The pangenome of hexaploid bread wheat

Juan D. Montenegro[1,†], Agnieszka A. Golicz[1,2,†,‡], Philipp E. Bayer[2,†], Bhavna Hurgobin[1,2], HueyTyng Lee[1,2], Chon-Kit Kenneth Chan[2], Paul Visendi[1], Kaitao Lai[3], Jaroslav Doležel[4], Jacqueline Batley[1,2,5] and David Edwards[1,2,5,*]

[1]*School of Agriculture and Food Sciences, University of Queensland, Brisbane, Australia,*
[2]*School of Plant Biology, University of Western Australia, Crawley, WA 6009, Australia,*
[3]*CSIRO, North Ryde, NSW 2113, Australia,*
[4]*Institute of Experimental Botany, Centre of the Region Haná for Biotechnological and Agricultural Research, Šlechtitelů 31, CZ-783 71 Olomouc, Czech Republic, and*
[5]*Institute of Agriculture, University of Western Australia, Crawley, WA 6009, Australia*

[*] For correspondence (e-mail dave.edwards@uwa.edu.au).

[†] These authors contributed equally to the manuscript.

[‡] Present address: Plant Molecular Biology and Biotechnology Laboratory, Faculty of Veterinary and Agricultural Sciences, University of Melbourne, Parkville, Melbourne, VIC, Australia.

SUMMARY

There is an increasing understanding that variation in gene presence–absence plays an important role in the heritability of agronomic traits; however, there have been relatively few studies on variation in gene presence–absence in crop species. Hexaploid wheat is one of the most important food crops in the world and intensive breeding has reduced the genetic diversity of elite cultivars. Major efforts have produced draft genome assemblies for the cultivar Chinese Spring, but it is unknown how well this represents the genome diversity found in current modern elite cultivars. In this study we build an improved reference for Chinese Spring and explore gene diversity across 18 wheat cultivars. We predict a pangenome size of 140 500 ± 102 genes, a core genome of 81 070 ± 1631 genes and an average of 128 656 genes in each cultivar. Functional annotation of the variable gene set suggests that it is enriched for genes that may be associated with important agronomic traits. In addition to variation in gene presence, more than 36 million intervarietal single nucleotide polymorphisms were identified across the pangenome. This study of the wheat pangenome provides insight into genome diversity in elite wheat as a basis for genomics-based improvement of this important crop. A wheat pangenome, GBrowse, is available at http://appliedbioinformatics.com.au/cgi-bin/gb2/gbrowse/WheatPan/, and data are available to download from http://wheatgenome.info/wheat_genome_databases.php.

Keywords: wheat, genome, pangenome, diversity, single nucleotide polymorphisms, database, *Triticum aestivum*.

INTRODUCTION

Wheat is one of the most important food crops in the world, and its continued improvement is essential to maintain food security in the face of a growing human population and the disturbance of agricultural production due to climate change (Abberton *et al.*, 2015; Batley and Edwards, 2016). Wheat was domesticated 8000–10 000 years ago (Dubcovsky and Dvorak, 2007), and today bread wheat (*Triticum aestivum*) provides roughly a fifth of the world's food. Genome analysis in bread wheat is a challenge because of its large (17 Gbp) genome, consisting of between 80 and 90% repetitive sequence (Wanjugi *et al.*, 2009; Šafář *et al.*, 2010). Bread wheat is also hexaploid, being derived from a combination of three diploid donor species which are proposed to have diverged from an ancestral diploid species between 2.5 and 6 million years ago (Huang *et al.*, 2002; Chantret *et al.*, 2005). There have been several efforts to sequence the genome of hexaploid bread wheat. The *de novo* assembly of sequence data from flow-sorted chromosome arms was initially performed for 7DS, demonstrating that it was possible to assemble all known 7DS genes (Berkman *et al.*, 2011). The same approach delimited a translocation between chromosome arms 7BS and 4AL (Berkman *et al.*, 2012), with a subsequent comparison of all group 7 chromosomes, highlighting genomic changes during the early evolution and domestication of this important crop (Berkman *et al.*, 2013). The application of a similar approach towards all chromosome arms with the exception of 3B (IWGSC, 2014), together with a whole-genome assembly of Roche 454 sequence data (Brenchley *et al.*, 2012), provided the first draft genome assemblies for the wheat cultivar Chinese Spring. Two additional cultivars, OpataM85 and W7984, have undergone whole-genome shotgun sequencing using Illumina data, and although gene presence comparisons were performed using cDNA mapping, these assemblies were not annotated (Chapman *et al.*, 2015), limiting their use for pangenome analysis. With the exception of Chapman *et al.* (2015), each of these studies has focused on the cultivar Chinese Spring.

Crop breeding increasingly benefits from the application of molecular tools such as marker-assisted selection (MAS), and more recently genomic selection (GS), and the increasing availability of genomic information supports these advanced breeding tools (Poland *et al.*, 2012; Crossa *et al.*, 2014; Simeao Resende *et al.*, 2014; Cros *et al.*, 2015; Sallam *et al.*, 2015). Modern molecular breeding tools apply single nucleotide polymorphism (SNP) molecular genetic markers, and numerous studies have discovered and validated large numbers of SNP markers across the wheat genome (Lai *et al.*, 2012, 2015; Wang *et al.*, 2014; Winfield *et al.*, 2015). SNPs have been used to find genes that are undergoing selective sweeps and population bottlenecks (Cavanagh *et al.*, 2013), and have also been used to map low-diversity regions which could have been targets of selection (Lai *et al.*, 2015).

The decreasing cost of DNA sequencing has accelerated genomics research in recent years (Edwards *et al.*, 2013; Visendi *et al.*, 2013). Most sequencing projects focus on reference genome assembly and the discovery of SNPs; however, the importance of structural variants is becoming increasingly acknowledged (Saxena *et al.*, 2014; Jordan *et al.*, 2015; Wendel *et al.*, 2016). Studies in several plant species have revealed the existence of extensive structural variation (Springer *et al.*, 2009; Xu *et al.*, 2012; Gordon *et al.*, 2014; Li *et al.*, 2014; Zhang *et al.*, 2014; Jordan *et al.*, 2015; Hardigan *et al.*, 2016). One form of structural variation, the presence or absence of genes or genomic regions between individuals of the same species, is being increasingly acknowledged as an important form of variation in plants, and the sum of core and variable regions of the genome for a species is known as the pangenome.

Several approaches to pangenome assembly and analysis have been developed (Golicz *et al.*, 2015a). The traditional approach, first applied in bacteria, involves whole-genome assembly of all genotypes, followed by individual annotation and comparison of the gene content (Tettelin *et al.*, 2005; Li *et al.*, 2014; Schatz *et al.*, 2014). An alternative is a read mapping and assembly approach, where sequence reads are first mapped to an existing reference and the unmapped reads are then assembled (Golicz *et al.*, 2015a, 2016; Yao *et al.*, 2015).

The first step towards the production of a pangenome for a crop species is the production of a suitable reference assembly, followed by the expansion of this reference with additional sequences from other varieties that are not present in the reference. In this study we have reassembled a draft Chinese Spring wheat genome reference and used this as the basis for a pangenome study, identifying core and variable genes across 18 cultivars, 16 of which were selected by a large Australian national initiative (Edwards *et al.*, 2012) with the remaining two being the only other public whole-genome data sets available of sufficient sequencing depth. We have also identified 36.4 million SNPs between these 18 cultivars. The Chinese Spring reference is different in gene content from the 18 cultivars, suggesting that this pangenome and the associated SNPs may provide a better reference for wheat crop improvement than the current Chinese Spring references.

RESULTS AND DISCUSSION

Wheat (cv. Chinese Spring) genome assembly

An assessment of the sequence duplication in the IWGSC draft Chinese Spring assembly (IWGSC, 2014) showed that 663 Mb (7%) of the assembly consisted of exact duplications greater than 1 kb, with more than 40% of chromosome arms 4AS and 4AL being duplicated (Figure S1, Table S6 in the Supporting Information). Following reassembly, producing 10.7 Gb of new reference, these duplications were reduced to of 0.4 Mb (0.004%). The high frequency of duplicated regions in the IWGSC assembly (Figure S1, Table S6) may be an artefact of using the parallelized de Bruijn graph assembler ABySS (Simpson *et al.*, 2009) as they were not observed in the previous assemblies of group 7 data (Berkman *et al.*, 2011, 2012, 2013) which used the non-parallel de Bruijn graph assembler Velvet (Zerbino and Birney, 2008).

A reassembly of the IWGSC data in this study using Velvet produced a reference with a larger assembly size and greatly reduced frequency of duplicated regions (Figure S1, Table S6) compared with the published draft genome (IWGSC, 2014). CEGMA analysis (Parra *et al.*, 2009) was performed to assess the completeness of the assembly, and it identified 245 (98.8%) of the 248 core eukaryotic genes compared with 243 genes identified in the IWGSC assembly.

Pangenome assembly

Whole-genome sequence reads from 18 wheat cultivars were mapped to the new Chinese Spring assembly, and unmapped reads assembled. The sequencing depth ranged from 8.4× to 20.0×, except for Chinese Spring which had a coverage that ranged from 60× to 200× for each of the chromosome arms. (Table S5). After removal of contaminant sequences, the newly assembled sequence contained 221 991 scaffolds with a total length of 350 Mb (Table S1) and a total of 21 653 predicted genes. Mapping of Chinese Spring sequence reads to the pangenome demonstrated that this sequence was not present in the Chinese Spring reference and represents a 3.3% increase in the size of the wheat reference genome. A similar approach was used by Yao *et al.* (2015) with 1483 rice accessions from the *japonica* and *indica* groups, where they assembled 15.8 Mb and 24.6 Mb of additional sequence for each subspecies, respectively, representing an increase of 4 and 6% in genome size. Similarly, local reassembly in *Brachypodium distachyon* identified 19.2 Mb of additional sequence in seven highly diverse inbred lines, a 5% increase in the size of the reference genome. Golicz *et al.* (2016) characterized the pangenome of *Brassica oleracea* using nine diverse morphotypes, and assembled an additional 99 Mbp of sequence. The relatively small increase in pangenome assembly size we observe reflects the high degree of relatedness of the cultivars sequenced (Lai *et al.*, 2015). The additional sequence identified in this study is likely to be an underestimate of the total sequence content present in the cultivars, as sequences present in only one or two of the cultivars are unlikely to have sufficient coverage to assemble, as IDBA-UD has 81% assembly efficiency for samples with a sequencing depth of 10× (Peng *et al.*, 2012).

Discovery of gene presence–absence

The presence or absence of each gene was predicted for each cultivar based on the mapping of reads from each cultivar to the new pangenome assembly (Table S2). The approach followed the method of Golicz *et al.* (2016) which demonstrates a 0.05% error rate using 10× read coverage. Based on Chinese Spring read mapping to the pangenome, none of the additional genes identified in the 18 cultivars were identified as present in Chinese Spring. On average, each cultivar contains 128 656 genes, with 89 795 (64.3%) shared by all 19 cultivars, while 49 952 genes represent the variable genome across these cultivars. Based on gene presence and absence in each of the 18 cultivars we estimate that the pangenome of modern wheat cultivars contains 140 500 ± 102 genes (Figure 1), with an average of 49 unique genes per cultivar. This is similar to the 37 unique genes per cultivar identified in a similar study in *B. oleracea* (Golicz *et al.*, 2016) This is likely to be an underestimate of the broader wheat pangenome as it is predicted from a relatively narrow set of cultivars, and extending the study to more diverse landraces and wild relatives will provide a more comprehensive measure of the gene content of this important crop species. In addition, further analysis is required to determine the source of the origin of the variable sequence.

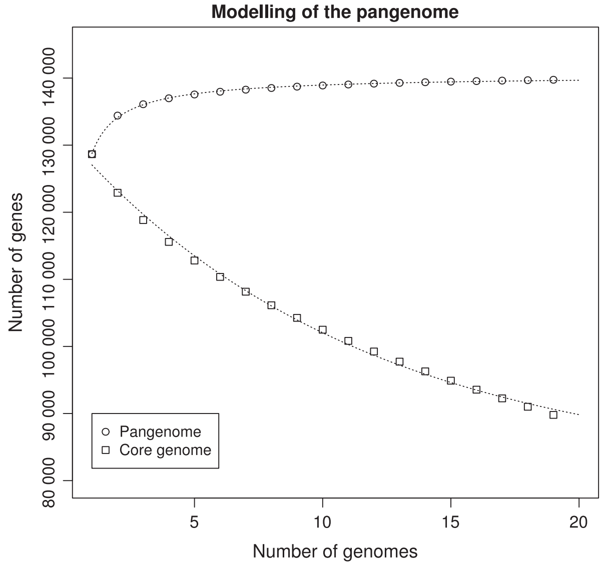

Figure 1. Modelling of the wheat pangenome and core genome. The modelling of the pangenome expansion predicts a closed pangenome with a total of around 140 000 genes. The core genome is predicted to contain around 81 000 of these genes.

Characterization of Chinese Spring gene content identified 245 genes in Chinese Spring which are absent from the 18 cultivars, while a further 12 150 genes were identified in all 18 cultivars but are not found in Chinese Spring (Table S2). A dendrogram reconstructed using gene presence–absence variation places Chinese Spring in a separate cluster at the base of the tree (Figure 2). This is similar to a previous study using simple sequence repeat markers where Chinese Spring was placed in the basal node away from most modern wheat cultivars (Plaschke *et al.*, 1995). Our results can also be explained by the history of Chinese Spring, which despite being a major source of cytogenetic stocks used in the discovery of the seven homoeologous chromosome groups and in early gene mapping efforts (Sears, 1966; Sharp *et al.*, 1988), and more recently in genome sequencing (IWGSC, 2014), is not widely used in breeding programmes due to its susceptibility to biotic and abiotic stress (Sears and Miller, 1985). Our identification of Chinese Spring as a genomic outlier with substantial sequence difference from current varieties is supported by recent paper by Liu *et al.* (2016), which shows that as much as 159.3 Mb of additional sequence is present in chromosome 3B of CRNIL1A and absent in Chinese Spring.

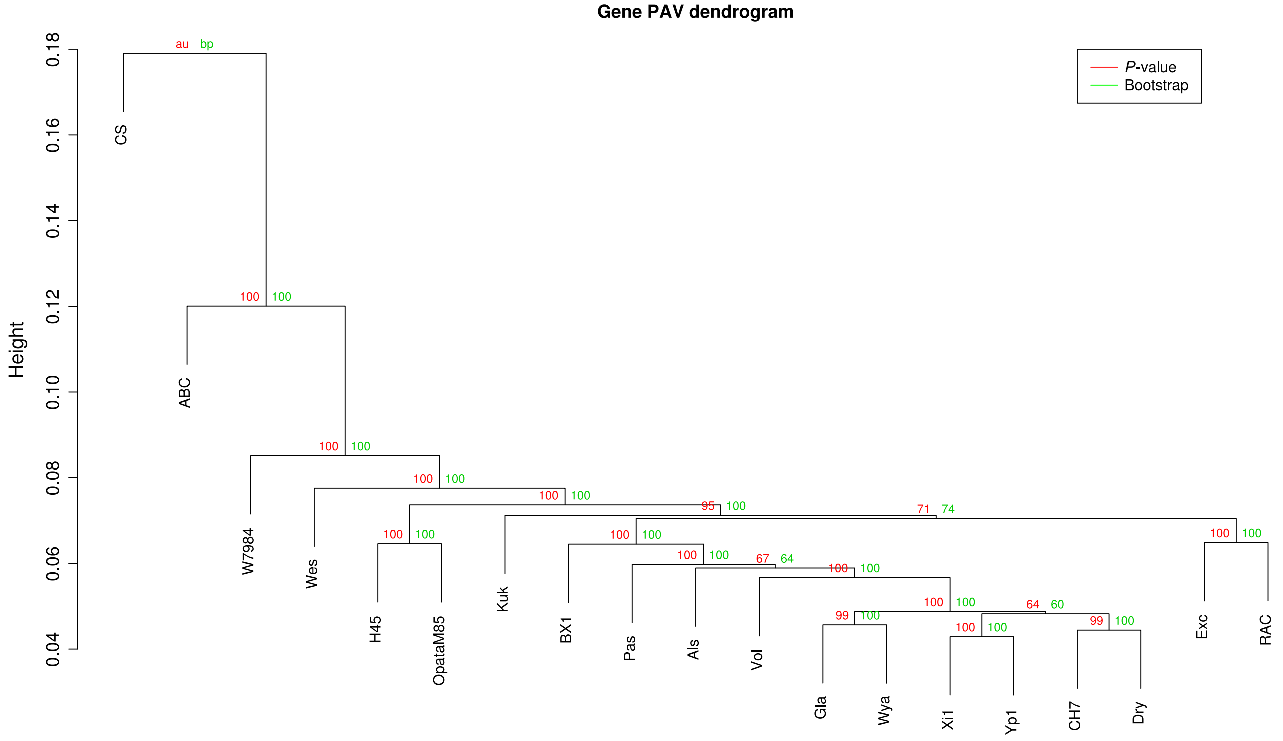

Figure 2. Dendrogram of the 19 cultivars based on the presence–absence variation of genes in each cultivar.
Five hundred iterations were performed for bootstrap and *P*-value calculations (see Figure S2). [Colour figure can be viewed at wileyonlinelibrary.com].

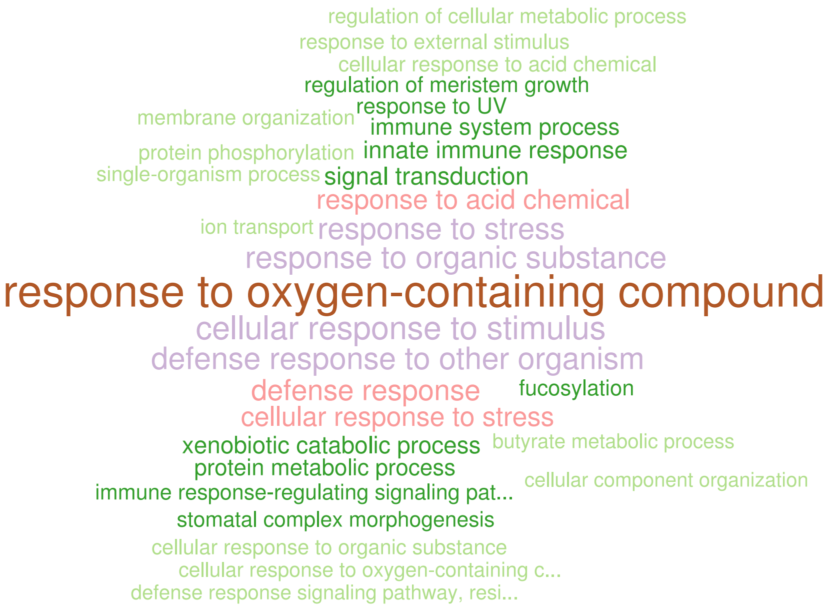

Figure 3. Functional enrichment analysis of the variable genome.
Graphical representation the 30 most enriched biological processes in the variable genome of wheat (see Table S3). [Colour figure can be viewed at wileyonlinelibrary.com].

Variable genes were annotated, and functional enrichment analysis suggests that the variable genome is enriched with genes involved in response to environmental stress and defence response (Figure 3, Table S3). Similarly, Yao *et al.* (2015) found that the variable genome of rice was enriched with genes related to defence to biotic stress, including NBS LRR genes and genes coding for protein kinases and abiotic stress tolerance (Yao *et al.*, 2015). Analysis of the *B. oleracea* pangenome by (2016) also found that variable genes were enriched for annotated genes related to major agronomic traits, including disease resistance.

SNP discovery

The capture and characterization of diversity are essential in the design and execution of breeding programmes. We have previously identified more than 4 million SNPs on the group 7 Chinese Spring chromosomes with a validation rate of 95% (Lai *et al.*, 2015). Using the same method, whole-genome shotgun reads from the 18 wheat cultivars were mapped to the pangenome assembly and SNPs were identified using SGSautoSNP (Lorenc *et al.*, 2012), leading to the identification of 36.4 million SNPs. Of these, 2.87 million were identified in scaffolds not present in the Chinese Spring assembly. The SGSautoSNP calls were compared with SNPs from a published Infinium array (Wang *et al.*, 2014). A total of 13 541 Infinium SNPs were identified as being at the same location as the SGSautoSNP calls. Of these, 96.3% were identified as polymorphic. This is similar to the validation rate observed by Lai *et al.* (2015) using the same approach. The majority of SNPs were found in intergenic regions, with only 392 142 (1%) SNPs located in coding regions. Of these 225 064 (57.4%) are predicted to be non-synonymous, resulting in a potentially different functional protein. These results are comparable to those obtained by Jordan *et al.* (2015), who found that 52.3% of the SNPs were non-synonymous (Jordan *et al.*, 2015). The dataset represents the most comprehensive SNP resource available for the improvement of elite bread wheat cultivars.

CONCLUSION

In this study, we constructed and analysed a draft wheat pangenome using a single reference and whole-genome sequencing data from 18 cultivars. The pangenome contains 128 656 predicted genes of which 64.3% are identified as core, which is present in all cultivars, while the remainder are variable and display presence–absence variation. Additionally, 12 150 genes are absent in the Chinese Spring reference sequence but present in all the other cultivars analysed. The pangenome sequence is a valuable resource for scientists involved in wheat genomics and breeding as understanding the diversity of genes is essential for their association with agronomic traits. Extending the genome reference and SNP content to regions which are not present in Chinese Spring provides a more complete resource for genomics-based improvement of wheat crops.

EXPERIMENTAL PROCEDURES

Genome assembly and annotation

Sequence data were downloaded from various repositories as described in Table S4. Clonal reads were removed using an in-house script (remove_clones.pl). Quality trimming and adapter clipping were performed using TRIMMOMATIC v.0.33 (Bolger *et al.*, 2014), and sequences shorter than 73 bp were removed. VELVET v.1.2.10 (Zerbino and Birney, 2008) was used for assembly using a kmer size of 71. RNA-seq reads were aligned to the reference genome using TOPHAT2 v.2.1.0.1 (Kim *et al.*, 2013). Accepted alignments were transformed into hints files with the script bam2hints from the AUGUSTUS package.

REPEATMASKER (Smit *et al.*, 2015) was used to mask repeated regions using RepBase version 20150807 (Jurka *et al.*, 2005) and 'viridiplantae' as species. AUGUSTUS v.2.1.0 (Keller *et al.*, 2011) predicted gene models using the hints produced from the RNA-seq alignments. Gene models were first filtered for size (≥300 bp). BEDOPS v.2.4.15 (Neph *et al.*, 2012) was used to identify and remove gene models that were not supported by TOPHAT2 annotation or overlapped repeat-masked regions. Finally, the protein sequences of the selected models were aligned to TE-related proteins with BLASTP and those with significant alignments (E-value $\leq 1 \times 10^{-5}$) were removed from the annotation. The protein sequences of the final gene set were aligned to the proteome of *Triticum urartu* to identify and merge split genes.

CEGMA (Parra *et al.*, 2009) was used to assess the completeness of the reference genome prior to annotation with default parameters.

Pangenome assembly and annotation

Reads from the 16 wheat cultivars were mapped to the new Chinese Spring assembly using BOWTIE2 v.2.2.5, and unmapped reads pooled. The sequencing depth per cultivar is shown in Table S5. TRIMMOMATIC v.0.33 removed adapter and low-quality sequence and the reads were assembled using IDBA_UD (Peng *et al.*, 2012) with standard parameters. The resulting scaffolds were compared with the NCBI non-redundant nucleotide database using BLASTn (E-value $\leq 1 \times 10^{-5}$) and the scaffolds with hits outside the seed plants taxonomy group were removed. REPEATMASKER v.4.0.6 masked repetitive elements using 'viridiplantae' as the species. Then, tBLASTx (Camacho *et al.*, 2009) was used to align the green plant expressed sequence tags (ESTs) from GenBank, and genes were predicted using AUGUSTUS v.2.1.0, supported by the EST alignments. The reads from W7984, OataM85 and 90 doubled haploid offspring were mapped to the full pangenome assembly and unmapped reads were processed and assembled as described above. Libraries with a mapping efficiency below 80% were not included for further analysis.

Gene presence–absence and pangenome prediction

BOWTIE2 v.2.2.5 was used to align the reads with standard parameters and an insert size between 0 and 1000 bp. Gene presence–absence was called as described by Golicz *et al.* (2015b). SAMTOOLS was used to calculate the coverage of the annotated genes, and an in-house script (pileup2cov.pl) predicted the presence–absence status of each gene based on the following requirements: coverage >2× and exon fraction covered >0.05. PVCLUST (Suzuki and Shimodaira, 2006) was used with the presence–absence binary matrix to estimate the relationship between the cultivars. One thousand resamplings were used for bootstrap calculations.

The program PANGP (Zhao *et al.*, 2014) was used to count the core and total genes present in all possible combinations of the 19 cultivars. The average resulting gene count from each iteration was plotted and used to model the wheat pangenome expansion using a power-law model ($f(x) = Ax^B + C$) (Tettelin *et al.*, 2005) by means of the R nls function. Assuming a closed pangenome, the C parameter was used as an estimator of the total gene content in the pangenome. The same approach was used to estimate the core genome, using the average gene count to fit the model $f(x) = Ae^{Bx} + C$.

SNP discovery

Reads were mapped to the pangenome using BOWTIE2 v.2.2.5 (–no-mixed –no-unal -I 0 -X 1000) (Langmead and Salzberg, 2012). Reads with mapping quality (MAPQ) < 20 and with low base qualities were removed from the alignments along with their mates. SAM files were further processed and duplicated reads removed with SAMTOOLS v.1.3.1 (Li *et al*., 2009). SGSAUTOSNP (Lorenc *et al*., 2012) was used to identify SNPs. SNPs were validated as described in Lai *et al*. (2015). SNPEFF v.4.2 (Cingolani *et al*., 2012) was used to predict the effect of the SNPs on the gene annotations.

SNP validation

The sequence tags from the 90k SNP Infinium array (Wang *et al*., 2014) were aligned to the reference wheat pangenome using NCBI BLAST Plus (Camacho *et al*., 2009). High-quality alignments (*E*-threshold $<1 \times 10^{-10}$ and ≥99% sequence identity) where used to identify common polymorphic SNPs as described in Lai *et al*. (2015).

ACKNOWLEDGEMENTS

This work was supported by the Australian Research Council (projects LP140100537 and LP130100925) and the Czech Republic National Program of Sustainability I (award 1204). Support is also acknowledged from the Australian Genome Research Facility (AGRF) and the Queensland Cyber Infrastructure Foundation (QCIF), the Pawsey Supercomputing Centre with funding from the Australian Government and the Government of Western Australia and resources from the National Computational Infrastructure (NCI), which is supported by the Australian Government. The authors declare no conflict of interest.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Comparison of duplicated sequence in the reference genome and the IWGSC assembly.

Table S1. Assembly statistics of the pooled unmapped reads of 18 wheat cultivars.

Table S2. Gene presence–absence variation in the wheat pangenome across the 18 wheat cultivars. (As this file is very large it can be downloaded from http://wheatgenome.info/.)

Table S3. Gene enrichment of the variable genome ($P < 0.01$).

Table S4. Source of data uses in analysis.

Table S5. Sequence depth of the wheat cultivars.

Table S6. Comparison of duplicated sequence in the reference genome and the IWGSC assembly.

REFERENCES

Abberton, M., Batley, J., Bentley, A. *et al.* (2015) Global agricultural intensification during climate change: a role for genomics. *Plant Biotechnol. J.* **14**, 1095–1098.

Batley, J. and Edwards, D. (2016) The application of genomics and bioinformatics to accelerate crop improvement in a changing climate. *Curr. Opin. Plant Biol.* **30**, 78–81.

Berkman, P.J., Manoli, S., McKenzie, M. *et al.* (2011) Sequencing and assembly of low copy and genic regions of isolated Triticum aestivum chromosome arm 7DS. *Plant Biotechnol. J.* **9**, 768–775.

Berkman, P.J., Skarshewski, A., Manoli, S. *et al.* (2012) Sequencing wheat chromosome arm 7BS delimits the 7BS/4AL translocation and reveals homoeologous gene conservation. *Theor. Appl. Genet.* **124**, 423–432.

Berkman, P.J., Visendi, P., Lee, H.C. *et al.* (2013) Dispersion and domestication shaped the genome of bread wheat. *Plant Biotechnol. J.* **11**, 564–571.

Bolger, A.M., Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Brenchley, R., Spannagl, M., Pfeifer, M. *et al.* (2012) Analysis of the bread-wheat genome using whole-genome shotgun sequencing. *Nature*, **491**, 705–710.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Cavanagh, C.R., Chao, S., Wang, S. *et al.* (2013) Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. *Proc. Natl Acad. Sci. USA*, **110**, 8057–8062.

Chantret, N., Salse, J., Sabot, F. *et al.* (2005) Molecular basis of evolutionary events that shaped the hardness locus in diploid and polyploid wheat species (*Triticum* and *Aegilops*). *Plant Cell*, **17**, 1033–1045.

Chapman, J.A., Mascher, M., Buluç, A. *et al.* (2015) A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome. *Genome Biol.* **16**, 1–17.

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Tung, N., Wang, L., Land, S.J., Lu, X. and Ruden, D.M. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w(1118); iso-2; iso-3. *Fly*, **6**, 80–92.

Cros, D., Denis, M., Sanchez, L. *et al.* (2015) Genomic selection prediction accuracy in a perennial crop: case study of oil palm (Elaeis guineensis Jacq.). *Theor. Appl. Genet.* **128**, 397–410.

Crossa, J., Perez, P., Hickey, J. *et al.* (2014) Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity*, **112**, 48–60.

Dubcovsky, J. and Dvorak, J. (2007) Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science*, **316**, 1862–1866.

Edwards, D., Wilcox, S., Barrero, R.A. *et al.* (2012) Bread matters: a national initiative to profile the genetic diversity of Australian wheat. *Plant Biotechnol. J.* **10**, 703–708.

Edwards, D., Batley, J. and Snowdon, R.J. (2013) Accessing complex crop genomes with next-generation sequencing. *Theor. Appl. Genet.* **126**, 1–11.

Golicz, A.A., Batley, J. and Edwards, D. (2015a) Towards plant pangenomics. *Plant Biotechnol. J.* **14**, 1099–1105.

Golicz, A.A., Martinez, P.A., Zander, M., Patel, D.A., Van De Wouw, A.P., Visendi, P., Fitzgerald, T.L., Edwards, D. and Batley, J. (2015b) Gene loss in the fungal canola pathogen Leptosphaeria maculans. *Funct. Integr. Genomics*, **15**, 189–196.

Golicz, A.A., Bayer, P.E., Barker, G.C. *et al.* (2016) The pangenome of an agronomically important crop plant Brassica oleracea. *Nat. Commun.* **7**, 13390.

Gordon, S.P., Priest, H., Marais, D.L.D. *et al.* (2014) Genome diversity in Brachypodium distachyon: deep sequencing of highly diverse inbred lines. *Plant J.* **79**, 361–374.

Hardigan, M.A., Crisovan, E., Hamilton, J.P. *et al.* (2016) Genome reduction uncovers a large dispensable genome and adaptive role for copy number variation in asexually propagated solanum tuberosum. *Plant Cell*, **28**, 388–405.

Huang, S., Sirikhachornkit, A., Su, X., Faris, J., Gill, B., Haselkorn, R. and Gornicki, P. (2002) Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the Triticum/Aegilops complex and the evolutionary history of polyploid wheat. *Proc. Natl Acad. Sci. USA*, **99**, 8133–8138.

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

Jordan, K.W., Wang, S., Lun, Y. *et al.* (2015) A haplotype map of allohexaploid wheat reveals distinct patterns of selection on homoeologous genomes. *Genome Biol.* **16**, 48.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O. and Walichiewicz, J. (2005) Repbase update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467.

Keller, O., Kollmar, M., Stanke, M. and Waack, S. (2011) A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics*, **27**, 757–763.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg, S.L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36.

Lai, K., Duran, C., Berkman, P.J. *et al.* (2012) Single nucleotide polymorphism discovery from wheat next-generation sequence data. *Plant Biotechnol. J.* **10**, 743–749.

Lai, K., Lorenc, M.T., Lee, H. *et al.* (2015) Identification and characterisation of more than 4 million inter-varietal SNPs across the group 7 chromosomes of bread wheat. *Plant Biotechnol. J.* **13**, 97–104.

Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Meth.* **9**, 357–359.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and **1000 Genome Project Data Processing Subgroup**. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li, Y.H., Zhou, G.Y., Ma, J.X. *et al.* (2014) De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat. Biotechnol.* **32**, 1045–1052.

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

Lorenc, M.T., Hayashi, S., Stiller, J. *et al.* (2012) Discovery of single nucleotide polymorphisms in complex genomes using SGSautoSNP. *Biology*, **1**, 370–382.

Neph, S., Kuehn, M.S., Reynolds, A.P. *et al.* (2012) BEDOPS: high-performance genomic feature operations. *Bioinformatics*, **28**, 1919–1920.

Parra, G., Bradnam, K., Ning, Z., Keane, T. and Korf, I. (2009) Assessing the gene space in draft genomes. *Nucleic Acids Res.* **37**, 289–297.

Peng, Y., Leung, H.C.M., Yiu, S.M. and Chin, F.Y.L. (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, **28**, 1420–1428.

Plaschke, J., Ganal, M.W. and Roder, M.S. (1995) Detection of genetics diversity in closely-related bread wheat using microsatellite markers. *Theor. Appl. Genet.* **91**, 1001–1007.

Poland, J., Endelman, J., Dawson, J. *et al.* (2012) Genomic selection in wheat breeding using genotyping-by-sequencing. *Plant Gen.* **5**, 103–113.

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

Sallam, A.H., Endelman, J.B., Jannink, J.L. and Smith, K.P. (2015) Assessing genomic selection prediction accuracy in a dynamic barley breeding population. *Plant Genome*, **8**, 1–15.

Saxena, R.K., Edwards, D. and Varshney, R.K. (2014) Structural variations in plant genomes. *Brief. Funct. Genomics*, **13**, 296–307.

Schatz, M.C., Maron, L.G., Stein, J.C. *et al.* (2014) Whole genome de novo assemblies of three divergent strains of rice, Oryza sativa, document novel gene space of aus and indica. *Genome Biol.* **15**, 1–16.

Sears, E. (1966) Nullisomic-tetrasomic combinations in hexaploid wheat. In *Chromosome Manipulations and Plant Genetics: The contributions to a symposium held during the Tenth International Botanical Congress Edinburgh 1964* (Riley, R. and Lewis, K.R., eds). Boston MA: Springer, pp. 29-45.

Sears, E.R. and Miller, T.E. (1985) The history of Chinese Spring wheat. *Cereal Res. Commun.* **13**, 261–263.

Sharp, P.J., Kreis, M., Shewry, P.R. and Gale, M.D. (1988) Location of β-amylase sequences in wheat and its relatives. *Theor. Appl. Genet.* **75**, 286–290.

Simeao Resende, R.M., Casler, M.D. and Vilela de Resende, M.D. (2014) Genomic selection in forage breeding: accuracy and methods. *Crop Sci.* **54**, 143–156.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J.M. and Birol, I. (2009) ABySS: a parallel assembler for short read sequence data. *Genome Res.* **19**, 1117–1123.

Smit, A., Hubley, R. and Green, P. (2015) RepeatMasker Open-4.0. http://www.repeatmasker.org.

Springer, N.M., Ying, K., Fu, Y. *et al.* (2009) Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genet.* **5**, e1000734.

Suzuki, R. and Shimodaira, H. (2006) Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, **22**, 1540–1542.

Tettelin, H., Masignani, V., Cieslewicz, M.J. *et al.* (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial “pan-genome”. *Proc. Natl Acad. Sci. USA*, **102**, 13950–13955.

Visendi, P., Batley, J. and Edwards, D. (2013) Next generation characterisation of cereal genomes for marker discovery. *Biology*, **2**, 1357–1377.

Wang, S., Wong, D., Forrest, K. *et al.* (2014) Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. *Plant Biotechnol. J.* **12**, 787–796.

Wanjugi, H., Coleman-Derr, D., Huo, N., Kianian, S.F., Luo, M.-C., Wu, J., Anderson, O. and Gu, Y.Q. (2009) Rapid development of PCR-based genome-specific repetitive DNA junction markers in wheat. *Genome*, **52**, 576–587.

Wendel, J.F., Jackson, S.A., Meyers, B.C. and Wing, R.A. (2016) Evolution of plant genome architecture. *Genome Biol.* **17**, 37.

Winfield, M.O., Allen, A.M., Burridge, A.J. *et al.* (2015) High-density SNP genotyping array for hexaploid wheat and its secondary and tertiary gene pool. *Plant Biotechnol. J.* **14**, 1195–1206.

Xu, X., Liu, X., Ge, S. *et al.* (2012) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat. Biotech.* **30**, 105–111.

Yao, W., Li, G., Zhao, H., Wang, G., Lian, X. and Xie, W. (2015) Exploring the rice dispensable genome using a metagenome-like assembly strategy. *Genome Biol.* **16**, 1–20.

Zerbino, D.R. and Birney, E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829.

Zhang, L.-M., Luo, H., Liu, Z.-Q., Zhao, Y., Luo, J.-C., Hao, D.-Y. and Jing, H.-C. (2014) Genome-wide patterns of large-size presence/absence variants in sorghum. *J. Integr. Plant Biol.* **56**, 24–37.

Zhao, Y., Jia, X., Yang, J., Ling, Y., Zhang, Z., Yu, J., Wu, J. and Xiao, J. (2014) PanGP: a tool for quickly analyzing bacterial pan-genome profile. *Bioinformatics*, **30**, 1297–1299.