De novo annotation of the wheat pan-genome reveals complexity and diversity of the hexaploid wheat pan-transcriptome

Authors

Benjamen White[1†], Thomas Lux[2†], Rachel Rusholme-Pilcher[1†], Angéla Juhász[6†], Gemy Kaithakottil[1], Susan Duncan[1,3], James Simmonds[3], Hannah Rees[1], Jonathan Wright[1], Josh Colmer[1], Sabrina Ward[1], Ryan Joynson[1,4], Benedict Coombes[1], Naomi Irish[1], Suzanne Henderson[1], Tom Barker[1], Helen Chapman[1], Leah Catchpole[1], Karim Gharbi[1], Utpal Bose6,9, Moeko Okada[5,16,17], Hirokazu Handa[18], Shuhei Nasuda[19], Kentaro K. Shimizu[5,16], Heidrun Gundlach[2], Daniel Lang[2], Guy Naamati[7], Erik J. Legg[8], Arvind K. Bharti[8], Michelle L. Colgrave[6,9], Wilfried Haerty[1], Cristobal Uauy[3], David Swarbreck[1], Philippa Borrill[3], Jesse A. Poland[10], Simon G. Krattinger[10], Nils Stein[11,15], Klaus F.X. Mayer[2,12], Curtis Pozniak[13], 10+ Wheat Genome Project, Manuel Spannagl^[2,20], Anthony Hall^[1,14]

[1]Earlham Institute, Norwich Research Park, Norwich,NR4 7UH, UK.
[2]PGSB Plant Genome and Systems Biology, Helmholtz Center Munich, German Research Center for Environmental Health, Neuherberg, Germany.
[3]John Innes Centre, Norwich Research Park, Norwich, NR4 7UH, UK.
[4]Limagrain Europe, Clermont-Ferrand, Auvergne-Rhône-Alpes, France.
[5]Department of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland.
[6]Australian Research Council Centre of Excellence for Innovations in Peptide and Protein Science, School of Science, Edith Cowan University, Joondalup, WA, 6027, Australia.
[7]EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridgeshire, UK.
[8]Syngenta Crop Protection, Research Triangle Park, NC, USA.
[9]CSIRO Agriculture and Food, St Lucia, QLD 4067, Australia.
[10]Plant Science Program, Biological and Environmental Science and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia
[11]Center of Integrated Breeding Research (CiBreed), Georg-August-University, Göttingen, Germany
[12]School of Life Sciences, Technical University Munich, Freising, Germany.
[13]Crop Development Centre, The University of Saskatchewan, Saskatoon, Canada.
[14]School of Biological Sciences, University of East Anglia, Norwich, UK
[15]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, Seeland, Germany
[16]Kihara Institute for Biological Research, Yokohama City University, Yokohama, Japan
[17]Graduate School of Science and Technology, Niigata University, Niigata, Japan
[18]Graduate School of Life and Environmental Sciences, Kyoto Prefectural University, Kyoto, Japan
[19]Graduate School of Agriculture, Kyoto University, Kyoto, Japan

[20]Centre for Crop & Food Innovation, Food Futures Institute, Murdoch University, Murdoch 6150, WA, Australia

[†]These authors contributed equally to this work.

[*]Correspondence: manuel.spannagl@helmholtz-muenchen.de, Anthony.Hall@earlham.ac.uk

Abstract
Wheat is the most widely cultivated crop in the world with over 215 million hectares grown annually. However, to meet the demands of a growing global population, breeders face the challenge of increasing wheat production by approximately 60% within the next 40 years. The 10+ Wheat Genomes Project recently sequenced and assembled to chromosome level the genomes of nine wheat cultivars to develop our understanding of genetic diversity and selection within the pan-genome of wheat. Here, we provide a wheat pan-transcriptome with *de novo* annotation and differential expression analysis for these wheat cultivars, across multiple different tissues and whole seedlings sampled at dusk/dawn. Analysis of these *de novo* annotations facilitated the discovery of genes absent from the Chinese Spring reference, identified genes specific to particular cultivars and defined the core and dispensable genomes. Expression analysis across cultivars and tissues revealed conservation in expression between a large core set of homeologous genes, but also widespread changes in subgenome homeolog expression bias between cultivars. Co-expression network analysis revealed the impact of divergence of sub-genome homeolog expression and identified cultivar-specific expression profiles. In a case study utilising both the newly constructed wheat pan-genome and pan-transcriptome we demonstrate prevalent variation in the prolamin superfamily and immune-reactive proteins across the pan-cultivars.In summary, this work provides both a valuable resource for the wider wheat community and reveals diversity in gene content and expression patterns between global wheat cultivars.

Introduction
Wheat (*Triticum aestivum*) is the most widely grown crop and is cultivated in 12 mega-environments across the world [1], with 777.7 metric tonnes harvested globally in 2021/22 (www.fao.org). Pressures of climate change, political instability, a move to more sustainable farming and a reduction in agricultural land are putting increasing demand on international wheat harvests [2]. Efforts to overcome these pressures can be accelerated by understanding the genetic diversity of global wheat cultivars and their pan-transcriptional variation.

Wheat has a large (15Gb) allohexaploid (BBAADD) genome, derived from a series of relatively recent hybridisation events [3]. Its size, evolutionary history, and high repeat content, despite hindering genome assembly, make wheat an interesting model for the evolution of large polyploid genomes. Step changes in technology have enabled the chromosome level assembly of nine high-quality wheat genomes by a global consortium. These genomes revealed evidence of widespread structural rearrangements, introgression from wild relatives and the impacts of parallel international breeding programmes [4], [5]. To date these genomes were annotated only by projecting Chinese Spring gene models across the new assemblies. The generation of *de*

novo annotations for these genomes provides a key insight into gene gain and loss, reveal novel gene models across wheat cultivars and facilitate comparative gene expression analysis between cultivars.

Previous analyses of the wheat transcriptional landscape described tissue-specific changes in gene expression in two cultivars, using a common Chinese Spring reference genome[6]. Polyploidy leads to complex effects on gene expression resulting from structural variation, gene duplication, deletion and neofunctionalization, ultimately increasing variation in gene expression and the plasticity of the species. In the allotetraploids Brassica napus [7] and cotton [8] "pan-transcriptome" identified asymmetry in differential expression between genomes and differential gene expression across and introgressed regions.

Here, we generate *de novo* gene annotations, incorporating long reads for the nine assembled wheat cultivars, providing a valuable resource for wheat researchers and breeders. The construction of a reference-agnostic wheat pan-genome yields evidence of widespread gene duplication and deletion, revealing the impact of repeated hybridisations from wild relatives and breeding programmes on genome architecture. We define the hexaploid wheat core and dispensable transcriptome and our analysis of gene expression and gene networks across different tissues and between cultivars reveals conservation and divergence in expression balance across homoeologous subgenomes. We exemplify the value of these analyses through in-depth investigation of the pan-genome variability of prolamin gene content and expression; a key trait for quality and health aspects in wheat.

Results

De novo gene annotations of the pan-cultivars define the core and accessory gene sets

To precisely assess the gene content and differences in gene expression, copy number and the presence/absence of genes between the wheat cultivars, we generated a *de novo* gene annotation for each of the nine pan-cultivars. A detailed description of origin, biological characteristics and diversity coverage of these pan-cultivars is provided as part of [4]. We used an established automated annotation pipeline which built evidence based gene model predictions using a comprehensive transcriptomic dataset. This dataset was made up of Iso-Seq data from roots and shoots (390-700 K reads per sample), and RNA-seq data (150 bp paired-end read, 56-85 M pairs of reads per sample) obtained for each cultivar from five distinct tissue types and whole aerial organs sampled at dawn and dusk (**Figure 1A,** see methods for a full description and **Figure S1** for details of quality control). In addition to the transcriptomic dataset, the gene annotation pipeline also used protein homology and *ab initio* prediction. Finally, a gene consolidation procedure (**Figure S2A**) was developed to identify and correct for missed gene models in each specific cultivar. This step ensures the best possible comparability between the wheat genomes and gene repertoire [9].

The number of high-confidence gene models identified ranges from 140,178 for CDC Landmark to 145,065 for Norin 61 (**Figure 1B**). Low-confidence genes, primarily representing gene fragments, pseudogenes and gene models with only weak support, are in the range of 315,390 (Mace) to 405,664 (SY Mattis). With a maximal difference of 3.5%, the number of high-

confidence (HC) genes appears to be similar across cultivars, whereas most of the differences in gene number observed can be attributed to the low-confidence gene set. For around 70% of the HC genes we obtained evidence for transcription in at least one condition.

We benchmarked the quality of the *de novo* gene predictions against BUSCO v5.1.2 with the poales_odb10 lineage dataset, representing 4,896 Poales near-universal single-copy orthologs. On average, we found more than 99.8% of the BUSCO genes represented as at least one complete copy and 86% by three complete copies (**Figure 1B**). This is an improvement in complete BUSCO genes over the gene projections from Chinese Spring used in the first study of the wheat pan-genome [4] and can be explained by the *de novo* gene annotation strategy applied here, which included comparable RNA-seq and Iso-Seq datasets and *ab initio* prediction, as well as the final consolidation step. Completeness and consistency of the gene sets were measured using OMArk with an average of 97.1% of high-confidence genes corresponding to gene families in the Pooideae clade and only 4.5% missing, likely classified as low-confidence (**Figure S2C**). The *de novo* annotations are available in Ensembl Plants release 52.

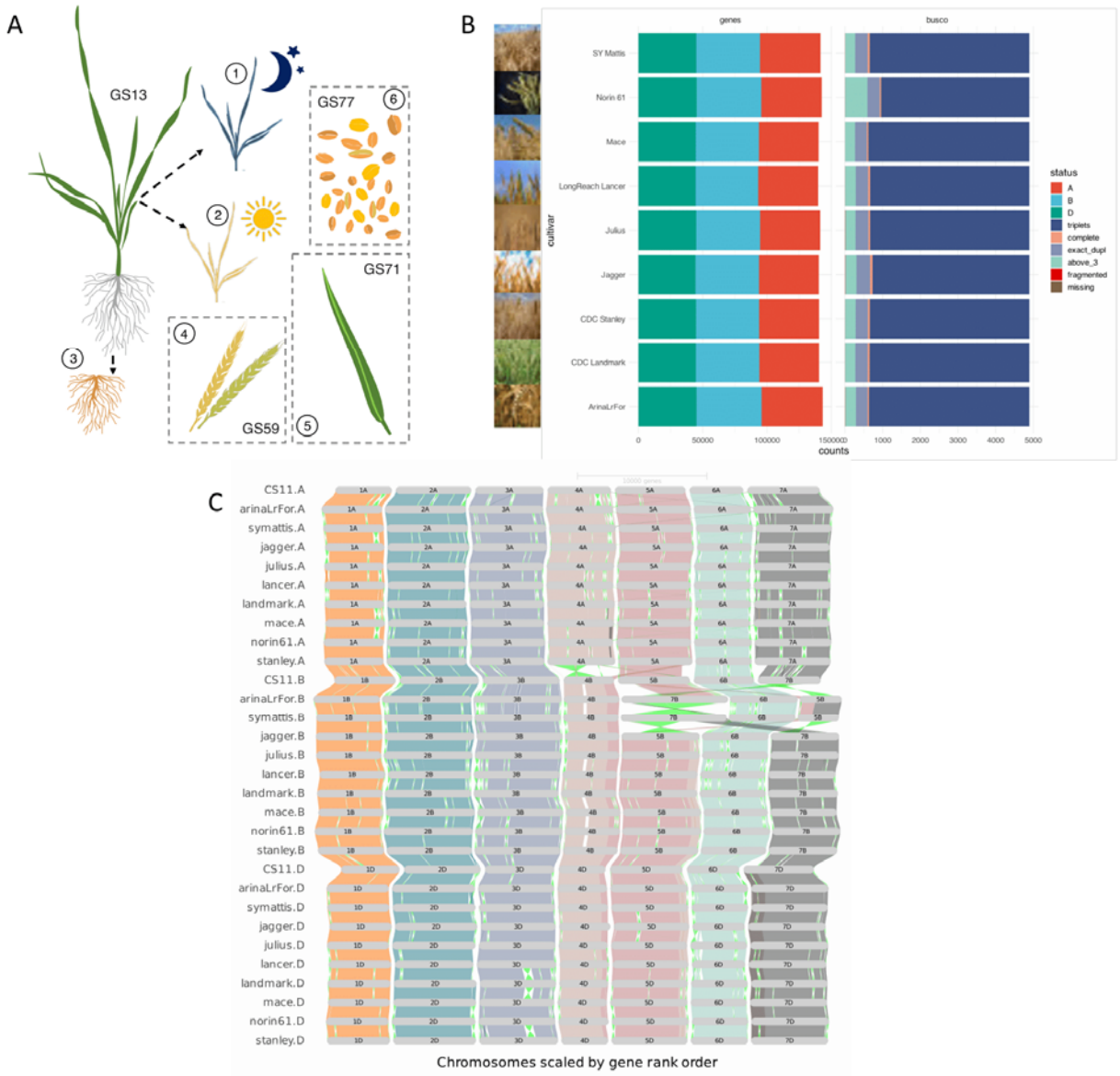

Figure 1. Study design, *de novo* gene annotations and orthologous framework. **A**. Overview of transcriptome data generated for this study of the wheat pan-cultivars. 1 & 2: whole aerial organs sampled at dawn and dusk, 3: root, 4: complete spike at heading (GS59), 5: flag leaf 7 days post anthesis (GS71), 6: whole grains 15 days post anthesis (GS77). **B.** *De novo* gene prediction results for each pan-cultivar (left side, "genes", separated for A, B and D subgenome) as well as summary of the BUSCO completeness assessment of gene models (right side, "BUSCO"). **C**. GENESPACE construction and visualisation of orthologous genes within the wheat pan-genome, using *de novo* predicted gene models.

The wheat pan-genome identifies gene duplications and introgressions

Genes and gene families exclusive or amplified in a specific cultivar are of major interest in a pan-genomic context [10]. While genes present in all compared cultivars are referred to as the core genome, cloud and shell genes are found only in one (cloud) or shared in a subset of

cultivars (shell). The improved gene annotation enabled the construction of a fully reference-agnostic pan-genome for bread wheat. It is noteworthy that our pan-genome differs in many ways from those constructed for natural populations [11], as we are investigating a hexaploid species of approximately 8,000 years old that has experienced population bottlenecks, and includes cultivated material originating mainly from breeding efforts rather than natural selection and/or drift. As a consequence, genes and genomic regions identified in the core, cloud and shell portions of our pan-genome mostly represent the result of targeted selection and could become important targets for wheat improvement and breeding. GENESPACE [12] was used to derive syntenic relationships between all chromosomes and subgenomes, allowing in-detail investigation of macro- and micro-synteny (**Figure 1C**) and gene copy number variations. While previously identified rearrangements such as the chromosome 5B/7B translocation in SY Mattis and Arina*LrFor* were confirmed, additional frequent small-scale structural variations can now be examined in the context of their gene content. We found a 16 Mb inversion, split into three segments of around 5Mb each, on chromosome 3D between Canadian cultivars CDC Stanley and CDC Landmark which coincides with the locations of QTLs related to biomass and grain weight [13].

We identified groups of orthologous genes (referred to as orthogroups) among the wheat high-confidence gene models of all cultivars. A total of 55,478 orthogroups contained 99.8% of all genes, with 112 orthogroups identified as cultivar-specific and 2,784 genes not clustered in any orthogroup - defining the cloud genome. Cloud and shell genes have previously been found to be associated with disease resistance [14], adaptation to new environments [15], or important agricultural traits [16]. Within the shell genome, our analysis identified orthogroups that are shared only between specific cultivars. Examples include CDC Stanley and CDC Landmark from Canada, Mace and LongReach Lancer from Australia or Arina*LrFor*, SY Mattis and Julius from Europe (highlighted in yellow in **Figure 2A**) which all share exclusive sets of genes. These observed patterns likely reflect the complex breeding history of the selected pan-cultivars which represent wheat lines from different regions, growth habits and breeding programs. Inspection of the chromosomal location of these gene groups identified multiple clusters (**Figure 2B and Figure S3**) that are likely associated with crosses to distinct material or hybridisations with wild or domesticated relatives; events common in wheat [17].

Proportions of core (genes present in all cultivars), shell (genes present in 2-8 cultivars) and cloud (genes found in only one cultivar and unclustered genes) genes were found to be similar across the pan-cultivars (**Figure 2C**). On average 62.52% of genes were classified as core, 36.61% as shell, and 0.86% as cloud (**Figure S2B**). These proportions differ from core-, cloud- and shell-gene sets reported for other plant species [18];[19] and can be attributed to the higher relatedness of the cultivated material investigated here. Amongst the core gene set we find biological functions associated with basic metabolic, catabolic and DNA repair/replication processes enriched (**Table S1**), while stress response and regulation of gene expression were overrepresented in the shell genes (**Table S2**). In the set of cloud genes, functions related to chromatin organisation and reproductive processes were found to be enriched (**Table S3**). These findings concur with gene functions predicted for the core-, shell- and cloud- gene sets in other crop plants [20]; [18]. Expression patterns of core, shell and cloud genes revealed pronounced

differences globally, but not between the subgenomes (**Figure 2D**). As observed in other pan-genomes [21], core genes tend to be higher expressed in all subgenomes and tissues, as compared to both shell and cloud genes.

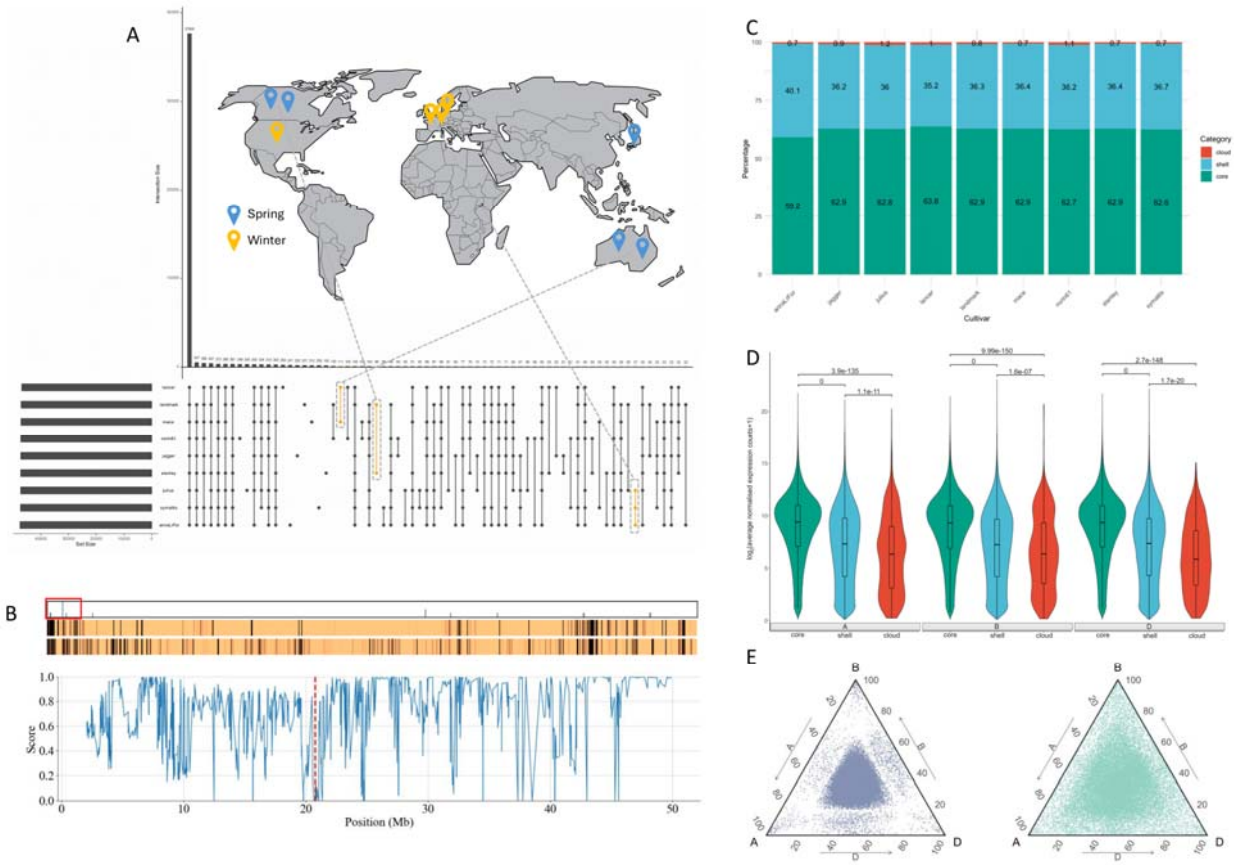

Figure 2. The wheat core-, shell-, and cloud- genome and homoelogous expression patterns. A. UpSet plot showing intersects of orthogroup conservation between pan-cultivars and the relation to their breeding programs and sowing season. **B.** A representation of CDC Stanley chromosome 3B showing the positions of Canadian-specific genes (top bar), heatmaps showing coverage scores between genes in CDC Stanley and CDC Landmark (middle bar) and coverage scores between CDC Stanley and Norin 61 (bottom bar). Coverage scores are calculated using kmers from each CDC Stanley gene to search the genomes of the other cultivar and range from 0 to 1 with values closer to 1 indicating greater similarity. Regions of greater difference are shown as darker bands. The plot shows the 0-50 Mb region of chromosome 3B (indicated by a red box). The mean of the coverage score between CDC Stanley genes in this region and genes in the non-Canadian lines is plotted. A cluster of four Canadian-specific genes (marked by a red dashed line) lies in a region which is noticeably different between CDC Stanley and the non-Canadian lines potentially representing an introgression. **C.** Percentage of genes belonging to the core-, shell- and cloud- orthologous groups across cultivars. **D.** Violin plots of core, shell and cloud log2 average gene expression across all combined cultivars and tissues, for each sub-

genome. Bonferroni adjusted p-values (Dunn test) from core, shell and cloud pairwise comparisons are shown above respective violin plots. Higher mean expression was observed in core genes across all sub-genomes. **E.** Ternary plots, of stable (left) and dynamic (right) 30-let (definition in main text) expression, where there is an homeolog present on each sub-genome, of all tissue in all cultivar, combined, showing more overall balanced expression in stable 30-lets and unbalanced expression in dynamic 30-lets.

In order to analyse cultivar-specific genes in the context of structural variation and expression, we extracted 592 genes that are unique to the Japanese cultivar Norin 61 based on our new de novo annotation (**Table S4**). The genomic positions of those genes are significantly enriched (p = 1.30E-27, Fisher's exact test, **Table S5**) in 176 Norin 61 unique genomic regions (**Table S6**). Furthermore, they were similarly enriched in 45 chromosomal segments that were previously shown to be specific to Norin 61 based on the unique pattern of transposable elements (TEs) and include putative alien introgressions [4]; [22]; [23] (p = 0.000264, Fisher's exact test; **Table S5, S7**). We next determined the expression patterns of those genes and found that 202 of the 592 Norin 61-specific genes were expressed at least in one of the examined tissues in Norin 61 (**Table S4**), and more than half of them (107) showed tissue-specific expression (tau > 0.8), excluding transposable elements (**Table S8**). Among the genes with tissue-specific expression patterns, functional descriptions related to defence response are enriched (Gene Ontology analysis; **Table S9**). Within the genomic regions with Norin 61 specific TE patterns we find genes related to disease resistance and plant immunity such as NONEXPRESSER OF PR GENES 1 and RGA2. These results demonstrate that the cultivar-specific chromosomal segments not only harbour unique coding and TE sequences consistent with introgressions but also encode genes with specific expression patterns, suggesting that these segments are promising targets for finding genes responsible for unique traits such as disease resistance.

Duplication of genes has been identified as a major driver of gene function evolution and adaptation in plants [15]. In wheat, a large number of tandem duplications was previously found both in the Chinese Spring (IWGSC v1.1) reference genome and the pan-genome assemblies[24], [4]. Our full de novo gene annotation of the wheat pan-genome, in combination with the extensive gene expression data presented in this study, allowed for an in-depth assessment of gene duplication dynamics across cultivars in hexaploid wheat.

We identified on average 7,011 tandem arrays (HC genes only) in each cultivar, with the lowest in CDC Landmark (6,889) and ArinaLrFor as the highest (7,196). We further classified these tandems whether they are true tandems (exactly two copies) or part of a tandem array (more than two genes). In the first category there are on average 2,918 and in the latter 4,093.
In addition, we tested whether there is a bias in expression towards one member of the array. We found that for the true tandem arrays 2,384 (88%) one of the two members was biased in its expression with respect to the other member, whereas for 534 (12%) arrays both copies were expressed at similar levels (**Figure S2D**). For arrays of tandems 904 (19%) showed a balanced expression, whereas 3,188 (81%) were imbalanced.
Amongst all tandemly duplicated genes in wheat, biological functions associated with phosphorylation, response to stimulus and stress and reproductive processes were enriched

(**Supp. Table 10**). Interestingly, biological processes for responses for stress, stimuli and toxic substances were enriched within the groups of balanced true tandems. Protein modification processes, regulation of biosynthetic processes and regulation of gene expression were enriched in the unbalanced true tandems.
We also investigated the conservation of tandem arrays across all pan-cultivars. MCScanX [25] was used to construct co-linear chains of tandems shared by two or more of the pan-cultivars. For the resulting 2,950 tandem chains (at least two, and up to nine pan-cultivars in the chain) we investigated the conservation of their expression profiles. We found 1,567 tandem chains with conserved/constant expression patterns across the tandems in the chain (that is either all balanced or imbalanced) and 1,383 tandem chains with variable expression profiles. These results highlight the impact of tandemly duplicated genes as a potential key driver of evolution and adaptation. Besides functional redundancy of homoeologous genes in hexaploid wheat, tandem genes and their expression (bias) are therefore an important target for breeding applications.

Conservation of Global Expression in the Wheat Pan-Transcriptome
To investigate changes in global gene expression across cultivars, biological replicates from whole aerial organs at dusk and dawn, and from flag leaf, root, spike and grain, were aligned to the corresponding de novo gene models to generate normalised gene expression counts. We observed from principal component analysis of the normalised counts that most of the variance is represented by the first principal component, representing the different developmental stages, and also similar grouping of expression overall (**Figure S1A, S1B**). We then used these normalised counts from the nine cultivars together with complete de novo annotations for the core, shell, and cloud group genes, to explore differences in expression between tissues across all cultivars. The patterns of expression observed in each individual orthologous class were consistent across tissues, and between sub-genomes, with core genes showing an overall higher mean expression than either shell or cloud (**Figure 2D, Figure S4A**). Cloud average gene expression was observed to be significantly lower than both core and shell, irrespective of tissue type or sub-genome biases (p-adj < 0.05). Overall, there is a globally conserved pattern of expression.

The tissue specific gene index (tau) was employed to assess the degree of gene expression specificity to flag leaf, root, spike or grain tissues across all cultivars (**Figure S4B**).
We observed the least number of tau genes in flag leaf (1,005 - 3,202 specific genes), that were significantly less (t-test; $p < 0.001$) overall compared to either root (4,736 - 8,974 specific genes), spike (5,453 - 9,323 specific genes) or grain (3,955 - 12,157 specific genes), that showed no significant difference between each other. However the number of specific genes showed least cultivar variability for flag leaf tissues, compared to the wide range in the number of grain-specific genes observed between cultivars. This could be the result of contrasting transcriptomic complexity between flag leaf and grain tissues, representing different developmental stages of maturity and metabolic activity. In polyploid crops agricultural traits are often modulated by an interaction of homoeologous copies of genes [6]. In wheat, previous studies have focused on tissue-specific expression across homoelogous triads, identifying sets of triads that are either balanced or unbalanced in their sub-genome expression. Here, we

compared variation in triad expression across cultivars using all 13,521 identified sets of 30-let genes (30-lets are defined as genes present as a triad in all nine cultivars and Chinese Spring) with a homoeolog present on each sub-genome of the de novo annotated cultivars. Using previously reported cut-off values [6], we observed similar sub-genome expression in these 30-lets, in each of the cultivars, to that reported previously in Chinese Spring, with 102 also being classed as not expressed (**Figure S4C**) [6]. However, when comparing the bias of sub-genome expression, we observed 8,028 (59.37%) of these 30-lets to have a conserved, 'stable', balanced expression between the three homoeologous copies across all cultivars (**Figure S4D**). Whereby 'stable' expression relates to a conserved sub-genome expression bias between cultivars, as opposed to a 'dynamic' expression where a change in sub-genome expression bias can be observed in one or more cultivars.

As well as conservation of the balanced state we also see conservation in dominance or suppression within triad groups with 276 showing stable suppressed expression and 63 stable dominant expression. Stably expressed 30-lets showed GO term enrichment for essential biological processes associated with photosynthesis, translation, DNA replication, exocytosis, glycolytic process and cell redox homeostasis (**Figure S4E**). Whilst the 5,052 37.36% 'dynamically' expressed 30-lets that showed a change in the bias of sub-genome expression in at least one cultivar were found to be significantly enriched for transmembrane transport, response to stress, response to oxidative stress, defence response and photosynthesis. These dynamic 30-lets were observed to be less fixed to a specific sub-genome expression pattern compared to stably expressed 30-lets, showing a further euclidean distance from a, b, c or centroid points (**Figure 2E**). Across these dynamically expressed 30-lets, 4,467 showed balanced expression in at least one cultivar, with B sub-genome suppression being the next most represented balance of expression occurring in 1,972 of the dynamic 30-let sets (**Figure S4F**). Overall, more suppression of expression was seen than dominance. The Kruskal-Wallis test, applied to assess differences in the mean values of the dynamic 30-let bias across the cultivars, revealed no significant differences when examining the total percentage of each expression bias ($p > 0.05$). This suggests that the bias of dynamic expression, whilst different for individual 30-lets, has been proportionally conserved across these cultivars.

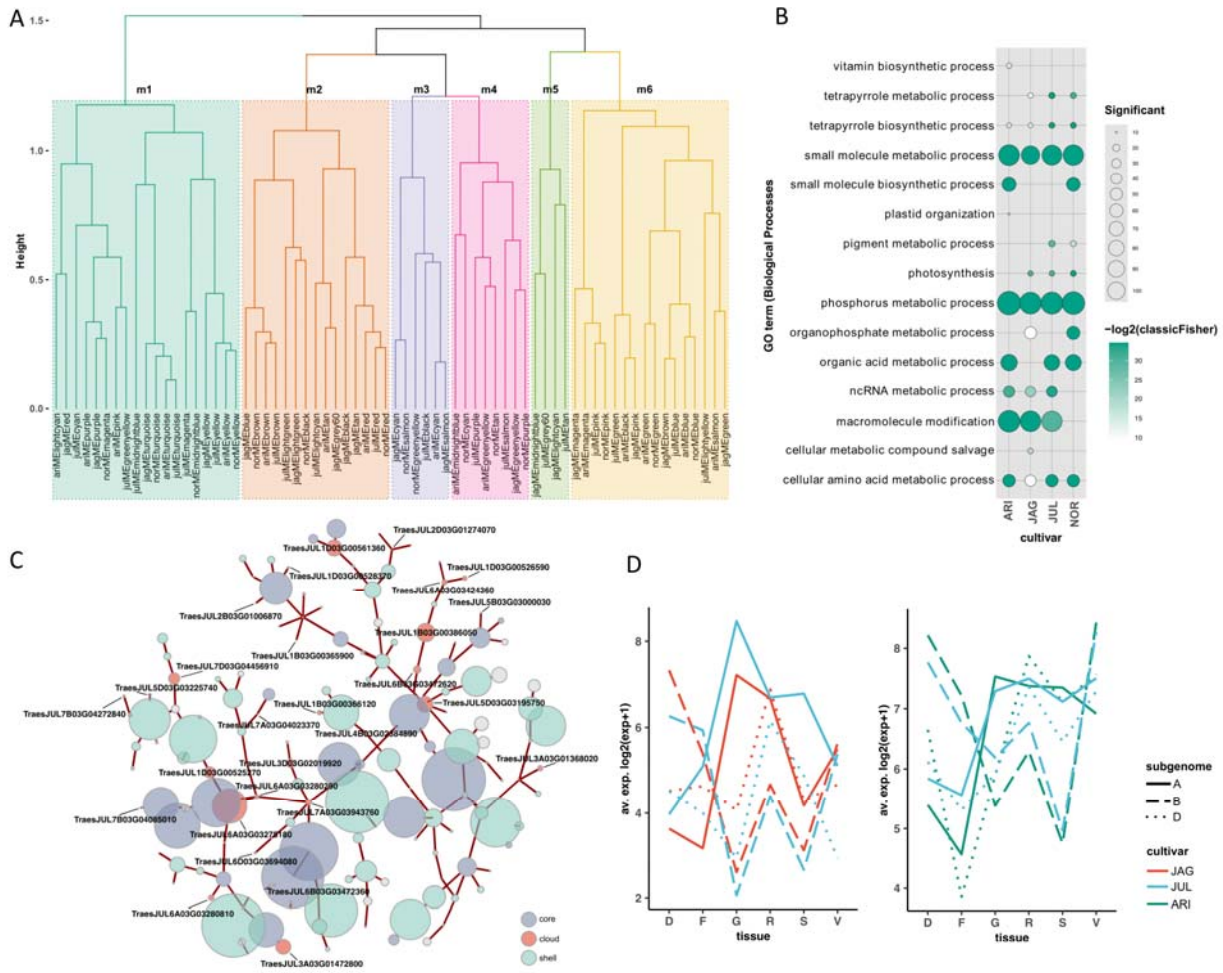

Figure 3. Components of the cultivar-specific networks with functional annotation and cultivar specific differences A. Hierarchical clustering of 68 module eigengenes from 4 cultivar networks identifying six metamodules (m1-m6). Each branch corresponds to a separate cultivar network module (ARI: ArinaLrFor, JAG: Jagger, JUL: Julius, NOR: Norin 61). **B.** GO terms of biological processes associated with genes in conserved metamodule one. Only terms with p<0.05 and >10 significant genes are shown. Bubble colour indicates the -log2p-value significance from Fisher's exact test and size indicates the frequency of the GO term in the underlying EBI Gene Ontology Annotation database (larger bubbles indicate more general terms). **C.** Network fragment from Julius module significantly enriched for cloud genes. Labelled nodes refer to cloud genes annotated as histones. The top five highly connected genes for each cloud gene are coloured according to core or shell genome membership (**Table S11**). Node size is scaled to the log2 average expression +1 of each gene across tissues and edge width reflects the weight of the connection between nodes. **D.** Expression of two divergent 30-let triads (L: HOG0029794, R: HOG0020263) with similarly divergent subgenome patterns of expression between Jagger and Julius (HOG0029794) and ArinaLrFor and Julius (HOG0020263). Annotated as F-box transcription factor and LRR protein, respectively. Tissues D: dawn, F: flag leaf, G: grain, R: root, S: spike, V: dusk.

Conserved patterns of co-expression

To explore how regulatory networks are conserved across tissues and cultivars of the pan-transcriptome, we constructed co-expression networks for ArinaLrFor, Jagger, Julius and Norin 61, representing a range of the genetic variation observed in the pan-genome [4]. For each cultivar we used alignments to the corresponding de novo high confidence gene models, retaining genes expressed at greater than five normalised counts in a minimum of two samples. The resulting genes (ArinaLrFor: 102,748; Jagger: 95,162; Julius: 98,435 and Norin 61: 97,734) were used to build four cultivar-specific networks. Between 14 (Norin 61) and 18 (Julius) modules were identified for each network, accounting for between 38.4% (Jagger) and 46.1% (ArinaLrFor) of the genes in each cultivar-specific dataset (**Table S12, S13**).

We used hierarchical clustering of 68 module eigengenes (ME) from all cultivar networks to identify six consensus metamodules (**Figure 3A**) spanning the four de novo cultivar networks. Five of the six metamodules contained modules from all cultivars, demonstrating conservation of expression patterns across both tissue type and cultivar. The orthologous framework described above was used to compare the gene content of these modules. Between 19.4-40.5% of orthogroups and 60.5-72.2% of GO terms present in the three largest metamodules (m1, m2 & m6) were common to ArinaLrFor, Jagger, Julius and Norin 61 modules (**Table S14**) demonstrating cross-cultivar conservation of gene content and function. GO term enrichment analysis revealed that these three conserved metamodules were involved in core processes such as photosynthesis, phosphorus metabolism, nucleosome assembly, cellular component organisation and primary metabolic processes (**Figure 3B, Figure S5A, Table S15**).

Two smaller metamodules, m3 and m4 had fewer GO terms in common (23.7 & 27.5%, respectively) and less than 5% of OG in common to all four cultivars. Whilst the expression patterns of genes belonging to all cultivars were similar, GO term analysis for metamodule three (m3) revealed that Jagger, Julius and Norin 61 were significantly enriched for genes involved in organic acid biosynthesis and ArinaLrFor genes were enriched for different biological processes including phenylpropanoid metabolism (**Table S15, Figure S5B**). Phenolic compounds are important for plant structural integrity and are implicated in the biosynthesis of key biotic and abiotic defence compounds [26]. Metamodule four (m4) had the smallest intersection of OG and GO terms between cultivars and this was reflected in diverse GO term enrichment for each cultivar (**Figure S5C**). The final metamodule (metamodule five; m5) only contained modules from Jagger and Julius and was enriched for genes involved in intracellular transport. Jagger genes in this metamodule were most significantly enriched for molybdopterin processes whereas Julius genes were primarily enriched for genes associated with cellular localisation and positioning.

Co-expression of cloud genes

We identified modules within our 4 cultivar-specific networks with a significant over-representation of previously defined cloud genes (**Table S16**). Four of these modules (one per cultivar), containing more than 10 cloud genes, demonstrated highly correlated MEs (Pearson correlation coefficient > 0.9, **Table S17**) and could be identified within the same metamodule

(m2), revealing their conservation in expression across tissue types and cultivar. The majority of cloud genes in these four modules (69-86.7%) were annotated as histones and were located in clusters on the genome. Histones have an essential role in transcriptional regulation and chromosome stability, with variants of paramount importance in the regulation of plant growth and development, and responses to both biotic and abiotic stresses [27]. Their co-expression in the cloud sector of all four cultivars could be indicative of related functions conferring cultivar-specific flexibility and adaptation. We were able to query our network resource to identify genes most highly connected to these histone orthologues for each cultivar. Visualising these network fragments for Julius revealed the connectivity of cloud genes to core and shell components and illustrated how the histone-annotated cloud genes were an integral part of the whole network (**Figure 3C, Table S18**).

Cultivar-specific expression of 30-let genes

For each cultivar network we compared inter-module relationships to identify network modules with divergent or similar patterns of expression. We then used these module relationships to compare how the 30-let triads were split across modules within each of the four cultivar networks. Of the 14,864 conserved 30-lets, we were able to identify between 39.1% (Norin 61) and 42.9% (ArinaLrFor) of 30-lets present as complete triads within each network. Of these triads, most were assigned to modules defined as the same or similar, reflecting triad conservation in both gene expression across tissues and orthology (ArinaLrFor 96.8%; Jagger 96.1%; Julius 97.1%; and Norin 61 96.7%).

The remaining triads (ArinaLrFor 3.2%; Jagger 3.9%; Julius 2.9%; and Norin 61 3.3%) spanned divergent network modules (**Table S19**). Divergent triads in the Jagger and Julius networks were significantly enriched for GO terms associated with phospholipid biosynthesis; compounds implicated in signalling pathways and regulation [28] andArinaLrFor and Julius were enriched for genes involved in transcription elongation (**Table S20**). Comparing these divergent triads sets across cultivars showed that these patterns of divergent subgenome expression were mostly cultivar-specific with >71% divergent triads demonstrating network specificity. The remaining 93 triads exhibited divergent subgenome expression in two or more cultivar networks (**Figure S6**) and contained genes with a range of functions, including leucine rich repeats, DNA double-strand break repair and transcription factors including five belonging to the F-box superfamily (**Figure 3D, Table S21**). F-box proteins play regulatory roles in protein degradation in response to cellular signals during plant development and growth, hormone responses, and biotic/abiotic stress responses [29].

Our work demonstrates the strength of a comparative network approach in identifying potentially biologically conserved pathways between cultivars. Combining these networks with our orthologous pan-genome framework enabled the identification of genes conserved across cultivars and tissues in both function and expression. We were also able to uncover co-expression modules demonstrating cultivar-specific patterns of expression, indicating diversity and scope within each cultivar for adaptation and flexibility. The resources we generated also enabled us to place these cultivar-specific variations in a wider genomic context, through the identification of highly connected network members.

A case study: Uncovering variation in the prolamin superfamily and immune-reactive proteins across the pan-cultivars

Prolamins represent a large superfamily in wheat involved in stress responses, cell growth and plant development, as well as end-use quality and protein content[30]. Additionally, along with HMW-glutenins prolamins trigger immune reactions in a subset of the population [31]. Here, we investigated the qualitative and quantitative differences in the 687 genes from the prolamin superfamily and HMW-glutenins to uncover their variations across the newly generated wheat pan-genome and pan-transcriptome data. We observed clear expression differences for individual developmental stages and between wheat cultivars for many genes from the prolamin superfamily, highlighting spatiotemporal variation in expression profile (**Figure 4A**).

Comparison of potential immune reactive genes identified in the Chinese Spring reference genome (IWGSC v1.1) and across the pan-genome cultivars [30]; [32]; [22] highlighted the challenges of precise annotation and characterisation of the dynamically expanded prolamin gene families in the pan cultivars [33]; [34]; [35]. This analysis therefore utilises the manually curated Chinese Spring prolamin annotations [30] as the common reference for comparative gene expression and adds the transcriptomes (generated in the same way as for the de novo annotated pan-cultivars; details in the methods) of five additional wheat cultivars without chromosome-scale genome assemblies and de novo gene predictions. The expression patterns of potentially immune-reactive gene products indicated differences in the major allergens and antigens (glutenins and gliadins). For example, SY Mattis and LongReach Lancer showed lower gene expression levels in alpha and gamma gliadins. Subsequent gene set enrichment analysis highlighted gamma gliadins as primarily enriched in the downregulated genes (**Figure S7, Table S22, S23**).

Celiac disease (CD) related epitope sequences encoded in the gliadin and glutenin genes show a significant variation in the wheat sub-genomes [30]; [32] and their generated immune response [31]. We analysed CD epitopes in the pan-genome cultivars and found varying expression levels of HLA-DQ epitope-containing genes. Notably, SY Mattis and LongReach Lancer exhibited lower expression, while Cadenza and Jagger showed higher expression. Cultivar-specific analysis reveals that ArinaLrFor and SY Mattis have lower alpha gliadin HLA-DQ epitope expressions due to differences in the expression activities of the three sub-genomes, possibly influenced by variations in cis-regulatory mechanisms and related transcription factor gene expressions (**Figure 4B, 4C, Table S24, S25, S26, Figure S8**). Although the sub-genome-specific expression patterns of gamma gliadin HLA-DQ epitopes did not show significant variation, the expression levels of alpha-gliadin genes with HLA-DQ epitopes from the A genome were lower in SY Mattis and LongReach Lancer. In addition, the highly immunogenic D genome alpha-gliadin epitope expression levels were lower in the ArinaLrFor cultivar (**Figure S9A**). Our results indicate that fine-tuned sub-genome-specific balance in the expression profiles may be associated with differences in the regulatory transcription factor profiles (**Figure 4B, 4C, Table S26**). Correlation analysis highlights common regulatory genes such as the PBF DOF triad (TraesCS5A02G155900, TraesCS5B02G154100, TraesCS5D02G161000) in complex with SPA-bZIP TFs regulate the

expression of a range of prolamin and starch synthesis genes through the endosperm box [36; 37], including the epitope-rich alpha and gamma gliadin genes, while different subsets of NAC or MYB genes impact the alpha and gamma gliadin gene expression (**Table S26**). TF binding site motif enrichment analysis shows gliadin-type and epitope group-specific motif enrichments, offering potential targets for genome editing (**Table S27**).

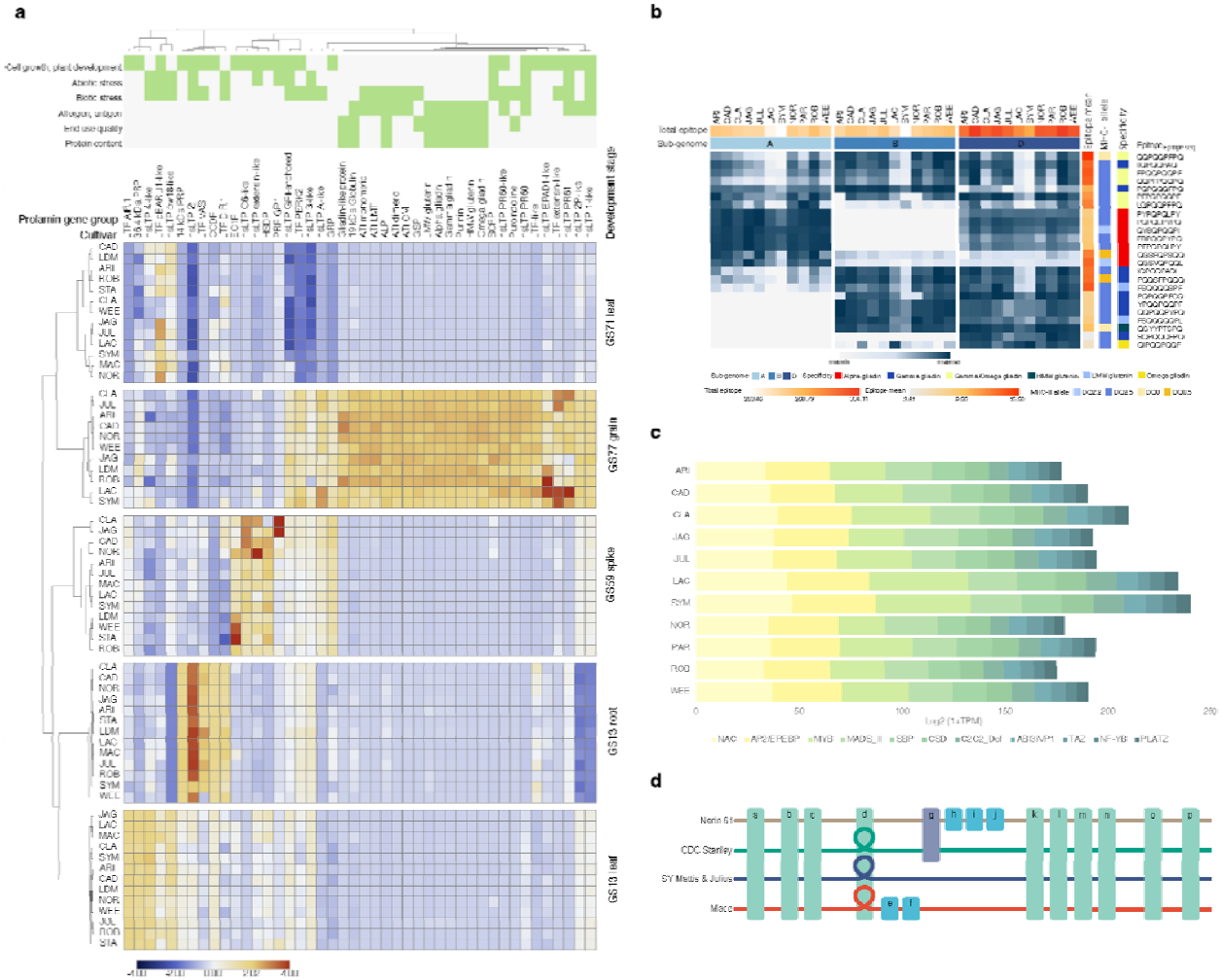

Figure 4. Gene and expression variation in the prolamin family across the wheat pan-cultivars. **A.** Prolamin superfamily gene expression across cultivars; **B.** Celiac disease epitope expression across cultivars. Epitope expression profiles were calculated as the sum of gene expression profiles with the highlighted HLA-DQ epitopes for each sub-genome. **C.** Relative proportion of cumulative expression profiles of transcription factor families showing strong co-expression patterns (Pearson correlation values > 0.8) with the epitope-coding prolamin genes. Results show significant differences in the NAC, AP2/EREBP and MYB transcription factor gene expressions, major storage protein gene expression regulators. **D.** Representation of the variation graph for the region of 6D containing the alpha-gliadin locus (**Figure S9B**). Horizontal coloured lines depict paths through the graph for each cultivar; Norin 61 (6D: 26,703,647-27,222,360 bp), CDC Stanley (6D: 28,164,601-28,660,350 bp) and Mace (6D: 26,808,846-

27,298,593bp), with SY Mattis (6D: 26,645,382-27,096,594 bp) and Julius (6D: 26,983,100-27,437,565 bp) sharing a single path. Rectangular blocks (a-p) represent individual genes at corresponding locations across cultivars (green: in common to all 4 cultivars, blue: occurring in one cultivar and purple: in common to 2 cultivars). Gene d is present as a single copy in Norin 61, and duplicated in CDC Stanley, SY Mattis, Julius and Mace. This duplication is represented as a loop in the path through the graph for these cultivars (**Figure S10, Table S29**).

Gliadin and glutenin loci were found to be highly conserved in all cultivars, with some variation due to the presence of pseudogenes and gene duplications (**Figure S9B**). Reverse translated consensus sequences of the known CD-specific T-cell epitopes were mapped to the genomes of all cultivars to determine the number and location of gliadin and glutenin genes containing CD-related immune reactive peptide regions (**Figure S9B, Table S28**). The number or combination of epitopes in the loci was not significantly different between the pan-cultivars. However, the gamma gliadin and alpha gliadin gene models with a high number of epitopes were found in cultivars ArinaLrFor, Norin 61 and Mace, respectively (**Figure S9B, Table S28**).

Although highly conserved in their locus structure on chromosome 6D, alpha gliadin genes encoding highly immunogenic proteins showed copy number variation within the wheat pan-genome. We constructed a localised pan-genome graph from five cultivars (Norin 61, CDC Stanley, SY Mattis, Julius, Mace) and extracted the subgraph of the alpha gliadin-containing locus (**Figure 4D, Figure S10**). Inspection of the subgraph helped to resolve the complex structure of the locus, with copy number variation observed as a loop in the paths of SY Mattis, Julius, CDC Stanley, Mace (2 copies of alpha gliadin genes) but not within the Norin 61 path (single alpha gliadin copy). While in total 4 to 6 epitopes were identified in the alpha-gliadins of the wheat pan-genome cultivars, 8 epitopes were detected in cultivars Mace and Norin 61 (**Figure S9B**). These results indicate that gene copy number expansion primarily affected the centre of the locus and resulted in the increase of gene variants with high epitope counts. While genome-wide construction and interpretation of pan-genome graphs remains a daunting task for complex genomes such as wheat, we found localised subgraphs, augmented by our de novo annotations, particularly helpful in resolving complex loci, and uncovering structural variation as also demonstrated in the current draft human pan-genome [38].

Discussion

We have built *de novo* gene annotations for nine wheat assemblies representative of global breeding programs [4]. Our consolidated gene annotation approach generated a robust set of core, high confidence genes shared across the pan-cultivars. It also identified genes and gene families that are found exclusively in or amplified in cultivars derived from specific breeding programmes. It is likely that some of this variation has come through widespread introgression events [39], often associated with adaptation to biotic or abiotic stress [16]. Our annotations also identified cultivar-specific variation in tandem gene duplication. Novel gene content, gene duplication and neo-functionalisation together with gene expression patterns will have impact

on researchers and breeders as they identify genes underlying traits, manipulate gene expression or incorporate and track new genetic variation.

Our analysis of global gene expression identified sets of genes with stable homoeologous expression patterns between cultivars, demonstrating tightly regulated key biological processes. We also identified homoeologous triads diverging in their expression patterns between cultivars, revealing genes enriched for processes associated with biotic and abiotic stress. Understanding the regulatory networks driving these altered patterns will provide important targets for manipulating these processes. Using network analysis, we identified widespread conservation of expression patterns across tissues and cultivars before focusing on cultivar-specific gene sets, to reveal networks of genes involved in regulation of plant growth and development and in responses to both biotic and abiotic stresses . These cultivar-specific network changes may be the result of wheat breeding programmes targeted to local environments. We also demonstrated the utility of our new resources by investigating genomic variation in the prolamin superfamily, focusing on immunogenic potential. Furthermore, the combination of the expression data with the sequence alignments using our *de novo* annotation made it possible to quantify the effect of gene expression on genome-wide pattern of selection (Halstead-Nussloch et al., accompanying paper).
In conclusion, this study reveals layers of hidden diversity spanning our modern wheat cultivars. Previously overlooked, this diversity is likely to underpin the agronomic success of wheat over a wide range of global mega-environments.

Materials and Methods

A detailed description of origin, biological characteristics and diversity coverage of the wheat pan-cultivars in this study is provided as part of [4].
We constructed full de novo gene predictions and an orthologous framework (defining core/shell/cloud gene sets) for 9 wheat pan-cultivars with chromosome-scale genome assembly sequences: Arina*LrFor* (abbreviated as "ARI" throughout the manuscript), Jagger (JAG), Julius (JUL), Lancer (LAC), CDC Landmark (LDM), Mace (MAC), Norin 61 (NOR), CDC Stanley (STA) and SY Mattis (SYM).
This dataset builds the foundation for the analyses in this study, unless specified otherwise. The Prolamin case study utilises the manually curated Chinese Spring prolamin annotations [30] as the common reference for comparative gene expression and adds the transcriptomes (generated in the same way as for the de novo annotated pan-cultivars) of five additional wheat cultivars without chromosome-scale genome assemblies and de novo gene predictions: Cadenza (CAD), Claire (CLA), Robigus (ROB), Weebil (WEE) and Paragon (PAR).

Plant Materials and Growth Conditions

The wheat cultivars were grown in a Controlled Environment Room (CER) (Conviron BDW80; Conviron, Winnipeg, Canada) set at 16 h day/8 h night photoperiod (300 µmol m^{-2} s^{-1}, lights on at 05:00, lights off at 21:00), temperatures of 20/16 °C, respectively, and 60% relative humidity. Plants were sampled in triplicate at the 3-leaf stage (Zadoks GS13), harvesting whole

roots and whole aerial organs separately, four hours after dawn (09:00). Whole aerial organs were also sampled two hours after dusk (23:00). Plants for subsequent adult plant sampling were treated according to their vernalisation requirements. In the case of spring wheat cultivars (CDC Landmark, CDC Stanley, Paragon, Cadenza, Mace and LongReach Lancer), seedlings were grown as described above. At 3-leaf stage, seedlings were transferred to 1 L pots containing Petersfield Cereal Mix (Petersfield, Leicester, UK) and maintained under the same CER conditions as described previously. For winter wheat cultivars (Julius, Jagger, Arina*LrFor*, Robigus, Claire and SY Mattis) and a facultative spring cultivar (Norin 61), seedlings were transferred in 40-well trays (7 days after sowing) to a vernalisation CER running at 6 ^{o}C with 8 h day/16 h night photoperiod for 61 days. After this period the plants were transferred to 1 L pots containing Petersfield Cereal Mix (Petersfield, Leicester, UK) and moved to the same CER and settings as described for the spring wheat cultivars. For both spring and winter wheat cultivars, three additional samples were harvested: complete spike at heading (GS59), flag leaf 7 days post anthesis (GS71) and whole grains 15 days post anthesis (GS77). All samples were harvested at four hours after dawn (09:00), and a single plant was used per each of the three biological replicates.

Sample Preparation and Sequencing

Total RNA was extracted using Qiagen RNeasy Plant Mini Kit (cat. no. 74904) and DNAse treated using an Invitrogen TURBO DNase kit (cat. no. AM2238) according to the manufacturer's protocol. Bead purification of the RNA was conducted using the Agencourt RNAClean XP beads.system (cat. no. A63987). Final sample concentrations were verified using a Qubit 4 Fluorometer, and the integrity of the RNA was checked on the Agilent 2100 Bioanalyzer, using the RNA 6000 nano kit (Agilent, 5067-1511), running the plant total RNA assay. The directional RNA-seq libraries were constructed using the NEBNext Ultra II Directional RNA Library prep for Illumina kit (NEB, E7760L) utilising the NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB, E7490L) and NEBNext Multiplex Oligos for Illumina (96 Unique Dual Index Primer Pairs) (cat. no. E6440S/L) at a concentration of 10 µM. The final libraries were equimolar pooled, a q-PCR was performed and the pool was sequenced on a Illumina NovaSeq 6000 with 150 bp paired-end reads.

The Iso-Seq libraries were constructed from 1 µg of total RNA per sample and full-length cDNA were then generated using the SMARTer PCR cDNA synthesis kit (Takara Bio Inc, 639506). The libraries were sequenced on the Sequel Instrument v1, using 1 SMRTcell v2 per library. All libraries had 600-minute movies, 120 minutes of immobilisation time, and 120 minutes pre-extension time.

Data Quality Control and Sample Validation

We used a set of cultivar specific SNPs to confirm the cultivar origin of each replicate and the developmental stage of each sample was validated through a machine learning approach trained using the pooled RNA-seq samples and then run on the entire set of biological replicates. Principal component analysis of the pooled samples shows them to cluster by developmental stage as expected.

Gene annotation

For the structural gene annotation of the chromosome-scale assembled pan-cultivars, we combined de novo gene calling and homology-based approaches with RNAseq, Isoseq, and protein datasets. The RNAseq data were mapped using STAR [40] (v2.7.8a) and further assembled into transcripts by StringTie [41] (v2.1.5, parameters -m 150-t -f 0.3). PacBio Iso-Seq transcripts were derived from the raw reads using PacBio SMRT Link software (v5.1.0.26412rev2, pbsmrtpipe.pipelines.sa3_ds_isoseq2, default parameters). The Iso-Seq transcripts were aligned to the genome assemblies using GMAP [42] (v2018-07-04). To assist the homology-based annotation approach, Triticeae protein sequences from publicly available datasets (UniProt, https://www.uniprot.org, 05/10/2016) were aligned against the genome sequence assemblies of all pan-cultivars using GenomeThreader [43] (v1.7.1; arguments -startcodon -finalstopcodon -species rice -gcmincoverage 70 -prseedlength 7 -prhdist 4). All transcripts derived from RNAseq, IsoSeq, and aligned protein sequences were combined using Cuffcompare [44] (v2.2.1). Stringtie (version 2.1.5, parameters --merge -m150) was employed to merge all sequences into a pool of candidate transcripts. To identify potential open reading frames and to predict protein sequences within the candidate transcript set, TransDecoder (version 5.5.0; http://transdecoder.github.io) was used.

We used Augustus [45] (v3.3.3) for the ab initio gene prediction. Guiding hints based on the RNAseq, protein, IsoSeq and TE datasets described above were used to counteract potential over-prediction (details in [46]). Augustus was run using the wheat model.

A consolidated set of gene models was selected using mikado [47], as implemented in the Minos pipeline (https://github.com/EI-CoreBioinformatics/minos), with models scored and selected based on a combination of intrinsic qualities and support from transcriptome and protein alignments.

BLASTP [48] (ncbi-blast v2.3.0+, parameters -max_target_seqs 1 -evalue 1e-05) was used to compare potential protein sequences with a trusted set of reference proteins (Uniprot Magnoliophyta, reviewed/Swissprot, downloaded on 3 Aug 2016; https://www.uniprot.org). This approach was employed to differentiate gene candidates into complete and valid genes, non-coding transcripts, pseudogenes, and transposable elements. This step was assisted by PTREP (Release 19; http://botserv2.uzh.ch/kelldata/trep-db/index.html), a database of hypothetical proteins containing deduced amino acid sequences in which internal frameshifts have been removed in many cases. We selected best hits for each predicted protein from each of the three databases used. Only hits with an e-value below 10e-10 were considered. Functional annotation of all protein sequences predicted in our pipeline was performed with the AHRD pipeline (https://github.com/groupschoof/AHRD).

We classified predicted proteins into two confidence classes: high and low confidence. Hits with subject coverage (for protein references) or query coverage (transposon database) greater than 80% were considered significant and protein sequences were classified as high-confidence based on following criteria: protein sequence was complete and had a subject and query coverage above the threshold in the UniMag database or no BLAST hit in UniMag but in UniPoa

and not PTREP; a low-confidence protein sequence was incomplete and had a hit in the UniMag or UniPoa database but not in PTREP. Alternatively, it had no hit in UniMag, UniPoa, or PTREP, but the protein sequence was complete. In a second refinement step, low-confidence proteins with an AHRD-score of 3* were promoted to high-confidence.

BUSCO [49] (v5.1.2.) software was used to evaluate the completeness and accuracy of structural gene predictions with the 'poales_odb10' database containing a total of 4896 single-copy genes. OMArk [50] (v0.3.0) was also used to evaluate the consistency of the gene models against gene families in the Pooideae clade. The evidence-based part of the annotation pipeline is deposited at https://github.com/PGSB-HMGU/plant.annot.

Consolidation

Pairwise whole genome alignments were generated using lastz [51]. The resulting alignments were stitched together into a single whole genome alignment using TBA/multiz [52]. The MAF output was converted into HAL format using maf2hal [53].
De novo gene annotation from one cultivar was lifted over to all other cultivars using the whole genome alignment and the halLiftover tool, whereas only full-length gene models were kept. Missing gene models in one cultivar were identified using bedtools [54].

Tandem array detection

Tandem arrays were identified using the tandem discovery model from the MCScanX package [25], with the following definitions: TrueTandems contain exactly two gene copies and TandemArrays are a sequential array of tandemly duplicated genes. Collinear tandems (chains) were detected using the detect_collinear_tandem_arrays tools provided by MCScanX and the results were filtered for TrueTandems. As described in the Alignment and Gene Expression Analysis section we used stringent mapping procedures to ensure only uniquely mapped reads were quantified, ensuring that we were able to correctly attribute reads to each respective gene copy. Expression bias was calculated using a modified method described previously [6]. Here we used normalised read counts instead of TPM values and a cut-off of 0.8. The following categories were assigned: unbalanced for tandems with only one gene expressed and no expression data for second gene; or for tandems in which only one gene is expressed under all RNASeq conditions; balanced, where both array members are equally expressed.

Orthogroup analysis

Two runs of OrthoFinder [55] were performed to construct the orthologous framework and both the triads and 30-lets datasets (see respective sections for definitions). The first OrthoFinder run used the full/unfiltered set of de novo predicted HC genes plus the Chinese Spring IWGSC RefSeq v1.1 annotation to construct 30-lets and triads. The second OrthoFinder run used the TE- and plastid-filtered set of de novo predicted HC genes to construct an orthologous framework for the pan-genome analyses. The resulting HOGs (Hierarchical Orthologous Groups) were used to determine the core-, shell-, and cloud- gene sets. The longest isoforms from high-confidence genes were used as input for Orthofinder [55]. We first applied filtering criteria for TE- and plasmid-related gene descriptions. Orthofinder was run using standard parameters. We used the UpSetR in the R package (http://gehlenborglab.org/research/projects/upsetr/) to

analyse and visualise how many orthogroups are shared between the cultivars or are unique to a single species. GENESPACE [12] was used to derive and visualise syntenic relationships between all chromosomes and subgenomes. Scripts for the definition of core-, shell-, and cloud- gene sets were deposited at https://github.com/PGSB-HMGU/BPGv2.

Analysis of Canadian-specific genes

Taking each genome in turn as a reference, kmers of length 51 were identified from genic regions using the annotation for that reference. These kmers were used to search the genomes of the other cultivars and a coverage score was computed [56] between each gene in the reference and every other genome. The coverage score (a value between 0 and 1) can be used as a proxy for sequence similarity/difference between genes in different cultivars where values closer to 0 indicate greater difference and values closer to 1 indicate similarity. Coverage scores for genes along chromosomes were plotted using the seaborn visualisation library [57] in Jupyter notebook. Coverage scores were also visualised as heatmaps with coverage scores close to 0 represented as dark bands.

Alignment and Gene Expression Analysis

Samples were aligned to the chromosome-scale assembled pan-cultivars, using HISAT2 v2.0.4 [58] and Stringtie v1.3.3 [59] was used to extract and quantify uniquely mapped reads at gene level to the respective *de novo* gene models. Normalised counts were generated using DESeq2 [60]..

GO term analysis

Functional enrichment of differentially expressed genes for biological processes was performed using the gene ontology enrichment analysis package, topGO [61] in R (v3.6.0, with the following parameters: nodeSize = 10, algorithm = "parentchild", classicFisher test $p < 0.05$). GO terms refer to ontology terms for biological processes unless otherwise specified and were obtained from Ensembl Plants 51, using the BioMart tool. Bubble plots were plotted using ggplot in R, adapting code from [62].

Tissue Specific Index

Specificity of gene expression to developmental stages was determined using the tissue specific index [63]. Where, N is the number of developmental stages (condition), and x_i is the expression profile component for a given gene in each condition, normalised by the maximal expression value of the given gene from all condition considered. This allowed us to classify genes as being highly specific to one condition (tau => 0.8). Assignment of tau values was performed in R using code adapted from previous work [64].

Subgenome Expression Bias

Analysis of sub-genome expression focused on 30-let homoeologs with a 1:1:1 relationship across all three sub-genomes. Of these, 13,521 were determined to be macrosyntenic, belonging to the same sub-genome in all cultivars (excluding UK cultivars which are not assembled), and 10,653 as microsyntenic, belonging to the same chromosome and sub-genome in all cultivars (excluding UK cultivars). From these 66 30-lets were not taken forward in the analysis due to low expression and/or quality filtering determined by DESeq2 (R package v

4.0.3) of at least one homoeolog in each set. Relative expression of 30-lets across homoeologs and associated sub-genome expression biases were calculated as previously reported, through use of our triad.expression R package (https://github.com/AHallLab/triad.expression).

Analysis of Norin 61-specific genes and expression pattern

Norin 61 unique genomic regions were defined as regions with conserved ratios of < 0.5 among nine cultivars: Stanley, SYMattis, ArinaLrFor, Jagger, Julius, Lancer, Landmark, Mace, and Norin 61. The average conservation ratio was calculated using a 2 Mbp window and 5 kbp step size. Fisher's exact test was used to determine whether Norin 61-specific genes were enriched in these Norin 61 unique genomic regions. Furthermore, unique TE regions of Norin 61 identified by the pattern of TEs were analyzed similarly [22]. Tissue specificity of gene expression (tau) was calculated using the expression levels of Norin 61 using the same method described above. Genes with a tau value > 0.8 were considered as genes with tissue-specific expression. GO enrichment analysis was performed by using R Bioconductor package topGO version 2.54.0 [61] using elim algorithm and Fisher's exact test with FDR < 0.05, with 195 expressed Norin 61-specific genes excluding transposable elements. Only GO terms associated with more than ten genes were considered.

Co-expression Analysis (WGCNA)

The WGCNA R package [65] (R version 3.6.0) was used to build co-expression networks for four cultivars (Arina*LrFor*,, Jagger, Julius and Norin 61) for which we had triplicate biological replicates for each developmental stage. These cultivars also span the range of genetic variation observed in the previously published pan-genome [4]. The expression matrices for each of the selected cultivars contained DESeq2 normalised counts of high confidence genes derived from alignment to the respective chromosome level assemblies and corresponding de novo annotations. These matrices were filtered and genes where the sum of counts across all samples was greater than 5 in at least 2 samples were retained. We used WGCNA to construct a signed networks for each cultivar using the blockwiseModules () function. A soft power threshold of 9 (ArinaLrFor, Jagger) or 10 (Julius, Norin 61) was used, together with the following parameters; minModuleSize = 30, corType = bicor, maxPOutliers = 0.05, mergeCutHeight = 0.3, minKMEtoStay = 0.4, maxBlockSize = 35,000. Eigengenes were then extracted for each module from each of the resulting four cultivar networks.

Defining threshold for classifying inter-module relationships

To classify inter-module relationships and identify modules with divergent or similar patterns of expression we defined a threshold of module similarity. Initially we calculated the distance between each pairwise module comparison for each cultivar network, using the Pearson correlation distance. For each cultivar we used the maximum distance of each of these pairwise comparisons, for each module and calculated the medians of these maximas. Next we investigated the proportion of 30-let triads identified as split across network modules, that would be classed as divergent using a module similarity threshold of 0-100%. From these results we selected a module similarity threshold of 85% the median of maximum distances, with distances above this classed as divergent and distances below, classed as similar.

Identifying metamodules

We used the R package clValid [66] to determine the optimal number of clusters for the 68 module eigengenes from across all networks. The resulting Dunn index [67] and silhouette width [68] indicated that the optimum number of clusters for our ME dataset was 6. We calculated the pairwise Pearson correlation coefficients for all our 68 cultivar ME (cor()) and converted this to a dissimilarity matrix (as.dist()). We used hierarchical clustering of this dissimilarity matrix (hclust()) to define metamodules.

Visualising highly connected cloud genes within the Julius network

We identified a network module both significantly enriched for and containing the highest number of cloud genes (29; JULbrown) and used the adjacency() function of WGCNA to determine the network adjacency of each of these cloud genes within the JULbrown module. We selected the five most highly connected module genes to each cloud gene and used the R package igraph [69] to visualise the integration of these cloud genes into the JULbrown module. Using the graph adjacency() function, graph adjacencies were created based on the Pearson correlation distances between genes in pairwise fashion. These directed graphs were simplified to remove multiple edges and loops, filtered to retain only those connections with an absolute Pearson correlation > 0.9. The mst() function using the prim algorithm was used to create a minimum spanning tree and the resulting subgraph was visualised using plot() with isolated nodes excluded.

Comparative analysis of immune reactive regions in the wheat pan-genome

Reference allergen identification and chromosome 6D comparison

Reference allergens in the wheat pan-genome were filtered using blastn algorithm against the identified sequences in the IWGSC v1 gene annotation v1.1 [30]. To identify unannotated gliadin and glutenin gene models and to compare the potential immune reactivity of the wheat cultivars, known CD-associated HLA-DQ T-cell epitopes were reverse translated, and the consensus nucleotide sequences were used for a motif search with 100% sequence identity. The mapped epitope-rich regions were used to compare the alpha-gliadin locus in chromosome 6D. Additional gene models representing complete gene models with DQ epitopes were manually annotated. The locus organisation was compared to the Chinese Spring chromosome 6D alpha gliadin locus in the IWGSC v1 reference genome assembly [30].

Epitope expression analysis

The epitope expression values were calculated by multiplying the DESeq normalised count values of genes where the reverse-translated consensus epitope sequence was detected by the number of epitopes in each sequence. The resulting values were then added together for each epitope type, as well as epitope types at the genome level.

Promoter motif enrichment analysis

Gene models representing the major alpha-gliadin and gamma-gliadin gene models were selected for the comparative analysis. To investigate the impact of transcriptional regulation at sub-genome levels, epitope group-specific 1000 bp promoter sequence lists were extracted

from the homoeologous chromosome 1 and 6 groups of the A, B and D sub-genomes using the gene models of the chromosome assembled pan cultivars.
Transcription factor binding site (TFBS) enrichment analysis was performed using the simple enrichment analysis (SEA) algorithm in the MEME suite [70], with the non-redundant plant-specific JASPAR 2022 motif collection for mapping. As a control promoter sequence list, 1000-bp promoter sequences were extracted from the Chinese Spring reference genome [24] high-confident gene models, and 1% randomly sampled promoter sequences representing gene models across the 21 chromosomes were used.
TFBS enrichment results were filtered at an adjusted p-value <0.05 using the method proposed by Benjamini and Hochberg [71] and motif hits characteristic of 100% of the analysed promoter sequences (100% true positive list), were used further analysed. Motifs showing gluten gene family type (alpha or gamma gliadin) and sub-genome specificity were also highlighted (**Table S27**). Motif enrichment ratios of each motif-epitope group pair were visualised in a clustered heatmap.

Gene co-expression analysis
Gene models with known TF functions were filtered using the existing gene annotations [24]; [6]. Correlation values between DESeq normalised counts>1 log2 transformed DESeq counts of glutenin and gliadin gene models and the TF gene models expressed in the grain tissue were calculated in R. Hits with r>0.8 and adj. p-value <0.01 were filtered and used for further analysis (**Table S2**). Wheat orthologs of representative TF genes identified through the enriched significant TFBSs were filtered and matched with the representative promoter motifs. The grain-specific co-expression network was created in Cytoscape (version 3.10.2) using a co-expression cut-off value of 0.8. The resulting network was annotated with the reference allergen-specific information for disease-relatedness and gene family. The first neighbour network was visualised in Cytoscape (**Figure S8**).

Pan-genome graph construction of 10Mb 6D region
We extracted a 10Mb region (20-30Mb) encompassing the alpha gliadin locus from the top of chromosome 6D for the cultivars Norin 61, CDC Stanley, SY Mattis, Julius, and Mace. To estimate the divergence of the input sequences, we used mash-2.2 [72], specifically the mash triangle command, to calculate a maximum sequence divergence of 0.039. To account for possible underestimation of sequence divergence and localised structural variants, we specified a minimum mapping identity value (-p 90) for pangraph construction using PGGB [73] together with segment size (-s 30kb), number of mappings (-n 6), minimum length of exact matches (-k 311), target sequence length for POA (-G 13117, 13219), mean length of each sequence pad for POA (-O 0.03) and k-mer size for mapping (-K 111). Default settings were used for all other parameters.

Extracting the alpha-gliadin locus sub-pangraph
Using ODGI toolkit [74] we extracted the subgraph of the alpha-gliadin locus from our 6D graph build. We used the odgi extract command together with coordinates of the Norin 61 gene models described in **Table S29** to extract the 520.7kb region encompassing the locus (6D: 26,703,647-27,222,360bp) and the corresponding paths intersecting with this region in CDC

Stanley (6D: 28,164,601-28,660,350 bp), SY Mattis (6D: 26,645,382-27,096,594 bp), Julius (JUL 6D: 26,983,100-27,437,565 bp), and Mace (6D: 26,808,846–27,298,593 bp). We used odgi sort to sort the resulting subgraph and odgi procbed to adjust the coordinates of the gene models for each cultivar to fit the resulting subgraph. odgi inject allowed us to visualise the placement of these gene models across the graph and identify cultivar-specific haplotypes. We generated a graphical fragment assembly (gfa) of this subgraph using odgi view (**Table S30**).

Data Availability

The genome sequence and gene annotations of all wheat cultivars can be viewed and downloaded in Ensembl Plants (https://plants.ensembl.org/index.html). This includes the *de novo* genes for the chromosome level cultivars generated within this study, and projected genes for all assemblies from the IWGSC RefSeq v1.1 annotation. All raw data used in this study is available at the European Nucleotide Archive under accession PRJEB51827.

Code Availability

Relevant code repositories are referenced throughout the Methods sections.

Supplementary Tables

All supporting tables and associated materials are available at https://opendata.earlham.ac.uk/wheat/under_license/toronto/Hall_2024-01-01_wheat_pantranscriptome.

Acknowledgements

We would like to acknowledge Technical Genomics and Core Bioinformatics Groups at Earlham Institute funded by BBSRC National Capability in Genomics and Single Cell Analysis (BBS/E/T/000PR9816) and BB/CCG1720/1 for the physical HPC infrastructure and data centre delivered via the NBI Research Computing group. Part of this work was delivered via the Earlham Institute Strategic Programme Grant Decoding Biodiversity BBX011089/1 (BBS/E/ER/230002B). Delivering Sustainable Wheat BB/X011003/1 (BBS/E/ER/230003C). Norwich Research Park Biosciences Doctoral Training Partnership grant BB/M011216/1. We also would like to thank the Australian Research Council Centre of Excellence for Innovations in Peptide and Protein Science (CE200100012) and Coeliac Australia (G1005443) as well as MEXT JSPS KAKENHI 22H05179, 22H02316, 22K21352, Swiss National Science Foundation 310030_212551, Japan Science and Technology Agency JPMJCR16O3, URPP Evolution in Action and UZH Global Strategy and Partnerships Funding Scheme of the University of Zurich. Helmholtz Munich would like to acknowledge support by BMBF project number 031B0190 (de.NBI). We would also like to thank the 10+ Wheat Genome Project for their feedback and guidance.

Author Contributions

CP arranged supplying seeds. JS & CU managed growing plants and supplying materials. SD & HR prepared samples and performed extractions for sequencing. NI, SH, TB, HC, LC and KG

generated and managed the iso-seq and RNA-seq data production. BW, RJ & JC performed quality control of biological replicate data. SW & BW reworked the existing package for sub-genome expression bias characterisation. BW carried out gene expression analysis, tissue specific indexing and investigation of sub-genome biases. TL, GK & DS performed *de novo* annotations. TL, DL & HG consolidated *de novo* annotations, generated orthogroups, and investigated introgressions and copy number variation. RRP performed co-expression analysis, generated networks and gliadin pangraph, JW performed genome kmer comparison. AJ & UB performed analysis of allergen and prolamin diversity, MG supervised the prolamin analysis. KKS, MO, HH and SN investigated Norin 61 specific gene expression patterns. GN imported and managed data for EnsemblPlants. BW, TL, RRP, JW, CU, MS & AH contributed to preparing the manuscript. PB, WH, CU, NS, SK, JP, KM & CP provided intellectual input. EL and AB provided funding. The 10+ Wheat Genome Project, EL, AB, AH and MS for conceptualisation of the project.

Ethics declarations

The authors declare no competing interests.

References

1. Braun, H.-J., Rajaram, S. & Ginkel, M. van. CIMMYT's approach to breeding for wide adaptation. *Euphytica* **92**, 175–183 (1996).
2. Erenstein, O. *et al.* Wheat Improvement, Food Security in a Changing Climate. 47–66 (2022) doi:10.1007/978-3-030-90673-3_4.
3. Levy, A. A. & Feldman, M. Evolution and origin of bread wheat. *Plant Cell* **34**, 2549–2567 (2022).
4. Walkowiak, S. *et al.* Multiple wheat genomes reveal global variation in modern breeding. *Nature* **588**, 277–283 (2020).
5. Bayer, P. E. *et al.* Wheat Panache: A pangenome graph database representing presence–absence variation across sixteen bread wheat genomes. *Plant Genome* **15**, e20221 (2022).
6. Ramirez-Gonzalez, R. H. *et al.* The transcriptional landscape of polyploid wheat. *Science (New York, NY)* **361**, eaar6089 (2018).
7. An, H. *et al.* Transcriptome and organellar sequencing highlights the complex origin and diversification of allotetraploid Brassica napus. *Nat. Commun.* **10**, 2878 (2019).
8. Sreedasyam, A. *et al.* Genome resources for three modern cotton lines guide future breeding efforts. *Nat. Plants* **10**, 1039–1051 (2024).
9. Thind, A. K. *et al.* Chromosome-scale comparative sequence analysis unravels molecular mechanisms of genome dynamics between two wheat cultivars. *Genome Biol.* **19**, 104 (2018).
10. Khan, A. W. *et al.* Super-Pangenome by Integrating the Wild Side of a Species for Accelerated Crop Improvement. *Trends Plant Sci.* **25**, 148–158 (2020).
11. Shang, L. *et al.* A super pan-genomic landscape of rice. *Cell Res.* **32**, 878–896 (2022).
12. Lovell, J. T. *et al.* GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. *eLife* **11**, e78526 (2022).
13. He, F. *et al.* Genomic variants affecting homoeologous gene expression dosage contribute to agronomic trait variation in allopolyploid wheat. *Nat. Commun.* **13**, 826 (2022).

14. Leister, D. Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance genes. *Trends Genet.* **20**, 116–122 (2004).
15. Hanada, K., Zou, C., Lehti-Shiu, M. D., Shinozaki, K. & Shiu, S.-H. Importance of Lineage-Specific Expansion of Plant Tandem Duplicates in the Adaptive Response to Environmental Stimuli . *Plant Physiol.* **148**, 993–1003 (2008).
16. Bayer, P. E., Golicz, A. A., Scheben, A., Batley, J. & Edwards, D. Plant pan-genomes are the new reference. *Nat. Plants* **6**, 914–920 (2020).
17. 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**, 896–904 (2019).
18. Wang, T. *et al.* Pan-genome analysis of 13 Malus accessions reveals structural and sequence variations associated with fruit traits. *Nat. Commun.* **14**, 7377 (2023).
19. He, Q. *et al.* A graph-based genome and pan-genome variation of the model plant Setaria. *Nat. Genet.* **55**, 1232–1242 (2023).
20. Jayakodi, M. *et al.* The barley pan-genome reveals the hidden legacy of mutation breeding. *Nature* **588**, 284–289 (2020).
21. Gordon, S. P. *et al.* Extensive gene content variation in the Brachypodium distachyon pan-genome correlates with population structure. *Nat. Commun.* **8**, 2184 (2017).
22. Shimizu, K. K. *et al.* De Novo Genome Assembly of the Japanese Wheat Cultivar Norin 61 Highlights Functional Variation in Flowering Time and Fusarium-Resistant Genes in East Asian Genotypes. *Plant Cell Physiol.* **62**, 8–27 (2020).
23. Keilwagen, J. *et al.* Detecting major introgressions in wheat and their putative origins using coverage analysis. *Sci. Rep.* **12**, 1908 (2022).
24. IWGSC, T. I. W. G. S. C. *et al.* Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science (New York, NY)* **361**, eaar7191 (2018).
25. Wang, Y. *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49–e49 (2012).
26. Glagoleva, A. Y. *et al.* Features of Activity of the Phenylpropanoid Biosynthesis Pathway in Melanin-Accumulating Barley Grains. *Front. Plant Sci.* **13**, 923717 (2022).
27. Wu, X., Zhang, X., Huang, B., Han, J. & Fang, H. Advances in biological functions and mechanisms of histone variants in plants. *Front. Genet.* **14**, 1229782 (2023).
28. Colin, L. A. & Jaillais, Y. Phospholipids across scales: lipid patterns and plant development. *Curr. Opin. Plant Biol.* **53**, 1–9 (2020).
29. Stefanowicz, K., Lannoo, N. & Damme, E. J. M. V. Plant F-box Proteins – Judges between Life and Death. *Crit. Rev. Plant Sci.* **34**, 523–552 (2015).
30. Juhász, A. *et al.* Genome mapping of seed-borne allergens and immunoresponsive proteins in wheat. *Sci. Adv.* **4**, eaar8602 (2018).
31. Tye-Din, J. A. *et al.* Comprehensive, Quantitative Mapping of T Cell Epitopes in Gluten in Celiac Disease. *Sci. Transl. Med.* **2**, 41ra51 (2010).
32. Halstead-Nussloch, G. *et al.* Multiple Wheat Genomes Reveal Novel Gli-2 Sublocus Location and Variation of Celiac Disease Epitopes in Duplicated α-Gliadin Genes. *Front. Plant Sci.* **12**, 715985 (2021).
33. Huo, N. *et al.* Rapid evolution of α-gliadin gene family revealed by analyzing Gli-2 locus regions of wild emmer wheat. *Funct. Integr. Genom.* **19**, 993–1005 (2019).

34. Huo, N. *et al.* Gene Duplication and Evolution Dynamics in the Homeologous Regions Harboring Multiple Prolamin and Resistance Gene Families in Hexaploid Wheat. *Front. Plant Sci.* **9**, 673 (2018).
35. Huo, N. *et al.* Dynamic Evolution of α-Gliadin Prolamin Gene Family in Homeologous Genomes of Hexaploid Wheat. *Sci. Rep.* **8**, 5181 (2018).
36. Merlino, M. *et al.* Wheat DOF transcription factors TaSAD and WPBF regulate glutenin gene expression in cooperation with SPA. *PLOS ONE* **18**, e0287645 (2023).
37. Juhász, A., Makai, S., Sebestyén, E., Tamás, L. & Balázs, E. Role of Conserved Non-Coding Regulatory Elements in LMW Glutenin Gene Expression. *PLoS ONE* **6**, e29501 (2011).
38. Liao, W.-W. *et al.* A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
39. Cheng, H. *et al.* Frequent intra- and inter-species introgression shapes the landscape of genetic variation in bread wheat. *Genome Biol.* **20**, 136 (2019).
40. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
41. Kovaka, S. *et al.* Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278 (2019).
42. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
43. Gremme, G., Brendel, V., Sparks, M. E. & Kurtz, S. Engineering a software tool for gene structure prediction in higher organisms. *Inf. Softw. Technol.* **47**, 965–978 (2005).
44. Ghosh, S. & Chan, C.-K. K. Plant Bioinformatics, Methods and Protocols. *Methods Mol. Biol.* **1374**, 339–361 (2015).
45. Hoff, K. J. & Stanke, M. Predicting Genes in Single Genomes with AUGUSTUS. *Curr. Protoc. Bioinform.* **65**, e57 (2019).
46. Nachtweide, S. & Stanke, M. Gene Prediction, Methods and Protocols. *Methods Mol. Biol.* **1962**, 139–160 (2019).
47. Venturini, L., Caim, S., Kaithakottil, G. G., Mapleson, D. L. & Swarbreck, D. Leveraging multiple transcriptome assembly methods for improved gene structure annotation. *GigaScience* **7**, giy093- (2018).
48. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology* **215**, 403–410 (1990).
49. Seppey, M., Manni, M. & Zdobnov, E. M. Gene Prediction, Methods and Protocols. *Methods Mol. Biol.* **1962**, 227–245 (2019).
50. Nevers, Y. *et al.* Quality assessment of gene repertoire annotations with OMArk. *Nat. Biotechnol.* 1–10 (2024) doi:10.1038/s41587-024-02147-w.
51. Kiełbasa, S. M., Wan, R., Sato, K., Horton, P. & Frith, M. C. Adaptive seeds tame genomic sequence comparison. *Genome Res.* **21**, 487–493 (2011).
52. Blanchette, M. *et al.* Aligning Multiple Genomic Sequences With the Threaded Blockset Aligner. *Genome Res.* **14**, 708–715 (2004).
53. Hickey, G., Paten, B., Earl, D., Zerbino, D. & Haussler, D. HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics* **29**, 1341–1342 (2013).
54. Quinlan, A. R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr. Protoc. Bioinform.* **47**, 11.12.1-11.12.34 (2014).

55. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015).
56. Clavijo, B. J., Accinelli, G. G., Yanes, L., Barr, K. & Wright, J. Skip-mers: increasing entropy and sensitivity to detect conserved genic regions with simple cyclic q-grams. *bioRxiv* (2017) doi:10.1101/179960.
57. Waskom, M. seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).
58. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
59. Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. & Salzberg, S. L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* **11**, 1650–1667 (2016).
60. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
61. *Bioconductor - topGO*.
62. Vega, J. J. D. *et al.* Differential expression of starch and sucrose metabolic genes linked to varying biomass yield in Miscanthus hybrids. *Biotechnol. Biofuels* **14**, 98 (2021).
63. Yanai, I. *et al.* Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650–659 (2005).
64. Kryuchkova-Mostacci, N. & Robinson-Rechavi, M. A benchmark of gene expression tissue-specificity metrics. *Brief. Bioinform.* **18**, 205–214 (2017).
65. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform.* **9**, 559 (2008).
66. Brock, G., Pihur, V., Datta, S. & Datta, S. clValid: An R Package for Cluster Validation. *Journal of Statistical Software* **25**, 1–22 (2008).
67. Dunn†, J. C. Well-Separated Clusters and Optimal Fuzzy Partitions. *J. Cybern.* **4**, 95–104 (1974).
68. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
69. Cśardi, G. & Nepusz, T. The igraph software package for complex network research. in.
70. Bailey, T. L. & Grant, C. E. SEA: Simple Enrichment Analysis of motifs. *bioRxiv* 2021.08.23.457422 (2021) doi:10.1101/2021.08.23.457422.
71. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc.: Ser. B (Methodol.)* **57**, 289–300 (1995).
72. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).
73. Garrison, E. *et al.* Building pangenome graphs. *Nature Methods* (2024) doi.org/10.1038/s41592-024-02430-3.
74. Guarracino, A., Heumos, S., Nahnsen, S., Prins, P. & Garrison, E. ODGI: understanding pangenome graphs. *Bioinformatics* **38**, btac308- (2022).