Title

A haplotype map of allohexaploid wheat reveals distinct patterns of selection on homoeologous genomes

Journal

Genome Biology, 16(1)

ISSN

1474-760X

Authors

Jordan, Katherine W

Wang, Shichen

Lun, Yanni

et al.

Publication Date

2015-12-01

DOI

10.1186/s13059-015-0606-4

Peer reviewed

Jordan et al. Genome Biology (2015) 16:48
DOI 10.1186/s13059-015-0606-4

Genome Biology

RESEARCH Open Access

A haplotype map of allohexaploid wheat reveals distinct patterns of selection on homoeologous genomes

Katherine W Jordan[1†], Shichen Wang[1†], Yanni Lun[1,2], Laura-Jayne Gardiner[3], Ron MacLachlan[4], Pierre Hucl[4], Krysta Wiebe[4], Debbie Wong[5], Kerrie L Forrest[5], IWGS Consortium, Andrew G Sharpe[6], Christine HD Sidebottom[6], Neil Hall[3], Christopher Toomajian[1], Timothy Close[7], Jorge Dubcovsky[8,9], Alina Akhunova[1,2], Luther Talbert[10], Urmil K Bansal[11], Harbans S Bariana[11], Matthew J Hayden[5], Curtis Pozniak[4], Jeffrey A Jeddeloh[12], Anthony Hall[3] and Eduard Akhunov[1*]

Abstract

Background: Bread wheat is an allopolyploid species with a large, highly repetitive genome. To investigate the impact of selection on variants distributed among homoeologous wheat genomes and to build a foundation for understanding genotype-phenotype relationships, we performed population-scale re-sequencing of a diverse panel of wheat lines.

Results: A sample of 62 diverse lines was re-sequenced using the whole exome capture and genotyping-by-sequencing approaches. We describe the allele frequency, functional significance, and chromosomal distribution of 1.57 million single nucleotide polymorphisms and 161,719 small indels. Our results suggest that duplicated homoeologous genes are under purifying selection. We find contrasting patterns of variation and inter-variant associations among wheat genomes; this, in addition to demographic factors, could be explained by differences in the effect of directional selection on duplicated homoeologs. Only a small fraction of the homoeologous regions harboring selected variants overlapped among the wheat genomes in any given wheat line. These selected regions are enriched for loci associated with agronomic traits detected in genome-wide association studies.

Conclusions: Evidence suggests that directional selection in allopolyploids rarely acted on multiple parallel advantageous mutations across homoeologous regions, likely indicating that a fitness benefit could be obtained by a mutation at any one of the homoeologs. Additional advantageous variants in other homoelogs probably either contributed little benefit, or were unavailable in populations subjected to directional selection. We hypothesize that allopolyploidy may have increased the likelihood of beneficial allele recovery by broadening the set of possible selection targets.

Background

Wheat genomic variation is shaped by the interplay of multiple factors including two recent polyploidization events [1-3] (Figure 1a), domestication [4], spread from the sites of origin to new geographic regions, gene flow from the populations of wild and domesticated ancestors [5], and post-domestication selection aimed at developing high-yielding locally adapted varieties. The eco-geographic habitats to which wheat is adapted span diverse environments ranging from low humidity regions in Nigeria, and the northern regions of Russia and Norway to the high-humidity regions of South America and Bangladesh [6]. It has been suggested that this broad adaptability likely results from the genetic diversity captured from the natural populations of its tetraploid ancestors [5,7] combined with a high rate of evolutionary changes in the wheat genome (particularly insertions and deletions), which are tolerated by its polyploid nature [8,9].

* Correspondence: eakhunov@ksu.edu
†Equal contributors
[1]Department Plant Pathology, Kansas State University, Manhattan, KS 66506, USA
Full list of author information is available at the end of the article

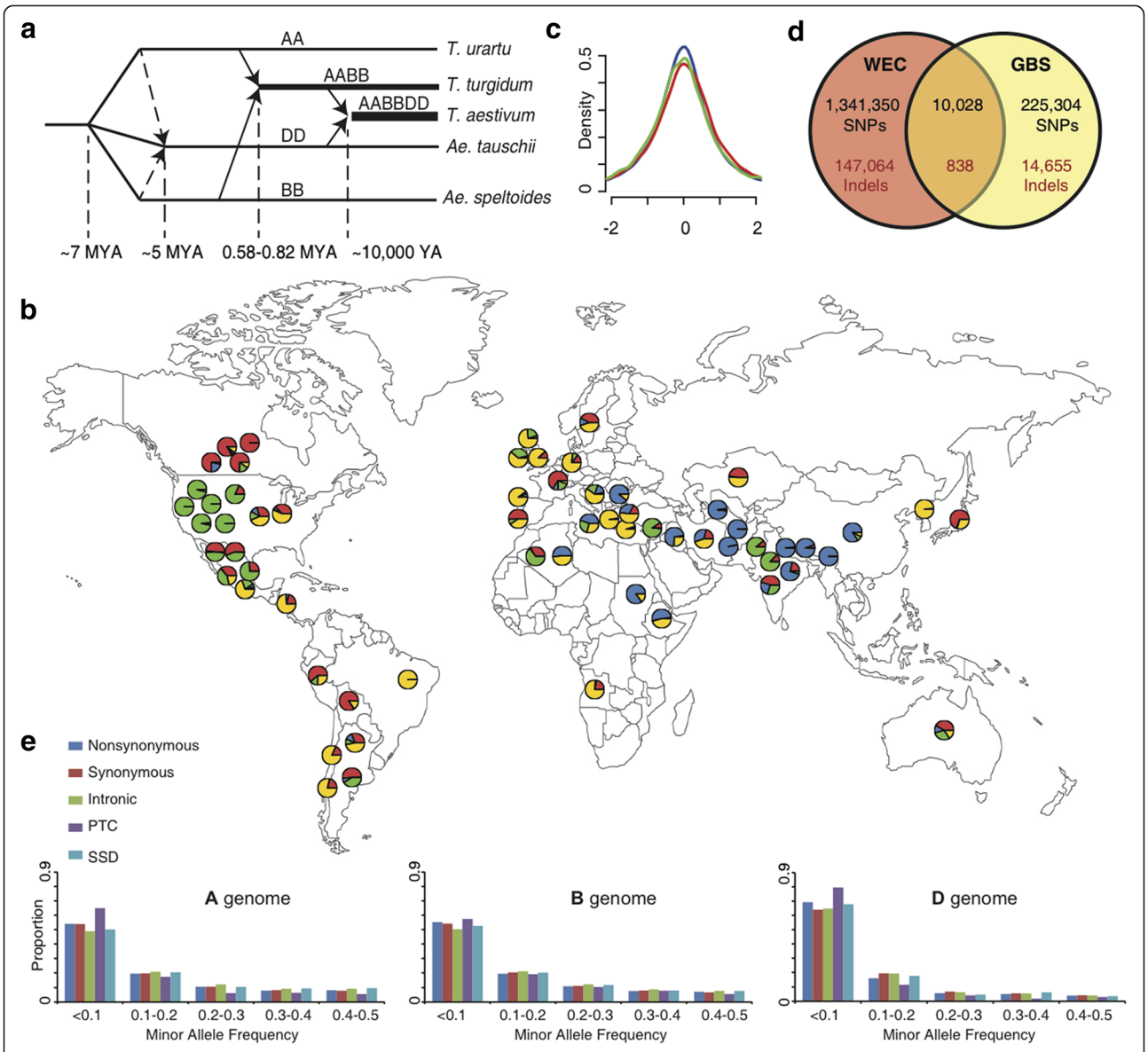

Figure 1 Summary of re-sequencing panel. (a) Evolution of the hexaploid wheat genome. The tetraploid wheat *T. turgidum* (AABB) originated by the hybridization of *T. urartu* with the close unidentified relative of *Ae. speltoides* occurred about 0.58 to 0.82 million years ago according to the genome-wide divergence time estimate [10]. The origin of hexaploid wheat occurred about 10,000 years ago [11] by the hybridization of *T. turgidum* with *Ae. tauschii* (DD) [12]. Marcussen et al. [10] suggested that *Ae. tauschii* might have originated by homoploid hybrid speciation (shown by dashed arrows). **(b)** Geographic distribution of 62 accessions of wheat accessions. Pie charts indicate the proportion of genetic ancestry for K = 4 inferred using Structure. **(c)** Efficiency of homoeologous gene capture. The depth of read coverage was extracted for each of the three copies of 47,739 homoeologous gene sets. The histogram of the $\log_2$ transformed ratio of read coverage between A and B (red), A and D (blue), and B and D (green) genomes was plotted. Each plot shows a normal distribution with the overall mean at 0. **(d)** Overlap between the SNP and indel datasets generated by WEC and GBS. **(e)** Minor allele frequency of different functional classes of SNPs as a proportion of total SNPs within each genome and class. PTC: premature termination codons; SSD: splice-site disruptions.

A detailed description of DNA sequence variation across the genome is a prerequisite for the systematic analysis of variants underlying trait variation in wheat and critical for understanding the role of various evolutionary factors in shaping genome diversity. Recently, low to medium density genotyping arrays were used to characterize SNP variation and linkage disequilibrium (LD) in wheat populations [13,14] and identify variants associated with phenotypic traits [15,16]. However, despite being a useful genotyping tool, these arrays are incapable of capturing the entire spectrum of DNA sequence variation and allele frequencies in wheat populations, and providing unbiased information that may help directly identify causal variants affecting phenotypes. Achieving this goal requires obtaining sequence data on a genome-scale from a diverse population of

lines. To date, this has only been performed on limited samples of wheat lines used to discover SNPs in the parental lines of mapping populations [17], or for SNP-based array design [18].

Genome sequencing of populations of individuals has been undertaken in a number of species including humans, *Arabidopsis*, and several crops [19-23] and helped to detect alleles contributing to phenotypic variation and adaptation. Despite recent advances in next-generation sequencing (NGS), performing similar analyses of genomic variation in wheat is substantially complicated by allopolyploidy and large genome size (approximately 17 Gb). However, sequencing of DNA samples subjected to complexity reduction by exome capture [18,24] and genotyping by sequencing [25,26] was shown to be an effective strategy to analyze the complex genomes. In addition, the recent release of the chromosome-specific wheat genome assemblies [27,28] can now help to alleviate the problems associated with variant calling in the allopolyploid genome, and allow us to describe the chromosomal distribution of variants and their potential effect on gene function.

Here we used the newly developed genome assembly of the cultivar Chinese Spring [27] based on flow-sorted chromosome survey sequence (CSS) contigs to create a diversity map of allopolyploid bread wheat. The data were generated by re-sequencing 62 diverse wheat lines using whole exome capture (WEC) and genotyping by sequencing (GBS) approaches. The panel of wheat lines was selected to capture the genetic diversity of the major global wheat growing regions and included landraces and cultivars (Figure 1b; Table S1 and Figure S1 in Additional file 1). We used the obtained data to describe the effect of genetic variation on gene function, gain insights into the effect of selection on duplicated genes, and explore the LD landscape in each of the three wheat sub-genomes to better understand the role of selection in shaping the genetic diversity of wheat.

Results and discussion

Re-sequencing the allopolyploid wheat genome

The WEC assay probes were designed with 107 Mb of non-redundant low-copy genic regions [29] targeting nearly 321 Mb of sequence in all three wheat genomes (Figure S2 in Additional file 1). The capture probes covered 78% of the 124,201 high-confidence protein-coding genes (at the 95% similarity threshold) in the CSS contigs [27]. The GBS approach generated sequence data primarily outside the genic regions. We produced roughly 4.7 billion paired-end reads (4.5 billion WEC reads and 0.2 billion GBS reads), and 62% of WEC reads and 51% of GBS reads uniquely mapped to the CSS contigs of the individual chromosomes (Figure S3 and Tables S2, S3 in Additional file 1), using alignment parameters optimized to separate reads from the different wheat genomes (Figures S4 and S5 in Additional file 1). In the WEC dataset, similar relative read coverage across homoeologous targets indicated that non-redundant capture probes are capable of recovering sequences from the different genomes with equal efficiency (Figure 1c, Table S4 in Additional file 1).

Variant calling was performed in the regions of the wheat genome covered by reads in >46 lines (>75%). We identified 1.57 million single nucleotide polymorphisms (SNPs) and 161,719 small insertions-deletions (indels) distributed across all 21 chromosomes, producing an average density of 1 variant every 175 bp (Figure 1d; Figure S6, Tables S5 and S6 in Additional file 1). Consistent with the previous estimates of genetic diversity [30], the A (649,522) and B (791,971) genomes contained about 2.5 times more variants than the D genome (286,880).

The overall genotype error rates for SNPs and indels assessed by comparing genotype calls generated for cultivar Chinese Spring with the CSS contigs of the same cultivar were 1.1% and 1.5%, respectively (for details see Materials and Methods). The error rate for rare SNPs and indels covered by >10 reads was 4.6% and 3.4%, respectively (Figure S7 in Additional file 1). The majority (77%) of variants in the GBS dataset were found in intergenic regions (Table S5 in Additional file 1), and only 4.3% of the variants overlapped with the WEC dataset (Figure 1d). This finding is supported by the *in silico* PstI digest of the CSS contigs, which showed that the regions targeted by the GBS cover only 6.8% of the regions targeted by the WEC.

Impact of purifying selection on genetic variation in the polyploid genome

One of the predicted consequences of whole genome duplication is functional redundancy that can result in the relaxation of purifying selection acting on duplicated copies of genes, thereby increasing the rate of accumulation of functional mutations. Previous studies suggested that polyploidy can result in the accelerated accumulation of premature termination codons in coding sequences [9] or an excess of non-synonymous changes in the polyploid lineage compared to the lineages of its diploid ancestors [27]. However, the ability of polyploid wheat to tolerate aneuploidy or large-scale deletions suggests that the duplicated homoeologs can be functional. It is not known whether this functionality is maintained by purifying selection or, as a consequence of redundancy and selection relaxation, subject to decay through the mutation process.

The wheat genome contains a number of variants that are predicted to impact gene function. We found 6,944 indels that could have a negative impact on gene function resulting from a predicted reading frame shift

(Table S7 in Additional file 1). The proportion of frame-shift indels, relative to in-frame indels, in the coding regions of the wheat genome (67%) was higher than that reported (57%) in the human genome [19]. There were twice as many indels with a length of a multiple of 3 located within the coding regions than within the introns or untranslated regions (Figure S8 in Additional file 1), indicative of purifying selection maintaining reading frame within the coding regions.

In coding sequences we identified 83,622 non-synonymous and 76,361 synonymous SNPs (Table S5 in Additional file 1). Based on high-confidence gene models in the CSS contigs, we determined that only 1,600 and 1,583 SNPs are predicted to produce premature termination codons (PTCs) and splice-site disruptions (SSDs), respectively, with a two- to three-fold lower incidence of functional mutations in the D genome than in the A and B genomes (Table S6 in Additional file 1). Out of the 6,230 genes that have homoeologous copies in the wheat genome and harbor coding sequence-disrupting mutations including frame-shift indels and PTCs, 4,870 (78%) have at least one intact homoeologous copy suggesting that the deleterious effects of these variants, if any, could be compensated. The compensatory potential of the duplicated homoeologous genes is consistent with a higher density of chemically induced mutations (five- to eight-fold) that can be achieved in wheat compared to other diploid plant species [31]. However, in spite of the presence of intact functional homoeologous copies of genes, we found a reduced number of non-synonymous, PTC, and SDS variants with a high derived allele frequency in the population (Figure 1e; Figure S9 in Additional file 1). This depletion of functional variants at higher allele frequencies is consistent with purifying selection acting against functional mutations, and suggests that the effect of purifying selection is not completely diminished by whole genome duplication. There are two plausible explanations for the retention of functional gene copies in young polyploids. First, selection acts on gene function that was partitioned among the homoeologs after the whole genome duplication. The functional partitioning is consistent with the studies of natural and artificial polyploids of wheat and other plants, often showing the tissue- and/or development-specific expression of homoeologs [32-35]. Alternatively, selection favors functional homoeologs to maintain the optimal stoichiometric ratios of gene products in macromolecular complexes, or in multistep regulatory cascades, which was proposed in the gene balance hypothesis [36].

The fraction of non-synonymous changes varied significantly among different protein families (Table S8 in Additional file 2) with the majority of PFAM domains involved in basic cellular functions showing a reduced proportion of non-synonymous changes compared to the genome-wide value, indicative of strong purifying selection. We detected a significant enrichment ($\chi 2$ test P value $<10^{-4}$, Table S8 in Additional file 2) for non-synonymous changes in the LRR and NB-ARC domains of disease resistance genes. The enrichment for major effect SNPs in these genes appears to be common for plant genomes and was also found in *Arabidopsis* [37] and peanut [38]. These observations are consistent with the hypothesis of an 'arms race' between the evolving populations of a pathogen and a plant defense system that results in fast evolution of genes with new disease-resistance specificities [39].

Global patterns of genetic variation

The global patterns of genomic variation and distribution of inter-variant associations are impacted by historic selection and demographic events, and by variation in recombination rate [40]. We found a non-random variant distribution along the chromosomes with reduced variation near the centromeres and elevated variation at the telomeres (Figure 2a and b; Figures S10-15 in Additional file 1), which is consistent with previous studies [28,30]. This pattern is similar to what was reported for maize and humans [19,41], but differs from *Arabidopsis* [37], where regions of high polymorphism were located near the centromeres. Our data also showed reduced diversity and an excess of rare alleles in the D genome when compared to the A and B genomes (Figure 2a; Table S9 in Additional file 1) [30]. These trends are consistent with the hypothesis that the limited number of ancestral genotypes of the D genome contributed to the origin of hexaploid wheat [42]. An elevated level of diversity in the A and B genomes, which otherwise would be expected to show the same levels of diversity as the D genome, could be attributed to the influx of allelic variation from the sympatric populations of wild tetraploid relatives [7,43].

Differentiation between landraces and cultivars (F_{ST}) varied along chromosomes with lower values found near the telomeres (Figure 2b; Figures S10-15 in Additional file 1). Long stretches of elevated F_{ST} were found along chromosome 4A and the short arm of chromosome 7B, two of the most structurally re-arranged chromosomes in the wheat genome [44] (Figure S16 in Additional file 1). Since these structural re-arrangements are fixed in wheat and, therefore, unlikely to affect gene flow between populations resulting in high F_{ST}, the overlap of differentiated genomic regions with those showing the signal of positive selection (Table S10 in Additional file 3) suggests that the detected differentiation could be associated with improvement selection.

Identification of genomic regions showing positive Tajima's D (excess of common alleles) and elevated diversity in the D genome (Figure 2b; Figures S10-15 in Additional file 1) is indicative of the presence of highly

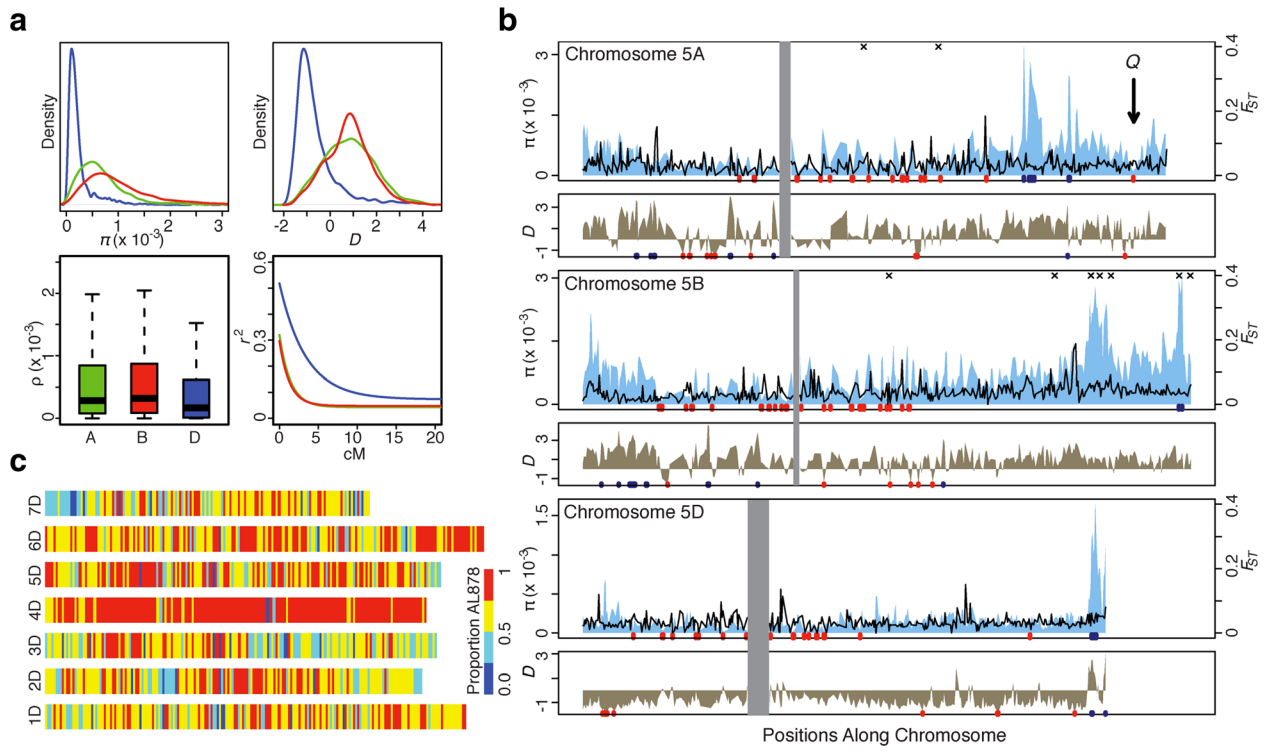

Figure 2 Distribution of genetic diversity, allele frequency, and recombination across the wheat genome. (a) Distribution of genetic diversity in the A (green), B (red), and D (blue) genomes: (π, top left), Tajima's measure of site frequency spectrum (D, top right), historic recombination (ρ, bottom left), and LD (bottom right). **(b)** Distribution of nucleotide diversity π (shaded polygon), F_{ST} between cultivars and landraces (solid black line), and site frequency spectrum (D) along chromosomes 5A (top panel), 5B (middle panel) and 5D (bottom panel). Gray shaded boxes represent the approximate location of the centromere. Rug plots represent lower (red) and upper (blue) 2.5% tails of test statistic distribution. Black X above the plot represents upper 2.5% tail of ρ statistic. The location of domestication (*Q*) gene is shown by arrow. **(c)**. Distribution of alleles of the AL8/78 genotype of *Ae. tauschii* along the chromosomes of the D genome in the 26 wheat landraces. The average frequency of AL8/78 alleles was calculated in a 3 Mb sliding window. The color scale shows the proportion of the AL8/78 alleles in each window (red - highest, blue - lowest).

diverged haplotypes in our sample most likely resulting from introgressions. To test this possibility, we compared the D genome haplotypes of 26 landraces not affected by modern breeding with the sequence of *Ae. tauschii* accession AL8/78 [45], which is considered the most closely related to the wheat D genome [46]. The high proportion of AL8/78 alleles (74%) and their distribution along the wheat chromosomes confirms the ancestry of the D genome (Figure 2c) [42]. However, the fine-scale haplotypic structure also reveals regions carrying highly divergent haplotypes (Figure 2c) suggestive of significant levels of introgression from either the diverged *Ae. tauschii* lines or independently originated hexaploid wheat lineages founded by diverged *Ae. tauschii* genotypes. The preferential localization of introgressions in the high-recombining regions of the chromosomes indicates that gene flow between the different D genome lineages was uneven along the chromosomes.

Using re-sequencing data, we now have the possibility to assess historic recombination rate (parameter ρ) [47], which is the product of meiotic recombination rate variation and effective population size. The median estimate of ρ in the D genome ($\rho = 1.7 \times 10^{-4}$/kb) was lower than in the A ($\rho = 2.9 \times 10^{-4}$/kb) and B ($\rho = 3.2 \times 10^{-4}$/kb) genomes (Figure 2a; Table S9 in Additional file 1). Consistent with the observations made in *Arabidopsis*, maize, and humans [40,41,48], we detected a positive correlation in the A (Spearman $r^2_{sp} = 0.23$, $P < 10^{-9}$) and B ($r^2_{sp} = 0.21$, $P < 10^{-12}$) genomes between the meiotic (R = genetic distance in cM / physical distance in Mb) and historic recombination rates suggesting stability of chromosomal recombination rates over time. However, no significant correlation was found between R and ρ in the D genome. This fact is most likely explained by the polyploidy-associated population bottleneck, which can impact the estimates of historic recombination [49] and also can result in reduced ρ in the D genome compared to that in the A and B genomes.

The historic recombination and diversity in wheat showed a positive correlation with relative distance from the centromere ($r^2_{sp} = 0.15$, $P < 10^{-4}$), a trend previously reported in maize, and humans [41,48]. These relationships

likely explain most of the inter-genomic diversity correlation along the homoeologous chromosomes (Table S11 in Additional file 1). Except for a few cases, we found low inter-genomic correlation of the window-based Tajima's D and F_{ST} estimates along the duplicated genomic regions of homoeologous chromosomes. This outcome is most likely a consequence of the sensitivity of these summary statistics to historic demographic and selection events that likely had different impacts on each wheat genome (Table S11 in Additional file 1).

Genotype imputation and GWAS

In genome-wide association studies (GWAS), marker density affects the probability of finding variants in linkage disequilibrium (LD) with a causal variant. A SNP-hiding test [40] showed that the probability of identifying high-LD SNPs ($r^2 > 0.8$) in our population within a 2-kb window was 70% to 71% for all three genomes. Consistent with the observed LD levels in the wheat genomes (Figure 2a), for SNPs located from 2 kb to 4 kb apart, the probability of finding high-LD SNPs in the D genome (56%) was higher than in the A and B genomes (48% to 49%).

Diversity maps have proven to be a powerful tool for imputing genotypes [19,20] allowing for an increase in marker density and the precision of trait mapping. Using a common set of SNP markers shared between the WEC data and the 90 K SNP assay [14] currently used by the community to genotype large numbers of wheat accessions, we tested the utility of our data for genotype imputation.

First, we sequentially selected each cultivar from our panel of 62 lines and, after 'hiding' all SNP sites besides those overlapping with the public 90 K SNP array [14], we used the WEC SNP data in the remaining 61 lines to predict the 'hidden' variants. Depending on the selected wheat line, using a genotype calling probability cutoff of 0.6, the accuracy of genotype predictions assessed by comparing with the observed data was in the range of 93% to 97% (Figure 3a; Table S12 in Additional file 4). This genotype probability cutoff value resulted in the removal of 5% to 15% of the data (Figure 3a), and allowed imputation of up to 549,918 SNPs. The accuracy of SNP imputation varied among the wheat genomes reflecting the inter-genomic differences in the extent of LD (Figure 2a). For example, the highest imputation accuracy was achieved for the D genome, which also showed the highest levels of inter-variant LD.

Second, we used our reference panel of 62 accessions to impute DNA polymorphisms in a GWAS. A panel of 678 diverse wheat landraces phenotyped for resistance to three rust diseases (Tables S13-S16 in Additional files 5, 6, 7, and 8) [15] was tested for marker-trait associations using genotyping data generated with the wheat 90 K SNP array. In this panel we were able to impute 344,544 SNPs, of which 210,017 SNPs with a MAF above 3% and the proportion of missing data less than 80% were used for GWAS. Three selected marker-trait associations were validated by mapping in the populations of recombinant inbred lines; two of these associations correspond to disease resistance loci *Lr67* and *Yr51*, and one represents an uncharacterized locus (Figure 3b) associated with stem rust resistance (Tables S13-S16 in Additional files 5, 6, 7, and 8). Two additional regions from our GWAS were shown to overlap with the positions of the previously mapped *Lr37/Yr17/Sr38* and *Lr68* disease resistance loci (Figure 3b). The markers associated with these loci showed highest similarity to CSS contigs 2AS-5264433 and 7BL-6748067, respectively [27,50]. Overall, comparison of marker-trait associations at non-imputed and imputed sites shows that imputed SNPs not only increase marker density but in most cases perform similar to or better than the SNPs directly genotyped using the 90 K assay (Figure 3b; Figure S17 in Additional file 1). These results demonstrate the value of having a more complete ascertainment of DNA polymorphisms for GWAS that is achieved utilizing the high-density SNP variation data developed from 62 lines and the public 90 K SNP genotyping array [14].

Signatures of selection in the polyploid genome

During the development of adapted lines, selection imposed by humans favored alleles controlling traits valuable for agriculture. When selection increases the frequency of beneficial alleles in a population, it impacts the standing variation of surrounding genomic regions resulting in reduced diversity, extended linkage disequilibrium, or strong inter-population allele frequency differentiation [51]. To detect these local patterns of variation, also referred to as 'selective sweeps', we investigated the patterns of genetic variation along the chromosomes and used two complementary approaches based on cross-population composite likelihood ratio (XP-CLR) [52] (Figure S18 in Additional file 1) and pair-wise haplotype sharing (PHS) [53] tests (Tables S17-S19 in Additional files 1, 9, and 10).

The reduced diversity observed near the wheat domestication genes *Q* and *Tg* was consistent with selection at the early stages of domestication (Figure 2b; Figure S11 in Additional file 1) [54,55]. For example, the CSS contig 1750512_5AL harboring the *Q* gene harbors only one intronic SNP at position 1249. Regions harboring genes known to be associated with local adaptation (*Ppd*, *Vrn*, *Rht*) also showed selection scan test statistic scores approaching the extremes (Table S20 in Additional file 1). To further validate that our selection scans detect genomic regions associated with candidate loci controlling agronomic traits targeted by humans during the development

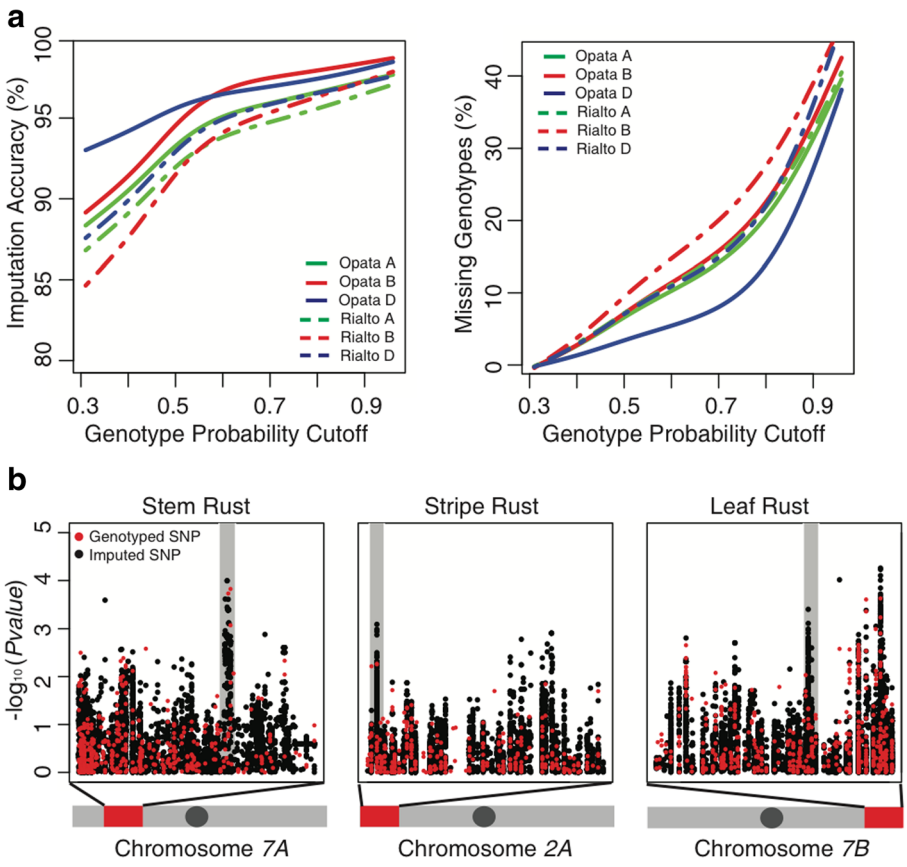

Figure 3 Genotype imputation. (a) Relationship between the accuracy of genotype imputation and the percentage of missing data, which is estimated after removing genotypes over a range of genotype calling probability thresholds. Imputation in Opata (solid lines) and Rialto (dashed lines) cultivars was performed using the reference panel of 60 lines (Opata and Rialto cultivars were excluded) genotyped using the 90 K iSelect assay. **(b)** Genotype imputation at disease resistance loci. Two GWAS regions overlapped with the positions of the previously mapped *Lr37/Yr17/Sr38* (middle panel) and *Lr68* (right panel) disease resistance loci; the markers associated with these loci showed highest similarity to CSS contigs 2AS-5264433 and 7BL-6748067, respectively. SNP sites directly genotyped using the 90 K SNP array are shown as red dots; imputed SNPs are shown as black dots.

of improved varieties, we compared selective sweeps with the marker-trait associations identified in mapping studies. We found 474 previously published associations that fell within the target regions (Table S21 in Additional file 11). These markers showed association with major domestication and agronomic traits including spike length, rachis fragility, and compactness [56,57], heading date and flowering time [16,58], grain shape and yield characteristics [59,60], and nitrogen use efficiency [61]. The selective sweep regions also included markers associated with resistance to stripe rust [50], bacterial leaf streak, and spot blotch [62,63]. The regions detected in the PHS scan overlapped with 459 marker-trait associations. Far fewer previously associated markers were located within the outliers of the F_{ST}, XP-CLR, and diversity scans.

Non-synonymous, that is, likely functional, variants were significantly enriched ($\chi 2$ test, P value $<5.4 \times 10^{-11}$) in the extreme tails of the selection scans compared to synonymous variants. Regions identified by multiple selection scans were not common (Table S22 in Additional file 1), and the regions where all three scans overlapped contained no genes with annotations. Among the overlapping regions detected using two different methods, one of the most common classes of genes were disease resistance genes (Table S10 in Additional file 3). Consistently, the NB-ARC and LRR encoding domains of genes involved in disease resistance pathways [64] were significantly over-represented in the extreme tail of the PHS scan, ($\chi 2$ test; FDR $<10^{-5}$ and $<10^{-2}$, respectively) (Table S23 in Additional file 1) suggesting that some targets of selection can be associated with selection for disease resistance. To confirm this hypothesis, we performed GWAS of resistance to leaf, stem, and stripe rust (Figure 3b, Materials and methods) and tested for enrichment of marker-trait associations in the extreme tail of the selection scan. In the upper 2.5% tail of the PHS scan, we found three-, four-, and five-

fold enrichment of SNPs ($\chi 2$ test, $P < 6.5 \times 10^{-6}$) associated with resistance to stem, leaf, and stripe rust, respectively, without the concomitant enrichment in the XP-CLR scan. Since the PHS test preferentially detects on-going selection events that have not reached fixation in a population [53], our results suggest that multiple disease resistance genes undergo selection across wheat populations, which is likely associated with the spatial and temporal variation in pathogen populations, and consistent with the 'arms race' hypothesis [39].

Among other candidates of selection is the WRKY transcription factor (Ta1dsLoc014113), identified by both selection scans and located on the short arm of chromosome 1D; its expression was shown to be associated with resistance to a fungal pathogen causing powdery mildew in wheat [65]. Glutathione-S-transferase encoding genes (Ta7dsLoc015003, Ta7bsLoc009692), that play an important role in drought response by reducing the toxicity of reactive oxygen species in wheat and other plants [66,67], were identified by the PHS and F_{ST} scans on chromosomes 7B and 7D (Table S10 in Additional file 3).

The development of the chromosome-specific wheat genome assemblies and population-scale whole exome re-sequencing data now provides the unique opportunity to investigate the impact of selection on duplicated copies of homoeologous genes. We have used the ordered sets of homoeologous genes to establish the syntenic relationships between the selection targets on different genomes. Inter-genomic comparison of selection targets associated with transition from landraces to cultivars revealed that only a single syntenic region shared the signature of selection between the B and D genomes (Table S24 in Additional file 1). This region on the long arm of chromosome 1 shared two annotated genes encoding cellulose synthase (Ta1dlLoc001027, Ta1blLoc007155, Ta1blLoc025394), which plays a central role in cellulose biosynthesis, and trehalase (TH) (Ta1dlLoc015131, Ta1blLoc007983). The latter gene contains a PFAM domain that was significantly enriched in the tail of the XP-CLR scan (Table S23 in Additional file 1), and along with trehalose phosphatase (TP), detected by both the XP-CLR and PHS scans (Table S10 in Additional file 3), is involved in trehalose metabolism. The overexpression of TH was shown to increase drought stress tolerance in *Arabidopsis* [68] and trehalose accumulation in transgenic rice expressing TP was associated with an increased tolerance to drought, salt, and cold stresses [69].

The regions subjected to recent selection detected by the PHS scan showed a more substantial overlap in pairwise comparisons between the genomes than the regions detected using the XP-CLR and F_{ST} scans (Table S24 in Additional file 1) suggesting that during wheat improvement selection most likely operated on standing variation. However, the presumed adaptive variants with high-PHS in the overlapping homoeologous regions under selection are rarely found in the same line (Figure S19 in Additional file 1). In the vast majority (75%) of overlapping homoeologous regions, no wheat lines in the panel possessed the high-PHS variants simultaneously in two genomes. Although there are known examples where allelic variation at the three wheat homoeologs affect the corresponding traits [70-72], our findings indicate that multiple parallel changes across homoeologous regions have rarely been favored by selection. Rather, it is likely that any favored variant at any one of the homoeologous regions may be sufficient to provide a fitness benefit, thereby expanding the target size of advantageous mutations. The rarity of individual genotypes with positively selected alleles in different homoeologous regions possibly indicate that additional advantageous mutations in other homoeologs either: (1) do not provide fitness benefit; or (2) are absent in a population subjected to improvement selection.

Gene expression studies have demonstrated that, in wheat, the homoeolog-specific transcriptional dominance affects up to 19% of genes [32] with different homoeologs being preponderant in different groups of functionally related genes and showing the tissue- or development-specific patterns of expression [33]. These data are consistent with partitioning of gene function among the duplicated homoelogous genes [34,35], which could also affect the distribution of selective sweeps among the wheat genomes. One possibility is that selection acts on the preferentially expressed homoeolog. For example, most of the natural variation impacting the vernalization requirement in wheat is located in the *VRN-A1* homoeolog [70], which is the homoeolog expressed at the highest level [73].

Conclusions

The sequence-based diversity map reported here is an important step towards the detailed characterization of DNA sequence polymorphism in the complex allopolyploid genome. The recently developed wheat genomic reference [27] allowed us to catalogue common SNPs and small-scale indel polymorphisms from the low-copy fraction of the genome, describe their patterns of chromosomal distribution and inter-variant association, and identify variants that may have an impact on gene function based on the available annotation. A developed haplotype map will be a valuable tool for imputing genotypes and transferring sequence-level variation data across multiple gene mapping projects, thereby increasing the power and precision of trait mapping in GWAS and helping to understand better the basis of complex phenotypic traits.

Our data helped us gain insights into historic selective events and identify candidate selection targets associated

with regions harboring genes controlling important agronomic traits or involved in response to biotic and abiotic stress stimuli. Our results suggest that directional selection in allopolyploids rarely acted on multiple parallel advantageous mutations across homoeologous regions. A favored variant at any one of the homoeologous regions appears to provide sufficient fitness benefit. By broadening the set of targets for selection, allopolyploidy may have played a critical role in the evolution of adaptation in wheat and contributed to wheat's success as a globally grown crop. Duplicated homoeologs may increase the likelihood of recovering beneficial alleles by expanding the advantageous mutation target size, and/or capturing allelic diversity present in different homoeologous genomes.

Materials and methods

Selection of wheat accessions

A total of 62 diverse hexaploid wheat lines (Table S1 and Figure S1 in Additional file 1) were selected to represent the genetic diversity of a large wheat collection that was previously genotyped using the 9 K wheat iSelect assay [13]. In addition, attempts were made to select accessions from major wheat growing areas (Figure 1b). The sample included 26 landraces, 29 cultivars, six breeding lines, and one synthetic wheat (broadly used in the breeding programs of CIMMYT, Mexico), among which 49 and 13 lines show spring and facultative/winter growth habits, respectively. The sample size of 62 lines allows for detection of most common variants present in populations at frequencies ≥1.6%.

Capture assay design

A wheat exome capture (WEC) assay targeted the 107 Mb of non-redundant low-copy regions in the wheat genome. The capture probes were designed using the low-copy number genome assembly (LGC) of the wheat cultivar Chinese Spring [74]. The LGC had chloroplast, mitochondria, and transposon sequences removed and also contained homoeologous copies of genes collapsed into a single contiguous sequence [29]. The LGC was 3.8 Gb in size. We adopted two strategies to reduce the size over which we could design probes and target exonic regions. First, we used the BLASTN program (e-value $<1e^{-10}$) to identify LGC contiguous sequences that were similar to *Brachypodium* exon sequences. Second, we used the same LGC sequence library (BLASTN e-value $<1e^{-20}$) to identify LGC contiguous sequences that matched a set of non-redundant wheat cDNA and EST sequences [18] and transcriptome assemblies generated by 454 sequencing of nine diverse wheat cultivars [13]. Finally, to remove sequence duplications from the contiguous sequences set, we compared the set against itself using the BLASTN program. Similar sequences were identified (95% identity over 100 bp) and the longest contiguous sequence of a matching pair was retained. This process resulted in a design space of 110 Mb that was used to design a final probe set covering 107 Mb currently available from Roche NimbleGen [75]. The WEC assay was designed by the wheat-barley exome capture consortium that also designed an assay for the enrichment of the barley exome [76].

DNA extraction and sequence capture

Each accession before DNA isolation was self-pollinated for two generations in the greenhouse. DNA was extracted using the DNeasy Plant Maxi Kit (Qiagen, USA) from a single 3-week-old seedling. One μg of DNA was fragmented with the Covaris S220 to obtain an average fragment length of 300 bp. The NEBNext DNA Library Prep Kit (NEB) for Illumina and Illumina TruSeq (TS) indexed (barcoded) adapters were then used for sample library preparation according to NEB protocol with the following exceptions. The PCR Enrichment step from the NEB was replaced with the Ligation Mediated PCR (LM-PCR). The TS-PCR Oligo1 (5′-AATGATACGGC GACCACCGAGA-3′) and TS-PCR Oligo2 (5′- CAAGC AGAAGACGGCATACGAG-3′) were used in the LM-PCR. The LM-PCR products were purified with the QIAquick PCR Purification Kit (Qiagen) followed by the size selection using Agencourt AMPure XP Beads (Beckman Coulter). The libraries were tested on 2100 Bioanalyzer (Agilent Technologies) and the NanoDrop Spectrophotometer (Thermo Scientific). Only samples with average fragment lengths of 200 to 400 bp, $A_{260/280}$ ratio 1.7 to 2.0, and LM-PCR yield >500 ng were pooled and used for sequence capture. Several levels of DNA sample pooling have been used in our study (Table S1 in Additional file 1). One μg of non-pooled or pooled DNA (for example, 1 μg, 500 ng, 333 ng, 250 ng, or 125 ng of each component DNA sample library were used for 0, 2×, 3×, 4×, or 8× pools, respectively) was used in each of the sequence capture hybridizations. The sequence capture was performed as previously described [77].

Genotyping by sequencing (GBS)

Complexity-reduced sequencing was performed according to the previously described protocol [25] modified from Poland *et al.* [78]. A pooled library was sequenced on one lane of Illumina HiSeq 2000 (2 × 100 bp paired-end). All subsequent analyses were carried out using the same approaches for exome capture and GBS datasets.

Selection of alignment parameters for polyploid genome

For mapping reads to the polyploid wheat genome we have developed a three-step iterative alignment strategy with parameters optimized to map reads uniquely to the different wheat subgenomes. Parameters were optimized

using a subset of 29,716 Illumina 2 × 100 bp reads generated for cultivar RAC875. These reads map to 100 homoeologous sets of genes (three copies per gene) from the wheat CSS assemblies [27]. We mapped Illumina reads to this reference set of 300 genes using various combinations of Bowtie's alignment parameters (Figure S4 in Additional file 1).

Alignment to the CSS assemblies

Raw paired-end Illumina reads were quality filtered using the default setting of NGS QC toolkit v2.3 [79], retaining reads if ≥70% of the bases had a quality score ≥20. Only paired-end reads were mapped to the CSS assemblies using Bowtie1 v.0.12.7 [80] and Bowtie2 v.2.0.0 [81]. We applied the three-step iterative mapping strategy using Bowtie to perform ungapped read alignment (Figures S4 and S5 in Additional file 1). Reads that do not align using more stringent criteria were reused for subsequent rounds of alignment with lower stringency. To find insertion/deletion (indel) polymorphisms, we performed gapped alignment using Bowtie2 v. 2.0.0 [81] with the following parameters: -N 1, -L 75, -D 20, -R 3, –no-mixed, –end-to-end.

More than 4.7 billion paired-end reads were generated from the population of 62 accessions. On average 62% of quality-filtered reads generated by sequence capture were mapped covering more than 321 Mb of the hexaploid wheat genome (Table S2; Figure S3 in Additional file 1). On average 51% of quality-filtered reads generated by the GBS approach were mapped covering more than 247 Mb of mostly intergenic space (Table S3 in Additional file 1).

Efficiency of homoeologous target capture

The WEC assay included probes covering 107 Mb of non-redundant target space in the wheat genome. To assess the ability of the capture assay to enrich for targets from the three homoeologous wheat genomes, we compared the WEC against the CSS assemblies and retained only those that had three best BLASTN hits (e-value $<1 \times 10^{-10}$); one hit per genome. There were 47,739 homoeologous gene sets that fit these criteria. The $\log_2$ ratio of average coverage depth for each of these gene sets in pair-wise genome comparisons (A vs. B, B vs. D, and A vs. D) was distributed normally with the mean centered at 0, suggesting that homoeologous targets are captured with equal efficiency (Figure 1c).

To assess *in silico* the total size of the regions targeted by the WEC assay in the wheat genome we used the BLAT program [82] to align the WEC design space against the CSS contigs (alignment length >100 bp, similarity >80-99%) (Figure S2 in Additional file 1).

Capture efficiency

Analysis of alignments generated using the bowtie [81] and BWA [83] aligners showed that 95% of the 107 Mb WEC design space was covered at 30× depth and 99% of the design space was covered by at least one read. Approximately 78% of the annotated high confidence exons in the CSS [27] were covered by at least one read. The average depth of read coverage for annotated exons in each accession was 8.1, 8.6, and 8.4 for the A, B, and D genomes, respectively (Table S4 in Additional file 1), further suggesting no bias in the efficiency of homoeologous genome capture.

Variant calling and filtering

Each accession's BAM file was sorted and indexed using Samtools version 0.1.18 [84] for variant calling. Next, GATK version 2.2-8 [85] was used to realign reads around indels. The program Picard v. 1.62 [86] was used to remove duplicate reads in the realigned BAM files. Finally, base quality recalibration was performed using the GATK program [85]. We identified 53.8 million raw variants using the Unified Genotyper from GATK following the GATK instructions for exome capture datasets. Subsequent recalibration of variant quality scores was performed using the default parameters of the VQSR tool in GATK [85]. As 'true variant' calls VQSR utilizes genotypes obtained with the wheat 90 K SNP assay [14]. All variants were filtered further to remove sites that had <46 (<75%) accessions genotyped or >2 alleles. We retained sites that had no more than one accession with a heterozygous call, or sites where the heterozygous call was due to a single read of the secondary allele (possibly sequencing error).

Due to the low-coverage depth obtained in the GBS dataset (on average of 1.04 read), genotype calling was performed only for alleles that were present in at least two different accessions in the population. The high level of genotype calling concordance between the WEC and GBS datasets was indicative of the high accuracy of the applied GBS genotype calling approach (see details in section 'Error rate estimation').

Error rate estimation

We assessed the accuracy of genotype calling using four datasets. First, we compared variant calls generated for the cultivar Chinese Spring in our sequence capture experiment with the CSS assemblies that have also been generated using the same wheat cultivar. Of the filtered sites we found that 1,505,400 SNPs had genotype calls for cultivar Chinese Spring and 16,736 of these genotype calls were different from the reference sequence for an error rate of 1.1%. Similarly we estimated an error rate of 1.5% for indels where we found 1,576 indels that disagreed with the 106,438 indels that had a genotype call for Chinese Spring. The rare variant (MAF ≤1.6%) calling accuracy assessed in a set of 12,426 singletons unique to cultivar Chinese Spring depended on the read coverage depth (Figure S7 in Additional file 1). By applying a ≥10 read coverage cutoff

we obtained a singleton error rate of 4.6% for SNPs and 3.4% for indels.

Second, to exclude the possibility that the assessed error rates are not impacted by the quality of CSS reference assembly, we compared our variant calls generated for Chinese Spring with the published Chinese Spring BAC sequences generated using the Sanger approach [87]. We found 260 discrepancies out of 85,973 SNPs for an error rate of 0.3%.

Third, we estimated the concordance of our genotype calls with the genotype calls generated using the 90 K iSelect genotyping assay [14]. We selected only those SNP assays from the 90 K assay that were polymorphic in our population and whose flanking sequences could be unambiguously mapped to a single location in the CSS assemblies. Out of 27,147 homozygous genotypes called for these SNPs in our alignments, 520 were different from the SNP assay dataset, for an overall concordance rate of 98.1%.

Fourth, we have assessed the concordance of genotype calls by comparing the WEC and GBS datasets including the shared 10,028 SNP and 838 indel sites (Figure 1d). The level of genotype calling concordance between these two datasets was 97.2% for SNPs and 95.4% for indels.

SNP annotation

The impact of SNPs on coding sequences was assessed with the SNPEffect program [88] using the 124,201 high confidence gene models predicted in the CSS contigs. We found that the proportion of SNPs identified in the intergenic regions is higher than that assigned to the coding regions. This pattern of SNP distribution is likely defined by several factors. First, WEC is designed using the low-copy fraction of wheat genome that overlapped with the sequences of wheat transcripts. These sequences do not always overlap with the high-confidence gene annotations of the CSS contigs used for annotating SNPs. It is possible that with better annotation of the wheat genome, new genes corresponding to these low-copy genomic regions could be found in WEC design. Second, the WEC co-captures regions outside the target region including introns and non-coding DNA. These regions are overall more genetically diverse and can also contribute to SNPs in the intergenic space.

Functional annotation and enrichment were determined using BLAST2GO software [89]. The ancestral state at SNP sites was inferred by comparing the SNP-flanking sequences among the A, B, and D genomes and with sequences of diploid wheat ancestors *T. urartu* [90] and *Ae. tauschii* [45]. We were able to infer the ancestral allelic states for 754,080 of the 1.57 million SNPs.

Genetic diversity analyses

Due to variation in the depth of read coverage across the genome genotype calling rate for rare variants present only once in the population can vary from region to region. To reduce the effect of false negative rare variant calls on the local estimates of diversity, all analyses were performed using the variants with MAF >1.6%.

Population structure was inferred using PCA and Structure analyses [91]. Structure was run 10 times using the admixture model with correlated allele frequencies for 20,000 burn-in 100,000 MCMC iterations. Results of independent runs were summarized using CLUMPP [92]. The optimal number of populations (K) in the dataset assessed by plotting the probability of data ln Pr (X|K) for each value of K was K = 4. The proportion of each accession's ancestry in one of the K populations is presented on Figure 1b as a pie chart.

Recently, it was demonstrated that the estimates of commonly used diversity statistics (F_{ST}, Tajima's D, and π) obtained using restriction enzyme-based sequencing can deviate from true values [93]. Ascertainment bias in GBS data can result from the mutations at the restriction enzyme cut sites, sensitivity of the PstI enzyme to DNA methylation, or a high proportion of missing genotypes in low-coverage re-sequencing datasets. Allele frequency estimates obtained for the same sites in the GBS and WEC datasets showed good correlation (Figure S20 in Additional file 1). However, we observe a slight decrease in correlation values with the increase of the proportion of missing data in the GBS dataset. To reduce the potential effect of missing data on our analyses this dataset was excluded from the estimates of diversity statistics.

The order of the CSS contigs along the wheat chromosomes was established using the combination of 'genome zipper' [94] and a high-density genetic map developed by the population sequencing (PopSeq) of 90 recombinant inbred lines [27]. The approximate physical positions of ordered CSS contigs on the chromosomes were inferred using a framework created by cross-linking high-density SNP wheat genetic map [14], wheat deletion bin map [95], and barley genome sequence [96]. Comparison of the inferred physical positions with the positions of CSS contigs on the published sequence of the chromosome 3B pseudomolecule [28] showed a high level of correlation (Pearson correlation $r^2 = 0.97$). The ordered CSS contigs were used to perform sliding window analyses. π and Tajima's D estimates for 2 Mb sliding windows (1 Mb step) with >10 kb of CSS contigs covered by reads were calculated using LDHat (v. 2.2) [97]. Outliers of π and Tajima's D were defined by the 2.5 percentiles of the test statistic distribution. F_{ST} was calculated by SNP using the R package adegenet (cran.us.r-project.org) by contrasting cultivars and landraces. A sliding window analysis (2 Mb with a 1 Mb step) was used to generate average estimates of F_{ST}. Outliers were calculated as the 97.5 percentile of all window values within the genome.

Adjacent outlier windows were merged. The thresholds for the cultivar-landrace comparison were 0.158 for the A genome, 0.108 for the B genome, and 0.0925 for the D genome.

For estimating recombination rate we selected 10,517 CSS contigs that contained at least 20 non-singleton SNPs. The historical recombination parameter ($\rho = 4Ne$) was estimated using a composite likelihood approach implemented in the Maxhap program [47]. Regions of the wheat genome showing elevated levels of historic recombination were defined as those falling above 97.5 percentile of the genome-wide ρ distribution. The critical values of ρ for each genome were 7.5×10^{-3}, 8.5×10^{-3}/kb, and 9.7×10^{-3}/kb for the A, B, and D genomes, respectively.

Pair-wise estimates of LD were obtained for SNPs with a MAF ≥0.05 by measuring r^2 as previously described [13]. The average length of pair-wise shared haplotypes in the population was calculated around each SNP according to the previously described procedure [98]. The genetic map distances were taken from the wheat SNP map developed using the 90 K SNP assay [14].

Distribution of PFAM domains in the CSS contigs

The nucleotide sequences of the CSS gene models for *T. aestivum* were mapped against the wheat survey sequence using GMAP [99]. Output was filtered for hits with ≥98% identity and ≥80% coverage. PFAM domains were identified in the amino acid sequences of the CSS gene models using PFAM hidden Markov models (release 26) from the Sanger Institute [100] and HMMER (version 3.0b3) from the Howard Hughes Medical Institute Janelia Farm Research Campus [101]. The PFAM searches were performed using the default parameters and the results were filtered for hits with ≥90% coverage.

Phylogenetic tree construction

A total of 20,000 SNPs were randomly selected from the dataset to estimate pairwise distances among the 62 accessions using the R package 'ape' V3.0.6 [102]. The bootstrap values for each node of the tree are the average of 10 separate bootstrap runs, each comprising 100 iterations. In total, 45 nodes of 60 have at least a 70% chance of being grouped into the same cluster, and 56 had at least a 50% chance of being grouped together. The neighbor-joining tree is shown on Figure S1B in Additional file 1.

Distribution of *Ae. tauschii* (genotype AL8/78) alleles across the D genome chromosomes

To determine which of the D genome alleles corresponds to the AL8/78 genotype, 100 bp flanking sequences of each D genome SNP were extracted and compared against the genomic sequence of *Ae. tauschii* AL8/78 genotype [45]. In total, we could determine AL8/78 alleles for 105,294 SNPs. The distribution of AL8/78 alleles along the wheat chromosomes was estimated in the population of the 26 landraces. The average frequency of the AL8/78 allele in a sliding window of 3 Mbp was plotted along the wheat chromosomes.

Detection of selective sweeps

The PHS statistic was calculated as described by Toomajian *et al.* [53]. Utilizing this statistic we have higher likelihood of detecting genomic regions harboring alleles present at intermediate frequencies in the population [53] than those alleles that have reached high frequency. Thresholds of the PHS statistic were determined by taking the 97.5 percentile of the distribution of PHS values for each SNP within different allele frequency classes (in a window of 0.05). For detecting selective sweeps the wheat genome was split into 50-kb windows; each window was assigned a maximum PHS value for SNPs in the window. Neighboring windows located within 1 Mb that contained outlier SNPs were merged. Annotated genes harboring the outlier SNPs were used for PFAM domain enrichment analyses. Enrichment analysis of GWAS SNPs included sites with a minor allele frequency >3% and a significance level $P < 10^{-3}$.

The selection scan using the XP-CLR approach is robust to assumptions regarding recombination rates and demography [52] and compares the allele frequency differentiation and the extent of linked variation between two populations (cultivars vs. landraces) to detect regions where change in frequency occurred too quickly to be caused by random drift. The XP-CLR scan was run with the grid size of 50 kb, the window size of 1 cM, and the maximum number of SNPs fixed at 500. The critical values for putative selection targets were estimated based on the 97.5 percentile of the test statistic distribution for each wheat genome. Since our population includes both spring and winter wheat lines, one of the concerns was that population differentiation between these growth habit groups can be mistaken for the signals of selection. However, we believe that at the genome-wide level, inclusion of spring and winter wheat lines should not have a large effect on the detected signals of selection because the level of genetic differentiation between the spring and winter wheat in our population was shown to be very low (mean genome-wide $F_{ST} = 0.03$). This low F_{ST} was observed previously [13] and attributed to the common practice to use lines from both growth habit groups in the same breeding programs, as well as to the heterogeneity of the genetic basis of flowering time regulation in wheat where the same phenotypic outcome can be obtained by mutations at several independent genetic loci. In addition, we have identified regions of the wheat genome showing extreme

genetic differentiation (F_{ST}) or XP-CLR test statistics between spring and winter wheat. Out of 372 XP-CLR outliers in the landrace-cultivar comparison only nine (2.4%) partially overlapped with the outliers in the spring-winter wheat comparison (shown in Table S18 in Additional file 10). Out of 168 outliers in the F_{ST} scan of cultivars and landraces only six (3.6%) overlapped with the F_{ST} outliers in the spring-winter wheat comparison. No overlap of genetically differentiated genomic regions between spring and winter wheat with the outliers of the genetic diversity scan was found. None of the genes located in the genomic regions that showed the signature of selection in at least two scans reported in Table S10 (Additional file 3) fall within the genomic regions differentiated between spring and winter wheat. Overall, these analyses suggest that the genetic differentiation between spring and winter wheat in our sample should not have significant impact on the results of our selection scans.

To test the significance of the observed overlap between the selective sweeps located on the homoeologous chromosomes, we have randomly permuted 10,000 times a genome-wide set of 50-kb windows. The proportions of windows that overlapped among each pairwise comparison of the wheat genomes was ranked and compared to observed data to calculate empirical *P* value.

Comparison of selective sweeps with previously characterized marker-trait associations

To check if the putative selective sweeps identified in our scans harbored loci associated with agronomic traits, we analyzed published marker-trait associations detected using the DArT markers [103], and 9 K and 90 K iSelect SNP assays [13,14]. The DArT markers were mapped to the CSS contigs using the BLAT program (best hit with minimum alignment length >150 bp). Out of 57 DArT markers that could be mapped to the ordered CSS contigs, 37 markers fell into the regions detected in the PHS scan, and one fell into a region detected by both the PHS and F_{ST} scans (Table S21 in Additional file 11). Out of 555 SNP markers mapped to the ordered CSS contigs, 422 mapped to the regions identified in the PHS scan. We found five SNPs in the XP-CLR regions, three of which overlapped with the PHS regions. One SNP marker was located within the region showing reduced level of diversity. There were 30 SNP-trait associations that fell into high F_{ST} regions, of which 18 overlapped with the high-PHS regions.

Imputation

Genotype imputation was performed using Beagle v.4 [104] with the following parameters: 'window = 5,000 overlap = 500 burns-its = 10 impute-its = 10'. To increase the accuracy of imputation, the settings of burns-its and impute-its have been increased from the default settings (burns-its = 5, impute-its = 5) to 10 (according to recommendations in user's manual). The accuracy of genotype imputation assessed in windows including from 1,000 to 5,000 markers for cultivars Avalon and Rialto showed no significant differences (Figure S21 in Additional file 1). A setting of window = 5,000 was selected because of its computational efficiency.

To test the accuracy of imputation, we have sequentially selected each cultivar from the panel of 62 lines and masked all variants, except approximately 14,000 SNPs overlapping between the WEC and 90 K SNP iSelect array. At these SNP sites at least 75% of accessions in both datasets had genotype calls. The remaining 61 cultivars were used as a reference panel for imputing 649,502 SNPs that were ordered along the wheat chromosomes. After imputation, genotypes were filtered using different thresholds of genotype probability assessed by Beagle. The filtered predicted genotypes in each cultivar were compared with the actual genotype calls obtained by WEC sequencing to assess the accuracy of imputation. Relationships between the genotype probability threshold, proportion of missing data after filtering, and imputation accuracy are presented on Figure 3a and Table S12 in Additional file 4. The number of imputed genotypes varied among chromosomes depending on the number of polymorphic SNPs from the 90 K iSelect assay on each chromosome. Because of the low level of polymorphism, a relatively low number of SNPs could be imputed in the wheat D genome. However, due to its high LD levels, imputation accuracy on the D genome chromosomes was higher than that in the A and B genomes.

A similar imputation strategy was used for imputing genotypes in the panel of 678 wheat lines used for GWAS of disease resistance.

Genome-wide association study (GWAS)

Plant materials

A total of 838 bread wheat accessions from the Arthur Watkins Collection were obtained from the Australian Winter Cereals Collection, Tamworth. This collection includes a large number of phenotypically diverse wheat landraces collected from 32 countries in the 1920s to 1930s [105]. The 838 accessions were grown under field conditions and single plant selections made for 678 accessions on the basis of plant type, rust resistance, and maturity. Despite some maturity differences, all 678 land race accessions flowered by the end of October. Seed from the purified 678 accessions can be obtained from the Plant Breeding Institute, Cobbitty, upon request (contact urmil.bansal@sydney.edu.au).

Phenotyping

Following seed bulk up for the single plant selections, the purified 678 accessions were grown under field

conditions during 2006 and 2007 at the Lansdowne site of the University of Sydney. Each accession was grown as a single 1 m row. Field trials were artificially inoculated with *Puccinia striiformis* f. sp. *tritici* (Pst) pathotype 134 E16A+; *P. triticina* (Pt) pathotypes 104-1,(2),3,(6),(7),11 + Lr37; 104-1,(2),3,(6),(7),11,13; 104-1,(2),3,(6),(7),9,11; 76-3,5,9,10 + Lr37; 10-1,3,(7),9,10,11,12 and *P. graminis* f. sp. *tritici* (Pgt) pathotypes 98-1,2,3,5,6 and 34-1,2,7 + Sr38. The pathotype designations are provided according to McIntosh *et al.* [106]. These pathotypes carry partial virulence for the genes given in the parenthesis. Stripe rust, leaf rust, and stem rust responses were recorded using a 1–9 scale, where 1 was highly resistant and 9 was highly susceptible [107]. For stripe rust disease, two records were taken within each year. The consistency of phenotypic evaluations across years was tested by calculating the Pearson correlation coefficient. For leaf, stripe, and stem rust phenotyping datasets the correlation coefficients were 0.75, 0.71-0.87, and 0.57, respectively.

Genotyping

The 678 landrace accessions were genotyped using the Infinium iSelect 90 K SNP assay, the content of which is reported to have minimal ascertainment bias for the analysis of diverse wheat landraces [14]. Genotyping was performed on the iScan instrument according to the manufacturer's protocols (Illumina). SNP genotype calling was performed using GenomeStudio v2011.1 software (Illumina) and the genotype calling algorithm reported in Wang *et al.* [14]. Monomorphic markers and SNPs with more than 10% missing data (due to the presence of null alleles or poor genotype call rates) were removed. The genetic map developed by genotyping multiple mapping populations with the 90 K array was anchored to the GenomeZipper and PopSeq maps by comparing the sequences of 90 K array SNPs with the sequences of wheat chromosome assemblies [27].

Association mapping

Mixed model variance component analysis of mean phenotypic values was performed using the R package GAPIT [108]. The information about the relationship among accessions in the population was provided as a kinship matrix (random effect). The effect of population structure was controlled by using the first three principal components from the principal component analysis (fixed effect). The P value $<1 \times 10^{-3}$ used to filter markers was selected based on the previous studies of agronomic traits in wheat and barley demonstrating that this threshold provides adequate accuracy for detecting marker-trait associations, as was validated independently in bi-parental mapping populations [50,63,109]. In a follow-up analysis we successfully validated five GWAS regions by comparing with previously published studies, or by mapping a trait in a bi-parental mapping population in our study (shown on Figure 3b and in Table S14 in Additional file 6).

Validation of GWAS signals in mapping populations

Plant materials

Mapping in the populations of recombinant inbred lines (RIL) was used to validate several marker-trait associations identified in the GWAS including two characterized and one previously uncharacterized disease resistance loci. The population segregating for leaf rust resistance gene *Lr67* included 124 F_3/F_4 lines derived from a cross between Thatcher and RL6077 [110]. The population segregating for stripe rust resistance gene *Yr51* was comprised of 89 F_6 RILs derived from a single heterozygous F3 plant #5515 from a cross between Watkins' line PBI769 and Westonia [111]. The population segregating for a new stem rust resistance gene on chromosome 7A was comprised of 96 F_6 RILs derived from a cross between Watkins' line PBI562 and Yitpi (henceforth, PBI562/Yitpi).

Identification of SNPs linked to rust genes

SNPs associated with rust resistance genes segregating in each mapping population were identified using bulk segregant analysis (BSA) [112] and selective genotyping (SG).

For BSA, resistant and susceptible bulks were prepared by pooling equal amounts of genomic DNA from at least 20 plants for each phenotypic class. An artificial F1 sample was prepared by combining an equal amount of DNA from each of the two bulks. The bulked DNA samples, artificial F1, and parental lines were genotyped using a custom Infinium iSelect bead chip assay on the iScan instrument following the manufacturer's instructions (Illumina Ltd.). The Thatcher/RL6077 and PBI No. 769/Westonia were genotyped using the 9 K and 90 K iSelect genotyping arrays, respectively. The SNPs were assessed for putative linkage by comparing the normalized theta values for each sample as described in Hyten *et al.* [113]. Polymorphism was considered to be linked to a rust resistance gene when the normalized theta values for the resistant bulk and resistant parent, and susceptible bulk and susceptible parent were similar, and when the normalized theta value for the artificial F_1 samples was about halfway between that of the other samples.

For SG, at least 15 resistant and 15 susceptible plants from PBI No. 562/Yitpi cross and its parents were genotyped using the 90 K iSelect bead chip assay. Polymorphism was considered to be linked to a rust resistant gene when the majority (>90%) of individuals within each

phenotypic class were fixed for the expected parental allele.

Validation of GWAS-linked SNPs using bi-parental mapping crosses

The genomic location of SNPs associated with rust disease resistance in the GWAS analysis were compared with those linked to mapped rust resistance genes in each of the bi-parental mapping crosses. GWAS-associated SNPs were considered to co-locate with the mapped resistance genes when the linked SNPs in each study were located at (or very near) the same position in the 90 K consensus SNP genetic map [14]. In instances where the linked markers in either study were not present in the 90 K consensus SNP map, the GWAS-associated SNPs were considered to co-locate with the mapped resistance genes when the CSS contigs tagged by the SNPs [14] co-located in the PopSeq map [27]. The co-location of SNPs associated with rust disease resistance in the GWAS analysis with mapped rust resistance genes in the bi-parental mapping populations is shown in Figure 3b and Table S14 in Additional file 6.

Data availability

The sequencing data have been deposited in the NCBI Short Read Archive under accession number SRP032974. Datasets used for diversity analyses, genotype imputation, and GWAS are available from the project website [114]. Variant calling datasets in the VCF format for both WEC and GBS are available from the USDA GrainGenes and URGI websites [115,116].

Additional files

Additional file 1: Tables S1-S7, S9, S11, S19-S20, S22-S24 and Figures S1-S21. Table S1. List of wheat lines sequenced. **Table S2.** Summary of exome capture results. **Table S3.** GBS of the wheat diversity panel. **Table S4.** Mean depth of read coverage. **Table S5.** Distribution of wheat exome capture and GBS variants among the genomic features. **Table S6.** Loss-of-function variants. **Table S7.** Distribution of indels among genomes and their effect on codon reading frame. **Table S9.** Quantile distribution of diversity statistics. **Table S11.** Inter-genomic correlations of diversity statistics. **Table S19.** Genomic regions showing the evidence of selection. **Table S20.** Percentiles of test statistic distribution around domestication and local adaptation genes. **Table S22.** Overlap of selective sweeps identified using different methods. **Table S23.** Distribution of over-represented PFAM domains. **Table S24.** Proportion of overlapping selective sweeps between the wheat genomes. **Figure S1.** PCA and NJ tree of the wheat diversity panel. **Figure S2.** Wheat genome sequences targeted by the WEC assay. **Figure S3.** Summary of read mapping. **Figure S4.** Selection of alignment parameters for bowtie. **Figure S5.** Flowchart of data processing. **Figure S6.** Variant distribution among various genomic features. **Figure S7.** Estimation of varinat calling error rates for singletons. **Figure S8.** Distribution of indel sizes. **Figure S9.** Enrichment of different functional classes of variants over synonymous variants. **Figure S10 - S15.** Diversity distribution along the wheat chromosomes. **Figure S16.** Genetic differentiation between cultivars and landraces. **Figure S17.** GWAS using imputed and non-imputed datasets. **Figure S18.** Manhattan plot of the XP-CLR statistics. **Figure S19.** The proportion of wheat lines in our sample that have the high-PHS variants in the both genomes of the overlapping homoeologous regions. **Figure S20.** Estimation of reference allele frequency in the GBS and WEC datasets. **Figure S21.** Impact of window size variation on genotype imputation accuracy.

Additional file 2: Table S8. The distribution of non-synonymous and synonymous variants among different genes classified according to GO terms and PFAM domains.

Additional file 3: Table S10. Gene models located in the regions detected by two selection scans (PHS/ XP-CLR, FST/PHS or FST/XP-CLR).

Additional file 4: Table S12. The impact of genotyping probability cutoff on imputation accuracy and proportion of missing data. The table shows the number of correctly imputed genotypes/number of genotypes after filtering using a given genotype probability threshold (proportion of missing in the filtered dataset in %).

Additional file 5: Table S13. The results of disease-resistance phenotyping performed for GWAS.

Additional file 6: Table S14. The results of GWAS of disease resistance in wheat.

Additional file 7: Table S15. The results of validation of three GWAS SNPs by mapping in the populations of recombinant inbred lines.

Additional file 8: Table S16. The comparison of marker-trait associations around the locus on chromosome 7A conferring resistance to stem rust pathogen.

Additional file 9: Table S17. The targets of selection identified in the PHS scan.

Additional file 10: Table S18. The targets of selection identified using the XP-CLR scan.

Additional file 11: Table S21. The overlap of selective sweep regions with published marker-trait associations detected for major agronomic traits in wheat.

Competing interests

JAJ recognizes a competing interest as an employee of Roche NimbleGen Inc.

Authors' contributions

KJ and SW performed most of the data analyses with support from RM, CP, CT and TC. AH, NH, LG, CP, EA, and JJ developed and tested the wheat exome capture assay. LT and JD performed selection of wheat lines and participated in manuscript writing. MH, LG, and AH assessed the accuracy of genotype calling and participated in manuscript writing. AA, YL, CS, KW, and AS performed sequence capture and DNA sequencing and contributed to preparation of manuscript. DW, KLF, and MH performed genotyping of GWAS panel. UKB, HSB, and MJH performed GWAS. EA proposed the idea, coordinated data analyses, and wrote the first draft of the manuscript with the help from KJ. The gene annotations, CSS contig assemblies, and CSS contig order along the chromosomes are provided by the IWGSC (www.wheatgenome.org). All authors read and approved the final manuscript.

Authors' information

International Wheat Genome Sequencing Consortium.

Acknowledgments

This project was supported by the National Research Initiative Competitive Grants 2011-68002-30029 (Triticeae-CAP) (JD) and 2012-67013-19401 from the USDA National Institute of Food and Agriculture (EA), by the Bill and Melinda Gates Foundation grant (EA), by Genome Prairie, Genome Canada, Saskatchewan Ministry of Agriculture, Western Grains Research Foundation (CP and PH), by a BBSRC Career Development Fellowship BB/H022333/1 (AH) and Doctoral Training Grant (LG), by the Howard Hughes Medical Institute and the Gordon and Betty Moore Foundation (JD), and by Kansas Agricultural Experiment Station contribution no. 15-338-J. We thank Cyrille Saintenac for help at the initial phases of the project and Daniel Andresen for assistance with the computing resources of KSU Beocat cluster funded by NSF grant ACI-144054.

Author details
[1]Department Plant Pathology, Kansas State University, Manhattan, KS 66506, USA. [2]Integrated Genomics Facility, Kansas State University, Manhattan, KS 66506, USA. [3]Institute of Integrative Biology, University of Liverpool, Liverpool L69 7ZB, UK. [4]Department Plant Sciences, University of Saskatchewan, Saskatoon, SK S7N 5A8, Canada. [5]Department Environment and Primary Industries, Bundoora, VIC 3083, Australia. [6]National Research Council Canada, 110 Gymnasium Place, Saskatoon, SK S7N 0 W9, Canada. [7]Department Botany & Plant Sciences, University of California, Riverside, CA 92521, USA. [8]Department Plant Sciences, University of California, Davis, CA 95616, USA. [9]Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA. [10]Department Plant Sciences & Plant Pathology, Montana State University, Bozeman, MT 59717, USA. [11]Plant Breeding Institute-Cobbitty, The University of Sydney, PMB4011, Narellan, NSW 2567, Australia. [12]Roche NimbleGen, Inc, Madison, WI 53719, USA.

Received: 26 September 2014 Accepted: 4 February 2015
Published online: 26 February 2015

References

1. Dvorak J, McGuire P, Cassidy B. Apparent sources of the A genomes of wheats inferred from the polymorphism in abundance and restriction fragment length of repeated nucleotide sequences. Genome. 1988;30:680–9.
2. Dvorák J, Terlizzi P, Zhang HB, Resta P. The evolution of polyploid wheats: identification of the A genome donor species. Genome. 1993;36:21–31.
3. Kilian B, Ozkan H, Deusch O, Effgen S, Brandolini A, Kohl J, et al. Independent wheat B and G genome origins in outcrossing Aegilops progenitor haplotypes. Mol Biol Evol. 2007;24:217–27.
4. Tanno K-I, Willcox G. How fast was wild wheat domesticated? Science. 2006;311:1886.
5. Luo M-C, Yang Z-L, You FM, Kawahara T, Waines JG, Dvorak J. The structure of wild and domesticated emmer wheat populations, gene flow between them, and the site of emmer domestication. Theor Appl Genet. 2007;114:947–59.
6. Ortiz R, Sayre KD, Govaerts B, Gupta R, Subbarao GV, Ban T, et al. Climate change: can wheat beat the heat? Agric Ecosyst Environ. 2008;126:46–58.
7. Dvorak J, Akhunov ED, Akhunov AR, Deal KR, Luo M-C. 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. 2006;23:1386–96.
8. Dubcovsky J, Dvorak J. Genome plasticity a key factor in the success of polyploid wheat under domestication. Science. 2007;316:1862–6.
9. Akhunov ED, Sehgal S, Liang H, Wang S, Akhunova AR, Kaur G, et al. Comparative analysis of syntenic genes in grass genomes reveals accelerated rates of gene structure and coding sequence evolution in polyploid wheat. Plant Physiol. 2013;161:252–65.
10. Marcussen T, Sandve SR, Heier L, Spannagl M, Pfeifer M, Jakobsen KS, et al. Ancient hybridizations among the ancestral genomes of bread wheat. Science. 2014;345:1250092–2.
11. Nesbitt M, Deutsch S. From staple crop to extinction? The archaeology and history of hulled wheats. In: Padulosi S, Hammer K, Heller J, editors. Int Work Hulled Wheats. Rome: Italy International Plant Genetic Resources Institute; 1996. p. 41–100.
12. McFadden ES, Sears ER. The origin of Triticum spelta and its free-threshing hexaploid relatives. J Hered. 1946;37:81–107.
13. Cavanagh CR, Chao S, Wang S, Huang BE, Stephen S, Kiani S, et al. Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. Proc Natl Acad Sci U S A. 2013;110:8057–62.
14. Wang S, Wong D, Forrest K, Allen A, Chao S, Huang BE, et al. Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. Plant Biotechnol J. 2014;12:787–96.
15. Daetwyler HD, Bansal UK, Bariana HS, Hayden MJ, Hayes BJ. Genomic prediction for rust resistance in diverse wheat landraces. Theor Appl Genet. 2014;127:1795–803.
16. Zanke C, Ling J, Plieske J, Kollers S, Ebmeyer E, Korzun V, et al. Genetic architecture of main effect QTL for heading date in European winter wheat. Front Plant Sci. 2014;5:217.
17. Edwards D, Wilcox S, Barrero RA, Fleury D, Cavanagh CR, Forrest KL, et al. Bread matters: a national initiative to profile the genetic diversity of Australian wheat. Plant Biotechnol J. 2012;10:703–8.
18. Winfield MO, Wilkinson PA, Allen AM, Barker GLA, Coghill JA, Burridge A, et al. Targeted re-sequencing of the allohexaploid wheat exome. Plant Biotechnol J. 2012;10:733–42.
19. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, et al. A map of human genome variation from population-scale sequencing. Nature. 2010;467:1061–73.
20. Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, et al. Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat Genet. 2011;43:956–63.
21. Chia J-M, Song C, Bradbury PJ, Costich D, de Leon N, Doebley J, et al. Maize HapMap2 identifies extant variation from a genome in flux. Nat Genet. 2012;44:803–7.
22. Xu X, Liu X, Ge S, Jensen JD, Hu F, Li X, et al. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nat Biotechnol. 2012;30:105–11.
23. Hufford MB, Xu X, van Heerwaarden J, Pyhäjärvi T, Chia J-M, Cartwright RA, et al. Comparative population genomics of maize domestication and improvement. Nat Genet. 2012;44:808–11.
24. Saintenac C, Jiang D, Akhunov ED. Targeted analysis of nucleotide and copy number variation by exon capture in allotetraploid wheat genome. Genome Biol. 2011;12:R88.
25. Saintenac C, Jiang D, Wang S, Akhunov E. Sequence-based mapping of the polyploid wheat genome. G3 (Bethesda). 2013;3:1105–14.
26. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One. 2011;6:e19379.
27. International Wheat Genome Sequencing Consortium. A chromosome-based draft sequence of the hexaploid bread wheat genome. Science. 2014;345:1251788.
28. Choulet F, Alberti A, Theil S, Glover N, Barbe V, Daron J, et al. Structural and functional partitioning of bread wheat chromosome 3B. Science. 2014;345:1249721–1.
29. Brenchley R, Spannagl M, Pfeifer M, Barker GL, D'Amore R, Allen AM, et al. Analysis of the bread wheat genome using whole-genome shotgun sequencing. Nature. 2012;491:705–10.
30. Akhunov ED, Akhunova AR, Anderson OD, Anderson JA, Blake N, Clegg MT, et al. Nucleotide diversity maps reveal variation in diversity among wheat genomes and chromosomes. BMC Genomics. 2010;11:702.
31. Uauy C, Paraiso F, Colasuonno P, Tran RK, Tsai H, Berardi S, et al. A modified TILLING approach to detect induced mutations in tetraploid and hexaploid wheat. BMC Plant Biol. 2009;9:115.
32. Akhunova AR, Matniyazov RT, Liang H, Akhunov ED. Homoeolog-specific transcriptional bias in allopolyploid wheat. BMC Genomics. 2010;11:505.
33. Pfeifer M, Kugler KG, Sandve SR, Zhan B, Rudi H, Hvidsten TR, et al. Genome interplay in the grain transcriptome of hexaploid bread wheat. Science. 2014;345:1250091–1.
34. Chen ZJ. Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. Annu Rev Plant Biol. 2007;58:377–406.
35. Hovav R, Udall JA, Chaudhary B, Rapp R, Flagel L, Wendel JF. Partitioned expression of duplicated genes during development and evolution of a single cell in a polyploid plant. Proc Natl Acad Sci U S A. 2008;105:6191–5.
36. Veitia RA, Bottani S, Birchler JA. Cellular reactions to gene dosage imbalance: genomic, transcriptomic and proteomic effects. Trends Genet. 2008;24:390–7.
37. Clark RM, Schweikert G, Toomajian C, Ossowski S, Zeller G, Shinn P, et al. Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. Science. 2007;317:338–42.
38. Ratnaparkhe MB, Wang X, Li J, Compton RO, Rainville LK, Lemke C, et al. Comparative analysis of peanut NBS-LRR gene clusters suggests evolutionary innovation among duplicated domains and erosion of gene microsynteny. New Phytol. 2011;192:164–78.
39. Birch PRJ, Rehmany AP, Pritchard L, Kamoun S, Beynon JL. Trafficking arms: oomycete effectors enter host plant cells. Trends Microbiol. 2006;14:8–11.
40. Kim S, Plagnol V, Hu TT, Toomajian C, Clark RM, Ossowski S, et al. Recombination and linkage disequilibrium in Arabidopsis thaliana. Nat Genet. 2007;39:1151–5.

41. Gore MA, Chia J-M, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL, et al. A first-generation haplotype map of maize. Science. 2009;326:1115–7.
42. Wang J, Luo M-C, Chen Z, You FM, Wei Y, Zheng Y, et al. Aegilops tauschii single nucleotide polymorphisms shed light on the origins of wheat D-genome genetic diversity and pinpoint the geographic origin of hexaploid wheat. New Phytol. 2013;198:925–37.
43. Berkman PJ, Visendi P, Lee HC, Stiller J, Manoli S, Lorenc MT, et al. Dispersion and domestication shaped the genome of bread wheat. Plant Biotechnol J. 2013;11:564–71.
44. Devos KM, Dubcovsky J, Dvorak J, Chinoy CN, Gale MD. Structural evolution of wheat chromosomes 4A, 5A, and 7B and its impact on recombination. Theor Appl Genet. 1995;91:282–8.
45. Jia J, Zhao S, Kong X, Li Y, Zhao G, He W, et al. Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation. Nature. 2013;496:91–5.
46. Dvorak J, Luo MC, Yang ZL, Zhang HB. The structure of the Aegilops tauschii genepool and the evolution of hexaploid wheat. TAG Theor Appl Genet. 1998;97:657–70.
47. Hudson RR. Two-locus sampling distributions and their application. Genetics. 2001;159:1805–17.
48. International HapMap Consortium. A haplotype map of the human genome. Nature. 2005;437:1299–320.
49. Pritchard JK, Przeworski M. Linkage disequilibrium in humans: models and data. Am J Hum Genet. 2001;69:1–14.
50. Zegeye H, Rasheed A, Makdis F, Badebo A, Ogbonnaya FC. Genome-wide association mapping for seedling and adult plant resistance to stripe rust in synthetic hexaploid wheat. PLoS One. 2014;9:e105593.
51. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. Genomic scans for selective sweeps using SNP data. Genome Res. 2005;15:1566–75.
52. Chen H, Patterson N, Reich D. Population differentiation as a test for selective sweeps. Genome Res. 2010;20:393–402.
53. Toomajian C, Hu TT, Aranzana MJ, Lister C, Tang C, Zheng H, et al. A nonparametric test reveals selection for rapid flowering in the Arabidopsis genome. PLoS Biol. 2006;4:e137.
54. Simons KJ, Fellers JP, Trick HN, Zhang Z, Tai Y-S, Gill BS, et al. Molecular characterization of the major wheat domestication gene Q. Genetics. 2006;172:547–55.
55. Faris JD, Zhang Z, Chao S. Map-based analysis of the tenacious glume gene Tg-B1 of wild emmer and its role in wheat domestication. Gene. 2014;542:198–208.
56. Faris JD, Zhang Q, Chao S, Zhang Z, Xu SS. Analysis of agronomic and domestication traits in a durum × cultivated emmer wheat population using a high-density single nucleotide polymorphism-based linkage map. Theor Appl Genet. 2014;127:2333–48.
57. Tzarfati R, Barak V, Krugman T, Fahima T, Abbo S, Saranga Y, et al. Novel quantitative trait loci underlying major domestication traits in tetraploid wheat. Mol Breed. 2014;34:1613–28.
58. Neumann K, Kobiljski B, Dencic S, Varshney R, Borner A. Genome-wide association mapping: a case study in bread wheat (Triticum aestivum L.). Mol Breed. 2011;27:37–58.
59. Rasheed A, Xia X, Ogbonnaya F, Mahmood T, Zhang Z, Mujeeb-Kazi A, et al. Genome-wide association for grain morphology in synthetic hexaploid wheats using digital imaging analysis. BMC Plant Biol. 2014;14:128.
60. Edae EA, Byrne PF, Haley SD, Lopes MS, Reynolds MP. Genome-wide association mapping of yield and yield components of spring wheat under contrasting moisture regimes. Theor Appl Genet. 2014;127:791–807.
61. Cormier F, Le Gouis J, Dubreuil P, Lafarge S, Praud S. A genome-wide identification of chromosomal regions determining nitrogen use efficiency components in wheat (Triticum aestivum L.). Theor Appl Genet. 2014;127:2679–93.
62. Adhikari T, Gurung S, Hansen J, Jackson E, Bonman J. Association mapping of quantitative trait loci in spring wheat landraces conferring resistance to bacterial leaf streak and spot blotch. Plant Genome. 2012;5:1–16.
63. Gurung S, Mamidi S, Bonman JM, Xiong M, Brown-Guedira G, Adhikari TB. Genome-wide association study reveals novel quantitative trait Loci associated with resistance to multiple leaf spot diseases of spring wheat. PLoS One. 2014;9:e108179.
64. Van der Biezen EA, Jones JD. The NB-ARC domain: a novel signalling motif shared by plant resistance gene products and regulators of cell death in animals. Curr Biol. 1998;8:R226–7.
65. Qin W, Zhao G-Y, Qu Z-C, Zhang L-C, Duan J-L, Li A-L, et al. Identification and analysis of TaWRKY34 gene induced by wheat powdery mildew (Blumeria graminis f. sp. tritici). ACTA Agron Sin. 2010;36:249–55.
66. Fleury D, Jefferies S, Kuchel H, Langridge P. Genetic and genomic tools to improve drought tolerance in wheat. J Exp Bot. 2010;61:3211–22.
67. Guo P, Baum M, Grando S, Ceccarelli S, Bai G, Li R, et al. Differentially expressed genes between drought-tolerant and drought-sensitive barley genotypes in response to drought stress during the reproductive stage. J Exp Bot. 2009;60:3531–44.
68. Van Houtte H, Vandesteene L, López-Galvis L, Lemmens L, Kissel E, Carpentier S, et al. Overexpression of the trehalase gene AtTRE1 leads to increased drought stress tolerance in Arabidopsis and is involved in abscisic acid-induced stomatal closure. Plant Physiol. 2013;161:1158–71.
69. Jang I-C, Oh S-J, Seo J-S, Choi W-B, Song SI, Kim CH, et al. Expression of a bifunctional fusion of the Escherichia coli genes for trehalose-6-phosphate synthase and trehalose-6-phosphate phosphatase in transgenic rice plants increases trehalose accumulation and abiotic stress tolerance without stunting growth. Plant Physiol. 2003;131:516–24.
70. Yan L, Loukoianov A, Tranquilli G, Helguera M, Fahima T, Dubcovsky J. Positional cloning of the wheat vernalization gene VRN1. Proc Natl Acad Sci U S A. 2003;100:6263–8.
71. Cockram J, Jones H, Leigh FJ, O'Sullivan D, Powell W, Laurie DA, et al. Control of flowering time in temperate cereals: genes, domestication, and sustainable productivity. J Exp Bot. 2007;58:1231–44.
72. Worland AJ, Börner A, Korzun V, Li WM, Petrovíc S, Sayers EJ. The influence of photoperiod genes on the adaptability of European winter wheats. Euphytica. 1998;100:385–94.
73. Loukoianov A, Yan L, Blechl A, Sanchez A, Dubcovsky J. Regulation of VRN-1 vernalization genes in normal and transgenic polyploid wheat. Plant Physiol. 2005;138:2364–73.
74. Plant Genome and Systems Biology Website [http://mips.helmholtz-muenchen.de/plant/wheat/uk454survey/download/index.jsp]
75. NimbleGen Website [http://www.nimblegen.com/products/seqcap/ez/designs/]
76. Mascher M, Richmond TA, Gerhardt DJ, Himmelbach A, Clissold L, Sampath D, et al. Barley whole exome capture: a tool for genomic research in the genus Hordeum and beyond. Plant J. 2013;76:494–505.
77. Henry IM, Nagalakshmi U, Lieberman MC, Ngo KJ, Krasileva KV, Vasquez-Gross H, et al. Efficient genome-wide detection and cataloging of EMS-induced mutations using exome capture and next-generation sequencing. Plant Cell. 2014;26:1382–97.
78. Poland JA, Brown PJ, Sorrells ME, Jannink J-L. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One. 2012;7:e32253.
79. Patel RK, Jain M. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PLoS One. 2012;7:e30619.
80. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10:R25.
81. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
82. Kent WJ. BLAT-the BLAST-like alignment tool. Genome Res. 2002;12:656–64.
83. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.
84. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.
85. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.
86. Picard Program Website. [http://broadinstitute.github.io/picard/]
87. Choulet F, Wicker T, Rustenholz C, Paux E, Salse J, Leroy P, et al. Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. Plant Cell. 2010;22:1686–701.
88. De Baets G, Van Durme J, Reumers J, Maurer-Stroh S, Vanhee P, Dopazo J, et al. SNPeffect 4.0: on-line prediction of molecular and structural effects of protein-coding variants. Nucleic Acids Res. 2012;40:D935–9.
89. Conesa A, Götz S. Blast2GO: A comprehensive suite for functional analysis in plant genomics. Int J Plant Genomics. 2008;2008:619832.

90. Ling H-Q, Zhao S, Liu D, Wang J, Sun H, Zhang C, et al. Draft genome of the wheat A-genome progenitor Triticum urartu. Nature. 2013;496:87–90.
91. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics. 2003;164:1567–87.
92. Jakobsson M, Rosenberg NA. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics. 2007;23:1801–6.
93. Arnold B, Corbett-Detig RB, Hartl D, Bomblies K. RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. Mol Ecol. 2013;22:3179–90.
94. Mayer KFX, Martis M, Hedley PE, Simková H, Liu H, Morris JA, et al. Unlocking the barley genome by chromosomal and comparative genomics. Plant Cell. 2011;23:1249–63.
95. Qi LL, Echalier B, Chao S, Lazo GR, Butler GE, Anderson OD, et al. A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat. Genetics. 2004;168:701–12.
96. Mayer KFX, Waugh R, Brown JWS, Schulman A, Langridge P, Platzer M, et al. A physical, genetic and functional sequence assembly of the barley genome. Nature. 2012;491:711–6.
97. LDHat Program Website [http://ldhat.sourceforge.net]
98. Mathews DJ, Kashuk C, Brightwell G, Eichler EE, Chakravarti A. Sequence variation within the fragile X locus. Genome Res. 2001;11:1382–91.
99. GMAP Program Website [http://research-pub.gene.com/gmap/]
100. PFAM Database Website [http://pfam.sanger.ac.uk/]
101. HMMER Website [http://hmmer.janelia.org]
102. Paradis E, Claude J, Strimmer K. APE: Analyses of phylogenetics and evolution in R language. Bioinformatics. 2004;20:289–90.
103. Sohail Q, Shehzad T, Kilian A, Eltayeb AE, Tanaka H, Tsujimoto H. Development of diversity array technology (DArT) markers for assessment of population structure and diversity in Aegilops tauschii. Breed Sci. 2012;62:38–45.
104. Browning BL, Browning SR. Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics. 2013;194:459–71.
105. Miller T, Ambrose M, Reader S. The Watkins collection of landrace derived wheats. In: Caligari PDS, Brandham PE, editors. Wheat taxon legacy of John Percival. London: Linnean Society; 2001. p. 113–20.
106. McIntosh RA, Wellings CR, Park RF. Wheat rusts: an atlas of resistance genes. Melbourne: CSIRO Press; 1995.
107. Bariana HS, Miah H, Brown GN, Willey N, Lehmensiek A. Molecular mapping of durable rust resistance in wheat and its implication in breeding. In: Buck HT, Nisi JE, Salomón N, editors. Wheat production in stressed environments. Rotterdam: Springer Netherlands; 2007. p. 723–8.
108. Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, et al. GAPIT: genome association and prediction integrated tool. Bioinformatics. 2012;28:2397–9.
109. Pasam RK, Sharma R, Malosetti M, van Eeuwijk FA, Haseneyer G, Kilian B, et al. Genome-wide association studies for agronomical traits in a world wide spring barley collection. BMC Plant Biol. 2012;12:16.
110. Forrest KL, Pujol V, Bulli P, Pumphrey M, Wellings C, Herrera-Foessel S, et al. Development of a SNP marker assay for the Lr67 gene of wheat using a genotyping by sequencing approach. Mol Breed. 2014;34:2109–18.
111. Randhawa M, Bansal U, Valárik M, Klocová B, Doležel J, Bariana H. Molecular mapping of stripe rust resistance gene Yr51 in chromosome 4AL of wheat. Theor Appl Genet. 2014;127:317–24.
112. Michelmore RW, Paran I, Kesseli RV. Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. Proc Natl Acad Sci U S A. 1991;88:9828–32.
113. Hyten DL, Smith JR, Frederick RD, Tucker ML, Song Q, Cregan PB. Bulked segregant analysis using the GoldenGate assay to locate the Rpp3 locus that confers resistance to soybean rust in soybean. Crop Sci. 2008;49:265–71.
114. Project Website describing datasets used in the paper [http://wheatgenomics.plantpath.ksu.edu/hapmap]
115. USDA GrainGenes Website [http://wheat.pw.usda.gov/pubs/2015/Akhunov]
116. URGI Website [http://wheat-urgi.versailles.inra.fr/Seq-Repository/Variations]