread wheat chromosome 3B. Science 322:101–104. https://doi.org/10.1126/science.1161847

Peng J, Ronin Y, Fahima T, Roder MS, Li Y, Nevo E, Korol A (2003) Domestication quantitative trait loci in *Triticum dicoccoides*, the progenitor of wheat. Proc Natl Acad Sci USA 100:2489–2494. https://doi.org/10.1073/pnas.252763199

Peng JH, Sun D, Nevo E (2011) Domestication evolution, genetics and genomics in wheat. Mol Breed 28:281–301. https://doi.org/10.1007/s11032-011-9608-4

Pont C, Leroy T, Seidel M, Tondelli A, Duchemin W, Armisen D, Lang D, Bustos-Korts D, Goue N, Balfourier F, Molnar-Lang M, Lage J, Kilian B, Ozkan H, Waite D, Dyer S, Letellier T, Alaux M, Russell J, Keller B, van Eeuwijk F, Spannagl M, Mayer KFX, Waugh R, Stein N, Cattivelli L, Haberer G, Charmet G, Salse J (2019) Tracing the ancestry of modern bread wheats. Nat Genet 51:905–911. https://doi.org/10.1038/s41588-019-0393-z

Sansaloni C, Franco J, Santos B, Percival-Alwyn L, Singh S, Petroli C, Campos J, Dreher K, Payne T, Marshall D, Kilian B, Milne I, Raubach S, Shaw P, Stephen G, Carling J, Pierre CS, Burgueno J, Crosa J, Li H, Guzman C, Kehel Z, Amri A, Kilian A, Wenzl P, Uauy C, Banziger M, Caccamo M, Pixley K (2020) Diversity analysis of 80,000 wheat accessions reveals consequences and opportunities of selection footprints. Nat Commun 11:4572. https://doi.org/10.1038/s41467-020-18404-w

Taagen E, Tanaka J, Gul A, Sorrells ME (2021) Positional-based cloning 'fail-safe' approach is overpowered by wheat chromosome structure variation. Plant Genome 14:e20106

Tanksley SD, Nelson JC (1996) Advanced backcross QTL analysis: a method for the simultaneous discovery and transfer of valuable QTLs from un-adapted germplasm into elite breeding lines. Theor Appl Genet 92:191–203. https://doi.org/10.1007/Bf00223376

Varshney RK, Bohra A, Yu JM, Graner A, Zhang QF, Sorrells ME (2021) Design future crops: genomics-assisted breeding comes of age. Trends Plant Sci 26:631–649. https://doi.org/10.1016/j.tplants.2021.03.010

Wang H, Sun S, Ge W, Zhao L, Hou B, Wang K, Lyu Z, Chen L, Xu S, Guo J, Li M, Su P, Li X, Wang G, Bo C, Fang X, Zhuang W, Cheng X, Wu J, Dong L, Chen W, Li W, Xiao G, Zhao J, Hao Y, Xu Y, Gao Y, Liu W, Liu Y, Yin H, Li J, Li X, Zhao Y, Wang X, Ni F, Ma X, Li A, Xu SS, Bai G, Nevo E, Gao C, Ohm H, Kong L (2020) Horizontal gene transfer of *Fhb7* from fungus underlies *Fusarium* head blight resistance in wheat. Science 368:eaba5435. https://doi.org/10.1126/science.aba5435

Wang Z, Hao C, Zhao J, Li C, Jiao C, Xi W, Hou J, Li T, Liu H, Zhang X (2021) Genomic footprints of wheat evolution in China reflected by a Wheat660K SNP array. Crop J 9:29–41. https://doi.org/10.1016/j.cj.2020.08.006

Würschum T, Langer SM, Longin CFH, Tucker MR, Leiser WL (2017) A modern Green Revolution gene for reduced height in wheat. Plant J 92:892–903. https://doi.org/10.1111/tpj.13726

Xu X, Wu K, Xu R, Wang S, Gao Z, Zhong Y, Li X, Liao H, Fu X (2019) Pyramiding of the *dep1-1* and *NAL1NJ6* alleles achieves sustainable improvements in nitrogen-use efficiency and grain yield in japonica rice breeding. J Genet Genomics 46:325–328. https://doi.org/10.1016/j.jgg.2019.02.009

Xie L, Zhang Y, Wang K, Luo X, Xu D, Tian X, Li L, Ye X, Xia X, Li W, Yan L, Cao S (2021) *TaVrt2*, an SVP-like gene, cooperates with *TaVrn1* to regulate vernalization-induced flowering in wheat. New Phytol 231:834–848. https://doi.org/10.1111/nph.16339

Xing L, Hu P, Liu J, Witek K, Zhou S, Xu J, Zhou W, Gao L, Huang Z, Zhang R, Wang X, Chen P, Wang H, Jones JDG, Karafiatova M, Vrana J, Bartos J, Dolezel J, Tian Y, Wu Y, Cao A (2018) *Pm21* from *Haynaldia villosa* encodes a CC-NBS-LRR protein conferring powdery mildew resistance in wheat. Mol Plant 11:874–878. https://doi.org/10.1016/j.molp.2018.02.013

Yan L, Loukoianov A, Tranquilli G, Helguera M, Fahima T, Dubcovsky J (2003) Positional cloning of the wheat vernalization gene *VRN1*. Proc Natl Acad Sci USA 100:6263–6268. https://doi.org/10.1073/pnas.0937399100

Yan L, Loukoianov A, Blechl A, Tranquilli G, Ramakrishna W, SanMiguel P, Bennetzen JL, Echenique V, Dubcovsky J (2004) The wheat *VRN2* gene is a flowering repressor down-regulated by vernalization. Science 303:1640–1644. https://doi.org/10.1126/science.1094305

Yan L, Fu D, Li C, Blechl A, Tranquilli G, Bonafede M, Sanchez A, Valarik M, Yasuda S, Dubcovsky J (2006) The wheat and barley vernalization gene *VRN3* is an orthologue of *FT*. Proc Natl Acad Sci USA 103:19581–19586. https://doi.org/10.1073/pnas.0607142103

Zhao G, Zou C, Li K, Wang K, Li T, Gao L, Zhang X, Wang H, Yang Z, Liu X et al (2017) The *Aegilops tauschii* genome reveals multiple impacts of transposons. Nat Plant 3:946–955. https://doi.org/10.1038/s41477-017-0067-8

Zhu T, Wang L, Rimbert H, Rodriguez JC, Deal KR, De Oliveira R, Choulet F, Keeble-Gagnère G, Tibbits J, Rogers J, Eversole K, Appels R, Gu YQ, Mascher M, Dvorak J, Ming-Cheng Luo M-C (2021) Optical maps refine the bread wheat *Triticum aestivum* cv Chinese spring genome assembly. Plant J 107:303–314. https://doi.org/10.1111/tpj.15289

7 Ancient Wheat Genomes Illuminate Domestication, Dispersal, and Diversity

Alice Iob, Michael F. Scott and Laura Botigué

Abstract

Ancient DNA (aDNA) promises to revolutionise our understanding of crop evolution. Wheat has been a major crop for millennia and has a particularly interesting history of domestication, dispersal, and hybridisation, summarised briefly here. We review how the fledgling field of wheat archaeogenomics has already contributed to our understanding of this complex history, revealing the diversity of wheat in ancient sites, both in terms of species and genetic composition. Congruently, ancient genomics has identified introgression events from wild relatives during wheat domestication and dispersal. We discuss the analysis of degraded aDNA in the context of large, polyploid wheat genomes and how environmental effects on preservation may limit aDNA availability in wheat. Despite these challenges, wheat archaeogenomics holds great potential for answering open questions regarding the evolution of this crop, namely its domestication, the different dispersal routes of the early domestic forms and the diversity of ancient agricultural practices. Not only will this research enhance our understanding of human history, but it will also contribute valuable knowledge about ancient selective pressures and agriculture, thus aiding in addressing present and future agricultural challenges.

Keywords

Archaeogenomics · Domestication · Genetic diversity

7.1 Shining a Light on the Past: The Promise of Ancient DNA

Ancient DNA (aDNA) has fostered a revolution in evolutionary genomics, as it allows direct observation of historical molecular diversity (Der Sarkissian et al. 2014). Previously, hypotheses were based solely on the observation of modern genetic diversity, which is the end effect of thousands of years of evolution, with the main caveat that the same pattern of genetic variation is often consistent with different historical scenarios (Lawson et al. 2018). The analysis of

A. Iob · L. Botigué (✉)
Centre for Research in Agricultural Genomics (CRAG), CSIC-IRTA-UAB-UB, Campus UAB, Bellaterra, Barcelona, Spain
e-mail: laura.botigue@cragenomica.es

A. Iob
e-mail: alice.iob@cragenomica.es

M. F. Scott
School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich, UK
e-mail: m.f.scott@ucl.ac.uk

R. Appels et al. (eds.), *The Wheat Genome*, Compendium of Plant Genomes,
https://doi.org/10.1007/978-3-031-38294-9_7

aDNA allows the genomic characterization of populations at different points in time, adding a fundamentally new dimension to evolutionary studies (Gutaker and Burbano 2017; Orlando et al. 2021).

The very first aDNA analysis was conducted on a mitochondrial sequence of a museum-preserved quagga (Higuchi et al. 1984). Since then, the field of archaeogenomics has rapidly flourished (Morozova et al. 2016), allowing for a better understanding of human, animal, and plant evolutionary history. Recent advances in this field include sedimentary, epigenetic, pathogens, and microbiome aDNA analysis (Key et al. 2020; Parducci et al. 2017; Pedersen et al. 2014; Spyrou et al. 2019; Warinner et al. 2014).

aDNA has already had a remarkable impact on our understanding of human history, shedding light on important patterns of migration (Lacan et al. 2011), admixture (Yang et al. 2020), adaptation (Marciniak and Perry 2017), population dispersal, expansion, and decline (Nielsen et al. 2017). Notably, aDNA gave fundamental contribution to our knowledge about the genetic relationships between modern humans and their extinct relatives Neanderthals (Weyrich et al. 2015) and Denisovans (Krause et al. 2010; Reich et al. 2010), the latter of which have only been identified through aDNA analysis. Similar insights have been gained in other animals, such as dogs (Botigue et al. 2017; Leathlobhair et al. 2018), cattle (Daly et al. 2018; Verdugo et al. 2019), pigs (Frantz et al. 2019), and horses (Gaunitz et al. 2018). These studies have led to a reassessment of previous evidence and an overturning of the existing narrative (Librado et al. 2021).

Now, aDNA promises a similar revolution in our understanding of how crops have been domesticated and spread around the globe, and the ways that these processes have shaped genetic diversity. By revealing how crops have adapted to new environments and what genetic diversity has been lost, aDNA can also set a basis for future breeding strategies (di Donato et al. 2018; Pont et al. 2019b). Crop archaeogenomics is still in its infancy, but aDNA from several important crops has been analysed, including maize (Ramos-Madrigal et al. 2016), barley (Mascher et al. 2016; Palmer et al. 2009), cotton (Palmer et al. 2012), bean (Trucchi et al. 2021), sunflower (Wales et al. 2018), sorghum (Smith et al. 2019), watermelon (Renner et al. 2019), and emmer wheat (Scott et al. 2019).

In this chapter, we first give a very brief overview of the history of wheat cultivation and the key genetic changes involved. The aDNA technology promises unique insights in this area. We review the wheat aDNA studies carried out so far and their contribution to understanding phenomena that have shaped wheat genomes. To conclude, we discuss the key open questions in this field and discuss the limitations posed by wheat's large polyploid genome and idiosyncratic preservation. Our goal is to give an overview of the important answered and unanswered questions in the history of wheat cultivation and the promise of aDNA for resolving them.

7.2 A Brief History of Wheat Cultivation

Human societies have relied on wheat for thousands of years. Thus, the history of wheat domestication, geographic expansion, and cultivation has cross-disciplinary significance (Fig. 7.1). Understanding how wheat genetic diversity has been shaped also has contemporary relevance due to its continued nutritional and economic importance. Archaeogenomic studies aim to give new information about at least three key aspects of this process: domestication, dispersal, and gene flow between different wheat species. To contextualize contributions from archaeogenomics, we briefly overview these basic tenets of wheat cultivation history.

7.2.1 Domestication

Wild tetraploid emmer wheat was one of the first species to be domesticated (Haas et al. 2018), during the so-called Neolithic Transition, in

Fig. 7.1 Wheat has been culturally important for millennia, and DNA extracted from ancient specimens can reveal how humans have shaped crop genetic diversity. Left: Facsimile of a vignette on the tomb of Sennedjem and Iineferti showing grain harvest in the abundant fields of the next life (painted by Charles K Wilkinson in 1922 CE, original ca. 1295–1213 BCE, public domain image from the Metropolitan Museum of Art). Right: Archaeological specimens of desiccated emmer wheat chaff from Egypt. Photo from Dorian Q. Fuller, University College London, Institute of Archaeology

parallel with humans' shift from hunting and gathering to agriculture and animal husbandry (Diamond 2002). The quintessential trait for cereal domestication is the loss of rachis brittleness: in wild cereals, the spikelets disarticulate spontaneously from the rachis upon maturity, ensuring seed dispersal and germination. In domestic cereals, the rachis is non-brittle; spikelets remain attached, allowing easier harvesting but requiring subsequent sowing in the following season in order to germinate. Because plants with a non-brittle rachis depend on human action for dispersal, this phenotype has been used to define domestication in cereals (Abbo et al. 2014; Snir et al. 2015). Loss-of-function mutations in the *TtBtr1-A* and *TtBtr1-B* genes on chromosomes 3A and 3B are the main determinants of such phenotype (Avni et al. 2017; Nave et al. 2019). Therefore, alleles at these two loci essentially distinguish wild from domesticated emmer wheat. Other traits that are favourable in the human-mediated environment and most likely deleterious in a wild environment (Kantar et al. 2017; Purugganan and Fuller 2009) give a more broad definition of the "domestication syndrome" (Larson et al. 2014), like the loss of seed dormancy and larger seed sizes (Haas et al. 2018; Zohary 2013).

Wild emmer wheat has a very restricted distribution, growing only in the Fertile Crescent region of Southwest (SW) Asia (Vavilov et al. 1992). The exact location of the emergence of domestic emmer has been a long-standing controversy. In the 2000s, early genetic studies started addressing this issue, with the so-called cradle of agriculture theory (Lev-Yadun et al. 2000). Further genetic studies had pointed to the Northern Fertile Crescent and specifically to the Karaca Dağ Mountain region as the centre of domestication of emmer wheat (Luo et al. 2007; Ozkan et al. 2002, 2005), mostly based on the higher similarities between the genomes of the modern domestic landraces and the wild emmer from the Northern Levant, compared to that of the Southern Levant (Avni et al. 2017).

However, this monophyletic origin has been challenged with increasing evidence that different wild populations have contributed to domestic wheats. Several authors argue that domestic emmer wheat arose from an admixed wild population and that mutations for domestication traits appeared in different chromosomes at different times and possibly in different places (Civáň et al. 2013; Jorgensen et al. 2017; Oliveira et al. 2020). This is in line with the observation that the domestic phenotype, which requires at least two independent recessive mutations, took millennia to be established (Avni et al. 2017; Fuller et al. 2014). As testified by the archaeological record, wild emmer wheat was first exploited in the Southern Levant, where increasing, even though small, proportions of phenotypically

domestic emmer wheat are found at different archaeological sites as early as during Early Pre-Pottery Neolithic B (8700–8200 BCE) (Arranz-Otaegui et al. 2018). However, domesticated emmer is found in very high proportions in the Northern Levant starting from the Middle/Late Pre-Pottery Neolithic B (8200–6300 BCE) (Arranz-Otaegui et al. 2016). This indicates that wild emmer was managed (a phenomenon often regarded as "pre-domestication cultivation") (Fuller et al. 2010) long before the domestic forms emerged, and that probably wild populations from across the Fertile Crescent contributed to the domestic pool (Feldman and Kislev 2007). The role of introgression from wild to domestic wheat has been demonstrated by several studies, e.g. (Cheng et al. 2019; Pont et al. 2019b; Przewieslik-Allen et al. 2021), even though the context in which these introgression events took place remains unknown.

Overall, archaeology and genetics point to a slow and geographically widespread domestication process in which both the Northern Levant and the Southern Levant played an important role.

7.2.2 Evolution

Domestic emmer wheat (*Triticum turgidum* subsp. *dicoccon*) gave rise to today's most economically important wheats: tetraploid durum wheat (*T. turgidum* subsp. *durum)* and hexaploid bread wheat (*T. turgidum* subsp. *aestivum*). These descendants differ from their ancestor in one character of great agricultural importance: the free-threshing phenotype. Emmer is a hulled, non-free-threshing wheat, and the extraction of seeds from husks requires substantial mechanical processing. On the other hand, durum and bread wheat are naked and free-threshing: as the spikelets disarticulate from the rachis they fall apart, releasing the seeds without further processing. While durum wheat is tetraploid (BBAA), bread wheat is hexaploid (BBAADD) and evolved from the hybridization of tetraploid wheat with the diploid wild goatgrass (*Aegilops tauschii*), donor of the D subgenome (Haas et al. 2018; Pont et al. 2019a). The tetraploid that contributed the B and A subgenomes to bread wheat has been a matter of debate (Sharma et al. 2019), but considering the need for multiple mutations to determine the free-threshing phenotype, the most supported (and most parsimonious) models indicate that hybridization with *A. tauschii* occurred with a free-threshing tetraploid (Zhou et al. 2020).

The emergence of modern wheat is therefore the result of three processes: (I) domestication of wild emmer wheat, associated with the loss of rachis brittleness; (II) crop evolution (often also referred to as crop improvement under cultivation), which includes the emergence of the free-threshing phenotype and adaptation to new ecological niches; (III) allopolyploidization between a free-threshing tetraploid with *A. tauschii*, giving rise to bread wheat. We summarize these changes in Fig. 7.2.

Perhaps surprisingly, hulled wheats continued to be used for thousands of years after the appearance of free-threshing durum wheat and bread wheat. The slow and regionally specific shifts in wheat usage probably reflect cultural practices and preferences (Nesbitt and Samuel 1996). Also, increasing archaeological evidence shows that early farmers relied on a wide range of other domestic wheats for their subsistence, including einkorn, spelt, and *Triticum timopheevii* alongside emmer and free-threshing wheats (Özbaşaran et al. 2018). This is in accordance with the evidence for intra and interspecific introgression that has been detected in modern wheat (Cheng et al. 2019; Zhou et al. 2020).

7.3 Archaeogenomics of Wheat

Wheat archaeogenomics is a powerful tool to investigate how wild wheat evolved into domestic forms and how these domestic wheat varieties adapted to different ecological niches and cultural preferences through history.

However, the limitations and the characteristics of ancient genomes have to some extent impacted the approach taken in this research

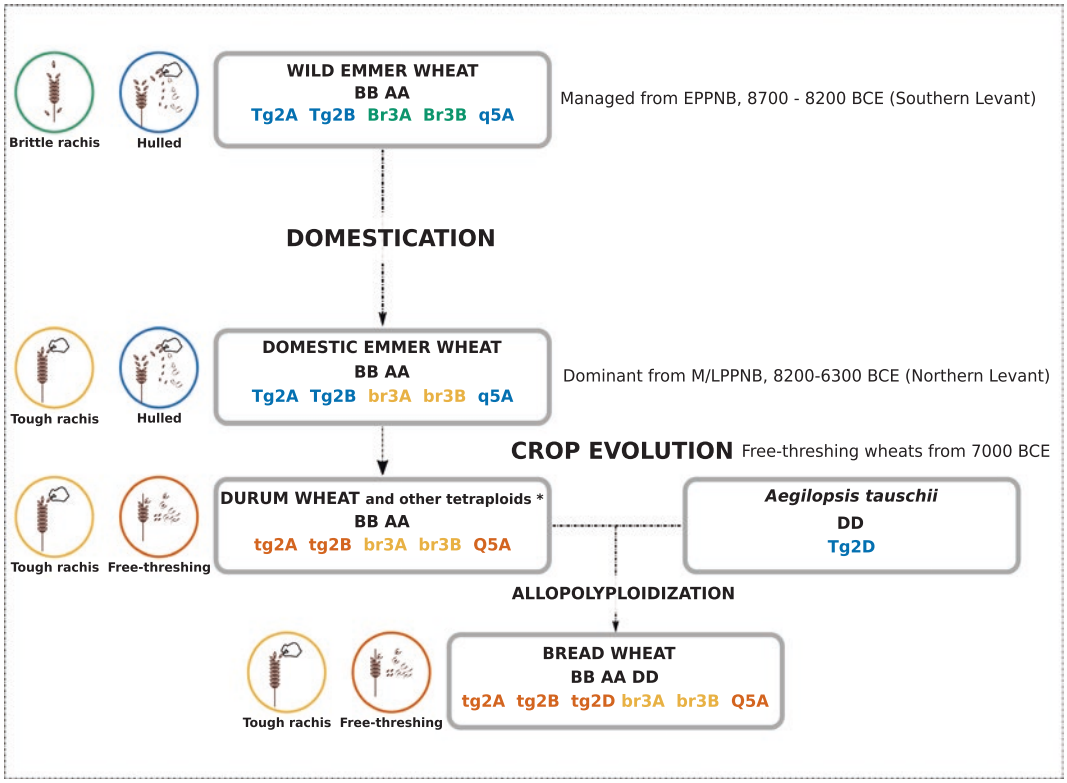


Fig. 7.2 Schematic representation of the domestication and evolution of the most economically important wheats today, showing important phenotypes and the mutations that determine them. Basic information about the appearance of the different wheats in the archaeological record is given on the right. The small white hand represents the investment of human labour in processing the harvest. *For simplicity, we use the common name "durum wheat" for all free-threshing tetraploids, but other common names are used for free-threshing tetraploids, and it is not known which was involved in this allopolyploid event. This scheme is an adaptation of the model proposed by Sharma et al. (2019)

field. Before high-quality reference genomes were available, most studies avoided whole-genome analysis and used a target and amplification strategy. This mitigates the challenges of a large genome but gives much less rich genomic information. Furthermore, the primers used for amplification mask the characteristic patterns of degradation that are useful for ruling out contamination by confirming the antiquity of the DNA. Unlike these amplification methods, whole-genome libraries can also be re-analysed to get more data without further destructive sampling of rare material. For these reasons, amplification approaches are no longer recommended for ancient samples (Gutaker and Burbano 2017; Prüfer and Meyer 2015).

We first overview wheat aDNA studies that use amplification and then describe the first two whole-genome analyses. Even though wheat archaeogenomics is in a germinal stage, the results have shifted our understanding of wheat genetics in important ways.

7.3.1 Target Gene Amplification

The most common use of target gene amplification has been to interrogate key genes or to identify wheat remains at the species level. The *x* and *y* copies of the *Glu1* loci were often the focus of early studies. These genes, present in all wheat subgenomes, are located in the long

arms of chromosome 1 and encode for the high molecular weight glutenin subunits (HMW-GSs), storage proteins present in the starchy endosperm cells of wheat. Allelic varieties in these genes impact the properties of dough for bread making. Because of its effect over bread quality, the evolution of the HMW genes can provide insights into the nature of human selective pressures during wheat evolution (Allaby et al. 1999). In this manuscript, authors surveyed these loci in a collection of modern and ancient wheats, constructed a phylogenetic tree, and obtained time estimates by using a substitution rate to calibrate the observed variation. By comparing the genetic variability for x and y copies in each genome, they were able to determine that the genetic variability in these loci for the cultivated species predates domestication, pointing to either incomplete lineage sorting, multiple domestication events, or introgression after domestication. Another study used a similar approach with the same loci to inquire about the origins of spelt (Blatter et al. 2002). They surveyed a collection of modern and ancient bread wheat and spelt specimens and determined that the high genetic variability of spelt compared to that of bread wheat in the A and B genomes are compatible with the origin of spelt being a hybridization event between bread wheat and hulled tetraploid emmer.

HMW genes have also been used to identify wheat remains at the subspecies level and inform about its dispersal. Without associated chaff, it is difficult to distinguish between free-threshing wheats (e.g. bread wheat or durum wheat). Bilgic et al. (2016) targeted the HMW promoter region in 8400-year-old specimens from a notorious Neolithic site in central Turkey, Çatalhöyük, to determine whether the genetic variability characteristic of the D genome could be recovered, as a proof of that wheat being hexaploid. The finding of HMW subunits from the A, B, and D genomes is quite remarkable, since it evidences the presence of hexaploid wheat at a very early point in time and highlights the importance of this settlement in the expansion of hexaploid wheat cultivation. Another study used the Internal Transcribed Spacer regions (ITS1 and ITS2) and the Inter-Genic Spacer region (IGS) from the nuclear ribosomal DNA for species level identification (Li et al. 2011). They also found early evidence for hexaploid wheat in Northwest China around 1760–1540 BCE.

These results highlight the high diversity of wheats consumed by humans during early agricultural expansion. Free-threshing naked wheats first appear in the archaeological record between 7000 and 5500 BCE (Feldman and Kislev 2007). Early naked wheats co-existed with domestic and wild emmer populations (Bilgic et al. 2016), giving opportunities for genetic exchange. Along with the protracted period of emmer domestication, this probably explains the higher genetic diversity on A and B subgenomes of modern bread wheat compared to the D subgenome (Cheng et al. 2019). This demonstrates how the details of agricultural history directly impact modern wheat diversity and breeding. Moreover, other wild *Triticum* species gave rise to domestic forms during the Neolithic. These include the diploid einkorn wheat, *Triticum monococcum* subsp. *monococcum,* that emerged from wild einkorn, *T. monococcum* subsp. *Aegilopoides* (Nesbitt and Samuel 1996), spelt (*Triticum spelta*), an hulled hexaploid, and tetraploid *T. timopheevii* (domesticated from *T. timopheevii araraticum*) (Wagenaar 1966), only recently classified thanks to aDNA analysis.

The position of *T. timopheevii* within the domestication process of wheat in SW Asia exemplifies the value of aDNA to gain insights on certain domestication processes. Briefly, due to the technical difficulties in the identification of *T. timopheevii,* for a long time its existence was questioned, and it was often unclassified, or ascribed to other wheat species, such as "New Glume Wheat". Recently, archaeological remains described as "New Glume Wheat" have been designated as domestic *T. timopheevii* based on aDNA evidence (Czajkowska et al. 2020). The authors used the *Ppd1* locus to identify G genome alleles in "New Glume Wheat" remains. This study has sparked the interest of the archaeobotanical community. Decades have passed since the first classification of

an archaeological specimen to "New Glume Wheat". It was not until numerous remains of this type of wheat were found in several Neolithic and Bronze Age archaeological sites in northern Greece and compared with other locations (Jones et al. 2000) that archaeologists were able to describe the distinctive features of this wheat (Ulaş and Fiorentino 2021). Nevertheless, identification based on grain morphology is still problematic. The identification of New Glume Wheat as domestic *T. timopheevii* thanks to ancient DNA analysis has had important ramifications on our understanding of the complexity of the domestication process in SW Asia and the confirmation that multiple species evolved into domestic forms, moving away from the "founder crops" theory. *T. timopheevii* was actually cultivated for a very long period of time in certain regions. New efforts are now being undertaken to revisit archaeobotanical assemblages and reassess the relative abundance of plant species, with the expectation that many grains classified as emmer wheat will now be classified as *T. timopheevii*.

The HMW loci were also used, together with the ribulose 1,5 biphosphate carboxylase (rbcL) and the chloroplast microsatellite WCT12 in the chloroplast genome to study the viability of DNA extraction on ancient plant specimens (Fernández et al. 2013). In this study, 126 grains of naked wheat in different preservation conditions (charred, partially charred, and waterlogged) were analysed (Fig. 7.3 shows different preservation conditions of ancient wheat samples). Results showed that DNA extraction from totally charred remains is virtually impossible, while DNA amplification of modern contaminants is pervasive. Unfortunately, almost all of the most ancient archaeological wheat specimens are charred, which is a severe limitation for future aDNA studies.

As mentioned above, one important limitation of amplification-based studies is the confidence with which one can rule out contamination. Commonly used indicators such as the fragment length distribution or deamination patterns are difficult to assess in target-specific PCR amplification studies. In addition, Allaby et al. (1999) reported PCR jumping, probably related with the shortness of some fragments. Their results showed patterns of linked diversity that did not exist in the modern pool and had to manually rearrange the observed diversity so it would match known modern haplotypes with the subsequent potential biases.

Different strategies have been used to increase confidence in the antiquity of the data. Allaby et al. replicated the results in situ with

Fig. 7.3 Examples of different preservation conditions of archaeobotanical wheat. *Left:* charred emmer wheat seeds from the Vinča culture in Serbia (middle/late Neolithic; c. 5400–4600/4500 BC), published in Filipovic (2014). *Right:* Waterlogged chaff remains of *Triticum* cf. durum/turgidum from the end of the 5th millennium BC at the site of Les Bagnoles. Photo by Raül Soteras, AgriChange Project, reproduced with permission

the same specimen and produced blanks with each extraction run. Czajkowska et al. (2020) performed the extractions in laboratory facilities where no wheat had been processed before, hoping to preclude contamination. Bilgic et al. (2016) processed all samples in two different facilities, so that replication of the results acts as a proof of authenticity. In spite of this, even if contamination can be ruled out, it is not possible to distinguish deamination patterns from true polymorphisms. Therefore, phylogenetic analyses and interpretation of the accumulation of variation through time should be taken with caution unless transitions (C/T or G/A SNPs) are excluded.

7.3.2 Whole-Genome Analyses

As with modern wheat samples, the genomic scale of archaeological wheat genetics has been expanded since the publication of reference genomes (Table 7.1). Nevertheless, only two studies have so far reported whole-genome sequence from archaeological wheat specimens. One has been the analysis of several bread wheat remains from China to infer dispersal into the region (Wu et al. 2019). The earliest bread wheat remains found in China date to approximately 4500 years ago in the north-western part of the country, but the most interesting aspect of its dispersal is that upon its arrival, wheat had to

Table 7.1 Genomic information available for wheats and relatives mentioned in the text

Species name	Genome(s)	Genome size	Common name	Key phenotypes	Reference genome(s)
Aegilops tauschii	D	4 Gb	Tausch's goatgrass		Luo et al. (2017)
Triticum urartu	A	4.5 Gb	Wild red einkorn	Brittle rachis, hulled	Ling et al. (2018)
Triticum monococcum	A^m	5.7 Gb	Wild einkorn	Brittle rachis, hulled	NA
			Einkorn	Non-brittle rachis, hulled	NA
Triticum turgidum	BA	12 Gb	Wild emmer	Brittle rachis, hulled	Avni et al. (2017), Zhu et al. (2019)
			Emmer	Non-brittle rachis, hulled	NA
			Durum	Non-brittle rachis, free-threshing	Maccaferri et al. (2019)
Triticum timopheevii	GA	5.7 Gb	Wild Timopheev's wheat	Brittle rachis, hulled	NA
			Timopheev's wheat	Non-brittle rachis, hulled	NA
Triticum aestivum	BAD	17 Gb	Spelt	Non-brittle rachis, hulled	Walkowiak et al. (2020)
			Bread/Common	Non-brittle rachis, free-threshing	Appels et al. (2018), Alonge et al. (2020), Walkowiak et al. (2020)

This is not a comprehensive list of wheat species/subspecies

be adapted to a wide variety of climatic conditions. Ancient wheat from two archaeological sites within the Xinjiang winter-spring wheat zone was analysed. Even though coverage was extremely low (0.25–0.01x), the authors were able to call more than 7000 SNP sites, compare them with modern data from neighbouring regions, and provide new evidence on wheat dispersal in China, a still controversial topic. Their results were consistent with one of the routes that had been previously suggested: an early dispersal into the Qinjianh Tibetan plateau, based on the highest genetic similarities between the ancient samples and the modern ones from that region. Conversely, another ancient route that advocated for an introduction towards the eastern region was not supported. However, more data is needed to determine whether different gene pools were introduced to China and to confirm that modern landraces correspond with ancient ones from the same area.

Another whole-genome analysis of archaeobotanical specimens looked at two desiccated samples of 3000-year-old emmer wheat chaff (Fig. 7.4) from Egypt (Scott et al. 2019) to investigate early wheat dispersal and introgression from wild populations. The ancient samples were used to genotype exonic SNPs that segregate in modern accessions, at which coverage was 0.48 X after quality control, yielding approximately 100,000 high confidence genotypes. The authors used a haplotype-based approach to overcome as much as possible the limitations of aDNA analysis of polyploid species. Nearby sites that are not broken apart by recombination form co-inherited blocks called haplotypes. A "haplotype reference panel" combines information from multiple modern genomes to characterise the haplotypic variation at each genomic location (McCarthy et al. 2016). In the analysis of ancient data, when a sufficient number of genotypes can be identified within a region, it is possible to assign a known haplotype (or no known haplotype, as may be the case when ancient diversity has been lost in existing populations) to the ancient sample. At this point, non-sequenced genotypes within the region can be deduced based on haplotype assignment, a method called imputation. Haplotypes are relatively long in wheat (Walkowiak et al. 2020) because selfing tends not to break apart haplotypes as much

Fig. 7.4 Desiccated emmer wheat chaff from Hememiah North Spur (Egypt) 14C dated 1300–1000 BC, analysed by Scott et al. 2019. Photo by Chris J. Stevens, reproduced with permission

as outcrossing. As a consequence, low coverage data is more likely to yield enough sites to assign an individual to a haplotype. This method allowed Scott et al. (2019) to identify genomic tracts tens of megabases long containing hundreds of genotypes that matched a modern sample in the haplotype reference panel. These included regions where important domestication QTLs had been identified, such that the domestication allele can be imputed and the phenotype inferred. In contrast, other genomic regions did not match anything in the haplotype reference panel.

The data essentially confirmed that genetic changes associated with domestication were completed by 3000 years ago, prior to emmer wheat dispersal to Egypt. Nevertheless, the ancient Egyptian sample carried more "unique" haplotypes than any other domesticated sample in the dataset, indicating regions where genetic diversity has been lost. It is not yet possible to state whether this lost variation is associated with adaptation to local environmental conditions or confers other useful traits. Nevertheless, these results highlight geographic and genomic regions that may harbour genetic diversity that has been used in the past and therefore might be useful in the present and future. Moreover, while the highly repetitive nature of the wheat genome increases the chances of misalignment issues and subsequent inflated heterozygosity, Scott et al. (2019) found that the estimated heterozygosity of the ancient sample fell within the range of the modern samples. This suggests that reliable genotypes can be obtained from ancient wheat, providing appropriate quality filters are used to restrict attention to sites that do not suffer from alignment problems.

Important results from this study concern early emmer wheat dispersal. Ancient routes of dispersal generally define modern population structure and overall genetic similarity but, with the changing usage of different wheat species and the adoption of modern elite varieties, we have little grasp of historical population dispersal and replacement. Contemporary emmer wheat subpopulations (landraces) reflect the dispersal outside of SW Asia to the West (Mediterranean), to the Balkans (Eastern Europe), to Transcaucasia (Caucasus) and towards India and the Arabian peninsula (Indian Ocean) (Avni et al. 2017). The authors found that the ancient sample from Egypt resembles modern cultivars from the Indian Ocean subgroup, indicating a connection between early emmer dispersal to the East (across the Iranian Plateau and into the Indus valley) and to the South-West (Nile Valley). This is particularly interesting in light of the fact that Ethiopia currently represents a region of genetic isolation and differentiation for tetraploid wheat. This ancient Egyptian sample also has signatures of gene flow with wild populations in the Southern Levant, which could have occurred during dispersal towards Egypt or during Egyptian conquests in the Ramesside era. We expect further aDNA studies to connect historical events with changes to wheat genetics. Answering these questions will not only bring a deeper understanding of wheat evolution, but also human history, which has been intimately linked to wheat cultivation for millennia.

Overall, the field of wheat archaeogenomics has yet to reach its full potential. However, the field is primed for new advances with the availability of reference genomes and a wealth of resequenced modern landraces for comparison. While the prospects for studying DNA from charred remains are poor, many desiccated or waterlogged samples have great potential for further study. Archaeological research on waterlogged sites is increasing, which promises new material to complement the specimens currently in museums and collections.

7.4 Analysing Degraded DNA from Ancient Polyploid Wheat

Degradation and contamination are key complications for the reliable analysis of ancient DNA. To mitigate these problems, specific methods have been developed for sample preparation and downstream analysis (reviewed in Orlando et al. 2021). Even with appropriate methodology, DNA from ancient and historical samples

cannot be used for all the applications that modern sequence data allows. We briefly overview these general principles of ancient DNA analysis, before discussing the specific issues posed by wheat, as all these factors should be considered during study design and analysis. We expect future methodological improvements to address these challenges, raising the possibility of resolving further important questions in the history of wheat domestication and evolution.

7.4.1 aDNA Damage

A prominent difference between ancient and modern DNA is that ancient DNA is much more fragmented prior to extraction (Fig. 7.5a). Most DNA fragmentation occurs rapidly after death (Kistler et al. 2017), as the DNA "backbone" breaks down through a process called "hydrolytic depurination", which is biochemically predicted to occur more rapidly with exposure to water and high temperatures (Lindahl 1993). Thus, local preservation and environmental conditions are key in determining DNA yield and quality in different samples. Nevertheless, fruitful DNA sequencing has been conducted from plant tissue that is thousands of years old and from tropical and warm environments (Fornaciari et al. 2018; Mascher et al. 2016; Ramos-Madrigal et al. 2016; Renner et al. 2019). Overall, excellent DNA preservation has been reported from plant remains in desiccated and waterlogged conditions (Kistler et al. 2020).

Besides fragmentation, the DNA sequence itself undergoes modifications. Notably, a proportion of cytosine residues lose an amine group, becoming uracil residues, which code as thymine during sequencing (Briggs et al. 2007). This hydrolytic deamination occurs more commonly on the single stranded overhangs of the fragmented DNA molecules. As a result, when aligned to a reference genome, sequenced ancient DNA has a higher proportion of C-to-T misincorporations at the 5′ end of each fragment. Double-stranded DNA libraries will also show a higher proportion of the complementary misincorporation, G-to-A, at the 3′ end of each fragment after alignment.

These characteristic patterns of degradation found in ancient samples can be useful to the analysis, as they are proof of the sample antiquity. Therefore, the most common approach is to carry out a protocol developed for partial UDG treatment (Rohland et al. 2015). With this method, uracil-DNA-glycosylase (UDG) is used to remove uracils (Briggs et al. 2010) in the inner region of the fragments, but not at their ends. In this way, some amount of damage is maintained, but it is confined to the fragment

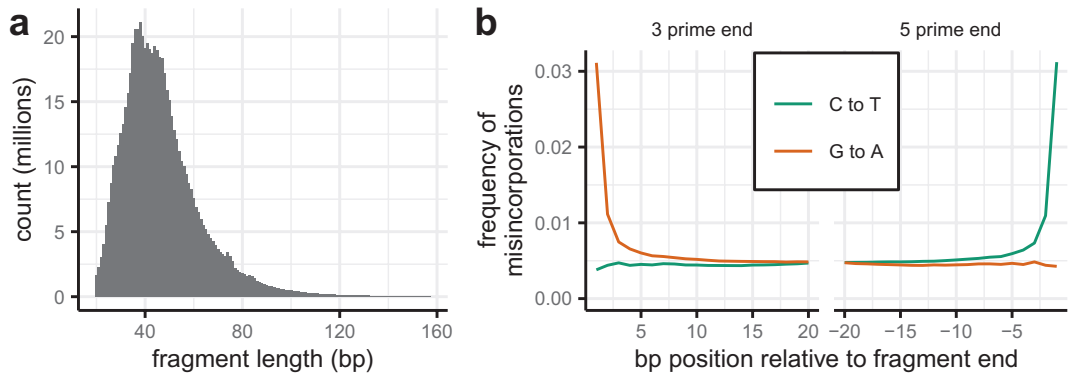


Fig. 7.5 Characteristic patterns of DNA degradation in sequence from a 3000-year-old emmer wheat sample (Scott et al. 2019). **a** Shows the raw distribution of fragments sizes and **b** shows misincorporations relative to the reference genome after alignment. In this case, the sequenced library was partially UDG treated such that the misincorporations caused by post-mortem damage are confined to a few base pairs at the fragment ends, which are removed for further analysis

ends (Fig. 7.5b). Similarly, the distribution of fragment lengths is used to confirm that the sequenced DNA is ancient, where large fragments may indicate contamination. Finally, paired-end sequencing of short fragments will often result in the same base pair being sequenced twice, which can be used to improve confidence in the sequence (Jonsson et al. 2014).

Standard bioinformatic protocols have been established for processing fragmented and damaged DNA. In general, standard approaches have been established for mapping short-read data to reference genomes and automated tools/pipelines are available for ancient genotypes calling for downstream analyses (Peltzer et al. 2016; Schubert et al. 2014). Common methods involve trimming off all the base pairs at the end of fragments that are potentially affected by damage (Jonsson et al. 2014) and verifying that analyses are unaffected when transitions (SNPs where the two alleles are either C/T or G/A and that can include post-mortem damage) are excluded (Korneliussen et al. 2014). We further note that "reference bias" (preferential alignment of reads carrying the same allele as the reference) is stronger in ancient data due to the shorter fragment size, so correction methods should be used (Günther and Nettelblad 2019).

For all these reasons, whole-genome sequencing has become the standard in ancient DNA studies, while PCR-based approaches are no longer considered unless for very specific goals such as genome identification, since they do not allow to verify the presence of these important patterns of post-mortem damage and to exclude contamination.

Contamination is a significant concern in ancient DNA studies. Because the amount of DNA preserved in ancient samples tends to be low, relatively small amounts of contamination from contemporary material can overwhelm the target DNA in the library (Renaud et al. 2019). Extraction and manipulation of ancient DNA therefore requires specialized facilities with protocols that minimize contamination by modern DNA (Fulton 2012). Standard practice is to create a control sequencing library without using the sample tissue (an "extraction blank"). The data from controls is analysed alongside the main sample to quantify the contamination and spurious signals likely to have been introduced during DNA extraction. Contamination can also come from microbial decomposers that invade tissues after death. A simple estimate for overall contamination is the percentage of reads that can be aligned to the reference genome of the targeted species, although other methods are available (Peyrégne and Prüfer 2020). So far, the percentage of endogenous DNA (the DNA of interest) reported in whole-genome studies of ancient plants has been high, compared to animal studies. For example, reported endogenous fractions have been 33–66% in emmer wheat (Scott et al. 2019), 5–90% in bread wheat (Wu et al. 2019), 7–54% (mean 44%) in common bean (Trucchi et al. 2021), and 70% in maize (Ramos-Madrigal et al. 2016).

Degradation and contamination limit the applications of ancient DNA, relative to modern DNA. Firstly, the fraction of endogenous DNA in well-preserved ancient DNA libraries is far below that of modern DNA (which usually is > 99%). Because endogenous fragments are short, the sequencer will often read through the DNA fragment and continue onto the adapter sequences used for library preparation. Sequenced adapter fragments must thus be discarded. Furthermore, if the sequencing has been performed for paired-ends, the forward and reverse reads will overlap (and are then collapsed into a consensus sequence). Given the low endogenous content and the short fragments, more sequence data is needed to reach reasonable coverage. Nevertheless, when small amounts of DNA are present in the sample, it may not be possible to keep sequencing to increase the coverage, since the library gradually yields diminishing returns as more duplicate reads are sequenced (Link et al. 2017). For all these reasons, coverage tends to be significantly lower in aDNA studies, when compared to the expectations for modern data.

Overall, due to low coverage and short fragments in ancient DNA, a typical approach is to identify variable sites (e.g. SNPs) using modern samples only, then use ancient DNA alignments

to genotype the ancient samples. Fortunately, this approach often yields sufficient high-quality genotypes to perform analyses of interest, such as estimating genome-wide relatedness, introgression, and population genetic parameters.

7.4.2 Large Polyploid Wheat Genomes

The large genome of wheat (17 gigabases for bread wheat) implies that whole-genome sequencing of each wheat sample requires more resources compared to other organisms with smaller genomes. This cost is exacerbated in ancient DNA studies by the lower fraction of endogenous DNA, which requires further sequencing effort to obtain the same genomic coverage. In wheat, pre-designed probes are available for exons and promoters (Gardiner et al. 2019; Jordan et al. 2015), which reduce sequencing costs by enriching for sequences that are captured by the probes used. In ancient DNA, capture can enrich endogenous DNA (Hofreiter et al. 2015) but increase clonality and introduce biases towards the sequence on the probes (Ávila-Arcos et al. 2011). Exome-wide capture has not been reported for an ancient wheat. However, targeted capture might be useful to avoid repetitive regions since short aDNA fragments give little information about this class of DNA.

Ploidy and the high identity between subgenomes, estimated to be as high as 97–98%, supposes another challenge for ancient DNA studies. Even with modern samples, wheat resequencing studies can only reliably observe genomic regions that can be unambiguously aligned using the read lengths available. The shorter fragment length of ancient DNA places a practical limit on the portion of the genome that can be directly observed by mapping to reference genomes.

Heterozygosity is commonly used as an indicator of misalignment problems. Because wheat is predominantly selfing (Golenberg 1988), most sites should be homozygous in most individuals. However, various structural variants can cause reads from different genomic regions in the sample to be aligned to the same position in the reference genome (Fig. 7.6) with high mapping-quality scores, thus passing quality filters. As a consequence, sample heterozygosity will be inflated after calling genotypes. A common solution is to remove variants that are heterozygous in multiple samples, e.g. (Gardiner et al. 2019; He et al. 2019). Recent data indicates that undetected gene duplicates are common within wheat subgenomes on reference assemblies (Alonge et al. 2020). In general, polyploid wheat resequencing data will suffer from additional misalignments due to homeologous sequences on

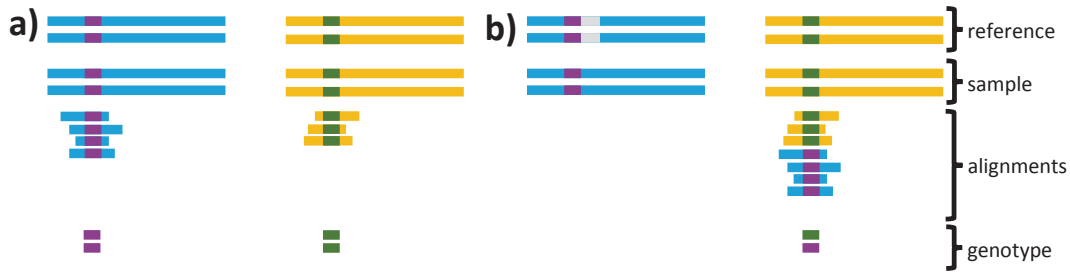


Fig. 7.6 False heterozygosity introduced by mis-mappings to the reference. Here, we consider two genomic regions (blue and yellow), which are homeologues or duplicated regions that are relatively similar to one another. A site in each region is genotyped (coloured purple and green). In **a**, the sample is similar to the reference so that reads can be aligned to the correct region, and the genotype calls are all homozygous, as expected for most sites in a largely selfing species. In **b**, there is a difference between the reference genome and sequenced genome (indicated in grey). The sample reads from the blue genomic region in b are best aligned to the yellow region of the reference. This results in a heterozygous genotype call, while all the true genotypes are homozygous. Thus, inaccurate reference genome assemblies, deletions, insertions, or duplications can all result in spurious heterozygous genotypes

different subgenomes, but reliable genotypes can be obtained from both modern and ancient wheat provided appropriate quality filters are used to restrict attention to sites that do not suffer from alignment problems. Nevertheless, we emphasize that care should be taken when measuring heterozygosity in polyploid wheats, especially from ancient genomes. The limitations in estimating heterozygosity are unfortunate because it is heterozygosity that is a common indicator of outcrossing and genetic variation in the population, changes to which are key questions in the history of cultivation practices (Smith et al. 2019; Trucchi et al. 2021).

7.5 The Future of the Past: Open Questions and Prospects for Wheat aDNA

Crop archaeogenomics has already proved to be a powerful tool to investigate phenomena such as domestication, crop dispersal, and subsequent adaptation (Kistler et al. 2020; Orlando et al. 2021). Studies on bean (Trucchi et al. 2021), sunflower (Wales et al. 2019), and sorghum (Smith et al. 2019) showed that the "domestication bottleneck" (i.e. the initial loss of genetic diversity associated with domestication) may not be as intense as previously assumed. Ancient DNA analysis has been used to trace the origin of some important winemaking grape cultivars (Ramos-Madrigal et al. 2019) and brought insights on the genetic basis of potato adaptation to the European climate (Gutaker et al. 2019). In maize, adaptation to climatic constraints (selected from ancient standing variation within the domestic forms) has been identified as the main driver of modern differentiation between populations (Da Fonseca et al. 2015; Swarts et al. 2017).

7.5.1 Open Questions in Domestication

In recent years, some paradigms of domestication have been challenged by new scientific discoveries, and wheat represents a good example of such changing perspectives. Because now we know that domestic forms took thousands of years to dominate archaeological assemblages and that different wild populations seem to contribute to modern diversity, it is likely that wheat domestication was not as severe, abrupt, or geographically restricted as expected under the assumption of a "domestication bottleneck" (see Sect. 7.2). The presence of peculiar haplotypes in an ancient emmer wheat sample from Egypt showed that possibly genetic diversity has been lost after emmer wheat domestication and dispersal to Egypt (Scott et al. 2019), in line with what has been found for other species, e.g. (Trucchi et al. 2021). In the case of wheat, more ancient samples are needed to determine the association (or lack of thereof) between domestication and losses of genetic diversity.

Second, it is unclear whether there is a monophyletic "centre of domestication" for emmer wheat in the Northern Levant. The contribution of the Southern Levant gene pool to domestic emmer has been detected in several studies, but its origin remains unsolved. Whether emmer was domesticated from a proto-domestic admixed population, or if early domestic populations benefited from extensive gene flow from the wild is still to be revealed. It has been proposed that the high genetic similarity of modern domestic to Turkish wild emmer could be explained by a feralization of the very first proto-domestic population (Civáň et al. 2013; Oliveira et al. 2020). The analysis of wild and domestic samples from this region dating back to Pre-Pottery Neolithic and Neolithic could help determine the origin of the domestic pool, and its relationships with ancient and extant wild populations.

The recent genetic identification of domesticated *T. timopheevii* has triggered a re-evaluation of its importance and abundance in the archaeological record. This effort will be greatly aided by a genetic survey of the modern wild specimens, together with ancient seeds. In general, it will be interesting to use ancient and modern genetic data to compare the origins in space and time of parallel domestication events in wheat (emmer wheat, einkorn wheat, and *T. timopheevii*).

Prospects for the analysis of DNA from fully charred remains are poor, which limits the direct genetic analysis to unveil some of the earliest and most crucial events in wheat domestication. Nevertheless, we expect that improvements in the modelling of genomic evolution and the increasing availability of waterlogged remains will allow to test alternative scenarios on top of addressing questions concerning adaptation and spread of wheat.

7.5.2 Open Questions in Dispersal and Adaptation

The dispersal of wheat was accompanied by adaptation to different environments, leading to the evolutionary success of this species. An interesting example is adaptation to altitude along certain dispersal routes. Wild emmer wheat from the Northern Levant, the closest to all domestic landraces, is always found at high altitude. Its dispersal towards Egypt entailed cultivation at sea level, but emmer wheat grown on the Ethiopian plateau is cultivated at high altitudes again. There are two possible routes of dispersal leading to Ethiopia, one through Africa and another through the Iranian plateau and the Arabian Peninsula. The first one would entail a second adaptation event to high altitudes. The other would have always been cultivated at high altitudes, but there would require a longer dispersal route. How did emmer wheat arrive to Ethiopia? The analysis of desiccated specimens from the Arabian Peninsula, Sudan, and ideally Iran could help to answer this question, as well as potentially unveiling genetic mechanisms for adaptation to high altitude.

7.5.3 Open Questions in Hybridization and Speciation

Archaeological data increasingly suggests that different wheat species were used in a complex geographical mosaic that shifted through time. Given that several wheat species, i.e. emmer, einkorn, naked wheats, and *T. timopheevi* (and wild relatives) co-existed in the same area for millennia, we can ask how much genetic exchange was ongoing in Neolithic settlements. While the vast majority of wheat cultivated today is bread wheat, other free-threshing hexaploids such as the Indian dwarf wheat or the Yunan wheat could have arisen from different hybridization events, since the phylogeny of the A and B genomes differs from that of the D genome (Zhou et al. 2020). Furthermore, forms such as *T. compactum* (Club Wheat) have been described (e.g. Kaplan et al. 1992), even though it is unclear whether these morphotypes are the product of different hybridizations events or the consequence of differential selective pressures. A comparison of the D subgenome in ancient hexaploids with modern *Aegilops* specimens could tackle this question and narrow down the geographic origin where these hybridizations occurred.

Even more intriguingly, we can speculate whether introgressed genetic variation between different wheats was important for crop evolution and adaptation to different environments such as adaptation to northern latitudes or to heat stress. Einkorn wheat and spelt were important crops in central and northern Europe. On the other hand, hexaploid free-threshing wheats such as Indian dwarf wheat and *T. compactum* are more commonly found in warm environments. Studying changes in allele frequencies with the spread of these crops into new environments would identify candidate adaptive regions, whose phenotypic effects and usefulness could be analysed through crossing and genetic mapping. Learning from the phylogenetic relationship between ancient wheat specimens would greatly increase the power to detect the genomic regions conferring adaptation to those traits.

Furthermore, besides the impact that archaegenomics has on our understanding of the past, it has also the potential to set the basis for future food security (Pont et al. 2019b), conservation and breeding strategies, in the current context of climate change (di Donato et al. 2018). During the dispersal of domestic plants, crops adapted to a multitude of environments,

and aDNA can reveal genetic diversity present in historical landraces but lost from the modern domestic pool (e.g. Scott et al. 2019). Detecting signals of positive selection in such lost diversity may therefore be particularly valuable, especially when it is the source of adaptations to extreme environments. After its identification, such diversity can be prioritized for preservation or introduced to modern cultivars via breeding if still present in seed banks, landraces, or wild relatives (di Donato et al. 2018). Plant aDNA studies can lead to the identification of lost crops and their wild relatives, revealing their genetic makeup. Such knowledge could set the ground for de novo domestications and ultimately aid in the diversification of our food system, which currently relies on a rather small number of domestic species (Estrada et al. 2018). Finally, aDNA can be informative of past plant-pathogens interactions and their co-evolution, e.g. (Yoshida et al. 2013), providing valuable insights for crop management (di Donato et al. 2018; Estrada et al. 2018; Przelomska et al. 2020).

In conclusion, archaeogenomics allows interrogation of a plethora of questions about wheat evolutionary history, such as population continuity and demographic changes through time, identification of climatic or cultural conditions that correspond to germplasm shifts, and relationships with other wheats. We expect these questions to be addressed in future aDNA studies. Overall, answering these questions will not only bring a deeper understanding of wheat evolution, but will also aid answering questions about human cultural evolution and trade.

Acknowledgements A. I. is a FPI fellow (PRE2018-083529). L. B. is a Ramón y Cajal Fellow. (RYC2018-024770-I) both fellowships funded by the Ministerio de Ciencia e Innovación—Agencia Estatal de Investigación/ Fondo Social Europeo. We acknowledge financial support from the Spanish Agencia Estatal de Investigación (Ministry of Science and Innovation-State Research Agency) (AEI), through the "Severo Ochoa Programme for Centres of Excellence in R&D" SEV-2015-0533 and CEX2019-000902-S. This work was also supported by the CERCA programme by the Generalitat de Catalunya.

M. F. S. is supported by a Leverhulme Trust Early Career Fellowship (ECF-2020-095).

References

Abbo S, Pinhasi van-Oss R, Gopher A, Saranga Y, Ofner I, Peleg Z (2014) Plant domestication versus crop evolution: a conceptual framework for cereals and grain legumes. Trends Plant Sci 19(6):351–360. https://doi.org/10.1016/j.tplants.2013.12.002

Allaby RG, Banerjee M, Brown TA (1999) Evolution of the high molecular weight glutenin loci of the A, B, D, and G genomes of wheat. Genome 42(2):296–307. https://doi.org/10.1139/g98-114

Alonge M, Shumate A, Puiu D, Zimin A, Salzberg SL (2020) Chromosome-scale assembly of the bread wheat genome reveals thousands of additional gene copies. Genetics. https://doi.org/10.1534/genetics.120.303501

Appels R, Eversole K, Feuillet C, Keller B, Rogers J, Stein N, Pozniak CJ, Choulet F, Distelfeld A, Poland J, Ronen G, Barad O, Baruch K, Keeble-Gagnère G, Mascher M, Ben-Zvi G, Josselin AA, Himmelbach A, Balfourier F et al (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science 361(6403):1–163. https://doi.org/10.1126/science.aar7191

Arranz-Otaegui A, Colledge S, Zapata L, Teira-Mayolini LC, Ibáñez JJ (2016) Regional diversity on the timing for the initial appearance of cereal cultivation and domestication in southwest Asia. Proc Natl Acad Sci USA 113(49):14001–14006. https://doi.org/10.1073/pnas.1612797113

Arranz-Otaegui A, González Carretero L, Roe J, Richter T (2018) "Founder crops" v. wild plants: assessing the plant-based diet of the last hunter-gatherers in southwest Asia. Quatern Sci Rev 186:263–283. https://doi.org/10.1016/j.quascirev.2018.02.011

Ávila-Arcos MC, Cappellini E, Romero-Navarro JA, Wales N, Moreno-Mayar JV, Rasmussen M, Fordyce SL, Montiel R, Vielle-Calzada JP, Willerslev E, Gilbert MTP (2011) Application and comparison of large-scale solution-based DNA capture-enrichment methods on ancient DNA. Sci Rep 1. https://doi.org/10.1038/srep00074

Avni R, Nave M, Barad O, Baruch K, Twardziok SO, Gundlach H, Hale I, Mascher M, Spannagl M, Wiebe K, Jordan KW, Golan G, Deek J, Ben-zvi B, Ben-zvi G, Himmelbach A, Maclachlan RP, Sharpe AG, Komatsuda T et al (2017) Wild emmer genome

architecture and diversity elucidate wheat evolution and domestication. Science 97(July):93–97. https://doi.org/10.1126/science.aan0032

Bilgic H, Hakki EE, Pandey A, Khan MK, Akkaya MS (2016) Ancient DNA from 8400 year-old Çatalhöyük wheat: implications for the origin of Neolithic agriculture. PLoS ONE 11(3):e0151974. https://doi.org/10.1371/journal.pone.0151974

Blatter RHE, Jacomet S, Schlumbaum A (2002) Spelt-specific alleles in HMW glutenin genes from modern and historical European spelt (*Triticum spelta* L.). Theor Appl Genet 104(2–3):329–337. https://doi.org/10.1007/s001220100680

Botigue LR, Song S, Scheu A, Gopalan S, Pendleton AL, Zeeb-lanz A, Arbogast R, Oetjens M, Taravella AM, Serege T, Burger J, Kidd JM, Veeramah KR, Bobo D, Daly K, Unterla M (2017) Ancient European dog genomes reveal continuity since the Early Neolithic. Nat Commun. https://doi.org/10.1038/ncomms16082

Briggs AW, Stenzel U, Johnson PLF, Pääbo S (2007) Patterns of damage in genomic DNA sequences from a Neandertal. Proc Natl Acad Sci 104:14616–14621. https://doi.org/10.1073/pnas.0704665104

Briggs AW, Stenzel U, Meyer M, Krause J, Kircher M, Pääbo S (2010) Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. Nucleic Acids Res 38. https://doi.org/10.1093/nar/gkp1163

Cheng H, Liu J, Wen J, Nie X, Xu L, Chen N, Li Z, Wang Q, Zheng Z, Li M, Cui L, Liu Z, Bian J, Wang Z, Xu S, Yang Q, Appels R, Han D, Song W et al (2019) Frequent intra- and inter-species introgression shapes the landscape of genetic variation in bread wheat. Genome Biol 20(1):1–16. https://doi.org/10.1186/s13059-019-1744-x

Civáň P, Ivaničová Z, Brown TA (2013) Reticulated origin of domesticated emmer wheat supports a dynamic model for the emergence of agriculture in the fertile crescent. PLoS ONE 8(11):1–11. https://doi.org/10.1371/journal.pone.0081955

Czajkowska BI, Bogaard A, Charles M, Jones G, Kohler-Schneider M, Mueller-Bieniek A, Brown TA (2020) Ancient DNA typing indicates that the "new" glume wheat of early Eurasian agriculture is a cultivated member of the *Triticum timopheevii* group. J Archaeol Sci 123(October):105258. https://doi.org/10.1016/j.jas.2020.105258

Da Fonseca RR, Smith BD, Wales N, Cappellini E, Skoglund P, Fumagalli M, Samaniego JA, Carøe C, Ávila-Arcos MC, Hufnagel DE, Korneliussen TS, Vieira FG, Jakobsson M, Arriaza B, Willerslev E, Nielsen R, Hufford MB, Albrechtsen A, Ross-Ibarra J, Gilbert MTP (2015) The origin and evolution of maize in the Southwestern United States. Nat Plants 1(January):1–5. https://doi.org/10.1038/nplants.2014.3

Daly KG, Delser PM, Mullin VE, Scheu A, Mattiangeli V, Teasdale MD, Hare AJ, Burger J, Verdugo MP, Collins MJ, Kehati R, Erek CM (2018) Ancient goat genomes reveal mosaic domestication in the Fertile Crescent. Science 361: 85–88, 24–27. https://doi.org/10.1126/science.aas9411

Der Sarkissian C, Allentoft ME, Avila-Arcos MC, Barnett R, Campos PF, Cappellini E, Ermini L, Fernandez R, da Fonseca R, Ginolhac A, Hansen AJ, Jonsson H, Korneliussen T, Margaryan A, Martin MD, Moreno-Mayar JV, Raghavan M, Rasmussen M, Velasco MS et al (2014) Ancient genomics. Philos Trans R Soc B: Biol Sci 370(1660):20130387–20130387. https://doi.org/10.1098/rstb.2013.0387

di Donato A, Filippone E, Ercolano MR, Frusciante L (2018) Genome sequencing of ancient plant remains: findings, uses and potential applications for the study and improvement of modern crops. Front Plant Sci 9(April). https://doi.org/10.3389/fpls.2018.00441

Diamond J (2002) Evolution, consequences and future of plant and animal domestication. Nature 418(6898):700–707. https://doi.org/10.1038/nature01019

Estrada O, Breen J, Richards SM, Cooper A (2018) Ancient plant DNA in the genomic era. Nat Plants 4(7):394–396. https://doi.org/10.1038/s41477-018-0187-9

Feldman M, Kislev EM (2007) Domestication of emmer wheat and evolution of free-threshing tetraploid wheat. Isr J Plant Sci 55:207–221

Fernández E, Thaw S, Brown TA, Arroyo-Pardo E, Buxò R, Serret MD, Araus JL (2013) DNA analysis in charred grains of naked wheat from several archaeological sites in Spain. J Archaeol Sci 659–670. https://doi.org/10.1016/j.jas.2012.07.014

Filipovic D (2014) Southwest Asian founder- and other crops at Neolithic sites in Serbia. Bul E-J Archaeol 4(2):229–247

Fornaciari R, Fornaciari S, Francia E, Mercuri AM, Arru L (2018) Panicum spikelets from the Early Holocene *Takarkori rockshelter* (SW Libya): Archaeomolecular and -botanical investigations. Plant Biosyst 152(1):1–13. https://doi.org/10.1080/11263504.2016.1244117

Frantz LAF, Haile J, Lin AT, Scheu A, Benecke N, Alexander M, Linderholm A, Mullin VE, Daly KG, Battista VM, Price M, Gron KJ, Alexandri P, Arbogast R, Adrian B, Barnett R, Bartosiewicz L, Baryshnikov G, Bonsall C et al (2019) Ancient pigs reveal a near-complete genomic turnover following their introduction to Europe. Pnas 116. https://doi.org/10.1073/pnas.2008793117

Fuller DQ, Allaby RG, Stevens C (2010) Domestication as innovation: the entanglement of techniques, technology and chance in the domestication of cereal crops. World Archaeol 42(1):13–28. https://doi.org/10.1080/00438240903429680

Fuller DQ, Denham T, Arroyo-Kalin M, Lucas L, Stevens CJ, Qin L, Allaby RG, Purugganan MD (2014) Convergent evolution and parallelism in plant domestication revealed by an expanding archaeological record. Proc Natl Acad Sci USA 111(17):6147–6152. https://doi.org/10.1073/pnas.1308937110

Fulton TL (2012) Setting up an ancient DNA laboratory. In: Shapiro B, Hofreiter M (eds) Ancient DNA: methods and protocols. pp 1–11. https://doi.org/10.1007/978-1-61779-516-9_1

Gardiner LJ, Brabbs T, Akhunov A, Jordan K, Budak H, Richmond T, Singh S, Catchpole L, Akhunov E, Hall A (2019) Integrating genomic resources to present full gene and putative promoter capture probe sets for bread wheat. GigaScience 8(4):1–13. https://doi.org/10.1093/gigascience/giz018

Gaunitz C, Fages A, Hanghøj K, Albrechtsen A, Khan N, Schubert M, Seguin-orlando A, Owens IJ, Felkel S, Bignon-lau O, Damgaard PDB, Mittnik A, Wallner B, Merz V, Merz I, Zaibert V, Willerslev E (2018) Ancient genomes revisit the ancestry of domestic and Przewalski's horses. 1–5

Golenberg EM (1988) Outcrossing rates and their relationship to phenology in *Triticum dicoccoides*. Theoret Appl Genetics 75:937–944. https://doi.org/10.1007/BF00258057

Günther T, Nettelblad C (2019) The presence and impact of reference bias on population genomic studies of prehistoric human populations. PLoS Genet 15(7):1–20. https://doi.org/10.1371/journal.pgen.1008302

Gutaker RM, Burbano HA (2017) Reinforcing plant evolutionary genomics using ancient DNA. Curr Opin Plant Biol 36:38–45. https://doi.org/10.1016/j.pbi.2017.01.002

Gutaker RM, Weiß CL, Ellis D, Anglin NL, Knapp S, Fernández-alonso JL, Prat S, Burbano HA (2019) The origins and adaptation of European potatoes reconstructed from historical genomes. Nat Ecol Evol 3(July). https://doi.org/10.1038/s41559-019-0921-3

Haas M, Schreiber M, Mascher M (2018) Domestication and crop evolution of wheat and barley: genes, genomics, and future directions. J Integr Plant Biol XXXX(Xxxx). https://doi.org/10.1111/jipb.12737

He F, Pasam R, Shi F, Kant S, Keeble-Gagnere G, Kay P, Forrest K, Fritz A, Hucl P, Wiebe K, Knox R, Cuthbert R, Pozniak C, Akhunova A, Morrell PL, Davies JP, Webb SR, Spangenberg G, Hayes B et al (2019) Exome sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome. Nat Genet 51(5):896–904. https://doi.org/10.1038/s41588-019-0382-2

Higuchi R, Bowman B, Freiberger M, Ryder OA, Wilson AC (1984) DNA sequences from the quagga, an extinct member of the horse family. Nature 312:282–284

Hofreiter M, Paijmans JLA, Goodchild H, Speller CF, Barlow A, Fortes GG, Thomas JA, Ludwig A, Collins MJ (2015) The future of ancient DNA: technical advances and conceptual shifts. BioEssays 37(3):284–293. https://doi.org/10.1002/bies.201400160

Jones G, Valamoti S, Charles M (2000) Early crop diversity: a "new" glume wheat from northern Greece. Veg Hist Archaeobot 9(3):133–146. https://doi.org/10.1007/BF01299798

Jonsson H, Korneliussen T, Margaryan A, Martin MD, Moreno-Mayar JV, Raghavan M, Rasmussen M, Velasco MS, Orlando L (2014) Ancient genomics. Phil Trans R Soc B 370. https://doi.org/10.1098/rstb.2013.0387

Jordan KW, Wang S, Lun Y, Gardiner LJ, MacLachlan R, Hucl P, Wiebe K, Wong D, Forrest KL, Sharpe AG, Sidebottom CHD, Hall N, Toomajian C, Close T, Dubcovsky J, Akhunova A, Talbert L, Bansal UK, Bariana HS et al (2015) A haplotype map of allohexaploid wheat reveals distinct patterns of selection on homoeologous genomes. Genome Biol 16(1):1–18. https://doi.org/10.1186/s13059-015-0606-4

Jorgensen C, Luo MC, Ramasamy R, Dawson M, Gill BS, Korol AB, Distelfeld A, Dvorak J (2017) A high-density genetic map of wild emmer wheat from the karaca dağ region provides new evidence on the structure and evolution of wheat chromosomes. Front Plant Sci 8(October):1–13. https://doi.org/10.3389/fpls.2017.01798

Kantar MB, Nashoba AR, Anderson JE, Blackman BK, Rieseberg LH (2017) The genetics and genomics of plant domestication. BioScience XX(X):1–12. https://doi.org/10.1093/biosci/bix114

Kaplan L, Smith MB, Sneddon LA (1992) Cereal grain phytoliths of Southwest Asia and Europe. In: Rapp G, Mulholland SC (eds) Phytolith systematics in archaeological and museum science. Springer US, pp 149–174. https://doi.org/10.1007/978-1-4899-1155-1_8

Key FM, Posth C, Esquivel-gomez LR, Hübler R, Spyrou MA, Neumann GU, Furtwängler A, Sabin S, Burri M, Wissgott A, Lankapalli AK, Vågene ÅJ, Meyer M, Nagel S, Tukhbatova R, Khokhlov A, Chizhevsky A, Hansen S, Belinsky AB et al (2020) Emergence of human-adapted *Salmonella enterica* is linked to the Neolithization process. Nat Ecol Evol 4(March). https://doi.org/10.1038/s41559-020-1106-9

Kistler L, Ware R, Smith O, Collins M, Allaby RG (2017) A new model for ancient DNA decay based on paleogenomic meta-analysis. Nucleic Acids Res 45:6310–6320. https://doi.org/10.1093/nar/gkx361

Kistler L, Bieker VC, Martin MD, Pedersen MW, Ramos Madrigal J, Wales N (2020) Ancient plant genomics in archaeology, herbaria, and the environment. Annu Rev Plant Biol 71:605–629. https://doi.org/10.1146/annurev-arplant-081519-035837

Korneliussen TS, Albrechtsen A, Nielsen R (2014) ANGSD: analysis of next generation sequencing data. BMC Bioinformatics 15:1–13. https://doi.org/10.1186/s12859-014-0356-4

Krause J, Fu Q, Good JM, Viola B, Shunkov MV, Derevianko AP (2010) The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. Nature 464(April):894–897. https://doi.org/10.1038/nature08976

Lacan M, Keyser C, Ricaut F, Brucato N, Duranthon F, Guilaine J (2011) Ancient DNA reveals male diffusion through the Neolithic Mediterranean

route.https://doi.org/10.1073/pnas.1100723108/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1100723108

Larson G, Piperno DR, Allaby RG, Purugganan MD, Andersson L, Arroyo-Kalin M, Barton L, Climer Vigueira C, Denham T, Dobney K et al (2014) Currrent perspectives and the future of domestication studies. Proc Natl Acad Sci 111(17):6139–6146. https://doi.org/10.1073/pnas.1323964111

Lawson DJ, van Dorp L, Falush D (2018) A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. Nat Commun 9(1):1–11. https://doi.org/10.1038/s41467-018-05257-7

Leathlobhair MN, Perri AR, Irving-pease EK, Witt KE et al (2018) The evolutionary history of dogs in the Americas. 85(July):81–85

Lev-Yadun S, Gopher A, Abbo S (2000) The cradle of agriculture. Science 361. https://doi.org/10.1126/science.aao4776

Li C, Lister DL, Li H, Xu Y, Cui Y, Bower MA, Jones MK, Zhou H (2011) Ancient DNA analysis of desiccated wheat grains excavated from a Bronze Age cemetery in Xinjiang. J Archaeol Sci 38(1):115–119. https://doi.org/10.1016/j.jas.2010.08.016

Librado P, Khan N, Fages A et al (2021) The origins and spread of domestic horses from the Western Eurasian steppes. Nature 598. https://doi.org/10.1038/s41586-021-04018-9

Lindahl T (1993) Instability and decay of the primary structure of DNA. Nature 362:709–715. https://doi.org/10.1038/362709a0

Link V, Kousathanas A, Veeramah K, Sell C, Scheu A, Wegmann D (2017) ATLAS: analysis tools for low-depth and ancient samples. BioRxiv 33(16):1–7

Ling HQ, Ma B, Shi X, Liu H, Dong L, Sun H, Cao Y, Gao Q, Zheng S, Li Y, Yu Y, Du H, Qi M, Li Y, Lu H, Yu H, Cui Y, Wang N, Chen C et al (2018) Genome sequence of the progenitor of wheat A subgenome *Triticum urartu*. Nature 557(7705):424–428. https://doi.org/10.1038/s41586-018-0108-0

Luo MC, Yang ZL, You FM, Kawahara T, Waines JG, Dvorak J (2007) The structure of wild and domesticated emmer wheat populations, gene flow between them, and the site of emmer domestication. Theor Appl Genet 114(6):947–959. https://doi.org/10.1007/s00122-006-0474-0

Luo MC, Gu YQ, Puiu D, Wang H, Twardziok SO, Deal KR, Huo N, Zhu T, Wang L, Wang Y, McGuire PE, Liu S, Long H, Ramasamy RK, Rodriguez JC, Van Sonny L, Yuan L, Wang Z, Xia Z et al (2017) Genome sequence of the progenitor of the wheat D genome *Aegilops tauschii*. Nature 551(7681):498–502. https://doi.org/10.1038/nature24486

Maccaferri M, Harris NS, Twardziok SO, Pasam RK, Gundlach H, Spannagl M, Ormanbekova D, Lux T, Prade VM, Milner SG, Himmelbach A, Mascher M, Bagnaresi P, Faccioli P, Cozzi P, Lauria M, Lazzari B, Stella A, Manconi A et al (2019) Durum wheat genome highlights past domestication signatures and future improvement targets. Nat Genet 51(5):885–895.https://doi.org/10.1038/s41588-019-0381-3

Marciniak S, Perry GH (2017) Harnessing ancient genomes to study the history of human adaptation. Nat Publ Group 18(11):659–674. https://doi.org/10.1038/nrg.2017.65

Mascher M, Schuenemann VJ, Davidovich U, Marom N, Himmelbach A, Hübner S, Korol A, David M, Reiter E, Riehl S, Schreiber M, Vohr SH, Green RE, Dawson IK, Russell J, Kilian B, Muehlbauer GJ, Waugh R, Fahima T et al (2016). Genomic analysis of 6000-year-old cultivated grain illuminates the domestication history of barley. Nat Genet 48(9):1089–1093. https://doi.org/10.1038/ng.3611

McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, Ripatti S (2016) A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet 48:1279–1283

Morozova I, Flegontov P, Mikheyev AS, Bruskin S, Asgharian H, Ponomarenko P, Klyuchnikov V, Kumar GPA, Prokhortchouk E, Gankin Y, Rogaev E, Nikolsky Y, Baranova A, Elhaik E, Tatarinova TV (2016) Toward high-resolution population genomics using archaeological samples. DNA Res 23(4):295–310. https://doi.org/10.1093/dnares/dsw029

Nave M, Avni R, Çakır E, Portnoy V, Sela H, Pourkheirandish M, Ozkan H, Hale I, Komatsuda T, Dvorak J, Distelfeld A (2019) Wheat domestication in light of haplotype analyses of the Brittle rachis 1 genes (BTR1-A and BTR1-B). Plant Sci 285(May):193–199. https://doi.org/10.1016/j.plantsci.2019.05.012

Nesbitt M, Samuel D (1996) From staple crop to extinction? The archaeology and history of the hulled wheat. In: Padulosi S, Hammer K, Heller J (eds) Hulled wheats, promoting the conservation and used of underutilized and neglected crops. pp 40–99

Nielsen R, Akey JM, Jakobsson M, Pritchard JK, Tishkoff S, Willerslev E (2017) Tracing the peopling of the world through genomics. Nature 541. https://doi.org/10.1038/nature21347

Oliveira HR, Jacocks L, Czajkowska BI, Kennedy SL, Brown TA (2020) Multiregional origins of the domesticated tetraploid wheats. PLoS ONE 15(1):1–20. https://doi.org/10.1371/journal.pone.0227148

Orlando L, Allaby R, Pontus S, Der Sarkissian C, Stockhammer PW, Ávila-Arcos MC, Fu Q, Krause J, Willerslev E, Stone A, Christina W (2021) Ancient DNA analysis. Nat Rev Genet 1–26. https://doi.org/10.1038/s43586-020-00011-0

Özbaşaran M, Duru G, Stiner MC, Esin U (2018) The early settlement at Aşıklı Höyük: essays in honor of Ufuk Esin. Ege Yayınları

Ozkan H, Brandolini A, Schafer-Pregl R, Salamini F (2002) AFLP analysis of a collection of tetraploid wheat indicated the origin of emmer and hard wheat domestication in south-eastern Turkey. Mol Biol Evol 19:1797–1801

Ozkan H, Brandolini A, Pozzi C, Effgen S, Wunder J, Salamini F (2005) A reconsideration of the domestication geography of tetraploid wheats. Theor Appl Genet 110:1052–1060

Palmer SA, Moore JD, Clapham AJ, Rose P, Allaby RG (2009) Archaeogenetic evidence of ancient nubian barley evolution from six to two-row indicates local adaptation. PLoS ONE 4(7):2–8. https://doi.org/10.1371/journal.pone.0006301

Palmer SA, Clapham AJ, Rose P, Freitas FO, Owen BD, Beresford-Jones D, Moore JD, Kitchen JL, Allaby RG (2012) Archaeogenomic evidence of punctuated genome evolution in gossypium. Mol Biol Evol 29(8):2031–2038. https://doi.org/10.1093/molbev/mss070

Parducci L, Bennett KD, Ficetola GF, Alsos IG, Suyama Y, Wood JR, Pedersen MW (2017) Ancient plant DNA in lake sediments. New Phytol 214(3):924–942. https://doi.org/10.1111/nph.14470

Pedersen JS, Valen E, Velazquez AMV, Parker BJ, Rasmussen M, Lindgreen S, Lilje B, Tobin DJ, Kelly TK, Vang S, Andersson R, Jones PA, Hoover CA, Tikhonov A, Prokhortchouk E, Rubin EM, Sandelin A, Gilbert MTP, Krogh A et al (2014) Genome-wide nucleosome map and cytosine methylation levels of an ancient human genome 454–466. https://doi.org/10.1101/gr.163592.113.Freely

Peltzer A, Jäger G, Herbig A, Seitz A, Kniep C, Krause J, Nieselt K (2016) EAGER: efficient ancient genome reconstruction. Genome Biol 17:60. https://doi.org/10.1186/s13059-016-0918-z

Peyrégne S, Prüfer K (2020) Present-day DNA contamination in ancient DNA datasets. BioEssays 42(1–11). https://doi.org/10.1002/bies.202000081

Pont C, Leroy T, Seidel M, Tondelli A, Duchemin W, Armisen D, Lang D, Bustos-Korts D, Goué N, Balfourier F, Molnár-Láng M, Lage J, Kilian B, Özkan H, Waite D, Dyer S, Letellier T, Alaux M, Russell J et al (2019a) Tracing the ancestry of modern bread wheats. Nat Genet 51(5):905–911. https://doi.org/10.1038/s41588-019-0393-z

Pont C, Wagner S, Kremer A, Orlando L, Plomion C, Salse J (2019b) Paleogenomics: reconstruction of plant evolutionary trajectories from modern and ancient DNA. Genome Biol 20(1):29. https://doi.org/10.1186/s13059-019-1627-1

Prüfer K, Meyer M (2015) Comment on "Late Pleistocene human skeleton and mtDNA link Paleoamericans and modern Native Americans." Science 347(6224):835. https://doi.org/10.1126/science.1260617

Przelomska NAS, Armstrong CG, Kistler L (2020) Ancient plant DNA as a window into the cultural heritage and biodiversity of our food system. Front Ecol Evol 8(March):1–8. https://doi.org/10.3389/fevo.2020.00074

Przewieslik-Allen AM, Wilkinson PA, Burridge AJ, Winfield MO, Dai X, Beaumont M, King J, Yang C, Griffiths S, Wingen LU, Horsnell R, Bentley AR, Shewry P, Barker GLA, Edwards KJ (2021) The role of gene flow and chromosomal instability in shaping the bread wheat genome. Nat Plants 7(2):172–183. https://doi.org/10.1038/s41477-020-00845-2

Purugganan MD, Fuller DQ (2009) The nature of selection during plant domestication. Nature 457(7231):843–848. https://doi.org/10.1038/nature07895

Ramos-Madrigal J, Smith BD, Moreno-Mayar JV, Gopalakrishnan S, Ross-Ibarra J, Gilbert MTP, Wales N (2016) Genome sequence of a 5310-year-old maize cob provides insights into the early stages of maize domestication. Curr Biol 26(23):3195–3201. https://doi.org/10.1016/j.cub.2016.09.036

Ramos-Madrigal J, Runge AKW, Bouby L, Lacombe T, Samaniego Castruita JA, Adam-Blondon AF, Figueiral I, Hallavant C, Martínez-Zapater JM, Schaal C, Töpfer R, Petersen B, Sicheritz-Pontén T, This P, Bacilieri R, Gilbert MTP, Wales N (2019) Palaeogenomic insights into the origins of French grapevine diversity. Nat Plants 5(6):595–603. https://doi.org/10.1038/s41477-019-0437-5

Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PLF, Maricic T, Good JM, Marques-bonet T, Alkan C, Fu Q, Mallick S, Li H, Meyer M, Eichler EE, Stoneking M (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. https://doi.org/10.1038/nature09710

Renaud G, Schubert M, Sawyer S, Orlando L (2019) Authentication and assessment of contamination in ancient DNA. In: Shapiro B et al (eds) Ancient DNA: methods and protocols. Springer US, pp 163–194. https://doi.org/10.2307/j.ctt183pd9z.20

Renner SS, Pérez-Escbar OA, Silber MV, Nesbitt M, Preick M, Hofreiter M, Chomicki G (2019) A 3500-year-old leaf from a Pharaonic tomb reveals that New Kingdom Egyptians were cultivating domesticated watermelon. BioRxiv. https://doi.org/10.1101/642785

Rohland N, Harney E, Mallick S, Nordenfelt S, Reich D (2015) Partial uracil–DNA–glycosylase treatment for screening of ancient DNA. Phil Trans R Soc B 370(1660). https://doi.org/10.1098/rstb.2013.0624

Schubert M, Ermini L, Sarkissian CD, Jónsson H, Ginolhac A, Schaefer R, Martin MD, Fernández R, Kircher M, McCue M, Willerslev E, Orlando L (2014) Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. Nat Protoc 9(5):1056–1082. https://doi.org/10.1038/nprot.2014.063

Scott MF, Botigué LR, Brace S, Stevens CJ, Mullin VE, Stevenson A, Thomas MG, Fuller DQ, Mott R (2019) A 3000-year-old Egyptian emmer wheat genome reveals dispersal and domestication history. Nat Plants 5(11):1120–1128. https://doi.org/10.1038/s41477-019-0534-5

Sharma JS, Running KLD, Xu SS, Zhang Q, Peters Haugrud AR, Sharma S, McClean PE, Faris JD (2019) Genetic analysis of threshability and other

spike traits in the evolution of cultivated emmer to fully domesticated durum wheat. Mol Genet Genomics 294(3):757–771. https://doi.org/10.1007/s00438-019-01544-0

Smith O, Nicholson WV, Kistler L, Mace E, Clapham A, Rose P, Stevens C, Ware R, Samavedam S, Barker G, Jordan D, Fuller DQ, Allaby RG (2019) A domestication history of dynamic adaptation and genomic deterioration in Sorghum. Nat Plants 5(April). https://doi.org/10.1038/s41477-019-0397-9

Snir A, Nadel D, Groman-Yaroslavski I, Melamed Y, Sternberg M, Bar-Yosef O, Weiss E (2015) The origin of cultivation and proto-weeds, long before neolithic farming. PLoS ONE 10(7):1–12. https://doi.org/10.1371/journal.pone.0131422

Spyrou MA, Bos KI, Herbig A, Krause J (2019) Ancient pathogen genomics as an emerging tool for infectious disease research. Nat Rev Genet 20(June). https://doi.org/10.1038/s41576-019-0119-1

Swarts K, Gutaker RM, Benz B, Blake M, Bukowski R, Holland J, Kruse-peeples M, Lepak N, Prim L, Romay MC, Ross-ibarra J, Sanchez-gonzalez JDJ (2017) Genomic estimation of complex traits reveals ancient maize adaptation to temperate North America. 515(August):512–515. https://doi.org/10.1126/science.aam9425Estimating

Trucchi E, Benazzo A, Lari M, Iob A, Vai S, Nanni L, Bellucci E, Bitocchi E, Raffini F, Xu C, Jackson SA, Lema V, Babot P, Oliszewski N, Gil A, Neme G, Michieli CT, De Lorenzi M, Calcagnile L et al (2021) Ancient genomes reveal early Andean farmers selected common beans while preserving diversity. Nat Plants 7(2):123–128. https://doi.org/10.1038/s41477-021-00848-7

Ulaş B, Fiorentino G (2021) Recent attestations of "new " glume wheat in Turkey: a reassessment of its role in the reconstruction of Neolithic agriculture. Veg Hist Archaeobotany 30(5):685–701. https://doi.org/10.1007/s00334-020-00807-w

Vavilov NI, Vavylov MI, Vavílov NÍ, Dorofeev VF (1992) Origin and geography of cultivated plants. Cambridge University of Press

Verdugo MP, Mullin VE, Scheu A, Mattiangeli V, Daly KG, Delser PM, Hare AJ, Burger J, Collins MJ, Kehati R, Hesse P, Fulton D (2019) Ancient cattle genomics, origins, and rapid turnover in the Fertile Crescent. 176(July):173–176. https://doi.org/10.5281/zenodo.3206663

Wagenaar EB (1966) Studies on the genome constitution of *Triticum timopheevi* Zhuk. II. The *T. timopheevi* complex and its origin. Soc Study Evol 20(2):150–164. https://www.jstor.org/stable/2406569

Wales N, Akman M, Watson RHB, Sánchez Barreiro F, Smith BD, Gremillion KJ, Gilbert MTP, Blackman BK (2018) Ancient DNA reveals the timing and persistence of organellar genetic bottlenecks over 3000 years of sunflower domestication and improvement. Evol Appl 12(December 2017):1–16. https://doi.org/10.1111/eva.12594

Wales N, Akman M, Watson RHB, Sánchez Barreiro F, Smith BD, Gremillion KJ, Gilbert MTP, Blackman BK (2019) Ancient DNA reveals the timing and persistence of organellar genetic bottlenecks over 3000 years of sunflower domestication and improvement. Evol Appl 12(1):38–53. https://doi.org/10.1111/eva.12594

Walkowiak S, Gao L, Monat C, Haberer G, Kassa MT, Brinton J, Ramirez-Gonzalez RH, Kolodziej MC, Delorean E, Thambugala D, Klymiuk V, Byrns B, Gundlach H, Bandi V, Siri JN, Nilsen K, Aquino C, Himmelbach A, Copetti D et al (2020) Multiple wheat genomes reveal global variation in modern breeding. Nature 588(7837):277–283. https://doi.org/10.1038/s41586-020-2961-x

Warinner C, Rodrigues JFM, Vyas R, Trachsel C, Shved N, Grossmann J, Radini A, Hancock Y, Tito RY, Fiddyment S, Speller C, Hendy J, Charlton S, Luder HU, Salazar-garcía DC, Eppler E, Seiler R, Hansen LH, Alfredo J et al (2014) Pathogens and host immunity in the ancient human oral cavity. Nat Genet 46(4). https://doi.org/10.1038/ng.2906

Weyrich LS, Dobney K, Cooper A (2015) Ancient DNA analysis of dental calculus. J Hum Evol 79:119–124. https://doi.org/10.1016/j.jhevol.2014.06.018

Wu X, Ding B, Zhang B, Feng J, Wang Y, Ning C, Wu H, Zhang F, Zhang Q, Li N, Zhang Z, Sun X, Zhang Q, Li W, Liu B, Cui Y, Gong L (2019) Phylogenetic and population structural inference from genomic ancestry maintained in present-day common wheat Chinese landraces. Plant J 99(2):201–215. https://doi.org/10.1111/tpj.14421

Yang MA, Fan X, Sun B, Chen C, Lang J, Ko Y, Tsang C, Chiu H, Wang T, Bao Q, Wu X, Hajdinjak M, Ko AM, Ding M, Cao P, Yang R, Liu F, Nickel B, Dai Q et al (2020) Ancient DNA indicates human population shifts and admixture in northern and southern China. Science 288(July):282–288. https://doi.org/10.1126/science.aba0909

Yoshida K, Schuenemann VJ, Cano LM, Pais M, Mishra B, Sharma R et al (2013) The rise and fall of the Phytophthora infestans lineage that triggered the Irish potato famine. Elife 2:1–25. https://doi.org/10.7554/eLife.00731.001

Zhou Y, Zhao X, Li Y, Xu J, Bi A, Kang L, Xu D, Chen H, Wang Y, Wang Y-G, Liu S, Jiao C, Lu H, Wang J, Yin C, Jiao Y, Lu F (2020) Triticum population sequencing provides insights into wheat adaptation. Nat Genet 52(12):1412–1422. https://doi.org/10.1038/s41588-020-00722-w

Zohary D (2013) Domestication of Crop Plants. In: Encyclopedia of biodiversity, 2nd edn. pp 657–664. https://doi.org/10.1016/B978-0-12-384719-5.00199-4

Zhu T, Wang L, Rodriguez JC, Deal KR, Avni R, Distelfeld A, McGuire PE, Dvorak J, Luo MC (2019) Improved genome sequence of wild emmer wheat Zavitan with the aid of optical maps. G3: Genes, Genomes, Genet 9(3):619–624. https://doi.org/10.1534/g3.118.200902

8 Gene Flow Between Tetraploid and Hexaploid Wheat for Breeding Innovation

Elisabetta Mazzucotelli, Anna Maria Mastrangelo, Francesca Desiderio, Delfina Barabaschi, Marco Maccaferri, Roberto Tuberosa and Luigi Cattivelli

Abstract

Durum and bread wheat are two related species with different ploidy levels but a high similarity between the common A and B genomes. This feature, which allows a continuous gene flow between the two species, can be exploited in breeding programs to improve key traits in both crops. Therefore, durum wheat, despite covering only 5% of cultivated wheat worldwide, also represents an asset for the genetic improvement of bread wheat. Tetraploid wheat, with a very large availability of wild and domesticated accessions, durum landraces, and cultivars, offers a large gene reservoir to increase the genetic diversity of A and B genomes in bread wheat. Moreover, thanks to the possibility of crossing durum wheat with *Aegilops tauschii*, synthetic hexaploid lines are generated which show a much larger genetic diversity also in the D genome compared to common wheat. The genome sequences of wild emmer, durum, and bread wheat provide power tools for gene cloning and comparative genomics that will also facilitate the shuttling of genes between tetraploid and hexaploid wheats.

Keywords

Tetraploid · Synthetic wheat · Gene flow · Selection signatures · Wild germplasm

E. Mazzucotelli · F. Desiderio · D. Barabaschi · L. Cattivelli (✉)
CREA, Research Centre for Genomics and Bioinformatics, Via San Protaso, 302, 29017 Fiorenzuola d'Arda, Italy
e-mail: luigi.cattivelli@crea.gov.it

E. Mazzucotelli
e-mail: elisabetta.mazzucotelli@crea.gov.it

F. Desiderio
e-mail: francesca.desiderio@crea.gov.it

D. Barabaschi
e-mail: delfina.barabaschi@crea.gov.it

A. M. Mastrangelo
CREA, Research Centre for Cereal and Industrial, SS 673, 71122 Foggia, Italy
e-mail: annamaria.mastrangelo@crea.gov.it

M. Maccaferri · R. Tuberosa
Department of Agricultural and Food Sciences, University of Bologna, Bologna, Italy
e-mail: marco.maccaferri@unibo.it

R. Tuberosa
e-mail: roberto.tuberosa@unibo.it

8.1 Introduction

Durum wheat (tetraploid) and bread wheat (hexaploid) are two closely related species with potentially different adaptation capacities and only a few distinct technological properties that make durum semolina and wheat flour more suitable for pasta or bread and bakery products, respectively (Mastrangelo and Cattivelli 2021).

R. Appels et al. (eds.), *The Wheat Genome*, Compendium of Plant Genomes,
https://doi.org/10.1007/978-3-031-38294-9_8

The history of wheat began with the domestication of wild emmer wheat (WEW, *Triticum turgidum* ssp. *dicoccoides)* in the mountains of the Fertile Crescent around 12–10 thousand years ago, which gave rise to the first domesticated form (domesticated emmer wheat, DEW, *T. turgidum* ssp. *dicoccum*) and a first domestication sweep related to *Brittle rachis (Btr)* trait. Then, human selection of natural mutations at a few loci associated with the domestication syndrome (i.e., *Tg tenacious glume* and *Q compact spike*) allowed for the selection of wheat forms with square-shaped spikes, soft glumes, and non-hulled grains improving with an improved threshing efficiency, grain size and uniformity, productivity, and suitable for a more widespread cultivation. This phenotypic evolution together with hybridization between different forms (Matsuoka 2011) led to free-threshing subspecies (*T. turgidum* ssp. *turgidum,* ssp. *turanicum,* ssp. *polonicum,* ssp. *carthlicum,* and ssp. *durum,* Fig. 8.1), all inter-fertile and sharing the same AABB genomic configuration. Hulled and free-threshing forms played a crucial role in the development of Mediterranean civilizations. They were at the base of early agricultural movements leading to agriculture systems based on tetraploid wheat. Among the different subspecies, durum wheat (*T. turgidum* ssp. *durum*) became the major cultivated form of tetraploid wheat during the last 3000 years. Nowadays, élite durum wheat cultivars (DWCs) and durum wheat landraces (DWLs) grow in different environments around the Mediterranean Basin and are of major importance for grain production and for staple food, respectively (Fig. 8.2).

The expansion of emmer cultivation toward the Transcaucasian corridor promoted an additional natural hybridization of tetraploid forms with *Aegilops tauschii* (genome DD) and the emergence of the hexaploid bread wheat (*T. aestivum L.* ssp. *aestivum,* genome AABBDD) (Dubcoksky and Dvorak 2007). As a result, durum and bread wheat share the A and B genomes and a long evolutionary history.

8.2 Tetraploid Genetic Resources

Most of wheat genetic diversity is contributed by tetraploid wheat genetic resources, particularly primitive tetraploids and wild and domesticated emmer. Indeed, the bottleneck effect, caused by the evolutionary recent hybridization events from which hexaploid wheat has evolved, has strongly limited its genetic diversity compared to tetraploid and diploid wheats (Cox 1997). Therefore, tetraploid wheat germplasm

Fig. 8.1 Examples of spikes of some subspecies belonging to the species *Triticum turgidum*

Fig. 8.2 Morphological and color variability for spikes of *T. turgidum* ssp *durum* cultivars

represents a strategic reservoir of alleles for both durum and bread wheat improvement (Marone et al. 2021). The International Maize and Wheat Improvement Center (CIMMYT) and the International Center for Agricultural Research in the Dry Areas (ICARDA) have the largest collections of tetraploid wheat, with approximately 27,500 and 22,500 accessions, respectively, including 22,000 and 20,000 durum wheat accessions. The ICARDA gene bank stores more than 15,700 accessions of DWL and traditional cultivars, while CIMMYT retains a larger collection of domesticated emmer wheat (around 3000 accessions). A wide tetraploid wheat diversity is conserved by the National Small Grains Germplasm Research Facility at USDA-ARS where approximately 12,500 accessions are conserved, with a large representation of primitive tetraploid subspecies (*T. turgidum* ssp. *polonicum,* ssp. *carthlicum,* ssp. *turanicum,* and ssp. *turgidum*) and more than 900 WEW accessions. Many national gene banks also retain important local germplasm resources which include historical materials and the predominant remaining landraces (Robbana et al. 2019).

Many studies have reported on the assembly and characterization of panels of tetraploid genotypes, from tens of genotypes to many hundreds of entries of wider origin. These germplasm collections are representative of (i) subspecies, (ii) specific geographic regions including local DWLs, historical cultivars, and modern DWCs, and (iii) breeding programs. They have been characterized for population structure, genome-wide molecular diversity, and linkage disequilibrium (LD)-decay rate as estimated with either multi-allelic (SSRs) and/or bi-allelic (DArT™, AFLPs) markers in earlier studies (Maccaferri et al. 2005; Mantovani et al. 2008; Laidò et al. 2013; Roncallo et al. 2019) or more recently with the Illumina iSelect 90K SNP array (Maccaferri et al. 2016; Saccomanno et al. 2018; N'Diaye et al. 2018) and the Axiom 35K array (Kabbaj et al. 2017). All these studies have generated an in-depth description of genetic diversity and differentiation within/among subspecies and subgroups with a focus on both temporal and spatial trends, particularly targeting the cultivated and DWL germplasm.

Following the publication of the first high-density SNP-based consensus map of tetraploid wheat (Maccaferri et al. 2015) and the first release of the durum wheat reference genome of cultivar SVEVO, many SSRs and iSelect 90K wheat SNPs have been anchored on the durum genome sequence, thus providing opportunities for genetic insights on relevant genomic regions (Maccaferri et al. 2019). Two recent collaborative studies provided germplasm panels and advanced in-depth analysis supporting a detailed knowledge at the molecular level of the historical loss of diversity events. The identification of favorable allelic combinations progressively accumulated over repeated

breeding cycles is instrumental for a more effective management of breeding. In the first study, the International Durum Wheat Sequencing Consortium supported a comprehensive analysis of genetic diversity in tetraploids which entailed the organization of the single seed descent Tetraploid Germplasm Collection (TGC) and its genetic diversity analysis using the Illumina iSelect 90K SNP array, projected onto the Svevo genome (Maccaferri et al. 2019; Fig. 8.3).

At the same time, the Wheat Initiative through the durum wheat-expert working group supported the development of the Global Durum Panel (GDP), a collection targeting mainly the cultivated durum germplasm, and fully genotyped with the Illumina iSelect 90K wheat SNP array. The GDP genetic diversity is described in Mazzucotelli et al. (2020). The two collections have been assembled, seed increased and made freely available for research with the aim to facilitate the inventory, molecular and phenotypic characterization, and use of tetraploid genetic resources for durum and bread wheat improvement. The collections are maintained at ICARDA (Morocco), University

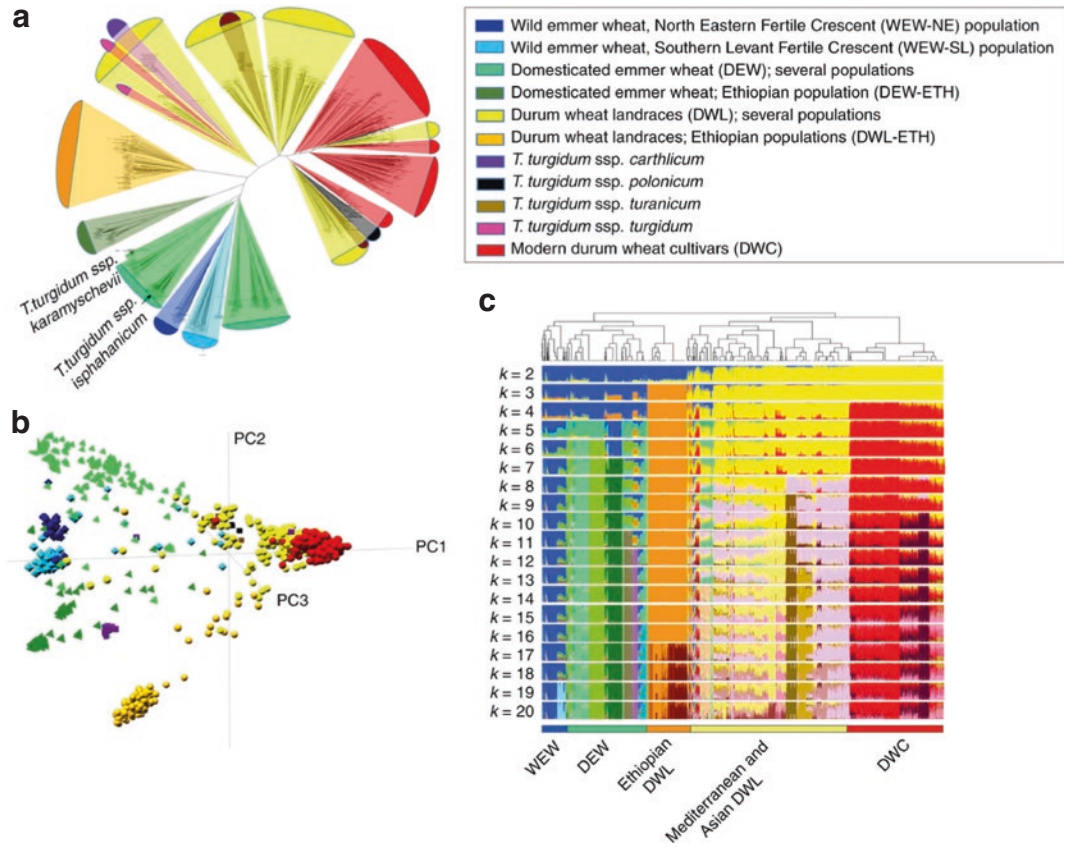


Fig. 8.3 Population structure of the Tetraploid Germplasm Collection (TGC) composed of 1856 accessions of tetraploid wheat. **a** Neighbor joining tree from Nei's genetic distances on the TGC. **b** Principal component analysis plot of the TGC calculated based on genome-wide linkage disequilibrium-pruned SNPs. **c** Admixture analyses of the TGC with *k* (number of populations assumed for the analysis) from 2 to 20. Correspondence between branches and main tetraploid wheat taxa/populations based on Nei's genetic distances, PCA and Admixture are indicated by color code (modified from Maccaferri et al. 2019). The analyses of population structure concord to highlight five major subpopulations: wild emmer wheat, domesticated emmer wheat, durum wheat landraces from Ethiopian, durum wheat landraces from Asian and Mediterranean regions, and durum wheat cultivars

of Bologna (Italy), and CREA-Research Centre for Genomics and Bioinformatics (Italy). Related information and genotypic data are accessible at the GrainGenes database (https://wheat.pw.usda.gov/GG3/global_durum_genomic_resources).

8.2.1 Wild Emmer Shows the Widest Range of Adaptation to Environment and Retains the Highest Level of Genetic Diversity Genome-Wide

WEW is an annual, predominantly self-pollinating allotetraploid species with large, elongated grains and brittle ears disarticulating at maturity into spikelets. Molecular data indicate that WEW is about 500,000 years old, resulting from a hybridization event between two wild diploid grasses that took place in the Fertile Crescent, probably in the vicinity of Mt. Hermon and the catchment area of the Jordan River where a center of WEW diversity has been reported (Dvorak and Akhunov 2005; Feldman and Kislev 2007). WEW is naturally distributed in the Near East Fertile Crescent with two major races which are geographically, morphologically, and genetically distinct: (1) the north-eastern part of the Fertile Crescent, with main populations found in north- and central-eastern Turkey, western Iran, and northern Iraq; (2) the western race found in the southern Levant, including Syria, Lebanon, Jordan, and Israel. Apart from dense and frequent natural populations found in the upper Jordan valley catchment area in Israel, and massive stands on the basalt slopes of the Karacadağ (Şanlıurfa and Diyarbakır provinces) in Turkey, WEW currently displays a patchy distribution in the region, with populations being semi-isolated or isolated. Its habitats range in altitude from 100 m below sea level up to 1800 m above sea level, with very different climatic regions from cool and humid Karacadağ Mountains to hot and dry valleys in Israel (Nevo et al. 2002).

The domestication dynamics of WEW are still unclear, though several pieces of the puzzle have been identified. In present days, the south-eastern Turkish subpopulations are more closely related to DEWs than any other wild emmer populations, but the monophyletic origin of DEW is still debated. Whole genome analysis based on multi-locus assays pointed out that the wild emmer populations from the Karacadağ region west of Diyarbakir and from the Sulaimanyia region along the Iraq/Iran border appeared the most closely related to DEW (Ozkan et al. 2011) with a further molecular indication in favor of the Diyarbakir WEW (Luo et al. 2007). Later analyses supported the reticulated origin including sharing phylogenetic signals with wild populations from all parts of the wild range (Civan et al. 2013). Recently, a study from Nave et al. (2021) focused on the *Brittle Rachis* gene (*BTR1*), a fundamental gene for wheat domestication present in the two homeologous copies *BTR1-A* and *BTR1-B*. Haplotype sequences showed that for the *BTR1-A* locus, the domestic *BTR1-A-hap11* is highly related to the WEW founder haplotype *BTR1-A-hap10* which is ubiquitous in both northern and southern Fertile Crescent, while for the *BTR1-B* copy, the domesticated haplotype *BTR1-B-hap8* was derived from the wild haplotype *BTR1-B-hap7* found only in the southern Levant. This indicated that at least part of the domestication process of WEW occurred outside of the "core area" of the northern part of the Fertile Crescent (Nave et al. 2021).

Each WEW race is genetically further subdivided in subpopulations with a pattern that mirrors the geographic origin (Luo et al. 2007; Ozkan et al. 2011; Badaeva et al. 2015; Maccaferri et al. 2019). Up to 12 well-distinct populations and subpopulations were identified by Admixture analysis in the TGC (Maccaferri et al. 2019; Fig. 8.4a). Moreover, it was shown that populations belonging to the eastern race were less diverse than those collected in the Levant. Notably, the genetic structure of the western population also correlates with differences in morphologic features. Indeed, most of the western populations belong to the *horanum* botanical variety and include accessions with a slender habit, while northern Israel is specifically inhabited by the subpopulation *judaicum*

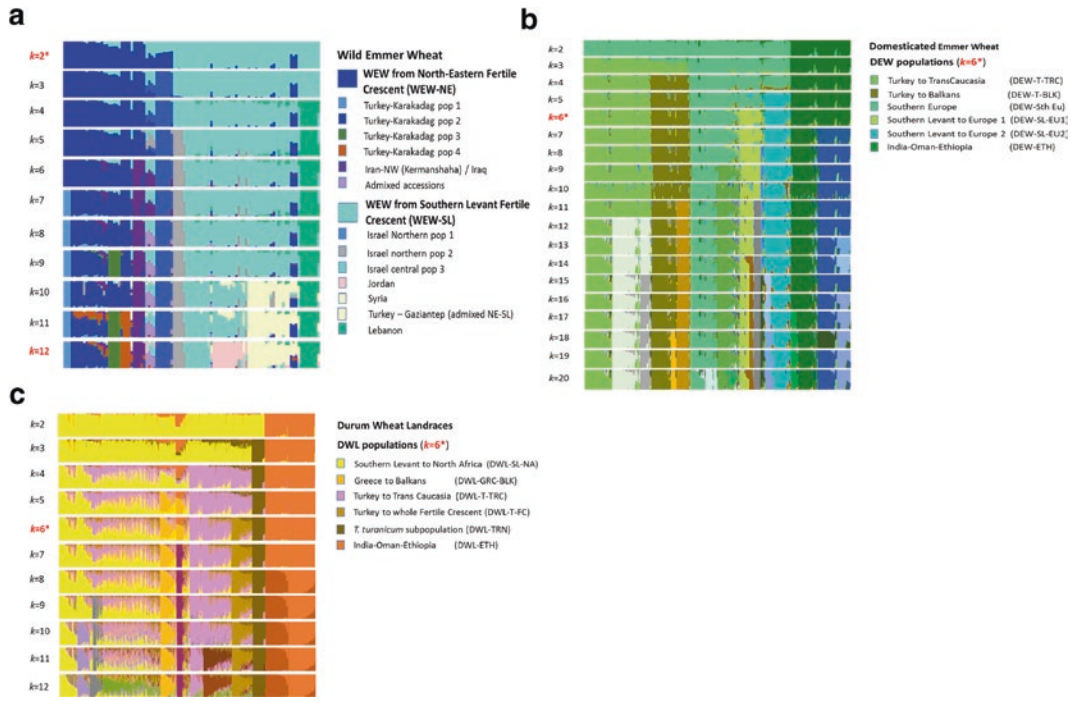


Fig. 8.4 Admixture analysis of wild emmer, emmer, and landraces included in the TGC, represented as bar plots of Q membership coefficients (modified from Maccaferri et al. 2019). More in detail results of: **a** wild emmer wheat accessions with K from 2 to 12; **b** domesticated emmer wheat accessions with K from 2 to 20; **c** durum wheat landrace accessions with K from 2 to 12

which includes tall accessions with upright habitus, wide spikes with large grains, and more fertile than the rest of WEWs in the western area (Poyarkova et al. 1991). Ecological variables also play an important role in shaping the genetic structure of WEW. Indeed, loci under positive selection significantly correlated with eco-geographical factors (e.g., geographic location, temperature, water availability, singly or in combination) for allele frequency suggesting that natural selection could have created regional divergence in WEW (Ren et al. 2013). An example of natural selection shaping WEW genetic diversity is provided by *Yr15,* a broad spectrum disease resistance gene cloned in WEW belonging to the family of tandem kinase-pseudokinase proteins (Klymiuk et al. 2018). Northern regions of Israel show climatic conditions more favorable for stripe rust pathogen development with respect to the southern regions. A large screening of wild emmer natural populations confirmed that *Yr15* gene is present only in northern Israeli populations and distributed along a narrow mountain ridge of about 100 km from Mt. Carmel to Mt. Hermon regions, mainly at an elevation higher than 500 m above sea level (Klymiuk et al. 2019a; He et al. 2020). Thus, it seems that selection pressure exerted by the pathogen is affecting the host-parasite interactions and co-evolution and shapes the distribution of resistance genes among wild emmer populations.

8.2.2 Emmer Wheat, the First Domesticated Wheat

DEW was a widely cultivated staple crop in the Near East, ancient Mesopotamia, and Egypt for over 7000 years during the Neolithic period. The decline started in Turkey during the Bronze Age about 5000 years ago when it was replaced

by naked wheats (*T. durum* and/or *T. aestivum*), while in Europe its cultivation continued until about 2000 years ago with a long and slow decline. Today, DEW can be found only in marginal areas and some isolated traditional farming communities in the Balkans and Mediterranean countries, Iran, Armenia, Ethiopia, Yemen, Oman, and India. Recently, it has been re-discovered by the organic food industry for bread and cookie production.

Geographical expansion of DEW was intimately associated with historical human migrations and spread from the Fertile Crescent with a typical *star-like dispersal mode*. Indeed, four major diffusion routes out of the Fertile Crescent have been postulated (Badaeva et al. 2015). The expansion of DEW was a long and complex process in which emmer genotypes became adapted to new habitats and climates. The genetic structure of DEW populations was affected, among other factors, by exchange of seed stock during migration and by gene flow between wild and domesticated wheats or between different locally adapted DEW populations.

Few studies have focused on the genetic structure of the DEW germplasm. A cluster analysis, based on karyotypic information on a comprehensive collection of 446 DEW lines from 47 countries, identified four groups (Balkan, Asian, European, and Ethiopian) that allowed the authors to postulate four major diffusion routes of the crop out of the Fertile Crescent (Badaeva et al. 2015). Notably, although specifically evolved in certain geographic regions, populations of DEW usually included representatives of more than one karyotypic group at different frequencies. This mixture of karyotypic groups probably originated from multiple crop introductions/exchanges by successive waves of colonizing civilizations, which swept across Europe, the Mediterranean, and Asia. This clustering partially agrees with the population structure highlighted by Liu et al. (2017a, 2017b) on a collection of 176 spring accessions representing a large portion of the worldwide genetic diversity in the gene pool of cultivated spring emmer wheat. Three major groups were recognized: an "African subpopulation" with mostly accessions from Ethiopia, Kenya, and Morocco, the "European subpopulation" with accessions from southern and western Europe, and an "Asian subpopulation" grouping accessions from eastern and western Asia. About DEW, the TGC collection included up to 335 unique, non-admixed DEWs comprehensively sampled from gene banks worldwide. Their analysis clearly evidenced the presence of at least six well-distinct main populations evolved based on the human-driven dispersal along the main already described migration routes (Maccaferri et al. 2019; Fig. 8.4b). The diversity analysis evidenced the already described high stratification level consisting of six main populations and up to 18 subpopulations corresponding to: (1–2) two distinct and ancestral populations from the *southern Levant*, (3) a population close relative of the southern Levant populations but distinct and evolved in *southern Europe*, (4) a population evolved along the dispersal route *Turkey-to-Balkans*, (5) a distinct population evolved along the *Turkey-to-Transcaucasia/Iran*, and (6) an early-separated population evolved and spread from *Oman-to-India/Ethiopia*.

All these data pointed out the presence of a considerable level of diversity naturally evolved post-domestication in adaptation to environments and well differentiated from the native Fertile Crescent (southern Levant and Turkey). The unique application of genome-wide association studies (GWASs) reported so far on DEW also indicated high genetic diversity. Indeed, the same collection showed to be a rich source of stripe rust resistance loci very useful for wheat improvement (Liu et al. 2017a, 2017b). Among the 51 loci for resistance including genes effective in multiple field environments or against multiple races, a large proportion mapped distantly from previously reported stripe rust resistance genes or QTLs and provide novel resistance loci. Notably, African germplasm showed a higher frequency of resistant genotypes to stripe rust than the other two subpopulations.

8.2.3 Variable Human and Environmental Pressures Have Affected Divergence of Durum Wheat Landraces

Similar to DEW, the southern Levant is the center of origin of *T. turgidum* ssp. *durum* (Vavilov 1951; Feldman 2001). The first evidence of durum wheat dates ~7500–6500 years ago (Faris 2014). Then, it spread throughout the same migration routes already described for DEW through substantially independent pathways with limited evidence of gene flow and/or admixture between DEW and DWL (Maccaferri et al. 2019).

The dispersal routes moved durum wheat west throughout the Mediterranean Basin up to the Iberian Peninsula, probably via trading by Phoenician merchants and along the caravan' routes along the Sahara desert or the North African coasts (Bozzini 1988), and east through the Silk Road to Asia (Waugh 2010). Following another early dispersal route to Ethiopia, an independent origin of durum wheat by a separate domestication of naked emmer has been suggested to have occurred in Ethiopia and have originated *T. durum* ssp. *abyssinicum* which is morphologically different from other durum wheat accessions, with uncompact spikes and small purple seeds (Mengistu et al. 2015, 2016). In addition, natural and anthropogenic selection in DEW during human migration resulted in the establishment of local DWLs specifically adapted to a diversity of agro-ecological zones (Nazco et al. 2012).

Local landraces were progressively abandoned starting from the early 1970s due to their replacement with the improved, more productive, and genetically uniform semi-dwarf cultivars derived from the Green Revolution. This notwithstanding, empiric breeding aimed to exploit the phenotypic variability of DWLs resulted in traditional varieties still preferred by smallholder farmers in traditional farming systems of rural/marginal areas where modern intensive ones cannot be adopted and/or where this germplasm provides the required higher stress tolerance. These traditional DWLs are usually tall plants and are often cultivated for both grain and straw, where in case yield is too low or even fail due to high temperature and drought the straw can still be harvested. Thus, crop diversity managed by smallholder farmers in traditional agro-systems is the outcome of historical and current processes interacting at various spatial scales and influenced by local factors such as farming practices and environmental constraints. Due to their evolutionary dynamics, landraces strongly represent the diversity of semiarid and marginal conditions. Indeed, evidence supports the hypothesis that DWLs harbor the largest source of biodiversity within the cultivated durum germplasm, including documented resilience to abiotic stresses and resistance to pests and diseases which could be used to enrich the modern wheat genetic repertoire for the improvement of commercially valuable traits (Lopes et al. 2015).

Many studies have focused on panels of landrace accessions from a restricted country/area, as those from southern Italy (Marzario et al. 2018; Mangini et al. 2018), Iran (Seyedimoradi et al. 2016), Spain (Giraldo et al. 2016; Ruiz et al. 2012), Tunisia (Robbana et al. 2019; Slim et al. 2019; Ouaja et al. 2021), Turkey and Syria (Baloch et al. 2017), Palestine, Jordan and Israel (Abu-Zaitoun et al. 2018), Morocco (Kehel et al. 2013; Sahri et al. 2014), and Ethiopia (Mengistu et al. 2016; Alemu et al. 2020).

Different drivers of population restructuring have emerged from the analysis of these collections. A collection of 91 DWLs originating from a wide range of ecological conditions of soil, temperature, and water availability in Turkey and Syria showed a grouping pattern not associated with the geographical distribution of durum wheat, suggesting a high mixing of Turkish and Syrian landraces due to large exchange of genetic material among farmers, as an alternative to the lack of commercial varieties (Baloch et al. 2017). Higher admixture among landraces was also observed in Ethiopia although it is a country characterized by a wide range of agro-ecological conditions coupled with diverse farmers' culture. Indeed, both the clustering of 167

DWLs by Alemu et al. (2020) and 287 Ethiopian DWLs by Mengistu et al. (2016) collected from major wheat-growing areas of the country did not reflect their geographical origin, suggesting admixture arose from the existence of historical seed exchanges involving regional and countrywide farming communities in Ethiopia. Moreover, Seyedimoradi et al. (2016) reported low correlation between genetic distances and geographical origin in a small panel of DWLs from different zones of Iran. Notably, the genetic analysis showed that the country of origin did not have any genetic footprint within the core collection of DWLs from Jordan, Palestine, and Israel, despite historical sociopolitical barriers present in this area during the last decades (Abu-Zaitoun et al. 2018). However, in the latter case adaptations to similar semiarid conditions might reflect no separation between neighboring countries. Conversely, the fingerprinting of a collection of the National Gene Bank of Tunisia (NGBT) for traditional varieties from different Tunisian agro-ecological zones has found a strong genetic stratification from north to south in Tunisia (Slim et al. 2019). Indeed, five subpopulations were identified, two of which appeared more strongly represented in germplasm collected in central and southern Tunisia, where environmental conditions at critical development phases of the plant are harsher. Notably, these subpopulations were underrepresented in modern varieties which were instead prevalent in the north, suggesting that traits for breeding more resilient varieties might be present in central and southern Tunisian traditional varieties. In Morocco, a stratification of the genetic diversity according to agro-ecological conditions (geography, but also water and temperature regimes) was recorded, related to the two distant regions Pre-Rif and Atlas Mountains which display very different environmental, cultural, and agronomic conditions (Kehel et al. 2013; Sahri et al. 2014). However, within each region, only a few patterns emerged from the genetic and morphologic characterization, as if distance does not represent a consistent barrier to genetic exchange (Sahri et al. 2014). Different hypotheses can explain these results, ranging from unconscious mixing by farmers in threshing areas, to unreliability of seed exchange networks (e.g., seed lots that do not correspond to the declared names) as well as limited farmers' interest for durum wheat cultivation and seed production. These mixtures create strong opportunities to generate diversity through cross-fertilization and recombination but also would homogenize the pool of traditional varieties in the absence of human or environmental divergent pressures that maintain some differentiation between them.

In the TGC collection (Maccaferri et al. 2019), up to 947 accessions of durum and durum-related unique and relatively low in admixture were analyzed for population structure (Fig. 8.4c). The results showed six main populations corresponding to: (1) a local *Turkey-to-Levant* (in particular Syria) population, (2) the main *Southern Levant-to-North Africa* and *Southern Europe* (Italy, Spain, and Portugal) migration route, (3) a highly differentiated *Ethiopian* population subdivided into two subpopulations, (4) a *Turkey-to-Transcaucasia/Russia* route, (5) a well-distinct *T. turgidum* ssp. *turanicum* population developed in Iran/Iraq up to Afghanistan, and (6) a localized *Greece-to-Balkans* population including representatives of the *T. turgidum* ssp. *turgidum*.

Other comprehensive studies considering both molecular and phenotypic data considered wide panels of landrace accessions from a larger geographic area as the collection of 172 DWLs from 21 Mediterranean countries characterized by Royo et al. (2014) and Soriano et al. (2016, 2018). Germplasm from the Mediterranean area is of interest for traits relevant for adaptation to the climate changes since in this region wheat is mainly grown under rain-fed conditions and yield is often constrained by water and heat stress that are common during the grain-filling period due to the low and unpredictable seasonal rainfall. Thus, the above collection highlighted an evident relationship between the genetic stratification and the eco-geographic patterning, which was suggested to be the result of different physiological and genetic strategies

to sustain yield according to prevalent climate conditions. Indeed, the 172 DWLs showed a genetic structure related to an eastern–western geographical pattern formed by four clearly defined groups: eastern Mediterranean, eastern Balkans and Turkey, western Balkans and Egypt, and western Mediterranean, in agreement with the dispersal pattern of wheat from east to west in the Mediterranean Basin (Soriano et al. 2016). Interestingly, this study also showed a reliable relationship between genetic and phenotypic population structures, the latter being based on yield, yield components, and crop phenology-related traits. A high number of spikes and harvest index were recorded in DWLs from the eastern Mediterranean Basin, in agreement with the findings of previous studies (Moragues et al. 2006; Royo et al. 2014) which demonstrated that durum wheat yield under warm and dry environments is determined mostly by the number of spikes per unit area, whereas kernel weight predominantly influences grain production in colder and wetter environments. Interestingly, using a subset of this collection, Soriano et al. (2018) identified 23 marker alleles with a differential frequency in DWLs from east and west regions of the Mediterranean Basin, which affected the mentioned agronomic traits. Eastern DWLs had higher frequencies than the western ones of alleles for increasing the number of spikes (chr. 1B), grains per m^2 (chr. 7B), grain-filling duration (several marker-trait associations), reduced cycle length, and lighter grains (chr. 4A, 5B, and 6B).

8.2.4 Main Breeding Gene Pools Within the DWC Germplasm

Durum wheat genetic makeup became more complex at the beginning of the twentieth century when conscious breeding started by applying artificial hybridization and selection pressure for commercial purposes (Autrique et al. 1996; Pecetti and Annicchiarico 1998; De Vita et al. 2007). The first durum wheat breeding program, setup in southern Italy by Nazareno Strampelli, was initially based on the selection of pure lines from local landraces (Scarascia Mugnozza 2005), which in 1915 led to the release of the cultivar CAPPELLI, a pioneer cultivar which had a major global impact in the following years and to which many modern varieties can be traced back to. A second major impact was provided by the deployment of lines carrying dwarfing genes to increase harvest index. This was first carried out by Nazareno Strampelli in Italy using *Rht8* from the variety AKAKOMUGI, from Japan, and several years later by Norman Borlaug in Mexico using *Rht-B1b* also from a Japanese variety called NORIN10. The dwarfing gene *Rht-B1b* was successfully transferred to the durum wheat CANDO in the 1960s and widely used in durum wheat breeding (Quick et al. 1976). The last decades have been characterized by several hybridizations occurring between different breeding programs or with relatives aiming at increasing productivity while ensuring genetic diversity and mega-cultivars that have crossed the boundaries of their country of origin (Ren et al. 2013). Thus, although Autrique et al. (1996) observed that a limited number of ancestral lines have contributed largely to the development of the modern durum wheat materials and that the molecular fingerprints of a few ancestors accounted for most of the molecular diversity detected in the cultivated gene pool, a more complex network has emerged from studies on the genetic diversity pattern of the most recent durum wheat germplasm.

Numerous studies reported on characterization of diverse panels made of DWCs (Maccaferri et al. 2005, 2006, 2011; Reimer et al. 2008; Condorelli et al. 2018) or related to a breeding program (N'Diaye et al. 2018). These works have identified a few main gene pools reflecting the genetic basis and breeding strategies involved in their development. Maccaferri et al. (2005) subdivided the DWCs into six major gene pools: (1) Italian group which includes varieties selected and released in Italy, (2) CIMMYT-ICARDA group with hallmark accessions derived from the CIMMYT-ICARDA breeding program and released in Mexico, Spain, Italy, and in several West Asia and North Africa countries, (3) French group encompassing lines released by French

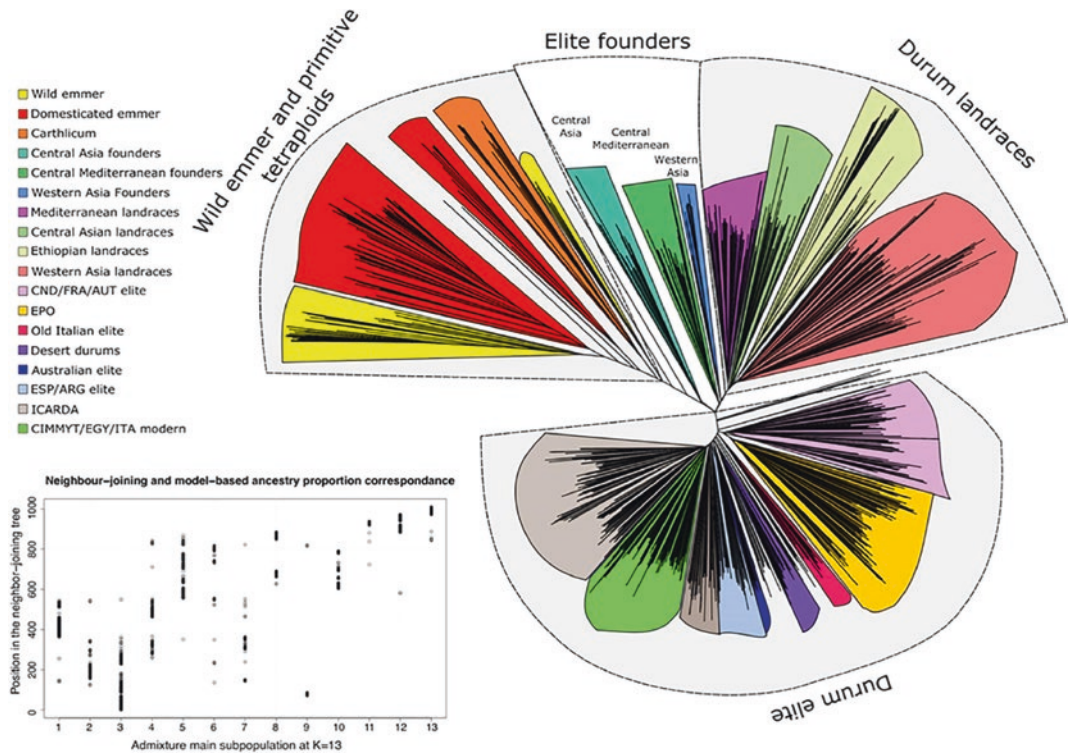


Fig. 8.5 Neighbor joining tree of the Global Durum Panel (GDP) collection (modified from Mazzucotelli et al. 2020)

breeders and well adapted to a range of environments throughout central and south Europe, (4) Austrian–Australian group derived from Austrian or Australian breeding programs, (5) North American group with accessions selected in the Great Plains of the USA and Canada, and (6) southwestern US group constituted by representative of the germplasm cultivated in the southwestern region of the US under irrigation and commonly referred to as desert durum.

More recently, the analysis of wider collections, as reported by Kabbaj et al. (2017) and by Mazzucotelli et al. (2020), provided the basis for separating the two CGIAR breeding programs (CIMMYT and ICARDA), as well as defining a subgroup of highly admixed varieties derived by exchange of materials among different breeding pools. This high exchange of materials was confirmed by the analysis of genetic diversity on the Global Durum Panel (GDP) clusters based on geography and breeding program of origin. Indeed, it was shown that most diversity remained among individuals within clusters, and only 13% of the total genetic variance could be captured by groups (Mazzucotelli et al. 2020). These insights also indicated that a good level of genetic diversity remains available within the breeding groups for direct exploitation, and there is even greater potential when considering exchanges between breeding groups. The 473 modern cultivars/breeding lines of the GDP were grouped into nine distinct groups organized as follows: old Italian elite, ICARDA, CIMMYT, Spanish/Argentinian elite germplasm, US desert durum, Australian elite, and at the opposite of diversity, the North American, Canadian, and French germplasm, the latter including the Evolutive Population (EPO, David et al. 2014) (Fig. 8.5). Founders of these modern cultivars were identified in western Asian durum landraces, North African Mediterranean landraces and central Asian, Turkey to Transcaucasia/

Russian landraces, indicating that some landraces and durum primitive groups did not contribute to the genetic makeup of modern cultivars.

Interestingly, the level of LD decay rate, an important feature to be assessed when implementing GWAS, is quite differentiated comparing modern durum cultivars to landraces (Fig. 8.6). On average, LD decays to $r^2 = 0.3$ (a generally accepted reference threshold in GWAS) in a range of physical distances of 4.21 Mb in modern durum, while the range decreases to 0.94 Mb in landraces. Thus, the landrace germplasm potentially offers higher genetic resolution in QTL mapping than modern varieties. However, these metrics are highly differentiated when referring to pericentromeric versus distal chromosome regions, with the physical-to-genetic ratio showing differences of 10^2–10^3 magnitude (Maccaferri et al. 2019).

Through GWAS on mentioned germplasm collections, specific phenotypic traits and/or the frequency of alleles at known loci for critical adaptation traits (vernalization requirement, response to photoperiod, heading date, plant height), for disease resistance, and for root morphology have been interestingly related to the population structure. For instance, Maccaferri et al. (2011) identified different patterns of allele frequency at the three major genes for wheat phenology and plant architecture (*Vrn-A1*, *Ppd-A1*, and *Rht-B1*) across the five subgroups present in a collection of elite durum mostly composed of Mediterranean germplasm. Most of the accessions were vernalization-insensitive and semi-dwarf, as expected for élite durum wheat materials. The vernalization-sensitive allele *vrn-A1* was present in only six accessions, all but one (CLAUDIO) from ICARDA germplasm.

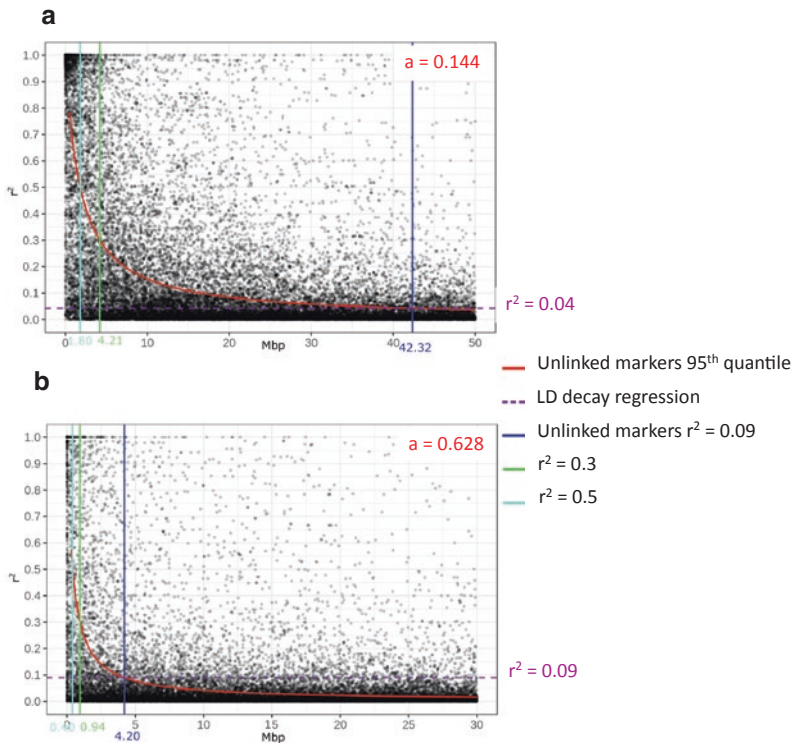


Fig. 8.6 Genome-wide linkage disequilibrium (LD) decay in respect to physical distance in the GDP collection for the two main groups of: **a** modern durum wheat germplasm and **b** durum landraces; critical distances are provided to three different r2 threshold values (0.5, 0.3, and 0.09 as for unliked markers) (Figure from Mazzucotelli et al. 2020)

Interestingly, in a subgroup made of cultivars bred for the semiarid areas, most genotypes carried the wild-type *Rht-B1a* allele and an almost fixed *Ppd-A1* wild-type allele. In addition to conferring a tall phenotype, the wild *Rht* allele also increases the coleoptile length hence allowing for deeper sowing and better exploitation of soil moisture, thus making genotypes better suited for drought prone areas (Rebetzke et al. 2007). As to the *Ppd* locus, the wild allele confers lateness through photoperiod sensitivity in the Mediterranean environments, while the photoperiod-insensitive alleles dominate in most of the modern germplasm accessions. The same collection was evaluated for root system architecture, with a focus on root growth angle which is considered a fundamental trait to enhance the genetic capacity of the plant to acquire soil resources (Sanguineti et al. 2007; Maccaferri et al. 2016). Indeed, a narrow and deep root in contrast to a shallow ideotype can contribute to drought resistance and was found correlated with grain yield under harsh rain-fed conditions. Alleles contributing a narrow root growth angle were found to be present at relatively high frequencies in the modern high-yielding germplasm including the most recent cultivars from CIMMYT/ICARDA programs, the Italian, and the desert durum cultivars. The major root growth angle QTL detected on chromosome 6AL (Maccaferri et al. 2016) was also reported in a study on Ethiopian germplasm (Alemu et al. 2021).

Resistance to diseases is a relevant trait for durum varieties. Notably, the analysis of panels made of cultivars has frequently identified the leaf rust-resistant gene *Lr14,* a locus originally transferred from DEW YAROSLAV to common wheat (McFadden 1930), then identified in the Chilean DWC LLARETA-INIA and in diverse loosely related genetic materials, such as the CIMMYT line Somateria (Herrera-Foessel et al. 2008a), the Italian cultivars COLOSSEO (Maccaferri et al. 2008), and CRESO (Marone et al. 2009). The resistant haplotype at the *Lr14* locus was found in many cultivars from Italian, CIMMYT and ICARDA breeding programs suggesting it was the most important source of resistance to leaf rust exploited by durum breeders. Another interesting example has been provided by the breeding for resistance to Hessian fly (Bassi et al. 2019). A major locus was identified on chromosome 6B in a group of Moroccan DWCs related to the cultivar NASSIRA. Pedigree analysis demonstrated kinship of these lines and traced back the origin of the locus to a resistant *T. araraticum* accession which had been used to introgress the resistance in locally adapted elite lines.

8.2.5 Selection Signatures

An exhaustive genome-wide analysis of changes in genetic diversity imposed by thousands of years of empirical selection and breeding was enabled by the Global Tetraploid Wheat Collection consisting of 1856 accessions representing the four main germplasm groups involved in tetraploid wheat domestication history and breeding (*T. dicoccoides, T. dicoccum,* DWLs, and DWCs) (Maccaferri et al. 2019). For each germplasm group, the pattern of diversity was assessed through a SNP-based gene diversity index (Fig. 8.7), then different metrics were used to detect selection signatures between evolutionary transitions, including both diversity reduction, and divergence/differentiation of allele frequency. WEWs showed the highest average diversity with only two pericentromeric regions (chr. 2A and 4A) with a lower-than-average diversity, thus the authors referred to WEW as the reference for assessing the reduction of diversity associated with domestication and breeding in tetraploid wheat. In total, 104 pericentromeric (average size 107.7 Mb) and 350 non-pericentromeric (average size 11.4 Mb) genomic regions reported co-occurrence of signals of selections in one or more evolutionary transitions.

Compared to WEW, each of the subsequently domesticated/improved germplasm group showed several strong diversity depletions that arose independently and were progressively consolidated through domestication and breeding. Consequently, the genome of DWCs revealed numerous regions showing near fixation of allelic diversity. Exceptions were observed for chromosomes 2A and 3A in

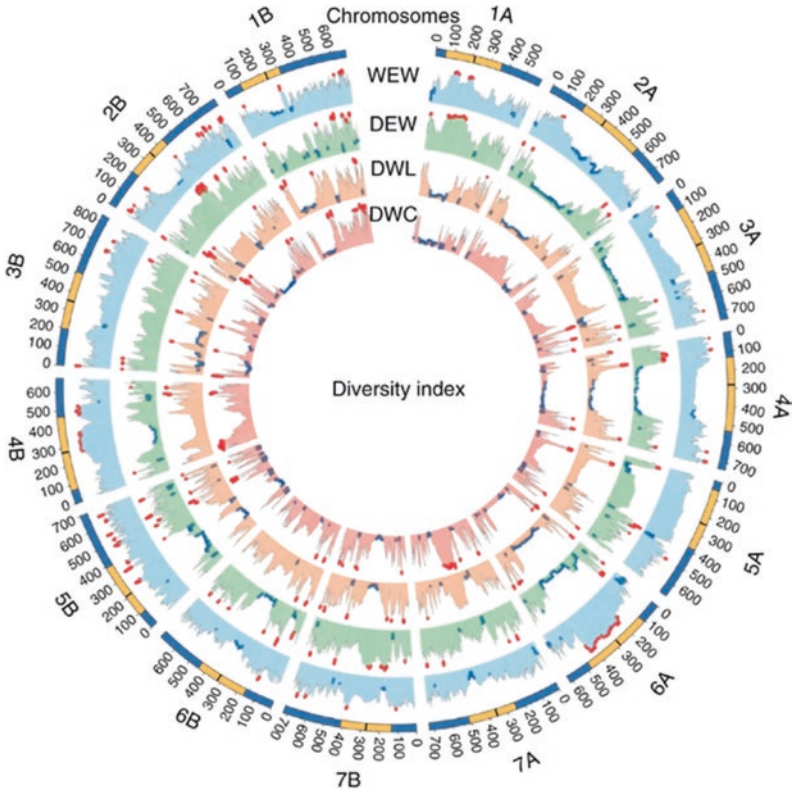


Fig. 8.7 SNP-based diversity index (DI) for the main germplasm groups identified in the TGC (WEW, DEW, DWL, and DWC). DI is reported as a centered 25 SNP-based average sliding window (single SNP step). Top and bottom 2.5% DI quantile distributions are highlighted as red- and blue-filled dots, respectively (modified from Maccaferri et al. 2019)

the pericentromeric region where the DWCs showed an increased diversity as compared to DWL and DEW groups. The pericentromeric regions showed extensive signals of divergence/selection in the WEW-DEW and WEW-DWL transitions which highlights that most of the loss of diversity and divergence signatures occurred during domestication. The combination of this analysis with availability of the durum wheat reference genome allowed higher resolution analysis for the non-pericentromeric regions based on a comparative alignment between selection signals and wheat genes and QTLs relevant for domestication/improvement. For instance, the transition from DEW to durum wheat showed different depletion of diversity related to technological quality improvements. Indeed, the locus *Glu-A1*, coding for glutenin subunits and located at 500.8 Mb on chromosome 1A, which was reported to be nearly fixed in modern germplasm for null allele, was associated to a local strong signal of diversity reduction. Analogously, other extreme reductions in diversity were found colocated with grain yellow pigment content loci, including *Psy-B1*. Interestingly, among a set of 41 previously cloned loci, that have been most probably the target of selection, many colocated with regions marked by strong selection metrics. Intriguing examples were detected at critical loci for grain weight, a trait that has been strongly modified across the domestication and selection history for its relationship with grain yield (Fig. 8.8, Desiderio et al. 2019). Co-location was found for *TaGW2* on chromosome 6A in the WEW-to-DEW transition and *TaGW2* on 2B for both WEW-to-DEW and DEW-to-DWL transitions. Additionally, *TaSus2-A1*, *TaSdr-A1*,

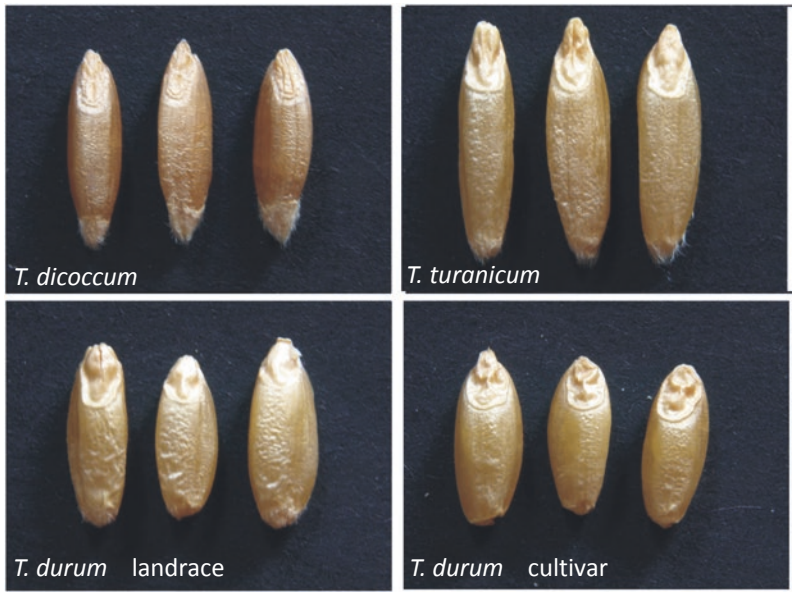


Fig. 8.8 Variation for kernel size and shape in tetraploid wheat (modified from Desiderio et al. 2019)

and *TaCWI-A1* on chromosome 2A and their homeologs on 2B were associated to multiple extended signals in WEW-to-DEW and in DEW-to-DWL transitions, while the durum germplasm showed extended regions of low diversity.

8.3 Tetraploid Wheat Genomes

The reference genome of the DWC (SVEVO v.1, Maccaferri et al. 2019) and of the WEW accession ZAVITAN (WEWseq v.1.0, Avni et al. 2017; assembly reviewed in Zhu et al. 2019 based on optical mapping) have been sequenced in recent years, hence allowing for a better exploitation of the genetic diversity of the tetraploid gene pool. The SVEVO genome sequence was assembled in 10.46 Gb including 0.5 Gb of unassigned scaffolds. Very similar numbers were produced for the genome sequence of ZAVITAN with an assembly size of 10.5 Gb including 0.4 Gb of unassigned scaffolds. The alignment of the durum wheat genome with high-density SNP genetic maps showed the typical pattern of recombination with highly recombinogenic distal chromosome regions and large pericentromeric regions nearly devoid of recombination.

A comparison of the two assemblies revealed strong overall synteny with high similarity in total gene number (66,559 high confidence genes in SVEVO vs. 65,012 in ZAVITAN) and repetitive element content (82.2% of the total assembly) and composition (Maccaferri et al. 2019). Nevertheless, a comparison of the orthologous gene pairs has highlighted several examples of presence–absence variations and of copy number variations as expected in a context of pangenome analysis where deletions and gene family expansions are frequently found (Walkowiak et al. 2020). For many years, genomic studies in tetraploid wheats were carried out based on the genomic resources developed in bread wheat thanks to the extensive sequence similarity and gene collinearity between the A and B genomes of the two species. As an example, the SNPs carried on the wheat 90K iSelect Infinium SNP assay (Wang et al. 2014), most of which originated from bread wheat, have been extensively used in genetic diversity and mapping studies in tetraploid wheat.

The availability of genome sequences for wild emmer and durum wheat is expected to facilitate genomic studies in tetraploid wheats. The projection of the 90K iSelect Infinium SNPs

to the SVEVO and ZAVITAN genomes allowed to map genes and QTLs with higher precision and resolution. Moreover, once identified, the QTL confidence interval can now be used to search for candidate genes directly in the tetraploid genome. Such approaches have been used in mapping studies based on biparental segregating populations, as in the case of the identification of the *SrKN* gene for stem rust resistance from the tetraploid DWC KRONOS (Li et al. 2021); GWAS for different traits (Saccomanno et al. 2018; Aoun et al. 2021), and ultimately to better refine QTL regions through meta-QTL analysis for quality, abiotic and biotic stresses in durum wheat (Soriano et al. 2021). The molecular characterization of gene families can greatly help in the search of candidate genes for a particular trait, in studies on gene mapping, functional analysis, and comparative genomics, as in the case of the analysis of *Hsp70* and *glutathione S-transferases* (GSTs) genes in different *Triticum* subspecies, with implications on evolution of these gene families and molecular mechanisms of their involvement in response to stress (Lai et al. 2021; Hao et al. 2021). The approach can also be focused on the characterization of a chromosomal locus, as seen for *Gli-2* locus regions containing *α-gliadin* genes on A and B genomes of WEW (Huo et al. 2019). The availability of genome sequences is also of particular interest in transcriptomic studies such as those based on RNA-seq, in which projecting the reads onto a high-quality genome can greatly improve the accuracy and completeness of the analysis (Arenas et al. 2022). In a recent study, both the analysis of the translatome, the collection of all open reading frames that are actively translated, and in vivo RNA structure profiling were carried out to investigate the complex wheat RNA structure landscape in durum wheat. The translatome revealed subgenome asymmetry at the translational level, due to the strong impact of mRNA structure on translation, independent of GC content (Yang et al. 2021).

Matching mapping results with information regarding the gene content and annotation of genomic regions provides a huge advantage for fine mapping and gene/QTL cloning in tetraploid wheat. With the fully assembled ZAVITAN genome, the causal mutations in *Brittle Rachis 1* (*TtBtr1*) genes controlling shattering, a key domestication trait, were identified (Avni et al. 2017). More recently, *TdHMA3-B1*, a gene encoding a metal transporter with a non-functional variant causing high accumulation of cadmium in grain, was rapidly cloned in SVEVO. Moreover, a wild functional allele, characterized by a very low frequency among DWCs, was rescued with great advantage for durum wheat breeding for cadmium accumulation in grain (Maccaferri et al. 2019). The utility of tetraploid wheat genome has also been shown for improvement of resistance to fungal diseases. The WEW derived *Yr15*, a gene for broad-spectrum resistance to stripe rust, was identified and cloned in a large mapping population developed by crossing the susceptible durum wheat line "D447" with introgression lines carrying *Yr15* in the genetic background of "D447" (Klymiuk et al. 2018). Both ZAVITAN and SVEVO genomes were used to clone and functionally characterize *Pm41*, a powdery mildew resistance gene derived from WEW, which encodes a coiled-coil, nucleotide-binding site, and leucine-rich repeat protein (CNL) (Li et al. 2020).

Interestingly, all available wheat genomes are important in comparative genomic approaches to precisely characterize a chromosomal region and its gene content in gene cloning studies. This way, tetraploid genomes are instrumental for fine mapping and cloning of genes not only in emmer or durum wheat, but also in bread wheat as in the case of *Ne2*, a typical CNL gene responsible for hybrid necrosis in wheat (Si et al. 2021).

All these data indicate that wheat genomes for bread (IWGSC 2018), durum (Maccaferri et al. 2019), and wild emmer (Avni et al. 2017) wheat once merged in a unique wheat pangenome will provide an excellent asset for genetic studies in both tetraploid and hexaploid wheat. At the same time, the tetraploid and hexaploid wheat germplasms could be considered a unique gene pool from which to recruit genes and alleles useful for breeding in both species (Mastrangelo and Cattivelli 2021).

8.4 Tetraploid Germplasm for Bread Wheat Improvements (and Vice Versa)

The increasing demand for food and more sustainable crops and the increasingly evident climatic changes require the selection of new wheat DWCs with improved grain yield, protein content, and resistance to biotic and abiotic stresses (Foley et al. 2011; Soares et al. 2019). Achievement of this goal is hindered by the limited genetic variability present in wheat modern cultivars resulting from multiple domestication bottlenecks and breeding involving a few selected progenitors. This situation prompted the attention toward the use of DWLs, non-cultivated wheat subspecies, and wild relatives to contribute genes conferring traits of interest and, more in general, to increase the genetic diversity of the cultivated gene pool (Reif et al. 2005). Nowadays, the identification and characterization of these genes and/or related regions (QTLs) are assisted and accelerated by the recent technological advances in wheat genomics (Tuberosa and Pozniak 2014; Vendramin et al. 2019; Rasheed and Xia 2019), and their introgression into elite cultivars can be performed both with marker-assisted selection (MAS) and/or with transgenic strategies (Cobb et al. 2019; Gadaleta et al. 2008; Mores et al. 2021).

The close phylogenetic proximity between tetraploid and hexaploid wheat allows for the transfer of specific genes between the two species, as already occurred during wheat evolution (Dubcoksky and Dvorak 2007). Crosses between the two species are feasible, overcoming the problems due to necrosis and low fertility of hybrids (Klymiuk et al. 2019b; Othmeni et al. 2019). Although the superior adaptability of the hexaploid genome has made bread wheat the most cultivated wheat worldwide, tetraploid wheat exhibits greater genetic diversity, a highly desirable feature for present and future wheat breeding programs (Mastrangelo and Cattivelli 2021). WEW is probably the most relevant reservoirs of genetic diversity for durum and bread wheat with durum acting as a bridge between the wild relative and the bread wheat to facilitate the introgression of WEW traits in modern bread wheat lines (Maccaferri et al. 2015; Klymiuk et al. 2019b).

8.4.1 Transfer of Disease Resistance Genes

A number of important genes for biotic stress resistance have been transferred into common wheat from the primary gene pool of tetraploid wheats (*T. turgidum* ssp. *dicoccoides*, ssp. *dicoccum*, and ssp. *durum*), such as those related to the most dreadful and economically important diseases of wheat: rust, namely yellow rust (Yr—*Puccinia striiformis* f. sp. *tritici*); leaf rust (Lr—*Puccinia triticina Eriks*); stem rust (Sr—*P. graminis* f. sp. *tritici*), powdery mildew (Pm—*Blumeria graminis* f. sp. *tritici*), and Fusarium head blight (FHB; *Fusarium graminearum*). The list of rust resistance genes identified and transferred from durum to hexaploid wheat includes: the yellow rust resistance genes *Yr53* (chr. 2BL, Xu et al. 2013), *Yr64,* and *Yr65* (chr. 1BS, Cheng et al. 2014); the leaf rust resistance genes *Lr23* (chr. 2BS, McIntosh et al. 1995; Sibikeev et al. 2020), *Lr61* (chr. 6BS, Herrera-Foessel et al. 2008b), and *Lr79* (chr. 3BL, Qureshi et al. 2018), and the stem rust resistance genes *Sr12* (chr. 3BL, Sheen and Snyder 1964), *Sr13* (chr. 6AL, Simons et al. 2011; Zhang et al. 2017), *Sr8155B1* (chr. 6AS, Nirmala et al. 2017), and *SrKN* (chr. 2BL, Li et al. 2021).

Other stem rust-resistant genes (*Sr2*, *Sr13*, and *Sr14*) have been transferred into common wheat from cultivated emmer. *Sr2* (3BS, McIntosh et al. 1995), a recessive and race non-specific adult plant resistance gene, was transferred from the emmer variety YAROSLAV into common wheat Hope and represents a major success in resistant wheat breeding which has been deployed in many cultivars in the last 80 years and still confers an effective rust resistance. *Sr14* (chr. 1BL) was identified in the cultivated emmer Khapli and introgressed into hexploid cultivar Steinwedel. Both *Sr2* and *Sr14* are currently important

sources of resistance to Ug99 lineage races of stem rust (Singh et al. 2011). *Sr13* (chr. 6AL), present in both in durum and cultivated emmer, was transferred to the common wheat variety KHAPSTEIN from the DEW KHAPLI. *Sr13* confers resistance to all races in the Ug99 group (Jin et al. 2007), but the resistant responses are influenced by temperature and genetic background (McIntosh et al. 1995; Roelfs and Mcvey 1979). *Lr53* and *Lr64* mapped on chromosome 6BS and 6AL, respectively, were transferred from WEW to common wheat (Kolmer 2008; Dadkhodaie et al. 2011; Huang et al. 2016). Several stripe rust resistance genes derived from WEW (*Yr15* on chr. 1BS, *Yr35*-6BS, *Yr36*-6BS, *YrH52*-1BS, and *YrSM139*-1BS) were mapped using *T. durum* × *T. dicoccoides* segregating populations and transferred into bread wheat using durum wheat as a "bridge" (Peng et al. 2000; Dadkhodaie et al. 2011; Hale et al. 2012; Yaniv et al. 2015; Zhang et al. 2016).

Durum wheat has been used as source of powdery mildew resistance genes (*Mld*, *Pm3h* and *PmDR147*) for bread wheat improvement (Miedaner et al. 2019). *Mld* (chr. 4B, recessive) was employed in wheat breeding in combination with other *Pm* resistance genes, such as *Pm2* (chr. 5DS, Bennett 1984) and *Pm3h* (chr. 1AS, dominant, Yahiaoui et al. 2006), and probably originated from an Ethiopian durum wheat accession (Srichumpa et al. 2005). *PmDR147* (chr. 2AL, dominant) was transferred into bread wheat cv. LAIZHOU 953 from the durum wheat accession DR147 (Zhu et al. 2004). Two powdery mildew resistance genes, formally named *Pm5a* and *Pm4a*, identified in cultivated emmer, were used for bread wheat improvement. *Pm5a* (chr. 7BL, recessive) (McIntosh et al. 1967) appeared in the varieties Hope and H-44 along with *Sr2*, while the dominant gene *Pm4a* (2AL, dominant; The et al. 1979) was transferred to bread wheat variety chancellor from the Indian emmer landrace Khapli (Briggle 1966). WEW is a main source of *Pm* resistance genes—twenty-one—for hexaploid wheat (Huang et al. 2016). A direct transfer from WEW into bread wheat was done for 13 of these, while for the others an identification/mapping after a crossing with durum wheat or a validation/mapping in durum background, followed by transfer into hexaploid wheat, was undertaken (Klymiuk et al. 2019b).

Even if several FHB resistance regions were identified in *Triticum turgidum* ssp., none provides a level of resistance comparable to that of *Fhb1* (3BS) in bread wheat. Until now, only two hard red spring wheat cultivars resistant to FHB, STEELE, and REEDER, were developed from crosses in which two cultivated emmer accessions resistant to FHB were involved (Mergoum et al. 2005; Stack et al. 2003). Hessian fly [Hf—*Mayetiola destructor* (Say) (*Diptera*: *Cecidomyiidae*)] is an important pest of durum and bread wheat (Stuart et al. 2012). To date, 37 Hf resistance genes have been identified (Bassi et al. 2019; Li et al. 2013; Zhao et al. 2020). Among them, 15 (*H6, H9-H11, H14-H19, H28,* and *H29* (all on chr. 1AS), *H31* on chr. 5BS and *H33* on chr. 3A) were identified in durum wheat, and one, *Hdic* (1AS), was derived from an accession of cultivated emmer wheat. Most of them, as for example *H9-H11* and *Hdic*, have been introgressed into common wheat (Patterson et al. 1994; Carlson et al. 1978; Stebbins et al. 1982; Liu et al. 2005), but only few have been deployed in commercial cultivars.

While all the genes described above were identified in tetraploid wheat and then introgressed into bread wheat, several disease-resistant genes moved in the opposite direction. Noteworthy is the case of the introgression and validation of the bread wheat locus *Fhb1* in three European durum wheat genotypes for resistance to Fusarium head blight (FHB) (Prat et al. 2017). Indeed, durum wheat is particularly susceptible to FHB, and limited genetic variation has been found so far within durum modern germplasm. On the contrary, *Fhb1* is a major determinant of FHB resistance found in the hexaploid wheat SUMAI-3 (Anderson et al. 2001). Another successful example is provided by the common wheat broad resistance gene *Lr34/Yr18/Sr57/Pm38/Ltn1* that was transferred into a Canadian cultivar by transgenesis (Rinaldo et al. 2017).

8.4.2 Transfer of Quality-Related Loci

Gpc-B1 (chr. 6BS, also known as *NAM-B1*), a gene responsible for high grain protein and mineral content, was first identified in WEW and cloned by a map-based approach (Uauy et al. 2006), then successfully introgressed into many durum and bread wheat cultivars (Tabbita et al. 2017). While many of the genes described above were identified in tetraploid wheat and then introgressed into bread wheat, several genes moved in the opposite direction. For instance, some glutenin loci responsible for gluten elasticity and extensibility were introgressed into durum to improve the quality of bread made with semolina. Two approaches have been undertaken to introgress bread wheat loci into durum wheat. The *Glu-D1* (chr. 1DL) alleles associated with good baking quality were introgressed in durum wheat either using lines carrying a mutation in *Pairing homolog-1* (*ph1b*) gene, thus promoting homoeologous recombination (Gennaro et al. 2012) or by standard crosses with triticale as intermediate (Lukaszewsky 2003). Similarly, the introgression of *Glu-D1-1d* and *Glu-D1-2b* from bread to durum wheat resulted in dough with stronger mixing features (Gadaleta et al. 2008). *Ph1b*-mediated chromosomal translocations (5DS–5BS) were also employed to produce a tetraploid wheat with soft grains by transferring the *Hardness* locus controlling kernel texture from bread wheat chromosome 5D (Boehm et al. 2017).

8.4.3 Transfer of Abiotic Stress-Related Loci

The introgression of specific alleles at *Vernalization* loci from winter bread wheat has led to durum wheat more adapted to cold environments (Longin et al. 2013). The tolerance to Al^{3+} in acidic soils of durum wheats was also improved by the transfer of *TaMATE1B* and *TaALMT1* (chr. 4B and 4D, respectively) which confer a large tolerance in Al^{3+} tolerance among bread wheats (Delhaize et al. 2012; Han et al. 2016). The *TaMATE1B* gene, responsible for constitutive citrate efflux from root tips (Tovkach et al. 2013), showed a positive effect also on grain yield, probably due to an increased root growth and proliferation. Indeed, the transgenic durum line JANDAROI–*TaMATE1B* compared to JANDAROI per se showed a significantly higher total root biomass and produced from 25.3 to 49.0% higher grain yield under both well-watered and terminal drought conditions (Pooniya et al. 2020).

The other way round, examples of gene transfer between durum and bread wheat for abiotic stress tolerance are more limited due to the complex genetic bases of this trait. The great genetic diversity present in collections of tetraploid wheat accessions, from WEWs to DWCs, represents an asset for future efforts aimed at the identification of loci explaining a good fraction of the phenotypic variation for resistance abiotic stresses, to be introgressed into hexaploid wheat.

8.5 Synthetic Wheats

Hexaploid wheat (*Triticum aestivum* L.), subgenomes AABBDD, is a natural amphiploid derived from interspecific cross between the tetraploid wheat species *T. turgidum* L. (AABB) and the diploid grass *Aegilops tauschii* Coss. (DD). The origin of common wheat is non-monophyletic, indeed, at the time of wheat's origin 10,000 years ago, the formation of more than one interspecific amphiploid contributed to the creation of common wheat (Caldwell et al. 2004). The resulting bottleneck effect has limited its genetic diversity compared with tetraploid and diploid wheats (Cox 1997). For this reason, breeders are looking at tools to increase the genetic diversity to be exploited in mainstream breeding on a worldwide scale. Based on those efforts, two distinct approaches were deployed: production of amphiploids, known as synthetic hexaploids, between *T. turgidum* and *Ae. tauschii*, and direct hybridization between *T. aestivum* and *Ae. tauschii*. Both approaches involve backcrossing to *T. aestivum* (Cox et al. 2017). The direct hybridization approach aims at increasing the genetic diversity for the D

genome, and this is an important issue in wheat breeding, as very low diversity values characterize this genome compared to A and B genomes (Poland et al. 2012; Maccaferri et al. 2015). Studies reporting the improvement of bread wheat for traits related to both agronomic performance (grain yield and quality) and resistance to fungal diseases and pests have been carried out via direct hybridization (Cox et al. 2017). Nevertheless, the reduced genetic diversity following the bottleneck at the origin of bread wheat also involves A and B genomes.

Accordingly, the development of synthetic hexaploid wheats (SHWs) allows for enhancing the diversity for all the three wheat genomes and for the direct transfer of loci for traits of interest from tetraploid to hexaploid wheat. The analysis of the population structure of a panel of 121 SHW lines genotypically characterized with 35,939 high-quality SNPs derived from genotyping-by-sequencing revealed that the percentage of SNPs on the D genome was nearly the same as the other two genomes (nearly 30%), demonstrating the effectiveness of this approach to enhance genetic diversity of the D genome (Bhatta et al. 2018a). When SHW and bread wheat groups were compared at level of the entire genome, the gene diversity of SHWs was from 33.2 to 50% higher compared with a sample of elite bread wheat cultivars in two distinct SHW panels (Bhatta et al. 2018a, 2019a). When these panels were used in GWAS, QTLs for yield and quality-related traits, as well as disease resistance, were identified on all the three genomes, underlying the importance of both durum wheat and *Aegilops* parents in increasing genetic diversity and providing alleles of interest for breeding of bread wheat (Bhatta et al. 2018b, c, 2019b). A relevant contribution of the durum parents has been revealed also for traits usually associated to the D genome. As an example, although tolerance to Al^{3+} toxicity has been mainly linked to the *TaALMT1* gene carried by the D genome (Han et al. 2016), a GWAS with 300 SHW lines besides the effect of *TaALMT1* has identified many other QTLs, mostly located on A and B genomes (Emebiri et al. 2020). Similar results were found in a GWAS with 173 SHWs for leaf, stem and yellow rusts, yellow leaf spot, *Septoria nodorum,* and crown rot (Jighly et al. 2016).

SHW lines have been also used as parents of segregating populations to identify QTLs for specific traits. In these studies, the SHW parent usually incorporates a variable number of loci from the *Ae. tauschii* parent, but some important QTLs are also contributed by the tetraploid parent of the SHW line. Some examples are available for root traits under drought stress conditions. Liu et al. (2020) analyzed a RIL population of 111 individuals derived from a bread wheat cultivar crossed to a SHW line, which incorporated mainly QTLs on the D genome, but also some QTLs, as those on chromosome 2B, which were probably derived from the tetraploid parent. A RIL mapping population derived from a cross between W7984 (synthetic) and OPATA 85 was evaluated for root length and root dry weight under water stress and control conditions. QTLs common to both water conditions and stress specific were identified on A, B, and D genomes (Ayalew et al. 2017). The same population together with a doubled haploid population derived from the same parents (W7984 and OPATA) were used to identify QTLs for number of crossover (Gutierrez-Gonzalez et al. 2019). Similar results, in terms of genetic contribution from A and B genomes, were found for traits related to resistance to stem rust (Dunckel et al. 2015; Sharma et al. 2021) and root rot (Mahoney et al. 2017).

Pshenichnikova et al. (2020) developed a SHW line from a cross between accessions of *T. dicoccoides* and *Ae. tauschii*. The resulting line (SYN6) was crossed with CHINESE SPRING to obtain a set of 21 substitution lines, each containing 20 chromosomes from CHINESE SPRING and one from SYN6. The 1A substitution resulted in a substantial reduction of root length and weight, while a chromosome 5D substitution led to a significant increase compared to the recipient and the donor lines under two contrasting irrigation regimes. Developing synthetic lines through interspecific crosses can have consequences on the genomic asset of the resulting lines. Sequence elimination can happen

after allopolyploidization, increasing the divergence among homoeologous chromosomes. This phenomenon has practical consequences on bread wheat breeding, as it can cause the loss of important genes. A recent study in which the differences between synthetic and natural hexaploid wheat lines were investigated by utilizing a large germplasm set of primary synthetics and synthetic derivatives revealed that reproducible segment elimination occurrence was highly dependent on the choice of diploid and tetraploid parental lines and, that the almost complete short arm of chromosome 1B carrying loci important for grain quality, was eliminated in one line (Jighly et al. 2019). In a different study in which 1862 mapped loci were compared between synthetic wheat SHW-L1 and its parental lines *Ae. tauschii* AS60 (DD) and *T. turgidum* AS2255 (AABB), the D genome of SHW-L1 showed a higher number of eliminated loci following the allopolyploidization compared to the A and B genomes (Yu et al. 2017). At a phenotypic level, hybrid chlorosis can be observed in SHW lines, and genetic loci involved in this phenomenon have been identified on D genome (Nakano et al. 2015; Nishijima et al. 2018).

Despite the possibility of losing some loci of interest, SHW lines have been extensively used in bread wheat breeding. An example is given by the importance of SHW lines in breeding programs at CIMMYT, where more than 1500 SHWs have been developed since 1980s and thousands of crosses have been generated with bread wheat to obtain synthetic lines. With this approach, advanced lines with excellent performance for yield and other traits have been obtained, and more than 80 have also been released as cultivars and are widely grown (Rosyara et al. 2019). Some very promising lines were obtained with adaptation to specific environments. As an example, a large breeding program was aimed at developing and evaluating SHW lines derived from winter durum wheat germplasm from Ukraine and Romania crossed with *Ae. tauschii* accessions from the Caspian Sea region at CIMMYT. These populations, subjected to rigorous pedigree selection under dry, cold, disease-affected environments of the Central Anatolian Plateau, provided superior lines characterized by resistance to leaf, stripe and stem rust, common bunt, and soil-borne pathogens, with the contribution of both durum and *Aegilops* parents (Morgounov et al. 2018).

The breeding programs involving SHWs are also deploying genomic selection. Ninety-seven populations were developed using first back-cross, biparental, and three-way crosses between 33 primary SHW genotypes and 20 spring bread wheat cultivars at CIMMYT. Genomic estimated breeding values (GEBVs) of parents and synthetic derived lines were estimated using a genomic best linear unbiased prediction (GBLUP) model, and higher GEBVs of progenies were related to introgression and retention of positive alleles from SHW parents (Jafarzadeh et al. 2016). Different results were shown by Dunckel et al. (2017), who analyzed selected lines from double haploid and RIL populations between six different primary synthetics and the elite cultivar OPATA M85 chosen for grain yield and other important agronomic traits. Overall, the prediction models had only moderate predictive ability, slightly lower than expected based on traits' heritability. Nevertheless, a more recent study, based on SHW populations and SHW derivatives coming from crosses between the primary SHWs and bread wheat cultivars, suggested that models with heterogeneous additive genetic variances may be suitable to predict breeding values in wheat crosses with variable ploidy levels (Puhl et al. 2021).

In conclusion, the loss of genetic diversity in bread wheat due to bottlenecks from polyploidy, domestication, and modern plant breeding can be compensated by introducing diversity in all the three wheat genomes from *Ae. tauschii* and durum wheat. The increasing use of SHWs and SHW derivatives worldwide and at CIMMYT indicates the success of these approaches in improving bread wheat for many traits of interest, from yield and yield-related traits to resistance to biotic and/or abiotic stresses. Considering the studies based on SHWs, a major limiting factor could be the low number of

durum wheat genotypes used as tetraploid parents of the crosses for the development of primary SHWs, ALTAR 84 and LANGDON being among the most frequently employed. This is in contrast with the large number of *Ae. tauschii* accessions used to broaden the genetic diversity of the D genome. As durum wheat genomics comes of age (Tuberosa and Pozniak 2014) and haplotype variation linked with target phenotypes of key traits becomes increasingly available (Maccaferri et al. 2019; Mazzucotelli et al. 2020), genomics-assisted breeding for durum wheat undergoes a notable evolution, one that will be accelerated by widening the basis of the tetraploid germplasm harnessed for SHW development by choosing the most suitable parents among the most recent and diverse products of the durum wheat breeding, hence increasing the contribution of the tetraploid gene pool to hexaploid wheat improvement.

8.6 Conclusions and Perspectives

Recent studies and breeding approaches clearly indicate that the tetraploid and hexaploid wheats are merging in a unique gene pool, where it will become increasingly easy to recruit genes and alleles for tetraploid and hexaploid wheat breeding regardless of the genome configuration of the donor genotype (Mastrangelo and Cattivelli 2021). Wild and domesticated emmer wheat and *T. turgidum* subspecies and landraces are resources of outstanding importance for breeding of both durum and bread wheat, and there is increasing evidence of élite cultivars carrying genes recruited from emmer and wheat landraces. The different wheat genomes will soon merge in a unique wheat pangenome that will change forever the current breeding strategies. Genes will "shuttle" easily from wild accessions to cultivars and between tetraploid and hexaploid wheats making possible the fast introgression of new traits (e.g., diseases resistance and traits for coping with the effect of climate change) and the selection of varieties increasingly based on genome knowledge which will provide a accurate and holistic view of the wheat genome, while safeguarding and ensuring an adequate level of food security for mankind.

References

Abu-Zaitoun SY, Chandrasekhar K, Assili S et al (2018) Unlocking the genetic diversity within a middle-east panel of durum wheat landraces for adaptation to semi-arid climate. Agronomy 8:233. https://doi.org/10.3390/agronomy8100233

Alemu A, Feyissa T, Letta T et al (2020) Genetic diversity and population structure analysis based on the high density SNP markers in Ethiopian durum wheat (*Triticum turgidum* ssp. *durum*). BMC Genet 21:18. https://doi.org/10.1186/s12863-020-0825-x

Alemu A, Feyissa T, Maccaferri M et al (2021) Genome-wide association analysis unveils novel QTLs for seminal root system architecture traits in Ethiopian durum wheat. BMC Genomics 22:1–16. https://doi.org/10.1186/s12863-020-0825-x

Anderson JA, Stack RW, Liu S et al (2001) DNA markers for Fusarium head blight resistance QTLs in two wheat populations. Theor Appl Genet 102:1164–1168. https://doi.org/10.1007/s001220000509

Aoun M, Rouse MN, Kolmer JA et al (2021) Genome-wide association studies reveal all-stage rust resistance loci in elite durum wheat genotypes. Front Plant Sci 12:640739. https://doi.org/10.3389/fpls.2021.640739

Arenas-M A, Castillo FM, Godoy D et al (2022) Transcriptomic and physiological response of durum wheat grain to short-term heat stress during early grain filling. Plants 11:59. https://doi.org/10.3390/plants11010059

Autrique E, Nachit MM, Monneveux P et al (1996) Genetic diversity in durum wheat based on RFLPs, morphophysiological traits, and coefficient of parentage. Crop Sci 36:735–742. https://doi.org/10.2135/cropsci1996.0011183X003600030036x

Avni R, Nave M, Barad O et al (2017) Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. Science 357:93–97. https://doi.org/10.1126/science.aan0032

Ayalew H, Liu H, Yan G (2017) Identification and validation of root length QTLs for water stress resistance in hexaploid wheat (*Triticum aestivum* L.). Euphytica 213:126. https://doi.org/10.1007/s10681-017-1914-4

Badaeva ED, Keilwagen J, Knüpffer H et al (2015) Chromosomal passports provide new insights into diffusion of emmer wheat. PLoS ONE 10(5):e0128556. https://doi.org/10.1371/journal.pone.0128556

Baloch FS, Alsaleh A, Shahid MQA et al (2017) Whole genome DArTseq and SNP analysis for genetic diversity assessment in durum wheat from central fertile crescent. PLoS ONE 12:e0167821. https://doi.org/10.1371/journal.pone.0167821

Bassi F, Brahmi H, Sabraoui A et al (2019) Genetic identification of loci for Hessian fly resistance in durum wheat. Mol Breed 39:24. https://doi.org/10.1007/s11032-019-0927-1

Bennett FGA (1984) Resistance to powdery mildew in wheat: a review of its use in agriculture and breeding programmes. Plant Pathol 33:279–300. https://doi.org/10.1111/j.1365-3059.1984.tb01324.x

Bhatta M, Morgounov A, Belamkar V et al (2018a) Unlocking the novel genetic diversity and population structure of synthetic hexaploid wheat. BMC Genomics 19:591. https://doi.org/10.1186/s12864-018-4969-2

Bhatta M, Morgounov A, Belamkar V et al (2018b) Genome-wide association study reveals novel genomic regions for grain yield and yield-related traits in drought-stressed synthetic hexaploid wheat. Int J Mol Sci 19:3011. https://doi.org/10.3390/ijms19103011

Bhatta M, Morgounov A, Belamkar V et al (2018c) Genome-wide association study reveals favorable alleles associated with common bunt resistance in synthetic hexaploid wheat. Euphytica 214:200. https://doi.org/10.1007/s10681-018-2282-4

Bhatta M, Shamanin V, Shepelev S et al (2019a) Genetic diversity and population structure analysis of synthetic and bread wheat accessions in Western Siberia. J App Genet 60:283–289. https://doi.org/10.1007/s13353-019-00514-x

Bhatta M, Morgounov A, Belamkar V et al (2019b) Genome-wide association study for multiple biotic stress resistance in synthetic hexaploid wheat. Int J Mol Sci 20:3667. https://doi.org/10.3390/ijms20153667

Boehm JD Jr, Zhang M, Cai X, Morris CF (2017) Molecular and cytogenetic characterization of the 5DS-5BS chromosome translocation conditioning soft kernel texture in durum wheat. Plant Genome 10:3. https://doi.org/10.3835/plantgenome2017.04.0031

Bozzini A (1988) Origin, distribution, and production of durum wheat in the world. In: Fabriani G, Lintas C (eds) Durum: chemistry and technology. American Association of Cereal Chemists, St. Paul, MN, pp 1–16

Briggle LW (1966) Transfer of resistance to *Erysiphe graminis* f. sp. *tritici* from Khapli emmer and Yuma durum to hexaploid wheat. Crop Sci 6:459–461. https://doi.org/10.2135/cropsci1966.0011183X000600050020x

Caldwell KS, Dvorak J, Lagudah ES et al (2004) Sequence polymorphism in polyploid wheat and their D-genome diploid ancestor. Genetics 167:941–947. https://doi.org/10.1534/genetics.103.016303

Carlson SK, Patterson FL, Gallun RL (1978) Inheritance of resistance to Hessian fly derived from *Triticum turgidum* L. Crop Sci 18:1011–1014. https://doi.org/10.2135/cropsci1978.0011183X001800060027x

Cheng P, Xu LS, Wang MN et al (2014) Molecular mapping of genes *Yr64* and *Yr65* for stripe rust resistance in hexaploid derivatives of durum wheat accessions PI 331260 and PI 480016. Theor Appl Genet 127:2267–2277. https://doi.org/10.1007/s00122-014-2378-8

Civáň P, Ivaničová Z, Brown TA (2013) Reticulated origin of domesticated emmer wheat supports a dynamic model for the emergence of agriculture in the fertile crescent. PLoS ONE 8(11):e81955. https://doi.org/10.1371/journal.pone.0081955

Cobb JN, Juma RU, Biswas PS et al (2019) Enhancing the rate of genetic gain in public-sector plant breeding programs: lessons from the Breeder's equation. Theor Appl Genet 132:627–645. https://doi.org/10.1007/s00122-019-03317-0

Condorelli GE, Maccaferri M, Newcomb M et al (2018) Comparative aerial and ground based high throughput phenotyping for the genetic dissection of NDVI as a proxy for drought adaptive traits in durum wheat. Front Plant Sci 9:893. https://doi.org/10.3389/fpls.2018.00893

Cox TS (1997) Deepening the wheat gene pool. J Crop Prod 1:145–168. https://doi.org/10.1300/J144v01n01_01

Cox TS, Wu J, Wang S et al (2017) Comparing two approaches for introgression of germplasm from *Aegilops tauschii* into common wheat. Crop J 5:355–362. https://doi.org/10.1016/j.cj.2017.05.006

Dadkhodaie NA, Karaoglou H, Wellings CR et al (2011) Mapping genes *Lr53* and *Yr35* on the short arm of chromosome 6B of common wheat with microsatellite markers and studies of their association with *Lr36*. Theor Appl Genet 122:479–487. https://doi.org/10.1007/s00122-010-1462-y

David J, Holtz Y, Ranwez V et al (2014) Genotyping by sequencing transcriptomes in an evolutionary pre-breeding durum wheat population. Mol Breed 34:1531–1548. https://doi.org/10.1007/s11032-014-0179-z

De Vita P, Li Destri Nicosia O, Nigro F et al (2007) Breeding progress in morpho-physiological, agronomical and qualitative traits of durum wheat cultivars released in Italy during the 20th century. Eur J Agr 26:39–53. https://doi.org/10.1016/j.eja.2006.08.009

Delhaize E, Ma JF, Ryan PR (2012) Transcriptional regulation of aluminium tolerance genes. Trends Plant Sci 17:341–348. https://doi.org/10.1016/j.tplants.2012.02.008

Desiderio F, Zarei L, Licciardello S et al (2019) Genomic regions from an Iranian landrace increase kernel size in durum wheat. Front Plant Sci 10:448. https://doi.org/10.3389/fpls.2019.00448

Dubcoksky J, Dvorak J (2007) Genome plasticity a key factor in the success of polyploid wheat under domestication. Science 316:1862–1866. https://doi.org/10.1126/science.1143986

Dunckel S, Crossa J, Wu S et al (2017) genomic selection for increased yield in synthetic-derived wheat. Crop Sci 57:713–725. https://doi.org/10.2135/cropsci2016.04.0209

Dunckel SM, Olson EL, Rouse MN et al (2015) Genetic mapping of race-specific stem rust resistance in the synthetic hexaploid W7984 x Opata M85 mapping population. Crop Sci 55:2580–2588. https://doi.org/10.2135/cropsci2014.11.0755

Dvorak J, Akhunov E (2005) Tempos of gene locus deletions and duplications and their relationship to recombination rate during diploid and polyploid evolution in the *Aegilops-Triticum* alliance. Genetics 171:323–332. https://doi.org/10.1534/genetics.105.041632

Emebiri LC, Raman H, Ogbonnaya FC (2020) Synthetic hexaploid wheat as a source of novel genetic loci for aluminium tolerance. Euphytica 216:135. https://doi.org/10.1007/s10681-020-02669-9

Faris J (2014) Wheat domestication: key to agricultural revolutions past and future. In: Tuberosa R, Graner A, Frison E (eds) Genomics of plant genetic resources. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-7572-5_18

Feldman M (2001) Origin of cultivated wheat. In Bonjean AP, Angus WJ (eds) The world wheat book: a history of wheat breeding. Intercept Limited, Andover, England, pp 3–58

Feldman M, Kislev EM (2007) Domestication of emmer wheat and evolution of free-threshing tetraploid wheat. Isr J Plant Sci 55:207–221. https://doi.org/10.1560/IJPS.55.3-4.207

Foley JA, Ramankutty N, Brauman KA et al (2011) Solutions for a cultivated planet. Nature 478:337–342. https://doi.org/10.1038/nature10452

Gadaleta A, Giancaspro A, Blechl AE et al (2008) A transgenic durum wheat line that is free of marker genes and expresses 1Dy10. J Cereal Sci 48:439–445. https://doi.org/10.1016/j.jcs.2007.11.005

Gennaro A, Forte P, Panichi D et al (2012) Stacking small segments of the 1D chromosome of bread wheat containing major gluten quality genes into durum wheat: transfer strategy and breeding prospects. Mol Breed 30:149–167. https://doi.org/10.1007/s11032-011-9606-6

Giraldo P, Royo C, González M et al (2016) Genetic diversity and association mapping for agro-morphological and grain quality traits of a structured collection of durum wheat landraces including subsp. *durum*, *turgidum* and *diccocon*. PLoS One 11:e0166577. https://doi.org/10.1371/journal.pone.0166577

Gutierrez-Gonzalez JJ, Mascher M, Poland J et al (2019) Dense genotyping-by-sequencing linkage maps of two Synthetic W7984×Opata reference populations provide insights into wheat structural diversity. Sci Rep 9:1793. https://doi.org/10.1038/s41598-018-38111-3

Hale I, Zhang X, Fu D et al (2012) Registration of wheat lines carrying the partial stripe rust resistance gene *Yr36* without the *Gpc-B1* high grain protein content allele. J Plant Regist 7:108–112. https://doi.org/10.3198/jpr2012.03.0150crg

Han C, Zhang P, Ryan PR et al (2016) Introgression of genes from bread wheat enhances the aluminium tolerance of durum wheat. Theor Appl Genet 129:729–739. https://doi.org/10.1007/s00122-015-2661-3

Hao Y, Xu S, Lyu Z et al (2021) Comparative analysis of the glutathione S-transferase gene family of four *Triticeae* species and transcriptome analysis of GST genes in common wheat responding to salt stress. Int J Genomics 6289174. https://doi.org/10.1155/2021/6289174

He Y, Feng L, Jiang Y et al (2020) Distribution and nucleotide diversity of *Yr15* in wild emmer populations and chinese wheat germplasm. Pathogens 9:212. https://doi.org/10.3390/pathogens9030212

Herrera-Foessel SA, Singh RP, Huerta-Espino J et al (2008a) Identification and molecular characterization of leaf rust resistance gene *Lr14a* in durum wheat. Plant Dis 92:469–473. https://doi.org/10.1094/PDIS-92-3-0469

Herrera-Foessel S, Singh R, Huerta-Espino et al (2008b) Molecular mapping of a leaf rust resistance gene on the short arm of chromosome 6B of durum wheat. Plant Dis 92:1650–1654. https://doi.org/10.1094/PDIS-92-12-1650

Huang L, Raats D, Sela H et al (2016) Evolution and adaptation of wild emmer wheat populations to biotic and abiotic stresses. Annu Rev Phytopathol 54:279–301. https://doi.org/10.1146/annurev-phyto-080614-120254

Huo N, Zhu T, Zhang S et al (2019) Rapid evolution of *α-gliadin* gene family revealed by analyzing *Gli-2* locus regions of wild emmer wheat. Func Int Genomics 19:993–1005. https://doi.org/10.1007/s10142-019-00686-z

IWGSC (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science 361:7191. https://doi.org/10.1126/science.aar7191

Jafarzadeh J, Bonnett D, Jannink J-L et al (2016) Breeding value of primary synthetic wheat genotypes for grain yield. PLoS ONE 11:e0162860. https://doi.org/10.1371/journal.pone.0162860

Jighly A, Joukhadar R, Sehgal D et al (2019) Population-dependent reproducible deviation from natural bread wheat genome in synthetic hexaploid wheat. Plant J 100:801–812. https://doi.org/10.1111/tpj.14480

Jighly A, Alagu M, Makdis F et al (2016) Genomic regions conferring resistance to multiple fungal pathogens in synthetic hexaploid wheat. Mol Breed 36:127. https://doi.org/10.1007/s11032-016-0541-4

Jin Y, Singh RP, Ward RW et al (2007) Characterization of seedling infection types and adult plant infection responses of monogenic *Sr* gene lines to race TTKS of *Puccinia graminis* f. sp *tritici*. Plant Dis 91:1096–1099. https://doi.org/10.1094/PDIS-91-9-1096

Kabbaj H, Sall AT, Al-Abdallat A et al (2017) Genetic diversity within a global panel of durum wheat (*Triticum durum*) landraces and modern germplasm reveals the history of alleles exchange. Front Plant Sci 8:1277. https://doi.org/10.3389/fpls.2017.01277

Kehel Z, Garcia-Ferrer A, Nachit MM (2013) Using bayesian and eigen approaches to study spatial genetic structure of Moroccan and Syrian durum wheat landraces. Am J Mol Biol 3:17–31. https://doi.org/10.4236/ajmb.2013.31003

Klymiuk V, Yaniv E, Huang L et al (2018) Cloning of the wheat *Yr15* resistance gene sheds light on the plant tandem kinase-pseudokinase family. Nat Commun 9:3735. https://doi.org/10.1038/s41467-018-06138-9

Klymiuk V, Fatiukha A, Fahima T (2019a) Wheat tandem kinases provide insights on disease-resistance gene flow and host-parasite co-evolution. Plant J 98:667–679. https://doi.org/10.1111/tpj.14264

Klymiuk V, Fatiukha A, Huang L et al (2019b) Durum wheat as a bridge between wild emmer wheat genetic resources and bread wheat. In: Miedaner T, Korzun V (eds) Applications of genetic and genomic research in cereals. Woodhead Publishing series in food science, technology and nutrition. Elsevier Ltd., Duxford, U.K, pp 201–230. https://doi.org/10.1016/B978-0-08-102163-7.00010-7

Kolmer J (2008) *Lr63*, *Lr64*. In: McIntosh RA, Dubcovsky J, Rogers WJ et al (eds) Catalogue of gene symbols for wheat: 2009 supplement, p 271 (Reference10550, p 273). Ann Wheat Newsl 55:256–278.

Lai D-l, Yan J, Fan Y et al (2021) Genome-wide identification and phylogenetic relationships of the *Hsp70* gene family of *Aegilops tauschii*, wild emmer wheat (*Triticum dicoccoides*) and bread wheat (*Triticum aestivum*). 3 Biotech 11:301. https://doi.org/10.1007/s13205-021-02639-5

Laidò G, Mangini G, Taranto F et al (2013) Genetic diversity and population structure of tetraploid wheats (*Triticum turgidum* L.) estimated by SSR, DArT and pedigree data. PLoS One 8:e67280. https://doi.org/10.1371/journal.pone.0067280

Li CL, Chen MS, Chao SM et al (2013) Identification of a novel gene, *H34*, in wheat using recombinant inbred lines and single nucleotide polymorphism markers. Theor Appl Genet 126:2065–2071. https://doi.org/10.1007/s00122-013-2118-5

Li M, Dong L, Li B et al (2020) A CNL protein in wild emmer wheat confers powdery mildew resistance. New Phytol 228:1027–1037. https://doi.org/10.1111/nph.16761

Li H, Hua L, Rouse MN et al (2021) Mapping and characterization of a wheat stem rust resistance gene in durum wheat "Kronos." Front Plant Sci 12:751398. https://doi.org/10.3389/fpls.2021.751398

Liu XM, Brown-Guedira GL, Hatchett JH et al (2005) Genetic characterization and molecular mapping of a Hessian fly-resistance gene transferred from *T. turgidum* ssp. *dicoccum* to common wheat. Theor Appl Genet 111:1308–1315. https://doi.org/10.1007/s00122-005-0059-3

Liu W, Maccaferri M, Chen X et al (2017a) Genome-wide association mapping reveals a rich genetic architecture of stripe rust resistance loci in emmer wheat (*Triticum turgidum* ssp. *dicoccum*). Theor Appl Genet 130:2249–2270. https://doi.org/10.1007/s00122-017-2957-6

Liu W, Maccaferri M, Rynearson S et al (2017b) Novel sources of stripe rust resistance identified by genome-wide association mapping in Ethiopian durum wheat (*Triticum turgidum ssp. durum*). Front Plant Sci 8 774. https://doi.org/10.1023/B:GRES.0000024164.80444.f0

Liu R-x, Wu F-k, Yi X et al (2020) Quantitative trait loci analysis for root traits in synthetic hexaploid wheat under drought stress conditions. J Int Agric 19:1947–1960. https://doi.org/10.1016/S2095-3119(19)62825-X

Longin CFH, Sieber A-N, Reif JC et al (2013) Combining frost tolerance, high grain yield and good pasta quality in durum wheat. Plant Breed 132:353–358. https://doi.org/10.1111/pbr.12064

Lopes MS, El-Basyoni I, Baenziger PS et al (2015) Exploiting genetic diversity from landraces in wheat breeding for adaptation to climate change. J Exp Bot 66:3477–3486. https://doi.org/10.1093/jxb/erv122

Lukaszewsky AJ (2003) Registration of six germplasms of bread wheat having variations of cytogenetically engineered wheat–rye translocation 1RS.1BL. Crop Sci 43:1137–1138

Luo MC, Yang ZL, You FM et al (2007) The structure of wild and domesticated emmer wheat populations, gene flow between them, and the site of emmer domestication. Theor Appl Genet 114:947–959. https://doi.org/10.1007/s00122-006-0474-0

Maccaferri M, Sanguineti MC, Noli E et al (2005) Population structure and long-range linkage disequilibrium in a durum wheat elite collection. Mol Breed 15:271–290. https://doi.org/10.1007/s11032-004-7012-z

Maccaferri M, Sanguineti MC, Natoli V et al (2006) A panel of elite accessions of durum wheat (*Triticum durum* Desf.) suitable for association mapping studies. Plant Genet Res 4:79–85. https://doi.org/10.1079/PGR2006117

Maccaferri M, Mantovani P, Tuberosa R et al (2008) A major QTL for durable leaf rust resistance widely exploited in durum wheat breeding programs maps on the distal region of chromosome arm 7BL. Theor Appl Genet 117:1225–1240. https://doi.org/10.1007/s00122-008-0857-5

Maccaferri M, Sanguineti MC, Demontis A et al (2011) Association mapping in durum wheat grown across a broad range of water regimes. J Exp Bot 62:409–438. https://doi.org/10.1093/jxb/erq287

Maccaferri M, Ricci A, Salvi S et al (2015) A high-density, SNP-based consensus map of tetraploid wheat as a bridge to integrate durum and bread wheat genomics and breeding. Plant Biotechnol J 13:648–663. https://doi.org/10.1111/pbi.12288

Maccaferri M, El-Feki W, Nazemi G et al (2016) Prioritizing quantitative trait loci for root system architecture in tetraploid wheat. J Exp Bot 67:1161–1178. https://doi.org/10.1093/jxb/erw039

Maccaferri M, Harris NS, Twardziok SO et al (2019) Durum wheat genome highlights past domestication signatures and future improvement targets. Nat Genet 51:885–895. https://doi.org/10.1038/s41588-019-0381-3

Mahoney AK, Babiker EM, See DR et al (2017) Analysis and mapping of *Rhizoctonia* root rot resistance traits from the synthetic wheat (*Triticum aestivum* L.) line SYN-172. Mol Breed 37:130. https://doi.org/10.1007/s11032-017-0730-9

Mangini G, Nigro D, Margiotta B et al (2018) Exploring SNP diversity in wheat landraces germplasm and setting of a molecular barcode for fingerprinting. Cereal Res Commun 46:377–387. https://doi.org/10.1556/0806.46.2018.033

Mantovani P, Maccaferri M, Sanguineti MC et al (2008) An integrated DArT-SSR linkage map of durum wheat. Mol Breed 4:629–648

Marone D, Del Olmo AI, Laidò G et al (2009) Genetic analysis of durable resistance against leaf rust in du-rum wheat. Mol Breed 24:25–39. https://doi.org/10.1007/s11032-009-9268-9

Marone D, Russo MA, Mores A et al (2021) Importance of landraces in cereal breeding for stress tolerance. Plants 10:1267. https://doi.org/10.3390/plants10071267

Marzario S, Logozzo G, David JL et al (2018) Molecular genotyping (SSR) and agronomic phenotyping for utilization of durum wheat (*Triticum durum* Desf.) ex situ collection from Southern Italy: a combined approach including pedigreed varieties. Genes 9:1–20. https://doi.org/10.3390/genes9100465

Mastrangelo AM, Cattivelli L (2021) What makes bread and durum wheat different? Trends Plant Sci 26:7. https://doi.org/10.1016/j.tplants.2021.01.004

Matsuoka Y (2011) Evolution of polyploid *Triticum* wheats under cultivation: The role of domestication natural hybridization and allopolyploid speciation in their diversification. Plant Cell Physiol 52:750–764. https://doi.org/10.1093/pcp/pcr018

Mazzucotelli E, Sciara G, Mastrangelo AM et al (2020) The Global Durum Wheat Panel (GDP): an international platform to identify and exchange beneficial alleles. Front Plant Sci 11:569905. https://doi.org/10.3389/fpls.2020.569905

McFadden ES (1930) A successful transfer of emmer characters to vulgare wheat. J Am Soc Agron 22:1020–1034

McIntosh RA, Luig NH, Baker EP (1967) Genetic and cytogenetic studies of stem rust, leaf rust, and powdery mildew resistances in Hope and related wheat cultivars. Aust J Biol Sci 20:1181–1192. https://doi.org/10.1071/BI9671181

McIntosh RA, Wellings CR, Park RF (1995) Wheat Rusts: An atlas of resistance genes (Jean K ed). CSIRO, Melbourne, VIC

Mengistu DK, Kiros AY, Pè ME (2015) Phenotypic diversity in Ethiopian durum wheat (*Triticum turgidum* var. *durum*) landraces. Crop J 3:190–199. https://doi.org/10.1016/j.cj.2015.04.003

Mengistu DK, Kidane YG, Catellani M et al (2016) High-density molecular characterization and association mapping in Ethiopian durum wheat landraces reveals high diversity and potential for wheat breeding. Plant Biotechnol J 14:1800–1812. https://doi.org/10.1111/pbi.12538

Mergoum M, Frohberg RC, Miller JD et al (2005) Registration of 'Steele-ND' wheat. Crop Sci 45:1163–1164

Miedaner T, Rapp M, Flath K et al (2019) Genetic architecture of yellow and stem rust resistance in a durum wheat diversity panel. Euphytica 215:71. https://doi.org/10.1007/s10681-019-2394-5

Moragues M, García del Moral LF, Moralejo M et al (2006) Yield formation strategies of durum wheat landraces with distinct pattern of dispersal within the Mediterranean Basin: I. Yield components. Field Crops Res 95:194–205. https://doi.org/10.1016/j.fcr.2005.02.009

Mores A, Borrelli GM, Laidò G et al (2021) Genomic approaches to identify molecular bases of crop resistance to diseases and to develop future breeding strategies. Int J Mol Sci 22(11):5423. https://doi.org/10.3390/ijms22115423

Morgounov A, Abugalieva A, Akan K et al (2018) High-yielding winter synthetic hexaploid wheats resistant to multiple diseases and pests. Plant Genet Res: Charact Utilization 16:273–278. https://doi.org/10.1017/S147926211700017X

N'Diaye A, Haile JK, Nilsen KT et al (2018) Haplotype loci under selection in Canadian durum wheat germplasm over 60 years of breeding: association with grain yield, quality traits, protein loss, and plant height. Front Plant Sci 9:1589. https://doi.org/10.3389/fpls.2018.01589

Nakano H, Mizuno N, Tosa Y et al (2015) Accelerated senescence and enhanced disease resistance in hybrid chlorosis lines derived from interspecific crosses between tetraploid wheat and *Aegilops tauschii*. PLoS ONE 10:e0121583. https://doi.org/10.1371/journal.pone.0121583

Nave M, Taş M, Raupp J et al (2021) The independent domestication of Timopheev's wheat: Insights from haplotype analysis of the *Brittle rachis 1* (*BTR1-A*) gene. Genes 12:338. https://doi.org/10.3390/genes12030338

Nazco R, Villegas D, Ammar K et al (2012) Can Mediterranean durum wheat landraces contribute to improved grain quality attributes in modern cultivars? Euphytica 185:1–17. https://doi.org/10.1007/s10681-011-0588-6

Nevo E, Korol AB, Beiles A et al (2002) Evolution of wild emmer and wheat improvement. Springer, Berlin

Nirmala J, Saini J, Newcomb M et al (2017) Discovery of a novel stem rust resistance allele in durum wheat that exhibits differential reactions to Ug99 isolates. G3-Genes Genom Genet 7:3481–3490. https://doi.org/10.1534/g3.117.300209

Nishijima R, Yoshida K, Sakaguchi K et al (2018) RNA sequencing-based bulked segregant analysis facilitates efficient D-genome marker development for a specific chromosomal region of synthetic hexaploid wheat. Int J Mol Sci 19:3749. https://doi.org/10.3390/ijms19123749

Othmeni L, Grewal S, Hubbart-Edwards S et al (2019) The use of pentaploid crosses for the introgression of *Amblyopyrum muticum* and D-genome chromosome segments into durum wheat. Front Plant Sci 10:1110. https://doi.org/10.3389/fpls.2019.01110

Ouaja M, Bahri BA, Aouini L et al (2021) Morphological characterization and genetic diversity analysis of Tunisian durum wheat (*Triticum turgidum* var. *durum*) accessions. BMC Genom Data 22:3. https://doi.org/10.1186/s12863-021-00958-3

Ozkan H, Willcox G, Graner A et al (2011) Geographic distribution and domestication of wild emmer wheat (*Triticum dicoccoides*). Genet Resour Crop Evol 58:11–15. https://doi.org/10.1007/s10722-010-9581-5

Patterson FL, Mass FB III, Foster JE et al (1994) Registration of eight Hessian fly resistant common winter wheat germplasm lines (Carol, Erin, Flynn, Iris, Joy, Karen, Lola, and Molly). Crop Sci 34:315–316. https://doi.org/10.2135/cropsci1994.0011183X003400010084x

Pecetti L, Annicchiarico P (1998) Agronomic value and plant type of Italian durum wheat cultivars from different eras of breeding. Euphytica 99:9–15. https://doi.org/10.1023/A:1018346901579

Peng JH, Fahima T, Roder MS et al (2000) High-density molecular map of chromosome region harboring stripe rust resistance genes *YrH52* and *Yr15* derived from wild emmer wheat, *Triticum dicoccoides*. Genetica 109:199–210. https://doi.org/10.1023/A:1017573726512

Poland JA, Brown PJ, Sorrells ME et al (2012) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping by-sequencing approach. PLoS ONE 7:e32253. https://doi.org/10.1371/journal.pone.0032253

Pooniya V, Palta JA, Chen Y et al (2020) Impact of the *TaMATE1B* gene on above and below-ground growth of durum wheat grown on an acid and Al^{3+}-toxic soil. Plant Soil 447:73–84. https://doi.org/10.1007/s11104-019-04231-6

Poyarkova H, Gerechteramitai ZK, Genizi A (1991) 2 Variants of wild emmer (*Triticum dicoccoides*) native to Israel: morphology and distribution. Can J Bot 69:2772–2789. https://doi.org/10.1139/b91-348

Prat N, Guilbert C, Prah U et al (2017) QTL mapping of Fusarium head blight resistance in three related durum wheat populations. Theor Appl Genet 130:13–27. https://doi.org/10.1007/s00122-016-2785-0

Pshenichnikova TA, Smirnova OG, Simonov AV et al (2020) The relationship between root system development and vernalization under contrasting irrigation in bread wheat lines with the introgressions from a synthetic hexaploid. Plant Growth Regul 92:583–595. https://doi.org/10.1007/s10725-020-00666-5

Puhl LE, Crossa J, Munilla S et al (2021) Additive genetic variance and covariance between relatives in synthetic wheat crosses with variable parental ploidy levels. Genetics 217:iyaa048. https://doi.org/10.1093/genetics/iyaa048

Quick JS, Miller JD, Donnelly BJ (1976) Cando North Dakota's first semidwarf durum. North Dakota Farm Res 33:15–18

Qureshi N, Bariana H, Kumran VV et al (2018) A new leaf rust resistance gene *Lr79* mapped in chromosome 3BL from the durum wheat landrace Aus26582. Theor Appl Genet 131:1091–1098. https://doi.org/10.1007/s00122-018-3060-3

Rasheed A, Xia X (2019) From markers to genome-based breeding in wheat. Theor Appl Genet 132:767–784. https://doi.org/10.1007/s00122-019-03286-4

Rebetzke GJ, Richards RA, Fettell NA et al (2007) Genotypic increases in coleoptile length improves stand establishment, vigour and grain yield of deep-sown wheat. Field Crop Res 100:10–23. https://doi.org/10.1016/j.fcr.2006.05.001

Reif JC, Zhang P, Dreisigacker S et al (2005) Wheat genetic diversity trends during domestication and breeding. Theor Appl Genet 110:859–864. https://doi.org/10.1007/s00122-004-1881-8

Reimer S, Pozniak CJ, Clarke FR et al (2008) Association mapping of yellow pigment in an elite collection of durum wheat cultivars and breeding lines. Genome 51:1016–1025. https://doi.org/10.1139/g08-083

Ren J, Sun D, Chen L et al (2013) Genetic diversity revealed by single nucleotide polymorphism markers in a worldwide germplasm collection of durum wheat. Int J Mol Sci 14:7061–7088. https://doi.org/10.3390/ijms14047061

Rinaldo A, Gilbert B, Boni R et al (2017) The *Lr34* adult plant rust resistance gene provides seedling resistance in durum wheat without senescence. Plant Biotechnol J 15:894–905. https://doi.org/10.1111/pbi.12684

Robbana C, Kehel Z, Ben Naceur M et al (2019) Genome-Wide genetic diversity and population structure of Tunisian durum wheat landraces based on DArTseq technology. Int J Mol Sci 20:1352. https://doi.org/10.3390/ijms20061352

Roelfs AP, McVey DV (1979) Low infection types produced by *Puccinia graminis* f. sp. *tritici* and wheat lines with designated genes for resistance. Phytopathol 69:722–730. https://doi.org/10.1094/phyto-69-722

Roncallo PF, Beaufort V, Larsen AO et al (2019) Genetic diversity and linkage disequilibrium using SNP (KASP) and AFLP markers in a worldwide durum wheat (*Triticum turgidum* L. var *durum*) collection. PLoS One 14:e0218562. https://doi.org/10.1371/journal.pone.0218562

Rosyara U, Kishii M, Payne T et al (2019) Genetic contribution of synthetic hexaploid wheat to CIMMYT's spring bread wheat breeding germplasm. Sci Rep 9:12355. https://doi.org/10.1038/s41598-019-47936-5

Royo C, Nazco R, Villegas D (2014) The climate of the zone of origin of Mediterranean durum wheat (*Triticum durum* Desf.) landraces affects their agronomic performance. Genet Resour Crop Evol 61:1345–1358. https://doi.org/10.1007/s10722-014-0116-3

Ruiz M, Giraldo P, Royo C et al (2012) Diversity and genetic structure of a collection of Spanish durum wheat landraces. Crop Sci 52:2262–2275. https://doi.org/10.1371/journal.pone.0166577

Saccomanno A, Matny O, Marone D et al (2018) Genetic mapping of loci for resistance to stem rust in a tetraploid wheat collection. Int J Mol Sci 19:3907. https://doi.org/10.3390/ijms19123907

Sahri A, Chentoufi L, Arbaoui M et al (2014) Towards a comprehensive characterization of durum wheat landraces in Moroccan traditional agrosystems: analysing genetic diversity in the light of geography, farmers' taxonomy and tetraploid wheat domestication history. BMC Evol Biol 14:264. https://doi.org/10.1186/s12862-014-0264-2

Sanguineti MC, Li S, Maccaferri M et al (2007) Genetic dissection of seminal root architecture in elite durum wheat germplasm. Ann Appl Biol 151:291–305 https://doi.org/10.1111/j.1744-7348.2007.00198.x

Scarascia Mugnozza GT (2005) The contribution of Italian wheat geneticists: from Nazareno Strampelli to Francesco D'Amato. In: Proceedings of International Congress on the wake of the double helix: from the green devolution to the gene revolution, Bologna, Italy, pp 52–75

Seyedimoradi H, Talebi R, Fayaz F (2016) Geographical diversity pattern in Iranian landrace durum wheat (*Triticum turgidum*) accessions using start codon targeted poly morphism and conserved DNA-derived polymorphism markers. Env Expl Bio 14:63–68. https://doi.org/10.22364/eeb.14.09

Sharma JS, Overlander M, Faris JD et al (2021) Characterization of synthetic wheat line Largo for resistance to stem rust. G3 11:jkab193. https://doi.org/10.1093/g3journal/jkab193

Sheen SJ, Snyder LA (1964) Studies on the inheritance of resistance to six stem rust cultures using chromosome substitution lines of a Marquis wheat selection. Can J Genet Cytol 6:74–82. https://doi.org/10.1139/g64-010

Si Y, Zheng S, Niu J et al (2021) *Ne2*, a typical CC–NBS–LRR-type gene, is responsible for hybrid necrosis in wheat. New Phytol 232:279–289. https://doi.org/10.1111/nph.17575

Sibikeev S, Druzhin A, Gultyaeva E et al (2020) Use of the durum wheat gene pool in breeding of spring bread wheat. Russ Agric Sci 46:432–436. https://doi.org/10.3103/S1068367420050201

Simons K, Abate Z, Chao S et al (2011) Genetic mapping of stem rust resistance gene *Sr13* in tetraploid wheat (*Triticum turgidum* ssp. *durum* L.). Theor Appl Genet 122:649–658. https://doi.org/10.1007/s00122-010-1444-0

Singh RP, Hodson DP, Huerta-Espino J et al (2011) The emergence of Ug99 races of the stem rust fungus is a threat to world wheat pro-duction. Annu Rev Phytopathol 49:465–481. https://doi.org/10.1146/annurev-phyto-072910-095423. (PMID: 21568701)

Slim A, Piarulli L, Chennaoui Kourda H et al (2019) Genetic structure analysis of a collection of Tunisian durum wheat germplasm. Int J Mol Sci 20:3362. https://doi.org/10.3390/ijms20133362

Soares JC, Santos CS, Carvalho SMP et al (2019) Preserving the nutritional quality of crop plants under a changing climate: importance and strategies. Plant Soil 443:1–26. https://doi.org/10.1007/s11104-019-04229-0

Soriano JM, Villegas D, Aranzana MJ et al (2016) Genetic structure of modern durum wheat cultivars and Mediterranean landraces matches with their agronomic performance. PLoS ONE 11:e0160983. https://doi.org/10.1371/journal.pone.0160983

Soriano JM, Villegas D, Sorrells ME et al (2018) Durum wheat landraces from east and west regions of the Mediterranean basin are genetically distinct for yield components and phenology. Front Plant Sci 9:80. https://doi.org/10.3389/fpls.2018.00080

Soriano JM, Colasuonno P, Marcotuli I et al (2021) Meta-QTL analysis and identification of candidate genes for quality, abiotic and biotic stress in durum wheat. Sci Rep 11:11877. https://doi.org/10.1038/s41598-021-91446-2

Srichumpa P, Brunner S, Keller B et al (2005) Allelic series of four powdery mildew resistance genes at the *Pm3* locus in hexaploid bread wheat. Plant Physiol 139:885–895. https://doi.org/10.1104/pp.105.062406

Stack RW, Frohberg RC, Hansen JM et al (2003) Transfer and expression of resistance to Fusarium head blight from wild emmer chromosome 3A to bread wheat. In: Lewis J, Ward RW (eds) Proceedings of national fusarium head blight forum, Bloomington, MN. Wheat and Bar-ley Scab Initiative, Washington, DC, p 232

Stebbins NB, Patterson FL, Gallun RL (1982) Interrelationships among wheat genes *H3*, *H6*, *H9*, and *H10* for Hessian fly resistance. Crop Sci 22:1029–1032. https://doi.org/10.2135/cropsci1982.0011183X002200050032x

Stuart JJ, Chen MS, Shukle R et al (2012) Gall midges (Hessian flies) as plant pathogens. Annu Rev Phytopathol 50:339–357. https://doi.org/10.1146/annurev-phyto-072910-095255

Tabbita F, Pearcec S, Barneix AJ (2017) Breeding for increased grain protein and micronutrient content in wheat: Ten years of the *GPC-B1* gene. J Cereal Sci 73:183–191. https://doi.org/10.1016/j.jcs.2017.01.003

The TT, McIntosh RA, Bennett FGA (1979) Cytogenetical studies in wheat. IX. Monosomic analyses, telocentric mapping and linkage relationships of genes *Sr21*, *Pm4* and *Mle*. Aust J Biol Sci 32:115–125. https://doi.org/10.1071/BI9790115

Tovkach A, Ryan PR, Richardson AE et al (2013) Transposon-mediated alteration of *TaMATE1B* expression in wheat confers constitutive citrate efflux from root apices. Plant Physiol 161:880–888. https://doi.org/10.1104/pp.112.207142

Tuberosa R, Pozniak C (2014) Durum wheat genomics comes of age. Mol Breed 34:1527–1530. https://doi.org/10.1007/s11032-014-0188-y

Uauy C, Distelfeld A, Fahima T et al (2006) A *NAC* gene regulating senescence improves grain protein, zinc, and iron content in wheat. Science 314:1298–1301. https://doi.org/10.1126/science.1133649

Vavilov N (1951) The origin, variation, immunity and breeding of cultivated crops. Chronicles Bot 13:1–36

Vendramin V, Ormanbekova D, Scalabrin S et al (2019) Genomic tools for durum wheat breeding: de novo assembly of Svevo transcriptome and SNP discovery in elite germplasm. BMC Genomics 20:1–16. https://doi.org/10.1186/s12864-019-5645-x

Walkowiak S, Gao L, Monat C et al (2020) Multiple wheat genomes reveal global variation in modern breeding. Nature 588:277–283. https://doi.org/10.1038/s41586-020-2961-x

Wang S, Wong D, Forrest K et al (2014) Characterization of polyploid wheat genomic diversity using a high-density 90,000 SNP array. Plant Biotechnol J 12:787–796. https://doi.org/10.1111/pbi.12183

Waugh DC (2010) The silk roads in history. Expedition 52:9–22

Xu LS, Wang MN, Cheng P et al (2013) Molecular mapping of *Yr53*, a new gene for stripe rust resistance in durum wheat accession PI 480148 and its transfer to common wheat. Theor Appl Genet 126:523–533. https://doi.org/10.1007/s00122-012-1998-0

Yahiaoui N, Brunner S, Keller B (2006) Rapid generation of new powdery mildew resistance genes after wheat domestication. Plant J 47:85–98. https://doi.org/10.1111/j.1365-313X.2006.02772.x

Yang X, Yu H, Sun W et al (2021) Wheat in vivo RNA structure landscape reveals a prevalent role of RNA structure in modulating translational subgenome expression asymmetry. Genome Biol 22:326. https://doi.org/10.1186/s13059-021-02549-y

Yaniv E, Raats D, Ronin Y et al (2015) Evaluation of marker-assisted selection for the stripe rust resistance gene *Yr15*, introgressed from wild emmer wheat. Mol Breed 35:43. https://doi.org/10.1007/s11032-015-0238-0

Yu M, Guan L-L, Chen G-Y et al (2017) Allopolyploidy-induced rapid genomic changes in newly generated synthetic hexaploid wheat. Biotechnol Biotechnol Equip 31:236–242. https://doi.org/10.1080/13102818.2016.1273797

Zhang H, Zhang L, Wang C et al (2016) Molecular mapping and marker development for the *Triticum dicoccoides*-derived stripe rust resistance gene *YrSM139-1B* in bread wheat cv. Shaanmai 139. Theor Appl Genet 129:369–376. https://doi.org/10.1007/s00122-015-2633-7

Zhang W, Chen S, Abate Z et al (2017) Identification and characterization of *Sr13*, a tetraploid wheat gene that confers resistance to the Ug99 stem rust race group. Proc Natl Acad Sci USA 114:E9483–E9492. https://doi.org/10.1073/pnas.1706277114

Zhao L, Abdelsalam NR, Xu Y et al (2020) Identification of two novel Hessian fly resistance genes *H35* and *H36* in a hard winter wheat line SD06165. Theor Appl Genet 133:2343–2353. https://doi.org/10.1007/s00122-020-03602-3

Zhu Z, Kong X, Zhou R et al (2004) Identification and microsatellite markers of a resistance gene to powdery mildew in common wheat introgressed from *Triticum durum*. Acta Botan Sin 46:867–872

Zhu T, Wang L, Rodriguez JC et al (2019) Improved genome sequence of wild emmer wheat Zavitan with the aid of optical maps. G3 (Bethesda) 9:619–624. https://doi.org/10.1534/g3.118.200902

9 Genome-Informed Discovery of Genes and Framework of Functional Genes in Wheat

Awais Rasheed, Humaira Qayyum and Rudi Appels

Abstract

The complete reference genome of wheat was released in 2018 (IWGSC in Science 361:eaar7191, 2018), and since then many wheats genomic resources have been developed in a short period of time. These resources include resequencing of several hundred wheat varieties, exome capture from thousands of wheat germplasm lines, large-scale RNAseq studies, and complete genome sequences with de novo assemblies of 17 important cultivars. These genomic resources provide impetus for accelerated gene discovery and manipulation of genes for genetic improvement in wheat. The groundwork for this prospect includes the discovery of more than 200 genes using classical gene mapping techniques and comparative genomics approaches to explain moderate to major phenotypic variations in wheat. Similarly, QTL repositories are available in wheat which are frequently used by wheat genetics researchers and breeding communities for reference. The current wheat genome annotation is currently lagging in pinpointing the already discovered genes and QTL, and annotation of such information on the wheat genome sequence can significantly improve its value as a reference document to be used in wheat breeding. We aligned the currently discovered genes to the reference genome, provide their position and *TraesID*s, and present a framework to annotate such genes in future.

Keywords

Wheat genomics · Single nucleotide polymorphisms (SNPs) · KASP markers · Gene discovery · Functional markers · Gene networks

9.1 Introduction

Wheat holds a central position among major food crops by providing 20% of the total caloric requirements for the humans around the world. Common wheat (*Triticum aestivum* L.) is an allohexaploid ($2n = 6x = 42$; AABBDD) crop successfully cultivated all over the world covering an area of approximately 220 million ha. Genetic improvement in wheat productivity,

A. Rasheed (✉) · H. Qayyum
Quaid-i-Azam University, Islamabad 45320, Pakistan
e-mail: arasheed@qau.edu.pk

A. Rasheed
Chinese Academy of Agricultural Sciences (CAAS), and CIMMYT-China Office, Beijing, China

R. Appels
University of Melbourne, Food and Nutrition, Parkville, and AgriBio (Latrobe University), Bundoora, Melbourne, Australia

R. Appels et al. (eds.), *The Wheat Genome*, Compendium of Plant Genomes,
https://doi.org/10.1007/978-3-031-38294-9_9

resilience to climate extremes, and quality are challenges to be met in continuing to feed the global population, mitigate the effects of climate change, and fulfill the end user quality preferences. Since the expansion of wheat production area will not be possible due to the continuous shrinking of arable land, the increase in the grain yield by improved agronomic practices and breeding are feasible approaches. It has been recognized that conventional crop breeding approaches are not able to deliver the target of 70% increase in crop productivity by the end of 2050 (Tester and Langridge 2010). The innovation required in all breeding components includes selection accuracy, selection intensity, deploying new genetic variations, and shortening of the breeding cycles in developing cultivars (Li et al. 2018).

Conventional plant breeding heavily has relied on the selection of key phenotypes related to yield-related traits such as harvest index in wheat (Lopes et al. 2012), and it seems impossible to further improve harvest index using conventional breeding. Secondly, the phenotypic-based selections are labor intensive and time consuming, and off-spring can only be selected at the certain homozygous generation at the later growth stages. The concept of genomics-assisted breeding (GAB) was proposed as an alternate to overcome the selection challenges associated with conventional breeding (Varshney et al. 2005). The marker-assisted selection component dominated in the breeding programs where the diagnostic markers for the genes with major phenotypic effects were developed and successfully used for selection (Liu et al. 2012). However, many complex traits such as yield and adaptability to stressed environments are controlled by many genes with minor effects or quantitative trait loci (QTL), further interacting with environment (Gao et al. 2015). Their individual effects are too small to be efficiently captured by one or few markers (Bernardo and Yu 2007). Therefore, a transition from marker to genome-based breeding is indispensable to achieve the productivity targets (Rasheed and Xia 2019).

The next-generation sequencing (NGS) has revolutionized plant genomics and resulted in development of techniques and resources amenable to plant breeding (Bevan et al. 2017). The ever-growing plant genomic resources have provided plethora of SNP information distributed throughout the plant genomes, which have made them markers of choice for a variety of research applications, especially in breeding and genetics research. Until now, the reference genome sequences are available for most of the crop species, including wheat, while pan-genome sequences are increasing with the rapid pace (see Chap. 14). Characterization of the pan-genome can rapidly identify variations within the candidate genes, which have a direct application in breeding. In this chapter, we discuss different genome-informed scenarios being pursued to discover genes underpinning important phenotypes (Blake et al. 2016). We also provide a framework of functional genes of wheat in the context of the recent reference genome sequence assembly and discuss database resources necessary to reduce redundancy in research.

9.2 Wheat Reference Genome Sequence and Other Genomic Resources

9.2.1 The Reference Genome Sequence of cv. CHINESE SPRING

Wheat has a history in being used a model plant for understanding cytogenetics, physical mapping of genes, and to facilitate pre-breeding to introduce inter-specific and intergenic diversity. For example, the array of wheat aneuploid stocks, unequaled in any other crop, was developed by Sears (1954). All these genetic stocks were developed using wheat cv. CHINESE SPRING (Sears and Sears 1978). Such aneuploids include all the possible chromosome addition or deletion lines in the form of nullisomics, trisomics, monosomics, and tetrasomics. These cytogenetic stocks greatly facilitated

the genetic studies which were not possible in many of the higher organisms at that time. These stocks were used to identify major genes controlling important traits and physically map their positions along chromosomes, including the genes related to waxiness, maturity, endosperm proteins, and vernalization (Driscoll and Jensen 1964; Shepherd 1968; Halloran and Boydell 1967; Law 1966). Later, these efforts provided the basis for starting a 'Catalogue of Gene Symbols for Wheat' to catalogue wheat genes (McIntosh 1973). Since a wide array of genetic stocks were available in the CHINESE SPRING wheat background, this cultivar was selected to develop the first reference genome sequence in wheat. The International Wheat Genome Sequencing Consortium (IWGSC) was established in 2005, and after 13 years of its establishment, the high-quality reference sequence was released in 2018 (IWGSC 2018).

9.2.2 Other Genomic Resources in Wheat

All genome sequence resources available in wheat to date are provided in Table 9.1 and include population-level whole-genome resequencing, exome sequencing, and to lesser extent some SNP genotyping resources. The analysis of the CHINESE SPRING reference genome is now complemented by de novo sequences of ten important wheat cultivars from global breeding programs and has allowed the documentation of breeding histories, wild introgressions in the cultivated wheat, and chromosomal structural rearrangements that facilitated wheat breeding (Walkowiak et al. 2020); Jayakodi et al. 2021). Apart from the sequencing efforts in cultivated wheat, the genome sequences of diploid and tetraploid progenitors of bread wheat including *Ae. Tauschii* (Zhao et al. 2017), *T. monococcum* (Ling et al. 2018), *Ae. Speltoides* (Avni et al. 2022), and *T. dicoccoides* (Avni et al. 2017) are available. Recently, a population-level genome sequence resource of global *Ae. Tauschii* accessions was provided for use in trait discovery and functional genetic validation of D-genome introgressions in bread wheat (Gaurav et al. 2022). The shared utility of all such resources is underpinning the assignment of functional attributes to genes through association genetics or by selective sweeps. For example, 120 Chinese wheat cultivars and landraces were resequenced, and it was identified that the D-subgenome of modern cultivars is mostly derived from landraces, while A- and B-subgenomes were mainly derived from European landraces (Chen et al. 2019). Strong signals of selective sweeps were restricted to 48 high-confidence (HC) genes selected during modern wheat breeding. The strongest signals were for genes *TaNPF6.1-6B*, *TaNAC24*, and *TaRVE3*, which are associated with nitrogen use efficiency, drought and heat stress tolerance, and flowering time, respectively (Chen et al. 2019).

The exome capture of more than 500 global wheat accessions was conducted to identify the genes underpinning selection of adaptation of modern-day bread wheat during last 10,000 years (Pont et al. 2019). The authors concluded that dispersion of wheat and human migration patterns were consistent with an origin out of the Fertile Crescent and Egypt to Maghreb (Northern Africa) with a coastal route. The major driving forces in wheat adaptation were the vernalization requirement, historical groupings, and geographic origins (Europe, Asia, Africa, and America) and thus resulted in the partitioning of the genetic diversity in wheat. Furthermore, a total of 168 Mb of genome regions on different chromosomes contained selective sweeps which were identical between the Asian and European germplasm, even though European wheats had more frequent introgressions compared to wheats from Eastern Asia (He et al. 2019; Zhou et al. 2020), based on the resequencing of 890 bread and durum wheat accessions and the identification of introgressions from wild species favoring global wheat adaptation. Another globally important genomic resource is the DArTseq database of 44,624 wheat accessions from the International Maize

Table 9.1 Wheat genomic resources post-reference genome sequence

Resource	Number of accessions	Sequencing strategy	Objective	Reference
Pan-genome	10	WGS	Build a pan-genome of wheat	Walkowiak et al. (2020)
Chinese accessions	120	WGR	Identify the selection regions during wheat breeding	Chen et al. (2019)
Global landraces and cultivated wheats	4506	280 K SNP array	Wheat phylogeography and genetic diversity	Balfourier et al. (2019)
Global wheat accessions	500	Exome sequencing	Years of hybridization, selection, adaptation, and plant breeding has shaped the genetic makeup of modern bread wheats	Pont et al. (2019)
Hexaploid/tetraploid accessions	890	Exome sequencing	Identify the wild-relative introgressions favoring global wheat adaptation	He et al. (2019)
Chinese wheat accessions	770	DArTseq/660 K	Dispersion history, adaptive evolution, and selection of wheat in China	Zhou et al. (2018)
CIMMYT germplasm	44,624	DArTseq	Genomic predictabilities of 35 key traits and demonstrate the potential of genomic selection for wheat end-use quality	Juliana et al. (2019)
Ae. tauschii global collection	242	WGR	D-genome diversity for gene discovery	Gaurav et al. 2022
Chinese minicore collection	287	Exome sequencing	Identify genetic regions associated yield and adaptability	Li et al. (2022a)
Elite cultivars of China	145	WGR	Seventy years of breeder-driven selection	Hao et al. (2020)
25 wild wheat populations	414	WGR	Introgression from wild populations	Zhou et al. (2020)
Aegilops tauschii	278	WGR	Novel haplotypes with potential applications in wheat improvement	Zhou et al. (2021)

WGS: Whole-genome sequencing; WGR: Whole-genome resequencing

and Wheat Improvement Center (CIMMYT) GenBank (Juliana et al. 2019). The DArTseq data was used to conduct genome-wide association studies (GWAS) for 50 different traits of breeding interest and identified important loci for end-use quality, biotic, and abiotic stress resistances. These studies provide a deep insight into genetic diversity and genetic regions in wheat under artificial and natural selection and will keep proving important resources for use of such information in breeding.

9.3 Wheat Functional Genes Discovery: Strategies and Inventory

Quantitative trait loci (QTL) mapping and GWAS have dominated wheat genomics research to date. These studies identify the favorable alleles and their diagnostic markers which can be then used in wheat breeding to introgress important QTL or genes (Rasheed and Xia 2019). In Table 9.2, we provide a near-to-complete framework of the functional genes discovered so far by such approaches. However, such genetic dissection especially in case of GWAS can be ambiguous due to the confounding effects of population structure or low-accuracy genotype calls at some loci (Browning and Yu 2009), or due to the small population size (Finno et al. 2014). It is, therefore, necessary to further validate the phenotypic effects of such loci in biparental mapping populations or other genetic backgrounds, as well as by other biological means such as genetic transformation, gene silencing or gene knockout, and gene editing. The population-level whole-genome resequencing or exome capture data facilitated the discovery of several genes for economically important traits. From the resequencing data of 145 Chinese wheat accessions, Hao et al. (2020) identified that *TaFRK2-7A* gene contained three non-synonymous mutations compared to CS allele and was strongly associated with starch and amylose contents in mature seeds. The exome sequence of 287 wheat accessions identified the causal variations in *TaARF12* encoding an auxin response factor and *TaDEP1* encoding the G-protein γ-subunit, pleiotropically regulating both plant height and grain weight in wheat (Li et al. 2022a, b).

In recent years, several loci were identified simultaneously by GWAS and biparental mapping strategies. Liu et al. (2017) identified marker-trait association for black point resistance. Loci underpinning flour color (Zhai et al. 2016), kernel number per spike (Shi et al. 2017), and thousand grain weight (Sehgal et al. 2020; Wang et al. 2021) were also identified following a similar strategy. A functional gene, *TaRPP13L1* associated with flour color, was identified by GWAS in wheat cultivars from China and two KRONOS wheat mutants carrying premature stop codons of the *TaRPP13L1* gene and was thus validated as a gene influencing flour color (Chen et al. 2019).

Another gene discovery approach which is now widely used is bulk segregation analysis (BSA), where DNA from individuals of a population showing contrasting, extreme, and phenotypes is pooled and then RNAseq, exome sequencing, or whole-genome resequencing is applied (Zou et al. 2016). This is a rapid method to identify consistent polymorphic regions between contrasting pools of wheat lines. In addition to the discovery of SNPs between contrasting pools, differentially expressed genes can also be identified in the case of RNAseq analysis of tissues. Using this approach, a QTL interval with four candidate genes has been discovered on chr4A underpinning resistance against orange wheat blossom midge (OWBM) affecting wheat production in many countries (Hao et al. 2019). Likewise, resistance to yellow rust in wheat cultivar ZHOUMAI 22 was delimited to a physical interval of 4 Mb using BSA and RNAseq approach (Wang et al. 2017a). Other studies where this approach has been effective in discovering candidate genes include

Table 9.2 Framework of functional genes characterized in wheat with positions in wheat genome and associated traits

Gene	Chr	Phenotype	Crop ontology	Position	Traes ID
TaNAAT1-A	1A	GFe/GZn	CO_321:0000224	chr1A:487330367.0.487333932	TraesCS1A02G291100
TaNAAT2-A	1A	GFe/GZn	CO_321:0000224	chr1A:487463045.0.487466385	TraesCS1A02G291200
Glu-A1	1A	Gluten/end-use quality	CO_321:0000152	chr1A:508723999.0.508726319	TraesCS1A02G317311
Glu-A3	1A	Gluten/end-use quality	CO_321:0000155	chr1A:4202215.0.4203588	TraesCS1A02G008000
TaPYL1-1B	1B	Drought tolerance	CO_321:0000131	chr1B:373628259.0.373629490	TraesCS1B02G206600
TOE-B1	1B	Flowering time	CO_321:0000007	chr1B:59192897.0.59197677	TraesCS1B02G076300
ELF3-B1	1B	Flowering time	CO_321:0000007	chr1B:685645287.0.685649392	TraesCS1B02G477400
TaFT3-B1	1B	Flowering time	CO_321:0000007	chr1B:581413558.0.581414952	TraesCS1B02G351100
TaNAAT1-B	1B	GFe/GZn	CO_321:0000224	chr1B:520925847.0.520929216	TraesCS1B02G300500
TaNAAT2-B	1B	GFe/GZn	CO_321:0000224	chr1B:520998902.0.521002315	TraesCS1B02G300600
Glu-B1-717	1B	Gluten/End-use quality	CO_321:0000153	chr1B:555765127.0.555766152	TraesCS1B02G329711
Glu-B3	1B	Gluten/End-use quality	CO_321:0000156	chr1B:5686611.0.5687693	TraesCS1B02G011700
AGP-L-1B	1B	Grain morphology	CO_321:0000040	chr1B:668129122.0.668132472	TraesCS1B02G449700
Elf3-D1	1D	Flowering time	CO_321:0000007	chr1D:493484553.0.493488588	TraesCS1D02G451200
Mot-D1	1D	Flowering time	CO_321:0000007	chr1D:492606158.0.492620025	TraesCS1D02G450200
TaNAAT1-D	1D	GFe/GZn	CO_321:0000224	chr1D:387796590.0.387800918	TraesCS1D02G289700
TaNAAT2-D	1D	GFe/GZn	CO_321:0000224	chr1D:387894784.0.387898194	TraesCS1D02G289800
Glu-D1	1D	Gluten/end-use quality	CO_321:0000154	chr1D:412160786.0.412163311	TraesCS1D02G317211
ZDS-A1	2A	Flour color	CO_321:0000214	chr2A:321150418.0.321156866	TraesCS2A02G238400
Ppd-A1	2A	Flowering time	CO_321:0000007	chr2A:36933684.0.36938202	TraesCS2A02G081900
TaVIT1-2A	2A	GFe	CO_321:0000222	chr2A:570192811.0.570195203	TraesCS2A02G336600
TaNAS1-A	2A	GFe/GZn	CO_321:0000224	chr2A:14976663.0.14978691	TraesCS2A02G033500
TaNAS3-A	2A	GFe/GZn	CO_321:0000224	chr2A:19162944.0.19164224	TraesCS2A02G049900
TaNAS9-A	2A	GFe/GZn	CO_321:0000224	chr2A:49221108.0.49222130	TraesCS2A02G095700
Sus2-2A	2A	Grain morphology	CO_321:0000040	chr2A:121141338.0.121145857	TraesCS2A02G168200
TaCwi-A1	2A	Grain morphology	CO_321:0000040	chr2A:508030243.0.508033950	TraesCS2A02G295400
TaCYP78A5	2A	Grain morphology	CO_321:0000040	chr2A:134273284.0.134275604	TraesCS2A02G175700
WFZP-A1	2A	Grain number	CO_321:0000391	chr2A:66848645.0.66849948	TraesCS2A02G116900
TaGS2-A1	2A	NUE	CO_321:0001671	chr2A:729293649.0.729297303	TraesCS2A02G500400
TaARF12	2A	Plant height	CO_321:0000020	chr2A:755768802.0.755776624	TraesCS2A02G547800
PPO-A1	2A	PPO activity	CO_321:0000214	chr2A:712187112.0.712189567	TraesCS2A02G468200
Ppo2-A1	2A	PPO activity	CO_321:0000214	chr2A:712344578.0.712346518	TraesCS2A02G468500
RMD-A1	2A	Root growth angle		chr2A:142707925.0.142709726	TraesCS2A02G182900
TaRSL4	2A	Root length		chr2A:162291365.0.162292945	TraesCS2A02G194200
Sdr-A1	2A	Seed dormancy/PHS	CO_321:0000081	chr2A:158452418.0.158453410	TraesCS2A02G191400
Ppd-B1	2B	Flowering time	CO_321:0000007	chrUn:293689186.0.293692375	TraesCSU02G196100
TaVIT1-2B	2B	GFe	CO_321:0000222	chr2B:492146188.0.492148400	TraesCS2B02G345300
TaNAS1-B	2B	GFe/GZn	CO_321:0000224	chr2B:23548049.0.23551608	TraesCS2B02G047100
TaNAS3-B	2B	GFe/GZn	CO_321:0000224	chr2B:29118956.0.29120236	TraesCS2B02G060800
TaNAS9-B	2B	GFe/GZn	CO_321:0000224	chr2B:72895029.0.72896639	TraesCS2B02G111100
TaSus2-2B	2B	Grain morphology	CO_321:0000040	chr2B:171030429.0.171034964	TraesCS2B02G194200
Tabas1	2B	Grain morphology	CO_321:0000040	chr2B:448904796.0.448907800	TraesCS2B02G313700
GNI	2B	Grain number	CO_321:0000391	chr2B:573974813.0.573975706	TraesCS2B02G405700
TaDA1-B	2B	Grain size	CO_321:0000040	chr2B:4646554.0.4654607	TraesCS2B02G007700
TaGS2-B1	2B	NUE	CO_321:0001671	chr2B:722629776.0.722634436	TraesCS2B02G528300
PPO-B1	2B	PPO activity	CO_321:0000214	chr2B:688478142.0.688480649	TraesCS2B02G491000
Ppo2-B1	2B	PPO activity	CO_321:0000214	chr2B:689764554.0.689766587	TraesCS2B02G491400
TaVSR-B1	2B	Root depth		chr2B:89554121.0.89558883	TraesCS2B02G122400
RMD-B1	2B	Root growth angle		chr2B:191742224.0.191744048	TraesCS2B02G209500
TaRSL4	2B	Root length		chr2B:197210852.0.197212507	TraesCS2B02G212700
Sdr-B1	2B	Seed dormancy/PHS	CO_321:0000081	chr2B:200572827.0.200573807	TraesCS2B02G215300
ZDS-D1	2D	Flour color	CO_321:0000214	chr2D:234144711.0.234150925	TraesCS2D02G236500
Ppd-D1	2D	Flowering time	CO_321:0000007	chr2D:33952224.0.33955766	TraesCS2D02G079600

(continued)

Table 9.2 (continued)

Gene	Chr	Phenotype	Crop ontology	Position	Traes ID
TaVIT1-2D	2D	GFe	CO_321:0000222	chr2D:419781553.0.419783725	TraesCS2D02G326300
TaNAS1-D	2D	GFe/GZn	CO_321:0000224	chr2D:12870350.0.12873858	TraesCS2D02G033000
TaNAS3-D	2D	GFe/GZn	CO_321:0000224	chr2D:18168587.0.18170017	TraesCS2D02G049200
TaNAS9-D	2D	GFe/GZn	CO_321:0000224	chr2D:45799198.0.45800220	TraesCS2D02G094200
TaCYP78A5	2D	Grain morphology	CO_321:0000040	chr2B:181118653.0.181120839	TraesCS2B02G201900
WFZP-D1	2D	Grain number	CO_321:0000391	chr2D:67496011.0.67496898	TraesCS2D02G118200
TaDA1-D	2D	Grain size	CO_321:0000040	chr2D:8281359.0.8289277	TraesCS2D02G016900
TaGS2-D1	2D	NUE	CO_321:0001671	chr2D:595161545.0.595165983	TraesCS2D02G500600
Rht8	2D	Plant height	CO_321:0000020	chrUn:24893964.0.24897255	TraesCSU02G024900
PPO-D1	2D	PPO activity	CO_321:0000214	chr2D:572952347.0.572954307	TraesCS2D02G468200
Ppo2-D1	2D	PPO activity	CO_321:0000214	chr2D:573903210.0.573905141	TraesCS2D02G468600
RMD-D1	2D	Root growth angle		chr2D:134790880.0.134792691	TraesCS2D02G190700
TaRSL4	2D	Root length		chr2D:138754346.0.138756038	TraesCS2D02G193700
Lyce-A1	3A	End-use quality	CO_321:0000214	chr3A:370233784.0.370237786	TraesCS3A02G208800
TaGS5-A1	3A	Grain morphology	CO_321:0000040	chr3A:176555776.0.176559839	TraesCS3A02G212900LC
Pod-A1	3A	POD activity/quality	CO_321:0000214	chr3A:730397626.0.730398805	TraesCS3A02G510600
Tamyb10-A1	3A	Seed color/PHS	CO_321:0000037	chr3A:703905707.0.703905910	TraesCS3A02G631500LC
Phs1	3A	Seed dormancy/PHS	CO_321:0000081	chr3A:7294435.0.7297613	TraesCS3A02G006600
Lyce-B1	3B	End-use quality	CO_321:0000214	chr3B:377418979.0.377422751	TraesCS3B02G239100
Fhb1_His	3B	FHB resistance	CO_321:0000651	chr3B:8526628.0.8529572	TraesCS3B02G019900
TaNAS5-B	3B	GFe/GZn	CO_321:0000224	chr3B:40773361.0.40778748	TraesCS3B02G068500
Tamyb10-B1	3B	Seed color/PHS	CO_321:0000037	chr3B:757918298.0.757920082	TraesCS3B02G515900
Vp1B1	3B	Seed dormancy/PHS	CO_321:0000081	chr3B:693338001.0.693342761	TraesCS3B02G452200
COMT-3B	3B	WSC/drought	CO_321:0000131	chr3B:829391763.0.829392973	TraesCS3B02G612000
CKX-D1	3D	Grain morphology	CO_321:0000040	chr3D:106736525.0.106740667	TraesCS3D02G143500
Myb10-D1	3D	Seed color/PHS	CO_321:0000037	chr3D:570801243.0.570803210	TraesCS3D02G468400
ALPb-4A	4A	End-use quality	CO_321:0000070	chr4A:718033180.0.718034037	TraesCS4A02G453800
PRR73-A1	4A	Flowering time	CO_321:0000007	chr4A:119083489.0.119087436	TraesCS4A02G105300
TaDMAS1-A	4A	GFe/GZn	CO_321:0000224	chr4A:74150821.0.74153009	TraesCS4A02G074800
TaNAS6-A	4A	GFe/GZn	CO_321:0000224	Chr4A:148780629.0.148781781	TraesCS4A02G127900LC
TaCYP78A5	4A	Grain morphology	CO_321:0000040	chr2D:127258537.0.127260686	TraesCS2D02G183000
TaGS1-A1	4A	NUE	CO_321:0001671	chr4A:60668121.0.60671232	TraesCS4A02G063800
MOR1-A1	4A	Root length		chr4A:685380302.0.685381598	TraesCS4A02G415400
Lox-B1	4B	Flour color	CO_321:0000214	chr4B:27248262.0.27252524	TraesCS4B02G037700
PRR73-B1	4B	Flowering time	CO_321:0000007	chr4B:427491684.0.427496233	TraesCS4B02G198700
TaDMAS1-B	4B	GFe/GZn	CO_321:0000224	Chr4B:481847465.0.481849531	TraesCS4B02G400500LC
TaNAS6-B	4B	GFe/GZn	CO_321:0000224	Chr4B:402432887.0.402433879	TraesCS4B02G183900
TaGS1-B1	4B	NUE	CO_321:0001671	chr4B:499898695.0.499901767	TraesCS4B02G240900
Pds-B1	4B	PDS activity/quality	CO_321:0000214	chr4B:586575839.0.586580177	TraesCS4B02G300100
Rht-B1	4B	Plant height	CO_321:0000020	chr4B:30861382.0.30863247	TraesCS4B02G043100
TaERF73-D1	4B	Root depth		chr4D:467792044.0.467801204	TraesCS4D02G406100LC
MOR1-B1	4B	Root length		chr4B:605691920.0.605693239	TraesCS4B02G316200
TaDMAS1-D	4D	GFe/GZn	CO_321:0000224	chr4D:392726584.0.392728858	TraesCS4D02G232200
TaNAS6-D	4D	GFe/GZn	CO_321:0000224	chr4D:323095782.0.323098145	TraesCS4D02G184900
TaD14-4D	4D	Grain yield	CO_321:0000013	chr4D:428116830.0.428119151	TraesCS4D02G258000
TaGS1-D1	4D	NUE	CO_321:0001671	chr4D:403145655.0.403148815	TraesCS4D02G240700
Rht-D1	4D	Plant height	CO_321:0000020	chr4D:18781062.0.18782933	TraesCS4D02G040400
TaERF73-A1	4D	Root depth		chr4A:3351141.0.3352418	TraesCS4A02G003300LC
MOR1-D1	4D	Root length		chr4D:478997945.0.478999338	TraesCS4D02G312800
Lr67	4D	Rust resistance		chr4D:405770870.0.405775112	TraesCS4D02G243100
Dro1-A1	5A	Drought tolerance	CO_321:0000131	chr5A:428994186.0.428997632	TraesCS5A02G213300
Vrn-A1a	5A	Flowering time	CO_321:0000007	chr5A:587411824.0.587423240	TraesCS5A02G391700
TaNAS4-A	5A	GFe/GZn	CO_321:0000224	chr5A:705402044.0.705403372	TraesCS5A02G552400
TaDep1-A1	5A	Grain morphology	CO_321:0000040	chr5A:430486331.0.430493530	TraesCS5A02G215100

(continued)

Table 9.2 (continued)

Gene	Chr	Phenotype	Crop ontology	Position	Traes ID
TaGL3.3-5A	5A	Grain morphology	CO_321:0000979	chr5A:26440090.0.26449927	TraesCS5A02G030300
Egt2-A1	5A	Root growth angle		chr5A:151732800.0.151736140	TraesCS5A02G102000
Dro1-B1	5B	Drought tolerance	CO_321:0000131	chr5B:381041995.0.381044714	TraesCS5B02G210500
Vrn-B1b	5B	Flowering time	CO_321:0000007	chr5B:573803238.0.573815903	TraesCS5B02G396600
TaDep1-B1	5B	Grain morphology	CO_321:0000040	chr5B:378517204.0.378520796	TraesCS5B02G208700
TaGL3.3-5B	5B	Grain morphology	CO_321:0000979	chr5B:27830119.0.27840027	TraesCS5B02G029100
Egt2-B1	5B	Root growth angle		chr5B:304265954.0.304269177	TraesCS5B02G164200
Dro1-D1	5D	Drought tolerance	CO_321:0000131	chr5D:327631371.0.327634216	TraesCS5D02G218700
Vrn-D1	5D	Flowering time	CO_321:0000007	chr5D:467176608.0.467184463	TraesCS5D02G401500
TaDep1-D1	5D	Grain morphology	CO_321:0000040	chr5D:326126003.0.326129557	TraesCS5D02G216900
TaGL3.3-5D	5D	Grain morphology	CO_321:0000979	chr5D:37321983.0.37331860	TraesCS5D02G038500
Pina-D1	5D	Grain texture	CO_321:0000072	chr5D:3591495.0.3592002	TraesCS5D02G004100
Pinb-D1	5D	Grain texture	CO_321:0000072	chr5D:3609640.0.3610146	TraesCS5D02G004300
Egt2-D1	5D	Root growth angle		chr5D:131504758.0.131508027	TraesCS5D02G113600
TaNAS2-A	6A	GFe/GZn	CO_321:0000224	chr6A:158316641.0.158317931	TraesCS6A02G163100
TaNAS7-A2	6A	GFe/GZn	CO_321:0000224	chr6A:603249197.0.603250189	TraesCS6A02G386200
TaNAS7-A1	6A	GFe/GZn	CO_321:0000224	chr6A:60971892.0.60973259	TraesCS6A02G093000
TaGW2-6A	6A	Grain morphology	CO_321:0000980	chr6A:237734835.0.237759808	TraesCS6A02G189300
TaT6P	6A	Grain morphology	CO_321:0000040	chr6A:461145380.0.461147406	TraesCS6A02G248400
SPL21-6A	6A	Grain morphology	CO_321:0000040	chr6A:136541506.0.136544204	TraesCS6A02G152000
NAM-A1	6A	Grain protein	CO_321:0000073	chr6A:77098570.0.77100127	TraesCS6A02G108300
Kat-2A	6A	Grain weight	CO_321:0000025	chr6A:606969628.0.606973059	TraesCS6A02G392400
Rht-24	6A	Plant height	CO_321:0000020	chr6A:413732327.0.413735532	TraesCS6A02G221900
Rht24	6A	Plant height	CO_321:0000020	chr6A:432253559.0.432257969	TraesCS6A02G229500
TaNAS2-B	6B	GFe/GZn	CO_321:0000224	chr6B:212158654.0.212159706	TraesCS6B02G186000
TaNAS7-B	6B	GFe/GZn	CO_321:0000224	chr6B:694258986.0.694259978	TraesCS6B02G425200
SPL21-6B	6B	Grain morphology	CO_321:0000040	chr6B:200509075.0.200512019	TraesCS6B02G180300
GW2-6B	6B	Grain morphology	CO_321:0000040	chr6B:291761397.0.291778503	TraesCS6B02G215300
NAM-B1	6B	Grain protein	CO_321:0000073	chr6B:134662733.0.134665065	TraesCS6B02G207500LC
KAT-2B	6B	Grain weight	CO_321:0000025	chr6B:701871007.0.701874630	TraesCS6B01G432600
1fehw3	6B	WSC/Drought	CO_321:0000131	chr6B:57283367.0.57288151	TraesCS6B02G080700
TaNAS2-D2	6D	GFe/GZn	CO_321:0000224	chr6D:121579210.0.121580540	TraesCS6D02G148600
TaNAS2-D1	6D	GFe/GZn	CO_321:0000224	chr6D:121225536.0.121228339	TraesCS6D02G148200
TaNAS7-D	6D	GFe/GZn	CO_321:0000224	chr6D:456540490.0.456541773	TraesCS6D02G370800
SPL21-6D	6D	Grain morphology	CO_321:0000040	chr6D:111567638.0.111570051	TraesCS6D02G142100
TaGS1a	6D	Nitrogen use efficiency	CO_321:0001671	chr6D:386290812.0.386294394	TraesCS6D02G383600LC
Moc-A1	7A	Agronomic traits/drought	CO_321:0000131	chr7A:557553815.0.557555303	TraesCS7A02G382800
ALPa-7A	7A	End-use quality	CO_321:0000070	chr7A:15697493.0.15698020	TraesCS7A02G035500
ALPb-7A	7A	End-use quality	CO_321:0000070	chr7A:15639003.0.15639854	TraesCS7A02G035200
PSY-A1	7A	Flour color	CO_321:0000214	chr7A:729397558.0.729401208	TraesCS7A02G557300
TEF-7A	7A	Grain morphology	CO_321:0000040	chr7A:66228020.0.66229066	TraesCS7A02G108900
Sus1-7A1	7A	Grain morphology	CO_321:0000040	chr7A:115204109.0.115208145	TraesCS7A02G158900
TaGW7	7A	Grain morphology	CO_321:0000980	chr7A:205459137.0.205465028	TraesCS7A02G233600
SPL20-7A	7A	Grain morphology	CO_321:0000040	chr7A:685212680.0.685214713	TraesCS7A02G495000
AGP-S-7A	7A	Grain morphology	CO_321:0000040	chr7A:342609326.0.34261711	TraesCS7A02G287400
WAPO-A1	7A	Grain number	CO_321:0000391	chr7A:674081462.0.674082918	TraesCS7A02G481600
VRT-A2	7A	Grain number	CO_321:0000391	chr7A:128826237.0.128833021	TraesCS7A02G175200
FRK2-7A	7A	Starch synthesis/grain morphology	CO_321:0001674	chr7A:459209231.0.459211266	TraesCS7A02G319000
PSY-B1	7B	Flour color	CO_321:0000214	chr7B:739442503.0.739445446	TraesCS7B02G482000
TaSus1-7B	7B	Grain morphology	CO_321:0000040	chr7B:68344330.0.68348404	TraesCS7B02G063400
WAPO-B1	7B	Grain number	CO_321:0000391	chr7B:649950255.0.649951851	TraesCS7B02G384000
PIN-B2	7B	Grain texture	CO_321:0000072	chr7B:699388914.0.699389366	TraesCS7B02G431200
TaCOL-B5	7B	Grain yield	CO_321:0000013	chr7B:667070044.0.667071768	TraesCS7B02G400600

(continued)

Table 9.2 (continued)

Gene	Chr	Phenotype	Crop ontology	Position	Traes ID
PSY-D1	7D	Flour color	CO_321:0000214	chr7D:636766504.0.636770671	TraesCS7D02G553300
Vrn-D3	7D	Flowering time	CO_321:0000007	chr7D:68416507.0.68417532	TraesCS7D02G111600
GS3-D1	7D	Grain morphology	CO_321:0000040	chr7D:6483394.0.6485745	TraesCS7D02G015000
SPL20-7D	7D	Grain morphology	CO_321:0000040	chr7D:592816295.0.592819560	TraesCS7D02G482400
Lr34	7D	Rust resistance		chr7D:47412273.0.47424077	TraesCS7D02G080300
TaNAS4-D	UNK	GFe/GZn	CO_321:0000224	chrUn:108595828.0.108597155	TraesCSU02G125200
TaDA1-A	UNK	Grain size	CO_321:0000040	chrUn:11740231.0.11748045	TraesCSU02G007800
TaERF73-B1		Root depth		chr4B:585962983.0.585964402	TraesCS4B02G299500

nitrogen-dependent lesion mimic gene *Ndhrl1* (Li et al. 2016), powdery mildew resistance gene *Pm4b* (Wu et al. 2018), leaf senescence gene *els1* (Li et al. 2018), stripe rust resistance gene *Yr26* (Wu et al. 2018), *YrMM58, YrHY1* (Wang et al. 2018a, b), dwarfing gene *Rht12* (Sun et al. 2019), and *Pm61* (Hu et al. 2019). It is likely that this approach will get more attention because it replaces the genotyping of complete populations (Zou et al. 2016).

Very few genes in wheat have been discovered using the traditional map-based cloning approach, and most of the genes have been identified by comparative genomics between wheat and related grass species due to the high collinearity and genetic organization among grass genomes (Rasheed and Xia 2019; Chen et al. 2020). According to the recent literature search, almost 33 genes related to grain morphology have been isolated by homology-based cloning and functional markers have been developed for use in breeding (Table 9.1). Likewise, genes related to other morphological and phenological traits have been isolated including *TaPRR73* (Zhang et al. 2016) and *TaZIM-A1* (Liu et al. 2018) underpinning flowering time; *TaPPH-7A* (Wang et al. 2018a; b) underpinning morphological traits; *TaARF4* (Wang et al. 2019b) controlling root growth and plant height; and *TaSnRK2.9-5A* (Ur Rehman et al. 2019) controlling drought tolerance.

9.3.1 Functional Genomics and Map-based Cloning in Wheat

The continuous development of new genomic resources in wheat including new reference genomes, transcriptome resources, wheat TILLING mutants with exome sequencing data, and high-density SNP database are conduits for carrying out map-based cloning to discover new genes in wheat. A QTL for head length and spikelet number was identified and then fine mapped to an interval of 0.2 cM (Yao et al. 2019). The map-based cloning identified that *Head Length 2 (HL2)* is the designated gene controlling head length and spikelet number. Zhang et al. (2018) fine mapped a heading time gene, *TaHdm605,* in an EMS mutant line. Spike architecture is an important yield-related attribute, and three genes *TaTFL1-2D*, *TaHOX2-2B,* and *TaAGLG1-5A,* controlling spike architecture were discovered analyzing a large-scale transcriptome data of 90 wheat lines (Wang et al. 2017b). The effects of these genes were validated by the transgenic assays. Another approach used for discovery of gene was the screening of a yeast cDNA library constructed from a heat- and drought-tolerant wheat cv. HANXUAN 10. Using this approach, *TaPR-1-1*, for tolerance to abiotic stress tolerance, was identified which encodes the pathogenesis-related (PR) protein family (Wang et al. 2019a).

The development of male sterile lines is an important component of hybrid wheat breeding program. Two studies simultaneously cloned *Male Sterile 2* (*Ms2*) gene underpinning male sterility in wheat (Ni et al. 2017; Xia et al. 2016). The causal mutation was identified to be a terminal-repeat retrotransposon in miniature (TRIM) element in the promoter of *Ms2*. The TRIM element was involved in the gene activation and causes male sterility. Liu et al. (2019) cloned *TaSPL8* gene controlling leaf angle and is an important component of auxin and brassinosteroid pathways and associated with cell elongation. The knockout mutants of *TaSPL8* had erect leaves due to the loss of the lamina joint, compact architecture, and increased spike number. *Pm21* is a durable disease resistance gene derived from *Haynaldia villosa* confers resistance against powdery mildew, and currently wheat cultivars with *Pm21* are cultivated on 4 m ha in China (Cao et al. 2011). Two complementary studies cloned *Pm21* and identified that it encodes a typical CC-NBS-LRR protein involved in broad spectrum resistance to powdery mildew (He et al. 2019).

Fusarium head blight (FHB) is one of the most important yield and quality limiting factors in wheat globally. There are very few resources providing durable resistance to FHB in wheat including some landraces from China like SUMAI 3, which is known to carry *Fhb1* gene. Rawat et al. (2016) used multiple approaches including positional cloning, development of overexpression lines, and gene silencing to report that a pore-forming toxin-like (*PFT*) gene was the candidate for *Fhb1*. However, it was later found that several FHB susceptible cultivars also carry *PFT* and its candidacy was doubted. Two new studies further established that a histidine-rich calcium-binding (*TaHRC* or *His*) gene adjacent to *PFT* is the actual *Fhb1* and was identified as a susceptibility factor (Su et al. 2019). In contrast, Li et al. (2019) concluded that *Fhb1* is a gain-of-function gene and that the newly generated protein acts as a regulator of host immunity.

9.3.2 Functional Genes and Their Diagnostic Markers

All the above examples show the discovery of genes following different strategies and include various validation approaches. Once a gene is discovered and its phenotypic effect is validated, it becomes important to identify and select the favorable alleles of those genes in breeding using functional markers (FMs). FMs are referred to the PCR-based diagnostic markers designed to identify causal polymorphism underpinning phenotypic differences. FMs are routinely used in crop breeding programs to identify and select the desirable allelic variations of specific functional genes (Liu et al. 2012; Rasheed et al. 2017; Rasheed and Xia 2019; Rouse et al. 2019). As mentioned earlier, FMs due to their high diagnostic value are ideal markers for use in breeding to identify and pyramid different genes in marker-assisted recurrent selection. FMs are also used in genomic selection to improve selection accuracy. Rasheed et al. (2016) converted a collection of 72 FMs to kompetitive allele-specific PCR (KASP) formats for their use in high-throughput platforms. This effort currently now includes 157 KASP markers to diagnose alleles of traits of breeding interest. These KASP markers have been used by various breeding programs, and a recent estimate from citation indicated that currently more than 35 wheat breeding and genetic programs all over the world used these markers. For example, CIMMYT elite lines were tagged with *TaGS3-D1*, *TaTGW6,* and *TaSus1* genes using these KASP markers (Sehgal et al. 2019). Zhao et al. (2019) screened 1152 diverse global wheat germplasm lines with KASP markers of 47 functional genes underpinning a number of important traits of breeding interest (Zhao et al. 2019). Favorable alleles of more than 39 genes of breeding importance were also identified in East African wheat germplasm using the aforementioned KASP markers (Wamalwa et al. 2020).

Several commercial alternatives to the KASP master mix are now available which have made SNP genotyping more cost effective. Apart from these commercial alternates to the KASP technology, some open-source SNP genotyping methods are also available. Two examples are the development of semi-thermal asymmetric reverse PCR (STARP) (Long et al. 2017) and Amplifluor (Jatayev et al. 2017) methods which can be used with wide range of commercial master mix. Several SNP markers were converted to STARP format to further reducing the cost of genotyping (Wu et al. 2020).

9.4 Mining Gene Networks Using Database Resources

We have outlined many genome sequencing projects carried out to generate genome variation data in wheat populations (Table 9.1). The amount of genome sequencing data being generated in wheat can often hinder scientists from translating complex and sometimes contradictory information into biological understanding and discoveries. Apart from using the data to investigate the genetic diversity, population-level genomic variation data provides a valuable resources and great opportunities for identifying trait-related genes, designing markers, constructing gene trees, exploring the evolutionary history, and assisting the design of molecular breeding. Mining the relevant information from the extensive genome variation datasets is a time-consuming and error-prone process if the proper tools are not used to explore the genes in questions. New tools are indispensable to develop for explaining how genes and gene networks might be implicated in a complex trait or disease. Another limitation is that tapping large and complex genome variation datasets requires computational skills exceeding the abilities of the most crop breeders. In nutshell, the reuse of genomic variation data plays an important role in driving current plant science research. We have provided an overview of the various genome variation tools and resources for quick analysis of gene and gene networks (Table 9.3).

9.4.1 Gene–gene Synteny Using PRETZEL

In defining a genetic framework at the genome level, the reliance on similarity searches with transcripts and proteins is of primary importance, and in this context, features of genome structure such as sequence/gene repetition impact on the capacity to identify the correct gene for detailed analysis. Sequence alignments underpin all the studies. The capacity to visualize genome features such as uneven repetition between loci aligned between several genomes (Fig. 9.1) can anticipate complications when gene alignments are carried out without this prior knowledge.

PRETZEL (https://plantinformatics.io; Keeble-Gagnere et al. 2019) is an online, interactive, and real-time visualization tool for analyzing and integrating genetic and genomic datasets. In Fig. 9.1, the alignments of the fructosyltransferase genes at the fructan synthesis locus on 7AS for the wheat cv. LANCER, cv. CHINESE SPRING, and cv. MACE are shown as a complex example where the IWGSC 7A-LANCE 7A alignment of the array of GH32 genes is fully syntenic between gene models within the LACER and CS loci. In contrast, the IWGSC 7A-MACE 7A alignment is evidently ambiguous as a result of small genome rearrangements possibly due to assembly errors. The software PRETZEL enables any locus of interest to be analyzed and potential issues to be identified.

The variations in fructosyltransferases on chromosomes 7A, 4A, 7D, 6A, 6B, and 6D are candidate genes in QTL that characterize fructan content in wheat grain and thus relate to quality/nutritional attributes of the grain (Zhang et al 2008; Huynh et al 2012; Langridge and Fleury 2012). The component fructosyltransferases genes in the 4A and 7D loci showed good alignment across LANCE, CS, and MACE except for an inversion relative the CS in the MACE locus similar to that shown for the 7A locus (Fig. 9.1). The 6B and 6D loci carried the component fructosyltransferases genes, referred to as fructan

Table 9.3 Genomics database in wheat for genome-informed characterization of wheat genes

Name	URL	Description	Referece
GrainGene	https://wheat.pw.usda.gov/GG3/	A comprehensive resource for molecular and phenotypic information for Triticeae and Avena	Odell et al. (2017)
MASWheat	http://maswheat.ucdavis.edu/	Marker-assisted selection database for wheat	NA
expVIP	http://wheat-expression.com/	Wheat transcriptome resources for expression analysis	Borrill et al. (2016)
WheatExp	https://wheat.pw.usda.gov/WheatExp/	Homoeologue-specific database of gene expression profiles for polyploid wheat	Pearce et al. (2015)
Cerealsdb	http://www.cerealsdb.uk.net/cerealgenomics/CerealsDB/indexNEW.php	Database for SNPs, genotyping arrays and sequences	NA
WheatIS	http://wheatis.org/	Wheat information system for wheat data, resources and bioinformatics tools	NA
OpenWildWheat	http://www.openwildwheat.org/	Sequencing resources of Ae. tauschii accessions	Gaurav et al. (2022)
IWGSC	http://www.wheatgenome.org/	Official website of IWGSC	NA
10+Wheat genomes	http://www.10wheatgenomes.com/	Wheat pan-genome resources	NA
Polymarker	http://polymarker.tgac.ac.uk/	SNP assay development tool	Ramirez-Gonzalez et al. (2015)
Triticeae tool box	https://triticeaetoolbox.org/wheat/	Repository of wheat data from wheat CAP	Blake et al. (2016)
Wheat Transcription factors	http://itak.feilab.net/	Database of wheat transcription factors	NA
TILLING	http://www.wheat-tilling.com/	Sequencing resource of CADENZA (6x) and KRONOS (4x) wheat TILLING population	Krasileva et al. (2017)
WGIN	http://www.wgin.org.uk/about.php	Wheat genetic improvement network	NA
URGI	http://wheat-urgi.versailles.inra.fr/	INRA-based resources for wheat sequence resources	NA
Gramene	http://www.gramene.org/	Open-source, integrated data resource for comparative functional genomics in crops and model plant species	Sun et al. (2022)
KnetMiner	http://knetminer.rothamsted.ac.uk/Triticum_aestivum/	Open-source software tools for integrating and visualizing large biological datasets	Hassani-Pak and Rawlings (2017)
Wheat SnpHub Portal	http://wheat.cau.edu.cn/Wheat_SnpHub_Portal/	A web interface to call variation data and map allele frequencies in global wheat populations based on exome capture and resequencing data	Wang et al. (2020)
Wheat Gmap	https://www.wheatgmap.org/	Bulk segregation analysis based on RNA or DNA sequencing data	Zhang et al. (2021)
WheatOmics	http://wheatomics.sdau.edu.cn/	Several wheat omics tools including blast, ID converter, sequence retriever, SNP marker	Ma et al. (2021)
WheatGene	http://wheatgene.agrinome.org	A Drupal-based interactive genome search database of wheat genomes and RNAseq	Garcia et al. (2021)

(continued)

Table 9.3 (continued)

Name	URL	Description	Referece
ggCOMP	http://wheat.cau.edu.cn/WheatCompDB/	A wheat resequencing database to enable unsupervised identification of pairwise germplasm resource-based identity by descent (gIBD) blocks	Yang et al. (2022)
ccnWHEAT	http://bioinformatics.cau.edu.cn/ccnWheat	A platform for searching and comparing specific functional co-expression networks, as well as identifying the related functions of the genes clustered therein	Li et al. (2022b)
TGT	http://wheat.cau.edu.cn/TGT/	A homology database, by integrating 12 Triticeae genomes and three outgroup model genomes and implemented versatile analysis and visualization functions	Chen et al. (2020)
Pretzel	https://plantinformatics.io/	An interactive, web-based environment for navigating multi-dimensional wheat datasets, including genetic maps and chromosome-scale physical assemblies	Keeble-Gagnere et al. (2019)
wheatQTL	http://wheatqtldb.net/	A QTL database of wheat	Singh et al. (2021)

1-exohydrolase (*1-FEH*) in Zhang et al (2008), and showed good alignment across LANCE, CS, and MACE. The 6A locus showed an inversion in MACE relative to CS and an absence of the locus in LANCER, consistent with the presence/absence polymorphism among wheat varieties for the 6A locus reported by Zhang et al. (2008).

In contrast to the locus carrying the fructosyltransferases, the wheat-*APO1* (*WAPO-A1*) locus on the long arm of 7A shows unambiguous alignments across the varieties examined (Fig. 9.2a, left-hand panel for entire chromosomes and right panel for the *WAPO1* locus region), and thus, the variation at the structural level that needs to be considered when gene functions are examined is not a significant factor. Interestingly, the h1 and h2 haplotypes at this locus (Fig. 9.2b) identified by Voss-Fels et al. (2018) using SNP variation in the genome sequence indicate striking sequence-level divergence in this *WAPO1* gene region that is not reflected at the gene–gene syntenic level shown in Fig. 9.2a.

The genome viewer in Fig. 9.2b is from DAWN (Watson-Haigh et al. 2018) and shows variation in SNP (colored drops) positions relative to the CHINESE SPRING refseq 2.1 as a reference and uses cv. LANCER and cv. MACE from the wheat 10Xgenome sequence dataset, and cv. XIAOYAN 54 and WESTONIA from Whole-Genome Shotgun (WGS) resequencing data (Watson-Haigh et al. 2018). In field trails, under rain-fed conditions, the SNP-based haplotype h2 was found to be significantly associated with increased grain yield compared to h1, conferring a 24% yield advantage relative to all other haplotypes, especially h1 which was the other prominent haplotype in the field trial (Voss-Fels et al. 2019).

PRETZEL aims to solve alignment problems and structural changes in cultivar sequences by providing an interactive, online environment for data visualization and analysis which, when loaded with appropriately curated data, can enable researchers with no bioinformatics training to exploit the latest genomic resources

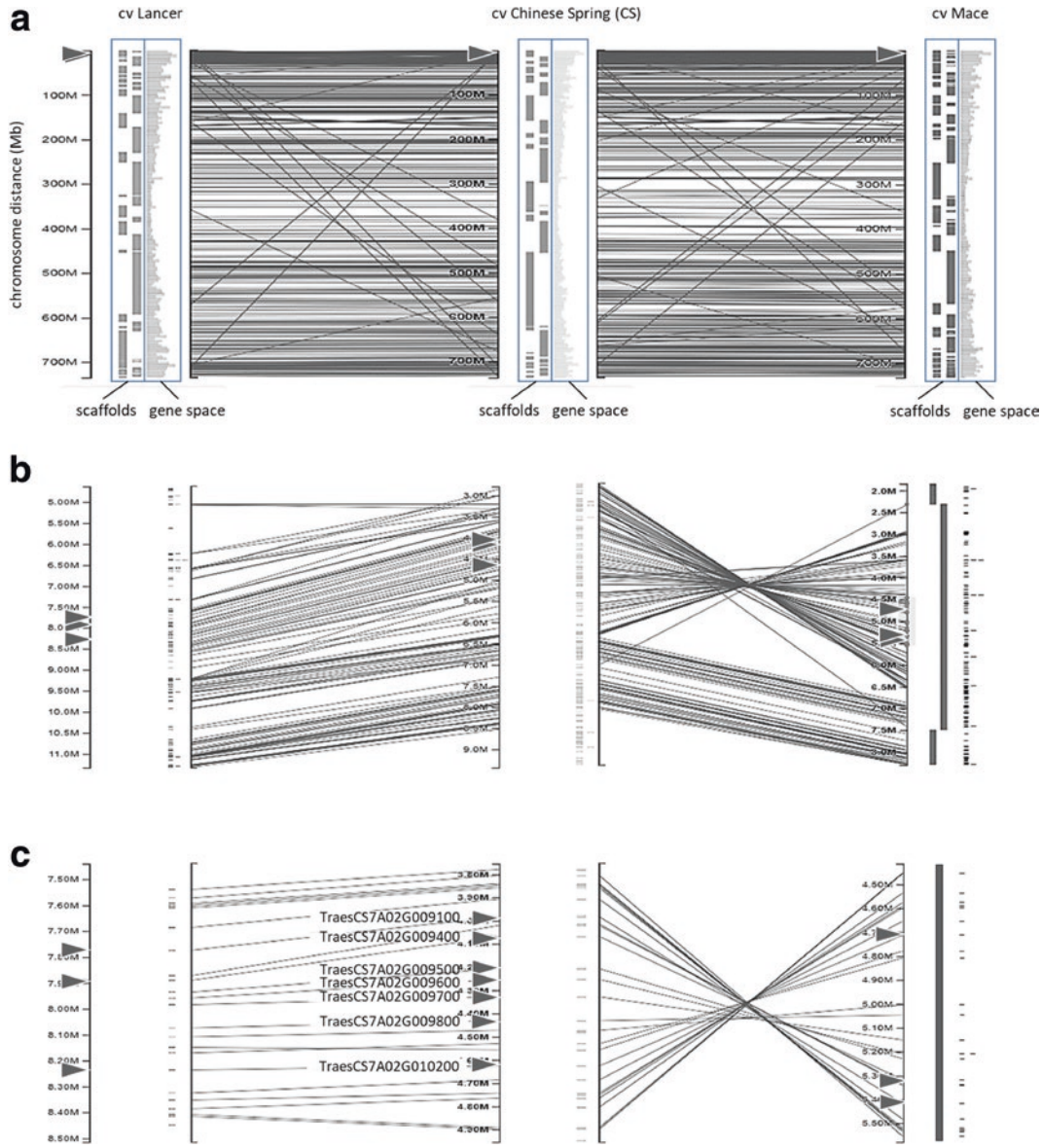


Fig. 9.1 Comparative analysis of 7AS fructan locus. In **a**, the arrows indicate the location of the locus within the entire chromosome, and **b** and **c** are the images resulting from ZOOMing into the locus. The marker genes *TraesCS7A02G009100*, *TraesCS7A02G009200* through *TraesCS7A02G010200* indicate the array of GH32 fructosyltransferases located at the locus in a ca 750 kb region (**c**). Scaffold columns to the right side of the PRETZEL maps are important for checking aberrations in colinearity (based on sequence similarity of 70% over 70% of the length of the sequence) as discussed in the text in terms of relating the boundaries of inverted regions to the boundaries of scaffolds in the assembly. In the region illustrated for MACE (**b**, **c**), the chance of the inverted region being an assembly error is reduced because the inversion is well within the respective scaffold

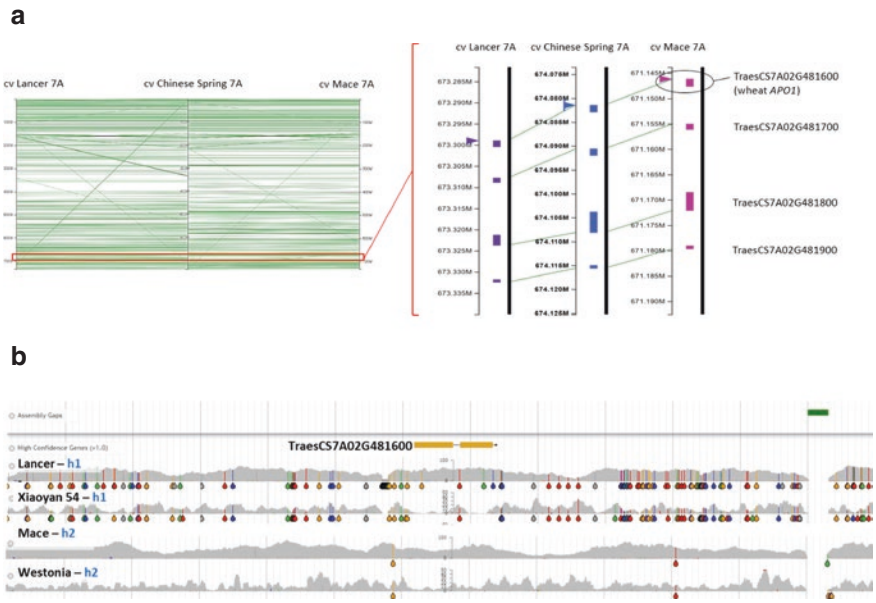


Fig. 9.2 **a** PRETZEL view of chr7A region (right panel) showing several genes including *WAPO-A1* (Voss-Fels et al. 2019; Kuzay et al. 2019, 2022) and structural changes in the *WAPO-A1* gene across three cvs. LANCER, CS, and MACE can be visualized with high-resolution (right panel). **b** is the genome viewer from DAWN (Watson-Haigh et al. 2018), and shows variation relative to the CHINESE SPRING refseq 2.1 as a reference and uses cv. LANCER and cv. MACE from the wheat 10Xgenome sequence dataset, and cv. XIAOYAN 54 and WESTONIA from Whole-Genome Shotgun (WGS) resequencing data

(Keeble-Gagnère et al. 2019). Apart from the visualization, PRETZEL can be used to retrieve the genome information (features including markers, genes, annotations, etc.) as dataset files of any selected chromosomal region for further downstream analysis.

9.4.2 Knowledge Graphs

Knowledge graphs (KG) are now extensively used to make search and information discovery more efficient. Knetminer is a data integration platform to visualize biological knowledge networks in an interactive web application (Hassani-Pak and Rawlings 2017). The data integration approach to build KGs has the ability to capture complex biological relationships between genes, traits, diseases, and many more information types derived from curated or predicted information sources. For example, *Rht24* is a new gene discovered associated with semi-dwarf phenotype in wheat and is present on chr6A. The Knetminer identified the gene network of *Rht24*, partially shown as Fig. 9.3 for clarity. The *Traes IDs* of both of the chr6B and chr6D homeologue are shown as interacting genes, and another gene, *TraesCS5B02G265400*, strongly interacts with *Rht24*. It can also be visualized that the gene interacts with bHLH27 transcription factor and physiologically influences the Gibberellin 20 pathway. Another feature is the identification of any stop/gain mutations in the CADENZA TILLING population, and mutant names and SNP positions can also be visualized.

The causal mutation of Rht24 on chr6A was identified in the exome capture data of the global hexaploid wheat collection (He et al. 2019). The target SNP was plotted for the frequency of wild-type and alternate SNP among global wheat accessions using SnpHub portal (Fig. 9.4).

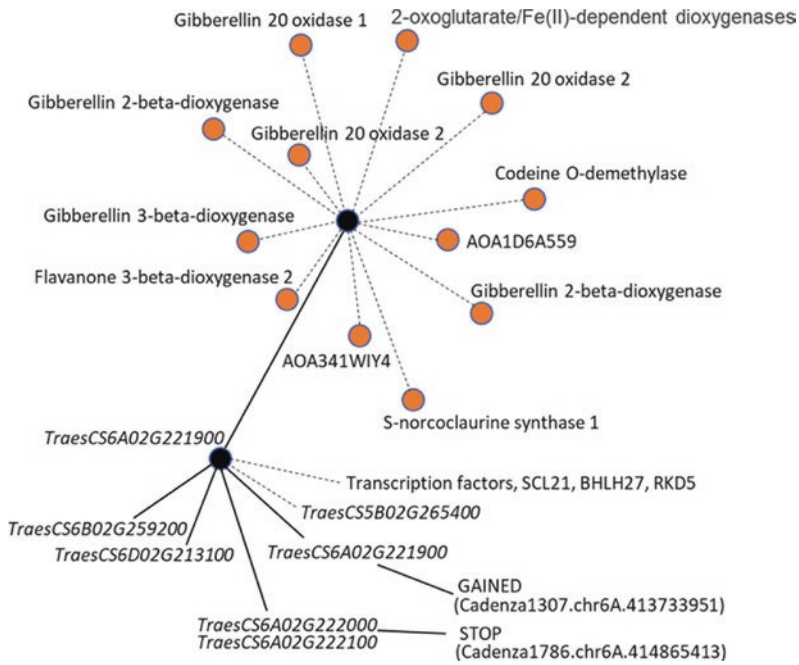


Fig. 9.3 KnetMiner network depicts connections with *Rht-24* on chr6A in wheat. This wheat reduced height gene, *Rht-24*, its homeologs on B- and D-genome along with other genes in cross-talk like TraesCS5B02G265400, associated transcription factors, and the mutations in the wheat TILLING population (e.g., two mutations in CADENZA TILLING population) can be visualized. Not all connections present in the KnetMiner network are depicted in the figure; only a subset is shown for clarity

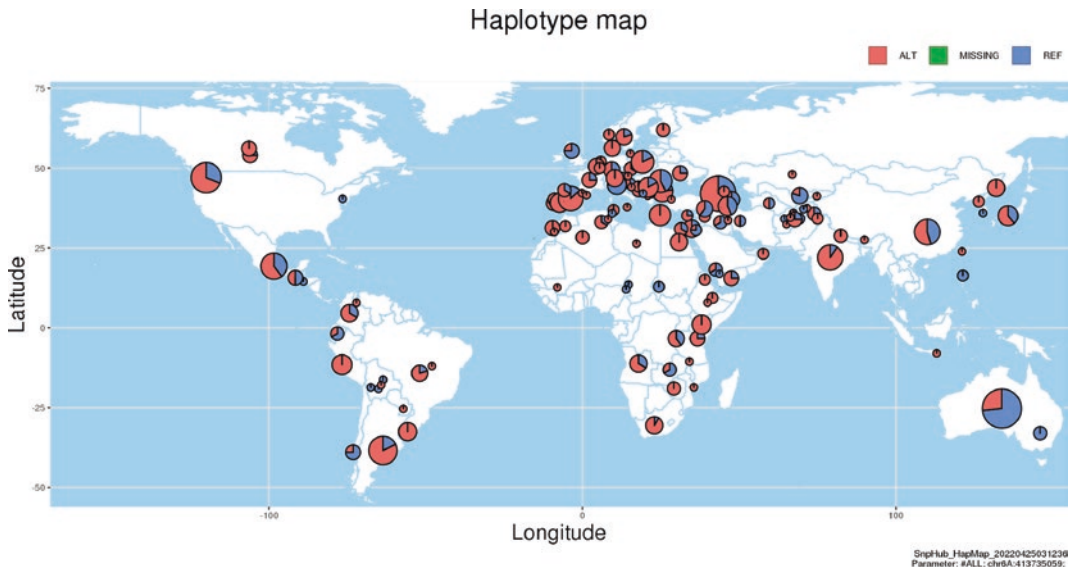


Fig. 9.4 SnpHub-based global haplotype map of non-synonymous mutation in Rht-24 is plotted based on the global exome sequencing data. In pie chart, the red proportion represents the frequency of wild-type mutation, while the blue proportion represents the frequency of non-synonymous mutation associated with reduced height

9.4.3 SnpHub Portal for Global Overview of Functional Gene Frequencies

SnpHub portal is a convenient way to identify mutations in the wheat genomes and then plotting the frequency of the SNPs country-wise in global what population (Wang et al. 2020). It is a Shing/R-based platform for mining and visualizing large genome variation data in wheat. Genome variation data in terms of .vcf files and genome annotation files can be accessed by a chromosomal interval of specific gene (*Traes ID*) to visualize genomic variation in heatmap, phylogenetic trees, haplotype networks, and haplotype geographic maps.

Apart from these platforms, several other platforms can be interactively used to mine useful genome variation and gene expression analysis (Table 9.3). The exVIP is an excellent resource for gene expression studies across various tissues and various experiments where the expression of certain genes can be visualized as heatmaps or as datafiles for further analysis. Similarly, WheatOmics (Ma et al. 2021) provides several features for analysis of genes including JBrowse with distinct track of several SNP genotyping and exome sequencing resources, *TraesID* converter, and sequence retriever. Last but not least, a wheat QTL database has been released recently which is an important resource to align QTL information with the IWGSC reference sequence (Singh et al. 2021).

9.5 Conclusion and Prospects

The complete annotation of functional genes in wheat is a challenge at multiple levels. For example, a first important intrinsic feature to impact annotation is the fragmentation level at the level of the number of exons per gene. As a CDS is fragmented into several exons, the difficulty to predict the correct intron/exon structure increases. In a detailed analysis of the wheat genome space by Choulet et al., (this volume, Chap. 4) it was emphasized that an important intrinsic feature of eukaryote gene structure that impacts on annotation is the fragmentation level at the level of the number of exons per gene. Choulet et al., (Chap. 4) noted that as a CDS is fragmented into several exons, the difficulty in predicting the correct intron/exon structure increases, although in wheat, (RefSeq Annotation v2.1) the average number of exons per CDS is only 4, and some genes (up to 10%) can have up to 17 exons. In this chapter, we have assigned genes and QTL to the reference genome and utilized available annotations to significantly improve the value of the outputs as reference documentation to be used in wheat breeding. The alignment of traits to annotated genes in the reference genome provides their position and *TraesIDs* to define a framework for establishing more informative markers for selecting lines to be deployed in crosses as well as for tracking targeted traits in segregating progeny from crosses.

Integration of a range of datasets has been emphasized in this chapter in order to deal with the complexity of the wheat genome and generating robust associations between genome haplotypes and agronomic traits for selecting parents for crossing and accurately tracking progeny from crosses. Since only 17% of genes are single copies, most key agronomic traits are likely to be the product of gene network interactions involving genes/gene families distributed across the chromosomes of the A-, B-, and D-subgenomes and genome signatures (haplotypes).

The sequencing data generated from cultivated and wild wheats, natural and breeding populations, and mutants is enabling the discovery of genes underpinning important traits of breeding interest. This information is useful to further develop and deploy the diagnostic markers for use in wheat breeding. The wheat genome variation is very complex for downstream analysis; therefore, the data analytics platforms have been developed to visualize genome variations and expression in heatmaps, haplotype and geographic maps, and gene

networks. We have provided an elucidated of gene frameworks discovered so far in wheat, and these need to be integrated with the thousands of QTL that have been discovered in wheat in different mapping populations and with many different marker platforms. The integration of wheat QTL information with genome visualization platforms for better understanding of gene networks and trait discovery is a key challenge.

Acknowledgements The authors are grateful for the perceptive comments provided by Dr Yuri Shavrukov (Flinders University, SthAustralia).

References

Avni R, Nave M, Barad O, Baruch K, Twardziok SO et al (2017) Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. Science 357:93–97

Avni R, Lux T, Minz-Dub A, Millet E, Sela H, Distelfeld A et al (2022) Genome sequences of three Aegilops species of the section Sitopsis reveal phylogenetic relationships and provide resources for wheat improvement. Plant J 110:179–192

Balfourier F, Bouchet S, Robert S, De Oliveira R, Rimbert H, Kitt J, Choulet F, Paux E (2019) Worldwide phylogeography and history of wheat genetic diversity. Sci Adv 5:eaav0536

Bernardo R, Yu J (2007) Prospects for genomewide selection for quantitative traits in Maize. Crop Sci 47:1082–1090

Bevan MW, Uauy C, Wulff BB, Zhou J, Krasileva K, Clark MD (2017) Genomic innovation for crop improvement. Nature 543:346–354

Blake VC, Birkett C, Matthews DE, Hane DL, Bradbury P, Jannink J-L (2016) The triticeae toolbox: combining phenotype and genotype data to advance small-grains breeding. Plant Genome 9:2. https://doi.org/10.3835/plantgenome2014.12.0099

Borrill P, Ramirez-Gonzalez R, Uauy C (2016) expVIP: a customizable RNA-seq data analysis and visualization platform. Plant Physiol 170:2172–2186

Browning BL, Yu Z (2009) Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. Am J Hum Genet 85:847–861

Cao A, Xing L, Wang X, Yang X, Wang W, Sun Y, Qian C, Ni J, Chen Y, Liu D, Wang X, Chen P (2011) Serine/threonine kinase gene Stpk-V, a key member of powdery mildew resistance gene Pm21, confers powdery mildew resistance in wheat. Proc Natl Acad Sci U S A 108:7727–7732

Chen H, Jiao C, Wang Y, Wang Y, Tian C, Yu H, Wang J, Wang X, Lu F, Fu X, Xue Y, Jiang W, Ling H, Lu H, Jiao Y (2019) Comparative population genomics of bread wheat (Triticum aestivum) reveals its cultivation and breeding history in China. bioRxiv:519587

Chen Y, Song W, Xie X, Wang Z, Guan P, Peng H, Jiao Y, Ni Z, Sun Q, Guo W (2020) A collinearity-incorporating homology inference strategy for connecting emerging assemblies in the Triticeae tribe as a pilot practice in the plant pangenomic era. Mol Plant 13:1694–1708

Driscoll CJ, Jensen N (1964) Chromosomes associated with waxlessness, awnedness and time of maturity of common wheat. Can J Genet Cytol 6:324–333

Finno CJ, Aleman M, Higgins RJ, Madigan JE, Bannascha DL (2014) Risk of false positive genetic associations in complex traits with underlying population structure: a case study. Vet J 202:543–549

Gao F, Wen W, Liu J, Rasheed A, Yin G, Xia X, Wu X, He Z (2015) Genome-wide linkage mapping of QTL for yield components, plant height and yield-related physiological traits in the Chinese wheat cross Zhou 8425B/Chinese Spring. Front Plant Sci 6:1099

Garcia DF, Wang Z, Guan J, Yin L, Geng S, Li A, Mao L (2021) WheatGene: a genomics database for common wheat and its related species. Crop J 9:1486–1491

Gaurav K, Arora S, Silva P, Sánchez-Martín J, Horsnell R et al (2022) Population genomic analysis of *Aegilops tauschii* identifies targets for bread wheat improvement. Nat Biotechnol 40:422–431

Halloran GM, Boydell CW (1967) Wheat chromosomes with genes for vernalization response. Can J Genet Cytol 9:632–639

Hao Z, Geng M, Hao Y, Zhang Y, Zhang L, Wen S, Wang R, Liu G (2019) Screening for differential expression of genes for resistance to Sitodiplosis mosellana in bread wheat via BSR-seq analysis. Theor Appl Genet 132:3201–3221

Hao C, Jiao C, Hou J, Li T, Liu H, Wang Y, Zheng J, Liu H, Bi Z, Xu F, Zhao J, Ma L, Wang Y, Majeed U, Liu X, Appels R, Maccaferri M, Tuberosa R, Lu H, Zhang X (2020) Resequencing of 145 landmark cultivars reveals asymmetric sub-genome selection and strong founder genotype effects on wheat breeding in China. Mol Plant 13:1733–1751

Hassani-Pak K, Rawlings C (2017) Knowledge discovery in biological databases for revealing candidate genes linked to complex phenotypes. J Integr Bioinform 14. https://doi.org/10.1515/jib-2016-0002

He F, Pasam R, Shi F, Kant S, Keeble-Gagnere G, Kay P, Forrest K, Fritz A, Hucl P, Wiebe K, Knox R, Cuthbert R, Pozniak C, Akhunova A, Morrell PL, Davies JP, Webb SR, Spangenberg G, Hayes B, Daetwyler H, Tibbits J, Hayden M, Akhunov E (2019) Exome sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome. Nat Genet 51:896–904

Hu J, Li J, Wu P, Li Y, Qiu D, Qu Y, Xie J, Zhang H, Yang L, Fu T, Yu Y, Li M, Liu H, Zhu T, Zhou Y, Liu

Z, Li H (2019) Development of SNP, KASP, and SSR markers by BSR-Seq technology for saturation of genetic linkage map and efficient detection of wheat powdery mildew resistance gene *Pm61*. Int J Mol Sci 20:750

Huynh B-L, Mather DE, Schreiber AW, Toubia J, Baumann U, Shoaei Z, Stein N, Ariyadasa R, Stangoulis JCR, Edwards J, Shirley N, Langridge P, Fleury D (2012) Clusters of genes encoding fructan biosynthesizing enzymes in wheat and barley. Plant Mol Biol 80:299–314

IWGSC (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science 361:eaar7191

Jatayev S, Kurishbayev A, Zotova L et al (2017) Advantages of Amplifluor-like SNP markers over KASP in plant genotyping. BMC Plant Biol 17:254

Jayakodi M, Schreiber M, Stein N, Mascher M (2021) Building pan-genome infrastructures for crop plants and their use in association genetics. DNA Res 28:dsaa030

Juliana P, Poland J, Huerta-Espino J, Shrestha S, Crossa J, Crespo-Herrera L, Toledo FH, Govindan V, Mondal S, Kumar U, Bhavani S, Singh PK, Randhawa MS, He X, Guzman C, Dreisigacker S, Rouse MN, Jin Y, Pérez-Rodríguez P, Montesinos-López OA, Singh D, Mokhlesur Rahman M, Marza F, Singh RP (2019) Improving grain yield, stress resilience and quality of bread wheat using large-scale genomics. Nature 51:1530–1539

Keeble-Gagnère G, Isdale D, Suchecki R, Kruger A, Lomas K, Carroll D, Li S, Whan A, Hayden M, Tibbits J (2019) bioRxiv. https://doi.org/10.1101/517953

Krasileva KV, Vasquez-Gross HA, Howell T, Bailey P, Paraiso F, Clissold L, Simmonds J, Ramirez-Gonzalez RH, Wang X, Borrill P, Fosker C, Ayling S, Phillips AL, Uauy C, Dubcovsky J (2017) Uncovering hidden variation in polyploid wheat. Proc Nat Acad Sci USA 114:E913–E921

Kuzay S, Xu Y, Zhang J, Katz A, Pearce S, Su Z, Fraser M, Anderson JA, Brown-Guedira G, DeWitt N, Haugrud AP, Faris JD, Akhunov E, Bai G, Dubcovsky J (2019) Identification of a candidate gene for a QTL for spikelet number per spike on wheat chromosome arm 7AL by high-resolution genetic mapping. Theor Appl Genetics 132:2689–2705

Kuzay S, Lin H, Li C, Chen S, Woods DP, Zhang J, Lan T, von Korff M, Dubcovsky J (2022) WAPO-A1 is the causal gene of the 7AL QTL for spikelet number per spike in wheat. PLoS Genet 18(1):e1009747

Langridge P, Fleury D (2012) Clusters of genes encoding fructan biosynthesizing enzymes in wheat and barley. Plant Mol Biol 80:299–314

Law CN (1966) The location of genetic factors affecting a quantitative character in wheat. Genetics 53:487–498

Li L, Shi X, Zheng F et al (2016) A novel nitrogen-dependent gene associates with the lesion mimic trait in wheat. Theor Appl Genet 129:2075–2084. https://doi.org/10.1007/s00122-016-2758-3

Li H, Rasheed A, Hickey L, He Z (2018) Fast-forwarding genetic gain. Trends Plant Sci 23:184–186

Li G, Zhou J, Jia H et al (2019) Mutation of a histidine-rich calcium-binding-protein gene in wheat confers resistance to Fusarium head blight. Nat Genet 51:1106–1112. https://doi.org/10.1038/s41588-019-0426-7

Li A, Hao C, Wang Z, Geng S, Jia M, Wang F, Han X, Kong X, Yin L, Tao S, Deng Z, Liao R, Sun G, Wang K, Ye X, Jiao C, Lu H, Zhou Y, Liu D, Fu X, Zhang X, Mao L (2022a) Wheat breeding history reveals synergistic selection of pleiotropic genomic sites for plant architecture and grain yield. Mol Plant 15:504–519

Li Z, Hu Y, Ma X, Da L, She J, Liu Y, Yi X, Cao Y, Xu W, Jiao Y, Su Z (2022b) ccnWheat: A database for comparing co-expression networks analysis of allohexaploid wheat and its progenitors. bioRxiv:2022b.2001.2017.476536

Ling H-Q, Ma B, Shi X, Liu H, Dong L, Sun H, Cao Y, Gao Q, Zheng S, Li Y, Yu Y, Du H, Qi M, Li Y, Lu H, Yu H, Cui Y, Wang N, Chen C, Wu H, Zhao Y, Zhang J, Li Y, Zhou W, Zhang B, Hu W, van Eijk MJT, Tang J, Witsenboer HMA, Zhao S, Li Z, Zhang A, Wang D, Liang C (2018) Genome sequence of the progenitor of wheat A subgenome *Triticum urartu*. Nature 557:424–428

Liu YN, He ZH, Appels R, Xia XC (2012) Functional markers in wheat: current status and future prospects. Theor Appl Genet 125:1–10

Liu J, He Z, Rasheed A et al (2017) Genome-wide association mapping of black point reaction in common wheat (*Triticum aestivum* L.). BMC Plant Biol 17:220 https://doi.org/10.1186/s12870-017-1167-3

Liu H, Li T, Wang Y, Zheng J, Li H, Hao C, Zhang Z (2018) TaZIM-A1 negatively regulates flowering time in common wheat (Triticum aestivum L.). J Integr Plant Biol 61:359–376. https://doi.org/10.1111/jipb.12720

Liu K, Cao J, Yu K, Liu X, Gao Y, Chen Q, Zhang W, Peng H, Du J, Xin M, Hu Z, Guo W, Rossi V, Ni Z, Sun Q, Yao Y (2019) Wheat TaSPL8 modulates leaf angle through auxin and brassinosteroid signaling. Plant Physiol 18:179–194. https://doi.org/10.1104/pp.19.00248

Long YM, Chao WS, Ma GJ et al (2017) (2017) An innovative SNP genotyping method adapting to multiple platforms and throughputs. Theor Appl Genet 130:597–607. https://doi.org/10.1007/s00122-016-2838-4

Lopes MS, Reynolds MP, Manes Y, Singh RP, Crossa J, Braun HJ (2012) Genetic yield gains and changes in associated traits of CIMMYT spring bread wheat in a "Historic" set representing 30 years of breeding. Crop Sci 52:1123–1131

Ma S, Wang M, Wu J, Guo W, Chen Y, Li G, Wang Y, Shi W, Xia G, Fu D, Kang Z, Ni F (2021) WheatOmics: a platform combining multiple omics data to accelerate functional genomics studies in wheat. Mol Plant 14:1965–1968

McIntosh R (1973) A catalogue of gene symbols for wheat. In: Missouri C (ed) Proceedings of 4th International Wheat Genetics Symposium

Ni F, Qi J, Hao Q et al (2017) Wheat Ms2 encodes for an orphan protein that confers male sterility in grass species. Nat Commun 8. https://doi.org/10.1038/ncomms15121

Odell SG, Lazo GR, Woodhouse MR, Hane DL, Sen TZ (2017) The art of curation at a biological database: principles and application. Curr Plant Biol 11–12:2–11

Pearce S, Vazquez-Gross H, Herin SY, Hane D, Wang Y, Gu YQ, Dubcovsky J (2015) WheatExp: an RNA-seq expression database for polyploid wheat. BMC Plant Biol 15:299

Pont C, Leroy T, Seidel M, Tondelli A, Duchemin W et al (2019) Tracing the ancestry of modern bread wheats. Nat Genet 51:905–911

Ramirez-Gonzalez RH, Uauy C, Caccamo M (2015) PolyMarker: a fast polyploid primer design pipeline. Bioinformatics 31:2038–2039

Rasheed A, Xia X (2019) From markers to genome-based breeding in wheat. Theor Appl Genet 132:767–784

Rasheed A, Wen W, Gao F, Zhai S, Jin H, Liu J, Guo Q, Zhang Y, Dreisigacker S, Xia X, He Z (2016) Development and validation of KASP assays for genes underpinning key economic traits in bread wheat. Theor Appl Genet 2016:1843–1860

Rasheed A, Hao Y, Xia XC, Khan A, Xu Y, Varshney RK, He ZH (2017) Crop breeding chips and genotyping platforms: progress, challenges and perspectives. Mol Plant 10:1047–1064

Rawat N, Pumphrey M, Liu S et al (2016) Wheat Fhb1 encodes a chimeric lectin with agglutinin domains and a pore-forming toxin-like domain conferring resistance to Fusarium head blight. Nat Genet 48:1576–1580. https://doi.org/10.1038/ng.3706

Rouse MN, Jin Y, Pérez-Rodríguez P, Montesinos-López OA, Singh D, Mokhlesur Rahman M, Marza F, Singh RP (2019) Improving grain yield, stress resilience and quality of bread wheat using large-scale genomics. Nature 51:1530–1539

Sears ER (1954) The aneuploids of common wheat. University of Missouri, College of Agriculture, Agricultural Experiment Station

Sears ER, Sears LMS (1978) The telocentric chromosomes of common wheat. In: 5th international wheat genetics symposium, pp 389–407

Sehgal D, Mondal S, Guzman C, Garcia Barrios G, Franco C, Singh R, Dreisigacker S (2019) Validation of candidate gene-based markers and identification of novel loci for thousand-grain weight in spring bread wheat. Front Plant Sci 10:1189

Sehgal D, Mondal S, Crespo-Herrera, Velu G, Juliana P, Huerta-Espino J, Shrestha S, Poland J, Singh R, Dreisigacker S (2020) Haplotype-based, genome-wide association study reveals stable genomic regions for grain yield in CIMMYT spring bread wheat. Front Genet 11. https://doi.org/10.3389/fgene.2020.589490

Shepherd K (1968) Chromosomal control of endosperm proteins in wheat and rye. In: Proceedings of 3rd International Wheat Genetics Symposium Australian Acadamic Science, pp 86–96

Shi W, Hao C, Zhang Y,Cheng J, Zhang Z, Liu J,Yi X, Cheng X, Sun D, Xu, Zhang X, Cheng S, Guo P, Guo J (2017) A combined association mapping and linkage analysis of kernel number per spike in common wheat (Triticum aestivum L.). Front. Plant Sci. Sec. Plant Breed 18. https://doi.org/10.3389/fpls.2017.01412

Singh K, Batra R, Sharma S et al (2021) WheatQTLdb: a QTL database for wheat. Mol Genet Genomics 296:1051–1056. https://doi.org/10.1007/s00438-021-01796-9

Su Z, Bernardo A, Tian B et al (2019) A deletion mutation in TaHRC confers Fhb1 resistance to Fusarium head blight in wheat. Nat Genet 51:1099–1105. https://doi.org/10.1038/s41588-019-0425-8

Sun J, Zhan K, Chu J, Zhang A (2018) A wheat dominant dwarfing line with Rht12, which reduces stem cell length and affects gibberellic acid synthesis, is a 5AL terminal deletion line (2019). Plant J 97:887–900. https://doi.org/10.1111/tpj.14168

Sun L, Yang W, Li Y, Shan Q, Ye X, Wang D, Yu K, Lu W, Xin P, Zhong X, Pei Z, Guo X, Liu D, Tello-Ruiz MK, Jaiswal P, Ware D (2022) Gramene: a resource for comparative analysis of plants genomes and pathways. In: Edwards D (ed) Plant bioinformatics: methods and protocols. Springer, US, New York, NY, pp 101–131

Tester M, Langridge P (2010) Breeding technologies to increase crop production in a changing world. Science 327:818–822

Ur Rehman S, Wang J, Chang X et al (2019) A wheat protein kinase gene TaSnRK2.9-5A associated with yield contributing traits. Theor Appl Genet 132:907–919

Varshney RK, Graner A, Sorrells ME (2005) Genomics-assisted breeding for crop improvement. Trends Plant Sci 10:621–630

Voss-Fels KP, Robinson H, Mudge SR, Richard C, Newman S, Wittkop B, Stahl A, Friedt W, Frisch M, Gabur I, Miller-Cooper A (2018) *VERNALIZATION1* modulates root system architecture in wheat and barley. Mol Plant 11(1):226–229

Voss-Fels KP, Keeble-Gagnère G, Hickey LT, Tibbits J, Nagornyy S, Hayden MJ, Pasam RK, Kant S, Friedt W, Snowdon RJ, Appels R, Wittkop B (2019) High-resolution mapping of rachis nodes per rachis, a critical determinant of grain yield components in wheat. Theor Appl Genet 132:2707–2719

Walkowiak S, Gao L, Monat C, Haberer G, Kassa MT, Brinton J et al (2020) Multiple wheat genomes reveal global variation in modern breeding. Nature 588:277–283

Wamalwa M, Tadesse Z, Muthui L et al (2020) Allelic diversity study of functional genes in East Africa bread wheat highlights opportunities for genetic improvement. Mol Breed 40:104. https://doi.org/10.1007/s11032-020-01185-x

Wang Y, Xie, Zhang H et al (2017a) Mapping stripe rust resistance gene YrZH22 in Chinese wheat cultivar Zhoumai 22 by bulked segregant RNA-Seq (BSR-Seq) and comparative genomics analyses. Theor Appl Genet 130:2191–2201. https://doi.org/10.1007/s00122-017-2950-0

Wang Y, Yu H, Tian C, Sajjad M, Gao C, Tong Y Wang X, Jiao Y (2017b) Transcriptome association identifies regulators of wheat spike architecture. Plant Physiol 175:746–757. https://doi.org/10.1104/pp.17.00694

Wang H, Wang S, Chang X et al (2018a) Identification of TaPPH-7A haplotypes and development of a molecular marker associated with important agronomic traits in common wheat. BMC Plant Biol 19:296. https://doi.org/10.1186/s12870-019-1901-0

Wang Y, Zhang H, Xie J, Guo B, Chen Y, Zhang H, Lu P, Wu Q, Li M, Zhang D, Guo G, Yang J, Zhang P, Zhang Y, Wang X, Zhao H, Cao T, Liu Z (2018b) Mapping stripe rust resistance genes by BSR-Seq: YrMM58 and YrHY1 on chromosome 2AS in Chinese wheat lines Mengmai 58 and Huaiyang 1 are Yr17. Crop J 6:91–98

Wang J, Mao X, Wang R, Jing R et al (2019a) Identification of wheat stress-responding genes and TaPR-1-1 function by screening a cDNA yeast library prepared following abiotic stress. Sci Rep 9. https://doi.org/10.1038/s41598-018-37859-y

Wang J, Wang R, Mao X, Li L Chang X, Zhang X, Jing R (2019b) TaARF4 genes are linked to root growth and plant height in wheat. Ann Bot 124:903–915. https://doi.org/10.1093/aob/mcy218

Wang W, Wang Z, Li X, Ni Z, Hu Z, Xin M, Peng H, Yao Y, Sun Q, Guo W (2020) SnpHub: an easy-to-set-up web server framework for exploring large-scale genomic variation data in the post-genomic era with applications in wheat. GigaScience 9:giaa060

Wang X et al (2021) Genome-wide association study identifies QTL for thousand grain weight in winter wheat under normal- and late-sown stressed environments. Theor Appl Genet 134:143

Watson-Haigh NS, Suchecki R, Kalashyan E et al (2018) DAWN: a resource for yielding insights into the diversity among wheat genomes. BMC Genomics 19:941. https://doi.org/10.1186/s12864-018-5228-2

Wu P, Xie J, Hu J, Qiu D, Liu Z, Li J, Li MM, Zhang H, Yang L, Zhou Y, Zhang Z, Li H (2018) Development of molecular markers linked to powdery mildew resistance gene Pm4b by combining SNP discovery from transcriptome sequencing data with bulked segregant analysis (BSR-Seq) in wheat. Front Plant Sci 9. https://doi.org/10.3389/fpls.2018.00095

Wu Y, Li M, He Z, Dreisigacker S, Wen W, Jin H, Zhai S, Li F, Gao F, Liu J, Wang R (2020) Development and validation of high-throughput and low-cost STARP assays for genes underpinning economically important traits in wheat. Theor Appl Genet 133:2431–2450

Xia C, Zhang L, Zou C, Gu Y, Duan J, Zhao G, Wu J, Liu Y, Fang X, Gao L, Jiao Y, Sun J, Pan Y, Liu X, Jia J, Kong X (2016) A TRIM insertion in the promoter of Ms2 causes male sterility in wheat. Nat Commun 8:15407. https://doi.org/10.1038/ncomms15407

Yang Z, Wang Z, Wang W, Xie X, Chai L, Wang X, Feng X, Li J, Peng H, Su Z, You M, Yao Y, Xin M, Hu Z, Liu J, Liang R, Ni Z, Sun Q, Guo W (2022) ggComp enables dissection of germplasm resources and construction of a multiscale germplasm network in wheat. Plant Physiol 188:1950–1965

Yao H, Xie Q, Xue S, Ma Z et al (2019) HL2 on chromosome 7D of wheat (Triticum aestivum L.) regulates both head length and spikelet number. Theor Appl Genet 132. https://doi.org/10.1007/s00122-019-03315-2

Zhai S, He Z, Wen W et al (2016) Genome-wide linkage mapping of flour color-related traits and polyphenol oxidase activity in common wheat. Theor Appl Genet 129:377–394. https://doi.org/10.1007/s00122-015-2634-6

Zhang J, Huang S, Fosu-Nyarko J, Dell B, McNeil M, Waters I, Moolhuijzen P, Conocono E, Appels R (2008) The genome structure of the 1-FEH genes in wheat (Triticum aestivum L): new markers to track stem carbohydrates and grain filling QTLs in breeding. Mol Breed 22:339–351

Zhang W, Zhao G, Gao L, Kong X, Guo Z, Wu B, Jia J (2016). Functional studies of heading date-related gene TaPRR73, a paralog of Ppd1 in common wheat. Front Plant Sci 7. https://doi.org/10.3389/fpls.2016.00772

Zhang X, Liu G, Zhang L Xia C, Zhao T, Jia J Liu X Kong X (2018). Fine mapping of a novel heading date gene, TaHdm605, in hexaploid wheat. Front Plant Sci. https://doi.org/10.3389/fpls.2018.01059

Zhang L, Dong C, Chen Z, Gui L, Chen C, Li D, Xie Z, Zhang Q, Zhang X, Xia C, Liu X, Kong X, Wang J (2021) WheatGmap: a comprehensive platform for wheat gene mapping and genomic studies. Mol Plant 14:187–190

Zhao G, Zou C, Li K, Wang K, Li T, Gao L, Zhang X, Wang H, Yang Z, Liu X, Jiang W, Mao L, Kong X, Jiao Y, Jia J (2017) The *Aegilops tauschii* genome reveals multiple impacts of transposons. Nat Plants 3:946–955

Zhao J, Wang Z, Liu H et al (2019) Global status of 47 major wheat loci controlling yield, quality, adaptation and stress resistance selected over the last century. BMC Plant Biol 19:5. https://doi.org/10.1186/s12870-018-1612-y

Zhou Y, Chen Z, Cheng M, Chen J, Zhu T, Wang R, Liu Y, Qi P, Chen G, Jiang Q, Wei Y, Luo M-C, Nevo E, Allaby RG, Liu D, Wang J, Dvorák J, Zheng Y (2018) Uncovering the dispersion history, adaptive evolution and selection of wheat in China. Plant Biotechnol J 16:280–291

Zhou Y, Zhao X, Li Y, Xu J, Bi A, Kang L, Xu D, Chen H, Wang Y, Wang Y-g, Liu S, Jiao C, Lu H, Wang J, Yin C, Jiao Y, Lu F (2020) Triticum population sequencing provides insights into wheat adaptation. Nat Genet 52:1412–1422

Zhou Y, Bai S, Li H, Sun G, Zhang D, Ma F, Zhao X, Nie F, Li J, Chen L, Lv L, Zhu L, Fan R, Ge Y, Shaheen A, Guo G, Zhang Z, Ma J, Liang H, Qiu X, Hu J, Sun T, Hou J, Xu H, Xue S, Jiang W, Huang J, Li S, Zou C, Song C-P (2021) Introgressing the *Aegilops tauschii* genome into wheat as a basis for cereal improvement. Nat Plants 7:774–786

Zou C, Wang P, Xu Y (2016) Bulked sample analysis in genetics, genomics and crop improvement. Plant Biotechnol J 14:1941–1955

Rapid Cloning of Disease Resistance Genes in Wheat

10

Katherine L. D. Running and Justin D. Faris

Abstract

Wheat is challenged by rapidly evolving pathogen populations, resulting in yield losses. Plants use innate immune systems involving the recognition of pathogen effectors and subsequent activation of defense responses to respond to pathogen infections. Understanding the genes, genetic networks, and mechanisms governing plant-pathogen interactions is key to the development of varieties with robust resistance whether through conventional breeding techniques coupled with marker selection, gene editing, or other novel strategies. With regards to plant-pathogen interactions, the most useful targets for crop improvement are the plant genes responsible for pathogen effector recognition, referred to as resistance (R) or susceptibility (S) genes, because they govern the plant's defense response. Historically, the molecular identification of R/S genes in wheat has been extremely difficult due to the large and repetitive nature of the wheat genome. However, recent advances in gene cloning methods that exploit reduced representation sequencing methods to reduce genome complexity have greatly expedited R/S gene cloning in wheat. Such rapid cloning methods referred to as MutRenSeq, AgRenSeq, *k*-mer GWAS, and MutChromSeq allow the identification of candidate genes without the development and screening of high-resolution mapping populations, which is a highly laborious step often required in traditional positional cloning methods. These new cloning methods can now be coupled with a wide range of wheat genome assemblies, additional genomic resources such as TILLING populations, and advances in bioinformatics and data analysis, to revolutionize the gene cloning landscape for wheat. Today, 58 R/S genes have been identified with 42 of them having been identified in the past six years alone. Thus, wheat researchers now have the means to enhance global food security through the discovery of R/S genes, paving the way for rapid R gene deployment or S gene elimination, manipulation through gene editing, and understanding wheat-pathogen interactions at the molecular level to guard against crop losses due to pathogens.

K. L. D. Running
Department of Plant Sciences, North Dakota State University, Fargo, ND, USA
e-mail: katherine.running@ndsu.edu

J. D. Faris (✉)
USDA-Agricultural Research Service, Cereal Crops Research Unit, Edward T. Schafer Agricultural Research Center, Fargo, ND, USA
e-mail: justin.faris@usda.gov

R. Appels et al. (eds.), *The Wheat Genome*, Compendium of Plant Genomes,
https://doi.org/10.1007/978-3-031-38294-9_10

Keywords

Wheat diseases · Resistance genes · Rapid cloning

10.1 Introduction

Pathogens and pests pose a significant threat to global food security, affecting not just primary yields, but also the stability and distribution of production and the quality of food (Savary et al. 2017). An estimated 21.47% of global wheat yields are lost annually due to pathogens and pests (Savary et al. 2019), equating to ~210 million metric tons of grain per year, enough to bake 290 billion loaves of bread (Wulff and Krattinger 2022). Combining agronomic practices that reduce the initial disease inoculum and infection rate with selection of genetically resistant varieties is an effective crop disease management strategy, and to develop genetically resistant wheat, resistance (R) genes need to be identified, characterized, and deployed. In some diseases, for example tan spot or septoria nodorum blotch, susceptibility is conferred by dominant genes. In these cases, the priority is to remove or disrupt susceptibility (S) genes rather than deploy novel R genes. Gene cloning is crucial to the efficient deployment of R genes and removal of S genes, requiring the identification of the nucleotide sequence of a target gene and validating its function. Diversity and functional studies can assess the effects of genetic variation within an R or S gene on their respective resistance/susceptibility, allowing researchers to develop molecular markers targeting the variants, which can then be used to select breeding lines with the most beneficial alleles. Cloned R genes can also be introduced into modern cultivars via gene complementation or cross-hybridization, and S genes can be removed through marker-assisted elimination or gene editing. The methods and resources used to clone R and S genes are shared, and as such R and/or S genes will be referred to as "R/S genes" in this chapter.

Although over 460 R/S genes in wheat have been described (Hafeez et al. 2021), only 58 have been cloned (Table 10.1). The genome of hexaploid bread wheat is large and repetitive due, in part, to its evolutionary history, making it challenging to clone R/S genes. The basic seven-chromosome Triticeae progenitor split into the *Triticum* and *Aegilops* branches about 3 million years ago (MYA) (reviewed by Faris 2014). Modern-day bread wheat (*Triticum aestivum* ssp. *aestivum* L., $2n=6x=42$, AABBDD) is an allohexaploid that evolved as a result of two amphiploidization events involving the hybridization of two different species followed by spontaneous chromosome doubling through meiotic restitution division, several mutations, and interspecific gene flow. Around 0.5 MYA the wild diploid species *T. urartu* Tumanian ex Gandylian ($2n=2x=14$, AA) hybridized with a species similar to *Aegilops speltoides* Tausch ($2n=2x=14$, SS) to form tetraploid wheat *Triticum turgidum* ssp. *dicoccoides* Thell ($2n=4x=28$, AABB), also known as wild emmer. *T. turgidum* ssp. *durum* ($2n=4x=28$, AABB), durum wheat, is a free-threshing derivative of *T. turgidum* ssp. *dicoccoides*, and it is today widely cultivated and used to make pasta and other semolina-based products. The second amphiploidization event occurred around 8000 years ago. A *T. turgidum* ssp. and the diploid wild goat grass *Aegilops tauschii* Coss. ($2n=2x=14$, DD) hybridized to form hexaploid (common or bread) wheat *T. aestivum* ($2n=6x=42$, AABBDD). Due to the differential presence of *Ae. tauschii* lineage specific sequences in modern cultivars, it is possible that more than one hybridization even occurred between *T. turgidum* spp. and *Ae. tauschii* (Gaurav et al. 2022). Together, bread and durum wheat provide about 18% of the caloric intake of humans worldwide, but in some regions of the world, wheat accounts for over a third of the caloric and protein intake (Erenstein et al. 2022).

Table 10.1 Cloned resistance and susceptibility genes effective in wheat

Gene[a]	Gene function	Class[b]	Cloning method	Validation method	Origin	Year[c]	Reference
TaMlo-B1	Powdery mildew susceptibility	Transmembrane domains	Homology-based	Gene complementation (Elliott et al. 2002) virus-induced gene silencing (Várallyay et al. 2012), TALEN-mediated gene knockout (Wang et al. 2014), TILLING (Acevedo-Garcia et al. 2017; Ingvardsen et al. 2019)	*T. aestivum/T. turgidum*	2002	Elliott et al. (2002)
Lr10	Leaf rust resistance	NLR	Mapping	Mutagenesis (EMS), gene complementation	*T. aestivum*	2003	Feuillet et al. (2003)
Lr21	Leaf rust resistance	NLR	Mapping	Gene complementation	*T. aestivum*	2003	Huang et al. (2003)
Pm3a, Pm3b, Pm3d, Pm3f	Powdery mildew resistance	NLR	Mapping	Transient expression, mutagenesis (γ-irradiation)	*T. aestivum*	2005/04	Srichumpa et al. (2005)/ Yahiaoui et al. (2004)
Lr1	Leaf rust resistance	NLR	Mapping	Virus-induced gene silencing, gene complementation	*T. aestivum*	2007	Cloutier et al. (2007)
Lr34/Yr18/Sr57/ Pm38/Ltn1	Leaf rust, stripe rust, stem rust, powdery mildew, and leaf tip necrosis resistance	Abscisic acid transporter	Mapping	Mutagenesis (γ-irradiation, sodium azide)	*T. aestivum*	2009	Krattinger et al. (2009)
Yr36 (*WKS1*)	Stripe rust resistance	START Kinase	Mapping	Mutagenesis (EMS), gene complementation	*T. turgidum* ssp. *dicoccoides*	2009	Fu et al. (2009)
Tsn1	Septoria nodorum blotch and tan spot resistance	S/TPK-NLR	Mapping	Mutagenesis (EMS), CRISPR-Cas9-mediated gene knockout (Poddar et al. 2023)	*T. turgidum* ssp. *durum*	2010	Faris et al. (2010)
TaMlo-A1	Powdery mildew susceptibility	Transmembrane domains	Homology-based	Virus-induced gene silencing (Várallyay et al. 2012), TALEN-mediated gene knockout (Wang et al. 2014), CRISPR-Cas9-mediated gene knockout (Wang et al. 2014), TILLING (Acevedo-Garcia et al. 2017; Ingvardsen et al. 2019)	*T. aestivum/T. turgidum*	2012	Várallyay et al. (2012)

(continued)

Table 10.1 (continued)

Gene[a]	Gene function	Class[b]	Cloning method	Validation method	Origin	Year[c]	Reference
TaMlo-D1	Powdery mildew susceptibility	Transmembrane domains	Homology-based	Virus-induced gene silencing (Várallyay et al. 2012), TALEN-mediated gene knockout (Wang et al. 2014), TILLING (Acevedo-Garcia et al. 2017)	*T. aestivum*	2012	Várallyay et al. (2012)
Sr33	Stem rust resistance	NLR	Mapping	Mutagenesis (EMS), gene complementation	*Ae. tauschii*	2013	Periyannan et al. (2013)
Sr35	Stem rust resistance	NLR	Mapping	Mutagenesis (EMS), gene complementation	*T. monoccocum*	2013	Saintenac et al. (2013)
Pm8	Powdery mildew resistance	NLR	Homology-based	Transient expression, gene complementation	*Secale cereale*	2013	Hurni et al. (2013)
Yr10 (Yr10cg)[d]	Stripe rust resistance	NLR	Mapping	Gene complementation	*T. aestivum*	2014	Liu et al. (2014)
Lr67/Yr46/Sr55/Pm46/Ltn3	Leaf rust, stripe rust, stem rust, powdery mildew resistance, and leaf tip necrosis resistance	Hexose transporter	Mapping	Mutagenesis (EMS), gene complementation	*T. aestivum*	2015	Moore et al. (2015)
Sr50	Stem rust resistance	NLR	Mapping	Mutagenesis (EMS), gene complementation	*Secale cereale*	2015	Mago et al. (2015)
Fhb1[e]	Fusarium head blight resistance	Pore-forming toxin-like gene	Mapping	Mutagenesis (EMS), RNAi-induced gene silencing, gene complementation	*T. aestivum*	2016	Rawat et al. (2016)
Snn1	Septoria nodorum blotch	WAK	Mapping	Mutagenesis (EMS), gene complementation	*T. aestivum*	2016	Shi et al. (2016)
Pm2a	Powdery mildew resistance	NLR	MutChromSeq	Mutagenesis (EMS)	*T. aestivum*	2016	Sánchez-Martín et al. (2016)
Sr22a, Sr22b	Stem rust resistance	NLR	MutRenSeq	Mutagenesis (EMS), gene complementation	*T. boeoticum T. monococcum*	2016	Steuernagel et al. (2016)
Sr45	Stem rust resistance	NLR	MutRenSeq	Mutagenesis (EMS) (Steuernagel et al. 2016), gene complementation (Arora et al. 2019)	*Ae. tauschii*	2016	Steuernagel et al. (2016)
Sr13	Stem rust resistance	NLR	Mapping	TILLING, gene complementation	*T. durum*	2017	Zhang et al. (2017)

(continued)

Table 10.1 (continued)

Gene[a]	Gene function	Class[b]	Cloning method	Validation method	Origin	Year[c]	Reference
Lr22a	Leaf rust resistance	NLR	Mapping and TACCA	Mutagenesis (EMS)	*Ae. tauschii*	2017	Thind et al. (2017)
Pm21[f]	Powdery mildew resistance	NLR	Mapping, MutRenSeq	EMS, gene complementation	*Dasypyrum villosum*	2017/18	He et al. (2017)/Xing et al. (2018)
Pm60, MlWE18	Powdery mildew resistance	NLR	Mapping	Virus-induced gene silencing, gene complementation, transient expression	*T. urartu*	2018	Zou et al. (2018)
Stb6	Septoria tritici blotch resistance	WAK-like protein	Mapping	Gene complementation, virus-induced gene silencing, TILLING	*T. aestivum*	2018	Saintenac et al. (2018)
Sr21	Stem rust resistance	NLR	Mapping	Mutagenesis (EMS), gene complementation	*T. monoccocum*	2018	Chen et al. (2018)
Yr15	Stripe rust resistance	Tandem kinase-pseudokinase	Mapping	Mutagenesis (EMS), gene complementation	*T. dicoccoides*	2018	Klymiuk et al. (2018)
Yr5a (*Yr5*), *Yr5b* (*YrSP*)	Stripe rust resistance	BED-NLR	MutRenSeq	Mutagenesis (EMS)	*T. aestivum*	2018	Marchal et al. (2018)
Yr7	Stripe rust resistance	BED-NLR	MutRenSeq	Mutagenesis (EMS)	*T. aestivum*	2018	Marchal et al. (2018)
Pm17	Powdery mildew resistance	NLR	Homology-based	Transient expression, gene complementation	*Secale cereale*	2018	Singh et al. (2018)
Sr46	Stem rust resistance	NLR	AgRenSeq	Mutagenesis (EMS), gene complementation	*Ae. tauschii*	2019	Arora et al. (2019)
YrAS2388R	Stripe rust resistance	NLR	Mapping	Mutagenesis (EMS), gene complementation	*Ae. tauschii*	2019	Zhang et al. (2019)
Sr60 (*WKS2*)	Stem rust resistance	Tandem kinase	Mapping	Gene complementation	*T. monoccocum*	2020	Chen et al. (2020)
Pm5e	Powdery mildew resistance	NLR	Mapping	Mutagenesis (EMS), gene complementation	*T. aestivum*	2020	Xie et al. (2020)
Pm24	Powdery mildew resistance	Tandem kinase	Mapping	Mutagenesis (EMS), gene complementation	*T. aestivum*	2020	Lu et al. (2020)
Pm41	Powdery mildew resistance	NLR	Mapping	Mutagenesis (EMS), gene complementation	*T. turgidum* ssp. *dicoccoides*	2020	Li et al. (2020)
YrU1	Stripe rust resistance	ANK-NLR-WRKY	Mapping	Gene complementation	*T. urartu*	2020	Wang et al. (2020b)

(continued)

Table 10.1 (continued)

Gene[a]	Gene function	Class[b]	Cloning method	Validation method	Origin	Year[c]	Reference
Sm1	Orange wheat blossom midge resistance	NLR-kinase-MSP domains	Mapping and haplotype analysis	Mutagenesis (EMS)	*T. aestivum*	2020	Walkowiak et al. (2020)
Fhb7	Fusarium head blight resistance	Glutathione S-transferase	Mapping	Mutagenesis (EMS), virus-induced gene silencing, gene complementation	*Thinopyrum elongatum*	2020	Wang et al. (2020a)
Pm1a	Powdery mildew resistance	NLR	Mapping, MutChromSeq	Mutagenesis (EMS), gene complementation	*T. aestivum*	2021	Hewitt et al. (2021a)
Sr26	Stem rust resistance	NLR	MutRenSeq	Mutagenesis (EMS), gene complementation	*Thinopyrum ponticum*	2021	Zhang et al. (2021a)
Sr61	Stem rust resistance	NLR	MutRenSeq	Mutagenesis (EMS), gene complementation	*Thinopyrum ponticum*	2021	Zhang et al. (2021a)
Lr14a	Leaf rust resistance	Ankyrin transmembrane domain protein	MutChromSeq	Mutagenesis (EMS), virus-induced gene silencing	*T. aestivum*	2021	Kolodziej et al. (2021)
Snn3-D1	Septoria nodorum blotch	PK-MSP	Mapping	Mutagenesis (EMS)	*Ae. tauschii*	2021	Zhang et al. (2021b)
Stb16q	Septoria tritici blotch resistance	CRK	Mapping	EMS, gene complementation	*Ae. tauschii*	2021	Saintenac et al. (2021)
Pm4a, Pm4b (Pm4c), Pm4d (Pm4e), Pm4f, Pm4g, Pm4h[g]	Powdery mildew resistance	MCTP-kinase	MutChromSeq	Mutagenesis (EMS), gene complementation, virus-induced gene silencing	*T. aestivum*	2021	Sánchez-Martín et al. (2021)
Sr27	Stem rust resistance	NLR	MutRenSeq	Mutagenesis (EMS), transient expression,	*Triticale (Secale cereale genome)*	2021	Upadhyaya et al. (2021)
Lr13, Yr27[h]	Leaf rust resistance, stripe rust resistance	NLR	MutRenSeq/ Mapping	Mutagenesis (EMS), virus-induced gene silencing, gene complementation	*T. aestivum*	2021/22	Hewitt et al. (2021b); Yan et al. (2021); Athiyannan et al. (2022)
SrTA1662	Stem rust resistance	NLR	*k*-mer GWAS	Gene complementation	*Ae. tauschii*	2022	Gaurav et al. (2022)
Sr62	Stem rust resistance	Tandem kinase	Mapping	Mutagenesis (EMS), gene complementation	*Ae. sharonensis*	2022	Yu et al. (2022)
TaPDIL5-1-4A	Wheat yellow mosaic virus susceptibility	Protein disulfide isomerase like	Homology	CRISPR-Cas9-mediated gene knockout	*T. aestivum*	2022	Kan et al. (2022)
TaPDIL5-1-4B	Wheat yellow mosaic virus susceptibility	Protein disulfide isomerase like	Homology	CRISPR-Cas9-mediated gene knockout	*T. aestivum*	2022	Kan et al. (2022)

(continued)

Table 10.1 (continued)

Gene[a]	Gene function	Class[b]	Cloning method	Validation method	Origin	Year[c]	Reference
TaPDIL5-1-4D	Wheat yellow mosaic virus susceptibility	Protein disulfide isomerase like	Homology	CRISPR-Cas9-mediated gene knockout	*T. aestivum*	2022	Kan et al. (2022)
Lr42	Leaf rust resistance	NLR	BSR-Seq[i]	Mutagenesis (EMS), gene complementation, virus-induced gene silencing	*Ae. tauschii*	2022	Lin et al. (2022)
TaYRG1	Stripe rust resistance	NLR	Transcriptomics	Virus-induced gene silencing, gene complementation	*T. aestivum*	2022	Zhang et al. (2022a)
TaSTP3	Stripe rust susceptibility	Sugar transporter	Transcriptomics	Virus-induced gene silencing, RNAi-induced gene silencing, gene complementation	*T. aestivum*	2022	Huai et al. (2022)
TaRPP13L1-3D	Powdery mildew resistance	NB-ARC	Transcriptomics	Gene complementation, virus-induced gene silencing	*T. aestivum*	2022	Zhang et al. (2022b)

[a] Alternate gene designations for resistance/susceptibility to the same pathogen are listed in parenthesis. Characterized alleles of a gene are separated by commas. Alternate gene designations for resistance/susceptibility to different pathogens are separated with a forward slash

[b] Class designations are abbreviated as follows: nucleotide-binding domain leucine-rich repeat containing (NLR), serine/threonine protein kinase (S/TPK), wall-associated receptor kinase (WAK), ankyrin-repeat (ANK), protein kinase (PK), major sperm protein (MSP), cysteine-rich receptor-like kinase (CRK), multiple C2 domain and transmembrane region proteins (MCTP), nucleotide binding (NB)

[c] Year cloned refers to the first reported functional validation of the gene

[d] *Yr10* provides race-specific resistance to yellow rust. A later analysis determined that *Yr10* does not provide race-specific resistance in the manner expected and therefore may not be *Yr10*. Instead, the authors refer to the cloned *Yr10* as *Yr10* candidate gene or *Yr10cg* (Yuan et al. 2018)

[e] Two later studies identified *Fhb1* as a histidine-rich calcium-binding protein (Li et al. 2019; Su et al. 2019)

[f] *Pm21* was initially reported as a Sr/Thr kinase (Cao et al. 2011)

[g] *Pm4f* and *Pm4g* appear to be susceptible alleles of *Pm4*

[h] *Yr27* and *Lr13* are distinct alleles of hybrid necrosis gene *Ne2*

[i] Bulked-segregant analysis and RenSeq

Despite their polyploid nature, bread and durum wheat behave like diploid plants genetically, with homologous chromosomes pairing and segregating during meiosis. The pairing of homoeologous chromosomes is prevented by genes *Ph1* and *Ph2* (Riley and Chapman 1958; Sears and Okamoto 1958; Mello-Sampayo and Lorente 1968) with the resulting diploid-like pairing of wheat chromosomes in meiosis simplifying segregation studies and genetic mapping of traits. Due to their formation through amphiploidization, hexaploid and tetraploid wheats often have three or two copies of each gene, respectively, called homoeologous genes. Homoeologous genes are often highly conserved, with ~97% identity across their coding regions (Schreiber et al. 2012), and this high sequence conservation among homoeologous genes hinders the development of homoeolog-specific molecular markers. Additionally, approximately 85% of the wheat genome is comprised repetitive elements (Wicker et al. 2018), making it difficult to design molecular markers that only target one locus for use in molecular mapping or marker-assisted selection.

Bread and durum wheat genomes are relatively large at 12 and 17 Gb, respectively (Bennett and Smith 1976). The sequencing and assembly of such large genomes are computationally challenging and further complicated by the highly repetitive nature of wheat genomes and interchromosomal gene duplications (IWGSC et al. 2014). The complexity of the wheat genome has hampered the generation of genomic data and bioinformatic analysis. Despite the challenges, multiple high-quality genome assemblies have been constructed (Table 10.2). Genome assemblies are used to design molecular markers and bait libraries, assess candidate genes, and evaluate structural variation as well as acting as a foundation for developing genomic resources and tools that aid in the cloning of R/S genes.

The first cloned S and R genes in wheat, *TaMlo-B1* and *Lr10*, were published in 2002 (Elliott et al. 2002) and 2003 (Feuillet et al. 2003), respectively. Since then, 48 more R/S genes have been cloned from *Triticum* or *Aegilops* species, and an additional eight R/S genes have been cloned from related species and shown to be functional in wheat (Table 10.1, current as of 8/1/2022). In just the last two years, more R/S genes were cloned than were cloned in the first decade of R/S gene cloning. Here, we review the surge of genomic resources and gene cloning methods that have contributed to the acceleration of R/S gene cloning in wheat.

10.2 Advances in Wheat Genome Sequencing

High-quality genomic sequences and assemblies act as the basis for gene cloning efforts in wheat, and the recognition of this requirement led to the formation of the International Wheat Genome Sequencing Consortium (IWGSC) in 2005. Several hexaploid, tetraploid, and diploid *Triticum* full genome assemblies have been released in the last five years (Table 10.2). The bread wheat variety CHINESE SPRING was selected for sequencing due to the extensive genetic and molecular resources developed using this variety (Gill et al. 2004), including aneuploid stocks developed by Ernie Sears that could be used to physically map genes and markers to specific chromosomes (Sears 1954, 1966; Sears and Sears 1978). Segmental deletion lines (Endo and Gill 1996) further specified physical regions within chromosomal arms and were used to map 16,000 expressed sequence tag (EST) loci (Qi et al. 2004).

Hexaploid wheat was estimated to be 17 Gb and included families of DNA sequences that were highly repetitive (Bennett and Smith 1976). A reduced-representation sequencing approach was used to reduce the genome complexity and size (IWGSC et al. 2014), making use of CHINESE SPRING ditelosomic stocks developed by Sears and Sears (1978) to isolate each chromosome arm by flow cytometry, and BAC libraries were subsequently constructed from the DNA of individual arms. The bin-mapped ESTs were used to assess the purity of the sorted chromosome fractions (Qi et al. 2004). Short read paired-end sequences of each BAC library were

Table 10.2 *Triticum* and *Aegilops* assemblies

Species	Genotype	Year	Genomes	Type	Reference	Doi or link
Ae. tauschii	AL8/78	2013	D	Scaffold	Jia et al. (2013)	https://doi.org/10.1038/nature12028
T. urartu	G1812/PI428198	2013	A	Scaffold	Ling et al. (2013)	https://doi.org/10.1038/nature11997
T. turgidum ssp. *durum*	CAPPELLI	2014	AB	Scaffold	IWGSC et al. (2014)	https://doi.org/10.1126/science.1251788
T. aestivum	CHINESE SPRING	2014	B	Pseudomolecule	Choulet et al. (2014)	https://doi.org/10.1126/science.1249721
T. aestivum	CHINESE SPRING	2014	ABD	Scaffold	IWGSC et al. (2014)	https://doi.org/10.1126/science.1251788
Ae. speltoides	ERX391140	2014	SS	Scaffold	IWGSC et al. (2014)	https://doi.org/10.1126/science.1251788
T. turgidum ssp. *durum*	STRONGFIELD	2014	AB	Scaffold	IWGSC et al. (2014)	https://doi.org/10.1126/science.1251788
Synthetic hexaploid	W7984	2015	ABD	Scaffold	Chapman et al. (2015)	https://doi.org/10.1186/s13059-015-0582-8
T. aestivum	CHINESE SPRING doubled haploid (Dv418)	2017	ABD	Scaffold	Zimin et al. (2017a)	https://doi.org/10.1093/gigascience/gix097
Ae. tauschii	AL8/78	2017	D	Pseudomolecule	Luo et al. (2017)	https://doi.org/10.1038/nature24486
Ae. tauschii	AL8/78	2017	D	Pseudomolecule	Zhao et al. (2017)	https://doi.org/10.1038/s41477-017-0067-8
Ae. tauschii	AL8.78	2017	D	Scaffold	Zimin et al. (2017b)	https://doi.org/10.1101/gr.213405.116
T. aestivum	CHINESE SPRING	2017	ABD	Scaffold	Clavijo et al. (2017)	https://doi.org/10.1101/gr.217117.116
T. turgidum ssp. *durum*	KRONOS	2017	AB	Scaffold	N/A	http://opendata.earlham.ac.uk/Triticum_turgidum/
T. aestivum ssp. *dicoccoides*	ZAVITAN	2017	AB	Pseudomolecule	Avni et al. (2017)	https://doi.org/10.1126/science.aan0032
T. aestivum	CHINESE SPRING	2018	ABD	Pseudomolecule	IWGSC et al. (2018)	https://doi.org/10.1126/science.aar7191
T. urartu	G1812/PI428198	2018	A	Pseudomolecule	Ling et al., (2018)	https://doi.org/10.1038/s41586-018-0108-0
T. turgidum ssp. *durum*	SVEVO	2019	AB	Pseudomolecule	Maccaferri et al. (2019)	https://doi.org/10.1038/s41588-019-0381-3
T. aestivum ssp. *dicoccoides*	ZAVITAN	2019	AB	Pseudomolecule	Zhu et al. (2019)	https://doi.org/10.1534/g3.118.200902
T. aestivum	2670/PI 190962	2020	ABD	Pseudomolecule	Walkowiak et al. (2020)	https://doi.org/10.1038/s41586-020-2961-x
T. aestivum	ARINA-LRFOR	2020	ABD	Pseudomolecule	Walkowiak et al. (2020)	https://doi.org/10.1038/s41586-020-2961-x
T. aestivum	CADENZA	2020	ABD	Scaffold	Walkowiak et al. (2020)	https://doi.org/10.1038/s41586-020-2961-x
T. aestivum	CDC LANDMARK	2020	ABD	Pseudomolecule	Walkowiak et al. (2020)	https://doi.org/10.1038/s41586-020-2961-x
T. aestivum	CDC STANLEY	2020	ABD	Pseudomolecule	Walkowiak et al. (2020)	https://doi.org/10.1038/s41586-020-2961-x
T. aestivum	CLAIRE	2020	ABD	Scaffold	Walkowiak et al. (2020)	https://doi.org/10.1038/s41586-020-2961-x

(continued)

Table 10.2 (continued)

Species	Genotype	Year	Genomes	Type	Reference	Doi or link
T. aestivum	JAGGER	2020	ABD	Pseudomolecule	Walkowiak et al. (2020)	https://doi.org/10.1038/s41586-020-2961-x
T. aestivum	JULIUS	2020	ABD	Pseudomolecule	Walkowiak et al. (2020)	https://doi.org/10.1038/s41586-020-2961-x
T. aestivum	LONGREACH-LANCER	2020	ABD	Pseudomolecule	Walkowiak et al. (2020)	https://doi.org/10.1038/s41586-020-2961-x
T. aestivum	MACE	2020	ABD	Pseudomolecule	Walkowiak et al. (2020)	https://doi.org/10.1038/s41586-020-2961-x
T. aestivum	NORIN 61	2020	ABD	Pseudomolecule	Walkowiak et al. (2020)	https://doi.org/10.1038/s41586-020-2961-x
T. aestivum	PARAGON	2020	ABD	Scaffold	Walkowiak et al. (2020)	https://doi.org/10.1038/s41586-020-2961-x
T. aestivum	ROBIGUS	2020	ABD	Scaffold	Walkowiak et al. (2020)	https://doi.org/10.1038/s41586-020-2961-x
T. aestivum	SY MATTIS	2020	ABD	Pseudomolecule	Walkowiak et al. (2020)	https://doi.org/10.1038/s41586-020-2961-x
T. aestivum	WEEBILL 1	2020	ABD	Scaffold	Walkowiak et al. (2020)	https://doi.org/10.1038/s41586-020-2961-x
T. aestivum ssp. *tibetanum* Shao	ZANG1817	2020	ABD	Pseudomolecule	Guo et al. (2020)	https://doi.org/10.1038/s41467-020-18738-5
Ae. tauschii	AL8/78	2021	D	Pseudomolecule	Wang et al. (2021)	https://doi.org/10.1093/g3journal/jkab325
Ae. tauschii (AY17)	AY17	2021	D	Pseudomolecule	Zhou et al. (2021)	https://doi.org/10.1038/s41477-021-00934-w
Ae. tauschii (AY61)	AY61	2021	D	Pseudomolecule	Zhou et al. (2021)	https://doi.org/10.1038/s41477-021-00934-w
T. aestivum	CHINESE SPRING (RefSeq v2.1)	2021	ABD	Pseudomolecule	Zhu et al. (2021)	https://doi.org/10.1111/tpj.15289
T. aestivum	FIELDER	2021	ABD	Pseudomolecule	Sato et al. (2021)	https://doi.org/10.1093/dnares/dsab008
T. aestivum	RENAN	2021	ABD	Pseudomolecule	Aury et al. (2022)	https://doi.org/10.1093/gigascience/giac034
Ae. tauschii (T093)	T093	2021	D	Pseudomolecule	Zhou et al. (2021)	https://doi.org/10.1038/s41477-021-00934-w
Ae. tauschii (XJ02)	XJ02	2021	D	Pseudomolecule	Zhou et al. (2021)	https://doi.org/10.1038/s41477-021-00934-w
Ae. longissima	AEG-6782-2	2022	S^l	Pseudomolecule	Avni et al. (2022)	https://doi.org/10.1111/tpj.15664
Ae. speltoides	AEG-9674-1	2022	S	Pseudomolecule	Avni et al. (2022)	https://doi.org/10.1111/tpj.15664
Ae. sharonensis	AS_1644	2022	S^{sh}	Pseudomolecule	Yu et al. (2022)	https://doi.org/10.1038/s41467-022-29132-8
T. aestivum	KARIEGA	2022	ABD	Pseudomolecule	Athiyannan et al. 2022)	https://doi.org/10.1038/s41588-022-01022-1
T. aestivum	SONMEZ	2022	ABD	Pseudomolecule	Akpinar et al. (2022)	https://doi.org/10.21203/rs.3.rs-1095548/v1
T. aestivum	ATTRAKTION	2022	ABD	Pseudomolecule	Kale et al. (2022)	https://doi.org/10.1111/pbi.13843
Ae. bicornis	TB01	2022	S^b	Pseudomolecule	Li et al. (2022)	https://doi.org/10.1016/j.molp.2021.12.019

(continued)

Table 10.2 (continued)

Species	Genotype	Year	Genomes	Type	Reference	Doi or link
Ae. searsii	TE01	2022	S^s	Pseudomolecule	Li et al. (2022)	https://doi.org/10.1016/j.molp.2021.12.019
Ae. sharonensis	TH02	2022	S^{sh}	Pseudomolecule	Li et al. (2022)	https://doi.org/10.1016/j.molp.2021.12.019
Ae. longissima	TL05	2022	S^l	Pseudomolecule	Li et al. (2022)	https://doi.org/10.1016/j.molp.2021.12.019
Ae. speltoides	TS01	2022	S	Pseudomolecule	Li et al. (2022)	https://doi.org/10.1016/j.molp.2021.12.019

assembled resulting in a 10.2 Gb draft assembly referred to as the Chinese Spring Survey Sequences (CSS) and represented 61% of the genome sequence (IWGSC et al. 2014).

A pseudomolecule level assembly of chromosome 3B was produced separately using a minimum tiling path of 8,452 BACs sequenced with Roche/454 paired-end reads (Choulet et al. 2014). After scaffold assembly, Illumina reads from flow sorted chromosome 3B were used to fill gaps. A detailed SNP-based genetic map from the CHINESE SPRING × RENAN population was used to orient and order scaffolds. Ultimately, the pseudomolecule level assembly represented 93% of chromosome 3B. A total of 124,201 high-confidence gene loci were annotated in the CSS and chromosome 3B assembly (IWGSC et al. 2014).

Whole genome shotgun (WGS) assemblies of the *Triticum turgidum* ssp. *durum* cultivars CAPPELLI and STRONGFIELD were released in 2014 alongside an assembly of *Ae. speltoides* accession ERX391140 (SS) (IWGSC et al. 2014). Although these assemblies consisted of numerous small contigs with unknown order, orientation, and space between contigs, partly due to the piling of repetitive elements, they offer a draft assembly of low-copy DNA and therefore can be used to identify alleles, design gene-specific markers, or compare genes and gene families among assemblies. Chapman et al. (2015) integrated WGS and genetic mapping to assemble and order contigs of the synthetic hexaploid W7984. Despite the WGS method and lack of chromosome isolation via flow sorting, the assembly was 9.1 Gb, just 1.1 Gb smaller than the CSS assembly.

With the growth of sequencing and assembly methods, more wheat scaffold and pseudomolecule level assemblies became available (Figs. 10.1 and 10.2). As of August 2022, 46 unique accessions have scaffold and/or pseudomolecule level assemblies (Table 10.2). In 2020, there was a significant increase in the number of hexaploid accessions with pseudomolecule or scaffold level assemblies. Through a large international collaborative effort, Walkowiak et al. (2020) published the 10+Wheat Genomes' paper, which included pseudomolecule assemblies of nine bread wheat lines and one *T. aestivum* ssp. *spelta* accession plus the scaffold level assemblies of five additional bread wheat

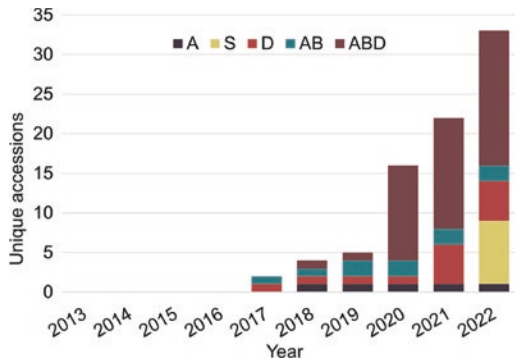


Fig. 10.1 Cumulative accessions with pseudomolecule level assemblies. Color corresponds to the subgenome of the accession

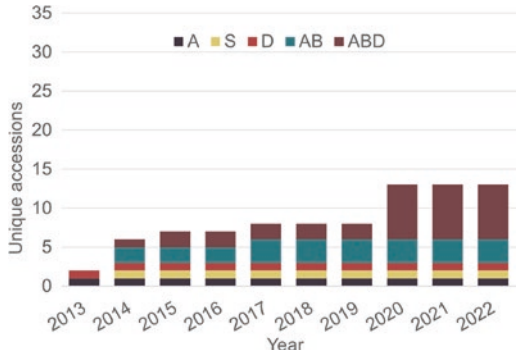


Fig. 10.2 Cumulative accessions with scaffold level assemblies. Color corresponds to the subgenome of the accession

lines. Prior to this, CHINESE SPRING and the synthetic hexaploid W7984 were the only hexaploids with either a pseudomolecule or scaffold level assembly. Principal component analysis of exome sequence capture alleles in ~1200 hexaploid accessions revealed that CHINESE SPRING was genetically distant from other hexaploid wheats (Walkowiak et al. 2020). The accessions included in the Walkowiak et al. (2020) paper were selected to more accurately represent the full diversity of hexaploid wheat allowing analysis of intergenome variability. The genome of the Tibetan semi-wild wheat (*T. aestivum* ssp. *tibetanum* Shao) accession ZANG1817 was also published the same year (Guo et al. 2020).

Most of the *Triticum* and *Aegilops* assemblies and genome browsers are hosted on websites. Not all assemblies are hosted on a single website and different assembly and annotation versions are available on different websites, so care should be taken when comparing assemblies or annotations from different sources. Many of these websites host additional resources that may be useful in the gene cloning and characterization process, such as molecular markers, exome capture data, varietal SNPs, and TILLING mutants.

The following are useful websites for accessing the genome assemblies:

- GrainGenes (Yao et al. 2022)—https://wheat.pw.usda.gov.
- Ensembl Plants—http://plants.ensembl.org/Triticum_aestivum.
- URGI—https://urgi.versailles.inrae.fr/blast/.
- Grassroots Infrastructure—https://grassroots.tools/service/blast-blastnCerealsDB.
- https://www.cerealsdb.uk.net/cerealgenomics/CerealsDB/blast_WGS.php.

10.3 Map-based Cloning

Map-based cloning was used to clone the first wheat R gene, *Lr10* (Feuillet et al. 2003). Since then, map-based cloning has been the most frequently used method to clone R/S genes in wheat (around 50%, Table 10.1). Map-based cloning uses the genetic relationship between a gene and molecular markers to place a gene on a genetic map. Originally, an iterative approach termed chromosome walking was used to define the candidate gene region. The two closest molecular markers were then used to screen large-insert libraries of cloned fragments of DNA (yeast artificial chromosomes or bacterial artificial chromosomes, YACs or BACs) to identify flanking clones, and new markers developed from the ends of the clones were used to rescreen the library and "walk" closer to the gene of interest until a clone containing the gene was identified. Sequencing of the clone(s) spanning markers defined by flanking genetic recombinants would reveal the nucleotide sequence of the R/S gene. While we still use the term "cloning," the development and screening of large-insert genomic clones are seldom still necessary to clone a gene. The development of molecular markers and subsequent high-density, or saturation, mapping of target R/S genes in segregating populations is a critical step in the map-based cloning process. Historically, high-density mapping was conducted on a low-throughput basis using markers such as restriction fragment length polymorphisms (RFLPs), amplified fragment length polymorphisms (AFLPs), or

simple-sequence repeats (SSRs, or microsatellites). Recent advances in high-throughput genotyping technologies such as Diversity Arrays Technology (DArT), DNA SNP arrays, custom Kompetitive allele-specific PCR (KASP) arrays, or genotyping by sequencing offer high-density genotyping at affordable costs. These genotyping technologies can also be used in combination with a bulked segregant analysis (BSA) approach to quickly find markers associated with a phenotype without having to genotype a large mapping population (see also Chap. 9).

The size of the candidate gene region, as defined by the genetic region between the closest markers flanking the R/S gene, is dependent on both the marker density and the recombination rate. In a population of fixed size, such as a recombinant inbred or doubled haploid population, there is a finite number of recombination events. Sometimes, there are not enough recombination events in a population to reduce the candidate gene region to a reasonable size. If the marker density is too low, recombination events can go undetected, resulting in a larger candidate gene region. Additional molecular markers in a region cosegregating with the gene will not increase resolution. Even in cases where marker density and recombination rate are high, a candidate gene region may be gene-rich, making it difficult to identify the trait-associated gene. Map-based cloning also requires access to the DNA sequence between the flanking markers. This need is often met by the multiple sequenced wheat genomes. It is important to remember that even if the sequenced wheat genotypes do not carry a functional allele of a target R/S gene, they may carry a nonfunctional allele. As such, it may be useful to identify candidate genes even in genotypes that do not display the desired resistant or susceptible phenotype. If the phenotypes of the sequenced wheat genomes are known, candidate genes may be eliminated based on a comparison of gene content between lines with and without the trait of interest (Running and Faris, unpublished).

If the sequenced wheat genotypes do not carry an allele of the R/S gene, or when the R/S gene is in an area of low recombination, such as an introgressed segment from a wild relative or near a centromere, alternate gene cloning methods may be more appropriate. Map-based cloning can be slow, dependent on the generation of the mapping population, and requires screening of 1000's of recombinant gametes.

10.4 Reduced-Representation Sequencing Methods

Reduced-representation sequencing (RRS) is a key step in the rapid cloning methods that are used in wheat (described below), and it can be combined with traditional map-based cloning methods to quickly identify candidate genes. RRS reduces genome complexity and therefore the cost and time of sequencing and analysis. The three main methods of RRS are transcriptome or RNA sequencing, exome capture, and chromosome flow sorting (Fig. 10.3). These methods allow preferential sequencing of more relevant spaces, either genic regions or promoters, or the specific chromosome containing an R/S gene. In some cases, RRS methods are incorporated into rapid cloning methods.

10.4.1 Exome Capture

In exome capture, the baits, or capture probes, hybridize to the targets and then are bound by streptavidin-coated magnetic beads. The magnetic beads are "captured" by a magnet, unbound DNA is washed away, and the remaining target-enriched library is amplified and sequenced. Capture probes' assays can target genes, promoters, and even specific types of genes like nucleotide-binding domain leucine-rich repeat containing genes (NLRs). Exome capture assays targeting the genic regions of wheat have been designed from the sequenced wheat genomes, each using an increasing design space size as additional wheat genome sequences became available.

Jordan et al. (2015) designed an exome capture probe assay called the "wheat exome capture" (WEC) using a design space of 110 Mb

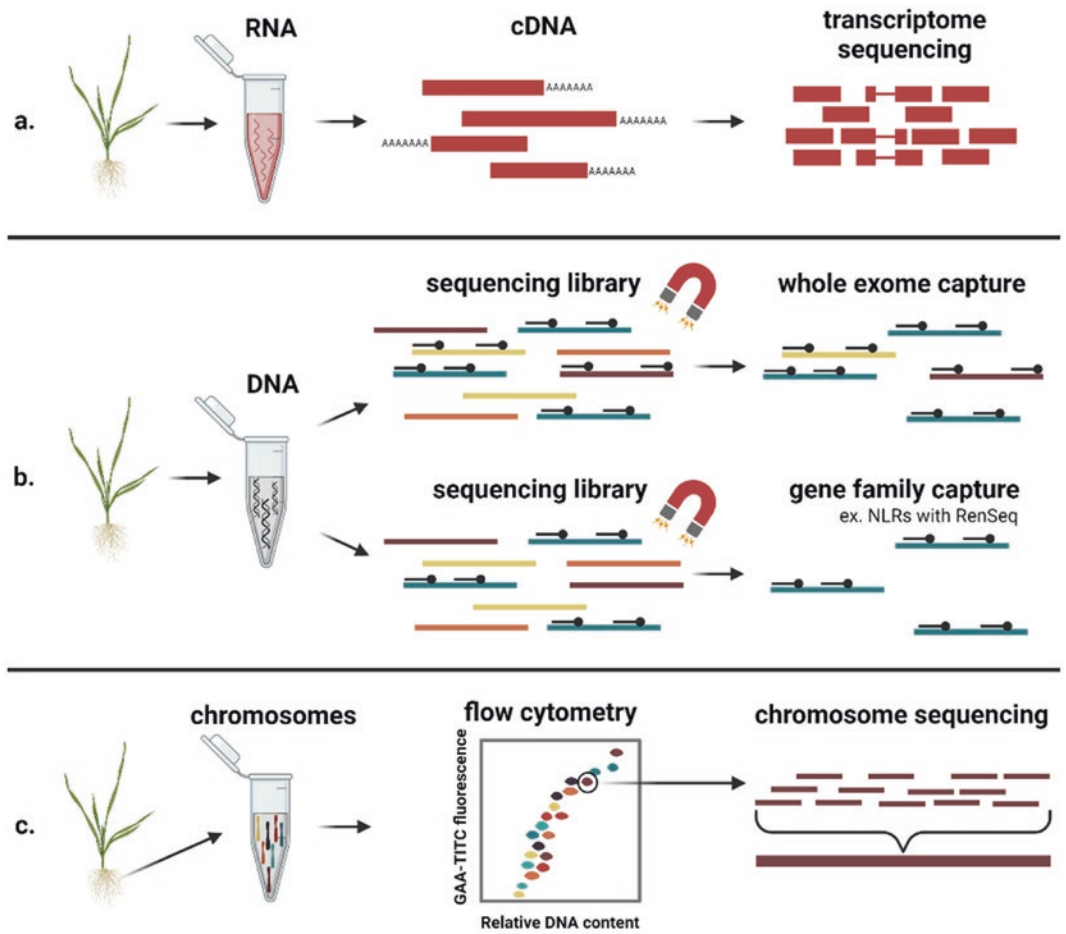


Fig. 10.3 Reduced sequencing methods. **a** Transcriptome sequencing. RNA is isolated from tissue and reverse transcribed into cDNA, which is sequenced and mapped to a reference assembly. **b** Exome sequencing. DNA is isolated from tissue and a DNA sequencing library is prepared. Short biotinylated baits complementary to the targets hybridize to the DNA, bind to magnetic beads, and are captured by a magnet, yielding a target-enriched sequencing library. Exome sequencing can target the whole exome or a particular gene family such as NLRs as is done in the RenSeq method. **c** Chromosome flow sorting. Liquid suspensions of mitotic chromosomes collected from dividing root cells are fluorescently labeled and separated using flow cytometry based on the fluorochrome signal and relative DNA content

from a 3.8 Gb low-copy number genome assembly of CHINESE SPRING (Brenchley et al. 2012). To identify genic regions, they aligned reported wheat cDNA and EST sequences and conducted a BLASTn search using *Brachypodium* exon sequences. Krasileva et al. (2017) designed *T. turgidum* and *T. aestivum* exome capture probes to target gene annotations from the CSS assembly, transcripts from transcriptome studies, and unannotated homologs of barley in wheat. The exome capture probes targeted 85 Mb. Following the publication of high-quality reference wheat genome assemblies and annotations in 2017 and 2018, Gardiner et al. (2019) discovered that the existing exome capture assay only targeted 32.6% of the high-confidence gene set of wheat. Using the high-confidence annotated genes in the CHINESE SPRING-TGACv1 and RefSeq.v1 genome assemblies, *Ae. tauschii* assembly Aet v4.0, and

the *T. turgidum* ssp. *dicoccoides* WEWSeq v1.0 assembly, they designed exome capture probe sets targeting genes and putative promoters. Probes of ~75 bp were designed approximately every 120 bp across 786 Mb of design space, of which 509 Mb was gene space, and 277 Mb was putative promoter sequences. The exome capture and promoter capture probe sets designed by Jordan et al. (2015), Krasileva et al. (2017), and Gardiner et al. (2019) were available through NimbleGen (Roche) but have since been discontinued. The most recent exome capture assay, the myBaits® Expert Wheat Exome capture, designed using the CHINESE SPRING-RefSeq v1.0 assembly, captures over 250 Mb of coding sequence (Daicel Arbor Biosciences).

To further reduce genome complexity, capture probes assays can be developed to target a particular gene class such as NLRs. NLRs are the most common class of cloned R/S gene in wheat (Table 10.1), and the wheat pangenome is estimated to contain 6–8 thousand NLR genes (Walkowiak et al. 2020). Exome capture of NLR genes and subsequent sequencing is termed Resistance gene enrichment Sequencing (RenSeq). The first R genes cloned using RenSeq were *Rpi-ber2* and *Rpi-rzc1*, which confer resistance against *Phytophthora infestans* infections in potato (Jupe et al. 2013). Since then, RenSeq has been incorporated into rapid cloning methods AgRenSeq and MutChromSeq (discussed below). RenSeq was also recently combined with BSA in a method termed BSR-Seq (Lin et al. 2022). RenSeq was applied to DNA pools of resistant and susceptible plants allowing the identification of SNPs in NLRs linked to resistance. RenSeq is a key method in multiple rapid cloning strategies, efficiently enriching NLR genes. Kale et al. (2022) found the Triticeae RenSeq Baits V3 probe set (Zhang et al. 2021a) resulted in target enrichment of 220-fold of 18 Mb of NLR genes annotated in CHINESE SPRING-RefSeq v1.0. However, because probes were designed to target previously annotated NLR genes, RenSeq captures are biased and may not capture unannotated NLRs, i.e., NLRs not present in the sequences and annotated genome assemblies. RenSeq also relies on the assumption that the target R/S gene is a member of the NLR class. If it is suspected that the target gene might belong to a different class, then other methods should probably be considered.

10.4.2 Transcriptome Sequencing

Transcriptome sequencing, or RNA-Seq, is a less biased RRS method as it is not limited to previously annotated genes and/or a gene family. RNA-Seq combined with BSA (BSR-Seq) was applied to two *Ae. tauschii* populations to map leaf rust resistance gene *Lr42*, yielding just three candidate genes (Lin et al. 2022). RNA-Seq is limited to detecting genes that are expressed at the time of RNA collection in sufficient levels, and assembly of transcripts can be challenged by the co-expression of homoeologs. Lin et al. (2022) avoided the latter challenge by conducting RNA-Seq on a diploid.

10.4.3 Chromosome Flow Sorting

Chromosome flow sorting separates an individual chromosome via flow cytometry based on the chromosome size and base-pair composition (Doležel et al. 2011). Following separation, the individual chromosome can be sequenced and assembled as was done to complete the CSS assembly (IWGSC et al. 2014). Chromosome flow sorting is a highly specialized skill requiring unique equipment available in few labs. Also, not all chromosomes are able to be sorted from all others at sufficient efficiency to obtain a sample with adequate purity, and the time and labor needed to develop cytogenetic stocks such as the ditelosomics developed by Sears and Sears (1978) in CHINESE SPRING preclude that from being a viable option. Therefore, it is important to first determine whether a target chromosome can be efficiently sorted using flow cytometry before embarking on a project that relies on it to be successful.

10.5 Rapid Cloning Methods

10.5.1 MutRenSeq

RenSeq is coupled with mutational genomics in the MutRenSeq rapid cloning strategy (Steuernagel et al. 2016). In the MutRenSeq method, a mutant population is screened to identify the expected mutant phenotype and then RenSeq is conducted on confirmed mutants (Fig. 10.4). Independent mutation events within a single NLR associated with the mutant phenotype reveal the candidate gene(s). *Sr22* and *Sr45* were the first wheat R/S genes cloned in wheat using MutRenSeq. *Sr22*, which provides stem rust resistance, resides in introgressions from *T. boeoticum* and *T. monococcum* that had poor agronomic performance due to linkage drag (Olson et al. 2010). Additionally, mapping efforts were hampered by reduced recombination in the *Sr22* region (Steuernagel et al. 2016). To clone stem rust resistance genes *Sr22* and *Sr45*, Steuernagel et al. (2016) developed EMS-mutant populations for each R gene and applied RenSeq to six mutants/population and the wild type. In each mutant population, comparative sequence analysis of the NLRs in the mutants and wild type revealed one gene with mutations in all six mutants. MutRenSeq effectively eliminated the need for high-resolution mapping, which is particularly difficult when the R/S gene of interest resides in a low recombination region. MutRenSeq has since been used to clone stem rust resistance genes *Sr26*, *Sr27*, and *Sr61*, stripe rust resistance genes *Yr5* and *Yr7*, leaf rust resistance gene *Lr13/Ne2*, and powdery mildew resistance gene *Pm21* (Xing et al. 2018; Marchal et al. 2018; Zhang et al. 2021a; Hewitt et al. 2021a, b; Yan et al. 2021; Upadhyaya et al. 2021).

MutRenSeq is a powerful tool to quickly clone NLR resistance genes and is particularly advantageous when trying to clone a gene in an area of low recombination. However, it is limited to genotypes that can be easily mutagenized and R genes in the NLR family. In general, higher ploidy levels tend to tolerate higher EMS levels. The lower tolerance of mutagen dose results in lower mutation density, increasing the number of mutants that must be generated and phenotypically evaluated to identify independent lines with mutant alleles. In some cases, mutagenesis of diploids can result in sterile plants making the MutRenSeq method a less attractive option.

10.5.2 AgRenSeq

To address the limitations of MutRenSeq, Association Genetics RenSeq (AgRenSeq) was developed (Arora et al. 2019) by combining association genetics and RenSeq. A diversity panel is phenotyped for disease reactions and RenSeq is conducted on the panel. *K*-mers within the sequenced NLR are identified and mapped to a reference assembly. Associations between *k*-mers and phenotypes are then calculated and plotted, similar to a Manhattan plot. Significant *k*-mers map to contigs that represent candidate genes. To test AgRenSeq, a panel of 174 *Ae. tauschii* ssp. *strangulata* accessions was genotyped and evaluated for reaction to six races of wheat stem rust pathogen *Puccinia graminis* f. sp. *tritici* (PGT). Two previously cloned genes, *Sr33* and *Sr45*, served as positive controls (Periyannan et al. 2013; Steuernagel et al. 2016). *K*-mers associated with resistance to PGT race RKQQC, which is avirulent to *Sr33*, resided on the contig containing the previously cloned *Sr33*. *Sr45*, which was previously identified using MutRenSeq (Steuernagel et al. 2016), was also identified via AgRenSeq. Candidate genes for *Sr46* and *SrTA1662* were also identified in this study, and the *Sr46* candidate was functionally validated by mutagenesis and gene complementation. Thus, Arora et al. (2019) demonstrated the ability of AgRenSeq to directly identify candidate genes. However, as with other RenSeq-based cloning methods, AgRenSeq is limited to cloning NLR genes.

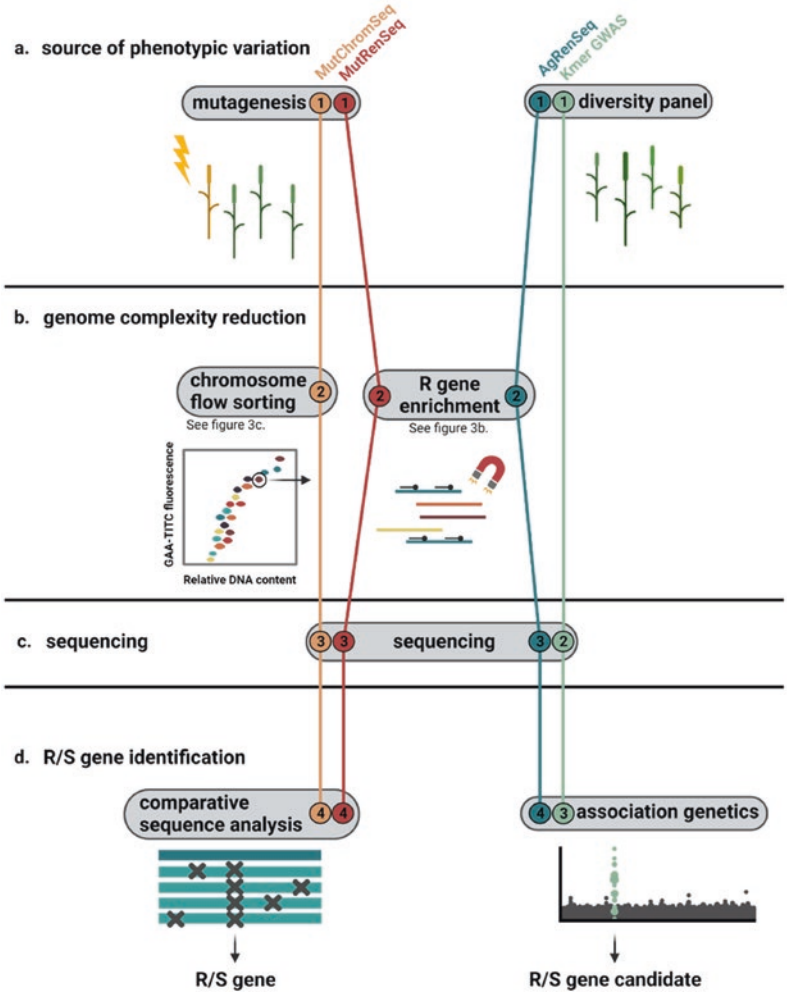


Fig. 10.4 Overview of R/S gene rapid cloning methods in wheat. The paths of MutChromSeq (orange), MutRenSeq (red), AgRenSeq (teal), and *k*-mer GWAS (seafoam) are shown with stops at particular methods numbered and connected with solid lines. **a** Source of phenotypic variation. Rapid cloning methods use one of two forms of phenotypic variation, induced phenotypic variation through mutagenesis (left) or natural variation in a diversity panel (right). **b** Genome complexity reduction. After phenotyping, the next step is genome complexity reduction through either chromosome flow sorting or R gene enrichment through gene family capture. Note, the *k*-mer GWAS path moves directly to sequencing. **c** Sequencing. Next, the flow sorted chromosome, captured genes, or diversity panel is/are sequenced. Depending on the target, personal preferences, and resources available, different sequencing methods may be used. **d** R/S gene identification. The final step involves identifying candidate genes. Left, candidate genes are identified through comparison of mutant (light teal) and wild-type sequences to identify regions with mutation overlap. Right, associations between particular NLRs or *k*-mers are identified with the highest associations being candidate genes or near candidate genes. Association genetics yields candidate genes that require functional validation while methods using induced variation through mutagenesis already include a functional validation step

10.5.3 K-mer GWAS

K-mer-based association mapping, or *k*-mer GWAS, is an extension of AgRenSeq, but it excludes the RenSeq step and is therefore not limited to the detection of only NLRs. Instead, *k*-mers are identified from whole-genome shotgun sequencing reads and projected onto a reference assembly. The analysis is similar to AgRenSeq, but because *k*-mers can be anywhere and not just within candidate genes, one must analyze the genes near the *k*-mers that were in linkage disequilibrium with the phenotype. Gaurav et al. (2022) conducted whole-genome shotgun sequencing on 242 *Ae. tauschii* accessions and used *k*-mer GWAS to identify a 50-kb linkage disequilibrium block containing two candidate genes for the stem rust resistance gene *SrTA1662*. Subsequent functional validation via gene complementation confirmed that *SrTA1662* is an NLR. The panel sequenced in Gaurav et al. (2022) is publicly available and can be used to rapidly clone R/S genes from *Ae. tauschii* accessions.

In 2020, Voichek and Wiegel published a reference-free *k*-mer GWAS method. In this method, the associations between *k*-mers and the phenotype were calculated prior to mapping the *k*-mers to a reference, allowing the identification of *k*-mers significantly associated with the trait, including those absent in a reference. In a case study in Arabidopsis, the authors identified *k*-mers significantly associated with two traits—growth in the presence of a flg22 variant and germination in darkness under low nutrient supply—neither of which mapped to their reference genome. Assembly and subsequent analysis of the short reads used to identify the significant *k*-mers revealed alternate structural variants of genes associated with the two traits. Although reference-free *k*-mer GWAS has not yet been used to clone R/S genes in wheat, it has been applied to map resistance to yellow rust and leaf rust (Kale et al. 2022). R/S genes display abundant presence/absence and copy number variation (Van de Weyer et al. 2019; Walkowiak et al. 2020), so the potential to detect structural variants not in a reference assembly via reference-free *k*-mer GWAS is appealing.

Both AgRenSeq and *k*-mer-GWAS require shot gun sequencing of an entire diversity panel, which can initially be expensive and laborious. However, once this has been completed, the same panel can be used to clone multiple R/S genes. Additionally, AgRenSeq and *k*-mer GWAS can be limited by the population structure of the diversity panel (Yu et al. 2006) and choice of the reference sequence can influence which associations are detected (Voichek and Weigel 2020; Kale et al. 2022).

10.5.4 MutChromSeq

In 2016, Sánchez-Martín et al. published the rapid cloning method MutChromSeq and used it to clone the powdery mildew resistance gene *Pm2a*, which had previously mapped to chromosome 6A (Huang and Röder 2004). Using the MutChromSeq method, which applied the RRS method chromosome flow sorting, chromosome 6A was sorted from six confirmed EMS-derived powdery mildew susceptible mutants and wild-type genotypes. The separated chromosomes were sequenced and assembled followed by sequence analysis to identify mutation overlap. Contigs with mutations in all or most of the mutant lines are most likely to contain the candidate gene. Two contigs were identified, although one was later discarded due to an abnormal SNV frequency, leaving just one contig with a NLR gene. MutChromSeq is similar to MutRenSeq, but it is not limited to NLR genes. MutChromSeq was also used to clone leaf rust resistance gene *Lr14a* with ankyrin transmembrane protein domains and *Pm4b*, which contains kinase, C2, and transmembrane domains (Kolodziej et al. 2021; Sánchez-Martín et al. 2021).

10.6 Validating Candidate Genes

Validating candidate genes is a critical step in proving a gene confers a particular phenotype. Forward and reverse genetics approaches

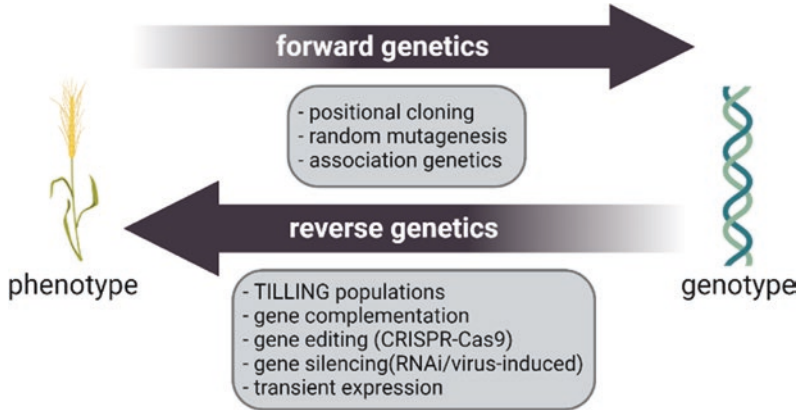


Fig. 10.5 Commonly used forward and reverse genetics methods to identify and/or validate R/S genes in wheat. Arrows indicate the direction of the genetic approaches with forward genetics approaches starting with a known phenotype and identifying the gene underlying the phenotype, while reverse genetics starts with a known gene sequence and identifies the phenotypic effects of genic or transcriptomic alternations. Common methods used to identify and/or validate R/S genes in wheat are listed under their approach type

can be used to identify and validate candidate genes. Forward genetics approaches start from a phenotype and identify the gene that confers the phenotype (Fig. 10.5). Many of the rapid cloning methods are considered forward genetics approaches as they start with variation in a phenotype, either natural or induced through mutagenesis. However, not all forward genetics approaches serve as functional validation methods. For example, map-based cloning and association genetics approaches often yield multiple candidate genes and must be followed up with functional validation to determine which candidate gene is the gene of interest. Because rapid cloning methods MutRenSeq and MutChromSeq use mutagenized populations, these methods both identify and validate candidate genes.

Reverse genetics approaches start with the gene sequence and identify the phenotypic effects of particular gene states (Fig. 10.5). Functional validation methods that use a reverse genetics approach include methods like RNA interference, gene complementation, or CRISPR/CAS9 gene editing, which alter the genetic or transcriptomic makeup of an individual to identify the phenotypic effect of the alteration. The Targeting Induced Local Lesions in Genomes (TILLING) resource can also be used to functionally validate genes in a reverse genetics manner. Krasileva et al. (2017) sequenced the exomes of 1200 CADENZA and 1535 KRONOS EMS mutants and characterized and cataloged the mutations relative to the CSS assembly. When the CHINESE SPRING-RefSeqv1.0 assembly was published, the TILLING raw reads were realigned to the new assembly. The TILLING resources expedite functional validation of genes as researchers do not need to create the genetic or transcriptomic alteration. Instead, mutant lines with known alterations in candidate genes can be selected on Ensembl Plants and ordered from SeedStor (https://www.seedstor.ac.uk/). However, the TILLING resource is limited to functionally validating genes present in CADENZA or KRONOS and annotated in the CHINESE SPRING-RefSeqv1.1 gene models.

Often both forward and reverse genetics approaches are applied to functionally validate R/S genes. The two most commonly used functional validation methods are mutagenesis and gene complementation, both of which have been used to validate around two-thirds of the cloned R/S genes. About 43% of the cloned R/S genes have been validated using both mutagenesis and gene complementation. Gene silencing, transient expression, gene editing, and the TILLING populations are less frequently used methods of

functional validation, with each being used to validate 15 or fewer R/S genes.

Clustered regularly interspaced short palindromic repeats (CRISPR) and its associated protein (Cas) can be used to produce site-specific double-stranded breaks, often resulting in gene knockout. Wang et al. (2014) used CRISPR-Cas9 and transcription activator-like effector nuclease (TALEN) technologies to knock out three homoeoalleles of the powdery mildew susceptibility gene *Mlo* in the cultivar Bobwhite, resulting in reduced susceptibility to *Blumeria graminis* f. sp. *tritici.* While CRISPR-Cas mediated gene knockout is a highly specific and targeted functional validation method, unlike random mutagenesis, it is somewhat limited to functionally validating genes present in easily transformable cultivars such as FIELDER or BOBWHITE. However, advancements in gene editing and transformation methods are expanding the definition of "transformable cultivars."

10.7 Conclusions and Future Outlook

The expansion of wheat genomic resources, genomic complexity reduction methods coupled with advanced sequencing technologies, and rapid cloning methods has enabled the accelerated cloning of R/S genes in wheat. In 2020 and 2021, sixteen cloned R/S genes were published, a feat that in the earlier years of R/S gene cloning took thirteen years to accomplish; it was not unheard of for cloning an R/S gene to take 10 years. Now, cloning an R/S gene is possible in less than a year. Undoubtedly, R/S gene cloning will continue to accelerate as more reference genomes are published, sequencing costs decrease, and cloning methods advance. The multiple sequenced wheat genomes that are currently available are a tremendous resource and make it relatively easy to assess gene content in a given R/S gene candidate gene region. However, given the common presence/absence and copy number variation displayed by R/S genes (Van de Weyer et al. 2019; Walkowiak et al. 2020), it is still possible for a gene of interest to be absent in all the wheat genomes currently available. We have not yet reached a true wheat pangenome, but costs for sequencing and assembly of entire wheat genomes continue to decline, and the data can be obtained in a matter of months. Therefore, it is now becoming more feasible to sequence and assemble the entire genome of a wheat line with the primary goal of cloning a single R/S gene, a feat that was nearly unthinkable when wheat genomics researchers met in 2003 to carve a path forward to sequence the first wheat genome (Gill et al. 2004).

Wheat's wild relatives offer a greater pool of R genes, as they have not undergone the genetic bottleneck characteristic of domestication. Association genetics methods, like *k*-mer GWAS and AgRenSeq, address some of the limitations of traditional map-based cloning, exploiting greater genetic diversity and ancestral recombination to identify unique disease resistance loci. Additionally, these diversity panels often allow the isolation of more than one R gene as they segregate for resistance to multiple isolates and/or races of multiple pathogens, whereas biparental mapping populations are often designed to segregate for only one R/S locus for ease of genetic mapping. The advances in sequencing technologies, cloning methods, and gene editing technologies will likely soon reshape the way R genes from wild relatives are deployed in adapted germplasm. Historically, chromosome engineering strategies involving cytogenetic methods to achieve chromosome substitutions, translocations, and ultimately introgressions of smaller segments containing target genes, were extremely laborious and time-consuming, and the end product usually suffered from deleterious linkage drag. The modern sequencing and cloning technologies discussed in this chapter may make it more feasible to clone the target gene in the wild relative accession itself. Although genetic transformation (GMO wheat) is currently not accepted, the acceptance of gene editing appears more promising. Thus, once a target R gene is cloned from a wild relative, it is conceivable that a homologous gene could be identified in wheat and edited to acquire the desired function.

With the availability of multiple reference wheat genomes, in some cases, the bottleneck of cloning R/S genes has shifted from candidate gene identification to functional validation. The use of the CADENZA- and KRONOS-TILLING populations offers rapid functional validation. However, a single bread wheat and durum wheat cultivar cannot feasibly represent the R/S gene content of all bread and durum wheat. Still, due to ease of use and affordability, the TILLING populations are an excellent resource worth considering.

Cloning and deploying R genes and removing S genes is a constant highly coordinated race to keep up with evolving pathogen populations. We suspect that as more R/S genes are cloned, more research will focus on identifying unique durable combinations of R/S genes. For example, Luo et al. (2021) transformed a five-gene cassette of stem rust resistance genes into bread wheat cultivar FIELDER, resulting in broad-spectrum resistance. Another benefit of cloning multiple R/S genes in a given system is that the cumulative knowledge acquired can begin to shed light on the essential components, which can lead to the development of designer genes that could operate to govern broad-spectrum resistance and perhaps resistance less prone to being overcome due to natural mutations occurring in the pathogen. The use of R gene cassettes, disruption of S genes, or the development and deployment of designer genes made possible through advancements in tissue culture, transformation methods, and gene editing technologies are promising directions to ensure stable wheat production enhancing global food security.

Acknowledgements Figures were created with Biorender.com.

References

Acevedo-Garcia J, Spencer D, Thieron H, Reinstädler A, Hammond-Kosack K, Phillips AL, Panstruga R (2017) mlo-based powdery mildew resistance in hexaploid bread wheat generated by a non-transgenic TILLING approach. Plant Biotechnol J 15:367–378

Akpinar BA, Leroy P, Watson-Haigh NS et al (2022) The complete genome sequence of elite bread wheat cultivar, "Sonmez". F1000Res 11:614. https://doi.org/10.12688/f1000research.121637.1

Arora S, Steuernagel B, Gaurav K et al (2019) Resistance gene cloning from a wild crop relative by sequence capture and association genetics. Nat Biotechnol 37:139–143. https://doi.org/10.1038/s41587-018-0007-9

Athiyannan N, Abrouk M, Boshoff WHP et al (2022) Long-read genome sequencing of bread wheat facilitates disease resistance gene cloning. Nat Genet 54:227–231. https://doi.org/10.1038/s41588-022-01022-1

Aury J-M, Engelen S, Istace B et al (2022) Long-read and chromosome-scale assembly of the hexaploid wheat genome achieves high resolution for research and breeding. GigaScience 11:giac034. https://doi.org/10.1093/gigascience/giac034

Avni R, Nave M, Barad O et al (2017) Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. Science 357:93–97. https://doi.org/10.1126/science.aan0032

Avni R, Lux T, Minz-Dub A et al (2022) Genome sequences of three *Aegilops* species of the section *Sitopsis* reveal phylogenetic relationships and provide resources for wheat improvement. Plant J 110:179–192. https://doi.org/10.1111/tpj.15664

Bennett MD, Smith J (1976) Nuclear DNA amounts in angiosperms. Phil Trans R Soc B 274:227–274

Brenchley R, Spannagl M, Pfeifer M et al (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. Nature 491:705–710. https://doi.org/10.1038/nature11650

Cao A, Xing L, Wang X et al (2011) Serine/threonine kinase gene *Stpk-V*, a key member of powdery mildew resistance gene *Pm21*, confers powdery mildew resistance in wheat. Proc Natl Acad Sci USA 108:7727–7732. https://doi.org/10.1073/pnas.1016981108

Chapman JA, Mascher M, Buluç A et al (2015) A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome. Genome Biol 16:26. https://doi.org/10.1186/s13059-015-0582-8

Chen S, Zhang W, Bolus S et al (2018) Identification and characterization of wheat stem rust resistance gene *Sr21* effective against the Ug99 race group at high temperature. PLoS Genet 14:e1007287. https://doi.org/10.1371/journal.pgen.1007287

Chen S, Rouse MN, Zhang W et al (2020) Wheat gene *Sr60* encodes a protein with two putative kinase domains that confers resistance to stem rust. New Phytol 225:948–959. https://doi.org/10.1111/nph.16169

Choulet F, Alberti A, Theil S et al (2014) Structural and functional partitioning of bread wheat chromosome 3B. Science 345:1249721. https://doi.org/10.1126/science.1249721

Clavijo BJ, Venturini L, Schudoma C et al (2017) An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. Genome Res 27:885–896. https://doi.org/10.1101/gr.217117.116

Cloutier S, McCallum BD, Loutre C et al (2007) Leaf rust resistance gene *Lr1*, isolated from bread wheat (*Triticum aestivum L.*) is a member of the large *psr567* gene family. Plant Mol Biol 65:93–106. https://doi.org/10.1007/s11103-007-9201-8

Doležel J, Kubaláková M, íhalíková J et al (2011) Chromosome analysis and sorting using flow cytometry. In: Birchler JA (ed) Plant chromosome engineering. Humana Press, Totowa, NJ, pp 221–238

Elliott C, Zhou F, Spielmeyer W, Panstruga R, Schulze-Lefert P (2002) Functional conservation of wheat and rice Mlo orthologs in defense modulation to the powdery mildew fungus. Mol Plant Microbe Interact 15:1069–1077

Endo TR, Gill BS (1996) The deletion stocks of common wheat. J Hered 87:295–307

Erenstein O, Jaleta M, Mottaleb KA et al (2022) Global trends in wheat production, consumption and trade. In: Reynolds MP, BraunHJ (eds) Wheat Improvement. Springer, Cham. https://doi.org/10.1007/978-3-030-90673-3_4

Faris JD (2014) Wheat domestication: key to agricultural revolutions past and future. In: Tuberosa R, Graner A, Frison E (eds) Genomics of plant genetic resources. Springer, Netherlands, Dordrecht, pp 439–464

Faris JD, Zhang Z, Lu H et al (2010) A unique wheat disease resistance-like gene governs effector-triggered susceptibility to necrotrophic pathogens. Proc Natl Acad Sci USA 107:13544–13549. https://doi.org/10.1073/pnas.1004090107

Feuillet C, Travella S, Stein N et al (2003) Map-based isolation of the leaf rust disease resistance gene *Lr10* from the hexaploid wheat (*Triticum aestivum* L.) genome. Proc Natl Acad Sci USA 100:15253–15258. https://doi.org/10.1073/pnas.2435133100

Fu D, Uauy C, Distelfeld A et al (2009) A kinase-START gene confers temperature-dependent resistance to wheat stripe rust. Science 323:1357–1360. https://doi.org/10.1126/science.1166289

Gardiner L-J, Brabbs T, Akhunov A et al (2019) Integrating genomic resources to present full gene and putative promoter capture probe sets for bread wheat. GigaScience 8:1–13. https://doi.org/10.1093/gigascience/giz018

Gaurav K, Arora S, Silva P et al (2022) Population genomic analysis of *Aegilops tauschii* identifies targets for bread wheat improvement. Nat Biotechnol 40:422–431. https://doi.org/10.1038/s41587-021-01058-4

Gill BS, Appels R, Botha-Oberholster A-M et al (2004) A workshop report on wheat genome sequencing. Genetics 168:1087–1096. https://doi.org/10.1534/genetics.104.034769

Guo W, Xin M, Wang Z et al (2020) Origin and adaptation to high altitude of Tibetan semi-wild wheat. Nat Commun 11:5085. https://doi.org/10.1038/s41467-020-18738-5

Hafeez AN, Arora S, Ghosh S et al (2021) Creation and judicious application of a wheat resistance gene atlas. Mol Plant 14:1053–1070. https://doi.org/10.1016/j.molp.2021.05.014

He H, Zhu S, Ji Y et al (2017) Map-based cloning of the gene *Pm21* that confers broad spectrum resistance to wheat powdery mildew. https://doi.org/10.1101/177857

Hewitt T, Müller MC, Molnár I et al (2021a) A highly differentiated region of wheat chromosome 7AL encodes a *Pm1a* immune receptor that recognizes its corresponding AvrPm1a effector from *Blumeria graminis*. New Phytol 229:2812–2826. https://doi.org/10.1111/nph.17075

Hewitt T, Zhang J, Huang L et al (2021b) Wheat leaf rust resistance gene *Lr13* is a specific *Ne2* allele for hybrid necrosis. Mol Plant 14:1025–1028. https://doi.org/10.1016/j.molp.2021.05.010

Huai B, Yuan P, Ma X et al (2022) Sugar transporter *TaSTP3* activation by *TaWRKY19/61/82* enhances stripe rust susceptibility in wheat. New Phytol. https://doi.org/10.1111/nph.18331

Huang X-Q, Röder MS (2004) Molecular mapping of powdery mildew resistance genes in wheat: a review. Euphytica 137:203–223. https://doi.org/10.1023/B:EUPH.0000041576.74566.d7

Huang L, Brooks SA, Li W et al (2003) Map-based cloning of leaf rust resistance gene *Lr21* from the large and polyploid genome of bread wheat. Genetics 164:655–664. https://doi.org/10.1093/genetics/164.2.655

Hurni S, Brunner S, Buchmann G et al (2013) Rye *Pm8* and wheat *Pm3* are orthologous genes and show evolutionary conservation of resistance function against powdery mildew. Plant J 76:957–969. https://doi.org/10.1111/tpj.12345

Ingvardsen CR, Massange-Sánchez JA, Borum F et al (2019) Development of mlo-based resistance in tetraploid wheat against wheat powdery mildew. Theor Appl Genet 132:3009–3022. https://doi.org/10.1007/s00122-019-03402-4

International Wheat Genome Sequencing Consortium (IWGSC) (2014) A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. Science 355:1251788. https://doi.org/10.1126/science.1251788

Jia J, Zhao S, Kong X et al (2013) *Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation. Nature 496:91–95. https://doi.org/10.1038/nature12028

Jordan KW, Wang S, Lun Y et al (2015) A haplotype map of allohexaploid wheat reveals distinct patterns of selection on homoeologous genomes. Genome Biol 16:48. https://doi.org/10.1186/s13059-015-0606-4

Jupe F, Witek K, Verweij W et al (2013) Resistance gene enrichment sequencing (RenSeq) enables reannotation of the NB-LRR gene family from sequenced plant genomes and rapid mapping of resistance loci in segregating populations. Plant J 76:530–544. https://doi.org/10.1111/tpj.12307

Kale SM, Schulthess AW, Padmarasu S et al (2022) A catalogue of resistance gene homologs and a chromosome-scale reference sequence support resistance gene mapping in winter wheat. Plant Biotechnol J pbi.13843. https://doi.org/10.1111/pbi.13843

Kan J, Cai Y, Cheng C et al (2022) Simultaneous editing of host factor gene *TaPDIL5-1* homoeoalleles confers wheat yellow mosaic virus resistance in hexaploid wheat. New Phytol 234:340–344. https://doi.org/10.1111/nph.18002

Klymiuk V, Yaniv E, Huang L et al (2018) Cloning of the wheat *Yr15* resistance gene sheds light on the plant tandem kinase-pseudokinase family. Nat Commun 9:3735. https://doi.org/10.1038/s41467-018-06138-9

Kolodziej MC, Singla J, Sánchez-Martín J et al (2021) A membrane-bound ankyrin repeat protein confers race-specific leaf rust disease resistance in wheat. Nat Commun 12:956. https://doi.org/10.1038/s41467-020-20777-x

Krasileva KV, Vasquez-Gross HA, Howell T et al (2017) Uncovering hidden variation in polyploid wheat. Proc Natl Acad Sci USA 114. https://doi.org/10.1073/pnas.1619268114

Krattinger SG, Lagudah ES, Spielmeyer W et al (2009) A putative ABC transporter confers durable resistance to multiple fungal pathogens in wheat. Science 323:1360–1363. https://doi.org/10.1126/science.1166453

Li G, Zhou J, Jia H et al (2019) Mutation of a histidine-rich calcium-binding-protein gene in wheat confers resistance to Fusarium head blight. Nat Genet 51:1106–1112. https://doi.org/10.1038/s41588-019-0426-7

Li M, Dong L, Li B et al (2020) A CNL protein in wild emmer wheat confers powdery mildew resistance. New Phytol 228:1027–1037. https://doi.org/10.1111/nph.16761

Li L-F, Zhang Z-B, Wang Z-H et al (2022) Genome sequences of five *Sitopsis* species of *Aegilops* and the origin of polyploid wheat B subgenome. Mol Plant 15:488–503. https://doi.org/10.1016/j.molp.2021.12.019

Lin G, Chen H, Tian B et al (2022) Cloning of the broadly effective wheat leaf rust resistance gene *Lr42* transferred from *Aegilops tauschii*. Nat Commun 13:3044. https://doi.org/10.1038/s41467-022-30784-9

Ling H-Q, Zhao S, Liu D et al (2013) Draft genome of the wheat A-genome progenitor *Triticum urartu*. Nature 496:87–90. https://doi.org/10.1038/nature11997

Ling H-Q, Ma B, Shi X et al (2018) Genome sequence of the progenitor of wheat A subgenome *Triticum urartu*. Nature 557:424–428. https://doi.org/10.1038/s41586-018-0108-0

Liu W, Frick M, Huel R et al (2014) The stripe rust resistance gene *Yr10* encodes an evolutionary-conserved and unique CC–NBS–LRR Sequence in wheat. Mol Plant 7:1740–1755. https://doi.org/10.1093/mp/ssu112

Lu P, Guo L, Wang Z et al (2020) A rare gain of function mutation in a wheat tandem kinase confers resistance to powdery mildew. Nat Commun 11:680. https://doi.org/10.1038/s41467-020-14294-0

Luo M-C, Gu YQ, Puiu D et al (2017) Genome sequence of the progenitor of the wheat D genome *Aegilops tauschii*. Nature 551:498–502. https://doi.org/10.1038/nature24486

Luo M, Xie L, Chakraborty S et al (2021) A five-transgene cassette confers broad-spectrum resistance to a fungal rust pathogen in wheat. Nat Biotechnol 39:561–566. https://doi.org/10.1038/s41587-020-00770-x

Maccaferri M, Harris NS, Twardziok SO et al (2019) Durum wheat genome highlights past domestication signatures and future improvement targets. Nat Genet 51:885–895. https://doi.org/10.1038/s41588-019-0381-3

Mago R, Zhang P, Vautrin S et al (2015) The wheat *Sr50* gene reveals rich diversity at a cereal disease resistance locus. Nat Plants 1:15186. https://doi.org/10.1038/nplants.2015.186

Marchal C, Zhang J, Zhang P et al (2018) BED-domain-containing immune receptors confer diverse resistance spectra to yellow rust. Nat Plants 4:662–668. https://doi.org/10.1038/s41477-018-0236-4

Mello-Sampayo T, Lorente R (1968) The role of chromosome 3D in the regulation of meiotic pairing in hexaploid wheat. EWAC Newslett 2:16–24

Moore JW, Herrera-Foessel S, Lan C et al (2015) A recently evolved hexose transporter variant confers resistance to multiple pathogens in wheat. Nat Genet 47:1494–1498. https://doi.org/10.1038/ng.3439

Olson EL, Brown-Guedira G, Marshall D et al (2010) Development of wheat lines having a small introgressed segment carrying stem rust resistance gene *Sr22*. Crop Sci 50:1823–1830. https://doi.org/10.2135/cropsci2009.11.0652

Periyannan S, Moore J, Ayliffe M et al (2013) The gene *Sr33,* an ortholog of barley *Mla* genes, encodes resistance to wheat stem rust race Ug99. Science 341:786–788. https://doi.org/10.1126/science.1239028

Poddar S, Tanaka J, Running KLD, Kariyawasam GK, Faris JD, Friesen TL, Cho M-J, Cate JHD, Staskawicz B (2022) bioRxiv. https://doi.org/10.1101/2022.04.05.487229

Poddar S, Tanaka J, Running KLD et al (2023) Optimization of highly efficient exogenous-DNA-free Cas9-ribonucleoprotein mediated gene editing in disease susceptibility loci in wheat (Triticum aestivum L.) Front Plant Sci 13:1084700. https://doi.org/10.3389/fpls.2022.1084700

Qi LL, Echalier B, Chao S et al (2004) A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat. Genetics 168:701–712. https://doi.org/10.1534/genetics.104.034868

Rawat N, Pumphrey MO, Liu S et al (2016) Wheat *Fhb1* encodes a chimeric lectin with agglutinin domains and a pore-forming toxin-like domain conferring resistance to Fusarium head blight. Nat Genet 48:1576–1580. https://doi.org/10.1038/ng.3706

Riley R, Chapman V (1958) Genetic control of the cytologically diploid behaviour of hexploid wheat. Nature 182:713–715

Saintenac C, Zhang W, Salcedo A et al (2013) Identification of wheat gene *Sr35* that confers resistance to Ug99 stem rust race group. Science 341:783–786. https://doi.org/10.1126/science.1239022

Saintenac C, Lee W-S, Cambon F et al (2018) Wheat receptor-kinase-like protein *Stb6* controls gene-for-gene resistance to fungal pathogen *Zymoseptoria tritici*. Nat Genet 50:368–374. https://doi.org/10.1038/s41588-018-0051-x

Saintenac C, Cambon F, Aouini L et al (2021) A wheat cysteine-rich receptor-like kinase confers broad-spectrum resistance against Septoria tritici blotch. Nat Commun 12:433. https://doi.org/10.1038/s41467-020-20685-0

Sánchez-Martín J, Steuernagel B, Ghosh S et al (2016) Rapid gene isolation in barley and wheat by mutant chromosome sequencing. Genome Biol 17:221. https://doi.org/10.1186/s13059-016-1082-1

Sánchez-Martín J, Widrig V, Herren G et al (2021) Wheat *Pm4* resistance to powdery mildew is controlled by alternative splice variants encoding chimeric proteins. Nat Plants 7:327–341. https://doi.org/10.1038/s41477-021-00869-2

Sato K, Abe F, Mascher M et al (2021) Chromosome-scale genome assembly of the transformation-amenable common wheat cultivar 'Fielder.' DNA Res 28:dsab008. https://doi.org/10.1093/dnares/dsab008

Savary S, Bregaglio S, Willocquet L et al (2017) Crop health and its global impacts on the components of food security. Food Sec 9:311–327. https://doi.org/10.1007/s12571-017-0659-1

Savary S, Willocquet L, Pethybridge SJ et al (2019) The global burden of pathogens and pests on major food crops. Nat Ecol Evol 3:430–439. https://doi.org/10.1038/s41559-018-0793-y

Schreiber AW, Hayden MJ, Forrest KL et al (2012) Transcriptome-scale homoeolog-specific transcript assemblies of bread wheat. BMC Genomics 13:492. https://doi.org/10.1186/1471-2164-13-492

Sears ER (1954) The aneuploids of common wheat. Mo Agr Exp Sta Res Bull 572:1–59

Sears ER (1966) Nullisomic-tetrasomic combinations in hexaploid wheat. In: Riley R, Lewis KR (eds) Chromosome manipulation and plant genetics. Oliver & Boyd, Edinburgh, pp 29–45

Sears ER, Okamoto M (1958) Intergenomic chromosome relationships in hexaploid wheat. Proc Int Congr Genet 2:258–259

Sears ER, Sears LMS (1978) The telocentric chromosomes of common wheat In: Ramanujam S (ed) Proceedings of the 5th international wheat genetics symposium. Indian Society of Genetics and Plant Breeding, New Delhi, pp 389–407

Shi G, Zhang Z, Friesen TL et al (2016) The hijacking of a receptor kinase–driven pathway by a wheat fungal pathogen leads to disease. Sci Adv 2:e1600822. https://doi.org/10.1126/sciadv.1600822

Singh SP, Hurni S, Ruinelli M et al (2018) Evolutionary divergence of the rye *Pm17* and *Pm8* resistance genes reveals ancient diversity. Plant Mol Biol 98:249–260. https://doi.org/10.1007/s11103-018-0780-3

Srichumpa P, Brunner S, Keller B, Yahiaoui N (2005) Allelic series of four powdery mildew resistance genes at the *Pm3* locus in hexaploid bread wheat. Plant Physiol 139:885–895. https://doi.org/10.1104/pp.105.062406

Steuernagel B, Periyannan SK, Hernández-Pinzón I et al (2016) Rapid cloning of disease-resistance genes in plants using mutagenesis and sequence capture. Nat Biotechnol 34:652–655. https://doi.org/10.1038/nbt.3543

Su Z, Bernardo A, Tian B et al (2019) A deletion mutation in TaHRC confers Fhb1 resistance to Fusarium head blight in wheat. Nat Genet 51:1099–1105. https://doi.org/10.1038/s41588-019-0425-8

The International Wheat Genome Sequencing Consortium (IWGSC), Mayer KFX, Rogers J et al (2014) A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. Science 345:1251788. https://doi.org/10.1126/science.1251788

The International Wheat Genome Sequencing Consortium (IWGSC), Appels R, Eversole K et al. (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science 361:eaar7191. https://doi.org/10.1126/science.aar7191

Thind AK, Wicker T, Šimková H et al (2017) Rapid cloning of genes in hexaploid wheat using cultivar-specific long-range chromosome assembly. Nat Biotechnol 35:793–796. https://doi.org/10.1038/nbt.3877

Upadhyaya NM, Mago R, Panwar V et al (2021) Genomics accelerated isolation of a new stem rust avirulence gene–wheat resistance gene pair. Nat Plants 7:1220–1228. https://doi.org/10.1038/s41477-021-00971-5

Van de Weyer AL, Monteiro F, Furzer OJ, Nishimura MT, Cevik V, Witek K, Jones JDG, Dangl JL, Weigel D, Bemm F (2019) A species-wide inventory of NLR genes and alleles in *Arabidopsis thaliana*. Cell 178(5):1260-1272.e14. https://doi.org/10.1016/j.cell.2019.07.038

Várallyay É, Giczey G, Burgyán J (2012) Virus-induced gene silencing of Mlo genes induces powdery mildew resistance in *Triticum aestivum*. Arch Virol 157:1345–1350

Voichek Y, Weigel D (2020) Identifying genetic variants underlying phenotypic variation in plants without complete genomes. Nat Genet 52(5):534–540. https://doi.org/10.1038/s41588-020-0612-7

Walkowiak S, Gao L, Monat C et al (2020) Multiple wheat genomes reveal global variation in modern breeding. Nature 588:277–283. https://doi.org/10.1038/s41586-020-2961-x

Wang Y, Cheng X, Shan Q, Zhang Y, Liu J, Gao C, Qiu J-L (2014) Simultaneous editing of three homoeoalleles in hexaploid bread wheat confers heritable resistance to powdery mildew. Nat Biotechnol 32:947

Wang H, Sun S, Ge W et al (2020a) Horizontal gene transfer of *Fhb7* from fungus underlies *Fusarium* head blight resistance in wheat. Science 368:eaba5435. https://doi.org/10.1126/science.aba5435

Wang H, Zou S, Li Y et al (2020b) An ankyrin-repeat and WRKY-domain-containing immune receptor confers stripe rust resistance in wheat. Nat Commun 11:1353. https://doi.org/10.1038/s41467-020-15139-6

Wang L, Zhu T, Rodriguez JC et al (2021) *Aegilops tauschii* genome assembly Aet v5.0 features greater sequence contiguity and improved annotation. Genes Genom Genet 11:jkab325. https://doi.org/10.1093/g3journal/jkab325

Wicker T, Gundlach H, Spannagl M et al (2018) Impact of transposable elements on genome structure and evolution in bread wheat. Genome Biol 19:103. https://doi.org/10.1186/s13059-018-1479-0

Wulff BB, Krattinger SG (2022) The long road to engineering durable disease resistance in wheat. Curr Opin Biotechnol 73:270–275. https://doi.org/10.1016/j.copbio.2021.09.002

Xie J, Guo G, Wang Y et al (2020) A rare single nucleotide variant in *Pm5e* confers powdery mildew resistance in common wheat. New Phytol 228:1011–1026. https://doi.org/10.1111/nph.16762

Xing L, Hu P, Liu J et al (2018) *Pm21* from *Haynaldia villosa* encodes a CC-NBS-LRR protein conferring powdery mildew resistance in wheat. Mol Plant 11:874–878. https://doi.org/10.1016/j.molp.2018.02.013

Yahiaoui N, Srichumpa P, Dudler R, Keller B (2004) Genome analysis at different ploidy levels allows cloning of the powdery mildew resistance gene *Pm3b* from hexaploid wheat: Positional cloning of *Pm3* from hexaploid wheat. Plant J 37:528–538. https://doi.org/10.1046/j.1365-313X.2003.01977.x

Yan X, Li M, Zhang P et al (2021) High-temperature wheat leaf rust resistance gene *Lr13* exhibits pleiotropic effects on hybrid necrosis. Mol Plant 14:1029–1032. https://doi.org/10.1016/j.molp.2021.05.009

Yao E, Blake VC, Cooper L et al (2022) GrainGenes: a data-rich repository for small grains genetics and genomics. Database 2022:baac034. https://doi.org/10.1093/database/baac034

Yu J, Pressoir G, Briggs W et al (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203–208. https://doi.org/10.1038/ng1702

Yu G, Matny O, Champouret N et al (2022) *Aegilops sharonensis* genome-assisted identification of stem rust resistance gene *Sr62*. Nat Commun 13:1607. https://doi.org/10.1038/s41467-022-29132-8

Yuan C, Wu J, Yan B et al (2018) Remapping of the stripe rust resistance gene *Yr10* in common wheat. Theor Appl Genet 131:1253–1262. https://doi.org/10.1007/s00122-018-3075-9

Zhang W, Chen S, Abate Z et al (2017) Identification and characterization of *Sr13*, a tetraploid wheat gene that confers resistance to the Ug99 stem rust race group. Proc Natl Acad Sci USA 114. https://doi.org/10.1073/pnas.1706277114

Zhang C, Huang L, Zhang H et al (2019) An ancestral NB-LRR with duplicated 3′UTRs confers stripe rust resistance in wheat and barley. Nat Commun 10:4023. https://doi.org/10.1038/s41467-019-11872-9

Zhang J, Hewitt TC, Boshoff WHP et al (2021a) A recombined *Sr26* and *Sr61* disease resistance gene stack in wheat encodes unrelated *NLR* genes. Nat Commun 12:3378. https://doi.org/10.1038/s41467-021-23738-0

Zhang Z, Running KLD, Seneviratne S et al (2021b) A protein kinase–major sperm protein gene hijacked by a necrotrophic fungal pathogen triggers disease susceptibility in wheat. Plant J 106:720–732. https://doi.org/10.1111/tpj.15194

Zhang L, Liu Y, Wang Q et al (2022a) An alternative splicing isoform of wheat *TaYRG1* resistance protein activates immunity by interacting with dynamin-related proteins. J Exp Bot erac245 https://doi.org/10.1093/jxb/erac245

Zhang X, Wang G, Qu X et al (2022b) A truncated CC-NB-ARC gene *TaRPP13L1-3D* positively regulates powdery mildew resistance in wheat via the RanGAP-WPP complex-mediated nucleocytoplasmic shuttle. Planta 255:60. https://doi.org/10.1007/s00425-022-03843-0

Zhao G, Zou C, Li K et al (2017) The *Aegilops tauschii* genome reveals multiple impacts of transposons. Nat Plants 3:946–955. https://doi.org/10.1038/s41477-017-0067-8

Zhou Y, Bai S, Li H et al (2021) Introgressing the *Aegilops tauschii* genome into wheat as a basis for cereal improvement. Nat Plants 7:774–786. https://doi.org/10.1038/s41477-021-00934-w

Zhu T, Wang L, Rodriguez JC et al (2019) Improved genome sequence of wild emmer wheat Zavitan with the aid of optical maps. Genes Genom Genet 9:619–624. https://doi.org/10.1534/g3.118.200902

Zhu T, Wang L, Rimbert H et al (2021) Optical maps refine the bread wheat *Triticum aestivum* cv. Chinese Spring genome assembly. Plant J 37:528–538. https://doi.org/10.1046/j.1365-313X.2003.01977.x

Zimin AV, Puiu D, Hall R et al (2017a) The first near-complete assembly of the hexaploid bread wheat genome, *Triticum aestivum*. GigaScience 6. https://doi.org/10.1093/gigascience/gix097

Zimin AV, Puiu D, Luo M-C et al (2017b) Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. Genome Res 27:787–792. https://doi.org/10.1101/gr.213405.116

Zou S, Wang H, Li Y et al (2018) The NB-LRR gene *Pm60* confers powdery mildew resistance in wheat. New Phytol 218:298–309. https://doi.org/10.1111/nph.14964

11 Genomic Insights on Global Journeys of Adaptive Wheat Genes that Brought Us to Modern Wheat

Deepmala Sehgal, Laura Dixon, Diego Pequeno, Jessica Hyles, Indi Lacey, Jose Crossa, Alison Bentley and Susanne Dreisigacker

Abstract

Since its first cultivation, hexaploid wheat has evolved, allowing for its widespread cultivation and contributing to global food security. The identification of adaptive genes, such as vernalization and photoperiod response genes, has played a crucial role in optimizing wheat production, being instrumental in fine-tuning flowering and reproductive cycles in response to changing climates and evolving agricultural practices. While these adaptive genes have expanded the range of variation suitable for adaptation, further research is needed to understand their mechanisms, dissect the pathways involved, and expedite their implementation in breeding programs. By analyzing data across different environments and over time, Meta-QTL analysis can help identify novel genomic regions and facilitate the discovery of new candidate genes. This chapter reports on two previously unknown Meta-QTL regions, highlighting the potential for further exploration in this field. Moving forward, it will be increasingly important to expand our understanding of how genetic regions influence not only flowering time but also other developmental traits and their responses to environmental factors. Advances in gene-based modeling hold promise for describing growth and development processes using QTL and other genomic loci analysis. Integrating these findings into process-based crop models can provide valuable insights for future research. Overall, the study of adaptive genes and their

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-38294-9_11.

D. Sehgal · D. Pequeno · J. Crossa · A. Bentley · S. Dreisigacker (✉)
International Maize and Wheat Improvement Center (CIMMYT), Texcoco, Mexico
e-mail: S.Dreisigacker@cgiar.org

D. Sehgal
e-mail: Deepmala.Sehgal@syngenta.com

D. Pequeno
e-mail: D.Pequeno@cgiar.org

J. Crossa
e-mail: J.Crossa@cgiar.org

A. Bentley
e-mail: a.bentley@cgiar.org

L. Dixon
School of Biology, University of Leeds, West Yorkshire, UK
e-mail: L.Dixon2@leeds.ac.uk

J. Hyles
The Plant Breeding Institute, University of Sydney, Cobbitty, NSW, Australia
e-mail: Jessica.Hyles@csiro.au

J. Hyles · I. Lacey
CSIRO Agriculture and Food, Canberra, ACT, Australia
e-mail: bs18isl@leeds.ac.uk

R. Appels et al. (eds.), *The Wheat Genome*, Compendium of Plant Genomes,
https://doi.org/10.1007/978-3-031-38294-9_11

impact on wheat production represents a vital area of research that continues to contribute to global food security.

Keywords

Hexaploid wheat · Adaptive genes · Novel genomic regions · QTL · Gene-based modeling · Process-based modeling · Global food security

11.1 Historical Perspective

Archaeological evidence suggests that hexaploid wheat was first cultivated in the Fertile Crescent of the Middle East around 7000 BC and that farming spread to Europe (former Yugoslavia, Bulgaria, Greece) approximately one thousand years later (Hillman 1972; Renfrew 1973). By approximately 4000 BC, wheat production had reached China, with archeological isotope analysis suggesting diets shifted from a dominance of C4 crop millet, to C3 cereals including wheat (Li et al. 2007; Cheung et al. 2019). The coincidence of changing climate, whereby conditions became colder and drier, is proposed to have led to the adoption of wheat due to its greater flexibility in sowing time to achieve yield (Cheung et al. 2019). The ancient Greek poet Hesiod described an awareness of the importance of the seasonal timing of wheat development as early as 800 BC in Greece (Aitken 1974), and approximately two thousand years later, French scientist Réaumur constructed a thermometer and showed that crop maturity was influenced by temperature (Réaumur 1735). In 1751 Carl von Linné published a floral calendar in Philosophia Botanica, observing that plant responses to the environment varied in different climates (Linné and Freer 2007), and since this time, multiple evidence of variation in flowering time due to temperature, daylength, and latitude has been reported (Aitken 1974). From the eighteenth century, bread wheat (*Triticum aestivum* L.) has grown on all continents except Antarctica, and to ensure successful cultivation in different environments, wheat breeding (hybridization and selection to achieve adaptation) began.

11.1.1 Early Breeding and Selection for Seasonal Adaptation

In France in 1743, the seed merchant Jeanne Claude Geoffroy and botanist Pierre d'Andrieux founded a seed company which began the Vilmorin family dynasty of wheat breeding that lasted more than 200 years. There is evidence that pedigree-based breeding was used at the Vilmorin company from 1840, with selection of seeds based on evaluation of progeny performance (Gayon and Zallen 1998). Henry de Vilmorin described the importance of wheat adaptation in Les meillieurs bles ("The best wheat") which illustrated the morphology, origin, adaptation, and best agronomic practice for different varieties. His astute preface included, "one of the best ways to increase harvests without increasing expenditure is to cultivate the breeds of wheat which are best suited to the circumstances in which the land is cultivated" and to "choose knowingly the most advantageous wheat in each locality" (Vilmorin 1880). Vilmorin had begun hybridization experiments in 1873, including the use of wheat which had been selected by Scottish agriculturalist Patrick Shirreff (Vilmorin 1880). Vilmorin's first variety DATTEL released 10 years later was the result of crossing an early maturing, short stature type from France with late maturing English wheat. DATTEL became widely adopted, as resistance to lodging and earliness created a uniform crop with high yield potential. A string of successful cultivars followed including VILMORIN 23, VILMORIN 27, and VILMORIN 29 and many others which feature in the ancestry of modern wheat (Lupton 1987).

At approximately the same time another European breeder, Wilhelm Rimpau was also crossing native types to English Squarehead wheat for improved yield, using North American varieties as donors of quality and winter hardiness. His most successful cultivar RIMPAU'S FRÜHER BASTARD was the most widely

grown in Germany for over 50 years after being released in 1889 (Porsche and Taylor 2001). That same year, pioneer breeder William Farrer made his first wheat crosses in Australia. Farrer also focused on introgression of wheat to improve quality and adaptation. He crossed European Purple Straw with Canadian Fife and Indian wheat, and the resulting early maturing cultivars were successful in Australia because the short life cycle avoided water-limiting conditions in summer and escaped rust infection. Farrer cultivars went on to dominate Australian wheat production in the early 1900s, on the basis that "He recognized that the characteristics of a variety limited its successful growth to certain localities, and therefore set himself the task of breeding varieties adapted to the different conditions" (Guthrie 1922).

The Canadian "hard" wheat FIFE used for crossing by Farrer created cultivars with increased dough strength relative to the soft white wheats traditionally used for baking. Initially, Farrer wheat was met with resistance from millers. From the Rust in Wheat Conference in Melbourne, 1896, "A prominent obstacle this Conference has met with has arisen from the objection of millers. The opinion this Conference has long held is that the opposition of millers to such wheats has no legitimate foundation but arises from either misconception or from conservatism" (Guthrie 1922). Australian millers realized the superior quality of Farrer wheat only when American wheat of the same type was imported to Australia to meet local demand (Guthrie 1922).

Other breeders were also crossing Canadian FIFE and INDIAN wheat. In Canada, Percy Saunders crossed RED FIFE and HARD RED CALCUTTA, and the resulting cultivar MARQUIS was selected and released by Charles Saunders in 1908. With excellent quality and adaptation through early maturity, MARQUIS dominated Canadian production and became the gold standard for quality classification (Lupton 1987; McCallum and DePauw 2008). The overwhelming popularity of MARQUIS (and two later releases, THATCHER and NEEPAWA) highlights a negative consequence if few adapted cultivars are widely used, that is, a decline of genetic diversity in the breeding pool over time (Fu and Dong 2015).

11.1.2 Expanding Knowledge of Seasonal Patterns

Fluctuating patterns of seasonal flowering time have been well documented across plant species (Andrés and Coupland 2012). Early work by Garner and Allard (1920) described the relationship between daily light duration, plant growth, and reproduction across several plant species. Their work demonstrated that daily light duration impacted both the rate and extent of growth as well as the time to reach and complete flowering and reproduction (Garner and Allard 1920). In wheat, Chinoy (1950) demonstrated that long days induced earlier onset of the reproductive phase with both cold (vernalization) and light (photoperiod) having measurable impact on development and growth. It was proposed that the first wheat (which were domesticated in the Fertile Crescent) shared both vernalization requirements and photoperiod responses with their progenitors, but that selection for alternative adaptation facilitated the spread of wheat throughout Europe, and then worldwide (reviewed by Cockram et al. 2007).

The detection of major genes controlling vernalization (positive vernalization response from wheat variety INSIGNIA 49 (Pugsley 1963)) and photoperiod (from Canadian variety SELKIRK (Pugsley 1965)) was demonstrated in segregating populations and provided evidence for simple inheritance. This offered the opportunity to apply selection for daylength specificity (Pugsley 1965), although additional genetic controllers were hypothesized. Genes for daylength duration, *Photoperiod-1* (*PPD1*), were mapped to the homoeologous group 2 chromosomes (Law et al. 1978) and studies by Martinic (1975) and Hunt (1979) demonstrated the prevalence of photoperiod-sensitive winter wheat in northern latitudes and photoperiod insensitivity in southern Europe.

Creation of near-isogenic wheat lines capturing *PPD1* variation by Worland and Law (1986) and Worland et al. (1998) confirmed genetic effects and allowed the understanding of their environmental performance throughout Europe. This demonstrated a yield disadvantage from earlier flowering in the UK, a moderate advantage in Germany, and a significant advantage in southern Europe (based on testing in the former Yugoslavia). These effects have been further elaborated with Börner et al. (1993) confirming that middle European varieties benefit from daylength sensitivity (conferred by *PPD1*), whereas insensitivity offers productivity-related increases where wheat experiences hot and dry summer conditions. Hoogendoorn (1985) assessed phenotypic response to photoperiod and vernalization in a collection of 33 wheat varieties from a range of geographies. This confirmed a prevalence of photoperiod sensitivity in varieties from Europe and North America and insensitivity in varieties originating from Mexico, India, and Australia.

Since the development of understanding the major controllers of photoperiod response and vernalization requirements in wheat, variation for the *PPD1* and *Vernalization-1* (*VRN1)* genes has aided wheat's adaptation to a wide range of global production environments (Sheehan and Bentley 2020). In many geographies, there is a documented progression of wheat cultivation across climatic features and areas including in North America (Olmstead and Rhode 2011), Asia, the Mediterranean, North Africa (Ortiz Ferrara et al. 1998), and China (Yang et al 2009).

11.1.3 Further Adaptive Progress Through Time

Farrer was first to target early maturity to breed adapted wheat for Australia, although the introduction of additional genetic diversity for phenology had been identified (Eagles et al. 2009). In 1945, Australian breeder Walter Lawry Waterhouse introgressed hexaploid wheat and an early maturing durum wheat, GAZA, producing an important cultivar, GABO. This daylength insensitive wheat was the leading cultivar in Australia for many years and sister line TIMSTEIN was also successfully cultivated in USA. This germplasm was utilized by the International Maize and Wheat Improvement Center (CIMMYT) in the breeding of cultivars such as CAJEME, MAYO, and NAINARI (Watson and Frankel 1972).

Other donors of photoperiod insensitivity have been traced to Japanese landrace AKAGOMUGHI (which also carried dwarfing gene, *RHT8*) and Chinese landraces MAZHAMAI and YOUZIMAI (Yang et al. 2009). Yang et al. (2009) showed that the distribution of alleles for daylength sensitivity depended upon the climate (temperature and latitude) where the wheat was cultivated. The adoption of photoperiod insensitive wheat by CIMMYT was key to the success of the shuttle-breeding program, whereby material could undergo selection in multiple environments due to broad adaptation (Trethowan et al. 2007). It is the subsequent sharing of germplasm during the "Green Revolution" which likely facilitated the spread of alleles for daylength insensitivity around the globe.

As the climate changes over time and new crop management practices are developed, it is probable that new genetic variation will be required for enhancing adaptation (Hunt et al. 2019). For instance, studies have highlighted a shift to early sowing which has meant that vernalization responsive, long season wheat are beneficial in some areas of southern Australia where spring types are traditionally cultivated (Hunt 2017; Cann et al. 2020). To expedite development of future adapted wheat cultivars, it is important to understand the genetic architecture of phenology and develop breeding tools such as molecular markers and simulation models for prediction.

11.2 Understanding the Genetic Control of the Synchrony of Flowering

11.2.1 The Three Known Gene Systems

As outlined above, within less than 10,000 years, wheat cultivation has expanded from its primary area of evolution within the Fertile Crescent to a broad spectrum of agro-ecology around the globe, adapting rapidly to a wide range of climatic conditions (Curtis 2002; Salamini et al. 2002). The essential path to achieve adaptation is the synchrony of flowering which in wheat is controlled by three major gene systems: (1) the *VRN* genes (exposure to cold temperature), (2) the *PPD* response genes (sensitivity to daylength), and (3) the autonomous earliness per se (*EPS*) genes (Kato and Yanagata 1988). The adaptation of a wheat genotype to a particular environment depends to a large extent on the interaction of these three systems.

11.2.2 Vernalization (VRN) Genes

Vernalization is the acquisition of a plant's ability to flower by exposure to cold (Chouard 1960). According to the vernalization requirement of a genotype, wheat is classified as having a winter or spring growth habit. Winter wheat has a considerable vernalization requirement, but spring wheat may be insensitive or only partly sensitive to vernalization (Trevaskis et al. 2003). The key element of the vernalization gene system is *VRN1* with its three orthologous genes (*VRN-A1*, *VRN-B1*, and *VRN-D1*) located on the long arms of chromosomes 5A, 5B, and 5D, respectively (Figs. 11.1 and 11.2a). *VRN1* is a member of the MADS-box transcription factor family, which has been shown to play a critical role in flowering gene models across crops (Zhao et al. 2006). The MIKC-type MADS-box proteins have a highly conserved MADS DNA-binding domain, an intervening (I) domain, a keratin-like (K) domain, and a C-terminal domain (C). The proteins bind as dimers to DNA sequences named "CArG" boxes and organize in tetrameric complexes (Li et al. 2019). The multimeric nature of these complexes generates many combinatorial possibilities with different targets and functions (Li et al. 2019; Honma and Goto 2001; Theißen et al. 2016).

Mutations in the promoter and deletions in the large first intron of *VRN1* are associated with increased expression of the genes in the absence of cold, accelerated flowering without vernalization and thus spring growth habit (Kippes et al. 2018). Additionally, single nucleotide polymorphisms (SNPs) in exons 4 and 7 have been identified in *VRN-A1* (Eagles et al. 2011; Muterko and Salina 2018). The exon 4 SNP results in an amino acid change (Leu117 → Phe117) in the conserved k-domain (Eagles et al. 2011; Chen et al. 2009; Díaz et al. 2012). This polymorphism was associated with a change in the number of days to stem elongation, vernalization requirement duration, frost tolerance, and flowering time in winter wheat (Chen et al. 2009; Muterko et al. 2016; Dixon et al. 2019). Another *VRN-A1* SNP that causes an amino acid substitution (Ala180 → Val180) in exon 7 in the C-terminal domain also regulates vernalization duration, via its regulation of a protein interaction with *TaHOX1* (Li et al. 2013).

Beyond regulation by alterations in nucleotide sequence (INDELS and SNPs), there is increasing evidence that vernalization in wheat is also regulated at the epigenetic level. The *VRN-A1* gene can be present as two or more copies with the assumption that the number of copies positively correlates with the vernalization requirement duration and flowering time of wheat (Díaz et al. 2012). The different nature of the diverse mutations (promoter insertions, intron deletions of different size, SNPs) in the three *VRN1* orthologs and gene duplication in the A genome are the most plausible explanation for varying gene actions observed (Li et al. 2013). Dominant alleles at *VRN-A1* have been shown to confer the largest effects leading to a lack of vernalization requirement relative to

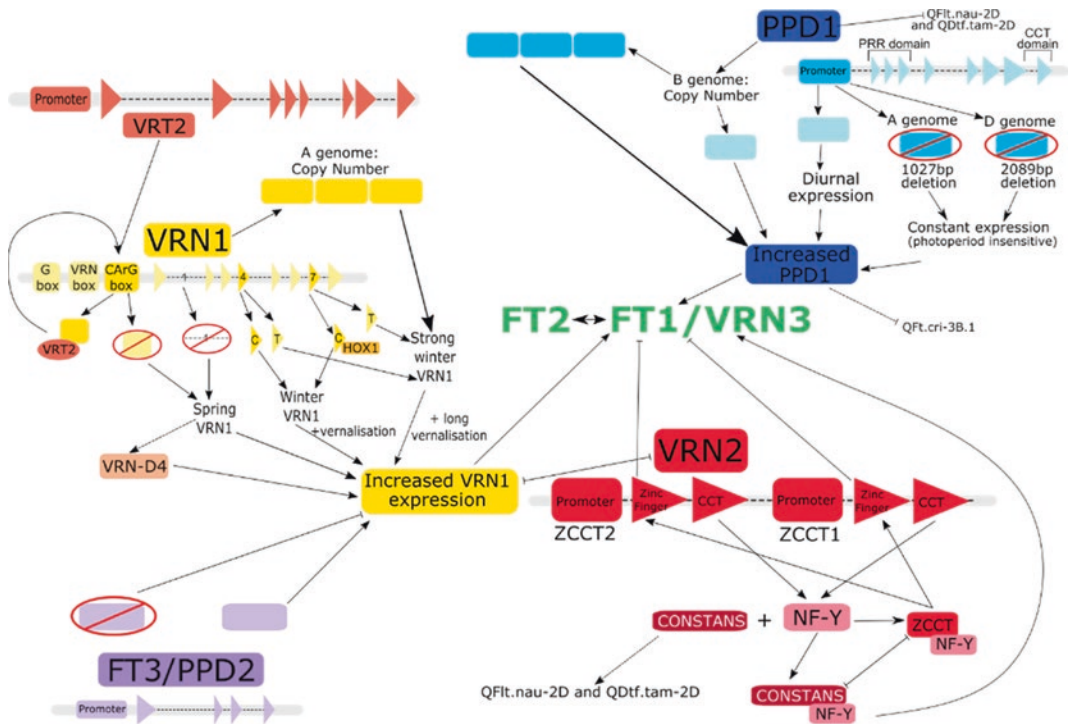


Fig. 11.1 Major flowering genes involved in photoperiod and vernalization response. The major genes involved with photoperiod and vernalization responses in wheat are highlighted in different colors. For each gene, known allelic variation is included and the effect of this variation on the level of expression or the flowering response is shown. The structure of each gene, along with the annotated domains, is represented on a gray background bar. Where interactions with uncharacterized QTL regions are known, these are also included on the network diagram. Deletions are indicated by a red oval with a line through it. Different *VRN1* alleles can determine the extent to which vernalization is required to increase expression, due to CArG box, *VRT2* interactions, and exon variants, including changes in copy number. A duplication and translocation of *VRN1*, in the form of *VRN-D4*, also promote spring habit. The locus *VRN2* (*ZCCT1* and *ZCCT2*) is a photoperiod-dependent repressor of *VRN1*, competing with CONSTANS and the nuclear transcription factor Y (NF-Y) proteins to activate *FT1* (also called *VRN3*) and potentially *FT2*. The *FT* genes interact with FD-like genes (*FDL2* or *FDL12*) to form a floral activating complex. Copy number variants, most notably of *VRN1* and *PPD1*, can determine heading date. *PPD1* determines flowering time through photoperiod sensitivity with variations in promoter deletions and copy number influencing expression levels. Short-day promotion of flowering is mediated through FT3 (also called PPD-2)

VRN-B1 and *VRN-D1*, which reduced vernalization requirement and defined semi-spring or facultative types (Trevaskis et al. 2003).

Other MADS-box genes also play a role in the regulation of wheat flowering. *VRN-D4* is a MADS-box transcription factor derived from the duplication and translocation of the *VRN-A1* gene to the short arm of chromosome 5D (Kippes et al. 2015). Being an extra gene copy, *VRN-D4* is associated with increased *VRN-A1* expression and thus reduced vernalization requirement. The *VEGETATIVE TO REPRODUCTIVE TRANSITION 2* (*VRT2*) gene belongs to the group of MADS-genes as *SHORT VEGETATIVE PHASE* in Arabidopsis and interacts with *VRN1*. The *VRT2* protein has been shown to bind to the CArG box in the *VRN1* promoter region, suggesting that *VRT2* represses the transcription of *VRN1* (Dubcovsky et al. 2008; Kane et al. 2007). More recently, Xie et al. (2021) corroborated an epistatic interaction between the two genes (including the ability of *VRT2* to bind to the promoter region of *VRN1*), but reported a shared upregulation of *VRN1* and *VRT2*.

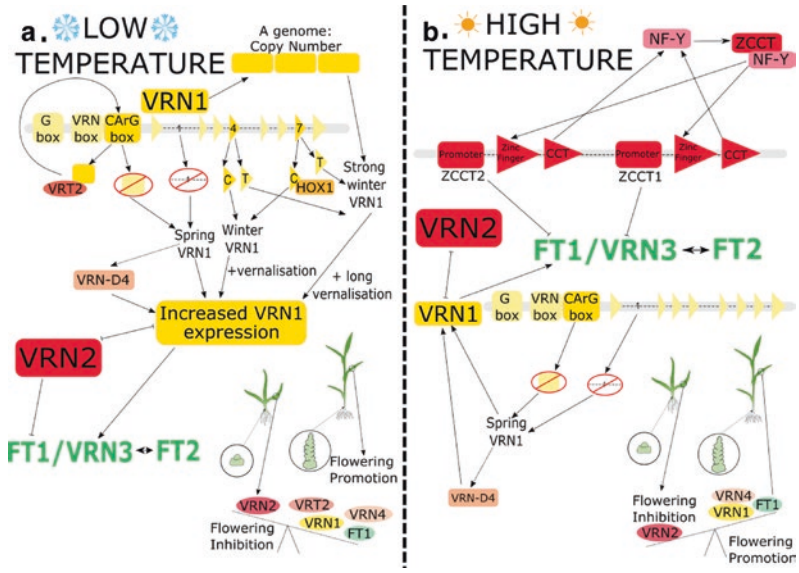


Fig. 11.2 Impact of major flowering genes responses to different temperatures, in the context of vernalization. This figure represents the role of temperature in the regulation of vegetative to reproductive meristem transition, and how this relates to the vernalization pathway. To indicate the different aspects of the flowering pathway and how responses which occur are more influenced by specific environmental conditions the pathway has been separated into **a** low temperatures and **b** high temperature (post or non-requiring vernalization), although it must be emphasized that each aspect does not act independently. The same gene structure and nomenclature are used as for Fig. 11.1. Additionally, the weighting of each gene signal is indicated in a seesaw schematic (see also Fig. 11.4)

In addition to *VRN1*, the *VRN2* and *VRN3* genes are located on the long arm of chromosome 5A and the short arm of chromosome 7B, respectively (Figs. 11.1 and 11.2). The *VRN*2 locus consists of two closely related genes (*ZCCT1* and *ZCCT2*) that encode proteins carrying a putative zinc finger and a CCT domain (Yan et al. 2004). The CCT domain is a 43-amino acid region, first described in protein sequences of CONSTANS (CO), CONSTANS-like (COL), and TIMING OF CAB1 (TOC1) (Putterill et al. 1995; Strayer et al. 2000; Robson et al. 2001) that is present in multiple regulatory proteins associated with light signaling, circadian rhythms, and photoperiodic flowering (Wenkel et al. 2006). *VRN2* is the major flowering repressor identified in wheat. Dominant gene action in combination with recessive *VRN1* and *VRN3* allele combinations confers winter wheat growth habit. Deletions or mutations involving positively charged amino acids at the CCT domain are associated with recessive *ZCCT1* and *ZCCT2* alleles for spring growth habit (Yan et al. 2004; Dubcovsky et al. 2005; Distelfeld et al. 2009). The CCT domains in ZCCT1, ZCCT2, and CO proteins further interact with proteins of the NUCLEAR FACTOR-Y (NF-Y) transcription factor family. Mutations in the CCT domain of ZCCT proteins also reduce the strength of ZCCT-NF-Y interactions and the ability of ZCCT1 to compete with CO to activate *VRN3* (Figs. 11.1 and 11.2b).

The *VRN3* gene encodes a RAF kinase inhibitor-like protein and has been mapped to the *FLOWERING LOCUS T-like* gene, often referred to as *FT1* in wheat. *VRN3/FT1* is expressed in long days in vernalized plants or spring types and thus triggers long-day-induced flowering. The VRN3/FT1 protein has been shown to travel through the phloem carrying the photoperiodic signal from the leaves to the shoot apex where it forms a protein complex binding to the promoter of *VRN1*, promoting its further expression. Dubcovsky et al. (2008)

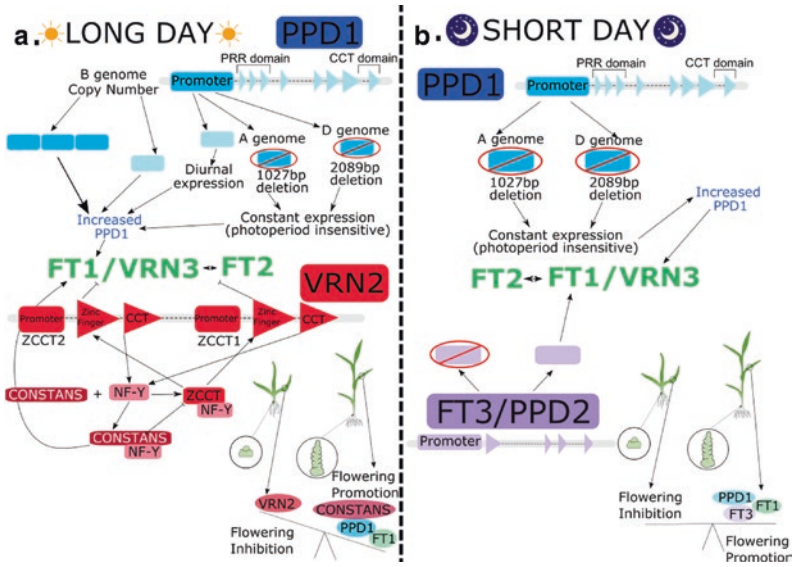


Fig. 11.3 Impact of major flowering genes responses to different daylengths. The role of daylength is represented in the regulation of vegetative to reproductive meristem transition. To indicate the different aspects of the flowering pathway and how responses which occur are more influenced by specific environmental conditions, the pathway has been separated into **a** long-day, **b** short-day, although it must be emphasized that each aspect does not act independently. The same gene structure and nomenclature are used as for Fig. 11.1. Additionally, the weighting of each gene signal is indicated in a seesaw schematic (see also Fig. 11.4)

demonstrated interaction of *Vrn3/FT1* with the FT2, FDL2, and FDL13 proteins. Transgenic plants showed that increased transcript levels of *FT2* (a *FT* paralogue) provide transcriptional activation of *VRN1*. *VRN3/FT1* therefore integrates the vernalization and photoperiod response gene systems. High levels of *VRN3/FT1* expression can overcome the vernalization requirement and are associated with spring growth habit (Figs. 11.1 and 11.2b) (Yan et al. 2006).

11.2.3 Photoperiod (PPD) Response Genes

Photoperiod genes promote the floral transition in response to long days (Searle and Coupland 2004). Photoperiod-sensitive wheat has a long-day phenotype. They flower earlier when the days are longer than a critical threshold. Photoperiod-insensitive wheat flowers largely independently of daylength and can be grown to maturity in long- or short-day environments. Photoperiod response is mainly controlled by the semi-dominant homoeologous *PPD1* genes on the short arm of chromosome group 2 (Law et al. 1978; Welsh et al. 1973). *PPD1* belongs to a pseudo-response regulator (*PRR*) gene family, which is characterized by a pseudo-receiver domain near the amino-terminus and a 43 amino acid CCT domain near the carboxy-terminus of the protein (Mizuno and Nakamichi 2005). Wild-type alleles of *PPD1 (PPD-1b)* have a rhythmic diurnal pattern of rather low gene expression and are associated with daylength sensitivity (Figs. 11.1 and 11.3). Non-wild-type alleles of *PPD1 (PPD-1a)* alter the expression of the gene, leading to elevated transcription throughout the day, and accelerated flowering through elevated *FT1* expression (Kitagawa et al. 2012). This can substitute for long days and reduce daylength sensitivity.

Several non-wild-type, photoperiod-insensitive alleles are known for *PPD1*. At the *PPD-D1* locus, a 2 kb deletion upstream of the coding

region of the gene confers photoperiod insensitivity of semi-dominant type (Beales et al. 2007). This mutation has been recognized as the major source of earliness in wheat varieties worldwide. Tanio and Kato (2007) described a *PPD-B1a* mutation from the Japanese cultivar FUKUWASEKOMUGI and Nishida et al. (2013) characterized a *Ppd-B1a* allele (a 308 bp insertion in the 5'-upstream region) derived from the Japanese landrace "SHIROBOR21". No genetic locus for *PPD-A1* has been defined in hexaploid wheat. However, Wilhelm et al. (2009) described two photoperiod-insensitive alleles from tetraploid wheat: "*GS-100*" *PPD-A1a* and "*GS-105*" *PPD-A1a*. These alleles have deletions of 1027 bp ("*GS-100*") and 1117 bp ("*GS-105*") in a similar region of the upstream promoter to *PPD-D1a*.

Nishida et al. (2013) described a *PPD-A1a* mutation (1085 bp deletion in the 5'-upstream region) in the Japanese hexaploid wheat cultivar CHIHOKUKOMOGI which is in a similar location to the deletions described by Wilhelm et al. (2009) but appears to be unique to Japanese wheat. In addition to photoperiod-insensitive mutations, Beales et al. (2007) identified candidate null alleles for *PPD-A1* and *PPD-D1* in photoperiod-sensitive cultivars. The loss of function alleles delays flowering time associated with reduced expression of *FT1*, similar to the wild-type alleles (Shaw et al. 2013).

Similar to *VRN1*, there is also variation among the potencies of the three *PPD-1a* loci, where plants with *PPD-A1a* and *PPD-D1a* are earlier in flowering than plants with *PPD-B1a.* Díaz et al. (2012) and Würschum et al. (2015) showed that alleles of *PPD-B1* were associated with increased copy number resulting in earlier flowering. These results, along with multiple copies of *VRN1*, confirm that copy-number variation is important for the adaptation of wheat.

More recently, three candidate genes for *PPD2* and *PPD3* (also designated as *FT3-B1*, *FT3-D1*, and *TOE-B1*) controlling short-day flowering pathway were identified on the long arm of chromosome group 1 in wheat (Zikhali et al. 2014, 2017; Halliwell et al. 2016). Four variations were observed for *FT3-B1* including the wild-type allele, a complete deletion of the gene, a SNP in the exon 3 causing an amino acid change (Gly → Ser), and copy-number variants. Both the deleted and mutated alleles confer delayed flowering under short-day photoperiod (Figs. 11.1 and 11.3). At the *FT3-D1* locus, a SNP in exon 4 was identified. The candidate gene for *PPD3*, *TOE-B1* is still speculative. SNPs in exons 1 and 9 of the *TOE1-B1* gene were shown to separate earlier flowering from later flowering cultivars suggesting the gene to be a putative flowering time repressor, while the mutant allele is expected to attribute earliness (Zikhali et al. 2017). A summary of the role of each gene on the different environmental signals on floral meristem development is again summarized in Fig. 11.4.

11.2.4 Earliness Per Se (EPS) Genes

The photoperiod and vernalization gene systems allow the coarse tuning of adaptation. However, there are still relatively minor variations in flowering time once requirements of vernalization and photoperiod are totally satisfied. These differences are regulated by earliness per se (*EPS*) genes, usually of small effect but critical for fine-tuning developmental phases in the crop cycle (Zikhali and Griffiths 2015; Griffiths et al. 2009). The genetics of *EPS* is still not well understood, and underlying genes with causal polymorphisms have only recently been identified in hexaploid wheat (Zikhali et al. 2017). In wild species *Triticum monococcum* L., a cereal ortholog of *Arabidopsis thaliana* circadian clock regulator *LUX ARRHYTHMO/PHYTOCLOCK 1 (LUX/PCL1)* was proposed as a promising candidate gene for the earliness per se 3 (*Eps-3A*^{m}) locus and the ortholog circadian clock regulator *EARLY FLOWERING 3* (*ELF3*) was identified as a candidate gene for the earliness per se *Eps-A*^{m}*1* locus (Gawroński et al. 2014; Alvarez et al. 2016). *ELF3* was suggested to be the best candidate gene within the *EPS-D1* locus in hexaploid wheat as a deletion containing *ELF3* is associated with advanced flowering (Zikhali et al. 2016). Recently, two

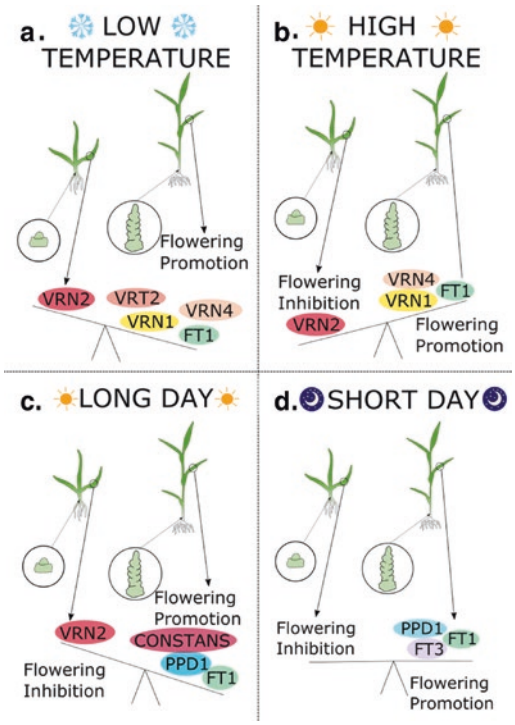


Fig. 11.4 Summary of the roles of different environmental signals on floral meristem development. Different environmental signals are used by plants to regulate the timing and rate of floral meristem development. The relative proportions of the expressed genes are shown in the seesaw summary figures, and the impact of these expression patterns on the floral developmental stage is indicated by the tipping of the seesaw balance. Environmental conditions which are considered are **a** low temperatures, **b** high temperatures (post or non-requiring vernalization), **c** long day, and **d** short day

additional *EPS* QTL in hexaploid wheat located on chromosomes 2B and 7D with the designated names *EPS-B2* and *EPS-D7* were identified (Basavaraddi et al. 2021a), and for the first time, interaction between both genes could be shown. *EPS* genes owe their name to the assumption that they act independent of environment. Despite this, Eps × temperature interaction was recently proven in some instances (Ochagavía et al. 2019; Prieto et al. 2020; Basavaraddi et al. 2021b). In barley, the *EPS* gene *ELF3* has been shown to play a role in the response of circadian clock genes to temperature (Ford et al. 2016).

11.3 Quantitative Trait Loci (QTL) for Flowering Time

The selection of spring and photoperiod-insensitive type cultivars during the evolutionary and breeding history of wheat preceded any methods for gene identification. However, the identification of the causal genes in the last two decades was important to enable targeted selection and therefore the potential for the directed development of new cultivars. One of the major methods utilized in the process of genetic mapping and gene identification is quantitative trait loci (QTL) analysis. QTL analysis is a powerful statistical tool used to calculate the probability of any marker within a genetic map contributing to the observed phenotype. The resolution and reliability of this method is increased via larger mapping populations, as these support higher levels of recombination. The resolution is also increased through an even distribution of markers in the genetic map; however, this is dependent on polymorphisms between the parent genotypes and can be severely limited when diversity is low. This is regularly observed for the D-genome of wheat or when mapping populations are generated between cultivars with a recent shared pedigree. Individual QTL analysis to identify flowering time genes has been conducted for a vast number of mapping populations under a large and diverse set of environmental conditions (for details refer to the next section). These have identified certain genetic hot spots for flowering time regulation, including the regions of major genes previously mentioned, e.g., on chromosomes 5A (*VRN-A1*) along with 7B (*VRN-B3*) and 2D (*PPD-D1*). Within these hot spot regions, it is apparent that multiple genes which regulate flowering time are closely genetically associated. The indication of these genetic hubs, combined with the dominance of the *PPD1* and *VRN1* genes in flowering regulation, suggests that there could be value in assessing the identified QTL for flowering through a meta-QTL (MQTL) analysis. This analysis would identify the number and

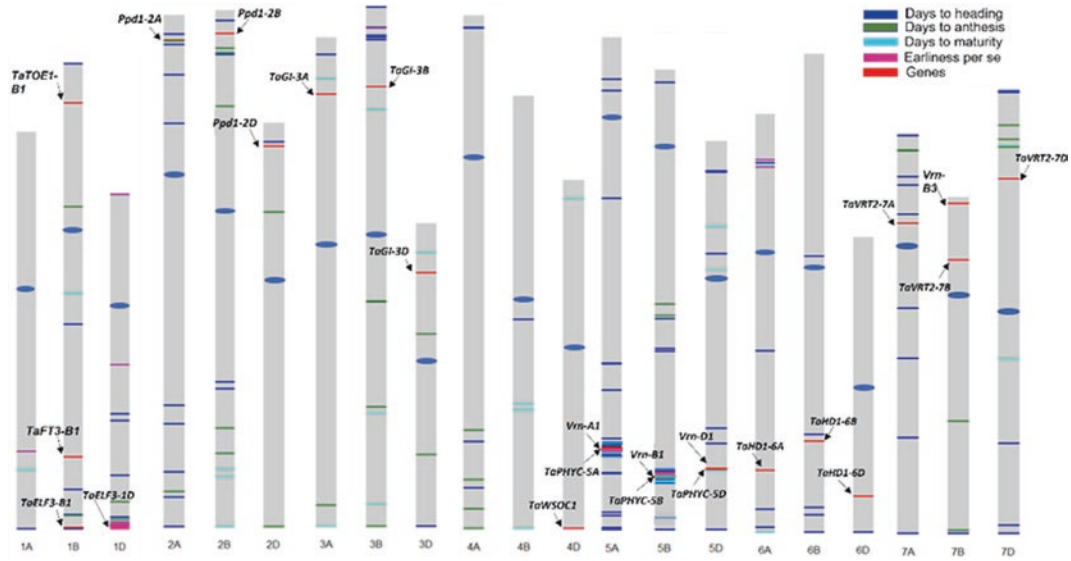


Fig. 11.5 QTL for flowering time-related traits and known major genes projected on the IWGSC RefSeq v1.0. The QTL are shown from short arm (top) to long arm (bottom). Centromeres are presented by blue ovals

genetic range of QTL beyond *PPD1* and *VRN1* and in combination with location information infer some climate-based associations for these additional QTL.

11.3.1 Meta-QTL (MQTL) Analysis

The results of a total of 18 QTL analyses and genome-wide association studies (GWAS) conducted on flowering time in bread wheat were utilized and aligned to the IWGSC-CHINESE SPRING reference sequence (IWGSC RefSeq v1.0) to identify genetic hot spots or MQTL. The studies consisted of 17 mapping populations and five GWAS panels. The traits included were days to heading, days to anthesis, days to maturity, and earliness per se (Supplementary Table S11.1). In addition, 24 flowering genes with known physical locations were integrated (Supplementary Table S11.2). We projected 201 flowering time QTL with 120, 27, 25, and 29 QTL related to days to heading, anthesis, maturity, and earliness per se, respectively (Fig. 11.5). QTL were projected on all chromosomes. The number of projected QTL per genome was 95 (47.3%), 71 (35.3%), and 59 (29.4%) for A, B, and D genomes, respectively. The number of QTL per chromosome ranged from 3 QTL on chromosome 1A to 50 QTL on chromosome 5A.

A window size of 30 Mb was used to infer MQTL. Seven MQTL for flowering time were detected that ranged from 10.5 Mb to 28.7 Mb across chromosomes (Table 11.1). On chromosome 1D, a MQTL1 was identified between 477.9 and 495.1 Mb (MQTL1) and has a size of 17.2 Mb. MQTL2 was located on chromosome 2A between the physical positions of 28.2–43.2 Mb and with a size of 15.0 Mb. On chromosome 2B, MQTL3 was located between 33.9 and 62.6 Mb and has the largest size of 28.7 Mb. MQTL4 was located between physical positions of 30.8–49.0 Mb on chromosome 3B. MQTL5 and MQTL6 were detected on chromosomes 5A and 5B with sizes of 13.8 and 15.1 Mb, respectively. MQTL5 had the maximum number of QTL (36) followed by MQTL6 (14). The MQTL7 was located on chromosome 6A between the physical positions of 67.0–77.5 Mb and had a size of 10.5 Mb.

The MQTL provides the advantage of readily separating QTL which are environmentally more stable, so might have relevance in many

Table 11.1 Summary of flowering time meta-QTL positioned on wheat reference genome IWGSC RefSeq v1.0

MQTL	Chromosome	Range of reference genome V 1.0 (Mb)	Size (Mb)	Number of QTL	Candidate gene
MQTL1	1D	477.9–495.1	17.2	7	*TaELF3-1D*
MQTL2	2A	28.2–43.2	15.0	3	*Ppd1-2A*
MQTL3	2B	33.9–62.6	28.7	3	*Ppd1-2B*
MQTL4	3B	30.8–49.0	18.2	4	
MQTL5	5A	581.1–594.9	13.8	36	*TaPHYC-5A, Vrn-A1*
MQTL6	5B	571.1–586.3	15.1	14	*TaPHYC-5B, Vrn-B1*
MQTL7	6A	67.0–77.5	10.5	4	

different locations globally from those which infrequently occur in QTL analyses. Using this distinction, marker and candidate gene identification can be targeted for specific environmental conditions and so enable the development of a deeper understanding and application of flowering time regulation.

The most frequently identified QTL identified in the MQTL analysis were located on chromosome 5 (A and B genomes) and associated with the *VRN1* region, along with the closely associated *PHYC* gene. A third very robust QTL region was identified on chromosome 1D, overlapping with the *EARLY FLOWERING 3* (*ELF3*) gene, and containing 7 QTL. Additionally, two regions were identified on chromosomes 3B and 6A where QTL were detected in multiple analyses and do not yet have a gene associated with them. Both chromosome regions on 3B and 6A are interesting targets for further investigation. Several QTL were further identified in, potentially, homoeologous regions. These may indicate that the same gene on homoeologous chromosomes contributes to the regulation on flowering time and, therefore, may represent a stable locus but with dosage effect, commonly seen in wheat. Examples for these QTL are in the proximal region of chromosome 5 and the distal region on chromosomes 5, 6, and 7.

11.4 The Effect of Major Genes on the Response to Vernalization and Photoperiod to Developmental Phases and Traits

As an essential trait, the mistiming of flowering can ultimately lead to partial or complete crop failure. However, the focus on time to flowering has meant that additional pleiotropic effects are also selected for, some of which are beneficial. The dominant regulator of vernalization, *VRN1*, is an important gene for the control of the vernalization response and also for the formation of the flower itself, highlighted by its homology to the Arabidopsis *AP1* gene (Yan et al. 2003). *VRN1*, in combination with its homologues *FUL2* and *FUL3*, contribute to the regulation of spikelet formation, plant height, and tiller progression (Li et al. 2019). Furthermore, the regulatory roles of *VRN1* are not limited to floral regulation. The growth of spring vs. winter near-isogenic lines for *VRN1* in barley identified that other traits including root density at specific soil depths were affected (Voss-Fels et al. 2018). In spring barley near-isogenic lines (NILs), root density during grain filling was increased at soil depths between 20 and 60 cm, compared to winter NILs (Voss-Fels et al. 2018).

Like *VRN1*, the regulator of photoperiod response, *PPD1*, is also linked to a number of additional phenotypes. Some of these are closely related to flowering time, for example, the rate of spikelet initiation is accelerated in *PPD1*-insensitive NILs, leading to a reduction in the number of spikelets per spike (Ochagavía et al. 2018). Likewise, the formation of additional or paired spikelets is also altered depending on the *PPD1* allele, a mechanism which is believed to be regulated through the strength of the *FT1* signal (Boden et al. 2015). Beyond the spike architecture, *PPD1* influences grain filling and dry mass production. In durum wheat (*Triticum turgidum* L. var *durum*), cultivars carrying *PPD1* alleles which conferred photoperiod insensitivity allowed earlier flowering and more robust grain filling, leading to enhanced yields. This correlation of effects may not be due to a direct effect of *PPD1* regulating these processes, but might be due to *PPD1* enabling optimal timing of flowering for the particular environment (Royo et al. 2016, 2018; Arjona et al. 2020).

Both the photoperiod and vernalization pathways are integrated through the cereal *FT1-like* gene. As such, allelic variation of *FT1* unsurprisingly shows variation in spikelet number, potentially linked with spikelet initiation rate. The link with spikelet initiation is supported as transgenic lines over-expressing *TaFT1* rapidly flower, while still on the callose regeneration media and produce a spike with only a few, infertile spikelets (Lv et al. 2014). In addition to *FT1*, cereals contain a vastly expanded family of *FT-like* genes, which are becoming a focus for characterization (Bennett and Dixon 2021). *FT2* has been linked with spikelet initiation (Gauley and Boden 2021), while *HvFT3* has also been associated with spikelet initiation in spring lines, independent of a photoperiod signal (Mulki et al. 2018). Interestingly, while *FT3* showed a role in photoperiod-independent spikelet initiation, plants were unable to complete floral development under short-day conditions, indicating that *FT3* alone cannot promote floral development (Mulki et al. 2018). Yet, in winter barley, over-expression of *FT3* could trigger the expression of *HvVRN1* and enable floral development in non-vernalized plants (Mulki et al. 2018). In contrast to this, *FT4* has been identified to function as a repressor of spikelet initiation in barley, with over-expression of *HvFT4* leading to a reduction in spikelet primordia and ultimately grains per spike (Pieper et al. 2021).

11.5 Extending Genetics to Prediction

Flowering time is a critical consideration in the adaptation of wheat to scenarios of changing environments. Future adaptation of any crop in their major producing countries must be forecast because of the substantial time lag in the planning, breeding, and release of new cultivars which can take between 6 and 10 years (Tanaka et al. 2015; Hammer et al. 2020). The delivery of climate-smart solutions for cultivars to be released in a time-reduced and cost-effective manner is a daunting challenge for current agricultural research (Ramirez-Villegas et al. 2020). Based on diverse climate change models, wheat yields will suffer climate change-related declines below current production rates in most regions, with the most negative impact projected to affect developing countries in warmer regions (Pequeno et al. 2021). For example, in a modeling study by Asseng et al. (2015), a decrease in wheat yield gain, namely a fall of 6% yield for each 1 ºC rise in temperature was predicted, with resultant uncertainty in production over space and time. More recently, Demirhan (2020) estimated a 90.4 million ton drop in global wheat production with a 1 ºC warming of surface temperature, but a 32.2 million ton increase in production associated with 1 ppm increase in CO_2 emissions. This emphasizes the complexity of climate change and its relationship with vital processes in nature.

To mitigate future uncertainties and to reduce the negative environmental impacts, exploratory simulation models or so-called "adaptation pathways" can be developed (Tanaka et al. 2015). Optimum flowering periods, defined by maximum grain yield potential, are explored by simulating interactions of genotype × environment × management (G × E × M) under current

and future climates for major crops including wheat (Pequeno et al. 2021; Zheng et al. 2012; Flohr et al. 2017; Chen et al. 2020). Thus, statistical and mechanistic models that enable prediction of the performance of plants cultivated in various environmental conditions will play a crucial role in breeding for environmental adaptability and optimization of crop management.

11.5.1 Finding Conceptional Ideotypes for Given Environments

Diagnostic molecular markers associated with the important regulatory genes and QTL related to wheat adaptation (as summarized above) provide a method for identifying existing allelic variation and estimating the effects of each of the alleles in a diverse target production environment. The estimated allele effects can be used to conceptualize ideotypes or genotypes that fit to a specific flowering time range or can predict outcomes of specific crosses in breeding.

Allelic variation in vernalization genes does not contribute to large differences in flowering time in environments where vernalization saturation occurs. A large worldwide panel of varieties was evaluated by Würschum et al. and revealed that a three-component system facilitated the adoption of heading date in winter wheat (Würschum et al. 2018). The *PPD-D1* locus was found to account for almost half of the genetic variance (the photoperiod-insensitive allele *PPD-D1a* mainly present in eastern and southern Europeans as well as in Eurasian cultivars), followed by copy-number variation at *PPD-B1*. Further fine-tuning to local climatic conditions was attributed to small-effect QTL. Sheehan and Bentley (2020) recently documented a dialog with UK wheat agronomists, outlining the requirement of greater flexibility of varietal flowering time (preferably earlier flowering genotypes) in UK winter wheat to find ideotypes for expected changing seasonal conditions, and increasing seasonal weather fluctuations.

In spring wheat, Cane et al. (2013) attempted to define a conceptual genotype or ideotype for environments in southern Australia characterized by variable rainfall in late autumn and early winter. The authors suggested a spring cultivar with slowed development from early sowing, followed by rapid development with increasing temperature and daylength, was an optimal type. The authors defined the allele combination (1) *PPD-B1* (3-copy variant) + *PPD-D1a* + *VRN-A1w* (WICHITA allele) + *VRN-B1* + *VRN-D1a* or (2) *PPD-B1* (3-copy variant) + *PPD-D1b* + *VRN-A1w* (WICHITA allele) + *VRN-B1* + *VRN-D1a* as most suitable. Overall, the variability present in modern Australian spring wheat cultivars was high and diverse combinations of alleles had been successful in the past and were widely grown (Eagles et al. 2009). Recently, Christy et al. (2020) developed a photoperiod-corrected thermal model that solely utilized the combination of *PPD* and *VRN* alleles to predict wheat phenology to identify the phenological suitability of germplasm across the cropping region in southern Australia. Similar to Cane et al. (2013), the authors used their model to identify the optimum allelic combinations required to target optimum flowering period for different locations when sown on different dates. By comparing a series of NILs with different major allele combinations and diverse phenology in the field, Bloomfield et al. (2019) however revealed that a model parameterized solely using multi-locus genotypes is not accurate enough to predict the adoption to flowering time under field conditions. For more accurate predictions, the authors suggested quantifying minor genetic drivers and including genotype × environment (G × E) interaction into models based on genetically derived parameter estimates.

In breeding programs, the major *VRN* and *PPD* loci are usually quickly fixed when targeted at a specific selection environment. In widely adapted CIMMYT spring bread wheat, bred mainly in Mexico but globally distributed through international nurseries and yield trials, the two spring alleles *VRN-B1a*, *VRN-D1a* and the *PPD-D1a*-insensitive allele are the most frequent (Van Beem et al. 2005; Dreisigacker et al. 2021a, b). Also apparent is a strong selection pressure against the spring allele, *VRN-A1a*,

which results in a strong negative effect on the accumulation of biomass and yield at the CIMMYT main selection site at CENEB, in North Mexico, suggesting that genotypes with some vernalization sensitivity are better adapted. Greater allelic variation was found at the *PPD-A1*, *PPD-B1* (including copy number variants), and *VRN-D3* loci. Further, alleles at the two more recently identified photoperiod genes, *TaTOE-B1* and *TaFT-B3*, positively promoted harvest index and yield (Dreisigacker et al. 2021a, b).

11.5.2 Genomic Prediction

With the swift development of next-generation sequencing technologies, whole-genome marker information is generated for all types of germplasm sets. Instead of using only several major loci, genomic prediction/selection aims to utilize whole-genome marker information to predict plant phenotypes (Meuwissen et al. 2001) and thus also includes minor genetic drivers of a trait. While the approach was initially proposed in animal breeding, studies on genomic prediction have been growing in crops including wheat and have become a practical tool in breeding (de Los Campos et al. 2009; Crossa et al. 2010, 2014; Dreisigacker et al. 2021a, b). Flowering time as an important agronomic trait has been predicted with genome-wide markers in wheat using different training and target populations. Within-environment and within single populations' genomic prediction accuracies for flowering time or heading date, measured as the correlation between genomic estimated breeding values and the observed traits, are in the range of 0.4 and 0.7 in the published literature guided by heritability (Charmet et al. 2014; Zhao et al. 2014; Liu et al. 2020; Haile et al. 2021; Crossa et al. 2016).

Predicting the performance of plant phenotypes across diverse environments is more difficult compared to within-environments because phenotypes of more complex traits are often influenced by G × E interaction. Multi-environment trials (METs) for assessing the G × E interaction are therefore common practice in plant breeding for selecting high-performing, well-adapted lines across environments. Models have been developed that evaluate G × E interaction in genomic prediction. Burgueño et al. (2012) were the first to use marker and pedigree genomic best linear unbiased prediction (GBLUP) models to assess G × E. Jarquín et al. (2014) proposed a reaction norm model where the main and interaction effects of markers and environmental covariates are introduced using highly dimensional random variance–covariance structures of markers and environmental covariables. A marker × environment (M × E) interaction model was proposed by Lopez-Cruz et al. (2015) and decomposed the marker effects into components that are common across environments (stability) and environment-specific deviations (interaction). Genomic prediction models that incorporate G × E or M × E interaction have shown to increase prediction accuracies by 10–40% with respect to within-environment analyses (Dreisigacker et al. 2021a, b; Crossa et al. 2017; Pérez-Rodríguez et al. 2017).

Another way to improve the prediction accuracy of G × E is to introduce secondary traits measured in each environment on both the training and target populations in multi-trait genomic prediction models. Recently, Guo et al. (2020) used days to heading as a fixed effect in a multi-trait model with additional yield components in a panel of USA soft facultative wheat. The multi-trait predictions demonstrated higher predictive accuracy than the single-trait models under a multiple-environmental analysis showing its capacity to predict the performance of a genotype for different target environments. Similarly, Gill et al. (2021) used multi-trait, multi-environment genomic prediction which performed best for all agronomic traits in their study including days to heading. Other studies introduce environmental covariates in genomic prediction models to predict the performance of lines in new environments (Jarquín et al. 2014; Heslot et al. 2014; Malosetti et al. 2016; Ly et al. 2018).

11.5.3 Integrating Crop Modeling with Genome-Based Prediction, Phenomics, and Environments

Modular crop model development approaches (Jones et al. 2001) and the rapid advance of QTL analyses conducted for a vast number of populations under diverse environments have opened up opportunities to integrate these two methods (see Box 11.1). This integration allowed the addition, modification, and maintenance of new components, including more recently gene-based functions into process-based crop models (Hoogenboom et al. 2004; White 2009; Zheng et al. 2013; Chenu et al. 2018; Hammer et al. 2019; Robert et al. 2020; Tardieu et al. 2021; Oliveira et al. 2011; Hu et al. 2021; Boote et al. 2021; Cooper et al. 2021; Potgieter et al. 2021; Cowling et al. 2020; Wallach et al. 2018; Hwang et al. 2017; Yin et al. 2018). The first simple gene-based model was developed by White and Hoogenboom (1996) linking gene information with genotype-specific parameters (GSPs) for a drybean model called BEANGRO (Hoogenboom et al. 2019), where seven genes were used to estimate 19 parameters simulating data for 32 cultivars.

Box 11.1 Crop models simulating flowering time

Most crop models use similar approaches to simulate the crop life cycle, integrating development rate over time, usually assuming a potential development rate driven by temperature and modified by several other factors such as photoperiod, vernalization, and other abiotic stresses that may accelerate or delay crop development (Oliveira et al. 2011). The rate of development used in many crop models is a function of a triangular or trapezoidal shape driven by time (TT), or growing degree days (GDD), that are calculated based on maximum and minimum air temperature. The temperature response for wheat has a base temperaure (below which no development occurs) of approximately 0 ºC, optimal temperature (maximum development rate) of approximately 26 ºC, and a maximum temperature (above which no development occurs) of approximately 34 ºC (Hu et al. 2021; Boote et al. 2021). These air temperature thresholds and calculations could vary depending on the crop model used, as some research articles have shown that base and optimal temperature could change during the wheat life cycle, besides soil mean crown temperature being adjusted by snow depth (Hoogenboom et al. 2004; Boote et al. 2021).

The day length effect on crop development is accounted for by a photoperiod sensitivity factor which results in a daily percent reduction of development rate, below the threshold of 20 hours of daylength. The vernalization effect is computed as a function of a vernalization sensitivity factor, or maximum development rate to reach the threshold number of accumulated vernalization days required for a specific cultivar. Vernalization is also lost when daily maximum temperature is above 30 ºC. Vernalization and photoperiod factors are used to modify accumulation of thermal time from emergence to floral initiation (Hoogenboom et al. 2004; Hu et al. 2021; Boote et al. 2021; Cooper et al. 2021).

The process-based modelling approaches mentioned above have been used to predict development of wheat and many crops with good accuracy across many years, having as input other genotype-specific parameters (GSPs) besides weather, soil, and crop management variables (Potgieter et al. 2021). However, only recently have these models started to incorporate true genetic information to capture differences among cultivars instead of empirical GSPs created and calibrated based on processes and observations from field and laboratory studies (Boote et al. 2021; Cowling et al. 2020) even though the idea and first studies started in the late 1990s (Wallach et al. 2018; Hwang et al. 2017; Yin et al. 2018).

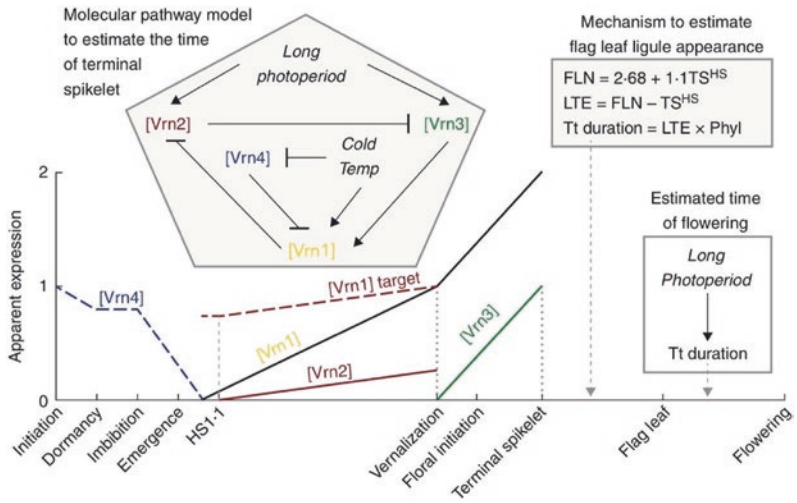


Fig. 11.6 Schematic representation of the integrated model. The crop must pass through each of the phases along the *x*-axis to reach anthesis. Temperature per se controls the progression through each phase in combination with the factors presented. Temperature and photoperiod control the expression of *VRN1*, *VRN2*, *VRN3*, and *VRN4* genes as demonstrated by the scheme within the pentagon (pointed arrows show promotion and flat arrows show repression) and subsequent amount of [Vrn1], [Vrn2], [Vrn3], and [Vrn4] protein expressed as demonstrated by the lines on the graph. The amount of these proteins controls the timing of vernalization and terminal spikelet (adapted from Brown et al. 2013)

Since then, there has been a rapid increase in the number of research studies including gene-based modeling applications, but most of them are still limited to crop phenology and other less complex traits. Brown et al. (2013) integrated molecular and physiological models to simulate time to anthesis using lines of spring and winter wheat under different temperature and photoperiod conditions. They linked the duration of phases to expressions of *VRN* genes to account for the effects of temperature during each developmental stage to develop a model (Fig. 11.6). This analysis framework was compared with CERES, ARCWHEAT1, and SIRIUS model approaches, suggesting the possibility of linking phenological parameters and anthesis time to the alleles or copy number of genes that control the expression of protein signals, relating anthesis genotype to phenotype.

Hu et al. (2021) used the APSIM wheat-G gene-based phenology model to identify the optimal flowering period of spring wheat and concluded that this type of model can identify the best combination of sowing dates and time to flowering to minimize frost and heat risk and achieve higher yields. Among the gene-based modeling applications, those that can be integrated with several other breeding tools have the greatest potential. Wang et al. (2019) reviewed necessary improvements for process-based crop models to simulate $G \times E \times M$ interactions and stated that the verification of temporal gene expression profiles, their environmental dependencies, and their expression levels are further required to trigger key phenological stages.

A growing body of research focuses on the benefits and challenges resulting from the integration of several modern technologies into breeding programs. This includes genomics using dense molecular markers, detailed trait analysis using advances in phenomics, image analyses, and the intense used of environmental covariables (environomics) and multi-trait analysis in order to accelerate genetic gains and

increase agricultural production (Crossa et al. 2019). Incorporating these newly available technologies, e.g., computer simulation for genomic-assisted rapid cycle population improvement, combining rapid genomic cycling with speed breeding, high-throughput phenotyping, and using historical climate and soil data, has potential to improve conventional breeding schemes. Integrating the machinery of crop modeling with that of genomic information and phenomics data together with environomics platforms can further increase the breeding efficiency. This in turn offers great promise to develop varieties rapidly since the selection of candidate individuals can be performed with higher accuracy.

There is evidence that crop models are useful for phenotypic prediction of relevant quantitative traits by simulating the behavior and growth of crops using solar radiation, water, nitrogen, etc., as input. Still, there is little empirical evidence that integration of this type of model with whole-genome prediction increases the prediction accuracy of unobserved cultivars. Two simulation studies (Technow et al. 2015; Messina et al. 2018) showed that integration to a combined model improved prediction accuracy relative to the genomic model alone.

Grain yield is the ultimate measure of crop adaptation due to phenology. Crop models can also be used for prediction of complex traits such as grain yield for different cultivars and location-year combinations within certain eco-geographical regions. It is necessary to incorporate the genetic variance of the traits and how these will change under different environmental conditions into the models. With the rapidly increasing availability of data on DNA sequences of individual cultivars or breeding lines, the use of crop models to improve crop model development and applications has been significantly fast. Similarly, advances in the understanding of the control of plant processes at the molecular level offer opportunities to strengthen how certain plant physiological mechanisms are incorporated into crop models.

It has been shown that crop models can be integrated with genomic prediction to enhance prediction accuracy using simulation data. For example, Heslot et al. (2014) employed crop models to derive stress covariates from daily weather data for predicted crop development stages, by means of the factorial regression model to genomic selection modeling of QTL × environment interaction on a genome-wide scale. The method was tested using a winter wheat dataset, and accuracy in predicting genotype performance in unobserved environments for which weather data were available increased by 11.1% on average. Furthermore, Cooper et al. (2016) used crop models with genomic-enabled prediction applied to an empirical maize drought data set. These authors found positive prediction accuracy for hybrid grain yield in two drought environments.

In general, crop models have been used for crop management decision support. The presence of G × E × M interactions for yield presents challenges for the development of prediction technologies for product development by breeding and product placement for different agricultural production systems. Messina et al. (2018) combined simulation and empirical studies to show how to use CGM with genome-enabled methodology for the application to maize breeding and product placement recommendation in the US corn-belt.

In plant breeding, genetic and environmental factors can interact in complex ways giving rise to substantial G × E interactions that can be used to select genotypes adapted to specific environments. Nevertheless, accurate predictions of future performances in environments are challenging and it requires consideration of the possible weather conditions that may occur within a region and how individual genotypes are expected to react to those conditions. Usually, METs occurring over many years and across multiple locations are utilized to facilitate such predictions. The major challenge is that MET is organized over few years and locations such that genotypes are often advanced without being tested under weather conditions that may critically affect their performance. To overcome this limited scope of the MET, de los Campos et al. (2020) proposed data-driven computer simulations that integrate field trial data, DNA

sequences, and historical weather records for predicting genotype performances and stability using limited years of field testing per genotype.

The data-driven simulation proposed by de los Campos et al. (2020) links modern genomic models that integrate DNA sequences (e.g., single nucleotide polymorphisms—SNPs) and environmental covariates (EC; Jarquín et al. 2014; Crossa et al. 2019) by means of Monte Carlo methods that integrate uncertainty about future weather conditions as well as model parameters (characterized using the their posterior distribution). The importance of this approach is to study ECs as a mechanism to characterize the environmental conditions prevailing during crop growing seasons on the current MET location-year but also in the past field trial data with historical (or simulated) weather records that describe environmental conditions that are likely to occur in a location or region. The results of de los Campos et al. (2020) results show that (1) it is possible to predict the performance of cultivar at environments where these cultivars have few (or none) testing data and (2) predictions that incorporate historical weather records are more robust with respect to year-to-year variation in environmental conditions than the ones that can be derived using only few field trials.

Further research is needed to add evidence that crop modeling together with genomic-enabled predictions can be of benefit in plant breeding together with phenomics and environomics. Three proposed directions for future research are: (a) to use historical data to complement the advantages of crop modeling with those of genomics and phenomics; (b) to conduct more simulation studies with different type of crop models, genomics, and phenomics models and (c) to conduct real experiments where the scientist can control the input of the crop model and measure as accurate as possible the output. Simulation studies should be conducted to benchmark the prediction performance of combined models (crop model + genomics) compared to stand-alone genomic prediction models. Comparing combinations of different types of crop and genomics models, which include random effects for G × E interaction terms, would be useful. New deep learning models that have been developed for dealing with big data sets should also be considered for incorporation with crop models for multi-trait, multi-environment predictions (Montesinos-Lopez et al. 2018, 2019).

11.6 Future Opportunities

Since its first cultivation in 7000 BC, hexaploid wheat has evolved and adapted, enabling expansion underpinning global food security. Adaptive genes (and their complex interactions) have played an important role in optimizing wheat production and will continue to play a significant role in fine-tuning flowering and reproductive cycles suited to changing climates and evolving agricultural production systems. As documented in this chapter, many QTL have been detected with robust effects within and across environments which have expanded the breadth of adaptive variation to be explored in future. However, additional work is required to identify underlying genes and dissect pathways to understand their mode of action and accelerate their validation and deployment in breeding. Likewise, MQTL can help to identify relevant genomic regions over space and time and facilitate the identification of new candidate genes. In the QTL comparison conducted here, we detected seven MQTL regions on chromosomes 1D, 2A, 2B, 3B, 5A, 5B, and 6A. While five MQTL were co-located with known flowering genes regions, candidate genes for two MQTL are not yet known. The identification of genes underpinning the two robust MQTL regions on chromosomes 3B and 6A and those identified to be in homoeologous regions (proximal on 5 and distal on chromosomes 5, 6, and 7) will offer new potential targets for exploitation. These genetic dissection efforts will be greatly aided by current and future developments in wheat genome sequencing and characterization of haplotypes across the wheat and progenitor pangenomes.

Moving beyond the identification of flowering time loci, it will also become increasingly important to understand how genetic regions influence other developmental traits and their responses to environmental factors. This depth of understanding will allow a more targeted "design" of adaptive ideotypes to suit current and future climates and is likely to influence the use of novel breeding methods. For example, the timing of flowering and regulation of distinct flowering stages may influence the efficiency of hybrid wheat seed production, supporting the development of mainstream hybrids. Similarly, genomic selection network approaches that can include multiple traits along with flowering time are likely to be useful in identifying high-performing, optimally adapted lines for breeding, selection, and release. Finally, much future potential exists in applying recently developed integrated genomics and crop modeling approaches. The advances with gene-based modeling in the future, if successful, should make it possible to describe growth and development processes with QTL and other genomic loci analysis, integrated in process-based crop models in a modular approach. This would potentially reduce the need for crop modeling calibration using phenotypic data after new cultivars are released to assess their response to genotype, environment, and management ($G \times E \times M$) conditions. This can both leverage extensive historical data (available in many breeding programs) to identify previously hidden environmental "clues" as well as providing novel targets for the design and deployment of further climate change adaptation strategies.

Acknowledgements This work was supported by the CGIAR Research Program WHEAT (CRP-WHEAT), Accelerating Genetics Gains in Maize and Wheat (AGG) and HeDWIC to CIMMYT, the Farrer Memorial Trust to Jessica Hyles, and the UKRI FLF MR/S031677/1 and the Rank Prize Funds New Lecturer Award to Laura Dixon. Funding for WHEAT comes from CGIAR and national governments, foundations, development banks, and other public and private agencies, in particular the Australian Centre for International Agricultural Research (ACIAR), the UK Foreign Commonwealth & Development Office (FCDO), and the USA Agency for International Development (USAID). AGG funding comes from the Bill and Melinda Gates Foundation, FCDO, USAID, and the Foundation for Food and Agriculture Research (FFAR). HeDWIC is supported by FFAR, CRP-WHEAT, the Sustainable Modernization of Traditional Agriculture (MasAgro) initiative from the Secretariat of Agriculture and Rural Development (SADER, Government of Mexico), AGG, FCDO, and USAID.

References[1]

Aitken Y (1974) Flowering time, climate and genotype. Melbourne University Press. ISBN 052284071X, 9780522840711

Alvarez MA, Tranquilli G, Lewis S, Kippes N, Dubcovsky J (2016) Genetic and physical mapping of the earliness per se locus Eps-A m 1 in Triticum monococcum identifies EARLY FLOWERING 3 (ELF3) as a candidate gene. Funct Integr Genomics 16:365–382

Andrés F, Coupland G (2012) The genetic basis of flowering responses to seasonal cues. Nat Rev Genet 13:627–639

Arjona JM, Royo C, Dreisigacker S, Ammar K, Subirà J, Villegas D (2020) Effect of allele combinations at Ppd-1 loci on durum wheat grain filling at contrasting latitudes. J Agron Crop Sci. https://doi.org/10.1111/jac.12363

Asseng S, Evert F, Martre P, Roetter RP, Lobell DB, Cammarano D (2015) Rising temperatures reduce global wheat production. Nat Clim Change 5:143–147

Basavaraddi PA, Savin R, Bencivenga S, Griffiths S, Slafer GA (2021a) Wheat developmental traits as affected by the interaction between eps-7d and temperature under contrasting photoperiods with insensitive ppd-d1 background. Plants 10:1–16

Basavaraddi PA, Savin R, Wingen LU, Bencivenga S, Przewieslik-Allen AM, Griffiths S, Slafer GA (2021b) Interactions between two QTLs for time to anthesis on spike development and fertility in wheat. Sci Rep 11:1–16

Beales J, Turner A, Griffiths S, Snape JW, Laurie DA (2007) A pseudo-response regulator is misexpressed in the photoperiod insensitive Ppd-D1a mutant of wheat (*Triticum aestivum* L.). Theor Appl Genet 115:721–733

Bennett T, Dixon LE (2021) Asymmetric expansions of FT and TFL1 lineages characterize differential evolution of the EuPEBP family in the major angiosperm lineages. BMC Biol 19:1–17

[1] Identifies references cited in Supplementary Table S11.1.

Bloomfield M, Hunt J, Trevaskis B, Ramm K, Hyles J (2019) Can allele variation at PPD1 and VRN1 gene loci predict flowering time in wheat under controlled conditions? Proc 2019 Agron Aust Conf 1:1–4

Boden SA, Cavanagh C, Cullis BR, Ramm K, Greenwood J, Jean Finnegan E, Trevaskis B, Swain SM (2015) Ppd-1 is a key regulator of inflorescence architecture and paired spikelet development in wheat. Nat Plants 1:1–6

Boote KJ, Jones JW, Hoogenboom G (2021) Incorporating realistic trait physiology into crop growth models to support genetic improvement vol 3, pp 1–19

Börner A, Worland AJ, Plaschke J, Schumann E, Law CN (1993) Pleiotropic effects of genes for reduced height (Rht) and day-length insensitivity (Ppd) on yield and its components for wheat grown in middle Europe. Plant Breed 111:204–216

Brown HE, Jamieson PD, Brooking IR, Moot DJ, Huth NI (2013) Integration of molecular and physiological models to explain time of anthesis in wheat. Ann Bot 112:1683–1703

Burgueño J, de los Campos G, Weigel K, Crossa J (2012) Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. Crop Sci 52:707–719

Cane K, Eagles HA, Laurie DA, Trevaskis B, Vallance N, Eastwood RF, Gororo NN, Kuchel H, Martin PJ (2013) Ppd-B1 and Ppd-D1 and their effects in southern Australian wheat. Crop Pasture Sci 64:100–114

Cann DJ, Schillinger WF, Hunt JR, Porker KD, Harris FAJ (2020) Agroecological advantages of early-sown winter wheat in semi-arid environments: a comparative case study from Southern Australia and Pacific Northwest United States. Front Plant Sci. https://doi.org/10.3389/fpls.2020.00568

Charmet G, Storlie E, Oury FX et al (2014) Genome-wide prediction of three important traits in bread wheat. Mol Breed 34:1843–1852

Chen Y, Carver BF, Wang S, Zhang F, Yan L (2009) Genetic loci associated with stem elongation and winter dormancy release in wheat. Theor Appl Genet 118:881–889

Chen C, Fletcher AL, Ota N, Flohr BM, Lilley JM, Lawes RA (2020) Spatial patterns of estimated optimal flowering period of wheat across the southwest of Western Australia. F Crop Res 247:107710

Chenu K, Oosterom EJ Van, Mclean G, Deifel KS, Fletcher A, Geetika G, Tirfessa A (2018) Integrating modelling and phenotyping approaches to identify and screen complex traits: transpiration efficiency in cereals, vol 69, pp 3181–3194

Cheung C, Zhang H, Hepburn JC, Yang DY, Richards MP (2019) Stable isotope and dental caries data reveal abrupt changes in subsistence economy in ancient China in response to global climate change. PLoS ONE 14:1–27

Chinoy J (1950) Effect of vernalization and photoperiod treatments on growth and development of wheat. Nature 165:882–883

Chouard P (1960) Vernalization and its relations to dormancy. Annu Rev Physiol 11:191–238

Christy B, Riffkin P, Richards R, Partington D, Acuña TB, Merry A, Zhang H, Trevaskis B, O'Leary G (2020) An allelic based phenological model to predict phasic development of wheat (*Triticum aestivum* L.). F Crop Res 249:107722

Cockram J, Jones H, Leigh FJ, O'Sullivan D, Powell W, Laurie DA, Greenland AJ (2007) Control of flowering time in temperate cereals: genes, domestication, and sustainable productivity. J Exp Bot 58:1231–1244

Cooper M, Technow F, Messina C, Gho C, Totir LR (2016) Use of crop growth models with whole-genome prediction: application to a maize multienvironment trial, pp 2141–2156

Cooper M, Powell O, Voss-Fels KP, Messina CD, Gho C, Podlich DW, Technow F, Chapman SC, Beveridge CA, Ortiz-Barrientos D, Hammer GL (2021) Modelling selection response in plant-breeding programs using crop models as mechanistic gene-to-phenotype (CGM-G2P) multi-trait link functions. Plants 3:1–21. https://doi.org/10.1093/insilicoplants/diaa016

Cowling WA, Gaynor RC, Antolín R, Gorjanc G, Edwards SM, Powell O, Hickey JM (2020) In silico simulation of future hybrid performance to evaluate heterotic pool formation in a self-pollinating crop. Sci Rep 10:1–8

Crossa J, de Los CG, Pérez P et al (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186:713–724

Crossa J, Pérez P, Hickey J et al (2014) Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity (edinb) 112:48–60

Crossa J, Jarquín D, Franco J et al (2016) Genomic prediction of gene bank wheat landraces. G3 Genes Genomes Genet 6:1819–1834

Crossa J, Pérez-Rodríguez P, Cuevas J et al (2017) Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci 22:961–975

Crossa J, Martini JWR, Gianola D, Pérez-Rodríguez P, Jarquin D, Juliana P, Montesinos-López O, Cuevas J (2019) Deep kernel and deep learning for genome-based prediction of single traits in multienvironment breeding trials. Front Genet 10:1–13

Curtis B (2002) Wheat in the world. In: Curtis B, Rajaram S, Macpherson G (eds) Bread wheat improved products food and agriculture organization of the United Nations Plant Protection Series, Rome, pp 1–17

de Los CG, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, Weigel K, Cotes JM (2009) Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182:375–385

de los Campos G, Perez-Rodriguez P, Bogard M, Gouache D, Crossa J (2020) A data-driven simulation platform to predict cultivarsâ€TM performances under uncertain weather conditions. Nat Commun 11:4876

Demirhan H (2020) Impact of increasing temperature anomalies and carbon dioxide emissions on wheat production. Sci Total Environ. https://doi.org/10.1016/j.scitotenv.2020.139616

Díaz A, Zikhali M, Turner AS, Isaac P, Laurie DA (2012) Copy number variation affecting the photoperiod-B1 and vernalization-A1 genes is associated with altered flowering time in wheat (*Triticum aestivum*). PLoS ONE. https://doi.org/10.1371/journal.pone.0033234

Distelfeld A, Tranquilli G, Li C, Yan L, Dubcovsky J (2009) Genetic and molecular characterization of the VRN2 loci in tetraploid wheat. Plant Physiol 149:245–257

Dixon LE, Karsai I, Kiss T, Adamski NM, Liu Z, Ding Y, Allard V, Boden SA, Griffiths S (2019) VERNALIZATION1 controls developmental responses of winter wheat under high ambient temperatures. https://doi.org/10.1242/dev.172684

Dreisigacker S, Burgueño J, Pacheco A, Molero G, Sukumaran S, Rivera-amado C, Reynolds M, Griffiths S (2021a) Effect of flowering time-related genes on biomass, harvest index, and grain yield in CIMMYT elite spring bread wheat susanne. Biology (Basel). https://doi.org/10.3390/biology10090855

Dreisigacker S, Crossa J, Pérez-rodríguez P et al (2021b) Implementation of genomic selection in the CIMMYT global wheat program, findings from the past 10 years. Crop Breeding Genet Genomics 3:e210005

Dubcovsky J, Loukoianov A, Bonafede M (2005) Regulation of flowering time in wheat and barley permalink. Comp Biochem Physiol A Mol Integr Physiol 141:263–264

Dubcovsky J, Li C, Pidal B, Tranquilli G (2008) Genes and gene networks regulating wheat development. In: 11th international wheat genetics symposium

Eagles HA, Cane K, Vallance N (2009) The flow of alleles of important photoperiod and vernalisation genes through Australian wheat. Crop Pasture Sci 60:646–657

Eagles HA, Cane K, Trevaskis B (2011) Veery wheats carry an allele of Vrn-A1 that has implications for freezing tolerance in winter wheats. Plant Breed 130(130):413–418

Flohr BM, Hunt JR, Kirkegaard JA, Evans JR (2017) Water and temperature stress define the optimal flowering period for wheat in south-eastern Australia. F Crop Res 209:108–119

Ford B, Deng W, Clausen J, Oliver S, Boden S, Hemming M, Trevaskis B (2016) Barley (Hordeum vulgare) circadian clock genes can respond rapidly to temperature in an EARLY FLOWERING 3-dependent manner. J Exp Bot 67:5517–5528

Fu Y-B, Dong Y-B (2015) Genetic erosion under modern plant breeding: case studies in Canadian crop gene pools, pp 89–104

Garner W, Allard H (1920) Effect of the relative length of day and night and other factors of the environment on growth and reproduction in plants. J Agric Res 18:553–603

Gauley A, Boden SA (2021) Stepwise increases in FT1 expression regulate seasonal progression of flowering in wheat (*Triticum aestivum*). New Phytol 229:1163–1176

Gawroński P, Ariyadasa R, Himmelbach A et al (2014) A distorted circadian clock causes early flowering and temperature-dependent variation in spike development in the Eps-3Am mutant of einkorn wheat. Genetics 196:1253–1261

Gayon J, Zallen DT (1998) The role of the Vilmorin company in the promotion and diffusion of the experimental science of heredity in France, 1840–1920. J Hist Biol 31:241–262

Gill HS, Halder J, Zhang J et al (2021) Multi-trait multi-environment genomic prediction of agronomic traits in advanced breeding lines of winter wheat. Front Plant Sci 12:1–14

Griffiths S, Simmonds J, Leverington M et al (2009) Meta-QTL analysis of the genetic control of ear emergence in elite European winter wheat germplasm. Theor Appl Genet 119:383–395

Guo J, Khan J, Pradhan S et al (2020) Multi-trait genomic prediction of yield-related traits in US soft wheat under variable water regimes. Genes (Basel) 11

Guthrie F (1922) William J. Farrer, and the results of his workcompiled by F.B. Guthrie, at the request of the Trustees of the Farrer Memorial Fund. Department of Agriculture, Sydney

Haile TA, Walkowiak S, N'Diaye A, Clarke JM, Hucl PJ, Cuthbert RD, Knox RE, Pozniak CJ (2021) Genomic prediction of agronomic traits in wheat using different models and cross-validation designs. Theor Appl Genet 134:381–398

Halliwell J, Borrill P, Gordon A, Kowalczyk R, Pagano ML, Saccomanno B, Bentley AR, Uauy C, Cockram J (2016) Systematic investigation of FLOWERING LOCUS T-like poaceae gene families identifies the short-day expressed flowering pathway gene, TaFT3 in wheat (*Triticum aestivum* L.). Front Plant Sci 7:1–15

Hammer G, Messina C, Wu A, Cooper M (2019) Opinion biological reality and parsimony in crop models—why we need both in crop improvement! pp 1–21

Hammer GL, McLean G, van Oosterom E, Chapman S, Zheng B, Wu A, Doherty A, Jordan D (2020) Designing crops for adaptation to the drought and high-temperature risks anticipated in future climates. Crop Sci 60:605–621

Heslot N, Akdemir D, Sorrells ME, Jannink JL (2014) Integrating environmental covariates and crop

modeling into the genomic selection framework to predict genotype by environment interactions. Theor Appl Genet 127:463–480

Hillman GC (1972) Papers in economic history. University Press

Honma T, Goto K (2001) Complexes of MADS-box proteins are sufficient to convert leaves into floral organs. Nature 409:525–529

Hoogenboom G, White JW, Messina CD (2004) From genome to crop: integration through simulation modeling, vol 90, pp 145–163

Hoogenboom G, Porter CH, Boote KJ et al (2019) The DSSAT crop modeling ecosystem. In: Boote KJ (ed) Advances in crop modelling for a sustainable agriculture. Burleigh Dodds Science Publishing, pp 173–216

Hoogendoorn J (1985) The physiology of variation in the time of ear emergence among wheat varieties from different regions of the world. Euphytica 34:559–571

Hu P, Chapman SC, Dreisigacker S, Sukumaran S, Reynolds MP, Zheng B (2021) Using a gene-based phenology model to identify optimal flowering periods of spring wheat in irrigated mega-environments. J Exp Bot erab326

Hunt L (1979) Photoperiodic responses of winter wheats from different climatic regions. J Plant Breed 82:70–80

Hunt JR (2017) Winter wheat cultivars in Australian farming systems: a review. Crop Pasture Sci 68:501–515

Hunt JR, Lilley JM, Trevaskis B, Flohr BM, Peake A, Fletcher A, Zwart AB, Gobbett D, Kirkegaard JA (2019) Early sowing systems can boost Australian wheat yields despite recent climate change. Nat Clim Change 9:244–247

Hwang C, Correll MJ, Gezan SA, Zhang L, Bhakta MS, Vallejos CE, Boote KJ, Clavijo-Michelangeli JA, Hyles J, Bloom MT, Hunt JR, Trethowan RM, Trevaskis B (2020) Phenology and related traits for wheat adaptation. https://doi.org/10.1038/s41437-020-0320-1

Jamil M, Ali A, Gul A, Ghafoor A, Napar AA, Ibrahim AMH, Naveed NH, Yasin NA, Mujeeb-Kazi A (2019)[s] Genome-wide association studies of seven agronomic traits under two sowing conditions in bread wheat. BMC Plant Biol 19:1–18

Jarquín D, Crossa J, Lacaze X et al (2014) A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theor Appl Genet 127:595–607

Jones JW, Keating BA, Porter CH (2001) Approaches to modular model development, vol 70, pp 421–443

Juliana P, Poland J, Huerta-Espino J et al (2019)[s] Improving grain yield, stress resilience and quality of bread wheat using large-scale genomics. Nat Genet 51(10):1530–1539

Kamran A, Iqbal M, Navabi A, Randhawa H, Pozniak C, Spaner D (2013)[s] Earliness per se QTLs and their interaction with the photoperiod insensitive allele Ppd-D1a in the Cutler × AC Barrie spring wheat population. Theor Appl Genet 126:1965–1976

Kane NA, Agharbaoui Z, Diallo AO, Adam H, Tominaga Y, Ouellet F, Sarhan F (2007) TaVRT2 represses transcription of the wheat vernalization gene TaVRN1. Plant J 51:670–680. https://doi.org/10.1111/j.1365-313X.2007.03172.x

Kato K, Yanagata H (1988) Method for evaluation of chilling requirement and narrow-sense earliness of wheat cultivars. Jpn J Breed 38:172–186

Kippes N, Debernardi J, Vasquez-Gross HA, Akpinar BA, Budak H, Kato K, Chao S, Akhunov E, Dubcovsky J (2015) Identification of the VERNALIZATION 4 gene reveals the origin of spring growth habit in ancient wheats from South Asia. Proc Natl Acad Sci USA 112:E5401–E5410

Kippes N, Guedira M, Lin L, Alvarez MA, Brown-Guedira GL, Dubcovsky J (2018) Single nucleotide polymorphisms in a regulatory site of VRN-A1 first intron are associated with differences in vernalization requirement in winter wheat. Mol Genet Genomics 293:1231–1243

Kitagawa S, Shimada S, Murai K (2012) Effect of Ppd-1 on the expression of flowering-time genes in vegetative and reproductive growth stages of wheat. Genes Genet Syst 87:161–168

Law C, Sutka J, Worland A (1978) A genetic study of day-length response in Wehat. Heredity (edinb) 41:185–191

Li X, Dodson J, Zhou X, Zhang H, Masutomoto R (2007) Early cultivated wheat and broadening of agriculture in Neolithic China. Holocene 17:555–560

Li G, Yu M, Fang T, Cao S, Carver BF, Yan L (2013) Vernalization requirement duration in winter wheat is controlled by TaVRN-A1 at the protein level. Plant J 76:742–753

Li C, Lin H, Chen A, Lau M, Jernstedt J, Dubcovsky J (2019) Wheat VRN1, FUL2 and FUL3 play critical and redundant roles in spikelet development and spike determinacy. Development 146:1–11

Lin F, Xue SL, Tian DG, Li CJ, Cao Y, Zhang ZZ, Zhang CQ, Ma ZQ (2008)[s] Mapping chromosomal regions affecting flowering time in a spring wheat RIL population. Euphytica 164:769–777

Linné C, Freer S (2007) Linnaeus' Philosophia Botanica. Oxford University Press

Liu C, Pinto F, Cossani CM, Sukumaran S, Reynolds MP (2019a)[s] Spectral reflectance indices as proxies for yield potential and heat stress tolerance in spring wheat: heritability estimates and marker-trait associations. Front Agric Sci Eng 6:296–308

Liu J, Wu B, Singh RP, Velu G (2019b)[s] QTL mapping for micronutrients concentration and yield component traits in a hexaploid wheat mapping population. J Cereal Sci 88:57–64

Liu C, Sukumaran S, Jarquin D, Crossa J, Dreisigacker S, Sansaloni C, Reynolds M (2020) Comparison of array- and sequencing-based markers for

genome-wide association mapping and genomic prediction in spring wheat. Crop Sci 60:211–225

Lopes MS, Dreisigacker S, Peña RJ, Sukumaran S, Reynolds MP (2015)[s] Genetic characterization of the wheat association mapping initiative (WAMI) panel for dissection of complex traits in spring wheat. Theor Appl Genet 128:453–464

Lopez-Cruz M, Crossa J, Bonnett D, Dreisigacker S, Poland J, Jannink J-L, Singh RP, Autrique E, de los Campos G (2015) Increased prediction accuracy in wheat breeding trials using a marker × environment interaction genomic selection model. G3 Genes Genomes Genet 5:569–582. https://doi.org/10.1534/g3.114.016097

Lupton FGH (1987) History of wheat breeding. In: Lupton FGH (ed) Wheat breed. Chapman and Hall Ltd., pp 51–70

Lv B, Nitcher R, Han X, Wang S, Ni F, Li K, Pearce S, Wu J, Dubcovsky J, Fu D (2014) Characterization of flowering locus T1 (FT1) gene in Brachypodium and wheat. PLoS One. https://doi.org/10.1371/journal.pone.0094171

Ly D, Huet S, Gauffreteau A et al (2018) Whole-genome prediction of reaction norms to environmental stress in bread wheat (*Triticum aestivum* L.) by genomic random regression. F Crop Res 216:32–41

Malosetti M, Bustos-Korts D, Boer MP, Van Eeuwijk FA (2016) Predicting responses in multiple environments: issues in relation to genotype × environment interactions. Crop Sci 56:2210–2222

Martinic ZF (1975) Life cycle of common wheat varieties in natural environments as related to their response to shortened photoperiod. Z Pflanzenzuchtung 75:237–251

McCallum BD, DePauw RM (2008) A review of wheat cultivars grown in the Canadian prairies. Can J Plant Sci 88:649–677

Messina CD, Technow F, Tang T, Totir R, Gho R, Cooper M (2018) Leveraging biological insight and environmental variation to improve phenotypic 2 prediction: integrating crop growth models (CGM) with whole genome prediction (WGP). Eur J Agron 100:151–162

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829

Mizuno T, Nakamichi N (2005) Pseudo-response regulators (PRRs) or true oscillator components (TOCs). Plant Cell Physiol 46:677–685

Molero G, Joynson R, Pinera-Chavez FJ, Gardiner L, Rivera-Amado C, Hall A, Reynolds MP (2019)[s] Elucidating the genetic basis of biomass accumulation and radiation use efficiency in spring wheat and its role in yield potential. Plant Biotechnol J 1–13

Montesinos-Lopez OA, Montesinos-López A, Crossa J, Gianola D, Hernández-Suárez CM, Martín-Vallejo J (2018) Multi-trait, multi-environment deep learning modeling for genomic-enabled prediction of plant traits. G3 Genes Genomes Genetics g3.200728.2018

Montesinos-López OA, Martín-Vallejo J, Crossa J, Gianola D, Hernández-Suárez CM, Montesinos-López A, Juliana P, Singh R (2019) A benchmarking between deep learning, support vector machine and Bayesian threshold best linear unbiased prediction for predicting ordinal traits in plant breeding. G3 Genes Genomes Genet 9:601–618

Mulki MA, Bi X, von Korff M (2018) Flowering locus T3 controls spikelet initiation but not floral development. Plant Physiol 178:1170–1186

Muterko A, Salina E (2018) Origin and distribution of the VRN-A1 exon 4 and exon 7 haplotypes in domesticated wheat species. Agronomy 8:1–14

Muterko A, Kalendar R, Salina E (2016) Novel alleles of the VERNALIZATION1 genes in wheat are associated with modulation of DNA curvature and flexibility in the promoter region. BMC Plant Biol. https://doi.org/10.1186/s12870-015-0691-2

Nguyen AT, Iehisa JCM, Kajimura T, Murai K, Takumi S (2013)[s] Identification of quantitative trait loci for flowering-related traits in the D genome of synthetic hexaploid wheat lines. Euphytica 192:401–412

Nishida H, Yoshida T, Kawakami K, Fujita M, Long B, Akashi Y, Laurie DA, Kato K (2013) Structural variation in the 5′ upstream region of photoperiod-insensitive alleles Ppd-A1a and Ppd-B1a identified in hexaploid wheat (*Triticum aestivum* L.), and their effect on heading time. Mol Breed 31:27–37

Ochagavía H, Prieto P, Savin R, Griffiths S, Slafer G (2018) Dynamics of leaf and spikelet primordia initiation in wheat as affected by Ppd-1a alleles under field conditions. J Exp Bot 69:2621–2631

Ochagavía H, Prieto P, Zikhali M, Griffiths S, Slafer GA (2019) Earliness per se by temperature interaction on wheat development. Sci Rep 9:1–11

Oliveira FAA, Jones JW, Pavan W, Bhakta M, Vallejos CE, Correll MJ, Boote KJ, Fernandes JMC, Hölbig Olmstead AL, Rhode PW (2011) Adapting North American wheat production to climatic challenges, pp 1839–2009. https://doi.org/10.1073/pnas.1008279108/,/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1008279108

Ortiz Ferrara G, Mosaad MG, Mahalakshmi V, Rajaram S (1998) Photoperiod and vernalisation response of mediterranean wheats, and implications for adaptation. Euphytica 100:377–384

Pánková K, Milec Z, Simmonds J, Leverington-Waite M, Fish L, Snape JW (2008)[s] Genetic mapping of a new flowering time gene on chromosome 3B of wheat. Euphytica 164:779–787

Pequeno DNL, Hernandez-Ochoa IM, Reynolds MP, Sonder K, Molero Milan A, Robertson RD, Lopes MS, Xiong W, Kropff M, Asseng S (2021) Climate impact and adaptation to heat and drought stress of regional and global wheat production. Environ Res Lett 16:054070

Pérez-Rodríguez P, Crossa J, Rutkoski J, Poland J, Singh R, Legarra A, Autrique E, Campos G de

los, Burgueño J, Dreisigacker S (2017) Single-step genomic and pedigree genotype × environment interaction models for predicting wheat lines in international environments. Plant Genome 10:plantgenome2016.09.0089

Pieper R, Tomé F, Pankin A, Von Korff M (2021) FLOWERING LOCUS T4 delays flowering and decreases floret fertility in barley. J Exp Bot 72:107–121

Pinto RS, Lopes MS, Collins NC, Reynolds MP (2016)[s] Modelling and genetic dissection of staygreen under heat stress. Theor Appl Genet 129(11):2055–2074

Porsche W, Taylor M (2001) German wheat pool. In: Bonjean A, Angus WJ (eds) The world wheat book. A history of wheat breeding. Intercept, pp 167–191

Potgieter AB, Zhao Y, Zarco-tejada PJ et al (2021) Evolution and application of digital technologies to predict crop type and crop phenology in agriculture, vol 3, pp 1–23

Prieto P, Ochagavía H, Griffiths S, Slafer GA (2020) Earliness per se×temperature interaction: consequences on leaf, spikelet, and floret development in wheat. J Exp Bot 71:1956–1968

Pugsley A (1963) The inheritance of a vernalization response in Australian spring wheats. Austr J Agric Res 14:622–626

Pugsley A (1965) Inheritance of a correlated day-length response in spring wheat. Nature 207:108

Putterill J, Robson F, Lee K, Simon R, Coupland G (1995) The CONSTANS gene of arabidopsis promotes flowering and encodes a protein showing similarities to zinc finger transcription factors. Cell 80:847–857

Ramirez-Villegas J, Molero Milan A, Alexandrov N et al (2020) CGIAR modeling approaches for resource-constrained scenarios: I. Accelerating crop breeding for a changing climate. Crop Sci 60:547–567

Réaumur RAF (1735) Observations du thermometre, faites a Paris pendant I'annee 1735, compares avec celles qui ont ete faites sous la ligne, a l'lsle de France, a Alger et en quelques-unes de nos isles de I' Amerique. Mem Acad des Sci

Renfrew JM (1973) Paleoethnobotany. The prehistoric food plants of the near East and Europe. Columbia University Press

Robert P, Le Gouis J, Rincent R (2020) Combining crop growth modeling with trait-assisted prediction improved the prediction of genotype by environment interactions. Front Plant Sci 11:1–11

Robson F, Costa MMR, Hepworth SR, Vizir I, Piñeiro M, Reeves PH, Putterill J, Coupland G (2001) Functional importance of conserved domains in the flowering-time gene CONSTANS demonstrated by analysis of mutant alleles and transgenic plants. Plant J 28:619–631

Royo C, Dreisigacker S, Alfaro C, Ammar K, Villegas D (2016) Effect of Ppd-1 genes on durum wheat flowering time and grain filling duration in a wide range of latitudes. J Agric Sci 154:612–631

Royo C, Ammar K, Alfaro C, Dreisigacker S, Fernando L, Villegas D (2018) Field crops research effect of Ppd-1 photoperiod sensitivity genes on dry matter production and allocation in durum wheat. Field Crop Res 221:358–367

Salamini F, Özkan H, Brandolini A, Schäfer-Pregl R, Martin W (2002) Genetics and geography of wild cereal domestication in the near east. Nat Rev 3:429–441

Searle I, Coupland G (2004) Induction of flowering by seasonal changes in photoperiod. EMBO J 23:1217–1222

Semagn K, Iqbal M, Chen H et al (2021)[s] Physical mapping of QTL associated with agronomic and end-use quality traits in spring wheat under conventional and organic management systems. Theor Appl Genet 134:3699–3719

Shaw LM, Turner AS, Herry L, Griffiths S, Laurie DA (2013) Mutant alleles of Photoperiod-1 in wheat (*Triticum aestivum* L.) that confer a late flowering phenotype in long days. PLoS One. https://doi.org/10.1371/journal.pone.0079459

Sheehan H, Bentley A (2020) Changing times: opportunities for altering winter wheat phenology, pp 1–11

Strayer C, Oyama T, Schultz TF, Raman R, Somers DE, Mas P, Panda S, Kreps JA, Kay SA (2000) Cloning of the Arabidopsis clock gene TOC1, an autoregulatory response regulator homolog. Science (80-) 289:768–771

Sukumaran S, Dreisigacker S, Lopes M, Chavez P, Reynolds MP (2015)[s] Genome-wide association study for grain yield and related traits in an elite spring wheat population grown in temperate irrigated environments. Theor Appl Genet 128:353–363

Sukumaran S, Lopes MS, Dreisigacker S, Dixon LE, Zikhali M, Griffiths S, Zheng B, Chapman S, Reynolds MP (2016)[s] Identification of earliness per se flowering time locus in spring wheat through a genome-wide association study. Crop Sci 56:2962–2972

Tanaka A, Takahashi K, Masutomi Y, Hanasaki N, Hijioka Y, Shiogama H, Yamanaka Y (2015) Adaptation pathways of global wheat production: importance of strategic adaptation to climate change. Sci Rep 5:2–11

Tanio M, Kato K (2007) Development of near-isogenic lines for photoperiod-insensitive genes, Ppd-B1 and Ppd-D1, carried by the Japanese wheat cultivars and their effect on apical development. Breed Sci 57:65–72

Tardieu F, Granato ISC, Van Oosterom EJ, Parent B, Hammer GL, Lepse I, De Montpellier U (2021) Are crop and detailed physiological models equally 'mechanistic' for predicting the genetic variability of whole-plant behaviour ? The nexus between mechanisms and adaptive strategies. Silico Plants 2:1–12

Technow F, Messina CD, Totir LR, Cooper M (2015) Integrating crop growth models with whole genome prediction through approximate Bayesian computation. PLoS ONE 10:e0130855

Theißen G, Melzer R, Rümpler F (2016) MADS-domain transcription factors and the floral quartet model of flower development: linking plant development and evolution. Development 143:3259–3271

Trethowan RM, Reynolds MP, Ortiz-Monasterio JI, Ortiz R (2007) The genetic basis of the green revolution in wheat production. Plant Breed Rev 28:39–58

Trevaskis B, Bagnall DJ, Ellis MH, Peacock WJ, Dennis ES (2003) MADS box genes control vernalization-induced flowering in cereals. Proc Natl Acad Sci U S A 100:13099–13104

Turuspekov Y, Baibulatova A, Yermekbayev K, Tokhetova L, Chudinov V, Sereda G, Ganal M, Griffiths S, Abugalieva S (2017)[s] GWAS for plant growth stages and yield components in spring wheat (*Triticum aestivum* L.) harvested in three regions of Kazakhstan. BMC Plant Biol 17(1):1–11

Van Beem J, Mohler V, Lukman R, Van Ginkel M, William M, Crossa J, Worland AJ (2005) Analysis of genetic factors influencing the developmental rate of globally important CIMMYT wheat cultivars. Crop Sci 45:2113–2119

Vilmorin H (1880) Les Meilleurs Bles. Vilmorin-Andrieux & Cie

Voss-Fels KP, Robinson H, Mudge SR et al (2018) VERNALIZATION1 modulates root system architecture in wheat and barley. Mol Plant 11:226–229

Wallach D, Hwang C, Correll MJ et al (2018) A dynamic model with QTL covariables for predicting flowering time of common bean (*Phaseolus vulgaris*) genotypes. Eur J Agron 101:200–209

Wang E, Brown HE, Rebetzke GJ, Zhao Z, Zheng B, Chapman SC (2019) Improving process-based crop models to better capture genotype × environment × management interactions. J Exp Bot 70:2389–2401

Watson IA, Frankel O (1972) Walter Lawry Waterhouse 1887–1969. Rec Aust Acad Sci 2(3)

Welsh J, Keim D, Pirasteh B, Richards R (1973) Genetic control of photoperiod response in wheat. In: Proceedings 4th international wheat genetic symposium. University of Missouri, pp 879–884

Wenkel S, Turck F, Singer K, Gissot L, Le Gourrierec J, Samach A, Coupland G (2006) CONSTANS and the CCAAT box binding complex share a functionally important domain and interact to regulate flowering of Arabidopsis. Plant Cell 18:2971–2984

White JW (2009) NJAS—Wageningen journal of life sciences combining ecophysiological models and genomics to decipher the GEM-to-P problem, vol 57, pp 53–58

White JW, Hoogenboom G (1996) Simulating effects of genes for physiological traits in a process-oriented crop model. Agron J 88:416–422

Wilhelm EP, Turner AS, Laurie DA (2009) Photoperiod insensitive Ppd-A1a mutations in tetraploid wheat (Triticum durum Desf.). Theor Appl Genet 118:285–294

Worland A, Law C (1986) Genetic-analysis of chromosome 2D of wheat. 1. The location of genes affecting height, day-length insensitivity, hybrid dwarfism and yellow-rust resistance. Z Pflanzenzuchtung 96:331–345

Worland AJ, Börner A, Korzun V, Li WM, Petrovíc S, Sayers EJ (1998) The influence of photoperiod genes on the adaptability of European winter wheats. Euphytica 100:385–394

Würschum T, Boeven PHG, Langer SM, Longin CFH, Leiser WL (2015) Multiply to conquer: copy number variations at Ppd-B1 and Vrn-A1 facilitate global adaptation in wheat. BMC Genet 16:1–8

Würschum T, Langer SM, Longin CFH, Tucker MR, Leiser WL (2018) A three-component system incorporating Ppd-D1, copy number variation at Ppd-B1, and numerous small-effect quantitative trait loci facilitates adaptation of heading time in winter wheat cultivars of worldwide origin. Plant Cell Environ 41:1407–1416

Xie L, Zhang Y, Wang K et al (2021) TaVrt2, an SVP-like gene, cooperates with TaVrn1 to regulate vernalization-induced flowering in wheat. New Phytol 231:834–848

Yan L, Loukoianov A, Tranquilli G, Helguera M, Fahima T, Dubcovsky J (2003) Positional cloning of the wheat vernalization, vol 100, pp 6263–6268

Yan L, Helguera M, Kato K, Fukuyama S, Sherman J, Dubcovsky J (2004) Allelic variation at the VRN-1 promoter region in polyploid wheat. Theor Appl Genet 109:1677–1686

Yan L, Fu D, Li C, Blechl A, Tranquilli G, Bonafede M, Sanchez A, Valarik M, Yasuda S, Dubcovsky J (2006) The wheat and barley vernalization gene VRN3 is an orthologue of FT. Proc Natl Acad Sci 103:19581–19586

Yang FP, Zhang XK, Xia XC, Laurie DA, Yang WX, He ZH (2009) Distribution of the photoperiod insensitive Ppd-D1a allele in Chinese wheat cultivars. Euphytica 165:445–452

Yin X, van der Linden CG, Struik PC (2018) Bringing genetics and biochemistry to crop modelling, and vice versa. Eur J Agron 100:132–140

Zhao T, Ni Z, Dai Y, Yao Y, Nie X, Sun Q (2006) Characterization and expression of 42 MADS-box genes in wheat (*Triticum aestivum* L.). Mol Genet Genomics 276:334–350

Zhao Y, Mette MF, Gowda M, Longin CFH, Reif JC (2014) Bridging the gap between marker-assisted and genomic selection of heading time and plant height in hybrid wheat. Heredity (edinb) 112:638–645

Zhao CH, Sun H, Liu C et al (2019)[s] Detection of quantitative trait loci for wheat (*Triticum aestivum* L.) heading and flowering date. J Agric Sci 157(1):20–30

Zheng B, Chenu K, Dreccer F, Chapman S (2012) Breeding for the future: what are the potential impacts of future frost and heat events on sowing and flowering time requirements for Australian bread

wheat (*Triticum aestivium*) varieties? Glob Chang Biol 18:2899–2914

Zheng B, Biddulph B, Li D, Kuchel H, Chapman S (2013) Quantification of the effects of VRN1 and Ppd-D1 to predict spring wheat (*Triticum aestivum*) heading time across diverse environments. J Exp Bot 64:3747–3761

Zikhali M, Griffiths S (2015) The effect of earliness per se (Eps) genes on flowering time in bread wheat. In: Ogihara Y, Takumi S, Handa H (eds) Advances in wheat genetics: from genome to field, proceedings 12th international wheat genetics symposium. Springer, pp 339–345

Zikhali M, Leverington-Waite M, Fish L, Simmonds J, Orford S, Wingen LU, Goram R, Gosman N, Bentley A, Griffiths S (2014) Validation of a 1DL earliness per se (eps) flowering QTL in bread wheat (*Triticum aestivum*). Mol Breed 34:1023–1033

Zikhali M, Wingen LU, Griffiths S (2016) Delimitation of the earliness per se D1 (Eps-D1) flowering gene to a subtelomeric chromosomal deletion in bread wheat (*Triticum aestivum*). J Exp Bot 67:287–299

Zikhali M, Wingen LU, Leverington-Waite M, Specel S, Griffiths S (2017) The identification of new candidate genes *triticum aestivum* flowering locus T3–B1 (TAFT3-B1) and target of EAT1 (TATOE1-B1) controlling the short-day photoperiod response in bread wheat. Plant Cell Environ 40:2678–2690

12 Genome Sequences from Diploids and Wild Relatives of Wheat for Comparative Genomics and Alien Introgressions

Adam Schoen, Gautam Saripalli, Seyedali Hosseinirad, Parva Kumar Sharma, Anmol Kajla, Inderjit Singh Yadav and Vijay Tiwari

Abstract

Bread wheat is an important food source worldwide, contributing ~20% of the caloric intake per person worldwide. Due to a domestication bottleneck and highly selective breeding for key traits, modern wheat cultivars have a narrow genetic base. Wheat production faces several challenges due to both abiotic and biotic stresses as well as changing climatic conditions and genetic improvement of wheat is generally considered to be the most sustainable approach to develop climate resilient cultivars with improved yield and end-use traits. Since wheat cultivars and landraces have been explored extensively to identify novel genes and alleles, one way to overcome these pitfalls is by looking into the proverbial treasure trove of genomic diversity that is present in wheat's wild relatives. These wild relatives hold reservoirs of genes that can confer broad-spectrum resistance to pathogens, increase yield, provide additional nutrition, and improve dough quality. Genetic approaches and techniques have existed to introgress wild chromatin to bread wheat, as well as trace introgressions present in the germplasm for over 7 decades. However with the availability of NGS technologies, it is now easier to detect and efficiently integrate the genetic diversity that lies within wheat's gene pools into breeding programs and research. This chapter provides a concise explanation of current technologies that have allowed for the progression of genomic research into wheat's primary, secondary, and tertiary gene pools, as well as past technologies that are still in use today. Furthermore, we explore resources that are publicly available that allow for insight into genes and genomes of wheat and its wild relatives, and the application and execution of these genes in research and breeding. This chapter will give an up-to-date summary of information related with genomic resources and reference

A. Schoen · G. Saripalli · S. Hosseinirad · P. K. Sharma · A. Kajla · I. S. Yadav · V. Tiwari (✉)
Department of Plant Science and Landscape Architecture, University of Maryland, College Park, MD, USA
e-mail: vktiwari@umd.edu

A. Schoen
e-mail: awschoen@umd.edu

G. Saripalli
e-mail: gautams@umd.edu

S. Hosseinirad
e-mail: hrad@umd.edu

P. K. Sharma
e-mail: pksharma@umd.edu

A. Kajla
e-mail: akajla97@umd.edu
I. S. Yadav
e-mail: isyadav@umd.edu

R. Appels et al. (eds.), *The Wheat Genome*, Compendium of Plant Genomes,
https://doi.org/10.1007/978-3-031-38294-9_12

assemblies available for wheat's wild relatives and their applications in wheat breeding and genetics.

Keywords

Wheat · Gene pool · Wild wheat relatives · Reference assemblies · Resistance genes · Exome · Yieldrelated traits · NGS · Comparative genomics · Alien introgressions

12.1 Introduction

Bread wheat is one of the most important staple crops and provides over 1/5th of the calories consumed by the world's population (FAOSTAT 2020). Global wheat production needs to be increased in light of the growing human population and changing climatic conditions (Hickey et al. 2019; Ray et al. 2012, 2013; Tilman et al. 2011). To cope with the numerous challenges that wheat faces, such as heat, drought, and diseases, it is important to find useful sources of genes and alleles for its improvement, and at the same time, develop approaches for efficient transfer of this useful genetic variability to cultivated wheat. Efforts have already been made in this direction, with the major and successful efforts that have been made after the wheat genome reference assembly using *T. aestivum* cv. CHINESE SPRING became available as a model in the 2018 (Appels et al. 2018). Since then, more and more resources have been added up to speed up breeding activities and the development of markers for important traits. For instance, in the years 2019 and 2020 a wheat pan-genome resource containing an assembly of 10+ wheat genomes including elite cultivars from across the globe and a 1 K exome capture data were generated (He et al. 2019; Walkowiak et al. 2020). In fact, although high-quality reference assembly is available for CHINESE SPRING, it does not capture the complete species-specific variation that can be exploited for variety development. Therefore, the above genomic resources including the pan-genome and exome capture data have proven to be highly useful. These resources have also been exploited for identification of useful wild introgressions in wheat followed by marker development for biotic and abiotic stress tolerance traits. The current pan-genome resource consists of ten genomes with pseudomolecules level assembly and five genomes with assemblies of hexaploid wheat.

One of the major objectives of any breeding program has been to develop resilient wheat varieties against environmental conditions as well as biotic stresses and significant progress has been made in the genetic improvement of wheat, mainly after the green revolution either using conventional or molecular breeding approaches through marker assisted selection. The introduction of dwarfing genes during the green revolution revolutionized wheat variety development and led to dramatic increase in wheat yield across the globe (Ali et al. 1973; Hedden 2003; Pingali 2012). Similarly, important genetic markers have also been identified for the QTL/genes providing resistance against different biotic and abiotic stresses (Saini et al. 2022; Singh et al. 2021). This has certainly led to the enhancement in the breeding populations of wheat; however, at the same time it has also narrowed down the genetic base thus resulting in reduced species variability. This ultimately necessitates the need to explore the wild and related species of wheat which are an important reservoir of useful genetic diversity as well as genes for biotic and abiotic stresses.

Based on the evolutionary distance between the species and the success rate of interspecies hybridization, Harlan and de Wet (1971) introduced the idea of wheat gene pools that included primary, secondary, and tertiary gene pool (Fig. 12.1) (Jiang et al. 1993; Mujeeb-Kazi et al. 2013). While, the genomes of primary and secondary gene pool share some homology with the wheat genome, the species in the tertiary gene pool do not share any homology with the wheat genome and, therefore, are sexually incompatible through homologous recombination. It is also difficult to cross the species of secondary and tertiary gene pool with hexaploid wheat

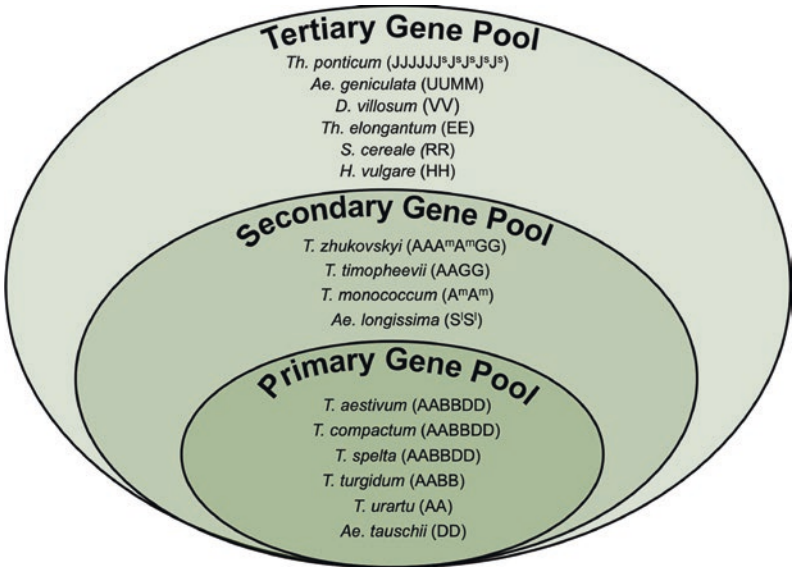


Fig. 12.1 Overview of bread wheat's gene pools with examples in each category

when compared to the species of primary gene pool (Mujeeb-Kazi et al. 2013).

The species in the primary gene pool include modern wheat cultivars and other *T. aestivum* landraces, *Triticum spelta* (AABBDD), tetraploid durum wheat *T. turgidum* (AABB), diploid wheat species *T. urartu* (AA), and *Aegilops tauschii* (DD). Examples of species in the secondary gene pool are tetraploid species *T. timopheevi*i (AAGG), and diploid species *T. monococcum* (AmAm) and *Ae. speltoide*s (SS). Species in the tertiary gene pool include cultivated species such as rye (RR) and barley (HH) as well as wild relatives of wheat. Importantly, wild relatives of wheat contain a treasure trove of variability that can overcome the genetic bottlenecks found in bread wheat (Tiwari et al. 2015). Examples of these are wild grasses such as diploid *Thinopyrum elongantum* (EE), tetraploid *Ae. geniculata* (UUMM), and octoploid *Leymus arenarius* (XXXXNNNN) (Pour-Aboughadareh et al. 2021; Anamthawat-Jónsson 2001). Due to the absence of pairing at meiosis between the tertiary pool chromosomes and those of wheat, techniques such as radiation induced chromosomal breaks or gene editing must be used to create introgression lines (Benlioğlu and Adak 2019; Jiang et al. 1993; Mujeeb-Kazi et al. 2013).

As mentioned above, the availability of genomic resources in hexaploid bread wheat has driven the development of useful markers leading to stress resilient wheat cultivars. However, looking at the complexity of the wheat genome owing to its large genome size and polyploid nature, it became necessary to develop genomic resources for the above wild relatives of wheat. Considerable progress has already been in this direction. For example, diploid relatives *Ae. longissima*, *Ae. speltoides*, and *Ae. sharonensis*, as well as several accessions of *Ae. tauschii* all have recently released reference quality assemblies available for BLAST and genome browsing (Avni et al. 2022; Gaurav et al. 2022; Zhou et al. 2021). Further, wild tetraploid species *T. turgidum* ssp. *dicoccoides* v. "ZAVITAN" have also recently had a high-quality assembly released with the use of optical maps for more accurate scaffolding.

The present chapter is mainly focused on providing an overview of the available reference assemblies, and genomic resources in wheat's wild relatives, which have been explored to identify useful introgressions in wheat. Some

examples include (i) *Fhb7* (from *T. elongatum*) providing resistance against Fusarium head blight in wheat (Guo et al. 2015); (ii) the well-known 1BL/1RS translocations from rye which has useful genes for improved grain yield and biomass especially under abiotic stress (Lukaszewski 1993), *Lr57* and *Yr40* from *Ae. geniculata* providing resistance against rust disease (Kuraparthy et al. 2007a, b). Recent developments in the next generation sequencing technologies have led to the development of low-cost sequencing reactions such as skim sequencing which provides a useful resource for the identification of alien introgressions with even a low coverage of less than 0.1x (Adhikari et al. 2022b). A comparative overview of synthetic relationships between wheat and wild relatives is also discussed. Overall, the present chapter will serve as a useful resource for the students and researchers working in alien wheat genomics and exploring useful alien wheat introgressions in development of wheat cultivars.

12.2 State of Reference Assemblies in Wheat and Its Wild Relatives

Wild and related species in wheat are a reservoir of important genes for different abiotic and biotic stress tolerances. Therefore, the availability of genomic resources for these wild relatives will prove to be an asset for identification of genes/QTLs and their linked markers which may be helpful in simplifying wheat genomics leading to development of elite wheat cultivars which is otherwise difficult due to complex and large wheat genome. Reference genome assemblies are now available for some of the important wild species belonging to all the three wheat gene pools. Reference assemblies for the important wheat relatives are explained in brief below.

12.2.1 Primary Gene Pool Reference Genomes

The first draft of the reference genome of bread wheat first became public in 2014, utilizing survey sequencing of individual chromosomes. Though this is considered a significant breakthrough in the world of wheat genomics, this initial draft sequence only accounted for ~61% of the entire wheat genome (Lukaszewski et al. 2014). Four years later, with the use of additional genetic data, including radiation hybrids, and sequence data, with the advancement of next generation sequencing (NGS) technologies, the fully annotated CHINESE SPRING reference genome was released with pseudomolecule assemblies for all 21 chromosomes (Appels et al. 2018). This reference genome has been continuously updated with the use of new technologies, both with the intent of more accurate contig establishment and scaffolding as well as annotation of genes not initially reported in the V.1.0. (Alonge et al. 2020; Zhu et al. 2021). Extensive comparative data shows that CHINESE SPRING is a genetic outlier when compared to domesticated species of *Triticum* sp. (Walkowiak et al. 2020).

The development of the pan-genome of wheat has allowed for more precise research and insight into the primary gene pool of wheat, including *T. spelta*. As of December 2022, 13 cultivars of wheat and one cultivar of *T. spelta* are available for BLAST as well as genome browsing. Interestingly, with the information gained by the 10+genome project, alien introgressions were able to be traced using reads derived from *T. timopheevii* and *T. ponticum* ($JJJJJJ^sJ^sJ^sJ^sJ^s$) in *T. aestivum* cv. LANCER, and *Ae. ventricosa* ($N^vN^vD^vD^v$) in *T. aestivum* cv. JAGGER in order to get more exact coordinates of these loci.

Tetraploid species of both cultivated (*T. durum*) and wild emmer (*T. dicoccoides*) wheat are also a part of the primary gene pool, due to the ability for homologous recombination to occur within the shared sub-genomes (A and B). When compared to hexaploid wheat, only 5% of wheat grown for human consumption is durum, and 95% is hexaploid. This may be attributed to the genome plasticity of hexaploid wheat which allowed for a broader potential for adaptation compared to tetraploid wheat (Mastrangelo and Cattivelli 2021). Also, compared to hexaploid wheat, the elite gene pool of durum wheat has

little genetic diversity, and most elite durum wheat cultivars are moderately to highly susceptible to disease resistance breeding (Clarke et al. 2010; Miedaner and Longin 2014). This is also not surprising due to the widely known fact that hexaploid bread wheat actually evolved from an inter-specific hybridization between *T. dicoccoides* and diploid species *Ae. tauschii* (Dvorak et al. 2012; Lukaszewski et al. 2014; Mcfadden and Sears 1946). However, it is evident from the published reports that wild emmer introgressions were responsible for significant gains in genetic diversity among the hexaploid lines as shown recently using the 1000 Wheat Exome Project (He et al. 2019). Similarly, the phenotypic variance contributed by several important traits including harvest weight, drought response, and plant height is largely attributed to these wild emmer introgressions (Nigro et al. 2022; Zhu et al. 2019).

Looking into the importance of wild emmer introgressions in hexaploid bread wheat, improved reference genomes of both wild emmer and cultivated durum wheat were published in 2019. The improved reference genome of wild emmer wheat cv. ZAVITAN (WEW) utilized optical maps as well as advancements in alignment technologies in order to increase the effective size of the reference genome by ~67 Mb, as well as adding over 2,000 high confidence genes. Additionally, between WEW_v1.0 and WEW_v.2.0, gaps of unknown size dropped from 2,767 to only 471 (Avni et al. 2017; Zhu et al. 2019). Later in 2019, a high-quality reference genome of *T. durum* cv. SVEVO was published, and by utilizing the WEW data, it was shown that the short-term evolutionary changes showed little change to synteny between WEW and durum. There were, however, lower copy numbers of important gene families such as NLRs in SVEVO in comparison with Zavitan, which implies a reduction of canonical R-genes (Maccaferri et al. 2019).

Diploid progenitor species of bread wheat genomes A (*T. urartu*) and D (*Ae. tauschii*), as well as close B genome relative *Ae. speltoides* (SS) all serve as a less complex system to work with for genomics research than the hexaploid bread wheat (Kerby and Kuspira 1987). Therefore, in recent years, reference genomes for all the three wheat genome donors (A genome; *T. urartu*, B genome: *Ae. speltoides*; D genome: *Ae. tauschii*) have been produced in order to help with wheat improvement. While the donors for A and D genome are included in the primary gene pool, the donors for the B genome are included in secondary gene pool. Therefore, the reference assemblies for the donors of A and D genome are discussed in more detail below, and the reference assemblies for the B genome donor (*Ae. speltoides*) are discussed in separate sub-heading in the next section involving secondary gene pool.

12.2.2 A Genome

The *T. urartu* reference genome was first published in 2018 (Ling et al. 2018), four months before the release of the CHINESE SPRING v.1.0 reference genome. In their analysis, done using the 2014 draft wheat genome v. 0.4, strong structural variations were observed between the *T. urartu* A genome and the bread wheat A genome, proposing evolutionary rearrangements. Within the diverse population of *T. urartu* accessions used for this study, and using the reference genome, three distinct groups were identified in the Fertile Crescent. The above diverse accessions were screened for powdery mildew resistance, and excitingly, after inoculation with powdery mildew (PM), one group (group 2) showed significant resistance against the pathogen. Further, analysis using the SNP data revealed a single putative candidate gene that was involved in providing resistance against powdery mildew. This resistance was perhaps due to the natural selection for powdery mildew resistance as well adaptation to grow at high altitudes.

12.2.3 D Genome

The D genome progenitor *Ae. tauschii* is a well of genetic variability in wheat, due to the

low level of variation seen within D genome of wheat (Dubcovsky and Dvorak 2007; Voss-Fels et al. 2015). This lack of variation is partially due to the small proportion of diversity that was obtained during polyploidization when hybridization between ancient, domesticated *T. turgidum* (AABB), and the small population of *Ae. tauschii* near the Caspian Sea (Dubcovsky and Dvorak 2007; Gaurav et al. 2022; Luo et al. 2017; Voss-Fels et al. 2015). However, due to the ability to develop synthetic wheat by hybridizing tetraploid species with *Ae. tauschii*, diversity in the D genome can be integrated into the breeding germplasm (Li et al. 2018). The first *Ae. tauschii* reference genome was released in 2017 in the background of accession AL8/78; the current version (*Aet* v.5.0) has been improved using optical maps as well as Pac-Bio long-read sequencing (Luo et al. 2017; Wang et al. 2021).

Since the initial release, several strides have been made in *Ae. tauschii* genomics. For instance, Zhou et al. (2021) developed reference quality genomes of four additional accessions representing four sub-lineages of *Ae. tauschii* with the intent to trace wild introgressions better in the germplasm. In the same year, the Open Wild Wheat Consortium (OWWC) generated whole genome sequencing (WGS) data for 242 non-redundant accessions of *Ae. tauschii*, to probe the evolution of bread wheat, determine the variation within the population, and perform genome-wide association studies (GWAS) for important traits using the AL8/78 reference genome (Gaurav et al. 2022). This study was able to show the two major lineages that make up the D genome in wheat, and using the wheat pan-genome, show the physical regions that come from these lineages. Additionally, a third lineage not associated with the evolution of bread wheat was also characterized.

Further, using *k*-mer-based GWAS, candidate genes for flowering time, stem rust (Sr) resistance, trichome number, spikelet number, PM resistance, and wheat curl mite resistance were also reported. Efforts are currently underway by the OWWC to develop a pan-genome resource for *Ae. tauschii* which will provide further information pertaining to the diversity prevailing in the genome sequences of diverse *Ae. tauschii* accessions (openwildwheat.org).

12.2.4 Secondary Gene Pool Reference Genomes

In comparison with the primary gene pool of wheat, genomic resources for members of the secondary gene pool are limited. Therefore, efforts are being made in this direction. For instance, (i) the development of reference assemblies for wild and cultivated *T. monococcum* accessions are available for public use (Ahmed et al. 2023).

Diploid wheat *T. monococcum* which is a close relative of *T. urartu* (A genome donor) is the only species with both domesticated (*T. monococcum* ssp. *monococcum*) and wild type (*T. monococcum* ssp. *aegilopoides*) accessions. Therefore, the reference assemblies for these species once available will certainly help in simplifying wheat genomics and may be an improvement over the reference assembly available for *T. urartu*. (ii) Transcriptome data for *T. monococcum* is also available from an earlier study (Fox et al. 2014). (iii) A core set of wild einkorn as well as domestic einkorn was also recently categorized by Adhikari et al. (2022a). Using GBS data, 145 domesticated einkorn accessions and 584 wild einkorn accessions were divided into α, ß, γ, and *monococcum*. A set of *T. urartu* accessions were also a part of this study, and as expected, they clustered together distally from *T. monococcum* accessions.

When compared to A and D genome, B genome of wheat has been difficult to study in a diploid species due the proposed extinction of the direct progenitor (Riley et al. 1958; Sarkar and Stebbins 1956). Researchers, however, have found a workaround this issue by working with species in the Sitopsis section of *Aegilops* (S^*S^*) due to their close relatedness with the B genome (Kerby and Kuspira 1987). In the last decade, reference quality genomes for five Sitopsis species were released to help with additional resources for not only the elucidation of the B

genome of wheat, but also as a further resource in researching the D genome of wheat (Li et al. 2022; Sandve et al. 2015; Yamane and Kawahara 2005; Yu et al. 2022). Recently, Avni et al. (2022) also communicated the release of three reference quality genomes in the same section (Sitopsis) which included two new assemblies for *Ae. sharonensis* (S^lS^l), and *Ae. speltoides* and one assembly for *Ae. sharonensis* which was in fact first communicated by Yu et al. (2022).

Alignments of the above assemblies with the different sub-genomes of wheat revealed a strong linear alignment of *Ae. sharonensis* and *Ae. longissima* with the D genome of bread wheat, and that of *Ae. speltoides* with the B genome which is obvious due to their strong relationship with the respective sub-genomes (Fig. 12.2). This was also further supplemented with the clustering of high confidence gene annotations of *Ae. sharonensis* and *Ae. longissima* with bread wheat's D genome as well as *Ae. tauschii*, and that of *Ae. speltoides* with WEW, durum wheat, and bread wheat's B genome.

In March 2022, reference assemblies of two additional "S" genomes (*Ae. bicornis* (S^bS^b) and *Ae. searsii* (S^bS^b)) were communicated, finally completing the Sitopsis section of the *Triticeae*. Both the above S genome assemblies also clustered with the D genome and D genome progenitors of wheat, in comparative alignments showing their closer association within the ancestry of wheat's evolution. Interestingly, with this complete information, it was found that the divergence of the D-related Sitopsis clade from the D progenitors was predicted to have happened around 5.23 Mya, whereas *Ae. speltoides* and the B genomes of both durum and bread wheat happened more recently at 4.44 Mya. With these available genomes, more precise genomic research can now be performed in the B genome of wheat, as well as diving deeper into the evolution of the D genome and its progenitors.

12.2.5 Tertiary Gene Pool Reference Genomes

The tertiary gene pool of wheat is underrepresented in terms of resource availability and research, due to the difficulty in defining these species, as well as limited genomic information (Qi et al. 2007; Schneider et al. 2008; Tiwari et al. 2015). Discussions in the literature have considered the Sitopsis section species as members of the tertiary gene pool, not including *Ae. speltoides*, but since the recent advancements in their genomic resources, it is more fitting to place them in the secondary genepool. Although some species, such as *T. elongantum* (EE), have had assemblies and annotations competed for attention with regards gene cloning, no reference genomes for wild grasses in the tertiary gene pool are currently available (Wang et al. 2020).

Two cultivated species, on the other hand, belonging to the tertiary gene pool, barley (*Hordeum vulgare*; $2n = 2x = 14$; HH) and rye (*Secale cereale*; $2n = 2x = 14$; RR), have had reference genomes published in the last ten years. The original barley genome was the first species in the Triticeae tribe to have a reference genome (Melonek and Small 2022; Mochida and Shinozaki 2013; Purugganan and Jackson 2021). Originally sequenced and annotated in

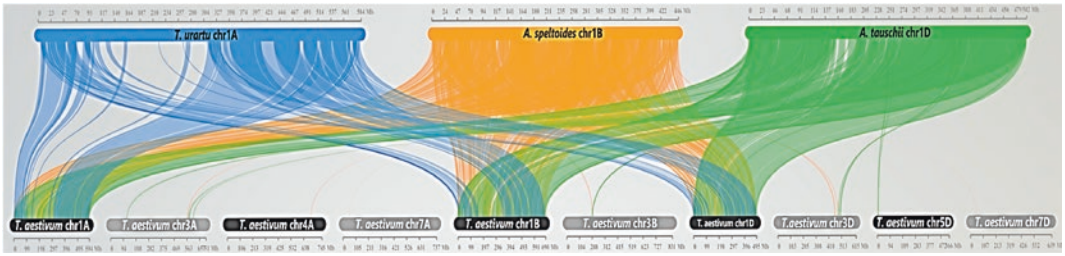


Fig. 12.2 Synteny between diploid wheat chromosomes 1A, 1S, 1D and hexaploid bread wheat's genome

2012, one of the biggest achievements in this assembly was overcoming the size and complexity of cereal genomes, due to the highly repetitive elements (Mayer et al. 2012). Since the release, updates have been made to properly order the chromosomes and create a better physical map, as well as reduce the unanchored sequences from ~250 to 83 Mb (Beier et al. 2017; Mascher et al. 2017; Monat et al. 2019). In a similar fashion to the achievements in wheat, a pan-genome project was also developed in barley, which included the sequencing and assembly of 19 additional barley lines including two highly transformable lines (GOLDEN PROMISE and IGRI) as well as a wild barley genotype (Jayakodi et al. 2020). This resource was, and is, an important milestone in the advancement of cereal crop genomics due to its early elucidation. Rye is an important member of the tertiary gene pool as a contributor of high tolerance for both biotic and abiotic stresses. Additionally, rye has been an important player in wheat breeding due to the importance of the 1BL/1RS and 1AL/1RS which confer resistance to multiple biotic diseases (Zeller and Sears 1973; Jung and Seo 2014). Moreover, synthetic hybrids of rye and wheat, named Triticale, have gained popularity due to their nutritional value as forage (Zhu 2018).

To better understand the underlying genetics behind the important aspects of rye, two reference quality genomes of wheat were released simultaneously in 2021. In the article by Rabanus-Wallace et al. (2021), a chromosome scale assembly was developed in the background of cv LO7, showing similar genomic makeup as other members of in Triticeae, and strong collinearity with the barley genome. Using this assembly, the researchers were able to determine a translocated region conferring frost tolerance in a 5A/5RL translocation line, first denoted using chromosome labeling and confirmed using read depth analysis on bread wheat's 5A chromosome and rye's 5R chromosome. In another article by Li et al. (2021), an additional genotype of rye, cv. WEINING, had a reference assembly created, which provided further support for the strong collinearity between the tertiary gene pool genomes. In their study, utilizing 2,517 single-copy orthologous genes, Li et al. (2021) developed a phylogenetic tree depicting 12 grasses and their evolutionary divergence. Although it is not necessarily new information, with the rye genome sequenced, the authors were able to compare rye with the other 11 sequenced genomes to deduce that rye had diverged from wheat ~5 Mya after barley and wheat's divergence, giving further evidence of rye's closer relationship with bread wheat and its progenitors. For a summary of the state of reference genomes in Triticeae from the past five years (see Fig. 12.3).

12.3 Alien Introgressions and Comparative Genomics

As described above, wild wheat relatives play an important role in the production of high performing wheat cultivars. Modern breeding techniques have reduced the genetic diversity in the breeding germplasm to select for higher yield (Keilwagen et al. 2022; Sansaloni et al. 2020; Schneider et al. 2008). Utilizing DNA segments from wild relatives that have been integrated into bread wheat's genome is a method to overcome this reduction in genetic diversity (Fig. 12.4); however, methods for detecting these introgressions are a must to properly trace these segments in breeding programs (Hao et al. 2020; Molnár-Láng et al. 2015). In this section, we will describe the methods, both old and new, that researchers utilize to detect and trace these introgressions, describe the important genes that come from these introgressions, as well as show the usefulness of modern technologies for comparative genomic analysis.

12.3.1 Methods for Detecting Alien Introgressions

Different methods for detecting the alien introgressions can be broadly classified into

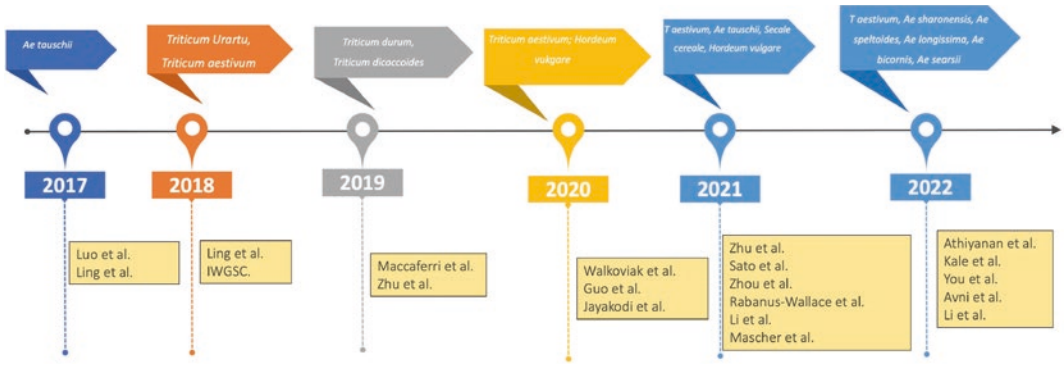


Fig. 12.3 Timeline of reference genomes in Triticeae from 2017 to 2022

cytological/cytogenetic, PCR-based markers and Recent Next Generation Sequencing (NGS)-based methods including skim sequencing.

12.3.2 Cytological Methods

Cytological methods for detecting chromosomal morphological differences have been used for almost 100 years (Gill and Friebe 1996). A popular method for observing different sizes and compositions of chromosomes was achieved by using centromeric heterochromatin staining, or C-banding, which allows for visualization of chromosomes and/or karyotypes of different species on a conserved scale (Endo and Gill 1996; Gill et al. 1991). This method was used for detecting rye/wheat hybrid pairing as far back as 1977 as well as determining *T. timopheevii* introgressions in *T. timopheevii* × *T. aestivum* hybrids (Badaeva et al. 1991; Dhaliwal et al. 1977). The C-banding method used alongside genomic in situ hybridization (GISH) also allowed for the detection of introgressions from *Ae. umbullata* (UU), *Ae. speltoides, Ae. comosa* (MM), *Ae. longissima*, and *T. timopheevii* as well as several others as far back as the early 90s (Friebe et al. 1996).

More recently, regions of *Leymus racemosus* DNA containing important Fusarium Head Blight (FHB) resistance gene *Fhb3* introgressed into bread wheat were traced using GISH and C-banding (Qi et al. 2008). Another, still popular, method of visualizing introgressions in wheat is the use of fluorescence in situ hybridization or FISH, which utilizes fluorescent-labeled DNA probes to detect important regions of chromosomes, such as introgressions (Campos-Galindo 2020; Jiang and Gill 2006). This method is still much in use today to provide further evidence of translocations in wheat, including the previously mentioned frost tolerance associated region in rye introgressed into wheat background (Rabanus-Wallace et al. 2021). This method has also been used to dissect introgressions coming from *T. elongantum, Ae. columnaris* ($U^cU^cX^cX^c$), *Ae. caudata* (CC), *T. timopheevii*, as well as many more not noted here (Badaeva et al. 2017; Devi et al. 2019; Grewal et al. 2020; Guo et al. 2022). Another use of this method was described in 2021, where FISH and GISH markers were utilized to visualize the recombination patterns of susceptible vs resistant genotypes of *Ae. geniculata* ($U^gU^gM^gM^g$) introgression lines in F_3 families (Steadham et al. 2021).

12.3.3 PCR-Based Markers

Another method to detect alien introgressions is by using PCR-based markers that are polymorphic between bread wheat and the wild species. The use of PCR-based markers for identifying alien introgressions in bread wheat dates back to early 90s when Rogowsky et al. (1993) designed PCR and RFLP markers to detect famous 1AS.1RL, 1BS.1RL, and 1DS.1RL rye

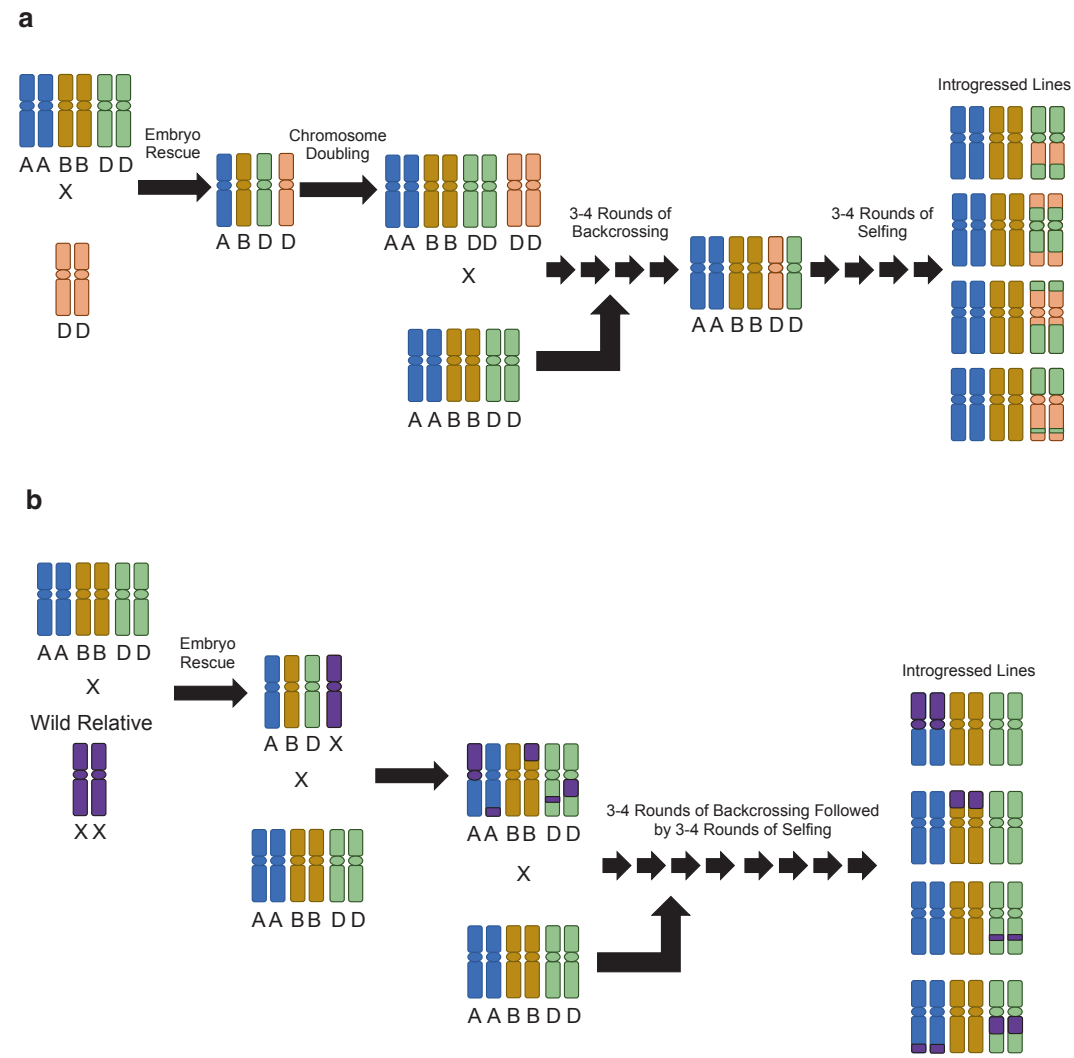


Fig. 12.4 Methods for developing introgression lines from wild relatives coming from **a** the primary gene pool and **b** the secondary and tertiary gene pools

introgressions in wheat background. Since then, PCR-based markers are continuously being implemented for identifying introgressions. More recently Li et al. (2019) designed markers to detect *Thinopyrum intermedium* ssp. *trichophorum* (JJJsJsStSt) introgressions in wheat that provide significant stripe rust resistance. To illustrate the importance of old and new technologies, these researchers utilized GISH, FISH, and C-banding in order to validate the effectiveness of the PCR markers, which now can be utilized in marker assisted breeding (MAS) to incorporate these genes into the breeding germplasm. Further, polymorphic SSR markers were also developed recently to detect introgressions from synthetic amphidiploid species *T. kiharae* (A^tA^tGGDD) which holds a reservoir of genes that have the potential to improve resistance to many diseases as well as increase the quality of flour production (Orlovskaya et al. 2020).

12.3.4 NGS Technology

With the advent of cost-effective NGS methods, researchers now have the ability to obtain sequence data coming from the transcriptome, exome, as well as the whole genome. This data can be generated from any species that the researchers are interested in, including the wild relatives of wheat. Examples of this have been mentioned in Sect. 12.2 of this chapter, in regard to whole genome assembly; however, data for wild relatives is constantly being generated for purposes of gene mapping and cloning, as well as diving deeper into wild relatives. One such example comes from Tiwari et al. (2015), where the 5 Mg chromosome of *Ae. geniculata* was sorted, sequenced, and assembled to gain insight into this important species. This information helped with the fine mapping of *Lr57* and *Yr40* in translocation wheat lines (Steadham et al. 2021).

In the past 5 years, NGS data has been utilized to detect introgressions in Triticeae species without the additional step of SNP calling, which can create artifacts as well as require more computational resources (Li and Wren 2014). Genotyping by sequencing (GBS) data provides short and low coverage genomic data, usually for the purpose of creating VCF files in order to genotype a population with relatively low computational and storage requirements (Perea et al. 2016). This data has now been shown to be able to discern introgressions in both wheat and barley. In a study by Keilwagen et al. (2019), they were able to detect putative introgressions from wild relatives in wheat, including the 1BL/1RS translocation. Interestingly, in the panel of 209 elite European winter wheat varieties in which GBS data was generated, many of the regions where introgressions were detected, these overlapped with important genes used in breeding programs such as *Yr17* from *Ae. ventricosa* ($N^vN^vD^vD^v$) and *Lr19* from *T. ponticum* as well as genes not yet known to be from wild relative introgressions such as *Glu-D1* and *Ppo-D1*. Due to the decrease in the cost of WGS data generation, one group set out to see the benefit of using resequencing data from multiple wild relatives to detect introgressions, utilizing the 10+ genomes described above. Keilwagen et al. (2022) used wild relative WGS data from both public repositories as well data generated from their own experiments to determine the regions of wild introgressions in 10 genotypes, gathered from the 10+ wheat genome project. In doing so, 9 introgressions coming from wild relatives *Ae. ventricosa, Ae. markgrafii* (CC), *Ae. speltoides*, *T. timopheevii*, *Ae. umbullata*, *Ae. uniaristata* (NN), and *T. ponticum* were found to be present on chromosomes 2A, 2B, 2D, 3D, and 4A. The researchers determined that within introgressions found on 2AS (from either *Ae. ventricosa* or *Ae. markgrafii*), 2B (from *T. timopheevii*) and 2DL (from *Ae. markgrafii* or, *Ae. umbullata*) contained genes that shared >90% amino acid similarity with genes coding for leaf rust and stripe rust genes, respectively. Fascinatingly, when checking the two introgressions that were present in all 10 genotypes, on 2A and 4AL coming from *Ae. speltoides*, in relatives of bread wheat, these introgressions were found in *T. urartu*, *T. boeoticum*, and *T. monococcum*, but not in *T. dicoccoides* or *T. spelta*. The studies also determined that these introgressions were able to be detected using only 1% of the total data.

To further save on computational cost, researchers have shown that skim sequencing of genomes can be used at a coverage as low as 0.025x, to determine introgressions, as described by Adhikari et al. (2022a, b). These authors used this method to determine barley introgressions on chromosomes 7A, 7B, and 7D in a population of 384 wheat–barley introgression lines. Additionally, they screened *T. intermedium*-durum wheat amphiploid lines to find not only lines where there were possible introgressions, but also certain lines containing whole wheat chromosomes. Due to the efficacy and precision of this method of detecting introgressions, this method is more than likely to define what the future of alien introgression mapping procedures looks like for researchers not only in wheat, but in all important crops.

12.3.5 Agronomically Important Genes Coming from Alien Introgressions

One of the most important alien introgressions in wheat is the 1BL/1RS translocation, in which the short arm of chromosome 1R in rye has replaced the short arm of 1B of wheat. This introgression has been used in wheat breeding not only for the disease resistance that is associated with this introgression, which has since become obsolete, but also because of the increased root biomass that has a positive effect on yield (Zeller and Hsam 1983; Sharma et al. 2011; Villareal et al. 1998). Despite the negative effects of this translocation on bread making quality, ~30% of modern cultivars contain the 1BL/1RS segment (Wang et al. 2017; Zeller et al. 1982). For a list of varieties containing 1R translocations visit http://www.rye-gene-map.de/rye-introgression/index.html (see also Ru et al. 2020). In recent years new 1BL/1RS lines have been developed to overcome some of the shortcomings of older introgressed lines, in which resistance against stripe rust, as well as drought tolerance was observed (Ren et al. 2022; Sharma et al. 2022; Gabay et al. 2020).

Ae. geniculata is also a genetic goldmine due to the strong disease resistance genes that are present in some accessions. The line TA10437, in which the 5 Mg chromosome was sequenced in 2015, contains important resistance genes against nefarious pathogens such as stripe rust and leaf rust (Tiwari et al. 2015). Recently, leaf and stripe rust resistance genes, *Lr57* and *Yr40* respectively, have been fine mapped in *Ae. geniculata* translocation lines utilizing mapping populations derived from a cross between resistant TA10437 derived introgression lines and susceptible disomic 5 Mg addition lines in the background of CHINESE SPRING (Steadham et al. 2021). In this study, *Lr57* and *Yr40* were not only fine mapped to a 1.5 Mb region of the introgressed *Ae. geniculata* 5 Mg segment, but through phenotyping of the mapping population and the donor parent of the 5 Mg segment, *Lr57* was shown to provide further evidence of its broad-spectrum resistance, confirming the results of an earlier study (Kuraparthy et al. 2007a, b). Moreover, this study showed that recombination is achievable in alien introgressions by crossing introgression lines with disomic lines containing homologous chromosomes of the alien species.

Sources of biotic disease resistance coming from wild relatives are unequivocally important for the sustenance and improvement of wheat; however, due to the associated linkage drag, their utilization in modern cultivars by durum and bread wheat breeders is limited for integrating "exotic" resistance genes from wild or cultivated relatives into their elite material (Hafeez et al. 2021; Steiner et al. 2019). But with the ever-increasing knowledge of wild wheat relatives, new genes that confer resistance are being integrated into the germplasm without a yield penalty. Powdery mildew and stripe rust resistance genes, *Pm5V* and *Yr5V* respectively, transferred from the annual diploid wheat relative *D. villosum* (VV) via amphiploid generation (*T. turgidum* × *D. villosum*, AABBVV) (Zhang et al. 2022). In order to integrate these genes into the germplasm, subsequent crossing with elite *D. villosum* introgression lines was performed and yielded lines with comparable yield to that of elite bread wheat lines. However, due to grain softness that is also associated with the 5 V chromosome, chemical mutagenesis was performed to knockout this undesirable trait, resulting in comparable yielding, hard grained genotypes for utilization in wheat breeding. A summary of disease resistance genes coming from wild relatives is described in Table 12.1.

Outside of resistance, genes controlling yield-related and end-use traits coming from wild relatives have also been utilized by researchers to further address the benefits of these species. Wild tetraploid wheat *Agropyron cristatum* (PPPP) has been used as a donor for abiotic and biotic disease resistance, as well as for yield-related traits for over 30 years (Chen et al. 1992; Zhang et al. 2015). In a study by Zhang et al. (2018), researchers found that

Pubing260, a T3BL.3BS/6PL translocation line containing a small terminal introgression from *Ag. cristatum* had increased grains per spike, spikelets per spike, thousand kernel weight, and flag leaf width in comparison with elite bread wheat genotypes without this segment. Additionally, in 2022, a high molecular weight glutenin subunit (HMW-GS) gene coming from *Ae. tauschii* was directly introduced into bread wheat, and although the dough quality was reduced slightly, the quality of Chinese steamed bread increased (Bo et al. 2022).

Table 12.1 Disease resistance genes coming from wild relatives

Trait	Wild relative	Gene	Chromosome	Reference
Powdery mildew resistance	*Triticum monococcum*	*Pm1b*, *Pm25*	7AL, 1AS	Hsam et al. (1998), Shi et al. (1998), Murphy et al. (1999)
	Triticum urartu	*Pm60*		Zhang et al. (2022)
	Triticum turgidum var. dicoccoides	*Pm16*, *Pm26*, *Pm30*, *Pm31*		Reader and Miller (1991), Rong et al. (2000), Liu et al. (2002), Xie et al. (2003)
	Aegilops speltoides	*Pm53*	5BL	Petersen et al. (2015)
	Dasypyrum villosum	*Pm55*, *Pm5V* and *Yr5V*	5AL, 5DL	Zhang et al. (2015), Zhang et al. (2022)
	Aegilops tauschii	*Pm35*	5DL	Miranda et al. (2007)
Leaf rust/strip rust resistance	*Triticum ventricosum*	*Yr17*, *Lr37* and *Sr38*		Delibes et al. (1993), Jahier et al. (1996)
	Agropyron elongatum	*Lr19/Sr25*		Sharma and Knott (1966)
	Aegilops geniculata	*Lr57* and *Yr40*	5DS	Kuraparthy et al. (2007a, b)
	Aegilops peregrina	*LrAp*	6BL	Narang et al. (2020)
	Aegilops caudata	*LrAC*	5DS	Riar et al. (2012)
	Aegilops markgrafii	*LrM*	2AS	Rani et al. (2020)
	Aegilops umbellulata	*Lr9*	6BL	Sears (1956)
	Aegilops triuncialis	*Lr58*		Kuraparthy et al. (2007a, b)
	Aegilops tauschii	*Lr21*, *Lr32*, *Lr41*, *Lr42*, *Lr22a*		Rowland and Kerber (1974), Kerber (1987), Cox et al. (1994)
	Thinopyrum ponticum	*Sr26* and *Sr61*	6AL	Zhang et al. (2021a, b)
Stem rust resistance	*Thinopyrum intermedium*	*Sr44*	7DL	Liu et al. (2013)
	Secale cereale	*Sr50*		Mago et al. (2015)
Fusarium head blight resistance	*Leymus racemosus*	*Fhb3*	7AS	Qi et al. (2008)
	Thinopyrum elongantum	*Fhb7*	7DL	Wang et al. (2020)

12.4 Available Resources for Sequence Data and Plant Material

Availability and accessibility of resources is paramount for the development of higher yielding, disease resistant cultivars of wheat. Fortunately, there exists web-based databases for the extraction of genomic and transcriptomic information regarding wheat and its relatives. Furthermore, there are avenues available for requesting seed material for many of the species mentioned above. In this section, we will provide an overview of the publicly available sites that can be utilized to not only browse and obtain genomic data from bread wheat and wheat's wild relatives but also where to request seeds from repositories across the world.

12.4.1 Web-Based Databases for Sequence Data

The National Center for Biotechnology Information is a resource for genetic research for almost any species that has had any type of sequence information generated (Sayers et al. 2022). Their user-friendly website allows for easy search for any topic, giving results for all 35 of their databases. A simple search for the term "*Triticum*" on December 12, 2022, yielded results in 26 of the 35 available databases. Over 4 million hits from this search go to their nucleotide database, whereas ~3 million hits come from the protein database. Moreover, NCBI's sequence read archive (SRA) is a significant repository of sequencing data coming from NGS reads from researchers across the globe. These SRAs are mostly publicly available and include genome and transcriptome data that is BLASTable. A search for *T. intermedium* in the SRA database yields over 4 thousand results, 184 of which are from reads coming from genome sequencing. Suffice to say, NCBI's website is a significant source of information, especially for those who may not have access to funding their own NGS studies. However, due to the abundance of avenues in which data is deposited into their databases, curated navigation for specific species may be overwhelming. Specifically, when BLASTing against their database, many of the hits received may be outdated, or repeats of similar information.

Ensembl overcomes some of the pitfalls of NCBI by allowing users to select specific organisms to browse (Cunningham et al. 2022). Moreover, Ensembl plant removes species coming from Animalia, Fungi, and prokaryotes are removed to deconvolute searches for specific species. Although their database is not as robust as NCBI, the navigation of certain aspects is made much easier. Their biomart and downloads tabs allow for easy access to nucleotide and protein data for the species hosted by the website, which can be downloaded from a single web page. Ensembl plant stays up to date with current versions of reference genomes, including the newest versions of *T. aestivum*, *Ae. tauschii*, and *H. vulgare*, although old versions are still available. Another significant feature of Ensembl is their variation track that is available for some species. This feature allows for users to find variants of specific genomic regions, either found naturally or induced via chemical mutagenesis. By clicking on this feature, users are able to browse either the effects of these variations, or in some cases such as in *T. aestivum*, find accession numbers for mutant genotypes. This is very important for researchers who are looking for variants of candidate genes in gene cloning projects, making it easy to find knockouts and/or missense mutations in candidate regions. Unfortunately, very few relatives of wheat are available for BLAST, genome browsing, or data acquisition. Currently, diploid species *T. urartu*, *Ae. tauschii*, rye, and barley are the only diploid relatives of wheat that are accessible using this website.

For researchers who work specifically in small grains, GrainGenes is a curated database that has many features that are useful (Yao et al. 2022). Genome browsers are easy to find and available for several wild relatives of wheat, including the five accessions of *Ae. tauschii*,

and three of the genomes in the Sitopsis section mentioned above. Additionally, BLASTing is robust, being able to select from many wild relatives, including all members of the Sitopsis section. Additionally, GrainGenes has an easy-to-use search for markers and probes found in literature. There also are some useful tools that are found in GrainGenes, including genome specific primer (GSP) design. The website, however, has become more cumbersome over the years as more and more data is being added to the site, though currently a more user-friendly interface is being developed.

A wheat specific database also exists in the form of URGI (Alaux et al. 2018; see also Chap. 2). This site allows for wheat-curated research in the form of BLASTs that can be performed on specific chromosomes for all available versions. This is important because many of the times, in the literature, different versions of reference genomes are used for research. This site, although not as user friendly as the previously mentioned databases, contains a significant amount of sequence data for wheat.

12.4.2 Germplasm Acquisition Resources

Researchers across the globe are willing to share material with one another for the greater good of assuring food security. Specifically in wheat research, seed requests can be performed from multiple sources. One such example is the Wheat Genetics Resource Center, hosted by Kansas State University. This site gives direct access to alien species coming from the aforementioned Sitopsis section, as well as multiple other species coming from *Aegilops*, such as *Ae. geniculata*. Along with this, there is access to *Triticum* species including diploid *monococcum* and *urartu*. WGRC also contains 95 unique accessions of *Dasypyrum villosum* coming from several different countries. Alien translocation lines with transfers coming from *Aegilops*, *Dasypyrum*, *Triticum*, *Secale*, and *Agropyron* species are directly accessible from this resource as well. This site links to other important germplasms and seed distributors such as CIMMYT and the USDA.

CIMMYT (The International Maize and Wheat Improvement Center) and the USDA utilize the Germplasm Resource Information Network (GRIN) or GRIN-global to give international institutions access to germplasms of several different species of plants, including wheat and some of its wild relatives. Although this resource is not specifically catered to wheat researchers, wild species belonging to *Aegilops* and *Triticum* are available. Similarly, Genesys is a resource for multiple different crop systems, but their user-friendly interface allows for easy search for species in *Triticum.* This site contains over 12 thousand accessions coming from *Aegilops* alone, and they are designated by subsets, including *Aegilops* core sets.

The OWWC, mentioned in Sect. 12.2.1, has their panel available through the Germplasm Resource Unit (GRU) hosted by the John Innes Centre. This resource has similar resources as the aforementioned sites; however, they have a core collection of Titiceae wild relatives that include *Dasypyrum*, *Aegilops*, *Triticum*, and *Eremopyron*. This site also contains seed resources for mutant, DH, and other mapping populations in wheat, as well as historical landraces.

12.5 Concluding Remarks

The ever-increasing breadth of knowledge coming from wheat and its relatives have large implications for improving the overall quality of cultivars in the coming years. This chapter gives an up-to-date overview of recent advances in genomic resources within wheat, highlighting the importance of wild relatives, and alien introgressions within the germplasm. The availability of the wheat pan-genome has allowed for researchers to trace introgressions that are present within cultivars across the world, some of these alien introgressions were found within the entire pan-genome, giving further evidence of the importance of the genetic diversities (Keilwagen et al. 2022). As more of these wild

genomes get reference quality assemblies associated with them, the more we can learn about the important genes that lie within these species. On the OWWC website, a pan-genome of *Ae. tauschii* is currently underway, allowing researchers to get a more in depth understanding of the diversities that are present in these progenitor species. High-quality reference genomes are still required for some important species, such as *T. elongatum*, *D. villosum*, and *Agrypyron* species. Researchers would also benefit from pan-genomes representing other important wild relatives that have been mentioned in this chapter, such as that of wild diploid *Triticum* species, tetraploid *Aegilops* species.

Extensive resources for obtaining both genomic data as well as seed material for these species are available for public use, making further novel research possible across the globe. It is an exciting time to work in the field of wheat research with the ability to obtain diverse populations of not only bread wheat and its primary gene pool, but also members of secondary and tertiary gene pools from collaborators in different countries. The web-based resources that exist now make it possible for quick turnaround for not only basic scientific knowledge but also for the integration of this diversity into local breeding programs. A future prospect that could make this process even more efficient is a localized database where these independent seed and data repositories can be accessed. CIMMYT and the USDA make it easy to find material from either establishment by utilizing systems like GRIN and GRIN-global, which share germplasm requests; however, many other institutions do not utilize this as a means for requests and distribution, and further many of these are not necessarily catered toward wheat-based research. A similar central database would be beneficial for the amount of sequence data that is becoming available in Triticeae. A system to search for data pertaining to specific gene pools could prove to be beneficial for future research, especially as more genomes are being sequenced. The examples and information provided here will hopefully make it easier for researchers, students, and curious minds alike to find information pertaining to wheat and the many species that make up its gene pools.

References

Adhikari L, Raupp J, Wu S, Wilson D, Evers B, Koo DH, Singh N, Friebe B, Poland J (2022a) Genetic characterization and curation of diploid A-genome wheat species. Plant Physiol 188(4):2101–2114. https://doi.org/10.1093/plphys/kiac006

Adhikari L, Shrestha S, Wu S, Crain J, Gao L, Evers B, Wilson D, Ju Y, Koo DH, Hucl P, Pozniak C, Walkowiak S, Wang X, Wu J, Glaubitz JC, DeHaan L, Friebe B, Poland J (2022b) A high-throughput skim-sequencing approach for genotyping, dosage estimation and identifying translocations. Sci Rep 12(1). https://doi.org/10.1038/s41598-022-19858-2

Ahmed HI, Heuberger M, Schoen A, Koo DH, Quiroz-Chavez J, Adhikari L, Raupp J, Cauet S, Rodde N, Cravero C, Callot C (2023) Einkorn genomics sheds light on history of the oldest domesticated wheat. Nature. https://doi.org/10.1038/s41586-023-06389-7

Alaux M, Rogers J, Letellier T, Flores R, Alfama F, Pommier C, Mohellibi N, Durand S, Kimmel E, Michotey C, Guerche C, Loaec M, Lainé M, Steinbach D, Choulet F, Rimbert H, Leroy P, Guilhot N, Salse J, Quesneville H et al (2018). Linking the international wheat genome sequencing consortium bread wheat reference genome sequence to wheat genetic and phenomic data. Genome Biol 19(1). https://doi.org/10.1186/s13059-018-1491-4

Ali CH, Khan A, Choudhari HA (1973) Income impact of the green revolution author. Pak Econ Soc Rev 11(1)

Alonge M, Shumate A, Puiu D, Zimin AV, Salzberg SL (2020) Chromosome-scale assembly of the bread wheat genome reveals thousands of additional gene copies. Genetics 216(2):599–608. https://doi.org/10.1534/genetics.120.303501

Anamthawat-Jónsson, K (2001) Molecular cytogenetics of introgressive hybridization in plants. Methods Cell Sci 23:141–150. https://doi.org/10.1023/A:1013182724179

Appels R, Eversole K, Feuillet C, Keller B, Rogers J, Stein N, Pozniak CJ, Choulet F, Distelfeld A, Poland J, Ronen G, Barad O, Baruch K, Keeble-Gagnère G, Mascher M, Ben-Zvi G, Josselin AA, Himmelbach A, Balfourier F, Wang L et al (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science 361(6403). https://doi.org/10.1126/science.aar7191

Avni R, Nave M, Barad O, Baruch K, Twardziok SO, Gundlach H, Hale I, Mascher M, Spannagl M, Wiebe K, Jordan KW, Golan G, Deek J, Ben-Zvi B, Ben-Zvi

G, Himmelbach A, MacLachlan RP, Sharpe AG, Fritz A, Distelfeld A et al (2017) Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. Science 357(6346):93–97. https://doi.org/10.1126/science.aan0032

Avni R, Lux T, Minz-Dub A, Millet E, Sela H, Distelfeld A, Deek J, Yu G, Steuernagel B, Pozniak C, Ens J, Gundlach H, Mayer KFX, Himmelbach A, Stein N, Mascher M, Spannagl M, Wulff BBH, Sharon A (2022) Genome sequences of three Aegilops species of the section Sitopsis reveal phylogenetic relationships and provide resources for wheat improvement. Plant J 110(1):179–192. https://doi.org/10.1111/tpj.15664

Badaeva ED, Budashkina EB, Badaev NS, Kalinina NP, Shkutina FM (1991) General features of chromosome substitutions in *Triticum aestivum* × *T. timopheevii* hybrids. Theor Appl Genet 82(2):227–232. https://doi.org/10.1007/BF00226218

Badaeva ED, Ruban AS, Aliyeva-Schnorr L, Municio C, Hesse S, Houben A (2017) In situ hybridization to plant chromosomes, pp 477–494. https://doi.org/10.1007/978-3-662-52959-1_49

Beier S, Himmelbach A, Colmsee C, Zhang X-Q, Barrero RA, Zhang Q, Li L, Bayer M, Bolser D, Taudien S, Groth M, Felder M, Hastie A, Šimková H, Staňková H, Vrána J, Chan S, Muñoz-Amatriaín M, Ounit R, Mascher M et al (2017) Construction of a map-based reference genome sequence for barley, *Hordeum vulgare* L. Sci Data 4(1):170044. https://doi.org/10.1038/sdata.2017.44

Benlioğlu B, Adak MS (2019) Importance of crop wild relatives and landraces genetic resources in plant breeding programmes. J Exp Agric Int 1–8. https://doi.org/10.9734/jeai/2019/v37i330268

Bo C, Fan Z, Ma X, Li A, Wang H, Kong L, Wang X (2022) Identification and Introgression of a Novel HMW-GS Gene from *Aegilops tauschii*. Agronomy 12(11):2709. https://doi.org/10.3390/agronomy12112709

Campos-Galindo I (2020) Cytogenetics techniques. In: Human reproductive genetics. Elsevier, pp 33–48. https://doi.org/10.1016/B978-0-12-816561-4.00003-X

Clarke JM, Clarke FR, Pozniak CJ (2010) Forty-six years of genetic improvement in Canadian durum wheat cultivars. Can J Plant Sci 90(6):791–801. https://doi.org/10.4141/cjps10091

Chen Q, Jahier J, Cauderon Y (1992) Production and cytogenetic analysis of BC1, BC2, and BC3 progenies of an intergeneric hybrid between *Triticum aestivum* (L.) Thell. and tetraploid *Agropyron cristatum* (L.) Gaertn. Theoret Appl Genet 84:698–703. https://doi.org/10.1007/BF00224171

Cox TS, Raupp WJ, Gill BS (1994) Leaf rust-resistance genes Lr41, Lr42, and Lr43 transferred from *Triticum tauschii* to common wheat. Crop Sci 34(2):339–343

Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Austine-Orimoloye O, Azov AG, Barnes I, Bennett R, Berry A, Bhai J, Bignell A, Billis K, Boddu S, Brooks L, Charkhchi M, Cummins C, da Rin Fioretto L, Flicek P et al (2022) Ensembl 2022. Nucleic Acids Res 50(D1):D988–D995. https://doi.org/10.1093/nar/gkab1049

Delibes A, Romero D, Aguaded S, Duce A, Mena M, Lopez-Brana I, Andrés MF, Martin-Sanchez JA, García-Olmedo F (1993) Resistance to the cereal cyst nematode (*Heteroder aavenae* Woll.) transferred from the wild grass *Aegilops ventricosa* to hexaploid wheat by a 'stepping–stone' procedure. Theor Appl Genet 87(3):402–408

Devi U, Grewal S, Yang CY, Hubbart-Edwards S, Scholefield D, Ashling S, Burridge A, King IP, King J (2019) Development and characterisation of interspecific hybrid lines with genome-wide introgressions from *Triticum timopheevii* in a hexaploid wheat background. BMC Plant Biol 19(1). https://doi.org/10.1186/s12870-019-1785-z

Dhaliwal HS, Gill BS, Waines JG, Egulation R (1977) Analysis of induced homoeologous pairing in a ph mutant wheat X rye hybrid. J Heredity 68. https://academic.oup.com/jhered/article/68/4/207/802618

Dubcovsky J, Dvorak J (2007) Genome plasticity a key factor in the success of polyploid wheat under domestication. Science 316(5833):1862–1866. https://doi.org/10.1126/science.1143986

Dvorak J, Deal KR, Luo MC, You FM, von Borstel K, Dehghani H (2012) The origin of spelt and free-threshing hexaploid wheat. J Hered 103(3):426–441. https://doi.org/10.1093/jhered/esr152

Endo TR, Gill BS (1996) The deletion stocks of common wheat. J Heredity 87. https://academic.oup.com/jhered/article/87/4/295/2186527

Food and Agriculture Organization of the United Nations (2020) FAOSTAT statistical database

Fox SE, Geniza M, Hanumappa M, Naithani S, Sullivan C, Preece J, Tiwari VK, Elser J, Leonard JM, Sage A, Gresham C, Kerhornou A, Bolser D, McCarthy F, Kersey P, Lazo GR, Jaiswal P (2014) De novo transcriptome assembly and analyses of gene expression during photomorphogenesis in diploid wheat Triticum monococcum. PLoS ONE 9(5). https://doi.org/10.1371/journal.pone.0096855

Friebe B, Badaeva ED, Kammer K, Gill BS (1996) Standard karyotypes of Aegilops uniaristata, Ae. mutica, Ae. comosa subspeciescomosa andheldreichii (Poaceae). Plant Syst Evol 202(3–4):199–210. https://doi.org/10.1007/BF00983382

Gabay G, Zhang J, Burguener GF, Howell T, Wang H, Fahima T, Lukaszewski A, Moriconi JI, Santa Maria GE, Dubcovsky J (2020) Structural rearrangements in wheat (1BS)–rye (1RS) recombinant chromosomes affect gene dosage and root length. Plant Genome. https://doi.org/10.1002/tpg2.20079

Gaurav K, Arora S, Silva P, Sánchez-Martín J, Horsnell R, Gao L, Brar GS, Widrig V, John Raupp W, Singh N, Wu S, Kale SM, Chinoy C, Nicholson P,

Quiroz-Chávez J, Simmonds J, Hayta S, Smedley MA, Harwood W, Wulff BBH et al (2022) Population genomic analysis of Aegilops tauschii identifies targets for bread wheat improvement. Nat Biotechnol 40(3):422–431. https://doi.org/10.1038/s41587-021-01058-4

Gill BS, Friebe B (1996) Plant cytogenetics at the dawn of the 21st century. Curr Opin Plant Biol 1:109–124. http://biomednet.com/elecref/1369526600100109

Gill BS, Friebe B, Endo TR (1991) Standard karyotype and nomenclature system for description of chromosome bands and structural aberrations in wheat (Triticum aestivum). Genome 34(5):830–839. https://doi.org/10.1139/g91-128

Grewal S, Othmeni M, Walker J, Hubbart-Edwards S, Yang CY, Scholefield D, Ashling S, Isaac P, King IP, King J (2020) Development of wheat-aegilops caudata introgression lines and their characterization using genome-specific KASP markers. Front Plant Sci 11. https://doi.org/10.3389/fpls.2020.00606

Guo J, Zhang X, Hou Y, Cai J, Shen X, Zhou T, Xu H, Ohm HW, Wang H, Li A, Han F, Wang H, Kong L (2015) High-density mapping of the major FHB resistance gene Fhb7 derived from Thinopyrum ponticum and its pyramiding with Fhb1 by marker-assisted selection. Theor Appl Genet 128(11):2301–2316. https://doi.org/10.1007/s00122-015-2586-x

Guo X, Shi Q, Yuan J, Zhang J, Wang M, Wang J, Wang C, Fu S, Su H, Liu Y, Huang Y, Liu C, Liu Q, Sun Y, Wang L, Wang K, Jing D, Zhang P, Li J, Han F (2022) Alien chromatin other than the GST-encoding Fhb7 candidate confers Fusarium head blight resistance in wheat breeding. https://doi.org/10.1101/2021.02.03.429547 (BioRxiv)

Hafeez AN, Arora S, Ghosh S, Gilbert D, Bowden RL, Wulff BBH (2021) Creation and judicious application of a wheat resistance gene atlas. Mol Plant 14(7):1053–1070. Cell Press. https://doi.org/10.1016/j.molp.2021.05.014

Hao M, Zhang L, Ning S, Huang L, Yuan Z, Wu B, Yan Z, Dai S, Jiang B, Zheng Y, Liu D (2020) The resurgence of introgression breeding, as exemplified in wheat improvement. In: Frontiers in plant science, vol 11. Frontiers Media S.A. https://doi.org/10.3389/fpls.2020.00252

Harlan JR, Wet JMJ (1971) Toward A rational classification of cultivated plants. Taxon 20(4):509–517. https://doi.org/10.2307/1218252

He F, Pasam R, Shi F, Kant S, Keeble-Gagnere G, Kay P, Forrest K, Fritz A, Hucl P, Wiebe K, Knox R, Cuthbert R, Pozniak C, Akhunova A, Morrell PL, Davies JP, Webb SR, Spangenberg G, Hayes B, Akhunov E et al (2019) Exome sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome. Nat Genet 51(5):896–904. https://doi.org/10.1038/s41588-019-0382-2

Hedden P (2003) The genes of the green revolution. Trends Genet 19(1):5–9. https://doi.org/10.1016/S0168-9525(02)00009-4

Hickey LTN, Hafeez A, Robinson H, Jackson SA, Leal-Bertioli SCM, Tester M, Gao C, Godwin ID, Hayes BJ, Wulff BBH (2019) Breeding crops to feed 10 billion. In: Nature biotechnology, vol 37, issue 7. Nature Research, pp 744–754. https://doi.org/10.1038/s41587-019-0152-9

Hsam SLK, Huang X, Ernst F, Hartl L, Zeller F (1998) Chromosomal location of genes for resistance to powdery mildew in common wheat (*Triticum aestivum* L. em Thell.). 5. Alleles at the *Pm1* locus. Theor Appl Genet 96:1129–1134

Jahier J, Tanguy AM, Abelard P, Rivoal R (1996) Utilization of deletions to localize a gene for resistance to the cereal cyst nematode, *Heterodera avenae*, on an *Aegilops ventricosa* chromosome. Plant Breed 115:282–284

Jayakodi M, Padmarasu S, Haberer G, Bonthala VS, Gundlach H, Monat C, Lux T, Kamal N, Lang, D, Himmelbach A, Ens J, Zhang XQ, Angessa TT, Zhou G, Tan C, Hill C, Wang P, Schreiber M, Boston LB, Stein N et al (2020) The barley pan-genome reveals the hidden legacy of mutation breeding. Nature 588(7837):284–289. https://doi.org/10.1038/s41586-020-2947-8

Jiang J, Gill BS (2006) Current status and the future of fluorescence in situ hybridization (FISH) in plant genome research. Genome 49(9):1057–1068. https://doi.org/10.1139/G06-076

Jiang J, Friebe B, Gill BS (1993) Recent advances in alien gene transfer in wheat. Euphytica 73(3):199–212. https://doi.org/10.1007/BF00036700

Jung WJ, Seo YW (2014) Employment of wheat-rye translocation in wheat improvement and broadening its genetic basis. J Crop Sci Biotechnol 17(4):305–313. Korean Society of Crop Science. https://doi.org/10.1007/s12892-014-0086-1

Keilwagen J, Lehnert H, Berner T, Beier S, Scholz U, Himmelbach A, Stein N, Badaeva ED, Lang D, Kilian B, Hackauf B, Perovic D (2019) Detecting large chromosomal modifications using short read data from genotyping-by-sequencing. Front Plant Sci 10. https://doi.org/10.3389/fpls.2019.01133

Keilwagen J, Lehnert H, Berner T, Badaeva E, Himmelbach A, Börner A, Kilian B (2022) Detecting major introgressions in wheat and their putative origins using coverage analysis. Sci Rep 12(1). https://doi.org/10.1038/s41598-022-05865-w

Kerber (1987) Resistance to leaf rust in hexaploid wheat: Lr32 A third gene derived from *Triticum tauschii*. Crop Sci 27(2):204–206

Kerby K, Kuspira J (1987) The phylogeny of the polyploid wheats *Triticum aestivum* (bread wheat) and *Triticum turgidum* (macaroni wheat). Genome 29(5):722–737. https://doi.org/10.1139/g87-124

Kuraparthy V, Chhuneja P, Dhaliwal HS, Kaur S, Bowden RL, Gill BS (2007a) Characterization and mapping of cryptic alien introgression from *Aegilops geniculata* with new leaf rust and stripe rust resistance genes Lr57 and Yr40 in wheat. Theor Appl Genet 114(8):1379–1389. https://doi.org/10.1007/s00122-007-0524-2

Kuraparthy V, Sood S, Chhuneja P, Dhaliwal HS, Kaur S, Bowden RL, Gill BS (2007b) A cryptic wheat-*Aegilops triuncialis* translocation with leaf rust resistance gene *Lr58*. Crop Sci 47(5):1995–2003

Li H, Wren J (2014) Toward better understanding of artifacts in variant calling from high-coverage samples. In: Bioinformatics, vol 30, issue 20. Oxford University Press, pp 2843–2851. https://doi.org/10.1093/bioinformatics/btu356

Li A, Liu D, Yang W, Kishii M, Mao L (2018) Synthetic hexaploid wheat: yesterday, today, and tomorrow. Engineering 4(4):552–558. https://doi.org/10.1016/j.eng.2018.07.001

Li J, Chen Q, Zhang P, Lang T, Hoxha S, Li G, Yang Z (2019) Comparative FISH and molecular identification of new stripe rust resistant wheat-Thinopyrum intermedium ssp. trichophorum introgression lines. Crop J 7(6):819–829. https://doi.org/10.1016/j.cj.2019.06.001

Li G, Wang L, Yang J, He H, Jin H, Li X, Ren T, Ren Z, Li F, Han X, Zhao X, Dong L, Li Y, Song Z, Yan Z, Zheng N, Shi C, Wang Z, Yang S, Wang D et al (2021) A high-quality genome assembly highlights rye genomic characteristics and agronomically important genes. Nat Genet 53(4):574–584. https://doi.org/10.1038/s41588-021-00808-z

Li LF, Zhang Z, Wang ZH, Li N, Sha Y, Wang XF, Ding N, Li Y, Zhao J, Wu Y, Gong L, Mafessoni F, Levy AA, Liu B (2022) Genome sequences of five Sitopsis species of Aegilops and the origin of polyploid wheat B subgenome. Mol Plant 15(3):488–503. https://doi.org/10.1016/j.molp.2021.12.019

Ling HQ, Ma B, Shi X, Liu H, Dong L, Sun H, Cao Y, Gao Q, Zheng S, Li Y, Yu Y, Du H, Qi M, Li Y, Lu H, Yu H, Cui Y, Wang N, Chen C, Liang C et al (2018) Genome sequence of the progenitor of wheat A subgenome Triticum urartu. Nature 557(7705):424–428. https://doi.org/10.1038/s41586-018-0108-0

Liu Z, Sun Q, Ni Z et al (2002) Molecular characterization of a novel powdery mildew resistance gene *Pm30* in wheat originating from wild emmer. Euphytica 123:21–29

Liu W, Danilova TV, Rouse MN, Bowden RL, Friebe B, Gill BS, Pumphrey MO (2013) Development and characterization of a compensating wheat-Thinopyrum intermedium Robertsonian translocation with Sr44 resistance to stem rust (Ug99). Theor Appl Genet 126(5):1167–1177

Lukaszewski AJ (1993) Reconstruction in wheat of complete chromosomes 1B and 1R from the 1RS.1BL translocation of "Kavkaz" origin. Genome 36(5):821–824. https://doi.org/10.1139/g93-109

Lukaszewski AJ, Alberti A, Sharpe A, Kilian A, Stanca AM, Keller B, Clavijo BJ, Friebe B, Gill B, Wulff B, Chapman B, Steuernagel B, Feuillet C, Viseux C, Pozniak C, Rokhsar DS, Klassen D, Edwards D, Akhunov E, Dubská Z et al (2014) A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. Science, 345(6194). https://doi.org/10.1126/science.1251788

Luo M-C, Gu YQ, Puiu D, Wang H, Twardziok SO, Deal KR, Huo N, Zhu T, Wang L, Wang Y, McGuire PE, Liu S, Long, H, Ramasamy RK, Rodriguez JC, Van SL, Yuan L, Wang Z, Xia Z, Dvořák J et al (2017) Genome sequence of the progenitor of the wheat D genome *Aegilops tauschii*. Nature 551(7681):498–502. https://doi.org/10.1038/nature24486

Maccaferri M, Harris NS, Twardziok SO, Pasam RK, Gundlach H, Spannagl M, Ormanbekova D, Lux T, Prade VM, Milner SG, Himmelbach A, Mascher M, Bagnaresi P, Faccioli P, Cozzi P, Lauria M, Lazzari B, Stella A, Manconi A, Cattivelli L et al (2019) Durum wheat genome highlights past domestication signatures and future improvement targets. Nat Genet 51(5):885–895. https://doi.org/10.1038/s41588-019-0381-3

Mago R, Zhang P, Vautrin S, Simkova H, Bansal U, Luo MC, Rouse M, Karaoglu H, Periyannan S, Kolmer J, Jin Y, Ayliffe MA, Bariana H, Park RF, McIntosh R, Dolezel J, Berges H, Spielmeyer W, Lagudah ES, Ellis JG, Dodds PN (2015) The wheat Sr50 gene reveals rich diversity at a cereal disease resistance locus. Nat Plants 1:15186

Mascher M, Gundlach H, Himmelbach A, Beier S, Twardziok SO, Wicker T, Radchuk V, Dockter C, Hedley PE, Russell J, Bayer M, Ramsay L, Liu H, Haberer G, Zhang XQ, Zhang Q, Barrero RA, Li L, Taudien S, Stein N et al (2017) A chromosome conformation capture ordered sequence of the barley genome. Nature 544(7651):427–433. https://doi.org/10.1038/nature22043

Mastrangelo AM, Cattivelli L (2021) What makes bread and durum wheat different? In: Trends in plant science, vol 26, issue 7. Elsevier Ltd., pp 677–684. https://doi.org/10.1016/j.tplants.2021.01.004

Mayer KFX, Waugh R, Langridge P, Close TJ, Wise RP, Graner A, Matsumoto T, Sato K, Schulman A, Ariyadasa R, Schulte D, Poursarebani N, Zhou R, Steuernagel B, Mascher M, Scholz U, Shi B, Madishetty K, Svensson JT, Stein N et al (2012) A physical, genetic and functional sequence assembly of the barley genome. Nature 491(7426):711–716. https://doi.org/10.1038/nature11543

Mcfadden ES, Sears ER (1946) The origin of *Triticum spelta* and its free-threshing hexaploid relatives. J Hered 37(3):81–89. https://doi.org/10.1093/oxfordjournals.jhered.a105590

Melonek J, Small I (2022) Triticeae genome sequences reveal huge expansions of gene families implicated in fertility restoration. In: Current opinion

in plant biology, vol 66. Elsevier Ltd. https://doi.org/10.1016/j.pbi.2021.102166

Miedaner T, Longin CFH (2014) Genetic variation for resistance to Fusarium head blight in winter durum material. Crop Pasture Sci 65(1):46. https://doi.org/10.1071/CP13170

Miranda LM, Murphy JP, Marshall D, Cowger C, Leath S (2007) Chromosomal location of *Pm35*, a novel *Aegilops tauschii* derived powdery mildew resistance gene introgressed into common wheat (*Triticum aestivum* L.). Theor Appl Genet 114(8):1451–1456

Mochida K, Shinozaki K (2013) Unlocking triticeae genomics to sustainably feed the future. Plant Cell Physiol 54(12):1931–1950. https://doi.org/10.1093/pcp/pct163

Molnár-Láng M, Ceoloni C, Doležel J (eds) (2015) Alien introgression in wheat. Springer International Publishing. https://doi.org/10.1007/978-3-319-23494-6

Monat C, Padmarasu S, Lux T, Wicker T, Gundlach H, Himmelbach A, Ens J, Li C, Muehlbauer GJ, Schulman AH, Waugh R, Braumann I, Pozniak C, Scholz U, Mayer KFX, Spannagl M, Stein N, Mascher M (2019) TRITEX: chromosome-scale sequence assembly of Triticeae genomes with open-source tools. Genome Biol 20(1). https://doi.org/10.1186/s13059-019-1899-5

Mujeeb-Kazi A, Kazi AG, Dundas I, Rasheed A, Ogbonnaya F, Kishii M, Bonnett D, Wang RRC, Xu S, Chen P, Mahmood T, Bux H, Farrakh S (2013) Genetic diversity for wheat improvement as a conduit to food security. In: Advances in agronomy, vol 122. Academic Press Inc., pp 179–257. https://doi.org/10.1016/B978-0-12-417187-9.00004-8

Murphy JP, Leath S, Huynh D, Navarro RA, Shi A (1999) Registration of NC96BGTA4, NC96BGTA5, and NC96BGTA6 wheat germplasm. Crop Sci 39:883–884

Narang D, Kaur S, Steuernagel B, Ghosh S, Bansal U, Li J, Zhang P, Bhardwaj S, Uauy C, Wulff BBH, Chhuneja P (2020) Discovery and characterisation of a new leaf rust resistance gene introgressed in wheat from wild wheat *Aegilops peregrina*. Sci Rep 10(1):7573. https://doi.org/10.1038/s41598-020-64166-2

Nigro D, Blanco A, Piarulli L, Signorile MA, Colasuonno P, Blanco E, Simeone R (2022) Fine mapping and candidate gene analysis of Pm36, a wild emmer-derived powdery mildew resistance locus in durum wheat. Int J Mol Sci 23(21):13659. https://doi.org/10.3390/ijms232113659

Orlovskaya O, Dubovets N, Solovey L, Leonova I (2020) Molecular cytological analysis of alien introgressions in common wheat lines derived from the cross of *Triticum aestivum* with *T. kiharae*. BMC Plant Biol 20. https://doi.org/10.1186/s12870-020-02407-2

Perea C, de La Hoz JF, Cruz DF, Lobaton JD, Izquierdo P, Quintero JC, Raatz B, Duitama J (2016) Bioinformatic analysis of genotype by sequencing (GBS) data with NGSEP. BMC Genomics 17. https://doi.org/10.1186/s12864-016-2827-7

Petersen S, Lyerly JH, Worthington ML, Parks WR, Cowger C, Marshall DS, Brown-Guedira G, Murphy JP (2015) Mapping of powdery mildew resistance gene *Pm53* introgressed from *Aegilops speltoides* into soft red winter wheat. Theor Appl Genet 128(2):303–312

Pingali PL (2012) Green revolution: impacts, limits, and the path ahead. Proc Natl Acad Sci USA 109(31):12302–12308. https://doi.org/10.1073/pnas.0912953109

Pour-Aboughadareh A, Sanjani S, Nikkhah-Chamanabad H, Mehrvar MR, Asadi A, Amini A (2021) Identification of salt-tolerant barley genotypes using multiple-traits index and yield performance at the early growth and maturity stages. Bull Natl Res Cent 45:117. https://doi.org/10.1186/s42269-021-00576-0

Purugganan MD, Jackson SA (2021) Advancing crop genomics from lab to field. In: Nature genetics, vol 53, issue 5. Nature Research, pp 595–601. https://doi.org/10.1038/s41588-021-00866-3

Qi L, Friebe B, Zhang P, Gill BS (2007) Homoeologous recombination, chromosome engineering and crop improvement. Chromosome Res 15(1):3–19. https://doi.org/10.1007/s10577-006-1108-8

Qi LL, Pumphrey MO, Friebe B, Chen PD, Gill BS (2008) Molecular cytogenetic characterization of alien introgressions with gene Fhb3 for resistance to Fusarium head blight disease of wheat. Theor Appl Genet 117(7):1155–1166. https://doi.org/10.1007/s00122-008-0853-9

Rabanus-Wallace MT, Hackauf B, Mascher M, Lux T, Wicker T, Gundlach H, Baez M, Houben A, Mayer KFX, Guo L, Poland J, Pozniak CJ, Walkowiak S, Melonek J, Praz CR, Schreiber M, Budak H, Heuberger M, Steuernagel B, Stein N et al (2021) Chromosome-scale genome assembly provides insights into rye biology, evolution and agronomic potential. Nat Genet 53(4):564–573. https://doi.org/10.1038/s41588-021-00807-0

Rani K, Raghu BR, Jha SK, Agarwal P, Mallick N, Niranjana M, Sharma JB, Singh AK, Sharma NK, Rajkumar S, Tomar SMS, Vinod (2020) A novel leaf rust resistance gene introgressed from *Aegilops markgrafii* maps on chromosome arm 2AS of wheat. Theor Appl Genet 133(9):2685–2694

Ray DK, Ramankutty N, Mueller ND, West PC, Foley JA (2012) Recent patterns of crop yield growth and stagnation. Nat Commun 3. https://doi.org/10.1038/ncomms2296

Ray DK, Mueller ND, West PC, Foley JA (2013) Yield trends are insufficient to double global crop production by 2050. PLoS ONE 8(6). https://doi.org/10.1371/journal.pone.0066428

Reader SM, Miller TE (1991) The introduction into bread wheat of a major gene for resistance to powdery mildew from wild emmer wheat. Euphytica 53:57–60

Ren T, Jiang Q, Sun Z, Ren Z, Tan F, Yang W, Li Z (2022) Development and characterization of novel wheat-rye 1RS 1BL translocation lines with high resistance to Puccinia striiformis f. sp. tritici. Phytopathology 112(6):1310–1315. https://doi.org/10.1094/PHYTO-07-21-0313-R

Riar AK, Kaur S, Dhaliwal HS, Singh K, Chhuneja P (2012) Introgression of a leaf rust resistance gene from *Aegilops caudata* to bread wheat. J Genet 91(2):155–161

Riley R, Unrau J, Chapman V (1958) Evidence on the origin of the B genome of wheat. J Hered 49(3):91–98. https://doi.org/10.1093/oxfordjournals.jhered.a106784

Rogowsky PM, Sorrels ME, Shepherd KW, Langridge P (1993) Characterisation of wheat-rye recombinants with RFLP and PCR probes. In: Theoretical applied genetics, vol 85. Springer

Rong JK, Millet E, Manisterski J, Feldman M (2000) A new powdery mildew resistance gene: introgression from wild emmer into common wheat and RFLP-based mapping. Euphytica 115:121–126

Rowland GG, Kerber ER (1974) Telocentric mapping in hexaploid wheat of genes for leaf rust resistance and other characters derived from *Aegilops squarrosa*. Can J Genet Cytol 16:137–144

Ru Z, Juhasz A, Li D, Deng P, Zhao J,Gao L, Wang K, Keeble-Gagnere G, Yang Z, Li G, Wang D, Bose U, Colgrave M, Kong C, Zhao G, Zhang X, Liu X, Cui G, Wang Y, Niu Z, Wu L, Cui D, Jia J, Appels R, Kong X (2020) 1RS.1BL molecular resolution provides novel contributions to wheat improvement. https://doi.org/10.1101/2020.09.14.295733 (bioRxiv preprint)

Saini DK, Srivastava P, Pal N, Gupta PK (2022) Meta-QTLs, ortho-meta-QTLs and candidate genes for grain yield and associated traits in wheat (*Triticum aestivum* L.). Theor Appl Genet 135(3):1049–1081. https://doi.org/10.1007/s00122-021-04018-3

Sandve SR, Marcussen T, Mayer K, Jakobsen KS, Heier L, Steuernagel B, Wulff BBH, Olsen OA (2015) Chloroplast phylogeny of Triticum/Aegilops species is not incongruent with an ancient homoploid hybrid origin of the ancestor of the bread wheat D-genome. New Phytol 208(1):9–10. https://doi.org/10.1111/nph.13487

Sansaloni C, Franco J, Santos B, Percival-Alwyn L, Singh S, Petroli C, Campos J, Dreher K, Payne T, Marshall D, Kilian B, Milne I, Raubach S, Shaw P, Stephen G, Carling J, Pierre CS, Burgueño J, Crosa J, Pixley K (2020) Diversity analysis of 80,000 wheat accessions reveals consequences and opportunities of selection footprints. Nat Commun 11(1):4572. https://doi.org/10.1038/s41467-020-18404-w

Sarkar P, Stebbins GL (1956) Morphological evidence concerning the origin of the B genome in wheat. Am J Bot 43(4):297–304. https://doi.org/10.1002/j.1537-2197.1956.tb10494.x

Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, Connor R, Funk K, Kelly C, Kim S, Madej T, Marchler-Bauer A, Lanczycki C, Lathrop S, Lu Z, Thibaud-Nissen F, Murphy T, Phan L, Skripchenko Y, Sherry ST et al (2022) Database resources of the national center for biotechnology information. Nucleic Acids Res 50(D1):D20–D26. https://doi.org/10.1093/nar/gkab1112

Schneider A, Molnár I, Molnár-Láng M (2008) Utilisation of Aegilops (goatgrass) species to widen the genetic diversity of cultivated wheat. Euphytica 163(1):1–19. https://doi.org/10.1007/s10681-007-9624-y

Sears ER (1956) The transfer of leaf rust resistance from *Aegilops umbellulata* into wheat. Brookhaven Symp Biol 9:1–21

Sharma D, Knott DR (1966) The transfer of leaf rust resistance from Agropyron to Triticum by irradiation. Can J Genet Cytol 8:137–143

Sharma S, Xu S, Ehdaie B, Hoops A, Close TJ, Lukaszewski AJ, Waines JG (2011) Dissection of QTL effects for root traits using a chromosome arm-specific mapping population in bread wheat. Theor Appl Genet 122(4):759–769. https://doi.org/10.1007/s00122-010-1484-5

Sharma P, Chaudhary HK, Kapoor C, Manoj N, Singh K, Sood VK (2022) Molecular cytogenetic analysis of novel wheat-rye translocation lines and their characterization for drought tolerance and yellow rust resistance. Cereal Res Commun 50(4):655–665. https://doi.org/10.1007/s42976-021-00212-7

Singh K, Batra R, Sharma S, Saripalli G, Gautam T, Singh R, Pal S, Malik P, Kumar M, Jan I, Singh S, Kumar D, Pundir S, Chaturvedi D, Verma A, Rani A, Kumar A, Sharma H, Chaudhary J, Gupta PK et al (2021) WheatQTLdb: a QTL database for wheat. Mol Genet Genom 296(5):1051–1056. https://doi.org/10.1007/s00438-021-01796-9

Shi AN, Leath S, Murphy JP (1998) A major gene for powdery mildew resistance transferred to common wheat from wild einkorn wheat. Phytopathology 88(2):144–147

Steadham J, Schulden T, Kalia B, Koo DH, Gill BS, Bowden R, Yadav IS, Chhuneja P, Erwin J, Tiwari V, Rawat N (2021) An approach for high-resolution genetic mapping of distant wild relatives of bread wheat: example of fine mapping of Lr57 and Yr40 genes. Theor Appl Genet 134(8):2671–2686. https://doi.org/10.1007/s00122-021-03851-w

Steiner B, Michel S, Maccaferri M, Lemmens M, Tuberosa R, Buerstmayr H (2019) Exploring and exploiting the genetic variation of Fusarium head blight resistance for genomic-assisted breeding in the elite durum wheat gene pool. Theor Appl Genet 132(4):969–988. https://doi.org/10.1007/s00122-018-3253-9

Tilman D, Balzer C, Hill J, Befort BL (2011) Global food demand and the sustainable intensification of

agriculture. Proc Natl Acad Sci USA 108(50):20260–20264. https://doi.org/10.1073/pnas.1116437108

Tiwari VK, Wang S, Danilova T, Koo DH, Vrána J, Kubaláková M, Hribova E, Rawat N, Kalia B, Singh N, Friebe B, Doležel J, Akhunov E, Poland J, Sabir JSM, Gill BS (2015) Exploring the tertiary gene pool of bread wheat: sequence assembly and analysis of chromosome 5Mg of Aegilops geniculata. Plant J 84(4):733–746. https://doi.org/10.1111/tpj.13036

Villareal RL, Bañuelos O, Mujeeb-Kazi A, Rajaram S (1998) Agronomic performance of chromosomes 1B and T1BL.1RS near-isolines in the spring bread wheat Seri M82. Euphytica 103

Voss-Fels K, Frisch M, Qian L, Kontowski S, Friedt W, Gottwald S, Snowdon RJ (2015) Subgenomic diversity patterns caused by directional selection in bread wheat gene pools; subgenomic diversity patterns caused by directional selection in bread wheat gene pools. Plant Genome 8(2). 10.3835/p

Walkowiak S, Gao L, Monat C, Haberer G, Kassa MT, Brinton J, Ramirez-Gonzalez RH, Kolodziej MC, Delorean E, Thambugala D, Klymiuk V, Byrns B, Gundlach H, Bandi V, Siri JN, Nilsen K, Aquino C, Himmelbach A, Copetti D, Pozniak CJ et al (2020) Multiple wheat genomes reveal global variation in modern breeding. Nature 588(7837):277–283. https://doi.org/10.1038/s41586-020-2961-x

Wang J, Liu Y, Su H, Guo X, Han F (2017) Centromere structure and function analysis in wheat–rye translocation lines. Plant J 91(2):199–207. https://doi.org/10.1111/tpj.13554

Wang H, Sun S, Ge W, Zhao L, Hou B, Wang K, Lyu Z, Chen L, Xu S, Guo J, Li M, Su P, Li X, Wang G, Bo C, Fang X, Zhuang W, Cheng X, Wu J, Kong L (2020) Horizontal gene transfer of Fhb7 from fungus underlies Fusarium head blight resistance in wheat. Science 368(6493). https://doi.org/10.1126/science.aba5435

Wang L, Zhu T, Rodriguez JC, Deal KR, Dubcovsky J, McGuire PE, Lux T, Spannagl M, Mayer KFX, Baldrich P, Meyers BC, Huo N, Gu YQ, Zhou H, Devos KM, Bennetzen JL, Unver T, Budak H, Gulick PJ, Dvorak J (2021). Aegilops tauschii genome assembly Aet v5.0 features greater sequence contiguity and improved annotation. G3 Genes|Genomes|Genetics 11(12). https://doi.org/10.1093/g3journal/jkab325

Xie CJ, Sun QX, Ni ZF, Yang TM, Nevo E, Fahima T (2003) Chromosomal location of a *Triticum dicoccoides*-derived powdery mildew resistance gene in common wheat by using microsatellite markers. Theor Appl Genet 106:341–345

Yamane K, Kawahara T (2005) Intra- and interspecific phylogenetic relationships among diploid *Triticum-aegilops* species (Poaceae) based on base-pair substitutions, indels, and microsatellites in chloroplast noncoding sequences. Am J Bot 92(11):1887–1898. https://doi.org/10.3732/ajb.92.11.1887

Yao E, Blake VC, Cooper L, Wight CP, Michel S, Cagirici HB, Lazo GR, Birkett CL, Waring DJ, Jannink JL, Holmes I, Waters AJ, Eickholt DP, Sen TZ (2022) GrainGenes: a data-rich repository for small grains genetics and genomics. Database. https://doi.org/10.1093/database/baac034

Yu G, Matny O, Champouret N, Steuernagel B, Moscou MJ, Hernández-Pinzón I, Green P, Hayta S, Smedley M, Harwood W, Kangara N, Yue Y, Gardener C, Banfield MJ, Olivera PD, Welchin C, Simmons J, Millet E, Minz-Dub A, Wulff BBH (2022) Aegilops sharonensis genome-assisted identification of stem rust resistance gene Sr62. Nat Commun 13(1). https://doi.org/10.1038/s41467-022-29132-8

Zeller FJ, Gunzel G, Fischbeck G, Gerstenkorn P, Weipert D (1982) Alteration of baking properties of wheat-rye chromosome 1B/1R translocation. Getreide Mehl Brot. 36(6):141–143

Zeller FJ, Sears ER (1973) 1B/1R wheat-rye chromosome substitutions and translocations. In: Fourth international wheat genetics symposium, pp 209–221

Zeller FJ, Hsam SL (1983) Broadening the genetic variability of cultivated wheat by utilizing rye chromatin. In Sixth international wheat genetics symposium, pp 161–173

Zhang J, Liu W, Han H, Song L, Bai L, Gao Z, Zhang Y, Yang X, Li X, Gao A, Li L (2015) De novo transcriptome sequencing of *Agropyron cristatum* to identify available gene resources for the enhancement of wheat. Genomics 106(2):129–136. https://doi.org/10.1016/j.ygeno.2015.04.003

Zhang J, Ma H, Zhang J, Zhou S, Han H, Liu W, Li X, Yang X, Li L (2018) Molecular cytogenetic characterization of an *Agropyron cristatum* 6PL chromosome segment conferring superior kernel traits in wheat. Euphytica 214(11). https://doi.org/10.1007/s10681-018-2276-2

Zhang J, Hewitt TC, Boshoff WHP, Dundas I, Upadhyaya N, Li J et al (2021a) A recombined Sr26 and Sr61 disease resistance gene stack in wheat encodes unrelated NLR genes. Nat Commun 12:3378

Zhang R, Lu C, Meng X, Fan Y, Du J, Liu R, Feng Y, Xing L, Cápal P, Holušová K, Doležel J, Wang Y, Mu H, Sun B, Hou F, Yao R, Xiong C, Wang Y, Chen P, Cao A (2022) Fine mapping of powdery mildew and stripe rust resistance genes Pm5V/Yr5V transferred from *Dasypyrum villosum* into wheat without yield penalty. Theor Appl Genet 135(10):3629–3642. https://doi.org/10.1007/s00122-022-04206-9

Zhang Q, Li Y, Li Y, Fahima T, Shen Q, Xie C (2021b) Introgression of the powdery mildew resistance genes *Pm60* and *Pm60b* from *Triticum urartu* to common wheat using durum as a 'bridge.' Pathogens 11(1):25

Zhou Y, Bai S, Li H, Sun G, Zhang D, Ma F, Zhao X, Nie F, Li J, Chen L, Lv L, Zhu L, Fan R, Ge Y, Shaheen A, Guo G, Zhang Z, Ma J, Liang H, Song C-P et al (2021) Introgressing the *Aegilops tauschii* genome into wheat as a basis for cereal improvement.

Nat Plants 7(6):774–786. https://doi.org/10.1038/s41477-021-00934-w

Zhu F (2018) Triticale: nutritional composition and food uses. In: Food chemistry, vol 241. Elsevier Ltd., pp 468–479. https://doi.org/10.1016/j.foodchem.2017.09.009

Zhu T, Wang L, Rodriguez JC, Deal KR, Avni R, Distelfeld A, McGuire PE, Dvorak J, Luo MC (2019) Improved genome sequence of wild emmer wheat Zavitan with the aid of optical maps. G3 Genes, Genomes, Genetics, 9(3):619–624. https://doi.org/10.1534/g3.118.200902

Zhu T, Wang L, Rimbert H, Rodriguez JC, Deal KR, de Oliveira R, Choulet F, Keeble-Gagnère G, Tibbits J, Rogers J, Eversole K, Appels R, Gu YQ, Mascher M, Dvorak J, Luo MC (2021) Optical maps refine the bread wheat *Triticum aestivum* cv. Chinese spring genome assembly. Plant J 107(1):303–314. https://doi.org/10.1111/tpj.15289

13 Haplotype Mapping Coupled Speed Breeding in Globally Diverse Wheat Germplasm for Genomics-Assisted Breeding

Rajib Roychowdhury, Naimat Ullah, Z. Neslihan Ozturk-Gokce and Hikmet Budak

Abstract

This century is facing huge challenges such as climate change, water shortage, malnutrition, and food safety and security across the world. These challenges can only be addressed by (i) the deliberate application and utilization of cutting-edge technologies and (ii) combining/using interdisciplinary, multidisciplinary, and even transdisciplinary tools and methods. For scientists to respond to these challenges in a timely manner, it is required the adoption of new tools and technologies and then transforming the technological outcomes into "knowledge". It is highly unlikely that we could maintain or meet the demands in year 2050 unless we use scientific and technological resources effectively and efficiently. Multidisciplinary and interdisciplinary approaches combined with all available tools are integral for academic and industry programs. This chapter summarizes wheat breeding and genetics coupled with genomics and speed breeding tools to assist with crop development and improvement.

Keywords

Genomics-aided breeding · Haplotype mapping · Speed breeding · Wheat genetic resources

R. Roychowdhury
Department of Plant Pathology and Weed Research, Institute of Plant Protection, Agricultural Research Organization (ARO)—Volcani Center, Rishon Lezion, Israel

N. Ullah
Institute of Biological Sciences (IBS), Gomal University, D. I. Khan, Pakistan
e-mail: naimat@alumni.sabanciuniv.edu

Z. N. Ozturk-Gokce
Ayhan Sahenk Faculty of Agricultural Sciences and Technologies, Department of Agricultural Genetic Engineering, Nigde Omer Halisdemir University, Nigde, Turkey
e-mail: zahideneslihan_ozturk@nigde.edu.tr

H. Budak (✉)
Montana BioAgriculture Inc., Missoula, MT, USA
e-mail: hikmet.budak@icloud.com

13.1 Sustainable Increase in Global Wheat Production

Wheat (*Triticum* spp.) is a major source of carbohydrates and is used as a staple food for global inhabitants. Genetically, diverse wheat resources show variable ploidy level (diploid, tetraploid, and hexaploid) as a result of prolonged evolution and the wheat domestication process (Jordan et al. 2015). As an allopolyploid crop, wheat breeding and genetics investigations are generally considered challenging and has

R. Appels et al. (eds.), *The Wheat Genome*, Compendium of Plant Genomes,
https://doi.org/10.1007/978-3-031-38294-9_13

provided for conventional breeding approaches to be complemented by genome-assisted breeding including the genomics toolbox with the available reference genomes to deal with the highly repetitive wheat genome and to decipher genotype–phenotype associations (Varshney et al. 2021a). More specifically, the increased sophistication of sequencing technologies/interpretation has led to extensive re-sequencing of low-copy genomic regions (Nyine et al. 2019) in diverse wheat haplotype mapping populations that are managed with reduced crop-cycle through speed breeding, or fast-forward breeding, toward the wheat improvement (Varshney et al. 2021b; Jordan et al. 2022). A key requirement is to understand diverse wheat genetic resources for trait improvement, environmental adaptations, and disease resistance under ongoing climate changing scenario.

Genomics-assisted breeding (GAB) has contributed to the enhancement of germplasm and the crop/cultivar development process to characterize allelic variation for important agronomic traits associated with crop production and quality attributes as well as tolerance to abiotic and biotic stresses (Varshney et al. 2005). With the advent of genome sequencing and the inclusion of genetic-based markers in sequencing repositories, a variety of genomic tools and approaches have become accessible for use in plant breeding. These methods and techniques include GAB which is capable of assisting growers in selecting appropriate parental lines for various crossing programs in the breeding platform, which will ultimately result in the creation of genetic variation for pyramiding into breeding lines (Varshney et al. 2005). A significant variety of molecular genetic markers, such as simple sequence repeat (SSR), diversity array technology (DArT), single feature polymorphism (SFP), and single nucleotide polymorphisms (SNP), are now available, as well as inter-specific and intra-specific mapping populations (Kover et al. 2009) for chromosome sequence-aided molecular markers-based selection strategies (Akpinar et al. 2017; Maccaferri et al. 2022).

13.2 Application of Genomic Breeding (GB) to the Development of Future Crops

Several GB methods, including marker-assisted selection (MAS), marker-assisted recurrent selection (MARS), haplotype-based breeding (HBB), marker-assisted backcrossing (MABC), promotion/removal of allele through genome editing (PAGE/RAGE), and genomic selection (GS), can be used in concurrently with speed breeding to design new varieties of crops (Varshney et al. 2021a).

13.2.1 Haplotype-Based Breeding (HBB) in Wheat

Recent developments in crop genomics have sparked the development of novel technologies that aim for diversifying the procedures of plant propagative strategies by combining desired phenotypes (Fig. 13.1) with the method of haplotype construction developed using information from sequencing genotypes (Varshney et al. 2005, 2021a). Aiming for haplotype construction, various crop species have made use of large SNP data sets obtained from genomic sequence-based technologies on multiple genotypes (Varshney et al. 2005) in order to define haplotype-linked biomarkers. Haplotype construction was initially challenging for the short-read sequences obtained through the second-generation sequencing because of the lower probability of the presence of allelic variations in the form of single nucleotide polymorphism (SNP) or insertion-deletions (InDel). In contrast, the definition of haplotypes using long-read sequences has become simpler, and in many specific crop species, the information is readily available from a large number of different individuals, including using single-cell approaches, and Pacific Biosciences (PacBio) and/or Oxford Nanopore Technology (ONT) based high-quality long-read sequencing technologies that show considerably greater genomic diversity (Torkamaneh and Belzile 2022). The method for constructing

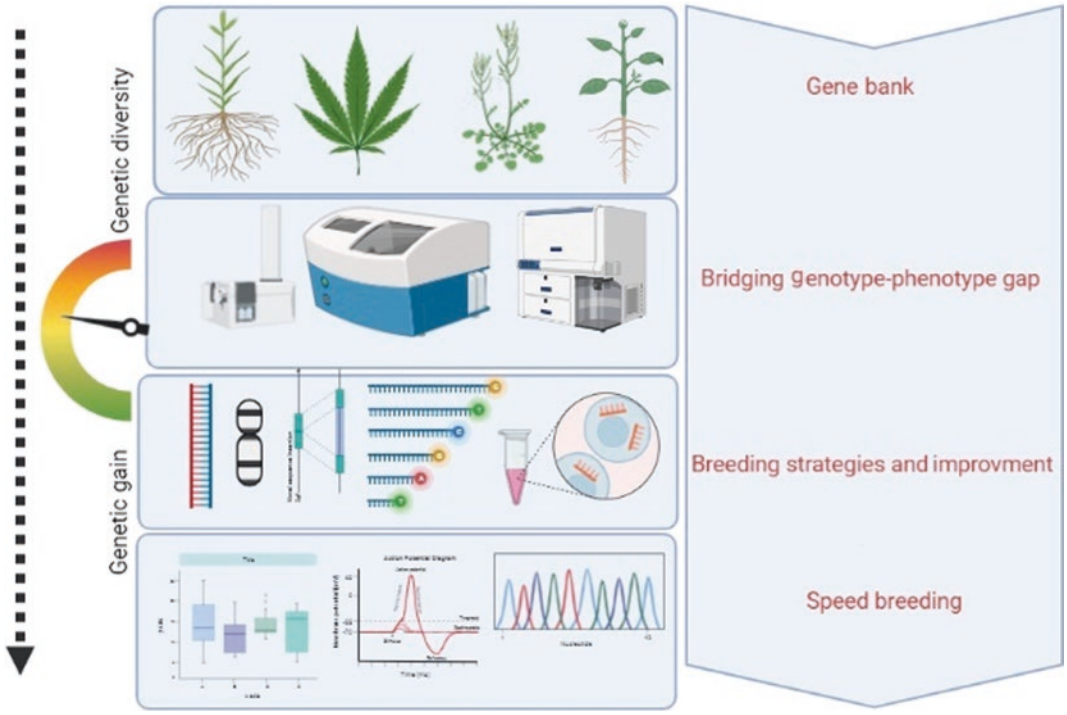


Fig. 13.1 Overview of breeding strategies for crop improvement through GAB. The image was created using BioRender (https://biorender.com/)

haplotypes using the breeding line sequencing data proceeds with the discovery and evaluation of the changes in the haplotype fingerprint using whole genome sequencing (WGS) data (Bevan et al. 2017; Bhat et al. 2021). Constructing haplotypes between adjacent SNPs on a chromosome is an alternate method that may be used to increase the genome-wide association study (GWAS) potential. Haplotypes, in this way, are particular collections of alleles that are detected on a single chromosome. They are passing throughout the generation of the population collectively, and there is a low possibility that they may recombine in the future.

Research on *Triticum* spp. has evinced that GWAS investigation based on haplotypes can be preferable to analysis based on a single marker in assessing the impacts of allelic variation (Sehgal et al. 2020) and allows HBB to produce a customized crop varieties by combining better haplotypes into a single plant, particularly novel combinational haplogroups. A wider pool of haplotype-linked genetic markers provides wheat breeders with a greater chance of developing high-performing, linkage-drag-free hybrids (Varshney et al. 2021b). The transmission of haplotypes within genetic populations must be monitored in order to pinpoint the best possible parents to cross and produce offspring with the beneficial adaptive and desired traits that are crucial for trying to create novel genetic compositions. Based on this premise, useful haplotypes have been identified by incorporating the combined results of extensive, entire, genome sequencing, and haplo-phenotyping database analysis (Bhat et al. 2021).

The construction of haplotype blocks typically makes use of the following three methods in order: (1) user-defined length, (2) sliding window, and (3) linkage disequilibrium (LD). The user-defined set length of haplotype blocks (2–15 bp) is the simplest way; however, the created haplotypes do not represent genomic factors such as crossover or LD (Sehgal et al.

2020), nor do they represent a common evolutionary process (Templeton et al. 2004). The second one is by far the most popular choice among GWAS researchers when it comes to the construction of haplotypes (Sehgal et al. 2020). This method is simple and straightforward to use; but, when neighboring SNPs are strongly linked to each other, it produces information that is redundant; hence, it is no-more helpful than using SNPs alone (Sehgal et al. 2020). It is challenging to determine the optimal window size for a genome-wide scan when LD frequencies differ throughout large genetic variants (Sehgal et al. 2020). This is similar to the previous point. In terms of finding instances of past integration in the population of interest, the LD-aided approach stands out as being the most effective (Qian et al. 2017; Sehgal et al. 2020).

According to an investigation by Brinton et al. (2020) on haplotype blocks in wheat, seven haplotypes (namely H1, H2,….,H7) were identified that included the gene *TaGW2-A* in the highly conserved genetic regions of chromosome 6A responsible for increased yield characteristics. As the two SNP markers based on the promoter regions of this gene could not discriminate the haplo-blocks, the haplotype block provided more gene-associated markers for complete reliability (Varshney et al. 2021b). Studies by Luján Basile et al. (2019) characterized haplotype blocks and GWAS in Argentinian bread wheats using genetic molecular markers and SNP profiling and revealed that several haplotype blocks span throughout the genome and including conserved genetic regions, e.g., 1BL/1RS wheat/rye translocation site on chromosome 1BS (e.g., in Chinese wheats; see Ru et al. 2020). Moreover, most of the haplotypes identified had significant effects on the yield attributes through multi-locational breeding trials. For spring wheat genetic resources, an approach of haplotype-based GWAS was targeted for epistatic interactions of multi-locational breeding trials in CIMMYT (Mexico) led by Sehgal et al. (2020). This study aimed to explore the stable genomic regions of the haplotypes for improved yield components and haplotype interactions and used LD approaches as numerous haplotype blocks were designed to span through >14 Mb of wheat genome. Haplotype-based GWAS revealed stable associations under drought stress environments with chromosomal hotspots. These studies support the need for developing genetic markers, and their deployment in agricultural crop development that are reliant on haplotypes rather than just single SNPs. Because full-genome sequencing data for the breeding lines collection in a variety of crops is expanding, it can be anticipated that the HBB method will continue to be used in the years to come (Varshney et al. 2021a).

Figure 13.1 provides a description of the integrative techniques that can be used to either add beneficial allelic variants to wheat genetic resources or remove harmful allelic variants from them in order to prepare future crop breeding techniques. The collections of germplasm that are stored in gene banks include both advantageous and detrimental impact alleles. Combining high-throughput sequencing with multi-omics assays and field phenotyping offers a valuable tool for connecting genomic variants with key phenotypes. The acquisition of knowledge about the genes that are responsible for important plant characteristics lays the path for haplotype-based genetic breeding or de novo domestication (Qian et al. 2017; Bhat et al. 2021; Varshney et al. 2021b). In this regard, speed breeding (SB) or fast-forward breeding approach will contribute to the acceleration of the advances made in crop breeding pipelines. The HBB strategy requires monitoring haplotype transfer via breeding lineages as a crucial step in creating novel genomic variants because it helps select the appropriate parents for breeding to create offspring with the desired traits. Incorporating genomic information into defining recombinants formed by mating distinct sets of parents can help simplify desired traits of interest, in particular for complex traits such as adaptation to harsh environments (Jensen et al. 2020) where it is necessary to distinguish between a correlation between different traits that are attributable to genuine linkage among the genes, or due to the pleiotropic actions of a given set

of genes (Bhat et al. 2021; Dixon et al. 2020). In the case of crops whose genomes include extensive linkage disequilibrium (LD) blocks, an HBB method becomes more pertinent since the LD blocks can be regions of conserved genetic variation.

13.2.2 Involvement of Speed Breeding in Haplotype Mapping for Wheat Genetic Resources

In plant breeding, generation time of a crop is a major factor to stabilize homozygote lines with enhanced genetic gain through hybridization and conventional breeding schemes. Some approaches such as double haploid, shuttle breeding, and tissue culture of embryo can help to minimize the generation time (Bhat et al. 2021). But to some extent, major key crops are intractable in double haploid techniques. Moreover, genetic linkages, recombination, and the lacuna of dedicated plant organ and tissue cultural infrastructure promote additional breeding avenues to fixing the genes. The development of a new and more sophisticated breeding method known as speed (fast) breeding (SB) has made it feasible to hasten agricultural innovation by shortening plant phenological cycle and gear up the progression of generational advancement (Ghosh et al. 2018; Watson et al. 2018). Speed or fast-forward breeding program deployed in several ways, such as by expanding light exposure time to the given crop species, instantaneously after it becomes available for grain harvesting, for fast propagation reduces the amount of time it takes for certain day-neutral and/or long-day plants to produce new generations (Ghosh et al. 2018; Watson et al. 2018). The basic fact in wheat SB is utilizing the early flowering period by manipulating the photoperiod (day length) and temperature (vernalization or cold requirement) under controlled condition (Ghosh et al. 2018). In this way, haplotypes and improved new varieties belonging to the same species can be developed through the synchronizing flowering time (anthesis) and introgressed into marker-assisted molecular breeding program coupled with abiotic stress tolerance (Song et al. 2022; Gahlaut et al. 2023). Under SB conditions, it could be possible to meet the flowering time of both wheat parents involved in the crossing experiments and propagation of future generations in very short time and space manner. Moreover, such accelerated generation times of this polyploid crop enable phenotypic screening of transformants for further selection and marker-aided investigation to improve grain yield, nutritional quality, improving beneficial traits, flowering time as well as adaptations to both environmental instabilities and disease pressures (Watson et al. 2018). Along with the screening of the wheat lines for abiotic and biotic stress response, SB protocols and techniques can be manipulated for rapid screening of the population even in the off season with the screening being done early in the life cycle of the plant generations (Alahmad et al. 2018; Ghosh et al. 2018). This is advantageous for breeding procedures especially for pyramiding beneficial/resistance genes for the production of climate-smart wheat. Speed breeding acts as a bridge to utilizing superior haplotype with exotic and adaptive alleles for haplotype-based breeding (HBB), genomic selection (GS), and genome editing. Using the SB approaches, accelerated generation can deliver the improved variety after going through high-throughput phenotyping, marker-assisted selection (MAS), genotyping, and sequencing (Fig. 13.2). In polyploid crops, haplotype phasing and scaffolding are becoming more advantageous as a result of increased chromosomal configuration monitoring (Zhang et al. 2019), sequencing, and Bionano Genomics (BNG) optical mapping-based genomic assemblies. SB coupled with single seed descent (SSD) for generation advancement of haplotypes and other bi- and/or multi-parental breeding populations enhances molecular marker-aided breeding (MAB) and precise genome editing for the desired trait(s).

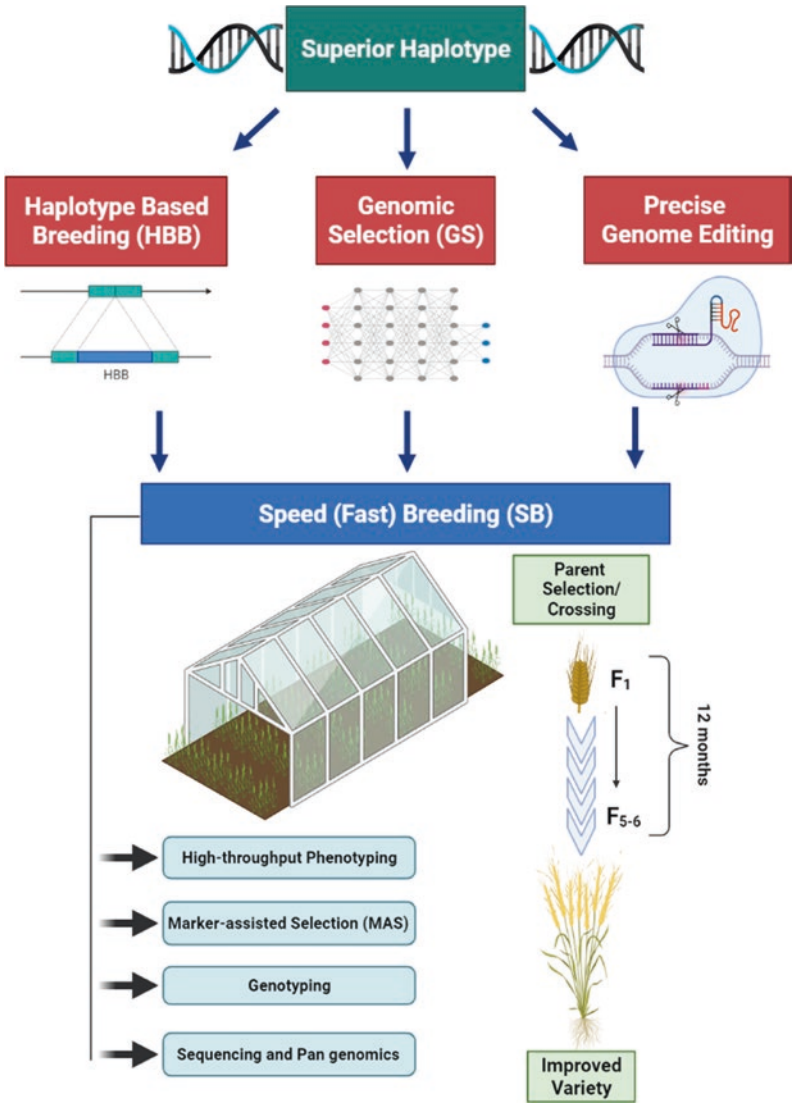


Fig. 13.2 Involvement of speed breeding in haplotype mapping to generate improved variety. This figure was prepared using BioRender application (https://biorender.com/)

13.3 Conclusion and Future Perspective

Breeding, especially breeding of main crops such as wheat, is as old as human history, and the focus on selection for mainly yield and high quality has tended to restrict the genetic diversity of modern wheat. The region in which domesticated wheat originated, namely in Mesopotamia in the Harran region of Turkey, has however a very large gene pool of Triticeae species with characteristics that provide for growth under challenging environmental conditions as well as to coping with multiple biotic factors. As detailed in Chap. 12, it is clear that these valuable abilities can be recovered in domesticated wheat varieties through alien introgression. For the present chapter, we have argued that molecular technologies can be captured in the form of haplotype mapping

combining selection based on haplotype signatures with speed breeding approaches as a primary genomics-assisted breeding strategy for complex traits.

References

Akpinar BA, Lucas S, Budak HA (2017) Large-scale chromosome-specific SNP discovery guideline. Funct Integr Genomics 17:97–105. https://doi.org/10.1007/s10142-016-0536-6

Alahmad S, Dinglasan E, Leung KM, Riaz A, Derbal N, Voss-Fels KP, Able JA, Bassi FM, Christopher J, Hickey LT (2018) Speed breeding for multiple quantitative traits in durum wheat. Plant Methods 14:36

Bevan MW, Uauy C, Wulff BBH, Zhou J, Krasileva K, Clark MD (2017) Genomic innovation for crop improvement. Nature 543(7645):346–354. https://doi.org/10.1038/NATURE22011

Bhat JA, Yu D, Bohra A, Ganie SA, Varshney RK (2021) Features and applications of haplotypes in crop breeding. Commun Biol 4(1):1–12. https://doi.org/10.1038/s42003-021-02782-y

Brinton J, Ramirez-Gonzalez RH, Simmonds J, Wingen L, Orford S, Griffiths S, Haberer G, Spannagl M, Walkowiak S, Pozniak C, Uauy C (2020) A haplotype-led approach to increase the precision of wheat breeding. Commun Biol 3(1):1–11. https://doi.org/10.1038/s42003-020-01413-2

Dixon LE, Pasquariello M, Boden SA (2020) TEOSINTE BRANCHED1 regulates height and stem internode length in bread wheat. J Exp Bot 71(16):4742–4750. https://doi.org/10.1093/JXB/ERAA252

Gahlaut V, Gautam T, Wani SH (2023) Abiotic stress tolerance in wheat (*Triticum aestivum* L.): molecular breeding perspectives. In: Wani SH, Wang D, Singh GP (eds) QTL mapping in crop improvement—present progress and future perspectives. Academic Press, pp 101–117

Ghosh S, Watson A, Gonzalez-Navarro OE, Ramirez-Gonzalez RH, Yanes L, Mendoza-Suárez M, Simmonds J, Wells R, Rayner T, Green P, Hafeez A, Hayta S, Melton RE, Steed A, Sarkar A, Carter J, Perkins L, Lord J, Tester M, Hickey LT et al (2018). Speed breeding in growth chambers and glasshouses for crop breeding and model plant research. Nature Protocols 13(12):2944–2963. https://doi.org/10.1038/S41596-018-0072-Z

Jensen SM, Svensgaard J, Ritz C (2020) Estimation of the harvest index and the relative water content—two examples of composite variables in agronomy. Eur J Agron112. https://doi.org/10.1016/J.EJA.2019.125962

Jordan KW, Wang S, Lun Y, Gardiner LJ, MacLachlan R, Hucl P, Wiebe K, Wong D, Forrest KL, Sharpe AG, Sidebottom CH et al (2015) A haplotype map of allohexaploid wheat reveals distinct patterns of selection on homoeologous genomes. Genome Biol 16(1):1–18

Jordan KW, Bradbury PJ, Miller ZR, Nyine M, He F, Fraser M, Anderson J, Mason E, Katz A, Pearce S, Carter AH (2022) Development of the wheat practical haplotype graph database as a resource for genotyping data storage and genotype imputation. G3 12(2):jkab390

Kover PX, Valdar W, Trakalo J, Scarcelli N, Ehrenreich IM, Purugganan MD, Durrant C, Mott R (2009) A multiparent advanced generation inter-cross to fine-map quantitative traits in *Arabidopsis thaliana*. PLOS Genetics 5(7):e1000551. https://doi.org/10.1371/JOURNAL.PGEN.1000551

Luján Basile SM, Ramírez IA, Crescente JM, Conde MB, Demichelis M, Abbate P, Rogers WJ, Pontaroli AC, Helguera M, Vanzetti LS (2019) Haplotype block analysis of an Argentinean hexaploid wheat collection and GWAS for yield components and adaptation. BMC Plant Biol 19(1). https://doi.org/10.1186/S12870-019-2015-4

Maccaferri M, Bruschi M, Tuberosa R (2022) Sequence-based marker assisted selection in wheat. In: Reynolds MP, Braun HJ (eds) Wheat improvement. Springer. https://doi.org/10.1007/978-3-030-90673-3_28

Nyine M, Wang S, Kiani K, Jordan K, Liu S, Byrne P, Haley S, Baenziger S, Chao S, Bowden R, Akhunov E (2019) Genotype imputation in winter wheat using first-generation haplotype map SNPs improves genome-wide association mapping and genomic prediction of traits. G3: Genes Genom Genet 9(1):125–133

Qian L, Hickey LT, Stahl A, Werner CR, Hayes B, Snowdon RJ, Voss-Fels KP (2017) Exploring and harnessing haplotype diversity to improve yield stability in crops. Front Plant Sci 8:1534. https://doi.org/10.3389/FPLS.2017.01534/BIBTEX

Ru Z, Juhasz A, Li D, Deng P, Zhao J,Gao L, Wang K, Keeble-Gagnere G, Yang Z, Li G, Wang D, Bose U, Colgrave M, Kong C, Zhao G, Zhang X, Liu X, Cui G, Wang Y, Niu Z, Wu L, Cui D, Jia J, Appels R, Kong X (2020). 1RS.1BL molecular resolution provides novel contributions to wheat improvement. https://doi.org/10.1101/2020.09.14.295733 (bioRxiv preprint)

Sehgal D, Mondal S, Crespo-Herrera L, Velu G, Juliana P, Huerta-Espino J, Shrestha S, Poland J, Singh R, Dreisigacker S (2020) Haplotype-based, genome-wide association study reveals stable genomic regions for grain yield in CIMMYT spring bread wheat. Front Genet 11:1427. https://doi.org/10.3389/FGENE.2020.589490/BIBTEX

Song Y, Duan X, Wang P, Li X, Yuan X, Wang Z, Wan L, Yang G, Hong D (2022) Comprehensive speed breeding: a high-throughput and rapid generation system for long-day crops. Plant Biotechnol J 20(1):13–15

Templeton AR, Maxwell T, Posada D, Stengård JH, Boerwinkle E, Sing CF (2004) Tree scanning: a method for using haplotype trees in phenotype/genotype association studies. Genetics 169(1):441–453. https://doi.org/10.1534/GENETICS.104.030080

Torkamaneh D, Belzile F (2022) Genome-wide association studies. Humana New York, p 371

Varshney RK, Graner A, Sorrells ME (2005) Genomics-assisted breeding for crop improvement. Trends Plant Sci 10(12):621–630. https://doi.org/10.1016/J.TPLANTS.2005.10.004

Varshney RK, Bohra A, Yu J, Graner A, Zhang Q, Sorrells ME (2021a) Designing future crops: genomics-assisted breeding comes of age. Trends Plant Sci 26(6):631–649

Varshney RK, Bohra A, Roorkiwal M, Barmukh R, Cowling WA, Chitikineni A, Lam HM, Hickey LT, Croser JS, Bayer PE, Edwards D, Crossa J, Weckwerth W, Millar H, Kumar A, Bevan MW, Siddique KHM (2021b) Fast-forward breeding for a food-secure world. Trends Genet TIG 37(12):1124–1136. https://doi.org/10.1016/J.TIG.2021.08.002

Watson A, Ghosh S, Williams MJ, Cuddy WS, Simmonds J, Rey MD, Asyraf Md Hatta M, Hinchliffe A, Steed A, Reynolds D, Adamski NM, Breakspear A, Korolev A, Rayner T, Dixon LE, Riaz A, Martin W, Ryan M, Edwards D, Hickey LT et al (2018) Speed breeding is a powerful tool to accelerate crop research and breeding. Nature Plants 4(1):23–29. https://doi.org/10.1038/S41477-017-0083-8

Zhang X, Zhang S, Zhao Q, Ming R, Tang H (2019) Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. Nature Plants 5(8):833–845. https://doi.org/10.1038/S41477-019-0487-8

14 Wheat Sequencing: The Pan-Genome and Opportunities for Accelerating Breeding

Amidou N'Diaye, Sean Walkowiak and Curtis Pozniak

Abstract

Wheat is a crucial crop globally, with widespread cultivation and significant economic importance. To ensure food security amidst the increasing human population and new production challenges, such as climate change, it is imperative to develop novel wheat varieties that exhibit better quality, higher yield, and enhanced resistance to biotic and abiotic stress. To achieve this, leveraging comprehensive genomic resources from global breeding programs can aid in identifying within-species allelic diversity and selecting optimal allele combinations for superior cultivars. While previous single-reference genome assemblies have facilitated gene discovery and whole-genome level genotype–phenotype relationship modeling, recent research on variations within the pan-genome of all individuals in a plant species underscores their significance for crop breeding. We summarize the different approaches and techniques used for sequencing the large and intricate wheat genome, while highlighting the challenge of generating high-quality reference assemblies. We discuss the computational methods for building the pan-genome and research efforts that are aimed at utilizing the wheat pan-genome in wheat breeding programs.

Keywords

Wheat breeding · Sequencing · Pan-genome · Accelerated breeding

14.1 Introduction

In the early 2000s, technological advances in DNA sequencing allowed the sequencing and the comparison of the genomes from several individuals of the same species (Medini et al. 2005). This helped fuel the notion that an individual genome is insufficient to serve as an appropriate genomic reference, since it does not capture the diversity that represents the species. The idea emerged of a "pan-genome" that encompasses the genomic information of several representative individuals. Pan-genomics was initially applied to many smaller and simple genomes of microbial species, particularly

A. N'Diaye · C. Pozniak (✉)
University of Saskatchewan, Crop Development Centre, Saskatoon, Saskatchewan, Canada
e-mail: curtis.pozniak@usask.ca

A. N'Diaye
e-mail: amidou.ndiaye@usask.ca

S. Walkowiak
Canadian Grain Commission, Grain Research Laboratory, Winnipeg, Manitoba, Canada
e-mail: sean.walkowiak@grainscanada.gc.ca

R. Appels et al. (eds.), *The Wheat Genome*, Compendium of Plant Genomes,
https://doi.org/10.1007/978-3-031-38294-9_14

to understand presence/absence variation (PAV) in genes (Medini et al. 2005). The idea of the pan-genome has since been applied to diverse species across all taxonomic kingdoms and has evolved to consider all possible variation present between genomes, including non-genic, PAV, copy number, and structural variation (Jayakodi et al. 2021). Pan-genomics has also been applied more broadly to groups of related species or genera, for "super pan-genomes." While still in its infancy, pan-genomics of crop species can be particularly valuable for harnessing genomic variants and increasing rates of crop improvement. The application of pan-genomes in crop breeding is gaining increased interest due to the importance of food security and the need for more efficient and effective breeding methods. To date, pan-genomes have been applied to the improvement of various crops, including barley, maize, rice, tomato, and soybean (Gao et al. 2019; Gui et al. 2022; Jayakodi et al. 2020; Liu et al. 2020; Shang et al. 2022; Zhao et al. 2018). Applications of pan-genomics for wheat improvement have also become possible since the completion and the public release of multiple high-quality reference genomes (Walkowiak et al. 2020).

Wheat is a crucial crop globally, with widespread cultivation and significant economic importance, supplying a fifth of global calories and protein (Dixon 2007; Shiferaw et al. 2013). To maintain food security in the context of exponential growth of the human population while facing new challenges (e.g., global warming and climate change) in production, it is essential to create new wheat varieties with increased yield, better quality, and resistance or tolerance to abiotic and biotic stress (Abberton et al. 2016; Batley and Edwards 2016). Early wheat improvement relied on traditional breeding methods, where wheat lines were phenotypically selected in field trials, which is both costly and labor intensive. As our understanding of wheat genetics improved, it became possible to identify major effect genes underlying qualitative traits and to select for these genes through marker-assisted selection (MAS, see also Chap. 9). Marker-assisted selection has been successfully applied to certain traits, particularly disease resistance (Miedaner and Korzun 2012). Unfortunately, many key traits, including yield, have a complex and polygenic determinism. Selection of quantitative traits that are more complex and are influenced by non-genic features, several genes, or gene interactions, require more advanced tools for making DNA-based selections. With the recent availability of high-quality genome assembly and gene annotations for wheat, it has been possible to apply high-throughput genotyping arrays or genotype-by-sequencing methods to gather genome-wide variation information and select for these complex traits at the whole-genome level, through genomic selection (GS) (Haile et al. 2021). Nevertheless, identifying key major effect genes as well as the mechanisms underpinning more complex traits requires a deeper understanding of the diversity of wheat and the impact of genomic variation on phenotypic traits. It is critical to understand the diversity within wheat that is available to breeders in order to make breeding more efficient, identify suitable parents to use in targeted crosses, and select for the best possible combination of genes for rapid trait enhancement.

Despite its importance for food security, the application of genomics and pan-genomics for wheat improvement has been challenged by the large size and the complexity of its genome. The genome is composed of three separated diploid subgenomes, resulting in allohexaploidy (genome AABBDD), where the 'A' subgenome was derived from *T. urartu*, the 'B' subgenome from a species related to *T. speltoides*, and the 'D' genome from *Ae. tauchii*. The genome of modern bread wheat is estimated to be 17 gigabase-pairs (Gb) in length and is composed of ~90% repetitive elements. Recent achievements in genome sequencing and assembly technologies have enabled the release of multiple wheat genomes and tools to create a pan-genome, which is inspiring a new age of wheat breeding. In this review, we explore the concept of pan-genomes and a pan-genome of wheat, the

history and evolution of the wheat genome and pan-genome, and the future outlook of wheat pan-genomics for research and applied breeding.

14.2 Motivations for Studying Pan-Genomes in Crop Breeding

During the last decade, there have been significant advancements in next-generation sequencing (NGS) technologies, which offer a direct view into DNA variation. These advancements have created numerous possibilities to investigate the connection between genotype and phenotype with greater precision than ever before. NGS has been used for various projects, including gene expression analysis, polymorphism detection, and the development of molecular markers (Barabaschi et al. 2012; Delseny et al. 2010). With the advent of affordable genome sequencing, breeders have started using NGS to sequence extensive groups of plants, which has enhanced the precision of identifying quantitative trait loci (QTL) and simplified the process of discovering genes. This has, in turn, formed the foundation for creating models to comprehend complex genotype–phenotype relationships at the whole-genome level. Over the past two decades, advancements in sequencing technologies, assembly techniques, and computational algorithms have enabled the release of genome sequences for over 700 plant species (Sun et al. 2022).

In parallel, advancements in using DNA-based tools for plant breeding, such as MAS and GS, have progressed significantly. Genomics approaches identified genomic markers associated with traits and were termed as QTL (Geldermann 1975). A single QTL can harbor many genes within the same locus (Beckmann and Soller 1983; Westman et al. 1997). MAS has been in use since the early 1990s and involves identifying genomic markers in silico, which are within causal genes for traits or are closely linked, which are then used to select individuals (Tanksley and Nelson 1996).

The development of reference genome assemblies has expedited the process of identifying candidate genes for in-demand traits. These assemblies serve as a basis for pinpointing single-nucleotide polymorphisms (SNPs), copy number variations (CNVs), and insertion–deletions (InDels) within an individual's DNA sequence. The markers were used as the basis for conducting genome-wide association studies (GWAS) and genomic selection (GS), which involve comparing diversity panels with reference genomes to identify statistical associations between markers and traits (Crossa et al. 2017; Hayes and Goddard 2010; Varshney et al. 2009). Despite providing a greater insight into the diversity of plant species, particularly at the SNP level (Gore et al. 2009; McNally et al. 2009), reference genomes cover only a limited portion of the overall genomic space of a species and are inadequate in capturing variation across every individual within a given crop species (Bayer et al. 2020). A paradigm shift is occurring due to new advancements in genomics, which now take into account the significance and amount of structural variations present in the pan-genome of crop species. This includes capturing all types of SVs such as PAVs, CNVs, and repetitive elements or TEs, present throughout the entire genome of all individuals belonging to a plant species (Danilevicz et al. 2020; Golicz et al. 2016; Tao et al. 2019). By cataloging this variation and linking it to phenotypic/trait information, it is then possible select parents and candidate wheat lines in breeding programs with more advanced knowledge and decision support tools, allowing for more efficient and targeted crop improvement.

14.3 Historical Challenges and Progress in Wheat Genome Sequencing and Assembly

Prior to the availability of NGS, whole-genome sequencing was performed using the Sanger sequencing technology. Due to a combination of several factors, including the cost and low throughput of Sanger sequencing, and the size and complexity of some large genomes, many genomes were first cloned into bacterial artificial

chromosomes (BACs) that included a few hundred thousand base-pairs per clone. This allowed for each BAC to be sequenced and assembled in parallel and then stitched together to assemble larger more complex genomes. After the release of the first human genome sequencing in 2000, which was achieved through the use of bacterial artificial chromosome (BAC) (Lander 2001; Venter et al. 2001), the *Arabidopsis* genome was the first plant genome to be sequenced using this approach. This was followed by the completion of multiple versions of the rice genome two years later (Goff et al. 2002; Yu et al. 2002). The wheat genome's larger size, almost 40 times that of rice, and its complexity, which included a high proportion of repetitive sequences and homoeologous DNA copies from three subgenomes, made it economically unfeasible to employ a standard sequencing method. To tackle this challenge, the International Wheat Genome Sequencing Consortium (IWGSC) was established in 2005. The consortium divided the immense task among 20 countries based on chromosomes and chromosome arms. The approach employed genetic stocks that could be differentiated by flow cytometry on an individual chromosome basis (Consortium et al. 2014). Physical maps and minimum tiling paths were produced by fingerprinting BAC libraries, which were subsequently sequenced and assembled (Safár et al. 2010). Although the chromosome-by-chromosome approach was adopted, it took nearly ten years to implement and was only partially accomplished for few chromosomes, including chromosome 3B (Paux et al. 2008). Due to the large size of the hexaploid wheat genome, certain researchers have opted to pursue a different approach by focusing on the genomes of related diploid species, such as *Ae. tauschii*. This species has a much smaller genome size, approximately one-third of that of hexaploid wheat (~ 4.792 Gb) and does not have any interference from homoeologous DNA copies during physical mapping and eventual sequence assembly. Despite implementing this method, the initial use of regular agarose gels made the task seem overwhelming. However, to anchor contigs, higher throughput technologies such as SNaPshot BAC fingerprinting and Illumina Infinium SNP array were utilized. It took a decade to produce the first version of the *Ae. tauschii* physical map, which involved fingerprinting 461,706 BAC clones and assembling them into 2263 contigs. Afterward, 7185 molecular markers were utilized to anchor these contigs onto a genetic map (Luo et al. 2013). Despite some success with *Ae. tauschii*, the BAC approach had limited achievement in hexaploid wheat and the approach was slowly abandoned for wheat once more advanced DNA sequencing, sequencing library preparation, and genome assembly technologies became available.

In the 2000s, wheat genome sequencing was boosted by Illumina sequencing technologies, which were able to perform short read paired-end sequencing at high depth and low cost. The sequencing was first done on the diploid ancestors of common wheat due to their smaller genome size and early challenges of applying short read data to large polyploidy genomes. The draft genome assembly for *Ae. tauschii*, the D genome donor for bread wheat, was completed using short read sequencing methods to about 90 × coverage (Jia et al. 2013). Approximately, 83.4% of the genome was covered by the assembled scaffolds, and out of these, 65.9% were identified as transposable elements (TEs). Using RNA-seq data from different tissues, a total of 43,150 protein-encoding genes were identified. A comparable approach was employed to construct the genome sequence of the A genome contributor, *T. urartu*. The assembly that was obtained had a total length of 3.92 Gb, which corresponds to 79.35% of the estimated size of the A genome (4.94 Gb). However, due to subgenome interactions and evolutionary processes spanning around 10,000 years, the genomes of the progenitors are not able to fully depict their counterparts in the common wheat genome. Therefore, the sequencing of the common wheat genome was yet to be achieved.

The first sequencing of the common wheat genome for the landrace CHINESE SPRING was accomplished using Roche 454

pyrosequencing, specifically the GS FLX Titanium and GS FLX1 platforms, which were used to sequence the wheat genome to about 5 × coverage. Sequencing of related progenitors was also performed using various platforms, such as Illumina methods for sequencing of *T. monococcum*, the A genome donor of bread wheat. Likewise, *Ae. tauschii* was sequenced using the Roche 454 sequencing platform. While whole-genome data was not yet available, cDNA sequences were sequenced from *Ae. speltoides*, which has a genome similar to the B genome. Using the SOLiD sequencing platform, additional short reads of CHINESE SPRING were generated. These yielded 95,000 predicted gene models, with most of them designated to either the A, B, or D subgenome. Despite its high degree of fragmentation, the draft genome was still considered valuable, as it was the first wheat genome available for community use (Brenchley et al. 2012).

As the IWGSC adopted the chromosome-based BAC sequencing approach, progress was consistently made. As NGS became available, it was possible to sequence the BACs using more high-throughput methods. The approach involved developing sequencing libraries from the DNA of individual chromosomes or their arms and subsequently sequencing pair-end reads on the Illumina HiSeq 2000 platform. The assembly obtained, which resembled the 454 assembly, comprised approximately 500,000 contigs with N50 values ranging from 1.7 to 8.9 kb. Its total size was 10.2 Gb. These contigs, taken together, make up 61% of the estimated hexaploid wheat genome. Predictions were made for a total of 133,090 high confidence genes, as well as 890,576 low confidence genes. Using a genetic map, just over half of the high confidence genes were assigned genetic positions (Mascher et al. 2013), allowing them to be considered within the context of the telosome-based assembly resources for each chromosome arm. This led to the completion of a draft genome assembly of wheat, known as the IWGSC chromosome survey sequence (CSS) assembly (Consortium et al. 2014).

The IWGSC also accomplished a noteworthy feat when they generated a reference-level sequence of chromosome 3B (Choulet et al. 2014). This high-quality sequence was created using a minimum tiling path consisting of 8452 BACs, spanning 774 Mb, and containing 5326 protein-coding genes as well as 85% of TEs. Additionally, a molecular-genetic map (CHINESE SPRING x RENAN) was used for long-range orientation of DNA sequences. The assembly of chromosome 3B demonstrated the success of the chromosome-based BAC sequencing strategy, although the assembly remained approximately 7% incomplete.

14.4 The Completion of a Chromosome-Scale Assembly of Hexaploid Wheat

While evidence suggested the BAC sequencing approach could work for achieving a chromosome-based wheat genome assembly, the complexity of the genome, high repeat content, high transposon activity, large genome size, and allopolyploidy were continuing to hamper assembly efforts. Meanwhile, third-generation sequencing technologies, which were created by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), surfaced and progressed quickly. These techniques produce reads with substantially longer lengths and have been extensively employed, in combination with established assembly algorithms, to construct intricate and sizable plant genomes with unparalleled precision (Cheng et al. 2021; Koren et al. 2017; Niu et al. 2022). This led to a paradigm shift away from BAC sequencing and toward the direct shotgun sequencing of the genome using more advanced sequencing technologies and assembly algorithms.

A new assembly method called MaSuRCA was used to assemble wheat using a hybrid approach that combined the strengths of both PacBio long reads, which have high error rates, and Illumina short reads, which are more accurate. This method was initially used to create a

genome assembly of *Ae. tauschii* (Zimin et al. 2017a). To obtain a comprehensive sequence coverage of the genome, a combination of sequencing methods was employed, including over 19 million PacBio reads providing approximately 38 × coverage of the D genome, 177 × coverage from Illumina HiSeq 2500 reads consisting of 200-base paired-end reads, and MiSeq reads consisting of 250-base paired-end reads. The sequencing libraries with a range of insert sizes yielded a total coverage of 200 × of the genome. The genome's quality was validated through a comparison with optical maps and BAC assemblies that were produced independently. Subsequently, the pipeline was utilized to produce the initial near-complete hexaploid wheat genome for CHINESE SPRING (Zimin et al. 2017b). Triticum 1.0 was a genome assembly consisting of 829,839 contigs with a total size of 17.05 Gb, with a contig and scaffold N50 of 76.3 kb and 101.2 kb, respectively. Another method involved assembling long reads directly with the FALCON assembler, which produced FALCON Trit1.0 with a size of 12.94 Gb. Although this version was shorter than the MaSuRCA-assembled version, it had a longer contig N50 of 215.3 kb. Using the genome alignment tool MUMmer (Kurtz et al. 2004), the combination of Triticum 1.0 and Trit1.0 resulted in a final assembly that spans almost the entire wheat genome, with a size of 15.3 Gb and a contig N50 of 232.6 kb.

At the same time, an alternative approach was also taken to create the CHINESE SPRING genome assembly using short reads (Clavijo et al. 2017). The approach involved 1.1 billion 250-bp paired-end reads (33 × genome coverage) from CHINESE SPRING short insert libraries, and 68 × coverage of long insert libraries, yielding the TGACv1 version of the wheat genome assembly. This version spanned 13.43 Gb and accounted for over 78% of the wheat genome. In addition to the improved assembly, strand-specific Illumina RNA-seq and PacBio full-length cDNAs were combined to achieve better annotation. Although chromosome-level assembly was not attained, this new wheat genome assembly was now available for the broader scientific community to utilize, bringing the prospect of a high-quality reference genome into focus.

Shortly thereafter, a breakthrough was made with the release of new short read assemblers. NRGene's DeNovoMagic (NRGene, Ness Ziona, Israel) algorithm and the TRITEX pipeline (Monat et al. 2019) for short read assemblies demonstrated that a shotgun whole-genome sequencing approach could be achieved when combining different Illumina library sizes and preparation methods. The AABB genome of wild emmer wheat (WEW), which represents the reference-level genome of polyploid wheat, was produced through the utilization of the DeNovoMagic algorithm (Avni et al. 2017). By sequencing on Illumina HiSeq 2500 machines, a total of 2.1 Terabase-pairs were generated, comprising 176 × genome coverage reads from five libraries. The insert sizes in the libraries ranged from 450 bp to 10 kb. The scaffolds were then consolidated using a high-density molecular-genetic linkage map and additional reads from a three-dimensional (3D) conformation capture Hi-C library. Ultimately, the final assembly was 10.5 Gb, accounting for 87.5% of the predicted tetraploid wheat genome. The annotation of 110,544 gene models provided strong evidence for the high quality of this genome assembly. Among these models, 58.8% (65,012) were identified as high confidence gene models, while the remaining 41.2% were of low confidence. This assembly successfully captured 98.4% of the total expected gene sets of WEW, as verified by BUSCO (Simão et al. 2015). Additionally, it was utilized for identifying the genes that played a role in the early domestication of wheat, as reported by Avni et al. (2017). After the completion of the WEW genome, bread wheat genome sequencing efforts quickly pivoted toward the same shotgun genomics approach. The successful completion of the bread wheat genome IWGSC RefSeq v1.0 was achieved using a combination of similar techniques and software. According to Consortium et al. (2018), DeNovoMAGIC2 utilized the complete genome as the primary framework and incorporated various sources of data such as physical maps,

genotyping-by-sequencing data, and Hi-C data. The common wheat genome was assembled into 21 pseudomolecules at the chromosome scale, which were assigned to the subgenomes A, B, and D. This resulted in a genome assembly with a super-scaffold N50 of 22.8 Mb, and total length of 14.5 Gb. Using a similar assembly approach, the genome sequencing of durum wheat (DW) was completed shortly after (Maccaferri et al. 2019).

14.5 Progress Toward a Wheat Pan-Genome

In 2018, the IWGSC released the first reference-quality genome sequence for the wheat landrace CHINESE SPRING, which marked a significant change in the use of genomics as a research tool for wheat. The publication enabled the wider research community to have easy access to this tool (Consortium et al. 2018). The CHINESE SPRING genome assembly was a major milestone in wheat genomics research, and within a few years, it has already laid the foundation for countless studies dissecting the genome to understand wheat biology. However, CHINESE SPRING shares only a distant ancestral connection with the majority of current wheat varieties. Additionally, due to the considerable diversity present within the species, a single genome sequence is insufficient for fully representing its genetic makeup. Additional pan-genome information is required to identify new genetic diversity that can enhance traits and understand the mechanism behind the traits present in elite wheat cultivars. Fortunately, with new short-read assembly algorithms capable of shotgun sequencing, the path forward to additional genomes would no longer be a technical limitation.

Choosing crop genotypes for pan-genome analysis is a challenging task as the objective is to encompass a wide range of genetic variations using a limited number of representative genotypes for the particular species. This selection procedure necessitates the acquisition of genome-wide genotypic data from either entire genebank collections or representative subgroups that cover all significant germplasm groups within the species. Recent reports have described several genebank genomics studies on rice (Wang et al. 2018), barley (Milner et al. 2019), and wheat (Juliana et al. 2019). Soleimani et al. (2020) have described different methods that can be used to choose core sets for pan-genome analysis. One tool that aims to maximize diversity, representativeness, and allelic richness of core sets is Core Hunter (De Beukelaer et al. 2018). It achieves this by using various algorithms that operate on genetic distance matrices. To further customize the selection process, clustering of the diversity space through principal component analysis (Patterson et al. 2006) or model-based ancestry estimation (Alexander et al. 2009) can be used. Pan-genome panels offer the possibility of incorporating not only cultivated plant varieties but also wild progenitors or ancestors of polyploid species. For example, teosinte as a wild progenitor of maize and wild emmer or *Aegilops tauschii* as progenitors of wheat. These wild relatives are valuable out-groups and represent diversity available in the secondary and tertiary gene pools. These relatives could be used to determine the ancestral states for SVs or because of their significance in introgression breeding (Harlan and de Wet 1971). Besides emphasizing on incorporating diverse global varieties in a crop, a pan-genome initiative might also choose specific genotypes that have a significant role in breeding and genetics. These could comprise founder genotypes of breeding programs, experimental population parents (Yu et al. 2008), or genotypes that can be genetically modified (Jain et al. 2019; Schreiber et al. 2020) to optimize the advantages for both research and breeding communities. These chosen accessions will serve as reference genotypes for future functional and genetic studies in pan-genomic research.

The International 10+Wheat Genomes Project (www.10wheatgenomes.com) was established in 2019 with the goal of creating reference-quality genome assemblies for at least ten diverse bread wheat cultivars. Using

genomic diversity analysis of 3800 wheat samples, ten wheat lines were chosen and sequenced utilizing Illumina short read sequencing technologies, and then assembled using NRGene's DeNovoMagic algorithm (NRGene, Ness Ziona, Israel). Subsequently, all these assemblies were organized into subgenome-aware pseudomolecules with the aid of Hi-C technology (van Berkum et al. 2010). Additionally, five other wheat varieties were also sequenced and assembled to the scaffold level using separate short-read assembly algorithms established at the Earlham Institute (Norwich, UK).

A gene projection strategy was implemented and applied to all assemblies to evaluate and compare the gene content of the newly sequenced lines in a fair and consistent manner, given the lack of genome-specific transcriptome data available at that time. This strategy involved using the CHINESE SPRING reference gene models and transferring them to all assemblies. Differences in gene content among the 10+wheat reference genomes were observed, likely due to the complex breeding histories of the selected lines. These variations in gene content were found to be linked with adaptation to different environments and with efforts to enhance grain yield, quality, and resistance to abiotic and biotic stresses. Significant structural rearrangements and introgressions from wild relatives were observed upon comparing the pseudomolecule structures of the reference sequences. This underscores the importance of having multiple reference genomes of quality (at pseudomolecule level) instead of relying on resequencing approaches, as only chromosome-level assemblies can provide information on large- and small-scale structural rearrangements with a high degree of resolution and accuracy. The study conducted by Walkowiak et al. (2020) illustrates how the wheat pan-genomes can be utilized to study causal genes for traits, as the genomes were used to uncover the gene *Sm1*, known for conferring resistance against midge. With the availability of recently sequenced and compiled wheat reference genomes, there is an unprecedented opportunity to identify functional genes and enhance wheat breeding. The subsequent phase of the project will involve generating de novo gene predictions for all chromosome-scale assemblies using extensive transcriptome data. These data will offer a comprehensive understanding of the functional and regulatory arrangement of the wheat pan-genome.

While the 10+Wheat Genomes Project provided the first insights into the wheat pan-genome, sequencing and assembly methods continued to evolve. Throughput increased for both PacBio and ONT sequencing platforms, leading to additional genome assemblies (Aury et al. 2022). Further, PacBio released its HiFi sequencing method based on circular consensus sequencing, which significantly improved sequencing and assembly accuracy. These long and accurate sequencing reads have led to the highest-quality genome assemblies of wheat achieved thus far. With the upcoming release of new long read sequencing technologies with high accuracy and output, such as the Revio platform from PacBio, it is expected that additional genomes for wheat will be released in the coming years. While no longer constrained by technological limitations in genome sequencing and assembly, the next chapter begins for integrating these data into a functional pan-genome that will drive future research and breeding.

14.6 A Functional Pan-Genome for Wheat Research and Applied Breeding

Pan-genome construction is the process of creating a comprehensive set of genetic information from a collection of related genomes. It is a complex task, requiring the use of multiple approaches and techniques. It involves assembling and annotating all genomic information and variants, can be used to understand genome and gene evolution, discover new genes and alleles, and investigate gene–gene interaction networks.

To construct a pan-genome, two primary methods can be utilized, whole-genome assembly and comparative genomics. Whole-genome

assembly involves assembling all of the reads from a collection of genomes into a single, contiguous genome. The steps for whole-genome assembly are well-documented (Jung et al. 2020). The approach is most appropriate for genomes that are closely related and possess significant sequence similarity. It offers the benefit of an all-encompassing perspective on the species' genetic variation, but it is often restricted by the number of genomes that can be sequenced. Comparative genomics (Pop et al. 2004), on the other hand, involves comparing and contrasting multiple genomes to identify shared and unique components. This method is most suitable for more distantly related genomes with lower sequence similarity.

The ability to assemble high-quality reference genomes for numerous plants simultaneously has been made possible by recent advancements in sequencing technologies and bioinformatic tools. Despite this progress, it is still challenging to perform combined analysis of multiple genomes or a subset of genomes and provide readily accessible genetic information to end-users, such as researchers and breeders (Li et al. 2020b). The comparison, analysis, and visualization of multiple reference genomes and their diversity necessitate powerful and specialized computational strategies and tools. De novo assembly, iterative assembly, and graph-based assembly methods have been employed to construct pan-genomes (Li et al. 2014; Liu and Tian 2020).

14.6.1 De Novo Assembly

Constructing a pan-genome can be achieved through the de novo assembly of genomes from multiple individuals, followed by comparative analysis to identify variant types and classify them as core or flexible genome components. This approach has been discussed by Mahmoud et al. (2019). Technological advancements in sequencing and assembly methods have enabled the generation of high-quality, chromosome-level plant genomes, including telomere-to-telomere genome assemblies (Miga et al. 2020). However, generating accurate genome assemblies can be costly, especially for large plant genomes, and may not be practical when dealing with hundreds of reference genomes for a single species (Hurgobin and Edwards 2017). Nevertheless, the 10+Wheat Genomes Project was successful at the construction of several chromosome-scale assemblies. Along with these genomes were tools to visualize haplotype blocks representing shared or unique regions between the assemblies (http://www.crop-haplotypes.com/) (Brinton et al. 2020). Likewise, many of the wheat genomes had major introgressions or large structural variants, which could be visualized using synteny viewers (https://kiranbandi.github.io/10wheatgenomes/, http://10wheatgenomes.plantinformatics.io/).

14.6.2 Iterative Assembly

The iterative assembly approach differs from de novo assembly in that it commences with the creation of a single-reference genome, which is then used as a framework for the sequential alignment of reads from other samples. Any unmapped reads are subsequently assembled and incorporated into the reference genome to form a non-redundant pan-genome (Golicz et al. 2016). This technique is less expensive than de novo assembly since low sequencing depths can be used for each sample, allowing for the pooling of numerous samples. Nevertheless, the iterative assembly method may struggle to handle genomes that contain many repeat regions and is not capable of detecting large structural variations that cannot be covered by individual short reads (Jiao and Schneeberger 2017). Resequencing and iterative assembly methods have been applied to wheat (Montenegro et al. 2017; Watson-Haigh et al. 2018). However, evidence suggests that wheat has a very plastic genome due to its allopolyploidy and has abundant PAV, CNV, and SV that are important for trait variation www.10wheatgenomes.com, (Nilsen et al. 2020). Therefore, iterative assembly approaches, particularly low-coverage reference-based analyses, are highly limiting when exploring wheat pan-genomics.

14.6.3 Graph-Based Assembly

Pan-genomes can also be constructed using graphs. The most commonly used graph for this purpose is the compacted de Bruijn graph, which integrates genetic information from different accessions of a species (Chikhi et al. 2016; Li et al. 2020a). In contrast, the bi-directed variation graphs capture genetic variations throughout a population and identify their potential positions on a reference genome. Compared to traditional linear genomes, graph-based pan-genomes have been shown to significantly mitigate reference bias (Garrison et al. 2018). However, graph-based pan-genomes are challenging to construct and apply due to several factors, including the intricate nature of plant genomes with their high repeat content and polyploidy. Additionally, there is a shortage of common downstream analysis tools and visualization techniques for the graph, which further adds to the limitations. Despite these challenges, graph-based genomes have strengths compared to other methods and may have more widespread applications for wheat research and breeding in the future, particularly as tools for graph-based assembly of more complex genomes improve.

14.6.4 Pan-Genome Annotation and Other Pan-Omics

Once the pan-genome has been assembled, there are several techniques that can be used to annotate it. One technique is to use gene prediction software to identify genes in the pan-genome. This can be done using homology-based or de novo gene prediction algorithms. There is a plethora of ab initio gene prediction software (Scalzitti et al. 2020), including Augustus (Stanke and Morgenstern 2005), Genscan (Burge and Karlin 1997), GeneID (Parra et al. 2000), GlimmerHMM (Majoros et al. 2004), and Snap (Korf 2004). Another technique to annotate the pan-genome is to use comparative genomics to identify conserved or novel gene families. This involves comparing the genomes of different species to identify shared and unique components. By comparing gene sequences between two species, it is possible to identify regions of similarity that may indicate similar functions. In wheat, comparative genomics has been used for identifying resistance genes (Marchal et al. 2020) and uncovering the molecular basis of nitrogen-use efficiency (Shi et al. 2022). In addition to annotating the gene space, there is increasing interest in expanding the annotation of the pan-genome to include the dynamics of gene expression (pan-transcriptomics), epigenomic modifications (epipan-genomics), as well as interaction networks between variants as well as genes, and associating these directly with biological traits. Such a complete atlas of biological information will equip researchers and breeders with unprecedented tools for wheat research and improvement.

14.6.5 Applying the Pan-Genome to Breeding

After constructing and annotating the pan-genome, the subsequent step involves utilizing it for crop enhancement. The effectiveness of next-generation breeding technologies, such as transgenics and CRISPR-Cas9 gene editing, has been proven for wheat (Nilsen et al. 2020). However, regulatory challenges exist that may limit the widespread adoption of these methods for delivering new wheat cultivars. As a result, wheat breeding will likely involve generating biparental populations and screening for progeny for some time to come. Gene discovery has certainly benefitted from the availability of pan-genomics resources for wheat, facilitating marker discovery that can be applied to MAS and making screening of parental lines and progeny more efficient (www.10wheatgenomes.com). With the availability of more genome assemblies that are representative of the genes and genomic variants that can be used in breeding, the need to generate additional high-quality genomes will likely lessen as genomes can be

imputed based on lower coverage haplotype information; for example, from genotype-by-sequencing or high-throughput SNP arrays (Alipour et al. 2019). Having genomic information available for the parental materials being used in crosses, even if imputed, will allow for breeders to make stronger associations between traits of interests and variants within the genome, allowing for more efficient and targeted genomic-based selections to be made in their resulting progeny through GS.

14.7 Conclusion and Future Directions

Owing to its ability to identify novel genetic variations that can enhance crucial traits, the pan-genome serves as a valuable asset for crop breeding, specifically in wheat. Through consistent pan-genome research in crops, more robust and productive varieties are expected to be developed, resulting in benefits for farmers and consumers worldwide. While it is difficult to predict all possible future applications of pan-genomics to wheat breeding, the resources are now available to innovate. With recent advances in GS, artificial intelligence, and deep learning, one can only imagine the possibilities when applying these tools to pan-genomics, particularly if the pan-genomes are well annotated and have associated phenotypic data generated through applied breeding. This may not only be able to predict the performance of parents or offspring but could potentially help optimize designer genomes for specific purposes, environments, or stresses.

References

Abberton M, Batley J, Bentley A, Bryant J, Cai H, Cockram J, Costa de Oliveira A, Cseke LJ, Dempewolf H, De Pace C, Edwards D, Gepts P, Greenland A, Hall AE, Henry R, Hori K, Howe GT, Hughes S, Humphreys M, Lightfoot D, Marshall A, Mayes S, Nguyen HT, Ogbonnaya FC, Ortiz R, Paterson AH, Tuberosa R, Valliyodan B, Varshney RK, Yano M (2016) Global agricultural intensification during climate change: a role for genomics. Plant Biotechnol J 14:1095–1098

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19:1655–1664

Alipour H, Bai G, Zhang G, Bihamta MR, Mohammadi V, Peyghambari SA (2019) Imputation accuracy of wheat genotyping-by-sequencing (GBS) data using barley and wheat genome references. PLoS ONE 14:e0208614

Aury J-M, Engelen S, Istace B, Monat C, Lasserre-Zuber P, Belser C, Cruaud C, Rimbert H, Leroy P, Arribat S, Dufau I, Bellec A, Grimbichler D, Papon N, Paux E, Ranoux M, Alberti A, Wincker P, Choulet F (2022) Long-read and chromosome-scale assembly of the hexaploid wheat genome achieves high resolution for research and breeding. GigaScience 11

Avni R, Nave M, Barad O, Baruch K, Twardziok SO, Gundlach H, Hale I, Mascher M, Spannagl M, Wiebe K, Jordan KW, Golan G, Deek J, Ben-Zvi B, Ben-Zvi G, Himmelbach A, MacLachlan RP, Sharpe AG, Fritz A, Ben-David R, Budak H, Fahima T, Korol A, Faris JD, Hernandez A, Mikel MA, Levy AA, Steffenson B, Maccaferri M, Tuberosa R, Cattivelli L, Faccioli P, Ceriotti A, Kashkush K, Pourkheirandish M, Komatsuda T, Eilam T, Sela H, Sharon A, Ohad N, Chamovitz DA, Mayer KFX, Stein N, Ronen G, Peleg Z, Pozniak CJ, Akhunov ED, Distelfeld A (2017) Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. Science (New York, NY) 357:93

Barabaschi D, Guerra D, Lacrima K, Laino P, Michelotti V, Urso S, Valè G, Cattivelli L (2012) Emerging knowledge from genome sequencing of crop species. Mol Biotechnol 50:250–266

Batley J, Edwards D (2016) The application of genomics and bioinformatics to accelerate crop improvement in a changing climate. Curr Opin Plant Biol 30:78–81

Bayer PE, Golicz AA, Scheben A, Batley J, Edwards D (2020) Plant pan-genomes are the new reference. Nature Plants 6:914–920

Beckmann JS, Soller M (1983) Restriction fragment length polymorphisms in genetic improvement: methodologies, mapping and costs. Theor Appl Genet 67:35–43

Brenchley R, Spannagl M, Pfeifer M, Barker GLA, D'Amore R, Allen AM, McKenzie N, Kramer M, Kerhornou A, Bolser D, Kay S, Waite D, Trick M, Bancroft I, Gu Y, Huo N, Luo M-C, Sehgal S, Gill B, Kianian S, Anderson O, Kersey P, Dvorak J, McCombie WR, Hall A, Mayer KFX, Edwards KJ, Bevan MW, Hall N (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. Nature 491:705

Brinton J, Ramirez-Gonzalez RH, Simmonds J, Wingen L, Orford S, Griffiths S, Haberer G, Spannagl M, Walkowiak S, Pozniak C, Uauy C, Wheat Genome P (2020) A haplotype-led approach to increase

the precision of wheat breeding. Communications Biology 3:712

Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. Journal of molecular biology 268

Cheng H, Concepcion GT, Feng X, Zhang H, Li H (2021) Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods 18:170–175

Chikhi R, Limasset A, Medvedev P (2016) Compacting de Bruijn graphs from sequencing data quickly and in low memory. Bioinformatics 32:i201–i208

Choulet F, Alberti A, Theil S, Glover N, Barbe V, Daron J, Pingault L, Sourdille P, Couloux A, Paux E, Leroy P, Mangenot S, Guilhot N, Le Gouis J, Balfourier F, Alaux M, Jamilloux V, Poulain J, Durand C, Bellec A, Gaspin C, Safar J, Dolezel J, Rogers J, Vandepoele K, Aury J-M, Mayer K, Berges H, Quesneville H, Wincker P, Feuillet C (2014) Structural and functional partitioning of bread wheat chromosome 3B. Science (New York, NY) 345:1249721

Clavijo BJ, Venturini L, Schudoma C, Accinelli GG, Kaithakottil G, Wright J, Borrill P, Kettleborough G, Heavens D, Chapman H, Lipscombe J, Barker T, Lu F-H, McKenzie N, Raats D, Ramirez-Gonzalez RH, Coince A, Peel N, Percival-Alwyn L, Duncan O, Trösch J, Yu G, Bolser DM, Namaati G, Kerhornou A, Spannagl M, Gundlach H, Haberer G, Davey RP, Fosker C, Palma FD, Phillips AL, Millar AH, Kersey PJ, Uauy C, Krasileva KV, Swarbreck D, Bevan MW, Clark MD (2017) An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. Genome Res 27:885–896

Consortium IWGS, Mayer KF, Rogers J, Doležel J, Pozniak C, Eversole K, Feuillet C, Gill B, Friebe B, Lukaszewski AJ (2014) A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. Science (New York, NY) 345:1251788

Consortium IWGS, Appels R, Eversole K, Stein N, Feuillet C, Keller B, Rogers J, Pozniak CJ, Choulet F, Distelfeld A (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science (New York, NY) 361:eaar7191

Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D, De Los CG, Burgueño J, González-Camacho JM, Pérez-Elizalde S, Beyene Y (2017) Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci 22:961–975

Danilevicz MF, Tay Fernandez CG, Marsh JI, Bayer PE, Edwards D (2020) Plant pangenomics: approaches, applications and advancements. Curr Opin Plant Biol 54:18–25

De Beukelaer H, Davenport GF, Fack V (2018) Core Hunter 3: flexible core subset selection. BMC Bioinformatics 19:203

Delseny M, Han B, Hsing Y (2010) High throughput DNA sequencing: the new sequencing revolution. Plant Science: An International Journal of Experimental Plant Biology 179:407–422

Dixon J (2007) The economics of wheat; Research challenges from field to fork. Wheat production in stressed environments. In: Proceedings of international wheat conference, 7; Mar de Plata (Argentina); 27 Nov–2 Dec 2005. ^ TWheat production in stressed environments Proceedings of International Wheat Conference, 7; Mar de Plata (Argentina); 27 Nov-2 Dec 2005^ ABuck, HT Nisi, JE Salomon, N^ ADordrecht (Netherlands)^ BSpringer^ C2007

Gao L, Gonda I, Sun H, Ma Q, Bao K, Tieman DM, Burzynski-Chang EA, Fish TL, Stromberg KA, Sacks GL, Thannhauser TW, Foolad MR, Diez MJ, Blanca J, Canizares J, Xu Y, van der Knaap E, Huang S, Klee HJ, Giovannoni JJ, Fei Z (2019) The tomato pangenome uncovers new genes and a rare allele regulating fruit flavor. Nat Genet 51:1044–1051

Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, Jones W, Garg S, Markello C, Lin MF, Paten B, Durbin R (2018) Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat Biotechnol 36:875–879

Geldermann H (1975) Investigations on inheritance of quantitative characters in animals by gene markers I. Methods. Theoretical and Applied Genetics 46:319–330

Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, Hadley D, Hutchison D, Martin C, Katagiri F, Lange BM, Moughamer T, Xia Y, Budworth P, Zhong J, Miguel T, Paszkowski U, Zhang S, Colbert M, Sun WL, Chen L, Cooper B, Park S, Wood TC, Mao L, Quail P, Wing R, Dean R, Yu Y, Zharkikh A, Shen R, Sahasrabudhe S, Thomas A, Cannings R, Gutin A, Pruss D, Reid J, Tavtigian S, Mitchell J, Eldredge G, Scholl T, Miller RM, Bhatnagar S, Adey N, Rubano T, Tusneem N, Robinson R, Feldhaus J, Macalma T, Oliphant A, Briggs S (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science (New York, NY) 296:92–100

Golicz AA, Bayer PE, Barker GC, Edger PP, Kim H, Martinez PA, Chan CKK, Severn-Ellis A, McCombie WR, Parkin IA (2016) The pangenome of an agronomically important crop plant *Brassica oleracea*. Nat Commun 7:1–8

Gore MA, Chia J-M, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL, Peiffer JA, McMullen MD, Grills GS, Ross-Ibarra J, Ware DH, Buckler ES (2009) A first-generation haplotype map of maize. Science (New York, NY) 326:1115–1117

Gui S, Wei W, Jiang C, Luo J, Chen L, Wu S, Li W, Wang Y, Li S, Yang N, Li Q, Fernie AR, Yan J (2022) A pan-Zea genome map for enhancing maize improvement. Genome Biol 23:178

Haile TA, Walkowiak S, N'Diaye A, Clarke JM, Hucl PJ, Cuthbert RD, Knox RE, Pozniak CJ (2021) Genomic

prediction of agronomic traits in wheat using different models and cross-validation designs. Theor Appl Genet 134:381–398

Harlan JR, de Wet JMJ (1971) Toward a rational classification of cultivated plants. Taxon 20:509–517

Hayes B, Goddard M (2010) Genome-wide association and genomic selection in animal breeding. Genome 53:876–883

Hurgobin B, Edwards D (2017) SNP discovery using a pangenome: has the single reference approach become obsolete? Biology 6(1):21. https://doi.org/10.3390/biology6010021. PMID: 28287462; PMCID: PMC5372014

Jain R, Jenkins J, Shu S, Chern M, Martin JA, Copetti D, Duong PQ, Pham NT, Kudrna DA, Talag J, Schackwitz WS, Lipzen AM, Dilworth D, Bauer D, Grimwood J, Nelson CR, Xing F, Xie W, Barry KW, Wing RA, Schmutz J, Li G, Ronald PC (2019) Genome sequence of the model rice variety KitaakeX. BMC Genomics 20:905

Jayakodi M, Padmarasu S, Haberer G, Bonthala VS, Gundlach H, Monat C, Lux T, Kamal N, Lang D, Himmelbach A, Ens J, Zhang X-Q, Angessa TT, Zhou G, Tan C, Hill C, Wang P, Schreiber M, Boston LB, Plott C, Jenkins J, Guo Y, Fiebig A, Budak H, Xu D, Zhang J, Wang C, Grimwood J, Schmutz J, Guo G, Zhang G, Mochida K, Hirayama T, Sato K, Chalmers KJ, Langridge P, Waugh R, Pozniak CJ, Scholz U, Mayer KFX, Spannagl M, Li C, Mascher M, Stein N (2020) The barley pan-genome reveals the hidden legacy of mutation breeding. Nature 588:284–289

Jayakodi M, Schreiber M, Stein N, Mascher M (2021) Building pan-genome infrastructures for crop plants and their use in association genetics. DNA Research 28:dsaa030

Jia J, Zhao S, Kong X, Li Y, Zhao G, He W, Appels R, Pfeifer M, Tao Y, Zhang X, Jing R, Zhang C, Ma Y, Gao L, Gao C, Spannagl M, Mayer KFX, Li D, Pan S, Zheng F, Hu Q, Xia X, Li J, Liang Q, Chen J, Wicker T, Gou C, Kuang H, He G, Luo Y, Keller B, Xia Q, Lu P, Wang J, Zou H, Zhang R, Xu J, Gao J, Middleton C, Quan Z, Liu G, Wang J, International Wheat Genome Sequencing C, Yang H, Liu X, He Z, Mao L, Wang J (2013) Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation. Nature 496:91

Jiao W-B, Schneeberger K (2017) The impact of third generation genomic technologies on plant genome assembly. Curr Opin Plant Biol 36:64–70

Juliana P, Poland J, Huerta-Espino J, Shrestha S, Crossa J, Crespo-Herrera L, Toledo FH, Govindan V, Mondal S, Kumar U, Bhavani S, Singh PK, Randhawa MS, He X, Guzman C, Dreisigacker S, Rouse MN, Jin Y, Pérez-Rodríguez P, Montesinos-López OA, Singh D, Mokhlesur Rahman M, Marza F, Singh RP (2019) Improving grain yield, stress resilience and quality of bread wheat using large-scale genomics. Nature Genetics

Jung H, Ventura T, Chung JS, Kim W-J, Nam B-H, Kong HJ, Kim Y-O, Jeon M-S, Eyun S-i (2020) Twelve quick steps for genome assembly and annotation in the classroom. PLoS Comput Biol 16:e1008325

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res 27:722–736

Korf I (2004) Gene finding in novel genomes. BMC Bioinformatics 5:59

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL (2004) Versatile and open software for comparing large genomes. Genome Biol 5:1–9

Lander ES (2001) Initial sequencing and analysis of the human germane. Nature 409:860–921

Li H, Feng X, Chu C (2020a) The design and construction of reference pangenome graphs with minigraph. Genome Biol 21:1–19

Li H, Feng X, Chu C (2020b) The design and construction of reference pangenome graphs with minigraph. Genome Biol 21:265

Li Y-h, Zhou G, Ma J, Jiang W, Jin L-g, Zhang Z, Guo Y, Zhang J, Sui Y, Zheng L, Zhang S-s, Zuo Q, Shi X-h, Li Y-f, Zhang W-k, Hu Y, Kong G, Hong H-l, Tan B, Song J, Liu Z-x, Wang Y, Ruan H, Yeung CKL, Liu J, Wang H, Zhang L-j, Guan R-x, Wang K-j, Li W-b, Chen S-y, Chang R-z, Jiang Z, Jackson SA, Li R, Qiu L-j (2014) De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nat Biotechnol 32:1045–1052

Liu Y, Du H, Li P, Shen Y, Peng H, Liu S, Zhou GA, Zhang H, Liu Z, Shi M, Huang X, Li Y, Zhang M, Wang Z, Zhu B, Han B, Liang C, Tian Z (2020) Pan-genome of wild and cultivated soybeans. Cell 182:162-176.e113

Liu Y, Tian Z (2020) From one linear genome to a graph-based pan-genome: a new era for genomics. Science China Life Sciences 63:1938–1941

Luo MC, Gu YQ, You FM, Deal KR, Ma Y, Hu Y, Huo N, Wang Y, Wang J, Chen S, Jorgensen CM, Zhang Y, McGuire PE, Pasternak S, Stein JC, Ware D, Kramer M, McCombie WR, Kianian SF, Martis MM, Mayer KF, Sehgal SK, Li W, Gill BS, Bevan MW, Simkova H, Dolezel J, Weining S, Lazo GR, Anderson OD, Dvorak J (2013) A 4-gigabase physical map unlocks the structure and evolution of the complex genome of *Aegilops tauschii*, the wheat D-genome progenitor. Proc Natl Acad Sci USA 110:7940–7945

Maccaferri M, Harris NS, Twardziok SO, Pasam RK, Gundlach H, Spannagl M, Ormanbekova D, Lux T, Prade VM, Milner SG (2019) Durum wheat genome highlights past domestication signatures and future improvement targets. Nat Genet 51:885–895

Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ (2019) Structural variant calling: the long and the short of it. Genome Biol 20:1–14

Majoros WH, Pertea M, Salzberg SL (2004) TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics 20:2878–2879

Marchal C, Wheat Genome P, Haberer G, Spannagl M, Uauy C (2020) Comparative genomics and functional studies of wheat BED-NLR loci. Genes (Basel) 11

Mascher M, Muehlbauer GJ, Rokhsar DS, Chapman J, Schmutz J, Barry K, Muñoz-Amatriaín M, Close TJ, Wise RP, Schulman AH, Himmelbach A, Mayer KFX, Scholz U, Poland JA, Stein N, Waugh R (2013) Anchoring and ordering NGS contig assemblies by population sequencing (POPSEQ). Plant J 76:718–727

McNally KL, Childs KL, Bohnert R, Davidson RM, Zhao K, Ulat VJ, Zeller G, Clark RM, Hoen DR, Bureau TE, Stokowski R, Ballinger DG, Frazer KA, Cox DR, Padhukasahasram B, Bustamante CD, Weigel D, Mackill DJ, Bruskiewich RM, Rätsch G, Buell CR, Leung H, Leach JE (2009) Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. Proc Natl Acad Sci 106:12273–12278

Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R (2005) The microbial pan-genome. Curr Opin Genet Dev 15:589–594

Miedaner T, Korzun V (2012) Marker-assisted selection for disease resistance in wheat and barley breeding. Phytopathology® 102:560–566

Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, Brooks S, Howe E, Porubsky D, Logsdon GA, Schneider VA, Potapova T, Wood J, Chow W, Armstrong J, Fredrickson J, Pak E, Tigyi K, Kremitzki M, Markovic C, Maduro V, Dutra A, Bouffard GG, Chang AM, Hansen NF, Wilfert AB, Thibaud-Nissen F, Schmitt AD, Belton J-M, Selvaraj S, Dennis MY, Soto DC, Sahasrabudhe R, Kaya G, Quick J, Loman NJ, Holmes N, Loose M, Surti U, Ra R, Graves Lindsay TA, Fulton R, Hall I, Paten B, Howe K, Timp W, Young A, Mullikin JC, Pevzner PA, Gerton JL, Sullivan BA, Eichler EE, Phillippy AM (2020) Telomere-to-telomere assembly of a complete human X chromosome. Nature 585:79–84

Milner SG, Jost M, Taketa S, Mazón ER, Himmelbach A, Oppermann M, Weise S, Knüpffer H, Basterrechea M, König P, Schüler D, Sharma R, Pasam RK, Rutten T, Guo G, Xu D, Zhang J, Herren G, Müller T, Krattinger SG, Keller B, Jiang Y, González MY, Zhao Y, Habekuß A, Färber S, Ordon F, Lange M, Börner A, Graner A, Reif JC, Scholz U, Mascher M, Stein N (2019) Genebank genomics highlights the diversity of a global barley collection. Nat Genet 51:319–326

Monat C, Padmarasu S, Lux T, Wicker T, Gundlach H, Himmelbach A, Ens J, Li C, Muehlbauer GJ, Schulman AH, Waugh R, Braumann I, Pozniak C, Scholz U, Mayer KFX, Spannagl M, Stein N, Mascher M (2019) TRITEX: chromosome-scale sequence assembly of Triticeae genomes with open-source tools. Genome Biol 20:284

Montenegro JD, Golicz AA, Bayer PE, Hurgobin B, Lee H, Chan C-KK, Visendi P, Lai K, Doležel J, Batley J, Edwards D (2017) The pangenome of hexaploid bread wheat. Plant J 90:1007–1013

Nilsen KT, Walkowiak S, Xiang D, Gao P, Quilichini TD, Willick IR, Byrns B, N'Diaye A, Ens J, Wiebe K (2020) Copy number variation of TdDof controls solid-stemmed architecture in wheat. Proc Natl Acad Sci 117:28708–28718

Niu S, Li J, Bo W, Yang W, Zuccolo A, Giacomello S, Chen X, Han F, Yang J, Song Y (2022) The Chinese pine genome and methylome unveil key features of conifer evolution. Cell 185(204–217):e214

Parra G, Blanco E, Guigó R (2000) GeneID in drosophila. Genome Res 10:511–515

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLOS Genetics 2:e190

Paux E, Sourdille P, Salse J, Saintenac C, Choulet F, Leroy P, Korol A, Michalak M, Kianian S, Spielmeyer W (2008) A physical map of the 1-gigabase bread wheat chromosome 3B. Science (New York, NY) 322:101–104

Pop M, Phillippy A, Delcher AL, Salzberg SL (2004) Comparative genome assembly. Brief Bioinform 5:237–248

Safár J, Simková H, Kubaláková M, Číhalíková J, Suchánková P, Bartos J, Dolezel J (2010) Development of chromosome-specific BAC resources for genomics of bread wheat. Cytogenet Genome Res 129:211–223

Scalzitti N, Jeannin-Girardon A, Collet P, Poch O, Thompson JD (2020) A benchmark study of ab initio gene prediction methods in diverse eukaryotic organisms. BMC Genomics 21:293

Schreiber M, Mascher M, Wright J, Padmarasu S, Himmelbach A, Heavens D, Milne L, Clavijo BJ, Stein N, Waugh R (2020) A genome assembly of the barley 'transformation reference' cultivar golden promise. G3 Genes, Genomes, Genetics 10:1823–1827

Shang L, Li X, He H, Yuan Q, Song Y, Wei Z, Lin H, Hu M, Zhao F, Zhang C, Li Y, Gao H, Wang T, Liu X, Zhang H, Zhang Y, Cao S, Yu X, Zhang B, Zhang Y, Tan Y, Qin M, Ai C, Yang Y, Zhang B, Hu Z, Wang H, Lv Y, Wang Y, Ma J, Wang Q, Lu H, Wu Z, Liu S, Sun Z, Zhang H, Guo L, Li Z, Zhou Y, Li J, Zhu Z, Xiong G, Ruan J, Qian Q (2022) A super pan-genomic landscape of rice. Cell Res 32:878–896

Shi X, Cui F, Han X, He Y, Zhao L, Zhang N, Zhang H, Zhu H, Liu Z, Ma B, Zheng S, Zhang W, Liu J, Fan X, Si Y, Tian S, Niu J, Wu H, Liu X, Chen Z, Meng D, Wang X, Song L, Sun L, Han J, Zhao H, Ji J, Wang Z, He X, Li R, Chi X, Liang C, Niu B, Xiao J, Li J, Ling H-Q (2022) Comparative genomic and transcriptomic analyses uncover the molecular basis of high nitrogen-use efficiency in the wheat cultivar Kenong 9204. Mol Plant 15:1440–1456

Shiferaw B, Smale M, Braun H-J, Duveiller E, Reynolds M, Muricho G (2013) Crops that feed the world

10. Past successes and future challenges to the role played by wheat in global food security. Food Security 5:291–317

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212

Soleimani B, Lehnert H, Keilwagen J, Plieske J, Ordon F, Naseri Rad S, Ganal M, Beier S, Perovic D (2020) Comparison between core set selection methods using different illumina marker platforms: a case study of assessment of diversity in wheat. Frontiers in Plant Science 11

Stanke M, Morgenstern B (2005) AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res 33:W465-467

Sun Y, Shang L, Zhu Q-H, Fan L, Guo L (2022) Twenty years of plant genome sequencing: achievements and challenges. Trends Plant Sci 27:391–401

Tanksley SD, Nelson JC (1996) Advanced backcross QTL analysis: a method for the simultaneous discovery and transfer of valuable QTLs from unadapted germplasm into elite breeding lines. Theor Appl Genet 92:191–203

Tao Y, Zhao X, Mace E, Henry R, Jordan D (2019) Exploring and exploiting pan-genomics for crop improvement. Mol Plant 12:156–169

van Berkum NL, Lieberman-Aiden E, Williams L, Imakaev M, Gnirke A, Mirny LA, Dekker J, Lander ES (2010) Hi-C: a method to study the three-dimensional architecture of genomes. Journal of Visualized Experiments: JoVE

Varshney RK, Nayak SN, May GD, Jackson SA (2009) Next-generation sequencing technologies and their implications for crop genetics and breeding. Trends Biotechnol 27:522–530

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA (2001) The sequence of the human genome. Science (New York, NY) 291:1304–1351

Walkowiak S, Gao L, Monat C, Haberer G, Kassa MT, Brinton J, Ramirez-Gonzalez RH, Kolodziej MC, Delorean E, Thambugala D (2020) Multiple wheat genomes reveal global variation in modern breeding. Nature 588:277–283

Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, Li M, Zheng T, Fuentes RR, Zhang F, Mansueto L, Copetti D, Sanciangco M, Palis KC, Xu J, Sun C, Fu B, Zhang H, Gao Y, Zhao X, Shen F, Cui X, Yu H, Li Z, Chen M, Detras J, Zhou Y, Zhang X, Zhao Y, Kudrna D, Wang C, Li R, Jia B, Lu J, He X, Dong Z, Xu J, Li Y, Wang M, Shi J, Li J, Zhang D, Lee S, Hu W, Poliakov A, Dubchak I, Ulat VJ, Borja FN, Mendoza JR, Ali J, Li J, Gao Q, Niu Y, Yue Z, Naredo MEB, Talag J, Wang X, Li J, Fang X, Yin Y, Glaszmann J-C, Zhang J, Li J, Hamilton RS, Wing RA, Ruan J, Zhang G, Wei C, Alexandrov N, McNally KL, Li Z, Leung H (2018) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature 557:43–49

Watson-Haigh NS, Suchecki R, Kalashyan E, Garcia M, Baumann U (2018) DAWN: a resource for yielding insights into the diversity among wheat genomes. BMC Genomics 19:941

Westman A, Kresovich S, Callow J, Ford-Lloyd B, Newbury J (1997) Biotechnology and plant genetic resources: conservation and use

Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, Cao M, Liu J, Sun J, Tang J, Chen Y, Huang X, Lin W, Ye C, Tong W, Cong L, Geng J, Han Y, Li L, Li W, Hu G, Huang X, Li W, Li J, Liu Z, Li L, Liu J, Qi Q, Liu J, Li L, Li T, Wang X, Lu H, Wu T, Zhu M, Ni P, Han H, Dong W, Ren X, Feng X, Cui P, Li X, Wang H, Xu X, Zhai W, Xu Z, Zhang J, He S, Zhang J, Xu J, Zhang K, Zheng X, Dong J, Zeng W, Tao L, Ye J, Tan J, Ren X, Chen X, He J, Liu D, Tian W, Tian C, Xia H, Bao Q, Li G, Gao H, Cao T, Wang J, Zhao W, Li P, Chen W, Wang X, Zhang Y, Hu J, Wang J, Liu S, Yang J, Zhang G, Xiong Y, Li Z, Mao L, Zhou C, Zhu Z, Chen R, Hao B, Zheng W, Chen S, Guo W, Li G, Liu S, Tao M, Wang J, Zhu L, Yuan L, Yang H (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science (New York, NY) 296:79–92

Yu J, Holland JB, McMullen MD, Buckler ES (2008) Genetic design and statistical power of nested association mapping in maize. Genetics 178:539–551

Zhao Q, Feng Q, Lu H, Li Y, Wang A, Tian Q, Zhan Q, Lu Y, Zhang L, Huang T, Wang Y, Fan D, Zhao Y, Wang Z, Zhou C, Chen J, Zhu C, Li W, Weng Q, Xu Q, Wang Z-X, Wei X, Han B, Huang X (2018) Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. Nat Genet 50:278–284

Zimin AV, Puiu D, Luo MC, Zhu T, Koren S, Marçais G, Yorke JA, Dvořák J, Salzberg SL (2017a) Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. Genome Res 27:787–792

Zimin AV, Puiu D, Salzberg SL, Hall R, Kingan S, Clavijo BJ (2017b) The first near-complete assembly of the hexaploid bread wheat genome, *Triticum aestivum*. GigaScience 6

15 Genome-Wide Resources for Genetic Locus Discovery and Gene Functional Analysis in Wheat

James Cockram

Abstract

Future wheat production faces considerable challenges, such as how to ensure on-farm yield gains across agricultural environments that are increasingly challenged by factors such as soil erosion, environmental change and rapid changes in crop pest and disease profiles. Within the context of crop improvement, the ability to identify, track and deploy specific combinations of genes tailored for improved crop performance in target environments will play an important role in ensuring future sustainable wheat production. In this chapter, a range of germplasm resources and populations are reviewed can be exploited for genetic locus discovery, characterisation and functional analysis in wheat. These include experimental populations constructed from two or more parents, association mapping panels and artificially mutated populations. Efficient integration of the knowledge gained from exploiting such resources with other emerging breeding approaches and technologies, such as high-throughput field phenotyping, multi-trait ensemble phenotypic weighting and genomic selection, will help underpin future breeding for improved crop performance, quality and resilience.

Keywords

Multi-parent populations · Plant genetic diversity · Sustainable crop production · Nested association mapping (NAM) · Multi-parent advanced generation intercross (MAGIC) · Targeting Induced Local Lesions in Genomes (TILLING)

15.1 Gene Discovery in the Context of Wheat Improvement and Breeding

If you compare two bread wheat (*Triticum aestivum* L.) cultivars, the chances are that you will find differences between them—and lots of them. Whether these differences are for agronomic traits, such as resistance to disease, for quality traits such as those important for bread making, or for a range of morphological traits such as those used to uniquely 'describe' a variety during varietal registration (Jones et al. 2013), such variation is abundant. It is the heritable component of these observable differences

J. Cockram (✉)
NIAB, 93 Lawrence Weaver Road, Cambridge, UK
e-mail: james.cockram@niab.com

R. Appels et al. (eds.), *The Wheat Genome*, Compendium of Plant Genomes,
https://doi.org/10.1007/978-3-031-38294-9_15

that is exploited via breeding to deliver new improved wheat varieties and deals with the complexities of pleiotropic effects resulting from the process. The question as to how best to do this is not a straightforward one. To give a simplified example, phenotypic selection for underlying combinations of genes and alleles that result in increased grain number per ear may result in fewer ears overall. Similarly, increasing the grain protein is often associated with a reduction in overall grain yield in wheat (Simmonds 1995; White et al. 2022) and other crop species (e.g. Dudley 2007), and increasing leaf size is thought to result in larger, but less dense stomata (Zanella et al. 2022). As the principal breeding target, grain yield represents the sum of all interacting genetic/epigenetic, environmental and management factors that occur from sowing to harvest. Selection for grain yield works well, with breeders having consistently delivered ~ 1% genetic gains per year in wheat yield potential over recent decades (e.g. Mackay et al. 2011). To some extent, wheat breeding practices focus on delivering performance under the assessment criteria and carefully managed growth conditions used by national bodies to determine subsets of the 'best' varieties marketed at a given time. In the United Kingdom (UK), for example, the annual AHDB 'Recommended List' provides performance data for such varietal subsets to help farmers choose which varieties to grow (www.ahdb.co.uk/knowledge-library/reccommended-lists-for-cereals-and-oilseeds-rl). However, on-farm wheat yields are increasingly falling behind the genetic potential of the varieties grown. Termed the 'yield-gap', and observed in wheat growing areas across the world (Senapati et al. 2022), this is likely to be due to the cost–benefit and practical considerations and trade-offs that take place under commercial farm conditions. Future wheat production will face additional challenges such as environmental change, soil degradation, increasing energy and input costs, and the effects of political conflict or instability. Thus, wheat genetic improvement will increasingly need to focus on yield stability under sub-optimal, fluctuating or unpredictable growth environments—delivered within the context of more sustainable food production systems. As the development of new wheat varieties is a relatively lengthy process (typically taking around 10 years), all available tools must be exploited to meet these challenges. As underpinning technologies advance, the ability to identify specific wheat genes or genetic loci, and understand how they function and interact within the context of crop performance, will play an increasingly important role towards delivering future wheat.

15.2 Genetic Variation in Hexaploid Bread Wheat: Luck, Bottlenecks and Breeding

If the foundation of gene discovery is heritable variation, then before exploring the germplasm and genomic resources currently in use to accelerate gene discovery and functional analysis in wheat, it is first worth briefly considering the history behind current wheat genetic variation. Collectively, the natural genetic variation present in modern day wheat represents the culmination of the speciation, domestication and breeding events and processes that have occurred in its past. Human selection and interventions have affected the wheat genome and the variation it contains, starting from its first origins in Neolithic farmers' fields, up to the current day. However, variation at the DNA level is not so evenly distributed across the bread wheat genome. To some extent, this is due to the order, age and nature of the polyploidisation events that occurred during its speciation. The bread wheat genome is hexaploid ($2n = 6x = 42$), which means it consists of three subgenomes that have merged via inter-species hybridisation events during its evolutionary history (reviewed by Levy and Feldman 2022; Fig. 15.1). Notably, the most recent event was a spontaneous hybridisation around 9000 years ago between the tetraploid progenitor of pasta wheat (the AA and BB subgenome donor) and a diploid wild wheat relative that grew alongside it called 'goat grass'

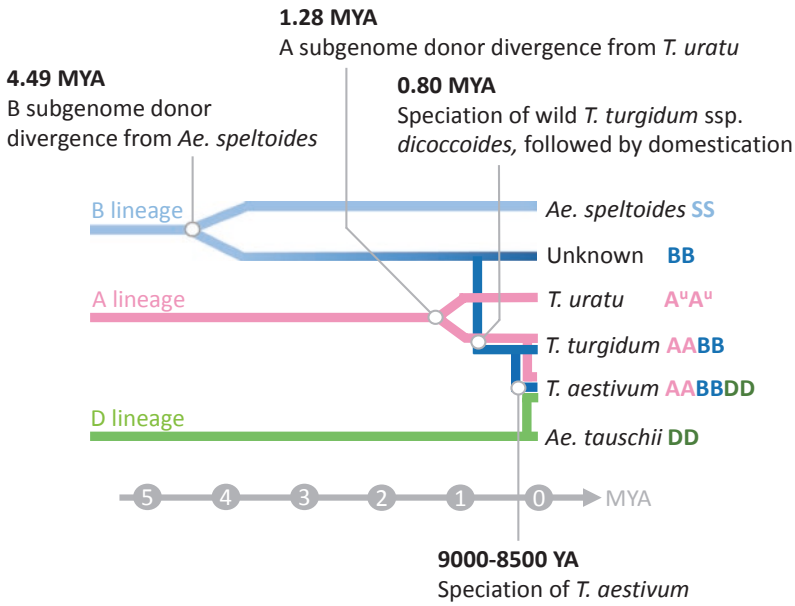


Fig. 15.1 Evolutionary history of hexaploid bread wheat (*Triticum aestivum*) from its diploid and tetraploid donors progenitors. The unknown or extinct wheat B subgenome donor is a derivative of the S-genome species of the section *Sitopsis*, which includes diploid *Ae. speltoides* (diploid SS genome), *Ae. bicornis* (S^bS^b), *Ae. longissima* (S^lS^l), *Ae. searsii* (S^sS^s) and *Ae. sharonensis* ($S^{sh}S^{sh}$). MYA = millions of years ago

(*Aegilops tauschii* Coss., DD subgenome donor) to create hexaploid bread wheat (AABBDD). Due to this event being rare, recent, and having occurred in a restricted *Ae. tauschii* sub-population close to the Caspian Sea (Wang et al. 2013), little D subgenome variation was captured, and there has been comparatively little time for genetic variation to subsequently accumulate via spontaneous mutation. The effect of this is evident in genetic analyses of bread wheat varieties from across the world (e.g. Wang et al. 2014; Walkowiak et al. 2020; Mellers et al. 2020), where D subgenome variation within genes is typically one-third to one-tenth of that seen on the A and B subgenomes. Consistent throughout the wheat subgenomes however is that gene density and gene variation are lower across the centromeric and adjacent pericentromeric chromosomal regions than in the remaining more distal chromosomal positions (IWGSC 2018). These centromeric and pericentromeric regions are associated with higher frequency of transposable elements (IWGSC 2018), higher levels of epigenetic modifications to DNA and histones associated with heterochromatin (tightly packed DNA), and lower genetic recombination (Gardner et al. 2016; Gardiner et al. 2019), which together are thought to result in the restricted rates of genome evolution observed in these regions (Akhunov et al. 2003). Against this genomic backdrop, in the ~9000 years since the speciation of bread wheat has been accumulating natural mutations which have either been retained or lost along the way due to a combination of selection, drift and geneflow. Such shifts in variation have underpinned the many generations of 'on-farm' selection that occurred from Neolithic times up until the advent of industrial breeding approaches at the end of the nineteenth century. Accordingly, wheat genetic variation was modulated across this time period by the interplay between human selection, be it conscious (such as selection for larger grains) or unconscious (such as selection for photoperiod insensitive lines; Jones and Lister 2022), and environmental factors such as prevailing climate and disease pressures. This ongoing domestication process resulted in the numerous locally

adapted 'landraces' that were grown across the world's wheat growing regions up until the end of the 1800s. Early breeders exploited these sources of genetic diversity by systematically selecting and evaluating such landraces, as well as the crosses made between them. The outcomes of this history are still evident in modern wheat varieties, as these first breeding programmes commonly exploited the landraces that were locally adapted to their regions at the time. Evidence of this history can be seen in modern day wheat. For example, genetic marker analysis of wheat from around the world shows clustering of Chinese landraces and cultivars in genetic diversity space (Cavanagh et al. 2013), while in an analysis of 180 UK varieties released since the year 2000, almost 90% include genetic contributions from the old Ukrainian landrace OSTKA-GALICYJSKA and the Mediterranean landrace from which the early UK variety SQUAREHEAD was developed (Fradgley et al. 2019). Over the years, there have been concerns that the industrial breeding era has resulted in genetic bottlenecks in numerous crops, and that this has restricted genetic diversity in modern wheat. While there are many approaches to measure genetic diversity loss (reviewed by Khoury et al. 2021), for wheat it is clear that more genetic diversity was present in the landraces versus pure-line bred cultivars (e.g. Winfield et al. 2018). The assumption of loss of diversity when within the modern breeding period is not necessarily so apparent, with changes in diversity depending on multiple factors, including the time period and region studied. One factor that has been noted is a reduction in genetic diversity at and soon after the introduction of the 'Green Revolution' semi-dwarfing genes across all international breeding programmes from the 1960s onwards (see Chap. 11). However, recent studies of on-farm wheat diversity indicate that at a national level, growers may now actually deploy a much more diverse portfolio of cultivars than was used 100 years ago. For example, in the USA the number of major commercially grown wheat cultivars has increased progressively, increasing fivefold from 1919 (33 cultivars) to 2019 (186 cultivars) with pedigree-based diversity measures of 1353 commercial USA varieties grown across this period indicating this increase in cultivar diversity is likely linked to increased genetic diversity (Chai et al. 2022). In the UK, combining measures of relatedness based on shared parentage (kinship), weighted by the proportional yearly acreage of cultivars over the last 30 years, found an increasing trend in the resulting landscape diversity index (Fradgley 2022). While the dominance of a very low number of varieties across national cropping landscapes is not as common as it once was (such as the use of cv. CAPPELLE-DESPREZ across more than 50% of the UK cropping area in the 1960s; Srinivasan et al. 2003), this is not necessarily the case throughout the wheat growing regions of the world. For example, between 2005 and 2010 the cultivar WYALKATCHEM represented more than 30% of the Australian wheat area sown, while more recently cv. MACE represented over 65% of the wheat cropping area in both 2015 and 2016 (Phan et al. 2020). Notably, these recent examples of low Australian landscape scale cultivar diversity are set against a wider background of a reduction in Australian wheat genetic diversity post Green Revolution (Joukhadar et al. 2017) and highlight the potential vulnerability of such landscape scale cultivar predominance to changes in pest and environmental pressures.

15.2.1 Systematic Broadening of the Wheat Genepool as Wild Wheats Are Deployed

A longstanding concern is that breeding results in loss of genetic diversity—however, as noted above this assumption is not a given. A good example in cereals is the maize long-term selection experiment, where continuous genetic gains within a closed population in response to selection for seed protein and oil content were observed across the 100-year programme, with no significant loss in genetic diversity (Dudley 2007). Presumably, this was achieved via continued selection for genetic loci of small additive

effect, as well as the fixing of epistatic interactions (i.e. instances where the allele of one gene hides or masks the phenotype of another gene) as additive effects. It is thus feasible to optimise existing variation present in wheat cultivars into new combinations, and to bring in additional genetic and functional diversity from systematic introgression and analysis of chromosomal regions originating from landraces and species related to wheat. When present in otherwise elite wheat genetic backgrounds, the chromosomal segments present in such 'wilder wheats' can often provide agronomic performance gains, despite the possible negative impacts of such chromosomal tracts (due, for example, to linkage drag or local effects on genetic recombination). Reminiscent of the activities at the start of the industrial breeding age, initiatives across the world are once again systematically screening variation captured in wheat landraces and are now supported by modern genetics, genomics, experimental population designs and analysis approaches. For example, the Watkins bread wheat landrace collection of 826 accessions from 32 countries has been genotyped using 41 microsatellite markers (Wingen et al. 2014), and selected accessions from across the genetic diversity space crossed to an elite spring cultivar to create a series of bi-parental genetic mapping populations (Wingen et al. 2017), termed a nested association mapping (NAM) panel. The benefits afforded by 'wilder wheats' created via introgressions from wheat relatives are illustrated by the UK cultivar ROBIGUS. Released in the UK in 2020, ROBIGUS delivered high yields and contained particularly novel genetics derived from a wheat wild relative (Gardner et al. 2016) and has been frequently used in the pedigrees of subsequent UK varieties (Fradgley et al. 2019)—without associated loss of wheat cultivar genetic diversity at landscape scale (Fradgley 2022). Genomic analyses now show that the presence of introgressions from wheat relatives is relatively common (e.g. Cheng et al. 2019; Keilwagen et al. 2022; Pont et al. 2019; Przewieslik-Allen et al. 2021; Scott et al. 2020a). Indeed, introgressions often underlie genomic regions conferring agronomically important traits—particularly disease resistance (Aktar-Uz-Zaman et al. 2017). For example, resistance to the wheat fungal disease yellow rust conferred by *Yr34* originated from a region of chromosome 5A introgressed over 200 years ago from einkorn wheat (*T. monococcum* L. ssp. *monococcum*; Chen et al. 2021), and still confers field resistance in the US (Chen et al. 2021) and UK (Bouvet et al. 2022b). The long breeding history of use and utility of introgression from wheat relatives is exemplified by the extensive use since the late 1980s of synthetic hexaploid wheats in the international wheat breeding programme run by the International Maize and Wheat Improvement Center (CIMMYT) (Das et al. 2016; see also Chap. 11). Synthetic hexaploid wheats address the lack of genetic diversity on the wheat D subgenome by recreating the ancient hybridization event between tetraploid wheat and *Ae. tauschii*. This is undertaken via inter-specific crosses followed either by embryo rescue, chromosome doubling (Li et al. 2018a, b) or use of specific cytogenetic stocks (Othmeni et al. 2022). While more than 1200 synthetic wheats have been generated by CIMMYT, historically these have sampled a relatively narrow range of *Ae. tauschii* diversity from the eastern Fertile Crescent. Systematic broadening of the diversity sampled in synthetic wheats is now being undertaken at pre-breeding initiatives at NIAB in the UK, where D subgenome *Ae. tauschii* genetic diversity from across its natural eco-geographic range is being captured in new synthetics and backcrossed into elite cultivars (Gaurav et al. 2022). While this and other initiatives (e.g. Zhou et al. 2021) are providing new sources of D subgenome genetic variation for breeding, similar approaches are systematically bringing in additional diversity from wheat A and B subgenome donors via the creation of inter-specific hybrids and subsequent backcrossing. For example, the generation of backcross-derived progenies from crosses between 59 diverse accessions of tetraploid *T. turgidum* ssp. *durum* with elite spring wheat cv. PARAGON (see also Chap. 8). Introgressions

into elite wheat varieties from more distantly related diploid and polyploid grass species are also being generated, including *Ambylopyrum muticum* (TT genome, Coombes et al. 2022) and *Thinopyrum* species (Li and Wang 2009; Grewal et al. 2018; Li et al. 2018a, b; Cseh et al. 2019; Baker et al. 2020). The utility of genetic loci originating from the tertiary wheat genepool has begun to lead in the identification of the underlying genes and genetic variants; for example, the wheat *Fhb7* locus conferring resistance to the fungal disease *Fusarium* head blight, and which originated from a *Th. elongatum* introgression, has been shown to encode an amino acid transferase that detoxifies toxins produced by the infecting fungus (Wang et al. 2020).

15.3 Current Genome-Wide Genotyping Approaches for Wheat

The history of speciation, domestication and breeding outlined above has shaped the heritable variation present across the wheat genome. At the DNA level, this variation includes changes to singe nucleotides (single nucleotide polymorphisms, SNPs), or via other rearrangements that typically involve DNA double strand break repair such as DNA insertions or deletions (InDels), gene copy number variation (CNV) and larger chromosomal rearrangements such as translocation and/or inversion of larger tracts of DNA. In the 2000s, advances in wheat research such as the sequencing across multiple tissues, developmental stages and cultivars of complimentary DNA (cDNA) transcribed from messenger RNA (mRNA), and subsequently the availability of genome assemblies for cv. CHINESE SPRING (the wheat reference genome; IWGSC 2018) and 15 additional wheat cvs. (Walkowiak et al. 2020; Chap. 14) (Table 15.1) have led to detailed catalogues of both genic and non-genic DNA variation. Due to their abundance and nature, wheat studies over the last 10 years have most commonly assayed genic single nucleotide polymorphisms (SNPs) for use in genetic mapping approaches. Since the publication of the first high-density wheat genotyping array in 2013 capable of assaying ~ 9000 SNPs (Cavanagh et al. 2013), several additional arrays ranging from 3000 to 850,000 features are now available (Table 15.2). While SNP genotyping arrays are relatively simple and cheap to use, one drawback is that only those variants that have been pre-selected to be present on the array can be assayed. Thus, if the SNP identification panel used to design the array does not contain adequate sampling of the variants in the target genepool, useful information on the variation present in a target set of germplasm cannot be adequately assessed. This is a common issue for example in synthetic hexaploid wheat and its derived germplasm, where much of the novel D subgenome variation captured in this germplasm may not be assayed. More recently, reductions in costs have meant that sequencing-based genotyping approaches have become increasingly used in wheat. These include complexity reduction approaches such as genotyping by sequencing (GbyS) (Poland et al. 2012), Diversity Array Technology sequencing (DArTseq™; Sansaloni et al. 2011) and exome and/or promotor capture followed by Illumina short-read (i.e. ~ 150 bp) sequencing (Table 15.2). More recently, whole genome low-coverage sequencing is beginning to be used for genotyping in wheat (Table 15.2) and is considered in more detail in Box 1. Natural variation in the form of InDels and CNV are also relatively abundant in the wheat genome (e.g. Pont et al. 2019; Walkowiak et al. 2020; Wang et al. 2022), and despite the relatively limited number of functionally characterised wheat genes to date (Chap. 9), such variation has been shown to be a relatively common source of functional variation. For example, just within the flowering time pathway, deletions across putative cis-regulatory sites caused by double-stranded DNA break repair via non-homologous recombination have been shown to result in at least seven functional alleles of the *VERNALIZATION1* (*VRN-1*) flowering time gene homoeologues in hexaploid and diploid wheat (Cockram et al. 2007), while CNV at the *PHOTOPERIOD-1* (*PPD-1*) homoeologues determine flowering time in tetraploid and hexaploid wheat (Díaz et al. 2012; Würschum et al. 2019;

Table 15.1 Bread wheat cultivars/lines with genome assemblies

Cultivar	Seasonal growth habit	Origin	Release year	Genome assembly type
CHINESE SPRING	Spring	China	NA[‡]	Reference genome[1]
ALCHEMY	Winter	UK	2006	PA[3]
ARINA*LR*FOR	Winter	Switzerland	NA	RQA[2]
BROMPTON	Winter	UK	2005	PA[3]
CADENZA	Spring	UK	1992[*]	Scaffold[2]
CDC LANDMARK	Spring	Canada	2015[†]	RQA[2]
CDC STANLEY	Spring	Canada	2009[*]	RQA[2]
CLAIRE	Winter	UK	1999	Scaffold[2], PA[3]
HEREWARD	Winter	UK	1991	PA[3]
JAGGER	Winter	USA	1994[*]	RQA[2]
JULIUS	Winter	Germany	2008	RQA[2]
LR LANCER[$]	Spring	Australia	2013[*]	RQA[2]
MACE	Spring	Australia	2008[*]	RQA[2]
NORIN 61	Facultative	Japan	1944[*]	RQA[2]
PARAGON	Spring	UK	1988	Scaffold[2]
RIALTO	Winter	UK	1994	PA[3]
ROBIGUS	Winter	UK	2003	Scaffold[2], PA[3]
SOISSONS	Winter	France	1995	PA[3]
SY MATTIS	Winter	France	2010	RQA[2]
WEEBILL 1	Spring	Mexico	1999[*]	Scaffold[2]
XI19	Facultative	UK	2002	PA[3]

RQA reference quality assembly. *PA* pseudomolecule assembly. *NA* not applicable. [*] From GRIS database. [1]IWGSC (2018). [2] Pre-publication BLAST access at https://www.cropdiversity.ac.uk/8magic-blast/. [3]Walkowiak et al. (2020). [†] Application for Plant Breeders' Rights date. [‡] Landrace. [$] LongReach Lacner. Additionally, a RQA is available for a winter accession of spelt wheat (*T. aestivum* ssp. *spelta*) accession PI190962 from Central Europe[2]

see also Chap. 11).*PPD-1*) homoeologues determine flowering time in tetraploid and hexaploid wheat (Díaz et al. 2012; Würschum et al. 2019; see also Chap. 11).

Box 1: Wheat genotyping via skim sequencing

As genotyping via genome skim sequencing is typically undertaken at significantly less than 1-times genome-wide sequence coverage per line assayed (termed 1 ×), multiple reads at any given chromosomal location are not expected for any single line. Therefore, this approach is suited for experimental populations with defined founders, such that confidence in the DNA variants identified from skim sequence in any one line is achieved via reads obtained from additional lines in the population that carry the same variant. For example, if there are 200 lines in a bi-parental population, with each line sequenced to 0.3 × coverage, we would expect on average 60 × coverage of any single locus, and therefore 30 × coverage of each allele at any bi-allelic locus, i.e. (200 × 0.3)/2. Thus, by cataloguing and the SNPs present at good coverage in the population as a whole, the presence of any of these SNPs identified via a single sequencing read in any given line can be called with good confidence. Pre-determining the sequence variants present in the population founders, for example by exome capture or whole genome assembly, may help the process of variant calling and the imputation of variants that are not directly sequenced in any given line. For example, Scott et al. (2020a) sequenced the 16 founders of a wheat multifounder population via exome + promotor capture

Table 15.2 Examples of recent high-density, high-throughput wheat genotyping approaches

Genome-wide genotyping approaches	DNA variation origin
SNP array	
9 k array (Cavanagh et al. 2013)	Genes from cultivars
90 k array (Wang et al. 2014)	Genes from cultivars
280 k array (Rimbert et al. 2018)	Genes and intergenic variants identified in whole genome sequence of 8 cultivars
660 k array (Cui et al. 2017)	Unknown
820 k array (Winfield et al. 2016)	Exomes of 23 bread wheat cvs./landraces, and 20 spp./accessions of diploid, tetraploid and decaploid wheat
35 k array (Allen et al. 2017)	Subset of SNPs from the 820 k array, above
DArTseq™	
(Sansaloni et al. 2011, e.g. as applied in wheat by Sansaloni et al. 2020)	DNA variants, including SNPs and SilicoDArT (presence/absence variation) identified via genomic complexity reduction (achieved via restriction enzyme digestion/ligation), PCR amplification of followed by DNA sequencing and bioinformatic analysis
Exome capture	
DNA probes covering 107 Mb of non-redundant exonic target space (Jordan et al. 2015), representing 33% of the RefSeq v1.0 high-confidence gene set	Genes identified from the wheat reference genome RefSeq v1.0 annotation (IWGSC 2018). Genes and DNA variants identified are dependent on the germplasm assayed
Exome + promotor capture sequencing	
DNA probes covering 509 Mb exonic and 277 Mb promotor space (Gardiner et al. 2019). > 20 samples can be multiplexed in a single capture	Genes and promotors identified from the reference genome annotations of wheat (RefSeq v1.0 annotation, IWGSC 2018; TGACv1 annotation, Clavijo et al. 2017), Emmer wheat (Avni et al. 2017) and *Ae. tauschii* (Luo et al. 2017). Genes and DNA variants identified are dependent on the germplasm assayed
Genotyping-by-Sequencing (GbyS)	
Complexity reduction via restriction enzyme digestion, adaptor ligation, PCR and sequencing (first applied to wheat by Poland et al. 2012)	DNA variants determined bioinformatically from the ~ 100–150 bp sequence data generated from restriction enzyme cleavage sites sampled from across the genome
Skim sequencing	
Whole genome low-coverage DNA sequencing (e.g. as applied to a 16-founder MAGIC population, Scott et al. 2020a)	DNA variants originate from single sequencing reads per genotype assayed. For experimental populations, sequencing depth is achieved via reads from all lines in the population that carry the same genomic region

PCR polymerase chain reaction

identifying 1.13 million SNPs across the 110,790 genes targeted by the capture probes. They then skim sequenced the 501 derived recombinant inbred lines (RILs) at 0.3 × coverage, which directly identified ~ 28% of these SNPs (i.e. 1.13 million SNPs × 0.3 = 339,000 SNPs). SNP imputation in the RILs was then undertaken using the software STICH (Davies et al. 2016), resulting in 94% of the 1.13 million founder SNPs to be called and founder haplotype dosage at each chromosomal location to be assigned for all RILs. Down-sampling the 0.3 × read coverage showed RILs could be accurately inferred from sequence coverage as low as 0.076 × per RIL. Notably, at sequence coverage of 0.076 × and above, imputation accuracy was not dependent on whether or not founder haplotypes were included as a reference panel. This means that accurate RIL haplotype mosaics in the RILs could be achieved without the need to

generate data on the 16 founders. In summary, imputation from low-coverage whole genome sequencing of experimental populations represents a relatively straightforward and cost-effective genotyping strategy for bi-parental and multifounder experimental wheat populations and does not suffer from the inherent bias of SNP array genotyping approaches that require the variants targeted to be pre-identified.

15.4 Genetic Mapping Resolution: Population Size, Genetic Recombination and Effect Size

Forward genetic mapping relies largely on the recombination fraction between a QTL and the genetic markers that have been genotyped in the population, and the heritability of the target trait. These considerations are reviewed in more detail elsewhere (e.g. Cockram and Mackay 2018), but in general greater genetic mapping resolution can be attained by increasing population size and/or undertaking additional rounds of crossing. Larger populations also have the benefit of providing greater QTL detection power. Important to consider is the heritability of the target trait and the effect size of the QTL detected. The more heritable a trait is, and the larger its effect size, the easier it is to detect and precisely locate. Indeed, most wheat QTL resolved to the underlying gene level are for highly penetrant major genes, such as gene-for-gene disease resistance loci (e.g. for a recent list of cloned wheat rust resistance genes, see Bouvet et al. 2022b), awn presence/absence (Huang et al. 2020), vernalization response (first undertaken in *T. monococcum*: Yan et al. 2003; Yan et al. 2004), plant height (Tian et al. 2022) and grain quality (Uauy et al. 2006). If trait heritability is low, phenotypic replication can increase line mean heritability and has been used to refine and update the genetic interval of a locus on chromosome 5A controlling ~ 10% variation for wheat grain size (Brinton 2017; Brinton et al. 2017). Aside from such highly penetrant genetic loci, the genetic architecture of most target traits in wheat is highly quantitative in nature. For example, the mean QTL effect size for grain size traits in wheat is less than 10%, compared to more than 20% in the diploid cereal rice, and is likely due to the buffering effect of homoeologues of overlapping function in hexaploid wheat (Brinton and Uauy 2019).

15.5 Population Types

The identification of functional gene variants via genetic mapping relies on the capture of sufficient genetic diversity and genetic recombination. Fundamentally, two broad experimental population types are employed by researchers interested in identifying genetic loci controlling traits of interest. Both exploit genetic variation, and the reshuffling of this variation via genetic recombination, in order to associate markers or groups of markers (haplotypes, see also Chap. 9) with target traits.

15.5.1 Experimental Populations

Experimental populations are derived from crossing two or more parents to produce progeny in which genetic loci can be identified by the strength of the associations between genetic markers and traits of interest. Examples of some commonly used experimental populations are listed below and are illustrated in Fig. 15.2.

15.5.1.1 Bi-parental

Bi-parental populations are most commonly used in wheat forward genetics research and are constructed by first crossing two parents to generate first filial (F_1) derived progeny lines. Inbred progeny are generated either by single seed descent (whereby individual F_2 lines are selfed over three or more generations to achieve acceptable levels of homozygosity genome-wide) or via doubled haploid approaches (where haploid F_1-derived gametes undergo chromosome doubling, resulting in completely inbred progeny in a single generation) (Fig. 15.2). Despite DH lines typically taking less time to

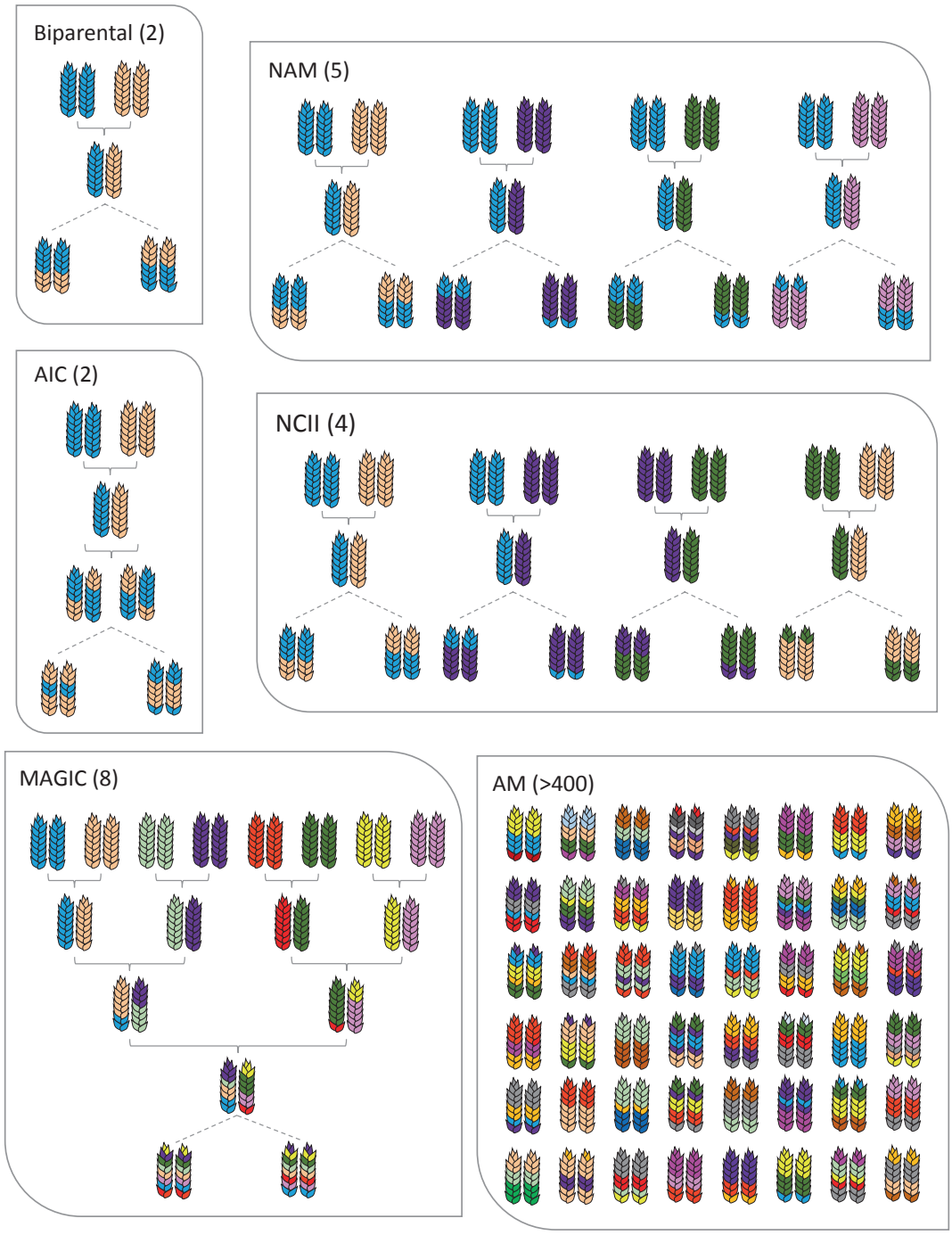


Fig. 15.2 Illustration of experimental population and association mapping panel designs. Number of founders illustrated in each panel is indicated in brackets. Dashed lines indicate inbreeding (via single seed descent or doubled haploid approaches) to produce multiple inbred lines. AIC = advanced intercross, two rounds of intercrossing illustrated, prior to the production of inbred lines. NAM = nested association mapping. NCII = North Carolina II model. MAGIC = multifounder advanced generation intercross. AM = association mapping population

create compared to RILs, DH populations capture less genetic recombination. This is because additional genetic recombination events can occur between regions of heterozygosity from the F_2 generation (25% heterozygous) until effective fixing at around the F_6 stage (1.6% heterozygous) or beyond, and which on average is equivalent to one additional round of crossing. Bi-parental populations are now beginning to be constructed from wheat cultivars with genome assemblies, such as the CHINESE SPRING × PARAGON population (Wingen et al. 2017).

15.5.1.2 Advanced Intercross

Even when bi-parental populations are created via single seed descent, the amount of genetic recombination captured can be relatively low. One way to increase the number of genetic recombinations is to continue random intercrossing of the F_2 for one or more generations before the production of inbred lines (Fig. 15.2). Such advanced intercross (AIC) populations (Darvasi and Soller 1995) designs provide greater precision compared to standard bi-parental populations of the same size. For example, Darvasi and Soller (1995) estimated that eight rounds of random intermating would reduce a QTL interval from 20 to 3.7 cM. While AIC have been used in species such as Arabidopsis (Fitz et al. 2014) and maize (Balint-Kurti et al. 2008), they have yet to be implemented in wheat—likely due to the time required to undertake additional rounds of crossing. However, the advent of 'speed breeding' approaches, that allow the generation time of both spring (Watson et al. 2018) and winter (Cha et al. 2022) wheat varieties to be reduced, means that for primary QTL screens, AIC approaches in wheat should become a more attractive prospect.

15.5.1.3 Near Isogenic Line Pairs, Introgression Lines and Chromosome Segment Substitution Lines

A near isogenic line (NIL) captures a relatively small chromosomal region from one 'donor' parent within the wider genomic context of a second 'recipient' parent (Fig. 15.2). NILs are generated via repeated rounds of backcrossing, often with the use of genetic markers to select for donor at the target chromosomal region, and for the recipient across the remainder of the genome. NILs are commonly used to target specific QTL of interest, allowing the effect of the contrasting alleles captured in the NIL pair to be evaluated using a single pair of lines, rather than a larger population in which additional genetic loci affecting the target trait may be segregating. Following this approach, individual genetic loci controlling a target trait can be investigated in detail, and the underlying physiology and pleiotropic effects on related traits can be assessed. Further, a NIL pair can be crossed to generate further genetic recombination and so further refine the genetic interval. For example, contrasting alleles at a major effect genetic locus for wheat grain weight identified in a bi-parental population of 192 inbred lines was subsequently assessed via phenotypic evaluation of BC_2- and BC_4-derived NILs, finding the ~7% increase in grain weight was (i) mediated predominantly by increased grain length, (ii) the maternal pericarp cell length was longer in the NIL carrying the high grain weight allele, and that (ii) increased grain length was detectable 12 days after fertilisation (Brinton et al. 2017). Additionally, the NILs were used to further refine the genetic interval to 4.3 cM (Brinton et al. 2017), with further analysis indicating that two genetic loci may be present at the locus (Brinton 2017). A series of NILs that capture chromosomal segments from wild and domesticated wheat relatives is termed introgression lines. Recent work in the UK has generated such germplasm resources for a range of wheat relatives. These include diploid *Ae. caudata* (CC genome. Grewal et al. 2020), *Am. muticum* (TT. Coombes et al. 2022), *Th. bessarabicum* (JJ. Grewal et al. 2018), and *T. uratu* (A^uA^u. Grewal et al. 2021), tetraploid *T. timopheevii* (A^tA^tGG. Devi et al. 2019), hexaploid *Th. intermedium* (JJ $J^{vs}J^{vs}$ S^tS^t. Cseh et al. 2019) and decaploid *Th. elongatum* (E^bE^b E^bE^b E^bE^b ES^tES^t ES^tES^t. Baker et al. 2020), with all introgression lines generated using the recipient wheat cv. PARAGON.

When a series of NILs is designed to collectively capture the entire donor background, the resulting resource is termed a chromosome segment substitution line (CSSL) population. In wheat, CSSLs populations that capture novel A, B and D subgenome diversity from wheat relatives have recently been developed using (i) a synthetic hexaploid wheat line (Horsnell et al. 2022) and (ii) a tetraploid *T. turgidum* ssp. *dicoccoides* accession (TTD-140). Not only do CSSL populations serve as useful sources of novel variation, they can also be used directly for genetic mapping, as recently illustrated in wheat by Horsnell et al. (2022).

15.5.1.4 Multifounder Populations: NAM

While bi-parental populations and derived NILs had long been the mainstay of forward genetic approaches, multifounder populations have recently become commonplace in plant research (reviewed by Scott et al. 2020b). Multi-parent mapping populations capture more variation than bi-parental populations and increase precision via joint linkage and association analysis. Nested association mapping (NAM) populations represent a series of bi-parental populations (termed 'families'), each of which has the same parent in common (Fig. 15.2). The first NAM population was made in maize (*Zea mays* L.) by crossing 25 diverse inbred lines with the inbred line B73 (termed here the 'linking' founder)—one of the most widely used lines in the history of maize breeding, and the line used for the maize reference genome (Yu et al 2008). Since then, the maize NAM parents have become extensively characterised, including provision of their genome assemblies (Gage et al. 2020). The genetic resolution obtained from NAM populations largely depends on the number of alleles present in the founders and the amount of genetic recombination captured in the progeny. The rarest alleles in any NAM population will be present in half of the progeny from the corresponding family. Therefore, in a NAM with 25 families and 200 progeny per family, rare alleles are expected to be present in 100 of the total 5000 progeny lines, i.e. a frequency of 2%. For NAM design, increasing the number of founders at the expense of family size should be preferable, as the decay of parental linkage disequilibrium for a given allele would likely, on average, be shared among more parents (Gage et al. 2020). NAM populations have now been made in many crop species and can be genetically analysed using association mapping approaches. At least part of the attraction of NAM design is that their composition (a series of bi-parental populations with a common parent) makes them more conceptually familiar to researchers experienced with bi-parental populations. Indeed, once a genetic locus has been identified in a NAM, it is straightforward to continue further analysis using one or more of the relevant constituent bi-parental populations. To date, several NAM populations have been created in wheat (Table 15.3; Fig. 15.3). The founders used include elite cultivars (e.g. Bajgain et al. 2016), genetically diverse landraces (Wingen et al. 2017), as well as germplasm that captures backcrossed chromosomal segments from wheat relatives via synthetic hexaploid wheat and wheat vs tetraploid durum wheat (*T. durum* ssp. *durum*) introgression lines. Further, a recent durum NAM has been constructed by crossing 50 durum landraces to an Ethiopian durum cultivar (Kidane et al. 2019). The largest wheat NAM currently available was constructed using 60 inbred worldwide landraces from the Watkins wheat landrace collection, backcrossed to the spring UK cultivar PARAGON, generating a population of 1192 RILs and a mean of 105 RILs per family (Wingen et al. 2017). Therefore, the rarest allele captured in the Watkins NAM would be expected to be present in 4% of the population—a frequency nominally sufficient for detection via genetic analysis.

15.5.1.5 Multifounder Populations: North Carolina II Model

A notable limitation of NAM populations is that while multiple founders are employed, a single 'linking' parent is used with which to cross to. The North Carolina II (NCII) design of Comstock and Robinson (1952) is conceptually an extension of NAM, whereby two or

Table 15.3 Examples of wheat multifounder populations and association mapping panels

Population type and name[1]	Genepool	Population details	Genotypic data	Germplasm availability
Diversity panels				
Watkins landraces Wingen et al. (2014) Wingen et al. (2017) Halder et al. (2019)	*T. aestivum* landraces from around the world	826 spring landrace accessions from 32 countries	41 microsatellites; 32,443 SNPs for 804 accessions	www.seedstor.ac.uk
Chinese landraces Zhou et al. (2017)	*T. aestivum* landraces from China	717 landraces accessions	9740 DArTseq and 178,803 SNPs	CAAS, China
WAGTAIL panel Downie et al. (2018) Fradgley et al. (2019) Sharma et al. (2022) White et al. (2022)	*T. aestivum* cvs., European	480 north-western European predominantly winter cvs. released 1916–2007	90 k SNP array	https://www.niab.com/research/agricultural-crop-research/resources
Seeds of Discovery Sansaloni et al. (2020)	*T. aestivum*	56,342 domesticated hexaploid wheats (additionally, 18,946 domesticated tetraploid wheats and 3903 wheat wild relatives)	>112,038 SNPs and SilicoDArT presence/absence markers	CIMMYT and ICARDA genebanks
Vavilov wheat collection Riaz et al. (2017)	*T. aestivum*	295 accessions, including 136 landraces, 32 cultivars, 10 breeding lines and 118 with unrecorded cultivation status	34,311 DArTseq markers	Australian Grains Genebank, Australia
MAGIC				
4-parent Australian MAGIC Huang et al. (2012)	*T. aestivum*. Australian spring cvs. BAXTER, CHARA, WESTONIA, YITPI	4 founders crossed in 3 funnels to generate 1579 RILs	826 DArT markers, 283 SNPs and 53 microsatellites across founders and 871 RILs	CSIRO, Australia
8-parent Australian MAGIC Shah et al. (2019)	*T. aestivum*. spring Australian (BAXTER, WESTONIA, YITPI), 4 worldwide spring (AC BARRIE, ALSEN, PASTOR, VOLCANI), and 1 Chinese winter (XIAOYAN54) cvs	8 founders crossed in 313 funnels followed by 0, 2 or 3 generations of intercrossing to produce 3412 RILs	27,687 genotyping array SNPs	CSIRO, Australia

(continued)

Table 15.3 (continued)

Population type and name[1]	Genepool	Population details	Genotypic data	Germplasm availability
NIAB Elite MAGIC Mackay et al. (2014) Bouvet et al. (2022b) Corsi et al. (2020) Downie et al. (2018) Lin et al. (2020a) Riaz et al. (2020) Wittern et al. (2022) Zanella et al. (2022)	*T. aestivum*. 7 winter UK cvs. (ALCHEMY, BROMPTON, CLAIRE, HEREWARD, RIALTO, ROBIGUS), 1 alternative UK (XI19), and 1 French (SOISSONS) cv	8 founders crossed in 180 funnels to produce > 1000 RILs	90 k SNP data for founders and 643 RILs (Mackay et al. 2014); Genome assembly for founders (Walkowiak et al., 2020), skim-seq for progeny[2]	NIAB, UK. https://www.niab.com/research/agricultural-crop-research/resources
NIAB Diverse MAGIC Scott et al. (2020a) Fradgley et al. (2022b)	*T. aestivum*. European cvs*, including 2 in common with the NIAB Elite MAGIC (ROGIBUS, SOISSONS)	16 winter-grown founders, 600 progeny		NIAB, UK. https://www.niab.com/research/agricultural-crop-research/resources
BMWpop Stadlmeier et al. (2018) Corsi et al. (2021) Lin et al. (2020b) Geyer et al. (2022) Stadlmeier et al. (2019)	*T. aestivum*. 7 German (BAYP4535, BUSSARD, EVENT, FRIL3565, FORMAT, JULIUS, POTENZIAL) and 1 Danish (AMBITION) winter cvs	8 founders crossed in 2 funnels with one further round of intercrossing to generate 516 RILs	5436 SNP genotyping array SNPs across founders and 394 RILs	Bavarian State Research Centre for Agriculture (LfL), Germany
WW-800 Sannemann et al. (2018) Lisker et al. (2022)	*T. aestivum*, 8 German cvs (BERNSTEIN, JB ASANO, JULIUS, LINUS, MEISTER, PATRAS, SAFARI, TOBAK)	8 founders crossed in 2 funnels to generate 910 RILs	7849 SNPs across the founders and 910 RILs	University of Halle, Germany
INRA MAGIC-like Thépot et al. (2014)	*T. aestivum*, 60 European/worldwide cvs	1 male-sterile line (cv. PROBUS) crossed and backcrossed with 59 European/worldwide lines before 12 generations of random intermating to generate 1000 RILs	8632 SNPs across 56 founders and 380 RILs	

(continued)

Table 15.3 (continued)

Population type and name[1]	Genepool	Population details	Genotypic data	Germplasm availability
NAM				
Watkins-60 Wingen et al. (2017) Khokhar et al. (2020)	*T. aestivum* landraces from around the world	60 bi-parental populations, created by crossing 60 spring Watkins landraces selected based on genetic diversity to the spring *T. aestivum* cv. PARAGON to generate 1192 RILs. Mean number RILs per population = 105	KASP markers and 31 microsatellite markers (for seven bi-parental populations)	JIC Seedstor, UK www.seedstor.ac.uk
Bajgain-10 Bajgain et al. (2016)	*T. aestivum* cvs., 10 spring stem rust resistant varieties (9 Kenyan, 1 US) versus Canadian stem rust resistant spring line LMPG-6	10 bi-parental populations, created by crossing the 10 founders to LMPG-6 to generate 852 RILs. Mean number RILs per population = 85	GbyS for 852 RILs	
Jordan-28 Jordan et al. (2018)	*T. aestivum* cvs. (3) and landraces (25), versus CIMMYT cv. Berkut	28 bi-parental populations, created by crossing the 29 founders to cv. BERKUT to generate 2100 RILs. Mean number RILs per population = 71 (estimated)	164,668 GbyS SNPs and 57,687 90 k array SNPs	
Kidane-50 Kidane et al. (2019)	*T. turgidum* ssp. *durum* landraces, 50 landraces versus Ethiopian durum cv. ASASSA	50 bi-parental populations, created by crossing the 50 founders to Asassa to generate 6280 RILs. Mean number RILs per population = 126	12,114 SNPs for 1280 RILs from 20 families	
NIAB_SW_TetHex_NAM[2]	58 *T. turgidum* ssp. *dicoccum, dicoccoides* and *durum* accessions versus spring wheat cv. PARAGON	58 bi-parental populations, created by crossing the 58 tetraploid accessions to cv. Paragon to generate 1784 RILs. Mean number RILs per population = 31	Axiom 35 k array datasets	NIAB, UK. https://www.niab.com/research/agricultural-crop-research/resources
NIAB_WW_SHW_NAM [3]	64 SHW accessions versus spring wheat cv. Paragon	64 bi-parental populations, created by crossing the 64 SHWs to cv. PARAGON to generate 4200 RILs. Mean number RILs per population = 66	None	
NIAB_WW_SHW_NAM [3]	54 SHW accessions versus winter wheat cv. ROBIGUS	54 bi-parental populations, created by crossing the 54 SHWs to cv. ROBIGUS to generate 3241 RILs. Mean number RILs per population = 60	Axiom 35 k array datasets	NIAB, UK. https://www.niab.com/research/agricultural-crop-research/resources

(continued)

Table 15.3 (continued)

Population type and name[1]	Genepool	Population details	Genotypic data	Germplasm availability
Wheat/wheat relative introgression lines				
NIAB_AB_CSSL Horsnell et al. (2022)	*T. turgidum* ssp. *dicoccoides* (tetraploid, AABB) accession TTD-140 versus *T. aestivum* cv. PARAGON	48 BC_4-derived inbred lines	Axiom 35 k array datasets	JIC Seedstor, UK www.seedstor.ac.uk
NIAB_D_CSSL Horsnell et al. (2022)	SHW (created via combining *T. durum* (tetraploid, AABB) accession Hoh-501 with *Ae. tauschii* (diploid, DD) accession Ent-336) versus *T. aestivum* (cv. PARAGON)	51 BC_4-derived inbred lines	Axiom 35 k array datasets	JIC Seedstor, UK www.seedstor.ac.uk
Ae. caudata introgression lines Grewal et al. (2020)	Introgression lines between *Ae. caudata* (diploid, CC) accession 2,090,001 and wheat cv. PARAGON	*Ae. caudata* versus wheat cv. PARAGON *ph1/ph1* mutant, F_1s backcrossed to wild-type PARAGON to generate BC_2, BC_3, BC_4 and BC_5 populations	620 KASP assays	JIC Seedstor, UK www.seedstor.ac.uk
Am. muticum introgression lines Coombes et al. (2022)	Introgression lines between *Am. muticum* accessions 2,130,004 and 2,130,012 crossed to wheat	*Am. muticum* versus wheat cv. PAVON or CHINESE SPRING, F_1s backcrossed to cv. PARAGON to produce different generation backcross lines	Whole genome sequencing, ~5 × coverage	JIC Seedstor, UK www.seedstor.ac.uk
Th. bessarabicum introgression lines Grewal et al. (2018)	Introgression lines between *Th. bessarabicum* (diploid, JJ) accession PI 531,712 and wheat cv. PARAGON	*Th. bessarabicum* versus (a) wheat cv. PARAGON *ph1/ph1* mutant, F_1s backcrossed to wild-type PARAGON to generate 12 BC-derived lines, (b) *T. turgidum* cv. CRESCO *ph1/ph1* mutant, F_1s backcrossed to wild-type PARAGON to generate 13 BC-derived lines	Axiom 35 k array datasets	JIC Seedstor, UK www.seedstor.ac.uk
Th. elongatum introgression lines Baker et al. (2020)	Introgression lines between *Th. elongatum* (decaploid E^bE^b E^bE^b E^bE^b ES^tES^t ES^tES^t) accession 401,007 and wheat cv. PARAGON	*Th. elongatum* versus wheat cv. CHINESE SPRING *ph1/ph1* mutant, F_1s backcrossed to wild-type PARAGON to generate 330 BC-derived lines	Axiom 35 k array datasets	JIC Seedstor, UK www.seedstor.ac.uk

(continued)

Table 15.3 (continued)

Population type and name[1]	Genepool	Population details	Genotypic data	Germplasm availability
Th. intermedium introgression lines Cseh et al. (2019)	Introgression lines between *Th. intermedium* (hexaploid, JJ $J^{vs}J^{vs}$ S^tS^t) accessions 401,141 and 440,016 and wheat cv. PARAGON	*Th. intermedium* versus wheat cv. PARAGON *ph1/ph1* mutant, F_1s backcrossed to wild-type PARAGON to generate 197 BC_2, BC_3 and BC_4-derived lines	Axiom 35 k array datasets	JIC Seedstor, UK www.seedstor.ac.uk
T. timopheevii introgression lines Devi et al. (2019) King et al. (2022) Steed et al. (2022)	Introgression lines between *T. timopheevii (tetraploid, A^tA^tGG)* accession P95-99.1–1 and wheat cv. PARAGON	*T. timopheevii* versus wheat cv. PARAGON *ph1/ph1* mutant, F_1s backcrossed to wild-type PARAGON to generate BC_2, BC_3 and BC_4 lines	Axiom 35 k array datasets	JIC Seedstor, UK www.seedstor.ac.uk
T. uratu introgression lines Grewal et al. (2021)	Introgression lines between *T. uratu* (diploid, A^uA^u) accessions 1,010,001, 1,010,002 and 1,010,006 and wheat cv. PARAGON	*T. timopheevii* accessions versus wheat cv. PARAGON *ph1/ph1* mutant, F_1s backcrossed to wild-type PARAGON to generate BC_3 lines	Axiom 35 k array datasets and 151 KASP markers	JIC Seedstor, UK www.seedstor.ac.uk

All wheat MAGIC populations are listed. For all other population types, notable examples of those available are listed. *CSSL* chromosome segment substitution line. *MAGIC* multifounder advanced generation intercross. *NAM* nested association mapping. *RIL* recombinant inbred line. *SNP* single nucleotide polymorphism. * BANCO, BERSEE, BRIGADIER, COPAIN, CORDIALE, FLAMINGO, GLADIATOR, KLOKA, MARIS FUNDIN, ROBIGUS, SLEJPNER, SOISSONS, SPARK, STEADFAST, STETSON) Introgressions were first undertaken by crossing *T. timopheevii* to wheat cv. PARAGON *ph1/ph1* mutant ($2n = 2x = 14$), and the resulting F_1 inter-specific hybrids backcrossed to wild-type PARAGON. [1]The first reference listed is the primary reference for the resource, subsequent references list examples of use of the resource. [2]J Cockram, personal communication. [3]Data repository at https://niab.github.io/niab-dfw-wp3/. *BC* backcross

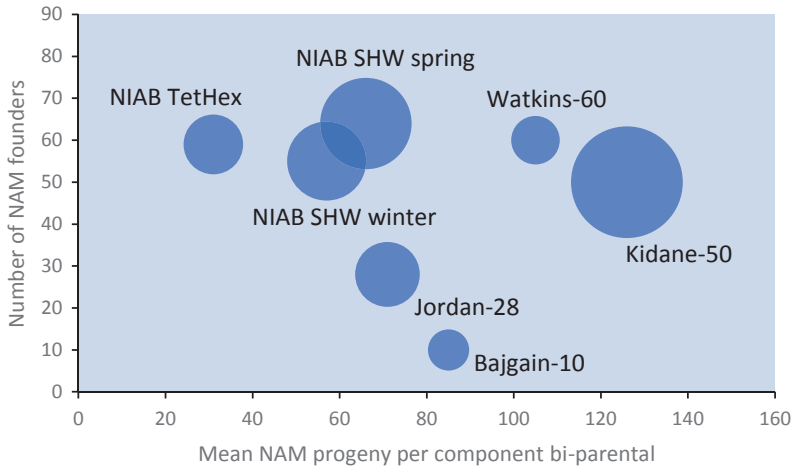


Fig. 15.3 Features of existing wheat nested association mapping (NAM) populations, comparing mean NAM progeny per component bi-parental population (*x*-axis) with the number of NAM founders (*y*-axis) and the size of the resulting population (proportional to the size of the circle)

more 'linking' parents are used such that every progeny family has half-sib relationships both through a common mother and through a common father (Fig. 15.2). Similarly, any combination of populations with founder links between them can be analysed together to undertake genetic analysis and to increase power and precision by increasing sample size (Cockram and Mackay 2018). However, such populations are more commonly used to confer detection of QTL in different genetic backgrounds and on the analysis of epistasis.

15.5.1.6 Multifounder Populations: MAGIC

While NAM and NCII populations capture more diversity than bi-parentals, they capture no additional genetic recombination than bi-parental populations of the same size. Since its pioneering use in mouse in 2002 (The Complex Trait Consortium 2002), the multi-parent advanced generation intercross (MAGIC) design has been applied to many crop species (Scott et al. 2020b). To aid crossing design, MAGIC populations typically use 4, 8 or 16 founders. However, unlike NAM or NCII populations, all MAGIC founders are intercrossed over multiple rounds of crossing to produce progeny that capture equal proportions of each founder genome (Fig. 15.2). Thus, MAGIC combines the benefits of increased genetic diversity afforded by NAM and NCII, with increased amounts of genetic recombination afforded by AIC, while minimising population structure via controlled crossing. In contrast to bi-parental populations, which are typically constructed to target a single target trait and are relatively quick to generate, MAGIC populations aim to capture and recombined multiple alleles across the genome and therefore take much longer to create. However, once complete, MAGIC, as well as other multi-parent populations, are well suited as community resources. In wheat, six MAGIC populations have been published, the first of which was the Australian spring wheat 4-parent MAGIC (Huang et al. 2012). Since then, four additional MAGIC populations have been created: 8-parent populations from Australia (Shah et al. 2019), the UK (Mackay et al. 2014) and Germany (Sannemann et al. 2018; Stadlmeier et al. 2018), as well as a 16-parent European wheat MAGIC (Scott et al. 2020a) (Fig. 15.4). Additionally, a MAGIC-like wheat population made between one male-sterile line crossed and backcrossed with 59 European/worldwide lines, followed by 12 generations of random intermating, has been generated (Thépot et al. 2014). To date, the 8-founder NIAB Elite MAGIC

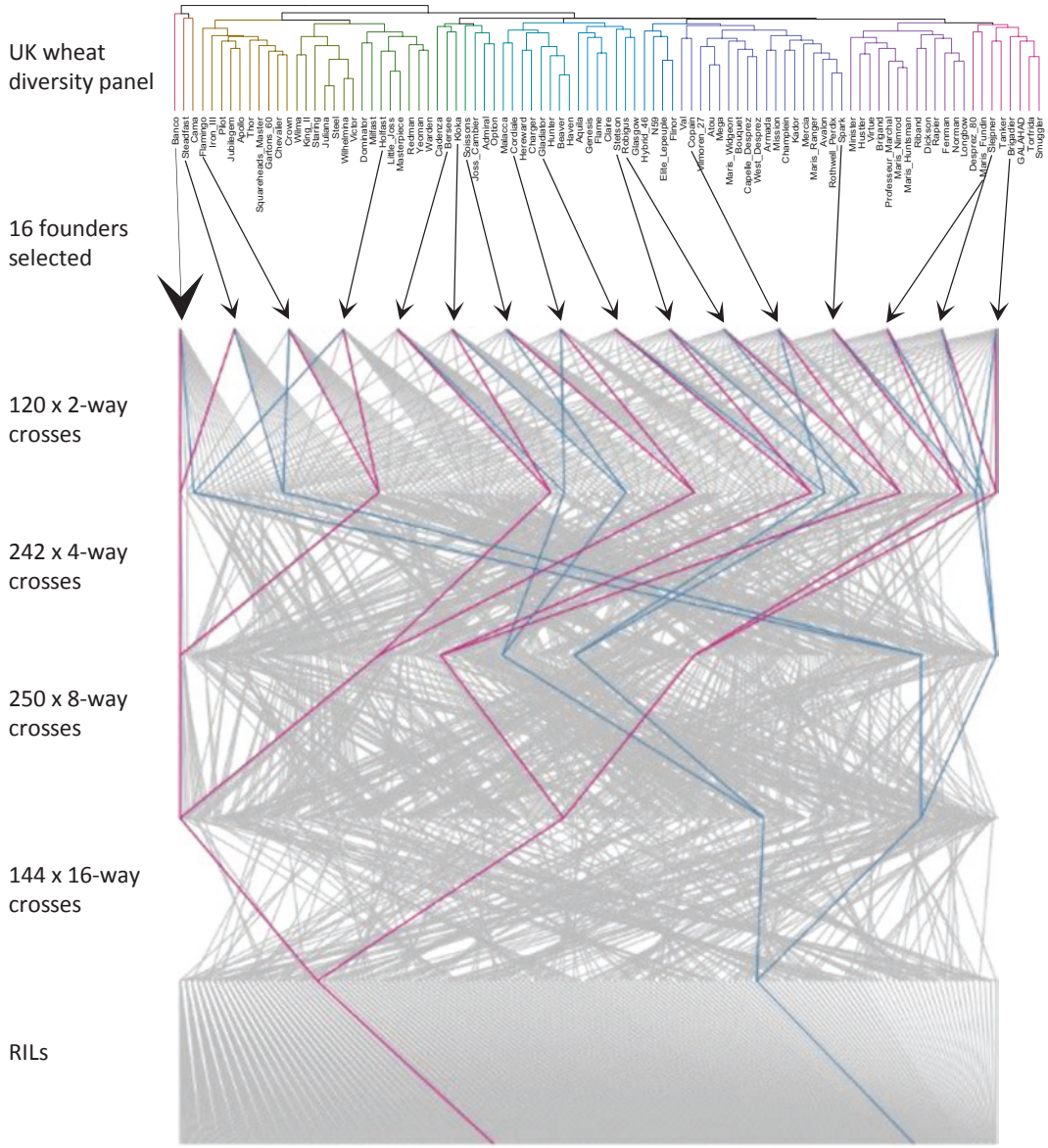


Fig. 15.4 Crossing diagram illustrating the founder selection and pedigree of the wheat 16-parent 'NIAB Diverse MAGIC' population. The red and blue lines each track the pedigree of a single recombinant inbred line (RIL) through the pedigree

population likely has the most publicly available resources available, including the population and associated 90 k array SNP data (Mackay et al. 2014) and genetic map (Gardner et al. 2016), genome assemblies for two of the founders (Walkowiak et al. 2020), and phenotypic and genetic data for numerous traits including disease (Bouvet et al. 2022a, c; Corsi et al. 2020; Lin et al. 2020a; Riaz et al. 2020) flowering time (Wittern et al. 2022), canopy architecture (Zanella et al. 2022), ear architecture (Dixon et al. 2018), end-use quality and mineral content (Fradgley et al. 2022a). Additionally, BLAST access to the genome assembles for the

remaining six founders is currently available ahead of publication (https://www.cropdiversity.ac.uk/8magic-blast/) and release of whole genome skim sequencing data for the RILs is imminent (J Cockram personal communication).

15.5.2 Founder Selection

Founder choice in any structured population is one of the first decisions addressed and depends to some degree on population type. For a bi-parental population, founders that contrast for a specific trait of interest are typically selected. In some cases, selection criteria will also include selection for specific traits that may otherwise confound the target phenotype. For example, founders with similar ear emergence date may be selected to avoid pleiotropic effects on diseases such as Fusarium head blight that affect the wheat ear. However, the differential presence of alleles of contrasting effect between founders may mean that while the parents may have been selected for similar phenotype, segregation for the phenotype may still be observed in the progeny. For NAM and MAGIC populations, founders should generally be selected to maximise genetic diversity, particularly in those designs that include larger founder numbers. For NAM populations, the selection of the 'linking' founder is notably important as each progeny line will sample 50% of its genome, and its genome will be highly represented in the population. 'Linking' founders typically represent a line which has been particularly well characterised, or is common in the wheat pedigree within the target geographical region. For example, the cultivar PARAGON has been selected as the 'linking' founder in three wheat NAM populations: Watkins-60 (Wingen et al. 2017), NIAB SHW and NIAB TetHex (data repository at https://niab.github.io/niab-dfw-wp3/). PARAGON is a spring UK variety released in 1988 which has a sequenced genome (Walkowiak et al. 2020), RNA sequence (RNA-seq) data from multiple tissues and a gamma-irradiated series of deletion lines (available via https://www.jic.ac.uk/research-impact/germplasm-resource-unit/). Similar considerations apply to the selection of linking founders in NCII population designs, although as two or more such founders are used, more flexibility is afforded.

If the aim of the population is to generate data under field conditions, founders should be suited for growth in the environments under which they will be phenotyped. When constructing populations using elite varieties, this should be relatively straightforward. For example, in the NIAB Diverse MAGIC population, the 16 founders were selected to sample maximum genetic diversity across a wider collection of 94 European winter wheat cultivars released over a 70 year period, for assessment under UK field conditions (Scott et al. 2020a). However, for populations that capture variation from landrace or species related to wheat, especially if these donors originate from geographic areas distant to the target environment, adaptability of the resulting populations to local field environments could be more problematic. In bi-parental or NAM populations, one way to address this is to generate populations from backcross-1 (BC_1) generation (where each progeny line contains on average 25% of the non-recurrent founder genome) or beyond, rather than from the F_1 which is expected to contain 50% contribution from each founder. This approach is logistically harder, as it involves an additional round of crossing and requires more progeny than an F_1-derived population to effectively sample non-recurrent founder genome. However, if the aim is to generate phenotypic data under field conditions, such approaches may be beneficial. For MAGIC designs, as each progeny line represents a balanced genomic mosaic of all founders, the inclusion of one, or possibly more, 'wilder' founder genomes is slightly less problematic. For example, in an 8-founder MAGIC which includes one 'wilder' founder, each RIL would be expected to contain a 1/8th genomic contribution from the 'wilder' founder. While no such MAGIC populations have been constructed to date in wheat, the most diverse is the INRA MAGIC-like population developed using one male-sterile line (cv. PROBUS) crossed and backcrossed with 59 European/worldwide lines before 12 generation of random

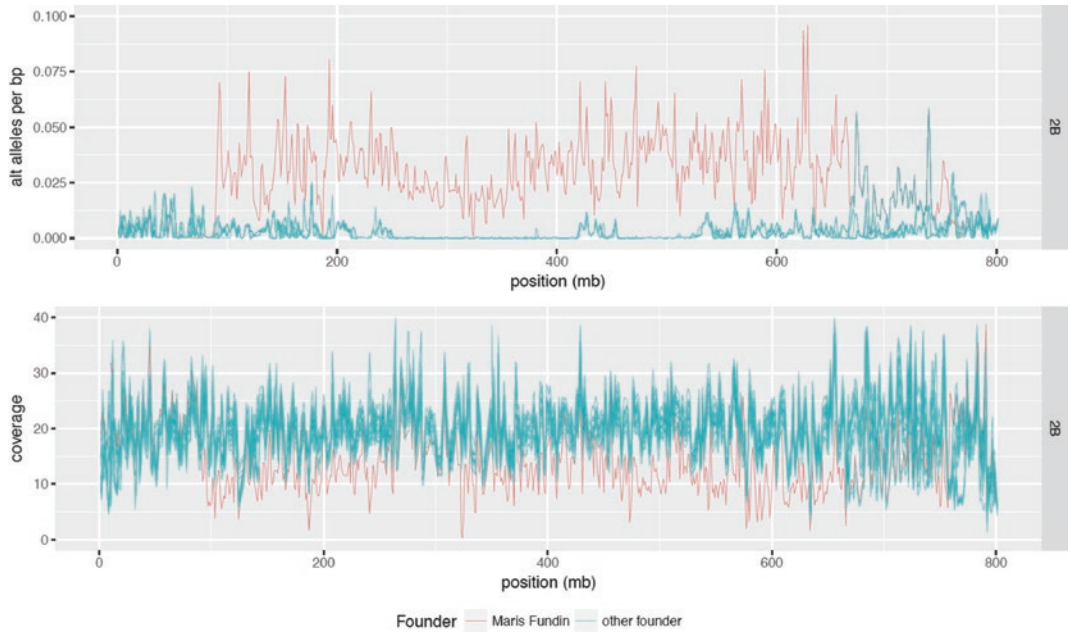


Fig. 15.5 ~540 Mb chromosome 2B introgression from *T. timopheevi* present in the NIAB Diverse MAGIC founder MARIS FUNDIN, as identified by analysis of exome-promotor capture sequence data of the 16 founders. The introgression is visualised here by the increase in non-reference (relative to chromosome 2B IWGSC RefSeq v1.0, cv. CHINESE SPRING) SNP variants (top) and as reduced sequence coverage (bottom) in MARIS FUNDIN, compared to the remaining 15 founders. Scott et al. (2020a) find the introgression to be substantially over-represented in the MAGIC progeny

intermating to generate 1000 lines (Thépot et al. 2014). Finally, for all population designs, it may be useful to consider the size and extent of any genomic rearrangements (e.g. the chromosome 5AL/7AL translocation Walkowiak et al. 2020) or chromosomal introgressions from wheat relatives, as their presence is likely to disrupt local genetic recombination rates. While such regions may specifically be sought, for example the *Ae. tauschii* (D) and *T. durum* ssp. *durum* (AB) genomic contributions captured in the NIAB SHW NAM, it is possible that one or more founders are unintentionally selected that contain such features. For example, in the 16-founder NIAB Diverse MAGIC population, cv. MARIS FUNDIN carries a large introgression of 540 Mb from *T. timopheevi* on chromosome 2B which is substantially over-represented in the MAGIC progeny (Fig. 15.5) (Scott et al. 2020a). Segregation distortion due to introgressions was also identified in the 8-founder NIAB Elite MAGIC, for example due to the chromosome 1B/1R wheat/rye introgression in cvs. BROMPTON and RIALTO and the presence of an introgression on the long arm of chromosome 4A in cv. ROBIGUS (Gardner et al. 2016).

15.5.3 Association Mapping Panels

The experimental populations described above take time to construct. However, it is possible to exploit the genetic variation and historical genetic recombination captured in existing collections of wheat varieties, landraces or accessions (Fig. 15.2). Such association mapping approaches aim to locate QTL based on the strength of the association between genetic markers and the target trait(s) and rely on the decay of linkage disequilibrium between markers and QTL over genetic distance (Cockram and Mackay 2018). Genetic analysis of association mapping panels can be conducted using

markers from candidate genes, or from across the genome using a whole genome association scan (GWAS) approach. Most commonly, single markers are regressed against the target trait. However, power can be increased by constructing haplotypes from the genotypic allele calls of two or more genetic variants that are closely physically or genetically linked within a defined region (haploblock). Use of haplotypes in GWAS can improve the estimation of allelic effects and increase statistical significance and is increasingly used in wheat. For example, linkage disequilibrium approach to defining haploblocks in a panel of 6333 wheat lines genotyped with 14,027 GbyS genetic markers resulted in the identification of 537 genome-wide haploblocks for downstream GWAS of grain yield (Sehgal et al. 2020). Alleles present at a frequency of less than 5% within the panel will typically not be detected, even if these alleles have relatively high effect sizes and/or the causative polymorphism is assayed. In human genetics, approaches that help identify rare alleles in GWAS are increasingly being used (reviewed by Lee et al. 2014), such as aggregation tests that evaluate cumulative effects of multiple genetic variants in a gene or region. The ability to generate experimental populations in plants means that such approaches are not as necessary to explore.

Unlike the case in most experimental populations in which allele frequency is relatively equally distributed among the progeny, association mapping panels are often characterised by notable levels of population substructure or subdivision. This is due to the differences in the shared ancestry of the lines over time, due to non-random mating. In cereal crops, population structure commonly arises from (i) physical separation, i.e. (geographic location), (ii) the contrasting germplasm preferences within different breeding companies, (iii) seasonal growth habit (i.e. spring or winter-sown) and (iv) traits underlying end-use quality (such as malting or feed in barley, or bread making versus in wheat) (Cockram et al. 2010; White et al. 2022) and yield (Sharma et al. 2022). For example, while relatively few major genetic determinants control the spring versus winter phenotype (Bentley et al. 2013), the common practice that spring cultivars are typically bred from other spring lines, while winters are bred from winters means that any genetic variants present at notably different frequencies between these two germplasm pools continue to show skews in their frequency in progeny lines. Thus here, if a favourable allele controlling a trait of interest happened to segregate predominantly in the spring pool, then the population structure inherent within spring varieties may lead to false-positive genotype-trait associations (termed Type-I errors) that are not due to close linkage of markers with the underlying QTL. It is possible to control statistically for population structure (Q) by using genetic markers to determine a Q-matrix of population membership estimates for each accession in the panel. Q-matrices can be determined using programmes such as STRUCTURE (Pritchard et al. 2000) or via principal component analysis (Zhao et al. 2007). Additional correction for more recent similarities due to close kinship (K) can also be included and can be determined using genetic markers. Indeed, approaches such as the Q+K mixed model (Yu et al. 2006) that account for multiple levels of relatedness between individuals have been shown to control well for false-positive as well as false-negative (Type-II error) associations and often lead to higher power than correction via Q or K alone (Yu et al. 2006). However, accounting for population structure/kinship sacrifices some level of experimental power to detect those genetic loci that are correlated with the adjustments made. Nevertheless, power and precision to detect genetic loci in association mapping panels can be high, compared to experimental populations of the same size. While improved power can be achieved by increasing the number of individuals in the panel, the inclusion of additional accessions may increase population substructure and/or kinship. Similarly, linkage disequilibrium may decay quite slowly in with genetic distance in cultivars (due to close kinship among all lines), which will reduce the precision to detect QTL (Cockram and Mackay 2018) but will increase

power. Conversely, linkage disequilibrium in panel's landraces is typically higher, enabling greater genetic mapping precision. Genotyped wheat landraces collections are now available that sample diversity with single countries (e.g. China, Zhou et al. 2017) or from around the world—such as the Watkins (Wingen et al. 2014) and Vavilov collections (Riaz et al. 2018). These are beginning to be used for GWAS of agronomic traits, such as disease resistance (tan spot, Halder et al. 2019; leaf rust. Riaz et al. 2018, stripe rust, Jambuthenne et al. 2022) and pre-harvest sprouting (Zhou et al. 2017). Given the multiple variables affecting GWAS in association mapping panels, it is useful to determine the efficacy of experimental design by undertaking power calculations, especially if population size is relatively small (e.g. White et al. 2022).

15.6 Reverse Genetics Germplasm Platforms

Functional validation of genes genetically mapped using experimental or association mapping populations can be undertaken using reverse genetics approaches. Transgenic methods aim to alter gene expression or function, typically via gene overexpression, gene silencing or gene editing (reviewed in wheat by Adamski et al. 2020). Alternatively, non-transgenic reverse transgenics approaches are available that exploit genetic variation induced by mutagenizing agents. In wheat, the most commonly used are Targeting Induced Local Lesions in Genomes (TILLING) populations, created by using an inbred donor line (termed the M_0 generation) and applying the chemical agent ethyl methanesulphonate (EMS). The resulting EMS treated seed is termed the M_1 generation, which can be subsequently selfed over several generations to generate a population of TILLING lines in which the EMS-generated mutations become progressively fixed in homozygous state. Bespoke experiment-specific TILLING populations are frequently used to determine genes underlying traits controlled by single major effect genes, such as gene-for-gene disease resistance. In such cases, a wheat line for which resistance to the target disease is controlled by a single major effect locus is mutated, and susceptible TILLING lines identified phenotypically. Assuming the underlying gene can be sequenced, relatively low numbers of TLLING lines with independent mutations at the target locus are generally sufficient to give a high statistical probability of identifying the causative gene. For example, Sánchez-Martín et al. (2016) estimated that the probability that the 12 kb gene containing contig of their target wheat gene (*Pm2* conferring resistance to powdery mildew) being mutated across all 12 identified powdery mildew susceptible TILLING mutants was 1 in 300,000,000,000. Several approaches to applying DNA sequencing to such gene identification approaches have been published: the first uses exome capture of pre-determined candidate gene families (termed resistance-gene enrichment sequencing, RenSeq, when applied to NRL disease resistance gene families; Jupe et al. 2013). The second, termed MutChromSeq, involves flow sorting and direct sequencing of the target chromosome in each of the phenotypically identified TILLING lines (Sánchez-Martín et al. 2016). In addition to such experiment-specific TILLING resources, exome capture followed by DNA sequencing of large numbers of TILLING lines generated from the spring bread wheat cv. CADENZA (1200 lines) and the tetraploid wheat cv. KRONOS (1535 lines) have been made publicly available (Krasileva et al. 2016). The resulting TILLING mutations have been aligned against the bread wheat reference genome of cv. CHINESE SPRING (RefSeq v1.1; IWGSC 2018) and searchable via the Ensembl plants (Cunningham et al. 2022) genome browser. The effects of mutations on protein sequence have been predicted in relation to CHINESE SPRING gene models, with deleterious mutations determined to be present in around 90% of the captured genes. The ability to identify and prioritise TILLING mutants in silico means these resources serve as useful genome-wide resources for gene functional validation in wheat. Considerations for the identification and validation of wheat TILLING

mutants in the CADENZA and KRONOS populations are listed in more detail by Adamski et al. (2020) and include the need to combine TILLING mutants in multiple homoeologues to overcome possible functional redundancy as well as the need to undertake sufficient rounds of backcrossing to remove background mutations. Examples of their use for gene functional characterisation include (i) wheat candidate genes orthologous to map-based cloned gene from model species (e.g. *TaGRAIN WIDTH2*, Simmonds et al. 2016), (ii) wheat genes identified via forward phenotypic screening followed by bulk segregant analysis of backcross derived progeny between mutant line and wild-type (e.g. *HOMEOBOX DOMAIN-2*, Dixon et al. 2022) and (iii) candidate genes underlying wheat genetic loci previously refined by fine-mapping (e.g. *WHEAT ORTHOLOG OF APO1*, Kuzay et al. 2022; *EARLY FLOWERING 3*, Wittern et al. 2022). While the ability to screen in silico the cv. CADENZA and KRONOS TILLING populations provide proven community resources for gene functional characterisation, they can only be used for those genes present in the two founding cultivars used. The availability of annotated genome assemblies for multiple wheat varieties now provides the underpinning knowledge from which it may in future be possible to develop additional sequenced TILLING resources that target genes not captured in cv. CADENZA and KRONOS.

15.7 The Future of Genetic Recombination

Genetic recombination in wheat is enriched in the telomeric regions and becomes progressively less frequent towards the pericentromeric and centromeric regions, with 80% of recombination events occurring in less than a quarter of the genome (e.g. Gardner et al. 2016; IWGSC 2018). As genetic mapping relies on the occurrence of recombination, being able to increase recombination at chromosomal regions of interest would help both genetic mapping precision, and the ability to recombine different haplotypes in breeding. Analysis of crossover events in RIL populations has identified QTL for genetic recombination frequency, such as a locus on chromosome 6A in the CHINESE SPRING × PARAGON population controlling around 6% of the variation (Gardiner et al. 2019). Further, recent work shows that recombination events in wheat pericentric regions can be increased in some chromosomes by increasing temperature during meiosis (Coulton et al. 2020), although this does come with reduced fertility (Draeger and Moore 2017). Transgenic approaches for altering genetic recombination rates and locations are also now being investigated. For example, transient virus induced gene silencing (VIGS) of wheat candidate genes homologous to genes in other species shown to control genetic recombination shows it is possible to alter the distribution of recombination along chromosomes (Raz et al. 2021). VIGS silencing of the durum wheat homologue of the anti-cross over gene *XRCC2* (a paralogue of *RAD51*) in F_1 plants ahead of meiosis resulted in increased genetic recombination across much of the pericentromeric region of chromosome 4B, as well the more distal pericentromeric regions of chromosome 5B (Raz et al. 2021). Such results indicate that it should be possible to increase genetic recombination in at least some of the pericentromeric landscape of wheat. The maturation of gene editing methodologies may soon enable the targeting of cross-overs and genetic recombination to more specific genomic locations.

15.8 Conclusions

In parallel to the efforts to provide wheat genomic and genotyping tools, the wheat community has generated extensive resources to support genetic locus and gene characterisation via forward and reverse genetics approaches. For highly penetrant wheat genetic loci originating from natural variants or via induced mutation, and where phenotype effectively acts as a genetic marker, various routes have been used to identify the underlying genetic loci, including

fine-mapping in bi-parental derived germplasm, as well as reverse genetics approaches such as RenSeq and MutChromSeq where the identification of multiple independent alleles rather that genetic recombination is required. For genetic loci of a more quantitative nature, to date it is those which account for an unusually high proportion of genetic variation that have been fine-mapped or map-based cloned, using bi-parental populations and also more recently via multifounder populations. The vast majority of remaining heritable variation in the wheat genepool is much more quantitative in nature, typically accounting for 3–5% of the phenotypic variation. For such loci, including those located in genomic regions with very low genetic recombination, identification of the underlying genes and variants via forward mapping approaches will continue to pose a challenge. However, genetic mapping approaches will allow their alleles and linked haplotypes to be determined, and increasingly, for the epistatic non-additive interaction effects of these loci to be characterised. For wheat breeding, advances in our knowledge of genetic loci and gene function will best be exploited within a quantitative genetics framework (Mackay et al. 2021). Trait improvement in the context of breeding over the next decade will likely focus on integration of multi-trait ensemble phenotypic weighting approaches (e.g. Fradgley et al. 2022b) combined with improved genomic selection methodologies and field-based phenotyping at increasing throughput and precision. The next decade will likely also see the maturation of approaches to engineer increased genetic recombination, and to design via gene editing new alleles with improved function. Finally, computer vision, artificial intelligence and machine learning and approaches are now maturing to the point at which they can more readily be applied to complex challenges such as crop phenotyping and plant breeding. Such approaches need to be efficiently combined to underpin future breeding for improved crop performance, quality and resilience.

Acknowledgements I thank Mike Scott for the images used in Figs. 15.4 and 15.5. My time was supported by Biotechnology and Biological Sciences Research Council (BBSRC) grants BB/P010741/1 and BB/R019231/1.

Glossary

$2n = 6x = 42$, AABBDD n is the gametic chromosome number, $2n$ is somatic chromosome number. x is the basic chromosome number, which for wheat is 7. Bread wheat is hexaploid with 6 chromosome sets in its genome (6x), termed the AA, BB and DD subgenomes. Thus, a somatic cell of the hexaploid bread wheat genome has a total of 42 chromosomes, summed across its AA BB and DD subgenomes.

Advanced intercross (AIC) A bi-parental populations, where F2 progeny are intercrossed over one or more generations before the generation of inbred lines.

Association mapping A method for genetic mapping of QTL that uses historic linkage disequilibrium to associated phenotype to genetic markers. Also known as 'linkage disequilibrium mapping'.

Copy number variation (CNV) Differences in the number of copies of a particular gene or chromosomal region. Where there is a presence or absence of a gene/region, it can also be termed presence/absence variation (PAV).

Genetic recombination The rearrangement of DNA sequences by the breakage and re-joining of chromosome segments.

Genome-wide association study (GWAS) A method for genetic mapping, using a collection of varieties, landraces or lines from an experimental population with phenotypic and genome-wide genotypic datasets.

Haplotype A set of DNA markers located sufficiently closely linked on the same chromosome to be frequently inherited as a single unit.

Linkage disequilibrium (LD) Non-random association of genetic markers at separate loci located that are typically located on the same chromosome.

Experimental population A population of lines created by crossing two or more founders.

Multi-parent advanced generation intercross (MAGIC) Experimental populations typically made by intercrossing 4, 8 or 16 founders over multiple generations so that the outputs of the crossing have contributions from each of the founders. Inbred lines are then derived by single seed descent.

Nested association mapping (NAM) A collection of two or more bi-parental populations, where all individual bi-parental populations share one founder in common (i.e. a single recurrent parent is used). E.g. Founder-1 × Founder-2, 1 × 3, 1 × 4.

North Carolina II (NCII) model A collection of three or more bi-parental populations, where any single bi-parental populations shares at least one founder in common with any other population, but where two or more recurrent parents are used. E.g. Founder-1 × Founder-2, 1 × 3, 2 × 3.

Population substructure Presence of a systematic difference in allele frequencies between groups of accessions, due to non-random mating.

Single nucleotide polymorphism (SNP) A genomic variant at a single base position in a DN

References

Adamski NM, Borrill P, Brinton, Harrington SA, Marchal C et al (2020) A roadmap for gene functional characterisation in crops with large genomes: lessons from polyploid wheat. eLife, 9:e55646

Akhunov ED, Goodyear AW, Geng S, Qi L-L, Echalier B et al (2003) The organization and rate of evolution of wheat genomes are correlated with recombination rates along chromosome arms. Genome Res 13:753–763

Aktar-Uz-Zaman M, Tuhina-Khatun M, Hanafi MM, Sahebi M (2017) Genetic analysis of rust resistance genes in global wheat cultivars: an overview. Biotechnol Biotechnol Equip 31:431–445

Allen AM, Winfield M, Burridge AJ, Rownie RC, Benbow HR, Barker GLA, Wilkinson PA, Coghill J, Waterfall C, Davassi A, Scopes G, Pirani A, Webster T, Brew F, Bloor C, Griffiths S, Bentley AR, Alda M, Jack P, Phillips AL, Edwards KJ (2017) Characterization of a wheat breeders' array suitable for high-throughput SNP genotyping of global accessions of hexaploid bread wheat (*Triticum aestivum*). Plant Biotechnol J 15:390–401

Avni R, Nave M, Barad O, Baruch K, Twardziok SO et al (2017) Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. Science 357:93–97

Bajgain P, Rouse MN, Tsilo TJ, Macharia GK, Bhavani S et al (2016) Nested association mapping of stem rust resistance in wheat using genotyping by sequencing. PLoS ONE 11:1–22

Baker L, Grewal S, Yang C-y, Hubbart-Edwards S, Scholefield D, Ashling S, Burridge AJ, Przewieslik-Allen AM, Wilkinson PA, King IP, King J (2020) Exploiting the genome of *Thinopyrum elongatum* to expand the gene pool of hexaploid wheat. Theor Appl Genet 133:2213–2226

Balint-Kurti PJ, Wisser R, Zwonitzer JC (2008) Use of an advanced intercross line population for precise mapping of quantitative trait loci for gray leaf spot resistance in maize. Crop Science, 48: 1696–1704.

Bentley AR, Jensen EF, Mackay IJ, Hönicka H. Fladung M, Hori K, Yano M, Mullet JE, Armstead IP, Hayes C, Thorogood D, Lovatt A, Morris R, Pullen N, Mutasa-Göttgens E, Cockram J (2013). Flowering time. In: Cole C (ed) Genomics and breeding for climate-resilient crops, vol 2. Springer, Berline, pp 1–67

Bouvet L, Percival-Alwyn L, Berry S, Fenwick P, Holdgate S, Mackay IJ, Cockram J (2022a) Genetic resistance to yellow rust infection of the wheat ear is controlled by genes controlling foliar resistance and flowering time. Crop Sci 62:1758–1770

Bouvet L, Holdgate S, James L, Thomas J, Mackay IJ, Cockram J (2022b) The evolving battle between yellow rust and wheat: implications for global food security. Theor Appl Genet 135:741–753

Bouvet L, Percival-Alwyn L, Berry S, Fenwick P, Mantello CC, Sharma R, Holdgate IJ, Mackay IJ, Cockram J (2022c) Wheat genetic loci conferring resistance to stripe rust in the face of genetically diverse races of the fungus *Puccinia striiformis* f. sp. *tritici*. Theor Appl Genet 135:301–319

Brinton J (2017) Deciphering the molecular mechanisms controlling grain length and width in polyploid wheat. PhD thesis, submitted to the University of East Anglia, Norwich, UK

Brinton J, Simmonds J, Minter F, Leverington-Waite M, Snape J, Uauy C (2017) Increased pericarp cell length underlies a major quantitative trait locus for grain weight in hexaploid wheat. New Phytol 215:1026–1038

Brinton J, Uauy C (2019) A reductionist approach to dissecting grain weight and yield in wheat. J Integr Plant Biol 61:337–358

Cavanagh CR, Chao S, Wang S, Huang BE, Stephen S et al (2013) Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. Proc Natl Acad Sci USA 110:8057–8062

Cha J-K, O'Connor K, Alahmad S, Lee J-H, Dinglasan E, Park H, Lee S-M, Hirsz D, Kwon S-W, Kwon Y, Kim K-M, Ko J-M, Hickey LT, Shin D, Dixon LE (2022) Speed vernalization to accelerate generation advance in winter cereal crops. Mol Plant 15:1300–1309

Chai Y, Pardey PG, Silverstein AT (2022) Scientific selection: a century of increasing crop varietal diversity in UK wheat. Proc Natl Acad Sci USA 119:e2210773119

Chen S, Hegarty J, Shen T, Hua L, Li H, Luo J, Li H, Bai S, Zhang C, Dubcovsky J (2021) Stripe rust resistance gene *Yr34* (synonym *Yr48*) is located within a distal translocation of *Triticum monococcum* chromosome $5A^{m}L$ into common wheat. Theor Appl Genet 134:2197–2211

Cheng H, Liu J, Wen J, Nie X, Xu L, Chen N, Li Z, Wang Q, Zheng Z, Li M, Cui L, Liu Z, Bia J, Wang Z, Xu S, Yang Q, Appels R, Han D, Song W, Sun Q, Jiang Y (2019). Frequent intra- and inter-species introgression shapes the landscape of genetic variation in bread wheat. Genome Biology 20:136

Clavijo BJ, Venturini L, Schudoma C, Accinelli GG, Kaithakottil G et al (2017) An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. Genome Res 27:885–896

Cockram J, Mackay I, O'Sullivan DM (2007) The role of double-stranded break repair in the creation of phenotypic diversity in cereal *VRN1* loci. Genetics 177:2535–2539

Cockram J, White J, Zuluaga DL, Smith D, Comadran J et al (2010) Genome-wide association mapping to candidate polymorphism resolution in the unsequenced barley genome. Proc Natl Acad Sci USA 107:21611–21616

Cockram J, Mackay I (2018) Genetic mapping populations for conducting high-resolution trait mapping in plants. In: Varshney R, Pandey M, Chitikineni A (eds) Plant genetics and molecular biology. Advances in biochemical engineering/biotechnology, vol 164. Springer, Cham

Comstock RH, Robinson HF (1952) In: Gowen JW (ed) Heterosis. Iowa State College Press, Ames, IA., pp 495–516

Coombes B, Fellers JP, Grewal S, Rusholme-Pilcher R, Hubbart-Edwards S, Yang CY, Joynson R, King IP, King J, Hall A (2022) Whole genome sequencing uncovers the structural and transcriptomic landscape of hexaploid wheat/*Ambylopyrum muticum* introgression lines. Plant Biotechnol J. https://doi.org/10.1111/pbi.13859

Corsi B, Percival-Alwyn L, Downie RC, Venturini L, Iagallo EM, Mantello CC, McCormick-Barnes C, See PT, Oliver RP, Moffat CS, Cockram J (2020) Genetic analysis of wheat sensitivity to the ToxB fungal effector from *Pyrenophora tritici-repentis*, the causal agent of tan spot. Theor Appl Genet 133:935–950

Corsi B, Obinu L, Zanella CM, Cutrupi S, Day R et al (2021) Analysis of a German multi-parental population identifies eight genetic loci controlling two or more yield components in wheat, including the genetic loci *Rht24, WAPO-A1* and *WAPO-B1* and multi-trait QTL on chromosomes 5A and 6A. Theor Appl Genet 134:1435–1454

Coulton A, Burrdidge A, Edwards KJ (2020) Examining the effects of temperature on recombination in wheat. Front Plant Sci 11:230

Cseh A, Yang C-Y, Hubbart-Edwards S, Scholefield D, Ashling S, Burridge AJ, Wilkinson PA, King IP, King J, Grewal S (2019) Development and validation of *Thinopyrum intermedium*–exome based SNP marker set for identification of the St, Jr and Jvs genomes in a wheat background. Theor Appl Genet 132:1555–1570

Cui F, Zhang N, Fan X-I, Zhang Wm Zhao C-H, Yang L-J, Pan R-Q, Chen M, Han J, Zhao X-Q, Ji J, Tong Y-P, Zhange H-X, Jia J-Z, Zhao G-Y, Li J-M (2017) Utilization of a Wheat660k SNP array-derived high-density genetic map for high-resolution mapping of a major QTL for kernel number. Sci Rep 7:3788

Cunningham F, Allen JE, Allen A, Alvarez-Jarreta J, Amode MR et al (2022) Ensembl 2022. Nucleic Acids Res 50:D988–D995

Das MK, Bai G, Mujeeb-Kazi A, Rajaram S (2016) Genetic diversity among synthetic hexaploid wheat accessions (*Triticum aestivum*) with resistance to several fungal diseases. Genet Resour Crop Evol 63:1285–1296

Darvasi A, Soller M (1995) Advanced intercross lines, and experimental population for fine genetic mapping. Genetics 141:1199–1207

Davies RW, Flint J, Myers S, Mott R (2016) Rapid genotype imputation from sequence without reference panels. Nat Genet 48:965–969

Devi U, Grewal S, Yang C-y, Hubbart-Edwards S, Scholefield D, Ashling S et al (2019) Development and characterisation of interspecific hybrid lines with genome-wide introgressions from *Triticum timopheevii* in a hexaploid wheat background. BMC Plant Biol 19:183. https://doi.org/10.1186/s12870-019-1785-z

Díaz A, Zikhali M, Turner AS, Isaac P, Laurie DA (2012) Copy number variation affecting the *Photoperiod-B1* and *Vernalization-A1* genes is associated with altered flowering time in wheat (*Triticum aestivum*). PLoS ONE 7:e33234

Dixon LE, Greenwood JR, Bencivenga S, Zhang P, Cockram J, Mellers G et al (2018) *TEOSINTE BRANCHED 1* regulates inflorescence architecture and development in bread wheat (*Triticum aestivum* L.). Plant Cell 30:563–581

Dixon LE, Pasquarello M, Badgami R, Levin KA, Poschet G, Ng PQ, Orford S, Chayut N, Adamski NM, Brinton J, Simmonds J, Steuernagel B, Searle IR, Uauy C, Boden SA (2022). MicroRNA-resistant alleles of *HOMEOBOX DOMAIN-2* modify

inflorescence branching and increase grain protein content of wheat. Science Advances 8:eabn5907

Downie RC, Bouvet L, Furuki E, Gosman N, Gardner KA, Mackay IJ, Mantello CC, Mellers G, Phan H, Rose GA, Tan K-C, Oliver R, Cockram J (2018) Assessing European wheat sensitivities to *Parastagonospora nodorum* necrotrophic effectors and fine-mapping of the *Sn3-B1* locus conferring sensitivity to the effector SnTox3. Front Plant Sci 9:881

Draeger T, Moore G (2017) Short periods of high temperature during meiosis prevent normal meiotic progression and reduce grain number in hexaploid wheat (*Triticum aestivum* L.). Theor Appl Genet 130:1785–1800

Dudley JW (2007) From means to QTL: the Illinois long-term selection experiment as a case study in quantitative genetics. Crop Sci 47:S20–S31

Fitz Gerald JN, Carlson AL, Smith E, Maloof JN, Weigel D, Chory J, Borevitz JO, Swanson RJ (2014) New Arabidopsis advanced intercross recombinant inbred lines reveal female control of nonrandom mating. Plant Physiol 165:175–185

Fradgley N (2022) Working towards a diverse UK landscape. In: NIAB landmark, issue 51, Winter 2022/2023, pp 6–7

Fradgley N, Gardner KA, Cockram J, Elderfield J, Hickey JM, Howell P, Jackson R, Mackay I (2019) A large-scale pedigree resource of wheat reveals evidence for adaptation and selection by breeders. PLoS Biol 17:e3000071

Fradgley NS, Gardner KA, Kerton M, Swarbreck SM, Bentley AR (2022a) Trade-offs in the genetic control of functional and nutritional quality traits in UK winter wheat. Heredity 128:420–433

Fradgley NS, Gardner KA, Bentley AR, Howell P, Mackay IJ, Scott MF, Mott R, Cockram J (2022b). Multi-trait ensemble genomic prediction and simulations of recurrent selection highlight importance of complex trait genetic architecture in long-term genetic gains in wheat. bioRxiv, https://doi.org/10.1101/2022.11.08.515457

Gage JL, Monier B, Giri A, Buckler ES (2020) Ten years of the maize nested association mapping population: impact, limitations and future directions. Plant Cell 32:2083–2093

Gardiner L-J, Wingen LU, Bailey P, Joynson R, Brabbs T, Wright J, Higgins JD, Hall N, Griffiths S, Clavijo BJ, Hall A (2019) Analysis of the recombination landscape of hexaploid bread wheat reveals genes controlling recombination and gene conversion frequency. Genome Biol 20:69

Gardner KA, Wittern LM, Mackay IJ (2016) A highly recombined, high-density, eight founder wheat MAGIC map reveals extensive segregation distortion and genomic locations of introgression segments. Plant Biotechnol J 14:1406–1417

Gaurav K, Arora S, Silva P, Sanchez-Martín J, Horsnell R et al (2022) Population genomic analysis of *Aegilops tauschii* identifies targets for bread wheat improvement. Nat Biotechnol 40:422–431

Geyer M, Mohler V, Hartl L (2022) Genetics of the inverse relationship between grain yield and grain protein content in common wheat. Plants 11:2146

Grewal S, Yang C, Edwards S, Scholefield D, Ashling S, Burridge AJ, King IP, King J (2018) Characterisation of *Thinopyrum bessarabicum* chromosomes through genome-wide introgressions into wheat. Theor Appl Genet 131:389–406

Grewal S, Hubbart-Edwards S, Yang C, Devi U, Baker L, Heath J, Ashling S, Scholefield D, Howells C, Yards J, Isaac P, King IP, King J (2020) Rapid identification of homozygosity and site of wild relative introgressions in wheat through chromosome-specific KASP genotyping arrays

Grewal S, Guwela V, Newell C, Yang C-y, Ashling S, Scholefield D, Hubbart-Edwards S, Burridge A, Stride A, King IP, King J (2021) Generation of doubled haploid wheat—*Triticum urartu* introgression lines and their characterisation using chromosome-specific KASP markers. Front Plant Sci 12:643636

Halder J, Zhang J, Ali S, Sidhu JS, Gill HS, Talukder SK, Kleinjan J, Turnipseed B, Sehgal SK (2019) Mining the genomic characterization of resistance to tan spot, Stagonospora nodorum blotch (SNB), and Fusarium head blight in Watkins core collection of wheat landraces. BMC Plant Biol 19:480

Horsnell R, Leight FJ, Wright TIC, Burrudge A, Ligeza A, Przewieslik-Allen MA, Howell P, Uauy C, Edwards KJ, Bentley AR (2022) A wheat chromosome segment substitution line series supports characterisation and use of progenitor genetic variation. Plant Genome e20288. https://doi.org/10.1002/tpg2.20288

Huang BE, George AW, Forrest KL, Kilian A, Hayden MJ, Morell MK, Cavanagh CR (2012) A multi-parent advanced generation inter-cross population for genetic analysis in wheat. Plant Biotechnol J 10:826–839

Huang D, Zheng Q, Melchkart T, Bekkaoui Y, Konkin DJF, Kagale S, Martucci M, You FM, Clarke M, Adamski NM, Chinoy C, Steed A, McCartney CA, Cutler AJ, Nicholson P, Feurtado JA (2020) Dominant inhibition of awn development by a putative zinc-finger transcriptional repressor expressed at the *B1* locus in wheat. New Phytol 225:340–355

International Wheat Genome Sequencing Consortium (IWGSC), Appels R, Eversole K, Stein N, Feuillet C, Keller B, Rogers J, Pozniak C et al (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science 361:eaar7191

Jambuthenne DT, Riaz A, Athiyannan N, Alahmad S, Ng WL, Ziems L, Afanasenko O, Periyannan SK, Aitken E, Platz G, Godwin I, Voss-Fels KP, Dinglasan E, Hickey LT (2022) Mining the Vavilov wheat diversity panel for new sources of adult plant resistance to stripe rust. Theor Appl Genet 135:1355–1373

Jones MK, Lister D (2022) The domestication of the seasons: the exploitation of variations in crop seasonality responses by later prehistoric farmers. Front Ecol Evol 10:907536

Jones H, Norris C, Cockram J, Lee D (2013) Variety protection and plant breeders' rights in the 'DNA era.' In: Lübberstedt T, Varshney R (eds) Diagnostics in plant breeding. Springer, Dordrecht, pp 369–402

Jordan KW, Wang S, Lun Y, Gardiner L-J, MacLachlan R, Hucl P et al (2015) A haplotype map of allohexaploid wheat reveals distinct patterns of selection on homoeologous genomes. Genome Biol 16:48

Jordan KW, Wang S, He F, Chao S, Lun Y et al (2018) The genetic architecture of genome-wide recombination rate variation in allopolyploid wheat revealed by nested association mapping. Plant J 95:1039–1054

Joukhadar R, Daetwyler HD, Bansal UK, Gendall AR, Hayden MJ (2017) Genetic diversity, population structure and ancestral origin of Australian wheat. Front Plant Sci 8:2115

Jupe F, Witek K, Verweij W, Sliwka J, Pritchard L et al (2013) Resistance gene enrichment sequencing (RenSeq) enables reannotation of the NB-LRR gene family from sequenced plant genomes and rapid mapping of resistance loci in segregating populations. Plant J 76:530–544

Keilwagen J, Lehnert H, Berner T, Badaeva E, Himmelbach A, Brner A, Kilian B (2022) Detecting major introgressions in wheat and their putative origins using coverage analysis. Sci Rep 12:1908

Khokhar JD, King J, King IP, Young SD, Foulkes MJ et al (2020) Novel sources of variation in grain zinc (Zn) concentration in bread wheat germplasm derived from Watkins landraces. PLoS ONE 15:e0229107

Khoury CK, Brush S, Costich DE, Curry HA, de Haan S, Engels JMM, Guarino L, Hoban S, Nercer KL, Miller AJ, Nabhan GP, Perales HR, Richards C, Riggins C, Thormann I (2021) Crop genetic erosion: understanding and responding to loss of crop diversity. New Phytol 233:84–118

Kidane YG, Gesesse CA, Hailemariam BN, Desta EA, Mengistu DK, Fadda C, Pè ME, Dell'Acqua M (2019) A large nested association mapping population for breeding and quantitative trait locus mapping in Ethiopian durum wheat. Plant Biotechnol J 17:1380–1393

King J, Grewal S, Othmeni M, Coombes B, Yang C-y, Walter N, Ashling S, Scholefield D, Walker J, Hunnart-Edwards S, Hall A, King IP (2022) Introgression of the *Triticum timopeevii* genome into wheat detected by chromosome-specific Kopetitive Allelle Specific PCR markers. Front Plant Sci 13:919519

Krasileva KV, Vasquez-Gross HA, Howell T, Bailey P, Paraiso F et al (2016) Uncovering hidden variation in polyploid wheat. Proc Natl Acad Sci USA 114:E913–E921

Kuzay S, Lin H, Li C, Chen S, Woods DP, Zhang J, Lan T, von Korff M, Dubcovsky J (2022) *WAPO-A1* is the coausal gene of the 7AL QTL for spikelet number per spike in wheat. PLoS Genet 18:e1009747

Lee S, Abecasis GR, Boehnke M, Lin X (2014) Rare-variant association analysis: study designs and statistical tests. Am J Hum Genet 95:5–23

Levy AA, Feldman M (2022) Evolution and origin of bread wheat. Plant Cell 34:2549–2567

Li H, Wang X (2009) *Thinopyrum ponticum* and *Th. intermedium*: the promising source of resistance to fungal and viral diseases of wheat. J Genet Genomics 36:557–565

Li A, Liu D, Yang W, Kishii M, Mao L (2018a) Synthetic hexaploid wheat: yesterday, today, and tomorrow. Engineering 4:552–558

Li DY, Long D, Li TH, Wu YL, Wang Y, Zeng J, Xu LL, Fan X, Sha LN, Zhang HQ, Zhou YH, Kang HY (2018b) Cytogenetics and stripe rust resistance of wheat–*Thinopyrum elongatum* hybrid derivatives. Mol Cytogenet 11:16

Lin M, Corsi B, Ficke A, Tan K-C, Cockram J, Lillemo M (2020a) Genetic mapping using a wheat multi-founder population reveals a locus on chromosome 2A controlling resistance to both leaf and glume blotch caused by the necrotrophic fungal pathogen *Parastagonospora nodorum*. Theor Appl Genet 133:785–808

Lin M, Stadlmeier M, Mohler V, Tan KC, Ficke A, Cockram J, Lillemo M (2020b) Identification and cross-validation of genetic loci conferring resistance to Septoria nodorum blotch using a German multi-founder winter wheat population. Theor Appl Genet 134:125–142

Lisker A, Maurer A, Schmutzer T, Kazman E, Cöster H, Holzapfel J, Ebmeyer E, Alqudah AM, Sannemann W, Pillen K (2022) A haplotype-based GWAS identified trait-improving QTL alleles controlling agronomic traits under contrasting nitrogen fertilization treatments in the MAGIC wheat population WM-800. Plants 11:3508

Luo MC, Gu YQ, Puiu D, Wang H, Twardziok SO et al (2017) Genome sequence of the progenitor of the wheat D genome *Aegilops tauschii*. Nature 551:498–502

Mackay IJ, Horwell A, Garner J, White J, McKee J, Philpott H (2011) Reanalyses of the historical series of UK variety trials to quantify the contributions of genetic and environmental factors to trends and variability in yield over time. Theor Appl Genet 122:225–238

Mackay I, Bansept-Basler P, Barber T, Bentley AR, Cockram J et al (2014) An eight-parent multiparent advanced generation intercross population for winter-sown wheat: creation, properties and validation. G3: Genes Genomes Genetics 4:1603–1610

Mackay IJ, Cockram H, Howell P, Powell W (2021) Understanding the classics: the unifying concepts of transgressive segregation, inbreeding depression and heterosis and their central relevance for crop breeding. Plant Biotechnol J 19:26–34

Mellers G, Aguilera JG, Bird N, Bonato ALV, Bonow S et al (2020) Genetic characterization of a wheat association mapping panel relevant to Brazilian breeding using a high-density single nucleotide polymorphism array. G3: Genes, Genomes, Genetics 10:2229

Naish M, Alonge M, Wlodzimierz P, Tock AJ, Abramson BW et al (2021) The genetic and epigenetic landscape of the *Arabidopsis* centromeres. Science 374:eabi7489

Othmeni M, Grewal S, Walker J, Yang C-Y, King IP, King J (2022) Assessing the potential of using the Langdon 5D(5B) substitution line for the introgression of *Aegilops tauschii* into durum wheat. Front Plant Sci 13:927728

Phan HTT, Jones DAB, Rybak K, Dodhia KN, Lopez-Ruiz FJ, Valade R, Gout L, Lebrun M-H, Brunner PC, Oliver RP, Tan K-C (2020) Low amplitude boom-and-bust cycles define the Septoria Nodorum Blotch interaction. Front Plant Sci 10:1785

Poland JA, Brown PJ, Sorrells ME, Jannink JL (2012) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS ONE 7:e32253

Pont C, Leroy T, Seidel M, Tondelli A, Duchemin W et al (2019) Tracing the ancestry of modern bread wheats. Nat Genet 51:905–911

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. Am J Hum Genet 67:170–181

Przewieslik-Allen AM, Wilkinson PA, Burridge A, Winfield MO, Dai X et al (2021) The role of gene flow and chromosomal instability in shaping the bread wheat genome. Nature Plants 7:172–183

Raz A, Dahan-Meir T, Melamed-Bessudo C, Leshkowitz D, Levy AA (2021) Redistribution of meiotic crossovers along wheat chromosomes by virus-induced gene silencing. Front Plant Sci 11:635139

Riaz A, Hathorn A, Dinglasan E, Ziems L, Richard C, Singh D, Mitrofanova O, Afanasenko O, Aitken E, Godwin I, Hickey L (2017) Into the vault of the Vavilov wheats: old diversity for new alleles. Genet Resour Crop Evol 64:531–544

Riaz A, Athiyannan N, Periyannan SK, Afanasenko O, Mitrofanova OP, Platz GJ, Aitken EAB, Snowdon RJ, Lagudah ES, Hickey LT, Voss-Fels KP (2018) Unlocking new alleles for leaf rust resistance in the Vavilov wheat collection. Theor Appl Genet 131:127–144

Riaz A, KockAppelgren P, Hehir J, Kang J, Meade F, Cockram J, Milbourne D, Spink J, Mullins W, Byrne S (2020) Genetic analysis using a multi-parent wheat population identifies novel sources of septoria tritici blotch resistance. MDPI Genes 11:887

Rimbert H, Barrier B, Navarro J, Kitt J, Choulet F et al (2018) High throughput SNP discovery and genotyping in hexaploid wheat. PLoS ONE 13:e0186329

Sansaloni C, Petroli C, Jaccoud D, Carling J, Detering F, Grattapaglia D, Kilian A (2011) Diversity arrays technology (DArT) and next-generation sequencing combined: genome-wide, high throughput, highly informative genotyping for molecular breeding of Eucalyptus. BMC Proc 5:P54

Sansaloni C, Franco J, Santos B, Percival-Alwyn L, Singh S et al (2020) Diversity analysis of 80,000 wheat accessions reveals consequences and opportunities of selection footprints. Nat Commun 11:4572

Sánchez-Martín J, Steueragel B, Ghosh S, Herren G, Hurni S et al (2016) Rapid gene isolation in barley and wheat by mutant chromosome sequencing. Genome Biol 17:221

Sannemann W, Lisker A, Maurer A, Léon J, Kazman E, Cöster H et al (2018) Adaptive selection of founder segments and epistatic control of plant height in the MAGIC winter wheat population WM-800. BMC Genomics 19:1–16

Scott M, Fradgley N, Bentley AR, Brabbs T, Corke F, Gardner KA, Horsnell R, Howell P, Ladejobi F, Mackay I, Mott R, Cockram J (2020a) Limited haplotype diversity underlies polygenic trait architecture across 70 years of plant breeding. Genome Biol 22:137

Scott MF, Ladejobi O, Amer S, Bentley AR, Biernaskie J et al (2020b) Multi-parent populations in crops: a toolbox integrating genomics and genetic mapping with pre-breeding. Heredity 125:396–416

Sehgal D, Mondal S, Crespo-Herrera L, Velu G, Juliana P, Huerta-Esoino J, Shrestha S, Poland J, Singh R, Dreisigacker S (2020) Haplotype-based, genome-wide association study reveals stable genomic regions for grain yield in CIMMYT spring bread wheat. Front Genet 11:589490

Senapati N, Semenov MA, Halford NG, Hawkesford MJ, Asseng S et al (2022) Global wheat production could benefit from closing the genetic yield gap. Nature Food 3:532–541

Shah R, Huang BE, Whan A, Newberry M, Verbyla K et al (2019) The complex genetic architecture of recombination and structural variation in wheat uncovered using a large 8-founder MAGIC population. bioRxiv, p 594317. https://doi.org/10.1101/594317

Sharma R, Cockram J, Gardner KA, Russell J, Ramsay L, Thomas WTB, O'Sullivan DM, Powell W, Mackay IJ (2022) Trends of genetic changes uncovered by Env- and Eigen-GWAS in wheat and barley. Theor Appl Genet 135:667–678

Simmonds NW (1995) The relation between yield and protein in cereal grain. J Sci Food Agric 67:309–315

Simmonds J, Scott P, Brinton J, Mestre TC, Bush M, del Blanco A, Dubcovsky J, Uauy C (2016) A splice acceptor site mutation in *TaGW2-A1* increases thousand grain weight in tetraploid and hexaploid wheat through wider and longer grains. Theor Appl Genet 129:1099–1112

Srinivasan CS, Thirtle C, Palladino P (2003) Winter wheat in England and Wales, 1924–1995: what do indices of genetic diversity reveal? Plant Genetic Resources 1:43–57

Stadlmeier M, Hartl L, Mohler V (2018) Usefulness of a multiparent advanced generation intercross population with a greatly reduced mating design for genetic studies in winter wheat. Front Plant Sci 871:1–12

Stadlmeier M, Nistrup Jørgensen L, Fejer Justesen A, Corsi B, Cockram J, Hartl L, Mohler V (2019) Genetic dissection of resistance to the three fungal plant pathogens *B. graminis*, *Z. tritici*, and *P. tritcirepentis* in the background of a multiparental winter wheat population. G3 9:1745–1757

Steed A, King J, Herwal S, Yang C-Y, Clarke M, Devi U, King IP, Nicholson P (2022) Identification of Fusarium head blight resistance in *Triticum timopheevii* accessions and characterization of wheat-*T. timopheevii* introgression lines for enhanced resistance. Frontiers in Plant Science 13:943211

The Complex Trait Consortium (2002) The collaborative cross, a community resource for the genetic analysis of complex traits. Nature Genetics 36:1133–1137

Thépot S, Restoux G, Goldringer I, Hospital F, Gouache D et al (2014) Efficiently tracking selection in a multiparental population: the case of earliness in wheat. Genetics 199:609–623

Tian X, Xia X, Xu D, Liu Y, Xie L, Hassan MA, Song J, Li F, Wang D, Zhang Y, Hao Y, Li G, Chu C, He Z, Cao S (2022) *Rht24b*, and ancient variation of *TaGA2ox-A9*, reduces plant height without yield penalty in wheat. New Phytol 233:738–750

Uauy C, Distelfeld A, Fahima T, Blechl A, Dubcovsky J (2006) A NAC gene regulating senescence improves grain protein, zinc, and iron content in wheat. Science 314:1298–1301

Walkowiak S, Gao L, Monat C, Haberer G, Kassa MT et al (2020) Multiple wheat genomes reveal global variation in modern breeding. Nature 588:277–283

Wang J, Luo M-C, Chen Z, You FM, Wei Y et al (2013) *Aegilops tauschii* single nucleotide polymorphisms shed light on the origins of wheat D-genome genetic diversity and pinpoint the geographic origin of hexaploid wheat. New Phytol 198:925–937

Wang S, Wong D, Forrest K, Allen A, Chao S et al (2014) Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. Plant Biotechnol J 12:787–796

Wang H, Sun S, Ge W, Zhao L, Hou B et al (2020) Horizontal gene transfer of Fhb7 from fungus underlies Fusarium head blight resistance in wheat. Science 368:6493

Wang Z, Deng Z, Kong X, Wang F, Guan J, Cui D, Sun G, Liao R, Fu M, Che Y, Hao C, Geng S, Zhang X, Zhou P, Mao L, Liu S, Li A (2022) InDels identification and association analysis with spike and awn length in Chinese wheat mini-core collection. Int J Mol Sci 23:5587

Watson A, Ghosh S, Williams MJ, Cuddy WS, Simmonds J et al (2018) Speed breeding is a powerful tool to accelerate crop research and breeding. Nature Plants 4:23–29

White J, Sharma R, Balding D, Cockram J, Mackay IJ (2022) Genome-wide association mapping of Hagberg falling number, protein content, test weight, and grain yield in U.K. wheat. Crop Sci 62:965–981

Winfield MO, Allen AM, Burridge AJ, Barker GLA, Benbow HR et al (2016) High-density SNP genotyping array for hexaploid wheat and its secondary and tertiary gene pool. Plant Biotechnol J 14:1195–1206

Winfield MO, Allen AM, Wilkinson PA, Burridge AJ, Baker GLA, Coghill J, Waterfall C, Wingen LU, Griffiths S, Edwards KJ (2018) High-density genotyping of the A. E. Watkins collection of hexaploid landraces identifies a large molecular diversity compared to elite bread wheat. Plant Biotechnology Journal 16:165–175

Wingen LU, Orford S, Goram R, Leverington-Waite M, Bilham L, Patsiou TS, Ambrose M, Dicks J, Griffiths S (2014) Establishing the A. E. Watkins landrace cultivar collection as a resource for systematic gene discovery in bread wheat. Theor Appl Genet 127:1831–1842

Wingen LU, West C, Leverington-Waite M, Collier S, Orford S, Goram R, Yang C-Y, King J, Allen AM, Burridge A, Edwards KJ, Griffiths S (2017) Wheat landrace genome diversity. Genetics 205:1657–1676

Wittern L, Steed G, Taylory LJ, Ramirez DC, Pingarron-Cardenas G, Gardner K, Greenland A, Hannah MA, Webb AAR (2022). Wheat *EARLY FLOWERING3* is a dawn-expressed circadian oscillator component that regulates heading date. bioRxiv. https://doi.org/10.1101/2021.09.03.458922.

Würschum T, Rapp M, Miedaner T, Longin CFH, Leiser WL (2019) Copy number variation of *Ppd-B1* is the major determinant of heading time in durum wheat. BMC Genet 20:64

Yan L, Loukoianov A, Tranquilli G, Helguera M, Fahima T, Dubcovsky J (2003) Positional cloning of the wheat vernalization gene *VRN1*. Proc Natl Acad Sci USA 100:6263–6268

Yan L, Loukoianov A, Blechl A, Tranquilli G, Ramakrishana W, SanMiguel P, Bennetzen JL, Echenique V, Dubcovsky J (2004) The wheat *VRN2* gene is a flowering repressor down-regulated by vernalization. Science 303:1640–1644

Yu J, Pressoir G, Briggs WH, Bi IV, Yamasaki M et al (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203–208

Yu J, Holland JB, McMullen MD, Buckler ES (2008) Genetic design and statistical power of nested association mapping in maize. Genetics 178:539–551

Zanella CM, Rotondo M, McCormick-Barnes C, Mellers G, Corsi B, Berry S, Ciccone G, Day R, Faralli M, Galle A, Gardner KA, Jacobs J, Ober E, Sanchez del Rio A, Van Rie J, Lawson T, Cockram J(2022) Longer epidermal cells underlie a quantitative source of variation for wheat flag leaf size. New Phytologist, online-first. https://doi.org/10.1111/nph.18676

Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C et al (2007) An Arabidopsis example of association mapping in structured samples. PLoS Genet 3:e4

Zhou Y, Tang H, Cheng M-P, Dankwa KO, Chen Z-X et al (2017) Genome-wide association study for pre-harvest sprouting resistance in a large collection of Chinese wheat landraces. Front Plant Sci 6:401

Zhou Y, Bai S, Li H, Sun G, Zhang D et al (2021) Introgressing the *Aegilops tauschii* genome into wheat as a basis for cereal improvement. Nature Plants 7:774–786