Plant Biotechnology Journal

Plant Biotechnology Journal (2015) **13**, pp. 97–104 doi: 10.1111/pbi.12240

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

Kaitao Lai[1,2], Michał T. Lorenc[1,2], Hong Ching Lee[1], Paul J. Berkman[1,2,3], Philipp Emanuel Bayer[1,2], Paul Visendi[1,2], Pradeep Ruperao[1,2,4], Timothy L. Fitzgerald[3], Manuel Zander[1,2], Chon-Kit Kenneth Chan[1], Sahana Manoli[1], Jiri Stiller[1,3], Jacqueline Batley[1,5] and David Edwards[1,2,5,*]

[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.

Summary

Despite being a major international crop, our understanding of the wheat genome is relatively poor due to its large size and complexity. To gain a greater understanding of wheat genome diversity, we have identified single nucleotide polymorphisms between 16 Australian bread wheat varieties. Whole-genome shotgun Illumina paired read sequence data were mapped to the draft assemblies of chromosomes 7A, 7B and 7D to identify more than 4 million intervarietal SNPs. SNP density varied between the three genomes, with much greater density observed on the A and B genomes than the D genome. This variation may be a result of substantial gene flow from the tetraploid *Triticum turgidum*, which possesses A and B genomes, during early co-cultivation of tetraploid and hexaploid wheat. In addition, we examined SNP density variation along the chromosome syntenic builds and identified genes in low-density regions which may have been selected during domestication and breeding. This study highlights the impact of evolution and breeding on the bread wheat genome and provides a substantial resource for trait association and crop improvement. All SNP data are publically available on a generic genome browser GBrowse at www.wheatgenome.info.

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 (AuAuBBDD) (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 AL8/78, with 454, Applied Biosystems SOLiD, and Illumina sequencing 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% (Wüschum *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 4 694 141 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.

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).

Table 1 Subgenomic varietal SNP density for 16 Australian wheat cultivars

	Total		Syntenic build	
	No. SNPs	SNPs/Mb	No. SNPs	SNPs/Mb
7A	1 486 040	4077	42 041	3212
7B	1 860 295	4737	38 508	3384
7D	671 976	1939	20 563	1088

Table 2 Pairwise intervarietal SNP matrix for chromosomes 7A, 7B and 7D between 16 Australian wheat varieties

AC Barrie	0															
Alsen	194 725	0														
Baxter	328 294	246 218	0													
Chara	592 193	438 075	146 171	0												
Drysdale	429 530	319 401	392 632	730 606	0											
Excalibur	346 557	273 217	324 087	567 179	367 279	0										
Gladius	529 898	327 659	472 457	906 611	616 253	491 885	0									
H45	385 753	265 113	339 227	627 589	298 414	280 576	519 690	0								
Kukri	245 356	208 666	290 506	541 524	428 134	318 029	480 575	345 358	0							
Pastor	302 731	289 053	340 269	603 323	336 029	284 559	552 119	309 025	302 231	0						
RAC875	412 818	257 630	390 967	722 089	429 038	368 152	158 973	386 145	418 037	375 137	0					
VolcaniDDI	508 175	413 676	412 553	808 658	696 467	600 478	813 067	633 916	498 017	586 694	643 205	0				
Westonia	354 599	276 490	310 192	623 591	500 461	362 800	557 464	405 842	346 683	349 542	403 411	678 631	0			
Wyalkatchem	525 289	341 043	433 228	800 300	560 759	327 888	386 213	449 614	436 777	442 941	235 924	800 137	505 345	0		
Xiaoyan 54	458 214	332 986	368 604	761 864	540 264	324 881	696 677	377 053	401 191	413 462	522 021	897 807	622 449	569 223	0	
Yitpi	544 440	328 216	468 743	968 088	690 017	548 694	233 539	587 310	530 687	580 060	287 648	951 537	654 967	444 084	844 785	0
	AC Barrie	Alsen	Baxter	Chara	Drysdale	Excalibur	Gladius	H45	Kukri	Pastor	RAC875	VolcaniDDI	Westonia	Wyalkatchem	Xiaoyan 54	Yitpi

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 SGSautoSNP. 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.

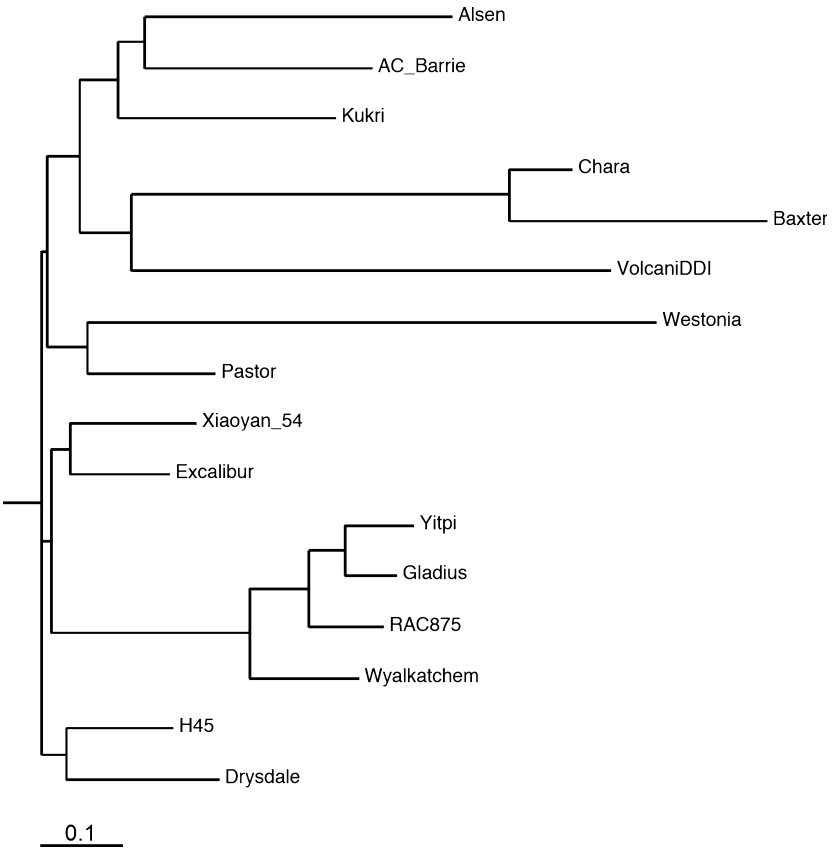

Figure 1 Phylogenetic relationships of 16 Australian wheat varieties based on SNP data obtained in this study.

Table 3 Summary of single nucleotide polymorphism validation

SNP ID	Forward primer	Reverse primer	Redundancy score	Chr.	SNP	Validation
UQ01TA7A495714	AGGTGTTTGTTCTTCACCGT	AAGGATCTTTGTGAAGTGGC	5	7A	A/G	True SNP
UQ01TA7A1381138	CATAACCGCTTCCTTTGTTT	ATCGGTAGACCTGCTCTTTG	36	7A	A/G	True SNP
UQ01TA7A781968	CAGATGAAGGCAGCAGTATG	TTTTCGTGACTACATCCGTG	23	7A	G/A	True SNP
UQ01TA7A14292	TTCTTATGTCGTGTTGTGCC	AAAAAGGACACGAAGAGGAA	28	7A	C/T	True SNP
UQ01TA7A19199	CCTACGCTTTAGCCACTGAT	CTCCCTTACAATGAACCAGC	4	7A	T/C	True SNP
UQ01TA7A19247	TAGGGATTTTGCATGGATTT	CCAACTTGTCGTCGTCATTA	16	7A	A/G	True SNP
UQ01TA7A04421	GGCGAGCTGACAATAAGTTT	TGTTTTGCAAATGAATGCTT	13	7A	T/C	True SNP
UQ01TA7A16656	ACAACCTCAGGTGAGAGAGC	TTGCCTGTCATGTCGATTAC	5	7A	C/T	True SNP
UQ01TA7A30454	CCATCATCATTGGAACAGAA	GATCAGATGTGGAAGAAGCC	84	7A	C/T	True SNP
UQ01TA7B299842	TTTTATCAGGCTAGTGGGGA	TGTCGTTGTGTAGGTATCCG	2	7B	C/T	Failed
UQ01TA7B64149	GTTGCATTCATCTTCGACAA	TGCTGGATCTTTGACTTGAA	25	7B	C/T	True SNP
UQ01TA7B588806	GTGCCCAGTTTTTCCATAAC	GTGACGGACTTGGAGAAGAC	12	7B	C/T	True SNP
UQ01TA7B1552504	GATGATCCTCGAAAAGGAAA	AAATAGTGGCCTTCATTCCA	4	7B	C/A	True SNP
UQ01TA7B1734345	TGCAAATGACATGCACATAA	TGCTAATGAGATGAAGAGCG	29	7B	G/A	True SNP
UQ01TA7B03509	AATGGGGATATTGTTTCGTG	ATGTCCTGGAGCTTTTTCAG	51	7B	A/G	True SNP
UQ01TA7B03075	TGGAATCATGTGATGTTGGT	GATATCCGTCCTCCATTCTG	50	7B	C/T	True SNP
UQ01TA7D349608	GAAAGAAGCGAATACCCAGA	GTCAAACTGATCCCAAGGAG	14	7D	T/C	True SNP
UQ01TA7D523654	GGGCTAAAAGAAATGGTCAA	CGAGATAATAGCCAGAGGGA	20	7D	A/C	True SNP
UQ01TA7D19459	GCCAGTGGAAGAAGAGTCAT	ACTTCCAGGTGTGTTTTGGT	25	7D	T/C	True SNP
UQ01TA7D09646	CGTGCTGGATAACTGTCTTG	GATCCCGTTTACCAAATGAC	22	7D	C/G	True SNP
UQ01TA7D05992	AGGGCAACATTTGTCTTCAT	GCAAGCTACGACATCTTTGA	4	7D	G/T	True SNP
UQ01TA7D12121	GGTCAGTTCTTTTGATGGCT	CGAAGAGAGTATTTTTCCGC	25	7D	C/T	True SNP

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 (Gegas *et al.*, 2010), consistent with domestication of the emmer wheat prior to the formation of the hexaploid. During the subsequent evolution of modern bread wheat, gene flow is predicted to have occurred between *T. aestivum* and its tetraploid progenitor *T. turgidum* (AuAuBB); however, no substantial gene flow is predicted to have occurred between the hexaploid and *Ae. tauschii* (DD) (Berkman *et al.*, 2013; Caldwell *et al.*, 2004; Dvorak *et al.*, 2006; Talbert *et al.*, 1998). This would be expected to result in a substantial increase in 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 (Coulondre *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 predication 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,

Alsen, Chara, Drysdale, Excalibur, Gladius, H45, Kukri, RAC875, VolcaniDDI, Xiaoyan 54 and Yitpi, according to a protocol adapted from Fulton *et al.* (Fulton *et al.*, 1995). PCR amplification of the 22 loci was performed using primers designed to bind to conserved sequence surrounding the SNPs (Table 3) in a 25 μL reaction volume containing 1 × iTaq PCR buffer (containing 100 mM Tris–HCl and 500 mM KCl, pH 8.3) (Bio-Rad, Hercules, CA), 200 μM each dNTP (Bio-Rad), 0.5 μM each primer, 1.5 U iTaq DNA polymerase (Bio-Rad), RNase- and DNase-free water (Gibco; Life Technologies, Carlsbad, CA) and 5 ng DNA. Thermocycling conditions for the reaction were 94 °C for 2 min, followed by 35 cycles of 94 °C for 30 s, annealing for 1 min at 59–62 °C and extension for 1 min at 72 °C. Final extension was performed at 72 °C for 10 min. Gel electrophoresis was performed on a 1% (w/v) agarose gel in 1 × TAE buffer (Sambrook and Russel, 2001) containing ethidium bromide-resolved products. The Australian Genome Research Facility's (AGRF) PD+ service was used to purify and subsequently sequence the PCR products. The purified PCR products were Sanger-sequenced using Big-Dye 3.1 (PerkinElmer, Waltham, MA), using forward and reverse PCR primers, and analysed using an ABI3730xl. The sequences for each locus and cultivar were aligned and compared using Geneious Pro v5.4.6 (Drummond *et al.*, 2011) with a cost matrix of 65%, a gap-open penalty of 6 and a gap-extension penalty of 3, and each of the alignments assessed visually to determine the SNP.

Flanking sequences for predicted SNPs from the recently published Infinium array (Wang *et al.*, 2014) were compared to the group 7 assemblies using WU-BLAST 2.0 (Gish, 1996-2006, http://blast.wustl.edu/) and an E-value of 1e-10. Where SNP positions matched predicted SNPs from SGSautoSNP, the Infinium SNP call was counted.

Acknowledgements

The authors would like to acknowledge funding support from Bioplatforms Australia and the Australian Research Council (projects LP0882095, LP0883462 and DP0985953). Support is also acknowledged from the Australian Genome Research Facility (AGRF), the Queensland Cyber Infrastructure Foundation (QCIF) and the Australian Partnership for Advanced Computing (APAC).

Conflict of Interest

None declared.

References

Allen, A.M., Barker, G.L.A., Berry, S.T., Coghill, J.A., Gwilliam, R., Kirby, S., Robinson, P., Brenchley, R.C., D'Amore, R., McKenzie, N., Waite, D., Hall, A., Bevan, M., Hall, N. and Edwards, K.J. (2011) Transcript-specific, single-nucleotide polymorphism discovery and linkage analysis in hexaploid bread wheat (*Triticum aestivum* L.). *Plant Biotechnol. J.* **9**, 1086–1099.

Allen, A.M., Barker, G.L., Wilkinson, P., Burridge, A., Winfield, M., Coghill, J., Uauy, C., Griffiths, S., Jack, P., Berry, S., Werner, P., Melichar, J.P., McDougall, J., Gwilliam, R., Robinson, P. and Edwards, K.J. (2013) Discovery and development of exome-based, co-dominant single nucleotide polymorphism markers in hexaploid wheat (*Triticum aestivum* L.). *Plant Biotechnol. J.* **11**, 279–295.

Altschul, S., Gish, W., Miller, W., Myers, E. and Lipman, D. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.

Barker, G., Batley, J., O'Sullivan, H., Edwards, K.J. and Edwards, D. (2003) Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. *Bioinformatics*, **19**, 421–422.

Batley, J. and Edwards, D. (2007) SNP applications in plants. In *Association Mapping in Plants* (Oraguzie, N., Rikkerink, E., Gardiner, S. and De Silva, H., eds), pp. 95–102. New York: Springer.

Batley, J., Barker, G., O'Sullivan, H., Edwards, K.J. and Edwards, D. (2003) Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. *Plant Physiol.* **132**, 84–91.

Berkman, P.J., Manoli, S., McKenzie, M., Kubaláková, M., Šimková, H., Batley, J., Fleury, D., Doležel, J., Edwards, D., Skarshewski, A., Lorenc, M.T., Lai, K., Duran, C., Ling, E.Y.S., Stiller, J., Smits, L. and Imelfort, M. (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., Lorenc, M.T., Stiller, J., Smits, L., Lai, K., Campbell, E., Kubalakova, M., Simkova, H., Batley, J., Dolezel, J., Hernandez, P. and Edwards, D. (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., Stiller, J., Manoli, S., Lorenc, M.T., Lai, K., Batley, J., Fleury, D., Simkova, H., Kubalakova, M., Weining, S., Dolezel, J. and Edwards, D. (2013) Dispersion and domestication shaped the genome of bread wheat. *Plant Biotechnol. J.* **11**, 564–571.

Bottley, A., Xia, G.M. and Koebner, R.M.D. (2006) Homoeologous gene silencing in hexaploid wheat. *Plant. J.* **47**, 897–906.

Buckler, E.S. and Holtsford, T.P. (1996) Zea ribosomal repeat evolution and substitution patterns. *Mol. Biol. Evol.* **13**, 623–632.

Caldwell, K.S., Dvorak, J., Lagudah, E.S., Akhunov, E., Luo, M.C., Wolters, P. and Powell, W. (2004) Sequence polymorphism in polyploid wheat and their D-genome diploid ancestor. *Genetics*, **167**, 941–947.

Cavanagh, C.R., Chao, S., Wang, S., Huang, B.E., Stephen, S., Kiani, S., Forrest, K., Saintenac, C., Brown-Guedira, G.L., Akhunova, A., See, D., Bai, G., Pumphrey, M., Tomar, L., Wong, D., Kong, S., Reynolds, M., da Silva, M.L., Bockelman, H., Talbert, L., Anderson, J.A., Dreisigacker, S., Baenziger, S., Carter, A., Korzun, V., Morrell, P.L., Dubcovsky, J., Morell, M.K., Sorrells, M.E., Hayden, M.J. and Akhunov, E. (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., Rahman, S., Bellec, A., Laubin, B., Dubois, I., Dossat, C., Sourdille, P., Joudrier, P., Gautier, M.F., Cattolico, L., Beckert, M., Aubourg, S., Weissenbach, J., Caboche, M., Bernard, M., Leroy, P. and Chalhoub, B. (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.

Chao, S., Zhang, W., Akhunov, E., Sherman, J., Ma, Y., Luo, M.-C. and Dubcovsky, J. (2009) Analysis of gene-derived SNP marker polymorphism in US wheat (*Triticum aestivum* L.) cultivars. *Mol. Breeding*, **23**, 23–33.

Charmet, G. (2011) Wheat domestication: lessons for the future. *C. R. Biol.* **334**, 212–220.

Coulondre, C., Miller, J.H., Farabaugh, P.J. and Gilbert, W. (1978) Molecular-basis of base substitution hotspots in *Escherichia coli*. *Nature*, **274**, 775–780.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G. and Durbin, R. and 1000 Genomes Project Analysis Group (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

Drummond, A.J., Ashton, B.S.B., Cheung, M., Cooper, A., Duran, C., Field, M., Heled, J., Kearse, M., Markowitz, S., Moir, R., Stones-Havas, S., Sturrock, S., Thierer, T. and Wilson, A. (2011) *Geneious v5.4.* http://www.geneious.com.

Duran, C., Appleby, N., Clark, T., Wood, D., Imelfort, M., Batley, J. and Edwards, D. (2009a) AutoSNPdb: an annotated single nucleotide polymorphism database for crop plants. *Nucleic Acids Res.* **37**, D951–D953.

Duran, C., Appleby, N., Edwards, D. and Batley, J. (2009b) Molecular genetic markers: discovery, applications, data storage and visualisation. *Curr. Bioinform.* **4**, 16–27.

Dvorak, J., Akhunov, E.D., Akhunov, A.R., Deal, K.R. and Luo, M.C. (2006) Molecular characterization of a diagnostic DNA marker for domesticated tetraploid wheat provides evidence for gene flow from wild tetraploid wheat to hexaploid wheat. *Mol. Biol. Evol.* **23**, 1386–1396.

Eckardt, N.A. (2001) A sense of self: the role of DNA sequence elimination in allopolyploidization. *Plant Cell*, **13**, 1699–1704.

Edwards, D., Henry, R.J. and Edwards, K.J. (2012a) Preface: advances in DNA sequencing accelerating plant biotechnology. *Plant Biotechnol. J.* **10**, 621–622.

Edwards, D., Wilcox, S., Barrero, R.A., Fleury, D., Cavanagh, C.R., Forrest, K.L., Hayden, M.J., Moolhuijzen, P., Keeble-Gagnère, G., Bellgard, M.I., Lorenc, M.T., Shang, C.A., Baumann, U., Taylor, J.M., Morell, M.K., Langridge, P., Appels, R. and Fitzgerald, A. (2012b) 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. (2013) Accessing complex crop genomes with next-generation sequencing. *Theor. Appl. Genet.* **126**, 1–11.

Feldman, M. and Levy, A.A. (2009) Genome evolution in allopolyploid wheat—a revolutionary reprogramming followed by gradual changes. *J. Genet. Genomics*, **36**, 511–518.

Fulton, T., Chunwongse, J. and Tanksley, S. (1995) Microprep protocol for extraction of DNA from tomato and other herbaceous plants. *Plant Mol. Biol. Report.*, **13**, 207–209.

Gegas, V.C., Nazari, A., Griffiths, S., Simmonds, J., Fish, L., Orford, S., Sayers, L., Doonan, J.H. and Snape, J.W. (2010) A genetic framework for grain size and shape variation in wheat. *Plant Cell*, **22**, 1046–1056.

Gilbert, D.G. (1999) *Phylodendron, an application for drawing phylogenetic trees*.

Giles, R. and Brown, T. (2006) GluDy allele variations in *Aegilops tauschii* and *Triticum aestivum*: implications for the origins of hexaploid wheats. *Theor. Appl. Genet.* **112**, 1563–1572.

Heffner, E.L., Sorrells, M.E. and Jannink, J.-L. (2009) Genomic selection for crop improvement. *Crop Sci.* **49**, 1–12.

Hernandez, P., Martis, M., Dorado, G., Pfeifer, M., Galvez, S., Schaaf, S., Jouve, N., Simkova, H., Valarik, M., Dolezel, J. and Mayer, K.F. (2012) Next-generation sequencing and syntenic integration of flow-sorted arms of wheat chromosome 4A exposes the chromosome structure and gene content. *Plant. J.* **69**, 377–386.

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.

Imelfort, M., Duran, C., Batley, J. and Edwards, D. (2009) Discovering genetic polymorphisms in next-generation sequencing data. *Plant Biotechnol. J.* **7**, 312–317.

Lai, J.S., Li, R.Q., Xu, X., Jin, W.W., Xu, M.L., Zhao, H.N., Xiang, Z.K., Song, W.B., Ying, K., Zhang, M., Jiao, Y.P., Ni, P.X., Zhang, J.G., Li, D., Guo, X.S., Ye, K.X., Jian, M., Wang, B., Zheng, H.S., Liang, H.Q., Zhang, X.Q., Wang, S.C., Chen, S.J., Li, J.S., Fu, Y., Springer, N.M., Yang, H.M., Wang, J.A., Dai, J.R., Schnable, P.S. and Wang, J. (2010) Genome-wide patterns of genetic variation among elite maize inbred lines. *Nat. Genet.* **42**, 1027–1058.

Lai, K., Berkman, P.J., Lorenc, M.T., Duran, C., Smits, L., Manoli, S., Stiller, J. and Edwards, D. (2012a) WheatGenome.info: an integrated database and portal for wheat genome information. *Plant Cell Physiol.* **53**, 1–7.

Lai, K., Duran, C., Berkman, P.J., Lorenc, M.T., Stiller, J., Manoli, S., Hayden, M.J., Forrest, K.L., Fleury, D., Baumann, U., Zander, M., Mason, A.S., Batley, J. and Edwards, D. (2012b) Single nucleotide polymorphism discovery from wheat next-generation sequence data. *Plant Biotechnol. J.* **10**, 743–749.

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

Li, R., Yu, C., Li, Y., Lam, T.W., Yiu, S.M., Kristiansen, K. and Wang, J. (2009b) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, **25**, 1966–1967.

Lorenc, M.T., Hayashi, S., Stiller, J., Lee, H., Manoli, S., Ruperao, P., Visendi, P., Berkman, P.J., Lai, K., Batley, J. and Edwards, D. (2012) Discovery of single nucleotide polymorphisms in complex genomes using SGSautoSNP. *Biology (Basel)*, **1**, 370–382.

McFadden, E. and Sears, E. (1946) The origin of *Triticum spelta* and its free-threshing hexaploid relatives. *J. Hered.* **37**, 81–89.

Nesbitt, M. and Samuel, D. (1998) Wheat domestication: Archaeobotanical evidence. *Science*, **279**, 1431.

Poland, J., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., Manes, Y., Dreisigacker, S., Crossa, J., Sánchez-Villeda, H., Sorrells, M. and Jannink, J.-L. (2012) Genomic selection in wheat breeding using genotyping-by-sequencing. *Plant Genome*, **5**, 103–113.

Pont, C., Murat, F., Guizard, S., Flores, R., Foucrier, S., Bidet, Y., Quraishi, U.M., Alaux, M., Doležel, J., Fahima, T., Budak, H., Keller, B., Salvi, S., Maccaferri, M., Steinbach, D., Feuillet, C., Quesneville, H. and Salse, J. (2013) Wheat syntenome unveils new evidences of contrasted evolutionary plasticity between paleo- and neoduplicated subgenomes. *Plant. J.* **76**, 1030–1044.

Rafalski, A. (2002) Applications of single nucleotide polymorphisms in crop genetics. *Curr. Opin. Plant Biol.* **5**, 94–100.

Rosenberg, N.A., Huang, L., Jewett, E.M., Szpiech, Z.A., Jankovic, I. and Boehnke, M. (2010) Genome-wide association studies in diverse populations. *Nat. Rev. Genet.* **11**, 356–366.

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.

Salamini, F., Ozkan, H., Brandolini, A., Schafer-Pregl, R. and Martin, W. (2002) Genetics and geography of wild cereal domestication in the near east. *Nat. Rev. Genet.* **3**, 429–441.

Sambrook, J. and Russel, D.W. (2001) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor, NY.: Cold Spring Harbor Laboratory Press.

Schlotterer, C. (2004) The evolution of molecular markers [mdash] just a matter of fashion? *Nat. Rev. Genet.* **5**, 63–69.

Shewry, P.R. (2009) Wheat. *J. Exp. Bot.* **60**, 1537–1553.

Simonsen, M., Mailund, T. and Pedersen, C.N.S. (2008) Rapid neighbour-joining. *Lect. Notes Bioinform.* **5251**, 113–122.

Talbert, L.E., Smith, L.Y. and Blake, M.K. (1998) More than one origin of hexaploid wheat is indicated by sequence comparison of low-copy DNA. *Genome*, **41**, 402–407.

Tian, F., Bradbury, P.J., Brown, P.J., Hung, H., Sun, Q., Flint-Garcia, S., Rocheford, T.R., McMullen, M.D., Holland, J.B. and Buckler, E.S. (2011) Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat. Genet.* **43**, 159–162.

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

Wang, S., Wong, D., Forrest, K., Allen, A., Chao, S., Huang, B.E., Maccaferri, M., Salvi, S., Milner, S.G., Cattivelli, L., Mastrangelo, A.M., Whan, A., Stephen, S., Barker, G., Wieseke, R. and Plieske, J., International Wheat Genome Sequencing Consortium, Lillemo, M., Mather, D., Appels, R., Dolferus, R., Brown-Guedira, G., Korol, A., Akhunova, A.R., Feuillet, C., Salse, J., Morgante, M., Pozniak, C., Luo, M.-C., Dvorak, J., Morell, M., Dubcovsky, J., Ganal, M., Tuberosa, R., Lawley, C., Mikoulitch, I., Cavanagh, C., Edwards, K.J., Hayden, M. and Akhunov, E. (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.X., Kianian, S.F., Luo, M.C., Wu, J.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.

Würschum, T., Langer, S.M., Longin, C.F., Korzun, V., Akhunov, E., Ebmeyer, E., Schachschneider, R., Schacht, J., Kazman, E. and Reif, J.C. (2013) Population structure, genetic diversity and linkage disequilibrium in elite winter wheat assessed with SNP and SSR markers. *Theor. Appl. Genet.* **126**, 1477–1486.

Xu, Y. and Crouch, J.H. (2008) Marker-assisted selection in plant breeding: from publications to practice. *Crop Sci.* **48**, 391–407.

You, F., Huo, N., Deal, K., Gu, Y., Luo, M.-C., McGuire, P., Dvorak, J. and Anderson, O. (2011) Annotation-based genome-wide SNP discovery in the large and complex Aegilops tauschii genome using next-generation sequencing without a reference genome sequence. *BMC Genomics*, **12**, 59.

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.