A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome

The International Wheat Genome Sequencing Consortium (IWGSC)

Science **345**, (2014);
DOI: 10.1126/science.1251788

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome

The International Wheat Genome Sequencing Consortium (IWGSC)

An ordered draft sequence of the 17-gigabase hexaploid bread wheat (*Triticum aestivum*) genome has been produced by sequencing isolated chromosome arms. We have annotated 124,201 gene loci distributed nearly evenly across the homeologous chromosomes and subgenomes. Comparative gene analysis of wheat subgenomes and extant diploid and tetraploid wheat relatives showed that high sequence similarity and structural conservation are retained, with limited gene loss, after polyploidization. However, across the genomes there was evidence of dynamic gene gain, loss, and duplication since the divergence of the wheat lineages. A high degree of transcriptional autonomy and no global dominance was found for the subgenomes. These insights into the genome biology of a polyploid crop provide a springboard for faster gene isolation, rapid genetic marker development, and precise breeding to meet the needs of increasing food demand worldwide.

Lists of authors and affiliations are available in the full article online.
Corresponding author: K. X. Mayer, e-mail: k.mayer@helmholtz-muenchen.de
Read the full article at http://dx.doi.org/10.1126/science.1251788

Ancient hybridizations among the ancestral genomes of bread wheat

Thomas Marcussen, Simen R. Sandve,* Lise Heier, Manuel Spannagl, Matthias Pfeifer, The International Wheat Genome Sequencing Consortium,† Kjetill S. Jakobsen, Brande B. H Wulff, Burkhard Steuernagel, Klaus F. X. Mayer, Odd-Arne Olsen

The allohexaploid bread wheat genome consists of three closely related subgenomes (A, B, and D), but a clear understanding of their phylogenetic history has been lacking. We used genome assemblies of bread wheat and five diploid relatives to analyze genome-wide samples of gene trees, as well as to estimate evolutionary relatedness and divergence times. We show that the A and B genomes diverged from a common ancestor ~7 million years ago and that these genomes gave rise to the D genome through homoploid hybrid speciation 1 to 2 million years later. Our findings imply that the present-day bread wheat genome is a product of multiple rounds of hybrid speciation (homoploid and polyploid) and lay the foundation for a new framework for understanding the wheat genome as a multilevel phylogenetic mosaic.

The list of author affiliations is available in the full article online.***Corresponding author.**
E-mail: simen.sandve@nmbu.no †The International Wheat Genome Sequencing Consortium (IWGSC) authors and affiliations are listed in the supplementary materials.
Read the full article at http://dx.doi.org/10.1126/science.1250092


Triticum monococcum

Triticum carthlicum

Ancestral wheat

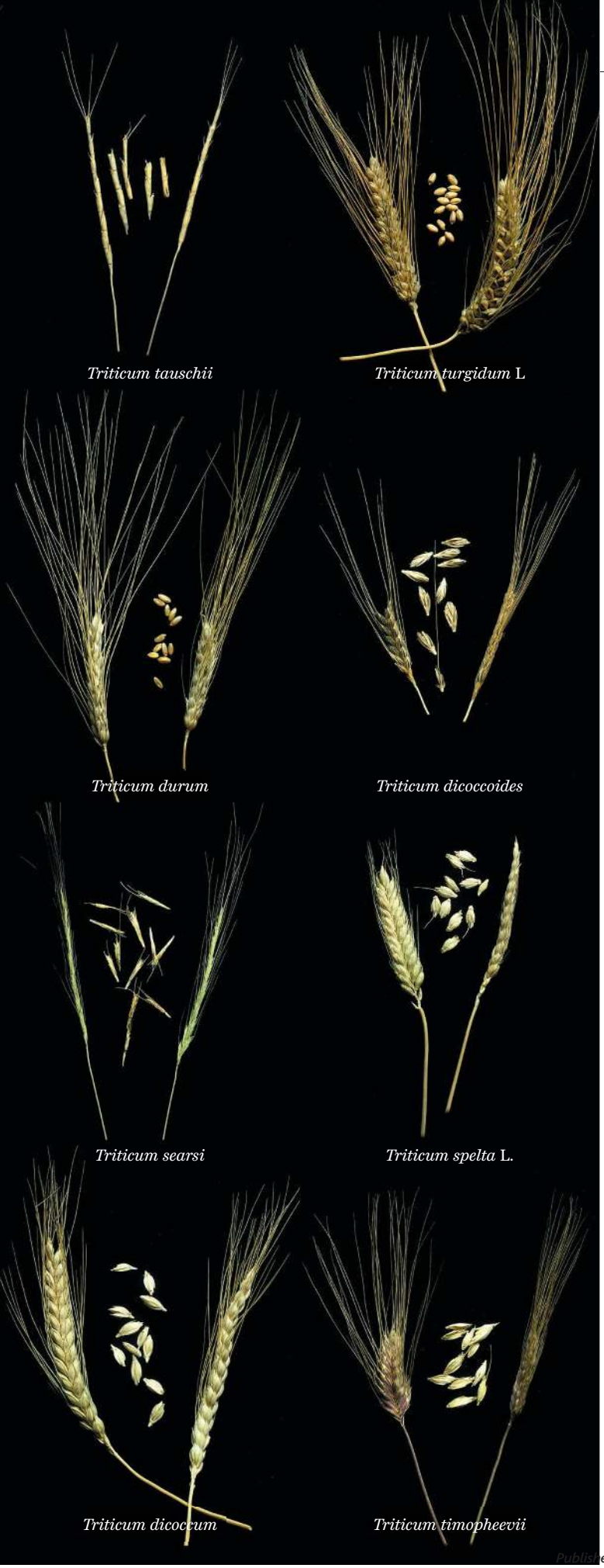
Wheat varieties and species (shown) believed to be the closest living relatives of modern bread wheat (*T. aestivum*). Multiple ancestral hybridizations occurred among most of these species, many of which are cultivated, and along with *T. aestivum* represent a dominant source of global nutrition.


Triticum boeticum


Triticum polonicum L.


Triticum macha


Triticum dicoccoides var. *araraticum*



PHOTOS: SUSANNE STAMP, ERNST MERZ/ETH ZURICH

Genome interplay in the grain transcriptome of hexaploid bread wheat

Matthias Pfeifer, Karl G. Kugler, Simen R. Sandve, Bujie Zhan, Heidi Rudi, Torgeir R. Hvidsten, International Wheat Genome Sequencing Consortium,* Klaus F. X. Mayer, Odd-Arne Olsen†

Allohexaploid bread wheat (*Triticum aestivum* L.) provides approximately 20% of calories consumed by humans. Lack of genome sequence for the three homeologous and highly similar bread wheat genomes (A, B, and D) has impeded expression analysis of the grain transcriptome. We used previously unknown genome information to analyze the cell type–specific expression of homeologous genes in the developing wheat grain and identified distinct co-expression clusters reflecting the spatiotemporal progression during endosperm development. We observed no global but cell type– and stage-dependent genome dominance, organization of the wheat genome into transcriptionally active chromosomal regions, and asymmetric expression in gene families related to baking quality. Our findings give insight into the transcriptional dynamics and genome interplay among individual grain cell types in a polyploid cereal genome.

The list of author affiliations is available in the full article online. *The International Wheat Genome Sequencing Consortium (IWGSC) authors and affiliations are listed in the supplementary materials. †**Corresponding author. E-mail: odd-arne.olsen@nmbu.no**

Read the full article at http://dx.doi.org/10.1126/science.1250091

Structural and functional partitioning of bread wheat chromosome 3B

Frédéric Choulet,* Adriana Alberti, Sébastien Theil, Natasha Glover, Valérie Barbe, Josquin Daron, Lise Pingault, Pierre Sourdille, Arnaud Couloux, Etienne Paux, Philippe Leroy, Sophie Mangenot, Nicolas Guilhot, Jacques Le Gouis, Francois Balfourier, Michael Alaux, Véronique Jamilloux, Julie Poulain, Céline Durand, Arnaud Bellec, Christine Gaspin, Jan Safar, Jaroslav Dolezel, Jane Rogers, Klaas Vandepoele, Jean-Marc Aury, Klaus Mayer, Hélène Berges, Hadi Quesneville, Patrick Wincker, Catherine Feuillet

We produced a reference sequence of the 1-gigabase chromosome 3B of hexaploid bread wheat. By sequencing 8452 bacterial artificial chromosomes in pools, we assembled a sequence of 774 megabases carrying 5326 protein-coding genes, 1938 pseudogenes, and 85% of transposable elements. The distribution of structural and functional features along the chromosome revealed partitioning correlated with meiotic recombination. Comparative analyses indicated high wheat-specific inter- and intrachromosomal gene duplication activities that are potential sources of variability for adaption. In addition to providing a better understanding of the organization, function, and evolution of a large and polyploid genome, the availability of a high-quality sequence anchored to genetic maps will accelerate the identification of genes underlying important agronomic traits.

The list of author affiliations is available in the full article online.
***Corresponding author. E-mail: frederic.choulet@clermont.inra.fr**

Read the full article at http://dx.doi.org/10.1126/science.1249721

WHEAT GENOME

A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome

The International Wheat Genome Sequencing Consortium (IWGSC)*†

An ordered draft sequence of the 17-gigabase hexaploid bread wheat (*Triticum aestivum*) genome has been produced by sequencing isolated chromosome arms. We have annotated 124,201 gene loci distributed nearly evenly across the homeologous chromosomes and subgenomes. Comparative gene analysis of wheat subgenomes and extant diploid and tetraploid wheat relatives showed that high sequence similarity and structural conservation are retained, with limited gene loss, after polyploidization. However, across the genomes there was evidence of dynamic gene gain, loss, and duplication since the divergence of the wheat lineages. A high degree of transcriptional autonomy and no global dominance was found for the subgenomes. These insights into the genome biology of a polyploid crop provide a springboard for faster gene isolation, rapid genetic marker development, and precise breeding to meet the needs of increasing food demand worldwide.

Rich in protein, carbohydrates, and minerals, bread wheat (*Triticum aestivum* L.) is one of the world's most important cereal grain crops, serving as the staple food source for 30% of the human population. Between 2000 and 2008, wheat production fell by 5.5% primarily because of climatic trends (*1*), and, in 5 of the past 10 years, worldwide wheat production was not sufficient to meet demand (*2*). With the global population projected to exceed 9 billion by 2050, researchers, breeders and growers are facing the challenge of increasing wheat production by about 70% to meet future demands (*3*, *4*). Concurrently, growers are facing rising fertilizer and other input costs, weather extremes resulting from climate change, increasing competition between food and nonfood uses, and declining annual yield growth (*5*). A rapid paradigm shift in science-based advances in wheat genetics and breeding, comparable to the first green revolution of the 1960s, will be essential to meet these challenges. As for other major cereal crops (rice, maize, and sorghum), new knowledge and molecular tools using a reference genome sequence of wheat are needed to underpin breeding to accelerate the development of new wheat varieties.

One key factor in the success of wheat as a global food crop is its adaptability to a wide range of climatic conditions. This is attributable, in part, to its allohexaploid genome structure, which arose as a result of two polyploidization events (Fig. 1). The first of these is estimated to have occurred several hundred thousand years ago and brought together the genomes of two diploid species related to the wild species *Triticum urartu* ($2n = 2x = 14$; AA; $2n$ is the number of chromosomes in each somatic cell and $2x$ is the basic chromosome number) and a species from the Sitopsis section of *Triticum* that is believed to be related to *Aegilops speltoides* ($2n = 14$; SS) (*6*). This hybridization formed the allotetraploid *Triticum turgidum* ($2n = 4x = 28$; AABB), an ancestor of wild emmer wheat cultivated in the Middle East and *T. turgidum* sp. *durum* grown for pasta today. A second hybridization event between *T. turgidum* and a diploid grass species, *Aegilops tauschii* (DD), produced the ancestral allohexaploid *T. aestivum* ($2n = 6x = 42$, AABBDD) (*6*, *7*), which has since been cultivated as bread wheat and accounts for over 95% of the wheat grown worldwide.

With 21 pairs of chromosomes, bread wheat is structurally an allopolyploid with three homeologous sets of seven chromosomes in each of the A, B, and D subgenomes. Genetically, however, it behaves as a diploid because homeologous pairing is prevented through the action of *Ph* genes (*8*). Each of the subgenomes is large, about 5.5 Gb in size and carries, in addition to related sets of genes, a high proportion (>80%) of highly repetitive transposable elements (TEs) (*9*, *10*).

The large and repetitive nature of the genome has hindered the generation of a reference genome sequence for bread wheat. Early work focused primarily on coding sequences that represent less than 2% of the genome. Coordinated efforts generated over 1 million expressed sequence tags (ESTs), 40,000 unigenes (www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html), and 17,000 full-length complementary DNA (cDNA) sequences (*11*). These resources have enabled studies of individual genes and facilitated the development of microarrays and marker sets for targeted gene association and expression studies (*12*–*14*). At least 7000 ESTs have been assigned to chromosome-specific bins (*15*), providing an initial view of subgenome localization and chromosomal organization and facilitating low-resolution mapping of traits. More recently, high-throughput low-cost sequencing technologies have been applied to assemble the gene space of *T. urartu* (*16*) and *Ae. tauschii* (*17*), two diploid species related to bread wheat (Fig. 1). About 60,000 genic sequences were also putatively assigned to the bread wheat A, B, or D subgenomes by using assembled Illumina (Illumina, Incorporated, San Diego, CA) sequence data for *Triticum monococcum* and *Ae. tauschii* and cDNAs from *Ae. speltoides* to guide gene assemblies of five-fold whole-genome sequence reads from *T. aestivum* 'Chinese Spring' (*18*). These resources have contributed information about the genes of hexaploid wheat and its wild diploid relatives and have underpinned the

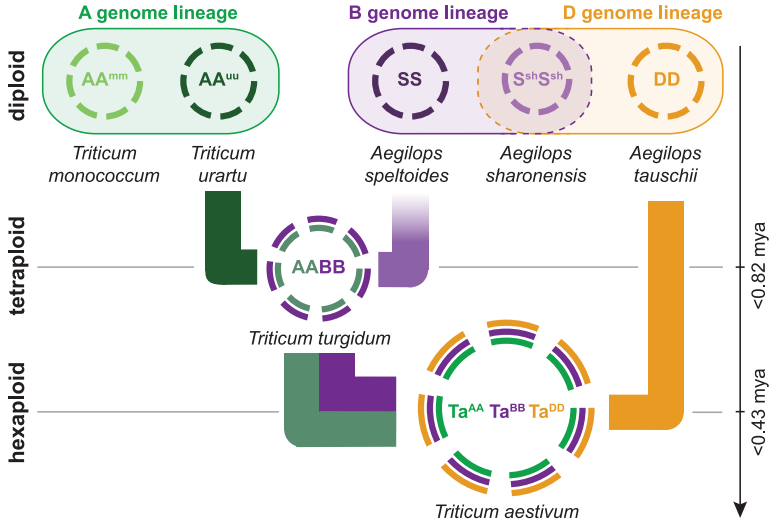


Fig. 1. Schematic diagram of the relationships between wheat genomes with polyploidization history and genealogy. Names and nomenclature for the genomes are indicated within circles that provide a schematic representation of the chromosomal complement for each species. Time estimates are from Marcussen *et al.* (*45*). mya, million years ago.

*All authors with their affiliations appear at the end of this paper.
†Corresponding author: K. X. Mayer (k.mayer@helmholtz-muenchen.de)

development of large sets of single-nucleotide polymorphism (SNP) markers (*19*–*21*). To date, however, relatively little is known about the position and distribution of genes on each of the bread wheat chromosomes and their evolution during the polyploidization events that resulted in the emergence of the hexaploid species.

Survey sequencing the bread wheat genome

We used aneuploid bread wheat lines derived from double ditelosomic stocks of the hexaploid wheat cultivar Chinese Spring (*22*) to isolate, sequence, and assemble de novo each individual chromosome arm [except for 3B, which was isolated and sequenced as a complete chromosome (*23*)]. This approach reduced the complexity of assembling a highly redundant genome and enabled the differentiation of genes present in multiple copies and highly conserved homologs. Each chromosome arm, representing between 1.3 and 3.3% of the genome (*24*), was purified by flow-cytometric sorting and sequenced to a depth of between 30× and 241× with Illumina technology platforms (*25*). The paired end sequence reads were assembled with the short-read de novo assembly tool ABySS (*25*, *26*). A high proportion of reads assembled into contigs of repetitive sequence less than 200 base pairs (bp) and were excluded from the final assembly of 10.2 Gb. The quality of the assemblies and purity of chromosome arm preparations were assessed by using alignment to bin-mapped ESTs (*15*) and to the virtual barley genome (*27*). Summary statistics for the chromosome arm assemblies are shown in Tables 1 to 3. Compared with cytogenetically estimated chromosome sizes (*24*), the sequence assemblies represent 61% of the genome sequence, with the L50 of repeat-masked assemblies ranging from 1.7 to 8.9 kb.

Repetitive DNA

We assessed the TE and sequence repeat space across the whole wheat genome and compared the repeat content of the A, B, and D subgenomes (*25*). From the frequency of mathematically defined repeats (MDRs; 20mers) (*28*), we estimated that 24 to 26% of the sequence reads contain high copy number repeats, represented by 20mers with more than 1000 copies. In total, 81% of raw reads and 76.6% of assembled sequences contained repeats, the latter showing reduced representation of Gypsy long terminal repeat (LTR) retrotransposons, as well as Mutator and Mariner-type DNA transposons.

Analysis of the distribution of transposons across the three subgenomes revealed that class I elements (retroelements) were more abundant in the A genome chromosomes relative to B or D (A > B > D), whereas class II elements (DNA transposons) showed the reverse (D > B > A). The most pronounced differences were observed between deteriorated and thus unclassifiable LTR retrotransposons, which showed a gradient of abundance across the subgenomes (A > D > B) distinct from other class I or class II elements. We hypothesize that unclassifiable LTR retrotransposons represent older (and thus more deteriorated) elements that were modified through polyploidization and ongoing TE amplification or degeneration. Assuming the amplification/degeneration dynamics are similar within each genome, the distribution of LTR retrotransposons across the three subgenomes suggest that the B genome progenitor contained a lower number of LTR retroelements and that transposon activity post-polyploidization has introduced a higher proportion of more recent amplifications into the B genome.

We observed a substantial reduction (down to 19.6%) in the TE content associated with the 0.8% (615 Mb) of the chromosomal survey sequences (CSSs) representing contigs containing high-confidence genes (for definition see below) (*25*). The analysis revealed a marked depletion of all class I elements in the neighborhood of genes, with the exception of non-LTR retrotransposons, which were enriched twofold. CACTA transposons accounted for the greatest proportion of the observed 67% reduction in class II elements, whereas minor components, especially Harbinger and miniature inverted-repeat TEs, were enriched. Selective exclusion of high-copy transposons that undergo epigenetic silencing and reduce expression by heterochromatin spreading (*29*) may result in depletion of repeat element types in the vicinity of genes.

miRNAs

A total of 270 different putative microRNA molecules (miRNAs) (49 not previously reported)

Table 1. Sequencing, assembly, and GenomeZipper statistics for wheat A genome chromosome arms. Sequence indicates the total assembled sequence (>200 bp); number of contigs is after filtering of highly repetitive sequence assemblies; syntenic loci is the number of gene loci anchored to reference gene; and the last row is the number of total anchored gene loci. Blank entries in all tables indicate data not applicable; fl-cDNA, full-length cDNA; nonred., nonredundant.

	1AS	1AL	2AS	2AL	3AS	3AL	4AS	4AL	5AS	5AL	6AS	6AL	7AS	7AL	Σ
Assembly															
Chromosome size (Mbp)	275	523	391	508	360	468	317	539	295	532	336	369	407	407	5,727
Sequence (Mbp)	178.1	250	255.2	328.2	201.8	247.2	282.3	362	198.8	318.1	219.2	214.4	198	252.4	3,505.7
Coverage (*x*-fold)	0.65	0.48	0.65	0.65	0.56	0.53	0.89	0.67	0.67	0.60	0.65	0.58	0.49	0.62	0.62
L50 (bp)	2,242	2,639	2,398	2,688	1,404	1,346	2,782	3,053	3,509	2,078	2,669	2,154	1,470	2,271	
Repeat															
No. of contigs	34,793	26,746	34,722	45,893	33,943	43,823	32,079	64,364	19,719	47,572	28,041	34,030	44,175	35,586	542,486
L50	4,769	6,369	6,678	6,677	3,846	3,789	7,499	6,601	8,713	5,355	7,091	6,589	4,397	5,849	
GenomeZipper															
No. of markers	147	380	139	278	106	332	167	200	150	309	174	286	169	278	3,115
No. of wheat fl-cDNAs	95	241	162	258	134	240	153	189	54	231	94	181	178	155	2,365
No. of nonred. contigs	937	1,750	1,673	2,499	1,323	2,300	848	2,613	574	2,495	811	1,422	2,100	1,600	22,945
No. of syntenic gene loci	544	1,515	1,155	1,816	850	1,628	842	1,642	405	1,821	647	1,073	1,228	1,049	16,215
No. of anchored gene loci	649	1,811	1,262	2,032	929	1,864	948	1,777	522	2,050	794	1,279	1,349	1,269	18,535
POP-Seq Positioning															
No. of contigs	38,940	45,649	34,853	32,941	31,094	49,586	25,068	27,248	5,578	35,333	28,234	30,828	31,628	32,435	449,415
No. of anchored gene loci	972	1,720	1,452	1,913	788	1,302	883	1,702	137	1,579	1,145	1,305	1,305	1,094	17,297
No. of anchored gene loci	618	1,257	1,408	1,903	769	1,469	778	1,116	678	2,432	995	1,458	1,405	1,711	17,997

were identified corresponding to 98,068 predicted miRNA-coding loci (25). Only 1668 loci (1.7%) evidenced expression on the basis of publicly available ESTs and of RNA sequencing (RNA-seq) data reported in this work, consistent with previous analyses in wheat (30, 31). Similarly, we observed that class II DNA transposons, specifically TcMar transposons, were predominantly found in miRNAs. For 87% of the putative miRNA-coding loci, at least one putative target gene was identified in the wheat CSS. A total of 6615 predicted miRNA-coding sequences (44 with evidence of expression) were characterized by at least one mature sequence and one target site covered by the same repeat element. This suggests that an active miRNA could arise when an advantageous regulatory niche evolves from a series of random

Table 2. Sequencing, assembly, and GenomeZipper statistics for wheat B genome chromosome arms. Sequence indicates the total assembled sequence (>200 bp); number of contigs is after filtering of highly repetitive sequence assemblies; syntenic loci is the number of gene loci anchored to reference gene; and the last row is the number of total anchored gene loci.

	1BS	1BL	2BS	2BL	3B	4BS	4BL	5BS	5BL	6BS	6BL	7BS	7BL	Σ
							Assembly							
Chromosome size (Mbp)	314	535	422	506	993	391	430	290	580	415	498	360	540	6,274
Sequence (Mbp)	212.8	299.4	292	404.5	638.6	308.2	248.7	174.5	415.2	210.2	257.4	206.1	259.6	3,927.2
Coverage (*x*-fold)	0.68	0.56	0.69	0.80	0.64	0.79	0.58	0.60	0.72	0.51	0.52	0.57	0.48	0.63
L50 (bp)	3,287	3,120	3,711	2,941	2,655	3,463	1,974	3,315	2,924	2,366	2,031	2,428	1,556	
							Repeat							
No. of contigs	26,050	29,783	35,743	75,879	75,022	38,515	46,576	18,001	75,887	29,566	35,727	24,119	58,554	569,422
L50	7,413	7,151	8,069	6,890	6,855	8,755	5,883	7,365	7,537	4,972	4,824	6,435	4,144	
							GenomeZipper							
No. of markers	78	348	278	428	500	46	145	167	404	217	245	140	198	3,194
No. of wheat fl-cDNAs	78	219	155	268	479	97	170	66	360	88	147	109	137	2,373
No. of nonred. contigs	776	1,927	1,859	3,079	5,011	893	1,634	576	3,296	915	1,525	1,172	1,890	24,553
No. of syntenic gene loci	485	1,485	1,181	1,973	3,123	788	1,155	426	2,315	565	1,003	733	1,050	16,282
No. of anchored gene loci	546	1,745	1,388	2,265	3,490	819	1,243	565	2,600	728	1,177	838	1,203	18,607
							POP-Seq Anchoring							
No. of contigs	31,038	50,219	33,603	54,522	99,341	50,927	41,135	19,794	49,140	30,962	38,064	48,514	50,397	597,656
No. of anchored gene loci	956	1,881	1,588	2,389	3,772	1,365	1,433	727	2,857	831	996	1,055	1251	21,101

Table 3. Sequencing, assembly, and GenomeZipper statistics for wheat D genome chromosome arms. Sequence indicates the total assembled sequence (>200 bp); number of contigs is after filtering of highly repetitive sequence assemblies; syntenic loci is the number of gene loci anchored to reference gene; and the last row is the number of total anchored gene loci.

	1DS	1DL	2DS	2DL	3DS	3DL	4DS	4DL	5DS	5DL	6DS	6DL	7DS	7DL	Σ
								Assembly							
Chromosome size (Mbp)	224	381	317	412	321	450	232	417	259	491	324	389	381	347	4,937
Sequence (Mbp)	128.2	254.4	166	261.6	145.4	186.5	142.1	347.6	148	236.8	156.6	199.8	209.1	222.9	2,805
Coverage (*x*-fold)	0.57	0.67	0.52	0.63	0.45	0.41	0.61	0.83	0.57	0.48	0.48	0.51	0.55	0.64	0.57
L50 (bp)	2,850	2561	1241	701	515	967	3278	1013	2,353	2,647	4,297	2,077	1,967	3,638	
								Repeat							
No. of contigs	17,725	35,770	43,044	110,446	46,795	69,259	18,245	197,398	22,449	34,622	16,077	26,236	36,701	26,737	701,504
L50	6,622	6,297	4,635	3,247	1,697	2941	7428	1855	5945	7049	8,904	6,821	5,031	7,399	
								GenomeZipper							
No. of markers	258	653	457	739	379	633	269	498	225	744	297	411	579	515	6,657
No. of wheat fl-cDNAs	89	251	177	323	128	244	130	255	99	375	103	208	200	212	2,794
No. of nonred. contigs	968	2,797	3,023	5,804	2,933	3,712	1,231	3,174	890	3,436	973	1,923	3,006	2,083	35,953
No. of syntenic gene loci	474	1,483	1,197	2,141	799	1,575	779	1,277	454	2,073	538	1,117	1,222	1,099	16,228
No. of anchored gene loci	642	1,882	1,475	2,542	1,051	1,923	912	1,582	598	2,482	758	1,347	1,592	1,423	20,209
								POP-Seq Anchoring							
No. of contigs	7,686	24,149	24,652	31,359	26,447	37,874	14,198	23,842	14,458	29,604	18,701	23,763	41,796	31,832	350,361

TE insertions and may represent a means by which a network of putative miRNAs and target genes may develop, even before miRNA activation (*32*).

Protein-coding genes

Annotation of protein-coding gene sequences in the CSS assemblies had its basis in comparisons to annotated genes in related grasses [*Brachypodium distachyon* (*33*), *Oryza sativa* (*34*), *Sorghum bicolor* (*35*), and *Hordeum vulgare* (*27*)], as well as publically available wheat full-length cDNAs (fl-cDNAs) (*11*) and RNA-seq data generated from five tissues of a Chinese Spring cultivar at three different developmental stages. Briefly, the reference grass coding sequences and wheat transcript resources were mapped separately to assembled CSS contigs, and the alignments were merged to define the exact coordinates of gene loci, alternative splicing forms, and transcripts with no similarity to related grass genes (*25*).

This analysis identified 976,962 loci with 1,265,548 distinct splicing variants. A total of 133,090 loci showing homology to related grass genes were classified as high confidence (HC) gene calls. These were further subdivided into four groups (HC1 to HC4) on the basis of the proportion of the length of the reference gene covered by a predicted locus. Of these, 124,201 (93.3%) genes were annotated on individual chromosome arm sequences, and the remaining 6.7% corresponded to wheat transcripts, which were not detected in the CSS assemblies (Fig. 2A). In total, 55,249 (44%) of the loci assigned to chromosomes were classified as HC1, that is, representing functional genes spanning at least 70% of the length of the supporting evidence (Table 4). The remaining 56% of HC genes comprised genes that were fragmented in the assembly and thus could only be partially structurally defined or were classified as gene fragments and pseudogenes. We expect that many of these will be merged as further sequencing improves the coverage and quality of genic sequences. On the basis of the level of completion of the assembly and the detection rate of HC1 genes (*25*), we estimated that the wheat genome contains 106,000 functional protein-coding genes. This supports gene number estimates ranging between 32,000 and 38,000 for each diploid subgenome in hexaploid wheat and is consistent with findings in related diploid species (*16*–*18*, *20*, *36*).

Consistent with observations of high levels of non–protein-coding loci in both plants (*27*, *37*) and animals (*38*), 890,576 loci did not share any, or only low, similarity with related grass genes. Loci with low sequence similarity (88,998) were defined as low-confidence (LC) genes, and the remainder were classified as repeat-associated, noncoding, or non–homology-supported loci (*25*). More than 96% of public wheat ESTs (HarvEST) mapped to the CSS gene sets (BLASTN; *E* value $<10^{-10}$), including 89% that correspond to HC gene-coding loci, demonstrating that the CSS assemblies contain a high representation of the current gene inventory of the bread wheat genome.

Our analysis revealed that 49% of the HC genes exhibit alternative splicing (AS) with an average of 2.6 transcripts per locus. This may be an underestimation, because 69% of the most complete gene loci (HC1) were alternatively spliced with an average of 3.5 transcripts per locus. Evidence that additional AS variants will be identified has already emerged from a preliminary assessment of gene structure prediction using proteomics analyses. In a study of 63 genes, 50 (81%) structures were confirmed, 8 (13%) provided evidence for alternative gene structures, whereas 5 were absent in the structural gene calls. Extrapolating these data to the whole genome, we estimate that hexaploid bread wheat encodes more than 300,000 distinctive protein-coding transcripts. The proportion of genes exhibiting AS appeared to be similar in all three subgenomes and is consistent with the transcriptional complexity reported for plant species such as *Arabidopsis thaliana* (*39*) and *H. vulgare* (*27*).

Gene distribution and order

Analysis of the gene distribution across the three subgenomes revealed a higher number of gene loci on the B subgenome (44,523; 35%) compared with the A and D subgenomes, which contained 40,253 (33%) and 39,425 (32%), respectively (Fig. 2A). This distribution was not consistent at the chromosomal level. For example, the gene distribution across homeologous group 3 chromosomes is 30% 3A, 42% 3B, and 28% 3D, whereas in homeologous group 7 the D genome contains the highest proportion of genes. These observations may reflect preexisting differences in the subgenomes before polyploidization or indicate that drivers determining the composition of the genome do not act at the subgenome level but regionally.

Up to 2.4-fold variation in gene density was observed on different chromosome arms, ranging from 4.4 loci per Mb (5AS) up to 10.4 loci per Mb (2DL) (Fig. 2B). Consistent with observations in rye (*40*) and the complete sequence of wheat chromosome 3B (*23*), on average 53.2% of the

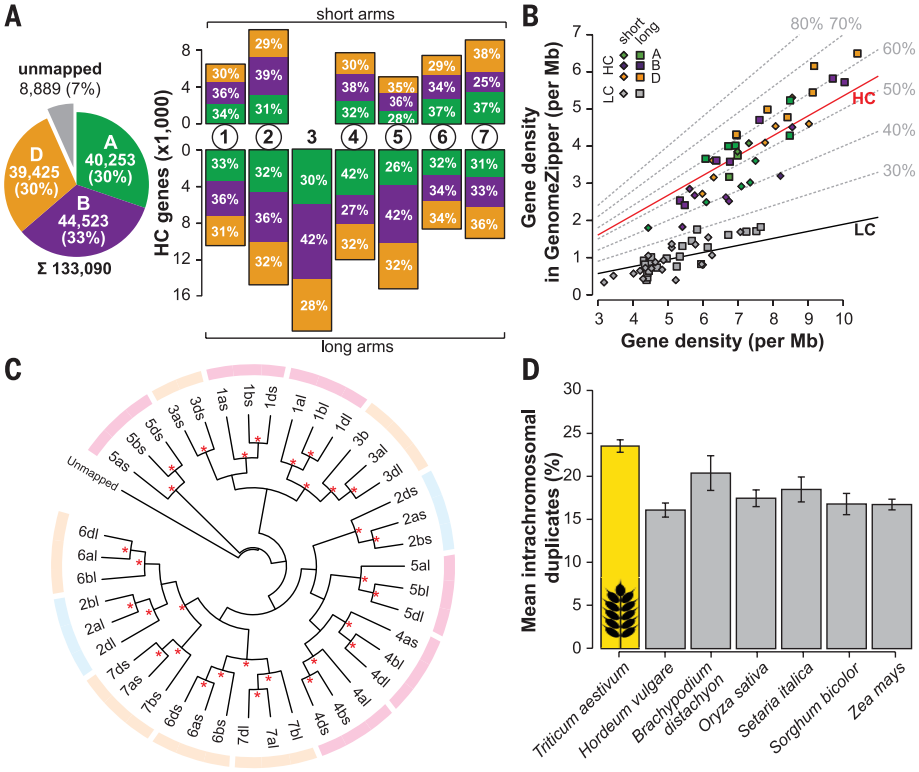


Fig. 2. Gene content, density, synteny, structural conservation, and tandemly duplicated genes. (**A**) Total number of HC bread wheat genes identified on the A (green), B (purple), and D (orange) subgenomes (left) and their distribution on individual chromosome arms or chromosomes (in the case of group 3) (right). (**B**) Syntenic conservation of HC and LC genes for each chromosome arm defined by the ratio of the number of genes anchored in the GenomeZipper and the number of annotated genes normalized per Mb of physical chromosome(-arm) size. Solid lines visualize average syntenic conservation for LC (black) and HC (red) genes, and dashed lines give isochores for different percentages of synteny. (**C**) Conservation of gene family composition between single chromosome arms. Color-coding in the outer ring indicates relatedness of the respective branches (A/D > B, light orange; A/B > D, light blue; B/D > A, light red). Red asterisks mark edges with boot-strapping values > 0.95. (**D**) Proportion of lineage-specific, intrachromosomally duplicated genes in the wheat genome compared with other grass genomes. Error bars indicate deviations among individual chromosomes.

Table 4. Characteristics of HC bread wheat genes. Distinct exons means that exons of two or more transcripts were counted once if they had identical start and stop positions; mean transcripts and mean exons are transcripts per locus and exons per locus, respectively; the second mean exons row shows exons per transcript.

	HC1	HC2	HC3	HC4	Σ
Gene loci	55,249	14,367	15,475	39,110	124,201
Single exon	9,181 (17%)	3,230 (22%)	4,906 (32%)	20,375 (52%)	37,692 (30%)
Multiple exon	46,068 (83%)	11,137 (78%)	10,569 (68%)	18,735 (48%)	86,509 (70%)
Alternatively spliced	38,059 (69%)	7,916 (55%)	6,465 (42%)	8,728 (22%)	61,168 (49%)
Mean size (bp)	3,319	2,204	1,608	901	2,216
Transcripts	194,624	37,116	31,957	61,450	325,147
Mean transcripts	3.52	2.58	2.07	1.57	2.62
Distinct exons	538,250	94,864	74,630	117,530	825,274
Mean exons	9.74	6.60	4.82	3.01	6.64
Mean exons^3	6.29	4.45	3.52	2.56	5.1
Mean size (bp)	321	315	314	281	314

HC genes were located on syntenic chromosomes compared to *B. distachyon* (Bd), *O. sativa* (Os), and *S. bicolor* (Sb). The average level of synteny for genes located on the D genome chromosomes (58%) was higher than the average for those on the A (51%) and the B (50%) chromosomes. Sequence conservation in LC genes is low, and, in comparison to HC genes, reduced syntenic conservation is observed. Thus, although the majority of LC genes are likely to result from the frequent generation of gene fragments by double-strand repair mechanisms or are deteriorated (pseudo)genes that were fragmented after the divergence from the other sequenced grass genomes (*10*), the retained synteny to other grass genomes suggests that some LC genes may be functional.

To determine the extent of gene conservation across homeologous chromosomes, we clustered the HC genes into protein families by sequence similarity (Fig. 2C) (*25*). With the exception of chromosome 4AL, the genes on all chromosome arms clustered with their corresponding homologs. The pattern of clustering observed for 4A is consistent with a known pericentromeric inversion and two translocations of segments from chromosome arms 5AL and 7BS (*41*, *42*). All possible cluster topologies were found between genes on the A, B, and D genomes. Overall, the patterns of conservation suggest that the gene content of the A and B homeologous chromosomes is more similar to the D genome chromosomes than to each other. This observation contradicts a model of bifurcating evolutionary relationships between the A, B, and D genomes but is consistent with models of interlineage hybridization (i.e., reticulate evolution) in the Triticeae (*43*, *44*) and corroborate phylogenomic analyses that suggest that the D genome is a product of homoploid hybrid speciation between A and B genome ancestors >5 million years ago (*45*). Although the potential for preexisting differences needs to be considered, the preservation of gene copies in each of the A, B, and D genomes provides evidence for their structural autonomy, a likely consequence of independent pairing during meiosis (*46*). A high degree of subgenome autonomy was also reflected in the observed patterns of gene expression (see below).

We used two independent but complementary approaches to generate an order for the many small contigs that comprise the chromosome arm assemblies (*25*). The GenomeZipper approach (*47*) combines the syntenic conservation of gene order in grasses (*48*) and the known gene orders of fully sequenced grass genomes (*33*–*35*) with high-density SNP-based genetic maps (*21*, *49*) to create a virtual gene order in wheat. The number of genes anchored per chromosome (chr.) ranged from 2125 (chr. 6B) to 4404 (chr. 2D) (Table 1). Overall, the GenomeZipper inferred positions of 21,221, 22,051, and 22,813 genes, respectively, in the A, B, and D genomes. To complement this, the POPSEQ approach (*50*) was used to build an ultradense genetic map comprising 13.3 million SNPs identified after shallow-coverage whole-genome sequencing of 90 doubled haploid individuals of the synthetic W7984 × Opata M85 population (*51*). This map assigned a partially overlapping set of 17,297, 21,101, and 17,997 HC genes, respectively, to the individual chromosomes of the A, B, and D genomes. The POPSEQ genetic map showed concordance with the gene assignments to flow-sorted chromosomes (99.4%) and the GenomeZipper (99.8%). The two inferred gene orders along chromosomes were also largely collinear (Spearman's correlation coefficient = 0.85). From both anchored data sets, we were able to position a nonredundant set of 75,183 HC genes on the 21 chromosomes of bread wheat by genetic mapping and/or syntenic conservation.

Gene duplication is frequently observed in plant genomes, arising from polyploidization or through tandem or segmental duplication associated with replication (*52*). For each wheat chromosome, the percentage of genes that have undergone lineage-specific intrachromosomal duplication was determined with OrthoMCL (*53*). By using the HC1 genes, we estimated that between 19.1% (chr. 7B) and 29.7% (chr. 2B) (23.6% average for all chromosomes) of the genes are duplicated on each chromosome (*25*). Comparison of the number of duplicated genes identified by this analysis for chr. 3B (25.3% of HC1 genes) with the 3B reference pseudomolecule (37% duplicated genes) (*23*) indicated that we are likely underestimating the number of duplicated genes. This is due to the fragmented nature of the assemblies obtained from whole-genome or chromosome-shotgun sequences that collapse highly conserved duplicates. No significant differences in the proportion of duplicates were observed between the three subgenomes (χ^2 test, $\chi^2 = 3.8$, $P = 0.15$).

For each chromosome, an average of 73% of the duplicates are located on one of the chromosome arms, suggesting that they may be tandem duplicates that arise through unequal crossing-over and replication-dependent chromosome breakage (*54*) or through the activity of transposable elements. When compared with the percentage of intrachromosomal duplicates found in rice, sorghum, barley, maize, and foxtail millet (17 to 20%) (*27*, *33*–*35*, *55*, *56*), the proportion of gene duplications in wheat was significantly higher (Fig. 2D; Tukey's honest significant difference, pairwise $P < 0.007$).

Comparisons with related species

We assembled sequence data from seven species related to progenitors of the bread wheat A, B, and D subgenomes (*25*). Illumina whole-genome sequence data and assemblies were generated from two tetraploid wheat cultivars (AABB) *T. turgidum* 'Cappelli' (originating from Italy) and *T. turgidum* 'Strongfield' (originating from Canada) as well as from the diploid genome of *Ae. speltoides* (SS). These data were combined with whole-genome sequence data from *T. urartu* (AA^{uu}) (*16*), *T. monococcum* (AA^{mm}), *Ae. tauschii* (DD) (*17*), and *Aegilops sharonensis* ($S^{sh}S^{sh}$). For the unannotated genomes of *T. turgidum*, *T. monococcum*, *Ae. speltoides*, and *Ae. sharonensis*, proteins of annotated grass genomes (*27*, *33*, *35*, *57*) and *T. aestivum* gene models were projected on the sequence assemblies.

Genes and gene families in the hexaploid, tetraploid, and diploid genomes were then compared to assess the dynamics of gene retention or loss after polyploidization and to define the core wheat genes. When comparing the sizes of gene families in *Ae. tauschii* (*17*) and *T. urartu* (*16*) diploid genomes with the individual subgenomes of hexaploid wheat (Fig. 3, A and B), we found that gene loss mainly affected genes belonging to expanded families, consistent with previous observations (*18*). In contrast, singletons (i.e., genes without paralogous copies within the same genome) were not usually subject to gene loss after polyploidization. Pronounced variations of gene copy retention or loss patterns were observed depending on the gene family considered.

Highly similar gene retention rates were found for all bread wheat subgenomes in comparison to *Ae. tauschii* and *T. urartu* [0.91 (A), 0.94 (B), and 0.89 (D) versus *Ae. tauschii* and 0.91 (A), 0.96 (B), and 0.91 (D) versus *T. urartu*] (Fig. 3, A and B). The extent of gene loss in the D subgenome, the most recent addition to the hexaploid genome, appeared slightly lower than the more ancient A and B subgenomes. Thus, as observed for

the gene content and structural similarities between individual chromosome arms, we found no evidence for a gradual gene loss induced by polyploidization. This may indicate that gene loss occurred rapidly after polyploid formation, followed by stabilization of gene content consistent with observations in newly created polyploids (*58*, *59*) and gene retention in cotton (*60*).

We conducted a clustering analysis of gene families and determined the number of genes in the bread wheat subgenomes that have an ortholog in the genomes from the A genome lineage (*T. urartu* and *T. monococcum*), the closest known relatives for the B lineage (*Ae. sharonensis* and *Ae. speltoides*), the D lineage (*Ae. tauschii*), as well as in the tetraploid *T. turgidum* genome (Fig. 3C). We found that the A, B, and D subgenomes contain very similar proportions of genes (60.1 to 61.3%) with orthologs in all the related diploid genomes. We also estimated the contribution of unique genes of the three subgenomes to the bread wheat genome. Because the absence of a particular gene in a single species could be due to incomplete sequence coverage or assembly errors, only lineage-specific gene family absence was considered in the analysis. Only a small fraction of the genes (1.3 to 1.7%) were specific to the A, B, or D lineages, demarcating the likely upper estimate of unique genes or gene families added to the bread wheat gene complement by the individual subgenomes.

High sequence similarity between genes in the bread wheat subgenomes impedes efficient marker development and the identification of nonsynonymous sequence variations that can potentially affect gene or protein functionality. We delineated single-nucleotide variations (SNVs) between the bread wheat genes and the diploid and tetraploid related genomes and reconstructed phylogenetic relationships by using unrooted parsimony (Fig. 4A) (*25*). In total, 11,435 SNVs within 6498 genes were specific to bread wheat and thus have likely been introduced after the second polyploidization event. Although most relationships support the known phylogeny of wheat, *Ae. sharonensis* was placed closer to the bread wheat D subgenome and *Ae. tauschii* than to *Ae. speltoides* and the B genome branch. This suggests that the Sitopsis group, which includes *Ae. sharonensis* and *Ae. speltoides*, is deeply furcated and related to both D and B genome branches.

The potential impact of all SNVs detected on proteins was measured by using Grantham amino acid substitution matrix scores (*25*, *61*). Most of the substitutions (80.8%) in gene sequences were conservative or moderately conservative and were randomly distributed across all chromosomes. However, bread wheat genes contained a higher proportion of substitutions with a predicted large impact on the protein functionality (i.e., moderately radical and radical changes) compared with their closest diploid or tetraploid relatives. This points to gene redundancy in hexaploid bread wheat enabling accelerated sequence evolution and potentially the evolution of novel protein functions.

We used the bread wheat gene annotation to analyze the introduction of likely premature stop codons in diploid and tetraploid related genomes as a measure for the rate and degree of pseudogenization (Fig. 4B). Using only the highest confidence genes (HC1), 290 (1.6%; *T. turgidum* A genome versus *T. aestivum* A genome) to 636 (3.6%; *Ae. sharonensis* versus *T. aestivum* D genome) gene loci had characteristics of pseudogenization in the respective related diploid genomes compared with the respective bread wheat A, B, and D subgenomes. Most of these likely pseudogenized loci were specific to the respective genomes, although overlapping candidate pseudogenized loci were also observed. However, the numbers of genes in these categories were small, ranging from 0.1 to 0.7%. Similar inferred pseudogenization rates were found in the A and B subgenomes of *T. turgidum* [290 (1.6%) in the A genome and 395 (2.0%) in the B genome, respectively], indicating no preferential pseudogenization or gene loss in any of the subgenomes. The number of pseudogenes observed in the D genome was similar to that of the A and B subgenomes and their diploid relatives, suggesting a rapid elimination process for pseudogenes. These findings are consistent with those from other plants, notably among *Arabidopsis* ecotypes (*62*), and smaller-scale analysis of pseudogenization dynamics within the bread wheat genome (*63*).

Earlier studies showed a high degree of gene sequence similarity between A, B, and D bread wheat subgenomes and their related diploid species (*6*). We analyzed the sequence conservation in bread wheat chromosomes compared to their diploid and tetraploid relatives to test for intergenomic translocations or introgressions (Fig. 4C). The sequences of genes were highly conserved, exceeding 99% identity, between the hexaploid subgenomes and their respective diploid relatives. High levels of conservation, averaging 97%, were also found between the A, B, and D lineages.

No gradients in sequence conservation were apparent along the chromosomes for the most closely related genomes. However, when comparing more distant genomes (e.g., *T. aestivum* D genome versus *T. urartu*), higher levels of sequence conservation were observed in genes located in proximal, pericentromeric, and centromeric regions. These results are consistent with findings for the 3B pseudomolecule analysis that demonstrated a partitioning of the chromosome with variable telomeric regions and a more conserved central chromosomal region (*23*). The most pronounced deviation in gene sequence similarity from the overall distribution is found for chr. 4A, which has undergone a recent inversion and translocations from chrs. 5A and 7B (*41*, *42*) (Fig. 4C). Other, smaller regions showing altered similarity profiles were also observed on other chromosomes (e.g., chrs. 2A and 7B) (*25*) suggesting the presence of further small translocations or introgressions that may have occurred after hybridization.

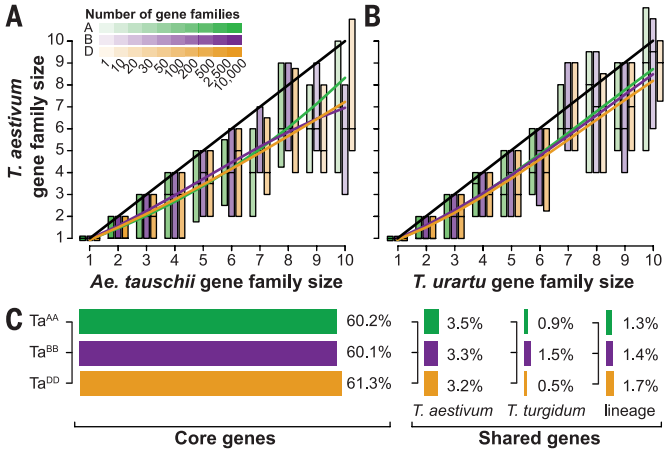


Fig. 3. Gene conservation and the wheat pan- and core genes. (**A** and **B**) Relationship between gene family sizes in diploid *Ae. tauschii* (A) and *T. urartu* (B) and each subgenome of hexaploid bread wheat (colors as in Fig. 2A). Boxes visualize the lower and upper quartiles of gene family sizes. Color intensity indicates the number of gene families in the respective bin. The black line shows a 1:1 gene copy number relationship for bread wheat, *Ae. tauschii*, and *T. urartu*, and colored lines show the regression fit for observed gene family size in the wheat subgenomes. (**C**) Percentages of genes of the bread wheat subgenomes that show significant sequence similarity to other genomes: Core genes correspond to genes with hits to all subgenomes as well as to *T. turgidum* and all diploid related progenitor genomes; shared genes–*T. aestivum* are genes with hits to any other *T. aestivum* subgenome but not to *T. turgidum* or any of the closest diploid relatives; shared genes–*T. turgidum* correspond to genes with hits to *T. turgidum* but not to any of the closest diploid relatives; shared genes–lineage, with hits to the subgenome's closest relative genome but not to *T. turgidum* or any of the other closest related genomes.

Hexaploid genome phylogeny

To further test the relatedness of the A, B, and D subgenomes across the entire wheat genome, we used syntenic gene alignments to estimate maximum likelihood phylogenetic trees. We obtained 2269 trees and analyzed them for topological variation. Across all chromosome groups, 40, 35, and 25% of the gene phylogenies supported AD,

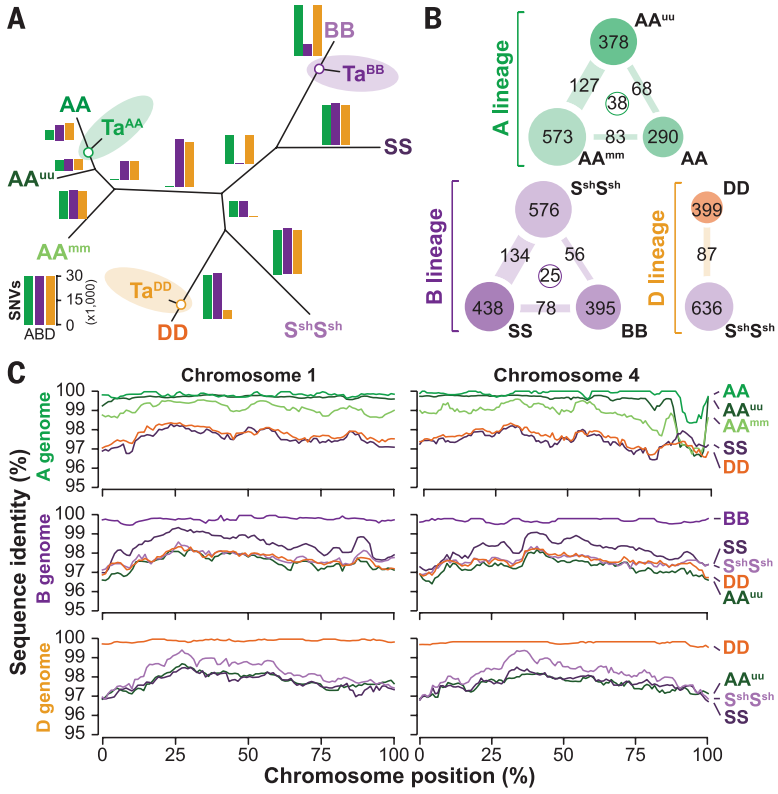


Fig. 4. Molecular evolution of the wheat lineage. SNVs were identified for coding sequences of bread wheat genes (TaAA, TaBB, and TaDD) against diploid *T. monococcum* (AA^{mm}), *T. urartu* (AA^{uu}), *Ae. speltoides* (SS), *Ae. sharonensis* (SshSsh), *Ae. tauschii* (DD), and tetraploid *T. turgidum* (AABB). (**A**) Unrooted phylogeny constructed on the basis of SNVs between bread wheat and its diploid or tetraploid relatives. The respective number of SNVs in each phylogenetic internodes is indicated with bar charts (scale at bottom left corner); colors indicate the respective bread wheat subgenome as in Fig. 2A. (**B**) Genes with stop codons in the respective related diploid genomes in comparison to the bread wheat A, B, and D subgenomes. Numbers in node connectors or in the center correspond to the number of introduced stop codons found in two (node connectors) or all (center) related genomes. (**C**) Chromosomal distribution of sequence identity between bread wheat genes and the diploid and tetraploid relatives for homeologous chromosomes.

BD, and AB as the closest pairs, respectively. This genome-wide observation supports previous findings of discordant phylogenetic signals within *Aegilops* and *Triticum* genera (*6*, *43*, *45*). Some variation in genome relationships was found among chromosomes: On group 4 chromosomes, most gene trees supported BD as closest pairs, whereas group 5 chromosomes had similar numbers of AD and BD topologies (AD = BD > AB). Distribution of variation in phylogenetic signals across homeologous chromosomes can help to better understand the nature of the evolutionary processes underlying such phylogenetic incongruence. Under incomplete lineage sorting and stochastic coalescence, levels of phylogenetic incongruence will be correlated with recombination rates, whereas single introgression events and limited recombination are expected to generate local chromosome blocks of homogenous phylogenetic signals. We used the inferred gene orders from the GenomeZipper to test for nonrandom distribution of phylogenetic signals along chromosomes. We were unable to consistently identify block structures larger than would be expected by chance. However, it is possible that the limitations of the inferred gene order hamper the ability to detect such patterns.

Gene expression

Our study did not reveal any pronounced bias in gene content, structure, or composition between the different wheat subgenomes. In paleopolyploid maize and soybean, transcriptional dominance of genes derived from one progenitor genome has been described (*64*–*66*). Previous analyses have shown that rapid initiation of differential expression of homeologous wheat genes occurs upon polyploidization with a predominantly additive mode (*13*, *67*). Sets of homeologous wheat genes with only one copy present in each of the subgenomes (triads) were used to test for differential expression at a genome-wide scale. Expression correlations were calculated for 6219 triads (18,657 genes) by using RNA-seq data from five organs (leaf, root, grain, spike, and stem) (Fig. 5A) (*25*). Whereas root-derived expression clustered separately, genes expressed in stem, leaves, grain, and spike clustered in a subgenome-specific manner. This indicates that the individual subgenomes exhibit a high degree of regulatory and transcriptional autonomy, with limited trans (inter-subgenome) regulation (*68*). At a global level, the overall pairwise expression correlation between subgenomes was very similar (Fig. 5B), and no evidence for genome-wide transcriptional dominance of an individual subgenome was observed.

By using hierarchical cluster analysis, we aggregated expressed genes into 13 distinct groups. These groups show predominant expression in particular organs (e.g., groups III and XIII in Fig. 5A) or in one of the subgenomes (e.g., groups II, IX, and X in Fig. 5A). Pairwise comparisons of individual expressed homeologous genes in the groups revealed abundant transcriptional dominance from specific subgenomes (Fig. 5B). Overall, 1333 (21%) of the homeologous gene triads showed an expression bias in one of the pairwise comparisons, and we detected a similar number of preferentially transcribed genes (378 to 393) in each subgenome (permutation test; $P < 0.05$). For the individual transcriptional groups, however, between 2% (groups I, IV, and V) and 20% (groups II and VI) of the genes were found to be transcriptionally dominant.

These patterns of gene expression across the three genomes contrast with patterns of gene expression reported in allopolyploid cotton (*69*, *70*); mesopolyploid *Brassica rapa* (*71*); synthetic allotetraploid *Arabidopsis* (*72*); and the paleopolyploid maize genome (*64*), where one of the genomes is more transcriptionally active than others. The apparent autonomy of the three wheat subgenomes may be explained by the relatively recent polyploidization. It may also be related to regulatory mechanisms that control the transcriptional interplay of homeologous genomes to balance expression of individual and groups of genes. While maintaining subgenome-specific expression profiles, a high degree of orchestration and functional partitioning between homeologous genes was also reported in grain development of bread wheat (*68*) and has been attributed to the rapid evolution of cis elements coupled to epigenetic mechanisms controlling gene expression (*68*, *73*, *74*).

Gene family size variation

The relationship between genes important to wheat adaptation, disease resistance, and end-use functionality in hexaploid wheat and its diploid relatives was examined for signs of adaptive evolution. These analyses identified three distinct patterns: gene expansion, gene loss, or independent gene evolution that may or may not include expansion or loss. In some cases, such as the genes containing a NB-ARC domain characteristic of many plant disease-resistance genes (*75*), we observed an expansion within a single subgenome (Fig. 6A). Indeed, a substantial expansion in *Ae. tauschii*, compared with the other diploid species and the D genome of hexaploid wheat, is consistent with the rich reservoir of disease-resistance genes known in this species

(*17*). In genes coding for the cysteine-rich gliadin domain, a functional domain characteristic of storage proteins, we observed a similar number of genes in all diploid genomes (except *T. monococcum*) that is higher than the number of genes found in each of the three hexaploid wheat subgenomes (Fig. 6B). This may indicate that gene loss occurred in hexaploid wheat and that there is a trend for the gliadin gene family to maintain some homeostasis with a similar global number of genes in polyploid and diploid wheat. In other cases, the patterns observed suggested independent evolution of gene families within the different genomes and subgenomes of wheat. This was seen for genes associated with abiotic stress tolerance. For example, for genes encoding the Apetala2 (AP2) DNA binding domain, associated with drought, heat, salinity, and cold stress–tolerance responses, we observed fewer AP2 genes in the A and D genomes of Chinese Spring compared with the diploid relatives or the B subgenome (Fig. 6C). Likewise, genes coding for MYB transcription factors, which have also been involved in abiotic stress response in plants (*76*), were underrepresented in the A subgenome of hexaploid wheat and *T. monococcum*, whereas a higher frequency was observed in *Ae. tauschii* (*17*) and *T. urartu* (*16*) (Fig. 6D).

In contrast, there was no evidence of expansion or loss of genes underlying phenology, such as the vernalization (*Vrn1*) and photoperiod response regulator (*Ppd1*) genes that differentiate spring and winter growth habits and sensitivity to day length, respectively. Similar numbers of genes were found in the diploids and hexaploid subgenomes coding for the two functional domains of *Vrn1*, a MADS-box and K-box domain (*77*) (Fig. 6E), and for genes containing the response regulator domain and CCT motif typical of cereal *Ppd* genes (*78*) (Fig. 6F). We identified an additional copy of a *Vrn1*-like gene in the hexaploid Chinese Spring A and D genomes and *T. urartu* (*16*) when compared with the remaining diploid species. An additional copy of a *Ppd1*-like gene was also identified in the Chinese Spring B genome relative to *Ae. sharonesis* and *Ae. speltoides* (Fig. 6F). Although only small differences were observed, small increases in copy number variation of *Vrn-A1* (A genome) and *Ppd-B1* (B genome) have been associated with longer periods of vernalization to potentiate flowering and an early flowering day neutral phenotype, respectively (*79*). Thus, the relative distribution of such patterns in ontology of these two genes is likely to reflect important factors that have allowed wheat to adjust its flowering time to adapt to a range of environmental conditions.

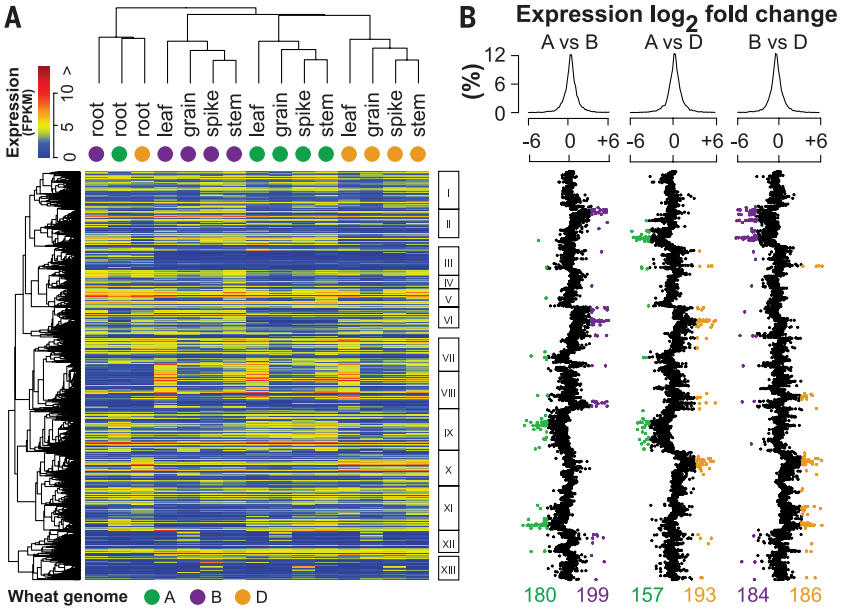


Fig. 5. Subgenome transcriptional profiling for individual wheat tissues. (**A**) Two-dimensional hierarchical cluster analysis of single-copy wheat homeologous gene expression (colors as in Fig. 2A) compared with organ-specific gene expression. (**B**) Analysis of $\log_2$-fold changes in pairwise gene expression between homeologous genes (averaged across organs). Top graphs depict the distributions of $\log_2$ fold changes. Dot plots show the fold changes for each triplet ordered as shown in the *y* axis in (A). Colored dots highlight homologs that show significant differential expression ($P < 0.05$). The numbers of differentially expressed triplets across all organs are shown at the bottom of the figure.

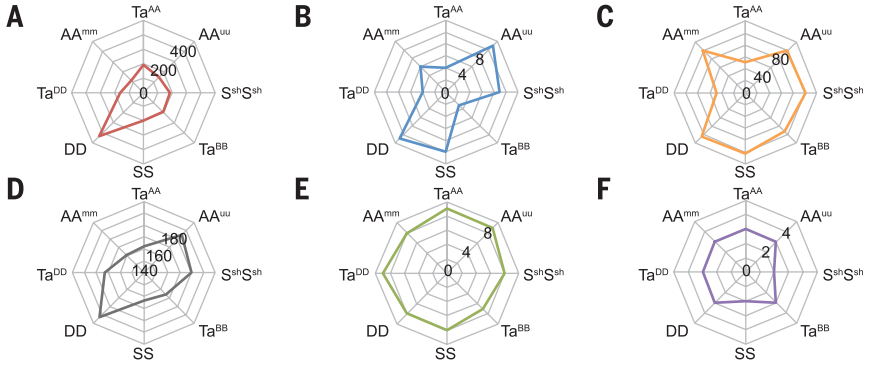


Fig. 6. Sizes of selected gene families and protein domains among hexaploid wheat and diploid relatives. (**A**) NB-ARC domain, (**B**) cysteine-rich gliadin domain, (**C**) AP2 domain, (**D**) MYB domain, (**E**) *Vrn1* (MADS-box/K-box domain), and (**F**) *Ppd* (photoperiod response regulator/CCT domain).

Molecular markers

Wheat improvement relies in part on the use of molecular markers to improve selection efficiencies and to allow the precise transfer of genes and QTL between different genetic backgrounds. To enhance the CSS as a genomic resource for the wheat genetics and breeding community, we anchored all publicly available DNA markers that are routinely used for genetic mapping and marker-assisted breeding in wheat. Because the majority of these markers are anchored to phenotypic maps, anchoring them to the CSS allows immediate association of CSS to traits targeted by breeders. In addition, insertion site–based polymorphism (ISBP) and SNP markers identified from recent whole-genome shotgun and transcriptome sequencing (*19*) and genotyping by sequencing (GBS) tags identified by using DArTSeq (Diversity Arrays Technology, Bruce, Australia) technology were also anchored. In total, over 3.6 million marker loci were anchored to the CSS, including 1,347,669 marker loci and 2,310,988 SNPs (Table 5).

Most marker types showed a distribution gradient across subgenomes, with the highest number associated with the B genome chromosomes and the lowest with the D genome, reflecting the differences in the level of polymorphism in these subgenomes. The proportions of ISBPs, SNPs detected from cultivar sequencing and GBS tags localized to the D genome ranged between 9.3 and 12%, with the lowest numbers mapping to the group 4 chromosomes (Table 5). Two hundred and ninety-two of 1867 simple sequence repeat (SSR) loci were successfully anchored to the CSS survey sequence. This low number is not surprising, given that these loci derive from repetitive AT- and GC-rich sequences that may be collapsed or

Table 5. Number and type of molecular markers mapped on individual chromosomes of the bread wheat genome.

	Bin mapped ESTs	EST-SSRs	Genomic SSRs	DArT Probes	Cereals DB	90K iSelect SNPs (87)	DArT Seq	ISBPs	Genic SNPs	Intergenic SNPs	Σ
Queries	18,771	2,926	1,867	7,552	7,228	81,987	29,375	Derived from cultivar sequencing			-
Mapped queries	16,876	2,435	282	5,228	5,136	80,820	18,515				
1A	1,325	156	8	414	479	13,093	1,371	68,074	13,980	127,663	226,563
2A	1,614	257	28	356	544	17,502	1,378	84,440	18,349	148,204	272,672
3A	1,136	75	14	252	302	12,172	1,008	44,740	10,770	94,975	165,444
4A	1,766	266	27	331	357	14,043	1,530	39,483	10,367	86,543	154,713
5A	1,189	155	46	256	343	13,099	893	62,193	12,624	115,085	205,883
6A	1,150	132	63	418	421	12,072	1,127	60,169	15,884	110,850	202,286
7A	1,240	146	120	321	326	13,168	1,474	71,597	15,516	154,748	258,656
Σ A genome	9,420	1,187	306	2,348	2,772	95,149	8,781	430,696	97,490	838,068	1,486,217
1B	1,379	226	15	378	618	13,776	1,846	66,994	14,447	131,682	231,361
2B	1,810	367	39	466	606	18,352	2,557	90,852	23,958	162,335	301,342
3B	1,845	188	29	406	444	14,471	2,294	108,810	22,032	208,306	358,825
4B	1,401	188	42	278	294	11,019	856	36,937	7,506	59,175	117,696
5B	1,911	343	86	399	527	17,087	2,112	84,179	21,389	159,359	287,392
6B	978	43	139	320	313	12,448	1,171	65,982	11,974	130,463	223,831
7B	999	107	151	270	205	11,635	1,123	72,307	10,997	136,932	234,726
Σ B genome	10,323	1,462	501	2,517	3,007	98,788	11,959	526,061	112,303	988,252	1,755,173
1D	1,165	149	13	378	380	12,093	660	17,366	5,004	36,457	73,665
2D	1,309	199	22	414	331	16,978	609	19,532	6,745	34,967	81,106
3D	854	104	14	428	151	11,699	420	10,920	1,403	18,078	44,071
4D	1,221	239	27	245	196	10,198	307	10,097	1,108	13,249	36,887
5D	1,584	408	78	400	289	13,308	488	13,629	3,582	22,957	56,723
6D	1,132	91	135	289	240	10,504	417	12,042	3,609	23,341	51,800
7D	1,461	230	139	862	243	12,826	767	18,174	3,969	34,344	73,015
Σ D genome	8,726	1,420	428	3,016	1,830	87,606	3,668	101,760	25,420	183,393	417,267
Σ	28,469	4,069	1,235	7,881	7,609	281,543	24,408	1,058,517	235,213	2,009,713	3,658,657

represented by uneven read coverage in Illumina sequences (*80*).

Well over 70 DNA markers are routinely deployed by breeders for agronomic, pest resistance, and end-use quality, and most are available in the public domain (http://maswheat.ucdavis.edu). Anchoring of these to the CSS would facilitate identification of SNP markers for development of high-density marker maps, as a resource of correlated markers, and to aid map-based cloning of genes underlying important traits. In total, we anchored 68 of these markers to 74 contigs in the CSS. The application of the CSS in marker improvement was demonstrated with the CAPS (cleaved amplified polymorphic sequence) marker *Usw47*, which is linked to *Cdu-B1*, a gene responsible for reduced grain cadmium content in tetraploid wheat (*81*, *82*). Although *Usw47* is routinely used in marker-assisted selection, it is not amenable to high-throughput genotyping. Alignment of the *Usw47* sequence against the CSS mapped it to contig 5BL-10759151. This and eight neighboring contigs in the GenomeZipper contained 33 SNP markers, of which 5 were polymorphic in a doubled haploid mapping population used previously to localize *Cdu-B1*. Of the five SNP markers, two co-segregated, and the remainder flanked the gene by a single recombination event. These SNP markers can be readily implemented now in a high-throughput fashion to select for reduced grain cadmium content within breeding programs.

Conclusion

We present the ordered and structured draft sequence of the bread wheat genome as well as a comparison between eight related wheat genomes. We defined a gene catalog for each of the 21 bread wheat chromosomes and positioned more than 75,000 genes along the chromosomes by using a combination of high-density wheat SNP mapping and synteny to sequenced grass genomes. In contrast to other species (*83*), polyploidization events in wheat did not cause a "genome shock" with subsequent rapid genome changes or functional dominance of one subgenome over the others. Intraspecific comparative analyses revealed a dynamic wheat genome with a high level of plasticity and a changing gene repertoire shaped by gene losses and gene-family expansions in all wheat genomes and subgenomes, with only a few species-specific genes. Through interspecific comparisons, we observed a higher abundance of intrachromosomal gene duplications in wheat compared with other grass genomes, which may be a mechanism for functional adaptation and underlie the global success of wheat as a cultivated crop.

The detection, chromosomal assignment, and description of a large proportion of the gene complement of bread wheat and their positional assignment on chromosome arms is a major milestone in facilitating the isolation of genes underlying agronomically important traits, providing a reference for future integration into systems biology, and improving wheat breeding efficiency. Already, the resources developed in this work have been used to support the analysis of selected wheat chromosomes (*20*, *41*, *84*–*86*). Last, as demonstrated by the completion of the reference sequence for chr. 3B (*23*), this draft genome sequence and complementary resources will support the assembly and annotation of the physical map–based reference sequences for the 21 bread wheat chromosomes.

REFERENCES AND NOTES

1. D. B. Lobell, W. Schlenker, J. Costa-Roberts, Climate trends and global crop production since 1980. *Science* **333**, 616–620 (2011). doi: 10.1126/science.1204531; pmid: 21551030
2. Food and Agriculture Organization (FAO) of the United Nations, FAO cereal supply and demand brief (2013); www.fao.org/worldfoodsituation/csdb/en/.
3. D. Tilman, K. G. Cassman, P. A. Matson, R. Naylor, S. Polasky, Agricultural sustainability and intensive production practices. *Nature* **418**, 671–677 (2002). doi: 10.1038/nature01014; pmid: 12167873
4. J. A. Foley *et al.*, Solutions for a cultivated planet. *Nature* **478**, 337–342 (2011). doi: 10.1038/nature10452; pmid: 21993620
5. Organisation for Economic Cooperation and Development (OECD)/FAO, *OECD-FAO Agricultural Outlook 2013* (OECD, Paris, 2013); doi: 10.1787/agr_outlook-2013-en.
6. G. Petersen, O. Seberg, M. Yde, K. Berthelsen, Phylogenetic relationships of *Triticum* and *Aegilops* and evidence for the

origin of the A, B, and D genomes of common wheat (*Triticum aestivum*). *Mol. Phylogenet. Evol.* **39**, 70–82 (2006). doi: 10.1016/j.ympev.2006.01.023; pmid: 16504543
7. M. Nesbitt, D. Samuel, "From staple crop to extinction? The archaeology and history of the hulled wheats," in *Hulled Wheat: Proceedings of the First International Workshop on Hulled Wheats*, S. Padulosi, K. Hammer, J. Heller, Eds. (International Plant Genetic Resources Institute, Rome, 1995), pp. 41–102.
8. E. Martinez-Perez, P. Shaw, G. Moore, The Ph1 locus is needed to ensure specific somatic and meiotic centromere association. *Nature* **411**, 204–207 (2001). doi: 10.1038/35075597; pmid: 11346798
9. T. Eilam *et al*., Genome size and genome evolution in diploid Triticeae species. *Genome* **50**, 1029–1037 (2007). doi: 10.1139/G07-083; pmid: 18059548
10. T. Wicker *et al*., Frequent gene movement and pseudogene evolution is common to the large and complex genomes of wheat, barley, and their relatives. *Plant Cell* **23**, 1706–1718 (2011). doi: 10.1105/tpc.111.086629; pmid: 21622801
11. 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; pmid: 19448038
12. A. N. Bernardo *et al*., Discovery and mapping of single feature polymorphisms in wheat using Affymetrix arrays. *BMC Genomics* **10**, 251 (2009). doi: 10.1186/1471-2164-10-251; pmid: 19480702
13. H. Chelaifa *et al*., Prevalence of gene expression additivity in genetically stable wheat allohexaploids. *New Phytol.* **197**, 730–736 (2013). doi: 10.1111/nph.12108; pmid: 23278496
14. T. E. Coram, M. L. Settles, M. Wang, X. Chen, Surveying expression level polymorphism and single-feature polymorphism in near-isogenic wheat lines differing for the Yr5 stripe rust resistance locus. *Theor. Appl. Genet.* **117**, 401–411 (2008). doi: 10.1007/s00122-008-0784-5; pmid: 18470504
15. L. L. Qi *et al*., A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat. *Genetics* **168**, 701–712 (2004). doi: 10.1534/genetics.104.034868; pmid: 15514046
16. H. Q. Ling *et al*., Draft genome of the wheat A-genome progenitor *Triticum urartu*. *Nature* **496**, 87–90 (2013). doi: 10.1038/nature11997; pmid: 23535596
17. J. Jia *et al*., *Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature* **496**, 91–95 (2013). doi: 10.1038/nature12028; pmid: 23535592
18. R. Brenchley *et al*., Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* **491**, 705–710 (2012). doi: 10.1038/nature11650; pmid: 23192148
19. A. M. Allen *et al*., Discovery and development of exome-based, co-dominant single nucleotide polymorphism markers in hexaploid wheat (*Triticum aestivum* L.). *Plant Biotechnol. J.* **11**, 279–295 (2013). doi: 10.1111/pbi.12009; pmid: 23279710
20. K. V. Krasileva *et al*., Separating homeologs by phasing in the tetraploid wheat transcriptome. *Genome Biol.* **14**, R66 (2013). doi: 10.1186/gb-2013-14-6-r66; pmid: 23800085
21. C. Saintenac, D. Jiang, S. Wang, E. Akhunov, Sequence-based mapping of the polyploid wheat genome. *G3* **3**, 1105–1114 (2013).
22. E. Sears, L. Sears, "The telocentric chromosomes of common wheat," in *Proceedings 5th International Wheat Genetics Symposium*, S. Ramanujam, Ed. (Indian Agricultural Research Institute, New Delhi, 1978) vol. 1, pp. 389–407.
23. F. Choulet *et al*., A reference sequence of wheat chromosome 3B reveals structural and functional compartmentalization. *Science* **345**, 1249721 (2014).
24. J. Šafář *et al*., Development of chromosome-specific BAC resources for genomics of bread wheat. *Cytogenet. Genome Res.* **129**, 211–223 (2010). doi: 10.1159/000313072; pmid: 20501977
25. Materials and methods are available as supporting materials on *Science* Online.
26. J. T. Simpson *et al*., ABySS: A parallel assembler for short read sequence data. *Genome Res.* **19**, 1117–1123 (2009). doi: 10.1101/gr.089532.108; pmid: 19251739
27. K. F. Mayer *et al*., A physical, genetic, and functional sequence assembly of the barley genome. *Nature* **491**, 711–716 (2012). pmid: 23075845
28. S. Kurtz, A. Narechania, J. C. Stein, D. Ware, A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. *BMC Genomics* **9**, 517 (2008). doi: 10.1186/1471-2164-9-517; pmid: 18976482
29. J. D. Hollister, B. S. Gaut, Epigenetic silencing of transposable elements: A trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Res.* **19**, 1419–1428 (2009). doi: 10.1101/gr.091678.109; pmid: 19478138
30. M. Kantar *et al*., Subgenomic analysis of microRNAs in polyploid wheat. *Funct. Integr. Genomics* **12**, 465–479 (2012). doi: 10.1007/s10142-012-0285-0; pmid: 22592659
31. 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; pmid: 22815845
32. G. M. Borchert *et al*., Comprehensive analysis of microRNA genomic loci identifies pervasive repetitive-element origins. *Mob. Genet. Elements* **1**, 8–17 (2011). doi: 10.4161/mge.1.1.15766; pmid: 22016841
33. International Brachypodium Initiative, Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **463**, 763–768 (2010). doi: 10.1038/nature08747; pmid: 20148030
34. International Rice Genome Sequencing Project, The map-based sequence of the rice genome. *Nature* **436**, 793–800 (2005). doi: 10.1038/nature03895; pmid: 16100779
35. A. H. Paterson *et al*., The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**, 551–556 (2009). doi: 10.1038/nature07723; pmid: 19189423
36. F. Choulet *et al*., Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. *Plant Cell* **22**, 1686–1701 (2010). doi: 10.1105/tpc.110.074187; pmid: 20581307
37. T. Lu *et al*., Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. *Genome Res.* **20**, 1238–1249 (2010). doi: 10.1101/gr.106120.110; pmid: 20627892
38. Y. Okazaki *et al*., Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**, 563–573 (2002). doi: 10.1038/nature01266; pmid: 12466851
39. Y. Marquez, J. W. Brown, C. Simpson, A. Barta, M. Kalyna, Transcriptome survey reveals increased complexity of the alternative splicing landscape in *Arabidopsis*. *Genome Res.* **22**, 1184–1195 (2012). doi: 10.1101/gr.134106.111; pmid: 22391557
40. M. M. Martis *et al*., Reticulate evolution of the rye genome. *Plant Cell* **25**, 3685–3698 (2013). doi: 10.1105/tpc.113.114553; pmid: 24104565
41. P. Hernandez *et al*., Next-generation sequencing and syntenic integration of flow-sorted arms of wheat chromosome 4A exposes the chromosome structure and gene content. *Plant J.* **69**, 377–386 (2012). doi: 10.1111/j.1365-313X.2011.04808.x; pmid: 21974774
42. J. Ma *et al*., Sequence-based analysis of translocations and inversions in bread wheat (*Triticum aestivum* L.). *PLOS ONE* **8**, e79329 (2013). doi: 10.1371/journal.pone.0079329; pmid: 24260197
43. J. S. Escobar *et al*., Multigenic phylogeny and analysis of tree incongruences in Triticeae (Poaceae). *BMC Evol. Biol.* **11**, 181 (2011). doi: 10.1186/1471-2148-11-181; pmid: 21702931
44. P. Cíváň, Z. Ivaničová, T. A. Brown, Reticulated origin of domesticated emmer wheat supports a dynamic model for the emergence of agriculture in the fertile crescent. *PLOS ONE* **8**, e81955 (2013). doi: 10.1371/journal.pone.0081955; pmid: 24312385
45. T. Marcussen *et al*., Ancient hybridizations among the ancestral genomes of bread wheat. *Science* **345**, 1250092 (2014).
46. S. Griffiths *et al*., Molecular characterization of Ph1 as a major chromosome pairing locus in polyploid wheat. *Nature* **439**, 749–752 (2006). doi: 10.1038/nature04434; pmid: 16467840
47. K. F. X. Mayer *et al*., Gene content and virtual gene order of barley chromosome 1H. *Plant Physiol.* **151**, 496–505 (2009). doi: 10.1104/pp.109.142612; pmid: 19692534
48. G. Moore, K. M. Devos, Z. Wang, M. D. Gale, Cereal genome evolution. Grasses, line up and form a circle. *Curr. Biol.* **5**, 737–739 (1995). doi: 10.1016/S0960-9822(95)00148-5; pmid: 7583118
49. M. C. Luo *et al*., 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; pmid: 23610408
50. M. Mascher *et al*., Anchoring and ordering NGS contig assemblies by population sequencing (POPSEQ). *Plant J.* **76**, 718–727 (2013). doi: 10.1111/tpj.12319; pmid: 23998490
51. M. E. Sorrells *et al*., Reconstruction of the synthetic W7984 x Opata M85 wheat reference population. *Genome* **54**, 875–882 (2011). doi: 10.1139/g11-054; pmid: 21999208
52. J. Zhang, Evolution by gene duplication: An update. *Trends Ecol. Evol.* **18**, 292–298 (2003). doi: 10.1016/S0169-5347(03)00033-8
53. L. Li, C. J. Stoeckert Jr., D. S. Roos, OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003). doi: 10.1101/gr.1224503; pmid: 12952885
54. R. Koszul, S. Caburet, B. Dujon, G. Fischer, Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments. *EMBO J.* **23**, 234–243 (2004). doi: 10.1038/sj.emboj.7600024; pmid: 14685272
55. J. L. Bennetzen *et al*., Reference genome sequence of the model plant *Setaria*. *Nat. Biotechnol.* **30**, 555–561 (2012). doi: 10.1038/nbt.2196; pmid: 22580951
56. P. S. Schnable *et al*., The B73 maize genome: Complexity, diversity, and dynamics. *Science* **326**, 1112–1115 (2009). doi: 10.1126/science.1178534; pmid: 19965430
57. T. Tanaka *et al*., The Rice Annotation Project Database (RAP-DB): 2008 update. *Nucleic Acids Res.* **36**, D1028–D1033 (2008).pmid: 18089549
58. H. Ozkan, A. A. Levy, M. Feldman, Allopolyploidy-induced rapid genome evolution in the wheat (*Aegilops-Triticum*) group. *Plant Cell* **13**, 1735–1747 (2001). doi: 10.1105/tpc.13.8.1735; pmid: 11487689
59. R. J. Buggs *et al*., Rapid, repeated, and clustered loss of duplicate genes in allopolyploid plant populations of independent origin. *Curr. Biol.* **22**, 248–252 (2012). doi: 10.1016/j.cub.2011.12.027; pmid: 22264605
60. A. H. Paterson *et al*., Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. *Nature* **492**, 423–427 (2012). doi: 10.1038/nature11798; pmid: 23257886
61. R. Grantham, Amino acid difference formula to help explain protein evolution. *Science* **185**, 862–864 (1974). doi: 10.1126/science.185.4154.862; pmid: 4843792
62. J. Cao *et al*., Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat. Genet.* **43**, 956–963 (2011). doi: 10.1038/ng.911; pmid: 21874002
63. E. D. Akhunov *et al*., Comparative analysis of syntenic genes in grass genomes reveals accelerated rates of gene structure and coding sequence evolution in polyploid wheat. *Plant Physiol.* **161**, 252–265 (2013). doi: 10.1104/pp.112.205161; pmid: 23124323
64. J. C. Schnable, N. M. Springer, M. Freeling, Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 4069–4074 (2011). doi: 10.1073/pnas.1101368108; pmid: 21368132
65. R. A. Rapp, J. A. Udall, J. F. Wendel, Genomic expression dominance in allopolyploids. *BMC Biol.* **7**, 18 (2009). doi: 10.1186/1741-7007-7-18; pmid: 19409075
66. B. Chaudhary *et al*., Reciprocal silencing, transcriptional bias and functional divergence of homeologs in polyploid cotton (Gossypium). *Genetics* **182**, 503–517 (2009). doi: 10.1534/genetics.109.102608; pmid: 19363125
67. M. Pumphrey, J. Bai, D. Laudencia-Chingcuanco, O. Anderson, B. S. Gill, Nonadditive expression of homoeologous genes is established upon polyploidization in hexaploid wheat. *Genetics* **181**, 1147–1157 (2009). doi: 10.1534/genetics.108.096941; pmid: 19104075
68. M. Pfeifer *et al*., Genome interplay in the grain transcriptome of hexaploid bread wheat. *Science* **345**, 1250091 (2014).
69. M. J. Yoo, E. Szadkowski, J. F. Wendel, Homoeolog expression bias and expression level dominance in allopolyploid cotton. *Heredity* **110**, 171–180 (2013). doi: 10.1038/hdy.2012.94; pmid: 23169565
70. K. L. Adams, R. Cronn, R. Percifield, J. F. Wendel, Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 4649–4654 (2003). doi: 10.1073/pnas.0630618100; pmid: 12665616
71. F. Cheng *et al*., Biased gene fractionation and dominant gene expression among the subgenomes of *Brassica rapa*. *PLOS ONE* **7**, e36442 (2012). doi: 10.1371/journal.pone.0036442; pmid: 22567157
72. J. Wang *et al*., Stochastic and epigenetic changes of gene expression in *Arabidopsis* polyploids. *Genetics* **167**, 1961–1973 (2004). doi: 10.1534/genetics.104.027896; pmid: 15342533
73. Z. J. Chen, Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. *Annu. Rev. Plant Biol.* **58**, 377–406 (2007). doi: 10.1146/annurev.arplant.58.032806.103835; pmid: 17280525
74. K. L. Adams, Evolution of duplicate gene expression in polyploid and hybrid plants. *J. Hered.* **98**, 136–141 (2007). doi: 10.1093/jhered/esl061; pmid: 17208934
75. G. van Ooijen *et al*., Structure-function analysis of the NB-ARC domain of plant disease resistance proteins. *J. Exp. Bot.* **59**, 1383–1397 (2008). doi: 10.1093/jxb/ern045; pmid: 18390848
76. A. Katiyar *et al*., Genome-wide classification and expression analysis of MYB transcription factor families in rice and

Arabidopsis. *BMC Genomics* **13**, 544 (2012). doi: 10.1186/1471-2164-13-544; pmid: 23050870
77. L. Yan *et al*., Positional cloning of the wheat vernalization gene VRN1. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 6263–6268 (2003). doi: 10.1073/pnas.0937399100; pmid: 12730378
78. A. Turner, J. Beales, S. Faure, R. P. Dunford, D. A. Laurie, The pseudo-response regulator Ppd-H1 provides adaptation to photoperiod in barley. *Science* **310**, 1031–1034 (2005). doi: 10.1126/science.1117619; pmid: 16284181
79. A. Díaz, M. Zikhali, A. S. Turner, P. Isaac, D. A. Laurie, Copy number variation affecting the Photoperiod-B1 and Vernalization-A1 genes is associated with altered flowering time in wheat (*Triticum aestivum*). *PLOS ONE* **7**, e33234 (2012). doi: 10.1371/journal.pone.0033234; pmid: 22457747
80. S. O. Oyola *et al*., Optimizing Illumina next-generation sequencing library preparation for extremely AT-biased genomes. *BMC Genomics* **13**, 1 (2012). doi: 10.1186/1471-2164-13-1; pmid: 22214261
81. R. E. Knox *et al*., Chromosomal location of the cadmium uptake gene (*Cdu1*) in durum wheat. *Genome* **52**, 741–747 (2009). doi: 10.1139/G09-042; pmid: 19935921
82. K. Wiebe *et al*., Targeted mapping of Cdu1, a major locus regulating grain cadmium concentration in durum wheat (*Triticum turgidum* L. var durum). *Theor. Appl. Genet.* **121**, 1047–1058 (2010). doi: 10.1007/s00122-010-1370-1; pmid: 20559817
83. L. Comai, The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.* **6**, 836–846 (2005). doi: 10.1038/nrg1711; pmid: 16304599
84. P. J. Berkman *et al*., Sequencing and assembly of low copy and genic regions of isolated *Triticum aestivum* chromosome arm 7DS. *Plant Biotechnol. J.* **9**, 768–775 (2011). doi: 10.1111/j.1467-7652.2010.00587.x; pmid: 21356002
85. P. J. Berkman *et al*., Sequencing wheat chromosome arm 7BS delimits the 7BS/4AL translocation and reveals homoeologous gene conservation. *Theor. Appl. Genet.* **124**, 423–432 (2012). doi: 10.1007/s00122-011-1717-2; pmid: 22001910
86. T. Tanaka *et al*., Next-generation survey sequencing and the molecular organization of wheat chromosome 6B. *DNA Res.* **21**, 103–114 (2013). pmid: 24086083
87. S. Wang *et al*., Characterization of polyploid wheat genomic diversity using a high-density 90, 000 single nucleotide polymorphism array. *Plant Biotechnol. J.* (2014). doi: 10.1111/pbi.12183; pmid: 24646323

ACKNOWLEDGMENTS

The authors would like to thank Graminor AS; Biogemma; Institut National de la Recherche Agronomique (INRA); International Center for Agricultural Research in the Dry Areas; Department of Biotechnology, Ministry of Science and Technology, Government of India (chr. 2A; grant no. BT/IWGSC/03/TF/2008); and the Biotechnology and Biological Sciences Research Council (BBSRC UK) for funding the chromosome sequencing at the Genome Analysis Centre. Chromosome sequencing at other centers was funded by the following: chr. 3A—U.S. Department of Agriculture Agriculture and Food Research Initiative (USDA AFRI) Triticeae-CAP (2011-68002-30029) and the Kansas Wheat Commission; chr. 3B—grants from the French National Research Agency (ANR-09- GENM-025 3BSEQ) and France Agrimer; chr. 6B—grants from the Ministry of Agriculture, Forestry and Fisheries of Japan "Genomics for agricultural innovation KGS-1003,1004", "Genomics based technology for agricultural improvement, NGB- 1003," and Nisshin Flour Milling Incorporated; chr. 6D and *Triticum durum* cv. Strongfield—grants from Genome Canada, Genome Prairie, University of Saskatchewan Ministry of Agriculture, Western Grains Research Foundation; chr. 7B—grant no. 199387 from the Norwegian Research Council and from Graminor AS; chr. 7A and 7D sequence reads were provided by D.E.. Chromosome flow sorting and DNA preparation was supported through grants P501/12/G090 and P501/12/2554 from the Czech Science foundation. Chromosome sequence assembly was supported by the BBSRC (UK). K.F.X.M. acknowledges grants from the German Ministry for Education and Research (BMBF) Plant2030, TRITEX, Deutsche Forschungsgemeinschaft (DFG) SFB 924, and EC Transplant. K.E. and J.R. are supported by sponsors of the IWGSC, which include Arcadia Biosciences, Australian Centre for Plant Functional Genomics, Biogemma, Bayer CropScience, Commonwealth Science and Industrial Research Organisation, Centro Internacional de Mejoramiento de Maíz y Trigo, Céréales Vallée, Dow AgroSciences, Dupont, Evogene, Florimond Desprez, Grains Research and Development Corporation, Graminor, Heartland Plant Innovation, INRA, KWS, Kansas Wheat Commission, Limagrain, Monsanto, RAGT, and Syngenta. N.G. is supported by European Commission Marie Curie Actions (FP7-MC-IIF-Noncollinear Genes). T.W. is supported by the Swiss National Foundation and P.F., M.C., A.M.S., and L.C. are supported by the Italian Ministry of Agriculture special project "MAPPA-5A." H.B. acknowledges funding from Sabanci University and the Scientific and Technological Research Council of Turkey. B.W. and B.S. were funded by the Gatsby Charitable Foundation and the BBSRC (UK) Grant BB/J003166/1. R.W. is a Trustee Director of TGAC, Norwich, UK, and A.K. is a shareholder of Diversity Arrays Technology Pty Ltd. The POPSeq analysis carried out by the U.S. Department of Energy Joint Genome Institute was supported by the Office of Science of the U.S. Department of Energy under contract no. DE-AC02-05CH11231. Additional support for the work was funded from the Triticeae-CAP, USDA AFRI (2011-68002-30029) to G.J.M.; the Scottish Government Rural and Environment Science and Analytical Services Division Research Programme to R.W.; and the German Ministry of Research and Education (BMBF TRITEX 0315954) to N.S. Sequence reads and assembled sequences are available at European Molecular Biology Laboratory/GenBank/DNA Data Bank of Japan short read archives and sequence repositories, respectively (PRJEB3955—whole-genome sequences of *T. aestivum* 'Chinese Spring,' *T. urartu*, *Ae. speltoides*, *Ae. tauschii*, *T. turgidum*; SRP004490.3—whole-genome sequencing of *T. monococcum*; SRP004490—whole-genome sequencing of *Ae. tauschii*; PRJEB4849—whole-genome sequences of *Ae. sharonensis*; PRJEB4750—*T. aestivum* RNA-seq data; SRP037990—*T. aestivum* SynOpDH mapping population; SRP037781—*T. aestivum* synthetic opata M85; SRP037994—*T. aestivum* synthetic W7984). All data can be accessed via the IWGSC repository at Unité de Rercherche Génomique Info: http://wheat-urgi.versailles.inra.fr/Seq-Repository/.

The International Wheat Genome Sequencing Consortium (IWGSC)
Authorship of this paper should be cited as "International Wheat Genome Sequencing Consortium." Participants are arranged by working group. Corresponding authors (*), major contributors (†), and equally contributing authors (‡) are indicated.
Principal Investigators: Klaus F. X. Mayer[1]* (k.mayer@helmholtz-muenchen.de), Jane Rogers[2]* (janerogersh@gmail.com), Jaroslav Doležel[3]* (dolezel@ueb.cas.cz), Curtis Pozniak[4]* (curtis.pozniak@usask.ca), Kellye Eversole[2]* (eversole@eversoleassociates.com), Catherine Feuillet[5]* (catherine.feuillet@bayer.com)
Provision of seed material for ditelosomic wheat lines: Bikram Gill,[6] Bernd Friebe,[6] Adam J. Lukaszewski,[7] Pierre Sourdille,[14] Takashi R Endo[8]
Chromosome sorting and DNA preparation: Jaroslav Doležel,[3]† Marie Kubaláková,[3] Jarmila Číhalíková,[3] Zdeňka Dubská,[3] Jan Vrána,[3] Romana Šperková,[3] Hana Šimková[3]
DNA sequencing: Jane Rogers,[2]† Melanie Febrer,[9] Leah Clissold,[10] Kirsten McLay,[10] Kuldeep Singh,[11] Parveen Chhuneja,[11] Nagendra K. Singh,[12] Jitendra Khurana,[13] Eduard Akhunov,[6] Frédéric Choulet,[14] Pierre Sourdille,[14] Catherine Feuillet,[5] Adriana Alberti,[15] Valérie Barbe,[15] Patrick Wincker,[15] Hiroyuki Kanamori,[16] Fuminori Kobayashi,[16] Takeshi Itoh,[16] Takashi Matsumoto,[16] Hiroaki Sakai,[16] Tsuyoshi Tanaka,[16] Jianzhong Wu,[16] Yasunari Ogihara,[17] Hirokazu Handa,[16] Curtis Pozniak,[4] P. Ron Maclachlan,[4] Andrew Sharpe,[18] Darrin Klassen,[18] David Edwards,[19] Jacqueline Batley,[19] Odd-Arne Olsen,[20,21] Simen Rød Sandve,[20] Sigbjørn Lien,[37] Burkhard Steuernagel,[22] Brande Wulff[22]
DNA sequence assembly: Mario Caccamo,[10]† Sarah Ayling,[10] Ricardo H. Ramirez-Gonzalez,[10] Bernardo J. Clavijo,[10] Burkhard Steuernagel,[22] Jonathan Wright[10]
Gene annotation: Matthias Pfeifer,[1] Manuel Spannagl,[1] Klaus F. X. Mayer[1]†
Genome Zipping: Mihaela M. Martis,[1] Eduard Akhunov,[6] Frédéric Choulet,[14] Klaus F. X. Mayer[1]†
POPSEQ analysis: Martin Mascher,[23] Jarrod Chapman,[24] Jesse A. Poland,[25] Uwe Scholz,[23] Kerrie Barry,[24] Robbie Waugh,[26] Daniel S. Rokhsar,[24] Gary J. Muehlbauer,[27] Nils Stein[28]
Repetitive DNA analysis: Heidrun Gundlach,[1] Matthias Zytnicki,[29] Véronique Jamilloux,[29] Hadi Quesneville,[29] Thomas Wicker,[30] Klaus F. X. Mayer[1]
miRNAs: Primetta Faccioli,[31]‡ Moreno Colaiacovo,[31]‡ Matthias Pfeifer,[1]‡ Antonio Michele Stanca,[31] Hikmet Budak,[32] Luigi Cattivelli[31]†
Genome structure and duplications: Natasha Glover,[14] Mihaela M. Martis,[1] Frédéric Choulet,[14] Catherine Feuillet,[5] Klaus F. X. Mayer[1]
Transcriptome sequencing and expression analysis: Matthias Pfeifer,[1] Lise Pingault,[14] Klaus F. X. Mayer,[1]† Etienne Paux[14]†
Gene family analysis: Manuel Spannagl,[1] Sapna Sharma,[1] Klaus F. X. Mayer,[1]† Curtis Pozniak[4]†
Proteogenomics analysis: Rudi Appels,[33]† Matthew Bellgard,[33] Brett Chapman,[33] Matthias Pfeifer[1]
Comparative analysis of diploid, tetraploid and hexaploid wheat: Matthias Pfeifer,[1] Simen Rød Sandve,[20] Thomas Nussbaumer,[1] Kai Christian Bader,[1] Frédéric Choulet,[14] Catherine Feuillet,[5] Klaus F. X. Mayer[1]†
Development and mapping of marker sets: Eduard Akhunov,[6] Etienne Paux,[14] Hélène Rimbert,[36] Shichen Wang,[6] Jesse A. Poland,[25] Ron Knox,[34] Andrzej Kilian,[35] Curtis Pozniak[4]†
Sequence repository: Michael Alaux,[29]† Françoise Alfama,[29] Loïc Couderc,[29] Véronique Jamilloux,[29] Nicolas Guilhot,[14] Claire Viseux,[29] Mikaël Loaec,[29] Hadi Quesneville[29]
Study design: Jane Rogers,[2] Jaroslav Doležel,[3] Kellye Eversole,[2] Catherine Feuillet,[5] Beat Keller,[30] Klaus F. X. Mayer,[1] Odd-Arne Olsen,[20,21] Sebastien Praud[36]

[1]Plant Genome and Systems Biology, Helmholtz Zentrum Munich, Ingolstädter Landstrasse 1, 85764 Neuherberg, Germany. [2]IWGSC, Eversole Associates, 5207 Wyoming Road, Bethesda, MD 20816, USA. [3]Institute of Experimental Botany, Center of Plant Structural and Functional Genomics, Šlechtitelů 31, 783 71 Olomouc, Czech Republic. [4]Crop Development Centre, Department of Plant Sciences, College of Agriculture and Bioresources, University of Saskatchewan, 51 Campus Drive, Saskatoon, SK, Canada. [5]Bayer Crop Science, 3500 Paramount Parkway, Morrisville, NC 27560, USA. [6]Kansas State University, Department of Plant Pathology, Manhattan, KS 66506–5502, USA. [7]College of Natural and Agricultural Sciences, Botany and Plant Sciences, University of California, Riverside, CA 92521, USA. [8]Laboratory of Plant Genetics, Graduate School of Agriculture, Kyoto University, Kyoto 606-8502, Japan. [9]Genomic Sequencing Unit, University of Dundee, Dow Street, Dundee DD1 5EH, UK. [10]Genome Analysis Centre, Norwich Research Park, Norwich, NR4 7UH, UK. [11]School of Agricultural Biotechnology, Punjab Agricultural University, Ludhiana 141 004, India. [12]National Research Centre on Plant Biotechnology, Indian Agricultural Research Institute, New Delhi 110 012, India. [13]Interdisciplinary Centre for Plant Genomics and Department of Plant Molecular Biology, University of Delhi, South Campus, New Delhi 110 021, India. [14]INRA–University Blaise Pascal UMR1095 Genetics, Diversity and Ecophysiology of Cereals, 5 chemin de Beaulieu, 63039 Clermont-Ferrand, France. [15]Commissariat à l'Energie Atomique Genoscope, Centre National de Séquençage, 2 rue Gaston Crémieux, CP5706, 91057 Evry, France. [16]Plant Genome Research Unit, National Institute of Agrobiological Sciences, 2-1-2, Kan-non-dai, Tsukuba 305-8602, Japan. [17]Kihara Institute for Biological Research, Yokohama City University, Maioka-cho 641-12, Totsuka-ku, 244-0813 Yokohama, Japan. [18]National Research Council Canada, 110 Gymnasium Place, Saskatoon, SK, S7N 0W9, Canada. [19]Australian Centre for Plant Functional Genomics, School of Agriculture and Food Sciences, University of Queensland, St. Lucia, QLD 4072, Australia, and School of Plant Biology, University of Western Australia, WA 6009, Australia. [20]Department of Plant Sciences, Center for Integrative Genetics (CIGENE), Norwegian University of Life Sciences, 1432 Ås, Norway. [21]Department of Natural Science and Technology, Hedmark University College, N-2318, Norway. [22]Sainsbury Laboratory, Norwich Research Park, Norwich, NR4 7UH, UK. [23]Bioinformatics and Information Technology, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), D-06466 Seeland OT Gatersleben, Germany. [24]U.S. Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA. [25]USDA-ARS Hard Winter Wheat Genetics Research Unit and Department of Agronomy, Kansas State University, Manhattan, KS 66506-5502, USA. [26]James Hutton Institute, Invergowrie, Dundee DD2 5DA, UK. [27]Department of Agronomy and Plant Genetics, Department of Plant Biology, University of Minnesota, St. Paul, MN 55108, USA. [28]Genome Diversity, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), D-06466 Seeland OT Gatersleben, Germany. [29]INRA, UR1164 URGI–Research Unit in Genomics-Info, INRA de Versailles, Route de Saint-Cyr, Versailles, 78026, France. [30]Institute of Plant Biology, University of Zurich, Zollikerstrasse 107, CH-8008 Zurich, Switzerland. [31]Consiglio per la Ricerca e la sperimentazione in Agricoltura–Genomics Research Centre, via San Protaso 302, I-29017 Fiorenzuola d'Arda, Italy. [32]Sabanci University Biological Sciences and Bioengineering Program, 34956 Istanbul, Turkey. [33]Centre for Comparative Genomics, Murdoch University, Perth, WA 6150, Australia. [34]Semiarid Prairie Agricultural Research Centre, Post Office Box 1030, Swift Current, Saskatchewan S9H 3X2, Canada. [35]Diversity Arrays Technology Pty Limited, 1 Wilf Crane Crescent, Yarralumla ACT2600, Australia. [36]Biogemma, Centre de Recherche de Chappes, Route d'Ennezat, 63720 Chappes, France. [37]Department of Animal and Aquicultural Sciences, CIGENE, Norwegian University of Life Sciences, Arboretvelen 6, 1432 Ås, Norway.

Supplementary Materials

www.sciencemag.org/content/345/6194/1251788/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S60
Tables S1 to S48
References (*88*–*160*)

5 February 2014; accepted 2 June 2014
10.1126/science.1251788