**Title: Genome interplay in the grain transcriptome of hexaploid bread wheat**

**Authors:** Matthias Pfeifer1†, Karl G. Kugler1†, Simen R. Sandve2, Bujie Zhan2, Heidi Rudi2, Torgeir R. Hvidsten3, IWGSC4, Klaus F. X. Mayer1 & Odd-Arne Olsen2\*

**Affiliations:**

1 Plant Genome and Systems Biology, Helmholtz Center Munich, 85764 Neuherberg, Germany

2 Department of Plant Sciences/Centre for Integrative Genetics, The Norwegian University of Life Sciences (NMBU), 1432 Åas, Norway

3 Department of Chemistry, Biotechnology and Food Science, The Norwegian University of Life Sciences(NMBU), 1432 Åas, Norway

4 International Wheat Genome Sequencing Consortium

\* Correspondence to: odd-arne.olsen@nmbu.no

†Authors contributed equally to the study.

**Abstract:**

Allohexaploid bread wheat (*Triticum aestivum L.*) provides ~ 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 novel 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 unprecedented insight into the transcriptional dynamics and genome interplay among individual grain cell types in a polyploid cereal genome.

**One Sentence Summary:**

The grain transcriptome of hexaploid bread wheat reveals regulatory genome asymmetry often allocated to chromosomal domains and associated with functional compartmentalization.

**Main Text:**

**Introduction**

The vast majority (90%-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. Due to 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 inter-genomic gene expression by 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 including 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 detailed 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 (TC) develop. Aleurone cells accumulate lipids and play an essential role in grain germination, whereas the transfer cells actively transport sucrose from the photosynthetic tissues to the endosperm and embryo. We applied RNA sequencing (RNA-seq) 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 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 novel transcripts and isoforms (*7*).

In total, 46,487 out of 85,173 high-confidence (HC) wheat genes (55%; HC1-3 (*6*)) were expressed during endosperm development (Fig. 1B), 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 (i.e. up-regulated compared to the other samples (*7*)) varied substantially from 136 in 20 DPA TC to 644 in 20 DPA AL (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 for lipid metabolism, structural development, carbohydrate metabolic processes, and amino acid biosynthesis (*4*), SE-specific genes for carbohydrate and saccharide metabolism (*11*) and TC-specific genes 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 (18-19%, 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 starchy endosperm cells are widely different in that aleurone cells accumulate lipid bodies and are desiccation-tolerant, whereas starchy endosperm cells accumulate starch granules and storage proteins and die towards grain maturation (*13*). The difference between aleurone cells and starchy endosperm cells was also apparent from our whole genome transcriptome analysis, where 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 DAP), while the clean starchy endosperm sample from 30 DAP (SE at 30 DAP) and endosperm from 10 DAP (W at 10 DAP) form separate clusters. Although transfer cells and starchy endosperm cells are highly different, in that transfer cells have extended cell walls to facilitate sucrose transport, the transfer cell sample clustered with the starchy endosperm samples, which is most likely due to tightly attached starchy endosperm cells. Taken together, these observations fit with characterized endosperm cell function and the progression of endosperm development towards 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, where the industrially important characteristics of the wheat grain are developed. They contained from 2,257 to 5,369 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 W when cell divisions are still occurring in the periphery of the endosperm and the transcription of storage proteins and the accumulation of starch have been initiated. Cluster II was expressed during the early differentiation phase, where storage protein and starch accumulation reach their maximum. Cluster IV included a significant proportion of genes preferentially expressed in 20 DPA AL, 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 & VI: aleurone layer, III & VII: starchy endosperm and V: transfer cells) in the intermediate (III, IV & V) and late phase (VI & VII).

We identified 6,576 homeologous gene loci that had exactly one representative member from each genome (referred to as homeologous triplets; 6,576 x 3= 19,728 genes). Among these homeologous triplets, 5,939 (88%) had at least one homeolog assigned to one of the co-expression clusters (*7*). We found that only 28% of the homeologous triplets (1,663) had all three homeoalleles assigned to the same co-expression cluster. For 41% (2,416) of the triplets, two out of three homeoalleles were assigned to the same co-expression cluster while for the remaining 31% (1,860) 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, 12%; B and D: 804, 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. Interestingly, clusters that are spatiotemporally related (e.g. early and intermediate development) often shared significant numbers of homeoalleles from the same triplets, while functionally different co-expression clusters only rarely did (one-sided Fisher’s exact test with Bonferroni adjusted p-value <0.05) (Fig. 2). Analyzing homeologous genes with single copies in only two of the genomes gave similar results (*7*). Therefore, while there is apparent expression divergence among homeologs, most of this divergence is due to non-radical 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**

Analysis of genome-wide expression patterns of the homeologous triplets was carried out using principal component analysis (PCA) 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 which 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 (turquoise­­ nodes in left panel of Fig. 3A) formed a large cluster of genes that were expressed at 20 DPA and 30 DPA, and were enriched for functions connected to e.g. energy metabolism, vitamin biosynthesis, and hydrolase activity. Starchy endosperm related modules (red nodes) were more scattered, and could be linked to e.g. polysaccharide catabolism, the glyoxylate cycle and autophagy. Transfer cells (yellow nodes) formed dense, separated clusters enriched for “response to stimulus” functionality. Transcriptional modules enriched for more general functionalities (e.g. transport, translation) without cell type or developmental phase specificity were also found (grey nodes). In total AL related modules constituted more than one third of the nodes (2,207), whereas the other cell types contributed to a lesser extent (SE: 658 (11%); TC 149 (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 panel) (*7*). Highly connected genes in the network (hub genes) were significantly more likely to show a genome bias than non-hub genes (one-sided Fisher’s exact test; p-value <0.001). They were significantly enriched in modules that serve as connecting layers among different regions of the network (highlighted red in the inset in Fig. 3A; one-sided Fisher’s exact test; p-value <0.001). As 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 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 is linked to chromosomal domains**

While 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 indicating local genome asymmetry affecting neighboring genes (Fig. 4B). One of these domains (indicated by a blue rectangle in Fig. 4) showed significantly 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-value <0.05), which excludes gene expression divergence due to 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, besides 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 & 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 finally the storage protein activators (SPA) (*27*) (Fig. 5A). 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%) supporting 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 storage protein activator (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 starchy endosperm 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). While gene copies of the B and D genome dominated total gene family expression, copies of the A genome accumulated fewer transcripts. For the α-gliadines, 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 approximately 200 genes) (Fig. 5C). Interestingly, we found the corresponding chromosomal region to be present in *Ae. 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 find 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 to provide a complete picture of relevant genes for a trait under study. Valuable information of associations between homeologous gene levels and traits of interest for wheat breeding, including baking quality, has to be learned from crosses between appropriate genotypes. While this asks for future experimentation, our work and findings build an important foundation that allows monitoring similarities and differences from a genomic, a transcriptional, and a network-derived angle.

**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 regulation within a polyploid genome. It enables studying the functional output in a wide range of wheat cultivars and under different environmental regimes to allow the identification of the underlying genetic and epigenetic factors and their interplay. This will impact 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 cv. 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 sequencing (RNA-seq) using paired-end libraries with 200 bp inserts for 30 samples with four biological replicates per cell type and time point (2 replicates per room) and two additional technical replicates. Libraries were sequenced using Illumina HiSeq2000.

*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 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 employing a reference-guided gene assembly with Cufflinks (*37*). Gene expression and tests for differentially expressed genes were computed on RNA-seq data using CuffDiff 2 (*38*) and expression levels were log2(FPKM+1)-transformed. To exclude small gene fragments and pseudogenes further statistical analysis was constrained on high-confidence bread wheat genes of the class levels HC1-3 (*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 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 in R. Preferentially expressed genes were defined by comparing the 95% confidence interval of gene expression and significance tests of differentially gene expression (FDR-adjusted p-value <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 (BLASTP, e-value ≤10-5, alignment identity ≥90%) requiring strictly persistent best-bidirectional hits in all pairwise combinations. Homeologous duplets (i.e. 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 using Pearson’s correlation distance on log2-transformed gene expression levels to group genes that showed similar spatiotemporal expression profiles. Different values for k were validated and k=10 selected as the most appropriate cluster number based on highest overall Silhouette coefficient. Thereby, all co-expression clusters with instable 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 using a one-sided Fisher's exact test (Bonferroni corrected p-value <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 (1,000 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 WGCNA approach (*42*) and network modules defined accordingly (*43*). Module-related genome dominance and cell type and stage-specific expression were assessed by presence of differentially expressed homeologs, module-wise gene expression profiles, and correlation of module eigengenes to pre-defined spatiotemporal expression patterns. We performed functional analysis for individual modules using the gene ontology terms inferred from the bread wheat gene annotation (*6*) and hypergeometric tests as implemented in the GOstats package (p-value <0.05) (*44*). These enrichments were summarized and visualized with the REVIGO webserver, 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 (KS), the number of nonsynonymous substitutions per nonsynonymous site (KA) based on the pairwise best scoring protein alignments (BLASTP (*45*)) using the yn00 module of the PAML 4 suite (*46*).

*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*) using gene order information from barley (*50*). Bread wheat genes were anchored on the scaffolds using BLASTP and a first best-hit criterion. Gene expression distribution along the chromosomes was analyzed using a sliding window approach and median gene expression.

*Identification and analysis of wheat grain quality genes* -Grain quality genes were identified using orthologous genes from *Ae. tauschii* (*41*) and publically available wheat sequence information deposited at the NCBI. The corresponding genes within the bread wheat genome and gene annotation were defined utilizing BLAST (*45*) and GenomeThreader (*51*) sequence homology searches as well as OrthoMCL gene family clustering (*39*). Based on this collective evidence and considering the RNA-seq alignment information, the candidate loci were manually refined. Gene expression for these genes was quantified as RPKM (*52*) using HTSeq (http://www-huber.embl.de/users/anders/HTSeq) and custom python scripts. Phylogenetic trees were inferred using multiple protein sequence alignments (CLUSTALW algorithm) (*53*) and 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 query proteins.

**References and Notes:**

1. 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).

2. A. R. Akhunova, R. T. Matniyazov, H. Liang, E. D. Akhunov, Homoeolog-specific transcriptional bias in allopolyploid wheat. *BMC Genomics* **11**, 505 (2010).

3. T. K. Pellny, A. Lovegrove, J. Freeman, P. Tosi, C. G. Love *et al.*, Cell walls of developing wheat starchy endosperm: comparison of composition and RNA-Seq transcriptome. *Plant Physiol* **158**, 612-627 (2012).

4. 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).

5. S. Drea, D. J. Leader, B. C. Arnold, P. Shaw, L. Dolan *et al.*, Systematic spatial analysis of gene expression during wheat caryopsis development. *Plant Cell* **17**, 2172-2185 (2005).

6. International Wheat Genome Sequencing Consortium, A chromosome-based draft sequence of the hexaploid bread wheat genome. *submitted*.

7. Supplementary materials for this text are available on *Science Online*.

8. M. F. Belmonte, R. C. Kirkbride, S. L. Stone, J. M. Pelletier, A. Q. Bui *et al.*, Comprehensive developmental profiles of gene activity in regions and subregions of the Arabidopsis seed. *Proc Natl Acad Sci U S A* **110**, E435-444 (2013).

9. N. Sreenivasulu, B. Usadel, A. Winter, V. Radchuk, U. Scholz *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).

10. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler *et al.*, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29 (2000).

11. W. H. Vensel, C. K. Tanaka, N. Cai, J. H. Wong, B. B. Buchanan *et al.*, Developmental changes in the metabolic protein profiles of wheat endosperm. *Proteomics* **5**, 1594-1611 (2005).

12. A. Serna, M. Maitz, T. O'Connell, G. Santandrea, K. Thevissen *et al.*, Maize endosperm secretes a novel antifungal protein into adjacent maternal tissue. *Plant J* **25**, 687-698 (2001).

13. O. A. Olsen, Nuclear endosperm development in cereals and Arabidopsis thaliana. *Plant Cell* **16**, S214-227 (2004).

14. P. R. Shewry, N. G. Halford, Cereal seed storage proteins: structures, properties and role in grain utilization. *J Exp Bot* **53**, 947-958 (2002).

15. S. Horvath, J. Dong, Geometric interpretation of gene coexpression network analysis. *PLoS Comput Biol* **4**, e1000117 (2008).

16. F. Supek, M. Bosnjak, N. Skunca, T. Smuc, REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* **6**, e21800 (2011).

17. T. Marcussen, S. R. Sandve, L. Heier, M. Spannagl, M. Pfeifer *et al.*, Ancient hybridizations among the ancestral genomes of bread wheat. *submitted*.

18. L. Comai, The advantages and disadvantages of being polyploid. *Nat Rev Genet* **6**, 836-846 (2005).

19. N. Shitsukawa, C. Tahira, K. Kassai, C. Hirabayashi, T. Shimizu *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).

20. Z. Hu, Z. Han, N. Song, L. Chai, Y. Yao *et al.*, Epigenetic modification contributes to the expression divergence of three TaEXPA1 homoeologs in hexaploid wheat (Triticum aestivum). *New Phytol* **197**, 1344-1352 (2013).

21. 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).

22. G. Moore, T. Foote, T. Helentjaris, K. Devos, N. Kurata *et al.*, Was there a single ancestral cereal chromosome? *Trends Genet* **11**, 81-82 (1995).

23. S. Bolot, M. Abrouk, U. Masood-Quraishi, N. Stein, J. Messing *et al.*, The 'inner circle' of the cereal genomes. *Curr Opin Plant Biol* **12**, 119-125 (2009).

24. 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).

25. C. Rustenholz, F. Choulet, C. Laugier, J. Safar, H. Simkova *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).

26. C. F. Morris, Puroindolines: the molecular genetic basis of wheat grain hardness. *Plant Mol Biol* **48**, 633-647 (2002).

27. C. Ravel, P. Martre, I. Romeuf, M. Dardevet, R. El-Malki *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).

28. 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).

29. 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).

30. N. Chantret, J. Salse, F. Sabot, S. Rahman, A. Bellec *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).

31. M. Wilkinson, Y. Wan, P. Tosi, M. Leverington, J. Snape *et al.*, Identification and genetic mapping of variant forms of puroindoline b expressed in developing wheat grain. *J Cereal Science* **48**, 722-728 (2008).

32. 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).

33. O. D. Anderson, N. Huo, Y. Q. Gu, The gene space in wheat: the complete gamma-gliadin gene family from the wheat cultivar Chinese Spring. *Funct Integr Genomics* **13**, 261-273 (2013).

34. 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).

35. 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).

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

37. C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan *et al.*, Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 511-515 (2010).

38. C. Trapnell, D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn *et al.*, Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* **31**, 46-53 (2013).

39. L. Li, C. J. Stoeckert, Jr., D. S. Roos, OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**, 2178-2189 (2003).

40. Arabidopsis Genome Initiative, Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**, 796-815 (2000).

41. J. Jia, S. Zhao, X. Kong, Y. Li, G. Zhao *et al.*, Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature* **496**, 91-95 (2013).

42. P. Langfelder, S. Horvath, WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).

43. 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).

44. S. Falcon, R. Gentleman, Using GOstats to test gene lists for GO term association. *Bioinformatics* **23**, 257-258 (2007).

45. 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).

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

47. International Brachypodium Initiative, Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature* **463**, 763-768 (2010).

48. International Rice Genome Sequencing Project, The map-based sequence of the rice genome. *Nature* **436**, 793-800 (2005).

49. A. H. Paterson, J. E. Bowers, R. Bruggmann, I. Dubchak, J. Grimwood *et al.*, The Sorghum bicolor genome and the diversification of grasses. *Nature* **457**, 551-556 (2009).

50. K. F. Mayer, M. Martis, P. E. Hedley, H. Simkova, H. Liu *et al.*, Unlocking the barley genome by chromosomal and comparative genomics. *Plant Cell* **23**, 1249-1263 (2011).

51. G. Gremme, V. Brendel, M. E. Sparks, S. Kurtz, Engineering a software tool for gene structure prediction in higher organisms. *Information and Software Technology* **47**, 965-978 (2005).

52. 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).

53. M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan *et al.*, Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948 (2007).

54. N. Darzentas, Circoletto: visualizing sequence similarity with Circos. *Bioinformatics* **26**, 2620-2621 (2010).

55. X. J. Min, G. Butler, R. Storms, A. Tsang, OrfPredictor: predicting protein-coding regions in EST-derived sequences. *Nucleic Acids Res* **33**, W677-680 (2005).

56. P. S. Schnable, D. Ware, R. S. Fulton, J. C. Stein, F. Wei *et al.*, The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112-1115 (2009).

57. R. L. Tatusov, E. V. Koonin, D. J. Lipman, A genomic perspective on protein families. *Science* **278**, 631-637 (1997).

58. T. W. Binsl, K. M. Mullen, I. H. van Stokkum, J. Heringa, J. van Beek, FluxSimulator: an R package to simulate isotopomer distributions in metabolic networks. *J. Stat. Softw* **18**, 1-18 (2007).

59. R. Suzuki, H. Shimodaira, Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **22**, 1540-1542 (2006).

60. P. J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Computational and Applied Mathematics* **20**, 53-65 (1987).

61. A. M. Yip, S. Horvath, Gene network interconnectedness and the generalized topological overlap measure. *BMC Bioinformatics* **8**, 22 (2007).

62. P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang *et al.*, Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-2504 (2003).

63. G. Csardi, T. Nepusz, The igraph software package for complex network research. *InterJournal, Complex Systems* **1695**, (2006).

64. K. F. Mayer, S. Taudien, M. Martis, H. Simkova, P. Suchankova *et al.*, Gene content and virtual gene order of barley chromosome 1H. *Plant Physiol* **151**, 496-505 (2009).

65. N. A. Eckardt, Grass genome evolution. *Plant Cell* **20**, 3-4 (2008).

66. P. Hernandez, M. Martis, G. Dorado, M. Pfeifer, S. Galvez *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).

67. Miftahudin, K. Ross, X. F. Ma, A. A. Mahmoud, J. Layton *et al.*, Analysis of expressed sequence tag loci on wheat chromosome group 4. *Genetics* **168**, 651-663 (2004).

68. K. M. Devos, J. Dubcovsky, J. Dvorak, C. N. Chinoy, M. D. Gale, Structural evolution of wheat chromosomes 4A, 5A, and 7B and its impact on recombination. *Theor Appl Genet* **91**, 282-288 (1995).

69. A. M. Waterhouse, J. B. Procter, D. M. Martin, M. Clamp, G. J. Barton, Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189-1191 (2009).

**Acknowledgments:**

Supported by research grant no. 199387 from the Norwegian Research Council to the Norwegian University of Life Sciences and Graminor AS to OAO. KFXM 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 Supporting Online Material section 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.

**Fig. 1. Overview of the wheat endosperm transcriptome.**

(**A**) At 10 days post anthesis (DPA), we sampled the whole endosperm (W). At 20 DPA, we also separated the starchy endosperm (SE), the aleurone layer (AL), and the transfer cell layer (TC) by manual dissection. At 30 DPA, the aleurone layer adheres tightly to the outermost starchy endosperm (SE) cells causing aleurone tissue to contain contamination from starchy endosperm cells (ALSE). (**B**) The number of wheat genes expressed, black outside line, and not expressed (shaded), 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 (hollow bars), and preferentially expressed (filled 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 to 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 and green: bootstrap probability (bp) p-value).

**Fig. 2. Spatiotemporal control of endosperm gene expression.**

Seven co-expression clusters (node I to VII) which span the phases of progression of starch and storage protein accumulation were identified with k-means clustering and spatially arranged 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 (e.g. homeologous A and B genome encoded genes were located in cluster I, while 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 towards spatiotemporal related clusters (e.g. 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 using 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.

**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 lower central inset depicts the segmentation into 25 co-expression network modules. The center panel 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 panel) and genome dominance (right panel) (*7*). Selected gene ontology (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-values >0.05. Circle radiuses depict the size of aggregated GO terms.

**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 starchy endosperm 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, e.g. the region indicated by the rectangle. (**B**) Heat maps display the pairwise log2-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-value <0.05,). In the marked area the D genome expression strongly dominates indicating at a chromosomal-specific regulation of expression.

**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 storage protein activator (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-gliadine 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 the IWGSC (*6*).