The transcriptional landscape of polyploid wheat

Borrill, P; International Wheat Genome Sequencing Consortium (IWGSC)

Document Version
Peer reviewed version

[Link to publication on Research at Birmingham portal](#)

The transcriptional landscape of polyploid wheat

R. H. Ramírez-González[1†], P. Borrill[1†*], D. Lang[2], S. A. Harrington[1], J. Brinton[1], L. Venturini[3], M. Davey[4], J. Jacobs[4], F. van Ex[4], A. Pasha[5], Y. Khedikar[6], S. J. Robinson[6], A. T. Cory[7], T. Florio[1], L. Concia[8], C. Juery[9], H. Schoonbeek[1], B. Steuernagel[1], D. Xiang[10], C. J. Ridout[1], B. Chalhoub[11], K. F. X. Mayer[2,12], M. Benhamed[8], D. Latrasse[8], A. Bendahmane[8], International Wheat Genome Sequencing Consortium[13], B. B. H. Wulff[1], R. Appels[14], V. Tiwari[15], R. Datla[10], F. Choulet[9], C. J. Pozniak[7], N. J. Provart[5], A. G. Sharpe[16], E. Paux[9], M. Spannagl[2], A. Bräutigam[17‡], C. Uauy[1*]

Affiliations:

[1] John Innes Centre, Norwich Research Park, NR4 7UH, Norwich, United Kingdom.

[2] Plant Genome and Systems Biology, Helmholtz Center Munich, Ingolstaedter Landstr. 1 85764 Neuherberg, Germany.

[3] Earlham Institute, Norwich Research Park, NR4 7UZ, Norwich, United Kingdom.

[4] Bayer Crop Science, Innovation Center, Technologiepark 38, 9052, Zwijnaarde, Belgium.

[5] Cell & Systems Biology/ Centre for the Analysis of Genome Evolution and Function, University of Toronto, 25 Willcocks St., Toronto, ON. M5S 3B2, Canada.

[6] Saskatoon Research and Development Centre, Agriculture and Agri-Food Canada, 107 Science Place, Saskatoon, SK, S7N 0X2, Canada.

[7] Crop Development Centre, University of Saskatchewan, Agriculture Building, 51 Campus Drive, Saskatoon SK, S7N 5A8, Canada.

[8] Institut of Plant Sciences Paris-Saclay (IPS2), UMR 9213/UMR1403, CNRS, INRA, Université Paris-Sud, Université d'Evry, Université Paris-Diderot, Sorbonne Paris-Cité, Bâtiment 630, 91405 Orsay, France

[9] GDEC, INRA, UCA, 5 Chemin de Beaulieu, Clermont-Ferrand 63039, France.

[10] Aquatic and Crop Resource Development, National Research Council Canada, 110 Gymnasium Place, Saskatoon, Saskatoon S7N 0W9, Canada.

[11] INRA, 2 rue Gaston Crémieux, Evry 9057, France

[12] School of Life Sciences Weihenstephan, Technical University Munich, Germany

[13] IWGSC, 5207 Wyoming Road, Bethesda, MD 20816, USA

[14] BioSciences, University of Melbourne/AgriBio-LaTrobe University/Veterinary and Ag Science, Murdoch University 90 South St, Perth Western Australia 6150, Australia.

[15] Plant Sciences and Landscape Architecture, University of Maryland, 4291 Field House Drive, College Park, Maryland, 20742 USA

[16] Global Institute for Food Security, University of Saskatchewan, 110 Gymnasium Place Saskatoon SK. S7N 4J8, Canada.

[17] Molecular Genetics, IPK Gatersleben, Corrensstrasse 3, 06466 Gatersleben, Germany.

*Correspondence to: Cristobal Uauy (cristobal.uauy@jic.ac.uk) and
Philippa Borrill (philippa.borrill@jic.ac.uk)

† These authors contributed equally to the work.

‡ Current address: Institute for Computational Biology, Faculty of Biology, University of Bielefeld, 33501, Bielefeld, Germany.

Abstract:

The coordinated expression of highly related homoeologous genes in polyploid species underlies the phenotypes of many of the world's major crops. Here we combine extensive gene expression datasets to produce a comprehensive, genome-wide analysis of homoeolog expression patterns in hexaploid bread wheat. Bias in homoeolog expression varied between tissues, with ~30% of wheat homoeologs showing non-balanced expression. We found expression asymmetries along wheat chromosomes, with homoeologs showing the largest inter-tissue, inter-cultivar, and coding sequence variation most often located in high-recombination distal ends of chromosomes. These transcriptionally dynamic genes potentially represent the first steps towards neo/sub-functionalization of wheat homoeologs. Co-expression networks revealed extensive coordination of homoeologs throughout development and, alongside a detailed expression atlas, provide a framework to target candidate genes underpinning agronomic traits in wheat.

One Sentence Summary: Homoeolog expression profiling across tissues reveals expression asymmetry along wheat chromosomes.

Main Text:

Polyploidy arises from whole-genome duplication or interspecific hybridization and is ubiquitous in eukaryotic plant and fungal lineages. Polyploidy has been proposed to confer adaptive plasticity, thereby shaping the evolution of plants, fungi and, to a lesser degree, animals (*1, 2*). This plasticity has facilitated the domestication and adaptation of several major crop species (*3*) including hexaploid wheat (*Triticum aestivum*; AABBDD genome), which is derived from relatively recent interspecific hybridizations between three different diploid species. In such polyploids, gene duplication alters the transcriptional landscape (*4*) by providing additional flexibility to adapt and evolve new patterns of gene expression for homoeologous gene copies (*5*). This flexibility has been suggested to be an important mechanism for controlling adaptive traits (*6, 7*), for example through neo-functionalization of duplicated genes (*8*) or tissue-specific expression (*9*). However, despite the likely importance of polyploidy in affecting gene expression, we have a limited understanding of the extent to which homoeologs share or differ in their expression patterns, the spatiotemporal dynamics of these relationships, and how epistatic interactions between individual homoeologs impact on biological traits. The new genomic resources available for wheat (*10*), along with its meiotic stability (*11*) and syntenic gene order (*12*), make wheat a particularly informative system for gaining insight into the effects of recent polyploidy on gene expression.

In this study, we leverage available RNA-Seq data (529 samples from 28 studies) and add 321 samples to explore global gene expression in hexaploid wheat across a diverse range of tissues, developmental stages, cultivars, and environmental conditions (*13*). We organized these sets of RNA-Seq samples into partially overlapping datasets from (i) a single developmental time course experiment (n = 209 samples), (ii) the reference accession Chinese Spring (CS) under non-stress conditions (n = 123 samples), (iii) four main tissue types under non-stress conditions (n = 537 samples), and (iv) seedling samples from abiotic (n = 50) and biotic (n = 163) stress experiments including controls (Table S1). These datasets, alongside a complete and annotated genome and transcriptome (*10*), provide an opportunity to conduct homoeolog-specific transcriptome profiling and to generate gene regulatory networks to better understand the spatiotemporal coordination of individual homoeologs underlying trait biology on a genome-wide scale.

A developmental gene expression atlas in polyploid wheat

We first assessed expression patterns through a developmental time course of the commercial wheat cultivar 'Azhurnaya' including 209 RNA-Seq samples comprising 22 tissue types representing grain, root, leaf, and spike samples across multiple timepoints (Fig. 1). We quantified expression using pseudoalignment of RNA-Seq reads to the RefSeqv1.0 transcriptome as implemented in kallisto (*14*), which accurately quantifies reads in a homoeolog-specific manner in polyploid wheat ((*13, 15*); Figs. S1-S2). We found evidence of expression for 83,741 (75.6%) high-confidence genes based on expression of >0.5 transcripts per million (TPM) in at least one of the 22 tissue types, and conducted complexity (Table S2) and differential expression analyses (Fig. S3). Tissue distinguished samples across development (Fig. S4; (*13*)) consistent with observations in other plant and animal species (*16, 17*). However, within similar tissue types, genome of origin also influenced expression patterns, consistent with previous results in wheat grain samples (*18*). This gene expression atlas provides a valuable resource for breeders and researchers to query and analyze their genes of interest through www.wheat-expression.com (*15*) and the Wheat eFP Browser (http://bar.utoronto.ca/efp_wheat/cgi-bin/efpWeb.cgi; Fig. S5 (*19*)).

Homoeolog expression patterns

In polyploid wheat, quantitative variation for many agronomic traits is modulated by genetic interactions between multiple sets of homoeologs in the A, B, and D genomes (*20*). These interactions range from buffering effects observed when gene homoeologs are functionally redundant (*21*), to dominance effects where variation in a single homoeolog can lead to dominant phenotypes (*22*). Understanding how these interactions influence gene expression will help inform strategies to target and manipulate individual or multiple homoeologs to quantitatively modulate trait responses for crop improvement (*20*).

To determine patterns of homoeolog expression, we analyzed 123 RNA-Seq samples representing 15 tissues under non-stress conditions (Table S1) from CS. This was the same accession used to generate the reference genome (*10*), thereby excluding cultivar-specific polymorphisms from our analysis. We found evidence of expression for 82,567 (74.5%) high-confidence genes, consistent with the cultivar Azhurnaya developmental time course. We

focused on 53,259 genes that had a 1:1:1 correspondence across the three homoeologous genomes, referred to as triads, and a summed expression of > 0.5 TPM across the triad (64.5% of expressed genes, 96.1% of all triads, Table S3). The majority of these expressed triads (94.3%) were in ancestral (i.e. syntenic) physical positions for at least two of the three genomes (50,238 genes corresponding to 16,746 syntenic triads (*10*)), whereas 5.7% (1,007 triads) had all genes in non-syntenic positions and are referred to as non-syntenic triads. For each of the 17,753 expressed triads, we standardized the relative expression of the A, B, and D genome homoeologs (Fig. S6) such that the sum was 1.0 in each individual tissue. In this way, the relative abundance of homoeolog expression is comparable within triads as well as across tissues, allowing the study of homoeolog expression bias (*23*).

We performed a global analysis combining data across all 15 tissues and focused on the 16,746 syntenic triads, but discuss patterns in non-syntenic triads where relevant. We found that the D genome had a subtle, yet significantly higher, relative abundance (33.65%) compared to the B (33.29%) and A (33.06%) genomes (Kruskal-Wallis P <0.001, Tables S4–S5). The homoeolog expression bias of the D genome is unlikely to reflect technical issues (Fig. S2) and was found in 11 of the 15 tissues, was consistent across multiple expression abundance cut-offs, and was also significant in the developmental time course (in all 22 tissues) from cultivar Azhurnaya (Figs. S7–S9, Table S5 (*13*)). This effect, however, is subtle when compared to the homoeolog expression bias observed in evolutionarily older polyploid crops such as cotton, in which genome doubling occurred at an earlier time point (*24*).

The relative expression of each homoeolog determined a triad's position in the ternary plot for the global analysis (Fig. 2A), as well as for analyses of individual tissues (Fig. S6, S10). From these plots, we defined seven homoeolog expression bias categories (*13*) including a balanced category with similar relative abundance of transcripts from the three homoeologs, and six homoeolog-dominant or homoeolog-suppressed categories on the basis of the higher or lower abundance of transcripts from a single homoeolog with respect to the other two (Fig. 2A). Most syntenic triads (72.5%) were assigned to the balanced category within each tissue, with balanced triads ranging from 62.6% in stigma and ovary to 78.9% in roots (Fig. 2B; Table S6). Triads with single-homoeolog dominance were infrequent (7.1%; range 4.7% to 11.3% between tissues), whereas syntenic triads classified as single-homoeolog suppressed were more common (20.5%; range 16.3% to 27.1%; Fig. 2B). These patterns shifted significantly in the 1,007 non-syntenic

triads, which had fewer balanced triads (58.9%) and a higher proportion of dominant (14.5%) and suppressed (26.6%) triads across tissues (Table S7-S8, χ^2 P <0.001). Across tissues, no differences were observed for the frequency of single-homoeolog dominance between genomes (Tables S6–S7). However, across all 15 tissues, D-homoeolog suppression was significantly less frequent (5.7%) than either A- or B-homoeolog suppression (7.5% and 7.2%, respectively; Kruskal-Wallis P <0.05), and this was consistent in non-syntenic triads and the developmental time course (Tables S6–S8). This pattern explains in part the subtle homoeolog expression bias observed for the D genome relative to the A and B homoeologs. This observation is consistent with the lower distribution of repressive H3K27me3 histone marks across the gene body of D-genome homoeologs compared to the A and B homoeologs (Fig. S11).

Genes from syntenic triads in the balanced category were expressed across a wider range of tissues and had higher absolute transcript abundance, on a per-genome basis (12.2 tissues; median 4.03 TPM), than genes in the suppressed (9.1 tissues; median 1.51 TPM) or dominant (6.9 tissues; median 0.57 TPM) categories (two-sample Kolmogorov–Smirnov, P <0.001; Fig. 2C-D, Tables S9–S10). The absolute transcript abundance data show that dominant triads are not the result of an overall increase in expression of a single homoeolog, but rather result from the relatively lower expression of the two other homoeologs.

To determine if the differences among homoeologs are a consequence of polyploidization, we analyzed RNA-Seq data from diploid and tetraploid progenitor species and newly created synthetic hexaploid wheat (SHW) lines (*25*). We found that 67.5% of non-balanced triads in modern-day wheat have a homoeolog expression bias category different than that observed in SHW, with all three genomes being equally affected (Table S11, Figs. S12-S14). Likewise, 47.1% of non-balanced triads in SHW are in a different category than would be expected on the basis of the progenitor species, with D-genome homoeologs most strongly influenced (*13*)(Table S12, Fig. S15). These results suggest that the polyploid context and the polyploidization process itself affect the relative expression of homoeologs compared to the baseline expression in the progenitor species (*13*), which has also been observed during the evolution of polyploid cotton (*26*) and monkeyflower (*27*).

We hypothesized that epigenetic mechanisms might be associated with differences in homoeolog expression patterns. To test this, we examined the associations between transposable elements

(TE), DNA methylation, and histone modifications with the relative expression of triads in leaves of CS. We found no clear relationship between the presence of TE in promoter regions and altered expression patterns between homoeologs in dominant and suppressed triads (Tukey HSD $P > 0.6$; Fig. S16; Table S13)(*13*). However, we identified significant differences in gene body DNA methylation and histone modifications among homoeologs (*13*).

Gene body CG methylation is widely conserved in angiosperms, although its functional significance is currently under debate (*28, 29*) given that two angiosperm species lack this epigenetic mark altogether (*30*). Here we find higher gene body CG methylation in constitutively expressed triads compared to more tissue specific triads (balanced > suppressed > dominant; Fig. S17). Within the non-balanced triads, homoeologs with higher expression had higher CG methylation than their corresponding non-dominant and suppressed homoeologs (Mann-Whitney $P <0.001$; Fig S17). These results are consistent with gene body CG methylation associated with housekeeping genes and its suggested role in homoeostatic gene expression (*29*). Similarly, the more highly expressed homoeologs within non-balanced triads had higher active histone marks (H3K36me3 and H3K9ac) and lower repressive (H3K27me3) marks in the gene body (Mann-Whitney $P <0.001$; Fig. S11). For H3K27me3, these differences were not limited strictly to the gene body and extended into the upstream and downstream regions for both dominant and suppressed triads (Fig. 2E), consistent with the tight association of H3K27me3 within inactive gene promoters (*31*). These results suggest that epigenetic status in gene bodies, as well as upstream and downstream regions, is associated with homoeolog expression bias in polyploid wheat, consistent with results in monkeyflower showing changes in DNA methylation upon polyploidization (*27*).

Breeders rely on recombination to generate new combinations of haplotypes for improving cultivars. In wheat, chromosome position strongly influences recombination rates, with relatively low recombination rates in the interstitial and proximal regions (R2a/C/R2b genomic compartments), but markedly higher rates towards the distal ends of the chromosomes (R1/R3 genomic compartments; (*32*)). In our analyses, syntenic triads in the balanced category were over-represented in the low-recombination regions (R2/C) which have higher levels of active histone marks (H3K36me3 and H3K9ac)(*10*), consistent with the higher expression of balanced triads. Homoeolog-dominant and homoeolog-suppressed triads were over-represented towards the high-recombination distal ends of chromosomes (R1/R3; χ^2 $P <0.001$; Table S14) which have

higher levels of repressive H3K27me3 histone marks (*10*), consistent with the lower expression of dominant and suppressed triads. This pattern was consistent in the developmental time course of cultivar Azhurnaya. However, when comparing these CS and Azhurnaya cultivars (nine tissues in common), we found that 84.5% of genes in the R2/C regions maintained the same expression category between cultivars, whereas only 72.2% of genes in the R1/R3 regions did so (χ^2 *P* <0.001; Table S15). These differences in homoeolog expression bias across cultivars have important implications for breeding as they suggest that through genetic crosses, breeders not only generate new combinations of haplotypes with differential expression of alleles, but also rearrange and select for homoeolog expression bias between cultivars.

Variation of triad expression patterns

Polyploidy may confer phenotypic plasticity by allowing homoeologs to be expressed differently across tissues and/or environmental conditions (*8*). Our analyses above provide a static overview of the relative homoeolog expression bias in individual tissues. Therefore, we explored whether syntenic triads retain their homoeolog expression bias category across the 15 tissues (Table S16). We found that 83.6% of balanced triads remained balanced in each of the 15 individual tissues, whereas dominant and suppressed triads tended to be more variable across tissues with only 73.4% and 62.2%, respectively, staying within their global dominance group across all 15 tissues (Fig. 3A). Both dominant and suppressed triads shifted most often to adjacent categories (16–20%) in the ternary plots and in few cases (<3.0%) changed to opposite categories (Fig. S18). These patterns were consistent in the developmental time course (Table S16). These data show that across tissues, triads most often remained consistent in their homoeolog expression bias classification, a phenomenon also seen across seven tissues in allotetraploid *Tragopogon mirus (33)*.

To complement this analysis, we determined the variation in behavior of each triad within the ternary plot across the 15 tissues by calculating the mean distance between the triad's position in each tissue to its global average position (*13*)(Fig. S19). This generated a distribution of mean distances (Fig. 3B); we focused on the 10% most stable triads (defined as those having the shortest mean distance across tissues) and the 10% most dynamic triads (largest mean distance; Fig. 3B–D). Stable triads were expressed more highly than dynamic triads (median 8.2 vs 3.2

TPM; P <0.001) and had a higher expression breadth, being expressed across almost all samples compared to dynamic triads, which were more tissue-specific (P <0.001; Fig. 3E; Table S17). Stable triads were enriched for high-level gene ontology (GO-slim) terms associated with housekeeping processes (e.g. translation, cell cycle), whereas dynamic triads were enriched for defense/external stimuli responses and secondary metabolic processes, functions which more frequently determine differences in individual fitness (Table S18)(6). In the global analysis, stable triads were significantly enriched for the balanced category (94.2%) whereas dynamic triads were almost equally spread between suppressed (37.9%), dominant (30.5%) and balanced (31.6%) categories (χ^2 P <0.001; Fig. 3F; Table S19). This pattern is consistent with stable triads being most frequently located in proximal regions (C) whereas dynamic triads tend to locate in distal ends of chromosomes (R1/R3; χ^2 P <0.001; Table S20). These results demonstrate expression asymmetry across wheat chromosomes, whereby the high-recombination distal ends of chromosomes have triads with higher homoeolog expression bias, being more dynamic across tissues, and having higher expression variation between cultivars than triads in the low-recombination proximal regions (Fig. 3G). This asymmetry is also reflected in contrasting distribution of histone marks and DNA methylation along chromosomes (10). The difference in epigenetic marks identified in leaves makes it tempting to speculate that epigenetic marks may also be associated with triad expression variation across multiple tissues.

We next investigated if divergence in spatial expression patterns (as measured by the mean distance statistic noted above) was coupled with 5' promoter and protein sequence divergence in syntenic triads (13). The 1.5 kb promoters of dynamic triads more frequently contained TEs (88.3% vs. 79.2%) which were closer to the translation start site (1,113 vs 1,234 bp), but were shorter than those in stable triads (median 220 bp vs. 259 bp) leading to equivalent TE densities (Fig. S20, Table S13; all comparisons Kruskal-Wallis P <0.001). These closer, more frequent, and shorter TEs could potentially act as novel cis-regulatory elements (34) or influence epigenetic marks (35). These results indicate that the promoter TE landscape relates more closely with the variation in the relative expression of homoeologs across tissues, rather than with a ubiquitous effect across all tissues (Table S13). Although only subtle differences were found in sequence identity between stable and dynamic triads (85.5 vs 85.0%; P =0.045)(Fig. 3H, Table S21), dynamic triads had fewer conserved TF binding site motifs across the three homoeologs (37% decrease, P <0.001; Fig. S21). Across coding sequences, we found a significant association

of both the nucleotide and protein identity distributions, which had a stepwise decrease in conservation from stable (average 97.2% CDS; 97.3% protein) to dynamic (95.0%; 93.4%; both *P* <0.001) triads (Table S21). We compared non-synonymous (Ka) to synonymous (Ks) substitution rates between homoeologs and observed that dynamic triads had significantly higher Ka/Ks compared to stable triads (0.33 vs. 0.21; Mann Whitney *P* <0.001, Fig. 3H) (Table S22). This higher ratio suggests that triads with greater divergence in spatial expression patterns are under more relaxed selection pressure, as seen for duplicated genes in humans (*9*), but not in soybean (*36*) and carp (*37*). This conclusion is supported by the observation that non-syntenic triads, which had greater expression divergence (10.5% larger mean distance; Mann Whitney *P* <0.001), also had significantly higher Ka/Ks (0.42; Mann Whitney *P* <0.001) compared to syntenic triads (Table S22). The above relationships were consistent using different percentage cut-offs to define stable and dynamic triads (5 and 25%), as well as in the cultivar Azhurnaya developmental time course (Tables S21–S22). These results show positive coupling of divergence in spatial expression patterns with divergence in TE and cis-regulatory elements in promoters and sequence divergence in coding sequence among wheat homoeologs. It is possible that divergence in spatial expression patterns, alongside relaxation of selection pressure, can lead to functional innovation through homoeolog neo- or sub-functionalization.

Coordinated expression of homoeolog triads

Our above analyses provide a framework to describe the relative expression of individual homoeologs between discrete triads in space and time. To understand how this coordination of homoeolog spatiotemporal expression may influence biological processes, we developed a series of co-expression networks to provide insight into tissue-specific developmental and stress-related processes.

We constructed four separate tissue-specific co-expression networks from non-stress RNA-Seq samples from grain (n = 119 samples), leaf (n = 245), root (n = 45), and spike (n = 128) using all genes expressed over 0.5 TPM in the given tissue (*13*). These networks were composed of 51–78 modules and contained 42.3% –88.0% of all expressed genes in each tissue (Fig. S22, Table S23, File S1). We found that across all tissue networks, homoeologs from 37.4% of the syntenic triads were in the same co-expression module, suggesting a highly coordinated expression pattern for

these triads (χ^2 P <0.0001 with respect to random triads, Table S24). However, the majority of triads (62.6%) had at least one homoeolog outside the same module.

To quantify whether homoeologs outside the module had 'similar' or 'divergent' expression patterns we calculated a threshold based on the pairwise distance between homoeologs (*13*). We found that 29.6% of syntenic triads had a divergent pattern wherein the expression of at least one homoeolog exceeded the distance threshold in the tissue network (Fig. 4A; Fig. S23). Conversely, 33% of triads had a similar pattern wherein all pairwise distances between homoeologs were lower than the threshold, suggesting a subtler variation in a single homoeolog. These values showed significant variation between tissue networks, ranging from 7% divergent triads in the leaf network to over 38% divergent triads in the root and grain networks (Fig. 4A). Non-syntenic triads had a higher proportion of divergent triads in all tissue networks compared to syntenic triads (mean 35.1% vs 29.6%, χ^2 P <0.001; Table S24). Using the same criteria as before (triad mean distance between tissues and global average position), we identified the 10% most stable and dynamic syntenic triads for each tissue-specific network. We found that dynamic triads were more frequently in divergent modules than stable triads for all four tissue networks (P <0.001; Fig. 4B; Fig. S24). These results are consistent with the above homoeolog expression bias analyses and support the idea that while many triads are expressed in a coordinated spatiotemporal pattern (with the same or similar profile), almost 30% of syntenic and 35% of non-syntenic triads have a divergent expression profile. Transcriptional divergence occurs both immediately upon polyploidization and following polyploidization (Figs. S12-S14) and may represent initial steps towards neo/sub-functionalization of wheat homoeologs.

Exploiting development and stress networks for biological discovery

To explore the potential for biological discovery, we first compared modules between networks to identify tissue-specific gene networks. Across the four networks, 73.2% of modules had significant overlap (P <0.05, Fisher's exact test) to modules in all four networks, with the root having the fewest conserved modules (61.1%) and the spike having the most (86.2%) (File S2). In the root, there were three modules that were not found in any other tissue, with the largest of these (root module 61; 82 genes) enriched for root-related plant ontology (PO) terms (e.g. root procambium P =3.3E-05 and central root cap of primary root P =4.5E-05; Table S25). We

hypothesized that transcription factors (TFs) controlling processes related to these PO terms would also be co-expressed within module 61. We found that four of the ten TFs in this root-specific module had known functions related to root development in Arabidopsis or rice (*38, 39*)(Table S26). Three of these TFs belonged to one homoeolog triad in the MADS_II family and one of their Arabidopsis orthologs (*AGL21*) has been shown to regulate lateral root development through auxin accumulation (*39*). To understand the target genes of these TFs in wheat we conducted a complementary network analysis using genie3, which predicted target genes of TFs across all 850 samples (*13*). Target genes of these three TFs were enriched for cell wall processes and lignification, consistent with their putative role in the differentiation zone where lateral roots emerge (Tables S27–S28). Closely related paralogs on chromosome group 6 in wheat were not located in root module 61; rather they were in modules 1 and 13 (Fig. 4C). These modules were conserved in all other tissue networks, implying a more general function for genes within them. Supporting this hypothesis, the rice ortholog of the chromosome 6 paralogs (*OsMADS57*) has been shown to play a role in tillering (*40*).

A significant challenge for wheat breeding is the selection of cultivars with tolerance to multiple stresses. Therefore, we focused on stress responses in seedlings and young vegetative plants, for which ten independent studies with twelve distinct abiotic and disease stresses were available (Table S1). We constructed gene co-expression networks for abiotic and disease stresses separately, including control samples from the same studies to allow for links between disease status and gene expression (*13*). We integrated the two networks to identify modules that might be common to both abiotic and disease responses. We found 84 pairs of modules between the two networks which had significantly overlapping gene content and were significantly correlated to both an abiotic and a disease stress (Tables S29–S31). The most significant overlap was between disease module 12 and abiotic module 2 (P =1.3E-94) which shared 355 genes (Fig. 5A). These two modules had similar enrichment for GO-slim terms relating to signal transduction and response to stimulus (Table S32), suggesting they might perform similar biological functions.

Among the 355 shared genes, there were 16 TFs, of which six have orthologs in rice or Arabidopsis with proven roles in abiotic or disease stress, and a further three have orthologs differentially expressed during stress in these species (Table S33). Furthermore, on the basis of the genie3 analysis, 11 of the 16 TFs have targets enriched in stress responses and seven have

targets enriched simultaneously in biotic and abiotic stress responses (Fig. 5B). Of these TFs, two homoeologs stood out as potential common regulatory components of abiotic and disease response: TraesCS5A01G237900 and TraesCS5B01G236400, which encode Heat Shock Factor (HSF) TFs. These two HSF TFs were in the top ten most central genes within disease module 12 (Table S34) as measured by intramodule connectivity, a value strongly correlated with the influence of a gene on a phenotype (*41*). The 387 predicted targets of these two TFs were frequently allocated to module 12 of the disease network (39.5%) and module 2 of the abiotic stress network (28.0%) (Table S35). The Arabidopsis ortholog of these genes, *TBF1*, was originally identified for its role in pathogen defense response (*42*), and has been shown to play a key role in the transition from growth to defense (*43*) while also positively regulating acquired thermotolerance (*44*). Recently, a '*TBF1* cassette' including the promoter and 5′ leader region of *TBF1* was used to engineer broad-spectrum disease resistance in both Arabidopsis and rice without a fitness cost (*45*). The fact that Arabidopsis *TBF1* is functional in rice suggests that this regulatory mechanism may be conserved across species, making the wheat orthologs identified here promising targets for further studies. These and other highly connected genes (Table S34) are strong candidates to control stress responses and the functions of their orthologs support this hypothesis. These results demonstrate the power of these datasets and that integrating gene networks from wheat, alongside phylogenetic relationships and knowledge of biological function from model species, can help identify target genes for further study in wheat.

Concluding statement

This study provides detailed insight into the spatiotemporal transcriptional landscape of polyploid wheat. We find evidence that the differences in relative expression among homoeologs observed in modern day wheat may have been established both upon the polyploidization of wheat itself and during the subsequent 10,000 years of polyploidy; this may have been determined through epigenetic changes affecting both DNA methylation as well as histone modifications. We identified asymmetries along wheat chromosomes for a series of features relating to homoeolog expression bias with important implications for breeding. Our work provides a framework for hypothesis generation in polyploid wheat which can now be experimentally tested using recent developments in sequenced mutant populations (*46*) and

genome editing approaches (*47*). Ultimately, this knowledge will help researchers and breeders modulate allelic variation across homoeologs to improve quantitative traits in polyploid wheat. This is an urgent task towards achieving global food security given that wheat provides over twenty percent of the protein and caloric intake (*48*) of humans.

Materials and Methods

RNA-seq samples

We included 246 samples previously described (*15*) and complemented this with 283 RNA-Seq samples which were deposited in the Short Reads Archive (SRA) between August, 2015 and January, 2017. An additional 321 RNA-Seq samples from six studies were used for this analysis and are detailed in Supplementary Materials (*13*).

Mapping of RNA-Seq reads to reference

For all 850 samples, metadata was assigned as described in (*15*), with high and low-level factors for tissue, age, variety, and stress. Due to the relatively large number of low level tissues (59), we further defined an intermediate level of tissues comprising 32 factors (average 26.5; median 12 replicates per factor) which was used for this study. We also assigned an intermediate level of stress comprising 15 factors (average 14.5; median 6 replicates per factor). We used kallisto v0.42.3 (*14*) to map the 850 RNA-Seq samples to the Chinese Spring RefSeqv1.0+UTR transcriptome reference. We used default parameters previously shown to result in accurate homoeolog-specific read mapping in polyploid wheat ((*15*); Fig. S1). We summarized expression levels from the transcript level to the gene level using tximport v1.2.0. We established the criteria that at least 1% of samples for a given gene all required to have expression values over 0.5 TPM for that gene to be considered expressed (*initial 850 filter*).

To confirm that kallisto enables homoeolog specific mapping (*15*) we analyzed expression of HC genes expressed >0.5 TPM in nulli-tetrasomic wheat lines from the publicly available study SRP028357 (*49*). The nulli-tetrasomic lines were missing an entire chromosome (1A, 1B, or 1D) which was replaced by a duplication of another homoeologous chromosome, e.g. Nulli1ATetra1B has 0 copies of 1A, 2 copies of 1B and 1 copy of 1D. We determined stringent homoeolog-specific mapping using a series of criteria detailed in Supplementary Materials (*13*).

Analyses of expressed genes

Starting from the subset of genes considered expressed using the *initial 850 filter* criterion, we determined genes which were expressed in at least one tissue within the Azhurnaya developmental time course (209 samples; 22 intermediate tissues) and Chinese Spring no stress (123 samples; 15 intermediate tissues) datasets. For this analysis, we first calculated the average

TPM expression of each gene in each of the intermediate tissue types (*average expression per tissue*). The number of samples that went into generating this *average expression per tissue* value varied for each intermediate tissue and are available in Table S1. We considered a gene expressed when its *average expression per tissue* was > 0.5 TPM in at least one intermediate tissue. For both datasets we focused on HC gene models (*10*). Whilst expression data was also assessed for LC genes, we excluded these from the main analysis to avoid confounding effects from pseudogenes and low-quality gene models. Through this analysis we found evidence of expression for 83,741 (75.6%) HC genes in Azhurnaya and 82,567 (74.5%) HC genes in Chinese Spring.

Using the *average expression per tissue* values, we also determined the global expression of each gene across all tissues in which it was expressed (based on the > 0.5 TPM criteria in the tissue). This generated an average value across tissues, rather than a geometric mean across all samples, to account for the variation in the number of samples per tissue. It also excludes tissues in which a gene is not expressed. This average across expressed tissues is referred to as either the "global analysis" or the "combined analysis (all tissues)" across the main text and in the supplementary materials and tables.

Relative expression levels of the A, B, and D genome homoeologs across triads

The analysis focused exclusively on the gene triads which had a 1:1:1 correspondence across the three homoeologous genomes, including 17,400 syntenic and 1,074 non-syntenic triads (total of 18,474 triads or 55,422 genes). Starting from the subset of genes considered expressed using the *initial 850 filter* criterion, we defined a *triad* as expressed when the sum of the A, B, and D genome homoeologs was > 0.5 TPM. This allowed us to include triads in which, for example, only a single homoeolog was expressed, and which could later be classified as a dominant triad. Using this criterion, we defined a total of 53,259 genes (17,753 triads) which were considered expressed (Table S3). To standardize the relative expression of each homoeolog across the triad, we normalized the absolute TPM for each gene within the triad as follows:

$$expression_A = \frac{TPM(A)}{TPM(A) + TPM(B) + TPM(D)}$$

$$expression_B = \frac{TPM(B)}{TPM(A) + TPM(B) + TPM(D)}$$

$$expression_D = \frac{TPM(D)}{TPM(A) + TPM(B) + TPM(D)}$$

Where A, B, and D represent the gene corresponding to the A, B, and D homoeologs in the triad. The normalized expression was calculated for each one of the intermediate tissues and for the average across all expressed tissues ("*combined analysis*" as described previously). Fig. S6 shows an example of these calculations for the roots and the combined analysis across three triads. The values of the relative contributions of each genome per triad were used to plot the ternary diagrams using the R package ggtern (*50*).

Definition of homoeolog expression bias categories

The ideal normalized expression bias for the seven categories was defined as:

	A	B	D
Balanced	0.33	0.33	0.33
A suppressed	0	0.5	0.5
B suppressed	0.5	0	0.5
D suppressed	0.5	0.5	0
A dominant	1	0	0
B dominant	0	1	0
D dominant	0	0	1

We calculated the Euclidean distance (rdist function from R 3.3.2) from the observed normalized expression of each triad to each of the seven ideal categories listed above. We assigned the homoeolog expression bias category for each triad by selecting the shortest distance. This was done for each of the intermediate tissue as well as for the average across all expressed tissues (*combined analysis*).

Analysis of the effects of polyploidy on homoeolog expression bias

We used RNA-Seq data (*25*) which consisted of two datasets based on RNA-Seq samples from the youngest leaf at fifth leaf stage. Dataset 1 (SHW1) included samples from tetraploid (BBAA) *Triticum turgidum* ssp. *turgidum* wheat accession AS2255, diploid *Ae. tauschii* (DD) accession AS60, and the synthetic hexaploid wheat (SHW1; BBAADD) resulting from the cross between the tetraploid and *Ae. tauschii* accessions. Dataset 2 (SHW2) consisted of tetraploid *T. turgidum* ssp. *durum* cv Langdon (BBAA), the same diploid *Ae. tauschii* (DD) accession AS60, and an independent synthetic hexaploid wheat (SHW2) derived from Langdon x AS60 (BBAADD). Note that AS2255 and Langdon are both *T. turgidum* ssp., but are defined as different subspecies based primarily on morphological features. These experiments recreate the polyploidization events that gave rise to modern bread wheat and the resulting SHW has the same genome composition as the CS and Azhurnaya datasets examined in this study.

We analyzed the RNA-Seq from both datasets by mapping reads to the CS RefSeqv1.0 transcriptome using the same bioinformatics pipeline as before (see 'Mapping of RNA-Seq reads to reference' section). However, for the tetraploid datasets we used only the A and B genome transcripts as a reference, for the diploid D genome datasets we used only D genome transcripts, and for the SHW datasets we used the complete RefSeqv1.0 transcriptome as the reference, as in CS and Azhurnaya. To generate the *expected* hexaploid wheat transcriptome based on progenitor species we weighted the TPM values from the tetraploid by 2/3 and the AS60 TPM values by 1/3 to maintain a total TPM of 10^6 in the combined dataset. The *in-silico* hexaploid wheat generated from the weighted tetraploid and diploid TPM values (referred to hereafter as the "*expected*" *in-silico* dataset) allows the direct comparison with the *observed* TPM values in SHW. We defined the seven homoeolog expression bias categories for both the *expected in-silico* and the *observed* SHW transcriptomes using the same methods as for CS and Azhurnaya and compared the classification of triads between the *observed* and *expected* datasets (Table S12). We next compared classifications to modern-day bread wheat CS and Azhurnaya. To enable a meaningful comparison across similar tissues from the Hao et al study (*25*) we used nine samples from the

PAMP Triggered Immune Response dataset from CS and six samples from the Azhurnaya dataset (Table S1). As before, we defined the seven homoeolog expression categories for the defined CS and Azhurnaya datasets and compared them with the SHW and the *in-silico* classifications (Table S11).

DNA methylation plant material and library preparation

Plants were grown as described in the *Chinese Spring tissues* study. The frozen leaves from the five samples at 3-leaf stage (Zadok stage 13) were ground and divided as input for the preparation of both RNA-Seq libraries (detailed in *Chinese Spring tissues* study) and whole genome bisulfite sequencing (WGBS) libraries. These samples enabled direct comparisons between the DNA methylation profile and homoeolog expression patterns in the same samples. WGBS libraries were constructed from purified nuclei prepared using the published methods (*51*). Input DNA was quantified using the Qubit high sensitivity DNA kit. A total of 500 ng of nuclear DNA was spiked with 270 pg of Lambda DNA to assess the conversion efficiency obtained using the EZ DNA Methylation-GoldTM Kit (Zymo research corp, Irvine, Ca, USA). WGBS libraries were prepared using the TruSeq DNA kit, (Illumina, Madison, WI) and 2×125 bp paired-end sequence reads was generated using the Illumina HiSeq 2500 v4 platform (Genome Quebec, Montreal,Qc, Canada). The data was deposited as SRP133674.

DNA methylation data analysis

Sequence quality and adaptor removal was performed using Trim_galore_v0.4.1 (*52*). High quality paired-end sequence reads were aligned to the RefSeqv1.0 Chinese Spring genome using Bismark version 0.16.1 (*53*) ensuring the removal of duplicate reads and only retaining unique unambiguous alignments. The data were processed to exclude regions with low coverage using a binomial test. The methylation data was annotated using the gene feature coordinates provided by the Chinese Spring v1.0 gene definitions. 5 kb flanking regions around the gene features were also extracted. The two flanking regions and the gene feature were each divided into 50 tiles (150 tiles in total) to summarize the observed methylation ratios. Data manipulation, statistical analysis and image generation were performed using the R language (*54*) utilizing the data.table (*55*), MethylKit v1.5.2 (*56*), genomation (*57*), and ggplot2 packages (*58*).

Comparison of RNA-Seq sample classification with DNA methylation

For the five RNA-Seq samples (from the same plants used for analyzing DNA methylation) we classified triads into the seven balanced, dominant, and suppressed categories using the same method as for previous analyses. We then classified homoeologs within dominant and suppressed triads into the "dominant" and "non-dominant" homoeologs, and "suppressed" and "non-suppressed" homoeologs. For example, in an A dominant triad the A genome is classified as "dominant" and B and D genomes are classified as "non-dominant" homoeologs. The DNA methylation patterns of genes in each of these categories were plotted using the methods described above (DNA methylation data analysis). Differences in DNA methylation levels between categories were tested pairwise using the non-parametric Mann-Whitney t-test using the wilcox.test() in R (Fig. S17).

Histone modification analysis

To study the role of histone modifications we carried out ChIP-Seq for three active marks (H3K36me3, H3K9ac, and H3K4me3) and one repressive mark (H3K27me3) (deposited under SRA accession number SRP126222). We used Chinese Spring at 3-leaf stage, however RNA-Seq data were not collected from these exact plants. To calculate the homoeolog expression bias we used Chinese Spring samples from a separate experiment (*PAMP-triggered immune response* study) in which the same tissue was collected at a similar stage (3-leaf stage). Whilst combining data from two separate experiments may introduce some noise into the analysis, the ChIP-Seq and RNA-Seq data are from similar tissues, at a similar growth stage, in the same wheat variety, and are thus highly comparable. Nevertheless, this confounding factor should be considered when interpreting these results. ChIP assays, DNA library preparation, and sequencing were performed as in (*10*).

Histone data analysis

Raw FASTQ files were preprocessed with Trimmomatic v0.36 (*59*) to remove Illumina sequencing adapters, trim 5' and 3' ends with quality score below 5 (Phred+33) and discard reads shorted than 20 bp after trimming. Paired-ends reads were aligned against IWGSC RefSeq v1.0 assembly using bowtie2 v2.3.3 with --very-sensitive settings (*60*). Alignments with MAPQ < 10 were discarded and duplicate reads removed with Picard MarkDuplicates (http://broadinstitute.github.io/picard/). Triad expression category was calculated using Chinese Spring samples from a separate experiment (*PAMP-triggered immune response* study) using the same method as in previous analyses. As in the DNA methylation analysis we classified homoeologs within dominant and suppressed triads into the "dominant" and "non-dominant" homoeologs, and "suppressed" and "non-suppressed" homoeologs.

We calculated meta-gene profiles for each category by computing the read density of each histone mark over different triads categories using Deeptools (*61*) *computeMatrix scale-regions* and plotted it with *plotProfile*. To make a statistical comparison, for each histone mark we scored the number of reads overlapping with gene bodies using *bedtools coverage –counts* (*62*). Only reads fully mapping within gene bodies were considered. To account for different gene size we divided the read counts over each gene by its length. The distributions of reads density over different triads categories were compared with a non-parametric t-test (Mann-Whitney U-test) using the function *wilcox.test* in R (Fig. S11).

Variation in homoeolog expression bias across tissues (Stable/Dynamic triads).

To define the variation in homoeolog expression bias of each triad across the intermediate tissues we calculated the Euclidean distance between the triad's global position (*combined analysis*) and each individual tissue in which the triad was considered expressed. We included only triads which were considered expressed in at least six tissues based on the *combined analysis* criteria outlined above. The average of these distances was defined as the "*triad mean distance*". We ranked triads by their *triad mean distance* and the percentile was calculated by:

$$percentile_i = truncate\left(\frac{rank(cmd_i)}{length(CMD)}\right)$$

where CMD is the vector containing all the '*triad mean distance*'. The first and last deciles where classified as stable 10% and dynamic 10% triads, respectively. A similar approach was used to define the correspond 5% and 25% extremes of the distribution. This analysis was conducted independently for the Chinese Spring no stress samples, the Azhurnaya developmental time course, and for each of the four tissue-specific networks. A visual representation is provided in Fig. S19.

TE presence in gene promoters

We extracted all TEs that were annotated to fall at least partly within 1.5 kb and 5 kb upstream of the canonical ATG start-codon for all genes. We then split the TEs into the relevant gene lists covering homoeolog expression bias variation (Stable 10%, Middle 80%, and Dynamic 10%) and homoeolog expression bias (Balanced, Dominant, Non-Dominant, Suppressed, and Non-Suppressed) based on the "combined analysis (all tissues)" for CS. We used these lists to identify the proportion of genes and triads in each category which contained at least one TE in the promoter region.

Enrichment of TE families in gene promoters

Using the GFF file of TE coordinates, we extracted TEs present in the promoter regions of HC genes. We retrieved all TE copies that are entirely or partially present in the 5 kb upstream of the ATG start-codon of the canonical transcript for each gene. We then calculated the number of genes in each of the Stable 10%, Middle 80%, and Dynamic 10% categories which contained specific TE families. We required the TE family to be present in at least 2% of the categorized genes for further analysis. We then found the deviation of this distribution from the expected 10-80-10 ratio using the χ^2 test, *P* values adjusted with Benjamini-Hochberg. We calculated the median length of each TE family based on all instances of that TE across the genome. We found fifteen TE families deviated significantly from the expected 10-80-10 distribution (Benjamini-Hochberg *P* < 0.01). However, the majority of these TE families were present in less than 5% of the genes considered, and showed very small variation in the number of promoters containing the TE, suggesting that the statistical significance may not be biologically relevant.

TE density in gene promoters

We calculated the density of TEs within 5 kb upstream of genes by calculating the proportion of TE bases in sliding windows of 100 bp with a step size of 10 bp. The mean of each window was then calculated for both the Stable 10%, Middle 80%, and Dynamic 10% triads and the genome dominance categories (Balanced, Dominant, Non-Dominant, Suppressed, and Non-Suppressed). Mann Whitney tests with Benjamini-Hochberg adjusted *P* values were used to test for differences in TE density between categories across each window.

Transcription factor binding site (TFBS) identification

The 1.5 kb of sequence upstream of the canonical ATG start-codon was used to identify transcription factor binding sites (TFBS) present in promoters of HC triads. The FIMO tool from

the MEME suite (v 4.11.4 (*63*)) was used with a position weight matrix (PWM) obtained from plantPAN 2.0 (*64*) to predict TFBS based on previously identified sites across multiple plant species. FIMO was run with a *P* value threshold of <1E-04 (default), --motif-pseudo set to 1E-08 as recommended for use with PWMs and a --max-stored-scores of 1,000,000 to account for the large size of the dataset. The background model was generated from all extracted promoter sequences using the MEME fasta-get-markov command. Details of the TFBS comparisons between homoeologs is presented in Supplementary Materials (*13*).

WGCNA Network construction

Co-expression networks were built for six separate sample sets: grain, leaf, spike, root, abiotic and disease (Table S1) using the WGCNA R package (*65*). For each network, we selected HC genes which were expressed > 0.5 TPM in three or more samples. The count expression level of each gene was normalized using variance stabilizing transformation from DESeq2 (*66*) to eliminate differences in sequencing depth between studies. The soft power threshold was calculated as the first power to exceed a scale-free topology fit index of 0.9 for each network separately. The soft powers used were: leaf =12, spike =12, roots =7, disease =7. For the abiotic and grain network the 0.9 threshold was not crossed until 15 and 20 respectively, which may be due to strong differences between samples within these datasets, therefore the soft power threshold was selected according to the number of samples, resulting in abiotic = 7 and grain = 6. Signed hybrid networks were constructed blockwise using the function blockwiseModules() with a maximum block size of 46,000 genes. The correlation type used was biweight mid-correlation "bicor" and the maxPOutliers was set to 0.05 to eliminate effects of outlier samples. The topographical overlap matrices (TOM) were calculated by the blockwiseModules function using TOMType = "unsigned" and the minimum module size was set to 30. The parameter mergeCutHeight =0.15 was used to merge similar modules.

Defining same, similar and divergent expression patterns of triads

For triads which had homoeologs within different modules in the WGCNA networks we developed a threshold to determine whether the different modules had a "similar" or "divergent" expression pattern. We calculated the Euclidean distance between module eigengenes using the R package dist() and with these values we calculate the distances between the homoeologs in each triad. Triads where the pairwise distances were zero were in the same module. Triads where the pairwise distances were over zero were in different modules. For these triads in different modules when the pairwise distance between any two homoeologs was > 50% of the median maximum distance between eigengenes, the triad was classified as having a "divergent" expression pattern. In cases where the pairwise distance was over zero between at least one pair of homoeologs and the distance between all three pairs of homoeologs were =< 50% of the median maximum distance, the triad was classified as having a "similar" expression pattern. The median maximum distance between eigengenes was averaged across all four tissue networks to give a final threshold (50% of median maximum distance) of 0.937431. This analysis was carried out for 1:1:1 syntenic and 1:1:1 non-syntenic triads expressed in each of the tissue networks (total triads = 9,599 grain, 5,378 leaf, 11,038 root and 6,173 spike). This excluded triads which had a putative transposable element (67 triads). Fig. S23 shows a graphic representation of this classification and the effect of altering the threshold in each of the four networks.

Module overlaps

Module overlap between networks was calculated using the R package GeneOverlap which calculates significant overlaps between modules using a Fisher's exact test. Modules were considered to have significant overlaps when the FDR adjusted *P* value <0.05.

Correlation to stress status

Modules within the abiotic and disease networks were tested for correlations to intermediate level stresses using the cor(function). The significant of correlations were calculated using the function corPvalueStudent() and corrected for multiple testing using p.adjust() using the Benjamini & Yekutieli method (*67*).

Genie3

HC genes expressed >5 TPM in at least one of the 850 sample were selected. Out of these 78,085 genes there were 3,386 transcription factors (methods described above). Random forest regression was estimated for each gene based on the transcription factors as inputs using the genie3 package (*68*) in R (version 3.3.2) with default parameters (K=sqrt, nb.trees=1000, input.idx=list of transcription factors, importance.measure=IncNodePurity, seed=NULL). For each transcription factor, all predicted target genes (connectivity >0.005) were extracted and functional enrichment within the target genes was determined using topGO (*69*)in R (version 3.3.2) with the following parameters (ontology="BP", nodeSize=10, classicFisher test $p<10^{-10}$). To summarize the results, the top three GO terms for each transcription factor, the *P* values for the strongest enrichment, and the direct blastx match in the Arabidopsis proteome (tair10) and well as the e-value and description were tabulated (Table S28). The complete list of GO term enrichments of the biological process ontology for each transcription factor and the list of transcription factors associated with each GO term in the ontology of biological process are published in e!DAL (https://doi.ipk-gatersleben.de/DOI/53148abd-26a1-4ede-802b-c2635af6a725/0dd8224a-34fc-4e3b-8ab8-883d07e52bd2/2/1847940088)

Identifying high connected hub genes

Hub genes within each module for the abiotic and disease stress networks were calculated using the WGCNA R package function signedKME(). This calculates the correlation between the expression patterns of each gene and the module eigengene. Genes which were more highly correlated to the eigengene were considered hub genes.

References and Notes (References (*70-88*) are only called out in the SOM).

1. W. Albertin, P. Marullo, Polyploidy in fungi: evolution after whole-genome duplication. *Proceedings of the Royal Society B: Biological Sciences*, (2012).
2. S. P. Otto, J. Whitton, Polyploid Incidence and Evolution. *Annual Review of Genetics* **34**, 401-437 (2000).
3. A. Salman-Minkov, N. Sabath, I. Mayrose, Whole-genome duplication as a key factor in crop domestication. *Nature Plants* **2**, 16115 (2016).
4. S. Renny-Byfield, J. F. Wendel, Doubling down on genomes: Polyploidy and crop plants. *American Journal of Botany* **101**, 1711-1725 (2014).

5.   M. Feldman, A. A. Levy, T. Fahima, A. Korol, Genomic asymmetry in allopolyploid plants: wheat as a model. *Journal of Experimental Botany* **63**, 5045-5059 (2012).
6.   Y. Van de Peer, E. Mizrachi, K. Marchal, The evolutionary significance of polyploidy. *Nat Rev Genet* **18**, 411-424 (2017).
7.   S. Ohno, *Evolution by Gene Duplication*. (Springer-Verlag, New York, 1970).
8.   F. A. Kondrashov, Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proceedings of the Royal Society B: Biological Sciences* **279**, 5048 (2012).
9.   K. D. Makova, W.-H. Li, Divergence in the Spatial Pattern of Gene Expression Between Human Duplicate Genes. *Genome Research* **13**, 1638-1645 (2003).
10.   IWGSC, Shifting the limits in wheat research and breeding using a fully annotated reference genome. (2018).
11.   E. Martinez-Perez, P. Shaw, G. Moore, The *Ph1* locus is needed to ensure specific somatic and meiotic centromere association. *Nature* **411**, 204-207 (2001).
12.   G. Moore, K. M. Devos, Z. Wang, M. D. Gale, Cereal Genome Evolution: Grasses, line up and form a circle. *Current Biology* **5**, 737-739 (1995).
13.   Materials and methods are available as supplementary materials at the *Science* website.
14.   N. L. Bray, H. Pimentel, P. Melsted, L. Pachter, Near-optimal probabilistic RNA-seq quantification. *Nat Biotech* **34**, 525-527 (2016).
15.   P. Borrill, R. Ramirez-Gonzalez, C. Uauy, expVIP: a customisable RNA-seq data analysis and visualisation platform. *Plant Physiology*, (2016).
16.   M. Melé *et al.*, The human transcriptome across tissues and individuals. *Science* **348**, 660 (2015).
17.   J. W. Walley *et al.*, Integration of omic networks in a developmental atlas of maize. *Science* **353**, 814 (2016).
18.   M. Pfeifer *et al.*, Genome interplay in the grain transcriptome of hexaploid bread wheat. *Science* **345**, (2014).
19.   D. Winter *et al.*, An "Electronic Fluorescent Pictograph" Browser for Exploring and Analyzing Large-Scale Biological Data Sets. *PLOS ONE* **2**, e718 (2007).
20.   P. Borrill, N. Adamski, C. Uauy, Genomics as the key to unlocking the polyploid potential of wheat. *New Phytologist* **208**, 1008-1022 (2015).
21.   R. Avni *et al.*, Functional characterization of *GPC-1* genes in hexaploid wheat. *Planta* **239**, 313-324 (2014).
22.   K. J. Simons *et al.*, Molecular Characterization of the Major Wheat Domestication Gene *Q*. *Genetics* **172**, 547 (2006).
23.   C. E. Grover *et al.*, Homoeolog expression bias and expression level dominance in allopolyploids. *New Phytologist* **196**, 966-971 (2012).
24.   S. Renny-Byfield *et al.*, Ancient Gene Duplicates in *Gossypium* (Cotton) Exhibit Near-Complete Expression Divergence. *Genome Biology and Evolution* **6**, 559-571 (2014).
25.   M. Hao *et al.*, The abundance of homoeologue transcripts is disrupted by hybridization and is partially restored by genome doubling in synthetic hexaploid wheat. *BMC Genomics* **18**, 149 (2017).
26.   M. J. Yoo, E. Szadkowski, J. F. Wendel, Homoeolog expression bias and expression level dominance in allopolyploid cotton. *Heredity* **110**, 171 (2012).
27.   P. P. Edger *et al.*, Subgenome dominance in an interspecific hybrid, synthetic allopolyploid, and a 140-year- old naturally established neo-allopolyploid monkeyflower. *The Plant Cell*, (2017).
28.   A. J. Bewick, R. J. Schmitz, Gene body DNA methylation in plants. *Current Opinion in Plant Biology* **36**, 103-110 (2017).
29.   D. Zilberman, An evolutionary case for functional gene body methylation in plants and animals. *Genome Biology* **18**, 87 (2017).
30.   A. J. Bewick *et al.*, On the origin and evolutionary consequences of gene body DNA methylation. *Proceedings of the National Academy of Sciences* **113**, 9111 (2016).
31.   X. Zhang *et al.*, Whole-Genome Analysis of Histone H3 Lysine 27 Trimethylation in Arabidopsis. *PLOS Biology* **5**, e129 (2007).
32.   E. D. Akhunov *et al.*, The Organization and Rate of Evolution of Wheat Genomes Are Correlated With Recombination Rates Along Chromosome Arms. *Genome Research* **13**, 753-763 (2003).
33.   J. A. Buggs Richard *et al.*, Tissue-specific silencing of homoeologs in natural populations of the recent allopolyploid Tragopogon mirus. *New Phytologist* **186**, 175-183 (2010).

34.   H. Zhao *et al.*, Proliferation of regulatory DNA elements derived from transposable elements in the maize genome. *Plant Physiology*, (2018).

35.   C. D. Hirsch, N. M. Springer, Transposable element influences on gene expression in plants. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1860**, 157-165 (2017).

36.   A. Roulin *et al.*, The fate of duplicated genes in a polyploid plant genome. *The Plant Journal* **73**, 143-153 (2013).

37.   J.-T. Li *et al.*, The fate of recent duplicated genes following a fourth-round whole genome duplication in a tetraploid fish, common carp (*Cyprinus carpio*). *Scientific Reports* **5**, 8199 (2015).

38.   C. Yu *et al.*, MADS-box Transcription Factor *OsMADS25* Regulates Root Development through Affection of Nitrate Accumulation in Rice. *PLOS ONE* **10**, e0135196 (2015).

39.   L.-H. Yu *et al.*, MADS-Box Transcription Factor AGL21 Regulates Lateral Root Development and Responds to Multiple External and Physiological Signals. *Molecular Plant* **7**, 1653-1669 (2014).

40.   S. Guo *et al.*, The interaction between OsMADS57 and OsTB1 modulates rice tillering via *DWARF14*. *Nature Communications* **4**, 1566 (2013).

41.   A. Ghazalpour *et al.*, Integrating Genetic and Network Analysis to Characterize Genes Related to Mouse Weight. *PLOS Genetics* **2**, e130 (2006).

42.   M. Kumar *et al.*, Heat Shock Factors HsfB1 and HsfB2b Are Involved in the Regulation of *Pdf1.2* Expression and Pathogen Resistance in *Arabidopsis*. *Molecular Plant* **2**, 152-165 (2009).

43.   K. M. Pajerowska-Mukhtar *et al.*, The HSF-like Transcription Factor TBF1 Is a Major Molecular Switch for Plant Growth-to-Defense Transition. *Current Biology* **22**, 103-112 (2012).

44.   M. Ikeda, N. Mitsuda, M. Ohme-Takagi, Arabidopsis HsfB1 and HsfB2b Act as Repressors of the Expression of Heat-Inducible *Hsfs*; But Positively Regulate the Acquired Thermotolerance. *Plant Physiology* **157**, 1243 (2011).

45.   G. Xu *et al.*, uORF-mediated translation allows engineered plant disease resistance without fitness costs. *Nature* **545**, 491-494 (2017).

46.   K. V. Krasileva *et al.*, Uncovering hidden variation in polyploid wheat. *Proceedings of the National Academy of Sciences* **114**, E913-E921 (2017).

47.   Y. Zhang *et al.*, Efficient and transgene-free genome editing in wheat through transient expression of CRISPR/Cas9 DNA or RNA. *Nature Communications* **7**, 12617 (2016).

48.   FAO, http://www.fao.org/faostat/.

49.   L. J. Leach *et al.*, Patterns of homoeologous gene expression shown by RNA sequencing in hexaploid bread wheat. *BMC Genomics* **15**, 276 (2014).

50.   N. Hamilton, ggtern: An Extension to 'ggplot2', for the creation of Ternary Diagrams. https://CRAN.R-project.org/package=ggtern, (2016).

51.   H. B. Zhang, X. Zhao, X. Ding, H. Paterson Andrew, A. Wing Rod, Preparation of megabase-size DNA from plant nuclei. *The Plant Journal* **7**, 175-184 (2002).

52.   S. Lindgreen, AdapterRemoval: easy cleaning of next-generation sequencing reads. *BMC Research Notes* **5**, 337 (2012).

53.   F. Krueger, S. R. Andrews, Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571-1572 (2011).

54.   R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/. , (2013).

55.   M. Dowle, M. Srinivasan, data.table: Extension of `data.frame`. R package version 1.10.4-3. . https://CRAN.R-project.org/package=data.table, (2017).

56.   A. Akalin *et al.*, methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biology* **13**, R87 (2012).

57.   A. Akalin, V. Franke, K. Vlahoviček, C. E. Mason, D. Schübeler, genomation: a toolkit to summarize, annotate and visualize genomic intervals. *Bioinformatics* **31**, 1127-1129 (2015).

58.   H. Wickham, ggplot2: Elegant Graphics for Data Analysis. *Springer-Verlag New York*, (2009).

59.   A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).

60.   B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357 (2012).

61.   F. Ramírez *et al.*, deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research* **44**, W160-W165 (2016).

62.   A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).

63. C. E. Grant, T. L. Bailey, W. S. Noble, FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017-1018 (2011).
64. C.-N. Chow *et al.*, PlantPAN 2.0: an update of plant promoter analysis navigator for reconstructing transcriptional regulatory networks in plants. *Nucleic Acids Research* **44**, D1154-D1160 (2016).
65. P. Langfelder, S. Horvath, WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
66. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550 (2014).
67. Y. Benjamini, D. Yekutieli, The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics* **29**, 1165-1188 (2001).
68. V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, P. Geurts, Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *PLOS ONE* **5**, e12776 (2010).
69. A. Alexa, J. Rahnenfuhrer, topGO: Enrichment Analysis for Gene Ontology. *R package version 2.30.0*, (2016).
70. H.-j. Schoonbeek *et al.*, Arabidopsis EF-Tu receptor enhances bacterial disease resistance in transgenic wheat. *New Phytologist* **206**, 606-613 (2015).
71. H. Webster, thesis, Murdoch University, (2014).
72. V. Chagué *et al.*, Genome-wide gene expression changes in genetically stable synthetic and natural wheat allohexaploids. *New Phytologist* **187**, 1181-1194 (2010).
73. R. Suzuki, H. Shimodaira, Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **22**, 1540-1542 (2006).
74. Y. Benjamini, Y. Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society* **57**, 289-300 (1995).
75. P. A. Wilkinson *et al.*, CerealsDB 3.0: expansion of resources and data integration. *BMC Bioinformatics* **17**, 256 (2016).
76. Y. Wang *et al.*, MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research* **40**, e49-e49 (2012).
77. S. Proost *et al.*, i-ADHoRe 3.0—fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Research* **40**, e11-e11 (2012).
78. B. J. Haas, A. L. Delcher, J. R. Wortman, S. L. Salzberg, DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics* **20**, 3643-3646 (2004).
79. D. Lang *et al.*, Genome-Wide Phylogenetic Comparative Analysis of Plant Transcriptional Regulation: A Timeline of Loss, Gain, Expansion, and Correlation with Complexity. *Genome Biology and Evolution* **2**, 488-503 (2010).
80. S. Lê, J. Josse, F. Husson, FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software* **1**, (2008).
81. K. P. Schliep, phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592-593 (2011).
82. M. D. Shirley, Z. Ma, B. S. Pedersen, S. J. Wheelan, Efficient "pythonic" access to FASTA files using pyfaidx. *PeerJ PrePrints* **3**, e970v971 (2015).
83. L. Venturini, S. Caim, G. Kaithakottil, D. L. Mapleson, D. Swarbreck, Leveraging multiple transcriptome assembly methods for improved gene structure annotation. *bioRxiv*, (2017).
84. N. Goto *et al.*, BioRuby: bioinformatics software for the Ruby programming language. *Bioinformatics* **26**, 2617-2619 (2010).
85. J. Daron *et al.*, Organization and evolution of transposable elements along the bread wheat chromosome 3B. *Genome Biology* **15**, 546 (2014).
86. K. Katoh, K. Misawa, K.-i. Kuma, T. Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* **30**, 3059-3066 (2002).
87. D. Charif, J. R. Lobry, in *Structural Approaches to Sequence Evolution: Molecules, Networks, Populations,* U. Bastolla, M. Porto, H. E. Roman, M. Vendruscolo, Eds. (Springer Berlin Heidelberg, Berlin, Heidelberg, 2007), pp. 207-232.
88. W.-H. Li, Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *Journal of Molecular Evolution* **36**, 96-99 (1993).

HS, BS, CJR, RA, VT, RD, FC, CP, NJP, AGS, and MS provided comments on the manuscript. All authors read and approved the final submission.

Competing interest: All authors declare no conflicts of interest.

Fig. 1. Developmental time course RNA-Seq expression profile in bread wheat. Schematic overview of tissues sampled for the RNA-Seq expression atlas across multiple growth stages (labelled in blue). Details of all samples are provided in Table S1.

Fig. 2. Homoeolog expression bias in syntenic homoeolog triads. (**A**) Ternary plot showing relative expression abundance of 16,746 syntenic triads (50,238 genes) in hexaploid wheat in the 15 tissues combined analysis from Chinese Spring. Each circle represents a gene triad with an A, B, D coordinate consisting of the relative contribution of each homoeolog to the overall triad expression (example in top left). Triads in vertices correspond to single-genome dominant categories, whereas triads close to edges and between vertices correspond to suppressed categories. Balanced triads are shown in gray. Boxplots indicate the relative contribution of each genome based on triad assignment to the seven categories. (**B**) Proportion of triads in each category of homoeolog expression bias across the 15 tissues. (**C**) Boxplot of absolute TPM expression abundance for each genome from the seven categories. (**D**) Number of tissues in which homoeolog-suppressed (brown), homoeolog-dominant (teal) and balanced (gray) triads are expressed. (**E**) Metagene profile for histone H3K27me3 marks from -2 kb upstream of the ATG until +2 kb downstream of the stop codon (normalized for gene length) for balanced triads (gray), dominant triads separated into dominant (teal) and non-dominant (pale blue) homoeologs, and suppressed triads separated into suppressed (tan) and non-suppressed (brown) homoeologs.

Fig. 3. Variation of triad expression patterns. (**A**) Variation of balanced, dominant, and suppressed triads (assigned based on global analysis) across 15 tissues. (**B**) Distribution of mean distance of triad variation across 15 tissues for 14,258 triads expressed in at least six tissues. The

10% most stable (blue) and 10% most dynamic (red) triads were defined. (**C**) Ternary plot of a representative stable and dynamic triad and (**D**) bar graph of the distance between the 15 individual tissues to the triad mean (horizontal line). (**E**) Number of tissues in which stable (blue) and dynamic (red) genes are expressed (Table S36). (**F**) Homoeolog expression bias classification of stable and dynamic triads in global analysis. (**G**) Schematic representation of wheat chromosome based on genomic compartments and features associated with distal (R1/R3) and interstitial and proximal (R2/C) regions. (**H**) Boxplots of percent coding and protein sequence identity (left) and Ka/Ks ratio (right) for stable 10% (blue), middle 80% (gray) and dynamic 10% (red) triads.

Fig. 4. Homoeolog co-expression patterns in tissue networks. (**A**) Triad assignment to the same, similar, and divergent modules in tissue co-expression networks. (**B**) Stable, middle, and dynamic triad assignment to the same, similar, and divergent modules in the root network. (**C**) Neighbor-joining phylogenetic tree of homologs for the Arabidopsis MADS_II gene *AGL21*. Wheat orthologs from chromosome group 2 are assigned to the root-specific module 61 (red), whereas chromosome group 6 orthologs are assigned to modules 1 and 13 (blue/purple).

Fig. 5. Overlapping modules within abiotic and disease stress networks. (**A**) Number of genes in abiotic module 2 and disease module 12 and the overlap between modules. The number of transcription factors is indicated in brackets. (**B**) Transcription factors found in both abiotic 2 and disease 12 (left) and the top five enriched GO terms of their targets as identified by genie3 (right).

Supplementary Materials:

Materials and Methods

Figs. S1 to S24

Tables S1 to S36

Data file S1 and S2

References (*70-88*)

coleoptile

seedling stage

1st leaf stage

1st leaf blade

1st leaf sheath

stem axis

radicle

root apical meristem

axillary roots

shoot apical meristem

3rd leaf stage

3rd leaf blade

3rd leaf sheath

5th leaf stage

5th leaf blade

5th leaf sheath

shoot apical meristem

tillering stage

shoot axis

root apical meristem

spikelet

awns

lemma

glumes

stigma

anthers

ovary

hard dough / ripening

embryo

endosperm

dough

ear emergence

milk grain / soft dough

anthesis

ligule

30% spike

peduncle

full boot

flag leaf sheath

flag leaf

flag leaf

shoot axis

internode

shoot axis

roots

A Balanced / Suppressed / Dominant ternary plots with Homoeolog Expression Bias Category legend:
- Balanced
- A suppressed
- B suppressed
- D suppressed
- A dominant
- B dominant
- D dominant

B Histogram of No. of triads vs Average distance between mean and tissue (%), with Stable 10%, Middle 80%, Dynamic10% categories.

C Stable / Dynamic ternary plots.

D Distance between mean and tissue (%) vs Tissue bar charts (Stable, Dynamic).

E Scaled density vs No. tissues (Stable 10%, Dynamic10%).

F Stable / Dynamic ternary plots.

G Schematic showing R1, R2a, C, R2b, R3 segments and trends:
	High	→	Infrequent
Recombination	High		Infrequent
Expression Breadth	Tissue Specific		Ubiquitous
GO Enrichment	Defence		Housekeeping
Homoeolog Expression Bias			
Within Tissue	Suppressed / Dominant		Balanced
Across Tissue	Dynamic		Stable
Inter-Cultivar	Variable		Conserved

H % Identity box plots (CDS, Protein) and Ka/Ks box plot (Protein) with Stable 10%, Middle 80%, Dynamic10%.

A

B

C

GLYMA09G33451
GLYMA01G02525
AT2G22630   *AGL17*
AT4G37940   *AGL21*
GLYMA02G38091
GLYMA14G36221
AT3G57230   *AGL16*
BRADI5G12450
BRADI5G12440
HORVU2Hr1G080490
TraesCS2A01G337900
TraesCS2B01G344000
TraesCS2D01G325000
Zm00001d003409
OS02G0579600   *MADS27*
BRADI3G46920
Zm00001d016957
Zm00001d050897
BRADI3G57017
HORVU6Hr1G073040
TraesCS6D01G273500
TraesCS6B01G322700
TraesCS6A01G292300
OS02G0731200   *MADS57*
Zm00001d017932
Zm00001d032218
OS04G0461300   *MADS61*
Zm00001d003408

A

Disease module 12

1338
(85)

355
(16)

4130
(200)

Abiotic module 2

B

TraesCS1D01G223600
TraesCS3D01G242500
TraesCS3D01G173300
TraesCS1A01G410500
TraesCS3B01G199000
TraesCS5B01G236400
TraesCS1D01G072900
TraesCS3A01G269900
TraesCS3A01G242500
TraesCS3D01G183900
TraesCS5A01G237900
TraesCS3A01G176500
TraesCS7D01G110600
TraesCS2B01G313600
TraesCS3B01G367500
TraesCS1D01G336400

response to abiotic stimulus
protein binding
peroxisome
response to extracellular stimulus
response to stress
response to external stimulus
response to biotic stimulus
response to endogenous stimulus
cell death
cell communication
signal transduction
kinase activity
endosome
signal transducer activity
receptor activity
cellular protein modification process
mitochondrion
growth
flower development
external encapsulating structure
endoplasmic reticulum
cell wall