A Review of Genome Sequencing in the Largest Cereal Genome, *Triticum aestivum* L.

Anandi Bierman, Anna-Maria Botha

Department of Genetics, Stellenbosch University, Stellenbosch, South Africa
Email: anandie@sun.ac.za

**Abstract**

Sequencing whole plant genomes has advanced rapidly with the development of next generation sequencing (NGS) technologies and bioinformatics, enabling the study of large and complex genomes such as that of the hexaploid cereal, *Triticum aestivum* L. (bread wheat). Despite advances, however, confounding factors such as repetitive elements and low polymorphism still hinder sequencing attempts. Isolation techniques such as sequencing of diploid progenitors and chromosome separation through flow cytometry have showed promise in reducing the size of the genome for sequencing. In this review we discuss the advances and stumbling blocks that have been encountered on the road toward the complete hexaploid wheat genome sequence. We also discuss the latest complimentary techniques and the progression of accumulation of sequence data relevant to wheat genome research.

**Keywords**

Hexaploid, Next Generation Sequencing, Flow Cytometry, Wheat, Genome

**1. Introduction**

The wheat genome is one of the largest (17,000 Mbp) and most repeat-rich of all crop plants (76.6% repetitive elements in assembled sequences; [1]). Wheat is an allohexaploid (*Triticum aestivum* L.; 2n = 6x = 42) of which the genome is divided into three sub-genomes (AA, BB and DD). These sub-genomes act as independent diploid genomes during meiosis due to the action of the *Ph* genes on chromosome 5B [2] [3]. The wheat genome was formed by two polyploidization events. The first polyploidization event formed *Triticum turgidum* (2n = 4x = 28; AABB), an allotetraploid which resulted from *Triticum urartu* (2n = 2x = 14; AA) and a species which is believed to be *Aegilops speltoides* (2n = 14; SS), from the *Sitopsis* section of *Triticum* [4]. The second hybridization event, which resulted in the ancestral *Triticum aestivum* (2n = 6x = 42; AABBDD), occurred...
between *Triticum turgidum* and the diploid grass, *Aegilops tauschii* (DD) [4][5]. These progenitor species often provide valuable genetic resources by serving as diploid model organisms for the study of wheat [6]. The sub-genomes in modern hexaploid wheat differ significantly from one another and many genes are not present in triplicate, but are chromosome specific [7]. Genes are also not distributed evenly, but are clustered in gene-rich regions, particularly at the distal regions of chromosomes [8].

The availability of the wheat reference genome sequence will play an absolutely important role in employing molecular tools within breeding programs. By enabling the identification, more efficiently, of genes and markers related to agronomically important traits, breeding schemes can be accelerated to employ and deliver the best stock.

2. First Steps toward Sequencing the Wheat Genome

2.1. Initial Focus on Markers and Genes

Early endeavors toward obtaining wheat sequence information focused on the coding regions of the genome such as cDNA and ESTs. Resources such as 17,000 cDNA sequences, 40,000 UniGenes, 1 million ESTs [9], of which 7000 have been assigned to chromosome-specific bins [10] and great numbers of SNP markers [11] [12] [13] were compiled before a draft genome sequence was available. These initial tools have allowed the development of marker sets and microarrays used in gene expression studies [14] [15] as well as the first views of genome localization and low resolution mapping of phenotypic traits.

Up until 2004, most completed genome sequences had been obtained through a clone-by-clone approach [16]-[22]. Though the data provided was comprehensive, it was an expensive approach, especially if considered for a large genome such as that of wheat. In turn, other methods also had their limitations as whole genome shotgun sequencing was hindered by computational power, especially for genomes with large numbers of repeats and genomic filtration techniques such as methylation filtration [23] and *Cot* selection [24] [25] were not well tested enough to guarantee high percentages of gene identification [7]. Ultimately, it was decided to sequence individual chromosomes of wheat, separated by flow cytometry [26] [27] and used to construct Bacterial Artificial Chromosome (BAC) libraries and filtration libraries.

2.2. From BAC Libraries to Isolated Wheat Chromosomes

Construction of BAC libraries from progenitor species provided the first steps toward a reduction of the wheat genome by focusing only on one sub genome within a BAC library. In 1999, Lijavetzky *et al.* [28] constructed a BAC library of the A genome using *T. monococcum* as template. This provided a 5.6 times coverage of the A genome and significantly increased the probability of finding a sequence of interest (>99.6%). However, many of the BAC ends produced included repetitive sequences, and in order to circumnavigate this, a low copy sequence selection step had to be introduced during the end-isolation procedure.
which subsequently enabled chromosome walking to obtain fragments of wheat sequence information.

The isolation of individual wheat chromosomes further increased the discriminatory power of BAC libraries in wheat. Kubaláková et al. (2002) [29] first demonstrated that wheat chromosomes could be flow sorted intact, though due to the lack of size differences between chromosomes, initially only 3B could be isolated with high purity. Estimated at 995 Mbp [1] [30], chromosome 3B is twice as large as the entire rice genome (370 Mbp; [31]), which made it an attractive candidate for initial attempts at physical mapping and sequencing [26]. After flow sorting ±1.8 million 3B chromosomes, a BAC library could be constructed representing 6.2 equivalents of the 3B chromosome [27]. As flow cytometric techniques improved, smaller chromosomes could be separated using telosomic wheat lines and lines carrying isochromosomes [29]. Aneuploid or ditelosomic wheat lines or wheat-rye addition lines [32] also provided select chromosomes with sizes differentiable from the remaining chromosomal complement.

### 2.3. Sequencing the Wheat Progenitors

In 2013, Jia et al. [33] published a draft genome sequence of the diploid D genome progenitor, *Ae. tauschii*, and obtained an assembly representative of 97% of the 4.36 Gbp genome using the Roche-454 sequencing platform. The estimated 65.9% repetitive elements in this genome are lower than that of hexaploid bread wheat, while it is estimated to contain 34,498 Protein Coding Genes (PCGs) which does correspond, roughly, to a third of the total number of genes in hexaploid bread wheat [1]. A draft genome assembly of *Triticum urartu* [34] was also produced through whole genome shotgun approaches.

During discussions at the International Wheat Genome Sequencing Consortium (IWGSC) meeting in 2003 though, many positive outcomes were listed for starting to sequence the hexaploid wheat genome in its entirety. Modern wheat has diverged substantially from its progenitors and its polyploid nature could reveal much about its evolution as well as polyploid speciation [7], which is seen in many agriculturally significant plant species. The final argument was that agriculturally important genes might be chromosome specific and possibly overlooked if only one of the progenitors were selected for sequencing [7].

As the sequencing of isolated wheat chromosomes and progenitor genomes became a reality, the first IWGSC was assembled and a workshop was held in Crystal City, Washington, DC in November 2003. This workshop would set into motion the unraveling of the genome of hexaploid wheat. The aim of the workshop was to discuss the status and future of wheat genomics with particular focus on the possibility of sequencing the complete wheat genome. At the meeting, the cultivar, “Chinese Spring” was selected as universal choice for sequencing of the wheat genome due to the availability of genetic stocks of this cultivar [7].

### 3. 2012 Draft Genome Sequence

Nine years after the initial conceptualization of a genome sequence for hexaploid
wheat, Brenchley et al. (2012) [35] published the first draft sequence obtained through whole genome shotgun sequencing. The authors obtained lower coverage (5 fold) with longer read lengths on the Roche 454 sequencing platform. Using diploid wheat relatives and progenitors the authors could classify homoeologous relationships and build orthologous gene family frameworks. Ultimately, a catalog of 132,000 SNPs was created in the various sub-genomes of wheat. Repeat elements were identified to make up 79% of the sequenced genome and the number of genes was estimated at 94,000 to 96,000. A representation of nearly all wheat genes could be identified as constructed orthologous groups matched 90% of metabolic genes in Arabidopsis and 92% of publicly available wheat full-length cDNA. Overall, a trend toward gene family size reduction was observed in hexaploid wheat despite its recent evolution as a hexaploid. Expanded gene families’ overrepresented categories included proteins involved in energy metabolism, defense, growth and nutrition. The authors also found gene families to be classified as A, B or D-genome derived, which indicates that transcriptional regulatory networks are maintained in a genome-specific manner in wheat.

4. 2014 IWGSC Genome

In 2014 the IWGSC published their paper on the wheat genome sequence [1]—an initiative that was started more than 10 years prior. The hexaploid nature of the wheat genome allows for a buffering effect in aneuploid wheat lines and viable mono-, tri and tetrasomic cytogenetic stocks as well as nullisomics have been developed in addition to aneuploids of every chromosome arm [36] [37]. Mayer et al. (2014) [1] used double ditelosomic wheat lines [37] of the cultivar, “Chinese Spring” to isolate individual chromosomes for sequencing on the Illumina platform. By sequencing the chromosomes individually, the sequencing consortium reduced much of the complexity of the genome and was able to differentiate genes which showed multiple copies as well as conserved homologs. Sequencing depth obtained ranged between 30x and 241x [1].

In their chromosome-based sequence data of the wheat genome, Mayer et al., (2014) [1] found 81% of raw sequencing reads and 76.6% of assembled sequences to contain repeats. Retro-elements were most abundant in the A sub-genome chromosomes (A > B > D in order of abundance) while transposons were most abundant in the D sub-genome (D > B > A in order of abundance). Of 270 miRNA molecules identified, 49 had not been reported previously and correspondence to 98,068 predicted miRNA-coding loci was identified. At least one target gene (putative) could be identified for 87% of these miRNA-coding loci. Gene annotation of the chromosome-based wheat genome sequence relied on comparisons to genes in related grass species such as Brachypodium distachyon [38], Oryza sativa [39], Sorghum bicolor [40], and Hordeum vulgare [41] in addition to full length wheat cDNAs [9] and RNA-seq data from “Chinese Spring” tissue at different stages of development. Mayer et al. (2014) [1] estimated a count of 106,000 protein coding genes in wheat, which is a slightly greater figure than that estimated by Brenchley et al. (2012) [35] though consistent
with estimates for the diploid sub-genomes, ranging between 32,000 and 38,000 as well as estimates for the diploid progenitors [33] [34] [35]. The distribution of genes across the sub-genomes was highest in the B sub-genome (35%) followed by A (33%) and D (32%) though this distribution pattern was not reflected at the level of the homeologous groups of chromosomes. Variation of up to 2.4 fold was observed among the different chromosome arms. This observation is also seen in rye [42].

Lineage specific intra-chromosomal duplication was determined for genes on each wheat chromosome. On average for all chromosomes, 23.6% of genes are duplicated, though Mayer et al. (2014) [1] state that this is likely an underestimation. Comparisons with intra-chromosomal duplicates in other species such as rice, sorghum, barley, maize and millet, show the gene duplications in wheat to be significantly higher (17% - 20%); [1] [38] [39] [40] [43] [44]. In comparing gene family sizes of hexaploid wheat with Ae. tauschii and T. urartu [33] [34], the authors found that genes belonging to expanded families were mainly affected by gene loss as was also reported by Brenchley et al. (2012) [35]. On the other hand, genes without paralogous copies, i.e. singletons, were not subject to gene loss and the retention rate of genes was very similar across the three sub-genomes of hexaploid wheat. The authors found no evidence for a gradual gene loss after polyploidization although the D genome, the most recent addition to hexaploid wheat, did show the lowest levels of gene loss of the three sub-genomes. In addition to their similarity in gene distribution, the gene content of the A, B and D sub-genomes was very similar with only a small number of truly unique genes (1.3% to 1.7%).

Sequence conservation in chromosomes of hexaploid wheat showed highly conserved gene sequences between the sub-genomes and their diploid relatives. In addition, the wheat sub-genomes have high regulatory autonomy with little regulation between the sub-genomes [45]. The patterns of gene expression in hexaploid wheat contrasts to that observed in other species such as allopolyploid cotton [46] [47], mesopolyploid Brassica rapa [48], synthetic allotetraploid Arabidopsis [49], and paleopolyploid maize [50], which all have one sub-genome more transcriptionally active than the others. Knowledge still outstanding, according to Mayer et al. (2014) [1], concerns the distribution of genes and their positions along the chromosomes of hexaploid wheat as well as gene evolution during wheat’s development.

5. New Developments

BioNano mapping or optical mapping (genome mapping in nanochannel arrays) [51] is fast becoming a useful tool as it generates short sequence maps along DNA that stretches thousands of base pairs in length, thus allowing for novel techniques for mapping and assembling gigabase sized genomes. As the IWGSC wheat reference genome is produced by the sequencing of physical maps from individually isolated chromosomes, applying the BioNano mapping strategy here seems practical. Šimková et al. (2016) [52] discussed their strategy of generating
a BioNano high resolution map for chromosome 7DS. The authors generated 371 contigs with N50 of 1.3 Mbp. Chromosome 7DS sequence assemblies obtained through clone-by-clone sequencing were anchored to the 7DS BioNano map. This proved valuable in improving BAC-contig physical maps and validating sequence assembly [52].

The BioNano technology has been implemented on a genome-wide scale by Zhu et al. (2016) [53] in order to study genome structure. The authors discuss the construction of whole genome BioNano maps for both wheat and Ae. tauschii and the subsequent comparison in order to detect structural differences between the D sub-genome of wheat and the Ae. tauschii genome. Their comparison yielded a large amount of indels and through using the BioNano maps, novel information on these differences could be obtained. Indels that occurred during the evolution of wheat could be discerned, located exactly and their role in recombination defined [53].

At the Plant and Animal Genome Conference in San Diego, 2016, Zimin et al. [54] made the remark that Illumina sequencing technologies, though the most popular sequencing technology, produces short reads (up to 300 bp) which often further complicates the analysis of already complicated genomes due to the fragmentary nature of the data. Longer reads would be advantageous and PacBio has recently stepped up to that challenge [55]. Though expensive, PacBio produces reads in the order of tens of thousands of base pairs. The error rate of PacBio is still quite significant and therefore, these authors propose a hybrid strategy leveraging the accuracy of the shorter Illumina reads with the length of the PacBio reads to construct an assembly of the 4.5 Gbp Ae. tauschii genome [54]. This hybrid approach making use of Illumina short reads in combination with the longer reads offered by PacBio was also used to construct the completed genome sequence of T. urartu, for which the draft genome was published in 2013 [34] and the genome sequence completed and presented 3 years later at the Plant and Animal Genome conference in 2016 [56].

Clavijo et al. (2016) [57] stated that despite all the improvements to date, an accurate and nearly complete assembly of the wheat genome is still out of reach. They state that the first draft sequence available [35], based on orthologous group assemblies of related grass protein sequences was highly fragmented and though the subsequent, chromosome-based assembly (IWGSC chromosome survey sequence (CSS); [1]) was able to identify homeologous relationships between the sub-genomes, it too remained fragmented. These authors, therefore, provide a sequence assembly and annotation of “Chinese Spring” wheat with more in-depth analysis of its sequence and structure. With 33x sequence coverage of the genome, the scaffolds produced through their study, contain 13.4 Gbp of the total genome with an N50 value of 88.8 kbp. Their annotation contains 104,091 protein coding genes of high confidence as it is supported by transcriptome sequence data and full length cDNA sequence data. This new assembly, termed TGAC1, was classified into chromosome arms by alignment to raw reads from the IWGSC-CSS data [1] and represents nearly 80% of the 17 Gbp wheat
A genome which provides a 60% improvement in genome coverage. In confirmation with previous studies, more than 80% of the TGAC1 assembly comprised of transposable elements of which 70% were retro-elements and 13% DNA transposons [57].

Miss-assembly and gaps in the sequence data are recurring problems in the assembly of any genome whether by whole genome shotgun approaches [58] [59] [60] or clone-by-clone approaches [61] [62]. Applications providing longer reads have fast become very popular for solving some of the issues associated with genome assemblies and platforms such as PacBio [55] and BioNano [51] show promise to improve the quality of shotgun assemblies. In an attempt at the wheat reference genome, the IWGSC adopted a clone-by-clone sequencing approach based on physical maps constructed from BAC libraries specific to individual chromosome arms. In order to reconstruct the genome from BAC sequences, contigs from the physical map are anchored-a process which relies heavily on many genetic markers present on the contigs, and with known positions on the chromosomes. Though high-density linkage maps are available for wheat, their resolution is often limited due to small sizes of mapping populations and regions of low recombination along wheat chromosomes [63] [64]. To circumvent this particular problem, approaches independent of recombination, such as the BioNano genome mapping system, have proven highly valuable [65].

6. What Is Currently Available

The IWGSC (http://www.wheatgenome.org/About) have released several whole genome assemblies of both wheat and its progenitors. As of January 2014 the IWGSC released genome survey sequence assemblies [1], population sequencing (POPSEQ) data and Genome Zipper data as well as annotated chromosome 3B pseudomolecule data for download by the public, followed shortly by the release of diploid and other progenitor wheat species’ whole genome sequences for download by the public. The initial wheat genome assembly (IWGSC1 + POPSEQ; http://archive.plants.ensembl.org/Triticum_aestivum/Info/Index) was a chromosome survey sequence of the “Chinese Spring” cultivar of hexaploid wheat, refined into chromosomal pseudomolecules using POPSEQ data [66]. By August 2014, a new version of the wheat survey sequence gene model was made publicly available (v2.2). Up until the release of the latest TGAC version of the wheat reference genome assembly, three versions of the IWGSC-CSS data had been released to the public with improvements such as the cleaning of data by removal of duplicates (version 2) and the incorporation of 185 Gbp of mate pair sequence data (version 3).

In April 2016 the TGACv1 wheat genome assembly was made publicly available on Ensembl Plants. The new assembly (TGAC, [57]) has a greater representation of the genome, at 78% with N50 for the scaffolds of 88.8 kbp. This new assembly combined Illumina and PacBio sequencing platforms to identify 104,091 protein coding genes and 10,156 non-coding RNA genes. In June of that same year the IWGSC WGA v0.4 was made publicly available for BLAST analy-
sis or download (IWGSC WGA v0.4; http://plants.ensembl.org/Triticum_aestivum/Info/Index; [57]). This new assembly consists of 13.4 Gbp in contigs greater than 500 bp with 99% of the total number of genes from the previous assembly, located in the new TGAC assembly. In addition, alignments of RNA-seq data have also been added to this assembly. The whole genome assembly for wheat, available through the IWGSC (https://wheat-urgi.versailles.inra.fr/Seq-Repository/; [1]) provides researchers with a chromosome-based draft sequence of “Chinese Spring” wheat annotated with 124,201 gene loci and representing 61% of the genome sequence. The Illumina short sequence reads produced scaffolds adding to a total of 14.5 Gbp (L50 7.1 Mbp).

7. Future Prospects and Conclusions

From the data generated on the hexaploid wheat genome thus far and the number of mapping studies on this cereal, it is clear that novel and combined approaches are required in order to characterize genes of agronomic importance. Novelty can be found in the advances of sequencing platforms such as PacBio (http://www.pacificbiosciences.com). The platform provides a method of sequencing individual DNA molecules in real time with read lengths greater than 1000 bp [67]. This already provides a ten-fold improvement in sequence read length and in combination with optical mapping (BioNano; [68]), may provide exciting advances to the analysis of complex genomes such as hexaploid wheat. This technology may provide the highly needed bridge between DNA sequence and chromosome, as the inability to de novo assemble these complex genomes often fail to bring the two together [69] [70]. Using optical mapping technology, Hastie et al. (2013) [70] assembled a 2.1 Mbp repetitive region of the Aegilops tauschii genome. The technique makes direct visualization of sequence motifs on long single DNA molecules possible, and previously unplaced sequence contigs can therefore be anchored. Following this approach, Hastie et al. (2013) [70] improved the genome assembly of Ae. tauschii significantly by increasing its completeness from 75% complete to 95% complete.

Genome zippers (linear gene order models) are fast gaining popularity in mapping studies [71]. The Genome Zipper approach [72] makes use of comparisons of shotgun sequences with reference grass genomes such as Brachypodium, rice and sorghum. Genes identified within the syntenic regions generated by these comparisons are used to construct a genomic build along a scaffold of markers, taking into account the order of the sequence tagged genes within their reference genomes and the order deduced from the scaffold of markers [73]. Vitulo et al. (2011) [74] generated a genome zipper for the short and long arms of chromosome 5A of wheat by identifying conserved homologies in Brachypodium, rice and sorghum. The authors stated that the basis of the genome zipper concept is that the virtual gene order in one species can be constructed based on synteny with closely related species. Also in 2011, Mayer et al. [72] were able to assemble 21,766 genes in a putative linear order in barley through the use of
Understanding wheat biology not only serves academic purposes but paves the way toward increasing of yields and the identification of phenotypes that protect against biotic and abiotic stresses. Genome sequencing of wheat is one way to increase our understanding of this hexaploid cereal [7]. Sequencing the wheat genome facilitates our understanding of related grass species through annotation based on conservation of orthologous groups between the grasses [7]. The availability of a polyploid, such as wheat’s, genome sequence also provides an attractive model toward the study of genome changes behind the evolution of polyploidy in plants [65].

The wheat genome sequence has posed such a challenge to the scientific community, that the importance of compiling its genome has allowed for the implementation and advancement of new sequencing technologies such as PacBio and BioNano [52] [54]. These rapid advancements in science and technology, so akin to the green revolution of the 1960’s, have been pivotal in the completion of the wheat genome sequence [1].

References


[54] Zimin, A. (2016) Assembly of the 4.5 Gb Ancestral Wheat D-Genome from Hybrid


Submit or recommend next manuscript to SCIRP and we will provide best service for you:

Accepting pre-submission inquiries through Email, Facebook, LinkedIn, Twitter, etc.
A wide selection of journals (inclusive of 9 subjects, more than 200 journals)
Providing 24-hour high-quality service
User-friendly online submission system
Fair and swift peer-review system
Efficient typesetting and proofreading procedure
Display of the result of downloads and visits, as well as the number of cited articles
Maximum dissemination of your research work

Submit your manuscript at: http://papersubmission.scirp.org/
Or contact as@scirp.org