The Bread Wheat Reference Genome Sequence

1

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

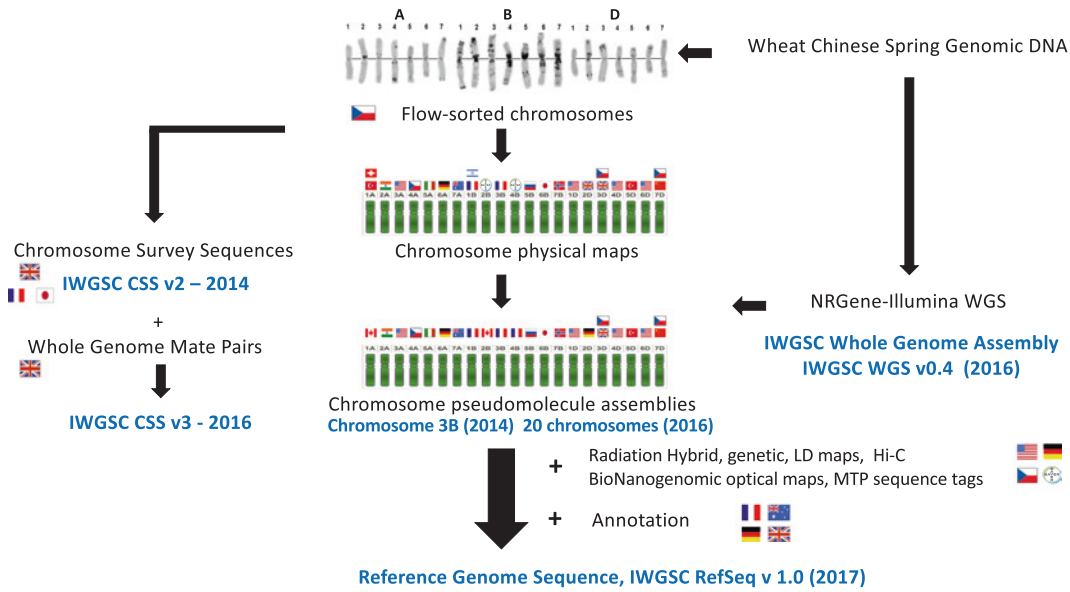

Fig. 1.1 Overview of the global community contributing to the sequencing of the wheat genome. National flags indicate the country-of-origin of the research groups contributing to the establishment of the high-quality *Triticum aestivum* cv. CHINESE SPRING reference genome assembly (IWGSC RefSeq v1.0) 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://triftdb.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