RESEARCH COMMENTARY

An overview of wheat genome sequencing and its implications for crop improvement

MEHANATHAN MUTHAMILARASAN and MANOJ PRASAD*

National Institute of Plant Genome Research, Aruna Asaf Ali Marg, New Delhi 110 067, India

[Muthamilarasan M. and Prasad M. 2014 An overview of wheat genome sequencing and its implication for crop improvement. *J. Genet.* **93**, 619–622]

Wheat (*Triticum aestivum* L.) serves as the staple food for 30% of the global population and is a rich source of proteins, minerals and other essential nutrients. But global warming is posing a serious threat to wheat productivity worldwide, and of note, wheat is extremely sensitive to heat, where ±2°C temperature variation has resulted in 50% decrease in wheat production (Asseng *et al.* 2011). Rise in greenhouse gases inflicts a steady increase in global temperature which has been projected to rise up to 4.5°C by 2080 (IPCC 2012; http://www.ipcc.ch/). This is expected to impose enormous negative impacts on productivity of wheat and substantial risks to global food production and security. This urged the scientific research community to work towards genetic improvement of wheat, so as to impart durable stress resistance and agronomic traits in this major cereal. Efforts have been invested on transgene-based approaches and molecular breeding programmes for improvement of wheat since times, but the progress is hindered due to the nonavailability of genome sequence information. Genome sequences are imperative for understanding the molecular basis of phenotypic traits and variation of a given crop plant. Though the genome sequence of model plants such as *Arabidopsis thaliana* and rice has revolutionized the understanding of plant biology over a decade, it has not been translated robustly into crop improvement for major cereals including wheat. Concurrently, less genomic conservation between rice and wheat has also restricted comparative genomic studies for genetic enhancement of wheat. This necessitated the sequencing of wheat genome, which would serve as the foundation for its improvement. Unfortunately, the size and complexity of wheat genome hindered the sequencing efforts, and this resulted in wheat becoming the only major crop whose genome remained unsequenced. With the advancements in next-generation sequencing (NGS) technologies and high-throughput sequence analysis platforms, whole genome sequencing of wheat has been made possible and recently its genome sequence has been released by two independent groups (table 1).

Brenchley *et al.* (2012) generated random shotgun libraries of total genomic and cDNA from *T. aestivum* cv. Chinese Spring (CS42) and sequenced them using the next-generation Roche 454 pyrosequencer. This generated ~85 Gb of sequence data (~220 million reads), which is equivalent to 5× coverage of estimated whole wheat genome (17 Gb). These whole genome NGS data identified around 94,000–96,000 genes that were further compared with sequence data from the progenitor genomes (*T. monoccum, Aegilops speltoides* and *Ae. Tauschii*), and about two-thirds of these genes were assigned to the A, B or D genomes (Brenchley *et al*. 2012). In addition, the genome of diploid wild relatives *Ae. tauschii* (DD) and *T. urartu* (AA) was also sequenced to identify 43,150 and 34,879 genes, respectively (Jia *et al.* 2013; Ling *et al.* 2013). Although these three studies have provided novel clues on the genes of hexaploid wheat and its wild diploid relatives, the location and distribution of genes on each of the bread wheat chromosomes and their evolution during the polyploidization events remained elusive. These bottlenecks have recently been resolved by the genome sequencing project led by International Wheat Genome Sequencing Consortium (2014). International Wheat Genome Sequencing Consortium (IWGSC) used aneuploid bread wheat lines derived from double ditelosomic stocks of the hexaploid wheat cultivar Chinese Spring to isolate, sequence and assemble *de novo* each individual chromosomes except chromosome 3B (IWGSC 2014). Chromosome 3B was isolated and sequenced as a complete chromosome by another independent study by Choulet *et al.* (2014).

IWGSC used the approach of separating the chromosomes using flow-cytometric sorting and then constructed paired-end sequence libraries with 500 bp target size. These libraries were sequenced on Illumina HiSeq 2000 and

*For correspondence. E-mail: manoj_prasad@nipgr.ac.in.

Keywords. genetic engineering; marker-assisted breeding; next-generation sequencing; wheat; whole genome sequence; SNP; miRNA.

Table 1. Comparative summary of wheat sequencing information of Brenchley *et al.* (2012) and International Wheat Genome Sequencing Consortium (2014).

	Parameters	Brenchley *et al.* (2012)	IWGSC (2014)
1	Cultivar	*T. aestivum* cv. Chinese Spring (CS2)	*Triticum aestivum* cv. 'Chinese Spring'; *T. urartu*; *Aegilops speltoides*; *T. turgidum* cv. *Cappelli*; *T. turgidum* cv. *Strongfield*
2	Nucleic acid sequence	Nuclear DNA and mRNA	Nuclear DNA and mRNA
3	Sequencing approach	Whole genome shotgun-next generation sequencing	Whole genome shotgun-next generation sequencing
4	Sequencing platform	Roche 454 pyrosequencing technology (GS FLX Titanium and GS FLX1 platforms)	Illumina HiSeq 2000 and Genome Analyser IIx
5	Insert size of DNA libraries	500–800 bp	500 bp
6	Genome assembly software	gsAssembler	ABySS
7	Draft genome size	17 Gb	17 Gb
8	Transposons	Covered ~80% genome; retroelements most abundant	Covered ~80% genome; class I (retroelements) and class II (DNA transposons) were more abundant in the A and D genome chromosomes, respectively
9	Small RNAs	No information	Two hundred and seventy miRNAs identified (49 are novel)
10	Total genes	94,000–96,000	124,201

Genome Analyzer IIx platforms to generate 100 or 150 base paired end reads (depth of between 30× and 241×). Similarly, the chromosome arms were also sequenced and the genome was assembled using different *in silico* approaches (to produce 17-Gb draft genome). The sequence read data and assembled sequence contigs were deposited in NCBI and European Nucleotide Archive under the accession ID 'PRJEB3955'. The data could also be retrieved from IWGSC database (http://wheat-urgi.versailles.inra.fr/Seq-Repository). Thus, the IWGSC decoded the nucleotide composition of all 21 chromosomes of bread wheat and identified 124,201 gene loci, with more than 75,000 loci positioned along the chromosomes. This study also compared the bread wheat gene sequences with the genome of wild diploid relatives (*T. urartu*, *Ae. speltoides*, *T. turgidum* cv. *Cappelli*, *T. turgidum* cv. *Strongfield*, *T. monococcum*, *Ae. sharonensis* and *Ae. tauschii*) and identified high sequence similarity and structural conservation with limited gene loss. It also predicted that none of the subgenomes dominated gene expression, resulting in a high degree of transcriptional autonomy. In addition, this project also identified 1,347,669 marker loci and 2,310,988 single nucleotide polymorphisms (SNPs) (IWGSC 2014).

Of the three homeologous sets of seven chromosomes (1A to 7A, 1B to 7B and 1D to 7D), chromosome 3B is the largest with a size of ~1 Gb. Being the first chromosome for which a BAC library was constructed (Safár *et al.* 2004) and a physical map generated (Paux *et al.* 2008), a detailed look at this chromosome was made by Choulet *et al.* (2014) using NGS approaches. A hybrid sequencing and BAC pooling strategy was used to sequence 8452 BAC clones from the minimal tiling path using Roche/454 and Illumina HiSeq2000 technologies. A final assembly of 2808 scaffolds representing 833 Mb with a N50 of 892 kb (i.e., half of the chromosome sequence is assembled in scaffolds larger than 892 kb) was produced. This analysis showed the size of chromosome 3B to be about 886 Mb, which is ~11% smaller than originally predicted. Further, a pseudomolecule of chromosome 3B was predicted by ordering 1358 scaffolds along the chromosome using an ordered set of 2594 anchor SNP markers. The pseudomolecule represented 774.4 Mb (93% of the complete sequence), with a scaffold N50 of 949 kb. The order of markers was estimated by linkage analysis of a recombinant inbred line population derived from a cross between *T. aestivum* cv. Chinese Spring (reference sequence) and cv. Renan (a French elite cultivar) and refined by incorporating linkage disequilibrium data from two panels and physical BAC contig information. Of note, the sequencing project identified 153,190 insertion site-based polymorphism markers and 35,579 microsatellite markers along the 3B chromosome. In addition, 121 quantitative trait loci (QTLs) for 50 different traits were also located on chromosome 3B (Choulet *et al.* 2014).

This NGS data was used to analyse the evolutionary relatedness and divergence times of the diploid genomes that have hybridized to form the A, B and D subgenomes of bread wheat (Marcussen *et al.* 2014). The study revealed that the A and B genomes diverged from a common ancestor ~7 million years ago and these genomes gave rise to the D genome through homoploid hybrid speciation 1 to 2 million years later (Marcussen *et al.* 2014). This showed that the current bread wheat genome is a product of multiple rounds of hybrid speciation (homoploid and polyploid) and this forms a base for a new framework for understanding the wheat genome as a multilevel phylogenetic mosaic (Marcussen *et al*. 2014). To gain insights into the transcriptome, Pfeifer *et al.* (2014)

sequenced the whole transcriptome of wheat grains (cv. Chinese Spring) using Illumina HiSeq 2000 platform and the analysis showed that 46,487 out of 85,173 high confidence wheat genes were expressed during endosperm development. The study also revealed that the transcriptional network delineates a complex and highly coordinated interplay between the individual wheat subgenomes, and further, it had also identified transcriptional active or inactive domains along the chromosomes that would be responsible for epigenetic control of grain development (Pfeifer *et al.* 2014).

Taken together, the availability of high-quality reference genome and transcriptome data would invariably expedite the crop improvement programmes. Using the existing genetic and genomic resources and tools, wheat researchers shall be able to integrate and apply the genome sequence information to explore the genes which are responsible for complex traits of agronomic importance. The genome sequence information will also enable the identification of important gene family members including transcription factors, peptide transporters, kinases, microRNAs, etc., through genomewide surveys. This would serve as a foundation for dissecting their cellular functions and understanding their roles in growth, development and stress response of wheat. Further, the availability of genome sequence will also facilitate the identification and positional cloning of genes responsible for traits selected during domestication, including the seed-shattering trait, which would lead to the identification of molecular changes selected during domestication.

An immediate application of genome sequence information is the development of high-density molecular markers which is of immense application in mapping agronomically desirable traits and also in identifying candidate genes within a region of interest. These traits could then be characterized and bred into elite wheat varieties. Further, the genome sequence data would also expedite sequence-enabled allele mining and accentuate it as an important prebreeding tool for wheat improvement programmes. The sequence-enabled allele mining exploits naturally occurring allelic variation at candidate genes controlling key agronomic traits and this strategy can assist in identifying the evolution of alleles, predicting novel useful haplotypes and also facilitate the development of allele-specific markers for use in marker-assisted selection. The availability of reference genome sequence will expedite resequencing the genomes of many wheat cultivars and aligning it to the reference genome for discovery of novel SNPs in large-scale. Eventually, these SNPs would assist in constructing high-resolution genetic maps and analysing the distribution of recombination and diversity along the wheat chromosomes. In addition to these, high-throughput genotyping-by-sequencing approach would also serve as an option for efficient marker assisted breeding, genomic selection and management of genetic resources at a higher resolution. Moreover, this would also open avenues for the researchers to explore the causal effects of other types of polymorphism that are related to structural variations, such as copy number variation, insertion–deletion polymorphisms and presence–absence variants.

In summary, the release of wheat whole genome sequence is considered to be a landmark in wheat genomics, which is anticipated to accelerate the research on this staple yet climate-vulnerable crop in both structural and functional genomics aspect (Muthamilarasan and Prasad 2013; Muthamilarasan *et al.* 2014). In view of this, the present study highlights the applications of wheat genome sequence information in wheat genetics and breeding, including the examination of genome variation (SNPs, InDels, etc.), genetic and association mapping, comparative genome mapping and identification of genes/QTLs responsible for stress tolerance and agronomic traits. The identified novel genes/QTLs and alleles (haplotypes) regulating stress tolerance and important agronomic characters can eventually be introgressed to cultivated wheat lines through marker-assisted genetic enhancement studies and transgene-based approaches. Further, the NGS data can be investigated to predict miRNAs and their roles in gene regulation which could be used in improving stress tolerance and productivity of wheat cultivars.

Acknowledgement

Mehanathan Muthamilarasan acknowledges University Grants Commission, New Delhi, India for providing Research Fellowship.

References

Asseng S., Foster I. and Turner N. C. 2011 The impact of temperature variability on wheat yields. *Global Change Biol.* **17**, 997–1012.

Brenchley R., Spannagl M., Pfeifer M., Barker G. L., D'Amore R., Allen A. M. *et al.* 2012 Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* **491**, 705–710.

Choulet F., Alberti A., Theil S., Glover N., Barbe V., Daron J. *et al.* 2014 Structural and functional partitioning of bread wheat chromosome 3B. *Science* **345** 1249721.

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

Jia J., Zhao S., Kong X., Li Y., Zhao G., He W. *et al.* 2013 *Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature* **496**, 91–95.

Ling H. Q., Zhao S., Liu D., Wang J., Sun H., Zhang C. *et al.* 2013 Draft genome of the wheat A-genome progenitor *Triticum urartu*. *Nature* **496**, 87–90.

Marcussen T., Sandve S. R., Heier L., Spannagl M., Pfeifer M., International Wheat Genome Sequencing Consortium *et al.* 2014 Ancient hybridizations among the ancestral genomes of bread wheat. *Science* **345** (doi: 10.1126/science.1250092).

Muthamilarasan M. and Prasad M. 2013 Wheat genome sequencing: a milestone in cereal genomics and its future potential. *Curr. Sci.* **104**, 286.

Muthamilarasan M., Parida S. K. and Prasad M. 2014 Advances in wheat genomics and its potential in ensuring food security in the scenario of climate change. *Proc. Indian Natl. Sci. Acad.* **80**, 325–331.

Paux E., Sourdille P., Salse J., Saintenac C., Choulet F., Leroy P. *et al.* 2008 A physical map of the 1-gigabase bread wheat chromosome 3B. *Science* **322**, 101–104.

Pfeifer M., Kugler K. G., Sandve S. R., Zhan B., Rudi H., Hvidsten T. R. *et al.* 2014 Genome interplay in the grain transcriptome of hexaploid bread wheat. *Science* **345** (doi: 10.1126/science.1250091).

Safár J., Bartos J., Janda J., Bellec A., Kubaláková M., Valárik M. *et al.* 2004 Dissecting large and complex genomes: Flow sorting and BAC cloning of individual chromosomes from bread wheat. *Plant J.* **39**, 960–968.

Received 24 July 2014; accepted 30 July 2014
Unedited version published online: 18 August 2014
Final version published online: 4 December 2014