



Supplementary Materials for

Shifting the limits in wheat research and breeding using a fully annotated reference genome

International Wheat Genome Sequencing Consortium (IWGSC)*†

*All authors with their affiliations are listed in the main article.

†Corresponding authors: Rudi Appels (rudi.appels@unimelb.edu.au); Kellye Eversole (eversole@eversoleassociates.com); Nils Stein (stein@ipk-gatersleben.de)

Published 17 August 2018, *Science* **361**, eaar7191 (2018)
DOI: 10.1126/science.aar7191

This PDF file includes:

Materials and Methods
Figs. S1 to S59
Tables S1 to S43
Captions for Databases S1 to S5
References

Other Supporting Online Material for this manuscript includes the following:

(available at www.sciencemag.org/content/361/6403/eaar7191/suppl/DC1)

Database S1. Metadata of 850 RNAseq samples used in the study (Excel file)
Database S2. SlimGO ubiquitous and Tissue-exclusive genes (Excel file)
Database S3. GO terms of the WGCNA850 analysis (Excel file)
Database S4. WGCNA Module Assignment (Excel file)
Database S5. Module 8 and 11 TF Arabidopsis and rice orthologs (Excel file)

Materials and Methods

Materials

Genome, transcriptome and other data resources

The IWGSC RefSeq v1.0 assembly, annotation data and related data are available in the IWGSC Data Repository hosted at URGI: <https://wheat-urgi.versailles.inra.fr/Seq-Repository>. The data can be downloaded but also analyzed using tools: BLAST (https://urgi.versailles.inra.fr/blast_iwgsc/blast.php), JBrowse (https://urgi.versailles.inra.fr/jbrowseiwgsc/gmod_jbrowse/?data=myData%2FIWGSC_RefSeq_v1.0) and InterMine (<https://urgi.versailles.inra.fr/WheatMine>). The data are linked to wheat genetic and phenomic data via the Wheat@URGI portal (56).

Chromosome-specific BAC libraries and physical maps

Individual clones and BAC libraries used to construct chromosome-specific physical maps can be obtained at <https://cnrgv.toulouse.inra.fr/en/Library/Wheat>. Physical maps are available at: https://urgi.versailles.inra.fr/download/iwgsc/Physical_maps/ and displayable at https://urgi.versailles.inra.fr/gb2/gbrowse/wheat_phys_pub/.

BAC sequence assemblies

Sequence assemblies of BAC clones representing complete or partial MTPs from 8 chromosomes and 2 chromosome arms are available at: https://urgi.versailles.inra.fr/download/iwgsc/BAC_Assemblies/.

Whole Genome Profiling (WGP™) tags of MTP BAC clones

WGP™ data generated from BAC clones representing the MTP of all 21 wheat chromosomes are available for download from IWGSC-BayerCropScience WGP™ tags: https://urgi.versailles.inra.fr/download/iwgsc/IWGSC_BayerCropScience_WGPTM_tags.

Bionano optical maps

Bionanogenomics optical maps for group 7 chromosomes are available at: https://urgi.versailles.inra.fr/download/iwgsc/IWGSC_RefSeq_Annotations/v1.0/iwgsc_refseqv1.0_optical_maps_group7.zip.

Radiation Hybrid (RH) maps

RH maps constructed to validate long-range order in physical maps and sequence assemblies for wheat chromosomes are available at: https://urgi.versailles.inra.fr/download/iwgsc/IWGSC_RefSeq_Annotations/v1.0/iwgsc_refseqv1.0_RH_map.zip.

Genetic maps

GBS-SNP marker coordinates and maps of genetic vs physical distances (SynOpDH88_ChineseSpring_v0.5.pdf) and (SynOpRIL993_ChineseSpring_v0.5.pdf) are available at: https://urgi.versailles.inra.fr/download/iwgsc/IWGSC_RefSeq_Annotations/v1.0/iwgsc_refseqv1.0_SynOpRIL993_GBS.zip. GBS data is available under SRA accession number SRP134280 for the submission SUB3762955.

Sequence reads used for whole genome assembly

The raw sequence data used for *de novo* whole genome assembly of genotype ‘Chinese Spring’ (CS) is available from the Sequence Read Archive under accession number SRP114784.

Sequence reads of RNAseq data

The raw sequence data from 321 RNAseq samples generated for this study are available under SRA accession IDs PRJEB25639, PRJEB23056, PRJNA436817, PRJEB25640, SRP133837, PRJEB25593.

Chromosome conformation capture (Hi-C) sequencing data

Hi-C sequence data generated from four independent Hi-C libraries are available under accession number PRJEB25248.

ChIP sequence data of histone marks and methylome data

All ChIP seq data are available under SRA study PRJNA420988 (SRP126222) and can be accessed using SRA RunSelector: <https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRP126222>. The CS bisulfite sequencing data is available under project ID SRP133674 (SRR6792673-SRR6792689 for the independent runs).

Chromosome-scale sequence assemblies (pseudomolecules)

Assemblies of all 21 wheat chromosomes are available at:
https://urgi.versailles.inra.fr/download/iwgc/IWGSC_RefSeq_Assemblies/.

Assemblies of organellar genomes

Assemblies representing the CS organellar genomes were assembled from the raw sequencing data produced for the whole genome assembly. A 135,905 bp circular chloroplast genome sequence and a 452,256 bp circular mitochondrial genome sequence were deposited at NCBI Genbank (MH051715, MH051716).

Genome annotations

All annotation data are available at:
https://urgi.versailles.inra.fr/download/iwgc/IWGSC_RefSeq_Annotations/.

Gene annotations are available here:
https://urgi.versailles.inra.fr/download/iwgc/IWGSC_RefSeq_Annotations/v1.1/iwgc_refseqv1.1_genes_2017July06.zip.

Pseudogene annotations are available here: :
https://urgi.versailles.inra.fr/download/iwgc/IWGSC_RefSeq_Annotations/v1.0/iwgc_refseqv1.0_PGSEB_annotation_files.zip.

Functional annotation data for all HC and LC protein-coding genes (wheatCS_IWGSC_FunctionalAnnotation_v1__HCgenes_v1.0.TAB, wheatCS_IWGSC_FunctionalAnnotation_v1__HCgenes_v1.0-repr.TEcleaned.TAB, wheatCS_IWGSC_FunctionalAnnotation_v1__LCgenes_v1.0.TAB) are available at:
https://urgi.versailles.inra.fr/download/iwgc/IWGSC_RefSeq_Annotations/v1.0/iwgc_refseqv1.0_FunctionalAnnotation_v1.zip.

Gene families

Information on expanded and contracted bread wheat gene families is available as database S6 (52) at: <http://dx.doi.org/10.5447/IPK/2018/5>.

Methods

1. Whole genome sequencing and assembly

1.1. DNA isolation and sequencing

Genomic DNA was isolated from fresh leaf tissue (~10 g) of cultivar CS using a phenol/chloroform large scale nuclei extraction protocol (57). The DNA was used to construct five sized sequencing libraries covering the range of 450 bp to 10 kb using standard protocols by the University of Illinois Roy J. Carver Biotechnology Center. The libraries were sequenced on an Illumina HiSeq 2500 to generate 3.69 Tb of data (equivalent to 234x genomic coverage, based on an estimated genome size of 15.76 Gb) using (Table S1).

1.2 Whole Genome Assembly (WGA)

The genome was assembled using the software package DenovoMAGIC2™ (NRGene, Nes Ziona, Israel) and scaffolded (Table S2) by applying gap-closure and error correction in three steps as described for the wild emmer wheat genome (57):

1. Read pre-processing and error correction: PCR duplicates, Illumina adaptors, Nextera linkers (for the mate pair (MP) libraries) and paired-end (PE) reads found to contain likely sequencing errors (i.e. sub-sequences of ≥ 23 bp not found in at least one other independent read) were removed. From the PE libraries, only pairs with at least 10 bp sequence overlap were merged to create 'stitched reads' (SRs).

2. *De novo* assembly of contigs: the SRs were used to build an initial De Bruijn graph of contigs (kmer=191 bp). By exploring graph structure, the software identified non-repetitive contigs and used SR information to resolve repeats and extend non-repetitive sequences of the contigs where possible (Table S2).

3. *De novo* assembly of scaffolds: a directed graph containing contigs as nodes and edges based on the PE and MP links as vertices was used for scaffolding. Erroneous connections were identified and filtered out to generate unconnected sub-graphs that were ordered into scaffolds. PE reads were used to find reliable paths in the graph for additional repeat resolving. This was accomplished through searching the graph for a unique path of contigs connecting pairs of reads mapping to two different non-repetitive contigs. The scaffolds were then ordered and linked using the MP libraries, estimating gaps between the contigs from the distances between MP links. Scaffold links were only accepted when confirmed with at least three filtered MPs, or at least one filtered MP with supporting confirmation from two or more filter failed MPs where the Nextera adaptor was not found. Scaffolds shorter than 380bp were masked and links between non-repetitive contigs mapping to the same scaffolds were united, generating a directed scaffold graph. A significant number of erroneous MPs linking long non-branched components in the scaffold graph were discarded by the scaffolding procedure that identified the non-branched components in the scaffolds graph. After filtering out the rare connections between them topological sorting based ordering of the initial scaffolds produced the final scaffolds.

1.3 Estimation of sequence accuracy

To estimate the sequence accuracy of the genome assembly a comparison was made between a scaffold from IWGSC RefSeq_v1.0 (scaffold4849-2) and a set of overlapping “finished” BAC clones spanning 760070 bp without inconsistencies from chromosome 3B TaaCsp3BFhA_0053M17, TaaCsp3BFhA_0078G10, TaaCsp3BFhA_0011O13, TaaCsp3BFhA_0149J15, TaaCsp3BFhA_0013G07, TaaCsp3BFhA_0061H24 (58). Alignment of the two sequences with BLAST2 identified 728782 bps aligned in high scoring pairs (HSPs)

larger than 20 kb. The aligned sequences contained 16 mismatches and 3 indels representing an error rate of 0.0026% i.e. an accuracy of 99.9974%.

In addition, we compared 5283 HC genes from IWGSC RefSeq v1.0 7A pseudomolecule [only genes with no assembly gaps (Ns), aligning across 100% length] with the BAC-based sequence assembly of chromosome 7A [GYDLE 7A, (59)] and the 7B-7D pseudomolecules using a minimum HSP of 30 bp with a sensitivity of 25 and 1 mismatch (~97% identity). Over 97% of the genes were identical between both independent 7A assemblies whereas 97% and 95% of the genes were dissimilar to homeologs on the 7B and 7D pseudomolecules. We conclude that the assembly process was highly subgenome-specific with very little evidence of mis-assembly of homeologous sequences.

1.4 Assembly of organellar genome sequences and detection of NUPT/NUMT

A sample of 150 million filtered and trimmed reads (50 or more consecutive bases with quality scores 20 or higher) of the CS WGA sequencing data (30 million from each of the five types of libraries) was used for the assembly of complete plastid and mitochondrial genomes of CS. Initial mapping of reads to reference sequences (cpDNA: NC_002762, mtDNA: NC_036024) was followed by iterative resolution steps combining local realignment, consensus resolution, segmental reorganisation, gap filling, dynamic remapping of all reads, interactive edition and assembly visualization all performed with NUCLEAR version 3.2.16 (GYDLE Inc., Québec, Canada) and VISION version 2.6.22 (GYDLE Inc., Québec, Canada). Organelle annotation was done using alignments with features extracted and curated from the GenBank records of the reference sequences. The read mapping parameters used to recruit sequences into the organelle assembly were: -l 40 -s 38 -m 1 --min-pct-cov 80. This selected High-scoring Segment Pairs (HSP = gapless local alignment meeting the requested criteria) of 40 bases or more, containing 38 consecutive identities and contain at most 1 mismatch every 40 bases (97.5% local similarity) that were combined into alignments (combination of HSPs with possible gaps between them) covering at least 80% of the fragment's sequence. The resolving phases of the assembly used local realignments (HSP length $\geq$ 30, 16 consecutive identities, 90% local similarity, alignment covering 50 or more bases) prior to resolving the consensus sequence. The assembler always keeps track of the mapping score of each read (in particular which mapped reads are perfectly aligned and which are not), therefore regions consistently connected and covered by strongly aligned reads are not subject to misassembly influenced by additional divergent reads, such as those carrying sequencing errors or representing insertions into the nuclear genome (which in addition to being divergent and unconnected to the main assembly have also much lower coverage). The organelle assemblies were completed in two rounds of resolve when the resulting sequence had no gap and all its bases covered by perfectly-mapped reads at high coverage (chloroplast ~10,000X, mitochondria ~200X). An independent assembly using a different sample of 150 million reads resulted in exactly the same organelle sequences.

For NUPT/NUMT analysis, the organelle genome were compared to the pseudomolecule sequences using NUCLEAR alignment parameters: -l 40 -s 21 -m 4 --max-gap-size 200 --min-score-cov 100 -F 3 --hsp-extend-masked (HSP length $\geq$ 40, 21 consecutive identities, 90% local similarity, alignment length $\geq$ 100 bp, inter-HSP gap $\leq$ 200 bp, masking of low-complexity regions). Alignments with more relaxed parameters (HSP length $\geq$ 30, 16 consecutive identities, 90% local similarity) yielded similar NUPT/NUMT results (less than 15% more alignments).

2. Data resources integrated into IWGSC RefSeq v1.0

2.1 Chromosome conformation capture sequencing (Hi-C)

Three dimensional (3D) chromosome conformation capture sequencing data were generated for physical ordering and orienting of sequence super-scaffolds. Four Hi-C libraries were generated for bread wheat cv. CS using a plant-adapted TCC version (60) of the original Hi-C protocol (61) essentially as described previously (62). The libraries were sequenced on an Illumina HiSeq2500 instrument according to manufacturer's instructions. Hi-C data pre-processing and reads were assigned to HindIII fragments (62) using the CS whole-genome assembly (WGA) as the reference sequence.

2.2 Physical map de novo contig assembly

BAC-based physical maps were assembled for the 21 bread wheat chromosomes between 2008 and 2016. With the exception of 3B (63) chromosome- or chromosome arm-specific BAC libraries were developed at the Institute of Experimental Botany, Olomouc, Czech Republic (64) (see <http://olomouc.ueb.cas.cz/dna-libraries/cereals>). BAC fingerprints generated using either High Information Content Fingerprinting (HICF) with SNaPSHOT technology (65) or Whole Genome Profiling (WGPTM)(66) were assembled with standard FingerPrinted Contig (FPC) (67) or Linear Topological Contig (LTC) software (68) (Table S3 and S4), which were also used to select minimal tiling path (MTP) for sequencing.

The quality of physical map contigs was checked and improved using the following approaches: 1) application of LTC in chromosome arms primarily assembled by FPC; 2) integration with deletion, genetic and radiation hybrid maps; 3) alignment with Bionano optical maps (7AS, 7AL, 7DS). Detailed information of procedures applied for particular chromosome arms is provided in Table S3 and publications cited therein.

2.3 BAC clone / MTP sequencing

BAC clones selected from physical maps to represent complete or partial MTPs from 8 chromosomes and 2 chromosome arms were shotgun sequenced using various sequencing technologies: (i) Roche-454 sequencing of pooled BACs (chromosome 3B) (18); (ii) Illumina paired end sequencing of pooled BACs (1B, 7A, 7DS) (69); (iii) Illumina paired end shotgun sequencing of individually tagged multiplexed BACs (1A, 3D, 4AL, 6B, 7B, 7DL); Ion Torrent paired end sequencing (5BS). Reads were assembled with assembly algorithms suited to the data type (Table S6) and in some cases initial assemblies were improved by manual assessment and targeted sequencing to fill gaps (3B,1B); or incorporation of BAC-based or whole chromosome mate-pair sequences (1A, 3DL, 6B, 7A, 7B, 7DL). The BAC sequence data used for pseudomolecule assembly are summarized in Table S6.

2.4 Whole Genome Profiling (WGPTM) of MTP clones

In addition to the construction of physical maps for chromosomes 2B, 2D, 4B, 5BL, 5DL, 6A and 6B (see Table S4), WGPTM was used, following the same procedures (70,71), to profile minimal tile path BACs identified from wheat chromosome physical maps constructed previously by HICF from chromosome-specific BAC libraries (Table S5).

2.5 Bionano optical maps for Group 7 chromosomes

Bionano optical maps of chromosome arms 7AS, 7AL, 7BS, 7BL, 7DS and 7DL were constructed using the protocol described for chromosome 7DS (69). A total of 2.8 million arms for each chromosome, corresponding to 2.0 – 3.1 µg DNA per telosome, were purified by flow cytometric sorting from cv. Chinese Spring ditelosomic lines (72) with purities ranging from 80 to 87% (Table S7). The DNA from each chromosome arm was nicked with Nt.BspQI (GCTCTTC recognition site), the nick sites were labelled and the DNA was stained with

IrysPrep® Reagent Kit and IrysPrep® DNA Stain, respectively. The labelled molecules were analyzed on the Irys platform (Bionano Genomics, San Diego, USA). A total of 78 to 248 Gb data in fragments greater than 150 kb were collected per chromosome arm (corresponding to 192 and 689 arm equivalents, respectively) (Table S7) and used to assemble optical maps *de novo* using pairwise comparison of all single molecules and graph building in IrysSolve software (Bionano Genomics). A p-value threshold of $1e^{-10}$ was used during the pairwise assembly, $1e^{-11}$ for extension and refinement steps, and $1e^{-15}$ for a final refinement.

2.6 Radiation Hybrid (RH) Maps

Using methods described previously (73) a CS-specific whole genome RH panel was produced from crosses between male cultivars of the reference hexaploid wheat line CS and the tetraploid cultivar Altar 84 (*Triticum turgidum* L. ($2n = 4x = 28$; AABB). The F1 (RH1) plants grown from seeds harvested from the pollinated tetraploid wheat spikes represented individual independent RH panel members with quasi-pentaploid (AABBD) genotypes. A whole genome RH panel of 500 RH1 plants was generated (240 from 10Gy and 260 from 15Gy gamma irradiation treatment).

RH lines were selected for high throughput genotyping by initial characterization of DNA isolated from 10-Gy and 15-Gy RH lines for retention of seven genome-specific SSR markers (one per chromosome), i.e. lines with the lowest marker retention and with at least one break per chromosome. The 228 lines identified were then genotyped together with three CS and three Altar controls and two F1 controls (CS x AL; non-irradiated cross) on the Axiom TaBW280K SNP array containing 280,226 genome-wide markers (74). SNPs were assigned to six categories according to cluster patterns produced by the Affymetrix software: Polymorphic High Resolution (PHR), Off-Target Variants (OTV), Monomorphic High Resolution (MHR), No Minor Homozygous (NMH), Call Rate Below Threshold (CRBT) and Others. For subsequent analyses, only OTVs were considered as these markers can be used to detect presence-absence variations (PAVs) and correspond to probes showing four clusters, including a null allele. After removal of markers representing poor genotyping data rather than true deletions, 20,677 SNP markers were retained for use in WGRH mapping of CS chromosomes. A consequence of using the tetraploid line Altar to create the RH panel was the D-genome contained the highest number of SNP markers, 9,472 SNPs compared to the A- and B-genomes, with 4,501 and 6,704 markers, respectively.

RH maps were constructed with the software package Carthagene (75). RH markers were grouped using minimum two-point LOD scores of 15.0 and a threshold distance of 0.3 to reduce the possibility of pseudo-linkage of markers on RH groups. Subsequently, individual RH groups were used to construct maps using LOD scores of 15.0 and a threshold distance of 0.3, applying Carthagene's map validation commands 'greedy', 'annealing', 'flip' and 'polish' with default settings to improve marker ordering. The maps for each chromosome/ arm were finalized using genetic maps to orient the order of markers on scaffolds. Due to the pentaploid nature of the WGRH panel (AABBD), D- genome specific markers were mapped on D- chromosomes without the need for marker polymorphism. The requirement for A- and B- genome chromosome markers to be polymorphic between CS and Altar, without any null allele in CS, significantly reduced the numbers of usable markers on A- and B- genome chromosomes. After critical filtering and application of very stringent grouping criteria for RH mapping, 8,521 markers were used to construct RH maps for the 21 CS chromosomes. (Table S8).

2.7 Genetic mapping

Two genetic maps for sequence scaffold anchoring and ordering were constructed by genotyping-by-sequencing (GBS) of two populations developed from crosses between the parental lines Synthetic W7984 (M6) and Opata M85 (76). GBS reads were generated from the populations SynOpDH88, comprising 88 double haploids and SynOpRIL993, comprising 993 Recombinant Inbred Lines, after treatment with restriction enzymes *Pst*I-*Msp*I and *Pst*I-*Hpa*I, respectively, as previously described (77). After demultiplexing, sequence reads were assigned to samples using the barcode followed by the TCGA overhang sequence and then aligned with bwa-mem(v0.7.12) to the IWGSC RefSeqv1.0 assembly. SNPs were called with samtools mpileup with `-uv -t DP` parameters and bcftools call `-c -v` (bcftools v1.2). Only high quality GBS-SNP markers were retained (quality score >40; <15% missing data; >10% heterozygous calls; minor allele frequency >20%) and the markers were given identifiers indicating the CS pseudomolecule designation and chromosomal position, e.g. chr1A_13829065. GBS-SNP markers that aligned to the “Unanchored” scaffolds (ChrUn) were removed for subsequent analysis. MSTMap on the R/ASMap package for R was used for genetic linkage analysis. 7,832 from a total of 12,492 GBS-SNP markers were incorporated into the genetic linkage map for the SynOpDH88 population and 4,745 from a total of 8,698 high quality GBS-SNP markers were incorporated into the genetic linkage map for the SynOpRIL993 population.

For the analysis of recombination rate distribution along chromosomes a third genetic map based on 430 Single Seed Descent (SSD) individuals derived from a cross between CS and Renan (CsRe) was generated from genotyping data produced using the TaBW280K SNP array (18, 74). Polymorphic SNPs were filtered to discard markers that deviated significantly ($P \leq 0.01$) from the expected 1:1 ratio in a chi-square test, markers with missing or heterozygous data in parents and markers with more than 15% missing data. SNPs were divided into 21 sets corresponding to the 21 chromosomes, based on CSS-based in silico assignment (6) prior to constructing genetic maps using MSTMap (78) with the default parameters: population type: RIL6; distance function: Kosambi; cut-off: 10^{-11} ; map dist.: 15; map size: 2; missing threshold: 0.20; estimation before clustering: yes; detect bad data: yes; objective function: ML. Once robust framework maps were obtained for each chromosome, a whole-genome map was built using selected markers covering all chromosomes and genetic bins. This map was used to place additional markers comprising unassigned markers from previous chromosomal analyses, as well as markers that were excluded during the filtration phase. Recombination patterns along the chromosomes were evaluated using 10-Mb sliding windows (step 1 Mb) (Fig. S24).

2.8 Genome size estimation

The size of the bread wheat genome was estimated from flow cytometric measurements of the amount of nuclear DNA in hexaploid wheat (*Triticum aestivum* cv. Chinese Spring). The genotype that was sequenced was analyzed to prevent errors due to copy number variation and intraspecific differences in genome size. As flow cytometry measures relative fluorescence intensity of cell nuclei stained by a DNA fluorochrome, determination of absolute DNA amounts requires comparison with a standard of known genome size (79). *Pisum sativum* cv. Ctirad was used as the reference standard in this work and its genome size (4,225,354,542 bp) was determined using human male leukocyte DNA as primary reference standard and the size of the human genome assembly GRCh38.11 released by the Genome Reference Consortium on June 14, 2017 (<https://www.ncbi.nlm.nih.gov/grc/human>), which reports a total of 3,253,848,404 bases. This approach yielded a mean genome size of hexaploid wheat of 15,764,430,570 bp, or 15.76 Gbp, indicating 92% genome coverage by the current 14.5 Gbp IWGSC RefSeq v1.0 wheat genome assembly.

An independent genome size estimate was obtained on the basis of read coverage decomposition as described for *Eucalyptus globulus* and *E. grandis* genome size estimates (80). The IWGSC RefSeq v1.0 average insert coverage at 1-fold was estimated using a subset of well assembled scaffolds (N=346; manually checked). Scaffolds containing regions of deep read coverage were excluded and thus the estimate range can be considered conservative. The 1x coverage estimate gained in this way was also checked against estimates from all scaffolds of length > 1M bp and was found to be almost identical for the mean but much reduced in variation (due to the leverage of the under-represented regions). The mean coverage estimate was then used as a scaling factor enabling scaffold contributions to be calculated from the observed coverage. This was carried out for the mean and 95% confidence interval (CI) range estimates. The mean estimate is 15.4 Gbp (s.d. 0.2 Gbp) and 95% CI was between 14.85 and 15.82 Gbp.

3. Assembly of chromosome-like pseudomolecules

Pseudomolecules of the 21 wheat chromosomes were assembled by integrating chromosome-specific and genome-wide map and sequence-based resources with IWGSC WGA v0.2 in a highly iterative procedure summarized in Fig. S1.

3.1 Genetic anchoring of scaffolds

Contigs of the chromosome shotgun sequence (CSS) assembly of bread wheat cv. CS (6) were mapped to the WGA with BWA mem version 0.7.13 (81). Primary alignments with a mapping quality ≥ 30 were extracted with SAMtools (82) and imported into R for further analysis. Chromosome arm assignments from flow-sorted chromosome data (6) and POPSEQ chromosome assignments (7) were lifted from aligned CSS contigs to the whole-genome assembly using a majority rule. POPSEQ genetic positions (centiMorgan coordinates) of WGS scaffolds were obtained by calculating the arithmetic mean across all POPSEQ-anchored CSS contigs aligned to a WGA scaffold. Chromosome assignment by Hi-C of a WGA scaffold was determined by tabulating the POPSEQ chromosome assignments of all other scaffolds with Hi-C links to this scaffold and applying a majority rule. Aggregation of map information was done using functionalities of the R package ‘data.table’ (<http://CRAN.R-project.org/package=data.table>). The CS x Renan genetic map (83) provided confirmation for the order of scaffolds and super-scaffolds in IWGSC RefSeq v1.0.

3.2 Detecting scaffold misjoins

WGA scaffolds misjoining two or more unlinked sequences (“chimeras”) were identified by inspecting genetic marker and Hi-C link information and the alignments of CSS contigs to the WGA scaffolds. Putative chimeras carried a higher than average number of at least two types of assigned sequences (POPSEQ, Hi-C, or CSS) to chromosomes other than the major one. Scaffold breakpoints were identified with the CE method as implemented in the R package “breakpoint” (<https://cran.r-project.org/web/packages/breakpoint/index.html>). Putative chimeric scaffolds and detected breakpoints were visually inspected and confirmed chimeras were split.

3.3 Ordering and orienting scaffolds by Hi-C

Scaffolds that (i) were assigned to chromosomes both by Hi-C and by POPSEQ, and (ii) harbored at least 50 *HindIII* fragments were used for Hi-C map construction. Scaffolds were ordered using a custom R implementation (62) of the algorithm described in (84), and considering only Hi-C links within a single chromosome. If both scaffolds were anchored to the POPSEQ genetic map, the genetic distance between them was required to be ≤ 20 cM. The accuracy of scaffold order and the completeness of the Hi-C map were confirmed by alignment to the POPSEQ genetic map. POPSEQ-confirmed scaffolds were oriented by Hi-C as described in (62) using a bin size of 1 Mb.

3.4 Superscaffolding

Superscaffolding proceeded in two steps. First, Whole Genome Profiling (WGPTM) tags, BAC sequence assemblies and optical maps were aligned to WGA scaffolds and evaluated using automated scripts to find putative links between adjacent WGA scaffolds. In a second step, potential scaffold joins were manually inspected, and high-confidence joins were accepted.

3.5 Automated pipeline for superscaffolding using chromosome-specific resources

WGPTM tags: WGPTM tags were aligned to the WGA scaffolds with BWA-mem version 0.7.13 (81). Alignments with a mapping quality ≥ 30 were imported into R for further analysis. WGPTM tags were assigned to fingerprinted (FP) contigs based on the physical maps of wheat chromosomes or chromosome arms downloaded from URGI (https://urgi.versailles.inra.fr/download/iwgsc/Physical_maps/). WGPTM tags that aligned within 200 bp of each other were grouped together and considered as a single tag for further analyses. Alignments were aggregated at the level of FP contigs. FP contigs with at least five tags aligned to two different WGS scaffolds were considered as potential joins. Pair-wise orientations of adjacent WGA scaffolds were determined by calculating correlation coefficients between the positions of WGPTM tags in the physical maps and their alignments to WGA scaffolds. For subsequent manual inspection tags were organized at the BAC level so that a complete set of tags was available for each BAC. For each chromosome, the following procedure was followed: 1) Tags were aligned to the WGA scaffolds assigned to chromosome pseudomolecules, requiring perfect alignment; 2) Alignments were processed with a custom Python script that assigned BACs to locations based on tag positions; 3) The script produced a tiling of the scaffolds with BACs, and produced an output that included the physical contig and index of each BAC.

BAC sequences: BAC sequence contigs were split into 5 kb fragments with a step size of 1 kb and aligned to WGA scaffolds with BWA-mem version 0.7.15 (81). The alignment results were converted into BED format with BEDTools (85) and imported into R. Only alignments with a length of at least 3 kb and a mapping quality ≥ 30 were considered. Aligned fragments were grouped (i) at the level of individual BAC sequence scaffolds or contigs, if multiple BACs were sequenced together in pools; or (ii) at the level of fingerprinted contigs if individually barcoded BACs were sequenced and assembled. BAC groups with at least five fragments aligned to two or more different scaffolds were considered as potential joins. Pair-wise orientations of adjacent WGA scaffolds were inferred by calculating correlation coefficients between positions of BACs in the physical maps and their alignment to WGA scaffolds, or by comparing the positions of 5 kb sequence fragments in the BAC assemblies and their alignments to the WGA scaffolds.

Optical maps: Contigs of the optical maps for 7AS, 7AL, 7BS, 7BL and 7DS were aligned to the WGA scaffolds with IrysView RefAligner (<http://www.bionanogenomics.com>) and mapping results were imported into R. Optical contigs with alignments to two different WGA scaffolds with a confidence score ≥ 20 were considered as potential joins. Pair-wise orientations of adjacent WGA scaffolds were inferred by comparing the orientations of the optical map alignments as provided by RefAligner.

Manual inspection

The final stage in the IWGSC RefSeq v1.0 assembly was manual inspection of scaffolds and superscaffold links using alignment with physical maps, optical maps, and genetic maps to confirm or identify additional superscaffold joins and to identify false joins. Chimeric joins and orientation errors were identified using genetic maps by mapping sequences of genetic markers to updated pseudomolecule positions for each chromosome and generating plots of genetic distance to physical distance. Any obvious long-range outlier markers were identified and, if contained in a super-scaffold defined by manual joins, the super-scaffold join was invalidated. Such chimeric joins can be caused by errors in the physical map or misassemblies in the WGA scaffolds. Incorrect orientations of large scaffolds that could be identified in the genetic/physical alignments

were corrected. If differences in the genetic and physical order were ambiguous and not supported by multiple genetic maps, no change was made. The results of the manual inspection stage have been documented in the AGP (“A Golden Path”) files for the chromosomes.

3.6 Gap filling using assembled BAC sequences

Chromosome specific BAC sequence assemblies available for 16 chromosome arms (1AS-L, 1BS-L, 3DS-L, 4AL, 5BS, 6BS-L, 7AS-L, 7BS-L, 7DS-L) and 3B were used to fill inter-scaffold gaps between adjacent WGA scaffolds. The strategy used sets of ISBP markers (13) (Insertion Site Based Polymorphism) shared between two scaffolds to identify overlapping scaffolds from two different assemblies, i.e. WGA vs BAC-based sequence assembly, to avoid performing a computationally expensive all-against-all alignment. TE models were identified for all WGA scaffolds using ClariTE (86) and the PERL script junctionDesigner.pl used to extract 150 bp ISBP markers (75 bp on each side of the junction between a given TE and its insertion site). Mapping 5,031,032 ISBPs with BWA (82) and requiring a perfect match to the BAC-based scaffold assemblies found 1,876,163 ISBPs that were then used to identify potential scaffold overlaps based on pairs of scaffolds sharing three or more ISBPs and residing on the same chromosome. 1,251 groups (accounting for a total of 2,233 scaffolds) of potentially overlapping scaffolds were identified by the PERL script slalomer.pl developed to perform this step. The scaffolds sharing common ISBPs were mapped fully with Nucmer (87) (cutoff 99% nucleotide identity over at least 5 kb) and the alignments were all manually inspected to keep only scaffolds that truly overlap, i.e. share identical sequences encompassing their extremities. The order of scaffolds along the WGA-based pseudomolecules was used as a template to identify potential joins for superscaffolds and to fill gaps between neighbors. For gap filling, the exact coordinates of overlapping segments were retrieved from the Nucmer output using the PERL script gapfiller.pl whose purpose is to update the AGP file (describing the positions of scaffolds on pseudomolecules) based on the Nucmer alignments. Gaps were filled if they met the criteria: (i) shared at least 5 kb contiguous aligned sequences with at least 99% nucleotide identity; (ii) alignments must include scaffold ends (while allowing unaligned ends of 10 kb at maximum); (iii) the difference in length between two aligned sequences does not exceed 10% of the length of the larger one. Using these criteria, 183 gaps were filled in the WGA-based pseudomolecules with 181 scaffolds originating from BAC-based sequencing projects.

4. Genome annotation

4.1 Centromere positioning

Centromeres were identified and positioned using published ChIP-seq data for CENH3 (19). Raw reads were downloaded from EMBL ENA (SRR1686799). After adapter trimming with cutadapt (88), reads were mapped to the CS pseudomolecules and unassigned scaffolds (“chrUn”) with BWA-MEM (arXiv:1303.3997). Alignments were converted to BAM format with SAMtools (82) and sorted with Novosort (<http://www.novocraft.com>). Uniquely mapped (MAPQ $\geq$ 20), non-duplicated reads were extracted with SAMtools and counted in non-overlapping 100 kb windows with BEDTools (85). The resultant count tables were imported into the R statistical environment for visualization. To define approximate centromere boundaries, read counts in regions closely to visually defined ChIP-seq peaks were manually inspected. An enriched region was called if at least three (two for chromosome 7B) consecutive 100 kb bins had more than three times the genomic average of ChIP-seq reads (320 reads per bin) mapped to them. The consecutiveness was not considered for bins of “chrUn”. Disjoint CENH3-enriched regions separated by fewer than 500 kb were merged.

4.2 Annotation of repetitive DNA

Transposable elements (TEs) were modeled using CLARITE and a wheat TE reference library named ClariTeRep as previously described (86). Briefly, CLARITE provides an annotation of TEs by analyzing the raw similarity search results, applying a defragmentation step, and reconstructing the nested insertion patterns.

4.3 Gene annotation

A federated gene prediction strategy (Fig. S8) was developed for the automated gene calling on IWGSC RefSeq v1.0 based on established and proven gene prediction pipelines and protocols at INRA-GDEC Clermont Ferrand, France, Helmholtz Zentrum, Munich, Germany (PGSB) and the Earlham Institute, UK.

4.3.1 Gene modeling using the TriAnnot pipeline

TriAnnot, a modular, customizable, and parallelized pipeline for plant genome sequence annotation (89), originally developed for the Triticeae and used for the annotation of wheat chromosome 3B (18) was used to model protein-coding genes in the wheat genome sequence. TriAnnot gene annotation of IWGSC RefSeq v1.0 combined the parallel annotation of individual RefSeq_v1.0 scaffolds and BAC-derived sequences used to fill gaps. Annotated features were positioned subsequently on the pseudomolecules using a custom script. The annotation process comprised four main steps: (1) TE repeat masking. TEs annotated with CLARITE were masked using RepeatMasker (cross_match engine, cutoff 250 - (90)). (2) Similarity search against transcripts and related proteomes. A similarity search was performed on the TE-masked sequence using BLAST (91) with the following datasets: wheat cv. CS transcripts (92); PacBio full-length transcripts of cv. Xiaoyan 81 (93); all available Triticeae ESTs and full-length cDNAs (EMBL release 126; tritfdb at <http://www.psc.riken.jp>); non-redundant proteomes of rice (UniProt Release 2016_03 and (16)), maize (UniProt Release 2016_03 and phytozome v11), sorghum (UniProt Release 2016_03 and phytozome v11), barley (UniProt Release 2016_03 and High Confidence proteins (94)), *Aegilops tauschii* (UniProt Release 2016_03 and (54)) and *Triticum urartu* (UniProt Release 2016_03 and (53)). Spliced alignments of BLAST hits generated by EXONERATE (95) were used to check gene models predicted in step#3 and for display as individual tracks in a genome browser. (3) Gene modeling. *Ab-initio* gene models were predicted using two gene finders trained with a wheat-specific matrix: FGeneSH (SOFTBERRY, <http://linux1.softberry.com/berry.phtml>) and AUGUSTUS (96). Evidence-driven gene model prediction was also computed with two different modules implemented in TriAnnot. The first module is based on BLASTX-EXONERATE spliced alignments of protein sequences from rice (16), Brachypodium (phytozome v11), sorghum (phytozome v11), maize (phytozome v11) and barley (High Confidence proteins (94)) and incorporated iterative extension over a range of ca. 200 codons to identify an in-frame start and stop codon in case of missing start and/or stop codons in the derived CDS model. Models with no start and/or stop codons are flagged as pseudogenes. The second module (named SIMsearch, derived from FPGP pipeline (97)) focused on similarity with wheat transcripts (full-length cDNAs (tritfdb at <http://www.psc.riken.jp/english/> and EMBL release 126) and RNA-Seq-derived transcripts (92)) to predict the CDS borders by comparison with known proteins from related Poaceae. (4) Selection of the best gene model at every locus. At each locus, the five gene models delivered by TriAnnot (derived from two *ab-initio* gene finders, BLASTX-EXONERATE, and two SIMsearch-derived models) were evaluated using a scoring system that considered the metrics of the alignment of each gene model against known proteins (taking into account percentage identity and coverage). The highest scoring models were retained and subjected to two further steps. First, models with a canonical structure but less than 70% similarity across the length of the best

BLAST hit were classified as pseudogenes (i.e. suggesting a large truncation). In the second step gene models that were ambiguous or doubtful were discarded based on comparisons between TriAnnot gene models, TGACv1 gene models (8), CSS gene models (7), and RNA-Seq derived transcripts (92) mapped on the RefSeq_v1.0 pseudomolecules with Gmap (98) made with cuffcompare (99) and InterProscan5. All gene models were removed that were not supported by evidence (i.e. *ab-initio* only; no similarity with proteins, transcripts or Interpro domain) and that do not correspond to a gene model from the CSS or TGACv1 datasets. The final Triannot output identified 180,270 gene models comprising 65,884 high confidence genes, 41,342 low confidence genes and 73,044 pseudogenes.

4.3.2 PGSB gene prediction pipeline

The annotation process of the PGSB pipeline comprised two main steps: (1) Mapping: The PGSB gene annotation pipeline combined information from mapping splice site-aware alignments with reference proteins, RNA-Seq based gene structure predictions, alignment of IsoSeq reads and alignments of full-length cDNAs (flcDNAs). First, publicly available RNA-Seq data sets were mapped with RNA-Seq read aligner Hisat2 (100) (version 2.0.4, arguments: --dta) to predict transcript structures on the wheat assembly (E-MTAB-2137, SRP045409, ERP004714/URGI, E-MTAB-1729, PRJEB15048 and PRJEB23081). Alignment files were sorted and converted into BAM files using Samtools (82) (version 1.3, arguments: sort -@ 8 -T). The BAM files from each data set were then combined into a single alignment file using Bamtools (101) (version 2.4.1, arguments: merge) and mapped reads were assembled into transcript sequences with StringTie (102) (version 1.2.3, arguments: -m 150 -t -f 0.3). StringTie was configured to include only isoforms expressed at levels at least as high as 30 % of the main isoform and to include only isoforms with a minimum length of 150 bp. Predicted protein sequences from five reference plant species (*A. thaliana*, *B. distachyon*, *O. sativa*, *S. bicolor* and *S. italica*), as well as all available complete Triticeae protein sequences (downloaded from Uniprot at 10/05/16) (103) were then aligned against the chromosome pseudomolecules using the splice-aware alignment software GenomeThreader (104) (version 1.6.6; parameters used: -startcodon -finalstopcodon -species rice -gcmincoverage 70 -prseedlength 7 -prhdist 4 -force). Mapping was performed for each chromosome separately to reduce running time and the resulting gene structure predictions were then combined using a custom python script. In addition, the GeMoMa tool (105) (version 1.4.2) was used with annotations of protein-coding genes from *A. thaliana* to generate high-quality homologous gene predictions using specifically the conservation of intron splice sites for protein detection. Full length cDNA sequences from wheat and barley (106), together with publicly available IsoSeq reads (8) were also included in the prediction pipeline. GMAP (98) (Version 2016-06-30; parameter: -f gff3_gene) was used to align sequences to the wheat assembly and the resulting transcript predictions were combined using a custom python script. (2) Prediction and selection of open-reading frames (ORFs). All predicted sequences from protein alignments, RNA-Seq mapping and long transcript sequence mapping were combined into a single structure file using Cuffcompare (107) (version 2.2.1). StringTie (version 1.2.3) was used to remove redundant transcript sequences and to merge fragments in the combined structure file. Using a script from Transdecoder suite (<https://github.com/TransDecoder/TransDecoder/releases>, version 3.0.0, `cufflinks_gtf_genome_to_cdna_fasta.pl`) transcript sequences were extracted and stored in a single fasta file. Similarly, Transdecoder longorfs (arguments: -p 0) was used to predict potential protein translations for each transcript sequence. Thereby, translations were forced to include the left-most methionine as a start codon if available. All potential protein sequences were then

mapped against a set of validated protein sequences from angiosperms (downloaded from Uniprot, 08/03/16) using protein BLAST (91) (version 2.3.0, -max_target_seqs 1 -evaluate 1e-05) and scanned for the presence of protein domains using Hmmer3 (108) (version 3.1b2). A table with BLAST results and protein domain information was then used by Transdecoder predict to select a single best translation for each transcript sequence. Finally, gene predictions from Transdecoder were merged with gene predictions from protein alignments, removing protein sequences that were fully included in another protein sequence at each locus to produce a non-redundant set of predicted protein sequences.

4.3.3 Production of IWGSC RefSeq Annotation v1.0 by gene model consolidation and integration

The two independent gene prediction pipelines generated alternative gene models for many loci. To evaluate and select the “best” representative gene model for each locus a rule-based approach was developed to use a combination of supporting evidence (e.g. high quality transcriptome data, protein homology) and intrinsic gene characteristics (e.g. CDS length, proportion of canonical introns). To enable an unbiased assessment of the alternative gene predictions, an independent set of quality filtered PacBio transcripts, RNA-Seq and cross species protein alignments to the genome assembly was generated (Tables S13, S14 and S15). RNA-Seq reads were assembled using three alternative approaches (Cufflinks, CLASS2 and StringTie) and integrated into a single set of transcripts using Mikado (109) (Table S16). A set of high confidence splice junctions was identified from RNA-Seq alignments using Portcullis (110). The gene model selection approach used heuristics including splicing agreement with wheat transcripts and cross-species BLAST alignment coverage. To measure the congruence between a gene model and its supporting evidence an average F1 score was calculated (using nucleotide, junction and exon F1) from the aligned PacBio transcripts, Mikado transcript assemblies and proteins. Homology support was indicated by determining query and target coverage using BLAST alignment of gene models to a plant protein database. Mikado was used to cluster genes from the two pipelines into loci, assign an overall score to each gene model and select a representative (highest scoring) gene model for each locus. Scoring utilized the externally generated F1 and homology scores in conjunction with selected Mikado metrics (Table S17). Gene models located on the opposite strand or intronic to another gene with no supporting evidence in the quality filtered data were excluded, the majority (3,520) being single exon models based on incorrectly aligned transcripts/proteins. Models with extremely long introns (>75 Kb) or exons (>30 kb) that represented alignment artefacts (233) were also removed.

Following selection of the representative gene model, Mikado evaluated overlapping transcripts to identify high quality alternative gene models (splice variants) based on defined criteria. Gene models that met the following requirements were retained:

- assigned to a single locus (to avoid selecting fusion transcripts)
- contain novel splicing (class codes j,J,G and h as determined by Mikado compare)
- cDNA overlap > 0.6 (relative to representative)
- CDS overlap > 0.6 (relative to representative)
- min score percentage > 0.5 (relative to representative)
- proportion of canonical introns = 1
- max exon length < 10,000 bp
- max intron length < 10,000 bp
- contain only verified introns (i.e. pass portcullis junction filter)
- do not contain any retained intron event, i.e. the CDS does not end within a CDS intron of another transcript in the locus

Gene models were refined by adding UTR features based on supporting cDNA, PacBio transcripts or RNA-Seq assemblies. PacBio transcripts and cDNAs were aligned to the genome with GMAP (Genome Mapping and Alignment Program). After filtering to reduce errors originating from alignment artefacts, chimeric transcripts and over-extended UTRs the Mikado transcript assemblies were assembled by PASA (The Program to Assemble Spliced Alignments) with the aligned cDNAs and PacBio transcripts. PASA takes existing annotations and compares them to the PASA assembly clusters to identify potential updates. UTR additions only (PASA status 13 and 14) were selected and the corresponding gene models updated ensuring CDS features remained unchanged. A second round of PASA UTR updates was performed using the Mikado PacBio assemblies. Following PASA additional checks were carried out and changes implemented to remove UTRs from models lacking start or stop codons, from models with greater than four 5' or 3' UTR features, or any model with UTR introns over 30kb (common alignment artefacts). UTR start and end positions were standardized across transcripts sharing UTR exons.

Comparison of the long transcripts generated by PacBio sequencing (ENA PRJEB15048) with the annotated gene models was used as a measure of testing annotation quality, noting that not all PacBio reads represent full length transcripts and the presence of retained introns and alignment errors can reduce the agreement between aligned data sets. Comparing TriAnnot, PGSB and RefSeq Annotation v1.1 gene models with PacBio transcripts aligned to the IWGSC RefSeq v1.0 sequence assembly showed that the integrated data in RefSeq Annotation v1.1 represents a higher proportion of more complete gene models than either of the outputs from the two automated pipelines alone. RefSeq Annotation v1.1 also provides more complete coverage of the bread wheat gene space than previous wheat genome annotations released in 2014 (IWGSC CSS, (6)) or 2017 (TGACv1, (8)), as judged by greater than 90% coverage of 1.5 million PacBio transcripts.

Over 80% of the genes defined as high confidence by either the TriAnnot or the PGSB pipelines overlapped a gene in the alternative annotation, with a substantial percentage of genes showing highly similar structures (67% TriAnnot, 48% PGSB). Less agreement was found between gene models classified as low confidence, in total combining low and high confidence genes 66,064 (37%) were specific to TriAnnot and 89,988 genes (44%) specific to PGSB. Differences between the two annotations reflect the challenge of annotating a transcriptionally complex polyploid species as well as differences in the evidence datasets utilized and individual characteristics of the annotation pipelines. The final RefSeq Annotation v1.1 (see Methods 4.3.5) incorporates 205,643 PGSB gene models (432,097 transcripts) and 180,270 TriAnnot gene models (180,270 transcripts) and removes aberrant transcripts that originate from incorrect alignments.

Overall 98% of RefSeq v1.1 genes aligned to the TGACv1 assembly and 78% to the IWGSC CSS using GMAP (version 20160923) with 95% identity and 80% coverage. Using BLAST alignment to assess the representation of previous wheat proteins in the RefSeq Annotation v1.1, a large proportion of high confidence genes identified in earlier studies (6) were found (TGAC v1 87%, IWGSC CSS 74%, minimum of 95% identity, 75% coverage), with the majority also classified as high confidence (81% of TGAC v1.0, 68% of IWGSC CSS). A more detailed comparison of RefSeq Annotation v1.1 gene structures with the previous gene models aligned to the IWGSC RefSeq v1.0 assembly revealed greater agreement with the TGACv1 genes (Fig. S11), as expected given the greater contiguity of the TGACv1.0 assembly compared with the IWGSC CSS. Overall, high levels of similarity were found between high confidence genes annotated in RefSeq annotation v1.1 TGAC v1 (identical gene structures for >50% genes and highly similar structures for a further 33%). Little agreement was seen for low confidence genes

with either the TGACv1 or IWGSC CSS annotation. By definition, low confidence genes have reduced homology support to help guide annotation and contain gene fragments and pseudogenes that are likely to show greater variability in gene structure between alternative annotation pipelines. A more detailed classification of these genes and other intergenic transcribed sequences will be a key part of future wheat annotation releases.

4.3.4 Gene confidence assignment

The IWGSC RefSeq Annotation v1.0 gene models were classified into HC (High-Confidence) or LC (Low-Confidence) genes using predictions based on their sequence homology to gene products in public databases [**UniPoa**: Annotated *Poaceae* proteins (SwissProt & trEMBL); Sequences were downloaded from Uniprot (Feb 2017, <https://www.ebi.ac.uk/uniprot>) and further filtered for complete sequences with start and stop codons; **UniMag**: Validated Magnoliophyta proteins (SwissProt); Sequences were downloaded from UniProt (Feb 2017) and further filtered for complete sequences with start and stop codons] and/or known repetitive sequences in **TREP** [The database of hypothetical proteins (“PTREP”) deduced from the non-redundant database of transposable elements (TE) within the TREP database (<http://botserv2.uzh.ch/kelldata/trep-db/index.html>)]. PTREP was useful for the identification of divergent TEs having no significant similarity at the DNA level. **HC genes** have complete gene models with very good sequence homology to experimentally verified plant proteins from SwissProt (HC1) or with very good sequence homology to annotated *Poaceae* proteins in SwissProt or trEMBL, and no good hits in the transposon database TREP (HC2). “Complete” gene models are defined as containing both start and stop codons (which do not necessarily represent the biologically “true” start and stop sites). **LC genes** have incomplete gene models with very good sequence homology to experimentally verified plant proteins from SwissProt (LC1), complete gene models with no significant homology to TREP, UniPoa or Unimag databases (see definition below) (LC2) or incomplete gene models with very good sequence homology to any annotated (also automatically) *Poaceae* protein in SwissProt or trEMBL but no good hits in the TREP database (LC1). Incomplete gene models with no significant homology to any of the three databases were classified as LC3. TREP genes have incomplete gene models with no significant homology to experimentally verified plant proteins from SwissProt, but good homology to TREP entries. For each transcript in the initial data set, sequence homology to each of the three databases was determined (BLASTP, e-value cutoff 10^{-10}) and classified according to the respective best hit: step 1: Alignment with maximal (and at least 90%) overlap between query and subject sequence for the two protein databases; alignment with maximal (and at least 90%) query coverage for TREP database; step 2: classification into distinct confidence groups and subgroups according to the following rules:

```
# Primary confidence class
unimag & complete <- "HC"
!unimag & (!trep & unipoa) & complete <- "HC"
(unimag | (!trep & unipoa)) & !complete <- "LC"
!unimag & !trep & !unipoa & complete <- "LC"
!(unimag & complete) & trep <- "TREP"

# Secondary confidence class
unimag & complete <- "HC1"
!unimag & (!trep & unipoa) & complete <- "HC2"
(unimag | (!trep & unipoa)) & !complete <- "LC1"
!unimag & !trep & !unipoa & complete <- "LC2"
!unimag & !trep & !unipoa & !complete <- "LC3"
!(unimag & complete) & trep <- "TREP"
```

The confidence group of each transcript is annotated in the respective GFF/GTF file. In cases where transcripts with different confidence classes are present at a single gene locus, the “best” confidence label of all transcripts present was assigned to the respective locus.

4.3.5 Annotation update and integration of manually curated genes

The automated annotation of eight gene families (CBF, NLR, WAK, RK, AATset9, Prolamin, and PPR) was updated to incorporate manual annotation. In addition, 3,318 HC genes were re-classified as LC-TE, following identification of TE elements from the functional annotation. The semi-automated process developed to integrate the manual curation data used a Python script based on common tools (GenomeTool, GFFCompare, pyBEDTools) to manage the integration process (Fig. S9). Gffcompare was used initially to check that the curated gene sets did not overlap each other and to identify the RefSeq Annotation v1.0 models to be updated. 5 types of modification were identified depending on the overlap between manually curated genes and the automated annotation (Fig. S9). The gene model updates included in RefSeq Annotation v1.1 are summarized in Table S12 and comprise 3,685 manually curated genes, of which 528 were not present in RefSeq Annotation v1.0 and 3,020 were replaced with a single manually annotated gene. The manually annotated genes have been designated HC3. In addition, 354 LC genes were removed from the LC gene set (161,537 genes) after gffcomp was used to identify LC genes overlapping the 3,685 manually curated genes. The final IWGSC Annotation v1.1 HC gene set (IWGSC_V1.1_HC.gff3) contains 107,891 genes including 528 genes not annotated in the v1.0 HC gene-set. The v1.1-LC gene set (IWGSC_V1.1_LC.gff3) contains 161,537 models, including 3,318 genes functionally designated TE that were re-classified. The characteristics of the predicted gene models are summarized in Table S18. To minimize the impact of artefacts arising from RNA-Seq alignment and assembly that can artificially inflate estimates of alternative splicing, a conservative approach was taken that excluded splice variants with non-canonical splicing, or with fragmented or truncated coding sequences. As a result, 133,745 distinct transcripts were associated with HC genes (average 1.24 transcripts per gene) with 16% of HC genes having an annotated splice variant.

4.4 Evaluation of the flanking / promoter regions of wheat genes

To assess the quality of the RefSeq v1.0 assembly in the immediate vicinity of genes, which is important for analysis of promoters and regulatory elements, the 5' upstream and 3' downstream flanking regions were compared in 269,428 HC and LC gene models of RefSeq v1.0, 217,907 HC and LC gene models from TGACv1.0 (8) and 99,386 HC gene models of the CSS assembly (6). The “bedtools flank” command (85) was used to extend gene coordinates from 1kb to 10kb upstream and downstream of the start/stop codon and corresponding nucleotide sequences were extracted using “bedtools getfasta”. In subsequent steps, nucleotide sequences shorter than the predefined length or containing any “N” sequences were discarded. The proportions of retained 5' upstream and 3' downstream sequences are shown in (Fig. S12).

4.5 Pseudogenes

4.5.1 Pseudogene annotation

Two complementary approaches were used to identify pseudogenes in IWGSC RefSeq v1.0: (1) *De novo* genome-wide identification of pseudogenes derived from HC genes: a specialized pseudogene detection and analysis pipeline was employed to identify and classify pseudogenes in a genome-wide manner, independently from elements annotated in the gene prediction process. Using a homology-based approach, pseudogenes derived from HC genes were identified by determining internal stop codons and other structural features that disrupt the gene models. This genome-wide analysis identified 288,839 pseudogenes, including 10,440 predictions overlapping with pseudogene annotations in the LC gene set. (2) Identification of disrupted or truncated gene

models in the LC gene set: Pseudogenes in the LC gene set were identified by detecting internal stop codons in the LC gene ‘open reading frames (ORF)’. 25,419 LC pseudogene predictions were identified of which 10,440 overlapped with the genome-wide pseudogene set. A significant overlap was defined as >96% coverage of the annotated pseudogene with the LC gene model, >15% coverage of the LC gene model with the annotated pseudogene and the presence of a premature stop codon (PTC) in the ORF of the LC gene model. Potential pseudogenes in the LC gene set were kept as LC genes in the v1.1 gene annotation, with a note indicating their potential pseudogene role.

4.5.2 Pseudogene classification

Intron sequences were used to classify pseudogenes as ‘duplicated’ or ‘retroposed’ where fragmentation did not disrupt splice sites. For the intron loss/retention criterion, six pseudogene classes were defined: (1) ‘duplicated’ pseudogenes still containing introns at each covered splice site; (2) ‘retroposed’ or ‘processed’ pseudogenes having lost all introns; (3) ‘chimeric’ pseudogenes having both retained and lost introns; (4) ‘single-exon gene’ pseudogenes are duplicates of genes with only one exon; (5) ‘single-exon splice variant’ pseudogenes are duplicates of isoforms with only one exon; (6) ‘fragmented’ pseudogenes are short gene fragments not sufficiently covering a gene splice site. A splice site was only considered covered if at least 10 base pairs of the exons on either side were present in the duplicate. The gap had to be at least 30 base pairs long to be considered a duplicated intron.

288,839 pseudogenes and gene fragments were identified in the bread wheat genome (Table S21, S22). 48,619 (16.8%) represent at least 80% of the CDS of a parent gene locus (high-coverage pseudogenes). Of these, 42.1% retained their exon-intron structure and are thus classified as “duplicated”. The distribution of pseudogenes on the 21 chromosomes of bread wheat followed the distribution of genes (Fig. S10) and, whilst the three sub-genomes have similar pseudogene content, fewer and less degenerated pseudogenes were found on the D sub-genome compared to sub-genomes A and B (Table S22).

4.6 Annotation of miRNA and tRNA

A two-step homology-based *in silico* method was applied to each chromosome pseudomolecule separately for miRNA identification. Chromosome sequences were compared with all known plant mature miRNA sequences retrieved from miRBase (v21, June 2014) (111) allowing at most a single base mismatch between mature miRNA and chromosome sequences using the in-house script SUMirFind (<https://github.com/hikmetbudak/miRNA-annotation/blob/master/SUMirFind.pl>) (112). Candidate precursors of mature miRNAs were extracted and assessed for their secondary structure formation features with UNAFold v3.8 (113) using the in-house script, SUMirFold (<https://github.com/hikmetbudak/miRNA-annotation/blob/master/SUMirFold.pl>). Additional criteria were applied for the precursor sequences which satisfied the folding evaluation (114). Finally, false-positives were eliminated from the results of SUMirFold by using SUMirScreen (<https://github.com/hikmetbudak/miRNA-annotation/blob/master/SUMirScreen.py>) and the genomic distribution of the final miRNA list was obtained with SUMirLocate python scripts (<https://github.com/hikmetbudak/miRNA-annotation/blob/master/SUMirLocate.py>).

tRNA annotation was performed as described previously (115) with tRNAscan-SE software (version 1.3.1), which uses a series of algorithms to predict tRNA sequences from the genome and examine their secondary structure features (116).

4.7 Functional annotation of genes

4.7.1 Automated functional annotation

Functions were assigned to annotated genes with the AHRD tool (Automated Assignment of Human Readable Descriptions, <https://github.com/groupschoof/AHRD>, version 3.3.3) to generate more informative annotation than can be derived for the Triticeae where the UniProt databases contain large numbers of uncharacterized genes and a best BLAST hit approach alone would return about 30% of 'unknown' descriptions. BLAST hits against the following databases were used as AHRD input: Swiss-Prot (version 02-15-17), *Arabidopsis* Araport 11 (version 201606) and custom made TrEMBL (version 02-15-17) *Viridiplantae* subset. The AHRD parameters re provided with the AHRD output files. Gene ontology, Pfam and InterPro assignments were derived from InterProScan Runs (version 5.23-62.0). A postprocessing filter (disTEG: distinction between TEs and Genes) on the AHRD and domain descriptions tagged the genes as either G for canonical genes with a nonTE description, TE for obvious transposons, TE? for potential transposons or U for unknowns, without any description. The numbers of genes assigned to these four categories in RefSeq annotation v1.0 are shown in Table S19. The HC genes with TE tags were moved subsequently from the HC to the LC category in RefSeq annotation v1.1.

4.7.2 Orthology-based ontology annotation of wheat genes

In addition to automated functional annotation, inferred orthologous relationships determined by phylogenomics, as described below, were used to transfer automatic and experimentally validated annotations from orthologous genes to the respective genes in the bread wheat genome. Gene Ontology (GO; (117)), Plant Ontology (PO; (118)) and Plant Trait Ontology (TO; (119)) term annotations were obtained and pooled from Gene Ontology (<http://geneontology.org/gene-associations>), TAIR (<https://www.arabidopsis.org/>), and Gramene (<ftp://ftp.gramene.org/pub/gramene/release52/data/ontology/>) resources. Gene identifiers were mapped to public resources using the UniProtKB mapping table (ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/). The pfam2GO mapping table available from the Gene Ontology resource (<http://geneontology.org/external2go/pfam2go>) was also employed to transfer GO terms based on the inferred domain architectures. The source evidence classes of the annotated, orthologous genes were translated into target evidence codes of wheat genes as follows: a) automatic annotations: IEA (Inferred by Electronic Annotation), b) experimental and reviewed computational analyses (for full list of evidence codes in these categories see <http://www.geneontology.org/page/guide-go-evidence-codes>): e.g. EXP (Inferred from Experiment) and e.g. RCA (Reviewed Computational Analysis), ISO (Inferred by Sequence Orthology) c) pfam2GO: ISM (Inferred from Sequence Model). The comprehensive ontology annotation including source term publication references and gene identifiers in GAF2 format comprising more than 5 million annotations is available from Table S20.

5. Transcriptome and gene co-expression network analysis

5.1 Expression analysis

The analysis used 529 previously published (described in Table S30) and 321 new RNAseq samples from six studies (Table S31) (23). Metadata was assigned as in (120) for tissue, age, variety and stress (Database S1). The tissues from where RNA was extracted for the RNA-Seq were defined as low level tissues. To facilitate analyses, we grouped these low level tissues based on their biological origin into a hierarchical structure with the highest hierarchy represented by the 4 high level tissues Roots, Leaves, Spike, Grain. Given the relatively large number of low level tissues (60), an intermediate level of tissues comprising 32 factors (median 12 replicates per factor) was defined and used for this study (Table S32). The 850 samples were mapped to the

RefSeqv1.0+UTR transcriptome (default parameters) with Kallisto v0.42.3 (121) and transcripts per million (tpm) were calculated per gene using tximport v1.2.0. To determine a minimum threshold required for a gene to be considered expressed, the 1% most highly expressed samples for each gene were examined. This criterion allowed tissue-specific expression to be accounted for, while still requiring at least eight samples to have a minimum expression abundance. This analysis identified a clustering of very low expressed genes between 0 and 0.5 tpm (corresponding to less than 10 reads/counts per gene, Fig. S30). Subsequent application of the criteria that over 1% of samples all required expression values over 0.5 tpm for a gene to be considered expressed across the 850 samples removed some of the expected low-level non-target homeolog mapping (which occurs when homeologs are identical across a long stretch of sequence), while retaining more lowly expressed genes. By applying this criterion to the 850 samples, expression was detected for 94,114 high confidence (HC) genes and 77,920 low confidence (LC) genes, corresponding to 84.95 % and 49.07 % of HC and LC genes, respectively. For the PCA analysis, tpm values for the 94,114 expressed HC genes were transformed ($\log_2(\text{tpm}+1)$) and the PCA calculated using the `prcomp` function in R.

Distribution of gene expression: The number of HC and LC genes expressed in each intermediate tissue (Database S1) were defined as genes for which three samples had expression values over 0.5 tpm. The tpm abundance of each gene was then averaged across its expressed tissues to determine an average expression of the gene across the intermediate tissues. The average expression values were assigned to bins for visualization. The values for all HC and LC genes are presented in Fig. S31.

Expression breadth: The number of expressed genes was determined for each tissue using the same criteria as for the distribution of gene expression (three samples with expression values over 0.5 tpm). Using this criterion, 89,444 genes were found to be expressed in at least one intermediate tissue (Table S32). Each chromosome was segmented into bins corresponding to 1% of the physical size and genes were assigned to the bins based on their physical position to generate comparable scales across chromosomes. Using the genomic compartment breakpoints (Fig. S25, Table S29) to delimit the R1, R2a, C, R2b and R3 regions for each scaled chromosome, a representative artificial chromosome was produced with the average breakpoint positions to define each genomic compartment. The original bins were rescaled to the artificial chromosomes with the following formula:

$$scaledBin_i = \left[\left(\frac{originalBin_i - originalBinStart_p}{originalBinSize_p} \right) * averageBinSize_p \right] + averageBinStart_p$$

where i correspond to the 1% bin and $i \in p$, where p is the partition group. The average of each bin was calculated and further grouped into bins of 2% for display purposes (Fig. S36).

Tissue exclusivity: To determine if a gene was expressed within a given tissue, the average expression of each gene was calculated for each of the 32 intermediate tissues (Table S32). Tissue exclusivity was defined as those genes whose average expression was over 0.5 tpm in a single intermediate tissue, and therefore had average expression < 0.5 tpm for all other intermediate tissues.

5.2 Co-expression analysis (WGCNA)

The WGCNA R package (122) was used to build the co-expression network with HC genes expressed > 0.5 tpm in $> 1\%$ of the 850 samples (94,114 genes). The expression count of each

gene was normalized using variance stabilizing transformation from DESeq2 (123) to eliminate differences in sequencing depth between studies. A soft power threshold of 6 was used because it was the first power to exceed a scale-free topology fit index of 0.9. A signed hybrid network was constructed blockwise in three blocks using the function `blockwiseModules` with a maximum block size of 46,000 genes and a biweight mid-correlation “bicor” with `maxPOutliers` = 0.05. The `blockwiseModules` function calculated the topographical overlap matrices (TOM) using `TOMType` = “unsigned” and the minimum module size was set to 30. Similar modules were merged by the parameter `mergeCutHeight`=0.15. Modules were tested for correlations to high level tissues using the `cor()` function. The significance of correlations were calculated using the function `corPvalueStudent()` and corrected for multiple testing using `p.adjust()` using the method of (124). The most central genes in modules were identified using the function `signedKME()` which calculates the correlation between the expression patterns of each gene and the module eigengene. The most highly correlated genes were considered central to the module.

6. Epigenomic analyses

6.1 Analysis of DNA methylation

Cytosine methylation was profiled in DNA extracted from two-week old CS leaf tissue in three different contexts: CpG dinucleotides, CHG and CHH (where H corresponds to A, T or C). 2.25 billion high quality whole genome bisulphite Illumina sequence (WGBS) reads were aligned with Bismark version 0.16.1 (125) to IWGSC RefSeq v1.0. Despite the complications of working with a highly repetitive complex polyploid species, unique bisulfite read alignment was found to be in excess of 60%. A total of 79,851,988,588 high confidence cytosine methylation calls representing 77.1% of the assembled genome with an average read depth of 16-fold and a minimum depth of 4-fold were used in subsequent analyses.

6.2 ChIP-seq analysis of histone marks

ChIP assays were performed with anti-H3K9ac (Millipore, ref. 07-352), anti-H3K27me3 (Millipore, ref. 07-449), anti-H3K4me3 (Millipore, ref. 07-473) and anti-H3K36me3 (Abcam, ab9050) antibodies, using a procedure adapted from (126). Briefly, after fixation of 10-day-old *in vitro* seedlings in 1% (v/v) formaldehyde, tissues were homogenized, nuclei isolated and lysed. Cross-linked chromatin was sonicated using Covaris S220. Protein/DNA complexes were immunoprecipitated with antibodies overnight at 4°C with gentle shaking and incubated for 1h at 4°C with 50 µL of Dynabeads Protein A (Invitrogen, Ref. 100-02D). The beads were washed 2 × 5 min in ChIP wash buffer 1 (0.1% SDS, 1% Triton X-100, 20 mM Tris-HCl pH 8.0, 2 mM EDTA pH 8.0, 150 mM NaCl), 2 × 5 min in ChIP Wash Buffer 2 (0.1% SDS, 1% Triton X-100, 20 mM Tris-HCl pH 8, 2 mM EDTA pH 8, 500 mM NaCl), 2 × 5 min in ChIP wash buffer 3 (0.25 M LiCl, 1% NP-40, 1% sodium deoxycholate, 10 mM Tris-HCl pH 8.0, 1 mM EDTA pH 8.0) and twice in TE (10 mM Tris-HCl pH 8.0, 1 mM EDTA pH 8.0). ChIPed DNA was eluted by two 15-min incubations at 65°C with 250µL elution buffer (1% SDS, 0.1 M NaHCO₃). Chromatin was reverse-crosslinked by adding 20µL of 5M NaCl and incubating overnight at 65°C. Reverse-cross-linked DNA was treated with RNase and proteinase K, and extracted with phenol-chloroform. DNA was precipitated with ethanol in the presence of 20µg of glycogen and resuspended in 20µL of nuclease-free water (Ambion) in a DNA low-bind tube. 10 ng of immuno precipitate (IP) or input DNA were used for ChIP-Seq library construction using NEB-Next Ultra II DNA Library Prep Kit for Illumina (New England Biolabs) according to manufacturer’s recommendations. For all libraries, ten cycles of PCR were used. The quality of the libraries was assessed with Agilent 2100 Bioanalyzer (Agilent), and the libraries were sequenced using 2x75bp pair-end reads on NextSeq 500 platform (Illumina). Raw FASTQ files were preprocessed with Trimmomatic v0.36 (127) to remove Illumina sequencing adapters, trim 5’ and 3’ ends with

quality score below 5 (Phred+33) and discard reads shorter than 20 bp after trimming. Paired-end reads were aligned against IWGSC RefSeq v1.0 using bowtie2 v2.3.3 with --very-sensitive settings (128). Alignments with MAPQ < 10 were discarded and duplicate reads removed with Picard MarkDuplicates (<http://broadinstitute.github.io/picard/>). Enrichment peaks were called with MACS2 v2.1.1 using the following settings: -f BAM -B --nomodel --to-large -q 0.05 --broad --bdg -g 17e9 --bw 300. ChIP-seq input data were used as a control. The peaks were annotated according to their position and overlap with HC genes (iwgsc_refseqv1.0_HighConf_2017Mar13.gff3). The distribution of histone marks along chromosomes was estimated as a function of the percentage of mark-associated genes in a sliding window of 10 Mb (step = 1 Mb).

7. Gene family analysis

7.1 Identification of gene families

Gene families were defined in an automated phylogenomics approach incorporating the predicted protein sequences from bread wheat (RefSeq Annotation v1.0), nine other grasses and four Viridiplantae genomes (Table S23; Fig. S13). Wheat sub-genomes were handled as independent taxa. Protein sequences were screened for known protein domains using HMMer3 `hmmsearch` (-cut_ga) against the PFAM 30.0 database and a set of custom HMMs representing gene families inferred previously (129). Protein sequences with domain matches from HMMs associated with TEs were classified as TE-related proteins. Domain matches were filtered and assembled into domain architectures using the DAMA software (130). Protein sequences were compared in all vs. all blastp searches using the NCBI blast+ toolkit as suggested in the OrthoFinder manual (131). Resulting blastp hits were filtered using the following criteria: A) both proteins have annotated domain architectures and the alignment covers the entire domain ranges of each protein; B) at least one of the proteins does not have annotated domains, but the alignment covers any contained domains and at least 60% of both proteins; C) the alignment covers at least 70% of the longer and 90% of the shorter protein. Links between TE-related proteins determined in the previous steps were discarded. The inferred domain architecture and protein alignment statistics were employed to discern the representative isoform for loci with multiple splice variants in any of the three bread wheat sub-genomes by choosing the best supported variant. The filtered protein matches were used as input for reconstruction of gene families in the form of orthologous groups using OrthoFinder (131). The resulting gene families and phylogenetic trees were analyzed to infer homologous relationships among gene family members (orthologs, homeologs, inparalogs and outparalogs) and definition of subfamilies by re-rooting the cladograms based on the ‘get_farthest_oldest_node’ (or as a fallback the ‘get_midpoint_outgroup’) method. Subsequent species tree reconciliation using the Species Overlap algorithm (132) was implemented using the ETE3 toolkit (133). During this procedure, orthologs among the three bread wheat sub-genomes were classified as homoeologs. Fig. S14 provides an example of the inferred speciation and duplication events in the resulting gene trees. Internal nodes representing the origin of subtrees i.e. subfamilies, initiated by pre-speciation duplication events and speciation events, were placed into taxonomic context using the species tree (Fig. S13) and the NCBI taxonomy database. The resulting 14,275 subtrees containing only orthologs were used to infer a species tree with branch lengths using the coalescence approach implemented in ASTRAL (134). The species tree was rooted using the four non-grass taxa and cut to contain only the *Poaceae* taxa under study (Fig. S15). The phylogenomics workflow identified 30,597 gene families including 26,080 families with members from at least one of the three wheat sub-genomes. Overall, 63,523,422 homologous relationships, comprising 2,555,680

orthologous, 123,588 homeologous, 1,797,522 inparalogous and 59,046,632 outparalogous links, were inferred.

7.2 Expansion and contraction of gene families

Gene family expansions and contractions in either one of the wheat sub-genomes, the wheat A- or D-lineage or the wheat genome in general were inferred using a previously established phylogenomic approach (135) (Fig. S42). Log-transformed gene family sizes among the 10 grasses were compared using phylogenetic comparative one-way ANOVA (136), correcting for phylogenetic interdependency using a species tree (Fig. S13) that was computed by coalescence analysis (134) of 14,275 orthologue-only subfamily clades extracted from the gene family trees derived by OrthoFinder. To account for hexaploidy in the bread wheat genome, all three sub-genomes were treated as separate entities i.e. taxa. False-positives from remnant chimeric gene families were excluded from the analysis by only considering gene families where at least 51% of the members harbored consistent domain architecture. This filter excluded 18 families. Gene family sizes were log-scaled.

It is important to note that our gene family expansion / contraction study followed a very conservative approach considering the limitations of incomplete draft progenitor genomes. We explicitly avoided reporting any direct contrast between *T. aestivum* and *T. urartu* or *A. tauschii*. Instead we only looked at the following two cases of lineage-specific gene copy number (GCNV) variations: a) wheat subgenome (*T.aestivum* A and D and *T.urartu* respectively *A. tauschii*)-lineage-specific GCNVs (need to be similarly expanded in D and *A. tauschii* or A and *T. urartu*), b) wheat lineage (*T. aestivum* A, B and/or D)-specific GCNVs that were detected for the bread wheat genome only. While the wheat lineage in this analysis included *T. aestivum*, *T.urartu* and *A. tauschii*, these expansions were not found in *tauschii/urartu*. This was either because of the limitation of the two progenitor genome assemblies or because they are *aestivum*-specific. To assess and quantify possible under-prediction of gene counts from the draft progenitor genomes of wheat A- and D-subgenomes used in this study we benchmarked the differences between gene sets obtained for a draft (54) and a high quality genome assembly (137) of the same *A. tauschii* genotype. With 33,928 genes predicted (54), the *A. tauschii* draft assembly contains ~14% fewer genes compared to the recently reported 39,622 gene calls (137). To be able to provide an estimate of the impact on our gene family reference set, we used hidden Markov models (HMMER) to extract gene models of selected sub-phylogenies of the prolamin gene family in the updated *A. tauschii* gene set (Table S39). Those gene models were then manually curated and verified and their number compared to those we discovered automatically in the draft genome gene sets (54). We found that – on average – only 0.43 genes were missing in the automatically annotated 2013 gene family set compared to the manually curated 2017 gene family set per subfamily. Based on this limited analysis there was no indication of severe under-prediction of gene counts for the draft D-progenitor genome used in our study.

7.3 Functional annotation of expanded and contracted wheat gene families

Sets of expanded/contracted gene families at 90% confidence (FDR) were assessed for enriched biological processes (GO), molecular functions (GO), cellular components (GO), plant structure developmental stages (PO), plant anatomical entities (PO) or plant morphology traits (TO) using ontology term enrichment with the Parent-Child method implemented in the Ontologizer software (138). To further dissect overlap and unique features in sets of enriched terms among the expanded and contracted gene families, the FGNet Bioconductor R package (139) was used to construct a functional ontology term network with the overlapping and distinct enriched GO, PO and TO terms. The resulting graph was clustered into 7 related subnetworks with maximal overlap by GLAY community clustering (140) (Figs. S45-S51).

7.4 Validation of phylogenomic gene family annotation

The orthologous groups of gene families derived from the automated phylogenomic approach were evaluated against seven manually curated gene families: Aquaporins, C-Repeat Binding Factors (CBF), Dehydrins, Nucleotide-Binding Site – Leucine-Rich Repeat genes (NLR), Pentatricopeptide Repeat genes (PPR), Prolamins, and Cell Wall Associated Kinase genes (WAK) (Table S37). Each orthologous group (OG), was assessed for expansion or contraction in the Poaceae taxa. The FDR-corrected p-values indicated that between 15% (PPRs) and 75% (CBFs) of the subfamilies were either expanded in wheat, in one of its sub-genomes, or in one of the ancestral lineages. Only a small number of examples of contraction of orthologous groups (OGs) was found. The uncorrected p-values indicated expansion – from 30% of the OGs in PPRs to 100.00% in CBFs. A more detailed phylogenomic analysis was undertaken for the CBFs (124 sequences) and Dehydrins (158 sequences). The domain composition was determined from the manually curated candidate sequence lists: CBFs – AP2 domain (PF00847); Dehydrins – Dehydrin domain (PF00257) and all proteins with at least one of these domains was extracted from the evograph dataset. The amino acid sequences were then aligned using UPP (141) and trees were constructed with FastTree (142) and visualized with iTOL (143), rooting the trees at mid-point. In these gene families, 96.97% of the OG were monophyletic (100% if those with a bootstrap support <0.75 are removed). The same is true when only considering expanded OGs (Table S38; FigS. S55, S56).

7.5 Comparative analysis of homeologous genomes

Homeologous groups (144) were defined based on the homologous relationships inferred from gene trees, and initially constructed for an unfiltered set of 260,162 genes comprising HC and LC genes. This set was iteratively analyzed using the results from all relevant analyses (AHRD, HMMER3/PFAM-domains, TE-overlap, pseudogene analysis) described above to identify potential pseudogenes and TE-related loci. This resulted in the final, filtered dataset comprising 181,036 genes (103,757 HC and 77,279 LC genes). As both HC and LC genes were included in the analysis, the resulting homeolog groups were classed as either ‘HC-only’, ‘LC-only’ or ‘mixed’. In addition, TE-related loci among the homeologs are included but tagged with ‘is_filtered=TRUE’. Homeologous groups harboring $\geq 50\%$ of these loci were considered as TE-related in their entirety. Each group of homeologous genes represents a clade of orthologous taxa in the gene tree of a family. Orthologs between sub-genomes were inferred as homeologs, allowing for sub-genome-specific inparalogs that emerged from post-hybridization duplication events. The homeologous groups were analyzed and annotated by comparing the chromosomal locations of group members and their orthologs in other species. In contrast to other methods used to define homeologs, e.g. best reciprocal BLAST hits, this definition of homeologs relies on the topology of a phylogenetic tree, resulting in less false-positive inferences due to a more stringent assessment of gene orthology beyond mere sequence similarity cut offs. The stringency, in turn, is limited by the scalability and performance of large-scale phylogenetic reconstruction, which is affected by the quality and completeness of the gene models in wheat and from the other species used in the phylogenomics approach. Thus, the total number of homeologs inferred by this approach should be considered a conservative estimate of the true number of homeologous loci in bread wheat.

Each homeologous group was assigned a theoretical and an absolute cardinality, based on the number of homeologs identified in each sub-genome. Table 3 shows an overview of the contributions of different homeolog cardinalities in the wheat genome.

7.6 Analysis of synteny

Microsynteny of the homeologs, i.e. the conservation and collinearity of local gene ordering between orthologous chromosomal regions, was analyzed using the MCScanX software (145) on the homeologous gene pairs. 1,256 blocks were analyzed with respect to their chromosomal

region, size, orientation (positive or negative) and relative chromosomal location (see below). In addition, the macrosynteny of homeologs was described. i.e. the conservation of chromosome macrostructure accompanied by overall conservation of chromosomal location and identity of genetic markers like homeologs, but not excluding the occurrence of local inversions, insertions or deletions. This was assessed by analyzing the chromosomal position of the individual homeolog groups with respect to their relative chromosomal position and compared to the chromosomal location of their respective orthologs in barley (146), *Aegilops tauschii* (54) and *Triticum urartu* (53). For both macro- and micro-synteny any regional bias was avoided by using a relaxed block size requirement (≥ 3 genes per block). The chromosomal identity of the majority of orthologous genes for each homeologous group was considered to be the ancestral state of that group. To account for different sizes and possible reorganization of chromosomes, the relative chromosomal position was analyzed in relation to the predicted centromere positions: A) short arm: $-(1-(\text{gene_start}/\text{posCEN}))$ B) $+(\text{gene_start}-\text{posCEN})/(\text{chr_length}-\text{posCEN}+1)$. These CEN-relative positions were subsequently compared by their standard deviations to further assess macrosynteny. In addition, a situation where any of the homeologs is not located on the inferred ancestral chromosomal location of a homeolog group, the maximal standard deviation of CEN-relative positions of all microsyntenic 1:1:1 homeologs (0.1983149) was used as a lower limit to test if any of the homeologs in a group substantially deviated in their relative chromosomal position or whether they displayed a conserved relative position indicative of macrosynteny. Both criteria were used to decide whether a homeologous gene pair is in macrosynteny (both on ancestral chromosome or both on same chromosome with conserved relative position) or not.

7.7 Error assessment for homeologous groups with a potential gene loss in one subgenome

The error of prediction of gene loss was assessed using BLASTP searches to determine the number of genes classified as absent in homeologous groups with an absolute cardinality of 1:0:1 or 1:1:0 or 0:1:1 but present in unanchored contigs compiled in chrUn. Protein sequences of these groups were aligned against each other and against the protein sequences of genes located on chrUn. Hits on chrUn were considered as putative homeologs if both members of a homeologous group could be aligned to the same chrUn protein sequence with BLASTP scores and query coverages $\geq 100\%$ of the average value of the reciprocal alignments. As the chromosomal origin of the additional copies could not be determined, they could either represent A) missing homeologs or B) additional subgenome inparalogs, or C) assembly artifacts. In order to correct for possible errors in our inference of gene losses, we assumed A) to be true for all cases (worst-case scenario) and removed all groups with additional copies from the set with potential gene losses and performed a `chisq.test` in R. Results are presented in Table S25.

7.8 Nucleotide divergence between homeologs in coding (CDS)/protein sequence

Divergence and synonymous (Ks) and non-synonymous substitution (Ka) rates were determined from pairwise comparisons of homeologs in each orthologous group and tandemly repeated genes in each tandem cluster. Dialign2 (147) was applied to generate pairwise protein alignments which were transformed into codon based nucleotide alignments using a custom python script. Only aligned positions with quality ≥ 1 were scored. Protein and nucleotide similarities were reported as the percentage of mismatches relative to the lengths of aligned protein and coding sequences. Substitution rates were computed as model-averaged values of several substitution models (parameter-‘MS’) applying KaKs Calculator 2.0 (148) on the codon alignments.

Homeolog sequence similarities at protein and CDS levels were measured as the percentage sequence identity of pairwise member comparisons for each homeologous group. Three

categories were analyzed separately: all ortholog clusters; strict 1:1 relationship, i.e. at most one homeologous copy is found in each sub-genome; and one-to-many relations for which, in at least one sub-genome, the copy number of the homeolog has been expanded. These categories were further subdivided into the three possible sub-genome permutations ('AB', 'AD', 'BD'). In addition, the one-to-many category sequence similarities were surveyed for the intra-sub-genome expanded homeologs, 'AA', 'BB' and 'DD'.

8. Breeding relevant case studies

8.1 IWGSC RefSeq v1.0 guided dissection of the QTL *SSt1*

Markers that were previously reported to be associated with the peak of the *SSt1* QTL in the Lillian/Vesper genetic map were downloaded from (49), and their IWGSC RefSeq v1.0 anchored physical positions were extracted. Markers were also anchored to TGACv1 (8) and Triticum 3.1 (9) and the corresponding scaffolds/contigs were aligned to IWGSC RefSeq v1.0 by NUCmer (Fig. S57). Sequence size and gene content were compared between Ta3B and chromosome 3B from RefSeq v1.0 (TableS S40, S41). All gene and marker mappings were performed using GMAP software with at least 80% coverage and 95% identity.

RNA-Seq was performed on wheat cultivars Lillian (solid-stemmed) and Vesper (hollow-stemmed), and durum wheat cultivars Langdon (LDN, hollow-stemmed) and LDN-GB-3B [solid-stemmed, chromosome 3B (from Golden Ball) substitution line of Langdon]. Tissue was extracted at Zadoks stage 32 (149) (entire lowermost node, extending approximately 0.5 cm up along the stem). Individually barcoded cDNA libraries were prepared using the Truseq v2 unstranded kit (Illumina) per the manufacturer's recommended protocol. Reads were aligned to the IWGSC RefSeq v1.0 reference sequence using STAR version 2.5 (150) with default parameters, except the maximum mismatch rate (--outFilterMismatchNmax) was set to 6 (minimum 96% sequence identity) and the maximum intron length (--alignIntronMax) was set to 10 Kb. BAM files containing aligned reads were inputted into Stringtie (151) to count reads mapping to genes in IWGSC RefSeq v1.0 annotation v1.0 HC gene models. A matrix of raw read counts was analyzed by DESeq2 (123) for analysis of differential expression between hollow by solid-stemmed comparisons. Data were filtered for differentially expressed genes using an adjusted $p < 0.01$.

A single-strand conformation polymorphism (SSCP) marker was developed to assay the promotor of one of the differentially expressed genes, TraesCS3B01G608800, which carries a repeating AG(n) element 71 bp upstream of the transcriptional start site (Fig. S58). This repeating element was targeted using the following primers and the amplicon size quantified by capillary electrophoresis (Table S43): SSR_F: CAAATCGCCACAAGCTAGAGA, SSR_R: GTGTTCCAGCAGCTTGATGAG.

We also developed a quantitative PCR (qPCR) assay to test for possible copy number variation (CNV) of TraesCS3B01G608800 (Fig. 5C). qPCR reactions were performed in 10 μ l reactions in a 384 well microtiter plate and quantification was performed using SYBR green fluorescence measured in a BioRAD CFX384 cycler. The relative number of copies of TraesCS3B01G608800 was determined using the $\Delta\Delta$ CT method (152) using the gene TraesCS3B01G61220 as an endogenous control gene with a single copy. Primer sequences used were as follows: TraesCS3B01G61220_F1: CTCAACGAACGACAACGAT, TraesCS3B01G61220_F2: AGATCACCAGCTGCTCTACACCT, TraesCS3B01G61220_R1: ATGCGTAGGAGTCCATGAG, TraesCS3B01G61220_R2: GGCACATATCATAGACGGCG, TraesCS3B01G608800_F1: GTTCCTGCACGCCATGGAC, TraesCS3B01G608800_F2: GATGTCCGGGAATCCTCAAT, TraesCS3B01G608800_R1: TCCCCATCGTCGCCATTA, TraesCS3B01G608800_R2: TAGTCCCTCTTGCCGGCT.

A high-throughput and low-cost KASP marker was developed that can be used for MAS for SSt1 in breeding (Fig. 5C). The marker was developed near TraesCS3B01G608800, within the gene TraesCS3B01G890300LC.1 that uses the primers: usw204-HF: GGCAAAGAACAAAAGCGGTAGAC, usw204-FF: GGCAAAGAACAAAAGCGGTAGAG, usw204_R: GGAGGCTGCGCTAAGAAATTC.

8.2 IWGSC RefSeq v1.0 guided genome-editing of putative flowering time regulating genes

Guide RNAs were designed to specifically target *TaAGL33* by CRISPR/Cas9-based gene editing, on all sub-genomes [TraesCS3A01G435000 (A-genome), TraesCS3B01G470000 (B-genome), and TraesCS3D01G428000 (D-genome)]. The guide RNAs were designed to target the I- and K-domains of *TaAGL33*, that are specifically used to delineate the *FLC* branch of the MADS box genes. Three guide RNAs targeted the following sequences in *TaAGL33* homeologs (note the PAM sequence is underlined): GGGTTCATTCACGTCCTTTCCGG (site for guide RNA g1), GGAAAGGACGTGAATGAACCCGG (site for guide RNA g2), and GTAAAGCCTCAAGGGCTACCAGG (site for guide RNA g3). Three T-DNA vectors each containing the bar selectable marker gene, CRISPR nuclease and one of three sgRNA sequences were used for *Agrobacterium*-mediated wheat transformation, essentially as described earlier (55). Transgenic plants were obtained with edits at the targeted positions in all *TaAGL33* homeologs. The three homeologs were sequenced in all plants obtained to describe eight gene edits in five independent events. Plants were subjected to vernalization conditions and monitored for expression and flowering time response as described before (50).

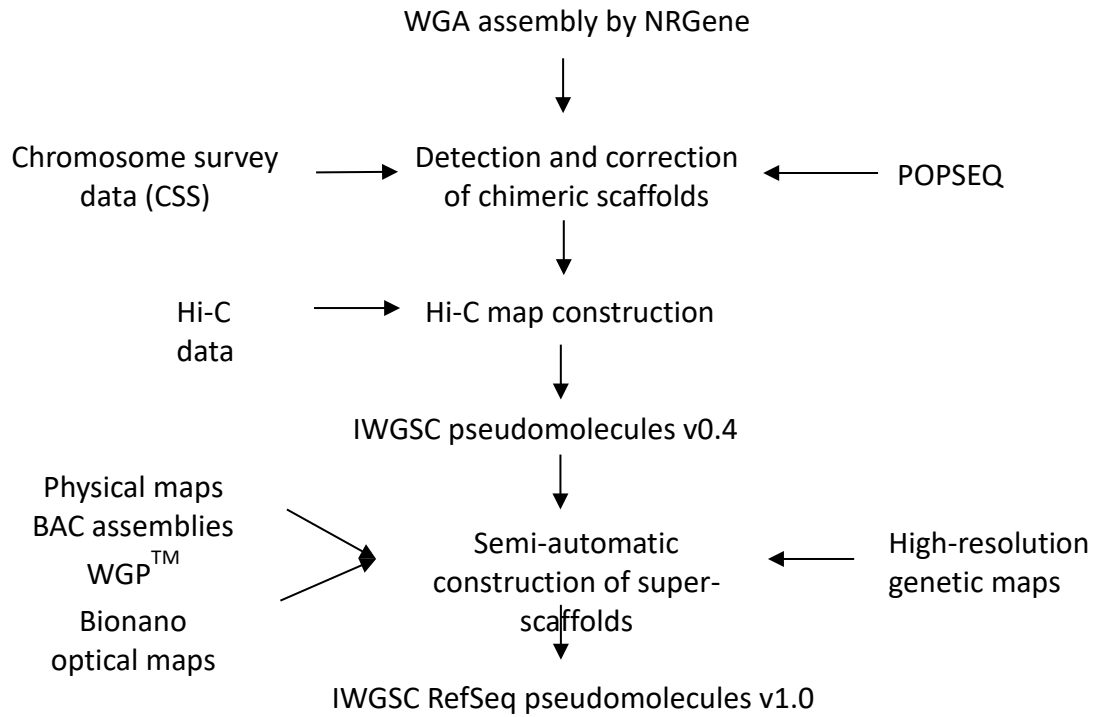


Fig. S1

Data integration pipeline for the assembly of IWGSC RefSeq v1.0. The whole genome assembly (WGA) used DNA from the cultivar CS.

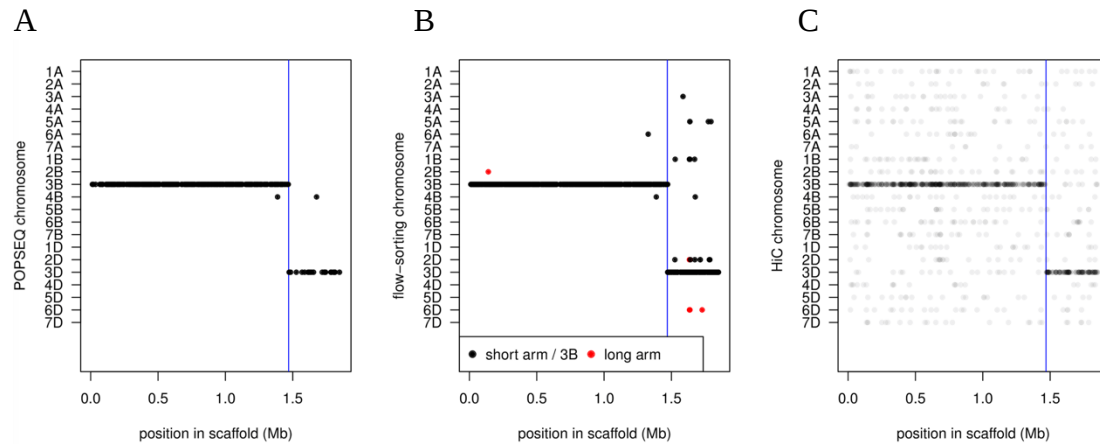


Fig. S2.

Example of a chimeric scaffold in the whole-genome assembly. The scaffold initially joined unlinked sequences from chromosomes 3B and 3D. The chimeric nature of the scaffold was detected A) by the POPSEQ genetic map, B) by the alignment of sequence contigs of the chromosome survey sequencing data (7), and C) by chromosome conformation capture data (Hi-C). The vertical blue lines indicates in all three panels the breakpoint at which the scaffold was split into two parts.

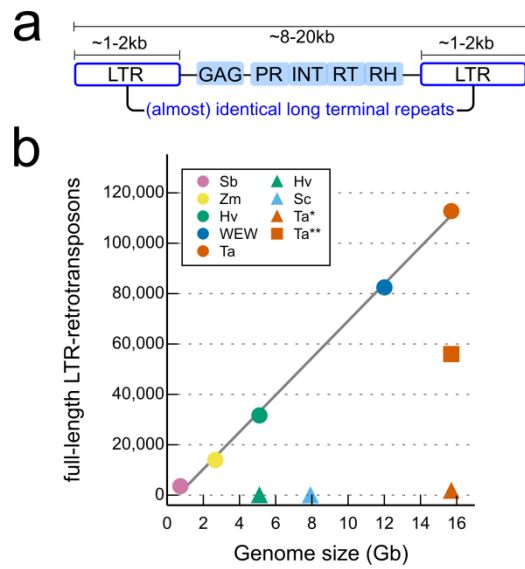


Fig. S3

Number of full length LTR-retrotransposons in different genome assemblies. (a) Schematic structure of a full length LTR-retrotransposon (fl-LTR). (b) fl-LTR content of different grass genome assemblies. The (almost) identical 1-2 kb long terminal repeats of fl-LTRs were often not correctly reconstructed in older contig assemblies (triangles). Circles denote more complete assemblies, triangles lower quality contig assemblies. Sb: *Sorghum bicolor* (153); Zm: *Zea mays* (154); Hv: *Hordeum vulgare*, triangle (94) circle (146); Sc: *Secale cereale* (155); WEW: wild emmer wheat (57); Ta: bread wheat, triangle IWGSC-2014 (6), square TGACv1 (8), circle IWGSC RefSeq v1.0.

All		84.7%
Class 1	Gypsy	46.7%
	Copia	16.7%
	Unclassified LTR-RTs	3.2%
	LINE	0.9%
	SINE	0.01%
Class 2	CACTA	15.5%
	Mutator	0.38%
	Mariner	0.16%
	Harbinger	0.16%
	hAT	0.01%
	Helitron	0.004%
	Unclassified Class 2	0.27%
Unknown	Unclassified repeats	0.68%

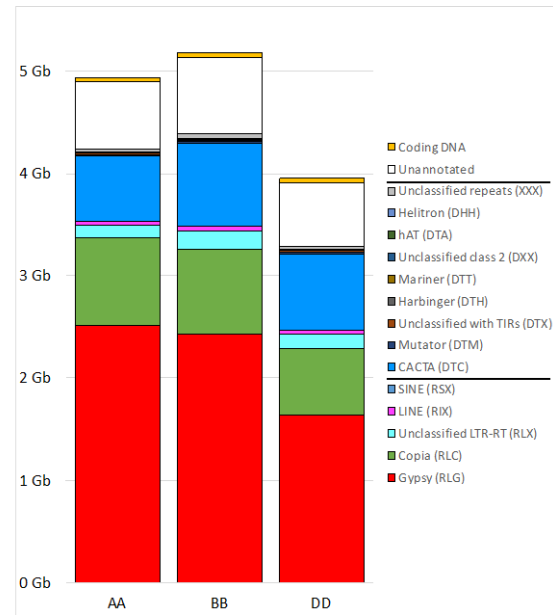


Fig. S4

Sub-genome distribution of the main TE classes in wheat. The right-hand side panel provides a summary of the relative TE distribution between the A, B and D sub-genomes of wheat.

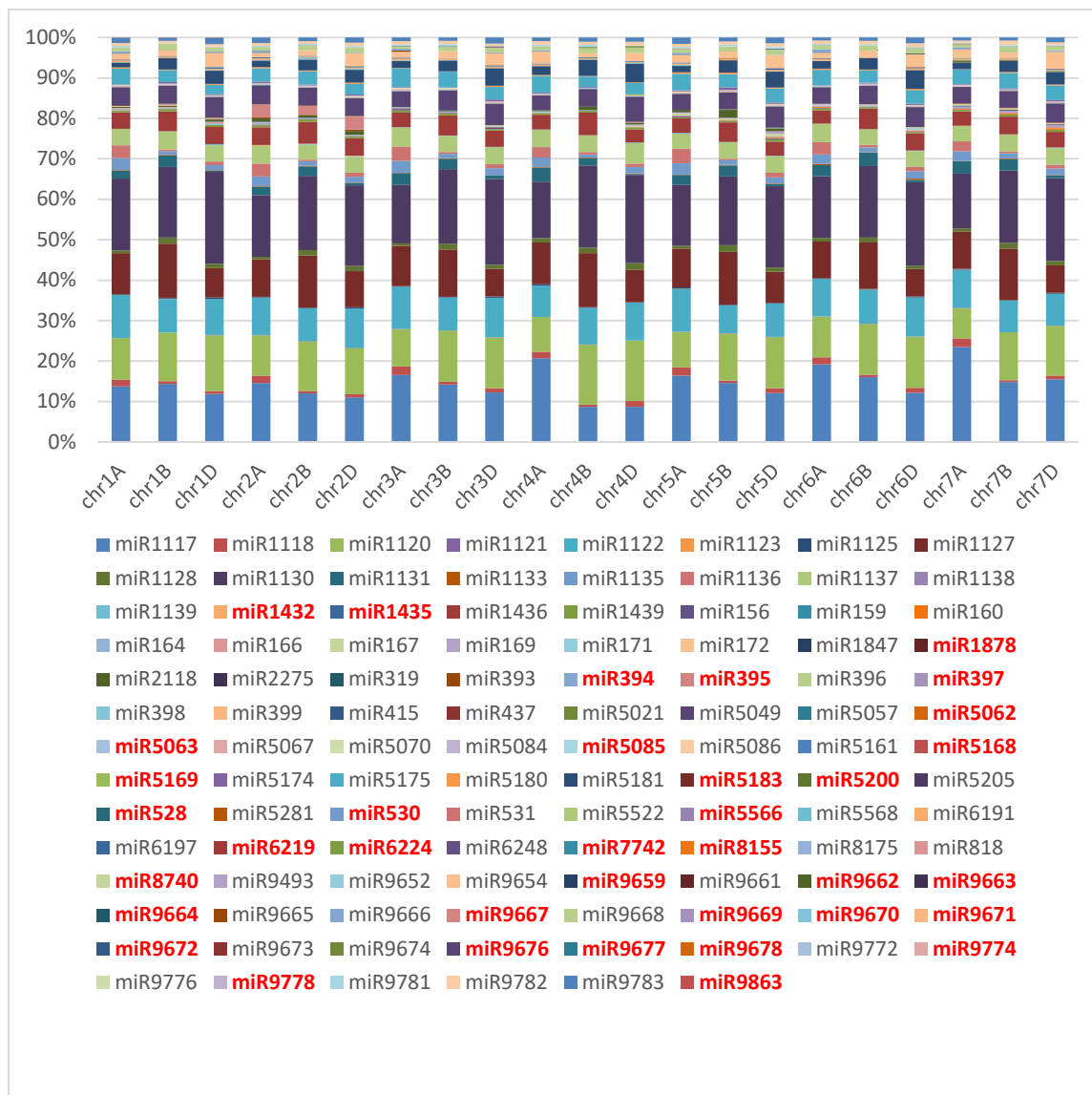


Fig. S5

Distribution of miRNA families in the 21 chromosomes of wheat. 36 miRNA families showed chromosome or homeolog specificity (highlighted in red): miR1432 (2ABD), miR1878 (5ABD), miR394 (6ABD), miR395 (2ABD), miR397 (6AB), miR5062 (5AB), miR5063 (3AB), miR5085 (4B), miR5168 (5ABD), miR5169 (4D), miR5183 (5A), miR5200 (7ABD), miR528 (5A), miR530 (2ABD), miR5566 (6D), miR6219 (6D), miR6224 (2D), miR7742 (4B), miR8155 (5B), miR8740 (7A), miR9659 (6B), miR9662 (6D), miR9663 (6B), miR9664 (1ABD), miR9667 (7ABD), miR9669 (3D), miR9670 (6ABD), miR9671 (7A), miR9672 (5D), miR9673 (3BD), miR9674 (3ABD), miR9676 (7ABD), miR9677 (2B), miR9678 (7ABD), miR9774 (2B), miR9778 (7ABD), miR9863 (1ABD). New miRNA families identified in the IWGSC RefSeq v1.0 (using all miRBase miRNAs): miR5021, miR5063, miR5067, miR5566, miR8155, miR8175, miR8740, miR9493.

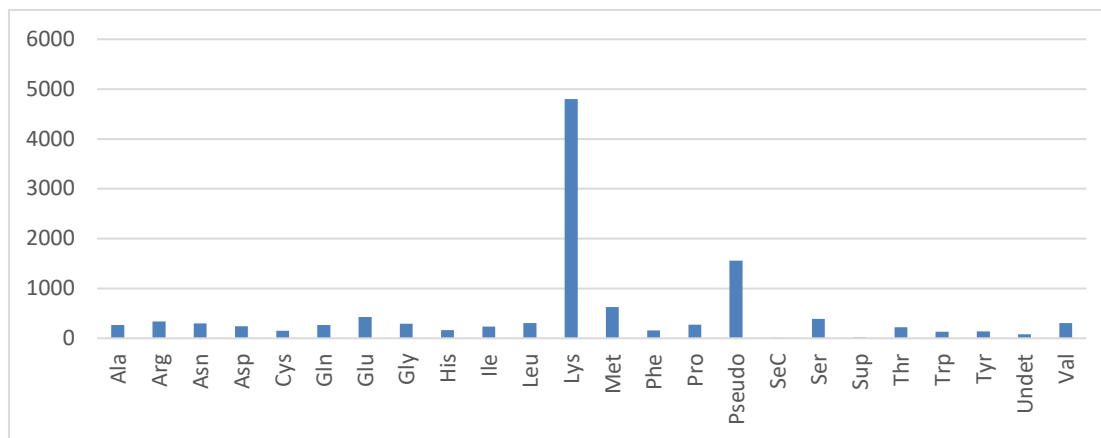


Fig. S6

tRNA gene frequency in IWGSC RefSeq v1.0. Numbers (Y axis) of tRNA coding genes for each amino acid (X axis). Pseudogenes and undet genes could not be assigned, and SeC (selenocysteine) and Sup (suppressor tRNAs) were detected in small amounts.

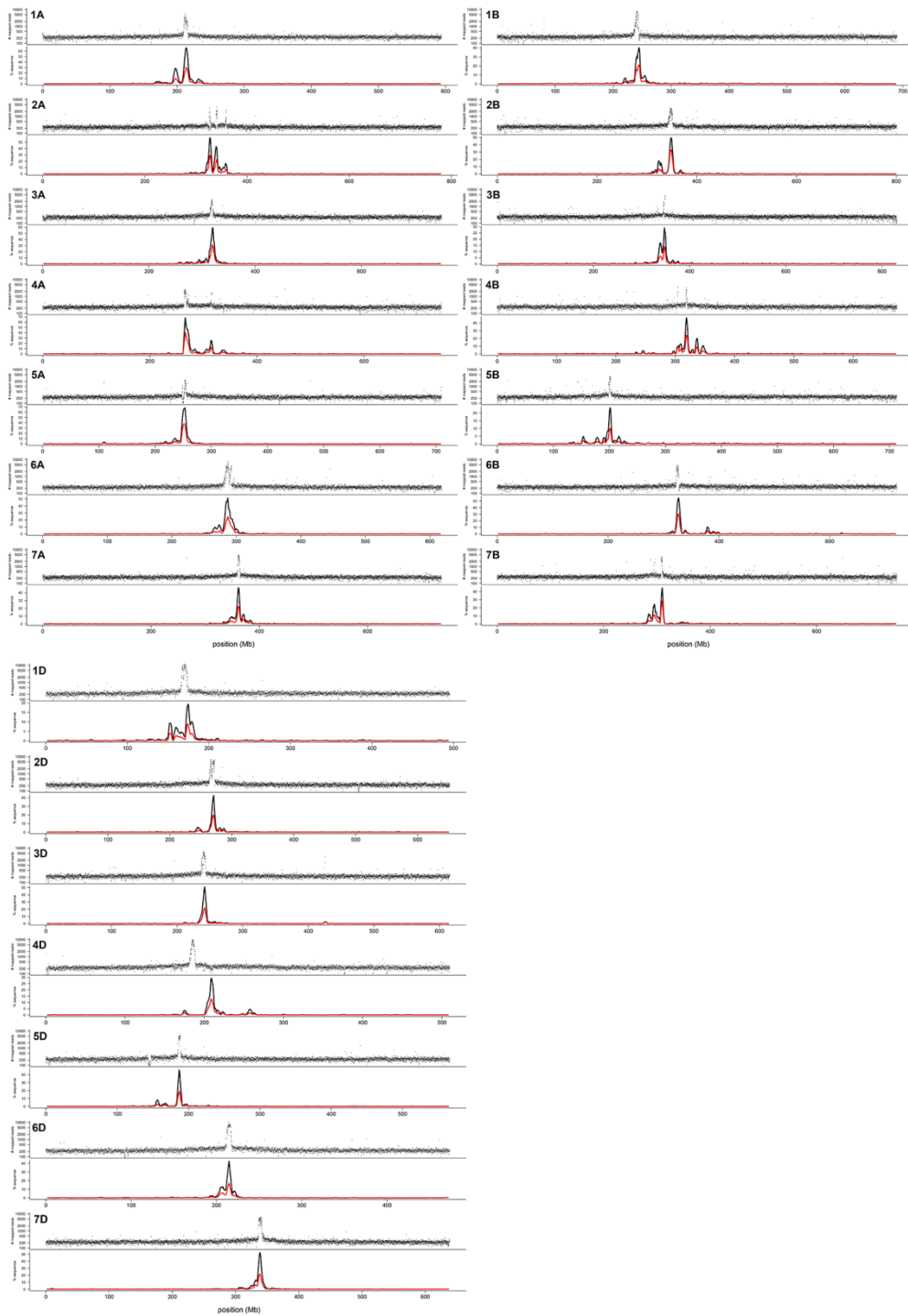


Fig. S7

Positioning of centromeres in wheat chromosome pseudomolecules. Distribution of CENH3 ChIP-Seq data across the 21 bread wheat chromosomes is shown in each upper panels. Lower panels show distribution and proportion of the total pseudomolecule sequence composed of TE of the Cereba (black)/Quinta (red) families. The most likely physical position of centromeres is at the peak of the CENH3 signal.

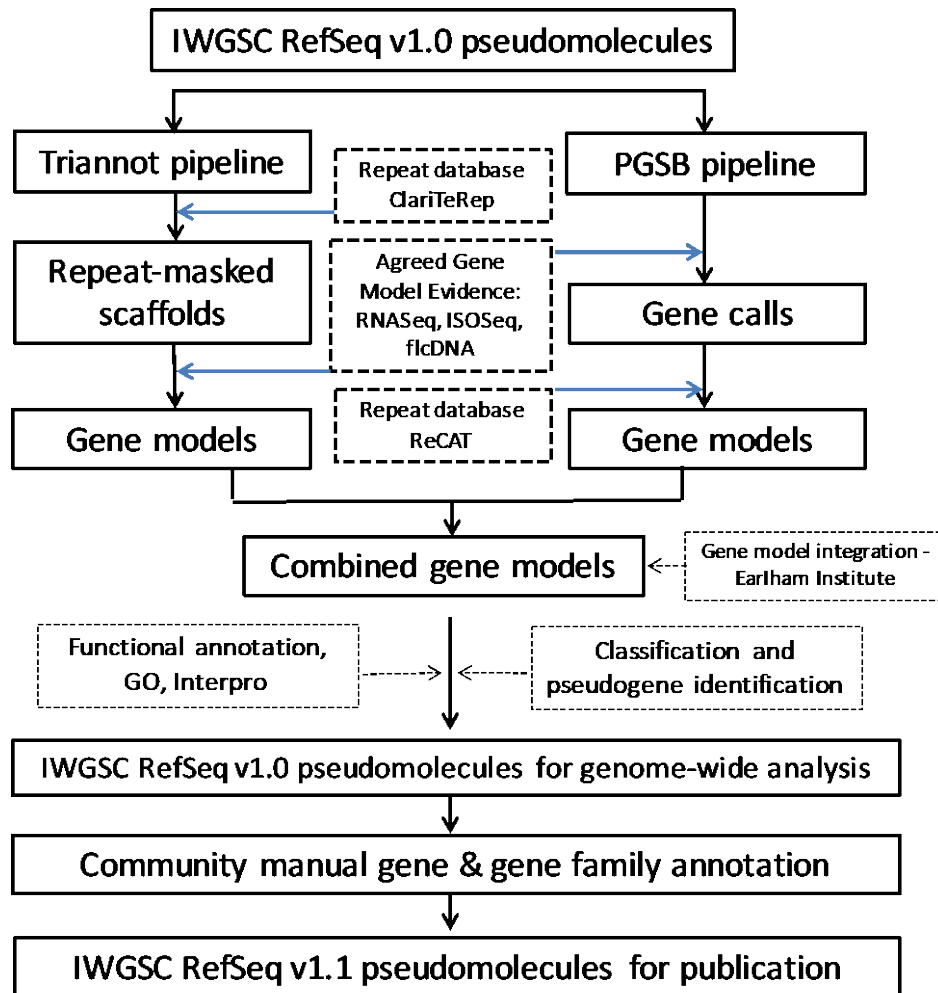
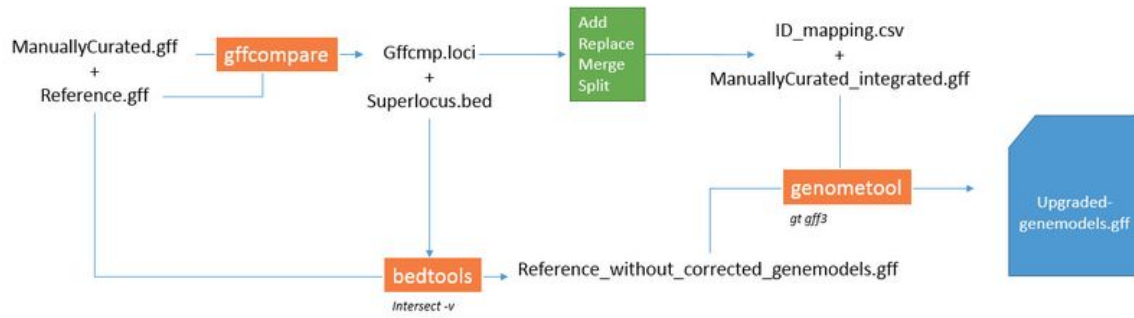


Fig. S8
IWGSC RefSeq v1.0 Genome annotation pipeline.

A**B****Fig. S9**

Semi-automated pipeline for integration of manually curated gene models. (A) Pipeline implementation and gene model modifications resulting from integration of 5 types of manual curation: (1) Add: new gene not annotated in automated annotation is annotated and a new gene-ID generated that reflects the position of the gene relative to the 2 surrounding genes (based on 9 bin intervals between the 2 genes.). (2) Replacement: the manually-curated gene corresponds to an already annotated gene. The gene ID is updated to correspond to the TraesCSCCVVGNNNNNN template, incorporating a change in confidence class, if necessary. (3): Split: manual curation indicates an existing gene model should be split. The new gene structure is annotated and the gene IDs modified accordingly. (4) Merge: the manual curation indicates existing gene models should be merged. The new gene structure is annotated and the gene IDs modified accordingly. (5) Multilocus: a special case where multiple reference and manual genes overlap is treated as an extension of the Merge/Split case. The new gene structures are annotated and the gene IDs updated accordingly. (B) Consequences of manual curation for gene models.

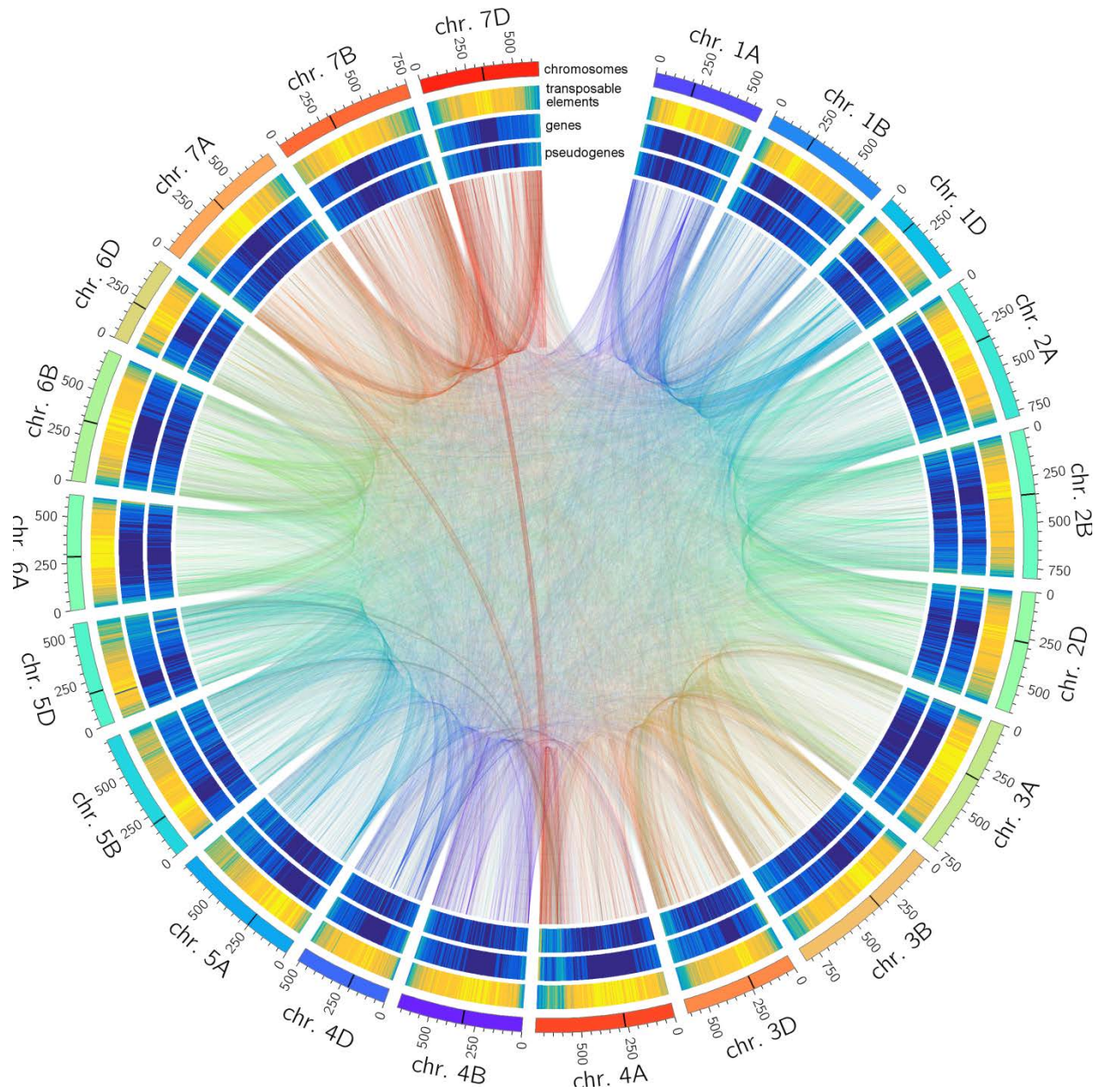


Fig. S10

Distribution of high-coverage pseudogenes (288,839; representing at least 80% of the CDS of a parent gene locus, see 4.5.2) in the IWGSC RefSeq v1.0 assembly. The four concentric circles visualize, from outside to inside, the 21 wheat chromosomes at physical scale, a heat map for the distribution of transposable elements, followed by the distribution of HC genes, and in the innermost circle the distribution of high-coverage pseudogenes. Lines in the center connect pseudogenes with their parent genes and are shown in the color of the chromosome harboring the parent gene.

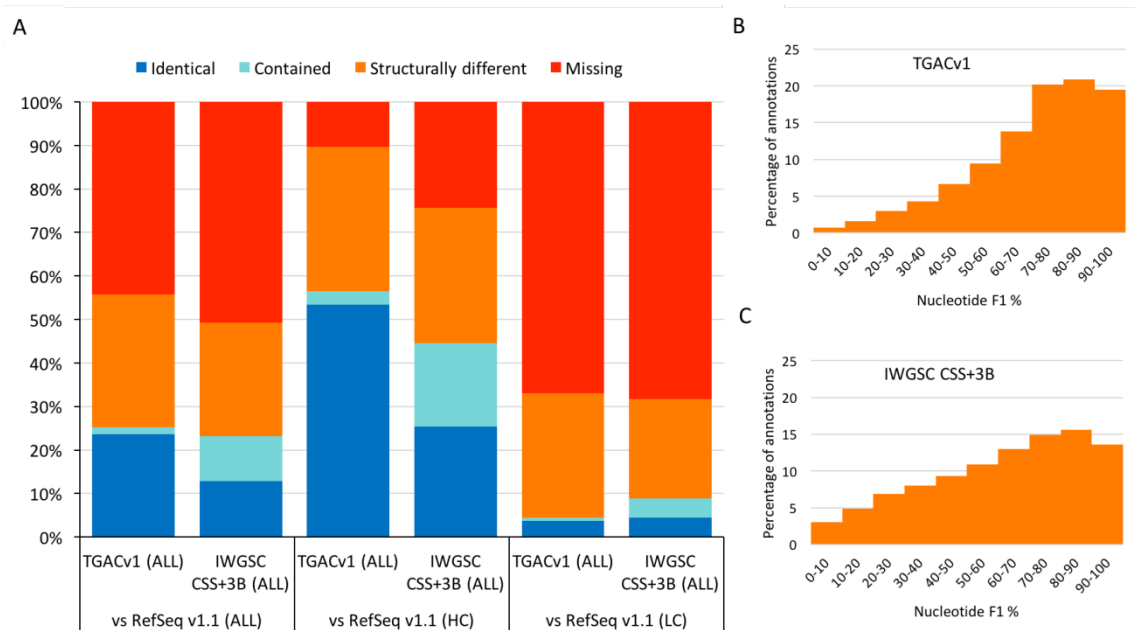


Fig. S11

Comparison of RefSeq annotation v1.1 to earlier wheat gene sets. (A) Identical indicates shared exon–intron structure; contained, exactly contained within the alternative gene; structurally different, alternative exon–intron structure; and missing, no overlap with the alternative gene. High confidence (HC), low confidence (LC) and the full set of RefSeq (ALL) gene models were compared to TGACv1 and earlier IWGSC gene sets aligned to the IWGSC v1.0 assembly. (B, C) Extent of nucleotide overlap (nF1) between RefSeq v1.1 and TGACv1 (B) or IWGSC CSS+3B (C) genes classified as structurally different.

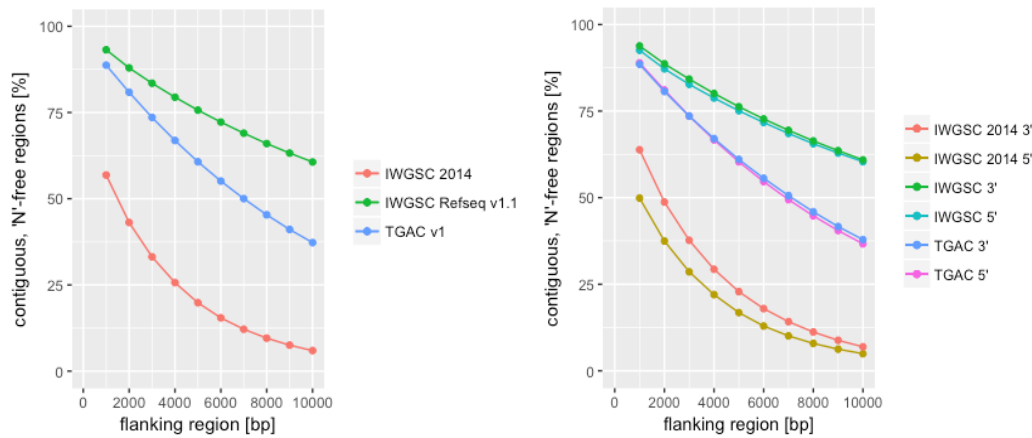


Fig. S12

Analysis of gene model flanking region length in different wheat genome assemblies. The completeness of flanking regions length was analysed in IWGSC RefSeq v1.0 compared to previous wheat assemblies TGACv1 (8) and IWGSC 2014 (6). Plotted are percentages of full-length “N”-free flanking regions (A: 5’-upstream and 3’-downstream combined; B: 5’-upstream and 3’-downstream separated) with increasing length. 93% of the predicted HC/LC gene models in RefSeq v1.0 contain a “N”-free flanking region of 1000 nt, compared to 88% in the TGACv1 assembly and 56% in the IWGSC CSS models. With increased flanking length of up to 10kb, 61% of RefSeq v1.0 gene models still retain a contiguous nucleotide sequence upstream and downstream of predicted gene models, whereas this number decreases to 37% in TGACv1 and to 5% in IWGSC CSS models.

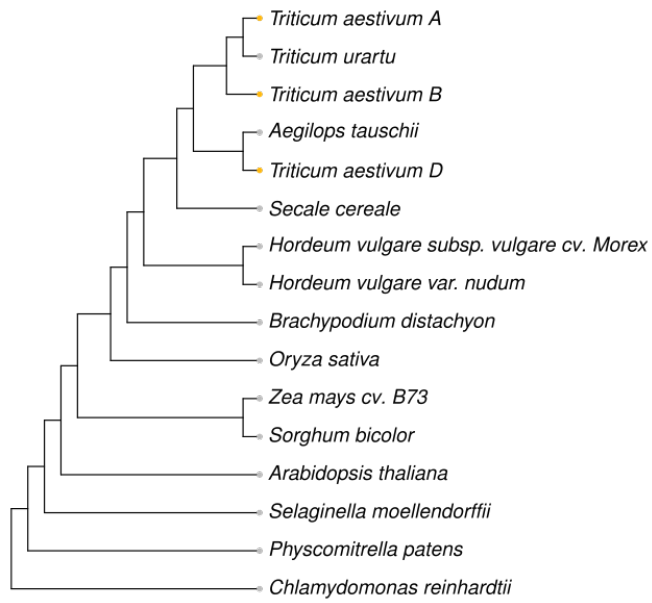


Fig. S13

Cladogram of the taxa used for phylogenomics.

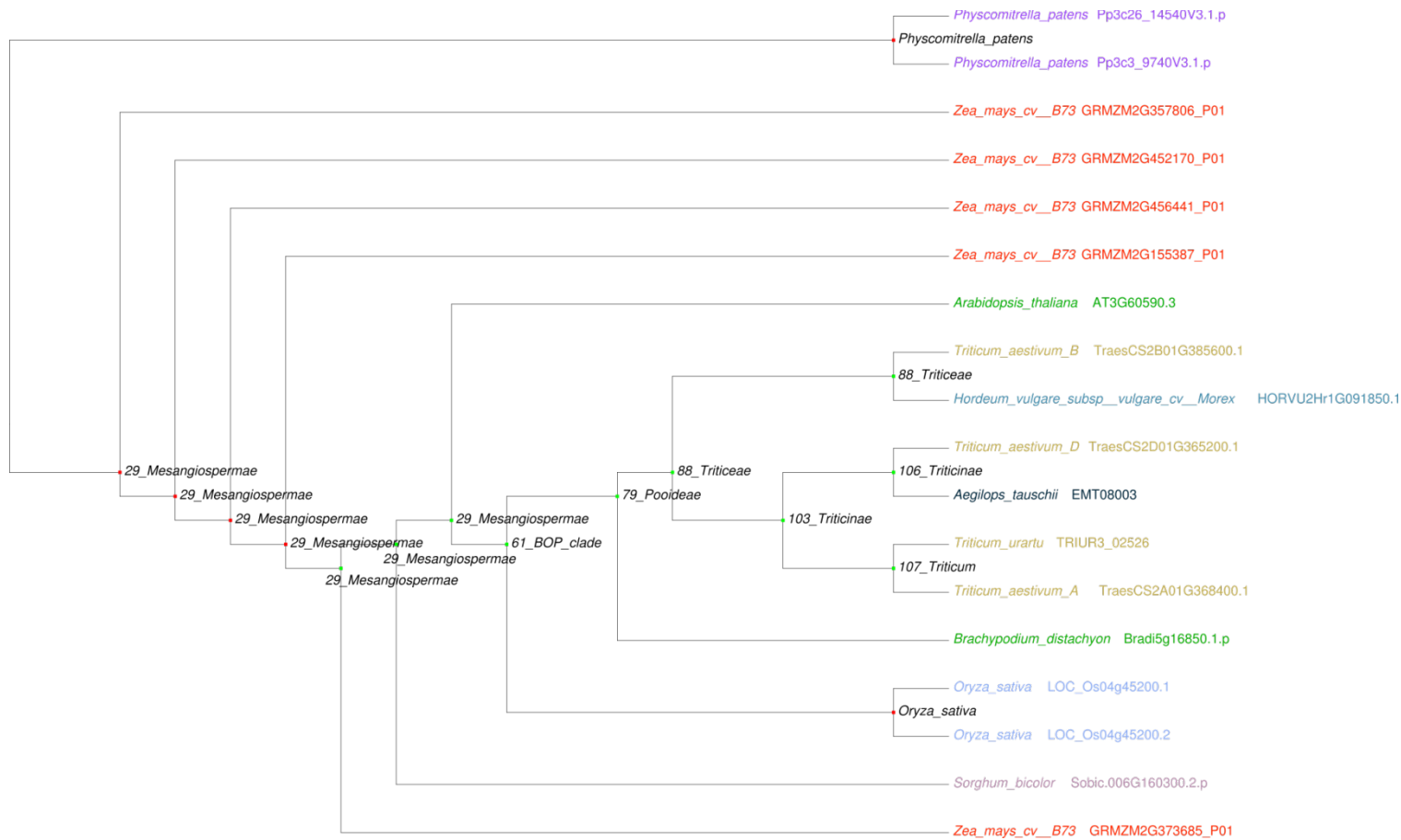


Fig. S14

Phylogenomic gene tree analysis of a yet uncharacterized plastid-localized cytochrome P450 subfamily. The family can be traced to the last common ancestor of land plants, showing the rooted, reconciled gene tree used for the inference of homolog relationships in the phylogenomics workflow. Where applicable, the originating data source for the selection of a representative isoform was used for phylogenetic inference. In this case the Phytozome database considered two splice variants in rice as two separate loci. Node colors encode the type of reconciled evolutionary event: green=speciation = orthologous relationships; red=duplication = inparalogous and outparalogous relationships. Inparalogs were called if descendant tips of a duplication node belong to the same taxon; Outparalogs derive from a node whose descendant tips harbor multiple taxa.

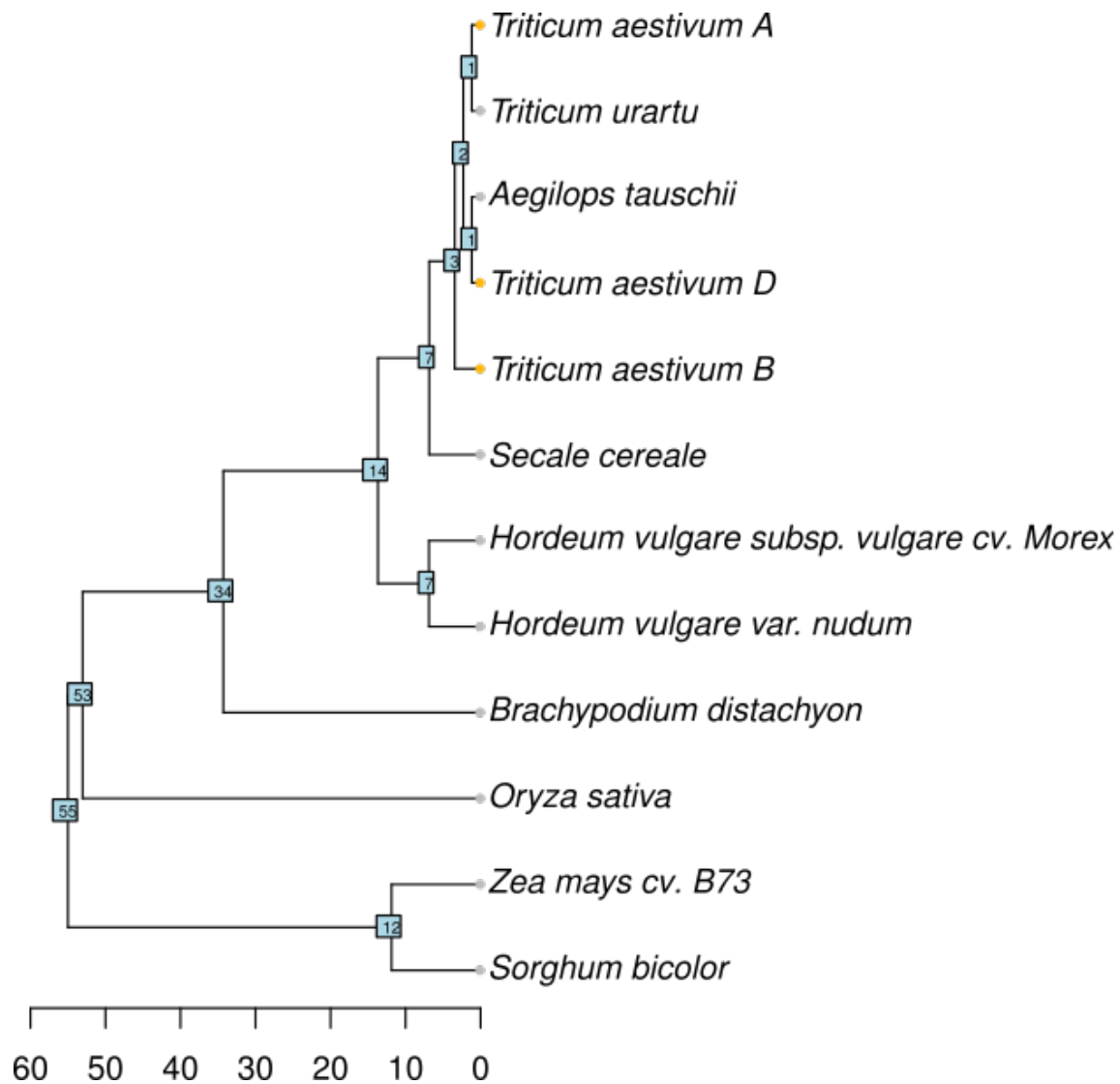


Fig. S15

- 5 Chronogram of the investigated Poaceae taxa. Divergence times were estimated using the penalized likelihood approach (156) and fixed calibration points inferred from a more comprehensive study on molecular dating of grasses (157).

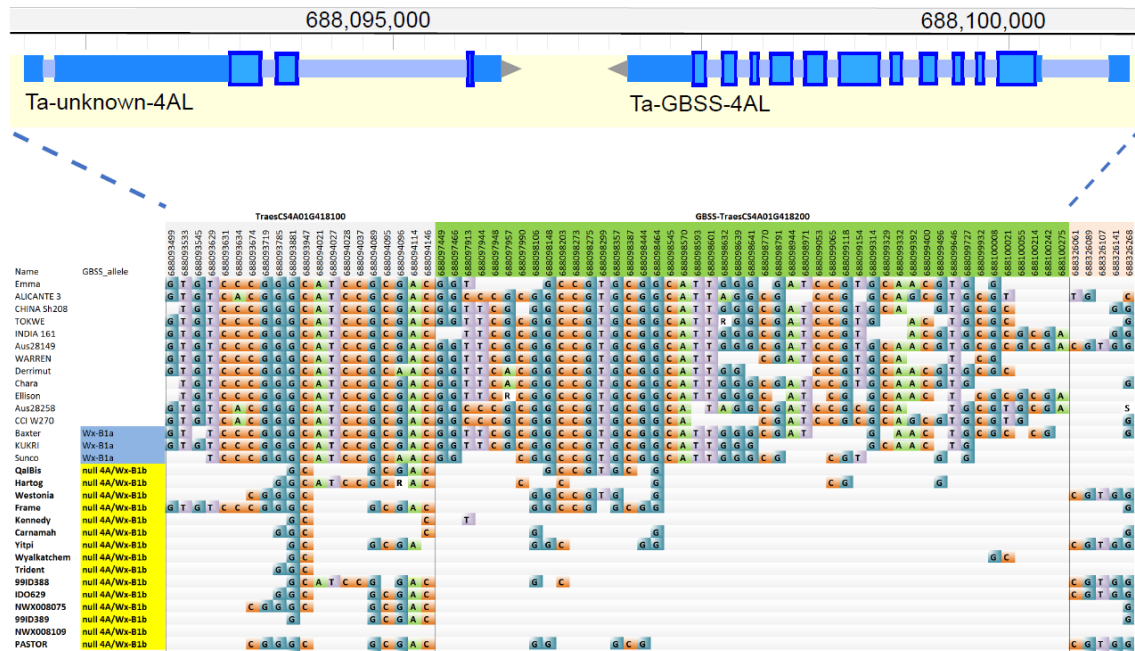


Fig. S16

Granule Bound Starch Synthase (GBSS; TraesCS4A01G418200) on chromosome 4A. The association of a null allele for GBSS-4AL (Wx-B1b, TraesCS4A01G418200) with udon noodle quality is well established (158). This gene model groups as an outparalog to a clade that is localized on chr 7 in *Triticeae*. The available data suggest the GBSS on chromosome 4A is a highly divergent translocated homeolog originally located in 7B. This gene shows an additional complexity in 3.9% (25) of a set of 644 (hexaploid) wheat varieties and landraces assessed using SNPs identified from snapshot exome sequence data ((159); a subset of 40 illustrated for wheat lines and accessions available from the Australian Germplasm Collection) where significant sections of TraesCS4A01G418200 within the green highlight target region to be deleted. Based on the distribution of missing SNP sequences, the lines highlighted in yellow that were independently confirmed GBSS-4AL/Wx-B1b null alleles, all carried deletions missing SNP sequences at positions 688098545-688099053 (6 exons at 3' end of gene) in chromosome 4A of the IWGSC RefSeq v1.0. A subset of deletions encompassed a larger region, 688098545-688099932 within 4A, but did not extend outside the gene model for TraesCS4A01G418200 and thus would not be expected to show any detrimental effects due deletion of the adjoining gene models (TraesCS4A01G418100, light grey highlight, and TraesCS4A01G418300, light orange highlight) and is consistent with varieties carrying the different forms of the deletion performing successfully at the agronomic level to satisfy the high value commercial udon noodle market. Lines that carried the Wx-B1a deletion did not show any missing SNP sequences in the GBSS-4L gene (TraesCS4A01G418200).

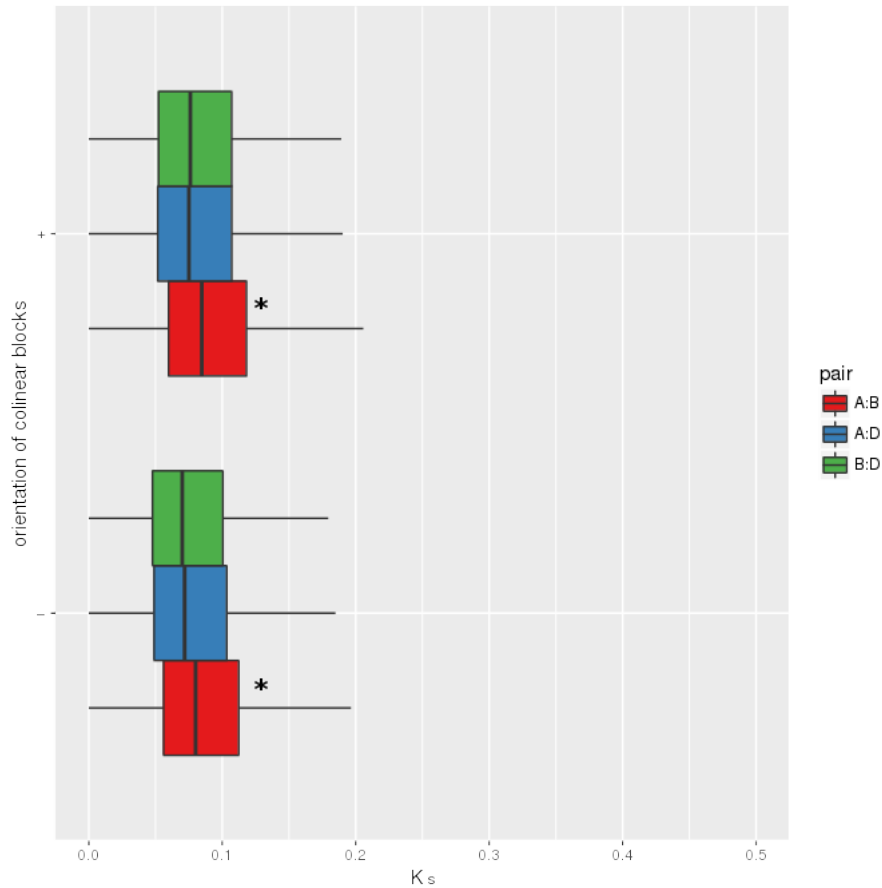


Fig. S17

Synonymous substitution rates (K_s) in homeologous gene pairs located in colinear blocks in the three wheat sub-genomes. The analysis of mutation rates within the protein-coding regions of homeologous gene pairs, reveals that the synonymous substitution rate is highest among homeologous A:B gene pairs (indicated by asterisks, significance: 99% confidence in Kruskal-Wallis rank sum test) as compared to A:D and B:D. This is consistent with the proposed sequence of polyploidization events in bread wheat, starting with hybridization of the A and B genome progenitors.

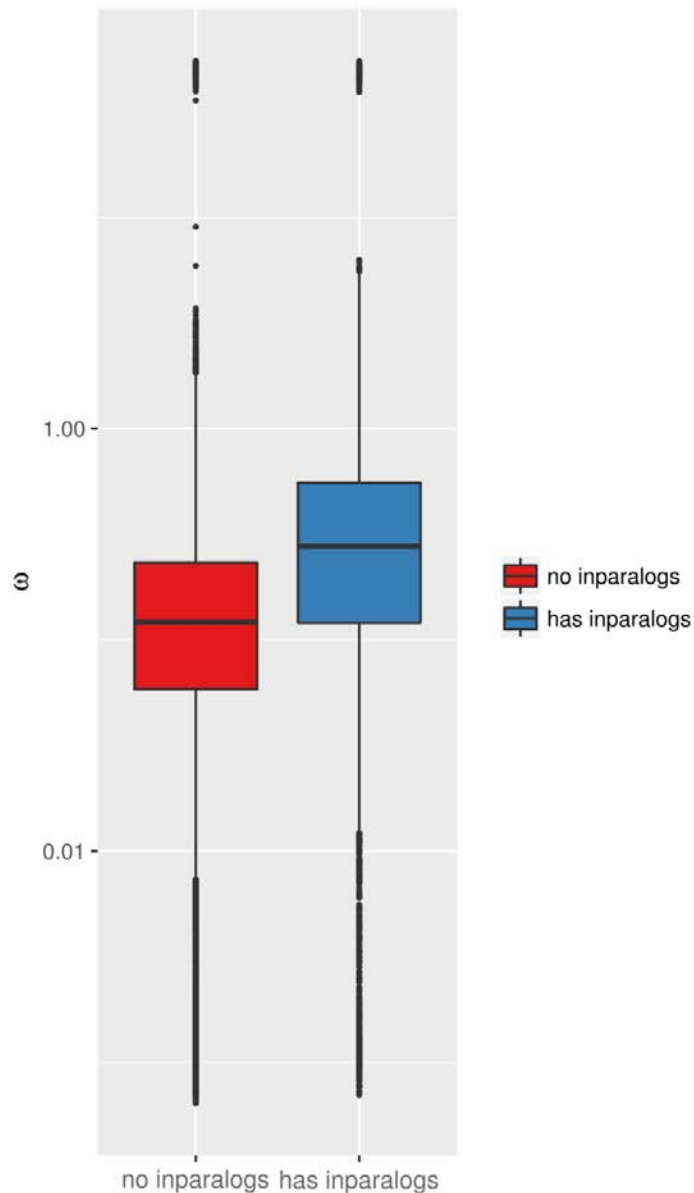


Fig. S18

Rate of non-synonymous to synonymous substitutions (K_a/K_s or ω) in homeologous gene pairs. Homeologs with sub-genome inparalogs (blue; gene duplications after hexaploidization) show higher levels of variation in ω along the chromosome and bear evidence of peaks of positive selection in contrast to homeologs without sub-genome inparalogs (red), which appear to be less variable and under purifying selection. Boxes represent lower (25%) and upper (75%) quartiles. The central line represents the median (50%) value. Outliers are plotted as individual dots.

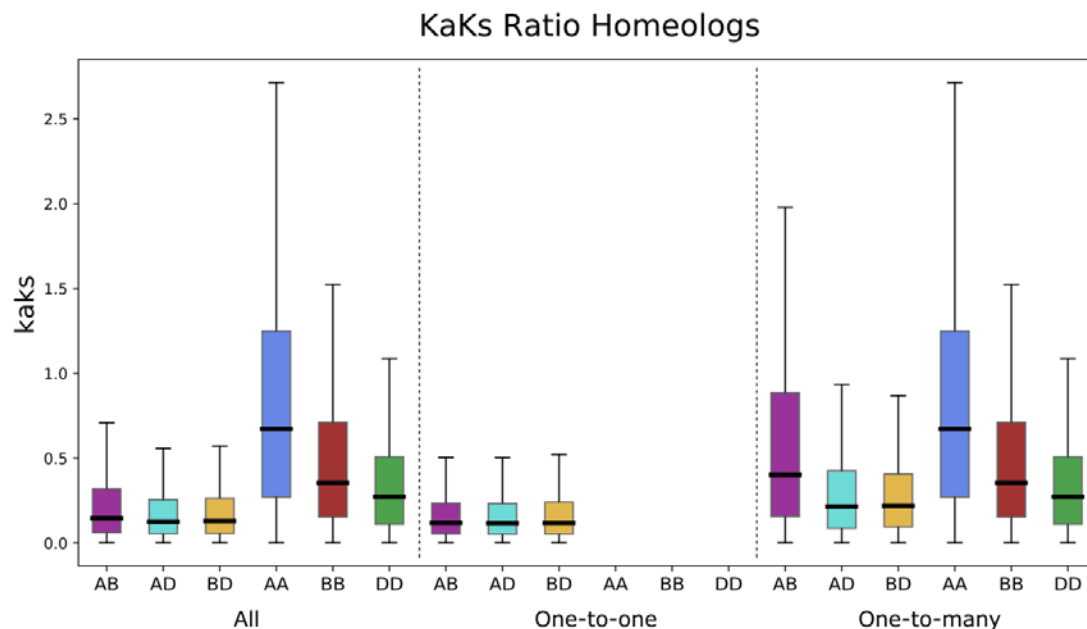


Fig. S19

Ka/Ks ratios of different homeolog categories and their subdivisions. Ka/Ks ratios are shown on the y-axis, median and spread (1st and 3rd quartile) for the three categories, 'all', 'one-to-one' and 'one-to-many' are illustrated as boxplots. Categories are further divided into pairwise sub-genome comparisons, AB, AD and BD, and intra-sub-genome comparisons if applicable. Boxes represent lower (25%) and upper (75%) quartiles. The central line within each colored box represents the median (50%) value. Lines represent the upper and lower whisker.

5

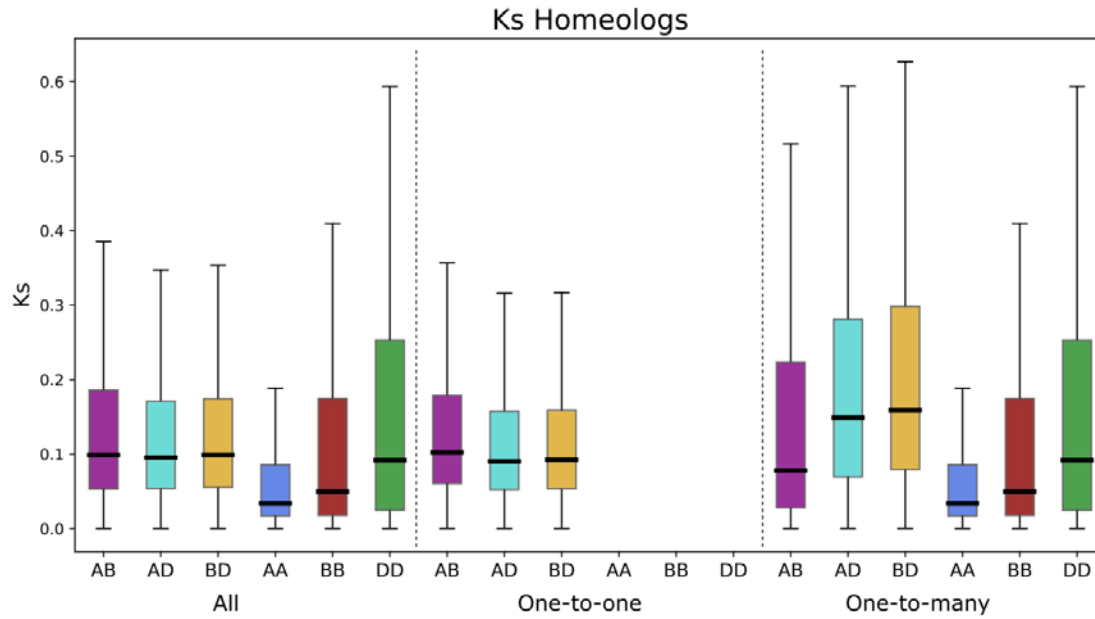


Fig. S20

Ks divergence in homeologous genes. Homeologous gene pairs with duplications in at least one of the sub-genomes (A:B, A:D, B:D) and the corresponding sub-genome inparalog pairs (A:A, B:B, D:D) bear evidence of elevated evolutionary rates. They showed higher ratios of non-synonymous to synonymous nucleotide substitutions (K_a/K_s or ω) than homeologous gene pairs from strict 1:1:1 or 1:1 groups. This phenomenon is particularly pronounced among sub-genome inparalogs, especially those found in the A or the B sub-genomes. The latter also have higher copy number than those in the D genome. Ks ratios are shown on the y-axis. Median and spread (1st and 3rd quartile) for the three categories, 'all', 'one-to-one' and 'one-to-many' are illustrated as boxplots. Categories are further divided into pairwise sub-genome comparisons, AB, AD and BD, and intra-sub-genome comparisons if applicable. Boxes represent lower (25%) and upper (75%) quartiles. The central line represents the median (50%) value. Lines represent the upper and lower whisker.

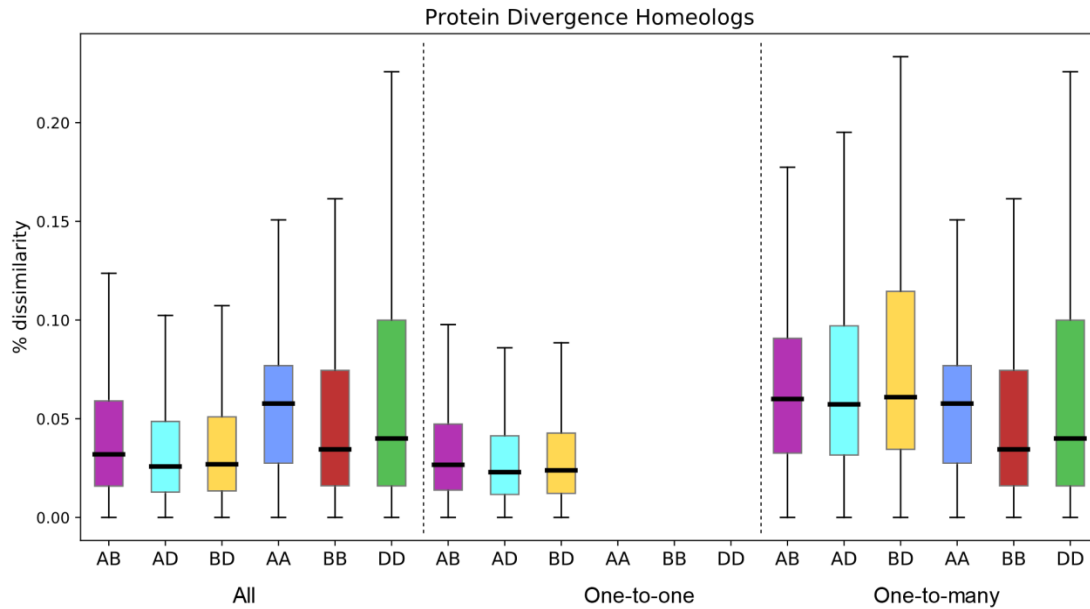


Fig. S21

- 5 Protein divergence in homeologous genes. Protein divergence in homeologous gene clusters is shown on the y-axis as percentage dissimilarity. Median values and spread (1st and 3rd quartile) for the three categories, 'all', 'one-to-one' and 'one-to-many' are illustrated as boxplots. Categories are further divided into pairwise sub-genome comparisons, AB, AD and BD, and intra-sub-genome comparisons if applicable. Median protein sequence identity was highest for
- 10 the one-to-one homeologs: 96.05% (SE±2.23%), 96.38% (SE±2.21%) and 96.56% (SE±2.12%) for AB, BD and AD subdivisions, respectively. Lower protein sequence identities (90.9%- 92%) were generally observed in the one/many-to-many categories for 'AB', 'AD' and 'BD'.
- 15 However, intra-sub-genomic homeologs of the one-to-many categories, i.e. tandem genes in 'AA', 'BB' and 'DD', showed higher similarities in all three subdivisions [92.2% ('DD'; SE±3.1%) to 93.8% ('BB'; SE±2.8%)]. Sequence identities at CDS levels exhibited a very similar trend (Fig. S22).

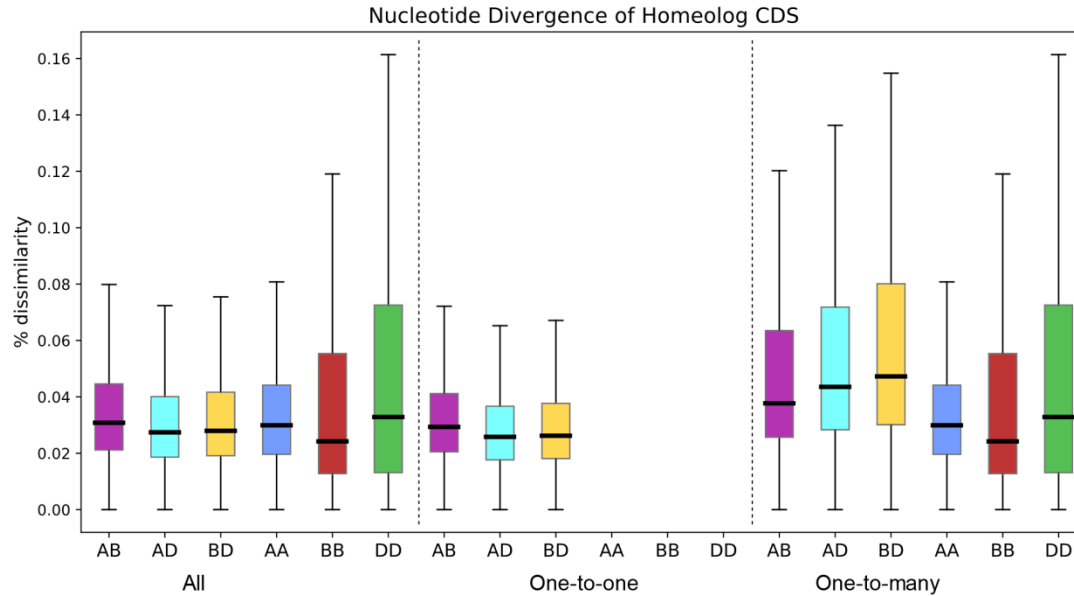


Fig. S22

CDS divergence of homeologous genes. CDS divergence is shown as percentage dissimilarity on the y-axis. Median values and spread (1st and 3rd quartile) for the three categories, 'all', 'one-to-one' and 'one-to-many' illustrated as boxplots. Categories are further divided into pairwise sub-genome comparisons, AB, AD and BD, and intra-sub-genome comparisons if applicable. One-to-one homeolog identities showed only small variation (96.3% ('AB'; $SE \pm 1.9\%$) - 96.8% ('AD'; $SE \pm 1.8\%$)), and were significantly higher than inter-sub-genomic one-to-many relations (92.8% ($SE \pm 2.8\%$)(BD), 93.3% ($SE \pm 2.7\%$)(AD) and 94.1% ($SE \pm 2.6\%$)(AB). Tandemly repeated homeologs were significantly more similar compared to the inter-sub-genomic gene pairs, with coding sequence identities ranging from 94.1% ($SE \pm 2.7\%$) for 'DD' to 95% ($SE \pm 2.7\%$) for 'AA' to 95.3% ($SE \pm 2.5\%$) for 'BB' subdivisions. In contrast to the protein divergence of the tandemly repeated homeologs in the one-to-many clusters which ranges between the one-to-one and one-to-many inter-sub-genome identities, nucleotide divergence of intra-sub-genome duplicated homeologs were comparable to the one-to-one relationships. These findings are supported by an increased level of non-synonymous substitutions and elevated KaKs ratios for tandem homeologs in comparison to the median ratios of homeologs between two sub-genomes in 1:N groups (Fig. S19). On the other hand, these latter subdivisions as well as 1:1 homeologs show higher Ks values compared to intra-sub-genomic homeologs of the one-to-many relations, indicating overall accelerated evolutionary rates and origins of tandem duplications occurring during the evolution of the hexaploid genome of bread wheat (Fig. S20).

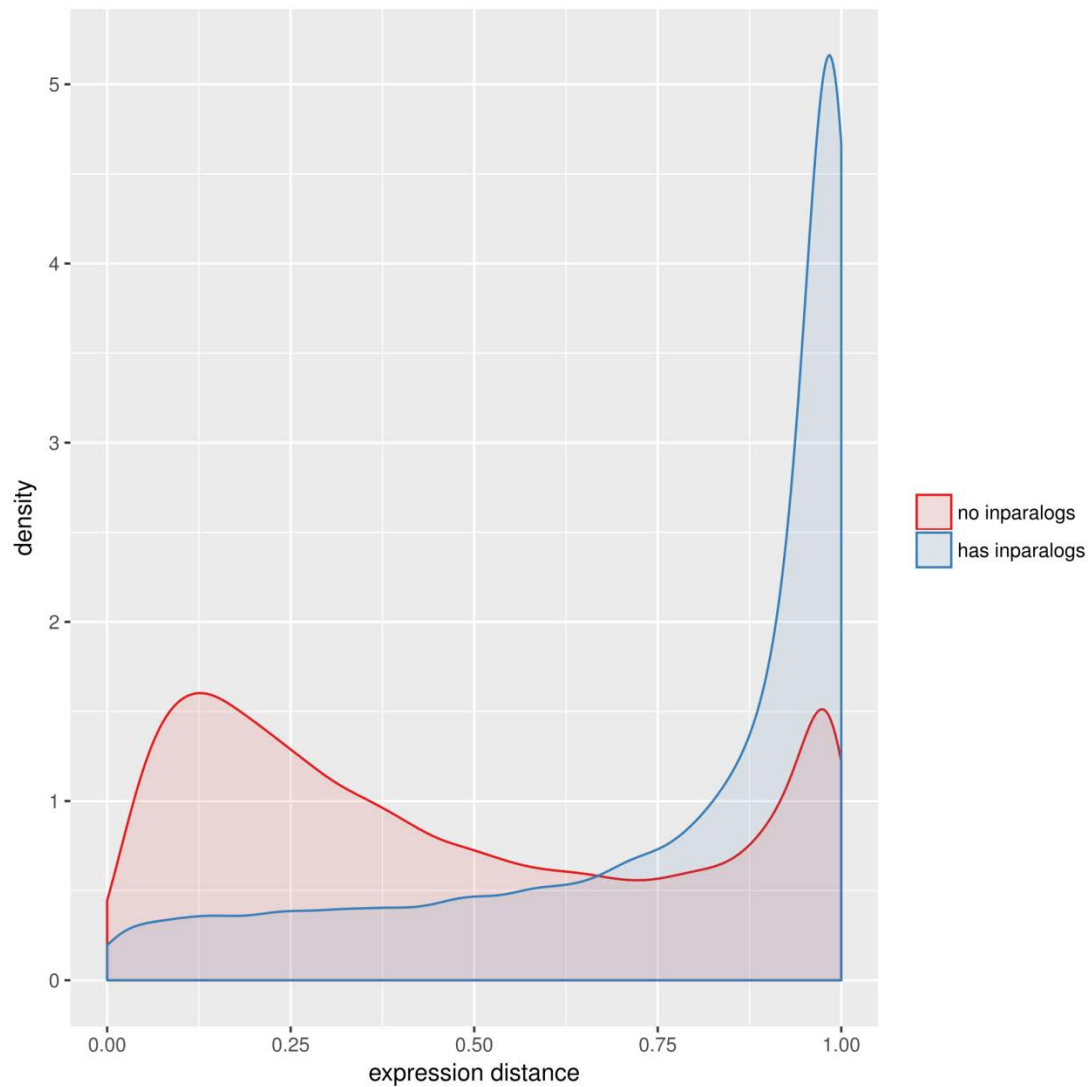


Fig. S23

Expression divergence of homeologous gene pairs with (blue) and without (red) sub-genome inparalogs (duplicates that arose after hexaploidization). Homeologous gene pairs with inparalogs exhibit a higher expression divergence. Shown is a density plot of the distribution of the expression divergence (x-axis) among homeologs. Expression divergence is expressed as the inverse of the bi-weight mid-correlations (bicor) between the expression vectors across all samples among a pair of homeologous genes ($1 - \text{bicor}$). Together with the patterns obtained for the mutation rates, these observations are consistent with gene duplication acting as an important motor of functional innovation enabling diversification of sub-genome inparalogs and resulting in sub- and neo-functionalization in addition to the dosage-effect provided by allohexaploidy.

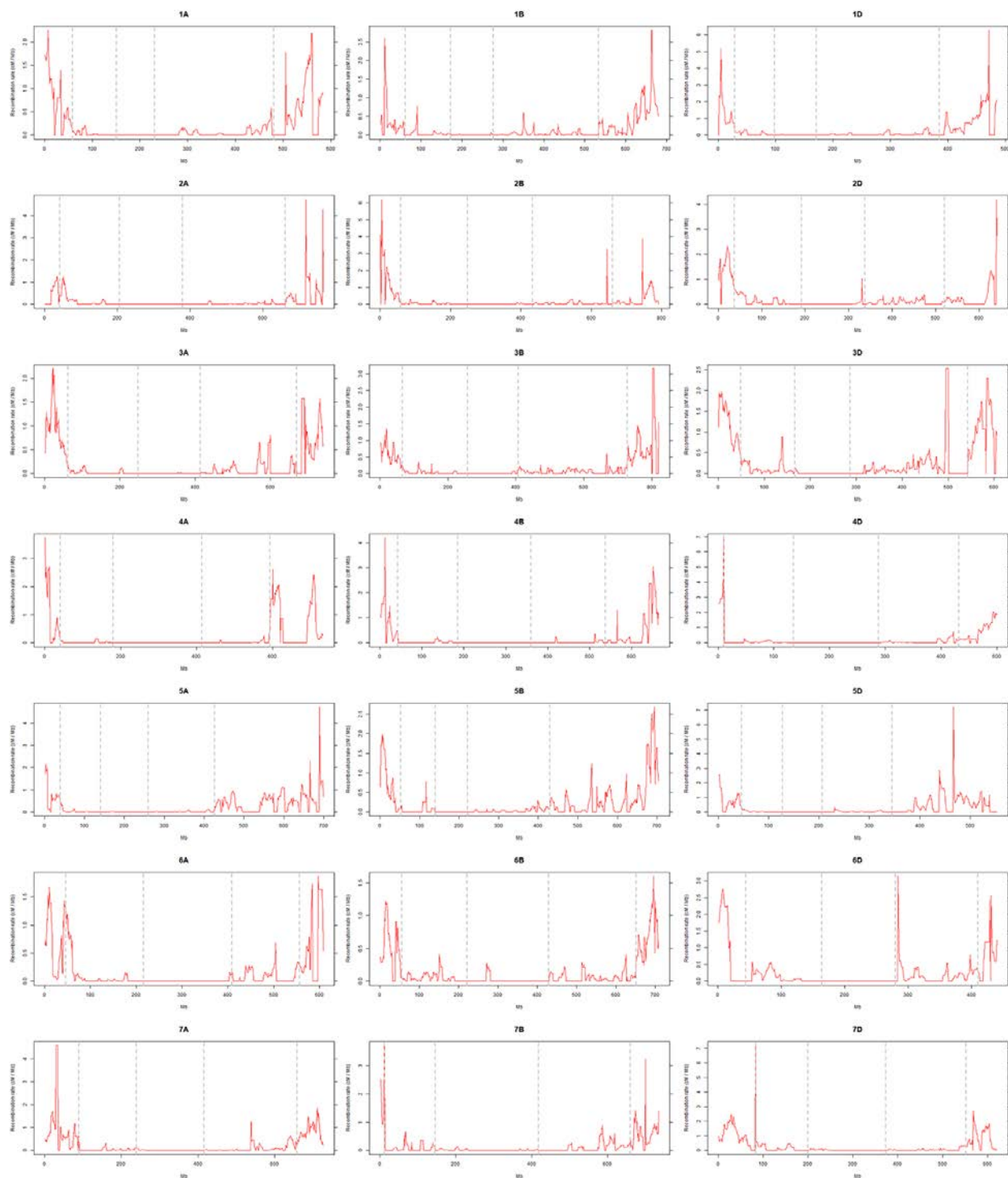


Fig. S24

Analysis of recombination rate distribution along wheat chromosomes. The CS x Renan (CsRe) map contained 146,602 SNPs genetically mapped in 21 linkage groups corresponding to the 21 chromosomes of bread wheat, with no unlinked markers. The D-genome contained the fewest mapped markers (18%), while the A- and B-genomes were similarly covered with 40 and 41% of mapped markers, respectively. Chromosome 3B has the most markers (10,638) and chromosome

4D (2,624) the least. The genetic map covers 3,592 cM with 25,385 unique genetic positions. The D-chromosomes had the longest genetic maps, with an average of 204 cM and a cumulative length of 1,427 cM, followed by the A-chromosomes (mean length = 164 cM; cumulative length = 1,146 cM) and the B-chromosomes (mean = 145 cM; cumulative = 1,018 cM). Average recombination rates (in red) were calculated in 10-Mb sliding windows (step 1 Mb) and estimated to be 0.26 cM/Mb for the whole genome (range 0.16 cM/Mb for chromosome 6B to 0.41 for chromosome 5D). The D-genome exhibited the highest recombination rate (0.36 cM/Mb) which reflects the low polymorphism level of this genome while A- and B-genomes had lower rates. Local rates varied between 0 and ~7 cM/Mb. Highest rates were observed for D chromosomes (1D, 4D, 5D, 7D) confirming that the D genome recombines more than the other two homeologous genomes. Cross-overs were found to occur mainly in the distal regions whereas most of the chromosomal proximal regions were recombination-poor. The X-axis represents the position on chromosomes (in Mb), the Y-axis the recombination rate (in cM / Mb).

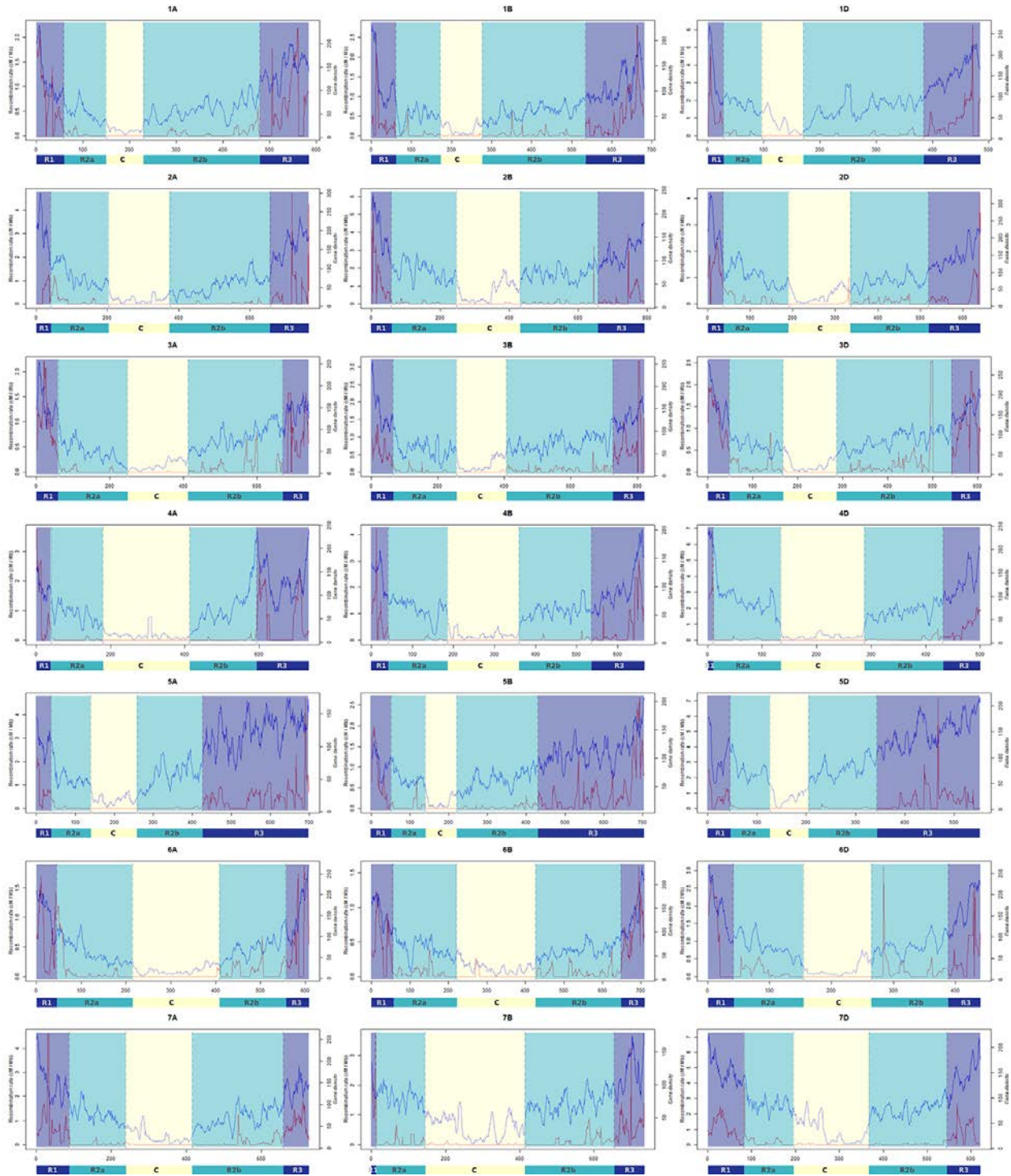


Fig. S25

Structural and functional partitioning of wheat chromosomes. The distribution of recombination rate (cM/Mb)(red line), gene density (genes/Mb)(blue line), together with TE density (%TE/Mb), expression breadth (number of conditions in which a gene is expressed) and proportions of genes associated with histone marks (% genes / Mb)(not shown) were determined using the R package *changepointv2.2.2* with Binary Segmentation method and BIC penalty on the mean change in

these features. Sliding window size: 10 Mb, step: 1 Mb. The five main regions of wheat chromosomes are represented: R1 and R3 (dark blue), R2a and R2b (dark cyan); C (yellow).

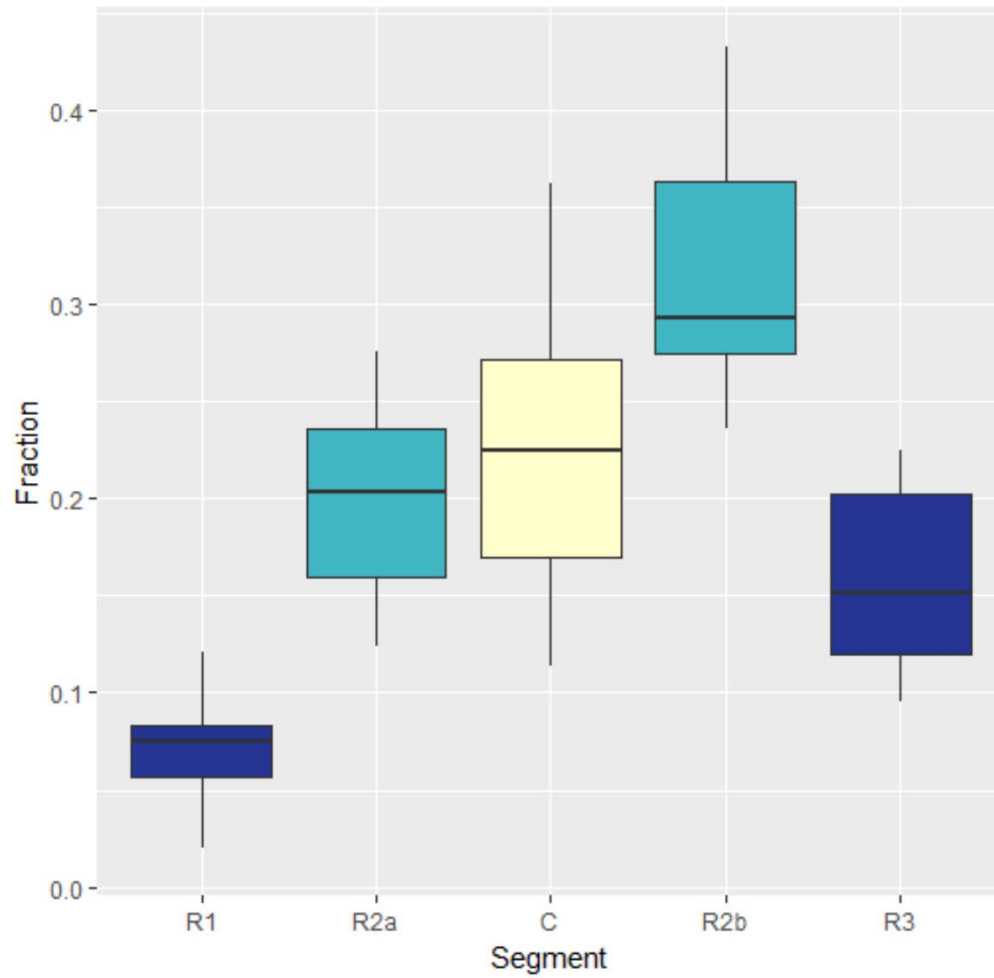


Fig. S26

Conservation of the genomic fractions covered by chromosomal segments R1 and R3 (dark blue), R2a and R2b (dark cyan) and C (yellow) across the genome.

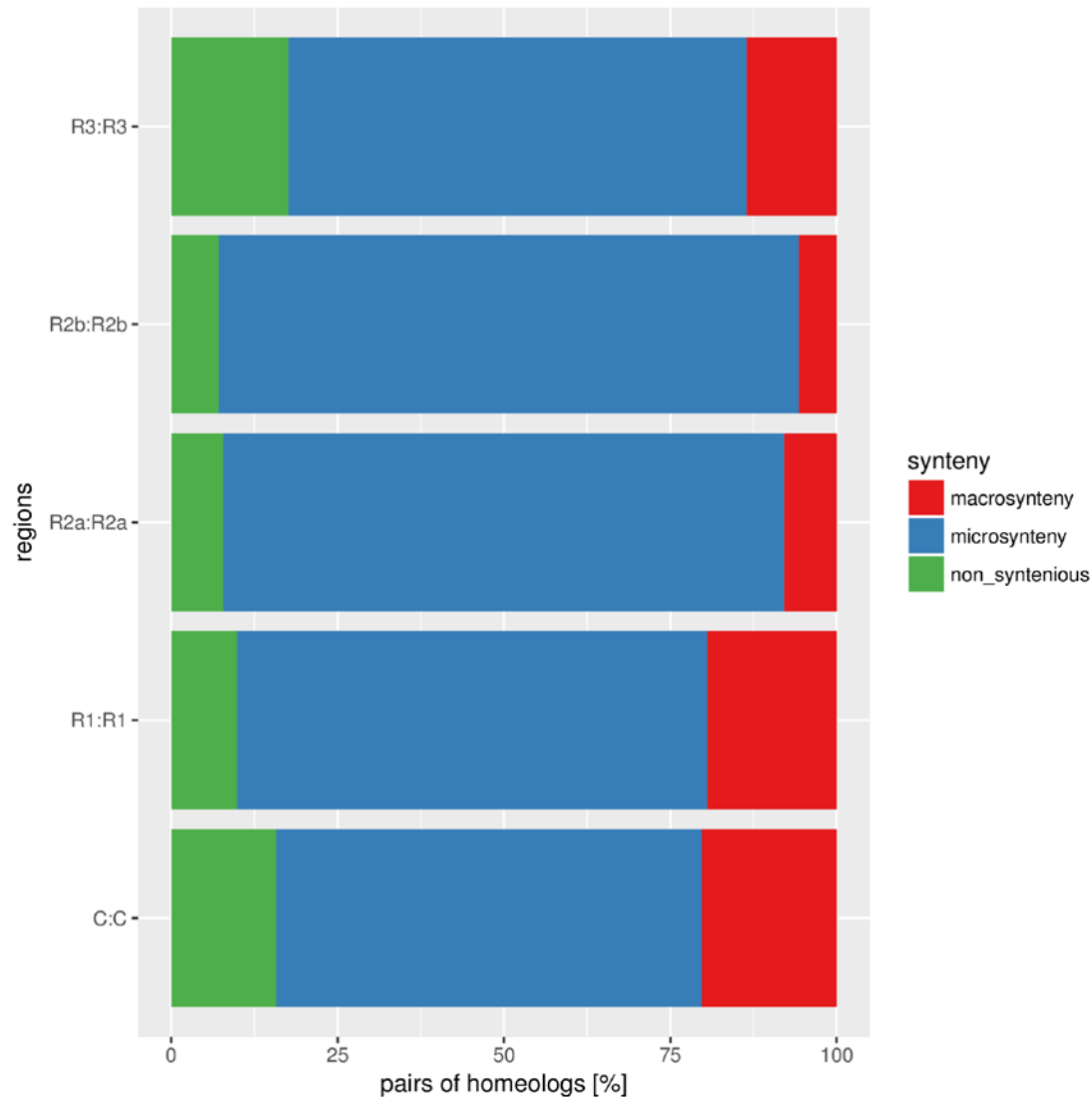


Fig. S27

Distribution of homeologous wheat genes in respect to chromosomal compartments. The evolution of homeologs in relation to their location in chromosomal compartments was assessed by comparing the percentage of genomic distribution of macrosyntenic (red), microsyntenic (blue) or non-syntenic (green) homeologous gene pairs in the distal R1 and R3 regions, the interstitial R2a and R2b regions and the proximal C region. Homeologous pairs located in the interstitial regions (R2a and R2b) showed higher levels of synteny compared to the distal (R1 and R3) and proximal regions (C). The comparison was limited to gene pairs located in the same compartment.

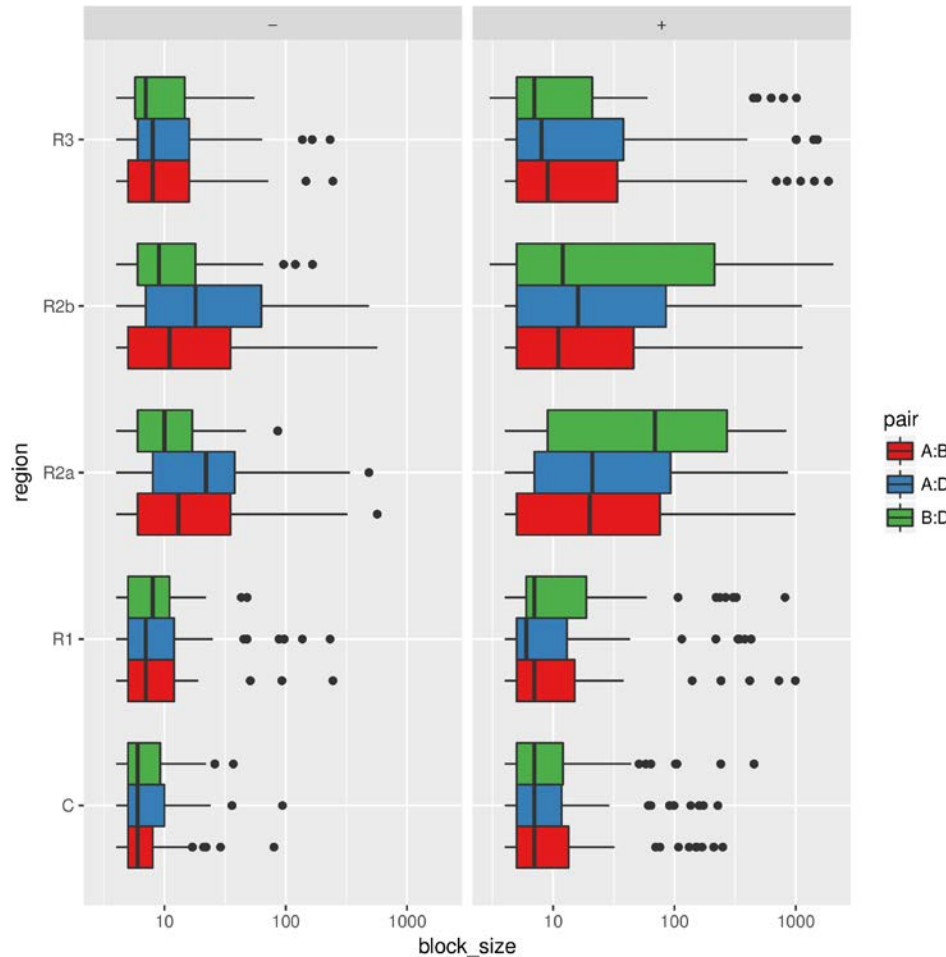


Fig. S28

Comparison of syntenic block sizes (number of syntenic genes) in same (+) or in inverse orientation (-) among the three wheat sub-genomes, separated for the chromosomal compartments (C, R1-R3). Box-whisker plots show the homeolog block sizes. Boxes represent lower (25%) and upper (75%) quartiles. The central line represents the median (50%) value. Outliers are plotted as individual dots, lines represent the upper and lower whisker. R2a and R2b compartments harbor larger syntenic blocks, with more microsyntenic genes, than the highly recombinant, distal R1 and R3 regions, although the latter are more gene dense. The R1 and R3 regions show similar ratios of homeologs in macrosyntenic or non-syntenic context as compartment C. Interestingly, distal and proximal regions of the chromosome harbor quite antagonistic features (distal: high recombination rate, gene and DNA transposon density and low LTR retrotransposon content; proximal: low gene density, low recombination and high LTR retrotransposon content). Although the distinct diversifying forces acting on both compartments might be acting on different timescales, surprisingly similar consequences for genome reorganization are observed. Further investigation of the size of syntenic blocks reveals an asymmetry of blocks sizes among the different sub-genome pairs particularly in the interstitial zones (R2a, R2b), with the sizes of syntenic blocks between B:D in R2a being significantly larger than A:D and A:B.

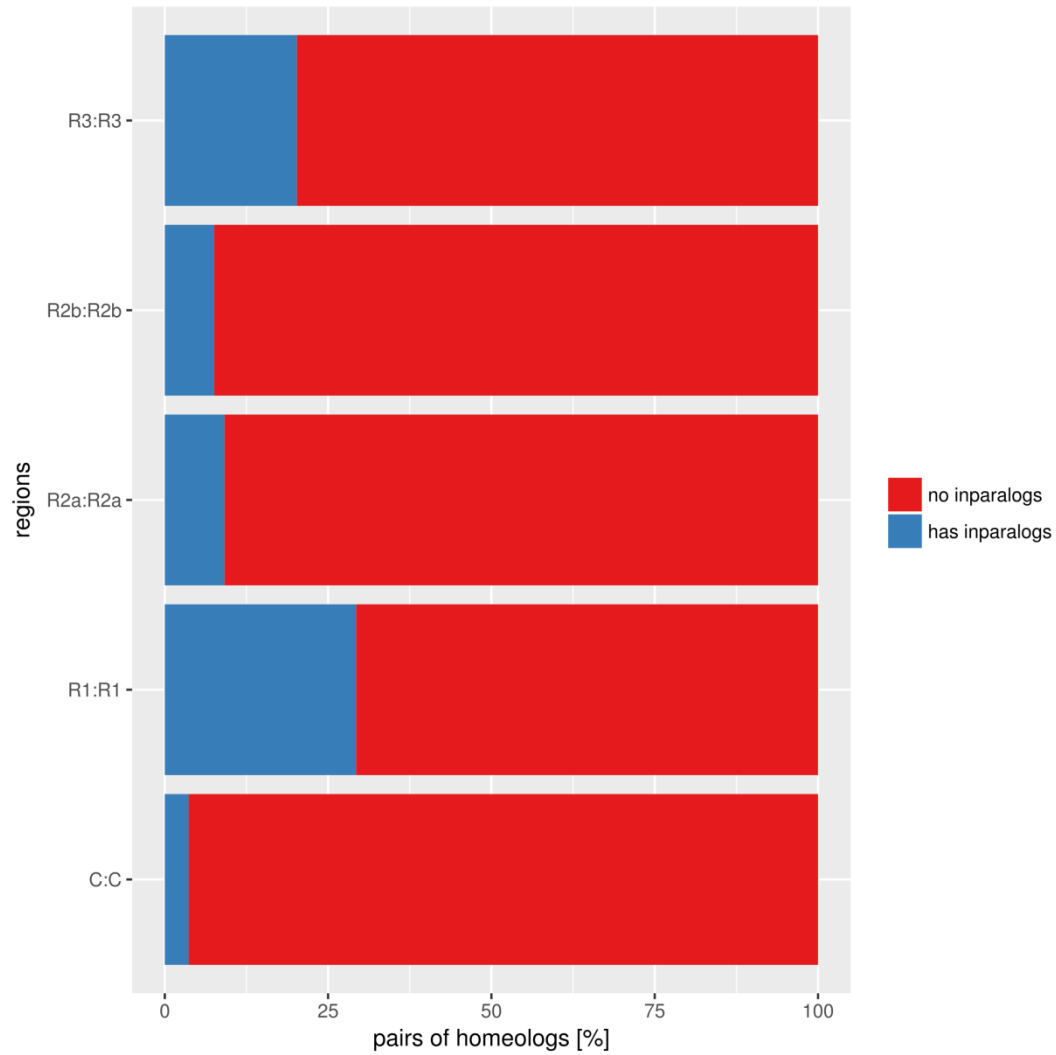


Fig. S29

Percentages of homeologous gene pairs with and without sub-genome inparalogs distributed to the different genomic compartments. Duplication of homeologs seems to occur at different rates in the different chromosomal compartments. Sub-genome inparalogs are much more frequent in the highly recombinant distal chromosomal regions R1 and R3.

5

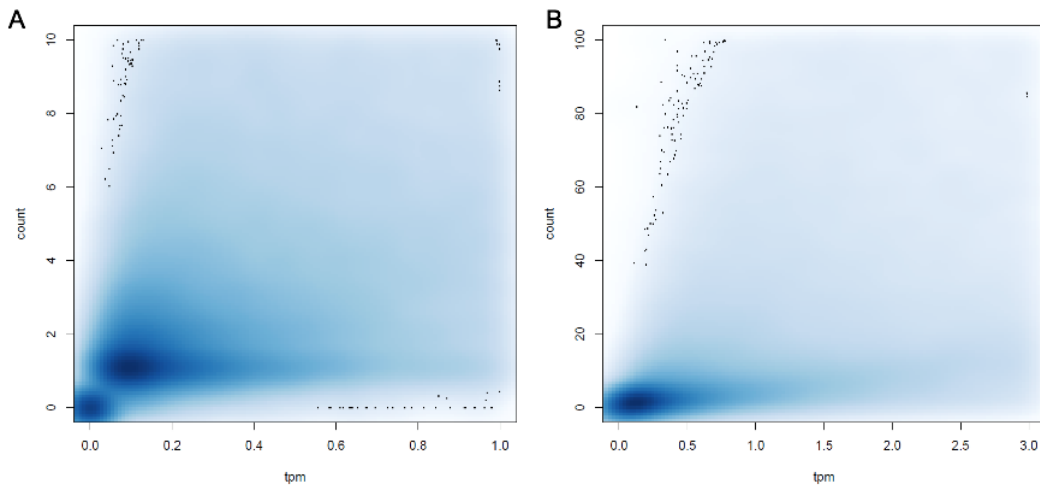


Fig. S30

Average expression abundance in transcript per million (tpm) and counts for the top 1% of samples from HC genes. (A) tpm values ranging from 0 to 1.0 are shown to illustrate clustering of non-expressed genes at the origin and the immediate vicinity. (B) tpm values ranging from 0 to 3.0 to show the wider context of very low expressed genes below the 0.5 tpm threshold used as the criterion for a gene to be considered expressed. Gene expression values are represented by the blue density plot, outliers (100 lowest density points) are indicated as black dots.

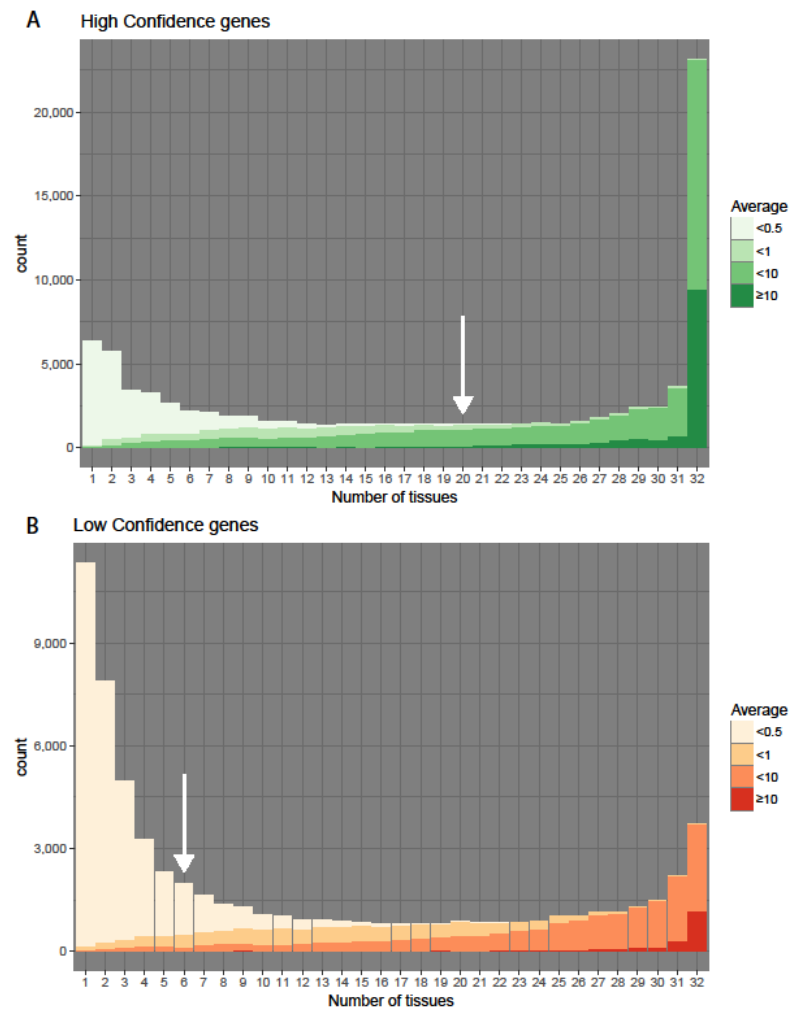


Fig. S31

Gene expression levels and number of intermediate tissues in which HC (A) and LC (B) genes are expressed. Arrows indicate the median number of tissues in which HC and LC genes are expressed.

5

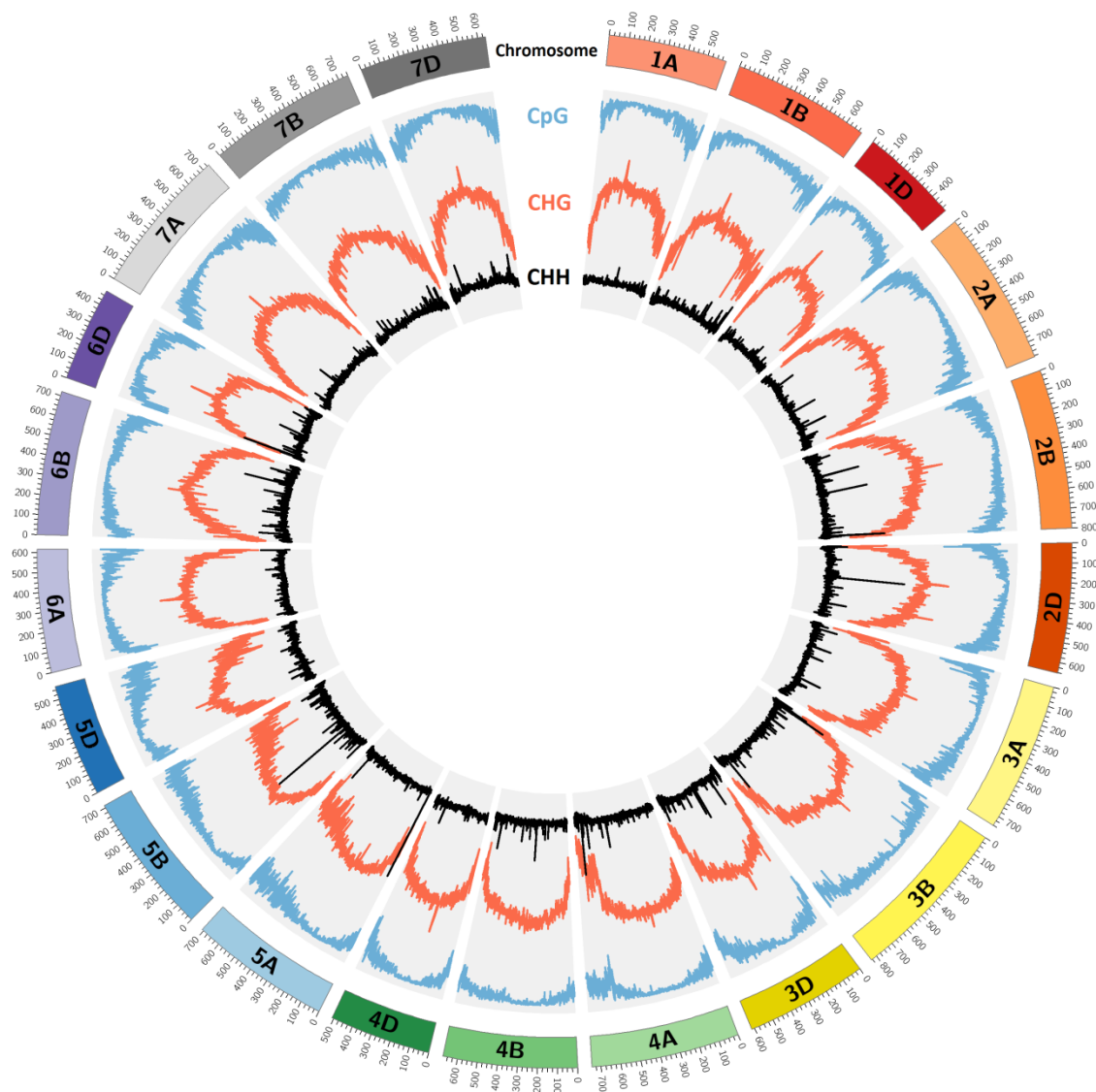


Fig. S32

DNA methylation distributions along wheat chromosomes. Wheat DNA methylation frequency of cytosines in the sequence contexts of CpG (average 92.7%, blue), CHG (average 51.3%, red) and CHH (average 2.7%, black) as revealed by whole genome bisulphite Illumina sequencing (WGBS) is shown from outside to inside in concentric circles of a Circos plot (160). The outermost circle is visualizing the 21 wheat chromosomes grouped according to homoeology. DNA methylation level frequencies are presented in 1 Mb windows. The observed levels of cytosine methylations are among the highest observed in angiosperms (161), likely reflecting the abundance of repetitive elements throughout the wheat genome. Methylation patterns in wheat largely follow those observed in other species, showing enrichment in CpG and CHG sequence contexts at pericentromeric regions (gene poor) and depletion toward the chromosome ends (gene rich).

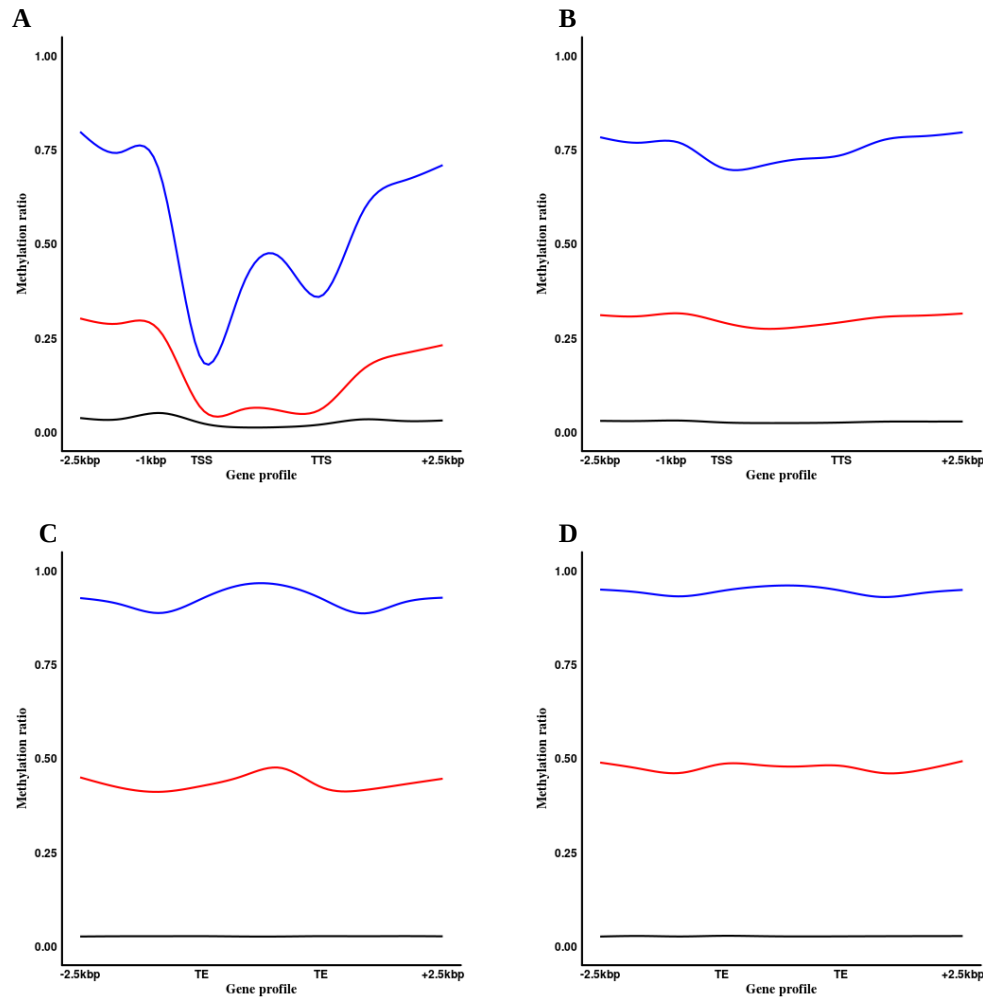


Fig. S33

Averaged DNA methylation profiles of annotated features in the wheat genome. Cytosine methylation in context of CpG (blue), CHG (red) and CHH (black) motifs is shown for A) high confidence (HC) genes (TSS = transcription start site; TTS = transcription termination site), B) low confidence (LC) genes, C) Copia (RLC) repeat elements (TE = transposable element start and stop sites), and D) Gypsy (RLG) repeat elements. High rates of DNA methylation likely serve to prevent transposition by restricting the expression of transposable elements. However, where repetitive elements are proximal to gene sequences, the enriched methylation can perform a regulatory function, predominantly silencing expression. The distinct and highly conserved methylation patterns observed in regions of HC genes and their regulatory regions showed higher levels of DNA methylation associated with the 5' regulatory regions in all contexts that diminished rapidly at the transcriptional start site (TSS). DNA methylation increased in the gene body where the CpG methylation formed a peak, whereas gene body methylation levels remained at extremely low levels at CHG and CHH sites. In the 3' regulatory region after the transcriptional termination site (TTS) methylation rapidly reverted to the levels in 5' sequences. This contrasted with the pattern observed for LC genes, where a near uniform level of methylation was observed in all sequence contexts. As a conclusion, many of the features

included in the LC annotation are either no genes, are truncated or have lost their function through mutation (i.e. pseudogenes).

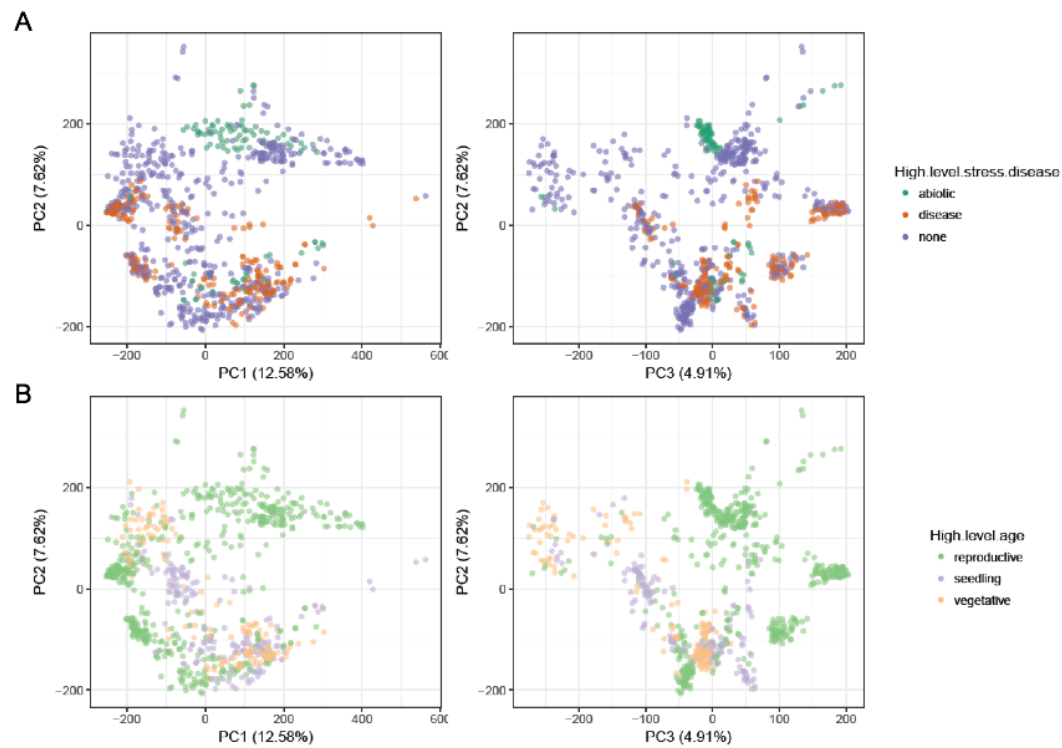


Fig. S34

Principal component analysis plots of the 850 RNA-Seq samples colored according to their high level stress/disease (A) or high level age (B).

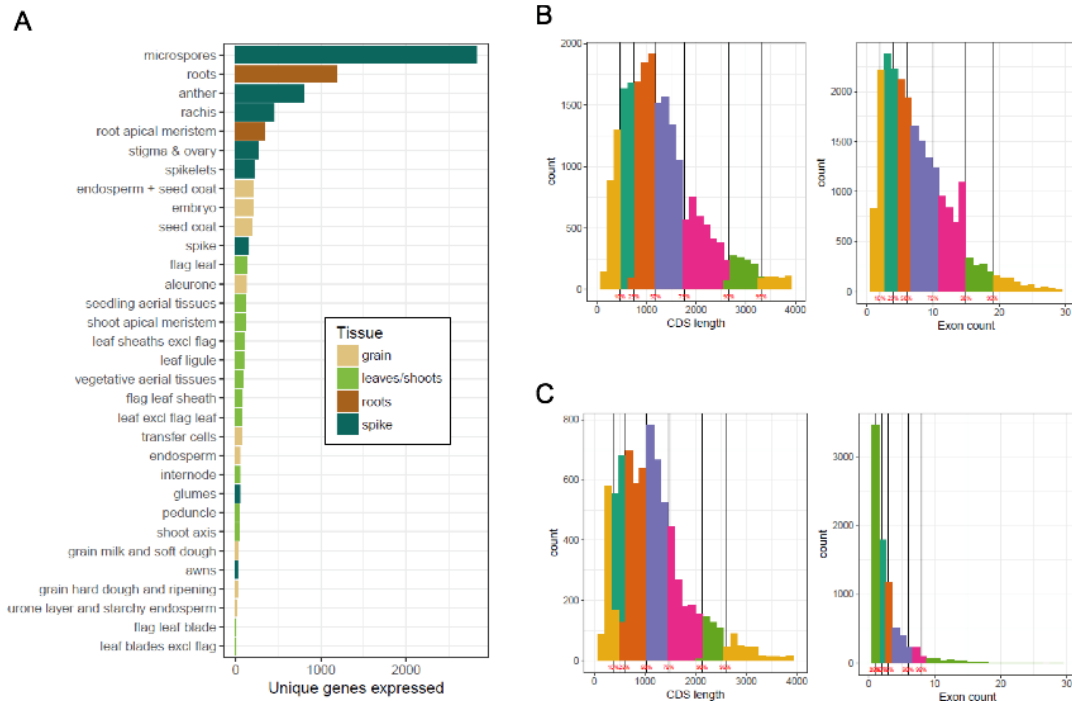


Fig. S35

Tissue specific gene expression in wheat. (A) Genes with tissue-exclusive expression across intermediate tissues. Comparison of coding sequence size and exon number between 23,146 ubiquitous (B) and 8,231 tissue-exclusive (C) genes. Bars are coloured according to quantiles which are defined by red percentages and solid vertical lines across each plot. From left to right quantiles are colored: 0-10% yellow, 10-25% green, 25-50% orange, 50-75% purple, 75-90% pink, 90-95% green, 95-100% yellow.

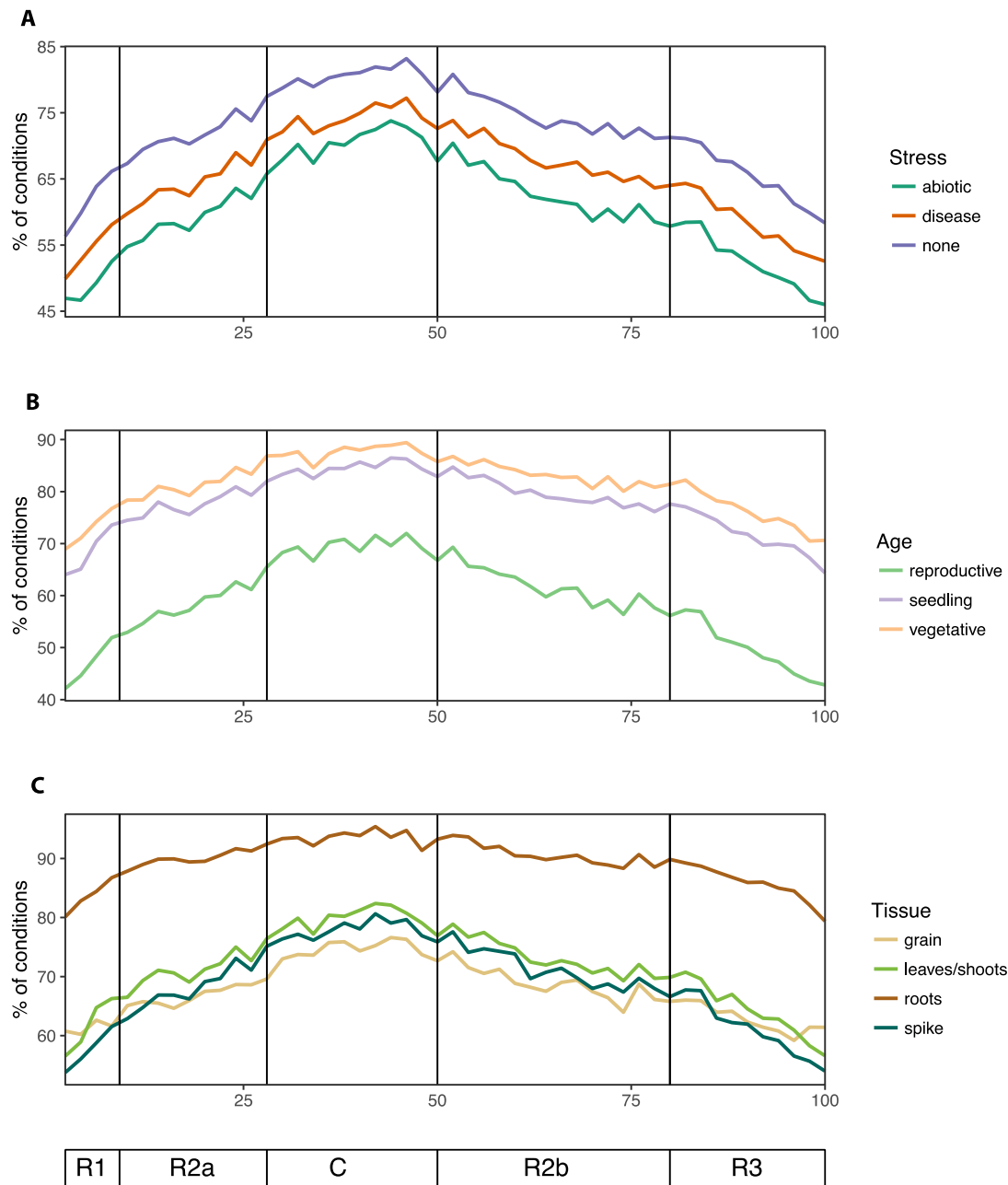


Fig. S36

Average percentage of conditions in which genes are expressed (expression breadth) based on physical position across wheat chromosomes. RNA-Seq samples were classified according to high level tissue (A), stress (B), and age (C) and the average expression across all 21 chromosomes was plotted based on their scale position within the corresponding genomic compartment.

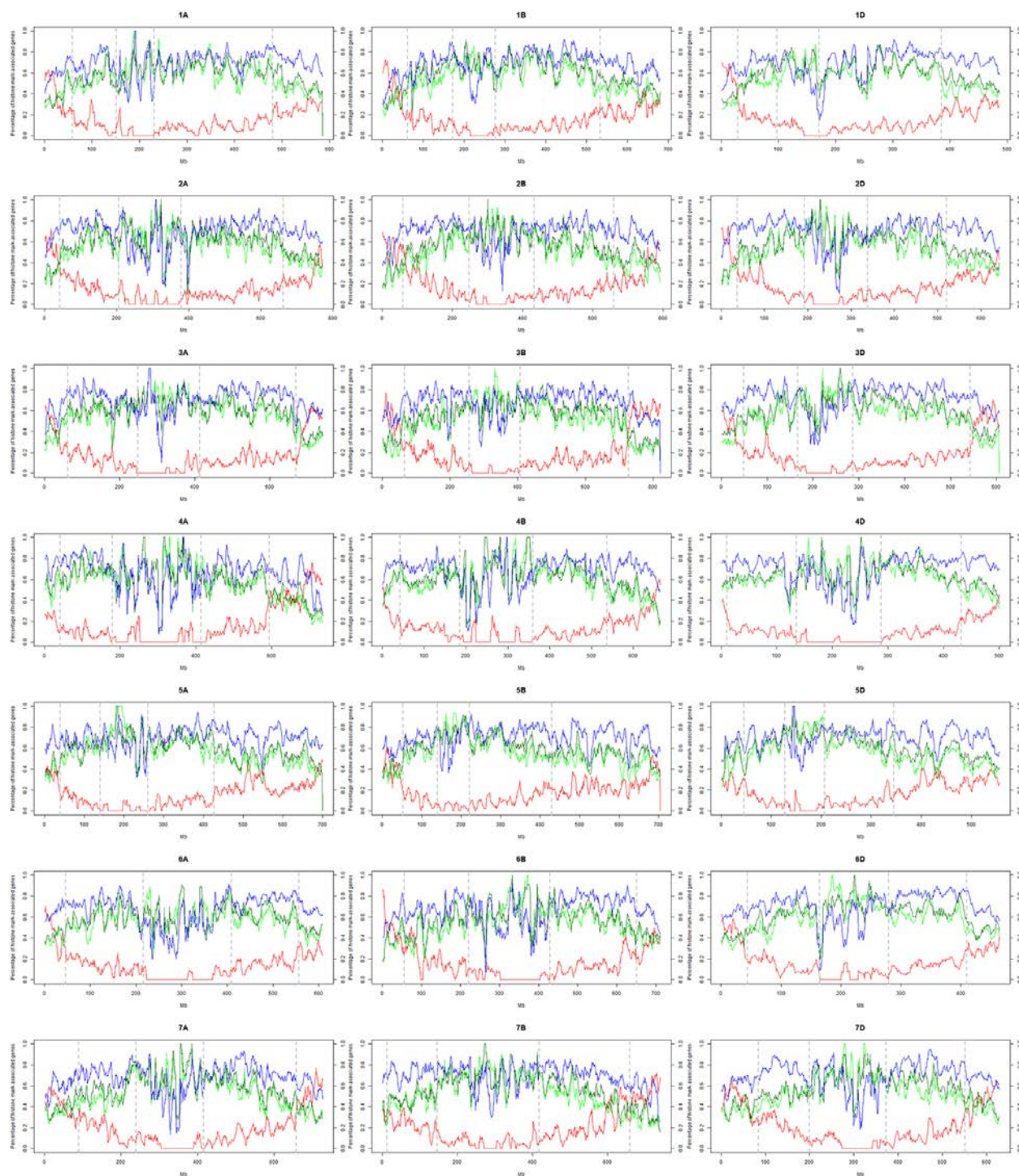


Fig. S37

Distribution of four histone marks along wheat chromosomes. Percentage of histone mark-associated genes are calculated in 10-Mb sliding windows (step 1 Mb). The distribution of histone marks along wheat chromosomes is following highly contrasting patterns. Whilst the repressive H3K27me3 mark (red) was enriched in the distal ends of the chromosomes with a pattern that is reminiscent to overall gene density, the active H3K36me3 (light green) and

H3K9ac marks (dark green) were more abundant within the proximal regions. In addition, the expression breadth pattern was correlated with the distribution of the repressive H3K27me3 ($R = -0.76$, $P < 2.2E-16$) and with the active H3K36me3 and H3K9ac ($R = 0.9$ and 0.83 , respectively, $P < 2.2E-16$) histone marks. The modification H3K4me3 is shown in blue. The X-axis represents the physical position on chromosomes.

5

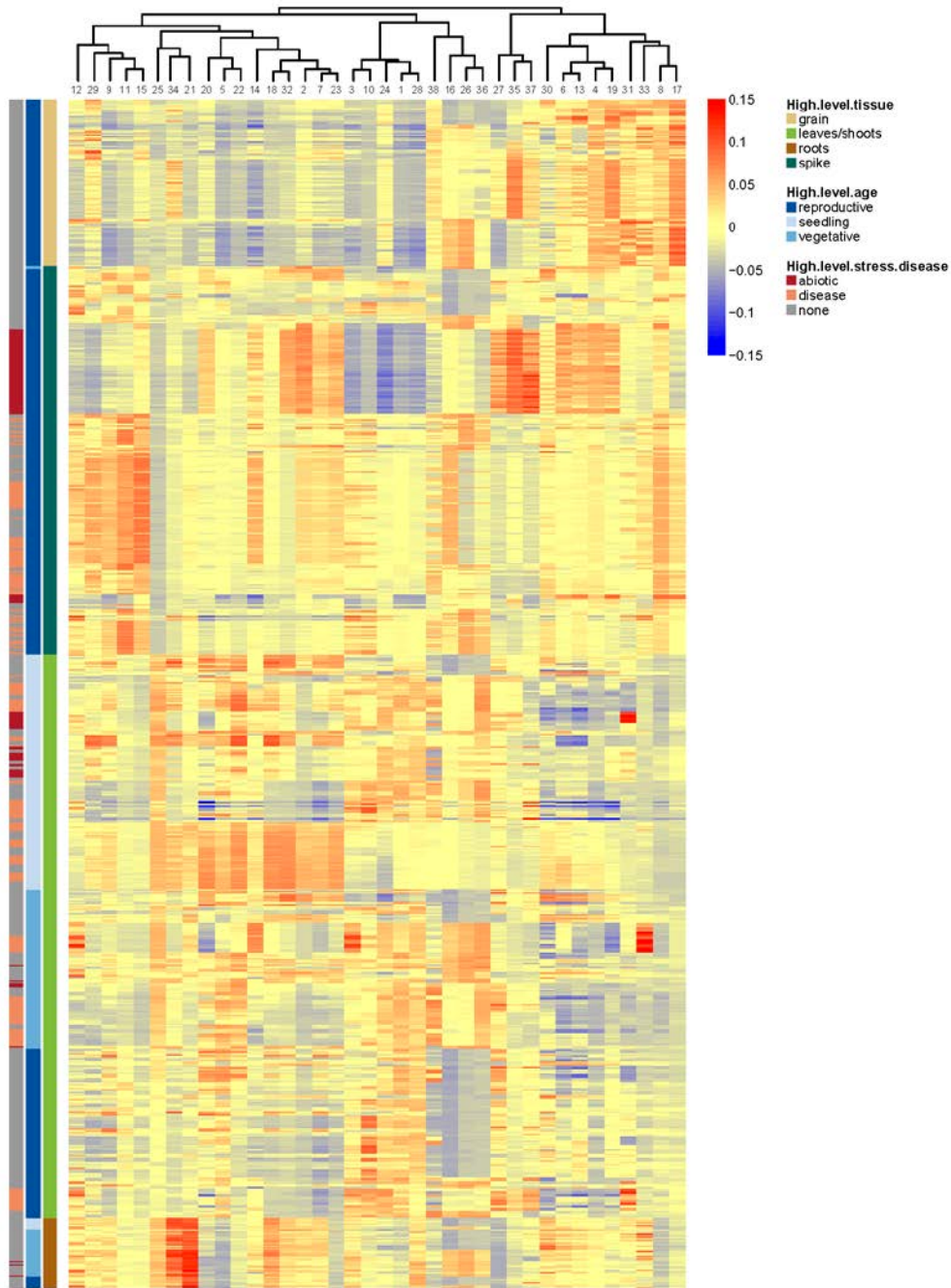


Fig. S38

Heatmap illustrating the expression of a representative gene (eigengene) per module. Modules are represented as columns, with the dendrogram illustrating eigengene relatedness. Each row represents one sample and the coloured bars beside the heatmap indicate the High Level Tissue, Age and Stress from which the sample originated. Expression levels were normalised by DESeq2 variance stabilising transformation. Values >0.15 or <-0.15 were capped at 0.15 or -0.15, respectively. This capped 16 out of 32,300 values (12 values >0.15 and 4 values <-0.15).

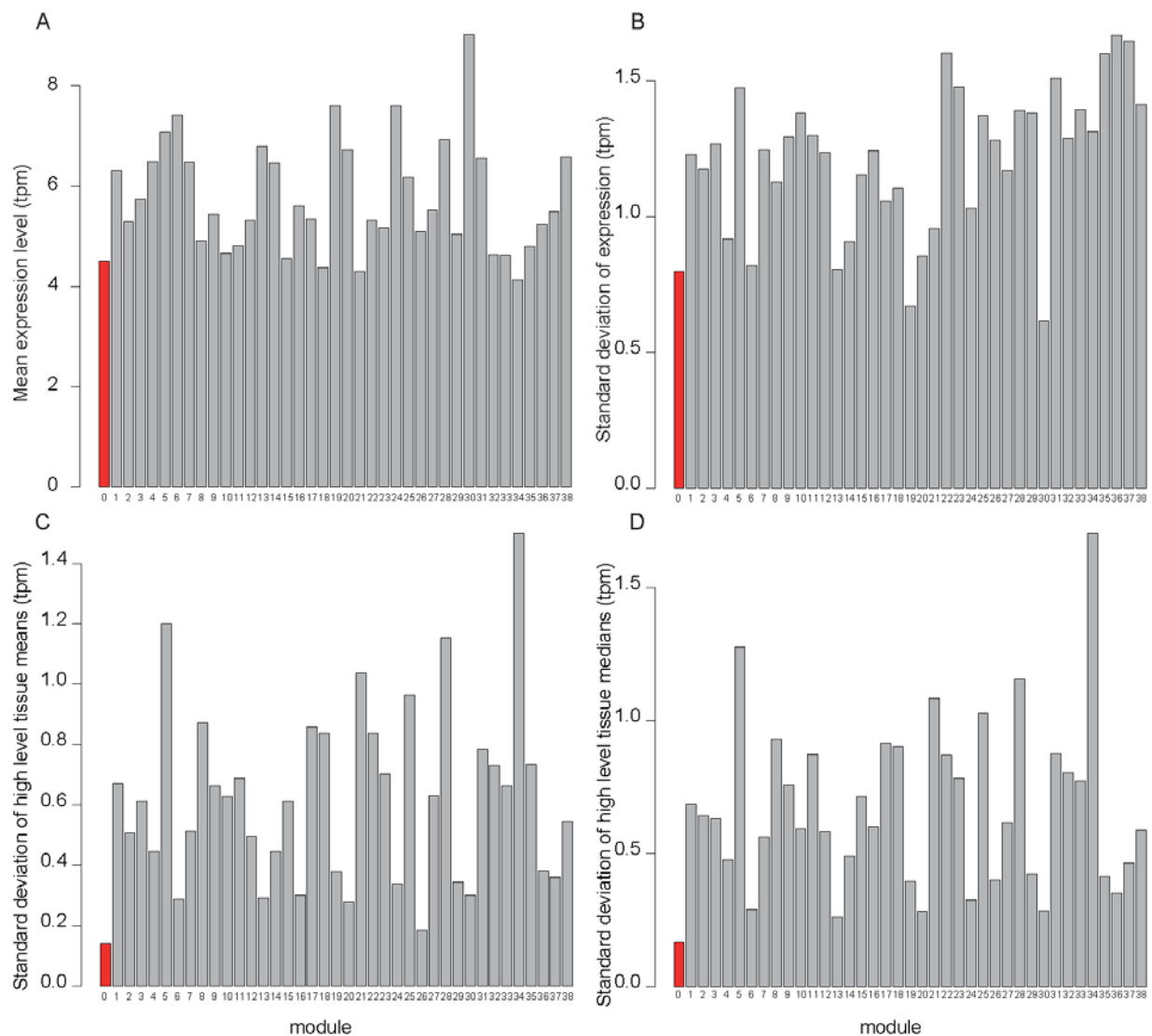


Fig. S39

Characteristics of genes assigned to WGCNA modules compared to unassigned genes (module 0, red). Mean expression level (tpm) (A) and standard deviation of expression (B) for all genes in each module. Standard deviation of the mean (C) and median (D) values calculated for each module for high level tissues. Unassigned genes (module 0) tend to be expressed at lower levels and have reduced variation in expression across high level tissues.

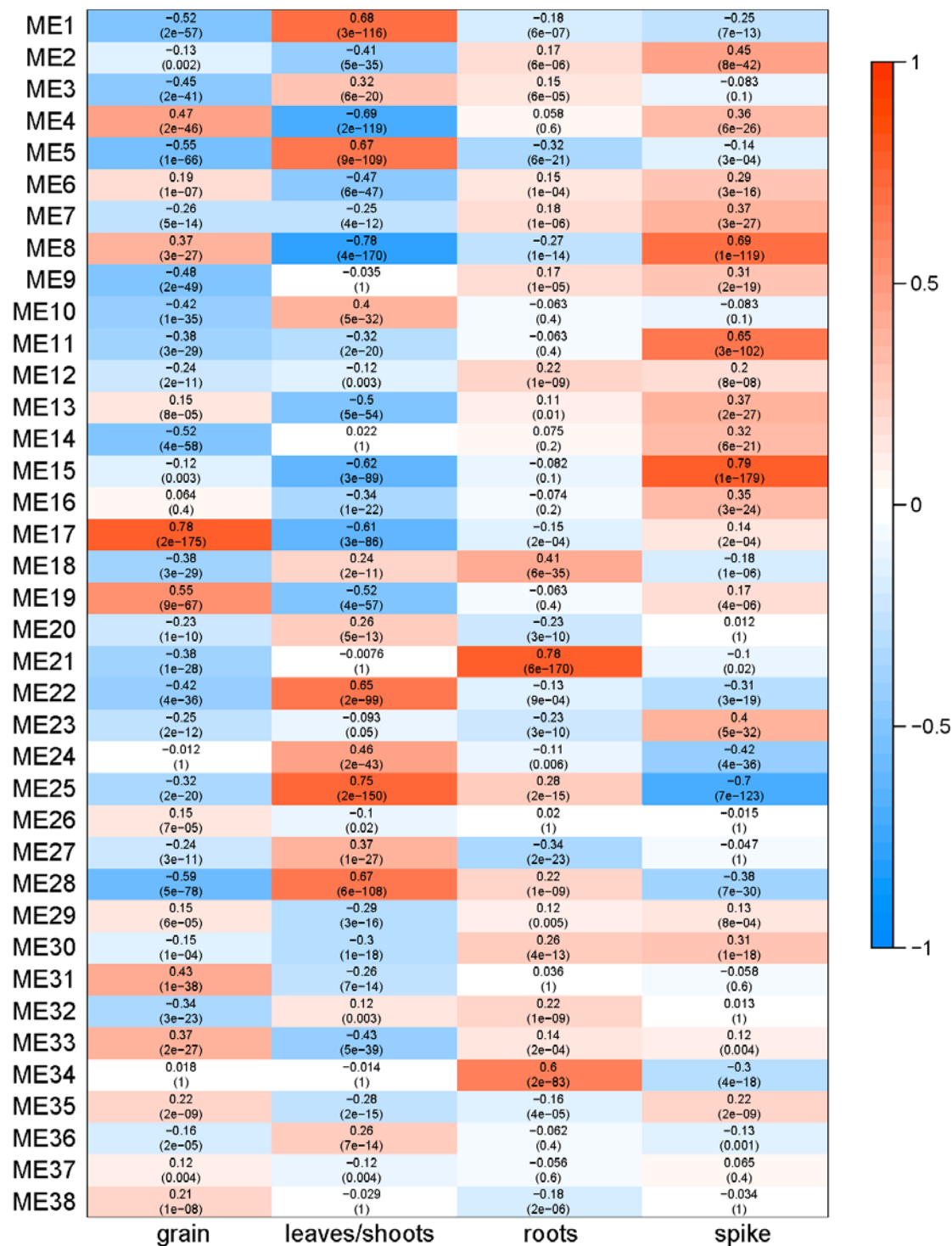


Fig. S40

Correlation between the module eigengene (kME; representative gene expression pattern) and the high level tissue for each module in the WGCNA co-expression network.

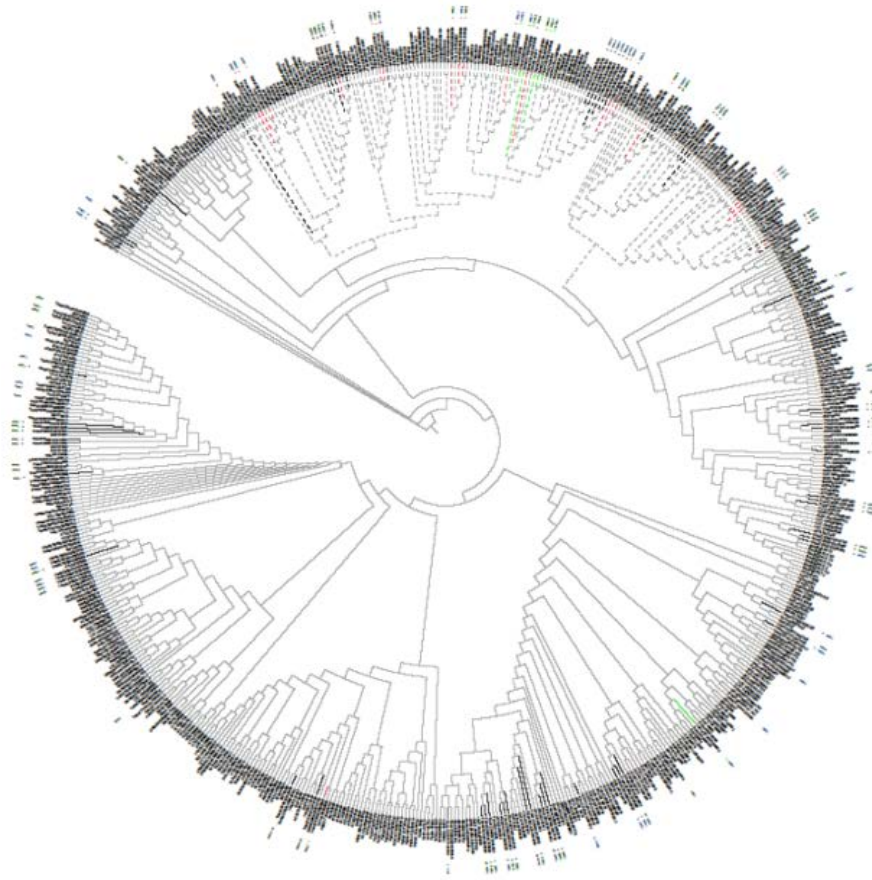


Fig. S41

Phylogenetic tree for MADS transcription factor gene family OG0000041. Red branches are genes within module 8, green branches are genes within module 11 and black branches are genes in other modules. Grey branches indicate non-wheat genes or wheat genes not allocated to a module in the WGCNA network. The dotted branches are the clades in which Arabidopsis and rice orthologues were found to be involved in flowering regulation. Numbers around the outside of the phylogenetic tree indicate the module to which the gene was allocated and the gene family is described in the outer ring. For a high resolution image see iTOL:

<http://itol.embl.de/shared/borrillp>

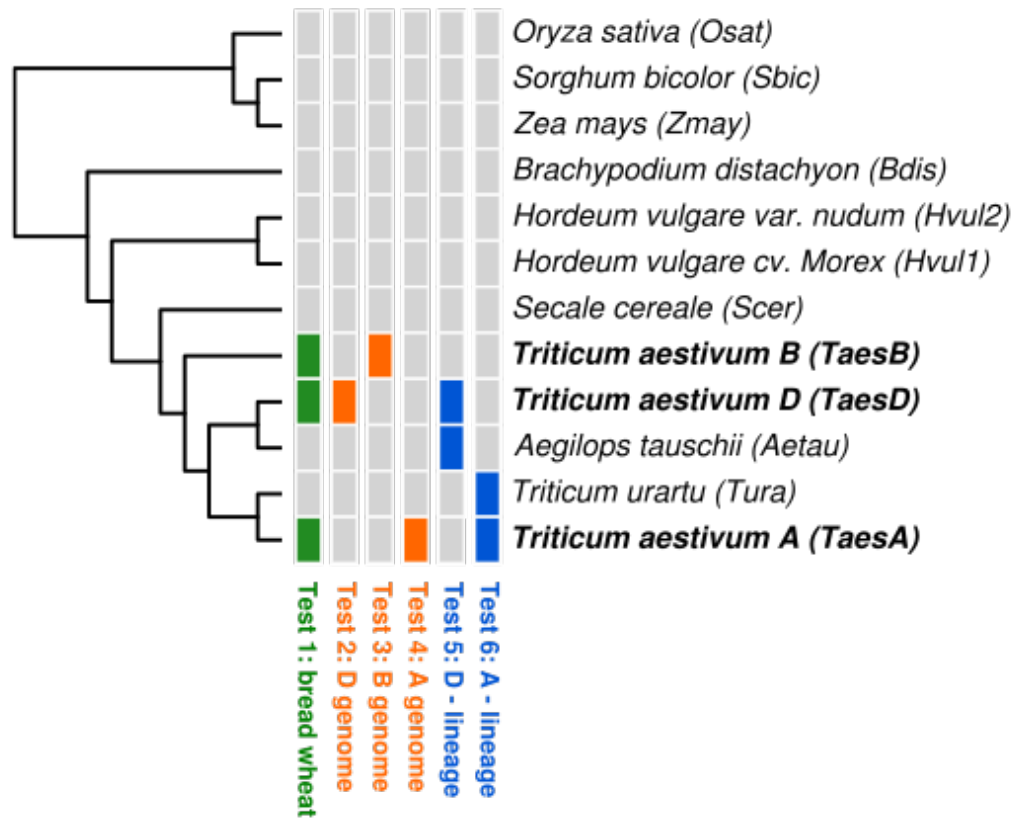


Fig. S42

Taxon sets compared in the phylogenetic comparative ANOVA (phyloANOVA) of gene family sizes. The cladogram on the left depicts the phylogenetic relationships of the 12 genomes under comparison (3 bread wheat sub-genomes and the genomes of 9 other grasses). The 6 columns in the colored matrix between the cladogram and the taxon labels depict each of the phyloANOVA tests used to infer significant expansions and contractions in the 26,080 gene families. Each set of colored cells (green, orange and blue) indicates the taxon set whose log-scaled gene family sizes were compared with the other grass taxa (grey cells). P-values of ANOVA test statistics were corrected for multiple testing by using the false-discovery rate (162). T-values of phyloANOVA tests with $FDR < 0.1$ were subsequently used to infer expansion ($t > 0$) and contraction ($t < 0$) of gene families in the respective genome set (Tests 1-6).

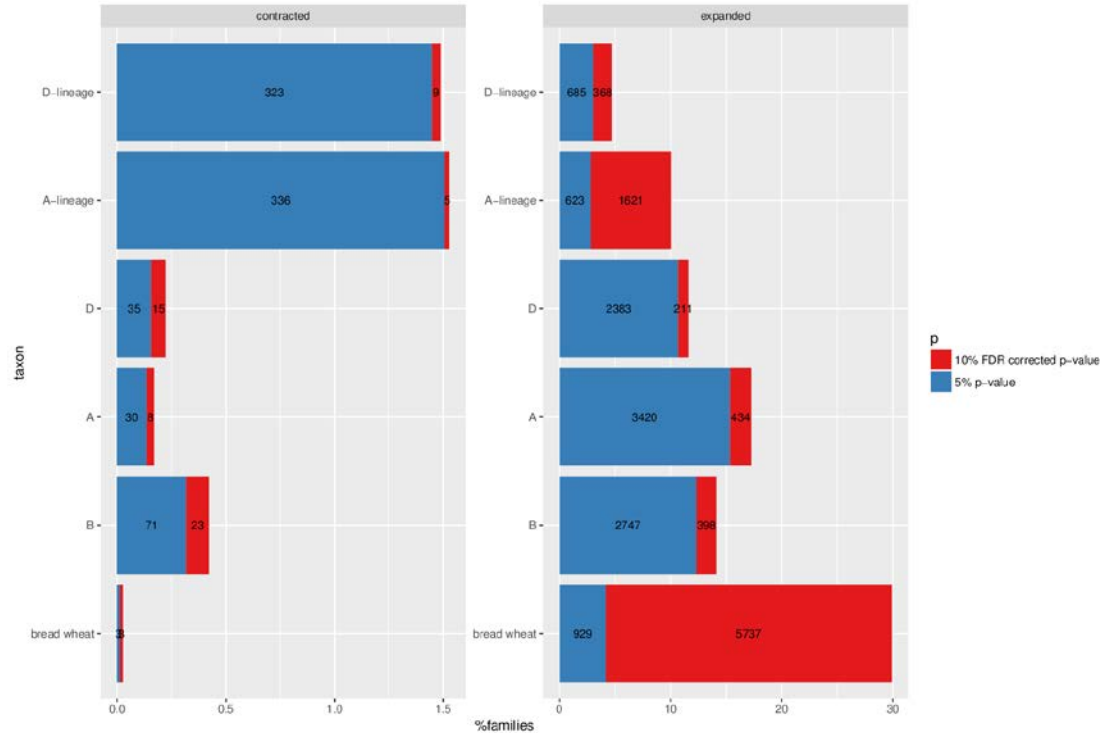


Fig. S43

Expansion and contraction of gene families in the bread wheat genome. Stacked barplot of the percentages of significantly contracted (left panel) or expanded (right panel) gene families [95% confidence level based on the uncorrected p-values (blue), 90% confidence based on the FDR-corrected p-values (red)]. Absolute numbers of gene families in each set are annotated as numbers in the bars. The values in each of the stacked columns are additive – families with a FDR-corrected p-value<0.1 also have a p-value <0.05. Only non-TE families and families with consistent domain architectures are shown. Gene families comprising only pseudogenes were excluded. Up to 25% of all gene families were equally expanded in all sub-genomes of bread wheat in comparison to the other grasses, while maximally 10-14% (D and A) of all gene families are expanded in one of the sub-genomes (using uncorrected p<0.05). Considering type I errors in (multiple) hypothesis testing and correcting for FDR, reduces these numbers to 21% and 1-2% (FDR<0.1). While the size changes that were only detectable using the uncorrected p<0.05 cutoff might still contain losses/gains of single loci representing size fluctuations without selective consequences, the FDR<0.1 should provide a more robust measure to pick up changes that go beyond these fluctuations, possibly reflecting selective processes and adaptive traits. All large-scale analyses reported in the main text rely on the sets defined at 10% FDR. The majority of recorded expansions affected the sub-genomes similarly (5,737 expanded non-TE, non-pseudogene families at FDR<0.1). The statistically significant differences between the sub-genomes in terms of sub-genome-specific expansions (χ^2 82.206, p-value < 2.2e-16; A>B>D) and contractions (χ^2 7.3478, p-value = 0.02538; B>D>A) are largely consistent with a scenario where the progenitors diverged 5-3 mya (Fig. S15); the A and B sub-genomes coexisted for a longer time period (1-0.5 my) before the tetraploid A:B progenitor hybridized with the D progenitor about 10,000 years ago. The A genome-lineage-specific expansions (1,621), were

greater than the number of expansions in the D genome and more losses and fewer gains were observed in B than in A, a possible consequence if the B progenitor was the maternal donor in the initial tetraploidization event. Employing the 10% FDR cutoff, there was no overlap between the sets of contracted and expanded gene families. By relaxing the criteria to $p < 0.05$, a maximum overlap of 31 gene families was found expanded in A and contracted in B. All other set intersections between the sub-genomes were either zero or do not go beyond four families. This suggests that unbalanced translocations among sub-genomes did not play a major role in the sub-genome-specific expansions, but processes like trans-duplication were more likely to give rise to the gene family expansions observed in bread wheat.

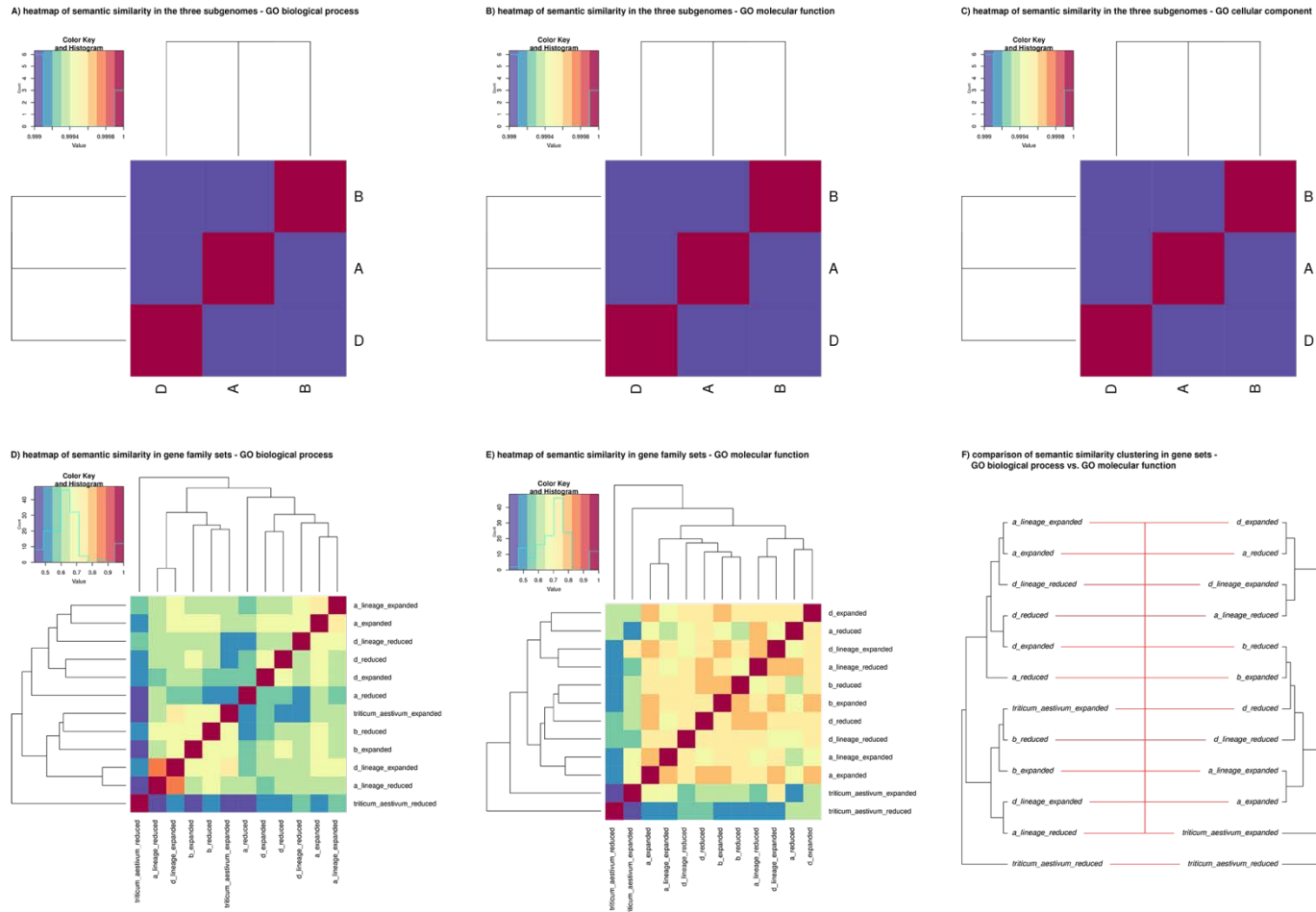


Fig. S44

Functional similarity of the three wheat sub-genomes (A-C) and the different sets of genes encoded by the expanded and contracted gene families (D-F) as measured by semantic similarity of Gene Ontology (GO) term annotations. Semantic similarity provides a quantitative measure for functional similarity of genes by measuring the distances of their assigned GO terms in the ontology graph.

Concepts that do not share any functional overlap have a semantic similarity of 0, highly related concepts show values close to 1 and the identical concept has a similarity of 1. Upper row: Heatmap matrices depicting the semantic similarity (163) of the three wheat sub-genomes for the GO biological process (BP; A), molecular function (MF; B) and cellular component (CC; C) category. Lower row (D-E): Heatmap matrices depicting the semantic similarity of the subsets of genes in significant expanded and contracted gene families (Fig. 4B) in the GO BP and MF category. F) Juxtaposition of the hierarchical clustering of semantic similarities of the gene family sets in the BP (left tree) and MF (right tree) GO category. The topology of the two trees often groups expansions in one sub-genome with contractions in another, indicating how losses in one are often balanced by gains in another sub-genome. Large numbers of significantly enriched Gene Ontology and Plant Ontology terms (1,169 distinct terms) were identified indicating that expanded families are involved in many aspects of wheat biology, morphology and development. Similarly, no distinct functional or morphological categories dominated the list of enriched GO and PO terms (272 distinct terms) for contracted gene families. But the assessment of functional overlap between the expanded and reduced gene family sets, as well as the overall gene complements of the sub-genomes indicated high overlap in terms of semantic similarity (0.999) for the sub-genomes and high semantic similarities between expansions in one genome and reductions in others, which overall suggests balancing of losses and gains potentially driven by gene transfer between the sub-genomes.

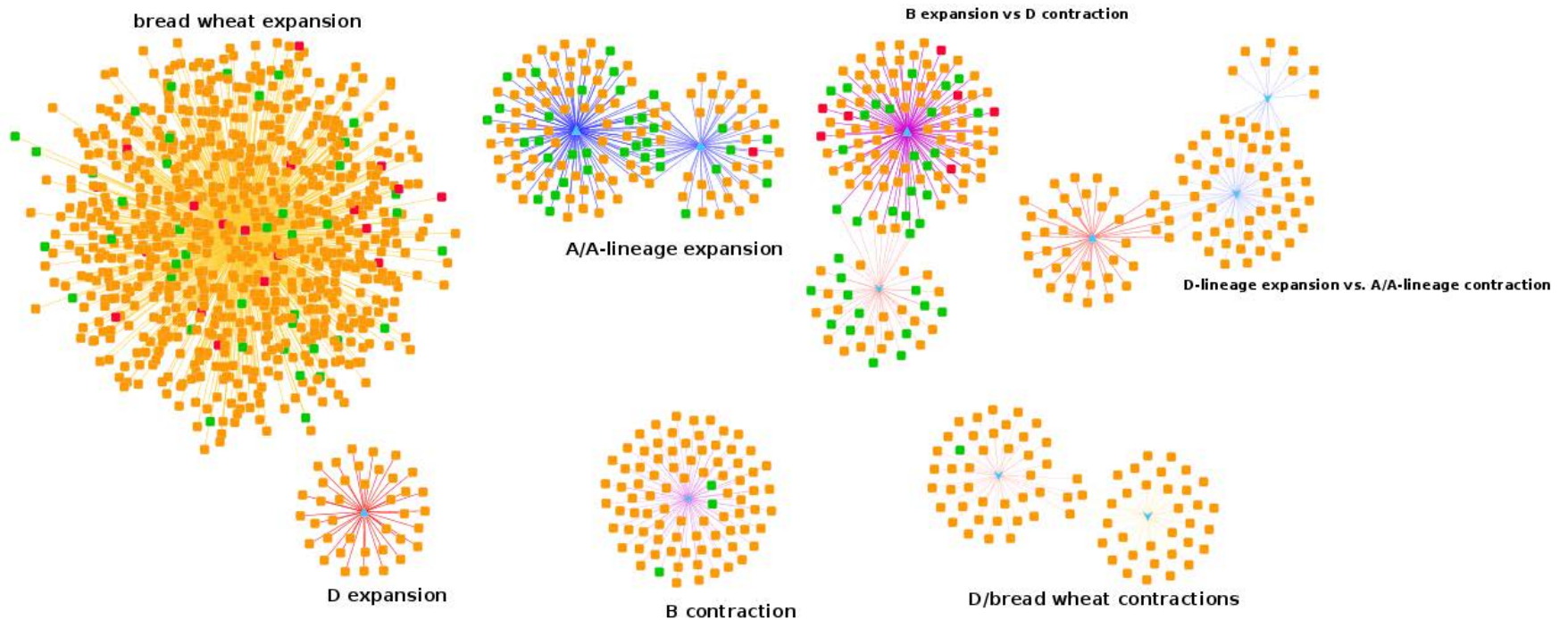


Fig. S45

Ontology term network for enriched terms among expanded and contracted wheat gene families. Each set is depicted as a blue, central node (upward facing triangles=expanded families; downward facing triangle=contracted families) linked to ontology terms [GO (orange nodes), PO (green nodes) and TO (red nodes)]. Overlapping enriched ontology terms were clustered using GLAY community clustering and the resulting subnetworks represent sets with significant conceptual overlap. Expanded views of the clusters are shown in Figs. S46-51. While the overall grouping of the community clustering reflects the results based on semantic similarity analysis of GO terms and is in line with overall balancing of losses and gains, some mild specialization can be inferred from the term composition of the sub-genome-specific expansion sets (D-genome and lineage: defense response, B-genome: plastid and housekeeping; A-genome and lineage: vegetative growth).



5

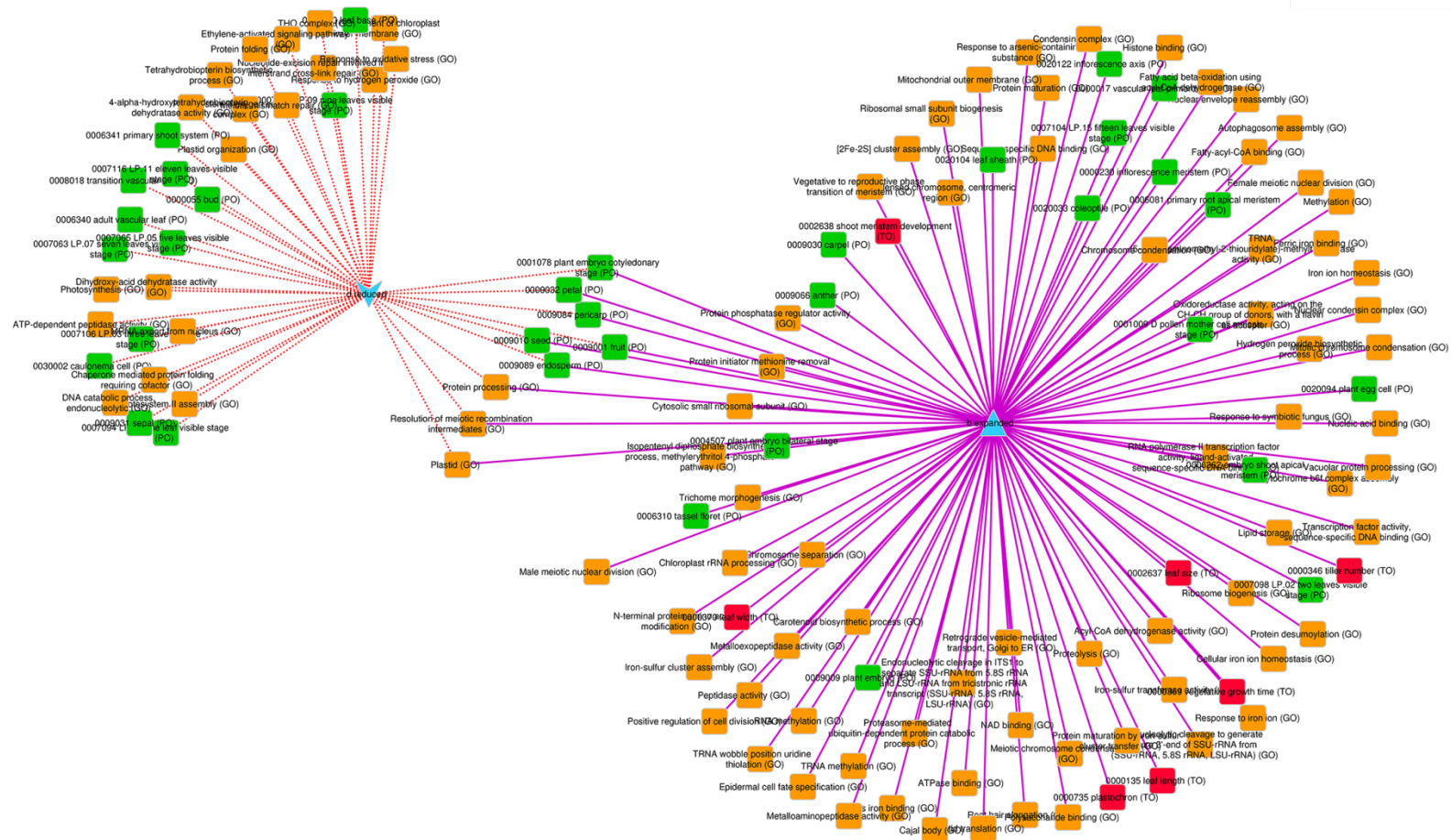


Fig. S47

Enriched ontology terms for the gene sets encoded by gene families contracted in the D genome and expanded in the B genome. The two sets share six PO terms and three GO terms.

5

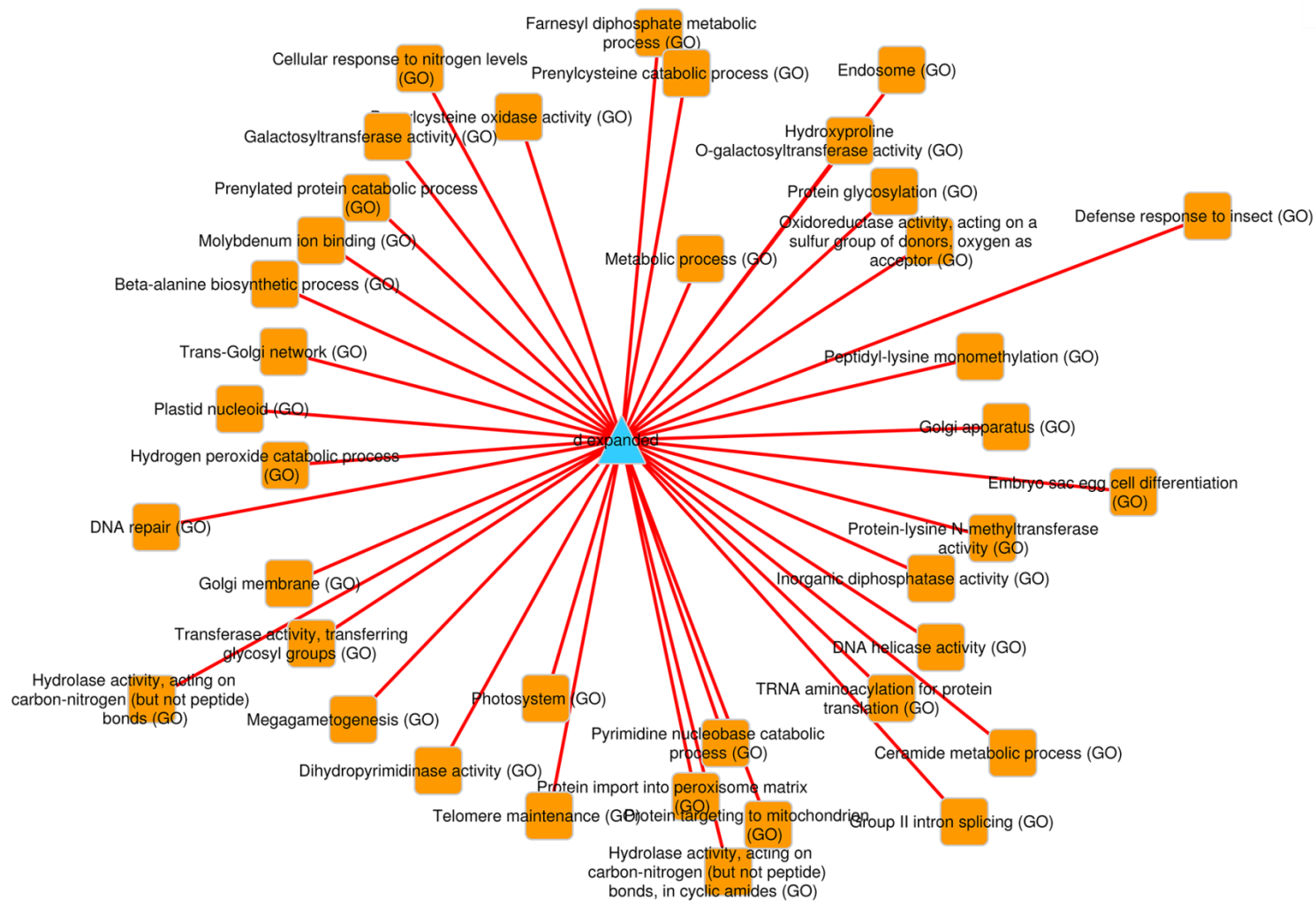
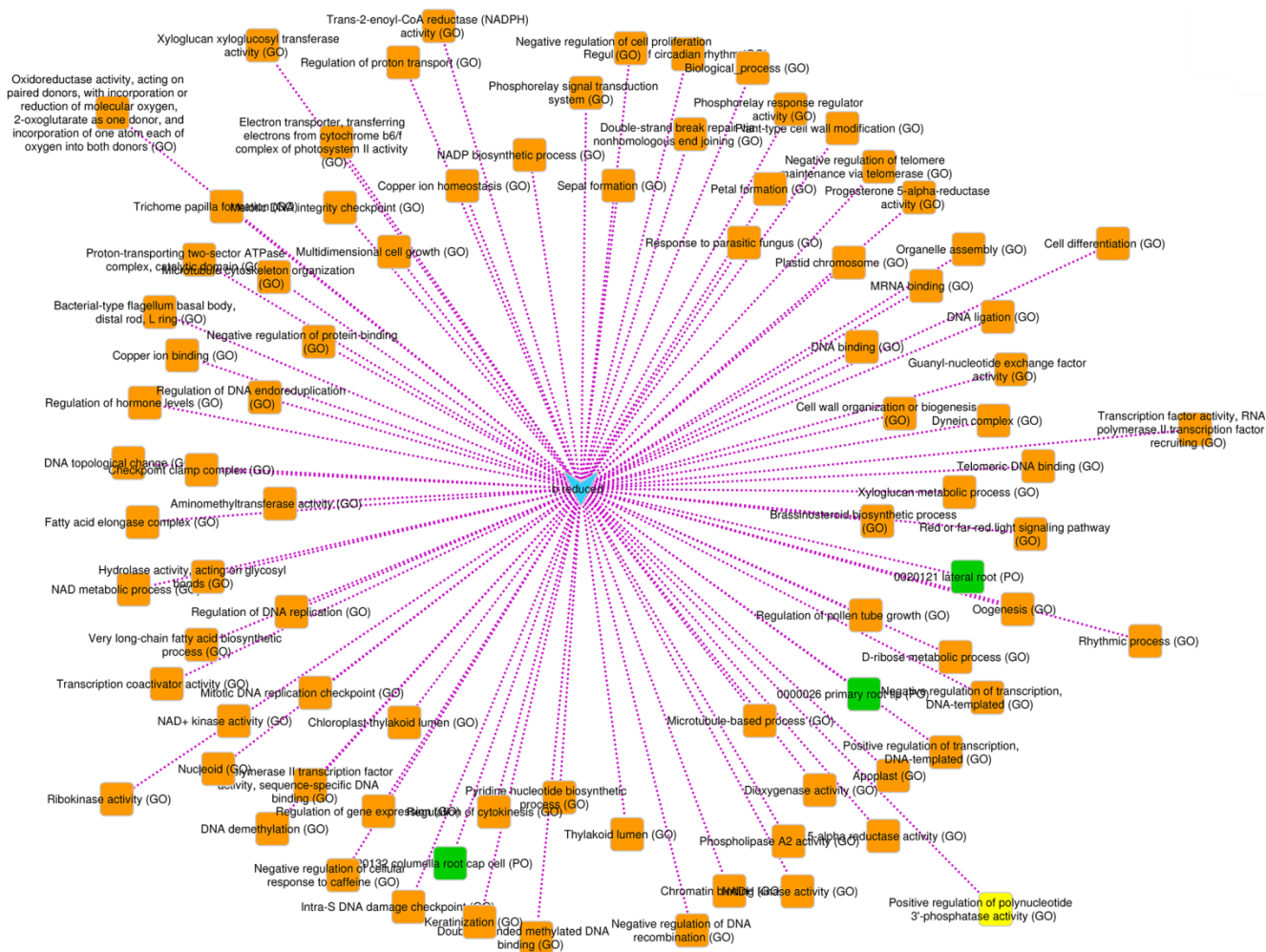


Fig. S50

Enriched ontology terms for the gene sets encoded by gene families expanded in the D genome.



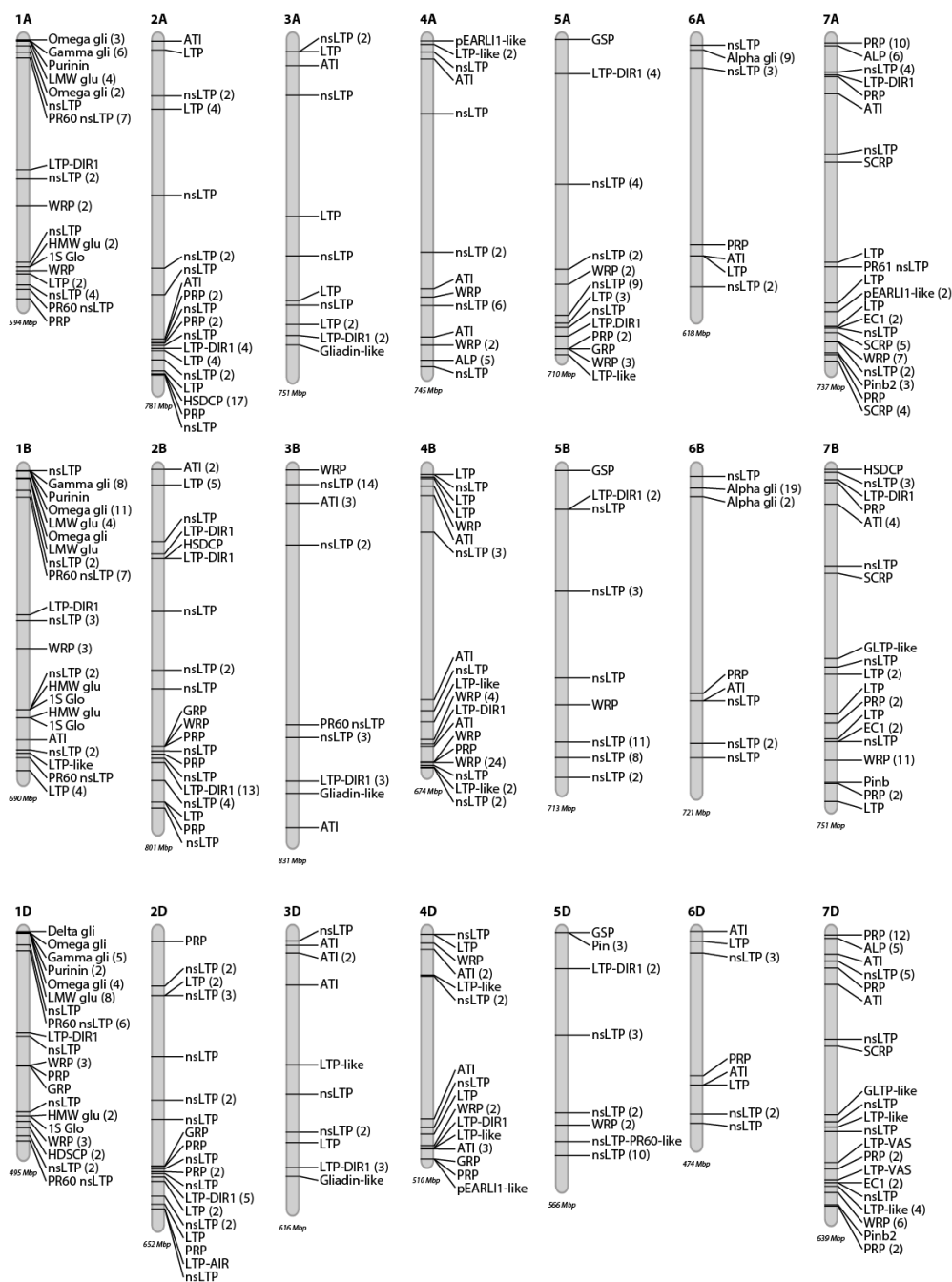


Fig. S52

Chromosomal distribution of the super-family of genes coding for the prolamins in wheat. The gene families are generally monophyletic in phylogenomic studies and occur in clusters in the genome assembly. All gene models have been analysed with respect to content of epitopes associated with coeliac disease and allergies (47).

5

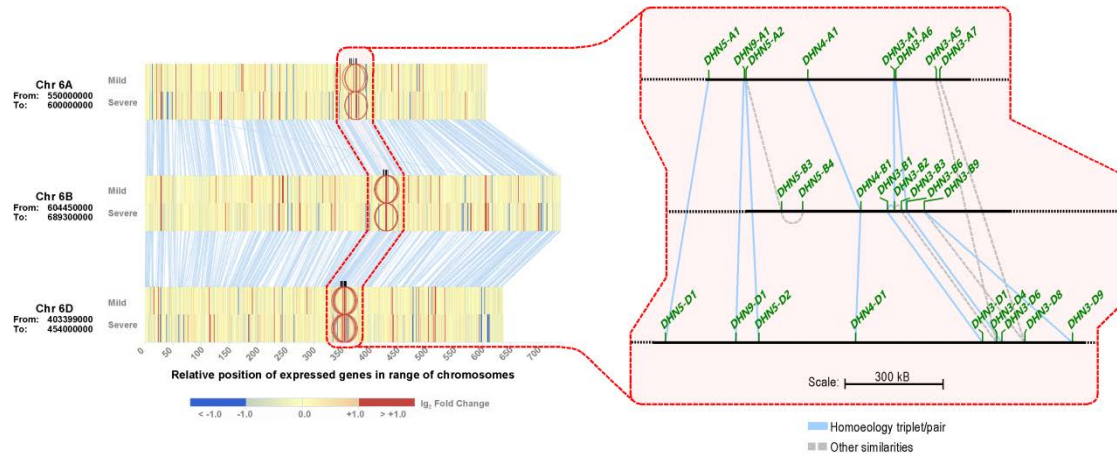


Fig. S53

Gene families and networks in drought tolerance. The left panel presents the differential expression of the dehydrins in a cluster on the long arms of chromosome 6. X axis provides gene positions in Mb. Each chromosome is represented by two bands of vertical lines: upper band is associated to Mild stress whereas the lower one is associated to Severe stress. Each vertical line corresponds to a sequential gene and its colour indicates differential expression (log Fold Change –lgFC). The cluster of dehydrins surrounded with a dotted red line are marked with short black lines on top of the Mild band in each chromosome. High levels of similarity are found between the dehydrin genes; a blue line connects homoeologous genes whereas a dotted grey line connects genes with an unassigned phylogenomic relationship. 6B dehydrin genes have the highest expression levels.

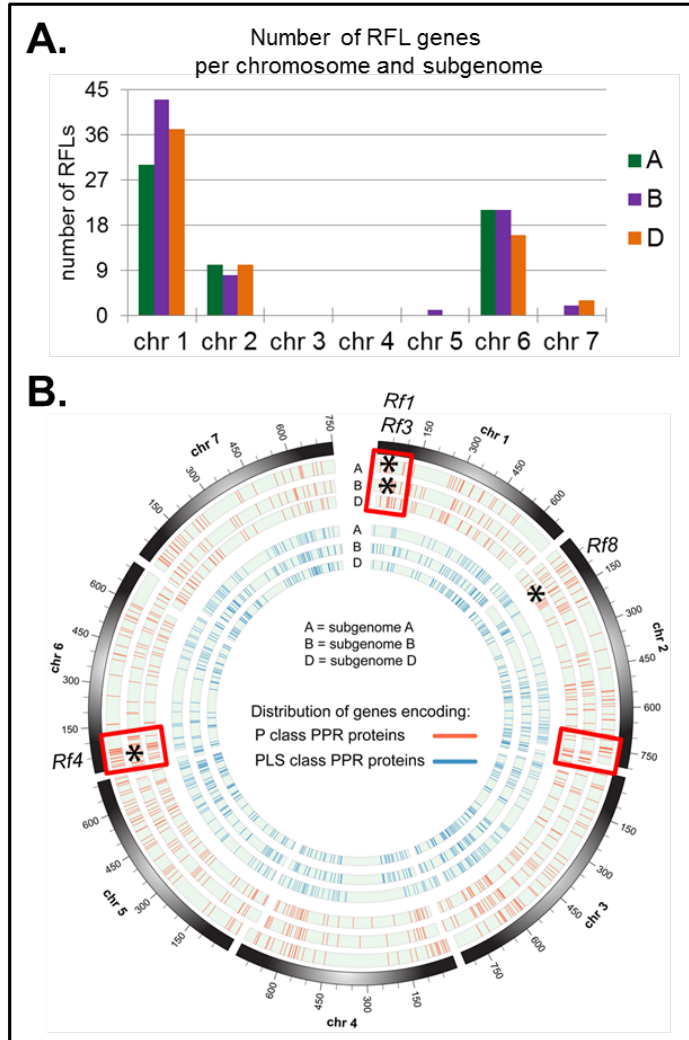


Fig. S54

Genome-wide distribution of RFL genes in IWGSC RefSeq v1.0. The RFL genes were identified by three methods: Orthofinder (131), OrthoMCL-DB (164) and phylogeny (45). In contrast to the ~500 PPR proteins encoded in flowering plant genomes (165), over 1600 PPR genes were found in wheat, a number nevertheless lower than expected from simply adding genes present in the progenitor diploid genomes (165). Evidence for gene inactivation and loss during polyploidization was found in truncated or frame-shifted gene fragments. Within the PPR gene family in IWGSC RefSeq v1.0 over 200 loci were identified as RFL genes that are organised in clusters on chromosomes 1, 2 and 6. The *Rf1* and *Rf3* restorer genes map to the largest cluster on chromosome 1 (166, 167), whereas *Rf4* maps to a cluster on chromosome 6 (166). *Rf8* mapped to chromosome 2D (168). **A.** Number of RFL genes per chromosome and sub-genome. **B.** Co-localisation of RFL clusters with the mapped restorer loci in wheat. The approximate position of the mapped *Rf* genes is indicated (black asterisk). Genomic regions carrying clusters of RFL genes are boxed in red.

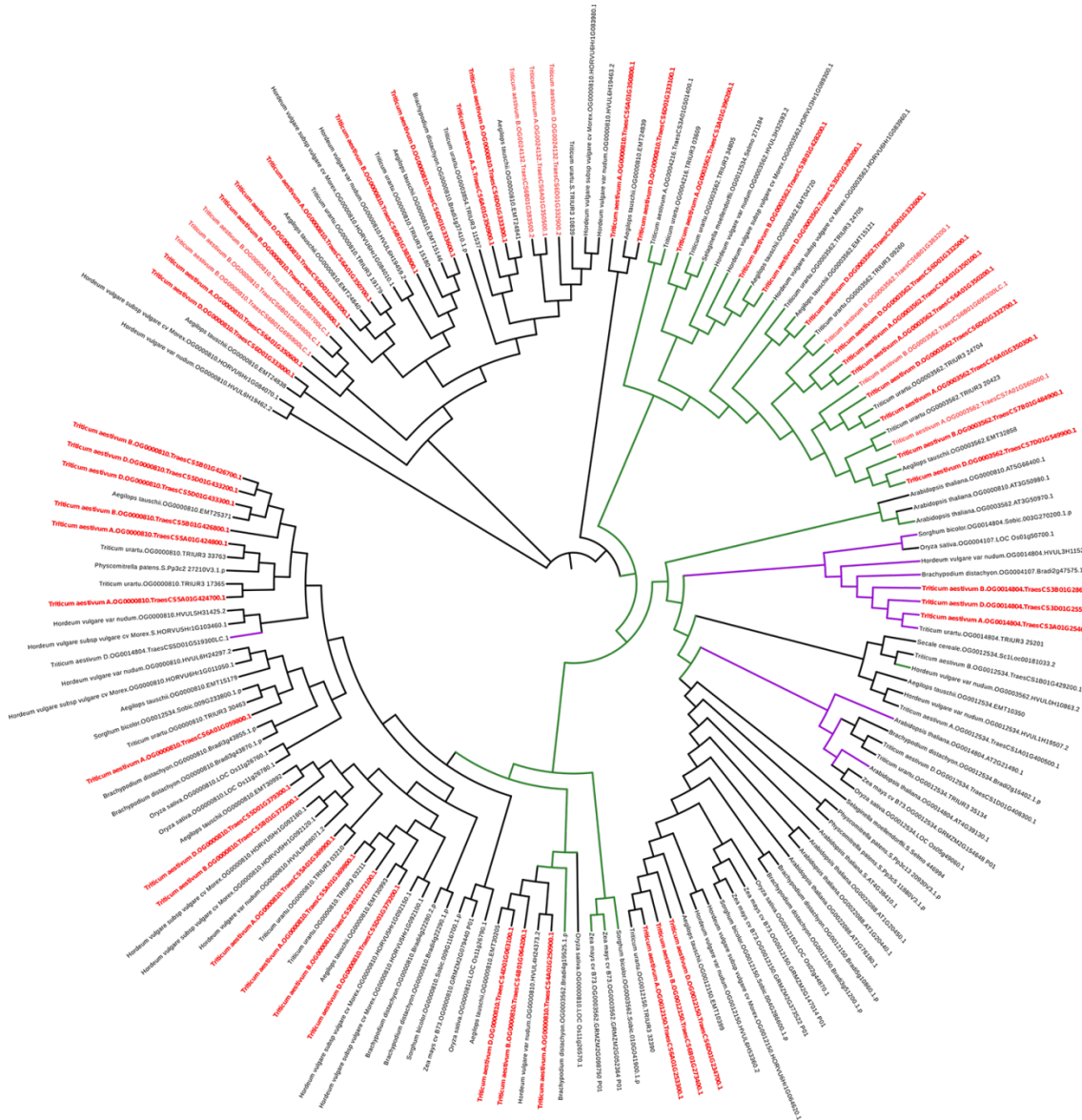


Fig. S55

Tree of expanded Dehydrin Orthologous Groups. The branches of the expanded orthologous groups are highlighted. OG0003562, green; OG0014804, purple. Wheat sequences are highlighted in red.

5

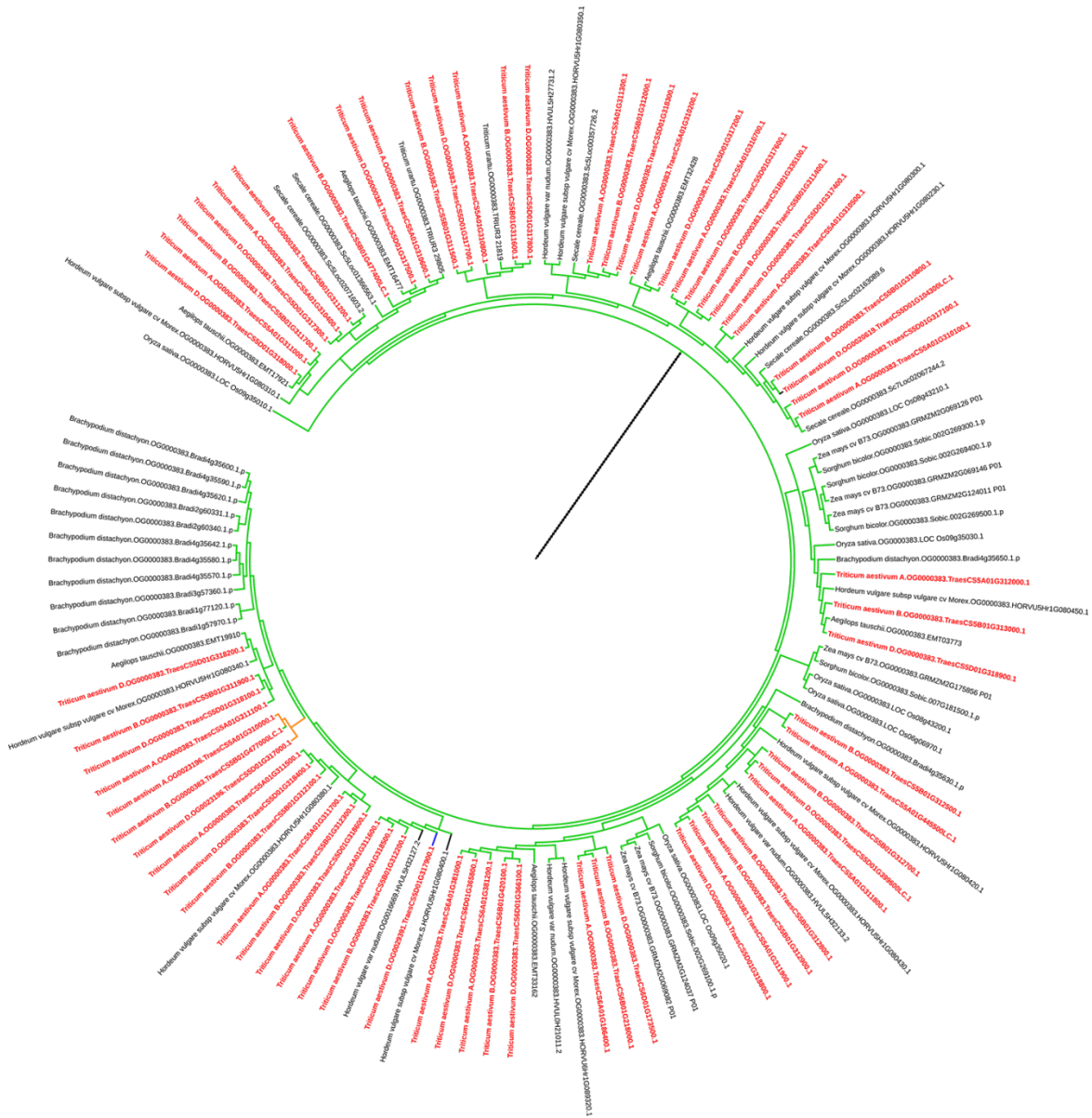


Fig. S56

The AP2 domain (Pfam: PF00847) was used to extract CBF orthologous groups from the bread wheat phylogenomics dataset (see Table S38) and to construct a tree. Sixty-one CBF genes were previously defined on group 5 chromosomes. Of those, 52 fall in OG0000383 (green clade). Seven sequences were not found in the AP2 tree and two fall into minor orthologous groups (OG0023196, orange branches; OG0029391, blue branch). OG0000383 is a monophyletic clade (bootstrap value = 1.00) consisting of 118 sequences of which 66 are from *T. aestivum* (highlighted in red). This orthologous group shows a highly significant expansion in *T. aestivum* ($p < 0.001$). Manual curation confirmed that the 62 of the CBF genes in OG0000383 included 17 CBF genes from 5A, 19 CBF genes from 5B, and 18 CBF genes from 5D. It also includes genes from 6A (3), 6B (2), 6D (3). Eight of the sequences were either partial genes or pseudogenes.

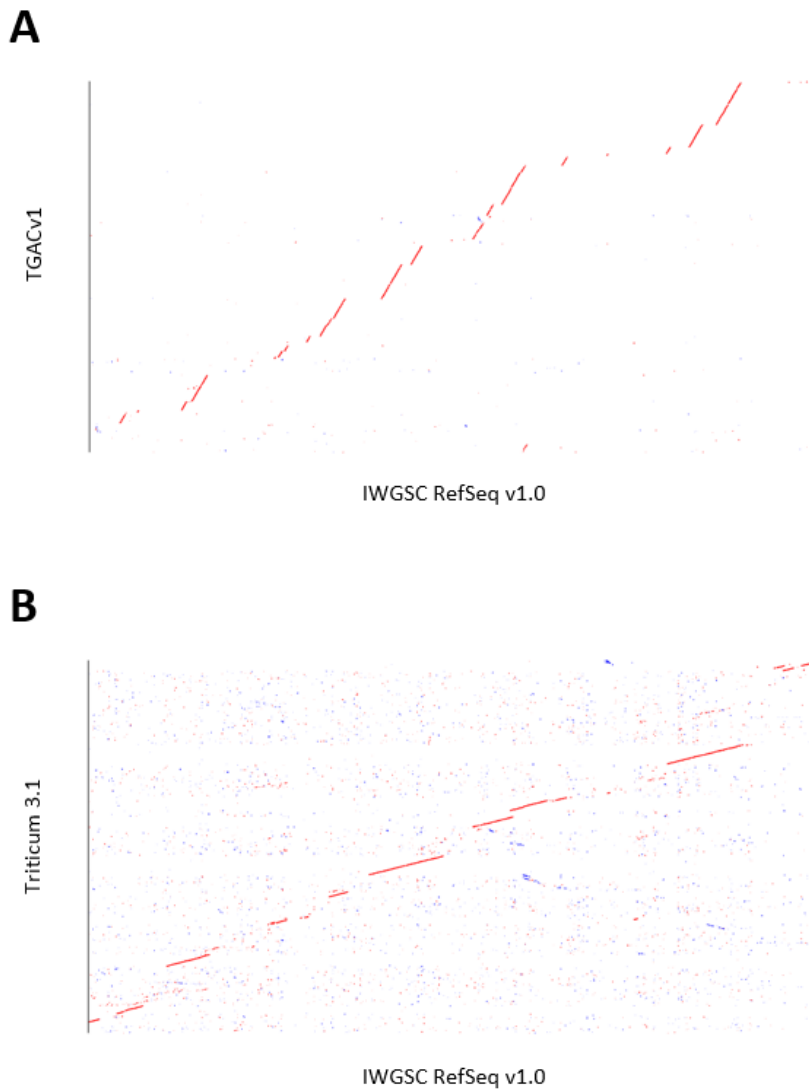


Fig. S57

NUCmer alignments of *SSt1* between IWGSC RefSeq v1.0 and available fragmented assemblies of CS. Scaffolds/contigs were retrieved from the fragmented assemblies using markers from the *SSt1* genetic map interval in Lillian/Vesper. A) Alignment of TGACv1 captures 39% of the interval, while B) alignment of Triticum 3.1 captures 61% of the interval. TGACv1 were prefiltered for placement in chromosome 3B, which was not possible for Triticum 3.1.

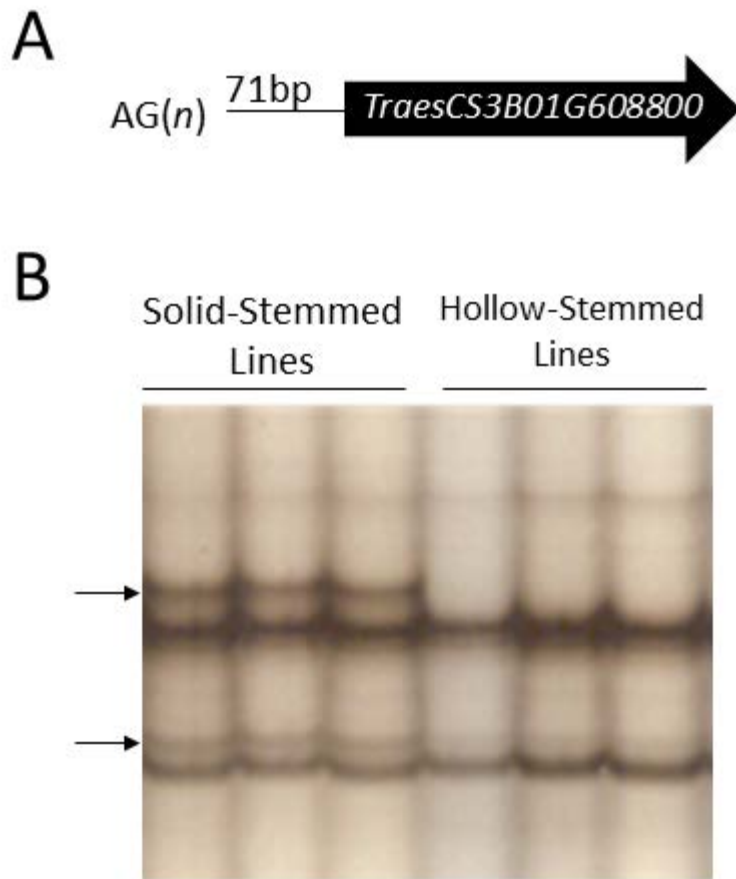
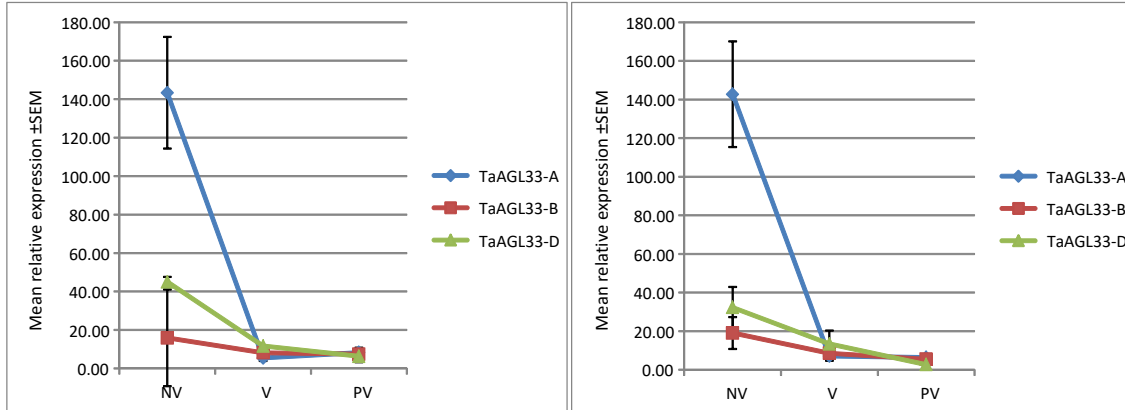


Fig. S58

- 5 Promoter variation in *TraesCS3B01G608800*. **A)** The gene *TraesCS3B01G608800* has a repeating AG(n) element 71 bp upstream of the transcription start site. **B)** Multiple variants of the AG(n) element are observed in solid-stemmed lines that are missing from hollow-stemmed lines (arrows), which were identified using single-strand conformation polymorphism (SSCP) gels.

10

A wild type**B genome edited****Fig. S59**

Gene expression analysis of gene editing events. Expression of TaAGL33A, TaAGL33B, and TaAGL33D prior (non-vernalized, NV), during (8 weeks vernalized, V) and post vernalization (2 weeks after transfer to greenhouse conditions, PV). Panel A summarizes wild-type segregants of 8 genome-editing events (with 3 biological repeats and 5 technical repeats per timepoint), panel B summarizes edited plants of 8 genome-editing events (with 3 biological repeats and 5 technical repeats per timepoint). Expression is measured with primers that locate downstream of the premature termination codons in the genome-edited plants. Vernalization experiments, RNA extraction, and qPCR were conducted as described before (50). Expression is expressed as mean $\pm$ SEM.

Table S1.

Sequencing library data for bread wheat whole genome assembly

#	Library type	Insert size	Sequencing Instrument	Read length	Minimal Coverage
1	PCR-free PE library	450bp	HiSeq 2500	PE265bp	X76
2	PCR-free PE library	800bp	HiSeq 2500	PE160bp	X35
3	MP (Nextera MP Gel Plus)	2-4kbp	HiSeq 2500	PE160bp	X36
4	MP (Nextera MP Gel Plus)	5-7kbp	HiSeq 2500	PE160bp	X34
5	MP (Nextera MP Gel Plus)	8-10kbp	HiSeq 2500	PE160bp	X36

Table S2.

Results of IWGSC whole genome assemblies (WGA) with NRGene software package DenovoMAGIC2TM before and after correction for scaffold mis-joins. A total of 227 putative chimeric scaffolds (166 detected by POPSEQ and 61 by the Hi-C data) were detected and broken in the second assembly version (IWGSC WGA v0.2).

5

	Initial WGA		IWGSC WGA v0.2	
	Contigs	Scaffolds	Contigs	Scaffolds
Total number of sequences	38,971,011	331,904	685,085	138,607
Assembly size (bp)	13,813,647,020	14,347,451,665	14,263,899,337	14,532,155,117
Gaps size (bp)	0	823,737,867	0	261,883,061
Gaps %	0	5.74%	0	1.80%
N₅₀-length (bp)	21,061	7,536,495	51,840	7,005,151
N₉₀-length (bp)	3,035	1,393,142	11,660	1,244,256
Maximal length (bp)	225,580	45,627,251	580,542	45,793,852

.

Table S3.

Summary of the IWGSC physical map data and associated publications.

Chromosome / chromosome arm	Project leader	Institution	BAC library size (clone number)	method	Assembly algorithm	Assembled contig length (Mb)	Contig N50 (kb)	Contig L50	MTP BACs	Reference
1AS	T. Wicker, B. Keller	U. Zurich, Switzerland	31104	HICF-SNaPShot	LTC	226	486	147	4155	(169)
1AL	H. Budak	Sabancı U., Turkey	92544	HICF-SNaPShot	LTC	461	1166	133	7470	(170)
1BS	T.Fahima, A. Korol	U. Haifa, Israel	55296	HICF-SNaPShot	LTC	261	2430	35	6447	(171)
1BL	E. Paux	INRA-GDEC, France	92160	HICF-SNaPShot	LTC	502	961	162	6023	(172)
1D,4D,6D	B. Gill	KSU, USA	286464	HICF-SNaPShot	LTC	1,690	1910	274	25650	
2AS	K.Singh	ICAR, Nat. Research Centre on Plant Biotechnology, New Delhi, India	55648	HICF-SNaPShot	LTC	496	2098	72	4442	
2AL			76800	HICF-SNaPShot	LTC	739	808	292	7708	
2BS	J. Jacobs	Bayer CropScience	67968	WGP TM	LTC	385	3800	33	4805	
2BL			70656	WGP TM	LTC	463	3900	34	6051	
2DS	J. Jacobs	Bayer CropScience	43008	WGP TM	LTC	246	6600	14	3025	
2DL			58368	WGP TM	LTC	359	4800	24	4840	
3AS	B. Gill	KSU, USA	110692	HICF-SNaPShot	FPC	432	674	144	5187	
3AL			78952	HICF-SNaPShot	FPC	515	986	136	5342	
3B	E. Paux	INRA Clermont Ferrand	150144	HICF-SNaPShot	FPC	811	857	390	9216	(173)

3DS	J. Bartoš, J. Doležel	IEB, Olomouc, Czech Republic	36864	HICF-SNaPSHOT	FPC	262	487	160	3823	(174)
3DL	J. Wright	Earlham Institute	64152	HICF-SNaPSHOT	FPC	637	897	138	6144	
4AS	M.Valarik, J. Doležel	IEB, Olomouc, Czech Republic	49152	HICF-SNaPSHOT	LTC	312	1359	60	4422	
4AL			92160	HICF-SNaPSHOT	LTC	479	480	273	8526	(175)
4BS	J. Jacobs	Bayer CropScience	58368	WGP™	LTC	348	4100	27	4870	
4BL			63744	WGP™	LTC	405	4300	30	5299	
5AS	D.Barabaschi, L. Cattavelli	CREA-GRC, Italy	46080	HICF-SNaPSHOT	LTC	330	820	128	4360	(176)
5AL			90240	HICF-SNaPSHOT	LTC	676	563	407	5528	
5BS	E. Salina	Inst.Cytology & Genetics, Novosibirsk, Russia	43776	HICF-SNaPSHOT	LTC	354	3078	34	3090	(177)
5BL	J. Jacobs	Bayer CropScience	76800	WGP™		567	4000	43	6166	
5DS	H. Budak	Sabancı U., Turkey	36864	HICF-SNaPSHOT	LTC	177	2173	27	2527	(178)
5DL	J. Jacobs	Bayer CropScience	72960	WGP™	LTC	352	4100	26	4753	
6AS	T.SchnurbuschN. Stein	IPK-Gatersleben, Germany	49152	WGP™	LTC	522	1090	1106	5139	(70)
6AL			55296	WGP™	LTC	543	945	921	5621	
6BS	H. Handa	NARO, Japan	57600	WGP™	FPC	492	1503	87	4983	(71)
6BL			76032	WGP™	FPC	495	2422	65	5981	
7AS	R. Appels	Murdoch U. Australia	58368	HICF-SNaPSHOT	FPC/LTC	353	1380	81	5280	(59)
7AL			61056	HICF-SNaPSHOT	FPC/LTC	402	1700	64	5832	
7BS	O-A. Olsen	NMBU, Ås, Norway	49152	HICF-SNaPSHOT	LTC	298	6367	14	3039	(179)

7BL			72960	HICF-SNaPSHOT	LTC	451	6820	22	5229	
7DS	H. Šimková, J. Doležal	IEB, Olomouc, Czech Republic	49152	HICF-SNaPSHOT	FPC	370	528	205	4608	(180)
7DL	S. Weining	Northwest A&F U., China	50304	HICF-SNaPSHOT	FPC	484	361	353	4457	

Table S4.

Summary of the Whole Genome Profiling Tags generated from chromosome-specific BAC libraries used for physical map construction.

Chromosome / chromosome arm	BAC library	Number of BACs with tags	Number of unique tags	Total number of tags (50bp)	Av. tags / BAC	Number of tags used for pseudomolecule assembly	Publication
2BS	TaaCsp2BShA	61,707	394,587	2,131,095	34.5	55947	
2BL	TaaCsp2BLhA	63,503	621,776	2,231,411	35.1	43084	
2DS	TaaCsp2DShA	37,634	136,792	1,129,020	30.0	19797	
2DL	TaaCsp2DLhA	50,586	212,681	1,410,847	27.9	28357	
4BS	TaaCsp4BShA	52,435	411,761	1,671,137	31.9	43957	
4BL	TaaCsp4BLhA	57,688	528,688	2,005,054	34.8	53669	
5BL	TaaCsp5BLhA	64,584	222,147	2,579,840	39.8	26666	
5DL	TaaCsp5DLhA	64,492	507,176	1,970,200	30.5	32153	
6AS	TaaCs6AShA	19,289	109,611	572,883	29.7	16634	(70)
6AL	TaaCsp6ALhA	18,660	108,811	533,676	28.6	16277	
6BS	TaaCsp6BShA	35,515	122,164	731609	20.6	17971	(71)
6BL	TaaCsp6BLhA	45,895	113,522	1,000,511	21.8	17933	

Table S5.

Summary of WGP™ tags derived from chromosome-specific BAC MTPs

Chromosome / chromosome arm	MTP BAC library	50 nucleotide tags				75 nucleotide tags				100 nucleotide tags				Number of tags used for pseudomolecule assembly
		Number of BACs with tags	Number of unique tags	Total number of tags	Av. tags / BAC	Number of BACs with tags	Number of unique tags	Total number of tags	Av. tags / BAC	Number of BACs with tags	Number of unique tags	Total number of tags	Av. tags / BAC	
1AS	TaaCsp1AShMTPv2	3959	75167	140619	35.5	3956	92424	166686	42.1	3950	102793	179461	45.4	24512
1AL	TaaCsp1ALhMTP	7336	148572	261979	35.7	7335	181130	304937	41.6	7336	197618	322605	44.0	45795
1BS	TaaCsp1BSHmTP	6144	128429	257233	41.9	6150	359236	302783	49.2	6151	181504	328151	53.3	31302
1BL	TaaCsp1BLhMTPv1	8305	198096	366881	44.2	8312	241343	421272	50.7	8314	268405	448763	54.0	43746
1D,4D,6D	TaaCsp146eAMTP	6252	188623	256932	41.1	6270	237117	307074	49.0	6270	261970	328183	52.3	50220
1D,4D,6D	TaaCsp146eBMTP	13138	288739	448047	34.1	1315	368206	533811	40.6	13158	417627	574062	43.7	62094
1D,4D,6D	TaaCsp146eCMTP	5099	148154	186276	36.5	5101	184322	224430	43.6	5098	204719	239829	47	41698
2AS	TaaCsp2AShMTP	4533	113824	200290	44.2	4537	140903	236978	52.2	4539	155757	254287	56.0	39996
2AL	TaaCsp2ALhMTP	7735	146803	307254	39.7	7742	187907	369520	47.7	7741	216251	403114	52.1	43833
3AS	TaaCsp3AShMTP	5142	110043	198397	38.6	5143	132253	229330	44.6	5144	143548	242196	47.1	35360
3AL	TaaCsp3ALhMTP	5286	128787	194163	36.7	5286	154840	224254	42.4	5277	167608	237126	44.9	37824
3DS	TaaCsp13DSHmTP	3780	92020	153511	40.6	3781	114850	184610	48.8	3782	130252	202109	53.4	27236
3DL	TaaCsp3DLhMTPv2	6081	117939	216120	35.5	6111	150115	261857	42.9	6130	172814	286789	46.8	37772
4AS	TaaCsp4AShMTP	4346	105516	192637	44.3	4346	122266	216065	49.7	4346	128183	223083	51.3	38543
4AL	TaaCsp4LShMTP	8240	176224	333128	40.4	8244	204680	368365	44.7	8244	215597	377772	45.8	43071
5AS	TaaCsp5AShMTPv2	5222	90394	191508	36.7	5228	104266	213930	40.9	5237	109488	220815	42.2	30741
5AL	TaaCsp5ALhMTPv2	8438	140730	300857	35.7	8438	167841	341026	40.4	8439	180282	356059	42.2	41496
5BS	TaaCsp5BSHmTP	3117	77581	122607	39.3	3120	88001	135864	43.5	3118	92062	140336	45.0	20803
5DS	TaaCsp5DSHmTP	2486	50401	95729	38.5	2489	60119	111671	44.9	2491	64947	118401	47.5	26666
7AS	TaaCsp7AShMTP	5712	151105	289804	50.7	5713	187408	341961	59.9	5716	208805	367791	64.3	40153
7AL	TaaCsp7ALhMTP	5706	138089	279147	48.9	5707	176141	337729	59.2	5707	203589	371525	65.1	38086
7BS	TaaCsp7BSHmTP	3261	86318	150529	46.2	3263	100159	170956	52.4	3263	106773	179148	54.9	27729
7BL	TaaCsp7BLhMTP	5325	140869	253035	47.5	5322	165648	286549	53.8	5323	178900	301495	56.6	33144
7DS	TaaCsp7DSHmTP	4545	107292	180326	39.7	4547	124412	202896	44.6	4547	130603	209492	46.1	33470
7DL	TaaCsp7DLhMTP	4358	117167	175779	40.3	4357	135725	197390	45.3	4358	142605	203679	46.7	32130

Table S6.

Chromosome-specific assembled BAC MTP sequences

Chromosome / chromosome arm	Project leader	Institution	Number of sequences	N50 (kb)	Total Length (Mb)	BAC format	Sequencing technology	Assembly algorithm	Publication
1A	C. Pozniak	U. Saskatchewan, Canada	45301	52	1218	single BACs	Illumina PE + MP	RAY, sspace	
1B	E. Paux	INRA-GDEC, France	13227	351	920	BAC pools	Illumina PE + 5kb MP	Newbler + sspace	
3B	C. Feuillet	INRA-GDEC, France	9158	167	1360	BAC pools	Roche-454 MP + finishing	Newbler	(18)
3DS	J. Bartoš,	IEB, Olomouc, Czech Republic	50760	49	498	single BACs	Illumina PE	RAY, CAP3	
3DL	M. Clark	Earlham Institute, UK	67447	28	728	single BACs	Illumina PE	ABYSS / SOAPdenovo	(181)
4AL	M. Valarik, J. Doležel	IEB, Olomouc, Czech Republic	4470	43	46	single BACs	Illumina PE	CLC Bio genomic software	(175)
5BS	E. Salina	Inst.Cytology & Genetics, Novosibirsk, Russia.	17700	4	26	single BACs	Ion Torrent	MIRA	(182, 183)
	N. Ravin	Research Center of Biotechnology RAS, Russia	12788	22	103	BAC pools	Roche-454	Newbler	
6B	H. Handa	NARO, Japan	24611	71	809	single BACs	Roche-454	GS assembler	
7A	R. Appels	Murdoch U., Australia	74190	28	878	BAC pools	Illumina PE	ABYSS	(59)
7B	O-A. Olsen	NMBU, Ås, Norway	928529	26	1869	single BACs	Illumina PE	SOAPdenovo	
7DS	H. Šimková	IEB, Olomouc, Czech Republic	19455	72	562	BAC pools	Illumina PE	SASSY	
7DL	S. Weining	Northwest A&F U., China	7237	137	764	single BACs	Illumina PE + PacBio reads	SOAPdenovo	

Table S7.

Summary of the Group 7 chromosome Bionano optical map assemblies

Chromosome arm	7AS	7AL	7BS	7BL	7DS*	7DL
Arm size (Mb)	407	407	360	540	381	346
No, of arms sorted (million)	2.8	2.8	2.8	2.8	1.6	2.8
Purity of sorted arm (%)	80	86	83	87	84	83
DNA amount (µg)	2.3	2.3	2.1	3.1	1.2	2.0
Raw data > 150 kb (Gb)	78	97	248	131	69	118
Filtered molecules N50 (kb)	206	232	225	231	354	210
Arm coverage	192x	238x	689x	243x	181x	341x
No, of contigs	783	330	254	626	371	364
Assembly length (Mb)	447	413	355	512	350	316
Contig N50 (Mb)	1.6	2.1	2.0	1.5	1.3	1.3
Average contig length (Mb)	0.57	1.25	1.4	0.82	0.94	0.87

* (69)

Table S8.

Summary of marker distribution in the sub-genomes of Wheat Whole Genome Radiation Hybrid Maps

Genome	Number of markers per sub-genome	Total map length in centiRay(cR) (sub-genome level)	Average map resolution (Mb/cR)
D-chromosomes	4,584	5,245	0.8
B-chromosomes	2,206	5,550	1.02
A-chromosomes	1,731	5,694	0.93
Whole genome	8,521	16,489	0.97

Table S9.

Molecular markers assigned to IWGSC RefSeq v1.0.

Marker type ¹	Total markers mapped	Number of unique marker positions
SSR	595	504
SNP	875279	205807
DArt	4215	3025
EST	14508	6689
ISBP	4607897	4512979

¹Details with respect to included markers and accessing GFF documents for the sequences can be found at

5 <https://wheat-urgi.versailles.inra.fr/Seq-Repository>

Table S10.

Summary of miRNA content of wheat chromosomes

Chromosome	Number of unique pre-miRNA/mature miRNA pairs	Number of miRNA families	Highest abundance	Chromosome-specific miRNAs
1A	2342	46	miR1130	-
1B	4064	47	miR1130	-
1D	2532	47	miR1130	-
2A	3101	51	miR1130	-
2B	5199	52	miR1130	miR9774
2D	3364	51	miR1130	miR6224
3A	2899	47	miR1117	-
3B	5262	52	miR1130	-
3D	3042	50	miR1130	miR9669
4A	3210	45	miR1117	-
4B	3694	46	miR1130	miR5085, miR7742
4D	2070	47	miR1130	miR5169
5A	2862	55	miR1117	miR5183, miR528
5B	4831	56	miR1130	miR8155
5D	2792	56	miR1130	miR9672
6A	2331	45	miR1117	-
6B	4383	51	miR1130	miR9659, miR9663
6D	2221	51	miR1130	miR5566, miR6219, miR9662
7A	3341	55	miR1117	miR8740, miR9671
7B	4858	49	miR1130	-
7D	3352	52	miR1130	-

Table S11.

Delimiting centromeres based on CENH3 ChIP-seq.

Chromosome	CENH3-enriched intervals				Total length (Mb)
	Pseudomolecule positions (Mb)				
1A	210.2-215.8				5.7
1B	237.7-243.5				5.9
1D	166.2-173.8				7.7
2A	326.3-327.0	339.4-342.0	359.3-359.5		3.8
2B	344.4-351.3				7.0
2D	264.4-272.5				8.2
3A	316.9-319.9				3.1
3B	345.8-347.0	348.5-349.0			1.9
3D	237.1-243.2				5.5
4A	264.1-267.9	315.1-315.7			4.6
4B	303.9-304.4	317.8-319.6			2.5
4D	182.3-188.2				6.0
5A	108.9-109.1	248.7-249.0	252.5-255.1		3.4
5B	198.9-202.5				3.7
5D	185.6-188.7				3.2
6A	283.3-288.7	290.7-292.5			7.4
6B	323.0-327.5				4.6
6D	211.9-217.4				5.6
7A	360.2-363.8				3.7
7B	288.2-288.3	294.4-294.6	296.4-296.5	308-310.1	2.9
7D	336.3-341.7				5.5
Average					4.9
chrUn					39.8
Average incl, chrUn					6.7
chrUn - unassigned scaffolds					

Table S12.

Summary of changes made to RefSeq Annotation v1.0 by integrating manually curated genes.

Integration status	IWGSC Annotation v1.0 HC genes	IWGSC Annotation v1.1 HC genes
No change	104,201	104,206
Add	-	528
Replace	3,020	3,020
Merge	224	94
Split	14	28
Multilocus	16	15
Total	107,475	107,891

Table S13.

Illumina RNA-Seq reads from the following datasets were used to support annotation integration: six different tissues described in (8) (PRJEB15048), grain-development samples from (40) (ERP004505), reads from (6) (ERP004714) collapsed into grain, leaf, root, spike and stem samples, seedling samples under normal condition and subjected to drought stress and heat stress (PRJNA257938) and reads from FHB challenged near isogenic wheat lines (PRJEB4202).

	PRJEB15048	ERP004505	ERP004714	PRJNA257938	PRJEB4202
Number of samples	6	7	5	7	20
Number of reads	731,931,657	873,550,049	1,412,029,174	921,578,806	1,827,362,091
Average no of reads per sample	121,988,610	124,792,864	282,405,835	131,654,115	91,368,105
Average overall alignment rate	95.93%	89.11%	93.84%	81.29%	87.12%

Table S14.

Pacific BioSciences transcript sequence (PacBio) reads (PRJEB15048) and alignments to IWGSC RefSeq v1.0 used to support annotation integration.

Stage	Leaf	Root	Seed	Seedling	Spike	Stem	Total
Total	199,119	315,137	219,965	209,923	277,704	287,474	1,509,322
Aligned	191,967	304,087	212,603	203,578	267,912	275,817	1,455,964
Aligned (%)	96.41%	96.49%	96.65%	96.98%	96.47%	95.95%	96.46%

Table S15.

Protein alignments used to support annotation integration, Protein sequences from 6 species were soft masked for low complexity (segmasker from NCBI BLAST+ 2,3,0) and aligned to the soft masked genome with exonerate v2.2.0 (95), Proteins were filtered at 50% identity and 80% coverage.

5

	<i>A. thaliana</i>	<i>B. distachyon</i>	<i>O. sativa</i>	<i>S. bicolor</i>	<i>S. italica</i>	<i>Z. mays</i>
Total Proteins	48,359	52,972	49,061	47,205	43,001	88,760
Proteins Aligned	15,683	36,710	27,359	27,335	27,146	44,252
Proteins Aligned (%)	32.43%	69.30%	55.77%	57.91%	63.13%	49.86%
Protein Alignments	82,321	181,573	141,023	137,605	138,783	224,775

Table S16.

Transcript assembly statistics.

Method	Loci	Transcripts	Average no of exons	Average cDNA size	Number of monoexonic transcripts
CLASS	342,738	5,821,796	5.61	1,335,32	628,375
Cufflinks	337,372	6,089,845	4.59	1,659,69	1,677,144
StringTie	499,560	5,999,887	4.73	1,683,71	1,682,569
Mikado PacBio	80,119	114,233	6.2	2,000,61	18,871
Mikado Illumina and PacBio	278,015	406,886	4	1,247,67	169,516

Table S17.

Metrics used for scoring transcripts.

Metric	Weighting	Metric type
Mean F1 across the base, exon and junction level vs. the best match among aligned PacBio transcripts	X 10	Evidence based
Mean F1 across the base, exon and junction level vs. the best match among Mikado RNA-Seq models	X 10	Evidence based
Mean F1 across the base, exon and junction level vs. the best match among aligned proteins (Table S14)	X 10	Evidence based
BLAST Query coverage of the best BLAST alignment across the protein databases (species described in Table S14 + uniprot Magnoliophyta)	X 10	Evidence based
BLAST Target coverage of the best BLAST alignment across the protein databases (species described in Table S14 + uniprot Magnoliophyta)	X 10	Evidence based
Proportion of Portcullis-verified introns assigned to the locus that can be found in the transcript	X 10	Evidence based
Proportion of introns that are canonical within the transcript	X 2	Intrinsic
Longest CDS	X 1	Intrinsic
Penalty on the number of introns greater than 50kbps	X -10	Intrinsic
Penalty for any transcript with introns over 10kbps	X -2	Intrinsic
Penalty on the presence of exons over 20kbps	X -10	Intrinsic

Table S18.

Characteristics of IWGSC RefSeq v1.1 wheat genes.

RefSeq V1.1	All	A	B	D	Unknown
Genes	269,428	86,930	94,002	79,047	9,449
Transcripts	298,775	96,383	103,931	88,631	9,830
Transcripts per gene	1.11	1.11	1.11	1.12	1.04
Transcript mean size cDNA (bp)	1,131,59	1,124,54	1,120,58	1,183,67	847,42
Exons per transcript	3.57	3.59	3.50	3.76	2.34
Exon mean size (bp)	317,32	313,52	320,00	315,22	362,32
Transcript mean size CDS (bp)	934,42	929,70	927,10	969,20	744,44
Single exon transcripts	131,105 (43.9%)	42,007 (43.58%)	46,387 (44.63%)	37,355 (42.15%)	5,356 (54.49%)
Genes with alternative splicing	19,762 (7.3%)	6,389(7.3%)	6,690(7.1%)	6,412(8.1%)	271(2.9%)
High Confidence (HC)	All	A	B	D	Unknown
Genes	107,891	35,345	35,643	34,212	2,691
Transcripts	133,745	43,697	44,221	42,828	2,999
Transcripts per gene	1.24	1.24	1.24	1.25	1.11
Transcript mean size cDNA (bp)	1,699,27	1,672,81	1,716,87	1,733,26	1,339,99
Exons per transcript	5.60	5.62	5.62	5.71	3.54
Exon mean size (bp)	303,34	297,65	305,70	303,38	378,69
Transcript mean size CDS (bp)	1,333,32	1,310,85	1,351,15	1,354,32	1,097,94
Single exon transcripts	26,973 (20.2%)	8,605 (19.69%)	8,872 (20.06%)	8,457 (19.75%)	1,039 (34.64%)
Genes with alternative splicing	16,961 (15.7%)	5,507(15.6%)	5,610(15.7%)	5,638(16.5%)	206(7.7%)
Low Confidence (LC)	All	A	B	D	Unknown
Genes	161,537	51,585	58,359	44,835	6,758
Transcripts	165,030	52,686	59,710	45,803	6,831
Transcripts per gene	1.02	1.02	1.02	1.02	1.01
Transcript mean size cDNA (bp)	671,52	669,85	678,96	669,78	631,16
Exons per transcript	1.92	1.90	1.94	1.92	1.81
Exon mean size (bp)	350,44	352,47	350,72	348,10	348,29
Transcript mean size CDS (bp)	611,15	613,52	613,10	609,15	934,42
Single exon transcripts	104,132 (63.1%)	33,402 (63.40%)	37,515 (62.83%)	28,898 (63.09%)	4,317 (63.20%)
Genes with alternative splicing	2,801 (1.7%)	882(1.7%)	1080(1.8%)	774(1.7%)	65(0.9%)

Table S19.

Automated Assignment of Human Readable Descriptions (AHRD) annotation of wheat gene models identified in RefSeq Annotation v1.0. Classification of HC genes by their functional description via disTEG (distinction between TEs and Genes). Note: for RefSeq Annotation v1.1 genes with obvious transposon descriptions (3,318 HC genes) were re-classified as LC-TE.

5

	disTEG tag	Wheat CS IWGSC HC genes v1.0		Wheat CS IWGSC LC genes v1.0		
		Number	%	Number	%	
	G	99,118	89.5	81,552	51.4	canonical genes
	U	301	0.3	595	0.4	unknown
	TE?	8,083	7.3	15,099	9.5	potential transposons
	TE	3,288	3.0	61,526	38.8	obvious transposons
	sum	110,790	100	158,772	100	
Without TEs (G-TE)		107,502	97.0	97,246	61.2	
G-TE: with good scoring functional annotation		90,919	82.1			

Table S20.

Ontology term annotation statistics. 5,182,416 associations to 8,133 unique ontology terms describing biological processes, molecular functions, anatomical entities, developmental stages or genetically linked phenotypic traits of orthologous genes in well-characterized plant models like Arabidopsis, rice and maize were obtained for 117,595 wheat genes in 16,662 gene families.

A) Summary of ontology term associations for each of the three source ontologies (Plant Ontology (PO), (118); Gene Ontology (GO), (117); Plant Trait Ontology (TO), <https://bioportal.bioontology.org/ontologies/PTO>). B) Summary of ontology term associations by evidence class (IEA-Inferred Electronic Annotation; EXP – Inferred from Experiment; ISM – Inferred from Sequence Model).

A) Ontology	Number of associations	Number of families	Number of sequences	Number of terms
PO	3,622,724	12,631	79,530	446
GO	1,560,102	15,983	113,815	7,408
TO	8,173	227	1,060	279
B) Evidence_code	Count			
ISO	4,989,419			
IEA	177,154			
ISM	155,326			

Table S21.

Genome-wide pseudogene analysis based on pseudogenes identified by *de novo* homology analysis of HC genes and disrupted or truncated gene models in the LC gene set: Sub-genome distribution and basic statistics for pseudogene set identified by *de novo* homology analysis.

Sub-genome	Genome-wide approach	LC-only	Total	%
Chr Un	13,062		13,062	
A	94,686	5,068	99,754	33%
B	103,353	5,744	109,097	36%
D	77,738	4,167	81,905	27%
All	288,839	14,979	303,818	100%

5

Table S22.

Genome-wide analysis based on pseudogenes identified by *de novo* homology analysis of HC genes and disrupted or truncated gene models in the LC gene set: Number and pseudogene types for all pseudogenes, subdivided by high and low coverage classes and sub-genome distribution. High coverage pseudogenes are represented by at least 80% of the CDS of a parent gene locus, low coverage below 80% (see also 4.5.2).

	number	%	mean length	mean coverage	mean identity	with ptc	% with ptc
all pseudogenes	288,839	100.0	342.8	38.5	89.5	168,096	58.2
-- duplicated	73,241	25.4	529.7	49.9	91.3	51,988	71.0
-- processed	7,199	2.5	347.8	42.2	89.8	5,062	70.3
-- single exon gene	85,754	29.7	348.1	52.2	88.6	47,040	54.9
-- single exon isoform	55	0.0	678.0	39.1	90.8	40	72.7
-- fragmented	118,587	41.1	213.1	20.8	88.9	60,471	51.0
-- chimeric	4,003	1.4	636.4	50.9	91.3	3,495	87.3
HighCov	48,619	16.8	738.8	94.6	92.1	33,284	68.5
-- duplicated	20,464	42.1	939.5	94.8	93.1	15,569	76.1
-- processed	1,372	2.8	623.5	94.1	92.0	1,029	75.0
-- single exon gene	23,589	48.5	564.1	94.7	91.3	14,217	60.3
-- single exon isoform	8	0.0	1,674.1	94.2	96.2	8	100.0
-- fragmented	1,974	4.1	708.5	91.6	91.6	1,363	69.1
-- chimeric	1,212	2.5	922.0	93.1	92.9	1,098	90.6
LowCov	240,220	83.2	262.6	27.1	88.9	134,812	56.1
-- duplicated	52,777	22.0	370.8	32.4	90.5	36,419	69.0
-- processed	5,827	2.4	282.9	29.9	89.3	4,033	69.2
-- single exon gene	62,165	25.9	266.1	36.1	87.5	32,823	52.8
-- single exon isoform	47	0.0	508.4	29.8	89.8	32	68.1
-- fragmented	116,613	48.5	204.7	19.6	88.9	59,108	50.7
-- chimeric	2,791	1.2	512.4	32.6	90.5	2,397	85.9
subgenome A	94,686	32.8	330.9	38.0	89.5	55,349	58.5
subgenome B	103,353	35.8	345.7	38.2	89.4	60,540	58.6
subgenome D	77,738	26.9	347.3	39.5	89.5	45,540	58.6
unknown	13,062	4.5	379.1	37.3	90.4	6,667	51.0
HighCov subgenome A	15,241	31.4	732.6	94.5	92.1	10,653	69.9
HighCov subgenome B	17,308	35.6	758.4	94.5	91.9	12,019	69.4
HighCov subgenome D	14,260	29.3	721.1	94.7	92.3	9,575	67.2
HighCov unknown	1,810	3.7	742.5	94.5	93.4	1,037	57.3

Table S23.

Species and genome annotation versions utilized for the phylogenomics analysis.

Taxon	NCBI taxono my ID	Source	Original filename
<i>Triticum aestivum</i> A	4565	This study	Triticum_aestivum.IWGSC_V1.0_March_2017.all.proteins.subgenome_A.representative.fasta
<i>Triticum urartu</i>	4572	(53)	Triticum_urartu.ASM34745v1.pep.representative.fasta
<i>Triticum aestivum</i> B	4565	This study	Triticum_aestivum.IWGSC_V1.0_March_2017.all.proteins.subgenome_B.representative.fasta
<i>Aegilops tauschii</i>	37682	(184)	Aegilops_tauschii.ASM34733v1.pep.all.fa
<i>Triticum aestivum</i> D	4565	This study	Triticum_aestivum.IWGSC_V1.0_March_2017.all.proteins.subgenome_D.representative.fasta
<i>Secale cereale</i>	4550	(155)	RyeMIPSV3final_PROT_mar14.fa
<i>Hordeum vulgare</i> subsp. <i>vulgare</i> cv. <i>Morex</i>	112509	(146)	Hvulgare.IBSC_PGSC_r1.only_representative.proteins_HighConf.fa
<i>Hordeum vulgare</i> var. <i>nudum</i>	112509	(185)	Tibetan_Hulless_barley.pep.fa
<i>Brachypodium distachyon</i>	15368	Phytozome 12	Bdistachyon_314_v3.1.protein_primaryTranscriptOnly.fa
<i>Oryza sativa</i>	4530	Phytozome 12	Osativa_323_v7.0.protein_primaryTranscriptOnly.fa
<i>Sorghum bicolor</i>	4558	Phytozome 12	Sbicolor_313_v3.1.protein_primaryTranscriptOnly.fa
<i>Zea mays</i> cv. <i>B73</i>	4577	Phytozome 12	Zmays_284_Ensembl-18_2010-01-MaizeSequence.protein_primaryTranscriptOnly.fa
<i>Arabidopsis thaliana</i>	3702	Phytozome 12	Athaliana_447_Araport11.protein_primaryTranscriptOnly.fa
<i>Selaginella moellendorffii</i>	88036	Phytozome 12	Smoellendorffii_91_v1.0.protein_primaryTranscriptOnly.fa
<i>Physcomitrella patens</i>	3218	Phytozome 12	Ppatens_318_v3.3.protein_primaryTranscriptOnly.fa
<i>Chlamydomonas reinhardtii</i>	3055	Phytozome 12	Creinhardtii_281_v5.5.protein_primaryTranscriptOnly.fa

Table S24.

Groups of homeologous genes in the wheat genome. A homeologous gene group represents a group of shared genes among the A, B and/or D sub-genomes that were inferred as “sub-genome orthologs” by species tree reconciliation in the respective gene family. Unless otherwise stated numbers include both HC and LC genes filtered for TEs. * Sub-genome-specific genes are supported by at least one ortholog in the reference grass species.

homeologous group (A:B:D)	# in wheat genome	% of groups	# genes in A	# genes in B	# genes in D	# total genes	% genes A	% genes B	% genes D	%genes total
1:1:1	21,603	55.1%	21,603	21,603	21,603	64,809	35.8%	33.8%	38.1%	35.8%
1:1:n	644	1.6%	644	644	1,482	2,770	1.1%	1.0%	2.6%	1.5%
1:n:1	998	2.5%	998	2,396	998	4,392	1.7%	3.7%	1.8%	2.4%
n:1:1	761	1.9%	1,752	761	761	3,274	2.9%	1.2%	1.3%	1.8%
1:1:0	3,708	9.5%	3,708	3,708	0	7,416	6.1%	5.8%	0.0%	4.1%
1:0:1	4,057	10.3%	4,057	0	4,057	8,114	6.7%	0.0%	7.1%	4.5%
0:1:1	4,197	10.7%	0	4,197	4,197	8,394	0.0%	6.6%	7.4%	4.6%
other ratios	3,270	8.3%	4,999	5,371	4,114	14,484	8.3%	8.4%	7.2%	8.0%
1:1:1 in microsynteny	18,595	47.4%	18,595	18,595	18,595	55,785	30.8%	29.1%	32.8%	30.8%
total in microsynteny	30,339	77.3%	27,240	27,063	28,005	82,308	45.2%	42.3%	49.3%	45.5%
1:1:1 in macrosynteny	19,701	50.2%	19,701	19,701	19,701	59,103	32.7%	30.8%	34.7%	32.6%
total in macrosynteny	32,591	83.1%	29,064	30,615	30,553	90,232	48.2%	47.9%	53.8%	49.8%
HC-only	26,446	67.4%	24,753	24,922	25,047	74,722	41.0%	39.0%	44.1%	41.3%
total in homeologous groups	39,238	100.0%	37,761	38,680	37,212	113,653	62.6%	60.5%	65.5%	62.8%
conserved subgenome orphans*			12,412	12,987	10,844	36,243	20.6%	20.3%	19.1%	20.0%
non-conserved subgenome singletons			10,084	12,185	8,679	30,948	16.7%	19.1%	15.3%	17.1%
non-conserved subgenome duplicated orphans			71	83	38	192	0.1%	0.1%	0.1%	0.1%
total (filtered)			60,328	63,935	56,773	181,036	100.0%	100.0%	100.0%	100.0%

Table S25.

Error assessment for homeologous groups with a potential gene loss in one subgenome. The observed differences in potential gene losses still remain significant after worst-case adjustment.

homeologous group (A:B:D)	# groups on pseudomolecules	% groups on pseudomolecules	# groups with additional copy on chrUn	% groups with additional copy on chrUn	# worst-case adjusted groups total	% worst-case adjusted groups total	# expected groups on pseudomolecules	# expected worst-case adjusted groups total
1:1:0	3,694	9.4%	576	1.5%	3,118	8.0%	3,968	3,464
1:0:1	4,036	10.3%	452	1.2%	3,584	9.2%	3,968	3,464
0:1:1	4,174	10.7%	484	1.2%	3,690	9.4%	3,968	3,464
total these groups	11,904	30.4%	1,512	3.9%	10,392	26.6%	11,904	10,392
total all groups	39,099					χ^2 test p-value	2.07E-07	2.46E-12

Table S26.

Phylogenomic determination of the origin of ZIP4 (within Ph1 locus on chromosome 5B). Phylogenomic analysis indicates the ancestral locus of ZIP4 is a gene model on chr3B. This inparalog has homoeologs on 3A (TraesCS3A01G401700) and 3D (TraesCS3D01G396500) and its orthologs, Sc3Loc00242129.3, HORVU3Hr1G090150.1, HVUL3H17755.2, EMT19842 are all localized on chromosome 3. The inparalog on 3B is part of syntenic blocks A:B ID=255 and B:D ID=985 (all in compartment R2b). ZIP4 is located in compartment R3 in a non-syntenic location and, although TEs are not implicated, it may be a transduplication of the original 3B locus. The 3B gene model and its homoeologs are all in transcription expression module 2, the ancestral state of the family. In contrast, the translocated copy on 5B does not associate with any of the modules, consistent with evidence of expression divergence (correlation coefficient of 5B with 3D and 3B: (TraesCS5B01G255100 | TraesCS3D01G396500 = 0.482 compared to TraesCS3B01G434600 | TraesCS3D01G396500 = 0.786). The 5B locus also has higher Ka/Ks ratios compared with the 3A and 3D homoeologs or the inparalog on 3B. Recently, a single mutant of the *ZIP4* 5B locus was sufficient to increase homoeologous crossovers when crossed with wild relatives (25), providing functional evidence for the unique role of the 5B locus.

x locus	y locus	Ka/Ks
TraesCS3A01G401700	TraesCS3B01G434600	0.401
TraesCS3A01G401700	TraesCS5B01G255100	0.586
TraesCS3B01G434600	TraesCS3D01G396500	0.400
TraesCS3B01G434600	TraesCS5B01G255100	0.626
TraesCS5B01G255100	TraesCS3D01G396500	0.645

Table S27.

Tandemly repeated genes and clusters in five monocotyledonous species: bread wheat (ta, v1.0 annotation), barley (hv), *Brachypodium distachyon* (bd), rice (os) and sorghum (sbi). For bread wheat HC genes an all-against-all blastp similarity table was used to construct an undirected graph with nodes representing genes and edges between two nodes if the following criteria were met: (i) an e-value $\leq 1e^{-20}$, (ii) a maximal distance of 10 genes in the genome between the gene pair, and (iii) a minimal alignment coverage $\geq 70\%$ of the shorter gene. Tandem genes and clusters were retrieved as connected components of this graph. The 9,616 clusters containing 29,737 high confidence genes detected in bread wheat were compared with other monocotyledonous species by repeating the approach with the proteomes of barley (Hv v1.0 (146)), rice (186), brachypodium and sorghum (v3.1 from <https://phytozome.jgi.doe.gov/pz/portal.html>). The percentages of tandem genes in column 5 are related to the total numbers (column 4) of genes located on pseudochromosomes of the respective species. For wheat, this excludes gene models located on chromosome “Unknown”. The table also shows the distribution of tandem genes across the A(ta-A), B(ta-B) and D(ta-D) sub-genomes.

	Number of genes	Number of clusters	Total genes	% Tandem genes
ta	29,737	9,616	108,061	27.5
hv	7,499	2,888	37,577	20.0
bd	5,235	1,879	34,303	15.3
os	6,702	2,139	38,864	17.2
sbi	6,226	2,130	34,027	18.3
ta-A	9,462	3,102	36,302	26.1
ta-B	10,735	3,385	36,738	29.2
ta-D	9,540	3,129	35,021	27.2

Table S28.

Inversions between homeologous chromosomes across the wheat genome. The numbers of inversions between chromosome pairs were determined after filtering ‘uncertain regions’ of pseudomolecules, all scaffolds with unknown (?) orientations in the AGP and removing inversions that have more than 30% of the genes in an ‘uncertain region’.

5

Chromosome pairs	Inversions with ≥ 10 genes
1A_1B	9
1A_1D	9
1B_1D	9
2A_2B	8
2A_2D	11
2B_2D	6
3A_3B	10
3A_3D	10
3B_3D	9
4A_4B	9
4A_4D	10
4B_4D	6
5A_5B	10
5A_5D	8
5B_5D	8
6A_6B	11
6A_6D	12
6B_6D	7
7A_7B	12
7A_7D	9
7B_7D	5

Table S29.

Structural and functional partitioning of wheat chromosomes. The segmentation is based on the distribution of recombination rate, gene and transposable element density, gene expression breadth, percentage of H3K36me3-, H3K9ac-, and H3K27me3-associated genes.

		Boundary location (Mb)				Region length (Mb)					Chromosomal fraction				
Chromosome	Chromosome length (Mb)	R1/R2a	R2a/C	C/R2b	R2b/R3	R1	R2a	C	R2b	R3	R1	R2a	C	R2b	R3
chr1A	594	59	151	231	480	59	92	80	249	114	9.9%	15.5%	13.5%	41.9%	19.2%
chr1B	689	62	172	277	534	62	110	105	257	155	9.0%	16.0%	15.2%	37.3%	22.5%
chr1D	495	29	98	171	385	29	69	73	214	110	5.9%	13.9%	14.7%	43.2%	22.2%
chr2A	780	42	206	379	662	42	164	173	283	118	5.4%	21.0%	22.2%	36.3%	15.1%
chr2B	801	59	248	433	660	59	189	185	227	141	7.4%	23.6%	23.1%	28.3%	17.6%
chr2D	651	37	192	338	520	37	155	146	182	131	5.7%	23.8%	22.4%	28.0%	20.1%
chr3A	750	62	249	414	670	62	187	165	256	80	8.3%	24.9%	22.0%	34.1%	10.7%
chr3B	830	66	257	407	728	66	191	150	321	102	8.0%	23.0%	18.1%	38.7%	12.3%
chr3D	615	49	167	287	543	49	118	120	256	72	8.0%	19.2%	19.5%	41.6%	11.7%
chr4A	744	41	180	414	594	41	139	234	180	150	5.5%	18.7%	31.5%	24.2%	20.2%
chr4B	673	42	186	360	537	42	144	174	177	136	6.2%	21.4%	25.9%	26.3%	20.2%
chr4D	509	10	135	288	432	10	125	153	144	77	2.0%	24.6%	30.1%	28.3%	15.1%
chr5A	709	39	140	260	427	39	101	120	167	282	5.5%	14.2%	16.9%	23.6%	39.8%
chr5B	713	52	140	221	430	52	88	81	209	283	7.3%	12.3%	11.4%	29.3%	39.7%
chr5D	565	46	128	207	345	46	82	79	138	220	8.1%	14.5%	14.0%	24.4%	38.9%
chr6A	617	46	216	409	556	46	170	193	147	61	7.5%	27.6%	31.3%	23.8%	9.9%
chr6B	720	56	221	429	651	56	165	208	222	69	7.8%	22.9%	28.9%	30.8%	9.6%
chr6D	473	44	164	280	410	44	120	116	130	63	9.3%	25.4%	24.5%	27.5%	13.3%
chr7A	736	89	239	416	659	89	150	177	243	77	12.1%	20.4%	24.0%	33.0%	10.5%
chr7B	750	12	146	418	660	12	134	272	242	90	1.6%	17.9%	36.3%	32.3%	12.0%
chr7D	638	84	200	373	552	84	116	173	179	86	13.2%	18.2%	27.1%	28.1%	13.5%

Table S30.

Publicly available RNA-Seq samples used in this study.

SRA	Samples	Study details	Analysed expVIP	Total reads	Mapped reads	%
DRP000768	12	Phosphate starvation in roots and shoots	Y	118,053,746	100,948,404	85.51
ERP004505	30	Grain tissue-specific developmental timecourse	Y	873,709,556	718,254,718	82.21
ERP008767	9	Grain tissue-specific expression	Y	45,213,827	36,308,713	80.30
SRP004884	3	Flag leaf downregulation of <i>GPC</i>	Y	102,103,001	65,221,796	63.88
SRP013449	6	Grain tissue-specific developmental timecourse	Y	132,702,451	105,655,016	79.62
SRP017303	1	Seedling leaf	Y	33,361,836	15,232,257	45.66
SRP022869	4	Seedling leaf	Y	49,619,979	41,417,307	83.47
ERP003465	60	Fusarium graminearum infected spikelets	Y	1,827,362,091	1,568,753,997	85.85
SRP028357	4	Fifth leaf shoots/roots	Y	133,881,441	119,685,598	89.40
SRP029372	7	Grain tissue-specific developmental timecourse	Y	101,477,759	38,334,982	37.78
SRP038912	3	Stamen, pistil and pistillody stamen	Y	217,315,378	189,209,346	87.07
SRP041017	21	Powdery mildew and yellow rust infection timecourse	Y	395,463,786	327,718,364	82.87
SRP045409	14	Drought and heat stress time-course in seedlings	Y	921,578,806	688,753,403	74.74
SRP056412	42	Grain developmental timecourse (4A dormancy QTL)	Y	615,077,336	419,124,669	68.14
ERP004714	30	Developmental time-course of CS	Y	1,536,051,415	1,320,450,267	85.96
ERP016738	6	Developmental time-course of CS	N	824,241,135	663,188,969	80.46
ERP013829	72	Fusarium graminearum timecourse (rachis, palea, lemma)	N	2,061,182,960	1,649,533,020	80.03
SRP078208	8	Fusarium pseudograminearum infected coleoptile	N	161,807,328	137,267,473	84.83
ERP015130	16	Magnaporthe oryzae infected leaf	N	351,588,058	207,187,100	58.93
SRP068165	24	PEG timecourse leaves	N	1,054,250,452	944,502,318	89.59
SRP068156	14	Fusarium graminearum + hormone infected spikelets	N	561,431,175	476,370,730	84.85
SRP067916	5	Flag leaf timecourses	N	165,495,247	142,565,852	86.14
SRP064598	9	Microspores cold treatment	N	608,252,566	401,980,553	66.09
ERP009837	30	Zymoseptoria tritici leaf infection timecourse	N	1,081,355,610	905,111,026	83.70
SRP060670	6	Fusarium graminearum infected rachis	N	61,289,682	40,916,842	66.76
SRP048912	48	Fusarium graminearum infected shoots	N	769,083,359	651,917,469	84.77
SRP043554	6	Cold stress leaves	N	224,280,385	194,356,916	86.66
ERP013983	39	Yellow rust infection timecourse	N	866,322,161	687,543,489	79.36

Table S31.

Summary of unpublished RNA-Seq samples used in this study

Study details	Samples	SRA	Total reads	Mapped reads	%
Development time-course	209	PRJEB25639	8,269,498,786	6,603,975,140	79.86%
PAMP Triggered Immune Response	21	PRJEB23056	2,436,781,618	2,156,752,129	88.51%
CS Spike	10	PRJNA436817	429,796,610	333,383,697	77.57%
Developing spike	61	PRJEB25640	780,374,855	362,666,932	46.47%
CS tissues	13	SRP133837	544,053,283	467,082,558	85.85%
Aneuploidy controls	7	PRJEB25593	752,040,275	667,605,170	88.77%

Table S32.

Number of samples and tissue complexity of intermediate tissue represented in the 850 RNA-Seq samples. ¹ Detailed description of wheat-specific terminology: <http://bio-gromit.bio.bris.ac.uk/cerealgenomics/WheatBP/glossary.htm>

Intermediate tissue ¹	Sample number	High-level tissue	Genes accounting for 50% tpm
Aleurone	12	Grain	803
Aleurone layer and starchy endosperm	4	Grain	47
Embryo	3	Grain	1,877
Endosperm	25	Grain	57
Endosperm + seed coat	3	Grain	247
Grain hard dough and ripening	22	Grain	289
Grain milk and soft dough	40	Grain	143
Seed coat	6	Grain	1,062
Transfer cells	4	Grain	46
Flag leaf	26	Leaves	891
Flag leaf blade	30	Leaves	1,646
Flag leaf sheath	12	Leaves	2,154
Internode	15	Leaves	3,909
Leaf blades excluding flag leaf	27	Leaves	1,345
Leaf excluding flag leaf	73	Leaves	1,576
Leaf ligule	3	Leaves	2,557
Leaf sheaths excluding flag leaf	12	Leaves	3,408
Peduncle	9	Leaves	2,194
Seedling aerial tissues	164	Leaves	2,321
Shoot apical meristem	6	Leaves	3,110
Shoot axis	12	Leaves	3,676
Vegetative aerial tissues	14	Leaves	1,473
Root apical meristem	6	Root	2,709
Roots	44	Root	4,569
Anther	5	Spike	3,182
Awns	6	Spike	1,985
Glumes	12	Spike	3,351
Microspores	9	Spike	1,042
Rachis	8	Spike	3,372
Spike	81	Spike	925
Spikelets	149	Spike	5,197
Stigma & ovary	8	Spike	3,652

Table S33.

Over- and under-represented GO slim terms based on chromosome compartment, Empty cells correspond to non-significant values.

Ontology	GO_slim	Description	Over-represented			Under-represented		
			Distal	Interstitial	Proximal	Distal	Interstitial	Proximal
BP	GO:0008219	cell death	1.4E-18				0.0E+00	
	GO:0009607	response to biotic stimulus	1.5E-16				0.0E+00	3.2E-03
	GO:0019748	secondary metabolic process	2.2E-13				0.0E+00	3.3E-03
	GO:0009605	response to external stimulus	9.0E-10				0.0E+00	
	GO:0006950	response to stress	4.8E-09				4.8E-09	2.3E-02
	GO:0009719	response to endogenous stimulus	5.8E-08				5.3E-07	3.4E-06
	GO:0007154	cell communication	3.1E-07				3.0E-06	
	GO:0009875	pollen-pistil interaction	2.1E-05				5.5E-05	
	GO:0009991	response to extracellular stimulus	2.7E-02				2.7E-02	
		nucleobase-containing compound metabolic process						
	GO:0006139	process		1.5E-47	1.4E-11	0.0E+00		
	GO:0016043	cellular component organization		3.1E-27	3.0E-10	0.0E+00		
	GO:0009987	cellular process		2.4E-24	7.4E-12	0.0E+00		
	GO:0008152	metabolic process		3.5E-22	5.7E-04	0.0E+00		
	GO:0009058	biosynthetic process		3.3E-20	2.0E-03	0.0E+00		
	GO:0006259	DNA metabolic process		5.6E-20	5.7E-07	0.0E+00		
	GO:0008150	biological_process		5.8E-18	9.4E-07	0.0E+00		
		generation of precursor metabolites and energy		6.9E-15	7.7E-12	0.0E+00		
	GO:0005975	carbohydrate metabolic process		9.4E-13		0.0E+00		
	GO:0007049	cell cycle		7.7E-10	1.2E-07	0.0E+00		
	GO:0030154	cell differentiation		2.8E-09	2.3E-02	8.9E-09		
	GO:0006810	transport		4.3E-09	1.2E-07	2.8E-08		
	GO:0007275	multicellular organism development		3.8E-08	1.8E-03	4.0E-07		
	GO:0006412	translation		1.3E-07	3.9E-08	5.1E-07		
	GO:0040029	regulation of gene expression, epigenetic		5.7E-07	8.6E-08	2.3E-06		
	GO:0009628	response to abiotic stimulus		5.4E-06		1.0E-05		
	GO:0000003	reproduction		3.0E-05	1.2E-02	3.4E-05		
	GO:0009606	tropism		6.9E-05	7.5E-03	8.6E-05		
	GO:0009791	post-embryonic development		5.5E-03		5.5E-03		
	GO:0015979	photosynthesis		7.8E-03	1.0E-12	7.9E-03		
	GO:0009653	anatomical structure morphogenesis		9.3E-03		9.2E-03		
	GO:0016049	cell growth		2.0E-02		2.0E-02		
	GO:0019538	protein metabolic process			1.2E-07			

	GO:0006464	cellular protein modification process		1.2E-03			
CC	GO:0030312	external encapsulating structure	4.2E-11			0.0E+00	1.5E-02
	GO:0005618	cell wall	3.1E-07			2.0E-06	2.0E-03
	GO:0005575	cellular component		2.9E-246	3.9E-57	0.0E+00	
	GO:0005623	cell		1.6E-225	6.5E-64	0.0E+00	
	GO:0005622	intracellular		1.4E-207	7.8E-60	0.0E+00	
	GO:0005737	cytoplasm		2.6E-104	9.8E-33	0.0E+00	
	GO:0009536	plastid		1.7E-22	3.8E-17	0.0E+00	
	GO:0016020	membrane		4.5E-19	1.3E-13	0.0E+00	
	GO:0005739	mitochondrion		1.7E-16	2.6E-08	0.0E+00	
	GO:0005634	nucleus		1.4E-14	2.4E-08	0.0E+00	
	GO:0005856	cytoskeleton		2.1E-14	3.9E-04	0.0E+00	
	GO:0005794	Golgi apparatus		4.6E-12	4.4E-04	0.0E+00	
	GO:0005654	nucleoplasm		1.2E-10	1.0E-04	0.0E+00	
	GO:0005635	nuclear envelope		1.3E-05		3.1E-05	
	GO:0005773	vacuole		4.6E-03		4.5E-03	
	GO:0005777	peroxisome		1.0E-02		1.0E-02	
	GO:0005840	ribosome			4.7E-08		
	GO:0009579	thylakoid			4.1E-07		
	GO:0005829	cytosol			2.3E-03		
MF	GO:0000166	nucleotide binding	1.6E-39			0.0E+00	
	GO:0030246	carbohydrate binding	6.8E-36			0.0E+00	3.2E-03
	GO:0016301	kinase activity	1.8E-19			0.0E+00	
	GO:0016740	transferase activity	1.2E-13			0.0E+00	8.7E-08
	GO:0003674	molecular function	9.6E-06			2.3E-05	
	GO:0004871	signal transducer activity	4.0E-04			4.2E-04	
	GO:0004872	receptor activity	7.9E-03			7.9E-03	
	GO:0030234	enzyme regulator activity	1.8E-02			1.8E-02	
	GO:0005488	binding	3.4E-02			3.4E-02	
	GO:0003676	nucleic acid binding		5.2E-23	5.6E-08	0.0E+00	
	GO:0016787	hydrolase activity		9.3E-23	1.7E-04	0.0E+00	
	GO:0005215	transporter activity		2.3E-17	4.4E-04	0.0E+00	
	GO:0003723	RNA binding		5.6E-14	6.7E-20	0.0E+00	
	GO:0004518	nuclease activity		5.9E-10	4.3E-04	0.0E+00	
	GO:0005198	structural molecule activity		3.7E-09	9.6E-10	1.0E-08	
	GO:0003700	transcription factor activity, sequence-specific DNA binding		1.4E-04		1.4E-04	
	GO:0008135	translation factor activity, RNA binding		2.3E-04	1.6E-03	2.4E-04	
	GO:0003774	motor activity		3.2E-03		3.2E-03	
	GO:0003677	DNA binding		4.9E-02		4.9E-02	

GO:0005515	protein binding	1.3E-03
GO:0003682	chromatin binding	1.2E-02

Table S34.

Module assignment of homeologs for syntenic and non-syntenic triads and duplets. ¹ Triads/duplets with all homeologs assigned to a module, ² Triads/duplets with two (or three) homeologs further away than 50 % of the median distance to the furthest eigengene, ³ Percentage based on total triads/duplets.

Type (A:B:D)	Syteny	Triad/ Duplet ¹	Homoeolog assignment		Divergent expression ²	Homoeologs in same module (%)	Homoeologs in different modules (%)	Divergent expression (%) ³
			Same module	Different modules				
1:1:1	syntenic	8,617	4,997	3,620	1,393 (78)	58.0	42.0	16.2
0:1:1	syntenic	591	342	249	99	57.9	42.1	16.8
1:0:1	syntenic	684	415	269	124	60.7	39.3	18.1
1:1:0	syntenic	435	245	190	91	56.3	43.7	20.9
1:1:1	non-syntenic	392	191	201	83 (4)	48.7	51.3	21.2
0:1:1	non-syntenic	72	37	35	17	51.4	48.6	23.6
1:0:1	non-syntenic	80	45	35	20	56.3	43.8	25.0
1:1:0	non-syntenic	71	36	35	18	50.7	49.3	25.4
Triads (1:1:1)	syntenic + non-syntenic	9,009	5,188	3,821	1,476 (82)	57.6	42.4	16.4
Duplets	syntenic + non-syntenic	1,933	1,120	813	369	57.9	42.1	19.1
Total		10,942	6,308	4,634	1,845 (1,927)	57.6	42.4	16.9

Table S35.

Top 10 hub genes ranked by correlation to the module eigengene (kME) in WGCNA co-expression modules 8 and 11 (enriched for floral organ identity GO:0010093).

Module	Gene	kME correlation	kME pvalue	Human Readable annotation
8	TraesCS2A01G017100	0.95	0	NADH-quinone oxidoreductase subunit F
8	TraesCS2A01G545900	0.93	0	RING/U-box superfamily protein
8	TraesCS2D01G122200	0.93	0	Photosystem II reaction center protein M
8	TraesCS2D01G122300	0.92	0	Delta(24(24(1)))-sterol reductase
8	TraesCS4D01G247000	0.93	0	Chorismate synthase; RING/FYVE/PHD zinc finger superfamily protein
8	TraesCS5D01G514200	0.93	0	Threonine--tRNA ligase
8	TraesCS6B01G083400	0.93	0	F-box family protein-like protein
8	TraesCS6D01G058000	0.93	0	F-box family protein
8	TraesCS7D01G070800	0.93	0	F-box family protein
8	TraesCSU01G225400	0.94	0	Ribulokinase
11	TraesCS1A01G439000	0.91	0	3-ketoacyl-CoA synthase
11	TraesCS2A01G467200	0.92	0	Homeodomain-like protein
11	TraesCS2D01G468300	0.90	0	Aminodeoxychorismate synthase
11	TraesCS3B01G030300	0.92	0	Cytochrome P450
11	TraesCS3B01G030400	0.94	0	Dirigent protein
11	TraesCS4D01G247200	0.90	0	Cytochrome P450
11	TraesCS5D01G234100	0.92	0	Sugar transporter, putative
11	TraesCS5D01G293100	0.90	0	Auxin efflux carrier component
11	TraesCS7D01G020900	0.90	0	Benzyl alcohol O-benzoyltransferase
11	TraesCS7D01G220700	0.91	0	Protein kinase-like protein

Table S36.

Gene families in the wheat genome. Overview of gene families identified in bread wheat in comparison to 13 organisms belonging to the green lineage. Unless stated otherwise, numbers are based on TE-filtered gene sets comprising both HC and LC loci, Family percentages are relative to the number of families with at least one gene from one of the wheat sub-genomes (first row). Contracted gene families do comprise also families with no genes in either sub-genome.

5

	# total	% of gene families	# genes in A	# genes in B	# genes in D	# total genes	% genes A	% genes B	% genes D	%genes total
wheat gene families	26,080	100.0%	53,527	55,726	51,183	160,436	88.7%	87.2%	90.2%	88.6%
wheat families conserved in other species	21,751	83.4%	49,559	51,372	47,756	148,687	82.1%	80.4%	84.1%	82.1%
wheat families conserved outside Triticeae	16,526	63.4%	44,737	47,919	44,177	136,833	74.2%	74.9%	77.8%	75.6%
wheat families with pseudogenes	12,265	47.0%	40,289	42,877	38,817	121,983	66.8%	67.1%	68.4%	67.4%
wheat families with pseudogenes conserved in other species	11,461	43.9%	39,310	41,727	38,006	119,043	65.2%	65.3%	66.9%	65.8%
wheat families with pseudogenes conserved outside Triticeae	8,725	33.5%	35,036	38,093	34,731	107,860	58.1%	59.6%	61.2%	59.6%
wheat pseudogene-only families	505	1.9%	251	247	253	751	0.4%	0.4%	0.4%	0.4%
wheat pseudogene-only families conserved in other species	424	1.6%	189	164	197	550	0.3%	0.3%	0.3%	0.3%

wheat pseudogene-only families conserved outside Triticeae	122	0.5%	64	60	51	175	0.1%	0.1%	0.1%	0.1%
expanded wheat gene families (FDR<0.1)	8,592	32.9%	22,962	24,492	20,974	68,428	38.1%	38.3%	36.9%	37.8%
expanded wheat gene families with pseudogenes (FDR<0.1)	3,547	13.6%	18,516	20,569	17,377	56,462	30.7%	32.2%	30.6%	31.2%
contracted wheat gene families (FDR<0.1)	78	0.3%	83	67	65	215	0.1%	0.1%	0.1%	0.1%
wheat gene families similarly expanded in all subgenomes (FDR<0.1)	6,216	23.8%	20,456	23,651	19,970	64,077	33.9%	37.0%	35.2%	35.4%
wheat gene families expanded in one subgenome (FDR<0.1)	1,109	4.3%	1,718	1,655	1,016	4,389	2.8%	2.6%	1.8%	2.4%
wheat gene families expanded in one subgenome with pseudogenes (FDR<0.1)	387	1.5%	720	576	384	1,680	1.2%	0.9%	0.7%	0.9%
wheat gene families expanded in one subgenome lineage (FDR<0.1)	2,102	8.1%	2,305	319	792	3,416	3.8%	0.5%	1.4%	1.9%
wheat gene families	986	3.8%	1,294	228	375	1,897	2.1%	0.4%	0.7%	1.0%

expanded in one subgenome lineage with pseudogenes (FDR<0.1)										
HC+LC subgenome orphan (non- pseudogenic/non- TE)			9,333	11,075	8,001	28,409	18.6%	21.3%	16.7%	18.9%
HC (non- pseudogenic/non- TE)			35,251	35,545	34,208	105,004	70.2%	68.5%	71.3%	70.0%
HC+LC (non- pseudogenic/non- TE)			50,224	51,888	47,993	150,105	83.3%	81.2%	84.5%	82.9%
total genes (non- TE)			60,328	63,935	56,773	181,036	100.0%	100.0%	100.0%	100.0%

Table S37.

Overlap of manually annotated gene families and gene families identified using the automated phylogenomic approach. 85.24-100.00% of the manually annotated gene models were found in the respective family using the automated approach. Those that were not discovered using the automated approach did not fall in orthologous groups (i.e. were singletons), had no conserved domain structure or were only added, or significantly changed within the gene curation efforts for annotation version 1.1 (as this automated analysis was based on v1.0 gene models)

5

Gene Family	Total number of genes manually assigned to gene family	Number of genes identified by automated approach in common with expert curation	% of genes identified by automated approach in common with expert curation
<i>Aquaporins</i>	158	146	92.41
<i>CBFs</i>	61	54	88.52
<i>Dehydrins</i>	49	49	100.00
<i>NLRs</i>	2,052	1,811	88.26
<i>PPRs</i>	147	130	88.44
<i>Prolamins</i>	828	706	85.27
<i>WAKs</i>	555	555	100.00

Table S38.

Summary of the detailed analysis of tree topologies for the orthologous groups (OGs) of the Aquaporins, CBFs and DHN candidate gene families.

Gene Family	Orthologous Group	Topology	Bootstrap Value	% of mono- phyletic OGs	% of monophyletic OGs low BS values removed	% of monophyletic expanded OGs
<i>Aquaporins</i>	OG0000154	split at root	NA	90.91	100.00	100.00
	OG0000427	split at root	NA			
	OG0001355	monophyletic	0.886			
	OG0002982	monophyletic	1.000			
	OG0003682	monophyletic	0.444			
	OG0007810	polyphyletic	0.444			
	OG0010867	monophyletic	0.884			
	OG0011484	split at root	NA			
	OG0011556	monophyletic	1.000			
	OG0011644	monophyletic	0.606			
	OG0012131	monophyletic	1.000			
	OG0014331	monophyletic	1.000			
	OG0021232	monophyletic	0.762			
	OG0027911	monophyletic	0.864			
<i>CBFs</i>	OG0000383	monophyletic	1.000	100.00	100.00	100.00
	OG0023196	monophyletic	0.999			
	OG0029391	(one sequence)	NA			
<i>Dehydrins</i>	OG0000810	monophyletic	0.938	100.00	100.00	100.00
	OG0003562	split at root	NA			
	OG0012150	monophyletic	1.000			
	OG0014804	monophyletic	0.968			

Table S39.

Comparison of gene family completeness in different *A. tauschii* genome assemblies. Selected sub-phylogenies of the prolamin gene family were evaluated for gene numbers in the two independent annotations of a draft assembly (54) and a recent high quality genome assembly (137) of *A. tauschii* as well as the IWGSC RefSeq Annotation v1.1 (HC gene models predicted on the D subgenome only). The sub-phylogenies comprise grain quality genes that are part of gluten (gamma gliadins and low molecular weight glutenins), and several alpha-amylase/trypsin inhibitor (ATI) subgroups - proteases involved in resistance to parasites.

gene family	<i>A. tauschii</i> 2013	<i>A. tauschii</i> 2017	IWGSC RefSeq Annotation v1.1 D-genome HC	Difference 2013-2017
Gamma gliadin	5	6	5	1
LMW glutenin	4	5	7	1
Mono-/dimeric ATI	3	3	3	0
ATI - CM1/CM16 group	2	2	2	0
ATI - CM3 group	1	1	1	0
mATI - Cmx group	5	5	5	0
ATI - Trypsin Inhibitor group	0	1	0	1
Total	20	23	23	3
Median	3	3	3	0
Mean	2.86	3.29	3.29	0.43

Table S40.

Comparison of gene models to Ta3B and IWGSC RefSeq v1.0 Chr.3B.

Gene Models	Ta3B ^a	RefSeq v1.0	Difference
MIPS 2.2 Annotations	86%	93.7%	+9.0%
RefSeq 1.0 Annotations	88%	100%	+11.7%
Ta3B Annotations	100%	98.9%	-0.6%
Pseudomolecule size	774 Mbp	831 Mbp	57 Mbp

^a Analysis was restricted to the pseudomolecule assembly of Ta3B (18).

Table S41.IWGSC RefSeq v1.0 genes from *SSt1* absent in Ta3B.

Gene	Functional Annotation	Position
<i>TraesCS3B01G598900</i>	Ankyrin repeat-containing protein	820,286,268
<i>TraesCS3B01G599000</i>	F-box domain containing protein, expressed	820,333,146
<i>TraesCS3B01G600100</i>	Beta-adaptin-like protein	820,789,324
<i>TraesCS3B01G600200</i>	Dual-specificity RNA methyltransferase RlmN	820,893,665
<i>TraesCS3B01G600300</i>	Protein kinase	820,902,164
<i>TraesCS3B01G605300</i>	ubiquitin carboxyl-terminal hydrolase-like protein, putative (DUF627 and DUF629)	823,761,895
<i>TraesCS3B01G605400</i>	Divalent ion symporter	823,894,387
<i>TraesCS3B01G606600</i>	.	825,644,346
<i>TraesCS3B01G606700</i>	Disease resistance protein (TIR-NBS-LRR class) family	825,710,109
<i>TraesCS3B01G606800</i>	rRNA N-glycosidase	825,771,506
<i>TraesCS3B01G606900</i>	Epoxide hydrolase 2	825,849,367
<i>TraesCS3B01G607000</i>	Aspartic proteinase nepenthesin-1	825,868,243
<i>TraesCS3B01G607100</i>	Aspartic proteinase nepenthesin-1	825,886,089
<i>TraesCS3B01G607200</i>	SNARE-interacting protein KEULE	825,950,185
<i>TraesCS3B01G607300</i>	Pectinesterase inhibitor	825,970,605
<i>TraesCS3B01G607400</i>	Plant/T31B5-30 protein	825,992,065
<i>TraesCS3B01G608200</i>	Vacuolar fusion protein MON1	826,645,944
<i>TraesCS3B01G608800</i>	Dof zinc finger protein	828,110,909
<i>TraesCS3B01G609000</i>	Dof zinc finger protein	828,292,221
<i>TraesCS3B01G612300</i>	Transcription factor, MADS-box	829,742,374
<i>TraesCS3B01G612500</i>	Transcription factor, MADS-box	829,935,275
<i>TraesCS3B01G612600</i>	Transcription factor, MADS-box	830,055,777
<i>TraesCS3B01G612800</i>	Transcription factor, MADS-box	830,320,830
<i>TraesCS3B01G612900</i>	Transcription factor, MADS-box	830,610,856

Table S42

RNA-Seq based expression analysis of high confidence genes within *SSt1* interval on Chromosome 3B.

Transcript ID	Functional Annotation	CSV1.0 Position	Log ₂ Fold Change		Adjusted p-value	
			LGB3B_vs_LDN	Lillian_vs_Vesper	LGB3B_vs_LDN	Lillian_vs_Vesper
TraesCS3B01G596700	2 Receptor-like protein kinase	819,502,902	-1.5	-1.1	0.0	0.0
TraesCS3B01G596800	Cytochrome P450	819,522,398	-2.1	-0.7	0.0	0.1
TraesCS3B01G596900	Receptor-like protein kinase	819,629,205	-1.8	0.0	0.0	1.0
TraesCS3B01G597000	Cytochrome P450	819,650,047	-2.4	-1.7	0.0	0.0
TraesCS3B01G597100	Phosphoenolpyruvate carboxykinase (ATP)	819,704,491	.	.	.	.
TraesCS3B01G597200	Cytochrome P450	819,740,077	.	.	.	.
TraesCS3B01G597300	Ae3	819,750,522	.	.	.	.
TraesCS3B01G597400	Pentatricopeptide repeat-containing protein	819,785,500	-0.2	0.3	0.9	0.8
TraesCS3B01G597500	Telomere-binding family protein	819,788,128	0.0	-0.3	1.0	0.5
TraesCS3B01G597600	Rab5-interacting family protein	819,809,430	0.1	-0.3	0.9	0.5
TraesCS3B01G597700	Haloacid dehalogenase-like hydrolase	819,854,349	0.1	-2.7	0.9	0.0
TraesCS3B01G597800	Vacuolar-processing enzyme	819,896,710	0.9	0.4	1.0	1.0
TraesCS3B01G597900	Vacuolar-processing enzyme	819,930,084	0.8	1.2	0.4	0.2
TraesCS3B01G598000	F-box family protein	819,948,218	.	-0.7	.	1.0
TraesCS3B01G598100	Pectinesterase	819,955,611	.	.	.	.
TraesCS3B01G598200	Glycosyltransferase	819,990,507	0.0	7.1	1.0	0.0
TraesCS3B01G598300	F-box family protein	820,074,306	-0.1	0.7	1.0	1.0
TraesCS3B01G598400	Glycosyltransferase	820,114,465	.	1.4	.	0.2
TraesCS3B01G598500	Glycosyltransferase	820,123,501	.	.	.	.
TraesCS3B01G598600	Ankyrin repeat-containing protein	820,130,589	-0.6	-1.7	0.8	0.2
TraesCS3B01G598700	Ankyrin repeat-containing protein	820,143,235	.	.	.	.
TraesCS3B01G598800	Ankyrin repeat-containing protein	820,163,917	-0.4	-1.0	0.8	1.0
TraesCS3B01G598900	Ankyrin repeat-containing protein	820,286,268	.	.	.	.
TraesCS3B01G599000	F-box domain containing	820,333,146	0.3	1.3	0.9	1.0
TraesCS3B01G599100	Ankyrin repeat-containing protein	820,397,835	.	.	.	.
TraesCS3B01G599200	Ankyrin repeat-containing protein	820,426,304	.	.	.	.
TraesCS3B01G599300	Ankyrin repeat-containing protein	820,446,972	0.3	1.8	0.9	0.1
TraesCS3B01G599400	Ankyrin repeat-containing protein	820,457,464	.	.	.	.
TraesCS3B01G599500	Kinetochore protein spc25	820,463,589	-0.3	0.5	0.8	0.6
TraesCS3B01G599600	Beta-adaptin-like protein	820,498,086	0.0	-1.9	1.0	0.0
TraesCS3B01G599700	NAC domain-containing protein	820,506,142	.	.	.	.
TraesCS3B01G599800	AP complex subunit	820,672,721	-0.4	-0.1	0.7	0.9
TraesCS3B01G599900	NAC domain-containing protein	820,761,513	.	.	.	.
TraesCS3B01G600000	NAC domain-containing protein	820,772,091	.	.	.	.
TraesCS3B01G600100	Beta-adaptin-like protein	820,789,324	0.0	1.6	1.0	0.1
TraesCS3B01G600200	Dual-specificity RNA methyltransferase	820,893,665	-0.2	0.1	0.9	0.9
TraesCS3B01G600300	Protein kinase	820,902,164	0.2	0.9	0.9	0.2
TraesCS3B01G600400	.	821,081,510	0.3	0.1	0.9	1.0
TraesCS3B01G600500	.	821,213,195	.	.	.	.
TraesCS3B01G600600	Calmodulin-binding protein, putative	821,215,369	-0.1	2.0	1.0	0.1
TraesCS3B01G600700	Calmodulin-binding protein, putative	821,276,027	.	.	.	.
TraesCS3B01G600800	Zinc finger family	821,554,671	.	5.6	.	0.0
TraesCS3B01G600900	Plant-specific domain TIGR01615	821,693,457	-0.4	0.2	0.6	0.6
TraesCS3B01G601000	Werner Syndrome-like exonuclease	821,804,687	.	.	.	.
TraesCS3B01G601100	Metallothionein	821,836,472	0.6	0.3	0.7	0.7

TraesCS3B01G601200	Vacuolar protein sorting	821,848,566	-0.3	-0.9	0.9	0.5
TraesCS3B01G601300	11S globulin seed	821,856,251	.	.	.	.
TraesCS3B01G601400	Sn1-specific diacylglycerol lipase	821,884,273	.	.	.	.
TraesCS3B01G601500	11S globulin seed	821,978,328	.	.	.	.
TraesCS3B01G601600	Metallothionein	822,039,201	0.7	-0.2	0.2	0.9
TraesCS3B01G601700	Metallothionein	822,067,205	0.8	1.0	0.3	0.1
TraesCS3B01G601800	Metallothionein	822,094,677	0.8	0.3	0.3	0.8
TraesCS3B01G601900	Plant invertase/pectin methylesterase	822,101,995	.	.	.	.
TraesCS3B01G602000	30S ribosomal protein	822,103,561	0.1	-0.6	1.0	0.3
TraesCS3B01G602100	30S ribosomal protein	822,114,565	-0.1	-0.7	1.0	0.0
TraesCS3B01G602200	SANT domain-containing protein	822,189,827	-0.1	-0.4	0.8	0.2
TraesCS3B01G602300	Mitochondrial ATP synthase	822,216,860	.	.	.	.
TraesCS3B01G602400	PR5-like receptor kinase	822,535,369	0.7	6.4	0.6	0.0
TraesCS3B01G602500	Lipoxygenase	822,548,132	-0.4	6.4	0.8	0.0
TraesCS3B01G602600	Extra-large guanine nucleotide	822,668,181	.	.	.	.
TraesCS3B01G602700	Protein kinase-like protein	822,676,261	2.5	1.6	0.0	0.0
TraesCS3B01G602800	Plant cadmium resistance	822,808,188	-1.8	-0.6	0.0	1.0
TraesCS3B01G602900	Dirigent protein	822,816,997	.	.	.	.
TraesCS3B01G603000	Gibberellin 2-beta-dioxygenase	822,819,355	-0.4	-4.3	0.8	0.0
TraesCS3B01G603100	Chaperone protein dnaJ	822,824,640	0.3	-7.0	0.7	0.0
TraesCS3B01G603200	SKP1-like protein	822,835,413	.	.	.	.
TraesCS3B01G603300	Kinase family protein	822,879,781	0.7	1.3	1.0	0.2
TraesCS3B01G603400	3 E3 SUMO-protein ligase	822,882,640	-0.1	0.2	0.9	0.6
TraesCS3B01G603500	Protein IQ-DOMAIN 1	822,903,506	0.0	-1.0	1.0	0.0
TraesCS3B01G603600	Serine/threonine-protein kinase ATM	822,968,033	.	.	.	.
TraesCS3B01G603700	Serine/threonine-protein kinase ATM	822,972,695	0.6	1.5	0.1	0.0
TraesCS3B01G603800	Kinase-like protein	822,977,716	0.7	-0.7	0.0	0.3
TraesCS3B01G603900	GATA transcription factor	822,984,216	2.5	2.7	0.0	0.0
TraesCS3B01G604000	Protein phosphatase 2C	823,028,880	0.5	0.4	0.5	0.5
TraesCS3B01G604100	NBS-LRR disease resistance	823,072,667	5.3	8.5	0.0	0.0
TraesCS3B01G604200	Ubiquitin carboxyl-terminal hydrolase	823,087,419	0.2	2.1	1.0	0.1
TraesCS3B01G604300	rRNA N-glycosidase	823,181,174	.	.	.	.
TraesCS3B01G604400	rRNA N-glycosidase	823,193,985	.	.	.	.
TraesCS3B01G604500	NBS-LRR resistance-like protein	823,290,293	0.1	0.7	1.0	1.0
TraesCS3B01G604600	2 Disease resistance protein	823,294,611	0.1	.	1.0	.
TraesCS3B01G604700	External alternative NAD(P)H-ubiquinone	823,328,283	-0.5	0.6	0.7	0.7
TraesCS3B01G604800	NBS-LRR-like resistance protein	823,425,742	6.7	7.2	0.0	0.0
TraesCS3B01G604900	Disease resistance protein	823,430,745	6.2	8.4	0.0	0.0
TraesCS3B01G605000	transmembrane protein, putative	823,551,897	-0.3	-0.8	1.0	1.0
TraesCS3B01G605100	Kinase interacting (KIP1-like)	823,573,764	0.2	1.0	0.9	0.0
TraesCS3B01G605200	Transmembrane protein, putative	823,592,606	.	.	.	.
TraesCS3B01G605300	ubiquitin carboxyl-terminal hydrolase-like	823,761,895	.	.	.	.
TraesCS3B01G605400	Divalent ion symporter	823,894,387	.	.	.	.
TraesCS3B01G605500	Cortactin-binding protein 2	823,984,063	0.2	1.5	0.9	1.0
TraesCS3B01G605600	AGAP002737-PA	824,006,142	0.4	1.2	0.9	1.0
TraesCS3B01G605700	Lipoxygenase	824,258,372	0.2	1.5	0.9	0.3
TraesCS3B01G605800	plant/protein	824,379,570	.	.	.	.
TraesCS3B01G605900	F-box protein	824,391,739	.	0.8	.	1.0
TraesCS3B01G606000	F-box protein	825,289,551	.	.	.	.
TraesCS3B01G606100	AGAP002737-PA	825,314,966	0.1	-2.1	1.0	0.1
TraesCS3B01G606200	AGAP002737-PA	825,492,715	0.2	-2.2	0.9	0.1
TraesCS3B01G606300	CAP-gly domain linker	825,495,597	-0.1	-1.8	1.0	0.2
TraesCS3B01G606400	Aspartic proteinase nepenthesin-1	825,539,090	.	.	.	.
TraesCS3B01G606500	plant/protein	825,571,873	1.1	0.6	0.1	0.7
TraesCS3B01G606600	.	825,644,346	.	.	.	.
TraesCS3B01G606700	Disease resistance protein	825,710,109	0.1	3.5	1.0	1.0
TraesCS3B01G606800	rRNA N-glycosidase	825,771,506	.	0.8	.	1.0
TraesCS3B01G606900	Epoxide hydrolase 2	825,849,367	0.1	0.6	1.0	1.0
TraesCS3B01G607000	Aspartic proteinase nepenthesin-1	825,868,243	.	.	.	.
TraesCS3B01G607100	Aspartic proteinase nepenthesin-1	825,886,089	.	.	.	.
TraesCS3B01G607200	SNARE-interacting protein KEULE	825,950,185	.	.	.	.
TraesCS3B01G607300	Pectinesterase inhibitor	825,970,605	.	.	.	.
TraesCS3B01G607400	Plant/T31B5-30 protein	825,992,065	.	0.8	.	1.0

TraesCS3B01G607500	Chaperone protein dnaJ	826,082,197	0.0	0.0	1.0	1.0
TraesCS3B01G607600	ABC transporter B	826,086,442	-0.2	-0.7	0.8	0.0
TraesCS3B01G607700	Disease resistance protein	826,101,725	.	.	.	.
TraesCS3B01G607800	ubiquinone biosynthesis protein	826,261,223	.	.	.	.
TraesCS3B01G607900	Dirigent protein	826,321,283	.	-1.4	.	0.3
TraesCS3B01G608000	Pro-apoptotic serine protease	826,363,262	.	.	.	.
TraesCS3B01G608100	Leucine-rich repeat (LRR)	826,484,373	.	.	.	.
TraesCS3B01G608200	Vacuolar fusion protein	826,645,944	-0.2	-0.9	0.8	0.0
TraesCS3B01G608300	Monopolar spindle protein	826,651,868	0.1	-1.0	1.0	0.5
TraesCS3B01G608400	GDSL esterase/lipase	826,711,412	0.2	-2.0	1.0	0.1
TraesCS3B01G608500	Aquaporin	826,725,085	0.9	1.6	0.2	0.0
TraesCS3B01G608600	MADS-box transcription factor	826,918,889	-0.6	-1.3	0.6	1.0
TraesCS3B01G608700	Agamous-like MADS-box protein	827,515,120	.	.	.	.
TraesCS3B01G608800	Dof zinc finger	828,110,909	-1.7	-2.3	0.0	0.0
TraesCS3B01G608900	Dof zinc finger	828,253,029	.	.	.	.
TraesCS3B01G609000	Dof zinc finger	828,292,221	.	.	.	.
TraesCS3B01G609100	Dof zinc finger	828,331,215	.	0.7	.	1.0
TraesCS3B01G609200	transmembrane protein	828,413,608	-0.2	-0.8	0.9	0.4
TraesCS3B01G609300	F-box protein-like protein	828,417,014	.	.	.	.
TraesCS3B01G609400	Cytochrome P450	828,420,001	-0.2	-0.7	1.0	0.7
TraesCS3B01G609500	Ankyrin repeat family	828,546,437	0.1	0.9	0.9	0.2
TraesCS3B01G609600	Cytochrome P450	828,763,303	-0.1	0.3	1.0	1.0
TraesCS3B01G609700	Nodulin MtN2/EamA-like	828,766,993	.	.	.	.
TraesCS3B01G609800	3-oxo-5-alpha-steroid 4-dehydrogenase family	828,770,429	-0.1	0.6	1.0	0.6
TraesCS3B01G609900	Pre-mRNA-splicing factor ISY1-like	828,781,704	.	.	.	.
TraesCS3B01G610000	2 Protein CHUP1, chloroplastic	828,825,898	.	.	.	.
TraesCS3B01G610100	2 Pectin acetyltransferase	828,917,506	-0.2	2.5	0.9	0.0
TraesCS3B01G610200	Pectin acetyltransferase	828,949,097	-1.1	-2.9	0.1	0.0
TraesCS3B01G610300	Pectin acetyltransferase	828,991,506	.	0.8	.	1.0
TraesCS3B01G610400	2 Pectin acetyltransferase	829,013,144	0.2	-0.4	0.9	0.3
TraesCS3B01G610500	91A protein	829,016,168	0.4	0.2	0.5	0.5
TraesCS3B01G610600	Pectin acetyltransferase	829,037,444	0.6	1.2	1.0	0.2
TraesCS3B01G610700	B3 domain-containing protein	829,054,314	.	.	.	.
TraesCS3B01G610800	Histone H2A	829,113,007	-0.1	-0.8	0.9	0.3
TraesCS3B01G610900	RNA-binding protein	829,114,036	0.0	.	1.0	.
TraesCS3B01G611000	.	829,117,291	0.0	-2.3	1.0	0.0
TraesCS3B01G611100	Receptor-like protein kinase	829,203,567	0.2	-2.6	0.9	0.0
TraesCS3B01G611200	DNA topoisomerase 3	829,223,901	.	.	.	.
TraesCS3B01G611300	Histone H2A	829,244,086	.	.	.	.
TraesCS3B01G611400	Outer-membrane lipoprotein LolB	829,246,632	-0.1	4.2	0.9	0.0
TraesCS3B01G611500	transmembrane protein, putative	829,252,291	0.5	-1.9	0.8	0.2
TraesCS3B01G611600	Soluble inorganic pyrophosphatase	829,273,469	-0.1	-0.3	0.9	0.6
TraesCS3B01G611700	Kelch repeat-containing F-box	829,283,466	-0.1	0.6	0.9	0.1
TraesCS3B01G611800	Soluble inorganic pyrophosphatase	829,288,040	1.0	1.1	0.0	0.1
TraesCS3B01G611900	Ubiquitin family protein	829,349,677	.	.	.	.
TraesCS3B01G612000	O-methyltransferase	829,391,763	0.9	2.6	1.0	0.0
TraesCS3B01G612100	Protein	829,508,771	-0.7	-1.0	0.6	0.3
TraesCS3B01G612200	2 MYB transcription factor	829,541,507	0.6	1.5	0.5	0.0
TraesCS3B01G612300	Transcription factor, MADS-box	829,742,374	.	.	.	.
TraesCS3B01G612400	Transcription factor, MADS-box	829,817,279	0.1	.	1.0	.
TraesCS3B01G612500	Transcription factor, MADS-box	829,935,275	.	.	.	.
TraesCS3B01G612600	Transcription factor, MADS-box	830,055,777	.	.	.	.
TraesCS3B01G612700	Transcription factor, MADS-box	830,112,178	.	.	.	.
TraesCS3B01G612800	Transcription factor, MADS-box	830,320,830	.	0.7	.	1.0
TraesCS3B01G612900	Transcription factor, MADS-box	830,610,856	.	.	.	.

Table S43.

Copy number variation analysis of the gene *TraesCS3B01G608800* in 96 diverse hexaploid cultivars

Sample	Solidness	<i>TraesCS3B01G608800</i>		<i>TraesCS3B01G608800</i> SSR - Band Sizes (bp)						
		Copy Number (qPCR)	SEM	396	408	410	416	418	420	422
Choteau	4.3	9.9	0.5			410		418		422
Fortuna	3.0	10.4	0.2			410		418		422
Lancer	3.0	10.3	0.4			410		418		422
AAC Bailey	2.5	9.6	0.9			410		418		422
AC Eatonia	2.5	8.5	0.5			410		418		422
Frontana	2.5	2.3	0.3				416	418		
G9608B1-L12J11BF02	2.5	6.0	0.2							
Janz	2.5	1.4	0.1	396			416			
Leader	2.5	3.7	0.2					418		
Lillian	2.5	10.1	0.7			410		418		422
LJP1091P	2.5	7.7	0.4	396	408			418		422
Mott	2.5	10.3	1.1	396		410		418		422
Rescue	2.5	6.2	0.1			410		418		422
S615	2.5	10.8	0.3			410		418		422
AC Abbey	1.9	6.4	0.4			410		418		422
CDC Landmark	1.9	7.7	0.4			409		418		422
Glencross	1.8	5.9	0.4						420	422
McKenzie	1.8	8.9	0.4							
CDC Rama	1.7	3.5	0.2						420	422
Unity	1.7	11.1	0.7			410		418		422
Glenlea	1.5	3.8	0.2						420	422
CDC Teal	1.3	2.7	0.1					418		
AC Crystal	1.2	2.9	0.2							422
AC Vista	1.2	3.1	0.2							422
Alvena	1.2	4.4	0.2							422
Burnside	1.2	6.0	0.2						420	422
AC Andrew	1.1	2.4	0.1				416	418		
AC Splendor	1.1	3.3	0.1					418		
AC Taber	1.1	2.6	0.1							422
CDC Bison	1.1	5.1	0.2						420	422
CDC Stanley	1.1	1.8	0.1							
CDC Walrus	1.1	2.4	0.0				416	418		
Kane	1.1	4.6	0.2					418		
Katepwa	1.1	2.8	0.1					418		
Laser	1.1	4.1	0.1					418	420	
5500HR	1.0	2.8	0.1						420	
5600HR	1.0	3.0	0.4					418		
5601HR	1.0	3.0	0.1					418		
5602HR	1.0	3.7	0.1				416	418	420	
5603HR	1.0	3.4	0.2				416		420	
5700PR	1.0	2.4	0.2							422
5701PR	1.0	2.7	0.1					418		
5702PR	1.0	2.1	0.2				416			
AC Barrie	1.0	2.9	0.1							
AC Cadillac	1.0	3.5	0.3							
AC Domain	1.0	3.1	0.2					418		
AC Elsa	1.0	3.0	0.2							422
AC Foremost	1.0	2.5	0.1							422
AC Intrepid	1.0	3.2	0.2							

AC Karma	1.0	2.5	0.2			422
Alikat	1.0	2.2	0.2		418	
Carberry	1.0	3.2	0.1		418	
CDC Abound	1.0	3.9	0.3		418	
CDC Alsask	1.0	2.8	0.2		418	
CDC Bounty	1.0	3.0	0.2		418	
CDC Go	1.0	3.1	0.1		418	
CDC Imagine	1.0	4.1	0.5			
CDC Kernen	1.0	2.1	0.2		418	
CDC Merlin	1.0	2.8	0.1		418	
CDC Osler	1.0	3.0	0.2		418	
CDC Thrive	1.0	2.5	0.1		418	
CDC Utmost	1.0	2.0	0.3		418	420
Chinese Spring	1.0	1.4	0.0			
Cutler	1.0	4.1	0.3			422
Glenn	1.0	4.0	0.4	416	418	
Goodeve VB	1.0	4.2	0.2		418	
GP069	1.0	2.7	0.1			422
Harvest	1.0	4.6	1.0			
Helios	1.0	4.8	0.3		418	
Infinity	1.0	3.8	0.2	416	418	
Journey	1.0	3.0	0.1		418	
Laura	1.0	3.2	0.2		418	
Lovitt	1.0	3.2	0.1		418	
Minnedosa	1.0	3.4	0.3			422
Muchmore	1.0	3.8	0.3		418	
Neepawa	1.0	4.2	0.3		418	
Park	1.0	2.4	0.2			420
Peace	1.0	4.6	0.6			420
Prodigy	1.0	4.1	0.3			
PT559	1.0	2.7	0.1		418	
Red Fife	1.0	2.6	0.1	416		
RL4137	1.0	0.8	0.1		418	
Roblin	1.0	3.1	0.1		418	
Sadash	1.0	3.2	0.1	416	418	
Selkirk	1.0	3.6	0.3			420
Snowbird	1.0	2.4	0.2		418	
Snowstar	1.0	3.7	0.2		418	
Somerset	1.0	4.5	0.3		418	
Stanley	1.0	3.6	0.3		418	
Stettler	1.0	3.7	0.2		418	
Sumai 3	1.0	3.2	0.1	416		420
Superb	1.0	3.1	0.1		418	
SY985	1.0	2.9	0.2			422
Vesper	1.0	2.9	0.2		418	420
Waskada	1.0	3.3	0.3		418	

Database S1 (separate .xls file)

Metadata of 850 RNAseq samples used in the study

Database S2 (separate .xls file)

SlimGO ubiquitous and Tissue-exclusive genes.

Database S3 (separate .xls file)

GO terms of the WGCNA850 analysis.

Database S4 (separate .xls file)

WGCNA Module Assignment.

Database S5 (separate .xls file)

Module 8 and 11 TF Arabidopsis and rice orthologs.

References and Notes

1. Food and Agriculture Organization of the United Nations, FAOSTAT statistics database, Food balance sheets (2017); www.fao.org/faostat/en/#data/FBS.
2. Food and Agriculture Organization of the United Nations, FAOSTAT statistics database, Crops (2017); www.fao.org/faostat/en/#data/QC.
3. G. N. Atlin, J. E. Cairns, B. Das, Rapid breeding and varietal replacement are critical to adaptation of cropping systems in the developing world to climate change. *Glob. Food Sec.* **12**, 31–37 (2017). [doi:10.1016/j.gfs.2017.01.008](https://doi.org/10.1016/j.gfs.2017.01.008) [Medline](#)
4. J. M. Hickey, T. Chiurugwi, I. Mackay, W. Powell; Implementing Genomic Selection in CGIAR Breeding Programs Workshop Participants, Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. *Nat. Genet.* **49**, 1297–1303 (2017). [doi:10.1038/ng.3920](https://doi.org/10.1038/ng.3920) [Medline](#)
5. K. Arumuganathan, E. D. Earle, Nuclear DNA content of some important plant species. *Plant Mol. Biol. Report.* **9**, 208–218 (1991). [doi:10.1007/BF02672069](https://doi.org/10.1007/BF02672069)
6. International Wheat Genome Sequencing Consortium (IWGSC), A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science* **345**, 1251788 (2014). [doi:10.1126/science.1251788](https://doi.org/10.1126/science.1251788) [Medline](#)
7. J. A. Chapman, M. Mascher, A. Buluç, K. Barry, E. Georganas, A. Session, V. Strnadova, J. Jenkins, S. Sehgal, L. Ollker, J. Schmutz, K. A. Yelick, U. Scholz, R. Waugh, J. A. Poland, G. J. Muehlbauer, N. Stein, D. S. Rokhsar, A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome. *Genome Biol.* **16**, 26 (2015). [doi:10.1186/s13059-015-0582-8](https://doi.org/10.1186/s13059-015-0582-8) [Medline](#)
8. B. J. Clavijo, L. Venturini, C. Schudoma, G. G. Accinelli, G. Kaithakottil, J. Wright, P. Borrill, G. Kettleborough, D. Heavens, H. Chapman, J. Lipscombe, T. Barker, F.-H. Lu, N. McKenzie, D. Raats, R. H. Ramirez-Gonzalez, A. Coince, N. Peel, L. Percival-Alwyn, O. Duncan, J. Trösch, G. Yu, D. M. Bolser, G. Namaati, A. Kerhornou, M. Spannagl, H. Gundlach, G. Haberer, R. P. Davey, C. Fosker, F. D. Palma, A. L. Phillips, A. H. Millar, P. J. Kersey, C. Uauy, K. V. Krasileva, D. Swarbreck, M. W. Bevan, M. D. Clark, An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. *Genome Res.* **27**, 885–896 (2017). [doi:10.1101/gr.217117.116](https://doi.org/10.1101/gr.217117.116) [Medline](#)
9. A. V. Zimin, D. Puiu, R. Hall, S. Kingan, B. J. Clavijo, S. L. Salzberg, The first near-complete assembly of the hexaploid bread wheat genome, *Triticum aestivum*. *Gigascience* **6**, 1–7 (2017). [doi:10.1093/gigascience/gix097](https://doi.org/10.1093/gigascience/gix097) [Medline](#)
10. T. R. Endo, B. S. Gill, The deletion stocks of common wheat. *J. Hered.* **87**, 295–307 (1996). [doi:10.1093/oxfordjournals.jhered.a023003](https://doi.org/10.1093/oxfordjournals.jhered.a023003)
11. M. E. Sorrells, M. La Rota, C. E. Bermudez-Kandianis, R. A. Greene, R. Kantety, J. D. Munkvold, Miftahudin, A. Mahmoud, X. Ma, P. J. Gustafson, L. L. Qi, B. Echallier, B. S. Gill, D. E. Matthews, G. R. Lazo, S. Chao, O. D. Anderson, H. Edwards, A. M. Linkiewicz, J. Dubcovsky, E. D. Akhunov, J. Dvorak, D. Zhang, H. T. Nguyen, J. Peng, N. L. Lapitan, J. L. Gonzalez-Hernandez, J. A. Anderson, K. Hossain, V. Kalavacharla, S. F. Kianian, D. W. Choi, T. J. Close, M. Dilbirligi, K. S. Gill, C. Steber, M. K. Walker-

- Simmons, P. E. McGuire, C. O. Qualset, Comparative DNA sequence analysis of wheat and rice genomes. *Genome Res.* **13**, 1818–1827 (2003). [Medline](#)
12. K. Eversole, J. Rogers, B. Keller, R. Appels, C. Feuillet, in *Breeding, Quality Traits, Pests and Diseases*, vol. 1 of *Achieving Sustainable Cultivation of Wheat*, P. Langridge, Ed. (Burleigh-Dodds Science Publishing, 2017), chap. 2.
 13. E. Paux, S. Faure, F. Choulet, D. Roger, V. Gauthier, J.-P. Martinant, P. Sourdille, F. Balfourier, M.-C. Le Paslier, A. Chauveau, M. Cakir, B. Gandon, C. Feuillet, Insertion site-based polymorphism markers open new perspectives for genome saturation and marker-assisted selection in wheat. *Plant Biotechnol. J.* **8**, 196–210 (2010). [doi:10.1111/j.1467-7652.2009.00477.x](https://doi.org/10.1111/j.1467-7652.2009.00477.x) [Medline](#)
 14. See supplementary materials.
 15. Arabidopsis Genome Initiative, Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000). [doi:10.1038/35048692](https://doi.org/10.1038/35048692) [Medline](#)
 16. International Rice Genome Sequencing Project, The map-based sequence of the rice genome. *Nature* **436**, 793–800 (2005). [doi:10.1038/nature03895](https://doi.org/10.1038/nature03895) [Medline](#)
 17. T. Wicker, H. Gundlach, M. Spannagl, C. Uauy, P. Borrill, R. H. Ramirez-Gonzalez, R. De Oliveira, K. F. X. Mayer, E. Paux, F. Choulet; International Wheat Genome Sequencing Consortium, Impact of transposable elements on genome structure and evolution in wheat. *Genome Biol.* 10.1186/s13059-018-1479-0 (2018).
 18. F. Choulet, A. Alberti, S. Theil, N. Glover, V. Barbe, J. Daron, L. Pingault, P. Sourdille, A. Couloux, E. Paux, P. Leroy, S. Mangenot, N. Guilhot, J. Le Gouis, F. Balfourier, M. Alaux, V. Jamilloux, J. Poulain, C. Durand, A. Bellec, C. Gaspin, J. Safar, J. Dolezel, J. Rogers, K. Vandepoele, J. M. Aury, K. Mayer, H. Berges, H. Quesneville, P. Wincker, C. Feuillet, Structural and functional partitioning of bread wheat chromosome 3B. *Science* **345**, 1249721 (2014). [doi:10.1126/science.1249721](https://doi.org/10.1126/science.1249721) [Medline](#)
 19. X. Guo, H. Su, Q. Shi, S. Fu, J. Wang, X. Zhang, Z. Hu, F. Han, De novo centromere formation and centromeric sequence expansion in wheat and its wide hybrids. *PLOS Genet.* **12**, e1005997 (2016). [doi:10.1371/journal.pgen.1005997](https://doi.org/10.1371/journal.pgen.1005997) [Medline](#)
 20. K. Wang, Y. Wu, W. Zhang, R. K. Dawe, J. Jiang, Maize centromeres expand and adopt a uniform size in the genetic background of oat. *Genome Res.* **24**, 107–116 (2014). [doi:10.1101/gr.160887.113](https://doi.org/10.1101/gr.160887.113) [Medline](#)
 21. Y. Jiao, P. Peluso, J. Shi, T. Liang, M. C. Stitzer, B. Wang, M. S. Campbell, J. C. Stein, X. Wei, C. S. Chin, K. Guill, M. Regulski, S. Kumari, A. Olson, J. Gent, K. L. Schneider, T. K. Wolfgruber, M. R. May, N. M. Springer, E. Antoniou, W. R. McCombie, G. G. Presting, M. McMullen, J. Ross-Ibarra, R. K. Dawe, A. Hastie, D. R. Rank, D. Ware, Improved maize reference genome with single-molecule technologies. *Nature* **546**, 524–527 (2017). [Medline](#)
 22. H. Yan, P. B. Talbert, H.-R. Lee, J. Jett, S. Henikoff, F. Chen, J. Jiang, Intergenic locations of rice centromeric chromatin. *PLOS Biol.* **6**, e286 (2008). [doi:10.1371/journal.pbio.0060286](https://doi.org/10.1371/journal.pbio.0060286) [Medline](#)
 23. R. H. Ramírez-González, P. Borrill, D. Lang, S. A. Harrington, J. Brinton, L. Venturini, M. Davey, J. Jacobs, F. van Ex, A. Pasha, Y. Khedikar, S. J. Robinson, A. T. Cory, T. Florio,

- L. Concia, C. Juery, H. Schoonbeek, B. Steuernagel, D. Xiang, C. J. Ridout, B. Chalhoub, K. F. X. Mayer, M. Benhamed, D. Latrasse, A. Bendahmane, B. B. H. Wulff, R. Appels, V. Tiwari, R. Datla, F. Choulet, C. J. Pozniak, N. J. Provart, A. G. Sharpe, E. Paux, M. Spannagl, A. Bräutigam, C. Uauy; International Wheat Genome Sequencing Consortium, The transcriptional landscape of polyploid wheat. *Science* **361**, eaar6089 (2018). [doi:10.1126/science.aar6089](https://doi.org/10.1126/science.aar6089)
24. F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, E. M. Zdobnov, BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015). [doi:10.1093/bioinformatics/btv351](https://doi.org/10.1093/bioinformatics/btv351) [Medline](#)
 25. M.-D. Rey, A. C. Martín, J. Higgins, D. Swarbreck, C. Uauy, P. Shaw, G. Moore, Exploiting the *ZIP4* homologue within the wheat *Ph1* locus has identified two lines exhibiting homoeologous crossover in wheat-wild relative hybrids. *Mol. Breed.* **37**, 95 (2017). [doi:10.1007/s11032-017-0700-2](https://doi.org/10.1007/s11032-017-0700-2) [Medline](#)
 26. F. Cheng, J. Wu, X. Cai, J. Liang, M. Freeling, X. Wang, Gene retention, fractionation and subgenome differences in polyploid plants. *Nat. Plants* **4**, 258–268 (2018). [doi:10.1038/s41477-018-0136-7](https://doi.org/10.1038/s41477-018-0136-7) [Medline](#)
 27. T. Marcussen, S. R. Sandve, L. Heier, M. Spannagl, M. Pfeifer, K. S. Jakobsen, B. B. Wulff, B. Steuernagel, K. F. Mayer, O. A. Olsen; International Wheat Genome Sequencing Consortium (IWGSC), Ancient hybridizations among the ancestral genomes of bread wheat. *Science* **345**, 1250092 (2014). [Medline](#)
 28. Y. Van de Peer, S. Maere, A. Meyer, The evolutionary significance of ancient genome duplications. *Nat. Rev. Genet.* **10**, 725–732 (2009). [doi:10.1038/nrg2600](https://doi.org/10.1038/nrg2600) [Medline](#)
 29. P. S. Soltis, D. E. Soltis, Ancient WGD events as drivers of key innovations in angiosperms. *Curr. Opin. Plant Biol.* **30**, 159–165 (2016). [doi:10.1016/j.pbi.2016.03.015](https://doi.org/10.1016/j.pbi.2016.03.015) [Medline](#)
 30. M. Melé, P. G. Ferreira, F. Reverter, D. S. DeLuca, J. Monlong, M. Sammeth, T. R. Young, J. M. Goldmann, D. D. Pervouchine, T. J. Sullivan, R. Johnson, A. V. Segrè, S. Djebali, A. Niarchou, F. A. Wright, T. Lappalainen, M. Calvo, G. Getz, E. T. Dermitzakis, K. G. Ardlie, R. Guigó; GTEx Consortium, The human transcriptome across tissues and individuals. *Science* **348**, 660–665 (2015). [doi:10.1126/science.aaa0355](https://doi.org/10.1126/science.aaa0355) [Medline](#)
 31. S. C. Stelpflug, R. S. Sekhon, B. Vaillancourt, C. N. Hirsch, C. R. Buell, N. de Leon, S. M. Kaeppler, An expanded maize gene expression atlas based on RNA sequencing and its use to explore root development. *Plant Genome* **9**, 0025 (2016). [Medline](#)
 32. F. He, S. Yoo, D. Wang, S. Kumari, M. Gerstein, D. Ware, S. Maslov, Large-scale atlas of microarray data reveals the distinct expression landscape of different tissues in Arabidopsis. *Plant J.* **86**, 472–480 (2016). [doi:10.1111/tpj.13175](https://doi.org/10.1111/tpj.13175) [Medline](#)
 33. X. Wang, U. Gowik, H. Tang, J. E. Bowers, P. Westhoff, A. H. Paterson, Comparative genomic analysis of C4 photosynthetic pathway evolution in grasses. *Genome Biol.* **10**, R68 (2009). [doi:10.1186/gb-2009-10-6-r68](https://doi.org/10.1186/gb-2009-10-6-r68) [Medline](#)
 34. L. Xia, D. Zou, J. Sang, X. Xu, H. Yin, M. Li, S. Wu, S. Hu, L. Hao, Z. Zhang, Rice Expression Database (RED): An integrated RNA-Seq-derived gene expression database for rice. *J. Genet. Genomics* **44**, 235–241 (2017). [doi:10.1016/j.jgg.2017.05.003](https://doi.org/10.1016/j.jgg.2017.05.003) [Medline](#)

35. R. J. Schaefer, J.-M. Michno, C. L. Myers, Unraveling gene function in agricultural species using gene co-expression networks. *Biochim. Biophys. Acta* **1860**, 53–63 (2017). [Medline](#)
36. J. Beales, A. Turner, S. Griffiths, J. W. Snape, D. A. Laurie, A *Pseudo-Response Regulator* is misexpressed in the photoperiod insensitive *Ppd-D1a* mutant of wheat (*Triticum aestivum* L.). *Theor. Appl. Genet.* **115**, 721–733 (2007). [doi:10.1007/s00122-007-0603-4](#) [Medline](#)
37. L. Yan, D. Fu, C. Li, A. Blechl, G. Tranquilli, M. Bonafede, A. Sanchez, M. Valarik, S. Yasuda, J. Dubcovsky, The wheat and barley vernalization gene *VRN3* is an orthologue of *FT*. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 19581–19586 (2006). [doi:10.1073/pnas.0607142103](#) [Medline](#)
38. K. V. Krasileva, H. A. Vasquez-Gross, T. Howell, P. Bailey, F. Paraiso, L. Clissold, J. Simmonds, R. H. Ramirez-Gonzalez, X. Wang, P. Borrill, C. Fosker, S. Ayling, A. L. Phillips, C. Uauy, J. Dubcovsky, Uncovering hidden variation in polyploid wheat. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E913–E921 (2017). [doi:10.1073/pnas.1619268114](#) [Medline](#)
39. S. Wang, J. Lu, X.-F. Song, S.-C. Ren, C. You, J. Xu, C.-M. Liu, H. Ma, F. Chang, Cytological and transcriptomic analyses reveal important roles of *CLE19* in pollen exine formation. *Plant Physiol.* **175**, 1186–1202 (2017). [doi:10.1104/pp.17.00439](#) [Medline](#)
40. M. Pfeifer, K. G. Kugler, S. R. Sandve, B. Zhan, H. Rudi, T. R. Hvidsten, K. F. Mayer, O. A. Olsen; International Wheat Genome Sequencing Consortium, Genome interplay in the grain transcriptome of hexaploid bread wheat. *Science* **345**, 1250091 (2014). [doi:10.1126/science.1250091](#) [Medline](#)
41. P. Borrill, N. Adamski, C. Uauy, Genomics as the key to unlocking the polyploid potential of wheat. *New Phytol.* **208**, 1008–1022 (2015). [doi:10.1111/nph.13533](#) [Medline](#)
42. F. A. Kondrashov, Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proc. Biol. Sci.* **279**, 5048–5057 (2012). [doi:10.1098/rspb.2012.1108](#) [Medline](#)
43. P. H. Schiffer, J. Gravemeyer, M. Rauscher, T. Wiehe, Ultra large gene families: A matter of adaptation or genomic parasites? *Life* **6**, 32 (2016). [doi:10.3390/life6030032](#) [Medline](#)
44. T. Sykes, S. Yates, I. Nagy, T. Asp, I. Small, B. Studer, In silico identification of candidate genes for fertility restoration in cytoplasmic male sterile perennial ryegrass (*Lolium perenne* L.). *Genome Biol. Evol.* **9**, 351–362 (2017). [Medline](#)
45. J. Melonek, J. D. Stone, I. Small, Evolutionary plasticity of restorer-of-fertility-like proteins in rice. *Sci. Rep.* **6**, 35152 (2016). [doi:10.1038/srep35152](#) [Medline](#)
46. T. Würschum, C. F. H. Longin, V. Hahn, M. R. Tucker, W. L. Leiser, Copy number variations of *CBF* genes at the *Fr-A2* locus are essential components of winter hardiness in wheat. *Plant J.* **89**, 764–773 (2017). [doi:10.1111/tpj.13424](#) [Medline](#)
47. A. Juhász, T. Belova, C. G. Florides, C. Maulis, I. Fischer, Gy. Gell, Zs. Birinyi, J. Ong, G. Keeble-Gagnère, A. Maharajan, W. Ma, P. Gibson, J. Jia, D. Lang, K. F. X. Mayer, M. Spannagl, J. A. Tye-Din, R. Appels, O.-A. Olsen, Genome mapping of seed-borne allergens and immune-responsive proteins in wheat. *Sci. Adv.* **4**, eaar8602 (2018). [doi:10.1126/sciadv.aar8602](#)
48. M. Feldman, A. A. Levy, in *Alien Introgression in Wheat: Cytogenetics, Molecular Biology, and Genomics*, M. Molnár-Láng, C. Ceoloni, J. Doležel, Eds. (Springer, 2015), pp. 21–76.

49. K. T. Nilsen, A. N'Diaye, P. R. MacLachlan, J. M. Clarke, Y. Ruan, R. D. Cuthbert, R. E. Knox, K. Wiebe, A. T. Cory, S. Walkowiak, B. L. Beres, R. J. Graf, F. R. Clarke, A. G. Sharpe, A. Distelfeld, C. J. Pozniak, High density mapping and haplotype analysis of the major stem-solidness locus *SSt1* in durum and common wheat. *PLOS ONE* **12**, e0175285 (2017). [doi:10.1371/journal.pone.0175285](https://doi.org/10.1371/journal.pone.0175285) [Medline](#)
50. N. Sharma, P. Ruelens, M. D'hauw, T. Maggen, N. Dochy, S. Torfs, K. Kaufmann, A. Rohde, K. Geuten, A flowering locus C homolog is a vernalization-regulated repressor in *Brachypodium* and is cold regulated in wheat. *Plant Physiol.* **173**, 1301–1315 (2017). [doi:10.1104/pp.16.01161](https://doi.org/10.1104/pp.16.01161) [Medline](#)
51. H. Puchta, Applying CRISPR/Cas for genome engineering in plants: The best is yet to come. *Curr. Opin. Plant Biol.* **36**, 1–8 (2017). [doi:10.1016/j.pbi.2016.11.011](https://doi.org/10.1016/j.pbi.2016.11.011) [Medline](#)
52. International Wheat Genome Sequencing Consortium, Gene family expansion and contraction in the genome of bread wheat cv. Chinese Spring. e!DAL (2018). [doi:10.5447/IPK/2018/5](https://doi.org/10.5447/IPK/2018/5)
53. H.-Q. Ling, S. Zhao, D. Liu, J. Wang, H. Sun, C. Zhang, H. Fan, D. Li, L. Dong, Y. Tao, C. Gao, H. Wu, Y. Li, Y. Cui, X. Guo, S. Zheng, B. Wang, K. Yu, Q. Liang, W. Yang, X. Lou, J. Chen, M. Feng, J. Jian, X. Zhang, G. Luo, Y. Jiang, J. Liu, Z. Wang, Y. Sha, B. Zhang, H. Wu, D. Tang, Q. Shen, P. Xue, S. Zou, X. Wang, X. Liu, F. Wang, Y. Yang, X. An, Z. Dong, K. Zhang, X. Zhang, M.-C. Luo, J. Dvorak, Y. Tong, J. Wang, H. Yang, Z. Li, D. Wang, A. Zhang, J. Wang, Draft genome of the wheat A-genome progenitor *Triticum urartu*. *Nature* **496**, 87–90 (2013). [doi:10.1038/nature11997](https://doi.org/10.1038/nature11997) [Medline](#)
54. J. Jia, S. Zhao, X. Kong, Y. Li, G. Zhao, W. He, R. Appels, M. Pfeifer, Y. Tao, X. Zhang, R. Jing, C. Zhang, Y. Ma, L. Gao, C. Gao, M. Spannagl, K. F. X. Mayer, D. Li, S. Pan, F. Zheng, Q. Hu, X. Xia, J. Li, Q. Liang, J. Chen, T. Wicker, C. Gou, H. Kuang, G. He, Y. Luo, B. Keller, Q. Xia, P. Lu, J. Wang, H. Zou, R. Zhang, J. Xu, J. Gao, C. Middleton, Z. Quan, G. Liu, J. Wang, H. Yang, X. Liu, Z. He, L. Mao, J. Wang; International Wheat Genome Sequencing Consortium, *Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature* **496**, 91–95 (2013). [doi:10.1038/nature12028](https://doi.org/10.1038/nature12028) [Medline](#)
55. Y. Ishida, M. Tsunashima, Y. Hiei, T. Komari, in *Agrobacterium Protocols: Volume 1*, K. Wang, Ed. (Springer, 2015), pp. 189–198.
56. M. Alaux, J. Rogers, T. Letellier, R. Flores, F. Alfama, C. Pommier, N. Mohellibi, S. Durand, E. Kimmel, C. Michotey, C. Guerche, M. Loaec, M. Lainé, D. Steinbach, F. Choulet, H. Rimbart, P. Leroy, N. Guilhot, J. Salse, C. Feuillet, E. Paux, K. Eversole, A.-F. Adam-Blondon, H. Quesneville; International Wheat Genome Sequencing Consortium, Linking the International Wheat Genome Sequencing Consortium bread wheat reference genome sequence to wheat genetic and phenomic data. *Genome Biol.* **19**, 10.1186/s13059-018-1491-4 (2018).
57. R. Avni, M. Nave, O. Barad, K. Baruch, S. O. Twardziok, H. Gundlach, I. Hale, M. Mascher, M. Spannagl, K. Wiebe, K. W. Jordan, G. Golan, J. Deek, B. Ben-Zvi, G. Ben-Zvi, A. Himmelbach, R. P. MacLachlan, A. G. Sharpe, A. Fritz, R. Ben-David, H. Budak, T. Fahima, A. Korol, J. D. Faris, A. Hernandez, M. A. Mikel, A. A. Levy, B. Steffenson, M. Maccaferri, R. Tuberosa, L. Cattivelli, P. Faccioli, A. Ceriotti, K. Kashkush, M. Pourkheirandish, T. Komatsuda, T. Eilam, H. Sela, A. Sharon, N. Ohad, D. A.

- Chamovitz, K. F. X. Mayer, N. Stein, G. Ronen, Z. Peleg, C. J. Pozniak, E. D. Akhunov, A. Distelfeld, Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science* **357**, 93–97 (2017). [doi:10.1126/science.aan0032](https://doi.org/10.1126/science.aan0032) [Medline](#)
58. F. Choulet, T. Wicker, C. Rustenholz, E. Paux, J. Salse, P. Leroy, S. Schlub, M.-C. Le Paslier, G. Magdelenat, C. Gonthier, A. Couloux, H. Budak, J. Breen, M. Pumphrey, S. Liu, X. Kong, J. Jia, M. Gut, D. Brunel, J. A. Anderson, B. S. Gill, R. Appels, B. Keller, C. Feuillet, Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. *Plant Cell* **22**, 1686–1701 (2010). [doi:10.1105/tpc.110.074187](https://doi.org/10.1105/tpc.110.074187) [Medline](#)
59. G. Keeble-Gagnère, P. Rigault, J. Tibbits, R. Pasam, M. Hayden, K. Forrest, Z. Frenkel, A. Korol, E. Huang, C. Cavanagh, J. Taylor, M. Abrouk, A. Sharpe, D. Konkin, P. Sourdille, B. Darrier, F. Choulet, A. Bernard, S. Rochfort, A. Dimech, N. Watson-Haigh, U. Baumann, D. Fleury, P. Eckermann, A. Juhasz, S. Boisvert, M.-A. Nolin, J. Doležal, H. Šimková, H. Toegelová, J. Šafář, M.-C. Luo, F. Camara, M. Pfeifer, D. Isdale, J. Nyström-Persson, D.-H. Koo, M. Tinning, D. Cui, Z. Ru, R. Appels; International Wheat Genome Sequencing Consortium, Optical and physical mapping with local finishing enables megabase-scale resolution of agronomically important regions on wheat chromosome 7A. *Genome Biol.* 10.1186/s13059-018-1475-4 (2018).
60. R. Kalhor, H. Tjong, N. Jayathilaka, F. Alber, L. Chen, Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat. Biotechnol.* **30**, 90–98 (2012). [doi:10.1038/nbt.2057](https://doi.org/10.1038/nbt.2057) [Medline](#)
61. E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, J. Dekker, Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009). [doi:10.1126/science.1181369](https://doi.org/10.1126/science.1181369) [Medline](#)
62. S. Beier, A. Himmelbach, C. Colmsee, X.-Q. Zhang, R. A. Barrero, Q. Zhang, L. Li, M. Bayer, D. Bolser, S. Taudien, M. Groth, M. Felder, A. Hastie, H. Šimková, H. Staňková, J. Vrána, S. Chan, M. Muñoz-Amatriaín, R. Ounit, S. Wanamaker, T. Schmutzer, L. Aliyeva-Schnorr, S. Grasso, J. Tanskanen, D. Sampath, D. Heavens, S. Cao, B. Chapman, F. Dai, Y. Han, H. Li, X. Li, C. Lin, J. K. McCooke, C. Tan, S. Wang, S. Yin, G. Zhou, J. A. Poland, M. I. Bellgard, A. Houben, J. Doležal, S. Ayling, S. Lonardi, P. Langridge, G. J. Muehlbauer, P. Kersey, M. D. Clark, M. Caccamo, A. H. Schulman, M. Platzer, T. J. Close, M. Hansson, G. Zhang, I. Braumann, C. Li, R. Waugh, U. Scholz, N. Stein, M. Mascher, Construction of a map-based reference genome sequence for barley, *Hordeum vulgare* L. *Sci. Data* **4**, 170044 (2017). [doi:10.1038/sdata.2017.44](https://doi.org/10.1038/sdata.2017.44) [Medline](#)
63. J. Šafář, J. Bartoš, J. Janda, A. Bellec, M. Kubaláková, M. Valárik, S. Pateyron, J. Weiserová, R. Tusková, J. Cíhalíková, J. Vrána, H. Šimková, P. Faivre-Rampant, P. Sourdille, M. Caboche, M. Bernard, J. Doležal, B. Chalhoub, Dissecting large and complex genomes: Flow sorting and BAC cloning of individual chromosomes from bread wheat. *Plant J.* **39**, 960–968 (2004). [doi:10.1111/j.1365-3113.2004.02179.x](https://doi.org/10.1111/j.1365-3113.2004.02179.x) [Medline](#)
64. J. Šafář, H. Šimková, M. Kubaláková, J. Cíhalíková, P. Suchánková, J. Bartoš, J. Doležal, Development of chromosome-specific BAC resources for genomics of bread wheat. *Cytogenet. Genome Res.* **129**, 211–223 (2010). [doi:10.1159/000313072](https://doi.org/10.1159/000313072) [Medline](#)

65. M.-C. Luo, C. Thomas, F. M. You, J. Hsiao, S. Ouyang, C. R. Buell, M. Malandro, P. E. McGuire, O. D. Anderson, J. Dvorak, High-throughput fingerprinting of bacterial artificial chromosomes using the snapshot labeling kit and sizing of restriction fragments by capillary electrophoresis. *Genomics* **82**, 378–389 (2003). [doi:10.1016/S0888-7543\(03\)00128-9](https://doi.org/10.1016/S0888-7543(03)00128-9) [Medline](#)
66. J. van Oeveren, M. de Ruiter, T. Jesse, H. van der Poel, J. Tang, F. Yalcin, A. Janssen, H. Volpin, K. E. Stormo, R. Bogden, M. J. T. van Eijk, M. Prins, Sequence-based physical mapping of complex genomes by whole genome profiling. *Genome Res.* **21**, 618–625 (2011). [doi:10.1101/gr.112094.110](https://doi.org/10.1101/gr.112094.110) [Medline](#)
67. C. Soderlund, S. Humphray, A. Dunham, L. French, Contigs built with fingerprints, markers, and FPC V4.7. *Genome Res.* **10**, 1772–1787 (2000). [Medline](#)
68. Z. Frenkel, E. Paux, D. Mester, C. Feuillet, A. Korol, LTC: A novel algorithm to improve the efficiency of contig assembly for physical mapping in complex genomes. *BMC Bioinformatics* **11**, 584 (2010). [doi:10.1186/1471-2105-11-584](https://doi.org/10.1186/1471-2105-11-584) [Medline](#)
69. H. Staňková, A. R. Hastie, S. Chan, J. Vrána, Z. Tulpová, M. Kubaláková, P. Visendi, S. Hayashi, M. Luo, J. Batley, D. Edwards, J. Doležel, H. Šimková, BioNano genome mapping of individual chromosomes supports physical mapping and sequence assembly in complex plant genomes. *Plant Biotechnol. J.* **14**, 1523–1531 (2016). [doi:10.1111/pbi.12513](https://doi.org/10.1111/pbi.12513) [Medline](#)
70. N. Poursarebani, T. Nussbaumer, H. Šimková, J. Safář, H. Witsenboer, J. van Oeveren, J. Doležel, K. F. X. Mayer, N. Stein, T. Schnurbusch, Whole-genome profiling and shotgun sequencing delivers an anchored, gene-decorated, physical map assembly of bread wheat chromosome 6A. *Plant J.* **79**, 334–347 (2014). [doi:10.1111/tbj.12550](https://doi.org/10.1111/tbj.12550) [Medline](#)
71. F. Kobayashi, J. Wu, H. Kanamori, T. Tanaka, S. Katagiri, W. Karasawa, S. Kaneko, S. Watanabe, T. Sakaguchi, Y. Hanawa, H. Fujisawa, K. Kurita, C. Abe, J. C. M. Iehisa, R. Ohno, J. Šafář, H. Šimková, Y. Mukai, M. Hamada, M. Saito, G. Ishikawa, Y. Katayose, T. R. Endo, S. Takumi, T. Nakamura, K. Sato, Y. Ogihara, K. Hayakawa, J. Doležel, S. Nasuda, T. Matsumoto, H. Handa, A high-resolution physical map integrating an anchored chromosome with the BAC physical maps of wheat chromosome 6B. *BMC Genomics* **16**, 595 (2015). [doi:10.1186/s12864-015-1803-y](https://doi.org/10.1186/s12864-015-1803-y) [Medline](#)
72. M. Kubaláková, J. Vrána, J. Cíhalíková, H. Šimková, J. Doležel, Flow karyotyping and chromosome sorting in bread wheat (*Triticum aestivum* L.). *Theor. Appl. Genet.* **104**, 1362–1372 (2002). [doi:10.1007/s00122-002-0888-2](https://doi.org/10.1007/s00122-002-0888-2) [Medline](#)
73. V. K. Tiwari, A. Heesacker, O. Riera-Lizarazu, H. Gunn, S. Wang, Y. Wang, Y. Q. Gu, E. Paux, D.-H. Koo, A. Kumar, M.-C. Luo, G. Lazo, R. Zemetra, E. Akhunov, B. Friebe, J. Poland, B. S. Gill, S. Kianian, J. M. Leonard, A whole-genome, radiation hybrid mapping resource of hexaploid wheat. *Plant J.* **86**, 195–207 (2016). [doi:10.1111/tbj.13153](https://doi.org/10.1111/tbj.13153) [Medline](#)
74. H. Rimbart, B. Darrier, J. Navarro, J. Kitt, F. Choulet, M. Leveugle, J. Duarte, N. Rivière, K. Eversole, J. Le Gouis, A. Davassi, F. Balfourier, M.-C. Le Paslier, A. Berard, D. Brunel, C. Feuillet, C. Poncet, P. Sourdille, E. Paux; International Wheat Genome Sequencing Consortium, High throughput SNP discovery and genotyping in hexaploid wheat. *PLOS ONE* **13**, e0186329 (2018). [doi:10.1371/journal.pone.0186329](https://doi.org/10.1371/journal.pone.0186329) [Medline](#)

75. S. de Givry, M. Bouchez, P. Chabrier, D. Milan, T. Schiex, CARHTA GENE: Multipopulation integrated genetic and radiation hybrid mapping. *Bioinformatics* **21**, 1703–1704 (2005). [doi:10.1093/bioinformatics/bti222](https://doi.org/10.1093/bioinformatics/bti222) [Medline](#)
76. M. E. Sorrells, J. P. Gustafson, D. Somers, S. Chao, D. Benscher, G. Guedira-Brown, E. Huttner, A. Kilian, P. E. McGuire, K. Ross, J. Tanaka, P. Wenzl, K. Williams, C. O. Qualset, Reconstruction of the synthetic W7984 x Opata M85 wheat reference population. *Genome* **54**, 875–882 (2011). [doi:10.1139/g11-054](https://doi.org/10.1139/g11-054) [Medline](#)
77. J. A. Poland, P. J. Brown, M. E. Sorrells, J.-L. Jannink, Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLOS ONE* **7**, e32253 (2012). [doi:10.1371/journal.pone.0032253](https://doi.org/10.1371/journal.pone.0032253) [Medline](#)
78. Y. Wu, P. R. Bhat, T. J. Close, S. Lonardi, Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. *PLOS Genet.* **4**, e1000212 (2008). [doi:10.1371/journal.pgen.1000212](https://doi.org/10.1371/journal.pgen.1000212) [Medline](#)
79. J. Doležel, M. Kubaláková, E. Paux, J. Bartoš, C. Feuillet, Chromosome-based genomics in the cereals. *Chromosome Res.* **15**, 51–66 (2007). [doi:10.1007/s10577-006-1106-x](https://doi.org/10.1007/s10577-006-1106-x) [Medline](#)
80. A. A. Myburg, D. Grattapaglia, G. A. Tuskan, U. Hellsten, R. D. Hayes, J. Grimwood, J. Jenkins, E. Lindquist, H. Tice, D. Bauer, D. M. Goodstein, I. Dubchak, A. Poliakov, E. Mizrahi, A. R. K. Kullán, S. G. Hussey, D. Pinard, K. van der Merwe, P. Singh, I. van Jaarsveld, O. B. Silva-Junior, R. C. Togawa, M. R. Pappas, D. A. Faria, C. P. Sansaloni, C. D. Petroli, X. Yang, P. Ranjan, T. J. Tschaplinski, C.-Y. Ye, T. Li, L. Sterck, K. Vanneste, F. Murat, M. Soler, H. S. Clemente, N. Saidi, H. Cassan-Wang, C. Dunand, C. A. Hefer, E. Bornberg-Bauer, A. R. Kersting, K. Vining, V. Amarasinghe, M. Ranik, S. Naithani, J. Elser, A. E. Boyd, A. Liston, J. W. Spatafora, P. Dharmawardhana, R. Raja, C. Sullivan, E. Romanel, M. Alves-Ferreira, C. Külheim, W. Foley, V. Carocha, J. Paiva, D. Kudrna, S. H. Brommonschenkel, G. Pasquali, M. Byrne, P. Rigault, J. Tibbits, A. Spokevicius, R. C. Jones, D. A. Steane, R. E. Vaillancourt, B. M. Potts, F. Joubert, K. Barry, G. J. Pappas, S. H. Strauss, P. Jaiswal, J. Grima-Pettenati, J. Salse, Y. Van de Peer, D. S. Rokhsar, J. Schmutz, The genome of *Eucalyptus grandis*. *Nature* **510**, 356–362 (2014). [doi:10.1038/nature13308](https://doi.org/10.1038/nature13308) [Medline](#)
81. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009). [doi:10.1093/bioinformatics/btp324](https://doi.org/10.1093/bioinformatics/btp324) [Medline](#)
82. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin; 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009). [doi:10.1093/bioinformatics/btp352](https://doi.org/10.1093/bioinformatics/btp352) [Medline](#)
83. R. Whitford, D. Fleury, J. C. Reif, M. Garcia, T. Okada, V. Korzun, P. Langridge, Hybrid breeding in wheat: Technologies to improve hybrid wheat seed production. *J. Exp. Bot.* **64**, 5411–5428 (2013). [doi:10.1093/jxb/ert333](https://doi.org/10.1093/jxb/ert333) [Medline](#)
84. J. N. Burton, A. Adey, R. P. Patwardhan, R. Qiu, J. O. Kitzman, J. Shendure, Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125 (2013). [doi:10.1038/nbt.2727](https://doi.org/10.1038/nbt.2727) [Medline](#)

85. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010). [doi:10.1093/bioinformatics/btq033](https://doi.org/10.1093/bioinformatics/btq033) [Medline](#)
86. J. Daron, N. Glover, L. Pingault, S. Theil, V. Jamilloux, E. Paux, V. Barbe, S. Mangenot, A. Alberti, P. Wincker, H. Quesneville, C. Feuillet, F. Choulet, Organization and evolution of transposable elements along the bread wheat chromosome 3B. *Genome Biol.* **15**, 546 (2014). [doi:10.1186/s13059-014-0546-4](https://doi.org/10.1186/s13059-014-0546-4) [Medline](#)
87. S. Kurtz, A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, S. L. Salzberg, Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004). [doi:10.1186/gb-2004-5-2-r12](https://doi.org/10.1186/gb-2004-5-2-r12) [Medline](#)
88. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011). [doi:10.14806/ej.17.1.200](https://doi.org/10.14806/ej.17.1.200)
89. P. Leroy, N. Guilhot, H. Sakai, A. Bernard, F. Choulet, S. Theil, S. Reboux, N. Amano, T. Flutre, C. Pelegriin, H. Ohyanagi, M. Seidel, F. Giacomoni, M. Reichstadt, M. Alaux, E. Gicquello, F. Legeai, L. Cerutti, H. Numa, T. Tanaka, K. Mayer, T. Itoh, H. Quesneville, C. Feuillet, TriAnnot: A versatile and high performance pipeline for the automated annotation of plant genomes. *Front. Plant Sci.* **3**, 5 (2012). [doi:10.3389/fpls.2012.00005](https://doi.org/10.3389/fpls.2012.00005) [Medline](#)
90. A. F. Smit, Identification of a new, abundant superfamily of mammalian LTR-transposons. *Nucleic Acids Res.* **21**, 1863–1872 (1993). [doi:10.1093/nar/21.8.1863](https://doi.org/10.1093/nar/21.8.1863) [Medline](#)
91. S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997). [doi:10.1093/nar/25.17.3389](https://doi.org/10.1093/nar/25.17.3389) [Medline](#)
92. L. Pingault, F. Choulet, A. Alberti, N. Glover, P. Wincker, C. Feuillet, E. Paux, Deep transcriptome sequencing provides new insights into the structural and functional organization of the wheat genome. *Genome Biol.* **16**, 29 (2015). [doi:10.1186/s13059-015-0601-9](https://doi.org/10.1186/s13059-015-0601-9) [Medline](#)
93. L. Dong, H. Liu, J. Zhang, S. Yang, G. Kong, J. S. C. Chu, N. Chen, D. Wang, Single-molecule real-time transcript sequencing facilitates common wheat genome annotation and grain transcriptome research. *BMC Genomics* **16**, 1039 (2015). [doi:10.1186/s12864-015-2257-y](https://doi.org/10.1186/s12864-015-2257-y) [Medline](#)
94. K. F. Mayer, R. Waugh, J. W. Brown, A. Schulman, P. Langridge, M. Platzer, G. B. Fincher, G. J. Muehlbauer, K. Sato, T. J. Close, R. P. Wise, N. Stein; The International Barley Genome Sequencing Consortium, A physical, genetic and functional sequence assembly of the barley genome. *Nature* **491**, 711–716 (2012). [doi:10.1038/nature11543](https://doi.org/10.1038/nature11543) [Medline](#)
95. G. S. C. Slater, E. Birney, Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005). [doi:10.1186/1471-2105-6-31](https://doi.org/10.1186/1471-2105-6-31) [Medline](#)
96. M. Stanke, S. Waack, Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19** (suppl. 2), ii215–ii225 (2003). [Medline](#)
97. N. Amano, T. Tanaka, H. Numa, H. Sakai, T. Itoh, Efficient plant gene identification based on interspecies mapping of full-length cDNAs. *DNA Res.* **17**, 271–279 (2010). [doi:10.1093/dnares/dsq017](https://doi.org/10.1093/dnares/dsq017) [Medline](#)

98. T. D. Wu, C. K. Watanabe, GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005). [doi:10.1093/bioinformatics/bti310](https://doi.org/10.1093/bioinformatics/bti310) [Medline](#)
99. C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, L. Pachter, Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010). [doi:10.1038/nbt.1621](https://doi.org/10.1038/nbt.1621) [Medline](#)
100. D. Kim, B. Langmead, S. L. Salzberg, HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015). [doi:10.1038/nmeth.3317](https://doi.org/10.1038/nmeth.3317) [Medline](#)
101. D. W. Barnett, E. K. Garrison, A. R. Quinlan, M. P. Strömberg, G. T. Marth, BamTools: A C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* **27**, 1691–1692 (2011). [doi:10.1093/bioinformatics/btr174](https://doi.org/10.1093/bioinformatics/btr174) [Medline](#)
102. M. Pertea, G. M. Pertea, C. M. Antonescu, T.-C. Chang, J. T. Mendell, S. L. Salzberg, StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015). [doi:10.1038/nbt.3122](https://doi.org/10.1038/nbt.3122) [Medline](#)
103. The UniProt Consortium, UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017). [doi:10.1093/nar/gkw1099](https://doi.org/10.1093/nar/gkw1099) [Medline](#)
104. G. Gremme, V. Brendel, M. E. Sparks, S. Kurtz, Engineering a software tool for gene structure prediction in higher organisms. *Inf. Softw. Technol.* **47**, 965–978 (2005). [doi:10.1016/j.infsof.2005.09.005](https://doi.org/10.1016/j.infsof.2005.09.005)
105. J. Keilwagen, M. Wenk, J. L. Erickson, M. H. Schattat, J. Grau, F. Hartung, Using intron position conservation for homology-based gene prediction. *Nucleic Acids Res.* **44**, e89 (2016). [doi:10.1093/nar/gkw092](https://doi.org/10.1093/nar/gkw092) [Medline](#)
106. K. Mochida, T. Yoshida, T. Sakurai, Y. Ogihara, K. Shinozaki, TriFLDB: A database of clustered full-length coding sequences from Triticeae with applications to comparative grass genomics. *Plant Physiol.* **150**, 1135–1146 (2009). [doi:10.1104/pp.109.138214](https://doi.org/10.1104/pp.109.138214) [Medline](#)
107. S. Ghosh, C.-K. K. Chan, in *Plant Bioinformatics: Methods and Protocols*, D. Edwards, Ed. (Springer, 2016), pp. 339–361.
108. S. R. Eddy, Accelerated profile HMM searches. *PLOS Comput. Biol.* **7**, e1002195 (2011). [doi:10.1371/journal.pcbi.1002195](https://doi.org/10.1371/journal.pcbi.1002195) [Medline](#)
109. L. Venturini, S. Caim, G. Kaithakottil, D. L. Mapleson, D. Swarbreck, Leveraging multiple transcriptome assembly methods for improved gene structure annotation. *Gigascience* **10.1093/gigascience/giy093** (2018).
110. D. Mapleson, L. Venturini, G. Kaithakottil, D. Swarbreck, Efficient and accurate detection of splice junctions from RNAseq with Portcullis. bioRxiv 217620 [Preprint]. 10 November 2017. <https://doi.org/10.1101/217620>.
111. A. Kozomara, S. Griffiths-Jones, miRBase: Integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.* **39**, D152–D157 (2011). [doi:10.1093/nar/gkq1027](https://doi.org/10.1093/nar/gkq1027) [Medline](#)

112. S. J. Lucas, H. Budak, Sorting the wheat from the chaff: Identifying miRNAs in genomic survey sequences of *Triticum aestivum* chromosome 1AL. *PLOS ONE* **7**, e40859 (2012). [doi:10.1371/journal.pone.0040859](https://doi.org/10.1371/journal.pone.0040859) [Medline](#)
113. N. R. Markham, M. Zuker, in *Bioinformatics: Structure, Function and Applications*, J. M. Keith, Ed. (Humana Press, 2008), pp. 3–31.
114. H. B. Cagirici, S. Biyiklioglu, H. Budak, Assembly and annotation of transcriptome provided evidence of miRNA mobility between wheat and wheat stem sawfly. *Front. Plant Sci.* **8**, 1653 (2017). [doi:10.3389/fpls.2017.01653](https://doi.org/10.3389/fpls.2017.01653) [Medline](#)
115. B. A. Akpinar, M. Kantar, H. Budak, Root precursors of microRNAs in wild emmer and modern wheats show major differences in response to drought stress. *Funct. Integr. Genomics* **15**, 587–598 (2015). [doi:10.1007/s10142-015-0453-0](https://doi.org/10.1007/s10142-015-0453-0) [Medline](#)
116. T. M. Lowe, S. R. Eddy, tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997). [doi:10.1093/nar/25.5.0955](https://doi.org/10.1093/nar/25.5.0955) [Medline](#)
117. The Gene Ontology Consortium, Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.* **45**, D331–D338 (2017). [doi:10.1093/nar/gkw1108](https://doi.org/10.1093/nar/gkw1108) [Medline](#)
118. L. Cooper, R. L. Walls, J. Elser, M. A. Gandolfo, D. W. Stevenson, B. Smith, J. Preece, B. Athreya, C. J. Mungall, S. Rensing, M. Hiss, D. Lang, R. Reski, T. Z. Berardini, D. Li, E. Huala, M. Schaeffer, N. Menda, E. Arnaud, R. Shrestha, Y. Yamazaki, P. Jaiswal, The plant ontology as a tool for comparative plant anatomy and genomic analyses. *Plant Cell Physiol.* **54**, e1 (2013). [doi:10.1093/pcp/pcs163](https://doi.org/10.1093/pcp/pcs163) [Medline](#)
119. E. Arnaud, L. Cooper, R. Shrestha, N. Menda, R. T. Nelson, L. Matteis, M. Skofic, R. Bastow, P. Jaiswal, L. Mueller, G. McLaren, “Towards a reference plant trait ontology for modeling knowledge of plant traits and phenotypes,” in *Proceedings of the International Conference on Knowledge Engineering and Ontology Development (KEOD)* (KEOD, 2012).
120. P. Borrill, R. Ramirez-Gonzalez, C. Uauy, expVIP: A customizable RNA-seq data analysis and visualization platform. *Plant Physiol.* **170**, 2172–2186 (2016). [doi:10.1104/pp.15.01667](https://doi.org/10.1104/pp.15.01667) [Medline](#)
121. N. L. Bray, H. Pimentel, P. Melsted, L. Pachter, Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016). [doi:10.1038/nbt.3519](https://doi.org/10.1038/nbt.3519) [Medline](#)
122. P. Langfelder, S. Horvath, WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008). [doi:10.1186/1471-2105-9-559](https://doi.org/10.1186/1471-2105-9-559) [Medline](#)
123. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014). [doi:10.1186/s13059-014-0550-8](https://doi.org/10.1186/s13059-014-0550-8) [Medline](#)
124. Y. Benjamini, D. Yekutieli, The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **29**, 1165–1188 (2001).
125. F. Krueger, S. R. Andrews, Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011). [doi:10.1093/bioinformatics/btr167](https://doi.org/10.1093/bioinformatics/btr167) [Medline](#)

126. A. Veluchamy, T. Jégu, F. Ariel, D. Latrasse, K. G. Mariappan, S.-K. Kim, M. Crespi, H. Hirt, C. Bergounioux, C. Raynaud, M. Benhamed, LHP1 regulates H3K27me3 spreading and shapes the three-dimensional conformation of the *Arabidopsis* genome. *PLOS ONE* **11**, e0158936 (2016). [doi:10.1371/journal.pone.0158936](https://doi.org/10.1371/journal.pone.0158936) [Medline](#)
127. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014). [doi:10.1093/bioinformatics/btu170](https://doi.org/10.1093/bioinformatics/btu170) [Medline](#)
128. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012). [doi:10.1038/nmeth.1923](https://doi.org/10.1038/nmeth.1923) [Medline](#)
129. A. D. Zimmer, D. Lang, K. Buchta, S. Rombauts, T. Nishiyama, M. Hasebe, Y. Van de Peer, S. A. Rensing, R. Reski, Reannotation and extended community resources for the genome of the non-seed plant *Physcomitrella patens* provide insights into the evolution of plant gene structures and functions. *BMC Genomics* **14**, 498 (2013). [doi:10.1186/1471-2164-14-498](https://doi.org/10.1186/1471-2164-14-498) [Medline](#)
130. J. S. Bernardes, F. R. J. Vieira, G. Zaverucha, A. Carbone, A multi-objective optimization approach accurately resolves protein domain architectures. *Bioinformatics* **32**, 345–353 (2016). [doi:10.1093/bioinformatics/btv582](https://doi.org/10.1093/bioinformatics/btv582) [Medline](#)
131. D. M. Emms, S. Kelly, OrthoFinder: Solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015). [doi:10.1186/s13059-015-0721-2](https://doi.org/10.1186/s13059-015-0721-2) [Medline](#)
132. J. Huerta-Cepas, H. Dopazo, J. Dopazo, T. Gabaldón, The human phylome. *Genome Biol.* **8**, R109 (2007). [Medline](#)
133. J. Huerta-Cepas, F. Serra, P. Bork, ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016). [doi:10.1093/molbev/msw046](https://doi.org/10.1093/molbev/msw046) [Medline](#)
134. S. Mirarab, T. Warnow, ASTRAL-II: Coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* **31**, i44–i52 (2015). [doi:10.1093/bioinformatics/btv234](https://doi.org/10.1093/bioinformatics/btv234) [Medline](#)
135. D. Lang, B. Weiche, G. Timmerhaus, S. Richardt, D. M. Riaño-Pachón, L. G. G. Corrêa, R. Reski, B. Mueller-Roeber, S. A. Rensing, Genome-wide phylogenetic comparative analysis of plant transcriptional regulation: A timeline of loss, gain, expansion, and correlation with complexity. *Genome Biol. Evol.* **2**, 488–503 (2010). [doi:10.1093/gbe/evq032](https://doi.org/10.1093/gbe/evq032) [Medline](#)
136. J. T. Garland, A. W. Dickerman, C. M. Janis, J. A. Jones, Phylogenetic analysis of covariance by computer simulation. *Syst. Biol.* **42**, 265–292 (1993). [doi:10.1093/sysbio/42.3.265](https://doi.org/10.1093/sysbio/42.3.265)
137. M.-C. Luo, Y. Q. Gu, D. Puiu, H. Wang, S. O. Twardziok, K. R. Deal, N. Huo, T. Zhu, L. Wang, Y. Wang, P. E. McGuire, S. Liu, H. Long, R. K. Ramasamy, J. C. Rodriguez, S. L. Van, L. Yuan, Z. Wang, Z. Xia, L. Xiao, O. D. Anderson, S. Ouyang, Y. Liang, A. V. Zimin, G. Pertea, P. Qi, J. L. Bennetzen, X. Dai, M. W. Dawson, H.-G. Müller, K. Kugler, L. Rivarola-Duarte, M. Spannagl, K. F. X. Mayer, F.-H. Lu, M. W. Bevan, P. Leroy, P. Li, F. M. You, Q. Sun, Z. Liu, E. Lyons, T. Wicker, S. L. Salzberg, K. M.

- Devos, J. Dvořák, Genome sequence of the progenitor of the wheat D genome *Aegilops tauschii*. *Nature* **551**, 498–502 (2017). [doi:10.1038/nature24486](https://doi.org/10.1038/nature24486) [Medline](#)
138. S. Grossmann, S. Bauer, P. N. Robinson, M. Vingron, Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis. *Bioinformatics* **23**, 3024–3031 (2007). [doi:10.1093/bioinformatics/btm440](https://doi.org/10.1093/bioinformatics/btm440) [Medline](#)
 139. S. Aibar, C. Fontanillo, C. Droste, J. De Las Rivas, Functional Gene Networks: R/Bioc package to generate and analyse gene networks derived from functional enrichment and clustering. *Bioinformatics* **31**, 1686–1688 (2015). [doi:10.1093/bioinformatics/btu864](https://doi.org/10.1093/bioinformatics/btu864) [Medline](#)
 140. G. Su, A. Kuchinsky, J. H. Morris, D. J. States, F. Meng, GLay: Community structure analysis of biological networks. *Bioinformatics* **26**, 3135–3137 (2010). [doi:10.1093/bioinformatics/btq596](https://doi.org/10.1093/bioinformatics/btq596) [Medline](#)
 141. N. P. Nguyen, S. Mirarab, K. Kumar, T. Warnow, Ultra-large alignments using phylogeny-aware profiles. *Genome Biol.* **16**, 124 (2015). [doi:10.1186/s13059-015-0688-z](https://doi.org/10.1186/s13059-015-0688-z) [Medline](#)
 142. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree 2—Approximately maximum-likelihood trees for large alignments. *PLOS ONE* **5**, e9490 (2010). [doi:10.1371/journal.pone.0009490](https://doi.org/10.1371/journal.pone.0009490) [Medline](#)
 143. I. Letunic, P. Bork, Interactive tree of life (iTOL) v3: An online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242–W245 (2016). [doi:10.1093/nar/gkw290](https://doi.org/10.1093/nar/gkw290) [Medline](#)
 144. N. M. Glover, H. Redestig, C. Dessimoz, Homoeologs: What are they and how do we infer them? *Trends Plant Sci.* **21**, 609–621 (2016). [doi:10.1016/j.tplants.2016.02.005](https://doi.org/10.1016/j.tplants.2016.02.005) [Medline](#)
 145. Y. Wang, H. Tang, J. D. Debarry, X. Tan, J. Li, X. Wang, T. H. Lee, H. Jin, B. Marler, H. Guo, J. C. Kissinger, A. H. Paterson, MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012). [doi:10.1093/nar/gkr1293](https://doi.org/10.1093/nar/gkr1293) [Medline](#)
 146. M. Mascher, H. Gundlach, A. Himmelbach, S. Beier, S. O. Twardziok, T. Wicker, V. Radchuk, C. Dockter, P. E. Hedley, J. Russell, M. Bayer, L. Ramsay, H. Liu, G. Haberer, X.-Q. Zhang, Q. Zhang, R. A. Barrero, L. Li, S. Taudien, M. Groth, M. Felder, A. Hastie, H. Šimková, H. Staňková, J. Vrána, S. Chan, M. Muñoz-Amatriaín, R. Ounit, S. Wanamaker, D. Bolser, C. Colmsee, T. Schmutzer, L. Aliyeva-Schnorr, S. Grasso, J. Tanskanen, A. Chailyan, D. Sampath, D. Heavens, L. Clissold, S. Cao, B. Chapman, F. Dai, Y. Han, H. Li, X. Li, C. Lin, J. K. McCooke, C. Tan, P. Wang, S. Wang, S. Yin, G. Zhou, J. A. Poland, M. I. Bellgard, L. Borisjuk, A. Houben, J. Doležel, S. Ayling, S. Lonardi, P. Kersey, P. Langridge, G. J. Muehlbauer, M. D. Clark, M. Caccamo, A. H. Schulman, K. F. X. Mayer, M. Platzer, T. J. Close, U. Scholz, M. Hansson, G. Zhang, I. Braumann, M. Spannagl, C. Li, R. Waugh, N. Stein, A chromosome conformation capture ordered sequence of the barley genome. *Nature* **544**, 427–433 (2017). [doi:10.1038/nature22043](https://doi.org/10.1038/nature22043) [Medline](#)
 147. L. Al Ait, Z. Yamak, B. Morgenstern, DIALIGN at GOBICS—Multiple sequence alignment using various sources of external information. *Nucleic Acids Res.* **41**, W3–W7 (2013). [doi:10.1093/nar/gkt283](https://doi.org/10.1093/nar/gkt283) [Medline](#)

148. D. Wang, Y. Zhang, Z. Zhang, J. Zhu, J. Yu, KaKs_Calculator 2.0: A toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteomics Bioinformatics* **8**, 77–80 (2010). [doi:10.1016/S1672-0229\(10\)60008-3](https://doi.org/10.1016/S1672-0229(10)60008-3) [Medline](#)
149. J. C. Zadoks, T. T. Chang, C. F. Konzak, A decimal code for the growth stages of cereals. *Weed Res. (Oxford)* **14**, 415–421 (1974). [doi:10.1111/j.1365-3180.1974.tb01084.x](https://doi.org/10.1111/j.1365-3180.1974.tb01084.x)
150. A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T. R. Gingeras, STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013). [doi:10.1093/bioinformatics/bts635](https://doi.org/10.1093/bioinformatics/bts635) [Medline](#)
151. M. Pertea, D. Kim, G. M. Pertea, J. T. Leek, S. L. Salzberg, Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* **11**, 1650–1667 (2016). [doi:10.1038/nprot.2016.095](https://doi.org/10.1038/nprot.2016.095) [Medline](#)
152. K. J. Livak, T. D. Schmittgen, Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta C_T}$ method. *Methods* **25**, 402–408 (2001). [doi:10.1006/meth.2001.1262](https://doi.org/10.1006/meth.2001.1262) [Medline](#)
153. A. H. Paterson, J. E. Bowers, R. Bruggmann, I. Dubchak, J. Grimwood, H. Gundlach, G. Haberer, U. Hellsten, T. Mitros, A. Poliakov, J. Schmutz, M. Spannagl, H. Tang, X. Wang, T. Wicker, A. K. Bharti, J. Chapman, F. A. Feltus, U. Gowik, I. V. Grigoriev, E. Lyons, C. A. Maher, M. Martis, A. Narechania, R. P. Olliar, B. W. Penning, A. A. Salamov, Y. Wang, L. Zhang, N. C. Carpita, M. Freeling, A. R. Gingle, C. T. Hash, B. Keller, P. Klein, S. Kresovich, M. C. McCann, R. Ming, D. G. Peterson, D. Mehboob-ur-Rahman, D. Ware, P. Westhoff, K. F. Mayer, J. Messing, D. S. Rokhsar, The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**, 551–556 (2009). [doi:10.1038/nature07723](https://doi.org/10.1038/nature07723) [Medline](#)
154. P. S. Schnable, D. Ware, R. S. Fulton, J. C. Stein, F. Wei, S. Pasternak, C. Liang, J. Zhang, L. Fulton, T. A. Graves, P. Minx, A. D. Reily, L. Courtney, S. S. Kruchowski, C. Tomlinson, C. Strong, K. Delehaunty, C. Fronick, B. Courtney, S. M. Rock, E. Belter, F. Du, K. Kim, R. M. Abbott, M. Cotton, A. Levy, P. Marchetto, K. Ochoa, S. M. Jackson, B. Gillam, W. Chen, L. Yan, J. Higginbotham, M. Cardenas, J. Waligorski, E. Applebaum, L. Phelps, J. Falcone, K. Kanchi, T. Thane, A. Scimone, N. Thane, J. Henke, T. Wang, J. Ruppert, N. Shah, K. Rotter, J. Hodges, E. Ingenthron, M. Cordes, S. Kohlberg, J. Sgro, B. Delgado, K. Mead, A. Chinwalla, S. Leonard, K. Crouse, K. Collura, D. Kudrna, J. Currie, R. He, A. Angelova, S. Rajasekar, T. Mueller, R. Lomeli, G. Scara, A. Ko, K. Delaney, M. Wissotski, G. Lopez, D. Campos, M. Braidotti, E. Ashley, W. Golser, H. Kim, S. Lee, J. Lin, Z. Dujmic, W. Kim, J. Talag, A. Zuccolo, C. Fan, A. Sebastian, M. Kramer, L. Spiegel, L. Nascimento, T. Zutavern, B. Miller, C. Ambroise, S. Muller, W. Spooner, A. Narechania, L. Ren, S. Wei, S. Kumari, B. Faga, M. J. Levy, L. McMahan, P. Van Buren, M. W. Vaughn, K. Ying, C.-T. Yeh, S. J. Emrich, Y. Jia, A. Kalyanaraman, A.-P. Hsia, W. B. Barbazuk, R. S. Baucom, T. P. Brutnell, N. C. Carpita, C. Chaparro, J.-M. Chia, J.-M. Deragon, J. C. Estill, Y. Fu, J. A. Jeddelloh, Y. Han, H. Lee, P. Li, D. R. Lisch, S. Liu, Z. Liu, D. H. Nagel, M. C. McCann, P. SanMiguel, A. M. Myers, D. Nettleton, J. Nguyen, B. W. Penning, L. Ponnala, K. L. Schneider, D. C. Schwartz, A. Sharma, C. Soderlund, N. M. Springer, Q. Sun, H. Wang, M. Waterman, R. Westerman, T. K. Wolfgruber, L. Yang, Y. Yu, L. Zhang, S. Zhou, Q. Zhu, J. L. Bennetzen, R. K. Dawe, J. Jiang, N. Jiang, G. G. Presting, S. R. Wessler, S. Aluru, R. A. Martienssen, S. W. Clifton, W. R. McCombie, R. A. Wing, R. K. Wilson,

- The B73 maize genome: Complexity, diversity, and dynamics. *Science* **326**, 1112–1115 (2009). [doi:10.1126/science.1178534](https://doi.org/10.1126/science.1178534) [Medline](#)
155. E. Bauer, T. Schmutzer, I. Barilar, M. Mascher, H. Gundlach, M. M. Martis, S. O. Twardziok, B. Hackauf, A. Gordillo, P. Wilde, M. Schmidt, V. Korzun, K. F. X. Mayer, K. Schmid, C.-C. Schön, U. Scholz, Towards a whole-genome sequence for rye (*Secale cereale* L.). *Plant J.* **89**, 853–869 (2017). [doi:10.1111/tpj.13436](https://doi.org/10.1111/tpj.13436) [Medline](#)
 156. M. J. Sanderson, Estimating absolute rates of molecular evolution and divergence times: A penalized likelihood approach. *Mol. Biol. Evol.* **19**, 101–109 (2002). [doi:10.1093/oxfordjournals.molbev.a003974](https://doi.org/10.1093/oxfordjournals.molbev.a003974) [Medline](#)
 157. P.-A. Christin, E. Spriggs, C. P. Osborne, C. A. E. Strömberg, N. Salamin, E. J. Edwards, Molecular dating, evolutionary rates, and the age of the grasses. *Syst. Biol.* **63**, 153–165 (2014). [doi:10.1093/sysbio/syt072](https://doi.org/10.1093/sysbio/syt072) [Medline](#)
 158. X. C. Zhao, P. J. Sharp, Production of all eight genotypes of null alleles at ‘waxy’ loci in bread wheat, *Triticum aestivum* L. *Plant Breed.* **117**, 488–490 (1998). [doi:10.1111/j.1439-0523.1998.tb01979.x](https://doi.org/10.1111/j.1439-0523.1998.tb01979.x)
 159. K. W. Jordan, S. Wang, Y. Lun, L.-J. Gardiner, R. MacLachlan, P. Hucl, K. Wiebe, D. Wong, K. L. Forrest, A. G. Sharpe, C. H. D. Sidebottom, N. Hall, C. Toomajian, T. Close, J. Dubcovsky, A. Akhunova, L. Talbert, U. K. Bansal, H. S. Bariana, M. J. Hayden, C. Pozniak, J. A. Jeddelloh, A. Hall, E. Akhunov; IWGS Consortium, A haplotype map of allohexaploid wheat reveals distinct patterns of selection on homoeologous genomes. *Genome Biol.* **16**, 48 (2015). [doi:10.1186/s13059-015-0606-4](https://doi.org/10.1186/s13059-015-0606-4) [Medline](#)
 160. M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, M. A. Marra, Circos: An information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009). [doi:10.1101/gr.092759.109](https://doi.org/10.1101/gr.092759.109) [Medline](#)
 161. C. E. Niederhuth, A. J. Bewick, L. Ji, M. S. Alabady, K. D. Kim, Q. Li, N. A. Rohr, A. Rambani, J. M. Burke, J. A. Udall, C. Egesi, J. Schmutz, J. Grimwood, S. A. Jackson, N. M. Springer, R. J. Schmitz, Widespread natural variation of DNA methylation within angiosperms. *Genome Biol.* **17**, 194 (2016). [doi:10.1186/s13059-016-1059-0](https://doi.org/10.1186/s13059-016-1059-0) [Medline](#)
 162. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
 163. G. Yu, F. Li, Y. Qin, X. Bo, Y. Wu, S. Wang, GOSemSim: An R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* **26**, 976–978 (2010). [doi:10.1093/bioinformatics/btq064](https://doi.org/10.1093/bioinformatics/btq064) [Medline](#)
 164. S. Fischer, B. P. Brunk, F. Chen, X. Gao, O. S. Harb, J. B. Iodice, D. Shanmugam, D. S. Roos, C. J. Stoeckert Jr., Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. *Curr. Protoc. Bioinformatics* **35**, 6.12.11–16.12.19 (2011). [doi:10.1002/0471250953.bi0612s35](https://doi.org/10.1002/0471250953.bi0612s35)
 165. S. Cheng, B. Gutmann, X. Zhong, Y. Ye, M. F. Fisher, F. Bai, I. Castleden, Y. Song, B. Song, J. Huang, X. Liu, X. Xu, B. L. Lim, C. S. Bond, S.-M. Yiu, I. Small, Redefining the structural motifs that determine RNA binding and RNA editing by pentatricopeptide repeat proteins in land plants. *Plant J.* **85**, 532–547 (2016). [doi:10.1111/tpj.13121](https://doi.org/10.1111/tpj.13121) [Medline](#)

166. Z. Li, W. Zhu, S. Ma, G. Zhang, X. Zhao, P. Zhang, N. Niu, J. Wang, SSR analysis and identification of fertility restorer genes *Rf1* and *Rf4* of *Triticum timopheevii* cytoplasmic male sterility (T-CMS) in wheat (*Triticum aestivum* L.). *J. Agric. Biotechnol.* **22**, 1114–1122 (2014).
167. M. Geyer, T. Albrecht, L. Hartl, V. Mohler, Exploring the genetics of fertility restoration controlled by *Rf1* in common wheat (*Triticum aestivum* L.) using high-density linkage maps. *Mol. Genet. Genomics* **293**, 451–462 (2018). [doi:10.1007/s00438-017-1396-z](https://doi.org/10.1007/s00438-017-1396-z) [Medline](#)
168. P. Sinha, S. M. S. Tomar, V. K. Vinod, V. K. Singh, H. S. Balyan, Genetic analysis and molecular mapping of a new fertility restorer gene *Rf8* for *Triticum timopheevii* cytoplasm in wheat (*Triticum aestivum* L.) using SSR markers. *Genetica* **141**, 431–441 (2013). [doi:10.1007/s10709-013-9742-5](https://doi.org/10.1007/s10709-013-9742-5) [Medline](#)
169. J. Breen, T. Wicker, M. Shatalina, Z. Frenkel, I. Bertin, R. Philippe, W. Spielmeyer, H. Simková, J. Šafář, F. Cattonaro, S. Scalabrin, F. Magni, S. Vautrin, H. Bergès, E. Paux, T. Fahima, J. Doležal, A. Korol, C. Feuillet, B. Keller; International Wheat Genome Sequencing Consortium, A physical map of the short arm of wheat chromosome 1A. *PLOS ONE* **8**, e80272 (2013). [doi:10.1371/journal.pone.0080272](https://doi.org/10.1371/journal.pone.0080272) [Medline](#)
170. S. J. Lucas, B. A. Akpınar, M. Kantar, Z. Weinstein, F. Aydınoglu, J. Šafář, H. Simková, Z. Frenkel, A. Korol, F. Magni, F. Cattonaro, S. Vautrin, A. Bellec, H. Bergès, J. Doležal, H. Budak, Physical mapping integrated with syntenic analysis to characterize the gene space of the long arm of wheat chromosome 1A. *PLOS ONE* **8**, e59542 (2013). [doi:10.1371/journal.pone.0059542](https://doi.org/10.1371/journal.pone.0059542) [Medline](#)
171. D. Raats, Z. Frenkel, T. Krugman, I. Dodek, H. Sela, H. Simková, F. Magni, F. Cattonaro, S. Vautrin, H. Bergès, T. Wicker, B. Keller, P. Leroy, R. Philippe, E. Paux, J. Doležal, C. Feuillet, A. Korol, T. Fahima, The physical map of wheat chromosome 1BS provides insights into its gene space organization and evolution. *Genome Biol.* **14**, R138 (2013). [doi:10.1186/gb-2013-14-12-r138](https://doi.org/10.1186/gb-2013-14-12-r138) [Medline](#)
172. R. Philippe, E. Paux, I. Bertin, P. Sourdille, F. Choulet, C. Laugier, H. Simková, J. Šafář, A. Bellec, S. Vautrin, Z. Frenkel, F. Cattonaro, F. Magni, S. Scalabrin, M. M. Martis, K. F. X. Mayer, A. Korol, H. Bergès, J. Doležal, C. Feuillet, A high density physical map of chromosome 1BL supports evolutionary studies, map-based cloning and sequencing in wheat. *Genome Biol.* **14**, R64 (2013). [doi:10.1186/gb-2013-14-6-r64](https://doi.org/10.1186/gb-2013-14-6-r64) [Medline](#)
173. E. Paux, P. Sourdille, J. Salse, C. Sautenac, F. Choulet, P. Leroy, A. Korol, M. Michalak, S. Kianian, W. Spielmeyer, E. Lagudah, D. Somers, A. Kilian, M. Alaux, S. Vautrin, H. Bergès, K. Eversole, R. Appels, J. Safar, H. Simkova, J. Dolezel, M. Bernard, C. Feuillet, A physical map of the 1-gigabase bread wheat chromosome 3B. *Science* **322**, 101–104 (2008). [doi:10.1126/science.1161847](https://doi.org/10.1126/science.1161847) [Medline](#)
174. K. Holušová, J. Vrána, J. Šafář, H. Šimková, B. Balcárková, Z. Frenkel, B. Darrier, E. Paux, F. Cattonaro, H. Berges, T. Letellier, M. Alaux, J. Doležal, J. Bartoš, Physical map of the short arm of bread wheat chromosome 3D. *Plant Genome* **10**, 0021 (2017). [Medline](#)
175. O. Shorinola, B. Balcárková, J. Hyles, J. F. G. Tibbits, M. J. Hayden, K. Holušova, M. Valárik, A. Distelfeld, A. Torada, J. M. Barrero, C. Uauy, Association mapping and haplotype analysis of the pre-harvest sprouting resistance locus *Phs-A1* reveals a causal

- role of *TaMKK3-A* in global germplasm. *Front. Plant Sci.* **8**, 1555 (2017). [doi:10.3389/fpls.2017.01555](https://doi.org/10.3389/fpls.2017.01555) [Medline](#)
176. D. Barabaschi, F. Magni, A. Volante, A. Gadaleta, H. Šimková, S. Scalabrin, M. L. Prazzoli, P. Bagnaresi, K. Lacrima, V. Michelotti, F. Desiderio, L. Orrù, V. Mazzamurro, A. Fricano, A. M. Mastrangelo, P. Tononi, N. Vitulo, I. Jurman, Z. Frenkel, F. Cattonaro, M. Morgante, A. Blanco, J. Doležel, M. Delledonne, A. M. Stanca, L. Cattivelli, G. Valè, Physical mapping of bread wheat chromosome 5A: An integrated approach. *Plant Genome* **8**, 0011 (2015). [doi:10.3835/plantgenome2015.03.0011](https://doi.org/10.3835/plantgenome2015.03.0011)
 177. E. A. Salina, M. A. Nesterov, Z. Frenkel, A. A. Kiseleva, E. M. Timonova, F. Magni, J. Vrána, J. Šafař, H. Šimková, J. Doležel, A. Korol, E. M. Sergeeva, Features of the organization of bread wheat chromosome 5BS based on physical mapping. *BMC Genomics* **19** (suppl. 3), 80 (2018). [doi:10.1186/s12864-018-4470-y](https://doi.org/10.1186/s12864-018-4470-y) [Medline](#)
 178. B. A. Akpinar, F. Magni, M. Yuce, S. J. Lucas, H. Šimková, J. Šafař, S. Vautrin, H. Bergès, F. Cattonaro, J. Doležel, H. Budak, The physical map of wheat chromosome 5DS revealed gene duplications and small rearrangements. *BMC Genomics* **16**, 453 (2015). [doi:10.1186/s12864-015-1641-y](https://doi.org/10.1186/s12864-015-1641-y) [Medline](#)
 179. T. Belova, L. Grønvold, A. Kumar, S. Kianian, X. He, M. Lillemo, N. M. Springer, S. Lien, O.-A. Olsen, S. R. Sandve, Utilization of deletion bins to anchor and order sequences along the wheat 7B chromosome. *Theor. Appl. Genet.* **127**, 2029–2040 (2014). [doi:10.1007/s00122-014-2358-z](https://doi.org/10.1007/s00122-014-2358-z) [Medline](#)
 180. Z. Tulpová, M.-C. Luo, H. Toegelová, P. Visendi, S. Hayashi, P. Vojta, E. Paux, A. Kilian, M. Abrouk, J. Bartoš, M. Hajdúch, J. Batley, D. Edwards, J. Doležel, H. Šimková, Integrated physical map of bread wheat chromosome arm 7DS to facilitate gene cloning and comparative studies. *N. Biotechnol.* 10.1016/j.nbt.2018.03.003 (2018).
 181. F.-H. Lu, N. McKenzie, G. Kettleborough, D. Heavens, M. D. Clark, M. W. Bevan, Independent assessment and improvement of wheat genome sequence assemblies using Fosill jumping libraries. *GigaScience* **7**, giy053 (2018). [doi:10.1093/gigascience/giy053](https://doi.org/10.1093/gigascience/giy053) [Medline](#)
 182. M. A. Nesterov, D. A. Afonnikov, E. M. Sergeeva, L. A. Miroshnichenko, M. K. Bragina, A. O. Bragin, G. V. Vasiliev, E. A. Salina, Identification of microsatellite loci based on BAC sequencing data and their physical mapping into the soft wheat 5B chromosome. *Russian J. Genet. Appl. Res.* **6**, 825–837 (2016). [doi:10.1134/S2079059716070078](https://doi.org/10.1134/S2079059716070078)
 183. E. M. Sergeeva, A. B. Shcherban, I. G. Adonina, M. A. Nesterov, A. V. Beletsky, A. L. Rakitin, A. V. Mardanov, N. V. Ravin, E. A. Salina, Fine organization of genomic regions tagged to the 5S rDNA locus of the bread wheat 5B chromosome. *BMC Plant Biol.* **17** (suppl. 1), 183 (2017). [doi:10.1186/s12870-017-1120-5](https://doi.org/10.1186/s12870-017-1120-5) [Medline](#)
 184. M.-C. Luo, Y. Q. Gu, F. M. You, K. R. Deal, Y. Ma, Y. Hu, N. Huo, Y. Wang, J. Wang, S. Chen, C. M. Jorgensen, Y. Zhang, P. E. McGuire, S. Pasternak, J. C. Stein, D. Ware, M. Kramer, W. R. McCombie, S. F. Kianian, M. M. Martis, K. F. X. Mayer, S. K. Sehgal, W. Li, B. S. Gill, M. W. Bevan, H. Šimková, J. Doležel, S. Weining, G. R. Lazo, O. D. Anderson, J. Dvorak, A 4-gigabase physical map unlocks the structure and evolution of the complex genome of *Aegilops tauschii*, the wheat D-genome progenitor. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 7940–7945 (2013). [doi:10.1073/pnas.1219082110](https://doi.org/10.1073/pnas.1219082110) [Medline](#)

185. X. Zeng, H. Long, Z. Wang, S. Zhao, Y. Tang, Z. Huang, Y. Wang, Q. Xu, L. Mao, G. Deng, X. Yao, X. Li, L. Bai, H. Yuan, Z. Pan, R. Liu, X. Chen, Q. WangMu, M. Chen, L. Yu, J. Liang, D. DunZhu, Y. Zheng, S. Yu, Z. LuoBu, X. Guang, J. Li, C. Deng, W. Hu, C. Chen, X. TaBa, L. Gao, X. Lv, Y. B. Abu, X. Fang, E. Nevo, M. Yu, J. Wang, N. Tashi, The draft genome of Tibetan hulless barley reveals adaptive patterns to the high stressful Tibetan Plateau. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 1095–1100 (2015).
[doi:10.1073/pnas.1423628112](https://doi.org/10.1073/pnas.1423628112) [Medline](#)
186. Y. Kawahara, M. de la Bastide, J. P. Hamilton, H. Kanamori, W. R. McCombie, S. Ouyang, D. C. Schwartz, T. Tanaka, J. Wu, S. Zhou, K. L. Childs, R. M. Davidson, H. Lin, L. Quesada-Ocampo, B. Vaillancourt, H. Sakai, S. S. Lee, J. Kim, H. Numa, T. Itoh, C. R. Buell, T. Matsumoto, Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice (N. Y.)* **6**, 4 (2013).
[doi:10.1186/1939-8433-6-4](https://doi.org/10.1186/1939-8433-6-4) [Medline](#)