



Genome interplay in the grain transcriptome of hexaploid bread wheat

Matthias Pfeifer *et al.*

Science **345**, (2014);

DOI: 10.1126/science.1250091

This copy is for your personal, non-commercial use only.

If you wish to distribute this article to others, you can order high-quality copies for your colleagues, clients, or customers by [clicking here](#).

Permission to republish or repurpose articles or portions of articles can be obtained by following the guidelines [here](#).

The following resources related to this article are available online at www.sciencemag.org (this information is current as of July 21, 2014):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:

<http://www.sciencemag.org/content/345/6194/1250091.full.html>

Supporting Online Material can be found at:

<http://www.sciencemag.org/content/suppl/2014/07/16/345.6194.1250091.DC1.html>

A list of selected additional articles on the Science Web sites **related to this article** can be found at:

<http://www.sciencemag.org/content/345/6194/1250091.full.html#related>

This article **cites 67 articles**, 28 of which can be accessed free:

<http://www.sciencemag.org/content/345/6194/1250091.full.html#ref-list-1>

This article has been **cited by** 2 articles hosted by HighWire Press; see:

<http://www.sciencemag.org/content/345/6194/1250091.full.html#related-urls>

This article appears in the following **subject collections**:

Botany

<http://www.sciencemag.org/cgi/collection/botany>

Genetics

<http://www.sciencemag.org/cgi/collection/genetics>

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 boeoticum

Triticum polonicum L.



Triticum macha

Triticum dicoccoides var. *araraticum*



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 adaptation. 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>

PHOTOS: SUSANNE STAMP; ERNST MERZ/ETH ZÜRICH

Genome interplay in the grain transcriptome of hexaploid bread wheat

Matthias Pfeifer,^{1*} Karl G. Kugler,^{1*} Simen R. Sandve,² Bujie Zhan,² Heidi Rudi,² Torgeir R. Hvidsten,³ International Wheat Genome Sequencing Consortium†
Klaus F. X. Mayer,¹ Odd-Arne Olsen^{2‡}

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 vast majority (90 to 95%) of wheat produced in the world is the allohexaploid bread wheat (*Triticum aestivum* L., $2n = 6x = 42$, AABBDD) accounting for ~20% of calories consumed by humans worldwide. Because of global economic and social trends, wheat-processing industries put great emphasis on enhancing specific grain-quality attributes. However, an allopolyploid genome composition requires regulatory mechanisms within the cell that can orchestrate the complex intergenomic gene expression through genetic or epigenetic modifications. This may result in genomic asymmetry and favored expression of individual genes from specific, single genomes [reviewed in (1)]. For allopolyploid wheat, partial or complete genome dominance has been indicated as affecting traits such as grain size and form. However, the extent and characteristics of gene expression divergence between genomes in different tissues at the whole-genome level is largely unknown. The lack of a genome sequence for bread wheat that could enable the measurement of A, B, and D genome-specific transcription has previously hindered examination (2) of the genetic control of grain components (3, 4).

The global landscape of endosperm gene expression

We surveyed the genome-wide patterns of cell-type-specific gene expression, providing de-

tailed insights into the contributions of the A, B, and D genomes to the grain transcriptome. The bread wheat endosperm is composed of three main cell types (Fig. 1A) (5). The starchy endosperm (SE) accumulates starch and storage proteins and forms the bulk of the grain. The aleurone layer (AL) is a single cell layer surrounding the SE, except on the ventral crease side, where transfer cells (TCs) develop. Aleurone cells accumulate lipids and play an essential role in grain germination, whereas the TCs actively transport sucrose from the photosynthetic tissues to the endosperm and embryo. We applied RNA sequencing (RNA-seq) in order to monitor gene expression in different cell types at three different developmental stages that reflect the progression of starch and storage protein accumulation [10, 20, or 30 days post anthesis (DPA)] (Fig. 1A). The International Wheat Genome Sequencing Consortium (IWGSC) bread wheat genome survey sequence and annotation (6) were used as a reference, enabling assignment of the endosperm transcriptome to the three homeologous genomes A, B, and D and allowing the identification of previously unidentified transcripts and isoforms (7).

In total, 46,487 out of 85,173 high-confidence (HC) wheat genes [55%; HC1 to HC3 (6)] were expressed during endosperm development (Fig. 1B), which is consistent with observations in *Arabidopsis thaliana* and barley (8, 9). The number of genes from any genome preferentially expressed in individual cell types and developmental stages [up-regulated compared with the other samples (7)] varied substantially from 136 in TCs 20 DPA to 644 in AL cells 20 DPA (Fig. 1C). Genes that were preferentially expressed in endosperm at 10 DPA were enriched for gene ontology (GO) (10) terms related to carbohydrate metabolic processes and glycolysis (4, 9); AL-specific genes were en-

riched for lipid metabolism, structural development, carbohydrate metabolic processes, and amino acid biosynthesis (4); SE-specific genes were enriched for carbohydrate and saccharide metabolism (11); and TC-specific genes were enriched for proteolysis and defense-response genes (12). Low numbers of preferentially expressed genes in grain development have also been observed in *A. thaliana* (8). This suggests that there may be conserved regulatory principles across >100 million years of angiosperm evolution. Moreover, the three wheat genomes contribute about equally to the number of expressed genes in the endosperm as a whole (18 to 19%) (Fig. 1B) in each of the individual cell types and developmental stages (Fig. 1C), as well as to the number of preferentially expressed genes in each cell type and stage (Fig. 1C).

Morphologically, as well as functionally, aleurone and SE cells are widely different in that aleurone cells accumulate lipid bodies and are desiccation-tolerant, whereas SE cells accumulate starch granules and storage proteins and die toward grain maturation (13). The difference between aleurone cells and SE cells was also apparent from our whole-genome transcriptome analysis, in which aleurone cells exhibit a separate expression cluster from all other samples (Fig. 1D). In addition, gene expression profiles related to the developmental stages were more similar than that of cell type-expression patterns; samples from the 20-DPA stage cluster together (SE, TC, and W at 20 DPA), whereas the clean SE sample from 30 DPA (SE at 30 DPA) and endosperm from 10 DPA (W at 10 DPA) form separate clusters. Although TCs and SE cells are highly different in that TCs have extended cell walls to facilitate sucrose transport, the TC sample clustered with the SE samples, which most likely is due to tightly attached SE cells. Taken together, these observations fit with characterized endosperm cell function and the progression of endosperm development toward grain maturation (5, 13, 14).

Spatiotemporal gene expression control

Endosperm development progresses through four phases: the syncytial phase, the cellularization phase, the differentiation phase, and the maturation phase (13). We identified one group of genes with spatiotemporal unspecific gene expression (cluster 0) and seven clusters of co-expressed genes (clusters I to VII) (Fig. 2). The latter reflected phases of endosperm differentiation and maturation, in which the industrially important characteristics of the wheat grain are developed. They contained from 2257 to 5369 genes (24,826 in total) without evident genome dominance. Each cluster showed clear and distinct characteristics in terms of gene expression profiles, enrichments of preferentially expressed genes, and gene function enrichment (7). For example, cluster I exhibited increased expression in 10 DPA whole endosperm (W) when cell divisions were still occurring in the periphery of the endosperm and the transcription of storage proteins and the accumulation of starch had been

¹Plant Genome and Systems Biology, Helmholtz Center Munich, 85764 Neuherberg, Germany. ²Department of Plant Sciences/Centre for Integrative Genetics, The Norwegian University of Life Sciences (NMBU), 1432 Ås, Norway.

³Department of Chemistry, Biotechnology and Food Science, NMBU, 1432 Ås, Norway.

*These authors contributed equally to this work. †International Wheat Genome Sequencing Consortium (IWGSC) authors and affiliations are listed in the supplementary materials.

‡Corresponding author. E-mail: odd-arne.olsen@nmbu.no

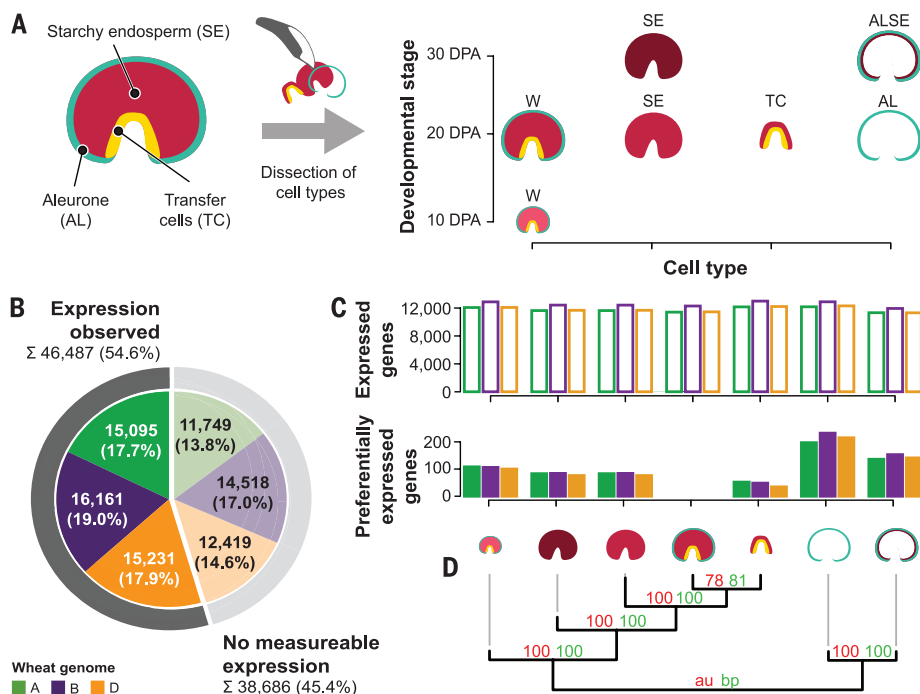


Fig. 1. Overview of the wheat endosperm transcriptome. (A) At 10 DPA, we sampled the whole endosperm (W). At 20 DPA, we also separated the SE, the AL, and the TC layer by means of manual dissection. At 30 DPA, the AL adheres tightly to the outermost SE cells, causing aleurone tissue to contain contamination from SE cells (ALSE). (B) The number of wheat genes expressed (dark gray outside line) and not expressed (lighter gray outside line) in endosperm (all cell types and developmental stages), distributed across the three genomes (green, A; purple, B; orange, D). (C) The number of genes expressed (open bars) and preferentially expressed (solid bars) in each cell type and developmental stage. Genes are considered preferentially expressed in a sample if they are up-regulated in that sample as compared with the other samples. (D) Hierarchical clustering of gene expression similarities of the samples across all genes. Bootstrap values indicate the strength of the clustering [red, approximately unbiased (au) *P* value; green, bootstrap probability (bp) *P* value].

initiated. Cluster II was expressed during the early differentiation phase, in which storage protein and starch accumulation reach their maximum. Cluster IV included a significant proportion of genes preferentially expressed in AL cells 20 DPA and was enriched for processes related to catalytic activity, lipid metabolic processes, and carbohydrate metabolism. The remaining clusters were characteristic for specific cell types (IV and VI, AL; III and VII, SE; and V, TCs) in the intermediate (III, IV, and V) and late phase (VI and VII).

We identified 6576 homeologous gene loci that had exactly one representative member from each genome (referred to as homeologous triplets; $6576 \times 3 = 19,728$ genes). Among these homeologous triplets, 5939 (88%) had at least one homeolog assigned to one of the co-expression clusters (7). We found that only 28% of the homeologous triplets (1663) had all three homeoalleles assigned to the same co-expression cluster. For 41% (2416) of the triplets, two out of three homeoalleles were assigned to the same co-expression cluster, whereas for the remaining 31% (1860), all homeoalleles fell in separate clusters. Co-expressed homeoalleles were uniformly distributed across genome pairs (A and B, 818 genes, 12%; A and D, 794 genes, 12%; B and D,

804 genes, 12%) (7). Therefore, if major expression divergence between A and B homeologs occurred in the ancestral tetraploid, our data suggests that this expression divergence has been reprogrammed in the hexaploid genome. Clusters that are spatiotemporally related (for example, early and intermediate development) often shared significant numbers of homeoalleles from the same triplets, whereas functionally different co-expression clusters only rarely did (one-sided Fisher's exact test with Bonferroni adjusted $P < 0.05$) (Fig. 2). Analyzing homeologous genes with single copies in only two of the genomes gave similar results (7). Therefore, although there is apparent expression divergence among homeologs, most of this divergence is due to nonradical alterations in spatiotemporal expression. Thus sub-functionalization, rather than neo-functionalization, appears to be occurring among the three bread wheat genomes.

Endosperm cell type function and module-associated genome dominance in the grain transcriptional network

We carried out analysis of genome-wide expression patterns of the homeologous triplets using principal component analysis and hierarchical clustering. Samples grouped according

to genomes rather than cell types (7)—a finding that is consistent with observations of transcriptomes from different wheat organs (6). This result demonstrated that genome-specific gene expression dominates over tissue-specific gene expression also during endosperm development. To gain deeper insights into the transcriptional dynamics of homeologs, we inferred a co-expression network that was partitioned into 25 modules (Fig. 3A) (7). These transcriptional modules display different spatiotemporal characteristics, as revealed by integrating the information on gene expression at different developmental stages and in different cell types. We found spatiotemporal clustering of the network that separated cell types, but also a clustering of grain developmental stages (7). AL-related modules (Fig. 3A, left, turquoise nodes) formed a large cluster of genes that were expressed at 20 and 30 DPA and were enriched for functions connected to, for example, energy metabolism, vitamin biosynthesis, and hydrolase activity. SE-related modules (Fig. 3A, left, red nodes) were more scattered and could be linked to, for example, polysaccharide catabolism, the glyoxylate cycle, and autophagy. TCs (Fig. 3A, yellow nodes) formed dense, separated clusters enriched for “response to stimulus” functionality. Transcriptional modules enriched for more general functionalities (such as transport or translation) without cell type or developmental phase specificity were also found (Fig. 3A, gray nodes). In total, AL-related modules constituted more than one third of the nodes (2207), whereas the other cell types contributed to a lesser extent (SE, 658 nodes, 11%; TC, 149 nodes, 2%). The remaining nodes grouped either with the early phase of endosperm development or unspecific clusters.

We tested the transcriptional network for genome asymmetry. Different genomes dominated expression for 23 of the modules (92%), and no single genome proved to be overly dominant (Fig. 3A, right) (7). Highly connected genes in the network (hub genes) were significantly more likely to show a genome bias than were non-hub genes (one-sided Fisher's exact test; $P < 0.001$). They were significantly enriched in modules that serve as connecting layers among different regions of the network (one-sided Fisher's exact test; $P < 0.001$) (Fig. 3A, inset, highlighted red). Because hub genes display characteristic expression profiles for network modules (15), our observations suggest that these genes might play an important role in orchestrating genome-specific expression in the grain transcriptome network.

We superimposed the observed genome asymmetry to a semantic aggregation (16) of significantly overrepresented GO categories. Both general cellular functions as well as specific functions during bread wheat grain development were attributable to distinct contributions from individual genomes (Fig. 3B). These findings illustrate at least partial functional compartmentalization of the endosperm transcriptome among the three genomes.

Sequence divergence analysis comparing homeologous gene pairs corroborated the hypothesis

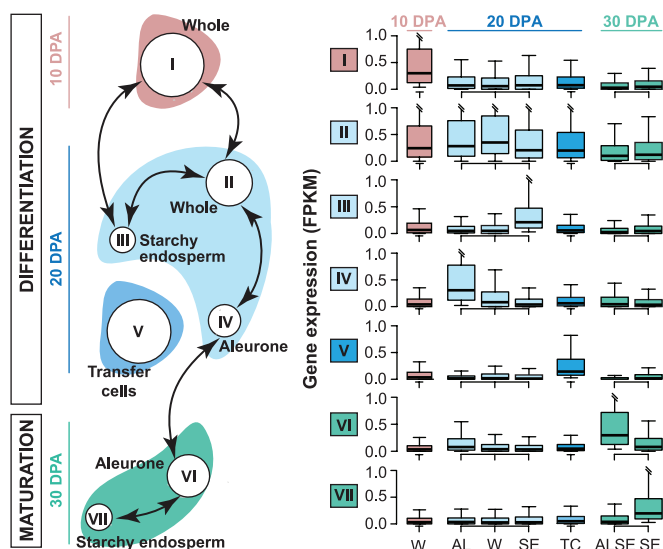


Fig. 2. Spatiotemporal control of endosperm gene expression. Seven co-expression clusters (nodes I to VII) that span the phases of progression of starch and storage protein accumulation were identified with *k*-means clustering and spatially arranged so as to reflect the phases of maturation. For homeologous triplets, spatiotemporal differences of expression patterns between individual (triplet) gene members were analyzed (7). Partitioning of expression for individual triplet genes was frequently observed (for example, homeologous A and B genome–encoded genes were located in cluster I, whereas the D genome copy clustered in cluster II). Bidirectional arrows connect pairs of clusters with a significant enrichment for expression partitioning among homeologous triplets. Assignment of different triplet members to different co-expression clusters is biased toward spatiotemporally related clusters (such as clusters IV and VI, late differentiation and maturation phase of aleurone cells). Gene expression profiles across developmental stages and cell types of each cluster are visualized by use of boxplots. Boxes span the data range between the first and the third quartiles, and the median is represented as horizontal lines. Whiskers extend to the most extreme data point, which is no more than 1.5 times the interquartile range away from the first and the third quartile.

of homoploid hybrid speciation of the D genome lineage (7, 17). The numbers of synonymous substitutions per synonymous site between A-D and B-D homeologous genes were similar and significantly smaller than for homeologous gene pairs from the A and B genomes. Expression profiles of homeologs and expression dominance within the co-expression modules were independent of evolutionary history and relatedness (7). This suggests other genetic or epigenetic regulatory mechanisms as the cause of genome asymmetry (18–20).

Regulation of endosperm gene expression linked to chromosomal domains

Although the regulation of gene expression is expected to affect genes at different loci in the genome, epigenetic mechanisms often influence neighboring genes (21). To investigate the impact of chromosomal positioning of genes on mRNA abundance, we mapped 57,903 (67%) bread wheat genes to seven Triticeae prototype chromosomes (7) that reflect the inferred ancestral linear gene order in the A, B, and D genomes (22, 23). Along all chromosomes, the gene expression oscillated to form chromosomal domains (24, 25) with only minor divergence between the spatiotemporal endosperm conditions (Fig. 4A) (7). This spatial distribution of gene expression was to a large extent similar between genomes. However,

we also observed chromosomal domains with asynchronous homeologous expression patterns, which indicates local genome asymmetry affecting neighboring genes (Fig. 4B). One of these domains (Fig. 4, blue rectangle) showed elevated gene expression in the D genome. We also found overrepresentation of differentially expressed homeologous triplets that were dominated by the D genome (Fisher's exact test; $P < 0.05$), which excludes gene expression divergence because of variation in local gene content between genomes. Genes located in this chromosomal domain encoded proteins involved in basal cellular processes such as DNA packaging and transport activity. The genome-wide contribution of each genome appears to be balanced, but as well as genetic effects such as variation in gene copy number, there is local genome asymmetry for neighboring genes. This suggests that epigenetic regulatory mechanisms act differently on particular corresponding domains of homeologous chromosomes.

Copy number and homeologous expression in baking-quality genes

Our data set enabled us to discriminate between transcripts from A, B, and D homeologous genes and to identify genes encoding major protein families known to affect dough quality (7). Genes identified include the high- and low-molecular-

weight glutenin genes (HMW-Glu and LMW-Glu); the α -, γ -, and ω -gliadins (Gli) (14); the grain-hardness locus genes encoding puroindoline A (pin-A) and puroindoline B (pin-B) (26); and last, the storage protein activators (SPAs) (Fig. 5A) (27). For both subunit types of glutenin genes, expression levels varied substantially, reflecting remarkable differences in the contribution of individual genomes to overall gene expression. Whereas the total gene expression of the LMW-Glu was dominated by the B genome (68%), genes of the D genome accounted for two thirds of the total expression of the HMW-Glu subunit. For genes of the A genome, we measured only marginal contribution (2%), which supports the observed inactivation of Glu-A homeoalleles in hexaploid wheat (28, 29). We found dominance of the D genome in the wheat hardness locus, including the pin-A and pin-B genes, which are exclusively encoded on the short arm of chromosome 5D (26, 30) and were expressed at high levels, as well as the three homeologous genes of the pin-B2 locus that were expressed at substantially lower levels (31). For the single-copy gene in each genome encoding the SPA (27), B genome expression strongly dominated over the expression of the A and the D homeologous gene copies. Although the B and the D genome copies were predominantly expressed at 10 DPA, the A genome–derived SPA allele was almost uniformly expressed at 10 DPA and in SE cells at 20 DPA. For the highly repetitive α -, γ -, and ω -gliadins (32–34), expression levels varied substantially both in the number of genes and their expression levels between the A, B, and D genomes (Fig. 5B). Whereas gene copies of the B and D genome dominated total gene family expression, copies of the A genome accumulated fewer transcripts. For the α -gliadins encoded on chromosome 6S, we found several gene copies for the A and B genomes, but strikingly, the entire α -Gli locus was absent from the D genome. Most likely, this lack was caused by a previously undescribed deletion of a small segment of chromosome 6DS (comprising ~200 genes) (Fig. 5C). We found the corresponding chromosomal region to be present in *Aegilops tauschii*, which suggests a partial chromosomal deletion that likely occurred after the hybridization event that formed hexaploid bread wheat.

For the inspected gene families, we found large genome-specific variation in the number and abundance of genes, which implies the presence of genomic asymmetry and preferential expression of mostly the B and D genomes. Moreover, we observed discrepancies in genome expression dominance in the absence of homeologous counterparts (α -Gli), potentially indicating trans-regulatory mechanisms and cross-talk between genomes [discussed in (1)]. This comprehensive analysis of major grain-quality gene families in Chinese Spring, although not a wheat used for making bread, illustrates the utility of measuring the number and expression levels of homologous genes in order to provide a complete picture of relevant genes for a trait under study. Valuable information of associations

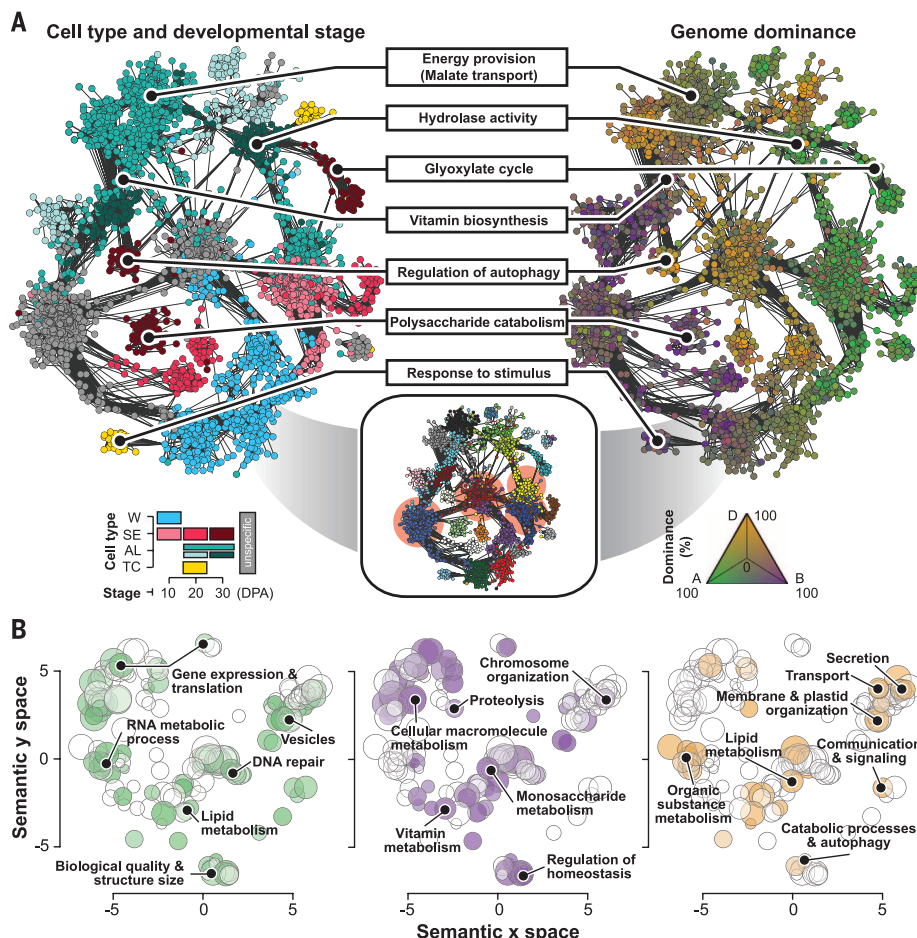


Fig. 3. Analysis of homeologous gene expression. (A) The dynamics of homeologous triplet gene expression patterns were captured in a co-expression network. Node coloring in the bottom central inset depicts the segmentation into 25 co-expression network modules. The inset denotes the network, with red circles around those modules that were significantly enriched for hub genes. The transcriptional network is colored for cell type and developmental stage (left) and genome dominance (right) (7). Selected GO categories in the center exemplify the relationship between the spatiotemporal layer and genome asymmetry (green, A; purple, B; orange, D). (B) Significantly enriched biological process GO categories were projected onto a two-dimensional semantic space (7, 16). Semantic representation of GO categories was colored according to the significant overrepresentation in co-expression modules (green, A; purple, B; orange, D). Color intensity reflects significance of enrichment test, with dark colors corresponding to lower P values and white to $P > 0.05$. Circle radii depict the size of aggregated GO terms.

between homeologous gene levels and traits of interest for wheat breeding, including baking quality, has to be learned from crosses between appropriate genotypes. This asks for future experimentation, and our work and findings build an important foundation that allows monitoring similarities and differences from genomic, transcriptional, and network-derived angles.

Conclusions

Our findings reveal the complex interplay in gene expression regulation during grain development in the hexaploid bread wheat at several levels. Genome autonomy, asymmetric contribution of genomes to particular functions, as well as indications for regulation of chromosomal domains were found. The study forms an important basis for addressing the inter- and intragenomic reg-

ulation within a polyploid genome. It enables studying the functional output in a wide range of wheat cultivars and under different environmental regimes so as to allow the identification of the underlying genetic and epigenetic factors and their interplay. This will affect the improvement of agronomical and industrial traits of one of the world's most important crops and contribute to global food security.

Materials and methods

Plant material and transcriptome sequencing

Grains of bread wheat cultivar Chinese Spring were of the same origin as those used for generating the reference genome sequence (6). Plants were grown in two phytotron chambers, and grains were harvested at 10, 20, and 30 DPA

and hand-dissected into individual cell types. RNA was isolated and prepared for RNA-seq by using paired-end libraries with 200-base pair inserts for 30 samples with four biological replicates per cell type and time point (two replicates per room) and two additional technical replicates. Libraries were sequenced with Illumina HiSeq. 2000 (Illumina, San Diego, California).

Gene annotation, RNA-seq mapping, and expression profiling

RNA-seq reads were aligned against the repeat-masked chromosome-sorted sequence (CSS) sequence assembly (6) with Bowtie2 (35) and Tophat2 (36). Returned alignments were stringently filtered so as to remove ambiguously mapped reads and read pairs with conflicting alignments. The transcriptome information was used to refine the IWGSC reference gene annotation and to identify endosperm-specific genes and alternative splicing transcripts following the procedure of (6) and using a reference-guided gene assembly with Cufflinks (37). Gene expression and tests for differentially expressed genes were computed on RNA-seq data by using CuffDiff 2 (38), and expression levels were $\log_2(\text{FPKM}+1)$ -transformed. To exclude small gene fragments and pseudogenes, further statistical analysis was constrained on HC bread wheat genes of the class levels HC1 to HC3 (6). Reproducibility of expression measurements were evaluated based on Pearson's correlation of gene expression between biological and technical replicates. The accuracy of computed expression levels was elucidated by means of an in silico simulation of an RNA-seq experiment of comparable size. Hierarchical clustering of gene expression was performed based on Pearson's correlation distance and average linkage method. Preferentially expressed genes were defined by comparing the 95% confidence interval of gene expression and significance tests of differential gene expression (false discovery rate-adjusted $P < 0.05$) between groups of samples.

Identification of homeologous gene triplets and duplets

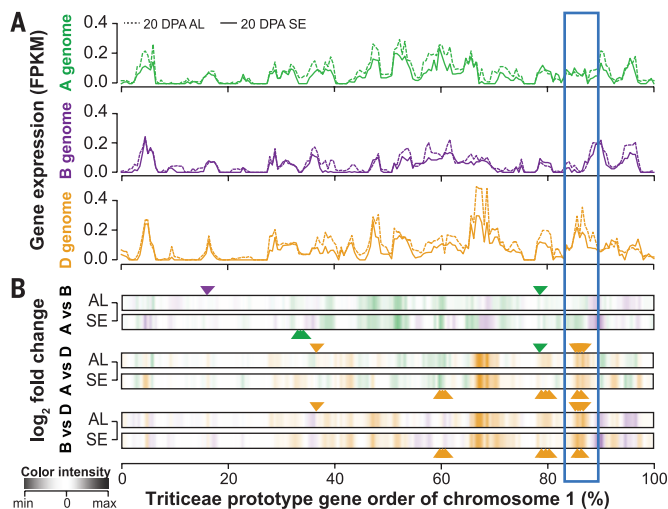
Homeologous gene triplets were defined by the protein sequence homology between the predicted protein sequences of genes from the A, B, and D genomes [Basic Local Alignment Search Tool, Protein (BLASTP), e value $\leq 10^{-5}$, alignment identity $\geq 90\%$], requiring strictly persistent best-bidirectional hits in all pairwise combinations. Homeologous duplets (one gene copy is absent in one genome) were determined from a global OrthoMCL (39) gene family clustering of individual wheat genomes and *A. thaliana* (40) and *Ae. tauschii* (41).

Identification of co-expression clusters and analysis of homeologous genes sub-functionalization

k -means clustering was performed by using Pearson's correlation distance on $\log_2$ -transformed gene expression levels to group genes that showed

Fig. 4. Local regulatory divergence at chromosomal domains.

Distribution of gene expression in 20 DPA AL and 20 DPA SE was analyzed along Triticeae prototype (Tp) chromosome 1 by using a sliding window algorithm (sliding window, size 50 Tp loci and shift 10 Tp loci). **(A)** Line charts show the median gene expression measured in aleurone and SE cells at 20 DPA. We observed a strong correlation between the expression within chromosomal domains among the three genomes. However, in several cases a single genome dominates the expression at specific domains, such as in the region indicated by the rectangle. **(B)** Heat maps display the pairwise $\log_2$ -fold changes in gene expression for each window between genomes. Triangles indicate chromosomal regions that are significantly enriched for homeologous triplets up-regulated in a single genome (Fisher's exact test; $P < 0.05$). In the marked area, the D genome expression strongly dominates, indicating a chromosomal-specific regulation of expression.



similar spatiotemporal expression profiles. Different values for k were validated, and $k = 10$ was selected as the most appropriate cluster number according to highest overall Silhouette coefficient. Thereby, all co-expression clusters with unstable gene-to-cluster assignments and negative Silhouette coefficients were combined to a “zero” cluster. The distribution of homeologous genes among the co-expression clusters was evaluated and tested for a significant number of transitions of homeologs between co-expression clusters by using a one-sided Fisher's exact test (Bonferroni corrected $P < 0.05$).

Analysis of single-copy homeologous genes

For identifying homeologous genes that were significantly differentially expressed between the genomes, we performed a permutation test (1000 iterations) for each condition and genome pairing. To analyze the expression of strict single-copy homeologous genes, expression values from the A, B, and D genomes were concatenated into a triplet expression matrix. A co-expression network was then inferred by using the Weighted Correlation Network Analysis (WGCNA) approach (42), and network modules were defined accordingly (43). Module-related genome dominance and cell type- and stage-specific expression were assessed by use of the presence of differentially expressed homeologs, module-wise gene expression profiles, and correlation of module eigengenes to predefined spatiotemporal expression patterns. We performed functional analysis for individual modules using the GO terms inferred from the bread wheat gene annotation (6) and hypergeometric tests as implemented in the GOSTATS package ($P < 0.05$) (44). These enrichments were summarized and visualized with the REVIGO Web server, which arranges semantically similar GO terms in a two-dimensional space (16). For each genome, we colored those terms, which were significantly enriched in the respective genome-dominated co-expression modules, according to the enrichment tests (P values). For sequence-divergence analysis of homeologous genes, we estimated the number of synonymous substitutions per synonymous site (K_S) and the number of nonsynonymous substitutions per nonsynonymous site (K_A) based on the pairwise best-scoring protein alignments (BLASTP) (45) using the yn00 module of the PAML 4 suite (46).

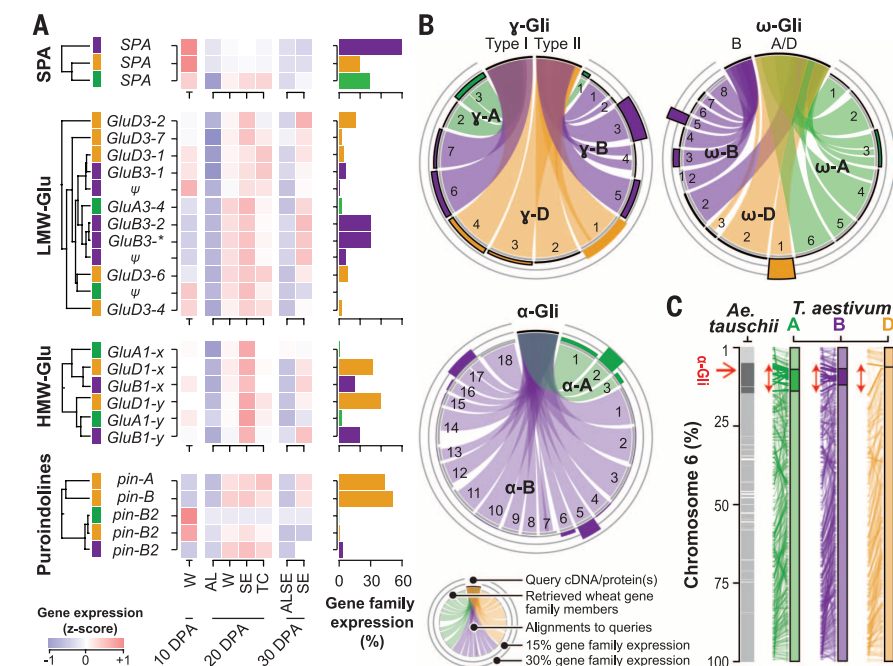


Fig. 5. Genome asymmetry in gene families affecting baking quality. Gene family composition and gene expression was analyzed for major seed storage proteins affecting baking quality of bread wheat. **(A)** Dendrograms depict phylogenetic trees for the SPA, low-molecular-weight glutenin subunit (LMW-Glu), high-molecular-weight glutenin subunit (HMW-Glu), and puroindoline (pin) A, B, and B-2 gene families. ψ indicates genes for which the predicted protein sequence is interrupted by premature stop codons, frame shifts, or repetitive elements (putative pseudogenes). Heat maps depict relative spatiotemporal gene expression, and bar charts visualize the contribution of individual genes and genomes to the total expression measured for each gene family. **(B)** The bread wheat genome and gene annotation were surveyed for α -, γ -, and ω -gliadin genes (7). Coverage of query gliadin gene sequences is indicated by the connectors. Bar height in the outer circle indicates the relative contribution of all samples to the overall gene expression of the respective family. **(C)** Comparative analysis between *Ae. tauschii* and bread wheat revealed a previously undescribed deletion of a small segment on chromosome 6DS in bread wheat “Chinese Spring,” including the locus encoding the α -gliadins. Lines connect putative orthologous gene pairs of *Ae. tauschii* (41) and the three bread wheat genomes as ordered by IWGSC (6).

Gene expression along the Triticeae prototype

The strong conservation of synteny between grass genomes and the availability of high-quality grass genomes allowed for an approximation of linear gene order. To position the A, B, and D genes in a common sequential order, they were mapped onto seven chromosome scaffolds named the Triticeae prototype chromosomes. These artificial scaffolds were generated by integrating syntenic genes from *Brachypodium* (47), rice (48), and sorghum (49) by using gene order information from barley (50). Bread wheat genes

were anchored on the scaffolds by using BLASTP and a first best-hit criterion. Gene expression distribution along the chromosomes was analyzed by use of a sliding window approach and median gene expression.

Identification and analysis of wheat grain-quality genes

Grain-quality genes were identified by use of orthologous genes from *Ae. tauschii* (41) and publically available wheat sequence information deposited at the National Center for Biotechnology Information. The corresponding genes within the bread wheat genome and gene annotation were defined with BLAST (45) and GenomeThreader (51) sequence homology searches as well as OrthoMCL gene family clustering (39). On the basis of this collective evidence and considering the RNA-seq alignment information, the candidate loci were manually refined. Gene expression for these genes was quantified as reads per kilobase per million (RPKM) (52) using HTSeq (www.huber.embl.de/users/anders/HTSeq) and custom python scripts. Phylogenetic trees were inferred by using multiple protein sequence alignments (CLUSTALW algorithm) (53), a neighborhood-joining algorithm, and the average percentage identity method. Circoletto (54) was used for comparative analysis of bread wheat genome sequences with known gliadin querry proteins.

REFERENCES AND NOTES

- M. Feldman, A. A. Levy, T. Fahima, A. Korol, Genomic asymmetry in allopolyploid plants: Wheat as a model. *J. Exp. Bot.* **63**, 5045–5059 (2012). doi: [10.1093/jxb/ers192](#); pmid: [22859676](#)
- A. R. Akhunova, R. T. Matniyazov, H. Liang, E. D. Akhunov, Homoeolog-specific transcriptional bias in allopolyploid wheat. *BMC Genomics* **11**, 505 (2010). doi: [10.1186/1471-2164-11-505](#); pmid: [20849627](#)
- T. K. Pellny et al., Cell walls of developing wheat starchy endosperm: Comparison of composition and RNA-Seq transcriptome. *Plant Physiol.* **158**, 612–627 (2012). doi: [10.1104/pp.111.189191](#); pmid: [22123899](#)
- S. A. Gillies, A. Futardo, R. J. Henry, Gene expression in the developing aleurone and starchy endosperm of wheat. *Plant Biotechnol. J.* **10**, 668–679 (2012). doi: [10.1111/j.1467-7652.2012.00705.x](#); pmid: [22672716](#)
- S. Drea et al., Systematic spatial analysis of gene expression during wheat caryopsis development. *Plant Cell* **17**, 2172–2185 (2005). doi: [10.1105/tpc.105.034058](#); pmid: [16006577](#)
- International Wheat Genome Sequencing Consortium, A chromosome-based draft sequence of the hexaploid bread wheat genome. *Science* **345**, 1251788 (2014).
- Materials and methods are available as supplementary materials on Science Online.
- M. F. Belmonte et al., Comprehensive developmental profiles of gene activity in regions and subregions of the *Arabidopsis* seed. *Proc. Natl. Acad. Sci. U.S.A.* **110**, E435–E444 (2013). doi: [10.1073/pnas.1222061110](#); pmid: [23319655](#)
- N. Sreenivasulu et al., Barley grain maturation and germination: Metabolic pathway and regulatory network commonalities and differences highlighted by new MapMan/PageMan profiling tools. *Plant Physiol.* **146**, 1738–1758 (2008). doi: [10.1104/pp.107.111781](#); pmid: [18281415](#)
- M. Ashburner et al., Gene ontology: Tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000). doi: [10.1038/75556](#); pmid: [10802651](#)
- W. H. Vensel et al., Developmental changes in the metabolic protein profiles of wheat endosperm. *Proteomics* **5**, 1594–1611 (2005). doi: [10.1002/pmic.200401034](#); pmid: [15800972](#)
- A. Serna et al., Maize endosperm secretes a novel antifungal protein into adjacent maternal tissue. *Plant J.* **25**, 687–698 (2001). doi: [10.1046/j.1365-3113x.2001.01004.x](#); pmid: [11319035](#)
- O. A. Olsen, Nuclear endosperm development in cereals and *Arabidopsis thaliana*. *Plant Cell* **16**, S214–S227 (2004). doi: [10.1105/tpc.017111](#); pmid: [15010513](#)
- P. R. Shewry, N. G. Halford, Cereal seed storage proteins: Structures, properties and role in grain utilization. *J. Exp. Bot.* **53**, 947–958 (2002). doi: [10.1093/jxbbot/53.370.947](#); pmid: [11912237](#)
- S. Horvath, J. Dong, Geometric interpretation of gene coexpression network analysis. *PLOS Comput. Biol.* **4**, e1000117 (2008). doi: [10.1371/journal.pcbi.1000117](#); pmid: [18704157](#)
- F. Supek, M. Bošnjak, N. Škunca, T. Šmuc, REVIGO summarizes and visualizes long lists of gene ontology terms. *PLOS ONE* **6**, e21800 (2011). doi: [10.1371/journal.pone.0021800](#); pmid: [21789182](#)
- T. Marcussen et al., Ancient hybridizations among the ancestral genomes of bread wheat. *Science* **245**, 1250092 (2014).
- L. Comai, The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.* **6**, 836–846 (2005). doi: [10.1038/nrg1711](#); pmid: [16304599](#)
- N. Shitsukawa et al., Genetic and epigenetic alteration among three homoeologous genes of a class E MADS box gene in hexaploid wheat. *Plant Cell* **19**, 1723–1737 (2007). doi: [10.1105/tpc.107.051813](#); pmid: [17586655](#)
- Z. Hu et al., Epigenetic modification contributes to the expression divergence of three TaEXPA1 homoeologs in hexaploid wheat (*Triticum aestivum*). *New Phytol.* **197**, 1344–1352 (2013). doi: [10.1111/nph.12131](#); pmid: [23360546](#)
- E. J. Finnegan, C. C. Sheldon, F. Jardinaud, W. J. Peacock, E. S. Dennis, A cluster of *Arabidopsis* genes with a coordinate response to an environmental stimulus. *Curr. Biol.* **14**, 911–916 (2004). doi: [10.1016/j.cub.2004.04.045](#); pmid: [15186749](#)
- G. Moore et al., Was there a single ancestral cereal chromosome? *Trends Genet.* **11**, 81–82 (1995). doi: [10.1016/S0168-9525\(00\)89005-8](#); pmid: [7732576](#)
- S. Bolot et al., The ‘inner circle’ of the cereal genomes. *Curr. Opin. Plant Biol.* **12**, 119–125 (2009). doi: [10.1016/j.pbi.2008.10.011](#); pmid: [19095493](#)
- K. Nakabayashi, M. Okamoto, T. Koshiba, Y. Kamiya, E. Nambara, Genome-wide profiling of stored mRNA in *Arabidopsis thaliana* seed germination: Epigenetic and genetic regulation of transcription in seed. *Plant J.* **41**, 697–709 (2005). doi: [10.1111/j.1365-3113x.2005.02337.x](#); pmid: [15703057](#)
- C. Rustenholz et al., A 3,000-loci transcription map of chromosome 3B unravels the structural and functional features of gene islands in hexaploid wheat. *Plant Physiol.* **157**, 1596–1608 (2011). doi: [10.1104/pp.111.183921](#); pmid: [22034626](#)
- C. F. Morris, Puroindolines: The molecular genetic basis of wheat grain hardness. *Plant Mol. Biol.* **48**, 633–647 (2002). doi: [10.1023/A:1014837431178](#); pmid: [11999840](#)
- C. Ravel et al., Nucleotide polymorphism in the wheat transcriptional activator Spa influences its pattern of expression and has pleiotropic effects on grain protein composition, dough viscoelasticity, and grain hardness. *Plant Physiol.* **151**, 2133–2144 (2009). doi: [10.1104/pp.109.146076](#); pmid: [19828671](#)
- P. Payne, G. Lawrence, Catalogue of alleles for the complex gene loci, Glu-A1, Glu-B1, and Glu-D1 which code for high-molecular-weight subunits of glutenin in hexaploid wheat. *Cereal Res. Commun.* **11**, 29–35 (1983).
- R. D. Thompson, D. Bartels, N. P. Harberd, R. B. Flavell, Characterization of the multigene family coding for HMW glutenin subunits in wheat using cDNA clones. *Theor. Appl. Genet.* **67**, 87–96 (1983). doi: [10.1007/BF00303930](#); pmid: [24258485](#)
- N. Chantret et al., Molecular basis of evolutionary events that shaped the hardness locus in diploid and polyploid wheat species (*Triticum* and *Aegilops*). *Plant Cell* **17**, 1033–1045 (2005). doi: [10.1105/tpc.104.029181](#); pmid: [15749759](#)
- M. Wilkinson et al., Identification and genetic mapping of variant forms of puroindoline b expressed in developing wheat grain. *J. Cereal Sci.* **48**, 722–728 (2008). doi: [10.1016/j.jcs.2008.03.007](#)
- F. M. Dupont, W. H. Vensel, C. K. Tanaka, W. J. Hurkman, S. B. Altenbach, Deciphering the complexities of the wheat flour proteome using quantitative two-dimensional electrophoresis, three proteases and tandem mass spectrometry. *Proteome Sci.* **9**, 10 (2011). doi: [10.1186/1477-5956-9-10](#); pmid: [21314956](#)
- O. D. Anderson, N. Huo, Y. Q. Gu, The gene space in wheat: The complete γ -gliadin gene family from the wheat cultivar Chinese Spring. *Funct. Integr. Genomics* **13**, 261–273 (2013). doi: [10.1007/s10142-013-0321-8](#); pmid: [23564033](#)
- O. D. Anderson, Y. Q. Gu, X. Kong, G. R. Lazo, J. Wu, The wheat omega-gliadin genes: Structure and EST analysis. *Funct. Integr. Genomics* **9**, 397–410 (2009). doi: [10.1007/s10142-009-0122-2](#); pmid: [19367421](#)
- B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009). doi: [10.1186/gb-2009-10-3-r25](#); pmid: [19261174](#)
- C. Trapnell, L. Pachter, S. L. Salzberg, TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009). doi: [10.1093/bioinformatics/btp120](#); pmid: [19289445](#)
- C. Trapnell et al., Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010). doi: [10.1038/nbt.1621](#); pmid: [20436464](#)
- C. Trapnell et al., Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* **31**, 46–53 (2013). doi: [10.1038/nbt.2450](#); pmid: [23222703](#)
- 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](#)
- The Arabidopsis Genome Initiative, Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000). doi: [10.1038/35048692](#); pmid: [11130711](#)
- 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](#)
- P. Langfelder, S. Horvath, WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008). doi: [10.1186/1471-2105-9-559](#); pmid: [19114008](#)
- P. Langfelder, B. Zhang, S. Horvath, Defining clusters from a hierarchical cluster tree: The Dynamic Tree Cut package for R. *Bioinformatics* **24**, 719–720 (2008). doi: [10.1093/bioinformatics/btm563](#); pmid: [18024473](#)
- S. Falcon, R. Gentleman, Using GOSTats to test gene lists for GO term association. *Bioinformatics* **23**, 257–258 (2007). doi: [10.1093/bioinformatics/btt567](#); pmid: [17098774](#)
- S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990). doi: [10.1016/S0022-2836\(05\)80360-2](#); pmid: [2231712](#)
- Z. Yang, PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007). doi: [10.1093/molbev/msm088](#); pmid: [17483113](#)
- International Brachypodium Initiative, Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **463**, 763–768 (2010). doi: [10.1038/nature08747](#); pmid: [20148030](#)
- International Rice Genome Sequencing Project, The map-based sequence of the rice genome. *Nature* **436**, 793–800 (2005). doi: [10.1038/nature03895](#); pmid: [16100779](#)
- 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](#)
- K. F. Mayer et al., Unlocking the barley genome by chromosomal and comparative genomics. *Plant Cell* **23**, 1249–1263 (2011). doi: [10.1105/tpc.110.082537](#); pmid: [21467582](#)
- G. Gremme, V. Brendel, M. E. Sparks, S. Kurtz, Engineering a software tool for gene structure prediction in higher organisms. *Inf. Softw. Technol.* **47**, 965–978 (2005). doi: [10.1016/j.infsof.2005.09.005](#)
- A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, B. Wold, Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008). doi: [10.1038/nmeth.1226](#); pmid: [18516045](#)
- M. A. Larkin et al., Clustal X and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948 (2007). doi: [10.1093/bioinformatics/btm404](#); pmid: [17846036](#)
- N. Darzentas, Circoletto: Visualizing sequence similarity with Circos. *Bioinformatics* **26**, 2620–2621 (2010). doi: [10.1093/bioinformatics/btq484](#); pmid: [20736339](#)

ACKNOWLEDGMENTS

This work was supported by research grant 199387 from the Norwegian Research Council to the Norwegian University of Life Sciences and Graminor AS to O.A.O. K.F.X.M. acknowledges generous funding by the Deutsche Forschungsgemeinschaft (DFG) to project SFB924 "Molecular mechanisms regulating yield and yield stability in plants," the German Federal Ministry of Education and Research (BMBF) in frame of the Synbreed project, and the European Commission's 7th Framework Program in frame of the TransPLANT project. The data reported in this paper are tabulated in the

supplementary materials and archived at the ArrayExpress database www.ebi.ac.uk/arrayexpress, accession no. E-MTAB-2137. No conflict of interest is known for the authors of this manuscript regarding data reported herein. S. I. Chi, E. I. Ako, A. Maharajan, and Ø. Jørgensen are thankfully acknowledged for grain dissection and plant cultivation. We thank P. Langridge, P. Shewry, B. Keller, and two anonymous reviewers for helpful suggestions on the manuscript. The RNA-seq was carried out at the Norwegian Sequencing Centre of University of Oslo. Graminor AS is acknowledged for generous support throughout this work.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/345/6194/1250091/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S25
Tables S1 to S31
References (55–69)
IWGSC Author List

23 December 2013; accepted 21 May 2014
10.1126/science.1250091