Identification and characterization of more than 4 million intervarietal SNPs across the group 7 chromosomes of bread wheat

Kaitao Lai1,2, Michal T. Lorenč1,2, Hong Ching Lee1, Paul J. Berkman1,2,3, Philipp Emanuel Bayer1,2, Paul Visendi1,2, Pradeep Ruperao1,2,4, Timothy L. Fitzgerald5, Manuel Zander1,2, Chon-Kit Kenneth Chan1, Sahana Manoli1, Jiri Stiller1,3, Jacqueline Batley1,5 and David Edwards1,2,5,*

1 School of Agriculture and Food Sciences, University of Queensland, Brisbane, Qld, Australia
2 Australian Centre for Plant Functional Genomics, University of Queensland, Brisbane, Qld, Australia
3 CSIRO Agriculture Flagship, Queensland Bioscience Precinct, St Lucia, Qld, Australia
4 International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, India
5 School of Plant Biology, University of Western Australia, WA, Australia

Received 5 January 2014; revised 7 July 2014; accepted 13 July 2014.
*Correspondence (Tel +61 7 3346 7084; fax +61 7 3365 1176; email Dave.Edwards@uq.edu.au)

Keywords: Triticum aestivum, diversity, single nucleotide polymorphisms, evolution.

Introduction

Wheat is a major food crop, ranked within the top four agricultural commodities globally by production and value according to the Food and Agricultural Organization of the United Nations (FAO; http://www.fao.org/home/en/), and used widely for making products including breads, pastries, noodles and dumplings. Substantial cultivation of tetraploid wheat occurs, primarily ‘durum’ wheat (Triticum durum). However, greater than 90% of the world’s cultivated wheat is the hexaploid species Triticum aestivum (Shewry, 2009), known as ‘common’ or ‘bread’ wheat. The hexaploid genome of bread wheat formed through two allopolyploidization events. Between 0.5 and 3 MYA, the diploid genomes of Triticum urartu (AuAu) and an unidentified species (BB) similar to Aegilops speltoides combined to produce the allotetraploid genome of wild emmer wheat or Triticum turgidum (AuAuBB) (Chantret et al., 2005; Eckardt, 2001; Huang et al., 2002). Approximately 8000 years ago, probably in a region close to the Caspian Sea, a second event combined the genomes of T. turgidum (AuAuBB) and Aegilops tauschii (DD), producing the allohexaploid T. aestivum genome (AuAuBDD) (McFadden and Sears, 1946). Genome analysis in bread wheat poses substantial challenges; in addition to the complexity associated with its hexaploid structure, the bread wheat genome is very large (~17 Gb; around 40 times the size of rice or nearly six times larger than the human genome) and consists of about 80–90% repetitive sequence (Šafář et al., 2010; Wanjugi et al., 2009).

Genome mapping using molecular markers has played a central role in genetics since the 1980s (Schlotterer, 2004), revolutionizing fundamental research approaches such as the definition of haplotypes, the discovery of genomic regions associated with specific traits and the assessment of evolutionary relationships between organisms. In addition to being critical for research in cereal crops such as wheat, molecular markers play a crucial role in modern cereal breeding (Duran et al., 2009b; Rafalski, 2002). For example, genotyping using molecular markers facilitates accurate identification and maintenance of genetic stocks and guides the development of genetically diverse populations for selection programs. In some instances, traditional marker-assisted selection, wherein selection for a specific trait is guided using a marker or markers that accurately predict the inheritance of that trait, has enabled rapid incorporation of favourable alleles into elite cereal cultivars (Xu and Crouch, 2008). Furthermore, an increase in the number of markers available for cereal crops and a decrease in the cost of genotyping are beginning to enable new approaches including genome-wide association studies (GWAS) (Rosenberg et al., 2010; Schlotterer, 2004; Tian et al., 2011) and genomic selection (Heffner et al., 2009; Poland et al., 2012) in cereals, with impressive results. Single nucleotide polymorphisms (SNPs) represent the most...
frequent type of genetic polymorphism and can therefore allow the development of the highest density of molecular markers (Batley and Edwards, 2007). Powerful next-generation sequencing (NGS) technologies provide the possibility of large-scale SNP discovery by comparing whole-genome shotgun sequences of individuals with high-quality reference genome sequences (Edwards et al., 2012a, 2013; Imelfort et al., 2009).

Expressed genes have been a traditional source of data for SNP discovery. AutoSNP (Barker et al., 2003; Batley et al., 2003) and the associated autoSNPdb (Duran et al., 2009a) are tools for this purpose and use redundancy and haplotype co-segregation to distinguish true polymorphism from sequence error. The large data volumes from NGS platforms provide the potential to discover very large numbers of SNPs both in expressed sequences and elsewhere throughout the genome (Visendi et al., 2013). For example, Lai et al. (2010) identified more than 1 million SNPs between six inbred maize lines; furthermore, the authors were able to detect a large number of presence/absence variations (PAVs) and suggested that this phenomenon may contribute to heterosis in this species.

High-throughput SNP discovery from NGS data has recently been applied to identify SNPs between two accessions of the diploid wheat genome progenitor Ae. tauschii. For this purpose, an ‘annotation-based genome-wide SNP discovery pipeline’ (AGSNP) (You et al., 2011) was developed to facilitate SNP discovery from species with large and complex genomes. Using this pipeline, the authors combined data from Roche 454 sequencing of Ae. tauschii accession ALB/78, with 454, Applied Biosystems SOLiD, and Illumina sequencing to of genomic DNA and cDNA from Ae. tauschii accession AS75, to identify a total of 497 118 candidate SNPs (You et al., 2011). In hexaploid wheat, Allen et al. (Allen et al., 2011) identified 14,078 putative SNPs in 6255 distinct reference sequences via de novo assembly of Illumina GAIIx cDNA sequence data from wheat lines Avalon, Cadenza, Rialto, Savannah and Recital, supplemented with publically available EST sequences. The authors obtained a validation rate of 67% for a subset of 1659 of these markers.

More recently, Allen et al. used targeted resequencing of the wheat exome to generate large amounts of genomic sequences from 8 bread wheat varieties and identified 95 266 putative SNPs (Allen et al., 2013), and of these, 10 251 were predicted to be genome-specific putative co-dominant SNP markers with a validation accuracy of 96%. A 9K Illumina Infinium SNP array was recently constructed and used to genotype a total of 8630 SNPs, with a validation accuracy of between 65% (Wüschem et al., 2013) and 90% (Cavanagh et al., 2013), with many of these SNPs contributing to a new 90K Illumina Infinium array (Wang et al., 2014). Furthermore, the autoSNPdb pipeline described above has recently been applied to discover 38 928 candidate SNPs from 4 951 141 reads of wheat 454 transcriptome data (Lai et al., 2012b). SGsautoSNP (second-generation sequencing autoSNP) is an additional SNP discovery pipeline designed specifically to predict SNPs from whole-genome Illumina shotgun sequence data. SGsautoSNP has recently been applied to identify more than 800 000 SNPs between four varieties of bread wheat with accuracy greater than 93% (Lorenc et al., 2012), using the wheat group 7 isolated chromosome arm assemblies as a reference (Berkman et al., 2011, 2012, 2013).

A large national initiative was established in Australia in 2010 to coordinate diverse wheat genetic and genomic activities and establish a resource for Australian crop improvement (Edwards et al., 2012b). This led to the production of whole-genome shotgun sequence data for 16 diverse Australian bread wheat varieties. In this study, we have discovered more than 4 million candidate intervarietal SNPs across the wheat group 7 chromosomes from these data, using the SGsautoSNP pipeline (Lorenc et al., 2012). This abundance of SNPs has permitted an assessment of SNP density variation across the length of these chromosomes and a comparison of homoeologous chromosomes representing the A, B and D genomes of wheat. Our results demonstrate the impact of evolution and breeding on bread wheat genome diversity and provide a valuable resource for the further characterization and improvement of this important crop.

Table 1 Subgenomic varietal SNP density for 16 Australian wheat cultivars

<table>
<thead>
<tr>
<th>Total</th>
<th>Syntenic build</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. SNPs</td>
</tr>
<tr>
<td>7A</td>
<td>1 486 040</td>
</tr>
<tr>
<td>7B</td>
<td>1 860 295</td>
</tr>
<tr>
<td>7D</td>
<td>671 976</td>
</tr>
</tbody>
</table>

Results

Whole-genome Illumina paired read sequence data were generated from 16 Australian bread wheat varieties (Berkman et al., 2012). After filtering to remove poor quality and clonal reads, a total of 13 642 million read pairs remained. Alignment of these read pairs to the wheat group 7 and 4AL chromosome assemblies (Berkman et al., 2011, 2013; Hernandez et al., 2012; Lorenc et al., 2012) using strict parameters resulted in 3.05%, 3.76% and 3.43% of read pairs mapping uniquely to chromosomes 7A, 7B and 7D, respectively. SNP calling using the SGsautoSNP pipeline (Lorenc et al., 2012) predicted a total of 4 018 311 intervarietal SNPs.

The majority of SNPs were identified on contigs which do not form part of the syntenic builds and are predominantly within intergenic regions, and a substantially greater number of SNPs were predicted on chromosomes 7A and 7B, compared to 7D (Table 1). Additionally, the SNP transition/transversion ratio (Tr/Tv) was determined for each of the three chromosomes. The average Tr/Tv ratios within the A and the B genomes were found to be significantly higher than those observed for the D genome (Figures S1 and S2; Table S1).

An intervarietal SNP matrix was constructed, which represents SNPs between each pair of the 16 Australian wheat varieties (Table 2). SNPs between varieties varied from 146 171 between Chara and Baxter to 968 088 between Chara and Yitpi. The average number of SNPs between varieties was 465 278, and the majority of pairwise wheat combinations (117 of 120) featured more than 200 000 SNPs. This matrix was used to produce a phylogenetic tree representing similarity between the 16 varieties (Figure 1).

In addition to SNP density variation between the chromosomes, SNP density also varied along the lengths of chromosome syntenic builds (Figure S3). To assess whether this variation is associated with selection for genes exhibiting specific characteristics, SNP density was calculated in regions 2 Kbp upstream and downstream of each predicted gene. A total of 146 genes were predicted to be in low-SNP-density regions, representing 40, 27 and 79 genes on the A, B and D genomes, respectively (Table S2).
These genes include MADS box and Myb transcription factors, signal transduction pathway genes, a sodium transporter, an iron-responsive transcription factor, a potassium transporter, callose synthase, sucrose synthase and sugar transporters. In contrast, a total of 14 genes were predicted to be in high-SNP-density regions, representing 10, 3 and 1 gene(s) on the A, B and D genomes, respectively (Table S3). These genes include cellulose synthase, argonaute and ethylene response factors.

Twenty-two candidate SNPs were amplified by PCR and Sanger-sequenced to assess the false discovery rate associated with the approach used in this study. SNPs were chosen to represent all three of the group 7 chromosomes, including syntenic builds and unplaced contigs and reflected a range of redundancy scores. Of the 22 SNPs, one assay failed to amplify a PCR product; of the 21 which amplified successfully, all were shown to be true intervarietal polymorphisms (Table 3). The validated SNPs had an average redundancy score of 23.6 (range 4–84); in contrast, the SNP which failed to amplify had a redundancy score of 2.

The SNPs from the recently published wheat Infinium array (Wang et al., 2014) were compared to those predicted by SGsAuoToSNP. A total of 850 SNPs were identified as having a match on the group 7 chromosomes at the same position as predicted in our study (Table S4). Of these, 482 (57%) were classified as polymorphic single locus, 316 (37%) as being polymorphic multilocus, while only 52 (6%) were monomorphic.

Discussion

We have identified more than four million candidate intervarietal SNPs across the group 7 chromosomes between 16 Australian bread wheat varieties. This represents the greatest number of SNPs identified to date for this important crop. By resequencing 22 loci in different varieties, we obtained a SNP validation rate of 95%, and comparison of SNPs with results from the recently published wheat Infinium study (Wang et al., 2014) shows that 94% of SNPs identified in both studies were polymorphic. This compares to an overall polymorphism rate across the Infinium assay of only 69% (Wang et al., 2014). Our results are similar to the 93% of SNPs we observed in a previous study examining four varieties (Lorenc et al., 2012). This is also similar to study in the diploid D genome of Ae. tauschii, which validated over 80% of predicted SNPs (You et al., 2011), and the recent 96% validation of 10 251 putative co-dominant SNP markers in bread wheat (Allen et al., 2013). This is significantly higher than the validation of SNPs on the 9K Illumina Infinium array where 65% of SNPs demonstrated accurate genotype calling (Würschum et al., 2013), although this rises to 90% across more diverse germplasm and following manual data curation (Cavanagh et al., 2013). Substantial variation in pairwise SNP numbers between varieties was observed with the greatest polymorphism identified between Chara and Yitpi and the least polymorphism identified between Chara and Baxter (Table 2). Understanding the level of genomic diversity in populations can facilitate breeding and selection, ensuring that crosses lead to progeny with high levels of sequence diversity for the mapping of segregating traits. The phylogenetic tree produced based on pairwise SNP similarity (Figure 1) reflects the known breeding history of these varieties (Berkman et al., 2012) and may assist the selection of varieties representing a high degree of diversity within this set, most suitable for the development of high-resolution genetic mapping populations.
The majority of SNPs were identified outside of the syntenic builds. The syntenic builds reflect gene containing contigs which display similarity with genes from syntenic regions of related species and represent only 4% of the total assembly. SNP densities on chromosomes 7A, 7B, and 7D were approximately 4077, 4737 and 1939 SNPs/Mb, respectively (Table 1). This
difference in SNP density is consistent with previous observations (Berkman et al., 2013; Chao et al., 2009) and reflects the early evolutionary history of this crop. In an evolutionary event believed to have occurred near the Caspian Sea around 8000 years ago, tetraploid emmer wheat crossed with wild D genome progenitor Ae. tauschii, to form the hexaploid species T. aestivum, which became common wheat (bread wheat) (Giles and Brown, 2006; Nesbitt and Samuel, 1998; Salamini et al., 2002); a greater number of genes for domestication traits are found on the A and B genomes which display favourable alleles. In contrast, genes in high-SNP-density regions may be associated with regions introgressed from related species. SNP density also varied across the lengths of the individual syntenic builds. Regions of low SNP density may reflect selection at loci associated with domestication or important agronomic traits, with a loss of diversity in and around genes which display transitions can thus be considered an ‘evolutionary footprint’ of polymorphism on the A and B genomes relative to the D genome in modern cultivated wheat, consistent with patterns of SNP diversity identified in this study.

In addition to the variation detected between chromosomes, SNP density also varied across the lengths of the individual syntenic builds. Regions of low SNP density may reflect selection at loci associated with domestication or important agronomic traits, with a loss of diversity in and around genes which display favourable alleles. In contrast, genes in high-SNP-density regions may be associated with regions introgressed from related species. To assess this, genes within low- and high-SNP-density regions were identified and analysed (Tables S2 and S3).

Recently, Cavanagh et al. found evidence for selection around a major ‘green revolution’ dwarfing gene Rht-B1 (Cavanagh et al., 2013). The genes identified here in low-SNP-density regions are good candidates for further assessment to explore possible contributions to desirable characteristics of cultivated wheat. It appears likely that assessment of SNP density around genes as performed in this study will identify alleles selected during breeding, some of which could be targets for further crop improvement. In contrast to the 146 predicted genes identified in low-SNP-density regions, 14 genes were identified in high-SNP-density regions. These may reflect natural variation in SNP density across the genome or may have been introgressed from other diverse lines or species leading to regions of high polymorphism in this population.

An additional observation made in this study was that chromosomes 7A and 7B feature a higher SNP transition/transversion ratio (Tr/Tv) than chromosome 7D (Figure S1 and S2). A relatively high frequency of C/T and A/G transitions has been observed in many species and is thought to be predominantly due to the tendency of methyl cytosine to mutate to uracil, which is then corrected to thymine (Couloumbre et al., 1978); transitions can thus be considered an ‘evolutionary footprint’ of methylation (Buckler and Holtsford, 1996). It has also previously been demonstrated that gene loss is greater in the A and B genomes than the D genome (Berkman et al., 2013; Pont et al., 2013). Genome-wide methylation and associated gene silencing (Bottley et al., 2006; Charmet, 2011) are immediate results of polyploidization (Feldman and Levy, 2009). It may be that the higher Tr/Tv ratio and frequency of gene loss observed in the A and B genomes are results of the additional polyploidy event involving these genomes compared to the D genome during the formation of hexaploid wheat.

Overall, this study has revealed a vast number of polymorphisms occurring within the chromosome 7 homoeologues of hexaploid wheat among elite Australian varieties. This resource is publically available to assist additional genetic analysis and breeding. Furthermore, observed patterns of SNPs across the homoeologous group 7 chromosomes have provided insight into the molecular consequences of the evolution and selection that resulted in modern hexaploid wheat.

Experimental procedures

SNP prediction

Whole-genome Illumina PE data for 16 Australian bread wheat varieties were downloaded from Bioplatforms website (https://downloads.bioplatforms.com/wheat_cultivars/), and clonal reads were removed using a custom Perl script. The remaining sequence for the 16 Australian wheat cultivars was mapped to the three group 7 wheat chromosome assemblies (Berkman et al., 2011, 2013; Lorenc et al., 2012) as well as an assembly of chromosome arm 4AL (Hernandez et al., 2012) using the alignment tool SOAP v2.21 (Li et al., 2009b) with default parameters, allowing up to 2 mismatches per read and only retaining read pairs mapping uniquely to the reference with parameter ‘-r 0’. Arm 4AL was included to prevent reads from the translocated 7BS/4AL region mapping to homoeologous locations on 7AS or 7DS. The resulting BAM files were merged using samtools v0.1.17 (r973:277) (Li et al., 2009a). SNP prediction was performed using SGSautoSNP (Lorenc et al., 2012), with output in snp format for subsequent analysis and gff format for presentation on a GBrowse genome viewer at www.wheatgenome.info (Lai et al., 2012a).

SNP matrix production and transition/transversion ratio analysis

The snp files generated by SGSautoSNP were parsed using a custom Python script to generate the SNP matrix file (Table 2). The SNP matrix was subsequently converted to Newick format (Simonsen et al., 2008), and a phylogenetic tree was constructed using the Philodendron web-based application (Gilbert, 1999). The transition/transversion ratio for each chromosome was calculated based on bins of 500 SNPs using VCFTools (Danecek et al., 2011).

SNP density and gene analysis

The SNP density plots for each chromosome were generated using a custom Python script that calculates relative density based on a window size of 50 000 bp. Subsequent analysis was conducted to identify genes in low-SNP-density regions, defined as those for which SNP density in the regions 2 kbp upstream and downstream was significantly lower than the mean for all genes on the chromosome.

Genes identified as being in low-SNP-density regions were compared with the Swissprot database (Release 2013.06) using BLASTX (BLASTALL 2.2.6) (Altschul et al., 1990) with an E-value cut-off 1e-5. The genes with minimum E-value has been identified in low/high SNP density regions with UniProtKB entry ID and protein names.

Validation

A total of 22 SNPs were selected from the three group 7 reference genomes for validation. These SNPs had a range of redundancy scores. Genomic DNA was isolated from 11 cultivars,


Supporting information

Additional Supporting information may be found in the online version of this article:

**Figure S1** The transition/transversion ratio and standard deviation for chromosomes 7A, 7B and 7D.

**Figure S2** Ts/Tv ratio across the 7A, 7B and 7D syntenic builds.

**Figure S3** SNP density across the 7A, 7B and 7D syntenic builds.

**Table S1** Transition and transversion SNPs for each variety on the 7A, 7B and 7D chromosomes.

**Table S2** Genes identified in low SNP density regions on chromosomes 7A, 7B and 7D.

**Table S3** Genes identified in high SNP density regions on chromosomes 7A, 7B and 7D.

**Table S4** SNPs from 90K SNP Array matched SNPs from 16 Australian Wheat.