Genomic variants affecting homoeologous gene expression dosage contribute to agronomic trait variation in allopolyploid wheat

Fei He, Wei Wang, William B. Rutter, Katherine W. Jordan, Jie Ren, Ellie Taagen, Noah DeWitt, Deepmala Sehgal, Sivakumar Sukumaran, Susanne Dreisigacker, Matthew Reynolds, Jytirmoy Halder, Sunish Kumar Sehgal, Shuyu Liu, Jianli Chen, Allan Fritz, Jason Cook, Gina Brown-Guedira, Mike Pumphrey, Arron Carter, Mark Sorrells, Jorge Dubcovsky, Matthew J. Hayden, Alina Akhunova, Peter L. Morrell, Les Szabo, Matthew Rouse & Eduard Akhunov

Allopolyploidy greatly expands the range of possible regulatory interactions among functionally redundant homoeologous genes. However, connection between the emerging regulatory complexity and expression and phenotypic diversity in polyploid crops remains elusive. Here, we use diverse wheat accessions to map expression quantitative trait loci (eQTL) and evaluate their effects on the population-scale variation in homoeolog expression dosage. The relative contribution of cis- and trans-eQTL to homoeolog expression variation is strongly affected by both selection and demographic events. Though trans-acting effects play major role in expression regulation, the expression dosage of homoeologs is largely influenced by cis-acting variants, which appear to be subjected to selection. The frequency and expression of homoeologous gene alleles showing strong expression dosage bias are predictive of variation in yield-related traits, and have likely been impacted by breeding for increased productivity. Our study highlights the importance of genomic variants affecting homoeolog expression dosage in shaping agronomic phenotypes and points at their potential utility for improving yield in polyploid crops.
Whole-genome duplications (WGD) can provide short-term evolutionary advantages and likely played an important role in the origin of most crops. Wheat experienced more than one round of WGD. Compared to their diploid relatives, polyploids have broader geographic distribution. This suggests the importance of WGD for improving the crops’ fitness in new environments, a factor that was critical for the spread of agriculture around the world. The success of polyploid crops can potentially be attributed to the adaptive genetic diversity contributed by ancestral species or generated de novo after WGD. The redundant genetic nature of polyploid genomes could promote the accumulation of novel variants without detrimental effects on fitness, consistent with the observed mutational robustness of polyploids.

Regulatory variants affecting gene expression levels play an important role in adaptive evolution and variation in complex traits. WGD greatly expands the range of possible interactions for regulatory variants controlling the expression of redundant homoeologous genes from different genomes. This contributes to substantial changes in the expression patterns of polyploids relative to their diploid ancestors. The redundant genetic nature of polyploid genomes is likely to result in increased genetic diversity and the potential for rapid adaptation to new environments.

Here, we perform expression quantitative trait loci (eQTL) analysis using geographically and genetically diverse allohexaploid wheat (genome formula AABBDD) accessions. The study use association mapping to identify cis- and trans-acting variants that explain the variance in homoeologous gene expression, with gene expression treated as a phenotype. We partition the genetic variation of gene expression traits using the approach of Yang et al. This approach can separate the genetic effects of portions of the genome on gene expression variance, and we use it to explore the relative contribution of variants from the three wheat genomes to population-scale transcript abundance variation from homoeologous genes. We show that the relative contribution of cis- and trans-acting variants to expression of homoeologous genes is affected by demographic events and selection. The relative expression dosage of homoeologous genes is primarily defined by the frequency of rare and common cis-regulatory variants whose accumulation is associated with biased homoeolog expression. The analyses of frequency, effect sizes, and levels of linkage disequilibrium between the cis-variants of homoeologous genes suggest that the relative homoeolog expression dosage is under selection. By investigating the distribution of cis- and trans-acting eQTL across genomic regions showing distinct epigenetic marks and chromatin architecture, we demonstrate that...

Fig. 1 Relative expression of homoeologous genes in the diverse panel of wheat lines. a Homoeologous gene pairs with matching and biased expression abundance of homoeolog A relative to homoeolog B. Red and green colors show low-expressing homoeologs in the A and B genomes, respectively. Increase in the frequency of accessions with a biased homoeolog is expected to reduce correlation between the levels of homoeolog expression measured in the panel. b Distribution of Spearman correlation coefficients (SCC) calculated between the AB, BD, AD gene pairs within the same homoeologous gene triplets (red) and random (black) pairs of genes using gene expression values from the 198 accessions. The mean of the sum of the total triplet expression (A + B + D) in groups of accessions with (y-axis) and without (x-axis) one of the gene copies downregulated. The red and blue dotted lines show the 1:1 and 2:3 combined expression ratios, respectively. Source data are provided as a Source Data file.
eQTL are enriched in the regions of active chromatin. Finally, summary-level eQTL and GWAS mapping data\(^29\), and gene co-expression networks (GCN) are analyzed jointly to study the role of variants linked with homoeologous gene regulation in shaping variation in major agronomic traits in wheat. We show that the frequency of genomic variants associated the relative expression dosage of homoeologous genes is predictive of variation in productivity traits in allopolyploid wheat and appear to be targeted by improvement selection. Thus, our study highlights the importance of WGD and emerging regulatory complexity in the evolution of phenotypic diversity that serves as a basis for the development of adapted crop varieties.

**Results**

Population-scale homoeologous gene expression variation. We collected RNA-seq data from total RNA isolated from 2-week-old seedlings of 198 diverse accessions (Supplementary Data 1) selected to represent the broad geographic and genetic diversity of bread wheat. An average of 65.7 million paired-end Illumina reads (2 × 100 bp) were collected for each sample, and after quality trimming mapped to the reference genome RefSeq v.1.0\(^3\). The proportion of reads unambiguously mapped to the individual wheat genomes was 81% (Supplementary Data 1). A simulation-based estimate suggested that the alignment settings used in our study provide 98% correct read mapping to the polyploid wheat genome (see “Methods”). Expression levels measured as Transcripts Per Million (TPM) were estimated for high-confidence (HC) genes in RefSeq v.1.0, with 52,511 transcripts (47,274 genes) showing TPM > 0.5 in at least three wheat lines (PRJNA670223) (Supplementary Data 2). In addition, we have analyzed RNA-seq previously generated for 90 wheat lines from spikes at the double-bread stage\(^30\).

In allopolyploid wheat, genes appear as homoeologs present in three (‘triplets’) or two copies or as singletons\(^18\). Compared to singletons, on average, homoeologs in triplets showed higher expression levels (ANOVA F-test = 87, df = 1, p = 2.2 × 10\(^{-16}\)), but lower expression variance (Supplementary Fig. 1a). The inter-genomic comparison of the population means of homoeolog expression in seedlings showed a positive correlation (Supplementary Fig. 1b), indicating that the relative expression levels of the most of homoeologs in our wheat panel tend to match. As a measure of the relative expression of homoeologs at the population level, we used Spearman Correlation Coefficient (SCC), which was calculated for each pair of homoeologs using their expression values in the panel of 198 accessions. While a strong positive correlation would be indicative of matching homoeolog expression levels (Fig. 1a) in most accessions in the panel, an increase in the proportion of accessions with biased homoeologs would decrease SCC (Fig. 1a). Compared to random pairs of genes selected from the distinct wheat genomes, the SCC distribution for the pairs of homoeologs was strongly shifted towards positive values (Fig. 1b, c), suggesting that the majority of accessions in the panel carry homoeologs with the matching levels of expression. The SCCs calculated for the same sets of homoeologs using RNA-seq data from both the seedlings and spike tissues\(^30\), collected from a distinct set of accessions were generally similar, suggesting that tissue-specific factors do not substantially affect co-expression of the majority of homoeologs at the population scale (Supplementary Fig. 1c).

In polyploids, the relative dosage of duplicated genes tends to be balanced\(^8,31,32\). To investigate whether the downregulation of one of the homoeologs in the population is compensated by increased expression of other homoeologs, we selected a set of 1443 gene triplets that met two criteria: (1) one out of three homoeologs was downregulated (TPM < 0.1) in at least two wheat lines, and (2) at least two wheat lines have all three homoeologs expressed (TPM > 2). We applied these criteria to each triplet to split 198 accessions into two groups, one group composed of accessions with one of the homoeologs downregulated and another group including accessions with all three homoeologs expressed. The sum of expression values from all three homoeologs (A + B + D) was calculated for each accession and used to derive the mean of total homoeologs’ expression for each group. In most cases, the mean expression ratio between these two groups across gene triplets (Fig. 1c) was below 1:1 but above 2:3, suggesting that decreased combined expression associated with homoeolog downregulation is not fully compensated by increase in the expression of other homoeologs.

Partitioning genetic variance for gene expression traits onto different wheat genomes. The genetic architecture of gene expression could be complex and driven by multiple cis- and trans-acting variants with a broad range of effect sizes. To account for the cumulative effect of all SNPs from the distinct wheat genomes on the expression of individual homoeologs, we performed partitioning of genetic variation (Fig. 2a)\(^3\). For this purpose, 2,021,936 SNPs with MAF > 0.05 identified in our panel were grouped into three genome-specific sets from the A, B, and D genomes. These sets were used to build genetic relationship matrices and estimate the genetic variance\(^25\) for the expression of each gene (Supplementary Data 3). SNPs located within the same genome as a gene of interest were referred to as cis-genomic SNPs, whereas SNPs located in other genomes were referred to as trans-genomic SNPs (Fig. 2a). For the 10,000 most highly expressed genes, the mean of gene expression variance explained by the entire SNP set from all three wheat genomes was 40.4% (Supplementary Fig. 2a and 2b). The D genome explained a lower proportion of variance (7.7%) in gene expression than either the A (19.1%) or B (13.6%) genomes (Mann–Whitney test W\(_{A/D} = 23,172,000\), p-value < 2.2 × 10\(^{-16}\); W\(_{B/D} = 41,122,000\), p-value < 2.2 × 10\(^{-16}\)) (Supplementary Figs. 2a and 2b).

We found that, on average, cis-genomic SNPs from the A or B genomes explained a higher proportion of gene expression variance (21.7% and 28.7%, respectively) than trans-genomic SNPs (5–17%) from other genomes (Fig. 2a). However, in the evolutionarily younger D genome\(^3\), the proportion of variance explained by the cis-genomic SNPs (12%) was only slightly higher than that explained by the trans-genomic SNPs from the A (11%) and B (9%) genomes (Fig. 2a). The results of variance partitioning in the developing spikes\(^30\) were consistent with the results obtained in seedlings (Fig. 2a).

The expression variance of a number of genes was largely explained by the trans-genomic SNPs (Supplementary Data 3). Among 34,691 genes with at least 20% of the total variance in gene expression explained jointly by SNPs from all three genomes, 6173 genes (17.8%) had <1% of variance explained by the cis-genomic SNPs, with the remaining variance explained by the trans-genomic SNPs. While this group of genes showed significantly reduced cis-regulatory diversity compared to trans-regulated genes (Fig. 2b), no significant reduction of diversity between cis- and trans-only regulated genes (Wilcoxon rank sum test p-value = 0.1) was found in wild emmer using data from the previously published study\(^8\), indicating that shift towards trans-regulation in these genes from the A and B genomes is likely associated with diversity loss during wheat improvement. Of these 6173 genes, 47.8% were located in the D genome, which shows reduced diversity (\(\pi_D = 0.0003\)) relative to the A and B genomes (\(\pi_A = 0.0007\) and \(\pi_B = 0.0010\))\(^3\). This result also indicates that polyploidization bottleneck, which disproportionally affected the D genome, explains most of the cis-regulatory diversity loss in wheat.
We further split all SNPs into cis-genic (SNPs within ±1 Mb region around genes) and trans-genic (SNPs outside of ±5 Mb region around genes and SNPs located on other wheat chromosomes) subsets. The partitioning of variance for gene expression showed that the average variance explained by trans-acting variants (19.3%) is substantially higher (Wilcoxon test, \( W = 3,228,268 \)) than that explained by cis-acting variants (12.3%). We also compared the means of total variance explained for two groups of genes: (1) genes with variance explained only by cis-genic SNP, where trans-genic SNPs contribute <1% to variance, and (2) genes with variance explained only trans-genic SNPs, where cis-genic SNPs contribute <1% to variance. The means of variance in these two groups of genes were similar (47.0% vs. 46.4%), suggesting that in the cases of cis-regulatory diversity loss in the allopolyploid genome, the contribution of trans-genic variants to expression variance is similar to the contribution of cis-genic variants.

The effects of cis- and trans-acting variants on expression correlation between homoeologs. The accumulation of rare and common mutations in gene promoters leads to dysregulation of gene expression in diploid genomes.\(^{35}\) Even though polyploidy was expected to increase the mutation load in wheat\(^{118}\) and result in expression dosage bias, the majority of homoeolog pairs in our study showed matching levels of expression. To better understand the genetic basis of homoeolog co-expression, we compared the proportions of expression variance in individual homoeologs explained by cis- and trans-genic SNPs among pairs showing the distinct levels of expression correlation (SCC) (Fig. 2c). An increase in SCC was accompanied by a decrease in the total variance explained, with the largest proportion of explained variance observed for homoeologs with SCC < 0 (Fig. 2c). While SCC increase was accompanied by thousand decrease in variance explained by trans-genic SNPs, more substantial 11-fold decrease in variance explained was observed for cis-genic SNPs, reaching only 1.6% for homoeologs showing high correlation in the expression levels (SCC > 0.90) (Fig. 2c). These results suggest that discordant expression of homoeologs in the panel is likely associated with the accumulation of cis- rather than trans-regulatory diversity affecting the homoeologous genes. This conclusion is consistent with a decrease in the number of common (MAF > 0.05) and rare (MAF < 0.05) SNPs around the homoeologs with discordant expression of homoeologs in the panel (Supplementary Fig. 3a and 3b), indicating that negative relationship between cis-genic diversity and SCC is independent of the levels of genetic diversity in individual genomes. The lack of strong relationship between the inter-genomic sequence divergence and SCC suggests that divergence in the regulatory regions inherited from diploid ancestors likely has substantial global impact on the relative levels of homoeolog expression (Supplementary Fig. 3c).

To assess the impact of rare cis-variants (MAF < 0.05) on the homoeolog expression levels, we investigated the relationship between the expression ranks of each homoeolog in the population, ordered from lowest to highest expression levels across wheat accessions, and the rare allele load in the upstream
5-kb regulatory regions. In humans and maize\textsuperscript{35,36}, the total number of rare alleles in the regulatory regions coincides with the extremely low or high levels of expression in the population. We observed a similar trend using the entire set of homoeologous genes (Supplementary Fig. 2c). However, a subset of homoeologs showing high levels of expression correlation (SCC > 0.8) showed no enrichment for down- or upregulating rare alleles (Supplementary Fig. 2d). Combined together, our results suggest that the biased expression of homoeologs in the panel is primarily associated with the accumulation of both common and rare cis-acting variants.

**Mapping and functional annotation of variants affecting gene expression variation.** We conducted genetic mapping of cis- and trans-acting variants (eQTL) associated with the expression variation of individual genes among 191 wheat lines (Supplementary Data 1). After LD-based merging ($r^2 \geq 0.2$), we identified 36,898 and 15,238 significant SNPs (FDR $< 10^{-5}$) in the RNA-seq datasets for wheat seedlings and spikes, respectively (Supplementary Data 4 and 5; Supplementary Figs. 4 and 5). A conservative criteria was applied to define trans-eQTL as eQTL located in different genomes or chromosomes relative to the target gene, and cis-eQTL as eQTL located $\pm 1$ Mb around a target gene. According to these criteria, in the seedlings, 8568 cis-eQTL and 14,645 trans-eQTL were associated with the expression of 8315 (8837 transcripts) and 8255 (8500 transcripts) genes, respectively. Out of these cis- and trans-eQTL in the seedlings, 247 eQTL affecting the expression of 1,469 genes (1500 transcripts) overlapped. In the RNA-seq data from spikes\textsuperscript{30}, we identified 3172 cis-eQTL for 3476 transcripts, and 9891 trans-eQTL for 7250 transcripts. The location of a cis-eQTL density peak averaged across all target genes relative to the coding sequence start site was similar between the A and B genomes but was $\sim$200 bp more distant for the D genome (Supplementary Figs. 6a, 6b) and likely to be a consequence of lower diversity and more extended LD in the D genome\textsuperscript{34}.

The functional properties of identified eQTL were evaluated by calculating their enrichment within the specific regions of the genome tentatively affecting the coding potential of a gene or its regulation. The effects of SNPs on coding sequences were assessed using SNPeffect\textsuperscript{37}. SNPs resulting in splice-site disruption and premature termination codons were considered as putatively deleterious. The regulatory regions were previously defined based on distinct epigenetic marks and open chromatin using a combination of MNase digest\textsuperscript{26}, DNaseI digest, and combined analyses of epigenetic variation, chromatin immunoprecipitation, and DNase-seq data\textsuperscript{27}. The greatest levels of cis- and trans-eQTL enrichment relative to all variants in the genome were found for putatively deleterious variants, followed by missense and synonymous variants (Fig. 3a). Both cis- and trans-eQTL were found enriched in the regulatory regions (Fig. 3b, c) and depleted in the regions of closed chromatin hyper-resistant to MNase treatment (Fig. 3b). The cis- and trans-eQTL showed similar levels of enrichment across the various epigenetic marks, with both types of eQTL enriched for epigenetic modifications associated with gene body (H3K4me1), transcription (H3K36me3), and active expression (H3K27ac, H3K4me3) (Fig. 3e, f)\textsuperscript{27,38,39}. Simultaneously, we observed a depletion of both cis- and trans-eQTL within epigenetic marks often associated with the repression of gene expression (H3K27me3) or transposable elements (H3K9me2) (Fig. 3c)\textsuperscript{40,41}.

Recent studies indicate that interaction between distant regulatory elements and their target genes could be facilitated by 3D chromatin contacts\textsuperscript{42}, which is consistent with the enrichment of cis-eQTL within regions involved in the formation of chromatin loops in humans and maize\textsuperscript{42,43}. To investigate the potential involvement of chromatin loops in gene expression regulation in wheat, we compared the distribution of eQTL-target gene pairs across the interacting regions identified by Hi-C\textsuperscript{28} in cultivar Chinese Spring, which was not part of our diversity panel. Although, the low resolution of wheat Hi-C data does not allow us to map precisely regulatory regions involved in interaction, it could be used to assess the enrichment of eQTL-target gene pairs within chromatin loops relative to randomized data. First, we found that both cis- and trans-eQTL p-values positively correlate with the frequency of Hi-C contacts. This suggests that regions harboring eQTL-gene pairs showing stronger association are also more likely to have a higher frequency of chromatin contacts than regions harboring eQTL-gene pairs showing weaker associations (Fig. 3d, e, Supplementary Data 6). Second, the regions harboring trans-eQTL between both homoeologous and non-homoeologous chromosomes showed elevated Hi-C contacts (log10[Hi-C] $= 1.24$) compared to a distribution based on the 100 randomized samples (mean log10[Hi-C] $= 0.92$) (Fig. 3g). This result indicates that the probability of trans-eQTL-target gene pair occurrence within the chromatin loops is substantially higher than within the randomly selected regions. Among trans-eQTL-target gene pairs with a Hi-C contact frequency $> 50$, 15% were located with the homoeologous chromosome regions, which are involved in chromatin interaction more frequently than non-syntetic regions (Fig. 3f, Supplementary Data 6)\textsuperscript{28}.

**Genetic architecture of homoeologous gene expression regulation.** To better understand the role of polyploidy in the regulation of homoeologous gene expression, we analyzed the genomic distribution of trans-eQTL and their gene targets. The total number of trans-eQTL in the A and B genomes for target genes in the same genomes was similar (Fig. 4a). However, the total number of trans-eQTL in the A and B genomes targeting genes in the D genome was 4.0 and 3.6 times higher than the total number of trans-eQTL in the D genome targeting genes in the A and B genomes, respectively (Fig. 4a). These results are consistent with the differences in the levels of genetic diversity between the wheat genomes\textsuperscript{34} that also contributed to differences in the proportions of genetic variance for gene expression explained by SNPs from different genomes (Fig. 2a).

Also, we observed a tendency for the co-localization of trans-eQTL and target genes in the syntenic regions of homoeologous chromosomes (Fig. 4b, Supplementary Figs. 4 and 5). These patterns of trans-eQTL-target gene distribution are likely associated with the presence of shared regulatory elements in homoeologous genes that influence their co-regulation by regulatory feedback loops conserved among homoeologs, as was demonstrated for the three homoeologs of the Vrn-1 gene\textsuperscript{44}. This hypothesis is supported the finding that, in a set of the 6,371 homoeologous gene triplets, 23% of homoeologs shared at least one eQTL, and that correlation in expression levels between the homoeologs\textsuperscript{34} that also contributed to differences in the proportions of genetic variance explained by SNPs from different genomes (Fig. 2a).

Prior studies showed that gene expression is under purifying selection\textsuperscript{45,46}. However, it remains unclear how genetic redundancy provided by polyploidy would affect selection on the expression of homoeologs. To answer this question, we compared the relationship between the minor allele frequency and the effect size of cis-eQTL for two groups of genes, homoeologs and singletons. Purifying selection acting against cis-eQTL with strong effects on gene expression results in negative correlation between cis-eQTL minor allele frequency and effect size\textsuperscript{45}. We argue that if genetic redundancy influences the strength of purifying selection on the expression of duplicated genes, we expect to see the
reduced or lack of correlation between the allele frequency and effect size in homoeologs compared to singletons. However, we found that both groups of genes showed significant negative correlations (homoeologs: SCC = −0.23, p < 2.2e−16; singletons: SCC = −0.20, p = 0.002) between the cis-eQTL minor allele frequency and effect size (Fig. 4d). There was no significant difference between the two correlation coefficients (Fisher’s z-test: z = 0.53, p-value = 0.60). The negative relationship between frequency and effect size was observed in homoeologs even for the subset of cis-eQTL whose effects are detectable at all frequencies (Supplementary Fig. 7). These results indicate that the expression levels of both homoeologs and single-copy genes are likely under purifying selection. Compared to homoeologs, the single-copy genes had cis-eQTL effect sizes higher across all MAF classes, likely due to the increased contribution of trans-acting variants to the expression variation of homoeologs compared to that of the single-copy genes.

To understand the effects of cis- and trans-eQTL on the relative levels of homoeolog expression in our panel, we investigated the distribution of expression correlation values between the pairs of homoeologs (SCC) for the sets of homoeologs grouped based on the following eQTL-target gene configurations: (1) a homoeolog pair has no eQTL, (2) a homoeolog pair is regulated by shared trans-eQTL and has no cis-eQTL, (3) each gene within a homoeolog pair is regulated only by cis-eQTL, and (4) one of the genes within a homoeolog pair has cis-eQTL that also acts as a trans-eQTL for another homoeolog (Fig. 4e, Supplementary Data 7). We found that homoeolog pairs with no eQTL associated with their expression levels showed high levels of expression correlation (mean SCC = 0.57), Fig. 3 Functional annotation of eQTL. The x-axis represents eQTL enrichment expressed as the log2 of the proportion of eQTL within specific classes of SNP variants (y-axis) relative to the proportion of eQTL within the random samples of SNPs. All analyses of enrichment and Hi-C contacts are based on 2,021,936 SNPs, 14,645 trans-eQTL and 8568 cis-eQTL. a Mean enrichment of cis-eQTL and trans-eQTL among SNP variants from different functional classes defined based on gene and transposable element (TE) annotation (N = 2,021,936 SNP sites). b Mean enrichment of cis-eQTL and trans-eQTL in the MNase Sensitive Footprints (MSF), MNase Resistant Footprints (MRF), and regulatory regions identified based on the sensitivity to DNase I treatment, epigenetic variation, and open chromatin marks. Enrichment was assessed relative to genome-wide patterns, except for MSF vs MRF, where enrichment was tested for eQTL located within MSF relative to MRF. ‘Reg’ region corresponds to regulatory elements, as defined in Li et al.27. ‘Reg2’ corresponds to regions classified as states 5–7 in the same study. c Mean enrichment of cis-eQTL and trans-eQTL in the epigenetically marked regions. d Relationship between the binned trans-eQTL p-values and frequency of Hi-C contacts between regions harboring a trans-eQTL and its target gene. Data are presented as mean values ± SEM. e Relationship between the binned distal (>2 Mbp) cis-eQTL-target gene p-values and frequency of Hi-C contacts between regions harboring a cis-eQTL and its target gene. Data are presented as mean values ± SEM. f Comparison of Hi-C contacts (log10-transformed) between regions harboring trans-eQTL-target gene pairs to Hi-C contacts between a random set of genes (two-sided Wilcoxon rank-sum test, W = 191,789,322, p-value = 8.8e−16). g Comparison of Hi-C contacts (log10-transformed) between the pairs of homoeologs and random pairs of genes.
consistent with the observed decrease in both SNP diversity (Fig. 2d) and expression variance explained by SNPs (Fig. 2c) with an increase in the homoeolog expression correlation. The expression correlation between homoeologs regulated only by shared trans-eQTL without effects detected from cis-eQTL showed bi-modal distribution with most homoeologs having a strong positive correlation (mean SCC = 0.59) (Fig. 4e); the subgroup with a higher correlation showed less diversity in the cis-regions than the subgroup with the lower correlations. These findings were consistent with the observation that the expression variance of the highly correlated homoeologs is decreased largely by the trans-genic variants (Fig. 2c). Compared to these two eQTL-target gene configurations (1 and 2), more than a threefold reduction in the level of expression correlation between the homoeologs (SCC = 0.17) was associated with the presence of cis-eQTL, which appear to change the relative expression dosage of homoeologs in the panel. The set of homoeologs having variants acting as cis-eQTL for one homoeolog and trans-eQTL for another homoeolog showed a negative expression correlation (SCC = −0.26) (Fig. 4e). These cases represent an extreme form of homoeolog expression bias, where a high-expressing homoeolog in one genome is often associated with a low-expressing homoeolog on another genome (Fig. 1a).

**Joint eQTL and GWAS analyses of agronomic traits.** Recent studies have demonstrated the utility of gene expression data for interpreting GWAS results and identifying candidate causal genes by jointly analyzing eQTL and SNPs linked with trait variation. Consistent with these findings, we found a significant enrichment of cis-eQTL among SNPs associated with variation in yield component and development traits in a diverse set of wheat lines (Supplementary Fig. 8, Supplementary Data 8). In addition, we used the results of QTL mapping in bi-parental populations and diversity panels to assess the overlap of eQTL detected in our study with significant marker-trait associations (MTAs) for a number of agronomic traits identified in WheatCAP (www.triticeaecap.org) and IWYP (iwyp.org) projects (Supplementary Note 1, Supplementary Data 10). Using strict criteria for overlap (±1 kb), out of 1,112 non-redundant MTAs, 70 and 36 MTAs had cis-eQTL and trans-eQTL located within ±1 kb, respectively. While for trans-eQTL, this overlap was not substantially different from the randomized control, for cis-eQTL, this overlap was nearly two times higher than the maximum overlap of 33 eQTL obtained in the randomized control (Supplementary Fig. 8d and 8e). Consistent with an earlier study, these results suggest that the trait-associated SNPs are more likely to be cis-regulatory rather than trans-regulatory variants.

Further, we used GWAS and eQTL summary-level Mendelian Randomization (SMR) data analysis to detect candidate genes whose expression levels co-vary with phenotypes due to pleiotropy or causal association (Fig. 5a). We obtained the SNP effects by performing GWAS for 14 productivity traits evaluated in two wheat populations, one characterized in this current study (see “Methods”) and another characterized in the 1000 wheat exomes project. By applying SMR to test for association between gene expression and productivity traits, we detected 971 and 424
and SNPs within a 2-Mb window, including another homoeologue. Only LD values above Supplementary Figs. 9 and 10, Supplementary Data 12 provided as a Source Data within different LD ranges (the dotted line shows a ratio of 1.0) estimated for the negatively correlated and random pairs of homoeologs. Source data are pFDR-corrected regulatory targets in the wheat eQTL map (see "Methods").

In rice, TaSPL14 was linked with increased panicle branching and yield, but no TaSPL14 alleles positively affecting productivity traits in wheat were reported. Here, we show that the TaSPL14 transcription factor (TaesCS5B01G512800) is associated with variation in spike compactness, grain length, and harvest weight (Fig. 5b, Supplementary Data 11), consistent with the decreased expression levels of these homoeologs in the panel (SCC = −0.18, p-value = 0.01) (Fig. 5c). The higher expression of TaElf3-B1 in lines with the TaElf3-D1 deletion than in the lines without the deletion (t-test p-value = 2 × 10^−4) (Supplementary Fig. 11b).

**Fig. 5 Joint eQTL and GWAS analysis of agronomic traits in wheat.** a Outline of the strategy used to integrate eQTL and GWAS data to investigate the genetic basis of yield component and development trait variation in wheat (see "Methods"). We used field-based phenotyping data collected for a diverse panel of ~800 wheat accessions from the 1000 Wheat Exomes project including grain filling period (GFP), harvest weight (HW), drought susceptibility index for harvest weight (HWS), heading date (HD) and plant height (PHT) traits. A set of phenotypic traits was collected for a diverse panel of 400 wheat accessions: grain length (GL), grain width (GW), thousand-grain weight (TGW), grain area (GRA), spike compactness (SPC), spikelet number per spike (SPN), awnedness (AWN), and height (PHT) (Supplementary Data 10). b Gene co-expression network (GCN) modules, including TaSPL14 and Elf3 genes, are enriched for genes associated with agronomic traits in SMR analyses. c eQTL located on chr. 1D acts as a cis-variant for Elf3-D1:TraesCS1D01G451200 (GWAS FDR-corrected p-value = 4e−54) and is tightly linked (r^2 > 0.8) with variants acting as trans-variants for Elf3-B1:TraesCS1B01G477400 (GWAS FDR-corrected p-value = 9e−10). A deletion of Elf3-D1 (Elf3-D1 locus) in wheat affects both HD and SPN traits. d Distribution of LD between negatively correlated homoeologs compared to LD between the random set of homoeolog pairs. LD was measured between the cis/trans-eQTL in one homoeologue and SNPs within a 2-Mb window, including another homoeologue. Only LD values above r^2 > 0.2 were plotted. The right y-axis shows the ratio of SNP pairs within different LD ranges (the dotted line shows a ratio of 1.0) estimated for the negatively correlated and random pairs of homoeologs. Source data are provided as a Source Data file.

**Methods**
suggests that the level of TaElf3-B1 expression could be conditioned by the allelic state of the TaElf3-D1 homoeolog. This result combined with the presence of 30 SNPs near TaElf3-B1 showing high LD ($r^2 > 0.4$) with the trans-eQTL on 1D indicates that the combinations of the TaElf3-D1 and TaElf3-B1 alleles in our panel could be non-random.

Allelic combinations of homoeologs showing negative expression correlation in the panel is not random. Previous studies indicate that the non-random combinations of alleles showing elevated inter-locus LD could arise due to population structure, selection, or epistasis. In our study, we identified a number of homozygous gene pairs whose expression levels, similar to the TaElf3-D1 and TaElf3-B1 homoeologs, show a negative correlation in our panel, which results from the presence of alleles carrying homoeologs showing both matching and biased expression levels in the panel (Fig. 1a). To test whether negative expression correlation between homoeologs is associated with the non-random combinations of homoeologous alleles, we estimated inter-chromosomal LD between SNPs located near the pairs of negatively correlated homoeologs showing SCC $< -0.4$ (59 homoeologs in total). LD was calculated between the cis-eQTL associated with variation in the expression of one of the homoeologs and SNPs located within a 2-Mb region around another homoeolog (Fig. 5d). To take into account the effects of population structure, which could also lead to elevated inter-locus LD, we calculated LD between SNPs from the randomly selected pairs of homoeologs (Fig. 5d). Compared to this random set of SNPs, the regions harboring the negatively correlated homoeologs still showed nearly a two-fold increase in the proportion of high-LD SNPs ($r^2 > 0.4$) (Fig. 5d), suggesting that the allelic combinations of negatively correlated homoeologs in the panel could result from selection.

Accumulation of homoelogous alleles with biased expression affects agronomic traits. Prior studies suggest that some adaptive traits in wheat could be impacted by the relative expression of the homozygous copies of causal genes. To better understand the overall impact of the biased homoeolog expression on phenotype, we investigated the relationship between productivity traits and the number of homoeologous alleles resulting in biased expression in the set of 59 negatively correlated homoeologs (SCC $< -0.4$), which were identified in the RNA-seq data from the seedlings (Supplementary Data 15). In the RNA-seq data from spikes, these homoeologs also showed the lack of coordinated expression (mean SCC $= 0.03 \pm 0.05$), although not as substantial as in the seedlings. The majority of these homoeologs had low-expressing alleles in at least two wheat lines. On average, for this set of homoeologs, we detected eight low-expressing alleles per line, ranging from 1 to 31 per line in the panel (Fig. 6a–c, Supplementary Data 15). The minor allele frequency (MAF) of cis-eQTL associated with these negatively correlated homoeologs was shifted towards common variants, with the mean MAF of 0.30 $\pm 0.01$ (Supplementary Fig. 12), indicating that the homoeologous alleles contributing to biased expression are present at high frequency in our panel. To assess whether the low-expressing alleles of homoeologs are associated with gene deletions, similar to the deletion of the TaElf3-D1 homoeolog (Fig. 5c), we compared the sequences of homoeologs with the wheat PanGenome. This analysis showed that the lack of transcripts from only two of 59 homoeologs could be linked with the presence/absence variation (Supplementary Data 16).

The majority of analyzed traits showed a significant positive or negative correlation with the total number of low-expressing alleles of homoeologs per line (Supplementary Fig. 13a), with the absolute correlation coefficients being higher than those obtained using the random control (Supplementary Fig. 13b). The accumulation of the low-expressing alleles of these homoeologs was associated with an increase in grain length (SCC $= 0.26$), width (SCC $= 0.41$) and weight (SCC $= 0.39$), and a decrease in heading date (SCC $= -0.29$), number of spikelets per spike (SCC $= -0.35$), spike length (SCC $= -0.19$), and plant height (SCC $= -0.18$) (Supplementary Fig. 13a). A similar analysis performed using the negatively correlated homoeologs detected in the spikes and the number of grains and the number of spikelets per spike also revealed negative correlation between the low-expressing homoeologous alleles and traits (SCC $= -0.25$ and SCC $= -0.16$, respectively).

We also tested for association between the negatively correlated homoeologs and phenotypes (Fig. 6a–c) by predicting productivity traits using the expression values of homoeologous alleles and ridge regression modeling. Except for spike length, spike compactness, and awnlessness traits, the correlation between predicted and observed trait values for major productivity traits ranged from 0.25 to 0.37 (Fig. 6d, Supplementary Table 1). For many traits, including spikelet number per spike and grain length, the correlation between the true traits and traits predicted using the expression levels of negatively correlated homoeologs was in the 99th percentile of distribution generated using expression data from the random sets of genes (Fig. 6e, Supplementary Table 1). This further confirms that the negatively correlated homoeologs are predictive of variation in productivity traits. These results combined with the observed correlation between the number of low-expressing alleles and the grain size/grain number traits (Supplementary Fig. 13b) suggest that the negatively correlated homoeologs could be connected with processes affecting variation in and trade-off between the productivity traits.

Association between the negatively correlated homoeologs and phenotype was independently validated using a panel of lines from the 1000 wheat exomes project. By correlating the number of cis-eQTL alleles associated with the low-expressing alleles in the set of 59 negatively correlated homoeologs and traits we showed that an increase in the number of these alleles is linked with an increase in grain yield and decrease in heading date (Supplementary Fig. 14). Combined together, our results indicate that the accumulation of eQTL variants linked with the relative changes in the homoeolog expression dosage have potential to affect traits of agronomic importance in polyploid wheat.

Discussion
We characterized the genetic variants associated with variation in homozygous gene expression measured in a panel of diverse allopolyploid wheat lines. The enrichment of detected variants in the regions of active chromatin and in the proximal transcription factor enrichment suggests our methodology is generalizable to other important traits. Using extensive trait mapping data from the WheatCAP and IWYP projects, we showed significant enrichment of cis-eQTL around the top marker-trait associations. Applying a transcriptome-wide SMR analysis, we identified a number of candidate genes whose expression is linked with variation in these traits. We showed that TaSPL14, the rice ortholog of transcription factor SPL14 (IPA1) controlling plant architecture, is associated with natural variation in spike and spikelet development traits in wheat. Our results suggest that joint modeling of GWAS and eQTL data using summary-level statistics has the potential to identify causal genes associated with trait variation in wheat or to prioritize candidate genes for further functional validation.
The relative contribution of cis- and trans-acting variants to population-scale variation in homoelogous gene expression and relative expression dosage appears to be strongly influenced by the demographic events accompanying wheat origin (polyploidization and gene flow) and human-mediated selection for domestication and improvement traits. The A and B genomes of hexaploid wheat are more genetically diverse than the D genome because of the post-polyploidization gene flow from tetraploid (AB genome) into hexaploid wheat. This factor contributed to a higher number of trans-acting variants in the A and B genomes compared to that in the D genome and likely accounts for the similar proportion of expression variance explained by cis- and trans-genomic variants for genes in the D genome. The latter is in contrast to the higher proportions of gene expression variance explained by cis- rather than trans-genomic variants in the genetically more diverse A and B genomes.

Our study highlights in importance of inter-genomic trans-effects in the regulation of genes that lost their cis-regulatory diversity in polyploid wheat due to either selection or polyploidization bottleneck. Among genes whose expression is largely explained by trans-genomic effects and that show the evidence of reduced genetic diversity, half were located in the D genome indicating that the loss of cis-acting diversity in these genes occurred during hybridization between the tetraploid wheat and the D genome ancestor. The loss of cis-acting variants in the A and B genomes is likely associated with domestication and improvement selective sweeps, which affected a significant portion of the wheat genome. The cumulative effect of trans-acting variants on the expression variance of genes that lost cis-regulatory diversity was comparable to the effects cis-acting variants on genes showing no evidence of trans-regulation, indicating that inter-genomic interactions should play significant role in regulating genes controlling domestication and improvement traits in wheat.

We showed that in allopolyploid wheat, the relative levels of homoelog expression are defined by the combination of cis- and trans-acting variants from all three genomes. On average, expression variance explained by trans-acting variants for all analyzed genes was 57% higher than variance explained by cis-acting variants. For homoelogous genes, the relative contribution of cis- and trans-acting variants for negatively correlated homoelogous (SCC<0) was also comparable. However, with increase in the levels of homoelog expression correlation contribution of cis-acting variants to expression variance significantly declined, indicating that these variants play more prominent role in creating expression dosage imbalance than trans-acting variants. The functional importance of cis-regulatory variants leading to homoelog expression bias was confirmed by earlier studies, which showed that causal variants underlying several adaptive and domestication traits also lead to changes in the homoelog expression dosage.

Previous study in maize linked the loss of fitness with gene expression dysregulation caused by rare mutations in the promoters of thousands of non-redundant genes. One of the
expected consequences of functional redundancy provided by polyploidy is increase in the mutation load in wheat\textsuperscript{8,10} that could disrupt the co-expression of homoeologous genes by creating low- or high-expressing alleles. The prevalence of homoeologs with the matching levels of expression in our study could be explained by the recent origin of allopolyploid wheat\textsuperscript{33}, which provided less time for the accumulation of dysregulating mutations. However, our results also suggest that the expression levels of homoeologous genes could be under purifying selection, consistent with the results previously reported for diploid organisms\textsuperscript{35,36}. The high levels of LD between the combinations of the high- and low-expressing homoeologous alleles showing negative expression correlation in our wheat panel, among others including the El\(\beta\)3 homoeologs, provides additional support for selection acting to maintain certain levels of homoeolog expression dosage. The \emph{trans}-acting variants appear could provide some level of robustness against the dysregulating effects of \emph{cis}-acting variants, as evidenced by the proportional increase in the number of shared \emph{trans}-eQTL. Nevertheless, this proportional robustness provided by the recent WGD in wheat\textsuperscript{37} is not sufficient to fully compensate homoeolog expression bias by increasing the expression levels of corresponding homoeologs from other genomes. If this expression imbalance affects the regulatory pathways controlling adaptive traits, the homoeolog expression dosage could be targeted by selection. This is in agreement with the previously reported evidence of deleterious variant removal from homoeologous genes in wheat\textsuperscript{8}.

Our study uncovered an association between the number of the common low-expressing alleles from homoeologs showing negative expression correlation and variation in productivity traits exhibiting trade-offs in wheat. A decrease in the total number of low-expressing alleles per line was accompanied by an increase in spikelet number per spike and a decrease in grain size and weight. Compared to a randomized control, the total number of low-expressing homoeologous alleles was more predictive of productivity trait variation, likely due to some connection of these homoeologs with trait-associated biological pathways. Recent yield increases in elite winter wheat cultivars were mostly linked with an increase in the number of spikelets and grains per spike\textsuperscript{58,69}, whereas in some cultivars from Asia, yield increase was mostly associated with an increase in grain size and weight\textsuperscript{70}. These trends are in agreement with studies suggesting that the contribution of different productivity traits to increased yield potential is environment-specific\textsuperscript{71}. Based on our results, we hypothesize that the low-expressing alleles of homoeologs creating dosage imbalance within the homoeologous gene sets, due to their impact on productivity traits in wheat, were targeted by improvement selection. Depending on which productivity trait was chosen as a breeding target, these homoeologous alleles were likely either purged (increases the number of spikelets/grains per spike) or accumulated (increases grain size/weight) in wheat lines. Identification of these homoeologous genes with imbalanced expression and associated pathways provides opportunities for targeted breeding or genome-editing strategies aimed at adjusting proportions of these alleles in the genome to maximize crop productivity.

\section*{Methods}

\subsection*{Plant material}

A panel of 400 diverse wheat lines (Supplementary Data 1) was selected from a larger worldwide sample of 2259 \textit{Triticum aestivum} accessions that were previously genotyped using the 9K Select SNP array\textsuperscript{22}. The seeds could be requested from the USDA National Small Grains Collection. Our panel was assembled to maximize: (1) genetic diversity, (2) representation of diverse geographic regions, and (3) representation of phenotypic response to the strains of fungal pathogen \textit{Puccinia graminis} f. sp. tritici (Pgt). The panel of 2259 lines was previously evaluated in the Wheat CAP project by infecting plants at the seedling stage using Pgt races TTKSK (Up99), TRTF, TTTTF, BCCBC, and a bulk of six races (TPMRC, RKRQC, RCRC5, QTHC, QCSC, and MCCFC). The pheno- type data is available from the Wheat CAP T3 database (https://triticasetoolbox.org). The Stakman infection types recorded on a 0–4 scale were converted to A through F grades where A corresponded to infection types 0′ to 1′; B to 1′ to 3′; mesothetic infection types; C to 2′ to 2′; D to 2′ to 3′ to 3′; and E to 3′ to 4′. We selected 50 wheat lines that showed an \( F \) grade to all five Pgt races. We also selected 350 additional lines that showed variable, race-specific responses to the Pgt races. When possible, 10 lines for a given pattern of infection type grades were selected. No lines possessed an A grade in response to all five races. During the selection of the lines within a given pattern of infection type grades, both geographic origin and PCGs were used to maximize the diversity of the panel. A subset of 204 wheat lines representing geographic and phenotypic diversity of this diversity panel was subjected to RNA-seq analysis and used for the genetic dissection of gene expression variation traits. The genetic relatedness analysis of this subset of lines was performed using an algorithm implemented in PLINK v.1.9. For this purpose, we have used genome-wide SNPs generated by the regulatory sequence capture and sequence-based genotyping approaches. This analysis shows that our panel does not contain highly related accessions, which otherwise might increase the chances of detecting spurious associations in GWAS (Supplementary Fig. 15).

\subsection*{RNA-seq data analysis}

Total RNA was isolated from 2-week-old seedlings of 204 lines. Three biological replicates for each line were performed. The three biological replicates were combined in equal amounts before RNA isolation using the RNaseasy Plant mini kit. RNA-seq libraries were prepared with TruSeq Sample Prep Kit (Illumina) using the Beckman’s Biomek\textsuperscript{8} FXP Laboratory Automation Workstation. Up to eight barcoded RNA-seq libraries were pooled per lane of NextSeq2000 flow cell to generate ≥100 bp reads. A total of 13,415,679,980 paired-end 2 × 100 bp reads were generated for 204 wheat accessions from the wheat diversity panel, with a mean of 65,763,137 reads per accession (GSE167479). The reads were mapped to the wheat RefSeq v.1.0 using HISAT2 (ver. 2.1.0) with the following parameters: --max-intronlen 70000, -- Dustin 70000. 90% of reads failing to map (Supplementary Data 1).

We generated RNA-seq data for 2-week-old seedlings from 204 wheat lines. We removed samples with a substantial amount of RNA contamination and samples with <40% uniquely mapped reads. The resulting set of 198 RNA-seq samples was used for further analysis (Supplementary Data 1). We used the Kallisto program that uses pseudolambdification of RNA-seq reads to reference gene models to assess the transcript abundance\textsuperscript{23}. Its performance has previously been evaluated in the polyploid wheat genome\textsuperscript{15}. We also evaluated the accuracy of Kallisto-based transcript abundance estimates for duplicated homoeologous genes. For this purpose, we have used RNA-seq data simulated for the A genome and then used all gene models from the wheat reference genome to calculate TPM values. Simulation performed for the A genome showed a high level of correlation between the real values and those estimated using Kallisto. We also evaluated the accuracy of Kallisto-based transcript abundance estimates for duplicated homoeologous genes. For this purpose, we have used RNA-seq data simulated for the A genome and then used all gene models from the wheat reference genome to calculate TPM values. Simulation performed for the A genome showed a high level of correlation (SCC = 0.98). We also evaluated the accuracy of Kallisto-based transcript abundance estimates for duplicated homoeologous genes. For this purpose, we have used RNA-seq data simulated for the A genome and then used all gene models from the wheat reference genome to calculate TPM values. Simulation performed for the A genome showed a high level of correlation (SCC = 0.92, N = 91,437) between the real values and those estimated using Kallisto. Only 0.1% of reads simulated using the A genome gene models were used to generate TPM estimates, achieving high accuracy of transcript abundance estimates for the homoeologous gene sets.

The same RNA-seq simulated dataset was used to estimate the accuracy of read mapping to the correct location in the wheat reference genome using HISAT2\textsuperscript{25}. We found that 98% of simulated reads could be unambiguously mapped by HISAT2 to the correct location in the wheat genome.
SNP genotyping of diverse wheat accessions used for eQTL mapping. We used a combination of different approaches to obtain genotyping data for the wheat diversity panel. First, we performed re-sequencing of the wheat genome using a Nimblegen capture assay, (2) wheat 90 K iSelect assay, (3) complexity-reduced genome sequencing, and (4) RNA-seq transcriptome dataset. SNPs were genotyped using the RNA-seq and RNA-seq transcriptome capture datasets for 203 wheat accessions were combined, and missing genotype calls were imputed using Beagle. This dataset was combined with the SNPs identified in the entire panel of 400 wheat accessions using the 90 K iSelect assay and complexity-reduced genome sequencing. Further, a common set of SNPs shared between our panel of 400 wheat accessions and 1000 wheat genome dataset, were used for genotype imputation (see details below).

Flanking sequences of a genetically mapped set of 46,977 SNPs from the 90 K iSelect assay were aligned to the IWGSC RefSeq v.1.0 using the BLAT program followed by filtering alignments using the following parameters: alignment coverage > 95%, sequence identity >97%, e-value < 1e−10. We identified genomic coordinates for 23,577 uniquely aligned SNPs, which also showed consistency with the marker order in the previously created genetic maps. For these SNP sites, we identified 16,037 SNPs segregating in our wheat panel of 400 wheat accessions. We used the wheat regulatory capture assay to re-sequence 203 wheat accessions used for transcriptome analyses. Study. Up to eight Illumina genomic libraries produced for each accession were pooled together for enrichment using the regulatory capture assay. A total of 9,418,016,463 paired-end 2 × 150 bp reads were generated for 203 accessions, with the mean of 46,394,170 reads per accession. Reads were aligned using HISAT2 (v. 2.1.0) with the following parameters: --max-intronlen 70000, --no-spliced-alignment. On average, all reads were aligned to the wheat genome uniquely. The recommended best practices were followed to call SNPs using GATK. Base quality recalibration was performed using genotyping data generated for the same set of lines using the 90 K iSelect assay. The genotype calls for sites with <3 reads depth of coverage were set as missing data. SNPs were filtered to remove sites with missing data >75% or heterozygosity >3%. Only biallelic sites were used in our analysis. Genotype calls generated for the 138,481 SNPs with MAF > 0.05 in a panel of 198 wheat lines were used for eQTL mapping. Reduced sequencing, resulting in a set of 4,449,989 SNPs. A total of 2,021,936 SNPs were removed due to the low proportion of mapped RNA-seq reads to the reference genome (Supplementary Data 1) and 227,922 SNPs from a set of 90 lines were grouped into three genome-specific sets (A, B, D genomes). Each set was used to build genotype relationship matrices using ‘--auto-num 30 –make-grm-inbred’ in GCTA. The genetic variance of expression traits was then calculated for three subsets jointly using ‘--mgmm –reml’. Out of the top 10,000 genes showing the highest levels of expression variance, 8698 gene expression traits in seedlings and 7090 gene expression traits in spikes were successfully processed (log-likelihood converged.) To remove the confounding effect of SNP density in different genomes, we used 1 SNP per 100 kb genomic window for the calculation of the genetic relationship matrix.

Detection of eQTL The association between SNPs and gene expression PEER RESOURCES (v. 21.0) were performed by modeling with a linear model — modelLINEAR: The set of 191 associations having matching RNA-seq and SNP genotyping data was used for final eQTL mapping (Supplementary Data 1). In addition to 6 lines removed due to low RNA-seq mapping quality (see above), we also removed 7 lines (GF294, GF342, GF366, GF380, GF381, GF383, GF387) that showed lack of good genotyping data concordance in the panel. The first three principal components (PCs) of the SNP matrix were used as covariates. Based on the estimates of genomic inflation factor (GIF), this approach was effective in controlling population structure for nearly 61% of genes, which showed no evidence of inflation of test statistics (61% of genes had GIF < 1.1) (Supplementary Fig. 16). While the remaining genes showed some effect of population structure on test statistic, these effects did not substantially inflate false discovery rate assessed by permutation of phenotypic data relative to genotypes. The expression values of each of the 52,060 genes in our seedling dataset were permuted relative to genotyping data (includes 2,021,937 SNPs) to generate 1000 randomized datasets. The SNP-gene expression association test statistic was calculated using Matrix eQTL. By applying p-value threshold corresponding to FDR ≤ 5e−5, on average, we detected 3,595 associations in the randomized datasets. In the real-life dataset, we have identified 11,421,859 associations (before LD merging) passing this significance threshold indicating that only 3.2 × 10^−6 associations passing our threshold are false positives. A similar permutation approach was applied for assessing the proportion of false positives among detected eQTL in the spikes. While in the original non-permuted dataset, 1,336,626 SNPs pass this p-value threshold (before LD merging), in the permuted datasets, on average we had only 10,858 SNPs passing threshold, suggesting that in spike eQTL the actual false discovery rate is around 3.2 × 10^−6.

All associations with FDR ≤ 5e−5 were considered as significant. For each transcript, significantly associated SNPs were merged based on LD (r^2 > 0.2) and distance (<100 kb) into genomic intervals. SNP with the strongest association signal within an interval was defined as an eQTL of the transcript. If an eQTL was located within ± 1 Mb around the target gene, it was defined as cis-eQTL. In our analyses we used a conservative definition of trans-eQTL, which was an eQTL significantly associated with the target gene located on a different chromosome. The eQTL effect size estimated by Matrix eQTL is based on the linear regression slope.

Analysis of eQTL and Hi-C data. Hi-C data for the hexaploid wheat cultivar Chinese Spring was downloaded from the NCBI database28. The Juicer Tools (v.21.01) was used to process all valid reads downloaded from NCBI GEO (GSM3929163_Wheat.s3u1tich10явлениеivalidPairs.txt.gz). We first generated a chic file using the ‘pre command, then the observed contact frequency map was calculated at 1 Mb resolution using the ‘dump’ command without normalization. The 3D chromatin contacts between a pair of eQTL–eGene was estimated based on the contact frequency between the two 1 Mb genomic intervals. For example, if eQTL is located at position chr7A:20,102,890 and its eGene TransC- gene is located at position chr1A:1,187,605, the contact frequency between cis-eQTL contact frequency was estimated between the genomic intervals 20–21 Mb on chromosome 7A and 1–2 Mb on chromosome 1A. We did not analyze the 3D chromatin contact between cis-eQTL and its eGene directly due to the relatively low depth of read coverage in the downloaded data. In order to evaluate the significance of the observed 3D chromatin contacts between cis-eQTL and eGene, we calculated the distribution of Hi-C contact frequency between randomly selected pairs of genomic intervals.
Phenotyping wheat for productivity traits. We have used field-based phenotyping data previously collected for a diverse panel of about 800 wheat accessions from the 1000 wheat exomes project^8, including a spring wheat allotetraploid (GPI), harvest weight (HW), drought susceptibility index for harvest weight (HWS), heading date (HD) and plant height (PHT) traits. Field data for these accessions were collected for two consecutive years under rainfed and irrigated conditions at the Agriculture Victoria research station located at Horsham, Victoria, Australia. Three replications of each accession were planted in a complete randomized design, with a seed spacing of 3.6 cm and row spacing of 65 cm. HD was recorded as the date on which 50% of the heads in a row fully emerged from the culms. PH was measured from the ground to the tip of the spike, excluding awns. The Best Linear Unbiased Estimators (BLUEs) were obtained using a model with fixed genotype effects and all other effects set as random in an individual year. The trait values from the rainfed and irrigated fields were used to calculate the drought susceptibility index for harvest weight (HWS), according to Fischer & Maure^85.

A panel of 400 spring wheat accessions were selected to represent genetic and geographic diversity of wheat (Supplementary Data 1). For phenotyping, plants were grown in Kansas State University greenhouse with 16 h light/8 h dark conditions with temperature set to 21 °C during the night and 24 °C during the day. Three plants of each accession were grown in the 1 gallon round pots filled with a self-made soil mix (volume ratio was 20 soil: 20 peat moss: 10 perlite: 1 CaSO₄). Plants were arranged according to a complete randomization design. Phenotyping data collected for each trait is listed in Supplementary Data 8. The heading date (HD) data was collected for two planting seasons when the plants reach stage 50, according to Zadoks scale^86. The date when the first spike in a pot appeared from the flag leaf sheath was recorded. The awn length (AWN) was measured after ripening. The accessions without awn, with short awn, and long awn were classified into three groups, and 2, respectively. Their data were collected in three consecutive planting seasons. The plant height (HT) was measured after ripening, from the base to the top of the main stem. The height of three plants for each accession was measured each season for four planting seasons. The spike length (SPL), spikelet number per spike (SPN), and spike compactness (SPC) were measured from the main spike of three plants for each accession. The ratio of each data was collected for three planting seasons. Grains from all the spikes of each plot were harvested and used for data collection. The MARVIN seed analyzer (GTA Sensornik GmbH, Germany) was used to estimate the Thousand Grain Weight (TGW) and grain weight (GW), length (GL), and area (GRA). In addition, the grain weight to length ratio (GWR) was calculated by dividing the grain length by grain width. The grain morphometric phenotypes were collected for three planting seasons.

Summary data-based Mendelian randomization analysis. We used GCTA (v.1.92.2beta) to perform genome association mapping^87 in a panel of 400 wheat lines. The genetic relationship matrix was calculated using the command ‘--make-grm-inbred –autosome-num 30’. The best linear unbiased predictions (BLUPs) were calculated for each phenotype by fitting a mixed linear model using the lmer function of R^88, and used for association mapping. On average, SCC between BLUPs and mean phenotypic values was 0.99. The ‘mima’ was used to calculate the association between SNPs and phenotypes. We used the first three principal components to control for population structure.

We applied summary data-based Mendelian randomization analysis (SMR)^89 to evaluate the association between gene expression and trait variation using summary-level data from our eQTL mapping study and two GWAS, one of which was performed in the current study, and another one was accomplished within the 1000 wheat exome project^90. The summary-level statistic of these two GWAS datasets was analyzed using the SMR commands ‘--trans-500000 --diff-freq-prop 0.2’ for trans-eQTL and ‘--diff-freq-prop 0.2 –cis-wind 10000’ for cis-eQTL. The eQTL identified using RNA-seq data from wheat seedlings and spikes were analyzed separately. Hardy test was used to separate functional associations from association due to linkage^91.

Construction of GCN connected with the SMR gene candidates. The candidate genes for each trait from SMR analysis (SMR p-value < 1e-4) based on seedling stage eQTL were used as a starting list of genes for network construction. In total, we used 1899 SMR gene candidates identified using the eQTL data identified in the wheat seedlings and summary-level data from GWAS conducted in the wheat 1000 exome project^92. We obtained co-expressed genes (SCC > 0.6) among the 198 seedling RNA-seq samples for all SMR gene candidates. All genes co-expressed with [SCC] > 0.6 with the SMR gene candidates or genes connected with the SMR gene candidates were included in this network. SMR candidate genes showing no co-expression with other genes in our dataset were excluded from this co-expression network. The Gephi (https://gephi.org) was used to visualize the GCN, which contains 3642 nodes and 57,837 edges. In addition, we used the seedling stage RNA-seq dataset to infer regulatory relationships between cis-eGene and its associated trans-eGene that are showing high levels of expression correlation at [SCC] > 0.6. For this purpose, we used all SNP-expression associations (FDR < 1e-5), which included 21,354,094 associations between 1,840,991 SNPs and 32,679 gene transcripts. We assumed that a SNP associated with both a cis-eGene and a trans-eGene at the same time, there is a potential regulatory relationship that exists between these two genes (Fig. 5a), and trans-eQTL effects are observed due to variation in the expression of a cis-eGene that acts as trans-factor. This scenario could be applied to the cases where a cis-eGene is a transcription factor, and a trans-eGene is its regulator target. In total, we detected 19,186 pairs of putative regulatory interactions among 5150 genes. The SCC values calculated using the PEER residues for each pair of genes showed a bimodal distribution (Supplementary Fig. 9). On the contrary, the background distribution of SCC values for random pairs of genes was unimodal, with a peak SCC value of 0.5, which was larger than 0.6, we predicted a potential regulatory relationship. In total, we predicted 1,200 regulations between 536 regulators and 294 targets. There were 903 negative regulations (SCC < −0.6) and 297 positive regulations (SCC > 0.6). Densely connected clusters of genes (network modules) were identified using a default routine implemented in Gephi^93. The network modules were also tested for GO enrichment (Supplementary Fig. 10, Supplementary Data 13–15). The network was also supplemented by information about genes previously characterized in wheat or rice^80.

Correlation between productivity traits and number of low-expressing alleles from the homoeologs with the biased expression. Among 21,807 homoeologous genes used in our eQTL analysis, we have identified 59 homoeologs showing evidence of biased expression dosage. We defined biased homoeologous genes (or negatively correlated homoeologs) using the following criteria: (1) negative correlation (SCC < 0) with both of its two homoeologs from other genomes, and (2) strong negative correlation (SCC < ~0.4) with at least one of its homoeologs. In most cases, negative inter-homoeolog expression correlation among these 59 homoeologs was associated with the presence of accession in the population that carry the downregulated gene variants. In each accession, a homoeologous gene was considered downregulated if it showed TPM < 3 and TPM mean - stdev, where ‘mean’ is the average TPM value among all wheat lines, and the ‘stdev’ is the standard deviation of TPM values among all wheat lines. Correlation between productivity trait BLUPs and the total number of low-expressing alleles from this set of 59 homoeologous genes in each accession was estimated using SCC. These SCC values were compared with the distributions of SCC values calculated between 1000 random sets of 59 homoeologous genes and each of the productivity traits.

Model-based prediction using the expression matrix of 59 homoeologous genes showing biased expression. Model-based prediction of productivity traits was performed using ridge regression method implemented in R package ‘glmnet’^48 using gene expression matrix with TPM values of 59 negatively correlated homoeologous genes for 198 wheat lines. Tenfold nested cross-validation was performed to test the accuracy of predictions^95. To assess the association between negatively correlated homoeologous genes and productivity traits, the accuracy of predictions obtained using these 59 genes was compared with the prediction accuracy obtained using the random sets of 59 genes from homoeologous gene triplets.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability. RNA-seq and genome resequencing data generated in this study are deposited to NCBI SRA PRJNA570223, PRJNA787276 and NCBI GEO GSE167479. All analyses were conducted using standard software. The settings of software used for analyses are described in the Methods. Source data are provided with this paper.

Received: 2 April 2021; Accepted: 26 January 2022; Published online: 11 February 2022.

References
1. Van De Peer, Y., Mizrahi, E. & Marchal, K. The evolutionary significance of polyploidy. Nat. Rev. Genet. 18, 411–424 (2017).
2. Conna, L. The advantages and disadvantages of being polyploid. Nat. Rev. Genet. 6, 836–846 (2005).
3. Salmon-Minkov, A., Sabath, N. & Mayrose, I. Whole genome duplication as a key factor in crop domestication. Nat. Plants 2, 16115 (2016).
4. Bao, Y. et al. Unraveling cis and trans regulatory evolution during cotton domestication. Nat. Commun. 10, 1–12 (2019).
5. The International Wheat Genome Sequencing Consortium (IWGSC). Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science 361, eaax2155 (2018).
6. Dubcovsky, J. & Dvorak, J. Genome plasticity a key factor in the success of polyploid wheat under domestication. Science 316, 1862–1866 (2007).
7. Kashkush, K., Feldman, M. & Levy, A. A. Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. Genetics 160, 1651–1659 (2002).
8. He, F. et al. Exome sequencing highlights the role of wild relative introgression in shaping the adaptive landscape of the wheat genome. *Nat. Genet.* 51, 1460–1466 (2019).

9. Hardigan, M. A. et al. Genome diversity of tuber-bearing *Solanum* uncovers complex evolutionary history and targets of domestication in the cultivated potato. *Proc. Natl Acad. Sci. USA* 114, E9999–E10008 (2017).

10. Krasileva, K. V. et al. Uncovering hidden variation in polyploid wheat. *Proc. Natl Acad. Sci. USA* 114, E913–E921 (2017).

11. Wang, W. et al. Gene editing of the wheat homologs of TONNEAU1–recruiting motif encoding gene confers grain shape and weight in wheat. *Plant J.* 100, 251–264 (2019).

12. Wang, W. et al. Gene editing and mutagenesis reveal inter-cultivar differences in wheat allohexaploids. *Nat. Genet.* 49, 202–213 (2017).

13. Pai, A. A., Pritchard, J. K. & Gilad, Y. The genetic and mechanistic basis for storage roots of sweet potato. *Nat. Genet.* 49, 1237–1249 (2017).

14. Aguet, F. et al. Genetic effects on gene expression across human tissues. *Nature* 550, 204–213 (2017).

15. Loukoianov, A., Yan, L., Blechl, A., Sanchez, A. & Dubcovsky, J. Regulation of VRN-1 Vernalization Genes in Normal and Transgenic Polyploid Wheat. *Plant Physiol.* 138, 2364–2373 (2005).

16. Joseph, B., Lee, Y. W., Stinchcombe, J. R. & Wright, S. I. Association mapping reveals the role of purifying selection in the maintenance of genetic variation in gene expression. *Proc. Natl Acad. Sci. USA* 112, 155390–155395 (2015).

17. Glassberg, E. C., Gao, Z., Harpak, A., Lan, X. & Pritchard, J. K. Evidence for weak selective constraint on human gene expression. *Genetics* 211, 757–772 (2019).

18. Barbier, A. N. et al. Exploiting the GTEx resources to decipher the genetic variants at GWAS loci. *Genome Biol.* 22, 49 (2021).

19. Yao, D. W., Connor, J. L. J., Price, A. L. & Gusev, A. Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nat. Genet.* 52, 626–633 (2020).

20. Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* 48, 245–252 (2016).

21. Nicolas, D. L. et al. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* 6, e1000888 (2010).

22. Miura, K. et al. OsSPL14 promotes panicle branching and higher grain productivity in rice. *Nat. Genet.* 42, 545–549 (2010).

23. Gao, J. et al. Arabidopsis FHY3/FAR1 gene family and distinct roles of its members in light control of arabidopsis development. *Plant Physiol.* 156, 4100–4102 (2004).

24. Kiseleva, A. A., Potokina, E. K. & Salina, E. A. Features of Ppd-B1 expression patterns of selection on homoeologous genomes. *Theor. Appl. Genet.* 119, 2463–2475 (2009).

25. Lin, R. & Wang, H. Arabidopsis FHY3/FAR1 gene family and distinct roles of its members in light control of arabidopsis development. *Plant Physiol.* 156, 4100–4102 (2004).

26. Kiseleva, A. A., Potokina, E. K. & Salina, E. A. Features of Ppd-B1 expression patterns of selection on homoeologous genomes. *Theor. Appl. Genet.* 119, 2463–2475 (2009).

27. Lin, R. & Wang, H. Arabidopsis FHY3/FAR1 gene family and distinct roles of its members in light control of arabidopsis development. *Plant Physiol.* 156, 4100–4102 (2004).

28. Kiseleva, A. A., Potokina, E. K. & Salina, E. A. Features of Ppd-B1 expression patterns of selection on homoeologous genomes. *Theor. Appl. Genet.* 119, 2463–2475 (2009).

29. Lin, R. & Wang, H. Arabidopsis FHY3/FAR1 gene family and distinct roles of its members in light control of arabidopsis development. *Plant Physiol.* 156, 4100–4102 (2004).

30. Kiseleva, A. A., Potokina, E. K. & Salina, E. A. Features of Ppd-B1 expression patterns of selection on homoeologous genomes. *Theor. Appl. Genet.* 119, 2463–2475 (2009).

31. Lin, R. & Wang, H. Arabidopsis FHY3/FAR1 gene family and distinct roles of its members in light control of arabidopsis development. *Plant Physiol.* 156, 4100–4102 (2004).

32. Kiseleva, A. A., Potokina, E. K. & Salina, E. A. Features of Ppd-B1 expression patterns of selection on homoeologous genomes. *Theor. Appl. Genet.* 119, 2463–2475 (2009).

33. Lin, R. & Wang, H. Arabidopsis FHY3/FAR1 gene family and distinct roles of its members in light control of arabidopsis development. *Plant Physiol.* 156, 4100–4102 (2004).

34. Kiseleva, A. A., Potokina, E. K. & Salina, E. A. Features of Ppd-B1 expression patterns of selection on homoeologous genomes. *Theor. Appl. Genet.* 119, 2463–2475 (2009).

35. Lin, R. & Wang, H. Arabidopsis FHY3/FAR1 gene family and distinct roles of its members in light control of arabidopsis development. *Plant Physiol.* 156, 4100–4102 (2004).

36. Kiseleva, A. A., Potokina, E. K. & Salina, E. A. Features of Ppd-B1 expression patterns of selection on homoeologous genomes. *Theor. Appl. Genet.* 119, 2463–2475 (2009).

37. Lin, R. & Wang, H. Arabidopsis FHY3/FAR1 gene family and distinct roles of its members in light control of arabidopsis development. *Plant Physiol.* 156, 4100–4102 (2004).

38. Kiseleva, A. A., Potokina, E. K. & Salina, E. A. Features of Ppd-B1 expression patterns of selection on homoeologous genomes. *Theor. Appl. Genet.* 119, 2463–2475 (2009).

39. Peng, Y. et al. Chromatin interaction maps reveal genetic regulation for quantitative traits in maize. *Nat. Commun.* 10, 1–11 (2019).

40. Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. *Nature* 518, 317–330 (2015).

41. Huyn, K., Jeon, J., Park, K. & Kim, J. Writing, erasing and reading histone lysine methylations. *Exp. Mol. Med.* 49, e324 (2017).

42. Ricci, W. A. et al. Widespread long-range cis-regulatory elements in the maize genome. *Nat. Genet.* 39, 102–109 (2007).

43. Nagy, J. et al. Genome-wide analysis of histone modifications: H3K4me3, H3K9ac, and H3K27ac in Oryza sativa L. *Japonica*. *Mol. Plant* 6, 1463–1472 (2013).
69. Würschum, T., Leiser, W. L., Langer, S. M., Tucker, M. R. & Longin, C. F. H. Phenotypic and genetic analysis of spike and kernel characteristics in wheat reveals long-term genetic trends of grain yield components. *Theor. Appl. Genet.* **131**, 2071–2084 (2018).

70. Zhang, Y. et al. Progress in genetic improvement of grain yield and related physiological traits of Chinese wheat in Henan Province. *F. Crop. Res.* **199**, 117–128 (2016).

71. Quintero, A., Molero, G., Reynolds, M. P. & Calderini, D. F. Trade-off between grain weight and grain number in wheat depends on GxE interaction: a case study of an elite CIMMYT panel (CIMCOG). *Eur. J. Agron.* **92**, 17–29 (2018).

72. Cavanagh, C. R. et al. Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. *Proc. Natl Acad. Sci. USA* **110**, 8057–8062 (2013).

73. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).

74. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–507 (2012).

75. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).

76. Gardiner, L.-J. et al. Integrating genomic resources to present full gene and putative promoter capture probe sets for bread wheat. *Gigascience* **8**, 1–13 (2019).

77. Wang, S. et al. Characterization of polyplloid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. *Plant Biotechnol. J.* **12**, 787–796 (2014).

78. Saintenac, C., Jiang, D., Wang, S. & Akhunov, E. Sequence-based mapping of the polyplloid wheat genome. *G3 (Bethesda)* **3**, 1105–1114 (2013).

79. Browning, B. L. & Browning, S. R. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**, 459–471 (2013).

80. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

81. Glaubitz, J. C. et al. TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS ONE* **9**, e90346 (2014).

82. Shabalin, A. A. Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012).

83. Fischer, R. A. & Maurer, R. Drought resistance in spring wheat cultivars: I. Grain yield responses. *Aust. J. Agric. Res.* **29**, 897–912 (1978).

84. Zadoks, J. C., Chang, T. T. & Konzak, C. F. A decimal code for the growth stages of cereals. *Weed Res.* **14**, 415–421 (1974).

85. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.* **46**, 100–106 (2014).

86. R Development Core Team, R. R: A Language and Environment for Statistical Computing. *R. Found. Stat. Comput.* **1**, 409 (2011).

87. Bastian, M., Heymann, S. & Jacomy, M. Gephi: an open source software for exploring and manipulating networks. in *International AAAI Conference on Weblogs and Social Media* (2009).

88. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).

**Acknowledgements**

This research was supported by Bål and Melinda Gates Foundation (INV-004438) to M.R., E.A., A.A., I.S., by the Agriculture and Food Research Initiative Competitive Grants 2022-68013-36439 (WheatCAP) to J.D. and 2019-67013-29017 from the USDA National Institute of Food and Agriculture to E.A., and by the International Wheat Yield Partnership (IWYP). We thank Dr. Li Lei for comments on an earlier version of the manuscript.

**Author contributions**

F.H.—bioinformatical and statistical analyses of data; F.H., W.W., J.D., P.L.M., and E.A.—data interpretation; M. Rouse—assembled diversity panel; W.W., M. Rouse, I.S., E.T., N.D., D.S., S.S., D.M. Reynolds, J.H., S.K.S., S.L., J. Chen, A.F., J. Cook, G.B., A.F., M.P., A.X.C., M.S., L.S., and M.J.H.—phenotypic characterization of the wheat populations and trait mapping. W.W., K.W.J., and A.A.—complexity-reduced sequencing and analysis; W.W., K.W.J., and M. Rouse—array-based SNP genotyping and analysis; W.R.—RNA-seq library construction and initial data processing. A.A. and J.R.—targeted sequence capture and RNA-seq data generation; F.H., W.W., P.L.M., and J.D.—editing the first draft of the manuscript. E.A.—conceived idea, performed data analyses, coordinated data collection, and wrote the manuscript.

**Competing interests**

The authors declare no competing interests.

**Additional information**

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-28453-y.

Correspondence and requests for materials should be addressed to Eduard Akhunov.

Peer review information *Nature Communications* thanks Guanjing Hu and the other anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

C {#fig:license}

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022