Running head: Evolution of genes in polyploid genome

Author for correspondence:
Eduard Akhunov
Department of Plant Pathology
4024 Throckmorton Plant Sciences Center
Kansas State University
Manhattan, KS 66506
e-mail: eakhunov@ksu.edu
phone: 785-532-1342
fax: 785-532-5692

Research area: Genome Analysis
Comparative analysis of syntenic genes in grass genomes reveals accelerated rates of gene structure and coding sequence evolution in polyploid wheat

Eduard D. Akhunov\textsuperscript{2,*}, Sunish Sehgal\textsuperscript{2}, Hanquan Liang, Shichen Wang, Alina R. Akhunova, Gaganpreet Kaur, Wanlong Li, Kerrie L. Forrest, Deven See, Hana Šimková, Yaqin Ma, Matthew J. Hayden, Mingcheng Luo, Justin D. Faris, Jaroslav Doležel, Bikram S. Gill

Department of Plant Pathology, Kansas State University, U.S.A. (E.D.A., S.S., H.L., S.W., A.R.A., G.K., B.S.G.)
Integrated Genomics Facility, Kansas State University, U.S.A. (H.L., A.R.A.)
Plant Science, Biology & Microbiology, South Dakota State University, Box 2140D, Brookings, SD 57007, U. S. A. (W.L.)
Department of Primary Industries Victoria, Victorian AgriBiosciences Center, 1 Park Drive, Bundoora, VIC 3083, Australia (K.L.F., M.J.H.)
USDA-Agricultural Research Service, Western Regional Small Grains Genotyping Lab, Johnson Hall, WSU, Pullman, WA, U.S.A. (D.S.)
Centre of the Region Haná for Biotechnological and Agricultural Research, Institute of Experimental Botany, Sokolovska 6, CZ-77200 Olomouc, Czech Republic (H. Š., J.D.)
Department of Plant Sciences, University of California, Davis, CA 95616, U.S.A. (Y.M., M.L.)
Department of Botany & Plant Sciences, University of California, Riverside, CA 92521, U.S.A. (Y.M.)
USDA-Agricultural Research Service, Northern Crop Science Laboratory, 1605 Albrecht BLVD, Fargo, ND 58102-2765, U.S.A. (J.D.F.)
King Abdulaziz University, Faculty of Science, Genomics and Biotechnology Section, Department of Biological Sciences, Jeddah 21589 Saudi Arabia (B.S.G.)
1 This project is funded by the USDA National Institute of Food and Agriculture grants 2006-35604-17248 (BSG), 2008-35300-04588 (BSG), CRIS0219050 (EDA), Kansas Wheat Commission (EDA), the Grains Research and Development Corporation, Australia (MH), HŠ and JD were supported by the Ministry of Education, Youth and Sports of the Czech Republic and the European Regional Development Fund (Operational Programme Research and Development for Innovations No. ED0007/01/01). Partial research support also came from KAES grants to WGGRC (BSG) and Integrated Genomics Facility (AA).

2 These authors made equal contribution to this work

* Author for correspondence: Eduard Akhunov (eakhunov@ksu.edu)

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Eduard D. Akhunov (eakhunov@ksu.edu).
Abstract

Cycles of whole genome duplication (WGD) and diploidization are hallmarks of eukaryotic genome evolution and speciation. Polyploid wheat (*Triticum aestivum*) has had a massive increase in genome size largely due to recent WGDs. How these processes may impact the dynamics of gene evolution was studied by comparing the patterns of gene structure changes, alternative splicing (AS), and codon substitution rates among wheat and the model grass genomes. In orthologous gene sets, significantly more acquired and lost exonic sequences were detected in wheat than in model grasses. In wheat, thirty five percent of these gene structure rearrangements resulted in frameshift mutations and premature termination codons (PTCs). An increased codon mutation rate in the wheat lineage compared to *Brachypodium* was found for 17% of orthologs. Discovery of PTCs in 38% of expressed genes was consistent with ongoing pseudogenization of the wheat genome. The rates of AS within the individual wheat subgenomes (21 - 25%) were similar to diploid plants. However, we uncovered a high level of AS pattern divergence between the duplicated homoeologous copies of genes. Our results are consistent with the accelerated accumulation of AS isoforms, non-synonymous mutations and gene structure rearrangements in the wheat lineage, likely due to genetic redundancy created by WGDs. Whereas these processes mostly contribute to degeneration of a duplicated genome and its diploidization, they have the potential to facilitate the origin of new functional variation, which, upon selection in the evolutionary lineage, may play an important role in the origin of novel traits.
Introduction

Evolution of protein coding sequences, gene exon-intron structure and alternative splicing define the diversity of proteome structure and function (Koonin and Wolf, 2010). Multiple studies performed mostly on diploid species or ancient polyploids suggested that gene duplication plays a major role in coding sequence evolution (Fan et al., 2008; Lynch and Conery, 2000; Ohno, 1970; Roux and Robinson-Rechavi, 2011). According to Ohno’s hypothesis, gene duplications lead to relaxed selection, which allows duplicated genes to accumulate mutations (Ohno, 1970). The evolutionary fate of redundant genes is defined by selection that can drive beneficial alleles to fixation resulting in the creation of novel gene functions through the processes of neo- and sub-functionalization (Ohno, 1970; Force et al., 1999; Hughes, 1994). Alternatively, genes can become non-functional and eventually lost from evolutionary lineages.

The discovery of paleopolyploidy in most of the analyzed plant genomes has highlighted the importance of WGD in the evolutionary success of flowering plants (Soltis et al., 2008; Patterson et al., 2003; Van de Peer et al., 2009). In ancient polyploids, it was shown that retained copies of duplicated genes are under selective constraint as estimated from the distribution of Ka/Ks values (Lin et al., 2010), or they show a high level of connectivity in gene networks (Severin et al., 2011). In maize, the probability of gene loss was associated with the reduced contribution of a duplicated copy to the total level of gene expression (Schnable et al., 2011). These facts are consistent with the “gene balance hypothesis” suggesting that dosage-sensitive genes are very likely to be retained after WGD (Birchler and Veitia, 2007; Freeling, 2009). While these studies have demonstrated the long-term consequences of polyploidization, the impact of recent WGD on genome coding potential is still not fully understood. The analysis of artificially synthesized polyploid plants suggests that WGD results in “genomic shock” accompanied by structural rearrangements (Kashkush et al., 2002; Gaeta et al., 2009), activation of transposons (Kashkush et al., 2003), expression changes (Akhunova et al., 2010; Pumphrey et al., 2009) and epigenetic modifications (Ni et al., 2009; Chen et al., 2008). These processes can become a source of new transgressive variation and provide the molecular basis for functional evolution (Chen, 2007; Comai, 2005; Osborn et al., 2003). Among plant species, wheat provides a unique opportunity to study early phases of duplicated gene evolution after WGD, the impact of increased genome size on the evolutionary dynamic of gene space, and understand how these processes shaped the coding potential of the wheat genome.
The hexaploid wheat genome (comprised of about 16,000 Mbp) resulted from the union of three diploid grass species. The first hybridization event, which resulted in the origin of tetraploid wheat, occurred 0.36-0.5 million years ago (MYA) (Huang et al., 2002; Dvorak and Akhunov, 2005; Chalupska et al., 2008; Dvorak et al., 1993; Dvorak and Zhang, 1990) and the second hybridization event, which resulted in the origin of hexaploid bread wheat, occurred about 8,000 years ago (Kihara, 1944; McFadden and Sears, 1944; 1946; Nesbitt and Samuel, 1996). Early analyses of EST maps show that about 20% of wheat genes are inter-chromosomally duplicated (Akhunov et al., 2003; Qi et al., 2004). Comparative analysis of model grass genomes (International Rice Genome Sequencing Project, 2005; Paterson et al., 2009; International Brachypodium Initiative, 2010; Schnable et al., 2009) revealed a more detailed picture of wheat genome structure evolution involving segmental duplications, chromosome inversions, translocations and small scale structural rearrangements (Akhunov et al., 2003; Mayer et al., 2011; Wicker et al., 2011; Salse et al., 2008; Choulet et al., 2010; Devos, 2010; Gu et al., 2004; Gu et al., 2006). These studies demonstrated an acceleration of genome structure evolution in wheat, which also had an impact on the evolution of the gene repertoire (Akhunov et al., 2007; Massa et al., 2011; Salse et al., 2008). However, the impact of these processes on the dynamics of gene structure evolution in the wheat genome has not been thoroughly investigated. The evolution of gene structure in eukaryotes was previously shown to be associated with exon shuffling, retroposition, transposon recruitment, or gene fusion (Long et al., 2003). Studies performed in plant species with small genomes such as rice (Wang et al., 2006; Fan et al., 2008) and poplar (Zhu et al., 2009) revealed the importance of exon shuffling, and retroposition in diversification of the coding potential of plant genomes. While recent studies showed evidence of transposon recruitment in the wheat genome (Akhunov et al., 2007; Wicker et al., 2011), the role of other mechanisms in gene structure evolution was not characterized.

Alternative splicing is commonly used by eukaryotic organisms to enhance proteome diversity and regulate transcript abundance (Ast, 2004; Reddy, 2007). The true scale of AS in plant genomes began to emerge only recently with the availability of deep RNA sequencing data. These studies have predicted >20% of genes in maize and rice, and >40% in *Arabidopsis*, undergo AS (Barbazuk et al., 2008; Filichkin et al., 2010; Zhang et al., 2010). In wheat, only a limited number of genes have been characterized for AS events (Baga et al., 1999; Terashima and Takumi, 2009) and the impact of WGD on AS patterns on a genome-wide scale remains
unknown.

Here, we performed intra- and inter-species comparative analyses of the wheat genome and model grass genomes to investigate the dynamics of gene structure, alternative splicing and coding sequence evolution in the wheat lineage. As part of the International Wheat Genome Sequencing Consortium effort (www.wheatgenome.org), we developed a high-density syntenic map of wheat chromosome 3A (827 Mb) using 9.2x coverage sequence data generated with 454 sequencing technology. This data was combined with publicly accessible wheat genomic sequences from the NCBI database to analyze the evolutionary dynamic of gene structure in the polyploid wheat genome and to describe homoeolog-specific patterns of AS. Developed gene models were annotated and used to reveal genes that had lost their function due to the accumulation of PTCs, and genes whose codons accumulate non-synonymous mutations at accelerated rates. Comparative analysis of orthologous gene sets between the syntenic regions of the model grass genomes and wheat was performed to estimate the rate, and to understand molecular mechanisms of gene structure evolution in wheat. The gene structure and patterns of AS were compared between homoeologous genes to assess the extent of functional differentiation between the wheat sub-genomes.

Results

Wheat chromosome 3A assembly

Out of 25.3 million 454 reads (7.65 Gb), 16.9 million (5.2 Gb) were assembled into 240,121 large contigs (≥500 bp; hereafter LC) with a total length of 331 Mb, representing 40% of the physical length of chromosome 3A (see Supplemental Table S1). The remaining 32% of reads were not included in the assembly due to having only partial overlap with contigs (588 Mb), absence of overlap with any other reads (335 Mb), high copy number (715 Mb), and/or artifacts such as chimeric sequences (842 Mb). The reduced level of chromosome coverage comes from the presence of highly abundant repetitive elements assembled into short contigs with deep sequence coverage (see Supplemental Fig. S1). The estimate of N50 for LC was 1,692 bp (Table 1). Using paired-end sequence reads, nearly 60% (195 Mb) of LCs were assembled into 38,036 scaffolds with an N50 size of 6,013 bp (Table 1) and a total gap length of 17 Mb. The final dataset included both scaffolds (≥2,000 bp) and LC (not included into the scaffolds) with a total length of 348 Mb (Table 1).
The fraction of targeted chromosomal arms in flow-sorted material was 86-92\% (see Supplemental Table S2). We confirmed these estimates by comparing LC sequences with the sequences of single-copy wheat ESTs mapped to the deletion-bins on wheat chromosome 3 (wheat.pw.usda.gov/wEST). A total of 374 ESTs were selected for this analysis, of which 87\% had significant BLASTN hits with chromosome 3A contigs and scaffolds (alignment length ≥ 100, identity ≥ 90\%). Comparison of 5,288 ESTs mapped to the wheat chromosomes other than chromosome 3 showed only 705 significant BLASTN hits suggesting that contamination of the flow-sorted material was less than 13\%. The distribution of alignment depth in the contig assemblies was consistent with genome coverage of ~7-8x (see Supplemental Fig. S2). This coverage is slightly lower than the expected 9.3x coverage and can be, at least partially, explained by contamination from other wheat chromosomes.

There is a possibility that the DNA contamination would impact the results of downstream analyses due to the presence of duplicated genes on other chromosomes including homoeologous genes on chromosomes 3B and 3D. The divergence of these genes from true orthologous genes may bias our estimates of the rates of gene evolution. Assuming that the fraction of chromosome 3A gene homologs located elsewhere in the wheat genome is about 14.7\% (see Supplemental Material), we estimated that, in our dataset, about 1.9\% of coding sequences could be mistaken for true chromosome 3A genes. However, we expect that this bias will likely be low due to the 100-fold lower level of coverage (0.07x) achieved for the contaminating fraction (13\%) of the wheat genome in our sample compared to 7x coverage obtained for chromosome 3A (see Supplemental Material). The low level of coverage obtained for contaminating DNA should reduce the contribution of contaminating reads to consensus sequences of the chromosome 3A contigs and the fraction of contaminating gene sequences assembled into contigs longer than 500 bp (see Supplemental Material).

The impact of flow-sorted chromosome contamination on evolutionary inferences should also be reduced by selecting syntenic genes sharing positional orthology among cereal genomes. However, this approach may also introduce bias in our evolutionary rate estimates because it has been shown that positionally conserved genes may be under stronger selection constraint (Notebaart et al., 2005). On contrast, transposed orthologous genes were shown to evolve faster (Han et al., 2009). In our case, gene selection using positional orthology will most likely result in an underestimation of evolutionary changes in the wheat genome, thereby producing more
The accuracy of chromosome 3A sequence assembly was assessed by comparing masked contigs and scaffolds with previously sequenced BAC clones from the wheat chromosome 3A BAC library. A total of 17 BAC clones (total length of 2,094,375 bp) spanning the region of wheat chromosome 3A homologous to the Fhb1 locus on chromosome 3B, were used in the comparison (see Supplemental Table S3). The error rate estimated from this analysis was 0.07% caused by small-scale indels in the regions of homopolymeric bases (the most common source of error in 454 sequence data) followed by single base errors. This result also suggests that the proportion of errors due to contamination with homoeologous chromosomes 3B and 3D should be relatively low in our sample. Chromosome 3A assemblies covered about 85% of repeat-masked portion of the wheat chromosome 3A BAC clones.

Syntenic gene order (SGO)

The chromosome 3A contigs and scaffolds were ordered based on the syntenic relationships with Brachypodium, rice and sorghum using a strategy similar to that used by Mayer et al. (2011) and described in detail in the Supplemental Material. The conserved syntenic regions of the three model grass genomes were used to recreate a hypothetical order of genes along wheat chromosome 3A (Fig. 1A). A total of 3,646 chromosome 3A contigs/scaffolds were ordered to construct the chromosome 3A SGO map (see Supplemental Table S8). The order of genes inferred using the grass genome synteny was largely consistent with the recombination-based SNP genetic map of Ae. tauschii, the diploid D-genome progenitor of polyploid wheat (Luo et al., 2009; see Supplemental Fig. S7). The SGO map, based on conserved orthologous gene sets, provides a robust framework for investigating the evolutionary relationships of 3A with other cereal genomes (Fig. 1) and among wheat chromosomes (see Supplemental Table S9, Supplemental Fig. S8). In addition, the selection of syntenic genes reduces the impact of flow-sorted chromosome contamination on evolutionary inferences made in our study.

Comparison of bin-mapped wheat ESTs with the genes from the 3A SGO map revealed 249 unique BLASTN hits detecting 10 segmental duplications involving chromosome 3A and other groups of wheat chromosomes (see Supplemental Table S9). There were three duplicated blocks on the chromosome combination w3-w1, two on the w3-w2, three on the w3-w5, and two on the w3-w7 (see Supplemental Fig. S8). Five of the detected duplications are new and five
have been described previously (Salse et al., 2008). Except for the ancient w3-w1 duplication shared between wheat and rice, all the duplicated regions occurred in the wheat lineage. All were shared among the three homoeologous chromosomes of each group suggesting that these duplication events originated before the split of diploid wheat ancestors.

We also used the SGO map to detect the centromeric region of chromosome 3A. We identified 510 SGO contigs/scaffolds mapped to the 3AS arm and 858 SGO contigs/scaffolds mapped to the 3AL arm. Eighty-two contigs and scaffolds, starting with SGO1163 and ending with SGO1394, mapped to either of the chromosome arms and spanned 10 Mb on the homologous region of rice chromosome 1 (see Supplemental Table S10). We further refined the location of the centromere using wheat ESTs (BF485348, BE637878, BG313557, BE404580) previously used for comparative analysis of centromere locations in wheat, rice and Brachypodium (Qi et al., 2010). Based on this analysis, the location of the centromere on wheat chromosome 3A was further reduced to a region flanked by SGO1221 and SGO1367, which corresponds to a homologous 7.7 Mb region in the rice genome (Fig. 1B).

**Evidence-based gene prediction**

Evidence-based annotation of gene exon-intron structure was performed using contig assemblies including chromosome 3A contigs generated in our study and 371 BAC contig sequences available in the NCBI database (see Supplemental Table S11). The latter dataset included 12 and 24 contigs originating from chromosome 3B (Choulet et al., 2010) and 3D (Bartoš et al., 2012), respectively, combined with contigs distributed across the wheat genome. The total length of the BAC contigs was 50.3 Mb, out of which 17.9 Mb and 1.6 Mb represented sequences from wheat chromosome 3B (Choulet et al., 2010) and 3D (Bartoš et al., 2012), respectively. The length of the wheat contigs varied from 20 Kb to 3.1 Mb.

For predicting gene models in the wheat genome, we used only those transcript sequences (see Supplemental Table S12) that met the similarity threshold (98.5%) established for separating homoeologs (see Materials and Methods for details). Assessment of previously published data for the distribution of inter-genomic sequence similarity levels in the genic regions of the wheat genome (Akhunov et al., 2010) suggests that this threshold can separate wheat homoeologs with 94.3% accuracy (see Supplemental Material). The remaining 5.7% of predicted genes could result from erroneously aligned transcripts caused by low levels of inter-
genome divergence. A total of 6,366 complete or partial gene structures were predicted on chromosome 3A contigs (Table 2). The average sizes of predicted introns (131 bp) and exons (171 bp) were also similar to previous estimates obtained by BAC sequencing (Choulet et al., 2010), suggesting that predictions of intron-exon structure using chromosome 3A assemblies do not result in biased estimates. By aligning 31,292 transcripts to wheat BAC contigs distributed across the wheat genome, we detected 1,049 genes, out of which 391 and 91 were predicted on the wheat chromosome 3B and 3D contigs, respectively (Table 2).

**Alternative splicing in the polyploid wheat genome**

AS is one of the major mechanisms used by eukaryotes to increase transcriptome and proteome diversity. We used extensive transcriptome sequence data assembled into homoeolog-specific transcripts to gain insights into the patterns of AS in the polyploid wheat genome (Table 3; see Supplemental Fig. S9). The use of 454 sequence data obtained from normalized cDNA libraries should ensure deep sampling of unique alternatively-spliced transcripts in the genome. More than 24% of genes predicted on wheat chromosomes 3A, 3B and 3D showed evidence of AS (Table 3). Only a slightly lower fraction of AS events (21%) was predicted in a sample of BAC clones distributed across the wheat genome, suggesting that trends observed on wheat chromosomes 3A, 3B and 3D do not depart dramatically from that of the whole genome.

Consistent with observations in other plant systems (Lu et al., 2010; Barbazuk et al., 2008; Marquez et al., 2012), the major classes of AS variants in the wheat genome were retained intron (RI, 30-37%) and alternative acceptor (AA, 26-43%) variants, followed by alternative donor (AD, 22-28%) and skipped exon (SE, 4-7%) spliced variants. In the majority (91%) of cases, intron retention in wheat resulted in PTCs.

**Evolution of AS forms**

Although, previous studies demonstrated that, in most cases, duplicated copies of genes in polyploid wheat are expressed (Akhunova et al., 2010; Bottley and Koebner, 2008; Nomura et al., 2005), the degree of AS pattern similarity between them remains unknown. We estimated AS divergence using a set of 86 and 20 homoeologous gene pairs identified between wheat chromosomes 3A and 3B (Choulet et al., 2010), and 3A and 3D (Bartoš et al., 2012), respectively. Even considering a 5.7% false positive rate in homoeologous transcript alignment,
the wheat genomes showed a substantial levels of AS divergence. Among identified gene pairs, 33 (37%) and 8 (40%) showed evidence of AS in the 3A-3B and 3A-3D genome comparisons, respectively. (Table 4; see Supplemental Table S13). We found that 39%, 28%, 28% and 5% of AS events in the 3A-3B comparison, and 32%, 39%, 21% and 7% of AS events in the 3A-3D comparison belong to RI, AD, AA and SE types, respectively (Table 4). The proportion of the AS events different between the homoeologous gene pairs in the 3A-3B and 3A-3D comparisons was 42% and 61%, respectively. The divergence between the wheat genomes was highest for RI events and smallest for AD events.

Previously, it was shown that the depth of sequence data impacts the probability of recovering AS events (Barbazuk et al., 2008). To assess the possibility of this type of bias in the estimation of AS divergence between homoeologs, we compared the proportion of AS events between shared and homoeolog-specific AS isoforms that are supported by a single EST/transcript. We found that these proportions among homoeolog-specific (7%) and shared (5%) AS events were similar, suggesting that the depth of sequencing did not have significant impact on our estimates of AS divergence between the wheat genomes.

Next, we compared the level of AS conservation observed in our study to previous estimates of AS conservation between rice and Arabidopsis (Wang and Brendel, 2006). Using a similar approach applied to 643 wheat-rice orthologous gene pairs, we identified 2,292 introns. Out of these gene pairs, 264 and 239 were alternatively spliced in wheat and rice, respectively. Among the 232 conserved introns identified in these genes, only 19 (8%) showed the same type of AS in both rice and wheat. Intron retention was the most conserved type of AS event between rice and wheat, with only 5% of the genes showing intron retention in both rice and wheat (Table 5). This estimate is two times lower than that obtained for Arabidopsis and rice (Wang and Brendel, 2006).

**Evolution of gene structure**

To investigate the dynamics of gene structure evolution in the polyploid wheat genome, we compared protein sequences of orthologous genes between wheat, Brachypodium and rice (Fig. 2A). A total of 1,059 orthologous gene pairs between wheat chromosome 3A and rice, and 839 orthologous gene pairs between wheat chromosome 3A and Brachypodium, were analyzed (Fig. 2B, see Supplemental Table S14). In addition, patterns of gene structure conservation and divergence were characterized between homoeologous genes on wheat chromosomes 3A, 3B and
3D and among orthologous gene sets identified in a sample of sequenced BAC clones distributed across the wheat genome and model genomes. Of 958 genes that were predicted in the wheat BAC clones, we identified 169 rice and 194 Brachypodium orthologs. Of the 391 genes predicted in wheat chromosome 3B BAC contigs, 86 had homoeologous copies on wheat chromosome 3A, and out of 91 genes predicted in the chromosome 3D BAC contigs, 20 had duplicates on wheat chromosome 3A.

The proportion of lost and acquired exonic sequences was higher in wheat-rice and wheat-Brachypodium than in rice-Brachypodium comparisons (Fig. 2B). This trend was consistent among all three sets of orthologous genes predicted in wheat chromosomes 3A, 3B, 3D and a sample of BAC clones distributed across the wheat genome. There was an average of 2.5 times fewer acquired exons and 5 times fewer lost exons discovered by comparing rice with Brachypodium than by comparing wheat and either of the two model genomes. Consistent with this observation, the proportion of conserved exons between rice and Brachypodium was from 25% to 37% higher than that in wheat-rice and wheat-Brachypodium comparisons (Fig. 2B). The distribution of exon loss and acquisition events along the wheat genes was similar. The mean of relative locations of exon acquisition and loss events along the coding sequence (from 0 to 1), estimated according to Fawcett et al. (2012) by dividing the distance of event from 5’-end by the total length of coding sequence, was 0.43 and 0.38, respectively.

We identified 68 genes that contained 130 acquired exons (AE; length ≥30 bp) specific to the wheat lineage (see Supplemental Table S15). In a randomly selected sample of 36 AEs, 72% were confirmed by RT-PCR (Supplemental Table S16). Searches of the NCBI database revealed that only 17 exons had significant BLASTN hits to coding sequences (e-value < 1e-10), suggesting that about 13% of genes in the wheat genome could have evolved by exon shuffling. Most of these exons were found only in the Triticeae lineage and showed similarity to barley or wheat ESTs/cDNAs. Nine of the AEs were similar to transposons in the TREP and/or GIRI repetitive databases, suggesting that 7% of exons might have originated by the recruitment of transposable elements (see Supplemental Table S17).

Comparative analysis of gene exon-intron structure revealed that out of 1,975 wheat genes, 168 (8.5%) were subjected to loss of intronic sequences that were otherwise present in the orthologous genes of the model genomes. Sixteen out of 168 genes (10%) were from non-syntenic regions. Among these genes, 105 (63%) contained PTCs, and were therefore,
considered to be pseudogenes. The potential mechanism for the origin of these genes could be retroposition, during which a copy of a parental gene is reverse-transcribed and inserted into a new genomic location (Wang et al., 2006; Kaessmann et al., 2009).

To determine the extent of structural divergence between homoeologous copies of wheat genes, we compared a set of 86 and 20 gene pairs between chromosomes 3A and 3B, and 3A and 3D, respectively. In spite of the recent divergence between the wheat genomes (~2.7 million years ago), these genes demonstrated a remarkable level of structural divergence (Fig. 2B). The proportion of lost and acquired exonic sequences between 3A and 3B (1.5 and 3 times higher, respectively), and 3A and 3D (6.5 and 5.3 times higher, respectively) was higher than that observed in the rice – *Brachypodium* comparison. The proportion of conserved exons between 3A and 3B, and 3A and 3D, respectively, was nearly 2 and 1.7 times higher than that between wheat and the model grass genomes.

**Rates of coding sequence evolution in syntenic regions of grass genomes**

Redundancy created by polyploidy can change the dynamic of coding sequence evolution by relaxing selection and providing opportunities for accumulation of new mutations that may impact gene function. Here we used two complimentary approaches to assess the rates of coding sequence evolution in the wheat genome. First, we assessed the $d_N/d_S$ rates from pair-wise comparisons of 1,022 orthologous sets of genes to identify genes showing evidence of directional selection ($d_N/d_S > 1$; Yang and Nielsen, 2000). The overall distribution of $d_N/d_S$ values in the three possible pair-wise comparisons among wheat, rice and *Brachypodium* was similar and showed no statistically significant differences according to a non-parametric Wilcoxon rank-sum test (Fig. 3A). The estimates of $d_N/d_S$ among pair-wise comparisons were highly correlated (Spearman rank correlation 0.82-0.91, $p < 2.2e^{-16}$), suggesting that the rates of mutations at synonymous and non-synonymous sites were gene-specific (Fig. 3D). Only four genes in the wheat lineage (casein kinase I isoform delta-like, rhomboid family protein, tetratricopeptide repeat protein 5-like; ABC transporter C family) had $d_N/d_S > 1$ (see Supplemental Table S18). BLASTN analysis (similarity $\geq 80\%$, alignment length $\geq 100$ bp) did not detect duplicated copies of these genes on chromosome 3A. Analysis of protein sequences of these genes revealed an accumulation of lineage-specific amino acid changes in wheat, consistent with an increased $d_N/d_S$ ratio. No PTCs were detected in the aligned portions of coding sequences.
To investigate the relationship between the chromosomal position of a gene and the rate of accumulation of non-synonymous mutations, we analyzed the distribution of the \( d_{SN}/d_S \) ratio along wheat chromosome 3A by plotting average \( d_{SN}/d_S \) ratios calculated in a sliding window of 5 genes. High variation in \( d_{SN}/d_S \) estimates was found along chromosome 3A, with only two chromosomal regions showing elevated \( d_{SN}/d_S \). One of these regions was located on the short arm and another one in the pericentromeric region of wheat chromosome 3A (Fig. 3D). Only one gene (tetratricopeptide repeat protein 5-like) in the pericentromeric region had \( d_{SN}/d_S > 1 \).

At the early stages of polyploid evolution, it is probable that many genes can become expressed pseudogenes whose function is impaired by an accumulation of PTCs. To test this possibility and estimate the frequency of non-functional expressed genes in the wheat genome, we identified orthologous ORFs for 1,085 wheat transcripts by performing reciprocal BLASTN comparisons and selecting best hits (>60% of cumulative identity percentage and >70% cumulative alignment length percentage) in the syntenic regions of rice and *Brachypodium* genomes (for details see Supplemental Material). Thirty-eight percent of these transcripts contained PTCs, suggesting that they are processed pseudogenes. No correlation was found between the occurrence of PTCs and upstream homopolymeric repeats (>5 bp) (see Supplemental Material) suggesting that overcall/undercall errors prevalent in 454 data were not the major source of PTC origin. About 36% of genes on 3B and 3D homoeologs have PTCs, which is similar to what we found on chromosome 3A.

The distribution of PTCs in orthologous sets of genes was compared with the distribution of exon loss/acquisition (LE/AE) events and \( d_{SN}/d_S \) estimates. Structurally rearranged genes tended to have more PTCs (56 out of 162) than genes that have not undergone LE/AE (53 out of 860) in the wheat lineage \( (\chi^2 \text{- test, } p < 0.001) \). Most of the PTCs were due to frameshift mutations. After correcting for frameshift, only 11% of genes still showed PTCs, which is comparable to 9.7%, the proportion of PTCs in genes that showed no exon gain or loss. No significant differences in pair-wise \( d_{SN}/d_S \) estimates were found between gene sets grouped according to the presence/absence of PTCs (Wilcoxon Rank Sum Test, \( p > 0.1 \)) or LE/AE events (Wilcoxon Rank Sum Test, \( p > 0.05 \)). The majority of structurally evolved genes did not show evidence for relaxation of selection; \( d_{SN}/d_S \) estimates of only 3.8% of these genes were higher than 0.5.

To gain better insights into the evolutionary dynamics of coding sequences in the wheat
genome, we compared the rates of codon evolution along the wheat and *Brachypodium* lineages using rice as an outgroup (Fig. 3B, C). The codon evolution models allow for explicit modeling of the evolution of a protein sequence by considering codons as units of evolution while taking into account changes in synonymous and non-synonymous sites (Muse and Gaut, 1994), rather than separately estimating substitution rates at the single nucleotide level (Fig. 3B). The relative rate test included a set of 986 orthologous triplets of genes from *Brachypodium*, rice, and wheat. Among these sets, 191 showed statistically significant differences ($p \leq 0.05$) in the rate of codon evolution (see Supplemental Table S19), out of which 167 (87%) showed increased codon mutation rates in the wheat lineage compared to that in *Brachypodium* ($\log_2(v_1/v_2) > 0$; Fig. 3C). This estimate, even taking into account the 1.9% probability of including non-orthologous genes into the analysis due to contamination, is still substantial. On average, the codon mutation rate was 1.4 times higher in wheat than in *Brachypodium*. The Gene Ontology (GO) annotation of genes with elevated rates of non-synonymous substitutions showed that the majority did not have a specific function assigned in the GO database. Those that could be classified (48 genes) belonged to various biological processes. Interestingly, out of the 15 genes that could be classified into GO categories according to their biological function, six were protein kinase members.

**Discussion**

We analyzed evolutionary processes that can impact the coding potential of the recently originated polyploid wheat genome through modifications of gene structure, coding sequence and alternative splicing. More than 370 Mb of wheat genomic sequence was analyzed to predict genes by aligning homoeolog-specific transcript assemblies. More than 6,600 complete and partial gene structures were predicted in chromosome 3A contig assemblies. Comparative genomics was used to establish syntenic relationships between wheat chromosome 3A and model grass genomes and to build a framework for the evolutionary analysis of coding regions. Even though a detectable level of contamination from other wheat chromosomes was found in the flow-sorted 3A material (up to 13%), and errors for separating homoeologous transcripts (5.7%) can introduce some bias in our evolutionary rate estimates, our results suggest that coding regions of the wheat genome have undergone complex historical changes and that the size of the wheat genome and polyploidy are important factors to consider for understanding its dynamic
Evolution.

Expansion of genome size in the Triticeae lineage associated with the proliferation of repetitive elements has impacted not only on the large-scale structural organization of the wheat genome (Massa et al., 2011; Choulet et al., 2010; Gu et al., 2006), but possibly has had a significant impact on gene structure evolution. A high frequency of LTR transposons in the wheat genome showing evidence of recent activity (Choulet et al., 2010; Charles et al., 2008) may have substantially accelerated the rate of retrogene origin. Our results are consistent with this possibility and are also supported by the earlier discovery of new genes in the wheat genome whose origin could be linked with transposon activity (Akhunov et al., 2007; Choulet et al., 2010; Wicker et al., 2010). Other studies also demonstrated that transposon activity could result in the origin of new genes (Lal et al., 2003; Wang et al., 2006; Zhu et al., 2009).

The high frequency of gene structural rearrangements in the wheat genome can be explained by the interplay of two other factors: the relaxation of selection resulting from WGD, and selection on dosage-sensitive regulatory pathways or gene complexes. According to the “gene balance model” (Birchler and Veitia, 2007; Freeling, 2009) dosage-insensitive genes have tendency to deteriorate over time through the accumulation of deleterious mutations whereas dosage-sensitive genes are subjected to purifying selection and tend to be retained. These predictions are consistent with findings that genes showing a higher connectivity level in networks (Severin et al., 2011), or contributing more to gene expression (Schnable et al., 2011), tend to be retained in ancient polyploids. In our study, selection against gene structure modifications was evident in the diploid model genomes, which showed a reduced number of LE/AE events compared to the wheat genome. This suggests that polyploidy can contribute to increased accumulation of non-functional structural variants of genes. Similarly, our finding that 38% of expressed genes in the wheat genome contain PTCs supports the relaxation of selection hypothesis. The high incidence of PTCs among structurally rearranged genes and retrogenes, as well as the preferential loss of exonic sequences, are indicative of relaxation of selection and ongoing degeneration of duplicated genes in the wheat genome. This is in contrast to finding that only 27% of retrogenes in the diploid rice genome are processed pseudogenes (Wang et al., 2006). It is possible that during early stages of polyploid evolution, the accumulation of new structural variants occurs almost neutrally. At the later stages of evolution when most of the genome becomes diploidized, functional constraints on structural variants of genes increase,
resulting in the preferential retention of functionally active genes.

Relaxation of selection in the wheat genome is consistent with results from the relative rate test, which showed an accelerated accumulation of non-synonymous mutations in the wheat lineage. An accelerated rate of codon evolution was observed for 19% of genes in either the wheat or *Brachypodium* lineages, suggesting significant levels of heterogeneity in the rates of gene evolution across the syntenic regions of grass genomes. No obvious effects of wheat chromosomal position or gene function were observed upon rates of codon evolution. According to the results of $d_S/d_S$ estimates, which showed a strong correlation among the three possible pair-wise genome comparisons, variation in the rates of non-synonymous mutations seems to be defined by gene-specific functional attributes. The differences in codon mutation rates along the *Brachypodium* lineage also suggest that genes in diploid genomes are under stronger selective constraint than in polyploid genomes. Certainly, we cannot exclude the possibility that some of the differences in codon mutation rates between the *Brachypodium* and wheat lineages were caused by an accelerated rate of mutation in the wild diploid ancestors of wheat. However, our previous analysis of mutation rates in diploid and polyploid wheat suggests the majority of non-synonymous mutations are associated with polyploid wheat evolution (Akhunov et al., 2008).

Estimates of AS obtained for each wheat subgenome using homoeolog-specific alignments were within the range (25-42%) previously predicted for the rice, *Arabidopsis* and *Brachypodium* genomes (Campbell et al., 2006; Filichkin et al., 2010; Wang and Brendel, 2006). Consistent with previous studies, the dominant type of AS events in wheat was intron retention, followed by alternative donor and acceptor splicing events (Campbell et al., 2006; Barbazuk et al., 2008; Wang and Brendel, 2006). Similar to other plant systems, exon skipping represented only a small (4-6%) fraction of AS events (Wang and Brendel, 2006) and intron retention mostly (91%) resulted in premature stop codons (Filichkin et al., 2010). It is suggested that the majority of transcripts resulting from this type of AS are targeted by the nonsense-mediated decay (NMD) and regulated unproductive splicing and translation (RUST) pathways, which play an important role in the regulation of functional transcript levels (Lewis et al., 2003; Maquat, 2004).

Comparative analysis of AS patterns between the wheat genomes revealed a significant number of new transcript isoforms in duplicated genes acquired since divergence from a common ancestor. Previous studies demonstrated the interdependence of gene duplication and AS (Roux and Robinson-Rechavi, 2011), and suggested that neutral processes associated with
the relaxation of selection can drive the accumulation of new splice isoforms in duplicated genes (Lev-Maor et al., 2007). Fast changes in RI splice variants previously reported for polyploid Brassica (Zhou et al., 2011) suggest that WGD may have substantial impact on divergence of AS patterns between the homoeologs. It is possible that WGDs in wheat have contributed to transcriptome diversity not only by combining transcriptomes of diverged diploid ancestors, but also by relaxing functional constraint on AS. Thus, even though we do not see significant differences in the level of AS in each individual wheat homoeolog, the divergence of AS between duplicated genes (42-61%) can increase the overall transcript isoform diversity in the polyploid genomes.

By analyzing AS conservation using the approach developed by Wang and Brendel (2006), we obtained results that were similar to that observed in a rice-Arabidopsis comparison, where only 8-9% of conserved introns had the same type of AS events in both wheat and rice. In spite of a higher divergence time between rice and Arabidopsis (130 MYA) than between rice and wheat (40 MYA), we observed a similar level of AS conservation. This could be a consequence of the approach used to predict the conserved AS events; by selecting conserved introns we may have enriched our sample for genes with AS events that are using evolutionarily conserved mechanisms of gene expression regulation (Lewis et al., 2003; Maquat, 2004). Another possibility is that the majority of AS events in plants are selectively neutral and their fate is defined by stochastic evolutionary processes, while only 8-9% of AS events are involved in biologically important function and, therefore, are widely conserved throughout diverged plant lineages. Alternatively, the rate of gene space evolution in the wheat genome may be accelerated due to polyploidy or high repetitive DNA content which, in turn, can result in a faster rate of AS evolution.

Conclusion

In conclusion, it seems that neutral processes largely define the evolution of the wheat genome coding regions. It is likely that genetic redundancy created by WGDs allows for an accelerated accumulation of AS isoforms, non-synonymous and PTC mutations, and gene structure rearrangements leading to degeneration of duplicated gene compliments. Even though these processes mostly contribute to genome diploidization, they have potential to accelerate the origin of new functional variation, providing raw material for selection. Therefore, it is plausible
that a high level of polyploid genome plasticity may have played an important role in broadening both genetic and phenotypic diversity in wheat. It will be of great interest to investigate the contribution of molecular processes described in our study to the origin of new variants of genes and assess their importance in phenotypic evolution in the context of crop domestication and improvement.

Materials and Methods

Sequencing of flow-sorted chromosomal DNA and assembly

The chromosome 3A double ditelosomic lines in the genetic background of wheat cultivar ‘Chinese Spring’ were used to isolate chromosome arm 3AS- and 3AL-specific DNA by flow-sorting (Kubaláková et al., 2002; Doležel et al., 2007). In addition to flow-sorted chromosomal DNA, single read sequencing was performed for a set of 2,743 BAC clones distributed among 1,104 BAC contigs assembled using the high information content fingerprint data obtained for 190,464 BAC clones from wheat chromosome 3A (http://wggrc.plantpath.ksu.edu/wheat/3A/3A_index.html). All genomic libraries were sequenced using 454 Titanium chemistry in the KSU Integrated Genomics Facility (www.ksre.ksu.edu/igenomics). The combined dataset generated for the flow-sorted chromosome arms and BAC clones corresponded to 9.3x coverage of wheat chromosome 3A (see Supplemental Table S1; NCBI accession number SRA045666). More detailed description of a sequencing protocol and assembly parameters could be found in the Supplemental Material.

Predicting exon-intron structure

The evidence-based prediction of exon-intron structure in wheat contigs was performed using wheat transcriptome sequence data, which was comprised of 454 sequence data (http://wheatgenomics.plantpath.ksu.edu/snp/, SRA048705), full-length cDNA (FL-cDNA) sequences (http://trifldb.psc.riken.jp/index.pl), and wheat ESTs from the NCBI database.

The 454 transcriptome data generated for eight wheat cultivars, consisting of about 5.8 million reads, was assembled using MIRA software (v3.03). This program uses an iterative sequence assembly approach that can refine contig assemblies after each iteration step by putting reads that contain mutations consistently differentiating them from the rest of the reads in a contig, into separate contigs (Chevreux et al., 2004). These settings are optimized for assembly
of ESTs (“est” option of MIRA assembler) and separation of transcript isoforms originating from different copies of homoeologous or paralogous genes in the wheat genome (Chevreux et al., 2004; see Supplemental Table S12). Since the average divergence of genes duplicated during wheat polyploidization is about 2-4% in coding regions (Dvorak et al., 2006), we would expect to have about 10 to 16 single-base substitutions per 500 bp read that differentiate one wheat sub-genome from another.

In order to infer the exon-intron structure of genes in the polyploid wheat genome, we pre-selected homoeolog-specific transcript assemblies, FL-cDNA and ESTs by aligning them against wheat contigs using BLAT (Kent, 2002). Transcripts showing sequence similarity >98.5% and minimum alignment length >200 bp were used for further analysis. These thresholds were chosen to preclude misalignment of duplicated genes to the wrong homoeologous copy of a gene in the wheat genome. The accuracy of genome assignment was based on counting the proportion of mutations that differentiate 86 homoeologous gene pairs identified in wheat chromosomes 3A and 3B (see Results, “Evolution of AS forms”), and that are also found to correctly match the respective genome in transcript/genomic sequence alignments. The results of this analysis suggest that 92% of homoeologous mutations can be correctly assigned to a genome. Spliced-alignments for gene prediction were created using the GMAP program integrated into the PASA gene prediction pipeline (http://pasa.sourceforge.net; see Supplemental Table S20).

**Alternative splicing**

Gene annotations produced by aligning homoeolog-specific transcript assemblies against wheat genomic sequences were extracted from the PASA database for estimating AS. Characterization of AS was performed separately in three sets of genomic sequence data including contigs/scaffolds generated for chromosome 3A in our study, contigs of chromosome 3B BAC clones generated by Choulet et al. (2010) and contigs of BAC clones distributed across the wheat genome (NCBI database; see Supplemental Table S11).

AS was compared between homoeologous copies of genes on wheat chromosomes 3A and 3B by counting the number of shared and different AS events including skipped exons, retained introns, alternative donor and acceptor variants.

The second method for comparing AS is based on the approach suggested by Wang and Brendel (2006). Briefly, conserved introns in orthologous sets of genes between rice and wheat
were identified by matching all rice introns and flanking 30 bp exons against wheat introns and flanking exons with TBLASTX (e-value < e^{-4}), and only those hits that contained at least 10 bp from both flanking exons were selected.

**Rates of molecular evolution**

The open reading frames (ORF) of orthologous sets of genes from rice, *Brachypodium* and wheat were aligned using the MUSCLE program (v3.8.31, [www.drive5.com/muscle/](http://www.drive5.com/muscle/)) to estimate rates of molecular evolution in coding sequences. Aligned ORFs were filtered to remove those that: 1) are shorter than 300 bp; 2) included gaps longer than 100 bp; or 3) included gaps covering more than 10% of alignment length. The rates of evolution were contrasted between the wheat and *Brachypodium* lineages using rice as an outgroup. Both the *Brachypodium* and wheat lineages diverged from rice about 50 MYA (Paterson et al., 2004). Two tests were applied to assess the rate of codon evolution in the alignments. The first test was performed using the yn00 program from the PAML package (v4.4d, Yang, 2007) with default parameters (icode = 0, weighting = 0, commonf3x4 = 0). The program was used to estimate $d_N/d_S$ ratios in wheat-rice and *Brachypodium*-rice pair-wise comparisons using the method of Yang and Nielsen (2000).

The relative rate test implemented in the HyPhy program (v2.0020110606beta, [www.hyphy.org](http://www.hyphy.org); Pond et al., 2005) was applied to compare the rates of codon evolution in the *Brachypodium* and wheat lineages using rice as an outgroup and the following options: Data type – Codon, Genetic Code – Universal, outgroup – rice, standard model - GY94, Model Options - Global. In this test, the codon evolution model was used for modeling nucleotide substitution rates (Muse and Gaut, 1994; Goldman and Yang, 1994). The model specifically focuses on estimating two separate parameters for synonymous ($\alpha$) and nonsynonymous ($\beta$) substitution rates. The relative rate test allows for the detection of significant changes in the relative rates of $\alpha$ and $\beta$ by comparing the two alternative codon evolution models using the likelihood ratio test (LRT): model $H_0$ assumes that the rates of codon evolution along the wheat ($v1$) and *Brachypodium* ($v2$) lineages are equal ($v1 = v2$), and model $H_1$ assumes that the rates of codon evolution along the wheat and *Brachypodium* lineages are different ($v1 \neq v2$). A more detailed description of the relative rate test using the codon evolution model can be found in a paper by Muse and Gaut (1994).
Evolution of gene structure

For this analysis, the alternative protein isoforms of *Brachypodium* and rice were splice-aligned against the genomic sequence of wheat, and the protein isoforms of rice were aligned against the genomic sequence of *Brachypodium*. Only those alignments that contained at least two exons were considered. We used the average proportion of acquired, lost and conserved exonic segments relative to the total number of coding segments among all possible pair-wise comparisons between orthologs (as defined on the Fig. 2A) as a measure of similarity between gene structures of the compared genomes.

First, we compared exon-intron structure of orthologous genes between wheat and the two grass genomes, rice and *Brachypodium*. Three types of evolutionary events resulting in exon-intron structural changes were recorded: 1) loss of exonic segments; 2) acquisition of exonic segments; and 3) conservation of exonic segments (CE) (Fig. 3A). For this purpose, the protein isoforms predicted in the model grass genomes were splice-aligned with the wheat contigs/scaffolds. The inferred exon-intron structure was compared with that predicted by aligning wheat ESTs/cDNAs. Annotations of exon-intron structure and corresponding CDSs for the rice and *Brachypodium* genomes were downloaded from public databases (www.brachypodium.org/ and rice.plantbiology.msu.edu). The translated sequences of all mRNA isoforms from syntenic regions of *Brachypodium* and rice genomes were splice-aligned against wheat genomic sequences using the GeneWise program (www.ebi.ac.uk/Tools/Wise2), with optimized settings (-splice flat; -null flat; -alg 2193; -both strands). The quality score of spliced alignments generated by GeneWise was kept above 35 as recommended by the software developer. The structures of orthologous genes in rice, *Brachypodium* and wheat were pair-wise compared. Differences in exon-intron structure were referred to as “coding segments”. Thus, each exon in our analysis was represented by a combination of several coding segments. Coding segments that were present in both compared species were referred to as constitutive coding segments, and coding segments that were present in only one of the two compared species were referred to as alternative coding segments. The proportion of alternative and constitutive coding segments between compared genes, relative to the total number of coding segments, was calculated for each orthologous gene set. The proportion of alternative coding segments per gene was averaged across all orthologous gene sets and used as the estimate of gene structure.
conservation between the compared genomes. The description of evolutionary changes in exon-intron structure used in this analysis is presented in Fig. 2A.

Changes in gene structure were further characterized to understand the molecular mechanisms involved in their origin. The sequences of newly acquired exonic fragments in the wheat lineage were compared against the databases of repetitive elements (GIRI, TREP) and proteins (NCBI) using BLAST. For identification of retroposition events resulting from the insertion of reverse transcribed mRNA, the structure of genes in the wheat genome was compared with that of the model genomes using the approach proposed by Fu et al. (2010).

The distribution of the \(d_{NS}/d_S\) ratio, premature termination codons and exon loss/acquisition events in the wheat lineage were plotted along chromosome 3A according to the syntenic gene order. The \(d_{NS}/d_S\) was calculated in a sliding window of 5 genes for each pair-wise comparison between wheat (W), rice (R) and Brachypodium (B). The \(d_{NS}/d_S\) ratio threshold for a window of 5 genes corresponds to the 95th-percentile of the mean \(d_{NS}/d_S\) ratio estimated for 100,000 sets of 5 genes, sampled randomly with replacement from the entire dataset.

**Accession Numbers**

Sequence data from this article can be found in the NCBI data libraries under accession numbers SRA012746, SRA045666, SRA048705 and PRJNA80879. Accession numbers of BAC clone sequences are JQ354542-JQ354560. The contig assemblies and alignments of orthologous genes are also available for download from [http://129.130.90.211/chr3A_data](http://129.130.90.211/chr3A_data). Wheat transcriptome assemblies can be searched using BLAST at the wheat SNP project website [http://wheatgenomics.plantpath.ksu.edu/snp/](http://wheatgenomics.plantpath.ksu.edu/snp/).

**Supplemental Material**

**Supplemental Figures**

**Supplemental Figure S1.** Relationship between chromosome 3A contig length and read coverage.

**Supplemental Figure S2.** Distribution of the number of alignment positions at each coverage depth (including the gaps at all positions) for chromosome 3A assemblies.

**Supplemental Figure S3.** Coverage of the wheat chromosome 3A gene space.

**Supplemental Figure S4.** Comparison of wheat chromosome 3A with the Brachypodium genome.
Supplemental Figure S5. Comparison of wheat chromosome 3A with the sorghum genome.
Supplemental Figure S6. Comparison of wheat chromosome 3A with the rice genome.
Supplemental Figure S7. Comparison of gene order between a SNP map of chromosome 3D and the SGO map of chromosome 3A.
Supplemental Figure S8. Intra-genomic duplications in wheat.
Supplemental Figure S9. Examples of alternative splicing types.
Supplemental Figure S10. Functional classification of genes on the 3A SGO map.
Supplemental Figure S11. Distribution of sequence similarity levels (%) in the coding regions of the wheat genome.

Supplemental Tables
Supplemental Table S1. Summary of sequence data used in the assembly of wheat chromosome 3A.
Supplemental Table S2. Purity assessment of amplified flow-sorted arms of wheat chromosome 3A.
Supplemental Table S3. Comparison of chromosome 3A BAC clone sequences with 454 contigs and scaffolds.
Supplemental Table S4. Classes of repetitive elements masked in contig assemblies
Supplemental Table S5. Comparison of chromosome 3A contigs with coding sequences (CDSs) of model genomes (rice, sorghum and Brachypodium).
Supplemental Table S6. Comparison of the repeat masked chromosome 3A contigs and scaffolds with the rice, Brachypodium and sorghum genomes.
Supplemental Table S7. Distribution of BLASTN hits between wheat chromosome 3A genes and three cereal genomes (rice, Brachypodium and sorghum).
Supplemental Table S8. Chromosome 3A contigs and scaffolds ordered using syntenic relationships between wheat and sequenced model grass genomes (SGO).
Supplemental Table S9. Inter-chromosomal duplications in the wheat genome.
Supplemental Table S10. Identification of chromosome 3A centromere boundaries.
Supplemental Table S11. List of BAC contig sequences downloaded from the NCBI database.
Supplemental Table S12. 454 transcriptome data generated for eight wheat cultivars.
Supplemental Table S13. Counts of alternative splicing events in a set of homoeologous genes.
identified between wheat chromosomes 3A and 3B.

**Supplemental Table S14.** Proportion of alternative and conserved coding segments between wheat and the model genomes.

**Supplemental Table S15.** Genes with acquired exonic sequences specific to the wheat lineage.

**Supplemental Table S16.** Validation of acquired exonic sequences by RT-PCR.

**Supplemental Table S17.** Characterization of acquired exonic sequences.

**Supplemental Table S18.** \( d_N/d_S \) ratio estimates for wheat - rice, wheat - *Brachypodium* and rice - *Brachypodium* pair-wise comparisons.

**Supplemental Table S19.** Comparison of the rates of codon evolution in the *Brachypodium* and wheat lineages using rice as an outgroup.

**Supplemental Table S20.** Prediction of exon-intron structure performed using the GMAP program integrated into the PASA gene prediction pipeline.
Literature cited

Akhunov ED, Akhunova AR, Anderson OD, Anderson JA, Blake N, Clegg MT, Coleman-Derr D, Conley EJ, Crossman CC, Deal KR, Dubcovsky J, Gill BS, Gu YQ, Hadam J, Heo HY, Huo N, Lazo GR, Luo MC, Ma YQ, Matthews DE, McGuire PE, Morrell P, Qualset CO, Renfro J, Reynolds S, Tabanao D, Talbert LE, Tian C, Toloeno D, Warburton M, You FM, Zhang W, Dvorak J (2008) Purifying selection and gene conversion in polyploid wheat evolution. Proc 11th Intl Wheat Genet Symp, Aug 24-29, Brisbane, Australia

Akhunov ED, Akhunova AR, Dvorak J (2007) Mechanisms and rates of birth and death of dispersed duplicated genes during the evolution of a multigene family in diploid and tetraploid wheats. Mol Biol Evol 24: 539-550

Akhunov ED, Akhunova AR, Linkiewicz AM, Dubcovsky J, Hummel D, Lazo G, Chao S, Anderson OD, David J, Qi L, Echalier B, Gill BS, Miftahudin, Gustafson JP, La Rota CM, Sorrells ME, Zhang D, Nguyen HT, Kalavacharla V, Hossain K, Kianian SF, Peng J, Lapitan NLV, Wennerlind EJ, Nduati V, Anderson JA, Sidhu D, Gill KS, McGuire PE, Qualset CO, Dvorak J (2003) Synteny perturbations between wheat homoeologous chromosomes caused by locus duplications and deletions correlate with recombination rates along chromosome arms. Proc Natl Acad Sci USA 100: 10836-10841

Akhunova AR, Matniyazov RT, Liang H, Akhunov ED (2010) Homoeolog-specific transcriptional bias in allopolyploid wheat. BMC Genomics 11: 505

Ast G (2004) How did alternative splicing evolve? Nat Rev Genet 5: 773-782

Bága M, Glaze S, Mallard CS, Chibbar RN (1999) A starch-branching enzyme gene in wheat produces alternatively spliced transcripts. Plant Mol Biol 40: 1019-1030

Barbazuk WB, Fu Y, McGinnis KM (2008) Genome-wide analyses of alternative splicing in plants: Opportunities and challenges. Genome Res 18: 1381-1392

Bartoš J, Vlček C, Choulet F, Džunková M, Cviková K, Sařář J, Simková H, Pačes J, Strnad H, Sourdille P, Bergès H, Cattonaro F, Feuillet C, Doležel J (2012) Intraspecific sequence comparisons reveal similar rates of non-collinear gene insertion in the B and D genomes of bread wheat. BMC Plant Biol 12: 155
Berkman PJ, Skarshewski A, Lorenc MT, Lai K, Duran C, Ling EY, Stiller J, Smits L, Imelfort M, Manoli S, McKenzie M, Kubaláková M, Simková H, Batley J, Fleury D, Doležel J, Edwards D (2011) Sequencing and assembly of low copy and genic regions of isolated Triticum aestivum chromosome arm 7DS. Plant Biotechnol J 9: 768-775

Birchler JA, Veitia RA (2007) The gene balance hypothesis: from classical genetics to modern genomics. Plant Cell 19:395-402.

Bottley A, Koebner RM (2008) Variation for homoeologous gene silencing in hexaploid wheat. Plant J 56:297-302

Campbell MA, Haas BJ, Hamilton JP, Mount SM, Buell CR (2006) Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis. BMC Genomics 7: 327

Chalupska D, Lee HY, Faris JD, Evrard A, Chalhoub B, Haselkorn R, Gornicki P (2008) Acc homoeoloci and the evolution of wheat genomes. Proc Natl Acad Sci USA 105: 9691-9696

Charles M, Belcram H, Just J, Huneau C, Viollet A, Couloux A, Segurens B, Carter M, Huteau V, Coritom O, Appels R, Samain S, Chalhoub B (2008) Dynamics and differential proliferation of transposable elements during the evolution of the B and A genomes of wheat. Genetics 180: 1071-1086

Chen ZJ (2007) Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. Annu Rev Plant Biol 58: 377-406

Chen M, Ha M, Lackey E, Wang J, Chen ZJ (2008) RNAi of met1 reduces DNA methylation and induces genome-specific changes in gene expression and centromeric small RNA accumulation in Arabidopsis alloplopolyploids. Genetics 178: 1845-1858

Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Müller WE, Wetter T, Suhai S (2004) Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res 14: 1147-1159

Choulet F, Wicker T, Rustenholz C, Paux E, Salse J, Leroy P, Schlub S, Le Paslier MC, Magdelenat G, Gonthier C, Couloux A, Budak H, Breen J, Pumphrey M, Liu S, Kong X, Jia J, Gut M, Brunel D, Anderson JA, Gill BS, Appels R, Keller B, Feuillet C (2010)
Megabase Level Sequencing Reveals Contrasted Organization and Evolution Patterns of the Wheat Gene and Transposable Element Spaces. Plant Cell 22: 1686-1701

Comai L (2005) The advantages and disadvantages of being polyploidy. Nat Rev Genet 6: 836-846

Devos KM (2010) Grass genome organization and evolution. Curr Opin Plant Biol 13: 139-145

Dolezel J, Kubaláková M, Paux E, Bartos J, Feuillet C (2007) Chromosome-based genomics in the cereals. Chromosome Res 15: 51-66

Dvorak J, Akhunov ED (2005) Tempos of gene locus deletions and duplications and their relationship to recombination rate during diploid and polyploid evolution in the aegilops-triticum alliance. Genetics 171: 323-332

Dvorak J, Akhunov ED, Akhunov AR, Deal KR, Luo MC (2006) Molecular characterization of a diagnostic DNA marker for domesticated tetraploid wheat provides evidence for gene flow from wild tetraploid wheat to hexaploid wheat. Mol Biol Evol 23: 1386-1396

Dvorák J, Terlizzi P, Zhang HB, Resta P (1993) The evolution of polyploid wheats: identification of the A genome donor species. Genome 36: 21-31

Dvorak J, Zhang HB (1990) Variation in repeated nucleotide sequences sheds light on the phylogeny of the wheat B and G genomes. Proc Natl Acad Sci USA 87: 9640-9644

Fan C, Zhang Y, Yu Y, Rounsley S, Long M, Wing RA (2008) The Subtelomere of Oryza sativa Chromosome 3 Short Arm as a Hot Bed of New Gene Origination in Rice. Mol Plant 1: 839-850

Fawcett JA, Rouzé P, Van de Peer Y (2012) Higher intron loss rate in Arabidopsis thaliana than A. lyrata is consistent with stronger selection for a smaller genome. Mol Biol Evol 29: 849-859

Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, Fox SE, Wong WK, Mockler TC (2010) Genome-wide mapping of alternative splicing in Arabidopsis thaliana. Genome Res 20: 45-58
Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151: 1531-1545

Freeling M (2009) Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. Annu Rev Plant Biol 60: 433-453

Gaeta RT, Yoo SY, Pires JC, Doerge RW, Chen ZJ, Osborn TC (2009) Analysis of gene expression in resynthesized Brassica napus Allopolyploids using arabidopsis 70mer oligo microarrays. PLoS One 4: e4760

Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11: 725-736

Gu YQ, Coleman-Derr D, Kong X, Anderson OD (2004) Rapid genome evolution revealed by comparative sequence analysis of orthologous regions from four triticeae genomes. Plant Physiol 135: 459-470

Gu YQ, Salse J, Coleman-Derr D, Dupin A, Crossman C, Lazo GR, Huo N, Belcram H, Ravel C, Charmet G, Charles M, Anderson OD, Chalhoub B (2006) Types and rates of sequence evolution at the high-molecular-weight glutenin locus in hexaploid wheat and its ancestral genomes. Genetics 174: 1493-1504

Huang S, Sirikhachornkit A, Faris JD, Su X, Gill BS, Haselkorn R, Gornicki P (2002) Phylogenetic analysis of the acetyl-CoA carboxylase and 3-phosphoglycerate kinase loci in wheat and other grasses. Plant Mol Biol 48: 805-820

Hughes AL (1994) The evolution of functionally novel proteins after gene duplication. Proc Biol Sci 256: 119-124

International Brachypodium Initiative (2010) Genome sequencing and analysis of the model grass Brachypodium distachyon. Nature 463: 763–768

International Rice Genome Sequencing Project (2005) The map based sequence of the rice genome. Nature 436: 793–800
Kaessmann H, Vinckenbosch N, Long M (2009) RNA-based gene duplication: mechanistic and evolutionary insights. Nat Rev Genet 10: 19-31

Kashkush K, Feldman M, Levy AA (2002) Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. Genetics 160: 1651-1659

Kashkush K, Feldman M, Levy AA (2003) Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. Nat Genet 33: 102-106

Kent WJ (2002) BLAT - The BLAST-Like Alignment Tool. Genome Res 4: 656-664

Kihara H (1944) Discovery of the DD-analyser, one of the ancestors of Triticum vulgare. (Japanese) Agric Hort 19: 889-890

Koonin EV, Wolf YI (2010) Constraints and plasticity in genome and molecular-phenome evolution. Nat Rev Genet 11: 487-498

Kubaláková M, Vrána J, Cíhalíková J, Simková H, Dolezel J (2002) Flow karyotyping and chromosome sorting in bread wheat (Triticum aestivum L.). Theor Appl Genet 104: 1362-1372

Lal SK, Giroux MJ, Brendel V, Vallejos CE, Hannah LC (2003) The maize genome contains a helitron insertion. Plant Cell 15: 381-391

Lev-Maor G, Goren A, Sela N, Kim E, Keren H, Doron-Faigenboim A, Leibman-Barak S, Pupko T, Ast G (2007) The "alternative" choice of constitutive exons throughout evolution. PLoS Genet 3: e203

Lewis BP, Green RE, Brenner SE (2003) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc Natl Acad Sci USA 100: 189-192

Lin JY, Stupar RM, Hans C, Hyten DL, Jackson SA (2010) Structural and functional divergence of a 1-Mb duplicated region in the soybean (Glycine max) genome and comparison to an orthologous region from Phaseolus vulgaris. Plant Cell 22: 2545-2561
Long M, Betrán E, Thornton K, Wang W (2003) The origin of new genes: glimpses from the young and old. Nat Rev Genet 4: 865-875

Lu T, Lu G, Fan D, Zhu C, Li W, Zhao Q, Feng Q, Zhao Y, Guo Y, Li W, Huang X, Han B (2010) Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. Genome Res 20: 1238-1249

Luo MC, Deal KR, Akhunov ED, Akhunova AR, Anderson OD, Anderson JA, Blake N, Clegg MT, Coleman-Derr D, Conley EJ, Crossman CC, Dubcovsky J, Gill BS, Gu YQ, Hadam J, Heo HY, Huo N, Lazo G, Ma Y, Matthews DE, McGuire PE, Morrell PL, Qualset CO, Renfro J, Tabanao D, Talbert LE, Tian C, Toleno DM, Warburton ML, You FM, Zhang W, Dvorak J (2009) Genome comparisons reveal a dominant mechanism of chromosome number reduction in grasses and accelerated genome evolution in Triticeae. Proc Natl Acad Sci USA 106: 15780-15785

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290: 1151-1155

Malko DB, Makeev VJ, Mironov AA, Gelfand MS (2006) Evolution of exon-intron structure and alternative splicing in fruit flies and malarial mosquito genomes. Genome Res 16: 505-509

Massa AN, Wanjugi H, Deal KR, O'Brien K, You FM, Maiti R, Chan AP, Gu YQ, Luo MC, Anderson OD, Rabinowicz PD, Dvorak J, Devos KM (2011) Gene space dynamics during the evolution of Aegilops tauschii, Brachypodium distachyon, Oryza sativa, and Sorghum bicolor genomes. Mol Biol Evol 28: 2537-2347

Maquat LE (2004) Nonsense-mediated mRNA decay: Splicing, translation and mRNP dynamics. Nat Rev Mol Cell Biol 5: 89-99

Marquez Y, Brown JW, Simpson C, Barta A, Kalyna M (2012) Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. Genome Res PMID: 22391557

Mayer KFX, Martis M, Hedley PE, Simkova H, Liu H, Morris JA, Steuernagel B, Taudien S, Roessner S, Gundlach H, Kubaláková M, Suchánková P, Murat F, Felder M, Nussbaumer T, Graner A, Salse J, Endo T, Sakai H, Tanaka T, Itoh T, Sato K, Platzer M, Matsumoto T, Scholz U, Dolezel J, Waugh R, Stein N (2011) Unlocking the Barley Genome
by Chromosomal and Comparative Genomics. Plant Cell 23: 1249-1263

McFadden ES, Sears ER (1944) The artificial synthesis of Triticum spelta. Rec Soc Genet Am 13: 26-27

McFadden ES, Sears ER (1946) The origin of Triticum spelta and its free-threshing hexaploid relatives. J Hered 37: 81-89, 107-116

Muse SV, Gaut BS (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol 11: 715-724

Nesbitt M, Samuel D (1996) From staple crop to extinction? The archaeology and history of the hulled wheats. In S Padulosi, K Hammer, J Heller, eds, Hulled wheats. Vol 4. International Plant Genetic Resources Institute, Rome, pp. 41-100

Ni Z, Kim ED, Ha M, Lackey E, Liu J, Zhang Y, Sun Q, Chen ZJ (2009) Altered circadian rhythms regulate growth vigour in hybrids and allopolyploids. Nature 457: 327-331

Nomura T, Ishihara A, Yanagita RC, Endo TR, Iwamura H (2005) Three genomes differentially contribute to the biosynthesis of benzoxazinones in hexaploid wheat. Proc Natl Acad Sci USA 102: 16490-16495

Ohno S (1970) Evolution by Gene Duplication. (New York: Springer)

Osborn TC, Pires JC, Birchler JA, Auger DL, Chen ZJ, Lee HS, Comai L, Madlung A, Doerge RW, Colot V, Martienssen RA (2003) Understanding mechanisms of novel gene expression in polyploids. Trends Genet 19: 141-147

Paterson AH, Bowers JE, Chapman BA (2004) Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proc Natl Acad Sci USA 101: 9903-9908

Paterson AH, Bowers JE, Peterson DG, Estill JC, Chapman BA (2003) Structure and evolution of cereal genomes. Curr Opin Genet Dev 13: 644-650
Paterson AH, et al (2009) The Sorghum bicolor genome and the diversification of grasses. Nature 457: 551-556

Pond SL, Frost SD, Muse SV (2005) HyPhy: hypothesis testing using phylogenies. Bioinformatics 21: 676-679

Pumphrey M, Bai J, Laudencia-Chingcuanco D, Anderson O, Gill BS (2009) Nonadditive expression of homoeologous genes is established upon polyploidization in hexaploid wheat. Genetics 181: 1147-1157

Qi LL, Echalier B, Chao S, Lazo GR, Butler GE, Anderson OD, Akhunov ED, Dvorák J, Linkiewicz AM, Ratnasiri A, Dubcovsky J, Bermudez-Kandianis CE, Greene RA, Kantety R, La Rota CM, Munkvold JD, Sorrells SF, Sorrells ME, Dilbirligi M, Sidhu D, Erayman M, Randhawa HS, Sandhu D, Bondareva SN, Gill KS, Mahmoud AA, Ma XF, Miftahudin, Gustafson JP, Conley EJ, Nduati V, Gonzalez-Hernandez JL, Anderson JA, Peng JH, Lapitan NL, Hosain KG, Kalavacharla V, Kianian SF, Pathan MS, Zhang DS, Nguyen HT, Choi DW, Fenton RD, Close TJ, McGuire PE, Qualset CO, Gill BS (2004) A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat. Genetics 168: 701-712

Qi L, Friebe B, Wu J, Gu Y, Qian C, Gill BS (2010) The compact Brachypodium genome conserves centromeric regions of a common ancestor with wheat and rice. Funct Integr Genomics 10: 477-492

Reddy ASN (2007) Alternative splicing of pre-messenger RNAs in plants in the genomic era. Annu Rev Plant Biol 58: 267-294

Roux J, Robinson-Rechavi M (2011) Age-dependent gain of alternative splice forms and biased duplication explain the relation between splicing and duplication. Genome Res 21: 357-363

Salse J, Bolot S, Throude M, Jouffe V, Piegu B, Quraishi UM, Calcagno T, Cooke R, Delseny M, Feuillet C (2008) Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. Plant Cell 20: 11-24

Schnable JC, Springer NM, Freeling M (2011) Differentiation of the maize subgenomes by
genome dominance and both ancient and ongoing gene loss. Proc Natl Acad Sci USA 108: 4069-4074

Schnable PS, et al (2009) The B73 maize genome: Complexity, diversity, and dynamics. Science 326: 1112–1115

Severin AJ, Cannon SB, Graham MM, Grant D, Shoemaker RC (2011) Changes in twelve homoeologous genomic regions in soybean following three rounds of polyploidy. Plant Cell 23: 3129-3136

Soltis DE, Bell CD, Kim S, Soltis PS (2008) Origin and early evolution of angiosperms. Ann NY Acad Sci 133: 3-25

Terashima A, Takumi S (2009) Allopolyploidization reduces alternative splicing efficiency for transcripts of the wheat DREB2 homolog, WDREB2. Genome 52: 100-105

Van de Peer Y, Maere S, Meyer A (2009) The evolutionary significance of ancient genome duplications. Nat Rev Genet 10: 725-732

Wang BB, Brendel V (2006) Genome wide comparative analysis of alternative splicing in plants. Proc Natl Acad Sci USA 103: 7175-7180

Wang W, Zheng H, Fan C, Li J, Shi J, Cai Z, Zhang G, Liu D, Zhang J, Vang S, Lu Z, Wong GK, Long M, Wang J (2006) High rate of chimeric gene origination by retroposition in plant genomes. Plant Cell 18: 1791-1802

Wei F, Stein JC, Liang C, Zhang J, Fulton RS, Baucom RS, De Paoli E, Zhou S, Yang L, Han Y, Pasternak S, Narechania A, Zhang L, Yeh CT, Ying K, Nagel DH, Collura K, Kudrna D, Currie J, Lin J, Kim H, Angelova A, Scara G, Wissotski M, Golser W, Courtney L, Kruchowski S, Graves TA, Rock SM, Adams S, Fulton LA, Fronick C, Courtney W, Kramer M, Spiegel L, Nascimento L, Kalyanaraman A, Chaparro C, Deragon JM, Miguel PS, Jiang N, Wessler SR, Green PJ, Yu Y, Schwartz DC, Meyers BC, Bennetzen JL, Martienssen RA, McCombie WR, Aluru S, Clifton SW, Schnable PS, Ware D, Wilson RK, Wing RA (2009) Detailed analysis of a contiguous 22-Mb region of the maize genome. PLoS Genet 5: e1000728
Wicker T, Buchmann JP, Keller B (2010) Patching gaps in plant genomes results in gene movement and erosion of colinearity. Genome Res 20: 1229-1237

Wicker T, Mayer KF, Gundlach H, Martis M, Steuernagel B, Scholz U, Simková H, Kubaláková M, Choulet F, Taudien S, Platzer M, Feuillet C, Fahima T, Budak H, Dolezel J, Keller B, Stein N (2011) Frequent gene movement and pseudogene evolution is common to the large and complex genomes of wheat, barley, and their relatives. Plant Cell 23: 1706-1718

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24: 1586-1591

Yang Z, Nielsen R (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol 17: 32-43

Zhang G, Guo G, Hu X, Zhang Y, Li Q, Li R, Zhuang R, Lu Z, He Z, Fang X, Chen L, Tian W, Tao Y, Kristiansen K, Zhang X, Li S, Yang H, Wang J, Wang J (2010) Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. Genome Res 20: 646-654

Zhou R, Moshgabadi N, Adams KL (2011) Extensive changes to alternative splicing patterns following allopolyploidy in natural and resynthesized polyploids. Proc Natl Acad Sci USA 108: 16122-16127

Zhu Z, Zhang Y, Long M (2009) Extensive structural renovation of retrogenes in the evolution of the Populus genome. Plant Physiol 151: 1943-1951
**Figure legends**

**Figure 1.** Comparative analysis between wheat chromosome 3A and the sequenced genomes of *Brachypodium*, rice and sorghum. A. Comparison of the wheat chromosome 3A SGO map (Ta3A) with sorghum chromosomes 3 and 9 (Sb3, Sb9), rice chromosomes 1 and 5 (Os1, Os5) and *Brachypodium* chromosome 2 (Bd2). Ancestral chromosome duplications predating the divergence of wheat, rice, *Brachypodium* and sorghum are colored in blue; chromosomal bins enriched for BLASTN hits with chromosome 3A contigs are colored in orange. B. Location of the wheat chromosome 3A centromere was inferred by BLASTN comparison of SGO with the 454 reads generated for flow-sorted 3AS and 3AL chromosomal arms. The locations of two chromosome 3A contigs on the SGO map (SGO1221 and SGO1367) showing similarity to previously mapped EST markers flanking the wheat centromere are shown.

**Figure 2.** Proportion of alternative and conserved coding segments between wheat and the model genomes. A) Classification of coding segments used for comparison of gene exon-intron structure among wheat, rice and *Brachypodium*. Only the wheat-rice comparison is shown. The protein isoforms of rice or *Brachypodium* were aligned to orthologous gene models in wheat. Differences in exon-intron structure in these pair-wise comparisons were referred to as “coding segments”. Thus, each exon in our analysis was represented by a combination of several coding segments. Coding segments that were present in both compared orthologs were referred to as constitutive coding segments; coding segments that were present in only one of the two compared orthologs were referred to as alternative coding segments. The following categories of mini-events resulting in gene structure changes were recorded: 1) Conserved Exons (CE); 2) Acquired Exons (AE); 3) Lost Exons (LE). Both AE and LE correspond to alternative coding segments; CE corresponds to a subset of constitutive coding segments. B) Proportion of AE, LE and CE between the compared genes, relative to the total number of coding segments, was used to assess the level of gene structure conservation among orthologous genes of wheat (W), *Brachypodium* (B) and rice (R). Gene structure evolution was compared between wheat chromosome 3A and model genomes (3A vs. R, 3A vs. Br), between two model grass genomes (B vs. R), between a set of wheat genes located on other chromosomes (non-3A) and model genomes (W vs. R, W vs. B), and between homoeologous genes on wheat chromosomes 3A, 3B.
and 3D (3A vs. 3B, 3A vs. 3D).

**Figure 3.** Evolution of coding sequences in the wheat genome. A) Distribution of pairwise estimates of $d_N/d_S$ in a set of orthologous genes from wheat (W), *Brachypodium* (B) and rice (R). B) $H_0$ and $H_A$ are null (codon mutation rates in the wheat and *Brachypodium* lineages are equal; $v_1 = v_2$) and alternative (codon mutation rates in the wheat and *Brachypodium* lineages are different; $v_1 \neq v_2$) hypotheses, respectively; $v_1$ and $v_2$ are expected codon mutation rates in wheat and *Brachypodium*, respectively. C) Log2 ratio of codon mutation rates in the wheat ($v_1$) and *Brachypodium* ($v_2$) lineages estimated for an orthologous set of 986 genes (see Supplemental Table S19). The significance threshold ($p$-value $\leq 0.05$) is shown by a red dashed line. D) Distribution of $d_N/d_S$ estimates, PTCs (blue bars) and exon loss/acquisition events (red bars) along wheat chromosome 3A. 95th-percentile of $d_N/d_S$ ratio distribution in a window of 5 genes (dashed lines) was calculated by permutation (see Materials and Methods).
## Tables

### Table 1. The chromosome 3A assembly statistics

|                        | Large Contigs (≥500 bp) | Scaffolds (≥2,000 bp) | Large Contigs (≥500 bp) & Scaffolds |
|------------------------|-------------------------|-----------------------|-------------------------------------|
| No. contigs/scaffolds  | 240,121                 | 38,036                | 213,660                             |
| Number of bases        | 330,997,554             | 195,458,009           | 348,362,074                         |
| Avg. size (bp)         | 1,378                   | 5,138                 | 1,630                               |
| N50 size (bp)          | 1,692                   | 6,013                 | 2,689                               |
| Largest size (bp)      | 43,015                  | 69,248                | 69,248                              |
| Ns in gaps (bp)        | -                       | 17,125,261            | 17,125,261                          |
|                              | 3A       | 3B<sup>a</sup> | 3D<sup>a</sup> | Whole genome sample<sup>b</sup> |
|------------------------------|----------|----------------|----------------|-------------------------------|
| Total transcript sequences   | 1,097,184| 8,945          | 1,138          | 31,292                        |
| Number transcripts with any  | 268,190  | 6,197          | 1,072          | 31,118                        |
| alignment                    |          |                |                |                               |
| Valid gmap alignments        | 60,301   | 2,715          | 531            | 22,006                        |
| Number of subclusters (genes)| 6,366    | 391            | 91             | 1,049                         |

<sup>a</sup> Sequences of BAC contigs from wheat chromosomes 3B (Choulet et al., 2010) and 3D (Bartoš et al., 2012).

<sup>b</sup> Sequences of BAC contigs downloaded from the NCBI database.
Table 3. Analysis of alternative splicing in the wheat genome

| Events<sup>a</sup> | 454 assemblies (3A) | 3B | 3D | Whole genome sample |
|-------------------|---------------------|----|----|---------------------|
| Involved in AS    | 1912 (24.8)         | 118 (24.8) | 27 (24.3) | 235 (20.9)         |
| AA                | 686 (35.9)          | 51 (43.2)  | 7 (25.9)   | 66 (28.1)          |
| AD                | 447 (23.4)          | 33 (28.0)  | 6 (22.2)   | 62 (26.4)          |
| RI                | 636 (33.3)          | 44 (37.3)  | 8 (29.6)   | 84 (35.7)          |
| SE                | 117 (6.1)           | 5 (4.2)    | 2 (7.4)    | 15 (6.4)           |

<sup>a</sup> AD – alternative donor site; AA – alternative acceptor site; AP – alternative position; SE – exon skipping; RI intron retention.
Table 4. Comparison of AS types between duplicated gene pairs on homoeologous chromosomes 3A, 3B and 3D.

| AS types<sup>a</sup> | 3A-3B comparison | 3A-3D comparison |
|----------------------|------------------|------------------|
|                      | Count in 3B | Count in 3A | Different AS events | Count in 3D | Count in 3A | Different AS events |
| AD                   | 14         | 10           | 5                   | 8           | 3           | 4                   |
| AA                   | 15         | 9            | 11                  | 1           | 5           | 4                   |
| RI                   | 19         | 14           | 19                  | 6           | 3           | 9                   |
| SE                   | 3          | 1            | 1                   | 1           | 1           | 0                   |

<sup>a</sup> AD – alternative donor site; AA – alternative acceptor site; AP – alternative position; SE – exon skipping; RI intron retention.
Table 5. Conservation of AS events in wheat and rice genes

| AS type a | Conserved introns b | Conserved events c |
|-----------|---------------------|--------------------|
| AD        | 2                   | 0                  |
| AA        | 28                  | 8                  |
| SE        | 0                   | 0                  |
| RI        | 216                 | 11                 |
| Total     | 232                 | 19                 |

a AD – alternative donor site; AA – alternative acceptor site; AP – alternative position; SE – exon skipping; RI - intron retention. b Conserved introns correspond to the number of introns that are conserved between wheat and rice and show the indicated AS type in wheat (not necessarily in rice). c Conserved events correspond to the number of conserved introns with the same AS type in both wheat and rice.
Figure 1. Comparative analysis between wheat chromosome 3A and the sequenced genomes of *Brachypodium*, rice and sorghum. **A.** Comparison of the wheat chromosome 3A SGO map (Ta3A) with sorghum chromosomes 3 and 9 (Sb3, Sb9), rice chromosomes 1 and 5 (Os1, Os5) and *Brachypodium* chromosome 2 (Bd2). Ancestral chromosome duplications predating divergence of wheat, rice, *Brachypodium* and sorghum are colored in blue; chromosomal bins enriched for BLASTN hits with the chromosome 3A contigs are colored in orange. **B.** Location of the wheat chromosome 3A centromere was inferred by BLASTN comparison of SGO with 454 reads generated for flow-sorted 3AS and 3AL chromosomal arms. The locations of two chromosome 3A contigs on the SGO map (SGO1221 and SGO1367) showing similarity to previously mapped EST markers flanking the wheat centromere are shown.
**Figure 2.** Proportion of alternative and conserved coding segments between wheat and the model genomes. **A)** Classification of coding segments used for comparison of gene exon-intron structure among wheat, rice and *Brachypodium*. Only wheat-rice comparison is shown. The protein isoforms of rice or *Brachypodium* were aligned to orthologous gene models in wheat. Differences in exon-intron structure in these pair-wise comparisons were referred to as “coding segments”. Thus, each exon in our analysis was represented by a combination of several coding segments. Coding segments that were present in both compared orthologs were referred to as constitutive coding segments; coding segments that were present in only one of the two compared orthologs were referred to as alternative coding segments. The following categories of mini-events resulting in gene structure changes were recorded: 1) Conserved Exons (CE); 2) Acquired Exons (AE); 3) Lost Exons (LE). Both AE and LE correspond to alternative coding segments; CE corresponds to a subset of constitutive coding segments. **B)** Proportion of AE, LE and CE between the compared genes, relative to the total number of coding segments, was used to assess the level of gene structure conservation among orthologous genes of wheat (W), *Brachypodium* (B) and rice (R). Gene structure evolution was compared between wheat chromosome 3A and model genomes (3A vs. R, 3A vs. Br), between two model grass genomes (B vs. R), between a set of wheat genes located on other chromosomes (non-3A) and model genomes (W vs. R, W vs. B), and between homoeologous genes on wheat chromosomes 3A, 3B and 3D (3A vs. 3B, 3A vs. 3D).
Figure 3. Evolution of coding sequences in wheat genome. A) Distribution of pairwise estimates of \( \frac{d_N}{d_S} \) in a set of orthologous genes from wheat (W), *Brachypodium* (B) and rice (R). B) \( H_0 \) and \( H_A \) are null (codon mutation rates in the wheat and *Brachypodium* lineages are equal; \( v1 = v2 \)) and alternative (codon mutation rates in the wheat and *Brachypodium* lineages are different; \( v1 \neq v2 \)) hypotheses, respectively; \( v1 \) and \( v2 \) are expected codon mutation rates in wheat and *Brachypodium*, respectively. C) Log 2 ratio of codon mutation rates in the wheat (\( v1 \)) and *Brachypodium* (\( v2 \)) lineages estimated for orthologous set of 986 genes (Supplemental Table S11). Significance threshold (\( p \)-value \( \leq 0.05 \)) is shown by red dashed line. D) Distribution of \( \frac{d_N}{d_S} \) estimates, PTCs (blue bars) and exon loss/acquisition events (red bars) along wheat chromosome 3A. 95\(^{th}\)-percentile of \( \frac{d_N}{d_S} \) ratio distribution in a window of 5 genes (dashed lines) was calculated by permutation (see Methods).