

Chapter 17

The Wheat Black Jack: Advances Towards Sequencing the 21 Chromosomes of Bread Wheat

Frédéric Choulet, Mario Caccamo, Jonathan Wright, Michael Alaux, Hana Šimková, Jan Šafář, Philippe Leroy, Jaroslav Doležel, Jane Rogers, Kellye Eversole and Catherine Feuillet

Contents

17.1	Introduction	406
17.2	What have we learned about the Wheat Genome so far?	407
17.2.1	The Wheat Genome: Three for One and One for All	407
17.2.2	Deciphering the Wheat Genome Composition and Organization: An ongoing Story	408
17.2.3	Tools and Technologies to Sequence, Assemble, and Annotate the Wheat Genome	413
17.2.4	Annotating the Wheat Genome	419
17.3	Strategies to Obtain a Reference Sequence of the Bread Wheat Genome	421
17.3.1	The Chromosome-Based Approach	421
17.3.2	MTP Sequencing of the 21 Wheat Chromosomes of Bread Wheat	423
17.3.3	Whole Genome Approaches can Support the Achievement of a Reference Wheat Genome Sequence	429

F. Choulet (✉) · P. Leroy · C. Feuillet
Genetics, Diversity and Ecophysiology of Cereals,
INRA Joint Research Unit 1095 Genetics, Clermont-Ferrand, France
e-mail: frederic.choulet@clermont.inra.fr

Genetics, Diversity and Ecophysiology of Cereals,
University Blaise Pascal Joint Research Unit 1095 Genetics,
Clermont-Ferrand, France

M. Caccamo · J. Wright · J. Rogers
The Genome Analysis Centre, Norwich Research Park, Colney, Norwich, UK

M. Alaux
INRA Centre de Versailles-Grignon, Unité de Recherche en Génomique-Info,
UR 1164, Versailles, France

H. Šimková · J. Šafář · J. Doležel
Centre of the Region Haná for Biotechnological and Agricultural Research,
Institute of Experimental Botany, Olomouc, Czech Republic

K. Eversole
International Wheat Genome Sequencing Consortium, Eversole Associates,
Wyoming Road 5207, Bethesda, USA

17.4 Integration of Wheat Sequence Information in Databases	430
17.4.1 Data Integration	430
17.4.2 Wheat Databases	431
References	434

Abstract Despite its socio-economic importance and the overall recognition that a reference genome sequence has great value for crop improvement, sequencing the wheat genome has long been considered “impossible” because of the sequencing cost and bioinformatic challenges associated with the assembly of the mostly repetitive 17 Gb hexaploid genome. In the past 5 years, however, new platforms and technologies have emerged that enabled the launching of an international effort to tackle the bread wheat genome sequence using a chromosome-by-chromosome approach. In this chapter, we review the features of the wheat genome as well as the tools and technologies that can be used to sequence, assemble, and annotate a large, complex, polyploid genome. We describe the strategies and current status of the efforts towards achieving a reference sequence for the 21 chromosomes of bread wheat. Finally, we present the databases that were established to support the integration of the sequence information with other genetic and biological information.

Keywords Wheat · Polyploid · Chromosome · Flow sorting · Physical map · Genome sequence · Next generation sequencing · Assembly · Annotation · Transposable elements · Database · Data integration

17.1 Introduction

The cultivation of wheat (*Triticum* spp.) reaches far back into history as it was one of the first domesticated food crops. For more than 8,000 years, wheat has been the basic staple food of the major civilizations of Europe, West Asia, and North Africa. Today, wheat is grown on more land area (255 million hectares) than any other crop and continues to be the most important food grain source for humans (<http://www.faostat.fao.org>). With changing diets and growing world populations, rising prices for fertilizers and phytosanitary products, increasing competition between food and non-food uses, and the negative effects of high temperature and drought resulting from climate change, food supplies must double in the next few decades to meet demand (Foley et al. 2011). Already, world wheat production has not been sufficient to meet demand in 6 out of the past 10 years (<http://faostat.fao.org>). Recent studies showed that annual, worldwide yield increases have slowed between 1995 and 2005 when compared with previous years (Brisson et al. 2010). Globally, models indicate that between 1980 and 2008 wheat production declined by 5.5 % as a result of climate trends (Lobell et al. 2011). Thus, to meet the challenge of delivering safe, high-quality, and health-promoting food and feed in an economically, environmentally sensitive, and sustainable manner, a paradigm shift is needed in wheat breeding and genetics.

A combination of new tools, methods, and germplasm resources must be established for wheat to facilitate this paradigm shift. One such resource that will underpin future wheat improvement is a high quality, reference genome sequence as it will 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. In the past decade, sequencing of model plant genomes, such as those of *Arabidopsis thaliana* and rice has revolutionized our understanding of plant biology and paved the way for the development of genome sequencing projects for a number of crops (Feuillet et al. 2011). Similar advances in wheat have been hampered by the giant size of the genome, ~ 17 Gb (haploid size), due to its high content in repeats ($> 90\%$) and its hexaploid nature ($2n = 6x = 42$; 21 pairs of homologous chromosomes originating from three homoeologous sets, referred to as the A, B, and D subgenomes, each with 7 chromosome pairs).

The international collaboration on wheat genome sequencing was launched after a USDA-NSF funded workshop that confirmed the need for sequencing the wheat genome and assessed different strategies and objectives (Gill et al. 2004). These objectives were to (i) construct an accurate, sequence-ready physical map (ordered BAC contigs) of the reference hexaploid wheat (*Triticum aestivum* L.) cultivar Chinese Spring for which large genetic stocks of aneuploid lines are available, (ii) assess the feasibility of a chromosome-specific approach (i.e., constructing maps of each of the 21 chromosomes using chromosome-specific BAC libraries), and (iii) explore different strategies for gene enrichment. Following these discussions, the International Wheat Genome Sequencing Consortium (IWGSC; <http://www.wheatgenome.org>) was launched in 2005 with the aim of advancing agricultural research for wheat production and utilization by developing DNA-based tools and resources that result from the complete sequence of the common (hexaploid) wheat genome (Feuillet and Eversole 2007).

In this chapter, we present current knowledge about the composition and organization of the wheat gene and transposable element (TE) spaces, the status of technologies and methods available to date to sequence, assemble, and annotate the wheat genome, and the complementary strategies that are being developed to obtain a reference sequence of the bread wheat genome. Finally, we present the different types of databases that are currently available to access the sequence information and link it to other types of datasets to provide scientists and breeders an opportunity for an optimal exploitation of this essential information in their programs.

17.2 What have we learned about the Wheat Genome so far?

17.2.1 *The Wheat Genome: Three for One and One for All*

One of the major features of the wheat genome is a high content in TEs ($\sim 80\%$; Smith and Flavell 1975) that resulted from massive amplifications in the ancestral genome before its divergence from related species of the *Triticeae* tribe around 15 million

years ago (MYA). Individual diploid genomes from different genera (e.g. *Aegilops*, *Triticum*, *Agropyrum*) subsequently evolved independently in the past 2.5–4 MY. During the evolution of the wheat lineage, spontaneous hybridizations occurred between diploid genomes resulting in different polyploid species. Hence, the ancestral allohexaploid *Triticum aestivum* species resulted from two hybridization events that brought together the diploid A, B, and D genomes ($2n = 6x = 42$; AABBDD) (McFadden and Sears 1946; Dubcovsky and Dvorak 2007). A first hybridization occurred about 0.5 MYA between a species related to *Triticum urartu* ($2n = 2x = 14$; A^uA^u) and one or more species from the Sitopsis section, including *Aegilops speltoides* ($2n = 2x = 14$; SS) which is the closest known relative to the B genome. The resulting fertile tetraploid ($2n = 4x = 28$; AABB) was domesticated over 10,000 years ago to become known as emmer wheat (*Triticum turgidum*), eventually giving rise to the subspecies *T. turgidum* ssp. *durum*, the ancestor of the durum wheat of today, also known as pasta wheat. Some 8,000 years ago, tetraploid emmer wheat reached the region south of the Caspian Sea and hybridized with *Aegilops tauschii* ($2n = 2x = 14$), a wild diploid species with a D genome leading to a fertile hexaploid species with an AABBDD genome – the ancestral *Triticum aestivum*, or bread wheat (Zohary and Hopf 2000). Subsequent to these hybridization and polyploidization events, a number of structural and functional rearrangements resulted in genome stabilization which was accompanied by slight reductions of 2–10 % in the size of the individual homoeologous genomes compared to their diploid ancestors (Feldman and Levy 2009).

As a result of this complex history, the bread wheat genome has one of the largest and TE-rich plant genomes (40 times larger than the rice genome; IRGSP 2005) and for a long time it was considered “impossible” to sequence. The organization and composition of the wheat genome can be depicted as two main components with different evolutionary dynamics and relative importance: a small conservative part that is subjected to selection pressure and mostly corresponds to the gene space, and a much larger and more variable component which is under more dynamic evolution and comprises the TE space as well as duplicated genes and gene fragments.

17.2.2 Deciphering the Wheat Genome Composition and Organization: An ongoing Story

17.2.2.1 The Gene Space

EST and mRNA Sequencing

In 1998, a major effort on sequencing polyA-tailed transcripts was undertaken by the wheat scientific community, mostly within the framework of the International Triticeae EST Cooperative (ITEC <http://avena.pw.usda.gov/genome/>) initiative, to compensate for the lack of wheat sequences in the public databases. In total 1,073,845 ESTs were produced (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html). In parallel, a major effort launched in Japan resulted in the production of 17,525

full length cDNAs (<http://trifldb.psc.riken.jp> and <http://srs.ebi.ac.uk>). cDNAs and gene-enriched genomic clones were hybridized on series of aneuploid lines lacking complete chromosomes or chromosome segments (Sears 1954; Sears and Sears 1978; Endo and Gill 1996) to locate genes along the wheat chromosomes, thereby proving first evidence of the gene space organization. A series of experiments conducted by Gill and collaborators (Gill et al. 1996a; Gill et al. 1996b; Sandhu and Gill 2002; Erayman et al. 2004) suggested that a large majority of the genes are located in a few regions (so called “gene-rich regions”) representing about 30 % of the genome. These gene-rich regions were found mainly in the distal part of the wheat chromosomes where recombination takes place. The rest of the genome, considered as “gene-poor” would be composed mostly of TEs and a few isolated genes (Sandhu and Gill 2002). Subsequently, Qi et al. (2004) mapped more than 7,000 wheat ESTs (16,099 loci) into chromosome bins and, although they confirmed previous observations on the increasing gene density towards the chromosome ends, they also found genes in proximal (centromeric) regions, suggesting a less contrasted pattern of gene distribution along the wheat chromosomes. Such experiments also enabled *in silico* comparative analyses with the rice genome sequence to study colinearity at high resolution. For example, using 4,485 wheat ESTs mapped in deletion bins, Sorrells and collaborators (Sorrells et al. 2003; La Rota and Sorrells 2004) performed a wheat-rice genome comparison and showed that wheat chromosome group 3 is the most conserved while chromosome group 5 is the least conserved compared to rice.

ESTs, however, cannot provide complete information about the wheat gene space composition. First, they do not represent the full gene content since they are limited to loci that are expressed at a level sufficient for cloning. Second, they provide only partial information on gene structure because of their limited read length (about ~550 bp on average). Finally, they provide very redundant information as many ESTs originate from highly expressed genes. A minimal gene set of 40,935 wheat unigenes has been defined by NCBI recently (<http://www.ncbi.nlm.nih.gov/unigene>) through EST clustering. However, such a clustering prevents the identification of duplicated copies of gene families (due to polyploidy or single gene duplication) and represents an underestimation of the wheat gene set. Recently, this unigene set was used to build a NimbleGen array (Rustenholz et al. 2011) and study the expression of 3,000 genes assigned to the chromosome 3B physical map. In conclusion, despite their limitations, wheat ESTs have been a useful resource for genetic and physical mapping, and for structural annotation of genes on genomic sequences.

End Sequencing of Short and Long DNA Clone Inserts

The first attempts to explore the wheat genome beyond its transcriptome were based on low-pass survey sequencing of random genomic fragments corresponding to the ends of plasmid and bacterial artificial chromosome (BAC) libraries. Analysis of 3 Mb of plasmid end sequences produced from *Ae. tauschii* (Li et al. 2004) indicated that 68 % of the sequence consisted of TEs whereas the annotation of 11 Mb of BAC end sequences (BES) from chromosome 3B of hexaploid wheat (Paux et al.

2006) revealed a repeat content of 86 %. Further, this study estimated a gene number of 6,000 for chromosome 3B which was extrapolated to suggest 36,000 genes per diploid genome in wheat. This contrasted dramatically with the predictions of Rabinowicz et al. (2005) that suggested as much as 98,000 genes for each of the three genomes of bread wheat based on the analysis of 1,597 plasmid ends from a methyl-filtration library. The results of these random analyses have provided some insights into the genome composition, but have not illuminated its organization (into gene islands for instance). Moreover, extrapolation of observations derived from such partial sequence datasets to the whole genome remains speculative. First, the insert end sequences may be biased, especially in wheat as repetitive sequences introduce bias in frequency of restriction sites that are used for cloning BAC and plasmid inserts. This was revealed during the analysis of the frequency and distribution of *Hind*III sites along 18 Mb of large sequenced contigs (see below) which showed that they were 1.5-fold overrepresented in TEs compared to the expected rate on random sequences. The bias in TE associated-*Hind*III sites that were used to build the BAC library thus resulted in an overrepresentation of TE sequences in the dataset and a bias towards repeats. Second, low-pass survey sequences cannot be assembled into sequence contigs thereby limiting the assessment of the organization of genes and TEs along the chromosomes. Finally, the limited length of sequence information precludes distinguishing pseudogenes from functional genes and discerning recently duplicated gene copies. Thus, insert-end sequencing can contribute to our knowledge of the wheat genome composition but cannot address questions related to organization such as gene and TE distribution, gene duplication, and Copy Number Variations (CNVs) that are increasingly relevant to understanding biological functions and polymorphisms associated with traits.

Individual BAC and BAC-contig Sequencing

During map-based cloning projects and comparative studies at disease resistance, storage protein, grain hardness, domestication or vernalization loci (for review see Feuillet and Salse 2009), a number of BAC clones and small BAC contigs were sequenced to obtain additional information on the composition and organization of the wheat genome. Analyses of the 3.8 Mb representing all wheat genomic sequences available in the public databases in 2005 revealed an average gene density of 1 gene per 24 kb and only ~55 % of TEs, thereby indicating a clear bias toward gene-rich regions in these first samples (Sabot et al. 2005). This was likely due to the fact that most of the regions targeted in the map-based cloning projects were telomeric and contained gene families. In contrast, sequencing randomly chosen BAC clones from the entire genome or sequencing the ends of BAC contigs representing complete chromosomes provided evidence for a more homogeneous gene distribution in the wheat genome. In a preliminary study of 4 BAC clones from a genomic BAC library of bread wheat cv. Chinese Spring, Devos et al. (2005) estimated an average gene density of 1/75 kb while Charles et al. (2008) reported a gene density of 1/100 kb homogeneously distributed after sequencing 10 BACs (1.43 Mb) randomly chosen

from a chromosome 3B-specific BAC library. Subsequently, draft sequences of 217 additional BAC clones from the Chinese Spring BAC library have been deposited in the databases (AC200765-851; AC207901-60; AC216550-85; AC232247-62; AC238983-88 and DQ767609-30) by Bennetzen and colleagues and, the automated annotation of 67 BACs (ses.library.usyd.edu.au) suggested a density of 1 gene/64 kb. While these analyses provided new insights into the organization and composition of the genome, the sampled regions represented only a minute fraction of the genome (< 0.1 %) and information remains limited to addressing TE dynamics and gene distribution as individual BAC sequences (100-150 kb) contain on average 1 or 2 genes and a few TEs most of which are sequenced only partially.

Recently, the sequencing of large Mb-sized BAC contigs from chromosome 3B (Choulet et al. 2010) and *Ae. tauschii* (Massa et al. 2011) provided novel information about the organization and content of the gene and TE spaces. In a first study, Choulet et al. (2010) sequenced and annotated 13 Mb-sized BAC contigs (18 Mb in total) selected from different regions of chromosome 3B of bread wheat cv. Chinese Spring, and performed Whole Chromosome Shotgun (WCS) sequencing. They showed that (i) genes are present along the entire chromosome and are clustered mainly (75 %) into numerous small islands of 3–4 genes separated by large blocks of TEs that are less than 1 Mb long, and (ii) genome expansion occurred homogeneously along the chromosome through specific TE bursts. In addition, the study revealed accelerated evolution through tandem or interchromosomal gene duplications in telomeric regions that led to an increase in the gene number in wheat compared to related grasses. Gene insertion activity did not disrupt dramatically the ancestral gene “backbone” but led to an increased number of non collinear genes in wheat compared to the other species. These gene rearrangements combined with the differential insertion or removal of specific TE families resulted in a contrasted sequence composition that is now observed between the proximal and distal regions of the wheat chromosomes. Based on the 148 protein coding genes identified in the 13 contigs, Choulet et al. suggested that about 8,400 genes are located on chromosome 3B and estimated a total number of 50,000 genes per diploid genome in bread wheat. In a second study, Massa et al. (2011) sequenced and annotated 9 different regions totaling 9.7 Mb of the 4 Gb D genome from the wild ancestor *Ae. tauschii* and suggested a total gene number of 36,371. Comparison with the rice, brachypodium and sorghum genomes indicated that *Ae. tauschii* had 7,813 more genes than the grass ancestral genome which was estimated to comprise 28,558 genes.

While these BAC and BAC contig analyses provided additional information about the organization and evolution of the wheat genome, the reliability of the extrapolations made from such partial data to the whole genome remains questionable. This is illustrated by the discrepancies in the estimation of gene density and total gene number observed between the different studies. The gene density estimates ranged from 1 in 75 kb in Devos et al. (2005) to about 1 in 100 kb in Charles et al. (2008) and Choulet et al. (2010), and up to 1 in 110 kb in Massa et al. (2011). All of these were higher than the 1 in 165 kb estimated by low-pass survey BAC end sequencing (Paux et al. 2006). Gene number estimates have had even larger differences ranging from

about 36,000 (Massa et al. 2011; Paux et al. 2006) to 50,000 genes (Choulet et al. 2010) and up to 98,000 genes per diploid wheat genomes (Rabinowicz et al. 2005).

17.2.2.2 The TE Space

Early studies of reannealing kinetics of single-stranded DNA fragments (Flavell et al. 1977) indicated that the wheat genome is composed of more than 83 % of repeated sequences. Staining of mitotic chromosomes with DNA fluorochromes such as DAPI, highlighted frequent heterochromatic bands that were shown to correspond to compacted chromatin containing an abundance of repeats. Estimates of the copy number of specific and highly repeated elements were obtained from sequencing and hybridization-based genome reconstruction calculations. Some retrotransposon families, such as the WIS family, were shown to be present in tens of thousands of copies (Muniz et al. 2001). Ultimately, the sequencing of individual BAC clones (Feuillet and Keller 1999; Wicker et al. 2001; Yan et al. 2003; Chantret et al. 2004), BAC end sequences (Li et al. 2004; Paux et al. 2006), and, more recently, large BAC contigs (Choulet et al. 2010; Massa et al. 2011) provided a more detailed view of the composition of the TE fraction. These studies confirmed that the TE fraction represents more than 80 % of the wheat genome with class I LTR retrotransposons comprising the majority. They also showed that less than 10 families (Fatima, Jorge, Angela, Laura, Sabrina, WIS, Wilma, and Nusif) account for > 50 % of the TE fraction and that different TE families are found in centromeric and telomeric regions (Choulet et al. 2010). The annotation of 18 Mb of large contigs of chromosome 3B (Choulet et al. 2010) revealed 3222 TEs including 800 new families indicating the potential for TE discovery in complete BAC contig sequences. These new elements were integrated into the TREP database that is dedicated to TE annotation in wheat, barley and rye (Wicker et al. 2002). In addition, the peak of transposition activity 1.4 MYA suggests that TEs have been mostly silent in wheat since that time. Interestingly, although very fragmented by nested insertions, the majority of transposable elements identified in the Mb-sized contigs were complete. Such features were missed by previous single BAC analyses because the extremities of the same element into which dozens of other TE have inserted can be as much as 200 kb apart i.e. larger than a typical BAC size (Choulet et al. 2010). These data indicate that the wheat genome expansion due to TE activity has mainly stopped and that no significant contraction has occurred yet.

Based on all of these data, a model of the organisation and evolution of the wheat genome at the chromosome level (Fig. 17.1) can be proposed. The main features of the model are (1) genes are found everywhere along the chromosomes, (2) there are no gene deserts larger than a few Mb, (3) the genes are mostly organised in small and numerous gene islands with a slightly higher density of those towards the telomeres, and (4) there are more genes and a significant number of non collinear genes than found in related grass models. While evidence for some of the mechanisms underlying the dynamics of the gene and TE spaces (e.g. TE insertions, illegitimate and unequal recombination, interchromosomal and tandem duplications) have been

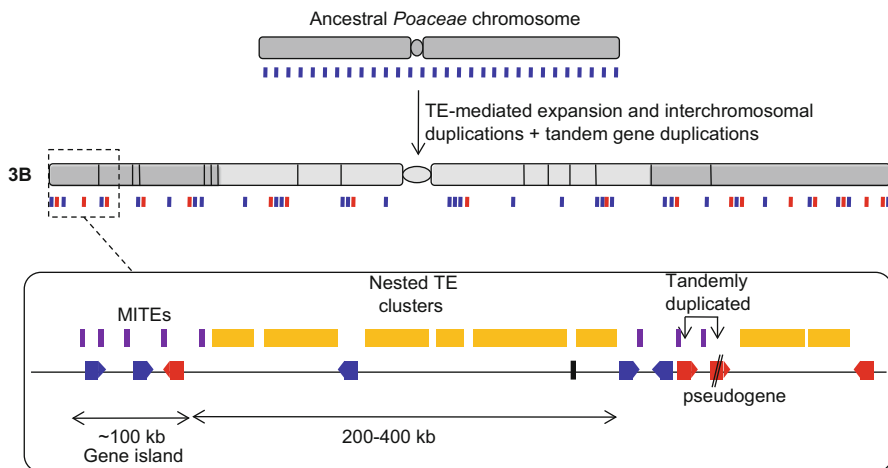


Fig. 17.1 A model for the evolution of the size, content, and organization of a wheat chromosome. Genes are represented by filled arrows: blue representing genes syntenic with other grass species and red representing locus-specific non syntenic wheat genes. TE clusters are represented by yellow rectangles while MITEs are in purple

obtained by comparative analyses of homoeologous loci (for review see Feuillet and Salse, 2009) and gene family studies (Akhunov et al. 2007), the full extent and relative impact of these mechanisms on the sequence organization remain unknown. Access to a fully annotated reference sequence will help significantly to improve our understanding of the organization and evolution of the wheat genome.

17.2.3 Tools and Technologies to Sequence, Assemble, and Annotate the Wheat Genome

17.2.3.1 Sequencing Technologies and Strategies

Prior to 2005, DNA sequencing was mostly based on the chain terminator DNA sequencing technique developed by F. Sanger (Sanger et al. 1977). Improvements in Sanger's original method that used fluorescently labeled dideoxynucleotides, automated fragment separation, and label detection facilitated the development of genomics and enabled the sequencing of entire genomes by the turn of the century (The Arabidopsis Genome Initiative 2000; Lander et al. 2001; Venter et al. 2001; Waterston et al. 2002). Sanger sequencing can produce reads of up to 1000 bases in length with low error rates, but it is relatively expensive and throughput is limited to ca. 1 million bases per day. In 2005, the emergence of the first of a second generation of sequencing technologies based on highly parallelised 'pyrosequencing' (Margulies et al. 2005) initiated a revolution in DNA sequencing and heralded the start of the competitive DNA sequencing field that exists today.

This was followed by the advent of massively parallel technologies such as the ‘sequencing-by-synthesis’ method developed by Solexa and now commercialized by Illumina (Bennett et al. 2005), and ‘Sequencing by Oligonucleotide Ligation and Detection’ (SOLiD) (Valouev et al. 2008) developed by Applied Biosystems (now Life Technologies). All of these second generation technologies rely on an amplification step to achieve the density of DNA fragments required to detect the signal of the individual bases in the DNA being sequenced. Generally, compared to Sanger sequencing, the second generation technologies produce much shorter sequencing reads, in much greater quantities, and at significantly reduced cost per sequenced base (Metzker 2009). While this reduced cost enabled the *de novo* sequencing of a large number of plant genomes (Feuillet et al. 2011), assembling large and complex repetitive genomes, such as wheat, using only short reads remains a significant challenge as the typical read lengths (100–500 bp) are not long enough to span the nested TEs that can be a few thousand bases long (Choulet et al. 2010). More recently, third-generation technologies arose that allow direct sequencing of single DNA strand and remove the dependency on an intermediate amplification step thereby reducing the execution times and potential artifacts that can arise during amplification. The Single Molecule Real Time (SRMT) technology developed by Pacific Biosciences (Eid et al. 2009) uses a polymerase attached within a nano-scale chamber to synthesize DNA complementary to a single stranded molecule attached to the polymerase within the chamber. As nucleotides are incorporated, fluorescent labels corresponding to each nucleotide are detected upon incorporation. Read lengths of up to 10 kb have been reported using this technology although reads are typically much shorter than this (1–3 kb) with the current chemistry. Another single-molecule sequencing technology is based upon detection of a change in electric current across a nanopore inserted into a lipid bilayer as individual bases are cleaved by an exonuclease from a DNA strand and pass through the nanopore (Astier et al. 2006). The identity of each base is detected by the amplitude of the change in electrical current as it passes through the nanopore. By combining high throughput and long read length, these technologies hold a great potential for sequencing large and complex genomes. Projects are underway to assess these technologies for sequencing the wheat genome (<http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=1032105>; Feuillet/Rogers/Wincker unpublished data).

The first plant genome sequencing projects of *Arabidopsis thaliana* (AGI 2000) and rice (International Rice Genome Sequencing Project 2005) were based on sequencing genome fragments cloned into vectors such as cosmids or bacterial artificial chromosomes (BACs). This is referred to as a “BAC-by-BAC” approach. In this approach, the clones are usually selected from a physical map that is assembled from the overlaps identified between BAC restriction fragment fingerprints to yield contigs representing the order of the cloned fragments in the original genome. By selecting clones from the physical map that contain fragments making up the genome sequence with minimal (~30 %) overlap between them, a minimal tiling path (MTP) of clones is defined to provide a template for sequencing. Then, each of these clones is sequenced using shotgun sub-cloning of random fragments, each clone sequence is reassembled from the shotgun reads, and the complete genome sequence is

reconstructed from the known positions and orientation of the sequenced clones. In the alternative whole genome shotgun (WGS) approach, the genome sequence is assembled from sequence reads generated from fragments of different sizes (generally ranging from 1 kb to 5 kb) obtained from the complete genome. The most effective assemblies have been built from sequence reads derived from fragments cloned into appropriately sized vectors and sequenced from both ends, yielding two reads from each clone, referred to as “mate-pairs”. As explained in more detail below, WGS sequencing depends on the development of assembly algorithms based on overlap graphs. In this approach, overlapping reads are first used to assemble regions of contiguous sequence called contigs, then mate-pair information is used to link contigs to generate longer scaffolds, thus providing a larger-scale framework for assembly of the smaller contigs of DNA (Fig. 17.2). With the emergence of the NGS technologies, the WGS approach has been established as the strategy of choice as reflected by a number of recently published genome sequences (Feuillet et al. 2011). Although the WGS approach is faster and generally cheaper, the assembly stage is much more complex increasing the risk of generating chimeric regions and misassemblies. Therefore, for sequencing large and complex plant genomes, such as wheat, the BAC-by-BAC approach is still preferred as it offers the best opportunity to generate a high quality reference sequence for the long term.

17.2.3.2 Assembling Genome Sequences

The Assembly Problem

Assembly is the reconstruction of the original genome sequence from the sequencing reads. The success of this task is limited by factors that range from the inherent complexity of genomes caused by repeat content and structure to technical issues related to specific sequencing technologies biases (e.g. error profiles and non-uniform coverage). In the general assembly pipeline, sequencing reads are assembled first into contigs, then long-range positional information, such as that obtained from mate-pairs, is used to join contigs together into scaffolds (Fig. 17.2). All assembly methods are highly dependent on the depth of coverage to which a genome is sampled, in other words, the degree of redundancy that is generated with respect to the target. For example, when enough sequence is generated to cover the target sample 10 times, the coverage is described as 10x. This sampling redundancy is essential for dealing with sequencing errors in the reads and ensures that reads will overlap sufficiently to allow accurate assembly. The first assembly algorithms were developed on the basis of the “overlap-layout-consensus approach”, i.e., the overlap information between sequencing reads is represented in a graph where the nodes are sequences and the edges connect overlapping reads. The paths in these graphs represent contigs from the underlying genome. Tools such as ARACHNE (Jaffe et al. 2003), ATLAS (Havlak et al. 2004) and more recently Newbler (<http://www.454.com>) utilize this strategy. One disadvantage of this approach is that the graph and therefore the required computer memory grow linearly with the number of reads. This is particularly

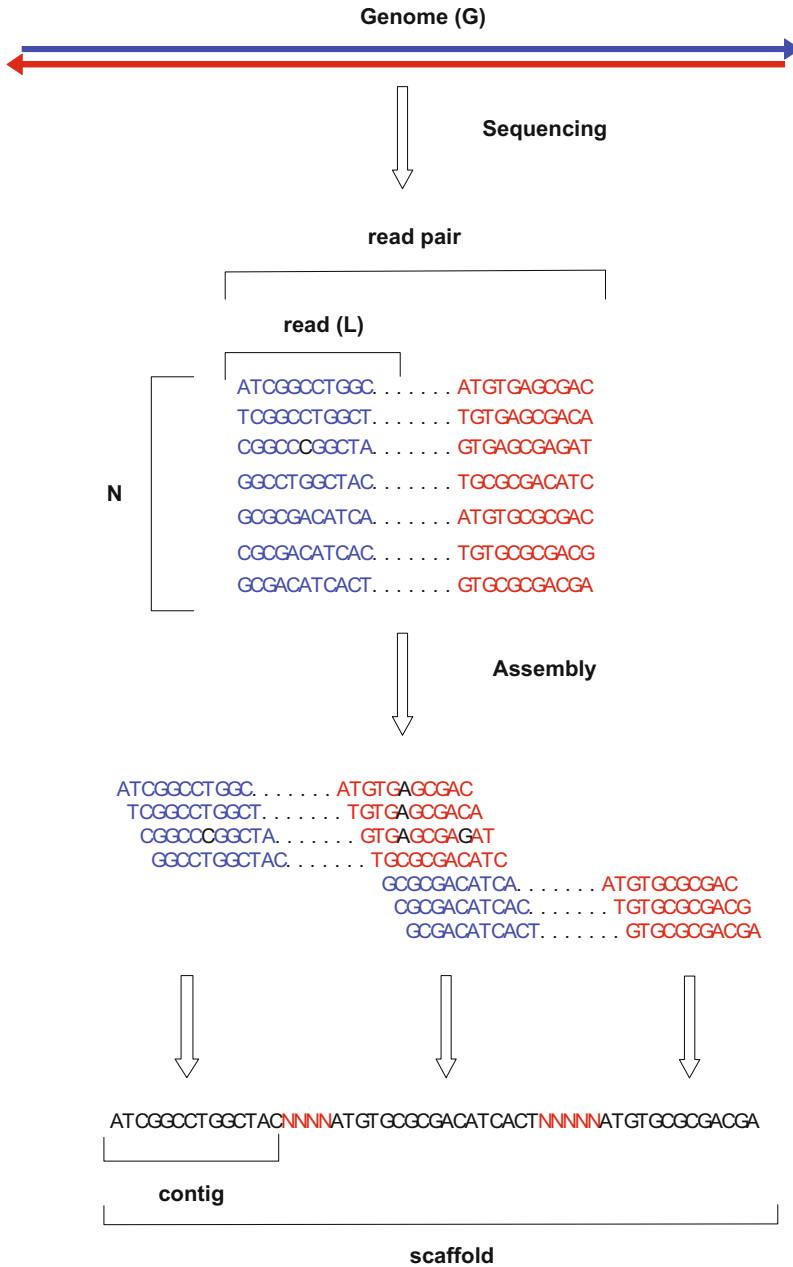


Fig. 17.2 Schematic representation of the sequence assembly process: N refers to the number of reads and L to the (average) length of the reads

relevant for NGS experiments that can generate billions of sequences per sample. The difficulty is exacerbated further when assembling large genomes as many sequencing runs are required to obtain adequate coverage. For instance, in order to generate 10x coverage of a typical 400 Mb wheat chromosome arm, 4×10^7 Illumina reads of 100 bp are required. A 10x coverage of the entire bread wheat genome represents 1.7 billion reads of 100 bp.

An alternative assembly approach pioneered by (Pevzner et al. 2001) is to represent the overlap information generated by looking only at words of fixed length k or k -mers, the so-called *de Bruijn* graphs (Fig. 17.3). The main advantage of this approach is that the graph scales with the number of observed k -mers in the dataset rather than the number of reads. In this way the graph can be built in linear time rather than the quadratic time taken with an overlap approach. In a *de Bruijn* graph, continuous stretches of nodes joined with edges represent sequence and sequencing errors appear as bubbles or tips that end abruptly (Fig. 17.3). Assembly tools generally simplify the graph by amalgamating the continuous stretches and correcting errors by bubble removal and tip clipping. Repetitive regions that share common k -mers, however, will be collapsed in single nodes generating cycles in the graph. Similarly to traditional assemblers based on sequence overlap, mate-pair reads are essential for traversing repeats in a genome and most NGS assembly algorithms incorporate this type of reads to resolve repetitive regions into scaffolds.

The Velvet assembly tool was one of the first software packages based on a *de Bruijn* graph approach to assemble short-read sequence data from bacterial genomes (Zerbino and Birney 2008). To assemble larger eukaryotic genomes, new techniques of memory optimization and parallelization were required. Algorithms such as those implemented in the open-source software packages ABySS (Simpson et al. 2009), SOAPdenovo (Li et al. 2009), ALLPATHS-LG (Gnerre et al. 2011), and Cortex (<http://cortexassembler.sourceforge.net>) allow the assembly process to be broken into multiple smaller chunks that can be distributed over multiple nodes in a high performance computing cluster. Commercial solutions are also available such as the *de novo* assembler from CLC Bio (<http://www.clcbio.com>), reputedly able to assemble a human sized genome on a single desktop computer. The SGA assembler is also designed to assemble large genomes and uses a string graph approach to determine overlaps between reads (Simpson and Durbin 2010). As well as being both memory efficient and relatively fast, SGA is the first algorithm that optimizes an overlap approach sufficiently such that it can be used to assemble large NGS datasets.

Generally, the Newbler assembler is used to assemble 454 reads as it is optimized for the error profiles and longer reads generated by the Roche platform. As such, it is being used in many of the wheat chromosome-based sequencing projects using the Roche platform such as the 3BSEQ project (<http://urgi.versailles.inra.fr/Projects/3BSeq>). In contrast, datasets generated on the Illumina platform are much larger due to the deeper coverage being generated and thus require a distributed assembly approach in addition to large amounts of computer memory. For example, as part of the IWGSC effort to obtain survey sequences of the 21 bread wheat chromosomes, the ABySS assembler is being used by The Genome Analysis Centre (<http://wheatdcc.tgac.bbsrc.ac.uk/index.php>) to produce homogeneous assemblies.

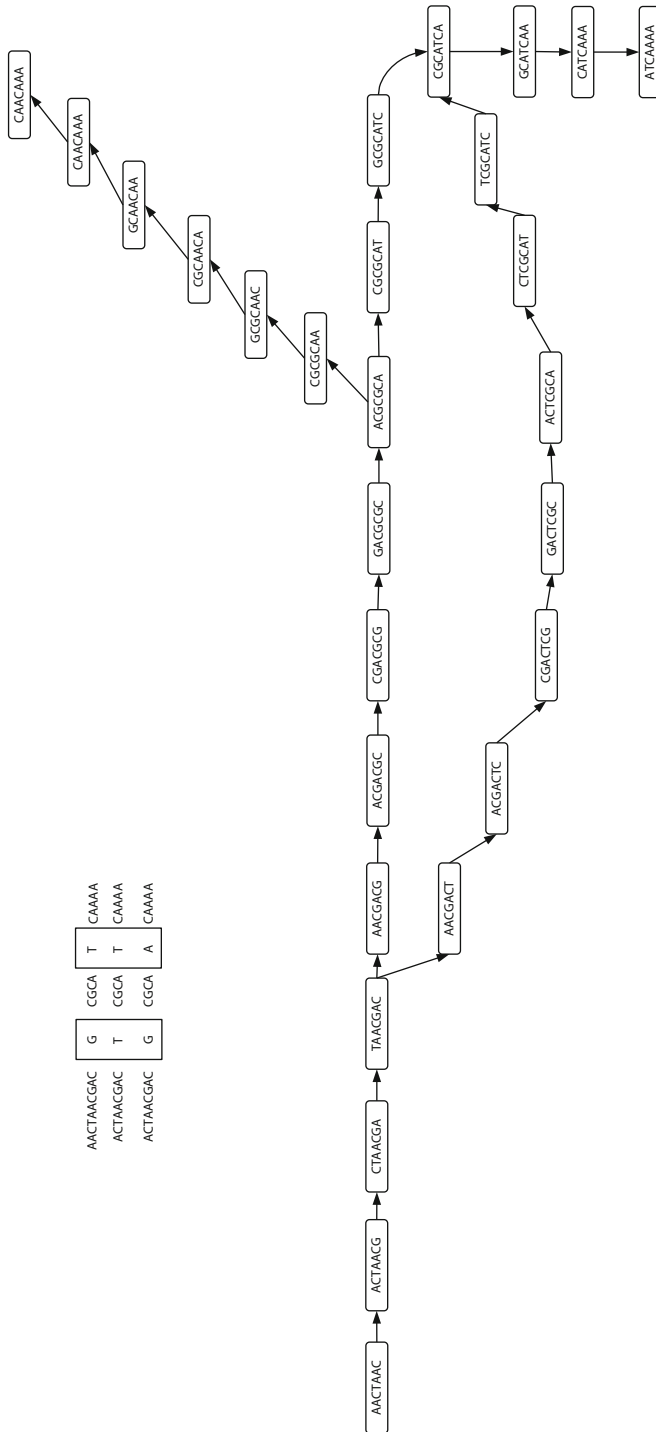


Fig. 17.3 The *de Bruijn* Graph structure. A *k-mer* length of 7 has been used to construct this graph with each *k-mer* shown as a node and joined to related nodes by edges. Errors in reads cause bubbles and tips in the graph

A Velvet-based strategy has also been adopted by other groups who have obtained comparable results on wheat chromosome group 7 (Berkman et al. 2011a, b).

17.2.3.3 Assessing the Quality of an Assembly

The N50 statistic is one of the most common metrics used to assess the quality of an assembly. After sorting the assembly contigs by size from largest to smallest, the N50 is defined as the length (L) of the contig such that half of the assembled bases are in contigs equal to or longer than L . Intuitively, the N50 represents in one figure a tradeoff between the number and median length of contigs; i.e., more complete assemblies with fewer and larger contigs will have a higher N50. The N50 statistic is further defined as ‘contig-N50’ for contigs and ‘scaffold-N50’ calculated using the scaffolds generated from the contigs. In a recent “Assemblathon” competition designed to evaluate various assembly tools on a simulated dataset (Earl et al. 2011), a NG50 statistic was introduced. Instead of using the total length of the assembled contigs as an estimation of the genome size, the NG50 used the average length of two simulated haplotypes to assess the assemblies. This allowed the authors to determine how well different assembly tools deal with haplotype-specific polymorphisms and to calculate the relative contributions of reads from each haplotype to the final assembly. Other metrics are also used to assess the quality of assembly such as sequence completeness, gene content, preservation of gene structures, and estimation of misassemblies. Generally, NGS sequencing methods are more prone to base calling errors than Sanger sequencing, thus the sequencing reads they produce are inherently “noisier” than Sanger reads. In addition to looking at the quality scores of called bases, “denoising” techniques such as k -mer frequency analysis or spectral correction alignment (Chaisson et al. 2004) can be applied to sequence data prior to assembly. The goal of these methods is to identify the k -mers (and reads) that occur at low frequency in the dataset and are likely to be erroneous as it is assumed that unique k -mers or reads should not occur with adequate sequencing depth.

17.2.4 Annotating the Wheat Genome

Achieving a robust structural and functional genome sequence annotation is essential to provide the foundation for further relevant biological studies. Genome annotation consists of identifying and attaching biological information to sequence features and it represents one of the most difficult tasks in genome sequencing projects. Genome annotation is generally a long and recursive process the difficulty of which increases with the size and complexity of the genome. It relies on a successive combination of software, algorithms, and methods, as well as the availability of accurate and updated sequence databanks. To manage the large amount of data generated by > 1 Gb genome size sequencing projects, sequence annotation needs to be automated, i.e. performed through a pipeline that combines all different programs and minimizes the subsequent long and laborious manual curation step. There are four different

categories of pipelines available. First, simple commercial softwares such as Vector NTI from invitrogen and DNASTAR from GATC can be used. However, these are not available on the web and they cannot be easily customized for specific needs. Second, suites of scripts that generate computational evidence for further manual curation have been developed in individual laboratories. For example, to annotate wheat BACs, DAWGPAWS was developed by (Estill and Bennetzen 2009) as a series of command line programs resulting in GFF output files. This type of pipeline, however, is also not available on the web and can only be used by skilled bioinformaticians. Third, “in house” automated pipelines have been developed by communities to annotate model plant genomes, e.g. rice (Ouyang and Buell 2004; IRGSP 2005) whereas major genomic resource centers such as the DOE/JGI (<http://www.phytozome.net/>), the MIPS (<http://mips.helmholtz-muenchen.de/plant/genomes.jsp>), Gramene (Liang et al. 2009), GenBank (<http://www.ncbi.nlm.nih.gov/genome/guide/build.shtml>), and the Ensembl project (Curwen et al. 2004) have developed their own pipelines to deal with multiple annotation projects. Although these pipelines are of high quality and generally are based on massive informatics resources, they are not accessible online. Finally, a number of automated annotation pipelines available on the web have been developed to assist communities in their efforts to annotate individual contigs of complete genomes. The first pipeline of this kind, RiceGAAS (Sakata et al. 2002), was developed originally for the annotation of the rice genome. Since then, a few others have been established such as DNA subway (iPlant, USA – <http://dnasubway.iplantcollaborative.org/>), FPGP (Amano et al. 2010), and MAKER (Cantarel et al. 2008). Each of these have user-friendly web interfaces; however, the online access does not enable the annotation of large genomes in a reasonable timeframe.

To support the annotation of the 17 Gb wheat genome sequence and to provide a useful resource to other communities coping with large and complex genomes, the “TriAnnot” pipeline was developed (Leroy et al. 2012). TriAnnot was conceived to: (1) enable automated, rapid, and robust structural and functional annotation of genes, transposable elements, and non-coding features; (2) be versatile, i.e. accessible through a user-friendly web interface to allow for the rapid analysis of a few hundred sequences or through a server for efficient and robust massive analysis in large scale projects (several thousand sequences); and (3) provide output files that can be retrieved and analyzed easily or that can be visualized directly on a web interface. Moreover, to ensure efficient use of the sequence information, TriAnnot enables links between the annotation and databases containing genetic and physical maps, markers, genes, QTL, phenotypes, ‘omics’ data, variomes, etc. that are available at the INRA URGI (<http://urgi.versailles.inra.fr/gnpis/>). To date, TriAnnot consists of four main panels for: (1) transposable element annotation and masking; (2) structural and functional annotation of protein-coding genes; (3) the identification of non-coding RNA genes and conserved non-coding sequences; and, (4) marker development. The performance of TriAnnot was evaluated in terms of sensitivity, specificity, and general fitness using curated reference sequence sets from rice and wheat. The results showed that TriAnnot predicted and annotated 83 and 93 % of the rice and wheat reference gene sets, respectively, with 54–67 % of those in accordance with the

reference annotations (Leroy et al. 2012). On the wheat dataset, TriAnnot demonstrated a higher fitness than three other pipelines that were not improved for wheat thereby proving its usefulness for the annotation of the wheat genome. The pipeline is accessible at <http://www.clermont.inra.fr/triannot> and is parallelized on a computing cluster that can run a 1 Gb sequence annotation in less than five days. Additional improvements such as the integration of modules to improve the annotation of TEs and non-coding RNA genes, automate the design of molecular markers, and permit online manual curation of the sequences are underway.

17.3 Strategies to Obtain a Reference Sequence of the Bread Wheat Genome

17.3.1 *The Chromosome-Based Approach*

All plant genomes sequenced so far are of diploid species and they were obtained either by sequencing BAC clones originating from the whole genome or by whole genome shotgun approaches (Feuillet et al. 2011). Because of the structural features of the wheat genome, in particular the high degree of identity between the homoeologous gene sets and the high amount of LTR transposable elements, sequencing the 21 chromosomes of wheat as a whole with the current technologies and tools would not deliver sufficient information to qualify as a reference genome sequence. Even if assembly of genic and low-copy regions is possible to some extent, such assemblies cannot be considered as finished reference genome sequences equivalent to those that have been produced for rice and *Arabidopsis*, or even the draft versions of other plant genomes (Feuillet et al. 2011). Critical information, such as recently duplicated gene families which have been shown to play a key role in speciation and non-coding intergenic sequences that also carry relevant biological information as illustrated in whole genome functional analyses of the human genome (Birney et al. 2007), would be lost in assemblies performed on the whole genome. In addition, ordering and assigning sequence scaffolds to individual chromosomes would be a daunting task.

Thus, to ensure the delivery of a reference sequence containing the information necessary for crop improvement and advanced biological studies, the IWGSC embarked on a chromosome- by-chromosome approach several years ago. The objective of this approach is to reduce the complexity of the wheat genome to manageable sizes by dealing with individual chromosomes and chromosome arms that range from 224–800 Mb and thus are comparable to already sequenced plant genomes (Dolezel et al. 2007). Previously, other strategies such as methyl filtration (Li and Gill 2004) and Cot fractionation (Lamoureux et al. 2005) have been proposed to reduce complexity and sequencing costs in wheat, however, their efficiency was too low to be applied at large scale. Unlike these strategies, the chromosome-based approach reduces sample size and complexity by sequencing smaller genome parts without sacrificing sample information content. A second important advantage of this approach is that without the

confounding effects of homoeologous sequences, the assembly of sequence reads is considerably simplified. Finally, it permits international collaboration in which chromosomes are sequenced by individual teams thereby facilitating cost sharing and rapid application into the numerous wheat breeding programs around the world.

The chromosome-based approach has been made possible due to the technological advancements obtained in the group of J. Doležel in Olomouc (Czech Republic) that enabled high-throughput chromosome isolation using flow cytometry (Vrána et al. 2000). Their approach involves preparation of liquid suspensions of intact chromosomes from root tips of hydroponically grown seedlings and classification of chromosomes stained with a DNA-specific fluorochrome according to fluorescence intensity (DNA content). Those chromosomes that can be resolved from others are then sorted. In a majority of wheat cultivars, including the IWGSC reference cultivar Chinese Spring, flow cytometric chromosome analysis (flow karyotyping) results in distribution of relative DNA content (flow karyotype) comprising three composite peaks (I – III) representing groups of chromosomes and a peak of chromosome 3B (Vrána et al. 2000). This is the only chromosome that can be purified from most of polyploid wheats (Kubaláková et al. 2002; Kubaláková et al. 2005) and thus the pioneering experiments with chromosome genomics were made with this chromosome. In some cultivars, other chromosomes than 3B can be resolved due to chromosome polymorphism (Kubaláková et al. 2002).

Cytogenetic stocks, such as telosomic and ditelosomic lines that carry chromosome arms as telocentric chromosomes (telosomes), have proven to be the most efficient method for further dissecting the polyploid wheat genome. Telosomic lines were created originally by Sears (1954) and have been used in flow cytometry to isolate chromosome arms as small as 224 Mbp (Šafář et al. 2010) which represents only 1.3 % of the genome. Of the 42 wheat chromosome arms, the long arms of chromosomes 3B (3BL) and 5B (5BL) are too large to be resolved from the peak of chromosomes 1D, 4D, and 6D. However, these two arms can be sorted as isochromosomes (chromosomes with both arms genetically identical). The complete set of telosomic lines described by Sears and Sears (1978) carry telocentric chromosomes originating from Chinese Spring with the exception of telosomes 2BS, 2DL, 4BS, 5DS, 6AL, and 7DL (designated later 7DS based on homology with 7S *Triticeae* chromosome arms) that originated from other varieties, and were backcrossed (up to 10 times) to Chinese Spring. A few of the telosomes are not truly telocentric and carry a very short second arm.

The potential utility of the chromosome-based approach depends on the sorting throughput, the purity of sorted fractions, and the quality of DNA prepared from flow-sorted chromosomes (Dolezel et al. 2004). DNA of wheat chromosomes purified according to Vrana et al. (2000) is intact and the chromosomes are suitable for preparation of high-molecular-weight DNA. The sample rate (and hence the sort rate) and the resolution of chromosome peaks on flow karyotype (and hence the purity in the sorted fractions) are inversely related. Typical rates at which wheat chromosomes are sorted range from 10–40/sec and it is realistic to sort between 200,000–400,000 chromosomes per working day using one flow sorter. For the largest chromosome 3B, this equals to 0.4–0.8 µg DNA, while for the smallest chromosome arm 1DS

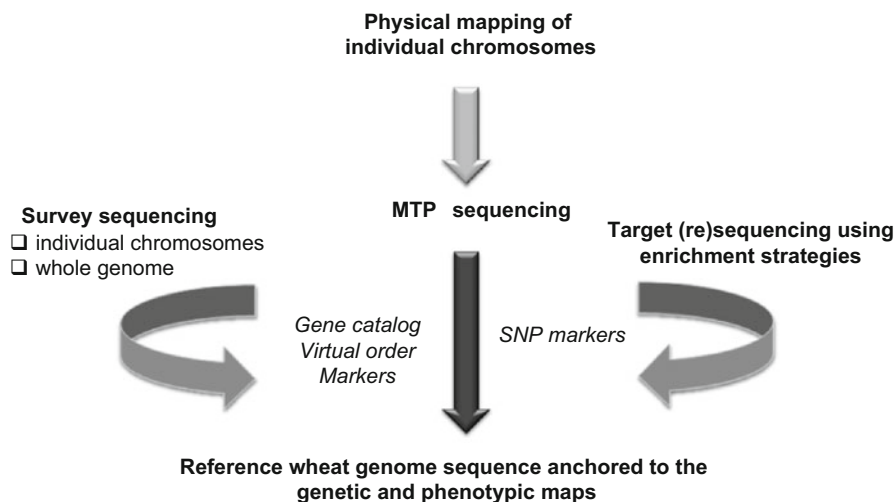


Fig. 17.4 Schematic representation of the combined strategies used by the International Wheat Genome Sequencing Consortium to obtain a reference sequence of the bread wheat genome (*T. aestivum* L. cv. Chinese Spring)

this translates to 0.09–0.18 μg DNA. The purity is best evaluated after sorting chromosomes onto a microscopic slide and fluorescent labeling of DNA repeats such as GAA microsatellites and *Afa* repeats, resulting in a chromosome-specific labeling pattern. In addition, a probe for telomeric sequence is used to confirm that intact telosomes have been sorted (Kubaláková et al. 2005). These analyses reveal that sorted wheat chromosome fractions are contaminated with fragments of chromosomes and chromatids, with no single prevailing chromosome. The purities determined during large-scale sorting experiments ranged from 80 % (telosomes 4AS, 4AL, 7AS) to 94 % (telosome 1AS). The fact that flow-sorted wheat chromosomes are intact and their DNA is of high molecular weight makes them suitable for applications ranging from BAC library construction to direct shotgun sequencing on amplified DNA (Dolezel et al. 2007).

17.3.2 MTP Sequencing of the 21 Wheat Chromosomes of Bread Wheat

17.3.2.1 Physical Mapping

In 2005, the International Wheat Genome Sequencing Consortium selected the chromosome-based approach as a foundation for its sequencing strategy for the hexaploid wheat genome (Fig. 17.4). In this strategy, physical maps are developed using chromosome-specific BAC libraries, a MTP is established for each chromosome/chromosome arm, and the BACs from the MTP are sequenced using any technology that will deliver a high quality reference sequence. To support this approach

and provide useful information (markers, gene content, etc.) while waiting for the MTP sequencing of each chromosome, whole chromosome shotgun sequences can also be produced using sorted chromosomes and combined with the physical map information (Fig. 17.4).

One of the first priorities of the IWGSC was to establish a physical map of the 21 wheat chromosomes. The physical maps, alone, facilitate the map-based isolation of genes and QTL for traits of agronomic importance and the identification of new and favorable alleles from the huge reservoir of genetic resources that are present in seed banks all over the world. Further, the development of genetically anchored physical maps can accelerate wheat improvement through the delivery of numerous markers for breeding. To reach this first priority, 37 chromosome/chromosome arm specific BAC libraries were constructed as follows: one library for chromosome 3B (Šafář et al. 2004), a composite library for chromosomes 1D, 4D, and 6D (Janda et al. 2004), and one for each arm of the remaining 34 chromosome arms. The small size of chromosome BAC libraries, typically consisting of only $3\text{--}10 \times 10^4$ clones for a chromosome coverage of $10\text{--}15 \times$ (Table 17.1), makes them easy to maintain and use (Šafář et al. 2010). As a comparison, a whole genome BAC library with similar parameters would include about 2×10^6 clones. The first chromosome-specific BAC library was constructed with only a few micrograms of chromosomal DNA from wheat chromosome 3B (Safar et al. 2004) and was used successfully to construct the first physical map of this chromosome (Paux et al. 2008), thereby validating the feasibility of constructing “sequence ready” physical maps of hexaploid wheat with a chromosome-by-chromosome approach. In total, 37 chromosome BAC libraries comprising 2,253,312 clones (Table 17.1) are available. While the use of chromosome/chromosome arm specific libraries provides important advantages, it is important to bear in mind potential limitations. The first relates to the fact that they have been constructed using a single restriction enzyme – *HindIII* (Šafář et al. 2010) which may result in uneven chromosome coverage and gaps in physical maps. The first results obtained with the 3B physical map, however, indicate 99% coverage of the chromosome (Rustenholtz et al. 2011) thereby suggesting that *HindIII* BAC libraries can provide a sufficient substrate for sequencing. One reason the coverage is sufficiently high may be related to a bias towards *HindIII* restriction sites in wheat that were observed during the sequencing of large contig sequences (Choulet et al. 2010). In fact, *HindIII* sites seem overrepresented in some of the highly repeated TE families in wheat and therefore in the whole genome. The results obtained from the first chromosome based physical maps also showed that the contamination of libraries with BAC clones originating from other chromosomes does not hamper physical map assembly as the contaminating clones are far less abundant and do not assemble into contigs. The wheat chromosome specific BAC libraries (Table 17.1) are currently being used to complete the physical map of the bread wheat genome (for a progress status as August 2013, see Fig. 17.5). Sequencing of the MTP is completed for chromosome 3B (C. Feuillet et al. unpublished) and is underway for chromosome 7B (O-A. Olsen, personal comm.).

Table 17.1 List of wheat cv. Chinese Spring chromosome-specific BAC libraries (as of January 2012)^a

Library code	Chromosome	Number of clones	kb size	Purity (%)	Coverage
TaaCsp146hA ^b	1D, 4D, 6D	87,168	85kb	91	3.4x
TaaCsp146hB ^c	1D, 4D, 6D	148,224	102kb	91	6.9x
TaaCsp146hC ^d	1D, 4D, 6D	138,240	116kb	91	7.4x
TaaCsp146eA ^{c,e}	1D, 4D, 6D	26,112	110kb	90	1.3x
TaaCsp1AShA ^c	1AS	31,104	111kb	94	11.8x
TaaCsp1ALhA ^c	1AL	49,536	103kb	83	8.0x
TaaCsp1ALhB ^d	1AL	43,008	109kb	87	7.7x
TaaCsp1BSShA ^d	1BS	55,296	113kb	81	15.7x
TaaCsp1BLhA ^d	1BL	92,160	114kb	81	15.4x
TaaCsp2AShA ^c	2AS	56,832	123kb	87	15.4x
TaaCsp2ALhA ^c	2AL	76,800	120kb	88	15.8x
TaaCsp3AShA ^b	3AS	55,296	80kb	89	10.9x
TaaCsp3AShB ^c	3AS	55,296	115kb	91	15.9x
TaaCsp3ALhA ^c	3AL	55,296	106kb	87	10.2x
TaaCsp3ALhB ^d	3AL	24,576	114kb	88	5.2x
TaaCsp3BFhA ^b	3B	67,968	103kb	89	6.2x
TaaCsp3BFhB ^c	3B	82,176	126kb	93	9.1x
TaaCsp3BFeA ^{c,e}	3B	21,120	107kb	93	1.9x
TaaCsp3DShA ^c	3DS	36,864	110kb	90	11.0x
TaaCsp3DLhA ^c	3DL	64,512	105kb	82	12.2x
TaaCsp4AShA ^d	4AS	49,152	131kb	82	16.6x
TaaCsp4ALhA ^d	4AL	92,160	126kb	81	17.3x
TaaCsp5AShA ^d	5AS	46,080	120kb	90	16.5x
TaaCsp5ALhA ^d	5AL	90,240	123kb	88	18.3x
TaaCsp5BSShA ^d	5BS	43,776	122kb	90	15.8x
TaaCsp5DShA ^d	5DS	36,864	137kb	88	17.0x
TaaCsp5DLhA ^d	5DL	72,960	128kb	87	16.0x
TaaCsp6AShA ^d	6AS	46,080	130kb	92	16.2x
TaaCsp6ALhA ^d	6AL	55,296	123 kb	86	15.7x
TaaCsp6BSShA ^d	6BS	57,600	132kb	85	15.3x
TaaCsp6BLhA ^d	6BL	76,032	130kb	92	18.0x
TaaCsp7AShA ^d	7AS	58,368	134kb	81	15.4x
TaaCsp7ALhA ^d	7AL	61,056	124kb	84	15.3x
TaaCsp7BSShA ^d	7BS	27,648	182kb	87	12.5x
TaaCsp7BLhA ^d	7BL	72,960	136kb	84	15.1x
TaaCsp7DShA ^c	7DS	49,152	114kb	84	12.2x
TaaCsp7DLhA ^c	7DL	50,304	115kb	89	14.8x
Total		2,253,312			

^aThe actual list and further details can be obtained from <http://olomouc.ueb.cas.cz/dna-libraries/cereals>

^bFirst-generation BAC libraries (one DNA size selection step)

^cSecond-generation BAC libraries (two DNA size selection steps resulting in larger inserts)

^dSecond-generation BAC libraries cloned in phage-resistant *E. coli* cells

^eBAC libraries prepared using *EcoRI* restriction endonuclease (other BAC libraries were constructed using *HindIII*)

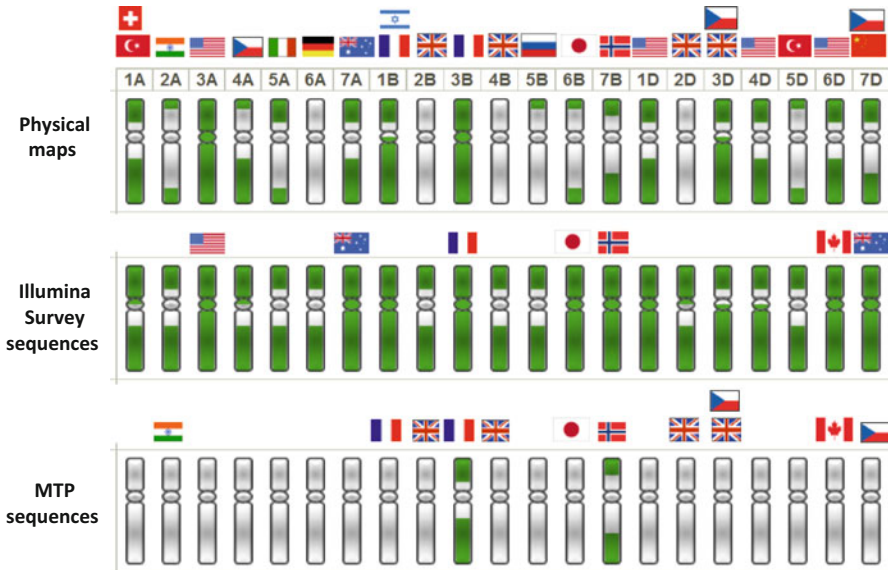


Fig. 17.5 Schematic representation of the current status (August 2013) of the IWGSC efforts to establish physical maps (a), sequence survey by Illumina (b), and reference sequences (c) of the 21 chromosomes of hexaploid wheat. The completion of each project is represented with green bars within each chromosome and chromosome arm. Four levels of completion are displayed (25, 50, 75 and 100%). The flags represent the country of origin of the laboratories in charge of the currently funded projects. When two flags are displayed, the upper flag corresponds to the short arm and the lower flag the long arm. Except for those chromosomes with a specific flag, Illumina survey sequences were obtained by The Genome Analysis Centre (TGAC) with support from BBSRC, Biogemma, Graminor, ICARDA and INRA. For more details and regular updates, see <http://www.wheatgenome.org/Projects>

17.3.2.2 Survey Sequencing

Whole chromosome shotgun sequencing efforts began simultaneously with the production of physical maps for a number of chromosomes. It began at low-coverage to develop markers and subsequently at greater coverage to perform assemblies and gain an initial picture of the gene content. In principle, chromosomes can be sorted in adequate quantities ($> 10^6$) to obtain a sufficient amount of DNA for sequencing. However, as this would require several weeks of sorting for each of the 42 arms, DNA amplified from smaller number of sorted chromosomes ($\sim 10^4$) has been used instead. The amplification was optimized first in barley by (Simkova et al. 2008) using a commercial kit for multiple displacement amplification that achieved a highly representative amplification as validated with 1426 SNP markers. This protocol was applied to obtain 1–2x sequence coverage of barley chromosome 1H (Mayer et al. 2009) and individual arms of the remaining barley chromosomes 2H–7H (Mayer et al. 2011) using the Roche-454 technology. Integration of such low coverage shotgun sequencing information with the gene order of orthologous rice, brachypodium, and

sorghum genes (“GenomeZipper analysis”) enabled the establishment of a virtual order along the barley chromosomes for 21,766 genes (Mayer et al. 2011).

Initial survey sequencing experiments in wheat were performed for chromosome arms of the homoeologous group 1 (Wicker et al. 2011) as well as chromosome arms 7DS and 7BS (Berkman et al. 2011a, b). Wicker et al. (2011) performed Roche-454 sequencing of individual arms of the 1A, 1B, and 1D chromosomes of bread wheat and the orthologous 1H barley chromosome at low coverage (1.3–2.2 ×). At this coverage, it is not possible to perform sequence assembly, predict genes *de novo*, or distinguish genes from pseudogenes. Nevertheless, a partial gene catalog could be established for each individual chromosome arm through comparisons with sequenced genomes and used to provide for the first time an extended comparison of the gene content between 3 homoeologous wheat chromosomes. This confirmed, at the whole chromosome scale, the high proportion of non collinear genes observed previously on a subset of physical contigs for chromosome 3B (Choulet et al. 2010). Indeed, while more than 85 % of the identified genes were found in at least one chromosome, the number of non collinear genes (6158) identified for each chromosome arm largely exceeded the number of collinear genes. Moreover, the fact that 2248 nonsynthetic genes are conserved between at least two of the 1A, 1B, and 1H chromosomes suggests selection pressure and functionality for those genes. Finally, in addition to providing information about the genome composition, Roche-454 chromosome survey sequencing provided a very useful resource for designing molecular markers, such as Insertion Site Based Polymorphism (ISBP) that require read lengths of more than 200 bp to properly identify both sides of a TE junction (Paux et al. 2011).

Survey sequences of chromosome arms 7BS and 7DS (Berkman et al. 2011a, b) were produced by Illumina with coverage of 30x and 34x, respectively, using paired-end reads of 35–100 bp. The main advantages of this approach are the high coverage that can be achieved for a limited cost and the possibility of assembling the reads. The 7DS assembly resulted in 571,038 contigs accounting for only 40 % of the expected length of the chromosome arm because of collapsing reads from identical repeated elements. Comparison with markers assigned to 7DS indicated that a majority of the expected genes are present in the assembly. However, with a contig N50 value of 1.2 kb, it is likely that most genes are split into several contigs and the capacity to distinguish recently duplicated genes and pseudogenes is limited. Similar to the group 1 chromosome analysis (Wicker et al. 2011), comparative sequence studies with other grass genomes confirmed the presence of a significant proportion (31 % of the genes identified on 7DS) of non syntenic genes. In the analysis of the 7BS assembly, the position of a previously reported translocation between chromosome arms 7BS and 4AL was delimited with a resolution of one or a few genes and it was estimated that approximately 13 % of the genes have been translocated from chromosome arm 7BS to 4AL. Finally, the gene content of the 7DS and 7BS syntenic builds was used to derive an estimate of about 77,000 genes in the hexaploid wheat genome which is significantly less than the predicted 100,000–150,000 genes based on random plasmid, BAC end and BAC contig analyses (see above). These pioneering studies were followed by Roche 454 sequencing of both arms of chromosomes 5A (Vitulo et al. 2011) and 4A (Hernandez et al. 2011) further contributing to unraveling the genome structure, gene content, and gene order in hexaploid wheat.

To complete these efforts and rapidly provide chromosome-specific sequences to scientists and breeders, the IWGSC decided to systematically produce at least 50x or greater coverage for all 21 bread wheat chromosomes (<http://www.wheatgenome.org/Projects/IWGSC-Bread-Wheat-Projects/Sequencing/Whole-Chromosome-Survey-Sequencing>) using Illumina paired-end reads (Fig. 17.4). Although these assemblies will suffer from the same limitations as described above and will provide only partial information, they will generate a unique resource to perform *in-silico* gene and SNP marker mapping, thereby reducing the costs and labor of identifying gene locations in the hexaploid wheat genome. The sequence is generated from flow-sorted DNA for all chromosomes/chromosome arms to a depth of between 50 and 200x for each arm. For the initial stage of the project the ABySS assembler (Simpson et al. 2009) has been used to assemble the sequence reads and find the optimal k -mer length. Based on an assessment of the assemblies generated using various k -mer lengths, 71 was chosen as the optimal length to use. The N50 length of the contigs larger than 200 bp was between 1 and 4 kb (mean N50 = 2.1 kb). It varied between chromosome arms due to sequence complexity, sequencing coverage, and quality of the dataset used to generate the assembly. The order of gene-containing contigs was then inferred from synteny with the fully sequenced rice, brachypodium, and sorghum genomes using the “GenomeZipper” approach (Mayer et al. 2009). The assemblies are stored in the wheat repository at the URGI website (<http://urgi.versailles.inra.fr/Species/Wheat/Sequence-Repository>) and are publicly available.

17.3.2.3 MTP Sequencing

Typically, with an average insert size of ~ 120 kb, 10 BACs from a MTP are required to cover about 1 Mb of sequence. Therefore, sequencing the MTP of the 17 Gb bread wheat genome will require sequencing about 170,000 BACs. To reduce sequencing cost without losing information, BAC pooling strategies have been proposed (Rounsley et al. 2009). A reference sequence of the 1 Gb chromosome 3B is currently under production (<http://urgi.versailles.inra.fr/Projects/3BSeq>) using a pooling strategy in which 8 kb paired-end libraries are built using barcoded pools of 10 BACs from the MTP (Paux et al. 2008; Rustenholz et al. 2011) and subsequently sequenced using the Roche-454 GSFLX Titanium technology. Less than a thousand pools were needed to cover the entire 3B chromosome, and approximately 150 runs were performed to reach at minimum 30x coverage of each genomic region. A similar pooling strategy can be employed with the Illumina sequencing technology; however, the number of barcodes used simultaneously will need to be very high to optimize the coverage of each pool. Currently, HiSeq 2000 instruments can produce about 200 Gb per run. Thus, assuming that 100x coverage is sufficient to correctly assemble wheat genomic regions, 2 Gb of MTP BAC pools (i.e., 1000 pools of 2 Mb each) could theoretically be sequenced in a single run. This will require 1000 different barcodes and the capacity to ensure that DNA concentration remains equimolar in all pools to ensure equal coverage. To date,

funding has been secured for the reference MTP sequencing of 10 of the 21 wheat chromosomes (www.wheatgenome.org) and the completion of the wheat genome sequence has been recently added to the action plan on food price volatility and agriculture of the G20 within the frame of the international initiative for wheat research (http://agriculture.gouv.fr/IMG/pdf/2011-06-23_-_Action_Plan_-_VFinale.pdf).

17.3.2.4 Towards the Gold Standard Reference Sequence

Sequencing ordered MTP BAC clones will not be enough to obtain one pseudo-molecule per chromosome in a complex genome such as wheat as physical maps likely will not cover fully the entire chromosomes and small gaps will remain. Further, within the MTP contigs, not all of the sequence scaffolds will be oriented and ordered. To address some of these limitations and achieve a high quality draft sequence of chromosome 3B, a hybrid strategy that combines high coverage Illumina shotgun sequencing of DNA from sorted 3B DNA with the Roche-454 Titanium sequencing of the MTP BACs has been deployed. Sanger sequencing of BAC-ends is also being used in this strategy to facilitate the assembly of super-scaffolds for each BAC contig. Finally, markers will be designed on every sequenced super-scaffold to build high density genetic maps that will be used to order as many scaffolds as possible along the chromosome using various mapping populations (recombinants, radiation hybrid panels). While such an anchored high quality draft sequence will provide invaluable information to the scientists and breeders, additional information is needed to achieve a gold standard reference of the wheat genome. The next step change required to obtain this is a sequencing technology that can provide reads long enough to help resolve the repetitive sequences at a throughput comparable to the current short read technologies. The Single Molecule Real Time (SMRT) sequencing technology developed by Pacific Biosciences (<http://www.pacificbiosciences.com>) may offer such opportunities and pilot projects are underway to determine the extent to which this can help to achieve the reference sequence of the bread wheat genome. Finally, because of the difficulty associated with ordering the physical and sequence contigs in wheat, additional resources may be needed to reach the gold standard assembly. Optical mapping has proven very useful for validating the rice genome assembly (Zhou et al. 2007) and greatly facilitated the assembly of the maize genome (Schnable et al. 2009; Wei et al. 2009). Preliminary data indicate the suitability of DNA prepared from flow-sorted chromosomes for optical mapping (Šimková et al., unpublished) providing a potential new resource for ordering and orientating the wheat sequence scaffolds into pseudomolecules.

17.3.3 Whole Genome Approaches can Support the Achievement of a Reference Wheat Genome Sequence

While waiting for the MTP chromosome-based reference genome sequence, whole genome shotgun approaches can be useful to accelerate marker development and estimate roughly the gene content. Such WGS sequences were produced recently for

Table 17.2 Current efforts in obtaining wheat genome sequences

Species	Ploidy level	Physical map	Survey sequence	Reference sequence
<i>T. urartu</i> (A genome progenitor)	2x	Planned	Yes	No
<i>Ae. speltoides</i> (B genome related progenitor)	2x	No	Yes	No
<i>Ae. tauschii</i> (D genome progenitor)	2x	Yes	Yes	Planned
<i>T. aestivum</i>	6x			
Chromosome based ^a		Yes	Yes	Yes
Whole Genome		No	Yes	No

^asee Fig. 17.5 for a detailed representation of the efforts on each of the 21 bread wheat chromosomes

the A and D genomes of the diploid related wild species *T. urartu* and *Ae. tauschii* (Table 17.2; Jia et al. 2013; Ling et al. 2013; Brenchley et al. 2012) as well as for the hexaploid wheat cv. Chinese Spring. In the latter instance, about 200 million Roche-454 reads were produced to achieve 5x coverage of the whole genome. This sample represents the first dataset of reads homogeneously covering the whole hexaploid genome.

17.4 Integration of Wheat Sequence Information in Databases

17.4.1 Data Integration

After production, the different types of sequence data are deposited into databases. Using the model of “three-tiers of database curation” (Parkhill et al. 2010), three database categories can be distinguished: (1) non-integrated, production databases that provide access to the raw sequence data in static repositories; (2) intra-species integrated databases enabling linking genome sequences with other data (e.g. genetic and physical maps, phenotypes, markers, proteomes, etc) from the same species; and (3) inter-species integrated databases that permit viewing genome sequences in relation to data from other species. Integrated databases enable the most efficient exploitation of genomic data in biological studies. To achieve integration of data from different production databases and provide users with a unified view of these data (Lenzerini 2002), two architectures can be used: data warehouses and virtual databases (also known as federated database systems). In a data warehouse, information is offloaded from one or several production databases, aggregated, and loaded into a single database. The raw data are then cleaned, transformed, catalogued, and made available for navigation and data mining. This architecture offers a high level of data consistency with data residing together in a single repository. In addition, it ensures referential integrity, i.e. no record can be deleted if it refers to another record. A federated database system is a meta-database management system that transparently integrates multiple dispersed database systems over a computer network. Through data abstraction, federated database systems can provide a uniform interface enabling

users to manipulate data in several dispersed databases with a single operation. To this end, the system is able to decompose the query into subqueries for submission to the individual databases. Subsequently, the system merges the result sets of the subqueries into a single set. While not offering the same level of consistency as data warehouses, federated databases are easier to maintain and update.

17.4.2 *Wheat Databases*

17.4.2.1 **Non-Integrated Wheat Sequence Databases**

Wheat data are found in different non-integrated sequence databases such as the nucleotide databases of Genbank (<http://www.ncbi.nlm.nih.gov/genbank/>) and the Sequence Read Archive (SRA; <http://trace.ncbi.nlm.nih.gov/Traces/sra/>) at NCBI, the European Nucleotide Archive (ENA) at EBI (<http://www.ebi.ac.uk/ena/home>) and the DNA Databank of Japan (DDBJ; <http://www.ddbj.nig.ac.jp/>). The three organizations exchange data on a daily basis. Some linked information is available in these databases but it is very limited because it is often restricted to metadata (e.g. authors, publication reference, species taxonomy, submission date) and occasionally contains limited annotation (e.g. genes, simple repeats). Genbank provides an annotated collection of all publicly available DNA sequences while the SRA is a repository dedicated to sequence reads delivered by second generation sequencing platforms. With regard to wheat, Genbank stores all of the EST and STS sequences and SRA stores the 5x coverage survey sequences of the cv. Chinese Spring genome from CerealsDB (<http://cerealsdb.uk.net>) Table 17.3. CerealsDB also provides BLAST search facilities against an assembly of these reads and gives access to wheat SNP, EST and DArT markers. In the fall of 2011, there were 107,3845 and 1775 *Triticum aestivum* ESTs and STSs, respectively, and approximately 85 Gb of low coverage survey sequence reads in these databases. Moreover, a wheat Gene Index comprising 93,508 ESTs is available for BLAST and downloading at the Dana-Farber Cancer Institute web site (<http://compbio.dfci.harvard.edu>). In addition, several clusterings of wheat transcript sequences were performed and are available at the Unigene web site of the NCBI (<http://www.ncbi.nlm.nih.gov/unigene>) and at the portal of TIGR Plant Transcript Assemblies (<http://plantta.jcvi.org>).

17.4.2.2 **Integrated Databases for Wheat**

Wheat specific integrated databases are those dedicated to wheat “omic” and genetic data. These databases integrate the wheat sequences with other wheat data using database-constrained data links or cross-references. The level of integration varies between the databases according to the quality, nature, and quantity of the linked information.

Table 17.3 Wheat sequence databases ordered by categories and sequence types

Category	DB name	Transcript seq	Physical maps	Survey seq	Ref seq (underway)
Non-integrated	NCBI Genbank and SRA	X		X ^a	
	TIGR Plant Transcript	X			
	DFCI Gene Index	X			
	CerealsDB	X		X ^a	
Intra-species chromosome specific	WheatDB		X		
	WGGRC	X	X		
Intra-species genome-wide	Wheatgenome.info			X ^b	
	GrainGenes	X	X		
Inter-species	URGI	X	X	X ^c	X ^e
	TriFLDB	X			
Inter-species	Komugi	X			
	PlantsDB			X ^d	
	Gramene	X	X		

^alow coverage

^bhigh coverage group 7 chromosomes

^chigh coverage genome-wide

^dGenomeZipper using high coverage genome-wide

^e3B reference chromosome sequence

Chromosome-Specific Databases

With the progress in physical mapping and sequencing individual chromosomes of wheat, specific databases devoted to sets of particular chromosomes have been established Table 17.3.

The INRA URGI hosts data for the bread wheat chromosome 3B (<http://urgi.versailles.inra.fr/Species/Wheat/Data>) and for chromosome arms 3DS, 3DL, 1BS, 1BL, 1AS and 1AL as part of the database established for the European project TriticeaeGenome (<http://urgi.versailles.inra.fr/Projects/TriticeaeGenome>). Physical maps, genetic neighbor maps with links to genetic markers and QTLs as well as BAC contig sequences are available. This will be completed soon with the chromosome 3B reference sequence that is currently under analysis in the framework of the 3BSEQ project (<http://urgi.versailles.inra.fr/Projects/3BSeq>). The Wheat Genome Database at Kansas State University, WGGRC, (<http://wggrc.plantpath.ksu.edu/default.html>) provides a GBrowse access to the physical map of chromosome 3A with links to the genetic markers and BLAST facilities. It also hosts information about the physical maps of chromosomes 1D, 4D, and 6D of Chinese Spring that were developed as part of the NSF funded project on the physical mapping of the wheat D-genome (see below). WheatGenome.info (<http://www.wheatgenome.info/>) which is developed by the Australian Centre for Plant Functional Genomics and the University of Queensland provides access to the draft survey sequence reads and assemblies of bread wheat chromosomes 7A, 7B, and 7D and to a BLAST portal for these sequences.

Genome Wide Databases

GrainGenes (<http://wheat.pw.usda.gov/>) was built to provide a suite of services for the *Triticeae* and oat communities, including databases, documents, tools, data files, announcements, curation, and community assistance Table 17.3. To date, GrainGenes stores 76 wheat genetic maps, more than 100,000 genetic markers, and approximately 271,000 wheat ESTs. These sequences can be searched through a BLAST server or by using queries to get additional information on genetic mapping data. GrainGenes also hosts the Triticeae Repeat (TREP) databank that comprises 1717 sequences of wheat transposable elements. As GrainGenes is dedicated to Triticeae and Avenae species, it presents some features of inter-species integrated database, including a GBrowse display of wheat, barley, and oat EST sequences mapped on the rice genome.

A specific website (<http://avena.pw.usda.gov/wheatD/>) is dedicated to the US National Science Foundation funded physical mapping project of *Ae. tauschii*, the bread wheat D genome progenitor. It provides links to physical mapping data and enables BLAST searches against survey sequences of *Ae. tauschii*.

The INRA URGI Wheat database (<http://urgi.versailles.inra.fr/Species/Wheat>) stores 26 wheat genetic maps, 19,029 markers, 324 QTLs, 10,819 SNPs and 544,529 ESTs Table 17.3. A GBrowse is available to display physical maps in relation with other datasets (e.g., genetic markers, reference sequences, QTLs and SNPs). Physical maps of cv. Chinese Spring chromosomes that are constructed under the framework of the IWGSC are being integrated regularly into the database (<http://www.wheatgenome.org/Projects/IWGSC-Bread-Wheat-Projects/Physical-mapping/>). To date, physical maps of chromosomes 1BL, 1AS, 3B, and 3DS have been integrated and a link to the 3A physical map at WGGRC is provided. In addition, URGI hosts the IWGSC sequence repository that provides access to the survey sequence assemblies of the 21 chromosomes of Chinese Spring (<http://urgi.versailles.inra.fr/Species/Wheat/Sequence-Repository>).

17.4.2.3 Inter-Species Sequence Integrated Databases

Inter-species integrated databases are designed to compare wheat sequences with related species sequences. The first way of performing inter-species sequence integration is through sequence comparisons that display the results as a repository with flat files to download or BLAST. For example, TriFLDB (<http://trifldb.psc.riken.jp>) provides access for download and BLAST against 8530 and 7341 full-length cDNAs of wheat and barley, respectively. These full-length cDNAs are further integrated into genome browsers of rice and sorghum. The second method for inter-species integration is to display the results in textual or graphical web pages that enable adding a limited level of intra-species integration when needed. The Komugi database (<http://www.shigen.nig.ac.jp>) has established a BLAST server on wheat ESTs and full length cDNAs, as well as a comparative map tool to display homologies between wheat and barley genic sequences along the rice chromosomes (<http://earth.lab.nig.ac.jp>).

PlantsDB developed by the Munich Information Centre for Protein Sequences (<http://mips.helmholtz-muenchen.de/plant/index.jsp>) has a tool called “GenomeZipper” that derives a putative chromosomal gene order for one species on the basis of its syntenic relationship with related species. Each syntenic gene presents additional information about genetic and physical maps and genomic annotation using a browser. Additional information that links the syntenic genes to the genetic and physical maps and to their annotation is available through a browser. A barley GenomeZipper is available already and a wheat GenomeZipper is underway in the framework of the IWGSC Survey Sequencing Initiative.

Gramene (<http://www.gramene.org/>) stores a number of wheat data (e.g. markers, genes) with their alignments to other cereal crops. In collaboration with EBI, efforts are underway to develop EnsemblPlants (<http://plants.ensembl.org>) which will provide access to individual genome data as well as comparative tools, such as Plant Compara. To date, EnsemblPlants does not contain any wheat sequences, but future genome sequences will likely be incorporated as they are completed.

Acknowledgments The authors want to thank Hadi Quesneville, Daphné Verdelet, Kirsley Chenen for their feedback on wheat databases. H.Š., J.S. and J.D. are supported by the Ministry of Education, Youth and Sports of the Czech Republic, the European Regional Development Fund (Operational Programme Research and Development for Innovations No. ED0007/01/01) and by the Czech Science Foundation (award no. P501/10/1740). F. C., P. L. and C. F. are supported by the European Community’s Seventh Framework Programme TriticeaeGenome (grant agreement n°FP7–212019), the Agence Nationale de la Recherche grant ANR(09-GENM-025), FranceA-griMer (201006-015-104) and the competitiveness cluster “Céréales Vallée” (http://www.cereales-vallee.org/default_gb.cfm).

References

- AGI (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408:796–815
- Akhunov ED, Akhunova AR, Dvorak J (2007) Mechanisms and rates of birth and death of dispersed duplicated genes during the evolution of a multigene family in diploid and tetraploid wheats. *Mol Biol Evol* 24:539–550
- Amano N, Tanaka T, Numa H et al (2010) Efficient plant gene identification based on interspecies mapping of full-length cDNAs. *DNA Res* 17:271–279
- Astier Y, Braha O, Bayley H (2006) Toward single molecule DNA sequencing: Direct identification of ribonucleoside and deoxyribonucleoside 5’-monophosphates by using an engineered protein nanopore equipped with a molecular adapter. *J Am Chem Soc* 128:1705–1710
- Bennett ST, Barnes C, Cox A et al (2005) Toward the \$1000 human genome. *Pharmacogenomics* 6:373–382
- Berkman P, Skarshewski A, Manoli S et al (2011a) Sequencing wheat chromosome arm 7BS delimits the 7BS/4AL translocation and reveals homoeologous gene conservation. *Theor Appl Genet*: 1–10
- Berkman PJ, Skarshewski A, Lorenc MT et al (2011b) Sequencing and assembly of low copy and genic regions of isolated *Triticum aestivum* chromosome arm 7DS. *Plant Biotech J* 9:768–775
- Birney E, Stamatoyannopoulos JA, Dutta A et al (2007) Identification and analysis of functional elements in 1 % of the human genome by the ENCODE pilot project. *Nature* 447:799–816

- Brenchley R, Spannagl M, Pfeifer M, Barker GL, 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 MC, Sehgal S, Gill B, Kianian S, Anderson O, Kersey P, Dvorak J, McCombie WR, Hall A, Mayer KF, Edwards KJ, Bevan MW, Hall N (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* 491:705–710
- Brisson N, Gate P, Gouache D et al (2010) Why are wheat yields stagnating in Europe? A comprehensive data analysis for France. *Field Crop Res* 119:201–212
- Cantarel BL, Korf I, Robb SMC et al (2008) MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res* 18:188–196
- Chaisson M, Pevzner P, Tang H (2004) Fragment assembly with short reads. *Bioinformatics* 20:2067–2074
- Chantret N, Cenci A, Sabot F, Anderson O, Dubcovsky J (2004) Sequencing of the *Triticum monococcum* hardness locus reveals good microcolinearity with rice. *Mol Genet Genomics* 271:377–386
- Charles M, Belcram H, Just J et al (2008) Dynamics and differential proliferation of transposable elements during the evolution of the B and A genomes of wheat. *Genetics* 180:1071–1086
- Choulet F, Wicker T, Rustenholz C et al (2010) Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. *Plant Cell* 22:1686–1701
- Curwen V, Eyras E, Andrews TD et al (2004) The Ensembl automatic gene annotation system. *Genome Res* 14:942–950
- Devos KM, Ma J, Pontaroli AC et al (2005) Analysis and mapping of randomly chosen bacterial artificial chromosome clones from hexaploid bread wheat. *Proc Natl Acad Sci U S A* 102:19243–19248
- Dolezel J, Kubalaková M, Bartos J, Macas J (2004) Flow cytogenetics and plant genome mapping. *Chromosome Res* 12:77–91
- Dolezel J, Kubalaková M, Paux E et al (2007) Chromosome-based genomics in the cereals. *Chromosome Res* 15:51–66
- Dubcovsky J, Dvorak J (2007) Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science* 316:1862–1866
- Earl D, Bradnam K, JJ St (2011) Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome Res* 21:2224–2241
- Eid J, Fehr A, Gray J et al (2009) Real-time DNA sequencing from single polymerase molecules. *Science* 323:133–138
- Endo TR, Gill BS (1996) The deletion stocks of common wheat. *J Hered* 87:295–307
- Erayman M, Sandhu D, Sidhu D et al (2004) Demarcating the gene-rich regions of the wheat genome. *Nucl Acids Res* 32:3546–3565
- Estill JC, Bennetzen JL (2009) The DAWGPAWS pipeline for the annotation of genes and transposable elements in plant genomes. *Plant Methods* 5:8
- 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, Eversole K (2007) Physical mapping of the wheat genome: A coordinated effort to lay the foundation for genome sequencing and develop tools for breeders. *Isr J Plant Sci* 55:307–313
- Feuillet C, Keller B (1999) High gene density is conserved at syntenic loci of small and large grass genomes. *Proc Natl Acad Sci. USA* 96:8265–8270
- Feuillet C, Salse J (2009) Comparative Genomics in the Triticeae. In: Feuillet C, Muehlbauer GJ (eds) *Plant Genetics and Genomics*. Springer, New York, pp 451–477
- Feuillet C, Leach JE, Rogers J et al (2011) Crop genome sequencing: lessons and rationales. *Trends Plant Sci* 16:77–88
- Flavell RB, Rimpau J, Smith DB (1977) Repeated sequence DNA relationship in four cereal genomes. *Chromosoma* 63:205–222
- Foley JA, Ramankutty N, Brauman KA et al (2011) Solutions for a cultivated planet. *Nature* 478:337–342

- Gill KS, Gill BS, Endo TR, Boyko EV (1996a) Identification and high-density mapping of gene-rich regions in chromosome group 5 of wheat. *Genetics* 143:1001–1012
- Gill KS, Gill BS, Endo TR, Taylor T (1996b) Identification and high-density mapping of gene-rich regions in chromosome group 1 of wheat. *Genetics* 144:1883–1891
- 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
- Gnerre S, MacCallum I, Przybylski D et al (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci. USA* 108:1513–1518
- Havlak P, Chen R, Durbin KJ et al (2004) The Atlas genome assembly system. *Genome Res* 14:721–732
- Hernandez P, Martis M, Dorado G et al (2011) Next-generation sequencing and syntenic integration of flow-sorted arms of wheat chromosome 4A exposes the chromosome structure and gene content. *The Plant J*: 69:377–386
- International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. *Nature* 436:793–800
- IRGSP (2005) The map-based sequence of the rice genome. *Nature* 436:793–800
- Jaffe DB, Butler J, Gnerre S et al (2003) Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res* 13:91–96
- Janda J, Bartoš J, Šafář J et al (2004) Construction of a subgenomic BAC library specific for chromosomes 1D, 4D and 6D of hexaploid wheat. *Theor Appl Genet* 109:1337–1345
- 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, Wang J; International Wheat Genome Sequencing Consortium, Yang H, Liu X, He Z, Mao L, Wang J (2013) *Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature* 496:91–95
- Kubaláková M, Vrána J, Čížalíková J et al (2002) Flow karyotyping and chromosome sorting in bread wheat (*Triticum aestivum* L.). *Theor Appl Genet* 104:1362–1372
- Kubaláková M, Kovářová P, Suchánková P et al (2005) Chromosome sorting in tetraploid wheat and its potential for genome analysis. *Genetics* 170:823–829
- La Rota M, Sorrells ME (2004) Comparative DNA sequence analysis of mapped wheat ESTs reveals the complexity of genome relationships between rice and wheat. *Funct Integr Genomics* 4:34–46
- Lamoureux D, Peterson DG, Li W et al (2005) The efficacy of Cot-based gene enrichment in wheat (*Triticum aestivum* L.). *Genome* 48:1120–1126
- Lander ES, Linton LM, Birren B et al (2001) Initial sequencing and analysis of the human genome. *Nature* 409:860–921
- Lenzerini M (2002) Data integration: a theoretical perspective. Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. ACM, Wisconsin, Madison, pp 233–246
- Leroy P, Guillhot N, Sakai H et al (2012) TriAnnot: a versatile and high performance pipeline for the automated annotation of plant genomes. *Frontiers in Plant Sciences* 3:1–14
- Li W, Gill B (2004) Genomics for cereal improvement. In: Gupta PK, Varshney RK (eds) *Cereal genomics*. Kluwer Academic Publishers, Dordrecht, pp 585–634
- Li W, Zhang P, Fellers JP et al (2004) Sequence composition, organization, and evolution of the core Triticeae genome. *Plant J* 40:500–511
- Li R, Yu C, Li Y et al (2009) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* 25:1966–1967
- Liang C, Mao L, Ware D, Stein L (2009) Evidence-based gene predictions in plant genomes. *Genome Res* 19:1912–1923
- 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, Wu H, Tang D, Shen Q, Xue P, Zou S, Wang X, Liu X, Wang F, Yang Y, An X, Dong Z, Zhang K, Zhang X, Luo MC, Dvorak J, Tong Y, Wang J, Yang H, Li Z, Wang D, Zhang A, Wang J (2013) Draft genome of the wheat A-genome progenitor *Triticum urartu*. *Nature* 496:87–90

- Lobell DB, Schlenker W, Costa-Roberts J (2011) Climate Trends and Global Crop Production Since 1980. *Science* DOI:10.1126/science.1204531
- Margulies M, Egholm M, Altman WE et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376–380
- Massa AN, Wanjugi H, Deal KR et al (2011) Gene Space Dynamics During the Evolution of *Aegilops tauschii*, *Brachypodium distachyon*, *Oryza sativa*, and *Sorghum bicolor* Genomes. *Mol Biol Evol* 28:2537–2547
- Mayer KF, Taudien S, Martis M et al (2009) Gene content and virtual gene order of barley chromosome 1H. *Plant Physiol* 151:496–505
- Mayer KF, Martis M, Hedley PE et al (2011) Unlocking the barley genome by chromosomal and comparative genomics. *Plant Cell* 23:1249–1263
- McFadden E, Sears E (1946) The origin of *Triticum spelta* and its free-threshing hexaploid relatives. *J Hered* 37:81–89107
- Metzker ML (2009) Sequencing technologies – the next generation. *Nat Rev Genet* 11:31–46
- Muniz LM, Cuadrado A, Jouve N, Gonzalez JM (2001) The detection, cloning, and characterisation of WIS 2–1A retrotransposon-like sequences in *Triticum aestivum* L. and *xTriticosecale* Wittmack and an examination of their evolution in related Triticeae. *Genome* 44:979–989
- Ouyang S, Buell CR (2004) The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. *Nucleic Acids Res* 32:D360–363
- Parkhill J, Birney E, Kersey P (2010) Genomic information infrastructure after the deluge. *Genome Biol* 11:402
- Paux E, Roger D, Badaeva E et al (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 et al (2008) A physical map of the 1-gigabase bread wheat chromosome 3B. *Science* 322:101–104
- Paux E, Sourdille P, Mackay I, Feuillet C (2011) Sequence-based marker development in wheat: Advances and applications to breeding. *Biotechnol Adv* 30:1071–1088
- Pevzner PA, Tang H, Waterman MS (2001) An Eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci U S A* 98:9748–9753
- Qi LL, Echalié 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
- Rabinowicz PD, Citek R, Budiman MA et al (2005) Differential methylation of genes and repeats in land plants. *Genome Res* 15:1431–1440
- Rounsley S, Marri P, Yu Y et al (2009) De Novo Next Generation Sequencing of Plant Genomes. *Rice* 2:35–43
- Rustenholtz C, Choulet F, Laugier C et al (2011) A 3000-loci transcription map of chromosome 3B unravels the structural and functional features of gene islands in hexaploid wheat. *Plant Physiol* 157:1596–1608
- Sabot F, Guyot R, Wicker T et al (2005) Updating of transposable element annotations from large wheat genomic sequences reveals diverse activities and gene associations. *Mol Genet Genomics* 274:119–130
- Šafař J, Bartoš J, Janda 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
- Šafař J, Šimková H, Kubaláková M et al (2010) Development of chromosome-specific BAC resources for genomics of bread wheat. *Cytogenet Genome Res* 129:211–223
- Sakata K, Nagamura Y, Numa H et al (2002) RiceGAAS: an automated annotation system and database for rice genome sequence. *Nucleic Acids Res* 30:98–102
- Sandhu D, Gill KS (2002) Gene-Containing Regions of Wheat and the Other Grass Genomes. *Plant Physiol* 128:803–811
- Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci of the United States of America* 74:5463–5467

- Schnable PS, Ware D, Fulton RS et al (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science* 326:1112–1115
- Sears ER (1954) The aneuploid of common wheat. *Mo Agr Exp Sta Res Bull* 572:1–58
- Sears ER, Sears L (1978) The telocentric chromosomes of common wheat In: Ramanujams S (ed) *Proc 5th Int Wheat Genetics Symp.* Indian Agricultural Research Institute, New Delhi, India., pp 389–407
- Simkova H, Svensson JT, Condamine P et al (2008) Coupling amplified DNA from flow-sorted chromosomes to high-density SNP mapping in barley. *BMC Genomics* 9:294
- Simpson JT, Durbin R (2010) Efficient construction of an assembly string graph using the FM-index. *Bioinformatics* 26:i367–i373
- Simpson JT, Wong K, Jackman SD et al (2009) ABySS: a parallel assembler for short read sequence data. *Genome Res* 19:1117–1123
- Smith DB, Flavell RB (1975) Characterization of Wheat Genome by Renaturation Kinetics. *Chromosoma* 50:223–242
- Sorrells ME, La Rota M, Bermudez-Kandianis CE et al (2003) Comparative DNA sequence analysis of wheat and rice genomes. *Genome Res* 13:1818–1827
- The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408:796–814
- Valouev A, Ichikawa J, Tonthat T et al (2008) A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res* 18:1051–1063
- Venter JC, Adams MD, Myers EW et al (2001) The Sequence of the Human Genome. *Science* 291:1304–1351
- Titolo N, Albiero A, Forcato C et al (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 et al (2000) Flow sorting of mitotic chromosomes in common wheat (*Triticum aestivum* L.). *Genetics* 156:2033–2041
- Waterston RH, Lindblad-Toh K, Birney E et al (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420:520–562
- Wei F, Zhang J, Zhou S et al (2009) The physical and genetic framework of the maize B73 genome. *PLoS Genet* 5:e1000715
- Wicker T, Stein N, Albar L et al (2001) Analysis of a contiguous 211 kb sequence in diploid wheat (*Triticum monococcum* L.) reveals multiple mechanisms of genome evolution. *Plant J* 26:307–316
- Wicker T, Matthews DE, Keller B (2002) TREP: a database for Triticeae repetitive elements. *Trends Plant Sci* 7:561–562
- Wicker T, Mayer KFX, Gundlach H et al (2011) Frequent gene movement and pseudogene evolution is common to the large and complex genomes of wheat, barley, and their relatives. *Plant Cell* 23:1706–18
- Yan L, Loukoianov A, Tranquilli G et al (2003) Positional cloning of the wheat vernalization gene VRN1. *Proc Natl Acad Sci USA* 100:6263–6268
- Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 18:821–829
- Zhou S, Bechner M, Place M et al (2007) Validation of rice genome sequence by optical mapping. *BMC Genomics* 8:278
- Zohary D, Hopf M (2000) *Domestication of plants in the old world*, 3rd edn. Oxford University Press, Oxford