doi:10.1038/nature12028

The *Aegilops tauschii* draft genome sequence reveals gene repertoire for wheat adaptation

Jizeng Jia*,#, Shancen Zhao*, Xiuying Kong*, Yingrui Li*, Guangyao Zhao*, Weiming He*, Rudi Appels*, Matthias Pfeifer, Yong Tao, Xueyong Zhang, Ruilian Jing, Chi Zhang, Youzhi Ma, Lifeng Gao, Chuan Gao, Manuel Spannagl, Klaus F.X. Mayer, Dong Li, Shengkai Pan, Fengya Zheng, Qun Hu, Xianchun Xia, Jianwen Li, Qinsi Liang, Jie Chen, Thomas Wicker, Caiyun Gou, Hanhui Kuang, Genyun He, Yadan Luo, Beat Keller, Qiuju Xia, Peng Lu, Junyi Wang, Hongfeng Zou, Rongzhi Zhang, Junyang Xu, Jinlong Gao, Christopher Middleton, Zhiwu Quan, Guangming Liu, Jian Wang, IWGSC, Huanming Yang, Xu Liu #, Zhonghu He#, Long Mao#, Jun Wang#

*These authors contributed equally to this work.
#To whom correspondence should be addressed.

Contents

S1. Genome sequencing details

S1.1 Genetic background of sequencing material

We sequenced the genome of *Aegilops tauschii* Coss (accession AL8/78), which is the ancestor of common wheat (*Triticum aestivum* L.) D genome. The *Ae. tauschii* accession AL8/78 was collected in Armenia by V. Jaaska, University of Estonia, Tartu, Estonia, and has been used to construct a D genome physical map by the American NSF-funded project (PI: J Dvorak, UC Davis. http://wheatdb.ucdavis.edu:8080/wheatdb/). Seeds of this accession were obtained from Dr. Jan Dvorak and Dr. Ming-Cheng Luo, UC Davis. Plants were grown in a plant growth chamber at 25℃ in dark condition for two weeks before DNA from leaves was purified using Chao's method[1].

S1.2 Library construction, sequencing and quality control

Libraries with insert sizes of 200, 500 and 700 bp (short insert size) and 2, 5, 10 and 20 Kb (long-insert size) were constructed following the manufacturer's instructions (Illumina, San Diego, CA). For DNA libraries with short-insert size, 5 μg of genomic DNA was fragmented by nebulization with compressed nitrogen gas. The ends of DNA fragments were blunted with an "A" base. Next, the DNA adaptors (Illumina, San Diego, CA) with a single "T" base overhang at the 3' ends were ligated to the above products. We then purified the ligation products on a 2% agarose gel, and excised and purified gel slices for each insert size with Gel Extraction Kit (Qiagen, Valencia, CA). In order to facilitate the assembly process, the insert size of a library needed to fall in a narrow range (±10% around the expected size). For long (≥2 Kb) mate-paired libraries, 10-30 μg of genomic DNA was fragmented by nebulization with compressed nitrogen gas, and then biotin labeled dNTPs were used for polishing, and gels were selected for the main bands among 2 Kb, 5 Kb, 10 Kb and 20 Kb. The DNA fragments were then circularized for self-ligation. The two ends of the DNA fragment were merged together and the linear DNA fragments were digested by DNA exonuclease. Then the circularized DNA was fragmented again, followed by enrichment of the "merged ends" with magnetic beads using a biotin/streptavidin, then the ends were blunted and "A" bases and adaptors were added.

After quality control of each DNA library (3.7 nmol/L for short inserts and 2.5 nmol/L for long inserts by qPCR), ssDNA fragments were hybridized to flow cells, amplified to form clusters, then subjected to Pair-End sequencing following the standard Illumina protocol. A base-calling pipeline (Solexa Pipeline-0.3) was applied to obtain sequences from the raw fluorescent images. In total, raw data containing 557.55 Gb was obtained (**Supplementary Table 1**).

Before *de novo* assembly, the reads were filtered out as follows: (1) Low quality ends (quality q ≤ 7) were trimmed directly based on the sequencing quality report; (2) Reads with Ns > 10% of the read length; (3) Reads with low quality bases (>50% bases with quality Q-value ≤ 8); (4) Reads with adaptor contamination; (5) PCR duplications (reads are considered duplications when read1 and read2 of the same paired end reads are identical). Raw data were reduced from 557.55 Gb to 378.86 Gb after filtering.

S1.3 Assembly procedures

SOAPdenovo (version 1.05; http://soap.genomics.org.cn)[2] was employed to assemble the genome from the filtered data. In the assembly process, all possible sequences from Illumina reads were assembled using a de Bruijn graph methodology, with a *K*-mer (*K* was set as 63 here) as a node and the *K*-1 bases overlap between two *K*-mers as an edge. Large *K*-mer helps us assemble some short repeats in the genome. To reduce the sequencing errors and limit branches, the tips and *K*-mers with low coverage were removed. The graph then was transformed to a contig graph through turning those linearly connected *K*-mers to a pre-contig node. Dijkstra's algorithm was used to detect bubbles, which were then merged into a single pathway if the sequences of branches were similar. With this method, the regions with repeat sequences were merged into consensus sequences.

The assembled contigs were linked to a scaffolding graph based on paired-end (PE) reads. Connections between contigs were defined as edges in this graph and the branch length was the gap size calculated from the insert size of PE reads. Then, the sub-graph linearization was applied to turn interleaving contigs into linear structure. PE reads were applied step by step with increased insert sizes of 170, 500 and 700 bp, as well as 2, 5, 10, 20 Kb. To fill gaps in the scaffolds, we aligned the PE reads and collected the ones with one end mapped to a contig and the other end falling in a gap,, and finally performed a local assembly with the retrieved reads. About 400 Mb gaps were closed with this method. Additional gaps were filled using approximately 18.4 Gb data sequenced by 454 sequencing platform. Approximately 79.7 Mb gaps were closed with this additional data. Finally, the scaffold N50 length achieved 57,585 bp with a total length of 4.23 Gb. The assembly consists of 6,995,685 scaffolds with 701 Mb Ns, and 111,337 scaffolds with length ≥1Kb, which account for 3.3 Gb of the genome (**Supplementary Table 2-3**).

S1.4 K-mer analysis

We adopted a method based on *K*-mer distribution to estimate the genome size with 28-fold high quality reads (~112 Gb) from short-insert size libraries (≤ 800bp). A *K*-mer refers to an artificial sequence division of K nucleotides. With this definition, a raw sequence read with *L* bp contains $(L - K + 1)$ *K*-mers. The frequency of each *K*-mer can be calculated from the reads used for analysis. It is proposed that the *K*-mer frequencies along the sequence depth gradient generally follow a Poisson

distribution for a given dataset. Thus, the genome size G is calculated as $G = K_{num} / K_{depth}$, where the K_{num} is the total number of K-mer and K_{depth} is the highest peak detected. K was set to 17 in our project based on our empirical analysis. We obtained the 17-mer depth distribution and observed that the peak depth was at 25. Thus, the genome size was estimated to be 4.36 Gb (**Supplementary Figure 1**). In empirical data, low-depth K-mer frequencies constitute a larger proportion due to sequence errors. We also observed a small peak at 50-fold depth which is probably caused by repetitive sequences in 436 Mb genomic regions.

S1.5 Evaluation of the assembly

The quality of the draft genome was comprehensive evaluated by assessing the sequencing depth and coverage using available EST and BAC sequences. Over 96% of the sequences were covered by ≥20 genomic reads with a peak depth of 76× (**Supplementary Figure 2**), which indicates that the draft genome had high single-base accuracy based on the reported accuracy of next-generation sequencing technologies.

To evaluate the quality of the assembled genome, we used the ESTs from two full-length cDNA libraries from leaf and root of *Ae. tauschii* (accession AL8/78) constructed with a modified CAP-trapper method[3]. We randomly sequenced 12,110 clones from the 3' end, of which 1,735 clones were also sequenced from 5'. Finally we obtained 13,780 EST sequences from *Ae. tauschii* and used them for genome evaluation and annotation. To determine whether these ESTs were contaminated, all the genomic reads and RNA-Seq reads were mapped to the ESTs by SOAPaligner[4], allowing 2 mismatches. We discarded 595 ESTs which were not covered by the genomic reads. We aligned the remaining 13,185 ESTs to the genome using BLAT[5] and 91.0% of all the ESTs could be mapped to the genome (identity > 95% and coverage > 90%). Besides, 87.80% of the seven BACs downloaded from GenBank were mapped against the assembled genome using BLAT with identity ≥ 95% (**Supplementary Table 4** and **Supplementary Figure 3**). It was consistent with 83.4% of genomic sequence content in the scaffolds. Thus, the draft sequences represent a considerable portion of the *Ae. tauschii* genome with high quality and coverage.

S2. Genome annotation

S2.1 RNA-Seq and expression analysis

To aid genome annotation and address a series of biological questions, we generated 53.21 Gb of RNA-Seq data from eight different organs: Pistil, pistils from spikes 2-4cm in length; Root, roots of 3-week-old seedlings cultivated in Hoagland solution; Seed, young seeds 5 days post anthesis; Spike, spikes less than 1cm in length during stem extension; Stamen, stamens from spikes 2-4cm in length; Stem,

ear stems in heading period; Leaf, leaves of 3-week-old seedlings cultivated in Hoagland solution; Sheath, sheaths of 3-week-old seedlings cultivated in Hoagland solution.

Total RNA was isolated with TRIzol (Invitrogen, Carlsbad, CA) from each sample according to the manufacturer's instructions. The recovered total RNA was first treated with RNase-free DNase I for 30 min at 37℃ (New England BioLabs, Beverly, MA, USA) to remove residual DNA. Beads with oligo(dT) were used to isolate poly(A) mRNA. The first strand cDNA was synthesized using random hexamer-primer and reverse transcriptase (Invitrogen, Carlsbad, CA). The second-strand cDNA was synthesized using RNase H (Invitrogen, Carlsbad, CA) and DNA polymerase I (New England BioLabs, Beverly, MA, USA). The cDNA libraries were prepared and sequenced according to Illumina's protocol, as described above. In total, we obtained 53.21 Gb raw data from 23 libraries.

After raw data processing, paired-end reads were merged into one using the overlap information when their total length is longer than the insert-size. With this approach, we obtained 169,378,164 single-end reads (23.5 Gb) with an average length of 138 bp, and then we assembled them using CAP3[6], a software finding the overlap from every two reads with high efficiency and accuracy. Due to its large memory consumption (30 times more than input data), we split all the reads into several parts and assembled each part separately. We put the results of each tissue together and assembled these reads using CAP3 in the same way. The redundancy in the transcriptome was removed by CD-HIT[7]. The final transcriptome size was 117 Mb, and average length was 932 bp (**Supplementary Table 5**).

To measure the gene expression level in eight tissues at different sequencing depths, we calculated the expression of each gene using RPKM (Reads Per Kilobase of exon model per Million mapped reads) value with the following formula:

$$RPKM = \frac{10^6 C}{NL / 10^3}$$

Set RPKM (A) to be the expression of gene A, C to be number of reads that uniquely aligned to gene A, N to be total number of reads that uniquely aligned to all genes, and L to be the base number in the CDS of gene A.

The RPKM method can be used to eliminate the influence of different gene length and sequencing discrepancy on the calculation of gene expression. Thus, the organ-specific index of genes – τ value can be estimated by the following formula based on RPKM value:

$$\tau_i = \frac{\sum_{j=1}^{n}(1-\log_{10}S_{(i,j)} / \log_{10}S_{(i,\max)})}{n-1}$$

In the formula, n was the number of wheat tissue sequenced; $S_{(i,j)}$ was the RPKM value of ith gene in jth tissue, and $S(i, max)$ was the highest RPKM value of gene i in

the *n* organs. The τ value ranges from 0 to 1, with higher values indicating a higher level of variation in expression across tissues or higher tissue specificity. If a gene expresses in only one tissue, τ approaches 0, whereas if a gene is equally expressed in all tissues, τ value equals 1.

S2.2 Gene modeling and prediction

To predict gene structures in this genome, multiple approaches were used including: *de novo*, homology-based, EST and RNA-Seq based predictions.

1) *De novo* prediction

We used FGENESH[8] (version 1.3) with model parameters for monocots and GeneID (version 1.4) to with 'wheat.param' as parameter profile. In total, 72,420 and 40,187 raw gene models were predicted, respectively.

2) Homolog prediction

We downloaded the predicted proteins of four close species from NCBI: *B. distachyon* (version, Bradi_1.0), *S. bicolor* (v1.0), *O. sativa* (version, IRGSP v5), *Z. mays* (version, ZmB73_AGPv1); and *Hordeum vulgare* from http://harvest.ucr.edu/. These proteins were firstly mapped to *Ae. tauschii* genome using TBLASTN (E-value ≤ 1e-5), and then the accurate spliceing pattern were built with GeneWise (version 2.0) [9]. We predicted 31072, 29252, 34079, 30767, and 27059 gene models from *B. distachyon, S. bicolor, O. sativa, Z. mays,* and *H. vulgare*, respectively.

3) EST prediction

We used BLAT[5] to align the EST to the genome with identity ≥98% and coverage ≥95%. We then used PASA (http://www.lerner.ccf.org/moleccard /qin/pasa/) to link the spliced alignments for accurate gene structures. We predicted 63,574 gene models after filtering out gene models with gaps in CDS regions.

4) RNA-Seq approach

The transcriptome assembled results were mapped onto the genome by BLAT with identity ≥ 99% and coverage ≥ 95%. A total of 38,561 candidate regions were identified. Second, we utilized TopHat[10] to identify exon-intron splicing junctions and refine the alignment of the RNA-Seq reads to the genome. The software Cufflinks[11] (Version, 1.2.0 release) was then used to define a final set of predicted genes.

To generate a final consensus gene set, we integrated evidence from four predictions with the method described as following:

Set the gene models from *de novo* prediction (FGENESH and GeneID) as the target dataset A. We aligned the protein sequences from the other three predictions to the target dataset A and retained those non-redundancy genes (34,498) that were

supported by at least one evidences (**Supplementary Table 6**). Set the gene models from RNA-Seq alignment as the target dataset B. We aligned the protein sequences from EST- and homolog-based predictions (query dataset) to the target dataset B and retained those non-redundancy genes (8,652) that were supported by at least one evidences. Note that only if a target gene has an overlap ≥60% with the predicted gene from the query dataset is considered as a high confidence gene. **Supplementary Figure 4** shows the flowchart used in the gene prediction. Combining the gene sets obtained above, we removed the redundancy genes and finally obtained a consensus gene set containing 43,150 genes. The characteristics of annotated genes were displayed in **Supplementary Figure 5.**

S2.3 Non-coding RNA annotation

1) Identification of tRNA genes

We searched the whole genome for tRNA with software tRNAscan-SE[12] using default settings. After removing the SINE-derived tRNA genes, we predicted 2,505 tRNA genes with the average length of 73 bp.

2) Identification of rRNA genes

We searched the whole genome for rRNA genes by aligning the *T. aestivum* 5S, 5.8S, 18S and 28S rRNA sequences obtained through searching the public domains (Accession numbers: AF150611, AY346115, AJ272181 and AY049041) in NCBI. The alignment was conducted by BLAST (E-value ≤ 1e-5, >85% identity and a match length ≥50bp). With this method, we predicted 358 rRNA genes with an average length of 228 bp.

3) Identification of other ncRNA gene

We aligned genome sequences against Rfam database (version 10.1, http://rfam.sanger.ac.uk) and predicted snRNA and snoRNA with INFERNAL[13]. We filtered the predicted snRNA and snoRNA genes under arbitrary criteria (≤1 mismatch & ≥85% identity compared with a combined snoRNA and snRNA database from *H. vulgare*, *O. sativa*, *Z. mays*, *T. aestivum* and *S. bicolor*). Finally, we identified 35 snRNA genes with an average length of 166 bp and 78 snoRNA genes with an average length of 121 bp.

In **Supplementary Table 7,** we summarized the statistics of non-coding RNAs identified in the *Ae. tauschii* genome.

S2.4 Repetitive sequence identification

We identified the repeat sequences in *Ae. tauschii* genome by searching tandem repeat sequences and transposable elements (TEs). The former was identified using Tandem Repeats Finder (TRF, v4.04)[14], and the latter was detected using a

combination of homolog-based and *de novo* approaches. The homology approach was based on the TE library combined with Repbase (v15.02) and TIGR (v3.0). We used RepeatMasker (v3.2.9, http://www.repeatmasker.org) to find TEs with the TE library. For *de novo* prediction, we used RepeatModeler (v1.0.3, http://www.repeatmasker.o-rg/RepeatModeler.html) to get TE consensus sequences, which was used as a library to predict the TEs by software RepeatMasker. TE sequences were classified based on the reported system[15] (**Supplementary Table 8**).

All scaffolds were used as queries in BLASTN searches against the TREP database (Release 11, beta-version, wheat.pw.usda.gov/ITMI/Repeats/). BLAST outputs were parsed by a Perl program which is available upon request. Regions that had homology to the same TE family and were separated by < 150 bp were assumed to belong to the same TE copy. The reason for this repeat merging step is that many TE families (especially Class 2 elements) contain highly variable regions that evolve rapidly and show virtually no sequence homology even between otherwise very closely related copies of the same family. Classification of TE sequences was based on the system proposed by Wicker *et al*[15]. For calculation of the total TE content, we only considered the total number of non-N bases in the sequences.

All 270,114 sequence scaffolds were screened for the presence of TE sequences, 84.6% of which were identified with TE sequences, indicating that approximately 15% of the assembled scaffolds do not contain any known repeat family. However, this does not necessarily mean that these 15% are mainly low-copy sequences because the wheat genome apparently contains a large proportion of yet uncharacterized repeat sequences. Known TE families contributed approximately 50% of all scaffold sequences.

The TE composition of the D genome assembly differs strongly from those of other Triticeae genomes such as barley and *T. aestivum*[16,17]. We assume that is because repeated sequences can cause mis-assembly leading to multiple copies being to collapse into the same sequence contig. It means that highly repetitive TE sequences tend to be under-represented in the final sequence assembly, thus limiting the possibilities of quantitative analyses. In contrast, the raw sequences produced by the shotgun method with next generation sequencing technologies were described to be less biased. Thus, we performed a quantitative analysis of TE sequences on ~5 million Illumina raw reads. As a result, 62.3% of all Illumina reads could be classified as TE sequences. This figure is comparable to findings of previous and ongoing studies on the composition of Triticeae genomes[16] (Middleton et al., in preparation). Gene space is assumed to contribute 2-3% to the entire genome. Thus, approximately one third of the sequences could not be classified. Bennett and Smith estimated that at least 80% of the Triticeae genomes are comprised of repetitive DNA[18]. One has therefore to assume that a large portion of the uncharacterized fraction is comprised of yet unknown repeat families.

In total, we found 410 different TE families. As in all Triticeae genomes studied

so far, the most abundant are *Copia* LTR retrotransposons of the Angela/BARE1 clade. As described for barley[16], the *Gypsy* LTR retrotransposons Sabrina and WHAM are also among the most abundant ones, counting for 8% and 5.6% of the whole genome, respectively. In contrast to barley, the CACTA transposon Jorge is the third most abundant TE family in *Ae. tauschii* while Jorge is found only in minuscule amounts in the barley genome[16]. On the other hand, BAGY2, which counts for over 5% of the barley genome, is virtually absent from the *Ae. tauschii* genome. In total, the 20 most abundant TE families count for over 50% of the *Ae. tauschii* genome (**Supplementary Figure 6**).

To confirm that there is less bias in whole-genome shotgun reads than in assembled scaffolds, we compared the results from the raw Illumina reads with a shotgun genome sample from *Ae. tauschii* which was produced with Roche/454 technology. We found that the representation of the different TE families is very similar in both datasets (**Supplementary Table 9**), which is consistent with our expectation.

S2.5 Dating the insertion time of LTR retrotransposons

LTR retrotransposons were clustered according to the internal sequence using CD-HIT[7], with a threshold of 90% global sequence identity. The longest sequence of each cluster was chosen as the representative sequence. To date the insertion time of LTR retrotransposons, we only considered the clusters with more than 10 copies, each covering at least 90% of the length of the representative sequence cluster. For each of these clusters, we aligned these long (>90 % query coverage) elements to the representative and selected those which aligned with at least 50 % of the length of the representative's LTRs. The two LTRs of each selected element were aligned and the date of divergence was calculated using Kimura's two-parameter method[19]: if P is the transition fraction in the aligned sequences, Q is the transversion fraction; K is the evolutionary distance, T is the time of divergence and k is the evolutionary rate, then $K = -1/2 * \ln[(1-2P-Q) * \sqrt{(1-2Q)}]$ and $T = K/2k$. We used a value of k as 1.3×10^{-8} substitutions/site/year, which was from the rate calculated for the *Adh* locus in grasses[20], and divided by two as LTR retrotransposons have a higher substitution rate than genes. To estimate the activity of repeats in genome expansion with respect to the genome structural change, we dated the insertion time of all LTR retrotransposons in families of 10 or more members, including both *Gypsy* and *Copia*. As shown in **Supplementary Figure 7**, a peak of increased insertion activity was found 3~4 mya, suggesting that the expansion of the D genome was relatively recent, coincident with climate change during the Pliocene Epoch[21].

We also calculated the insertion time of *Gypsy* and *Copia* during *Ae. tauschii* genome evolution (**Supplementary Figure 8**). We investigated the divergence rate of TEs against TE RepBase library and the pattern showed that most of copies of TE had a >10% divergence rate (**Supplementary Figure 9**). The lack of low rate TEs indicates that nearly all of the TEs had a long divergence time, supporting our results

of the LTR dating time.

S2.6 Genome duplication and phylogenetic tree

After the identification of syntenic blocks, pairwise protein alignments for each gene pair were first constructed with MUSCLE (http://www.drive5.com/muscle/). Nucleotide alignment was then created according to the protein alignment. 4DTv was then calculated on concatenated nucleotide alignments with HKY substitution models. The 4DTv distribution of duplicate gene pairs in *Ae. tauschii* genome, *Brachypodium*, rice and sorghum was displayed in **Supplementary Figure 10.** It indicates no recent duplication events happened in these species.

We also performed BLASTP with E-value 1e-5 and identity ≥ 30% to estimate duplication in genome. Pairwise Ks values of homologous genes in *Ae. tauschii* genome were used to infer the time of whole genome duplication event. The bottom histogram plot shows pairwise Ks values for gene family sizes ≥ 7 (in total 3,082 genes). The peak at ~0.36 indicates an ancient duplication in *Ae. tauschii* genome about 60 mya, considering a substitution rate $\lambda = 6.1 \times 10^{-9}$ (mean of $6.1–7.1 \times 10^{-9}$) per site per year[22] (**Supplementary Figure 11**).

We constructed a phylogeny tree for *Ae. tauschii, Brachypodium, Z. mays, S. bicolor* and *O. sativa* using single copy orthologous genes with Arabidopsis as an outgroup. The alignment of each gene was conducted by MUSCLE (http://www.drive5.com/muscle). Four-fold degenerate sites were then extracted from the alignment and concatenated to a super gene for each species, which were subjected to Mrbayes (version3.2)[23] for constructing the phylogenic tree with a best substation model (GTR+gamma+I). Species divergence time was estimated using MCMCTREE in PAML[24] under JC69 nucleotide substitution model and correlated rates molecular clock model. The divergence time between *Ae. tauschii* and *Brachypodium* was estimated to be 31 million years ago (mya; **Supplementary Figure 12**). The calculation time was chosen as follows: *Arabidopsis thaliana vs. Ae. tauschii* (150-250 mya); *S. bicolor vs. Z. mays* (5-15 mya); *S. bicolor vs. Ae. tauschii* (50-70 mya); *Ae. tauschii vs. Brachypodium* (15-40 mya); *Ae. tauschii vs. O. sativa* (30-50 mya)[25,26].

S3. Analysis of gene families in *Ae. tauschii*

S3.1 Genetic map construction and scaffolds anchoring

Anchoring of genomic scaffolds onto chromosomes requires a combination of resources: a high-resolution consensus genetic map and a new genetic map built specifically to aid in scaffold anchoring. Firstly, we constructed a high-resolution genetic map using a F_2 mapping population consisting of 490 plants from the cross of two *Ae. tauschii* accessions Y2280 and AL8/78. The later was sequenced for genome

assembly in this project. We genotyped the 490 F_2 plants by using restriction-site associated genomic DNA (RAD) tag sequencing on HiSeq2000 and produced 850 Gb data, from which we identified 151,083 SNP markers as the following procedure. After filtering the low quality reads, clean data from individual lines were aligned to the reference genome using Burrows-Wheeler Aligner (BWA)[27]. Then SOAPsnp[28] was used to call SNPs using the criteria as follows: (1) The Quality score of consensus genotype must be larger than 20; (2) The sequencing depth of the site must be between 2 and 200; (3) The average copy number of nearby regions must be less than 2; (4) The distance of two nearby SNPs must be larger than 1 bp. According to individual genotypes, we grouped F_2 plants using JoinMap4.0 and then ordered them using MSTmap. At last we obtained a genetic map using *kosambi* function (**Supplementary Figure 13**). This new genetic map occupied a total of 1059.806 cM and contained 13,688 scafflolds, which contained sequence information of 1.277 Gb and was an *Ae. tauschii* genetic map of highest density heretofore (**Supplementary Table 10-11)**.

The *Ae. tauschii* genome sequences were also aligned to known genetic maps, by downloaded 838 SSR sequences on three published genetic maps (Ta-SSR-2004[29]; Ta-Synthetic/Opata-GPW[30]; wheat-Composite2004, http://wheat.pw.usda.gov). The SSR sequences were aligned against the assembled genome of *Ae. tauschii* using BLAST with default parameters. For this dataset a total of 422 scaffolds were anchored to a genetic map (**Supplementary Figure 14**). The map can be viewed at http://ccg.murdoch.edu.au/cmap/ccg-live/ at a greater magnification to show all of the markers. Four additional molecular genetic maps with sequenced-based genetic loci, namely ESTs[31] and single nucleotide polymorphisms[32,33] (SNPs; Zhang et al submitted), were also used to position scaffolds into genetic maps (**Supplementary Figure 15**). The maps examined included those each developed from an *Ae. tauschii* cross[31], *Synthetic x Opata*[33], *Avalon x Cadenza*[32] and *Westonia x Kauz* (Zhang et al submitted).

To put *Ae. tauschii* scaffolds into the genetic order, a map of collinear orthologous relationships was constructed in the following steps: We mapped ESTs and SNP markers onto the scaffolds to locate the scaffold positions using BLAT. The orthologous genes between *Brachypodium* and *Ae. tauschii* were identified using the CIP-CALP method[34]. The CIP (Cumulative Identity Percentage) and CALP (Cumulative Alignment Length Percentage) statistics were used to identify the best pairwise alignment. Appropriate values (60% CIP, 70% CALP) were used to identify true orthologous genes between *Ae. tauschii* and other species, following the previous criteria[34]. We defined an orthologous block when more than five pairs of orthologous genes were present along the same chromosome of *Brachypodium* in order. The distribution of anchored scaffolds on each chromosome is shown in the **Supplementary Tables 12-13.** The scaffolds with orthologous genes within an orthologous block were put in the map with the gene order on *Brachypodium*, sorghum and rice chromosomes to produce the final alignments shown in **Supplementary Figure 16**.

The rate of non-synonymous (Ka) versus synonymous (Ks) substitutions was calculated using software KaKs_Calculator[35]. A total of 628 *Ae. tauschii* genes exhibited Ka/Ks values >0.8 when compared with one or two species. These genes were assigned a wide range of molecular functions in Gene Ontology (GO) analyses (**Supplementary Table 14**).

S3.2 Construction of gene families

We used the OrthoMCL (software version 2.0)[36] to define gene family clusters for *Ae. tauschii*, *Brachypodium*, *S. bicolor*, *O. sativa* and *H. vulgare* gene models (datasets used see below). In a first step, pairwise sequence similarities between all input protein sequences were calculated using BLASTP with an e-value cut-off of 1e-05. Markov clustering of the resulting similarity matrix[37] was used to define orthologous cluster structure, using an inflation value (-I) of 1.5 (OrthoMCL default).

The input datasets were:

B. distachyon: v1.2 MIPS[25].

S. bicolor: v1.4 MIPS[38].

O. sativa: MSU7[39].

H. vulgare: fl-cDNAs (28,592), clustered with CD-HIT[7] to remove redundancies.

Ae. tauschii: gene models described in this paper.

Splice variants were removed from the data set, keeping the longest protein sequence prediction, and data sets were filtered for internal stop codons and incompatible reading frames. A total of 115,666 coding sequences from these five grasses were clustered into 23,202 gene families. 8,443 clusters contained sequences from all five genomes.

The syntenic relationship of the *Ae. tauschii* genome to other grasses was also investigated using the *in silico* "chromosome painting" described by Mayer *et al* (2011) *versus Brachypodium*, rice, sorghum and barley[40]. The barley chromosomes were compiled by concatenation of barley fl-cDNAs as anchored by the barley genome zipper[40]. Therefore, we identified the *Ae. tauschii* gene models grouped together with at least one *Ae. tauschii* gene as determined by standard OrthoMCL clustering procedures. The gene density distributions of clustered genes were computed against the reference grass genomes *Brachypodium*, and rice, and sorghum (**Supplementary Figure 17).** Following the protocols established by Mayer *et al* (2011), the number of clustered reference genes was counted for each genome position (500 kb window and 100 kb shift for *Brachypodium*, rice and sorghum; 250 kb window and 50kb shift for barley) and visualized as heatmap.

The barley map reported by Mayer *et al* (2011) was aligned to the *Ae. tauschii* genome scaffold based genetic maps via putative orthologous gene pairs between *Ae. tauschii* genes and barley gene sequences that were defined by best bidirectional blast

hit comparisons within each OrthoMCL group. The alignments to the Poland *et al* (2012) and Luo *et al* (2009) maps are illustrated because the former map had the highest density of *Ae. tauschii* genome scaffolds and was based on the presence of gene sequences. Although apparent breaks in syntenic regions are evident, the overall alignment contributes to validating the assignments of *Ae. tauschii* genome scaffolds to genetic locations. In each panel, barley chromosomes 1H through to 7H are shown at the base and above each barley chromosome the *Ae. tauschii* genome genetic maps 1D through to 7D are arranged in order (**Figure 1**). The colored lines (matching the color of the respective D genome chromosome) indicate putative orthologous pairing of an *Ae. tauschii* gene and a gene sequence in the barley map. Analogous comparision of *Ae. tauschii* against the *Brachypodium* chromosomes Bd1 through Bd5 was performed.

S3.3 Over- and under-representation analysis using GO and Pfam

To assess the functional gene repertoire of *Ae. tauschii* in contrast to the reference grass organisms *Brachypodium* and rice, we computed GO Slim functional categories for all three organisms. The distribution of the different GO Slim molecular function categories among the total gene content is plotted in **Supplementary Figure 18** for each species. Additionally, we computed the ratio of each GO category and Pfam domain in the total GO/Pfam reservoir of a particular organism for *Ae. tauschii*, rice and *Brachypodium*. We then compared these ratios pairwise between genomes, e.g. *Ae. tauschii* and *Brachypodium*, and plotted them as total and normalized differences in **Supplementary Figure 19**. For each species combination and species we extracted the 10 most deviating terms, i.e. the ones differ the most in their ratio between the two organisms. The results for the comparison of Pfam terms in *Ae. tauschii* and *Brachypodium* are given in **Supplementary Tables 15-16**.

We computed Pfam domain signatures and GO terms for the gene annotations of the organisms *Ae. tauschii*, *B. distachyon*, rice, barley and sorghum using InterproScan[41]. Only GO terms from the category of molecular function were considered because of better transferability and comparability. To identify GO terms over- and under-represented in the gene sets of expanded and contracted *Ae. tauschii* gene families, we used the GOstats R package from Bioconductor (http://www.bioconductor.org/packages/release/bioc/html/GOstats.html). Significant terms are reported up to a p-value of smaller than 0.05. To identify Pfam domains over- and under-represented we used in-house software using Bonferroni correction for multiple testing. To assess GO Slim terms (molecular function category only) for lists of plain GO terms we used AgBase[42] (http://agbase.msstate.edu/cgibin/tools/goslimviewer_select.pl) with the 'Plant Slim/TAIR version Aug.2011' GO Slim set.

The gene members of orthologous families for *Ae. tauschii*, *Brachypodium*, rice, sorghum and barley (see **Supplementary Information S3.2**) were used to determine the gene family sizes of *Ae. tauschii* by counting the incorporated *Ae. tauschii* genes for each cluster. We compared the observed *Ae. tauschii* gene family size relative to

their gene family size in the sequenced reference grass species. For 14,970 OrthoMCL groups that contain at least one *Ae. tauschii* sequence and at least one sequence of *Brachypodium*, rice or sorghum, we defined the gene family size as number of incorporated reference grass genes. Following the completeness of the reference grass genome and its evolutionary distance criteria we used the following hierarchy for selecting the representative gene family size:

Number of clustered *Brachypodium* genes (13,381 orthologous gene clusters);
Number of clustered rice genes (764 orthologous gene clusters);
Number of clustered sorghum genes (238 orthologous gene clusters);
Number of clustered barley fl-cDNAs (587 orthologous gene clusters);

Then, we calculated the number difference of the gene family size and the observed gene copy number of *Ae.* tauschii. For gene families with ≤10 members the median of the observed *Ae. tauschii* gene count was determined and a polynomial fit of these values was calculated using locally-weighted polynomial regression (http://stat.ethz.ch/R-manual/R-patched/library/stats/html/ lowess.html). We compared the observed *Ae. tauschii* gene family size relative to their gene family size in the sequenced reference grass species (**Supplementary Figure 20**). In total, 80% (11,980 out of 14,970) of the groups that contain at least one *Ae. tauschii* sequence and at least one sequence of *Brachypodium*, rice, sorghum or barley showed no difference in gene family size. We identified 471 significantly expanded gene families located above the 95th percentile of the gene copy number frequency distributions. This distribution indicates that the degree of expansion and contraction of the *Ae. tauschii* gene families is similar to other grass species.

S3.4 Identification of NBS-LRR encoding genes

NBS-LRR encoding genes (*R*-genes) and their homologues were identified in an iterative process using HMM and BLAST. First, a model of NB–ARC domain (Pfam PF00931) representing the nucleotide binding site was selected to search against the predicted proteins of *Ae. tauschii* genome using hmmer3.0 (http://hmmer.janelia.org/) with default parameters[43]. To verify the result of HMM search, another protein database containing sequences retrieved from GenBank (http://www.ncbi.nlm.nih.gov) using key words "NBS, NBS-LRR, disease resistance" was constructed. Sequences with significant hits from the HMM search were used as queries to BLAST against the newly constructed protein database to verify that they encode NBS domain (E-value cutoff 1e-10). All verified *R*-gene sequences were used as queries to BLAST the *Ae. tauschii* genome for more homologues that were failed to be identified in the first step due to their divergence or incompleteness. The newly identified *R*-genes were verified if they had significant hits with any R protein in the database. The above process was repeated until no new *R*-gene-related sequences were identified. The physical distribution of disease resistance gene-analogs were analyzed through the coordinates in the scaffold.

The D genome of *Ae. tauschii* is considered an important gene pool for genetic

improvement of wheat in disease resistance[44]. A total of 1,219 *Ae. tauschii* gene models were identified with significant similarity to the NBS-LRR genes or so called R gene analogues (RGAs)[45,46]. This number is twice of that in rice (623) and six times of that in maize (216)[47], indicating that the RGA family has significantly expanded in *Ae. tauschii*. These RGAs can be grouped into 567 sub-families in which members were >80% identical in nucleotide sequences to at least one other member (**Supplementary Figure 21**), with the largest sub-family of 34 members and 322 single member sub-families. A total of 360 RGA sub-families contained at least one member significantly similar (E-value <1e-5) to a rice RGA, whereas the remaining 216 did not. In total 112 sub-families showed no significant similarity to *Brachypodium* RGAs. We found that 283 *Ae. tauschii* RGAs were organized in tandem forming more than 100 clusters (**Supplementary Figure 22**), a feature also found in the Arabidopsis and rice genomes[48,49]. Domain search also showed over-representation of protein kinase domains among predicted *Ae. tauschii* proteins (**Supplementary Table 15**), integral components of disease resistant genes, such as the stripe-rust resistant gene *Yr36* and the necrotrophic pathogen sensitive gene *Tsn1*.

S4. The contribution of *Ae. tauschii* to hexaploid wheat

S4.1 Cold resistance genes

Wheat, one of the most cold tolerant plant species in grass, can tolerate approximately −20°C[50]. One of major objectives for most winter wheat breeding programs in regions subject to severe winters is to select lines that minimize the effect of freeze damage during the vegetative phase. Frost-tolerant wheat varieties show an increase in freezing tolerance after exposure to low, non-freezing, temperatures, a phenomenon known as cold acclimation[51]. During cold acclimation, winter wheat adjusts its metabolism to low temperature and protects critical cell structures against the effect of freezing temperatures. Research results show that a large number of genes are being altered during the process of cold acclimation[52]. Recently functional genomics research results involved in plant response to low temperatures resulted in the characterization of several genes directly involved in stress perception, signal transduction and transcriptional regulation of cold regulated (COR) genes[53-55]. CBF transcription factors constituted a regulatory hub for cold acclimation[56]. These transcription factors and their target genes were the fundamental part of the signal cascade leading to acclimation and acquisition of frost tolerance in many different plant species[54]. Genes involved in this pathway include other transcription factors, late embryogenesis abundant (LEA) proteins, cold-regulated (COR) and cold-inducible (KIN) proteins, osmoprotectant biosynthesis proteins, carbohydrate metabolism-related proteins, phospholipase C enzymes, sugar transport proteins and so on[53,57]. Another pathway involved in cold acclimation in cereals was vernalization in which *VRN1* was the connection between freezing tolerance and flowering[58]. We collected information for 178 genes derived from the above two pathways and analyzed and compared those genes from other four sequenced grass species by

BLASTP search using the threshold value as following: query coverage > 50%, e-value < le-30, identity > 50%.

We investigated the sequenced *Ae. tauschii* genome and found 216 cold-related genes, markedly more than that in other grasses such as 164 genes in *B. distachyon*, 132 genes in *O. sativa*, 159 genes in *S.* bicolor and 148 genes in *Z. mays*. The copy numbers of three genes (chlorophyll a/b-binding protein WCAB precursor, Xyloglucan endotransglucosylase and fructan 6-fructosyltransferase) are observably higher than those from other grass species. According to the identification result of cold related genes from five sequenced grass genomes, if a gene occurred only in *Ae. tauschii* but not in other four grass species, we called it wheat-specific gene. If a gene occurred in both *Ae. tauschii* and *Brachypodium* but absent in other three grass species, we named it Pooideae-specific gene. Some genes were found to be Pooideae species specific such as those encoding ice recrystallization inhibition protein 1 precursor, DREB2 transcription factor alpha isoform and cold-responsive LEA/RAB-related COR protein. Of the 216 cold-related genes, 20 genes are wheat specific (**Supplementary Table 17**). The expression heatmap of *Ae. tauschii*-specific and Pooideae-specific cold-related genes was shown in **Supplementary Figure 23**, indicating that most of these cold-related genes were constitutively expressed in eight tissues. But some of these cold-related genes were specifically expressed in root, stem or seed.

S4.2 Transcription factors in *Ae. tauschii* genome

Transcription factors (TFs) are key regulators for transcriptional expression of genes in biological processes. We identified transcription factor families by searching for known DNA-binding domain and other domains such as auxiliary domain and forbidden domain, as described in literature[59] and classified according to the scheme reported in the plantTFDB (http://planttfdb.cbi.edu.cn/). In total, 1,489 predicted TFs were identified, including 56 families (**Supplementary Tables 18-19**) and representing 3.45% of the 43,150 predicted protein-coding loci. The most highly represented TF families were bHLH (130 genes), NAC (120 genes), MYB-related (103 genes), B3 (102 genes), MYB (103 genes), WRKY (95 genes), C2H2 (82 genes), ERF (74 genes), bZIP (71 genes), and GRAS (50 genes). TFs in *Ae. tauschii* were almost the same with those in *Brachypodium* (1,479), *O. sativa* (1,490), but fewer than those in *S. bicolor* (1,776) and *Z. mays* (1,975).

To compare the enrichment of transcription factor classes among these grass genomes, a statistical test was applied considering the ratio of each TF family to their total gene number in *Ae. tauschii* as compared with that in each of the other sequenced grass species (**Supplementary Figure 24**). The pairwise scatter plots support a strong correlation of *Ae. tauschii* TF families with *B. distachyon*. With probability set at P = 0.01, statistical analysis identified B3, M-type MADS and MYB-related as gene families more represented in *Ae. tauschii* than that in the *B. distachyon* genome. It can be concluded that the overall distribution of *Ae. tauschii*

transcription factor genes among the various known protein families is very similar between *B. distachyon* and *Ae. tauschii*. However, some families are relatively sparser or more abundant in *Ae. tauschii,* perhaps reflecting differences in biological function. And we obtained the similar results when we compared *Ae. tauschii* TFs with those from *O. sativa*, *S. bicolor* and *Z. mays*. The differences observed in relative transcription factor gene abundance may indicate that regulatory pathways in *Ae. tauschii* may differ from those described in other grass species.

We sequenced the transcripts of eight tissues in *Ae. tauschii*, and calculated RPKM of genes in different tissues to analyze co-expression of genes. We selected 14 transcription factors (TFs) in wheat which contained one drought tolerance gene[60] and applied the ARACNe algorithm[61] to infer transcriptional interactions with other genes. It was used to reconstruct accurate cellular network by inferring target TF interactions and direct interactions from large sets of gene expression profiles. We identified 71 transcriptional interactions among the TFs and 1212 transcriptional interactions between TFs and other genes. According to the result of ARACNe, we completed the network of those genes by Cytoscape (**Supplementary Figure 25** and **Supplementary Table 20**).

S4.3 miRNA analysis

S4.3.1 Small RNA library development and sequencing

Mixed tissues of leaf, root, stem, and spikes were used for total RNA isolation using Trizol (Invitrogen, Carlsbad, CA). Small RNAs were enriched by polyethylene glycol precipitation, separated on 15% denaturing PAGE, and visualized by SYBR-gold staining. Small RNAs of 16–28 nt were gel-purified. Small RNAs were ligated to a 5' adaptor and a 3' adaptor sequentially, reverse-transcription polymerase chain reaction (RT-PCR) amplified, and used for sequencing directly. Sequencing was performed on Illumina HiSeq2000 platform. Raw sequences were processed by removing adaptors and mapped to the D genome assembly. Conserved miRNAs were identified by comparing with miRBase plant miRNA sequences (http://www.mirbase.org/) and secondary structures were predicted using mFold [62]. Similar protocol was followed for novel miRNA discovery with additional consideration of the presence of miRNA*s.

S4.3.2 miRNA gene and target prediction

Primer sequences were removed from raw sequences using either cross-match[63] or Perl scripts. Small RNA sequences of 16 to 25 nt were collected for analysis. Identical sequences were removed using an in-house Perl script. For stem-loop structure prediction, Repbase repeats[64], TIGR wheat repeats, RFam RNA sequences, and known miRNAs from miRBase were first screened from scaffold genomic sequences using RepeatMasker. Novel miRNAs were predicted by MIREAP software (http://sourceforge.net/projects/mireap/) with the parameters as following: Minimal vs.

Maximal miRNA sequence length (18 vs. 25); Minimal vs. Maximal miRNA reference sequence length (20 vs. 23); Maximal copy number of reference miRNAs (20); Maximal free energy allowed for a miRNA precursor (-18 kcal/mol), Maximal space between miRNA and miRNA* (300); Minimal base pairs of miRNA and miRNA* (16); Maximal bulge of miRNA and miRNA* (4); Maximal asymmetry of miRNA/miRNA* duplex (4); Flank sequence length of miRNA precursor (20). Target genes of miRNAs were predicted using the criterion first used by Allen *et al*. (2005), in which mismatched bases were penalized according to their location in the alignment[65]. All other data processing and graphical display were performed using in-house Perl scripts. The distribution of miR395 and miR2118 on the scaffold was shown using circos software[66].

S4.3.3 GO enrichment of the target genes

The functional enrichment of miRNA targets was performed using BiNGO software[67] and Cytoscape plugin[68] was used to display GO hierarchy tree. For enrichment p-value calculation (at a significance level of < 0.05), hypergeometric test method was applied. For multiple hypotheses testing, false discovery rate (FDR) correction of Benjamini and Hochberg method was used to reduce false negatives[69]. The GO annotation of target genes is available from the agriGO website (http://bioinfo.cau.edu.cn/agriGO/).

S4.3.4 miRNA analysis

A total of 185 conserved miRNAs from 27 families were present in *Ae. tauschii*. While the family sizes of miR159-166-167 were reduced, the sizes of miR169-399-1436-2118-2275 families were expanded, with members of miR399-1436-2118 families nearly doubled compared with *O. sativa* (**Supplementary Table 21**). Segmental and tandem duplications were found to be the major mechanism for miRNA gene family expansion. For instance, 42 members of the miR2118 family were organized as two groups on 15 scaffolds (**Supplementary Figure 26**). There are 26 members with high sequence similarity (higher than 80%). The high sequence similarity among the miRNAs indicated that these expansions took place at the recent stage. Although the functions of these phasiRNAs are still unknown, the ubiquity of these types of siRNAs in the inflorescence of a number of crop species suggests their importance in spike development. Further, conserved miRNAs also appeared to evolve new functions in response to stimulus as shown by the enriched Gene Ontology (GO) terms (**Supplementary Figure 27**), indicating overall elevation in miRNA repertoires for abiotic stress responses in this wild grass. Finally, a total of 159 (133 families) novel miRNAs were predicted. Target analysis showed strong enrichment for genes in cell death process (such as *NB-ARC*, *NBS-LRR*, *RGH1A*, *MLA*, *Yr10*, and *Xa1*)[70,71] suggesting enhanced disease resistant capability (**Supplementary Figure 28**). Together, our miRNA data supported the notion that the DD genome from the diploid ancestor *Ae. tauschii* contributed significantly to the adaptation for more robust hexaploid wheat.

S4.4 Genes involved in grain quality

Using the gene protein sequences downloaded from GenBank as queries and searching our gene model nucleotide database with TBLASTN program, we identified the candidate genes for wheat quality gene. To verify the result of above BLAST search, we BLASTed these candidate sequences against nr database online and obtained 12 wheat grain quality genes including 2 genes for *HMW-GS*, 5 for *LMW-GS*, 2 for *Pinb*, and each for *Pina*, *GSP* and *SPA* genes. Using transcriptome data from eight tissues, we analyzed the expression levels for these 12 quality genes (**Supplementary Figure 29)**.

S5. Application in wheat molecular breeding

S5.1 Agronomically important loci in *Ae. tauschii*

As noted in **Supplementary section S3.1** the physical map of *Ae. tauschii* was aligned to 838 SSR sequences on three reported genetic maps (see http://ccg.murdoch.edu.au/cmap/ccg-live/cgi-bin/cmap/viewer and GrainGenes: http://wheat.pw.usda.gov/GG2/index.shtml). In total, 422 scaffolds were anchored to the genetic map. For QC (quality control) of our physical map, we searched 36 QTLs/genes on the composite map (wheat-composite2004, http://wheat.pw.usda.gov) and positioned them to the map; we found that all of them were located on our constructed physical map (**Supplementary Figure 14).** The *Ae. tauschii* genome sequence also provides a reference to integrate multiple published genetic maps based on different types of markers. The co-localization of dense scaffolds and genetically mapped QTLs/genes should assist in candidate gene-based gene cloning. For example, there were 33 QTLs/genes located on chromosome 2D previously, and all of them were integrated in our 2D co-location scaffold map (**Figure 3** with details in **Supplementary table 22**)

Though many agronomical traits have been modified in the process of domestication using the traditional breeding methods, some genes were cloned by map-based cloning methods in wheat: *Lr1, Lr10, Lr21, Lr34, Pm3, Yr36, Tsn1* associated with disease resistance; *Vrn1*, *Vrn2* and *Vrn3* associated with vernalization; *Gpc-B1* is responsible for grain protein content; *Rht1, Rht2* associated with plant height; and *Q* gene conferring free-threshing (see details in **Supplementary table 23)**. To identify more genes associated with the traits of interests in *Ae. tauschii*, we collected reported genes from common wheat, *O. sativa*, *H. vulgare* and *Z. may,* and then searched their orthologous in *Ae. tauschii*. With this method, we identified 28 genes associated with several important agronomical traits in the *Ae. tauschii* genome.

We primarily summarized the functions of these genes as following:

While *Lr1* is not very effective in controlling leaf rust due to the fact that most *P.*

triticina virulence phenotypes are virulent to *Lr1*, the *Lr1/Avr1* interaction is a good example of a classical gene-for-gene system. The overexpression of *Lr10* resulted in enhanced resistance with a complete prevention of rust sporulation. *Lr21* is a potentially durable and highly effective leaf rust resistance gene in wheat. *Lr34*/*Yr18* provides durable resistance for leaf rust and stripe rust. *Pm3* confers a specific resistance to Bgt races in an allelic manner. *Pm21* confers durable and broad spectrum resistance to wheat powdery mildew. *Ppd-D1* on chromosome 2D is the major photoperiod response locus in hexaploid wheat. In hexaploid wheat, the requirement for vernalization is mainly regulated by the vernalization gene *VRN1*. Wheat vernalization gene *VRN2* is a dominant repressor of flowering that is down-regulated by vernalization. *VRN3* is a promoter of flowering up-regulated by long days. *Rht1* reduces stem elongation in varieties by causing limited response to the phytohormone gibberellin (GA), resulting in improved resistance to stem lodging and yield benefits through an increase in grain number. The *Q* gene is largely responsible for the widespread cultivation of wheat because it confers the free threshing character. It also pleiotropically influences many other domestication-related traits such as glume shape and tenacity, rachis fragility, spike length, plant height, and spike emergence time. The presence versus absence of *GSPs* in single seed starch preparations is co-inherited with grain softness versus hardness. *Puroindoline a* and *Puroindoline b* (*Pina* and *Pinb*, respectively) genes together compose the wheat (*T. aestivum* L.) *Ha* locus that controls grain texture and many wheat end-use properties as well. Grain protein content (GPC) is important for human nutrition and has a strong influence on pasta and bread quality. The dominant allele at the *DEP1* locus is a gain-of-function mutation causing truncation of a phosphatidylethanolamine-binding protein-like domain protein. The effect of this allele is to enhance meristematic activity, resulting in a reduced length of the inflorescence internode, an increased number of grains per panicle and a consequent increase in grain yield. A single locus (*nud*) controls the covered/naked caryopsis phenotype of barley. *Vrs1* gene controls the development and fertility of the lateral spikelets of barley. *GW2* is involved in rice grain development, influencing grain width and weight. *Erect panicle2* (*EP2*) regulates panicle erectness in *indica* rice. *GS5* plays an important role in regulating grain size and yield in rice. *IPA1* defines ideal plant architecture in rice. *qSH-1* causes the reduction of seed shattering during rice domestication. *sh4* is involved in the degradation of the abscission layer between the grain and the pedicel, affecting seed shattering. *MOC1* controls tillering in rice. *tga1* exposes the kernel on the surface of the ear such that it could be readily utilized as a food source by humans. The *tb1* gene largely controls the increase in apical dominance in maize relative to teosinte.

S5.2 SSR analysis

Simple Sequence Repeats (SSRs) in the *Ae. tauschii* genome were predicted using SSRLocator[72]. The predicted SSRs were classified into six types according to the copy number they tandemly arranged: monomer (one copy), dimer (two copies), trimer (three copies), tetramer (four copies), pentamer (five copies), hexamer (six copies). Each type was classified into two subgroups according to the SSR length:

Class I (≥ 20 bp) and Class II (≥ 12 and < 20 bp). The statistics of SSRs (mono- up to hexamers) were shown in **Supplementary Table 24**. In *Ae. tauschii*, trimers (37.7%) and tetramers (27.5%) composed more than half of the SSRs (65.2%), which is more than that in *A. thaliana* (50.0%), *O. sativa* (62.0%) but smaller than that in *B. distachyon* (70.3%)[25]. We observed that SSRs are overwhelmingly present in intergenic (88.9%) regions compared with that in exonic (1.4%) and intronic (9.7%) in *Ae. tauschii* genome. Furthermore, trimers predominate in exons (57.0%) while trimers and tetramers predominate in introns (60.7%) and intergenic regions (65.6%). Besides, *Ae. tauschii* genome possesses much more class I SSRs than class II.

S5.3 SNP calling

SNPs was detected using SOAPsnp (version 1.05), a program inferring the genotype with highest posterior probability at each site on Bayes' theorem. After filtering low quality reads, 15 Gb of the *Ae. tauschii* accession Y2280 were aligned to the reference genome with the scaffolds that are smaller than 5 Kb excluded using Burrows-Wheeler Aligner (BWA). As a result, 93.10% bases were aligned to the reference genome. Then SOAPsnp[28] was used to call SNPs for this accession. The candidate SNPs retrieved were further filtered using the criteria as follows: (1) The Quality score of consensus genotype must be larger than 20; (2) The sequencing depth of the site must be between 4 and 1000; (3) The average copy number of nearby regions must be less than 2; (4) The distance of two nearby SNPs must be larger than 5. The dataset was saved as tab-separated file in text format. In total, 711,907 high-quality SNPs were identified.

Supplementary Tables

Supplementary Table 1: Summary of sequencing data for *Ae. tauschii* genome.

Insert-size	Libraries	GA lanes	Raw data (Gb)	Usable data (Gb)	Effective Depth*	Physical depth*
~200 bp	13	18	184.46	130.45	32.45	78.25
~500 bp	7	16	104.93	92.59	18.08	119.82
~700 bp	5	13	70.41	47.41	11.79	101.55
~2 Kb	7	22	73.06	53.64	13.34	503.84
~5 Kb	6	13	78.16	51.59	12.83	828.28
~10 Kb	6	9	38.98	20.88	5.19	865.80
~20 Kb	1	1	7.55	2.20	0.55	248.43
Total	45	92	557.55	398.76	94.24	2745.97

*The genome size was estimated as 4.02 Gb [73]

Supplementary Table 2: Summary of the *Ae. tauschii* genome assembly.

	Contig		Scaffold	
	Size (bp)	**Number**	**Size (bp)**	**Number**
N90	122	4,595,168	126	3,262,222
N80	128	1,681,351	403	199,377
N70	1,122	484,898	19,551	42,993
N60	2,638	280,635	39,546	28,281
N50	4,521	179,145	57,585	19,455
Longest	115,061		720,471	
Total size	3,528,022,538		4,229,254,522	
Total number (>=1kb)		516,176		111,337
Total number (>=2kb)	338,083		82,564	

Supplementary Table 3: Distribution of scaffold length for the *Ae. tauschii* genome assembly.

Scaffold length (bp)	Number	Subtotal length (bp)	Average length (bp)	Percentage (%)
>100,000	8,235	1,265,562,262	153,680	29,92
>50,000	22,732	2,290,622,661	100,766	54,16
>30,000	34,474	2,751,775,829	79,821	65,07
>20,000	42,567	2,952,107,324	69,352	69,80
>10,000	55,685	3,140,980,208	56,406	74,27
>1000	111,243	3,325,568,862	29,894	78,63

Supplementary Table 4: Statistics of BACs from public database matched by scaffolds of *Ae. tauschii* genome. The genomic sequence covered 87.80% of the total BACs.

BACs	BAC size (kb)	Matched scaffolds	Scaffold size(kb)	BAC region matched by scaffolds (kb)	Scaffold region matched by BACs (kb)	Mapping region gene number*	Maping region gene length (bp)
108D7_2008-10-3_223835_contig534	66	scaffold55989	103	0-56	46-103	0	
95F14_2008-10-16_234908	78	scaffold9085	196	0-78	27-106	0	
C4_2008-10-2_212410.seq	189	scaffold552	120	7-128	0-120	1(1)	5595
		scaffold66242	44	142-188	0-44	0	
gi|188038067|gb|EU660891.1|	89	scaffold13648	157	0-89	60-148	0(4)	
gi|188038087|gb|EU660897.1|	123	scaffold16044	77	0-9	65-77	0	
		scaffold19158	216	9-30	39-60	0	
		scaffold22383	123	32-68	2-46	0	
		scaffold15750	78	67-123	15-78	0	
84G23_2008-9-19_170424	44	scaffold55989	103	0-26	0-26	0	
		scaffold48675	20	26-43	7-17	0	
BAC32I6-12D11	146	scaffold41959	80	0-44	0-42	3(3)	10523
		scaffold2339	63	45-84	30-52	0(2)	
		scaffold8852	36	86-104	0-10	2(2)	3562
		scaffold2880	54	104-128	6-32	0	
		scaffold47315	76	127-146	0-20	0	

*Numbers in brackets represent gene number located on scaffolds.

Supplementary Table 5: Transcriptome analysis of different tissues of *Ae. tauschii* by RNA-seq.

Organs	Clean data (Gb)	Transcripts	Average length (bp)	Maximum length (bp)	Total size (Mb)
Pistil	5.97	40648	714	5834	29.0
Root	7.17	64852	864	15252	56.0
Seed	6.06	45048	881	11985	39.7
Spike	6.09	55101	905	15276	49.8
Stamen	5.16	41164	765	7178	31.5
Stem	8.06	38962	955	10539	37.2
Leaf*	7.27	NA	NA	NA	NA
Sheath*	7.44	NA	NA	NA	NA
Integration[#]	53.21	126218	932	15277	117.7

[#]all RNA-seq data were put together and assembled.

*reads from this organ were not used for assembly.

Supplementary Table 6: Comparison of gene numbers and features of five monocot genomes.

Species	***Z. mays***	***S. bicolor***	***B. distachyon***	***O. sativa***	***Ae. tauschii***
Total size	34467198	40003016	33149479	32931411	41505222
Maximum length	11148	14487	15384	15714	15360
Gene number (>100 bp)	32497	34496	25528	33189	34498
Gene number (>1Kb)	14084	16966	14293	13436	17202
Gene number (>2Kb)	3354	4548	4073	3061	5191
mRNA length*	1584/2595	1810/2616	2227/2953	1478/2144	2105/2931
CDS length*	879/1059	987/1159	1098/1298	810/990	981/1200
Exon length*	138/260	143/270	133/251	140/261	134/243
Exon number*	3/4.1	3/4.3	3/5.2	2/3.8	3/4.9
Intron length*	144/500	143/442	153/396	148/412	192/442
Intron number*	2/3.1	2/3.3	2/4.2	1/2.8	2/3.9
Gene GC	0.478	0.464	0.458	0.458	0.461
Exon GC	0.562	0.545	0.543	0.557	0.531
Intron GC	0.418	0.399	0.391	0.373	0.408

*median value/average value.

Supplementary Table 7: Summary of non-coding RNAs in the *Ae. tauschii* genome.

Type	Subtype	Copies	Average length (bp)	Total length (bp)
tRNA	-	2505	72.64	181953
rRNA	18S	42	551.74	23173
	28S	64	503.17	32203
	5.8S	18	369.33	6648
	5S	234	84.86	19857
	subtotal	358	228.72	81881
snRNA	-	35	166.49	5827
snoRNA	CD-box	75	121.27	9095
	HACA-box	3	117.00	351
	subtotal	78	121.11	9446

Supplementary Table 8: Statistics of repeat contents in the assembled genome of *Ae. tauschii* and other four monocots.

	Percentage of genome (%)				
	B. distachyon	***S. bicolor***	***O. sativa***	***Z. mays***	***Ae. tauschii***
Class I :Retrotransposon	21.58	50.77	21.00	76.35	44.03
LTR-Retrotransposon	18.38	49.70	19.85	75.52	41.35
LTR/*Gypsy*	13.77	42.85	16.39	48.43	31.25
LTR/*Copia*	4.46	6.81	3.08	26.55	9.91
Other	0.15	0.04	0.38	0.54	0.19
Non-LTR Retrotransposon	3.20	1.07	1.16	0.84	2.68
SINE	0.26	0.08	0.05	0.03	0.11
LINE	2.94	0.98	1.11	0.80	2.57
Class II DNA Transposon	5.33	7.17	5.82	5.39	11.09
DNA Transposon Superfamily	3.32	4.73	2.75	3.37	7.52
DNA-CACTA	1.44	3.67	2.38	2.06	6.01
hAT	0.43	0.26	0.27	0.75	0.48
Harbinger	0.26	0.20	0.08	0.22	0.30
Tc1/Mariner	1.19	0.61	0.03	0.07	0.73
MITE	1.95	2.31	3.07	0.77	1.96
Tourist	0.28	1.47	1.11	0.12	0.45
Stowaway	0.14	0.09	0.60	0.00	0.05
Unclassified MITE	1.53	0.74	1.37	0.66	1.46
Helitron	0.06	0.13	0.00	0.54	0.02
Tandem repeat	1.89	2.49	2.90	0.86	1.47
Low complexity	0.27	0.19	0.82	0.12	0.12
Unclassified	8.41	5.21	0.23	0.74	10.79
Total content	37.48	65.83	30.78	82.48	65.91

Supplementary Table 9: Comparison of representatives of TE families in the whole genome shotgun sequence datasets produced by Illumina and Roche/454 technology.

TE family	Superfamily	Illumina reads [%]	454 reads [%]
Angela	*Copia*	12.79	13.15
Sabrina	*Gypsy*	7.99	6.81
Jorge	CACTA	5.59	4.93
Wilma	*Gypsy*	2.92	2.42
WHAM	*Gypsy*	2.73	2.95
Romani	CACTA	1.95	2.07
Caspar	*Gypsy*	2.83	1.47
Ifis	*Gypsy*	1.48	0.81
Pavel	CACTA	1.36	1.18
Fatima	*Gypsy*	1.36	3.06
Egug	*Gypsy*	1.21	1.03
Laura	*Gypsy*	1.17	1.6
Hawi	*Gypsy*	1.13	0.64
Derami	*Gypsy*	1.11	0.99
Latidu	*Gypsy*	0.87	1.43
Sumaya	*Gypsy*	0.85	0.27
Lila	*Gypsy*	0.84	0.72
TAT1	CACTA	0.8	0.59
Xalax	RLX	0.78	0.38
Cereba	*Gypsy*	0.71	0.77

Supplementary Table 10: Summary for the new *Ae. tauschii* genetic map.

Chromosome	# of SNPs	# of bins	Genetic distance (cM)	# of scaffolds	Length of scaffolds (Mb)	# gene
1D	28740	218	133.468	2369	199.558	2801
2D	26046	301	179.614	2323	219.142	3446
3D	18424	257	159.673	1690	157.238	2945
4D	13997	151	98.738	1384	151.963	1802
5D	25582	288	186.867	2357	209.467	3122
6D	14072	166	116.772	1410	139.325	2101
7D	24222	292	184.674	2155	200.875	2952
total	151083	1680	1059.806	13688	1277.568	19169

Supplementary Table 11: Scaffold information anchored on the constructed genetic map.
(Please see the supplementary Excel file *Scaffold information anchored on the constructed genetic map*)

Supplementary Table 12: Statistics of scaffolds and genes anchored on seven chromosomes.

Chr	Scaffolds	Length(Mb)	Genes
1D	4201	237.3	4084
2D	4849	273.4	5040
3D	5114	269.2	4865
4D	2568	180.0	2915
5D	5077	289.6	5720
6D	3530	197.1	3635
7D	4964	275.1	4438
total	30303	1.72 Gb	30697

Supplementary Table 13: Scaffolds and genes anchored on seven chromosomes.
(Please see the supplementary Excel file *Scaffolds and genes anchored on seven chromosomes*)

Supplementary Table 14: Gene Ontology analysis of those 628 genes potentially under selection. (Please see the supplementary Excel file *Gene Ontology analysis of those 628 genes potentially under selection*)

Supplementary Table 15: PFAM domains enriched in *Ae. tauschii* compared with *Brachypodium*.

PFAM domain	PFAM description	*Ae. tauschii* (Number)	*Brachypodium* (Number)	Fisher exact (p-value 2-tail)
PF00931	NB-ARC domain	738	251	4,92E-043
PF00560	Leucine Rich Repeat	652	362	2,91E-012
PF00069	Protein kinase domain	1270	853	3,04E-009
PF00067	Cytochrome P450	485	262	2,89E-010
PF03478	Protein of unknown function (DUF295)	276	60	2,95E-028
PF07762	Protein of unknown function (DUF1618)	166	59	6,38E-010
PF04578	Protein of unknown function, DUF594	168	61	9,30E-010

Supplementary Table 16: PFAM terms enriched in *Brachypodium* vs. *Ae. tauschii*.

PFAM domain	PFAM description	*Ae. tauchii* gene number	*Brachypodium* gene number	Fisher exact p-value 2-tail
PF00097	Zinc finger, C3HC4 type (RING finger)	170	297	1,06E-013
PF00847	AP2 domain	92	141	1,71855E-05
PF00076	RNA recognition motif	215	239	0,008136397
PF03101	FAR1 DNA-binding domain	20	66	1,31E-008
PF04434	SWIM zinc finger	7	43	1,08E-008
PF10551	MULE transposase domain	19	52	6,6035E-06
PF00010	Helix-loop-helix DNA-binding domain	105	123	0,023715857
PF00226	DnaJ domain	82	103	0,012068945
PF00561	alpha/beta hydrolase fold	54	78	0,003850653
PF03168	Late embryogenesis abundant protein	35	61	0,000954915

Supplementary Table 17: Summary of the 216 cold-related genes in *Ae. tauschii*. (Please see the supplementary Excel file *Ae. tauschii cold-related genes summary*). By using 178 manually collected cold acclimation related genes as query and searching gene sets of *Ae. tauschii*, we found 216 cold-related genes in the *Ae. tauschii* genome. This table described the annotation information about 216 cold-related genes.

Supplementary Table 18: Transcription factors present in sequenced grass plant genomes.

Category	*Ae. tauschii*	*B. distachyon*	*O. sativa*	*S. bicolor*	*Z. mays*
bHLH	130	139	127	166	182
NAM	120	87	104	124	116
MYB-related	103	66	70	89	95
B3	102	41	39	61	42
MYB	103	67	95	113	149
WRKY	95	73	81	93	114
C2H2	82	90	89	97	122
ERF	74	99	114	132	174
bZIP	71	85	80	93	107
GRAS	50	45	50	76	74
M-type MADS	58	23	16	43	34
G2-like	50	52	43	56	54
C3H	38	41	43	44	45
FAR1	40	61	36	128	2
HD-ZIP	32	40	31	22	48
MIKC	25	30	33	12	29
AP2	20	25	20	22	23
HB-other	19	14	12	24	16
SBP	19	18	18	18	24
LBD	17	24	31	36	42
GATA	16	26	21	29	34
ARF	17	24	24	5	31
HSF_DNA-bind	17	24	23	24	23
TALE	15	22	24	12	25
NF-YB	12	17	11	13	18
trihelix	11	20	18	20	26
Dof	11	27	28	29	42
Nin-like	12	16	8	13	17
ARR-B	11	7	7	2	8
CPP	10	9	10	8	10
NF-YC	11	14	15	15	15
GeBP	9	14	11	15	21
WOX	9	8	14	5	16
GFR	7	4	8	3	10
E2F/DP	8	7	9	10	17
NF-YA	6	7	9	9	11
TCP	5	21	28	28	42
ZF-HD	6	15	15	14	21
CAMTA	5	7	5	7	6
DBB	5	15	14	11	15

EIL	5	6	5	7	9
SRS	4	5	5	5	9
YABBY	4	8	8	8	10
LSD	3	5	7	5	7
BES1	3	5	4	8	8
CO-like	3	7	6	3	8
BBR-BPC	2	3	5	5	4
HB-PHD	2	2	1	1	3
LFY	2	1	1	1	1
NF-X1	2	2	2	3	4
VOZ	2	2	2	2	4
Whirly	2	2	2	3	2
HRT-like	1	1	1	1	1
RAV	1	4	4	2	3
S1Fa-like	1	1	2	1	1
STAT	1	1	0	0	1
NZZ/SPL	0	0	1	0	0
total	1489	1479	1490	1776	1975

Supplementary Table 19: Summary of *Ae. tauschii* transcription factor genes. (Please see the supplementary Excel file *Ae. tauschii* TF gene summary).

Supplementary Table 20: TF gene InterProScan annotation from co-expression analysis.

gene ID	InterProScan annotation
AEGTA21796	IPR001471; Pathogenesis-related transcriptional factor/ERF, DNA-binding
	IPR016177; DNA-binding, integrase-type
AEGTA05972	IPR001005; SANT, DNA-binding
	IPR009057; Homeodomain-like
	IPR012287; Homeodomain-related
	IPR014778; Myb, DNA-binding
	IPR017884; SANT, eukarya
	IPR018117; DNA methylase, C-5 cytosine-specific, active site
AEGTA20129	IPR003340; Transcriptional factor B3
	IPR010525; Auxin response factor
AEGTA32267	IPR004827; Basic-leucine zipper (bZIP) transcription factor
	IPR011616; bZIP transcription factor, bZIP-1
AEGTA02173	IPR001005; SANT, DNA-binding
	IPR009057; Homeodomain-like
	IPR012287; Homeodomain-related
	IPR014778; Myb, DNA-binding
	IPR015495; Myb transcription factor
	IPR017930; Myb-type HTH DNA-binding domain
AEGTA27104	IPR001471; Pathogenesis-related transcriptional factor/ERF, DNA-binding
	IPR002194; Chaperonin TCP-1, conserved site
	IPR002423; Chaperonin Cpn60/TCP-1
	IPR012720; T-complex protein 1, eta subunit
	IPR016177; DNA-binding, integrase-type
AEGTA21259	IPR001005; SANT, DNA-binding
	IPR009057; Homeodomain-like
	IPR014778; Myb, DNA-binding
	IPR017884; SANT, eukarya
AEGTA30301	IPR001356; Homeobox
	IPR009057; Homeodomain-like
	IPR012287; Homeodomain-related
AEGTA07407	IPR001005; SANT, DNA-binding
	IPR009057; Homeodomain-like
	IPR012287; Homeodomain-related
	IPR014778; Myb, DNA-binding
	IPR015495; Myb transcription factor
	IPR017930; Myb-type HTH DNA-binding domain
AEGTA32664*	IPR001005; SANT, DNA-binding
	IPR009057; Homeodomain-like
	IPR012287; Homeodomain-related
	IPR014778; Myb, DNA-binding
	IPR015495; Myb transcription factor

	IPR017930; Myb-type HTH DNA-binding domain
AEGTA31644	IPR000571; Zinc finger, CCCH-type
	IPR001269; tRNA-dihydrouridine synthase
	IPR013785; Aldolase-type TIM barrel
	IPR018517; tRNA-dihydrouridine synthase, conserved site
AEGTA02201	IPR003347; Transcription factor jumonji/aspartyl beta-hydroxylase
	IPR007087; Zinc finger, C2H2-type
	IPR013087; Zinc finger, C2H2-type/integrase, DNA-binding
	IPR013129; Transcription factor jumonji
	IPR015880; Zinc finger, C2H2-like
AEGTA15908	IPR003316; Transcription factor E2F/dimerisation partner (TDP)
	IPR011991; Winged helix repressor DNA-binding
	IPR015633; E2F Family
AEGTA31223	IPR000571; Zinc finger, CCCH-type

*The gene is an ortholog of wheat drought tolerance gene[60].

Supplementary Table 21: Numbers of conserved miRNA in *Ae. tauschii* (DD), *T. aestivum* (TAE), *H. vulgare* (HVU), *O. sativa* (OSA), *B. distachyon* (BDI), *S. bicolor* (SBI) and *Z. mays* (ZMA).

microRNA	DD	TAE	HVU	OSA	BDI	SBI	ZMA
MIR156	9	1	1	12	4	9	12
MIR159	2	2	2	6	1	2	11
MIR160	4	1	-	6	5	6	7
MIR164	4	1	-	6	6	5	8
MIR166	6	-	3	14	6	11	14
MIR167	6	2	-	10	4	9	10
MIR168	1	-	1	2	1	1	2
MIR169	20	-	1	17	11	17	18
MIR171	7	2	1	9	4	11	14
MIR172	3	-	-	4	3	6	5
MIR319	1	1	-	2	2	2	4
MIR390	1	-	-	1	2	1	2
MIR393	1	-	-	2	2	2	3
MIR394	1	-	-	1	1	2	2
MIR395	25	2	-	25	14	12	16
MIR396	6	-	-	9	6	5	8
MIR398	1	1	-	2	3	1	2
MIR399	20	1	1	11	2	11	10
MIR408	1	1	-	1	1	1	2
MIR528	1	-	-	1	1	1	2
MIR530	1	-	-	1	-	-	-
MIR2118	42	-	-	18	-	-	7
MIR2275	8	-	-	2	-	-	4
MIR1432	2	-	-	1	-	1	1
MIR1436	10	-	1	1	-	-	-
MIR827	1	-	-	3	1	-	1
MIR1878	1	-	-	1	1	-	-
MIR1120	58	1	1	-	-	-	-
MIR1127	4	1	-	-	1	-	-
MIR1128	2	2	-	-	2	-	-
MIR1130	4	1	-	-	-	-	-
MIR1132	4	1	-	-	1	-	-
MIR1135	5	1	1	-	1	-	-
MIR2002	5	1	-	-	-	-	-
MIR2006	2	1	1	-	-	-	-
MIR2009	2	3	5	-	-	-	-
MIR2010	1	1	-	-	-	-	-
MIR2012	1	1	2	-	-	-	-
MIR2015	1	1	-	-	-	-	-

MIR2016	2	1	1	-	-	-	-
MIR2018	7	1	2	-	-	-	-
MIR2020	2	1	2	-	-	-	-
MIR2026	1	1	-	-	-	-	-
MIR2027	2	1	-	-	-	-	-
MIR2031	67	1	-	-	-	-	-
MIR2032	3	1	1	-	-	-	-
MIR5048	1	-	1	-	-	-	-
MIR5050	1	-	1	-	-	-	-
MIR5057	22	-	1	-	1	-	-
MIR5062	3	-	2	-	2	-	-
MIR5064	5	-	1	-	1	-	-
MIR5067	36	-	1	-	1	-	-
MIR5070	5	-	1	-	1	-	-
MIR5071	1	-	1	1	-	-	-
MIR5084	5	1	1	-	-	-	-
MIR5169	4	-	-	-	1	-	-
MIR5175	3	-	-	-	2	-	-
MIR5176	1	-	-	-	1	-	-
MIR5181	46	-	1	-	2	-	-
MIR5200	3	-	-	-	1	-	-
MIR5203	30	-	-	-	1	-	-
MIR2024	1	2	-	-	-	-	-
MIR2025	2	1	-	-	-	-	-
MIR3711	1	-	-	-	-	-	-

Supplementary Table 22: The details of 33 QTLs/genes mapped in chromosome 2D.

QTL*	Description	Original source map	Reference
QGne.nfcri-2D.1	Grain number per ear	Heshangmai*Yu8679	74
QGfrmax.nfcri-2D	Maximum grain filling rate	Heshangmai*Yu8679	74
QGt.orst-EF00	Free threshing habit	Syn x Opata Jantasuriyarat	75
Lr39	Leaf rust resistance	TA4186*TA1675	76
Rht8	Dwarfing gene	Cappelle*Cappelle(Mara)	77
QYld.ksu-2D	Grain yield	Syn*Opa SO	78
QFhs.pur-2D	Fusarium head blight resistance	Ning894037 x Alondra	79
QDta.umc-2D	Days to anthesis	Ernie*MO 94-317	80
Ppd-D1	Photoperiod response	Cappelle*Cappelle(Mara)	77
QSnb.fcu-2D	Resistance to SNB caused by isolate Sn6	BR34*Grandin	81
QGt.orst-2D.1	Glume tenacity	Cs*Cs(2D)	82
Snn2	Resistance to SNB	BR34*Grandin	83
Tg1	Tenacious glumes gene	Cs*Cs(2D)	82
QGba.orst-2D	Size (area) of detached glume base scars	Cs*Cs(2D)	82
CID-2D	Carbon isotope discrimination	Cranbrook*Halberd 07	84
Qnos.umc-2D	Number of spikeltes on inoculated head	Ernie*MO 94-317	80
QGfc.aww-2D.1	Grain fructan concentration (% of dry weight)	Berkut*Krichauff	85
QHt.crc-2D	Plant height	RL4452*AC Domain SO 05/08	86
bh-D1	Multi row spike recessive allele (alias mrs1)	Ruc163-1-02 x So149-1-02 2DS	
Acph-D2	Electrophoretically 'fast' acid phosphatase involved in intraspecies variation	*Ae. tauschii* E1**Ae. tauschii* S1 2D	87
QCr.W2Me-2D	Seedling resistance to crown rot	W21MMT70*Mendos	88
QGpc.ccsu-2D.1	Grain protein content	WL711 x PH132 Gupta	89
QGpc.ccsu-2D.2	Grain protein content	WL711 x PH132 Gupta	89
CL-2D	Coleoptile length	Cranbrook*Halberd 07	90
QGne.nfcri-2D.2	Grain number per ear	Heshangmai*Yu8679	74
QGPht.nfcri-2D	Plant height	Heshangmai*Yu8679	74
QTgw.nfcri-2D	Thousand grain weight	Heshangmai*Yu8679	74
QGnu.ipk-2D	Grain number	Syn x Opata Roder 031003	91
Pm43	Dominant powdery mildew resistance gene	CH5025 x CH5065 2DL	92
QGwe.ipk-2D.4	Grain weight/colour	Syn x Opata Roder 031003	91
Ppo-D1	Enzyme activity of polyphenol	Zhongyou9507*CA9632	93,94

	oxidase		
QHt.ipk-2D	Plant height	Syn x Opata Roder 031003	91
QGwe.ipk-2D.1	Grain weight/colour	Syn x Opata Roder 031003	91
QEet.ipk-2D	Ear emergence time	Syn x Opata Roder 031003	91
QGwe.ipk-2D.3	Grain weight/colour	Syn x Opata Roder 031003	91
QRg.ipk-2D	Glume colour	Syn x Opata Roder 031003	91
QGyld.agt-2D	Grain yield	Trident*Molineux	95
TGWM	Thousand-grain weight at the grain-filling stage	Hanxuan10*Lumai14-2D	96
SWSCF	Stem water-soluble carbohydrates at the flowering stage	Hanxuan10*Lumai14	96
Q.Sng.pur-2DL.2	Resistance to SNB	P91193D1*P92201D5-2D	
QTwt.crc-2D	Test weight	RL4452*AC Domain SO 05/08	86

*QTLs in light blue also occurred in Figure 3 in main text.

Supplementary Table 23: Homologous genes associated with important agronomical traits in *Ae. tauschii* genome.

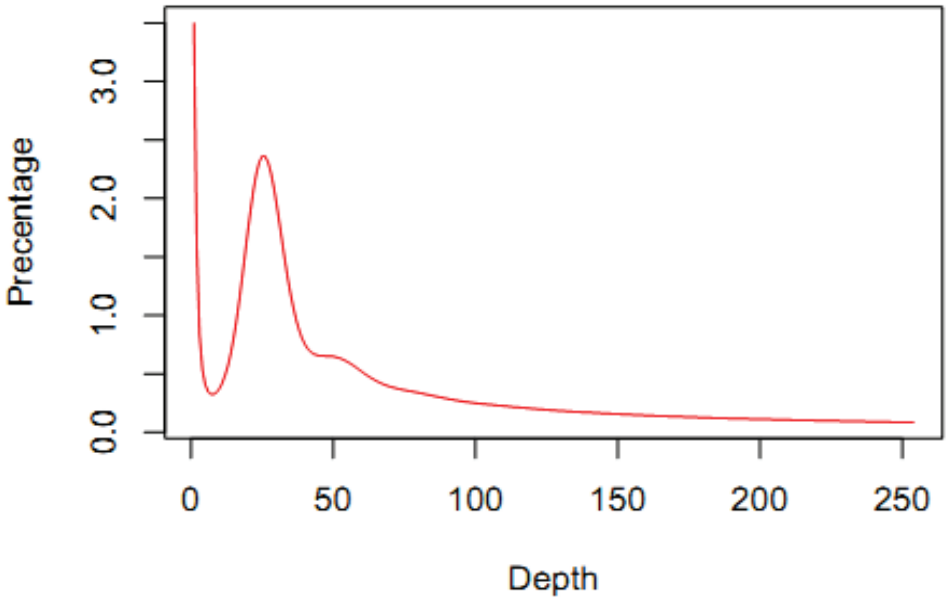
Gene name	Accession No.	Species	Query size (aa)	Scaffold ID	Length (bp)	E-value	Reference
Lr1	ABS29034	*T. aestivum*	844	scaffold68617	79977	e-135	[97]
Lr10	AAQ01784	*T. aestivum*	921	scaffold2223	78251	0	[98]
Lr21	ACO53397	*T. aestivum*	1080	scaffold25563	154615	0	[99]
Lr34	ADK62371	*T. aestivum*	1402	scaffold47033	133693	e-165	[100]
Pm3	ADH04488	*T. aestivum*	848	scaffold40896	137383	2e-097	[101]
Pm21 Stpk-A	AEF30548	*T. aestivum*	401	scaffold48600	47723	e-109	[102]
Pm21 Stpk-B	AEF30550	*T. aestivum*	401	scaffold48600	47723	e-113	[102]
Pm21 Stpk-D	AEF30549	*T. aestivum*	401	scaffold48600	47723	e-112	[102]
Pm21 Stpk-V	AEF30547	*Dasypyrum*	401	scaffold48600	47723	e-113	[102]
Ppd-D1	AEJ88051	*Ae. tauschii*	59	scaffold38896	148429	1e-020	[103]
Vrn-1	AAZ76883	*T. monococcom*	244	scaffold16679	103710	4e-028	[104]
Vrn2 ZCCT1	AAS60238	*T. monococcom*	213	scaffold12030	260836	1e-054	[105]
Vrn2 ZCCT2	AAS60252	*T. turgidum*	212	scaffold12030	260836	4e-038	[105]
Vrn3	ABK32208	*T. aestivum*	177	scaffold40898	134954	9e-056	[106]
Rht-1	Q9ST59	*T. aestivum*	623	scaffold6108	117751	e-100	[107]
Q	AAU94926	*T. aestivum*	447	scaffold55689	29428	3e-040	[108]
Gsp-B1	AEE25802	*T. aestivum*	164	scaffold25552	215673	3e-084	Direct submission
Pina	BAD22739	*T. aestivum*	148	scaffold25552	215673	4e-081	[109]
Pinb	AAT40245	*T. aestivum*	148	scaffold25552	215673	8e-082	Direct submission
DEP1	ACI25444	*T. aestivum*	295	scaffold37072	52557	5e-022	[110]
Gpc-B1	ABY67950	*H. vulgare*	406	scaffold14626	114998	e-138	[111]
Nud	BAG12386	*H. vulgare*	227	scaffold2283	100235	2e-074	[112]
Vrs1	BAK09316	*H. vulgare*	222	scaffold25608	136944	6e-082	Direct submission
GW2	ABO31101	*O. sativa*	425	scaffold64601	45836	7e-086	[113]
EP2	ACZ62640	*O. sativa*	1356	scaffold101819	58786	0	[114]

GS5	AEO37083	*O. sativa*	482	scaffold73832	133753	2e-090	[115]
IPA1	ADJ19220	*O. sativa*	417	scaffold29236	167324	8e-070	[116]
qSH-1	BAI78225	*O. sativa*	612	scaffold50665	18290	3e-058	Direct submission
sh4	ADK25385	*O. sativa*	390	scaffold11245	197276	4e-032	[117]
MOC1	AAP13049	*O. sativa*	441	scaffold11436	126951	1e-056	[118]
tga1	AAX83763	*Z. mays*	229	scaffold22570	132832	5e-057	[119]
tb1	ACI43570	*Z. mays*	335	scaffold1179	194208	2e-035	[120]

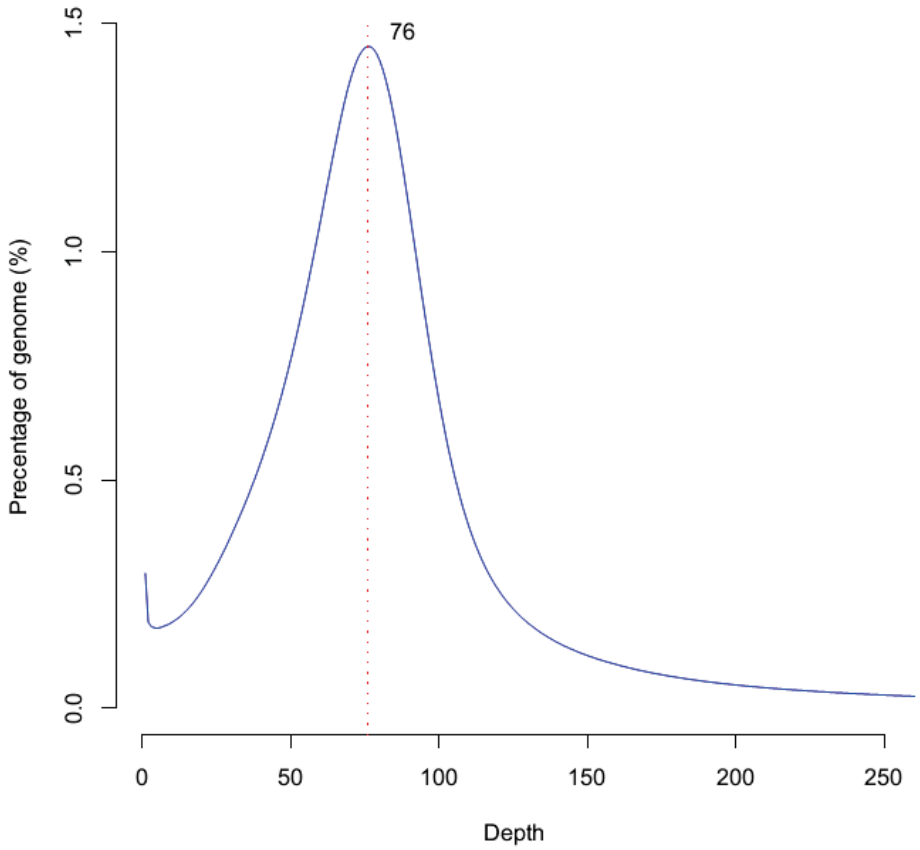
Supplementary Table 24: Summary of simple sequence repeats (SSR) types and numbers in the *Ae. tauschii* genome.

Type	Class	Total loci	Total repeats	Total length(bp)	Average length(bp)
Monomers	I	7635	191345	191345	25.1
Monomers	II	41954	589957	589957	14.1
Monomers	**total**	**49589**	**781302**	**781302**	**15.8**
Dimers	I	34185	481376	962752	28.2
Dimers	II	129782	902700	1805400	13.9
Dimers	**total**	**163967**	**1384076**	**2768152**	**16.9**
Trimers	I	24566	260808	782424	31.8
Trimers	II	299963	1279070	3837210	12.8
Trimers	**total**	**324529**	**1539878**	**4619634**	**14.2**
Tetramers	I	5908	33884	135536	22.9
Tetramers	II	230908	710677	2842708	12.3
Tetramers	**total**	**236816**	**744561**	**2978244**	**12.6**
Pentamers	I	4840	20235	101175	20.9
Pentamers	II	46895	140685	703425	15.0
Pentamers	**total**	**51735**	**160920**	**804600**	**15.6**
Hexamaers	I	3535	16323	97938	27.7
Hexamaers	II	29955	89865	539190	18.0
Hexamaers	**total**	**33490**	**106188**	**637128**	**19.0**
	I	80669	9.4%	2271170	28.2
Total	II	779457	90.6%	10317890	13.2
	total	**860126**	1	**12589060**	**14.6**

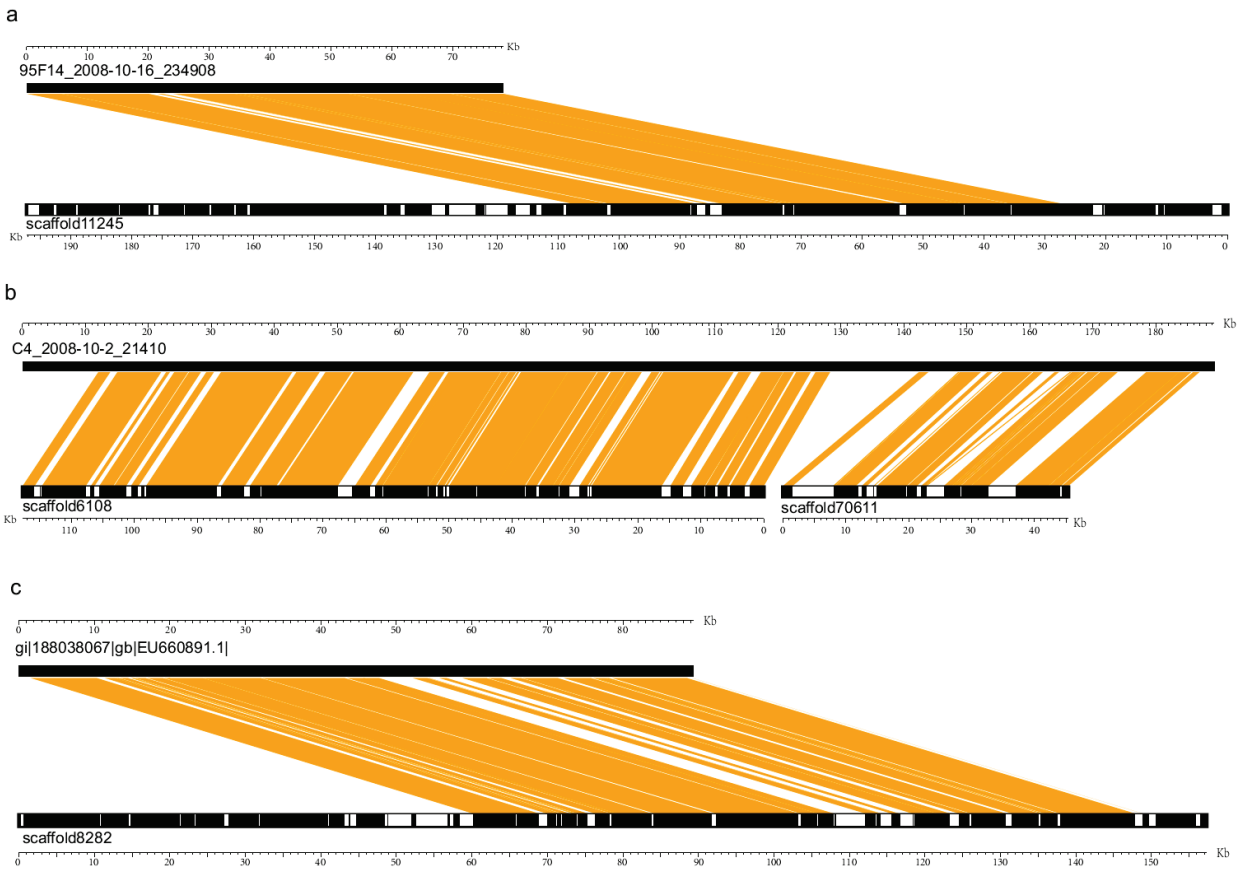
Supplementary Figures



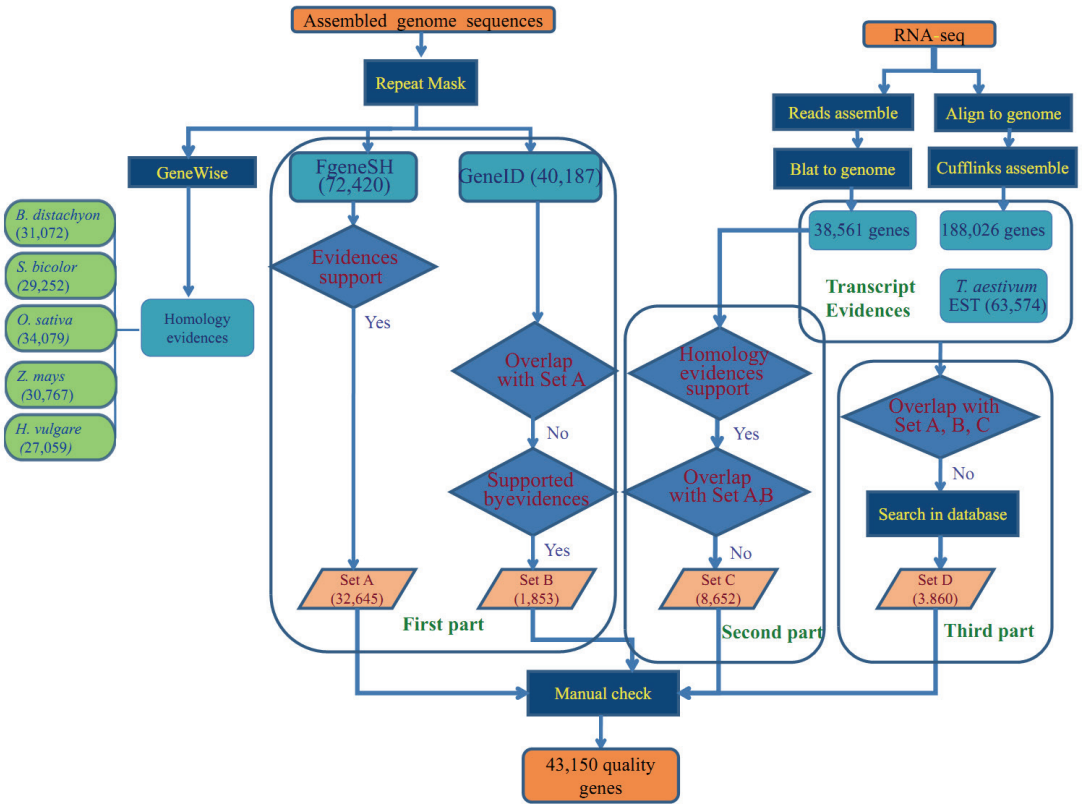
Supplementary Figure 1: The distribution of 17-mer depth of the high quality reads. The X-axis represents the sequencing depth and the Y-axis represents the proportion of a *K*-mer at a given sequencing depth.



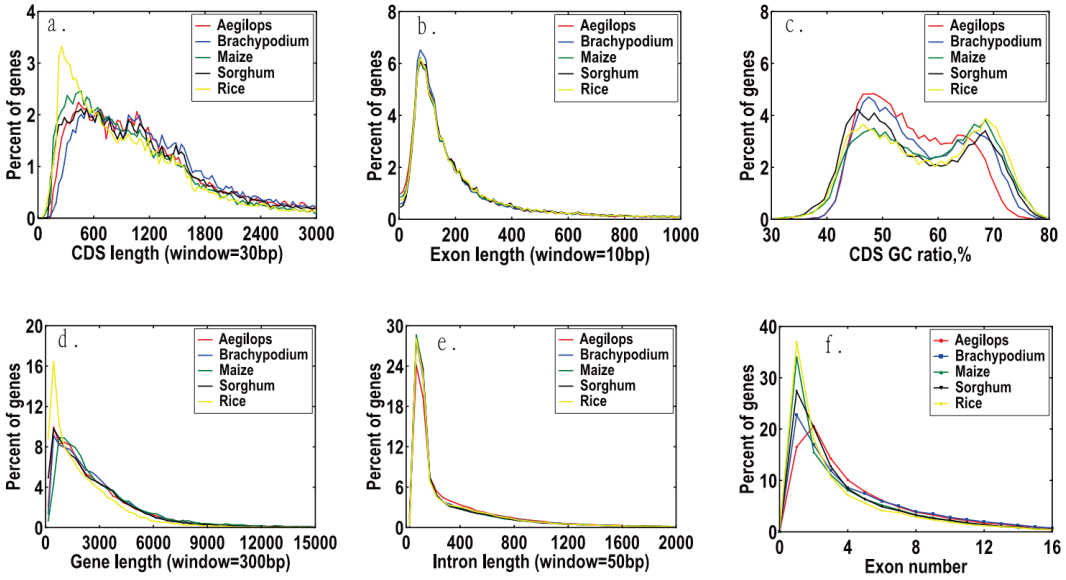
Supplementary Figure 2: The distribution of sequencing depth across the assembled genome. The Y-axis represents the proportion of the genome at a given sequencing depth.



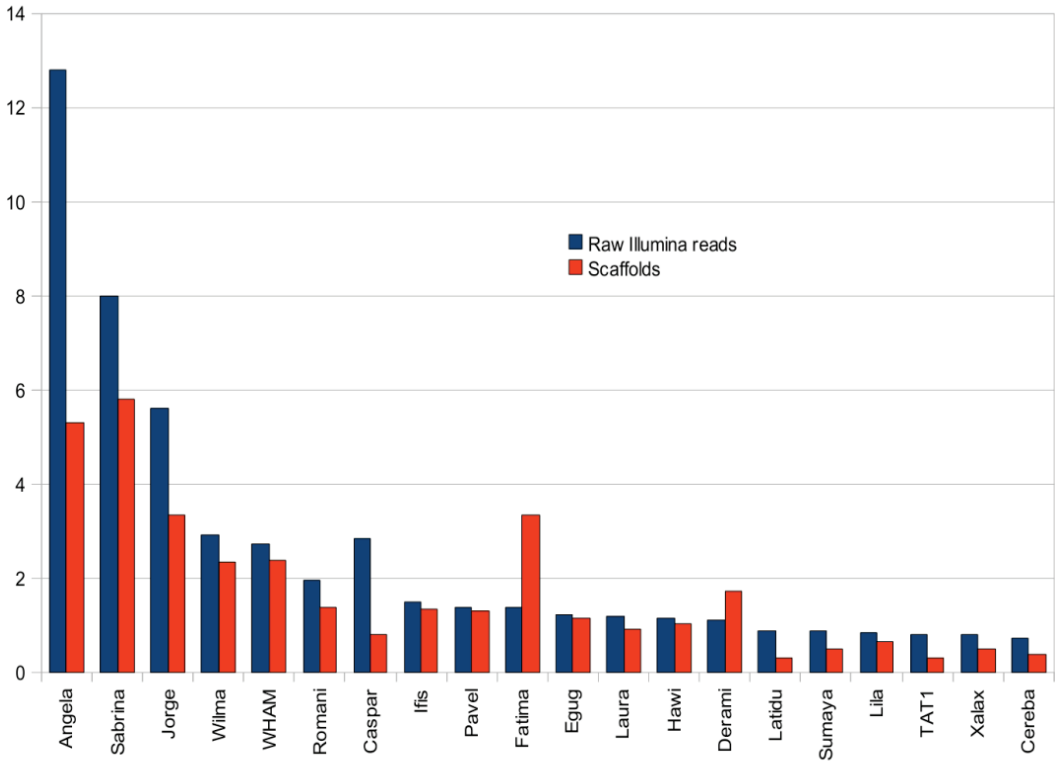
Supplementary Figure 3: Examples of three BACs well matched by the scaffolds. Note that in each figure, each BAC end is denoted on the top with the BAC name on the left, while each scaffold is denoted on the bottom with the scaffold name on the left. Those regions with 'N' in the scaffold are noted with blanks.



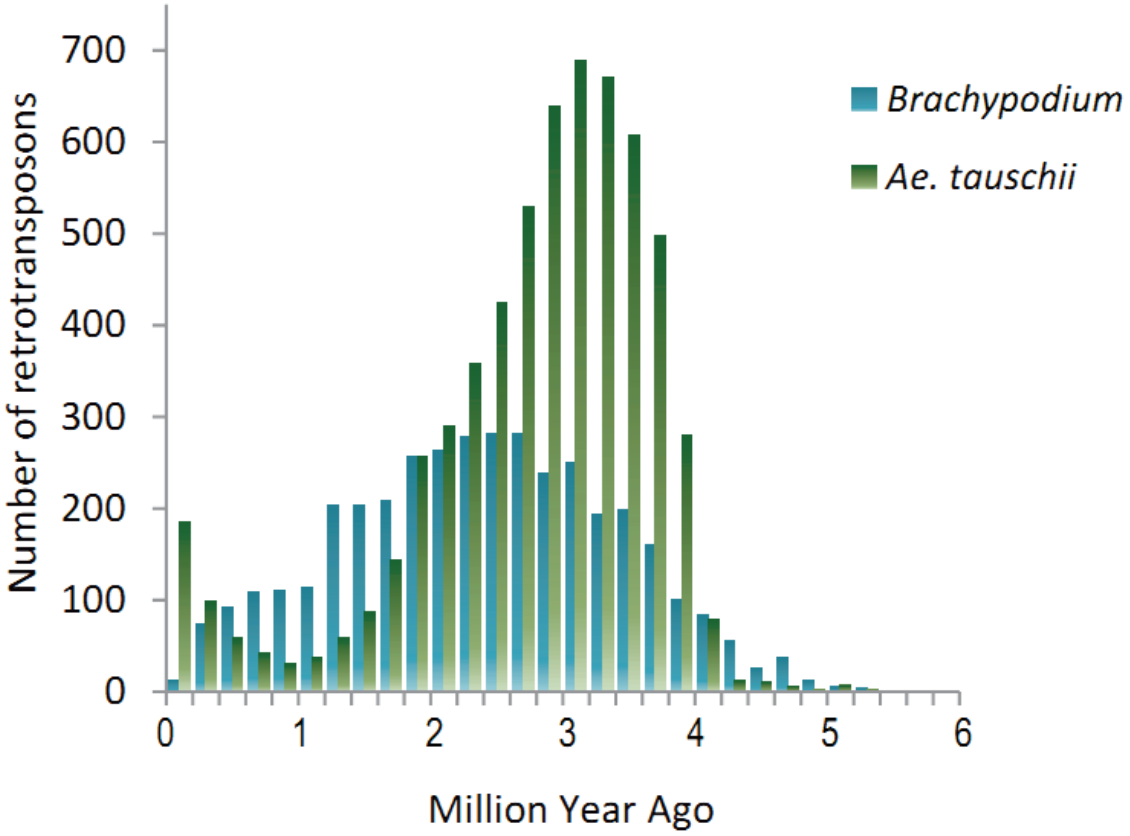
Supplementary Figure 4: Workflow used in our gene predictions combining *de novo* and evidence-based approaches.



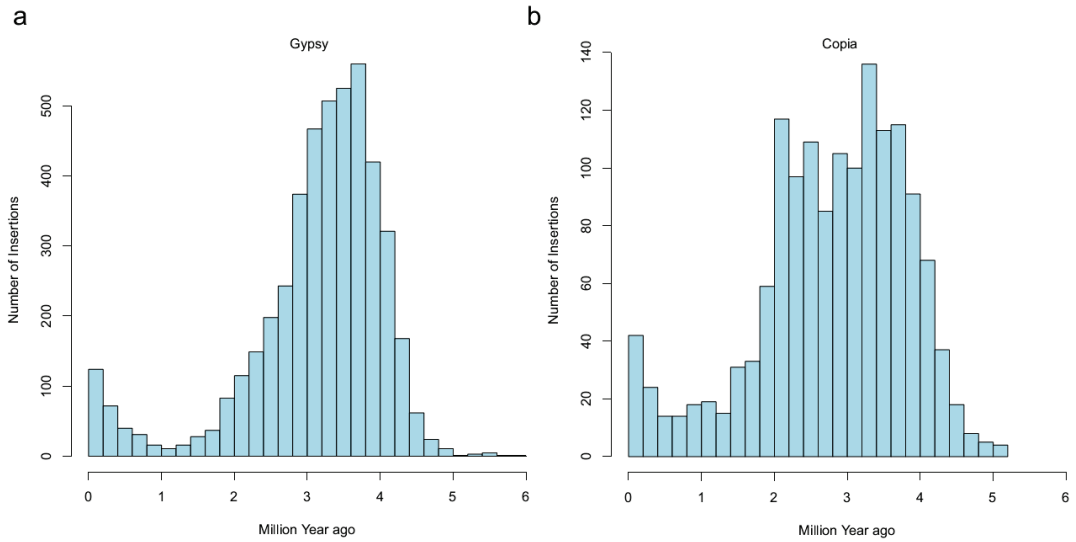
Supplementary Figure 5: Distribution comparison of (a) CDS length, (b) exon length, (c) CDS GC ratio, (d) gene length, (e) intron length and (f) exon number of *Ae. tauschii* to the four sequenced monocots.



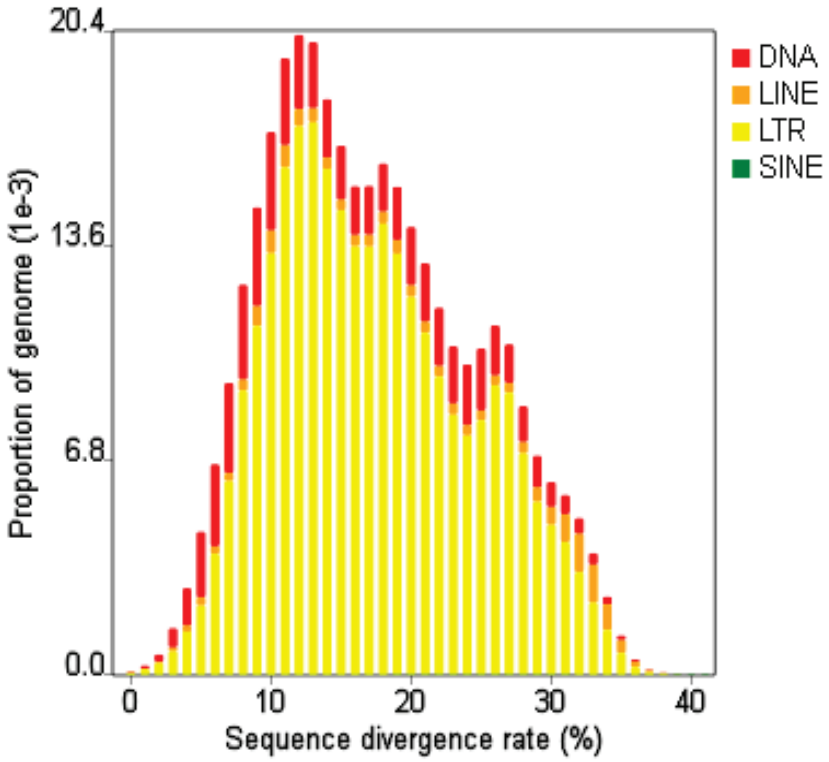
Supplementary Figure 6: Representation of the 20 most abundant TE families in single Illumina reads (blue series) was compared with their contribution to assembled scaffolds (red series). Most TE families belong to the *Gypsy* superfamily of LTR-retrotransposons. Angela belongs to the *Copia* superfamily while Jorge, Caspar, and TAT1 belong to the CACTA superfamily.



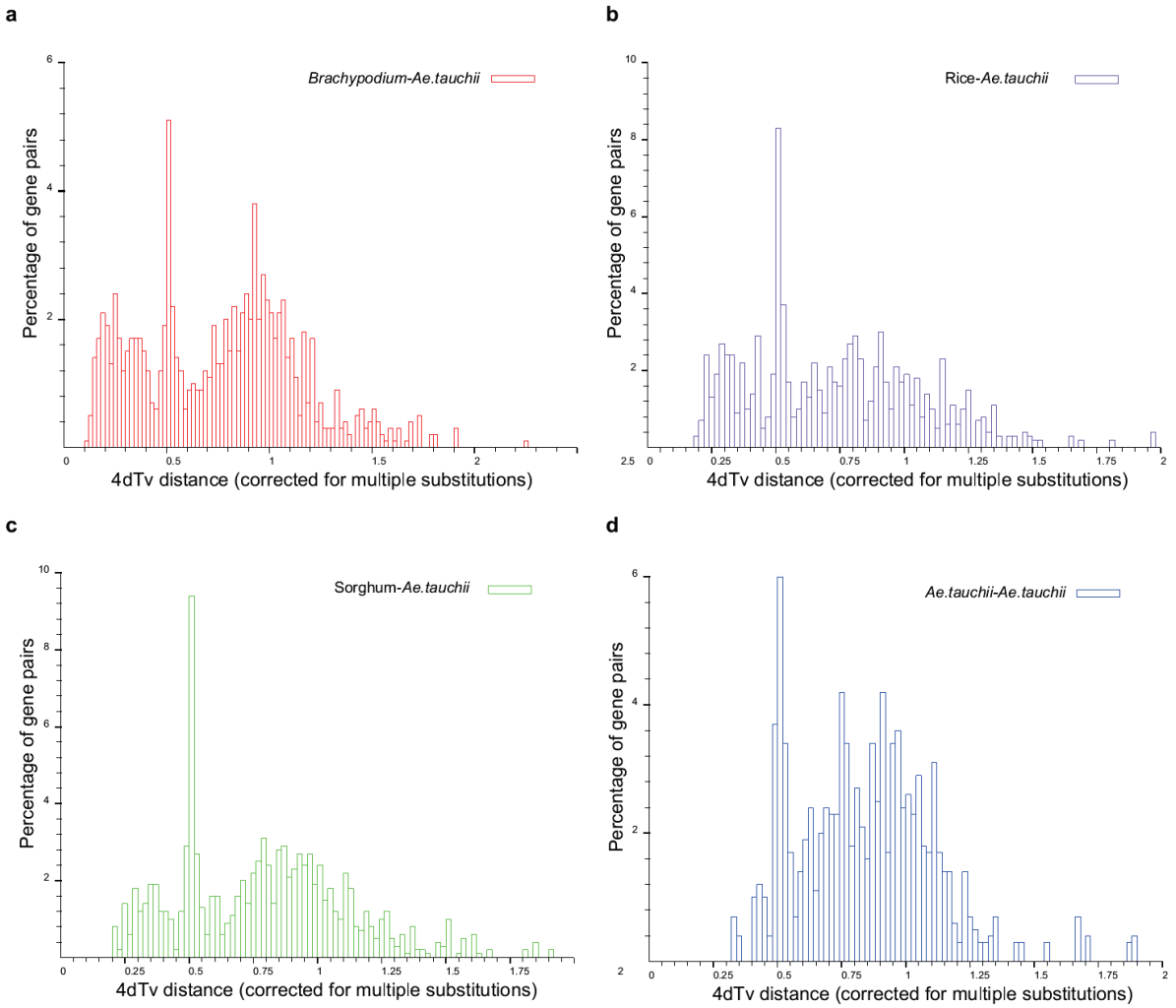
Supplementary Figure 7: Dating the LTR retrotransposon insertion time. All LTR retrotransposon families with 10 or more copies were considered. Dating of *Brachypodium* LTR retrotransposons was used as a comparison.



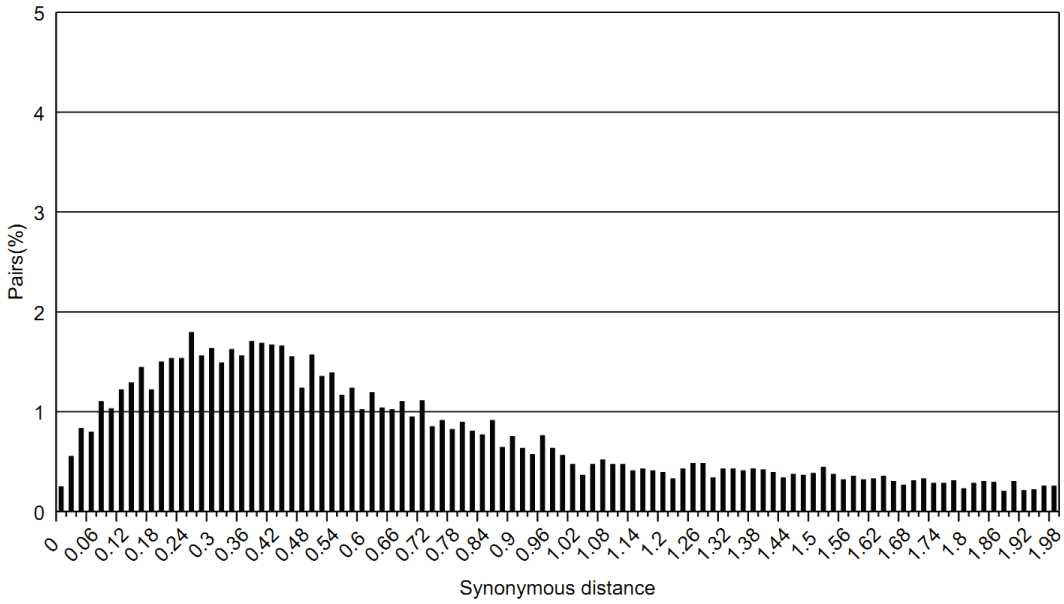
Supplementary Figure 8: Dating LTR retrotransposon insertion of *Gypsy* and *Copia* during *Ae. tauschii* genome evolution.



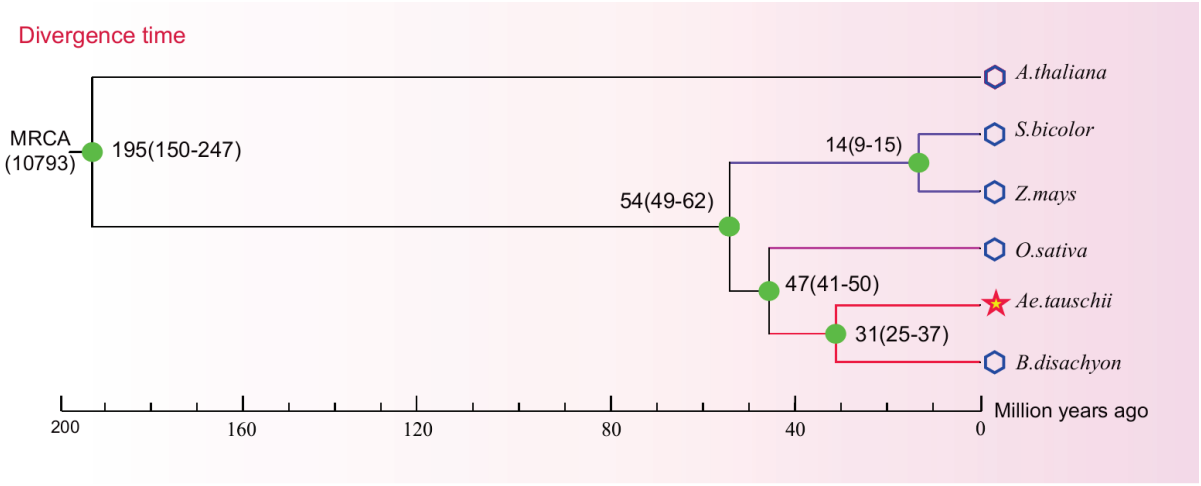
Supplementary Figure 9: Divergence distribution of classified TE families in *Ae. tauschii* genome. To analyze divergence, different TE families were aligned onto the Repbase library. DNA: DNA elements; LINE: long interspersed nuclear elements; LTR: long terminal repeat transposable element; SINE: short interspersed nuclear elements.



Supplementary Figure 10: The 4DTv distribution of duplicate gene pairs in *Ae. tauschii* genome, *Brachypodium*, rice and sorghum.



Supplementary Figure 11: Homologous relationships in *Ae. tauschii* genome. The bottom histogram plot shows pairwise Ks values for gene family sizes ≥ 7. The peak at ~0.36 indicates an ancient duplication in *Ae. tauschii* genome.

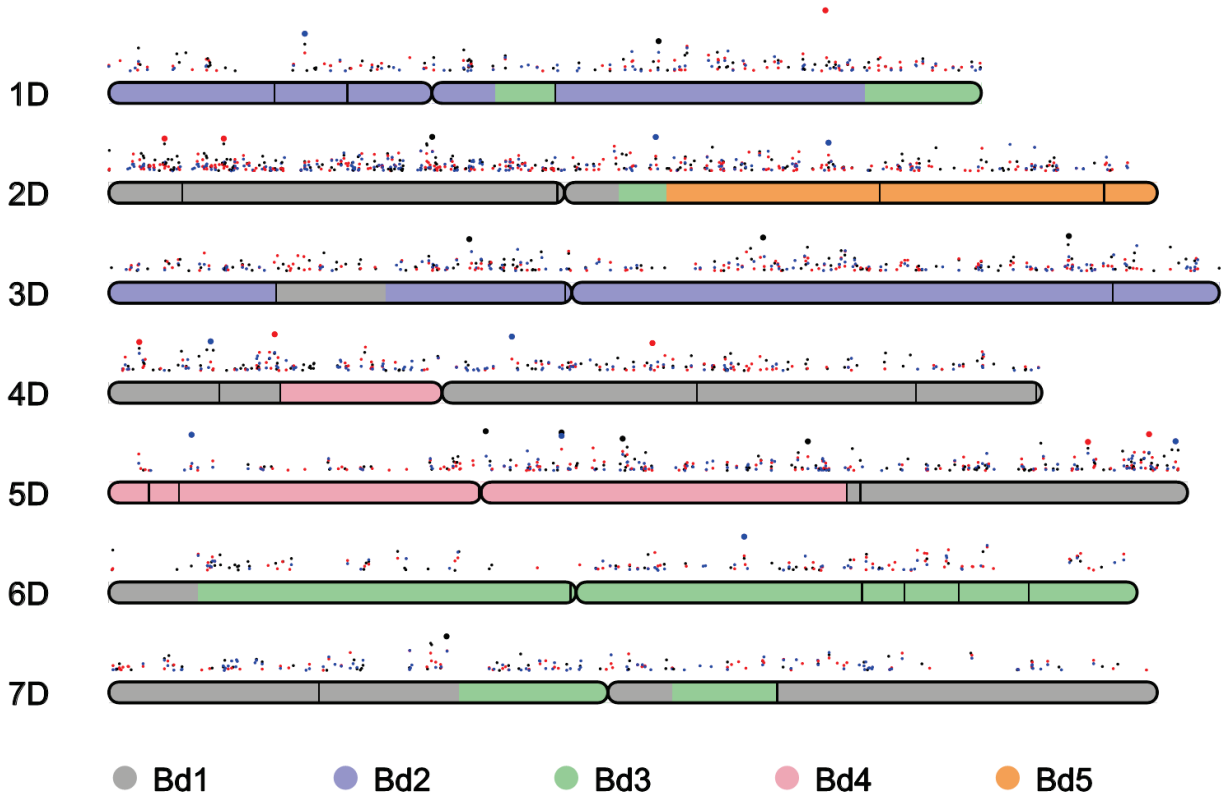


Supplementary Figure 12: Phylogenetic relationship of *Ae. tauschii* and four sequence monocots, with *A. thaliana* as outgroup. The time of divergence was estimated based on orthologs.

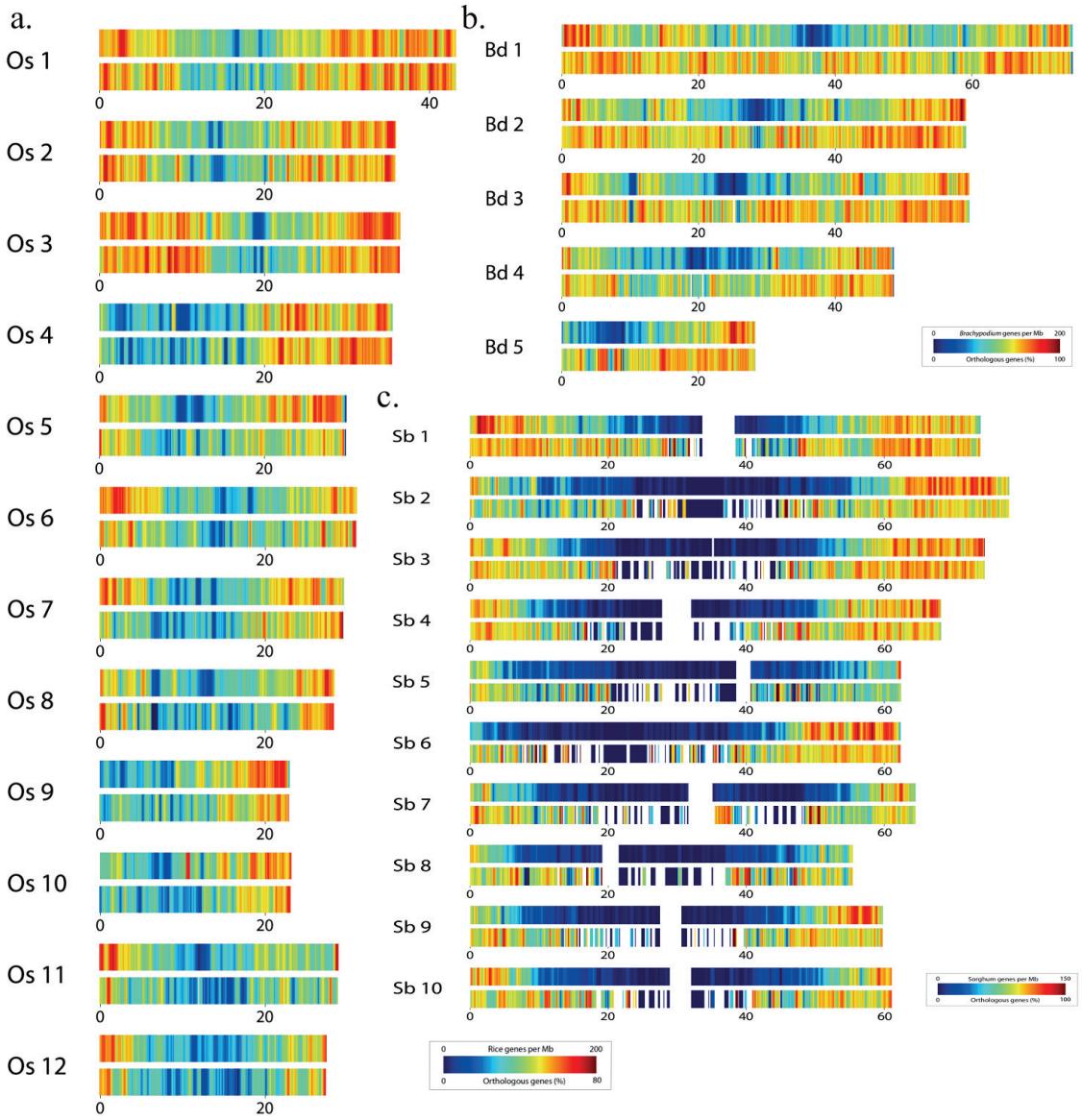
Supplementary Figure 13: The genetic map constructed using restriction site associated DNA (RAD) tag sequencing technology.
(Please see the supplementary figure *the genetic map constructed using RAD tag sequencing technology*).

Supplementary Figure 14: Anchoring of *Ae. tauschii* scaffolds to three SSR-based genetic maps. A selection of 36 genes/QTLs is denoted on the right of chromosomes in blue ovals. The markers were generally in a consistent order between maps although some cross-overs in location are evident and reflect either the distribution of repetitive regions in the genome or small inversions. The markers in red are within sets of markers that are shared between maps.
(Please see the supplementary figure *Anchoring of Ae. tauschii scaffolds to three SSR-based genetic maps*).

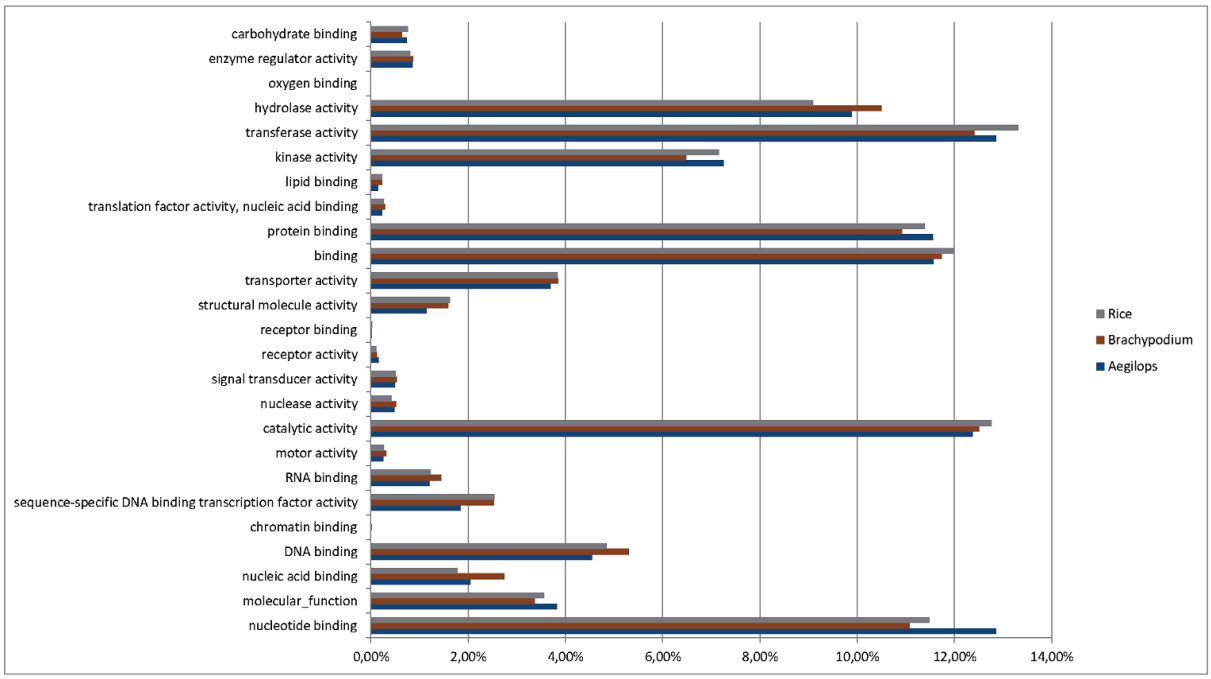
Supplementary Figure 15: CMap based comparisons of three SNP-based maps where sequenced based genetic markers could be anchored to *Ae. tauschii* genome scaffolds. The figure illustrates map alignments where scaffolds are shared between maps using the *Ae. tauschii*, *Synthetic x Opata*, and *Avalon x Cadenza* maps except for chromosome 4D where the *Westonia x Kauz* map is shown instead of *Avalon x Cadenza* due to a paucity of shared scaffolds across the maps of 4D. The shared scaffolds were generally in a consistent order between maps although some cross-overs for the location of scaffolds are evident and reflect either the distribution of repetitive regions in the genome or small inversions.
(Please see the supplementary figure *CMap based comparisons of three SNP-based maps*).



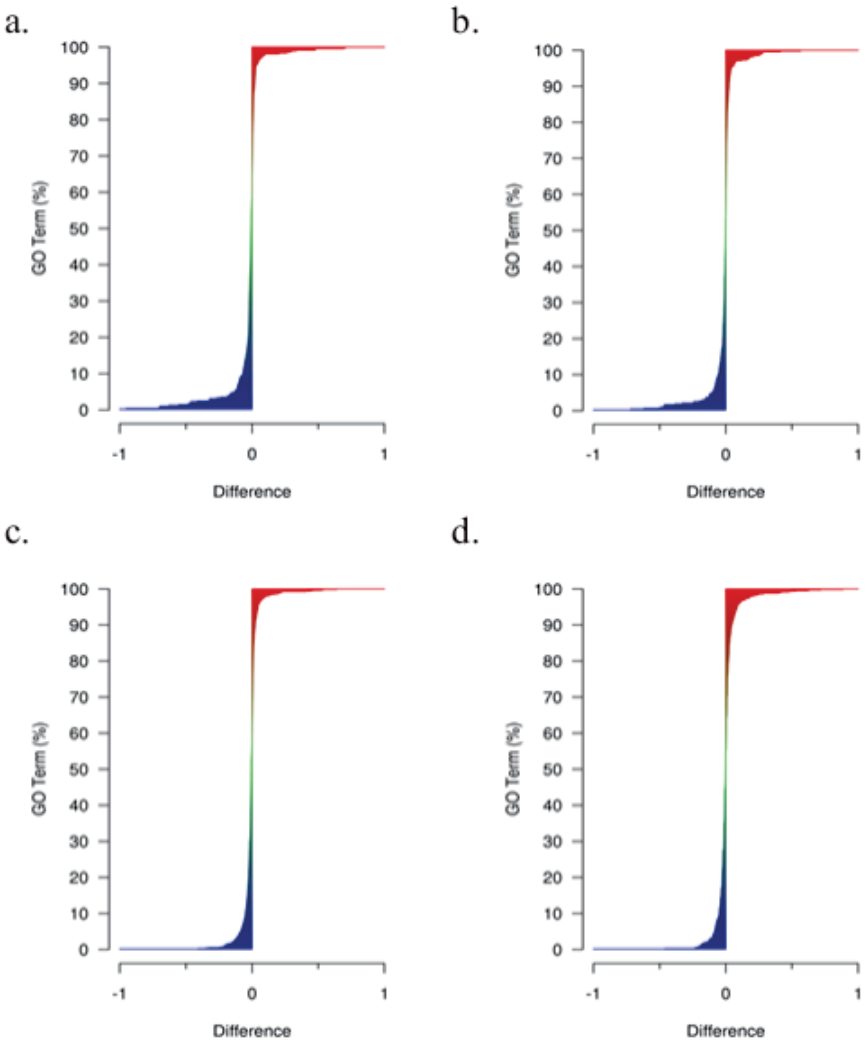
Supplementary Figure 16: Orthologous gene relationships between *Brachypodium* and *Ae. tauschii*. 4,531 orthologous relationships were identified using genetically mapped ESTs and SNP markers. The points represent the Ka/Ks ratios >0.3 between *Ae. tauschii* genes and their orthologs in *Brachypodium* (Bd), rice (Os), and sorghum (Sb). Large points depict putative rapid evolution genes with the Ka/Ks ratios >0.8.



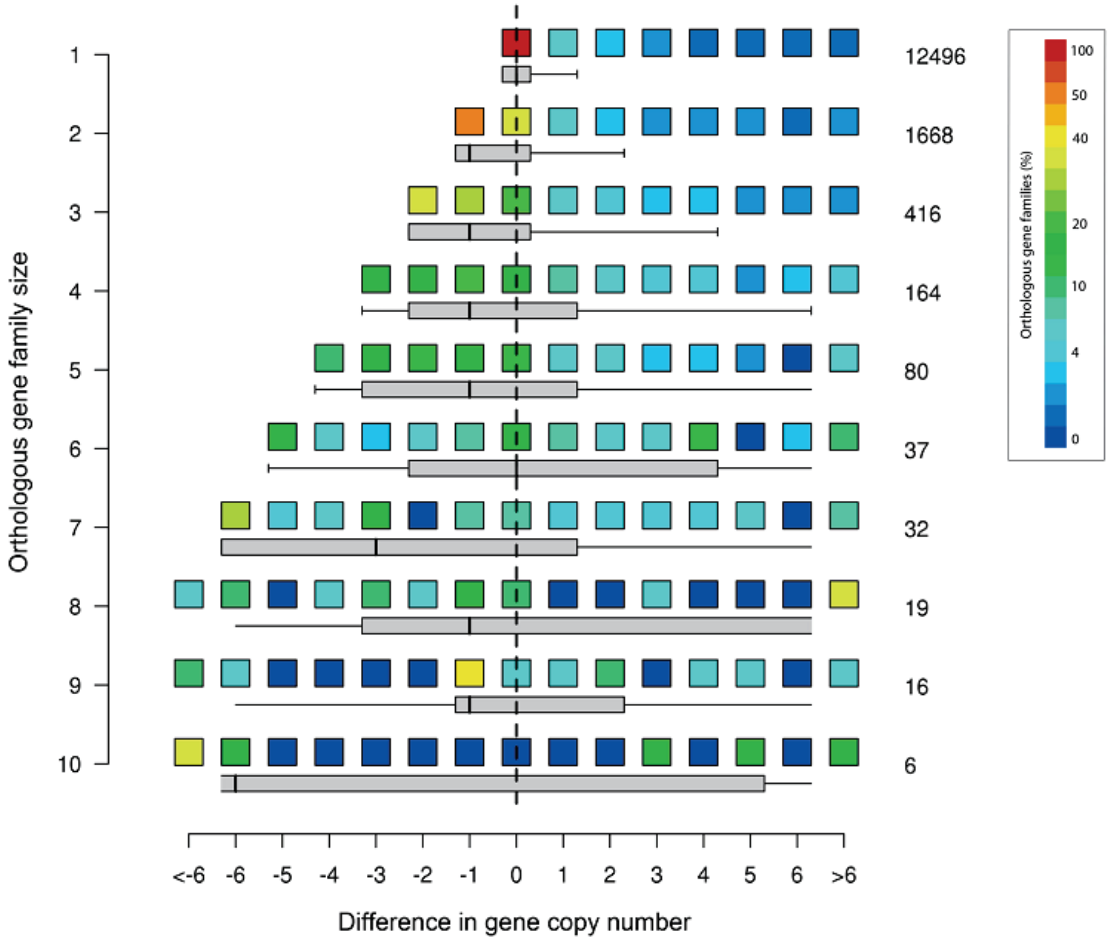
Supplementary Figure 17: *In-silico* "staining" of *Ae. tauschii* gene models against *Brachypodium*, rice and sorghum. Using a sliding window approach total gene density (upper track) and the relative distribution orthologous genes (lower track) were calculated for rice (a) *Brachypodium* (b) and sorghum (c). The heatmap scale is given in Mb and the coloring in the upper tracks in each panel shows the number of matched genes in per Mb as described in Mayer *et al* (2011). The lower tracks in each panel show the percentage of orthologous genes relative to the absolute number of clustered genes in a window.



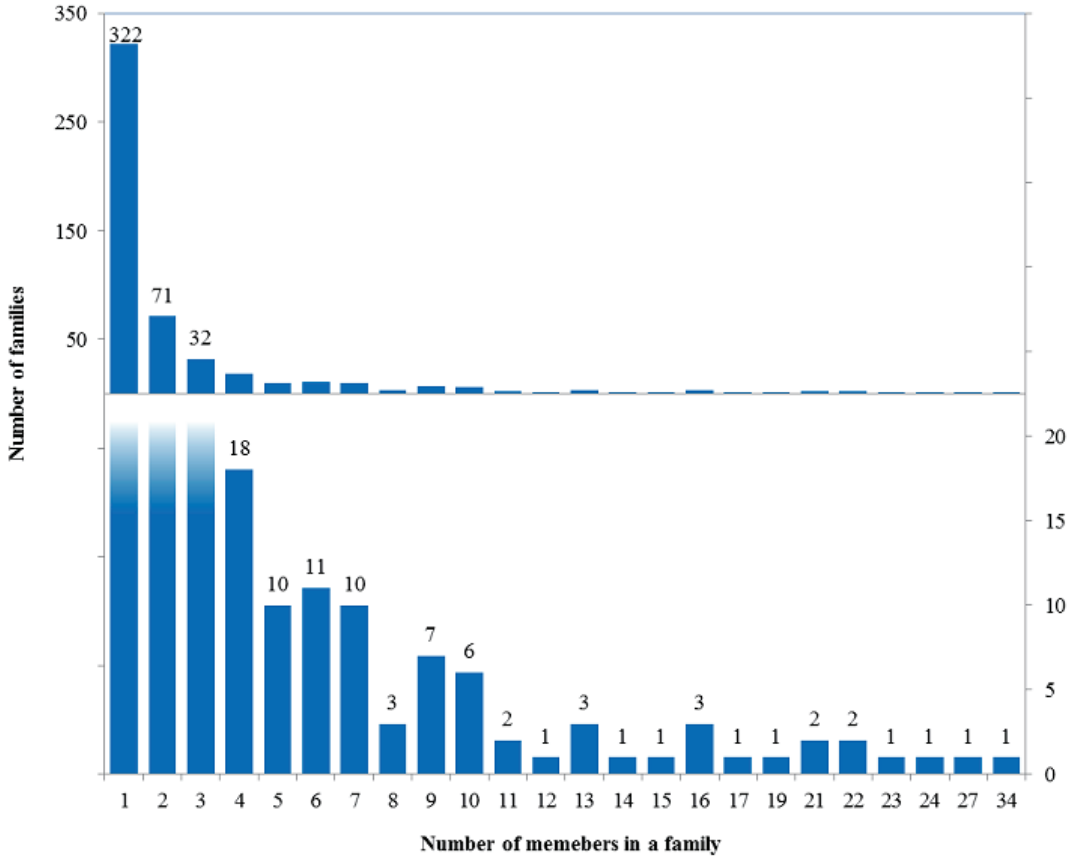
Supplementary Figure 18: Distribution of GO Slim molecular function categories for *Ae. tauschii*, *Brachypodium* and rice.



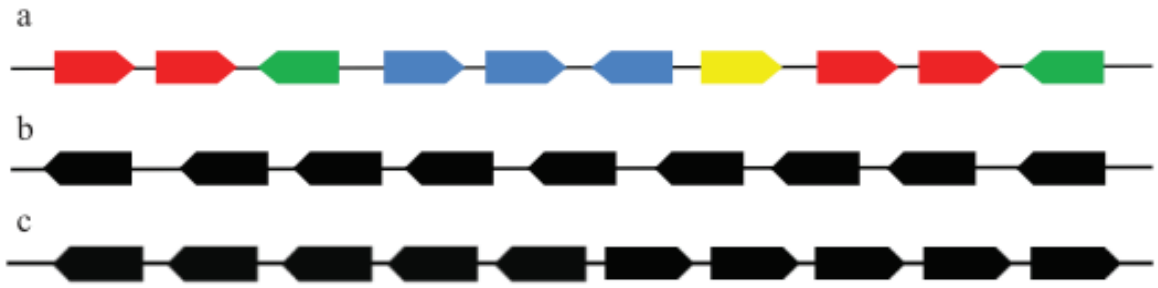
Supplementary Figure 19: Difference of GO categories and PFAM domains between *Brachypodium* and rice and *Ae. tauschii*. Difference of ratio of GO categories and PFAM domain in *Ae. tauschii* to the total GO/PFAM reservoir of a particular organism was computed and normalized. Negative values indicate enriched GOs/PFAMs in the reference organism, whereas positive values show GOs/PFAMs in *Ae. tauschii*. (A) GO *Brachypodium* vs. *Ae. tauschii*; (B) GO Rice vs. *Ae. tauschii*; (C) PFAM *Brachypodium* vs. *Ae. tauschii*; (D) PFAM Rice vs. *Ae. tauschii*.



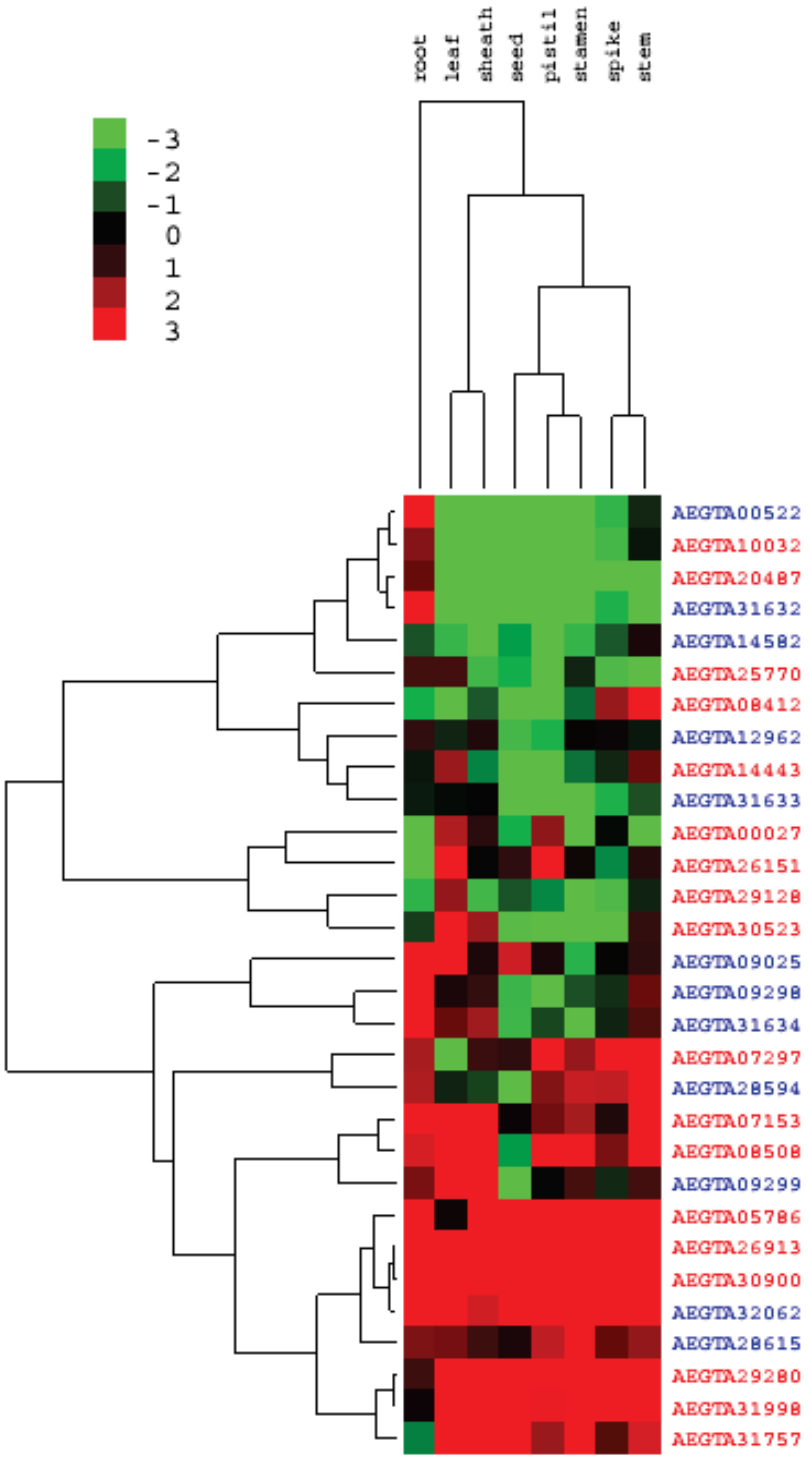
Supplementary Figure 20: Conservation of *Ae. tauschii* gene families. Differences in gene family size between *Ae. tauschii* and the reference grass species are shown. Negative values indicate a decreased gene copy number in *Ae. tauschii* and positive values an increased gene copy number in *Ae. tauschii*, respectively. The dashed line indicates conservation of gene copy number (1:1 relationship). For each reference gene family size category the total number of gene families within is denoted and the boxes color visualizes the relative frequency of clusters. Boxplots are restricted to lower quartile and upper quartiles and whiskers contain 90% of the observed values.



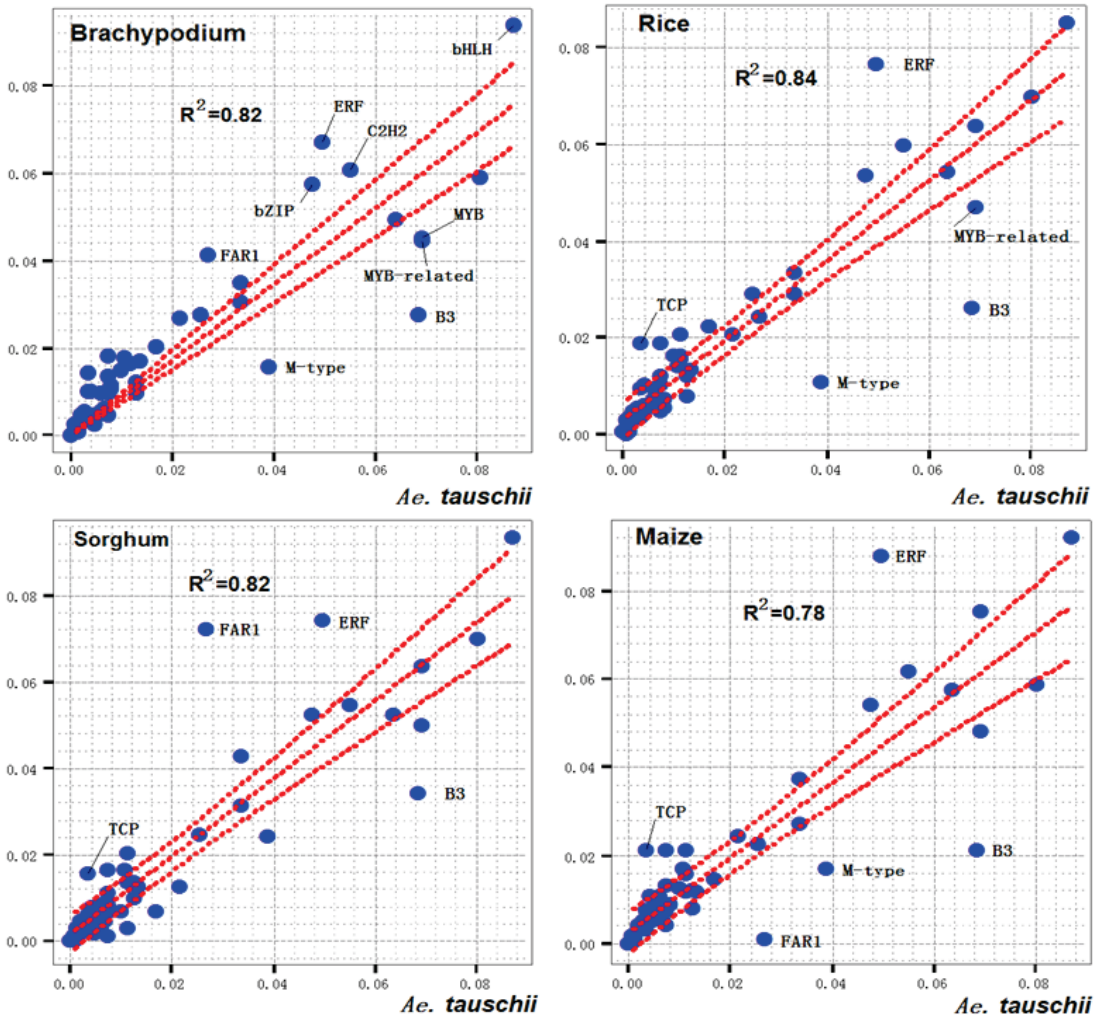
Supplementary Figure 21: Categorization of *Ae. tauschii* resistance gene analogs (RGAs). The figure was split into two sections to give a better resolution.



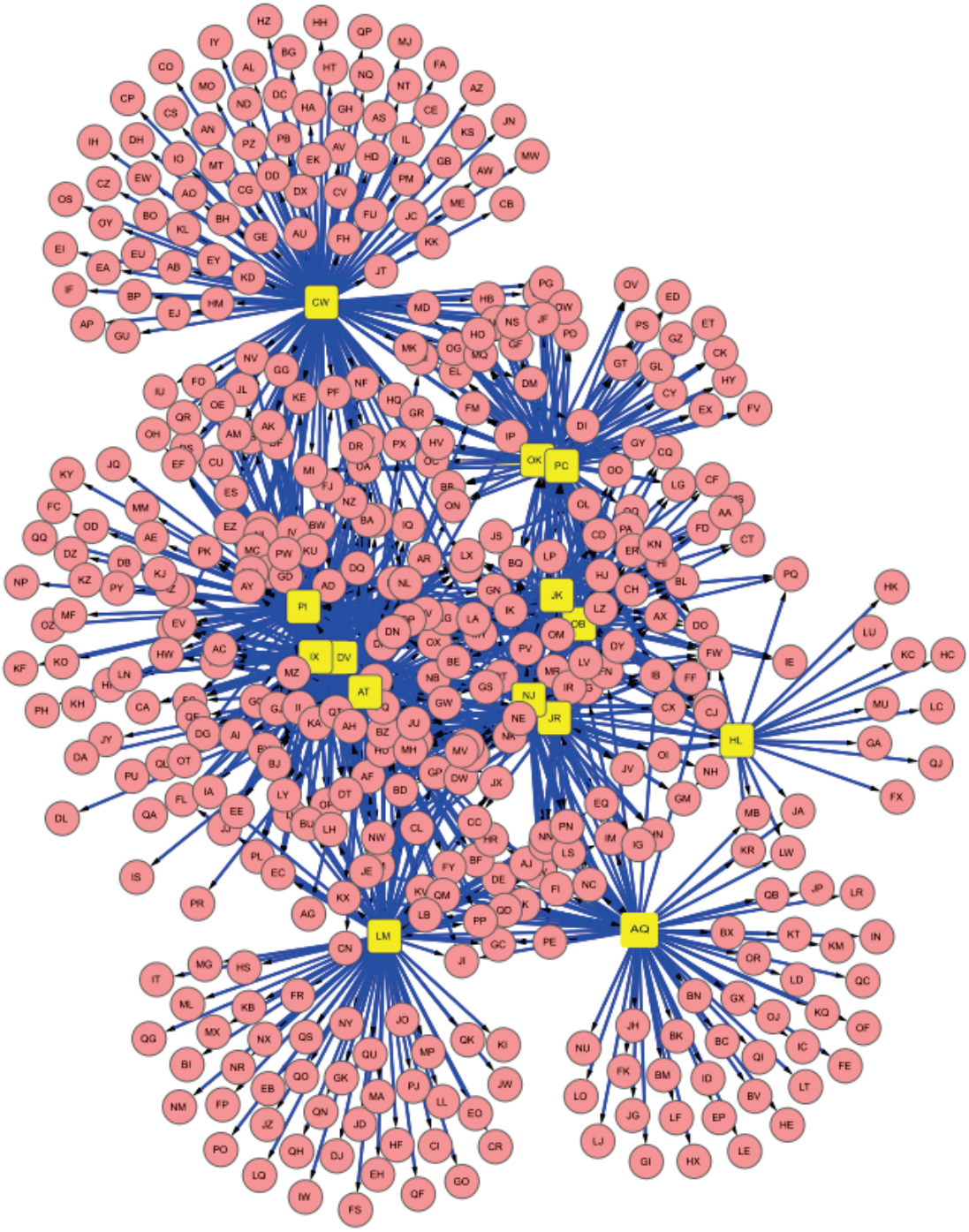
Supplementary Figure 22: The genomic organization of *R* gene clusters in *Ae. tauschii* genome. (a) A common cluster of R genes in a range of 227 Kb on scaffold20283; (b) An example of tandem duplication of R genes in a range of 202 Kb of scaffold3915; (c) An example of tandem duplication of R genes in 393 Kb of scaffold4725. The genes in same color come from the same sub-family with based on >80% nucleotide identity.



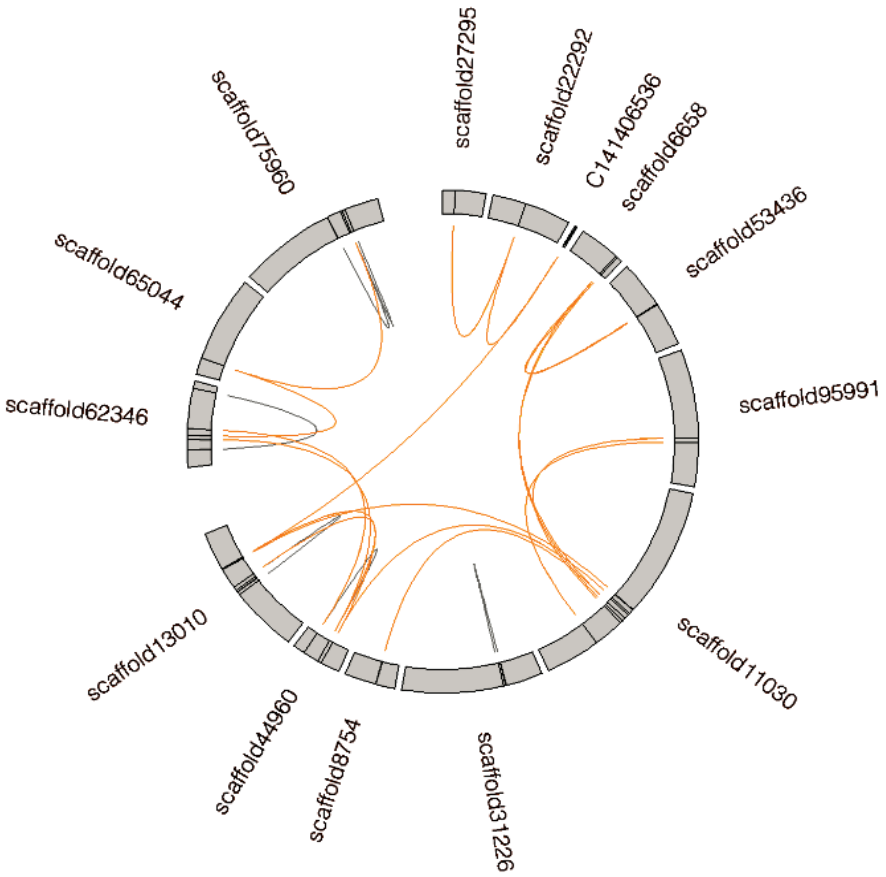
Supplementary Figure 23: Expression of selected cold-related genes in *Ae. tauschii*. The genes shown in this figure comprise of two categories of cold-related genes: *Ae. tauschii*-specific genes (red), Pooideae-specific genes (blue).



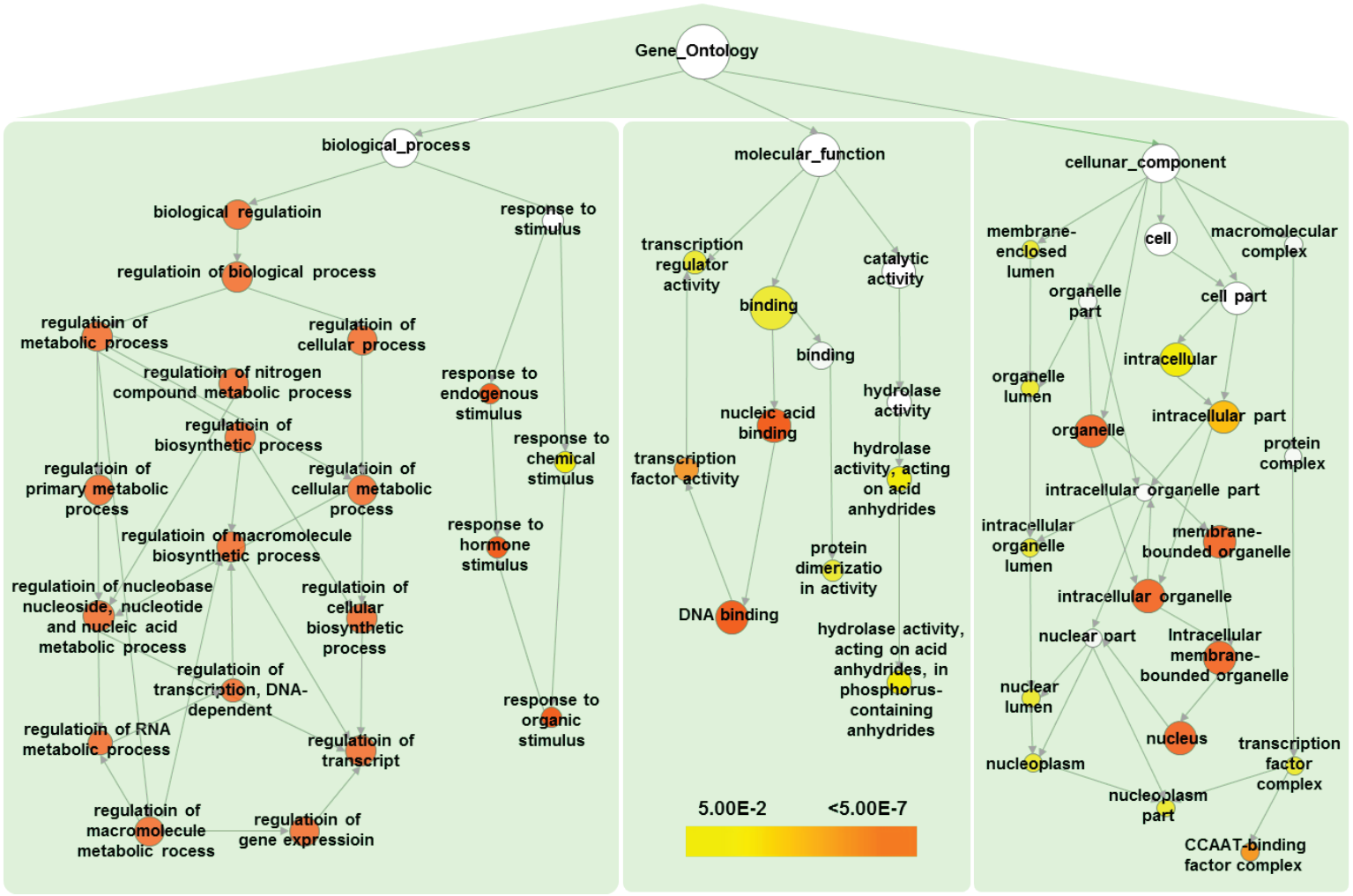
Supplementary Figure 24: Scatter plots of the *Ae. tauschii* transcription factors. Comparison between two genomes displays the *Ae. tauschii* transcription factors (TFs) as linear regressions. The confidence interval (p =0.01) of the regression is reported (dotted lines). Initial data for calculating the regressions were the percentages of each TF family over the total number of TFs present in each genome that was assessed. TF families that deviate significantly from the regression are indicated.



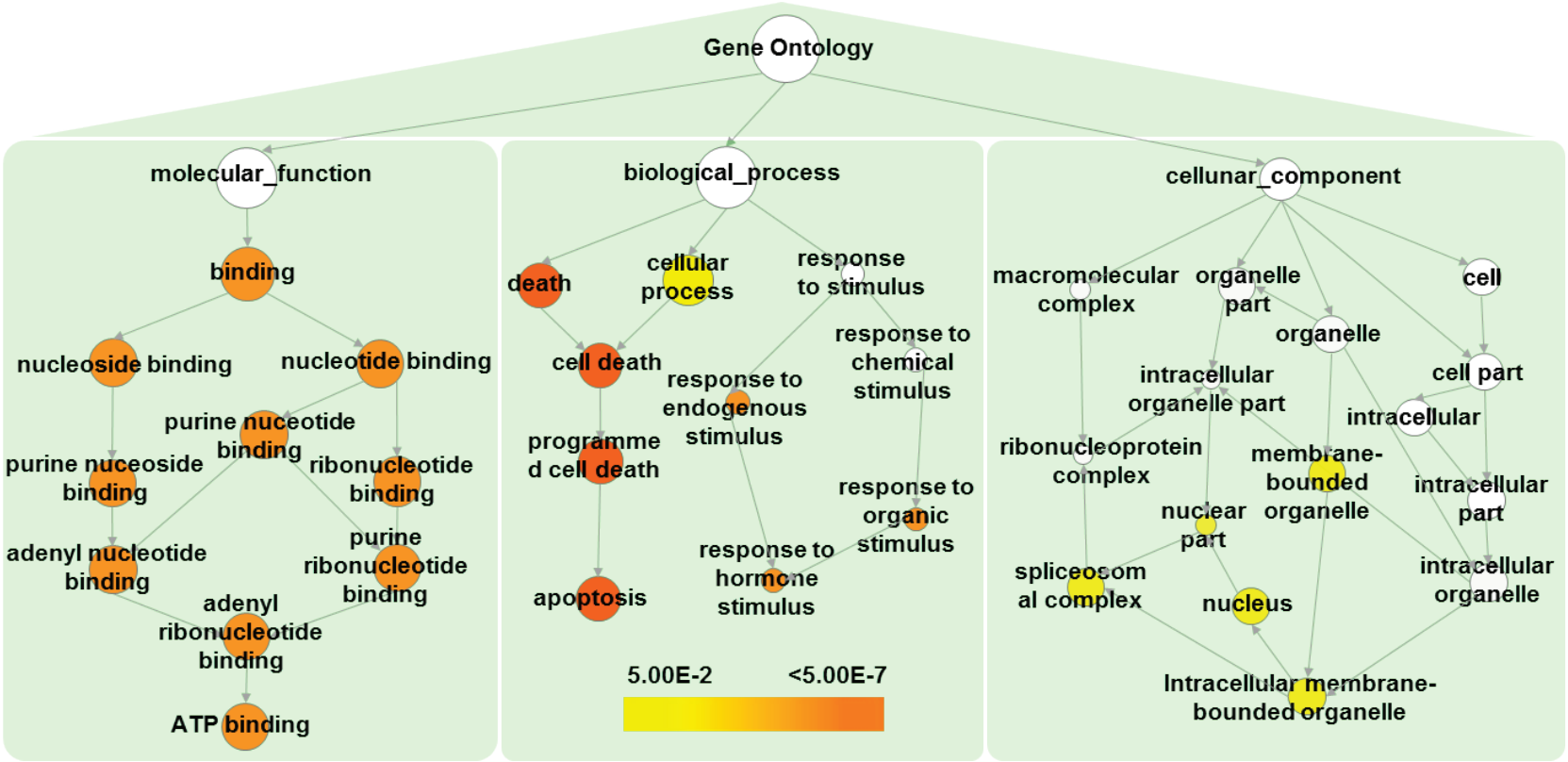
Supplementary Figure 25: Co-expression genes in *Ae. tauschii* genome. Blue edge corresponds to interaction between genes; each node presents a gene and the yellow interaction hub nodes were 14 transcript factors, in which AEGTA32664 (PI) involved in drought resistance; each gene represented by two capital letters. Please see the supplementary Excel file *The gene ID in the TF co-expression figure.*



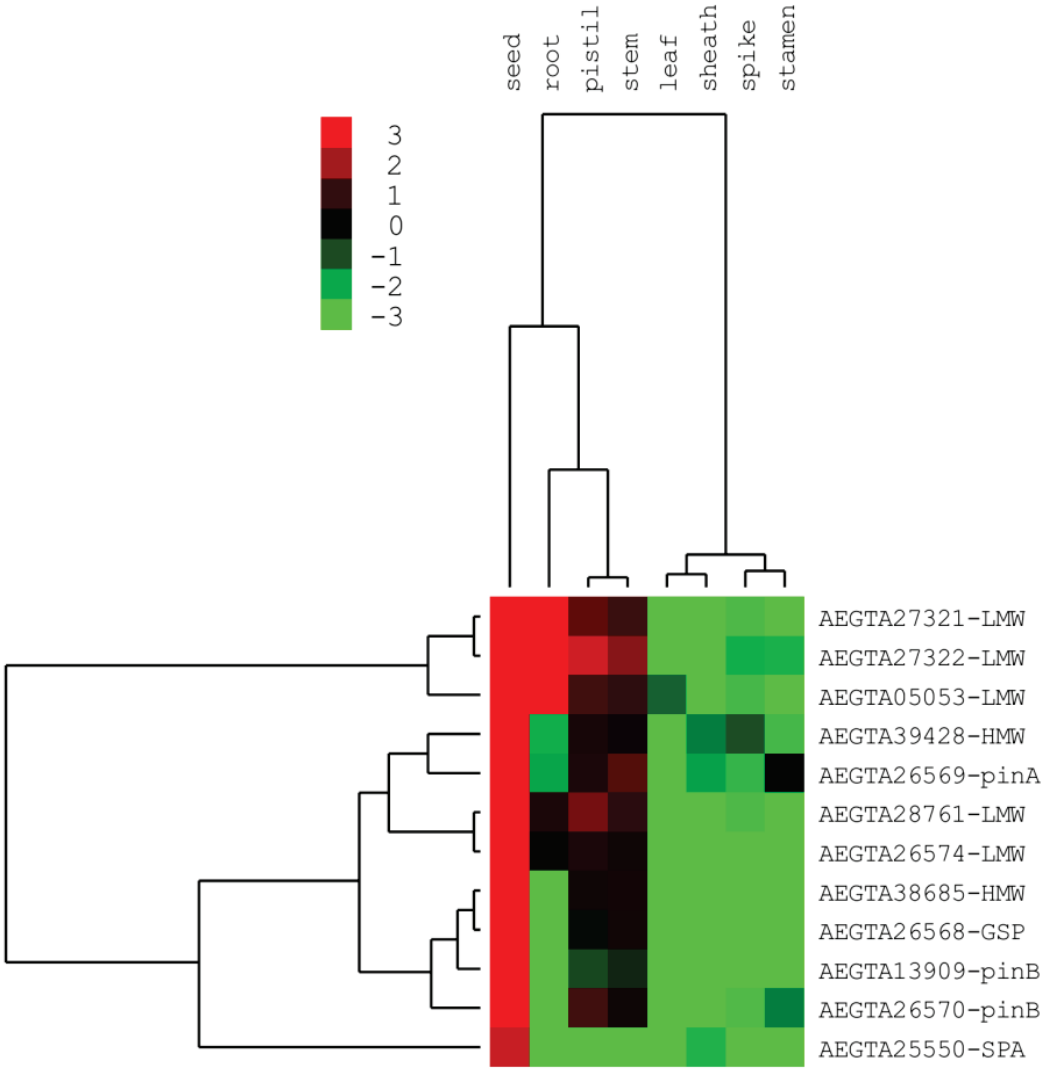
Supplementary Figure 26: The distribution of miR2118 family genes on the scaffolds of *Ae. tauschii*.



Supplementary Figure 27: Enriched GO biology process terms in conserved miRNA targets of *Ae. tauschii*. Significantly over-represented GO biology processes were visualized in Cytoscape. The size of the node is proportional to the number of targets in the GO category. The color represents enrichment significance (Orange solid nodes represent GO term enriched by duplicated miRNA target genes; yellow solid nodes represent GO terms enriched by non-duplicated miRNA target; white nodes represent non-enriched GO term for the hierarchical relationship). The color represents enrichment significance, the deeper the color on a color scale, the higher the enrichment significance. (Note: enrichment significance level < 0.05, and the false discovery rate (FDR) < 0.05).



Supplementary Figure 28: Enriched GO biology process terms in novel miRNA targets of *Ae. tauschii*. Significantly overrepresented GO biology processes were visualized in Cytoscape. The size of the node is proportional to the number of targets in the GO category. The color represents enrichment significance (Orange solid nodes represent GO term enriched by duplicated miRNA target genes; yellow solid nodes represent GO terms enriched by non-duplicated miRNA target; white nodes represent non-enriched GO term for the hierarchical relationship). The color represents enrichment significance, the deeper the color on a color scale, the higher the enrichment significance. (Note: enrichment significance level < 0.05, and the false discovery rate (FDR) < 0.05).



Supplementary Figure 29: The expression of twelve genes involved in grain quality in different tissues.

References:

1 Chao, S. *et al.* RFLP-based genetic maps of wheat homoeologous group 7 chromosomes. *Theor. Appl. Genet.* **78**, 495-504 (1989).

2 Li, R. *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265-272 (2010).

3 Carninci, P. *et al.* High efficiency selection of full-length cDNA by improved biotinylated CAP trapper. *DNA Res.* **4**, 61-66 (1997).

4 Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966-1967 (2009).

5 Kent, W. J. BLAT—The BLAST-Like Alignment Tool. *Genome Res.* **12**, 656-664 (2002).

6 Huang, X. & Madan, A. CAP3: a DNA sequence assembly program. *Genome Res.* **9**, 868-877 (1999).

7 Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659 (2006).

8 Salamov, A. A. & Solovyev, V. V. Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res.* **10**, 516-522 (2000).

9 Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988-995 (2004).

10 Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111 (2009).

11 Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protocols* **7**, 562-578 (2012).

12 Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 0955-0964 (1997).

13 Nawrocki, E. P., Kolbe, D. L. & Eddy, S. R. Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25**, 1335-1337 (2009).

14 Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**, 573-580 (1999).

15 Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* **8**, 973-982 (2007).

16 Wicker, T. *et al.* A whole-genome snapshot of 454 sequences exposes the composition of the barley genome and provides evidence for parallel evolution of genome size in wheat and barley. *Plant J.* **59**, 712-722 (2009).

17 Choulet, F. *et al.* Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. *Plant Cell* **22**, 1686-1701 (2010).

18 Bennett, M. D. & Smith, J. B. Nuclear DNA amounts in angiosperms. *Philosophical transactions of the royal society of London. B, biological sciences* **274**, 227-274 (1976).

19 Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111-120 (1980).

20 Gaut, B. S., Morton, B. R., McCaig, B. C. & Clegg, M. T. Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcL*. *Proc. Natl Acad. Sci. USA* **93**, 10274-10279 (1996).

21 Williams, M. *et al.* Pliocene climate and seasonality in North Atlantic shelf seas. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences* **367**, 85-108 (2009).

22 Wolfe, K. H., Sharp, P. M. & Li, W.-H. Rates of synonymous substitution in plant nuclear genes. *J. Mol. Evol.* **29**, 208-211 (1989).

23 Ronquist, F. *et al.* MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large

model space. *Syst. Biol.* **61**, 539-542 (2012).
24 Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586-1591 (2007).
25 Initiative, T. I. B. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **463**, 763-768 (2010).
26 Bolot, S. *et al.* The 'inner circle' of the cereal genomes. *Curr. Opin. Plant Biol.* **12**, 119-125 (2009).
27 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
28 Li, R. *et al.* SNP detection for massively parallel whole-genome resequencing. *Genome Res.* **19**, 1124-1132 (2009).
29 Somers, D. J., Isaac, P. & Edwards, K. A high-density microsatellite consensus map for bread wheat (*Triticum aestivum* L.). *Theor. Appl. Genet.* **109**, 1105-1114 (2004).
30 Song, Q. *et al.* Development and mapping of microsatellite (SSR) markers in wheat. *Theor. Appl. Genet.* **110**, 550-560 (2005).
31 Luo, M. C. *et al.* 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 (2009).
32 Allen, A. M. *et al.* Transcript-specific, single-nucleotide polymorphism discovery and linkage analysis in hexaploid bread wheat (*Triticum aestivum* L.). *CORD Conference Proceedings* **9**, 1086-1099 (2011).
33 Poland, J. A., Brown, P. J., Sorrells, M. E. & Jannink, J. L. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS One* **7**, e32253-e32253 (2012).
34 Salse, J. *et al.* Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. *Plant Cell* **20**, 11-24 (2008).
35 Zhang, Z. *et al.* KaKs_Calculator: Calculating Ka and Ks Through model selection and model averaging. *Genomics, Proteomics & Bioinformatics* **4**, 259-263 (2006).
36 Li, L., Stoeckert, C. J., Jr. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178-2189 (2003).
37 van Dongen, S. M. *Graph Clustering by Flow Simulation*, University of Utrecht, The Netherlands, (2000).
38 Paterson, A. H. *et al.* The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**, 551-556 (2009).
39 Ouyang, S. *et al.* The TIGR rice genome annotation resource: improvements and new features. *Nucleic Acids Res.* **35**, D883-D887 (2007).
40 Mayer, K. F. *et al.* Unlocking the barley genome by chromosomal and comparative genomics. *Plant Cell* **23**, 1249-1263 (2011).
41 Hunter, S. *et al.* InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.* **40**, D306-D312 (2012).
42 McCarthy, F. M. *et al.* AgBase: a functional genomics resource for agriculture. *BMC Genomics* **7**, 229-229 (2006).
43 Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput Biol* **7**, e1002195 (2011).
44 Dubcovsky, J. & Dvorak, J. Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science* **316**, 1862-1866 (2007).
45 McHale, L., Tan, X., Koehl, P. & Michelmore, R. W. Plant NBS-LRR proteins: adaptable guards. *CORD Conference Proceedings* **7**, 212-212 (2006).
46 Yue, J. X., Meyers, B. C., Chen, J. Q., Tian, D. & Yang, S. Tracing the origin and evolutionary history of plant nucleotide-binding site-leucine-rich repeat (*NBS-LRR*) genes. *New Phytol.* **193**, 1049-1063 (2012).

47 Luo, S. *et al.* Dynamic nucleotide-binding-site and leucine-rich-repeat- encoding genes in the grass family. *Plant Physiol.* (2012).

48 Meyers, B. C., Kozik, A., Griego, A., Kuang, H. & Michelmore, R. W. Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis*. *Plant Cell* **15**, 809-834 (2003).

49 Leister, D. Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance gene. *Trends Genet.* **20**, 116-122 (2004).

50 Gusta, L. & Fowler, D. *Cold resistance and injury in winter cereals*. 160-178 (John Wiley and Sons, 1979).

51 Sakai, A. & Larcher, R. *Frost survival of plants - responses and adaptation to freezing stress*. (Springer, 1987).

52 Thomashow, M. F. PLANT COLD ACCLIMATION: freezing tolerance genes and regulatory mechanisms. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **50**, 571-599 (1999).

53 Heidarvand, L. & Maali Amiri, R. What happens in plant molecular responses to cold stress? *Acta Physiol. Plant.* **32**, 419-431 (2010).

54 Tondelli, A., Francia, E., Barabaschi, D., Pasquariello, M. & Pecchioni, N. Inside the *CBF* locus in *Poaceae*. *Plant Sci.* **180**, 39-45 (2011).

55 Yamaguchi-Shinozaki, K. & Shinozaki, K. Transcriptional regulatory networks in cellular responses and tolerance to dehydration and cold stresses. *Annu. Rev. Plant Biol.* **57**, 781-803 (2006).

56 Thomashow, M. F. Molecular basis of plant cold acclimation: insights gained from studying the CBF cold response pathway. *Plant Physiol.* **154**, 571-577 (2010).

57 Nakashima, K., Ito, Y. & Yamaguchi-Shinozaki, K. Transcriptional regulatory networks in response to abiotic stresses in Arabidopsis and grasses. *Plant Physiol.* **149**, 88-95 (2009).

58 Dhillon, T. *et al.* Regulation of freezing tolerance and flowering in temperate cereals: the VRN-1 connection. *Plant Physiol.* **153**, 1846-1858 (2010).

59 Zhang, H. *et al.* PlantTFDB 2.0: update and improvement of the comprehensive plant transcription factor database. *Nucleic Acids Research* **39**, D1114-D1117 (2011).

60 Mao, X. *et al.* Transgenic expression of *TaMYB2A* confers enhanced tolerance to multiple abiotic stresses in *Arabidopsis*. *Funct. Integr. Genomics* **11**, 445-465 (2011).

61 Margolin, A. *et al.* ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* **7 Suppl 1**, S7 (2006).

62 Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31**, 3406-3415 (2003).

63 Gordon, D., Desmarais, C. & Green, P. Automated finishing with autofinish. *Genome Res.* **11**, 614-625 (2001).

64 Jurka, J. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet.* **16**, 418-420 (2000).

65 Allen, E., Xie, Z., Gustafson, A. M. & Carrington, J. C. microRNA-directed phasing during trans-acting siRNA biogenesis in plants. *Cell* **121**, 207-221 (2005).

66 Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639-1645 (2009).

67 Maere, S., Heymans, K. & Kuiper, M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **21**, 3448-3449 (2005).

68 Kohl, M., Wiese, S. & Warscheid, B. Cytoscape: software for visualization and analysis of biological networks. *Methods Mol. Biol.* **696**, 291-303 (2011).

69 Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N. & Golani, I. Controlling the false discovery rate in behavior genetics research. *Behav. Brain Res.* **125**, 279-284 (2001).

70 Wei, F., Wing, R. A. & Wise, R. P. Genome dynamics and evolution of the *Mla* (powdery mildew) resistance locus in barley. *Plant Cell* **14**, 1903-1917 (2002).

71 Yoshimura, S. *et al.* Expression of *Xa1*, a bacterial blight-resistance gene in rice, is induced by bacterial inoculation. *Proc. Natl Acad. Sci. USA* **95**, 1663-1668 (1998).

72 da Maia, L. C. *et al.* SSR Locator: Tool for simple sequence repeat discovery integrated with primer design and PCR simulation. *Int J Plant Genomics* **2008**, 1-9 (2008).

73 Arumuganathan, K. & Earle, E. D. Nuclear DNA content of some important plant species. *Plant Mol. Biol. Rep.* **9**, 211-215 (1991).

74 Wang, R. X. *et al.* QTL mapping for grain filling rate and yield-related traits in RILs of the Chinese winter wheat population Heshangmai x Yu8679. *Theor. Appl. Genet.* **118**, 313-325 (2009).

75 Jantasuriyarat, C., Vales, M. I., Watson, C. J. & Riera-Lizarazu, O. Identification and mapping of genetic loci affecting the free-threshing habit and spike compactness in wheat (*Triticum aestivum* L.). *Theor. Appl. Genet.* **108**, 261-273 (2004).

76 Raupp, W. J., Sukhwinder, S., Brown-Guedira, G. L. & Gill, B. S. Cytogenetic and molecular mapping of the leaf rust resistance gene *Lr39* in wheat. *Theor. Appl. Genet.* **102**, 347-352 (2001).

77 Korzun, V., Roder, M. S., Ganal, M. W., Worland, A. J. & Law, C. N. Genetic analysis of the dwarfing gene (*Rht8*) in wheat. I. Molecular mapping of *Rht8* on the short arm of chromosome 2D of bread wheat (*Triticum aestivum* L.). *Theor. Appl. Genet.* **96**, 1104-1109 (1998).

78 Narasimhamoorthy, B., Gill, B. S., Fritz, A. K., Nelson, J. C. & Brown-Guedira, G. L. Advanced backcross QTL analysis of a hard winter wheat x synthetic wheat population. *Theor. Appl. Genet.* **112**, 787-796 (2006).

79 Shen, X., Zhou, M., Lu, W. & Ohm, H. Detection of Fusarium head blight resistance QTL in a wheat population using bulked segregant analysis. *Theor. Appl. Genet.* **106**, 1041-1047 (2003).

80 Liu, S. *et al.* QTL associated with Fusarium head blight resistance in the soft red winter wheat Ernie. *Theor. Appl. Genet.* **115**, 417-427 (2007).

81 Friesen, T. L., Meinhardt, S. W. & Faris, J. D. The *Stagonospora nodorum*-wheat pathosystem involves multiple proteinaceous host-selective toxins and corresponding host sensitivity genes that interact in an inverse gene-for-gene manner. *Plant J.* **51**, 681-692 (2007).

82 Nalam, V. J., Vales, M. I., Watson, C. J., Johnson, E. B. & Riera-Lizarazu, O. Map-based analysis of genetic loci on chromosome 2D that affect glume tenacity and threshability, components of the free-threshing habit in common wheat (*Triticum aestivum* L.). *Theor. Appl. Genet.* **116**, 135-145 (2007).

83 Zhang, Z., Friesen, T., Simons, K., Xu, S. & Faris, J. Development, identification, and validation of markers for marker-assisted selection against the *Stagonospora nodorum* toxin sensitivity genes *Tsn1* and *Snn2* in wheat. *Molecular breeding* **23**, 35-49 (2009).

84 Rebetzke, G. J., Condon, A. G., Farquhar, G. D., Appels, R. & Richards, R. A. Quantitative trait loci for carbon isotope discrimination are repeatable across environments and wheat mapping populations. *Theor. Appl. Genet.* **118**, 123-137 (2008).

85 Huynh, B. L. *et al.* Quantitative trait loci for grain fructan concentration in wheat (*Triticum aestivum* L.). *Theor. Appl. Genet.* **117**, 701-709 (2008).

86 McCartney, C. A. *et al.* Mapping quantitative trait loci controlling agronomic traits in the spring wheat cross RL4452x'AC Domain'. *Genome* **48**, 870-883 (2005).

87 Kirby, J., Vinh, H. T., Reader, S. M. & Dudnikov, A. J. Genetic mapping of the *Acph1* locus in *Aegilops tauschii*. *Plant breeding* **124**, 523-524 (2005).

88 Bovill, W. D. *et al.* Identification of novel QTL for resistance to crown rot in the doubled haploid wheat population 'W21MMT70' x 'Mendos'. *Plant breeding* **125**, 538-543 (2006).

89 Prasad, M. *et al.* QTL analysis for grain protein content using SSR markers and validation studies using NILs in bread wheat. *Theor. Appl. Genet.* **106**, 659-667 (2003).

90 Rebetzke, G. J., Ellis, M. H., Bonnett, D. G. & Richards, R. A. Molecular mapping of genes for coleoptile growth in bread wheat (*Triticum aestivum* L.). *Theor. Appl. Genet.* **114**, 1173-1183 (2007).

91 Börner, A. *et al.* Mapping of quantitative trait loci determining agronomic important characters in hexaploid wheat (*Triticum aestivum* L.). *Theor. Appl. Genet.* **105**, 921-936 (2002).

92 He, R. *et al.* Inheritance and mapping of powdery mildew resistance gene *Pm43* introgressed from *Thinopyrum intermedium* into wheat. *Theor. Appl. Genet.* **118**, 1173-1180 (2009).

93 Sun, D. J. *et al.* A novel STS marker for polyphenol oxidase activity in bread wheat. *Molecular breeding* **16**, 209-218 (2005).

94 He, X. Y. *et al.* Allelic variation of polyphenol oxidase (PPO) genes located on chromosomes 2A and 2D and development of functional markers for the *PPO* genes in common wheat. *Theor. Appl. Genet.* **115**, 47-58 (2007).

95 Kuchel, H., Williams, K. J., Langridge, P., Eagles, H. A. & Jefferies, S. P. Genetic dissection of grain yield in bread wheat. I. QTL analysis. *Theor. Appl. Genet.* **115**, 1029-1041 (2007).

96 Yang, D. L., Jing, R. L., Chang, X. P. & Li, W. Identification of quantitative trait loci and environmental interactions for accumulation and remobilization of water-soluble carbohydrates in wheat (*Triticum aestivum* L.) stems. *Genetics* **176**, 571-584 (2007).

97 Cloutier, S. *et al.* Leaf rust resistance gene *Lr1*, isolated from bread wheat (*Triticum aestivum* L.) is a member of the large psr567 gene family. *Plant Mol. Biol.* **65**, 93-106 (2007).

98 Feuillet, C. *et al.* Map-based isolation of the leaf rust disease resistance gene *Lr10* from the hexaploid wheat (*Triticum aestivum* L.) genome. *Proc. Natl Acad. Sci. USA* **100**, 15253-15258 (2003).

99 Huang, L. *et al.* Evolution of new disease specificity at a simple resistance locus in a crop-weed complex: reconstitution of the *Lr21* gene in wheat. *Genetics* **182**, 595-602 (2009).

100 Krattinger, S. G. *et al. Lr34* multi-pathogen resistance ABC transporter: molecular analysis of homoeologous and orthologous genes in hexaploid wheat and other grass species. *CORD Conference Proceedings* **65**, 392-403 (2011).

101 Bhullar, N. K., Zhang, Z., Wicker, T. & Keller, B. Wheat gene bank accessions as a source of new alleles of the powdery mildew resistance gene *Pm3*: a large scale allele mining project. *CORD Conference Proceedings* **10**, 88-88 (2010).

102 Cao, A. *et al.* Serine/threonine kinase gene *Stpk-V*, a key member of powdery mildew resistance gene *Pm21*, confers powdery mildew resistance in wheat. *CORD Conference Proceedings* **108**, 7727-7732 (2011).

103 Huang, L. *et al.* Haplotype variations of gene *Ppd-D1* in *Aegilops tauschii* and their implications on wheat origin. *Genet. Resour. Crop Evol.* **59**, 1027-1032 (2012).

104 Dubcovsky, J. *et al.* Effect of photoperiod on the regulation of wheat vernalization genes *VRN1* and *VRN2*. *Plant Mol. Biol.* **60**, 469-480 (2006).

105 Yan, L. *et al.* The wheat *VRN2* gene is a flowering repressor down-regulated by vernalization. *Science* **303**, 1640-1644 (2004).

106 Yan, L. *et al.* The wheat and barley vernalization gene *VRN3* is an orthologue of *FT*. *Proc. Natl Acad. Sci. USA* **103**, 19581-19586 (2006).

107 Peng, J. *et al.* 'Green revolution' genes encode mutant gibberellin response modulators. *Nature* **400**, 256-261 (1999).

108 Simons, K. J. *et al.* Molecular characterization of the major wheat domestication gene *Q*. *Genetics* **172**, 547-555 (2006).

109 Chang, C. *et al.* Identification of allelic variations of puroindoline genes controlling grain hardness in wheat using a modified denaturing PAGE. *Euphytica* **152**, 225-234 (2006).

110 Huang, X. *et al.* Natural variation at the *DEP1* locus enhances grain yield in rice. *CORD Conference Proceedings* **41**, 494-497 (2009).

111 Distelfeld, A. *et al.* Colinearity between the barley grain protein content (GPC) QTL on chromosome arm 6HS and the wheat *Gpc-B1* region. *Mol. Breed.* **22**, 25-38 (2008).

112 Taketa, S. *et al.* Barley grain with adhering hulls is controlled by an ERF family transcription factor gene regulating a lipid biosynthesis pathway. *Proc. Natl Acad. Sci. USA* **105**, 4062-4067 (2008).

113 Song, X. J., Huang, W., Shi, M., Zhu, M. Z. & Lin, H. X. A QTL for rice grain width and weight encodes a previously unknown RING-type E3 ubiquitin ligase. *Nat. Genet.* **39**, 623-630 (2007).

114 Zhu, K. *et al. Erect panicle2* encodes a novel protein that regulates panicle erectness in *indica* rice. *CORD Conference Proceedings* **184**, 343-350 (2010).

115 Li, Y. *et al.* Natural variation in *GS5* plays an important role in regulating grain size and yield in rice. *Nat. Genet.* **43**, 1266-1269 (2011).

116 Jiao, Y. *et al.* Regulation of *OsSPL14* by OsmiR156 defines ideal plant architecture in rice. *Nat. Genet.* **42**, 541-544 (2010).

117 Thurber, C. S. *et al.* Molecular evolution of shattering loci in U.S. weedy rice. *Mol. Ecol.* **19**, 3271-3284 (2010).

118 Li, X. *et al.* Control of tillering in rice. *Nature* **422**, 618-621 (2003).

119 Wang, H. *et al.* The origin of the naked grains of maize. *Nature* **436**, 714-719 (2005).

120 Camus-Kulandaivelu, L. *et al.* Patterns of molecular evolution associated with two selective sweeps in the *Tb1-Dwarf8* region in maize. *Genetics* **180**, 1107-1121 (2008).

International Wheat Genome Sequencing Consortium

Catherine Feuillet[1], Kellye Eversole[2], Beat Keller[3] Jan Dvorak[4], Bikram Gill[5], Yasunari Ogihara[6] & Rudi Appels[7]

[1]Institut National de la Recherche Agronomique (INRA), UMR INRA/UBP 1095 GDEC, 63100 Clermont-Ferrand, France.
[2]Eversole Associates, Bethesda, Maryland 20816, USA.
[3]Institute of Plant Biology, University of Zurich, CH-8008 Zurich, Switzerland.
[4]Univeristy of California, Davis, California 95616, USA.
[5]Kansas State University, Manhattan, Kansas 66506, USA.
[6]Kihara Institute for Biological Research, Yokohama City University, Maioka-cho 641-12, Totsuka-ku, 244-0813 Yokohama, Japan.
[7]Centre for Comparative Genomics, Murdoch University, Perth, WA 6150, Australia.