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

Authors

Benjamen White[1†], Thomas Lux[2†], Rachel L Rusholme-Pilcher[1†], Angéla Juhász[6†], Gemy Kaithakottil[1], Susan Duncan[1,3], James Simmonds[3], Hannah Rees[1], Jonathan Wright[1], Joshua 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], 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 Krattinger[10], Nils Stein[11,15], Klaus F.X. Mayer[2,12], Curtis Pozniak[13], 10+ Wheat Genome Project, Manuel Spannagl^[2], 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

[†]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 the genomes of 15 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 nine of 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 homoeologous genes, but also widespread changes in sub-genome homoeolog expression bias between cultivars. Co-expression network analysis revealed the impact of divergence of sub-genome homoeolog expression and identified tissue-associated cultivar-specific expression profiles. 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]. 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, reveals novel gene models across wheat cultivars and facilitates 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[5].

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. To date, studies of plant pan-transcriptomes either rely on read alignment to a single reference genome which can result in reference bias, or generate *de novo* transcript assemblies from short read data that can accumulate errors and technical artefacts[6].

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. We identify evidence of widespread gene duplication and deletion, revealing the population structures imposed by repeated hybridisations from wild relatives and different breeding programmes. 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 sub-genomes. We exemplify the value of these analyses through an 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. 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 **Extended Data Figure 1** 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 (**Extended Data Figure 2A**) was developed to identify and correct for missed gene models. This step ensures the best possible comparability between the wheat genomes and gene repertoire[7].

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 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. The *de novo* annotations are available in Ensembl Plants release 52.

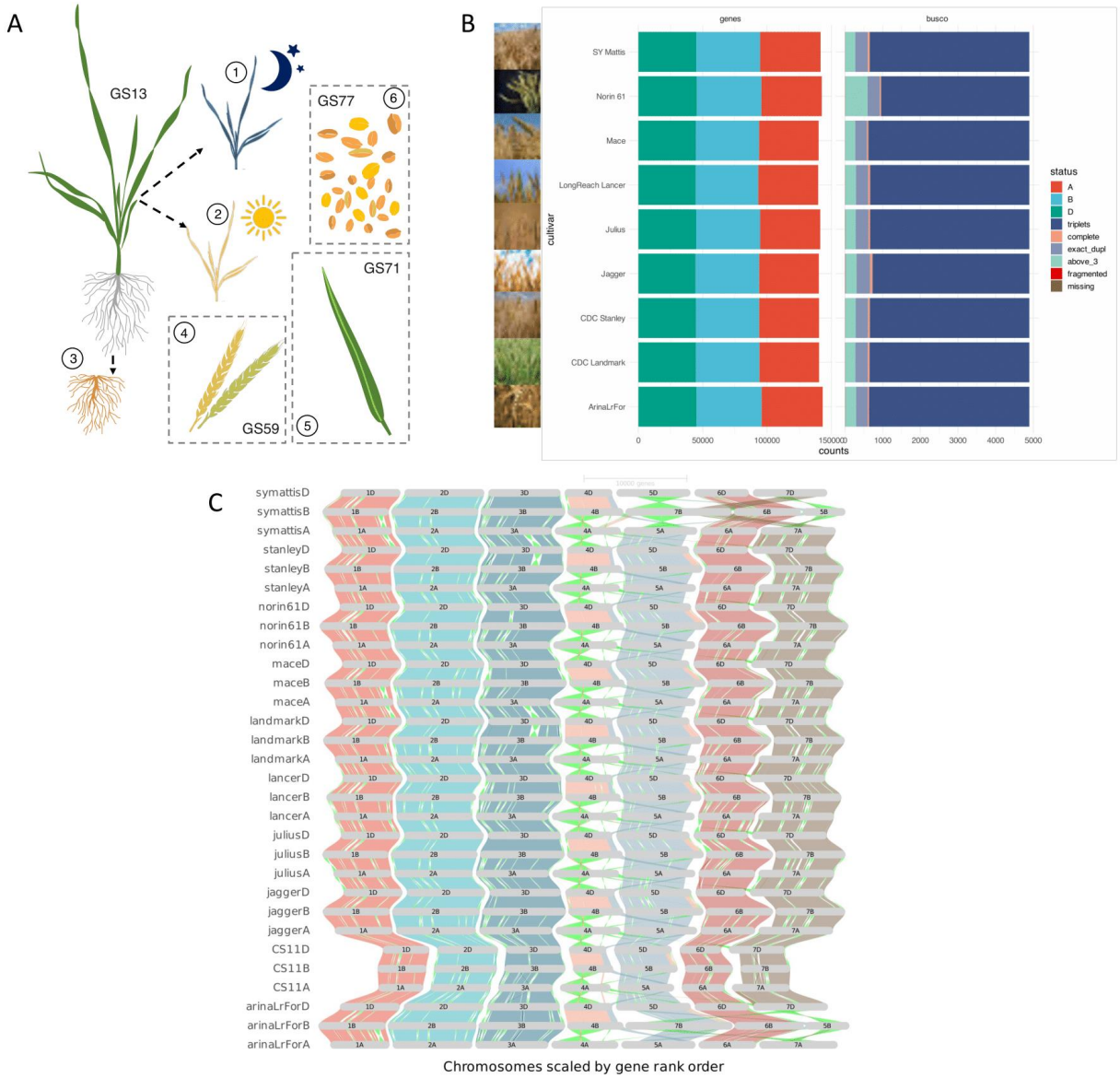

Figure 1. Study design, de-novo gene annotations and orthologous framework. **A**. Overview of transcriptome data generated for this study of the wheat pan-cultivars. 1 & 2: whole aerial organs sampled at dawn and dusk, 3: root, 4: complete spike at heading (GS59), 5: flag leaf 7 days post anthesis (GS71), 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 sub-genome) as well as a 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.

Genes and gene families exclusive or amplified in a specific cultivar are of major interest in a pan-genomic context[8]. 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 high-density orthologous framework for wheat. GENESPACE[9] was used to derive syntenic relationships between all chromosomes and sub-genomes, 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[10].

We identified groups of orthologous genes (referred to as orthogroups) among the wheat high-confidence gene models of all cultivars. A total of 54,865 orthogroups contained 99.7% of all genes, with 173 orthogroups identified as cultivar-specific and 3,756 genes not clustered in any orthogroup - defining the cloud genome. Cloud and shell genes have previously been found to be associated with disease resistance[11], adaptation to new environments[12], or important agricultural traits[13]. 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 Extended Data Figure 3**) that are likely associated with crosses to distinct material or hybridisations with wild or domesticated relatives; events common in wheat[14].

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 76.34% of genes were classified as core, 23.32% as shell, and 0.33% as cloud (**Extended Data Figure 2B**). Amongst the core gene set, we found biological functions associated with basic metabolic, catabolic and DNA repair/replication processes enriched (**Supp. Table 1**), while stress response and regulation of gene expression were overrepresented in the shell genes (**Supp. Table 2**). In the set of cloud genes, functions related to chromatin organisation and reproductive processes were found to be enriched (**Supp. Table 3**). Expression patterns of core, shell and cloud genes revealed pronounced differences globally, but not between the sub-genomes (**Figure 2D**). As observed in other pan-genomes[15], core genes tend to be higher expressed in all sub-genomes and tissues, as compared to both shell and cloud genes.

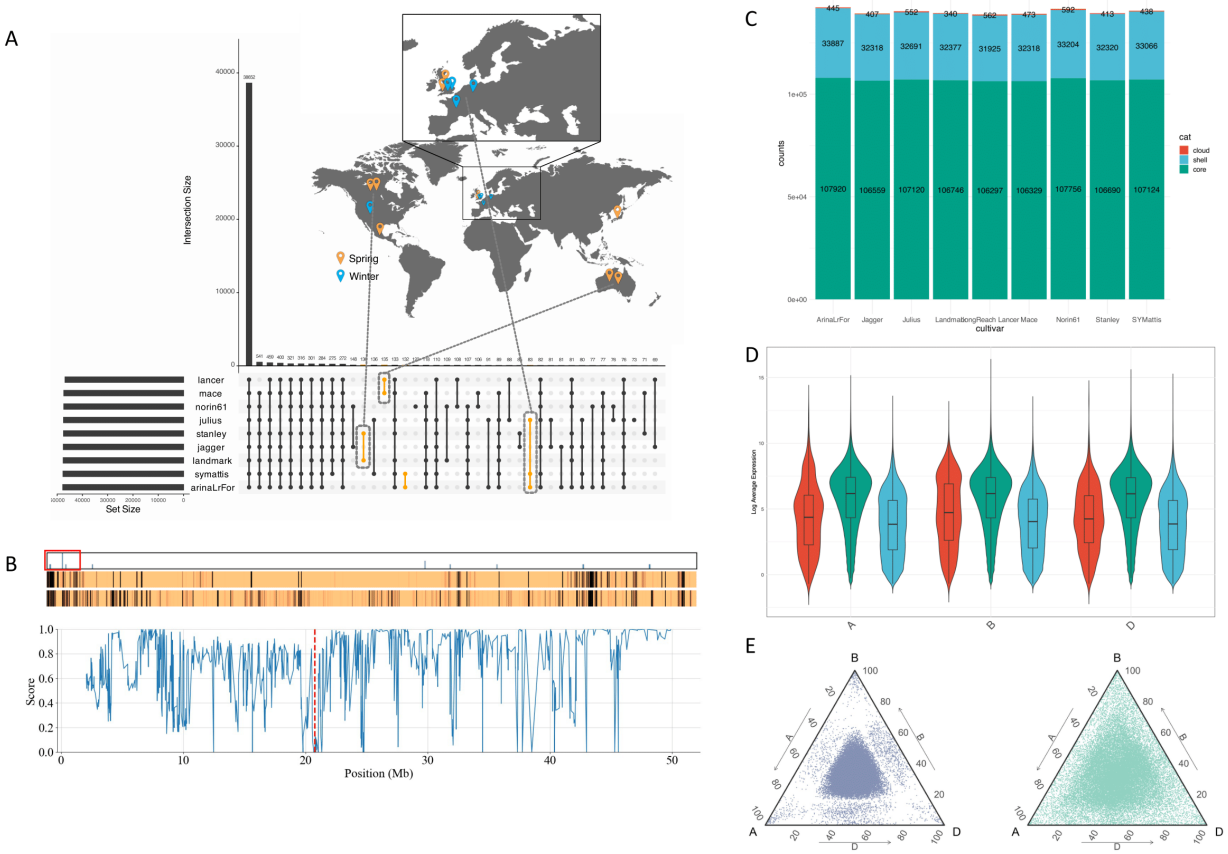

Figure 2. **The wheat core-, shell-, and cloud- genome and homoelogous expression patterns.** **A.** UpSet plot showing intersects of orthogroup conservation between pan-cultivars and the relation to their breeding programs and sowing season. **B.** A representation of CDC Stanley chromosome 3B showing 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.** Number of genes belonging to core, shell and cloud ortholog groups across cultivars. **D.** Violin plots of core, cloud and shell log average gene expression across all combined cultivars and tissues, in each sub-genome. Higher mean expression was observed in core genes across all sub-genomes. **E.** Ternary plots, of stable (left) and dynamic (right) 30-let expression, where there is a homoeolog present on each sub-genome, of all tissue in all cultivars, combined, showing more overall balanced expression in stable 30-lets and unbalanced expression in dynamic 30-lets.

Duplication of genes has been identified as a major driver of gene function evolution and adaptation in plants[12]. 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[16,4]. 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 5,040 tandem arrays (HC genes only) in each cultivar, with the lowest in CDC Landmark (4,914) and Arina*LrFor* as the highest (5,172). In addition, we tested whether there is a bias in expression towards one member of the array. We found that for 2,520 arrays one of the two members was biased in its expression with respect to the other member, whereas for 1,800 arrays both copies were expressed at similar levels or varied depending on the tissue (**Extended Data Figure 2C**). 719 tandem arrays were not expressed under the investigated conditions. Amongst all tandemly duplicated genes in wheat, biological functions associated with phosphorylation, response to stimulus and stress and reproductive processes were enriched (**Supp. Table 4**). We also investigated the conservation of tandem arrays across all pan-cultivars. Around 69% of the tandem arrays identified in a specific cultivar were found to be shared with all other pan-cultivars, with the remaining 31% tandems showing varying conservation (**Extended Data Figure 2D**). 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 used 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 (**Extended Data Figure 1A, B**). 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 also showing an overall higher mean expression than either shell or cloud (**Figure 2D, Extended Data Figure 4A**). Indicating a global conservation of expression, irrespective of tissue type or sub-genome biases.

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 (**Extended Data Figure 4B**). 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 the 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[5]. In wheat, previous studies have focused on tissue-specific expression across homelogous 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 with a homoeolog present on each sub-genome of the *de novo* annotated cultivars. Using previously reported cut-off values[5], 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 (**Extended Data Figure 4C**)[5]. 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 (**Extended Data Figure 4D**). 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.(**Extended Data Figure 4E**). 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 (**Extended Data Figure 4F**). 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.

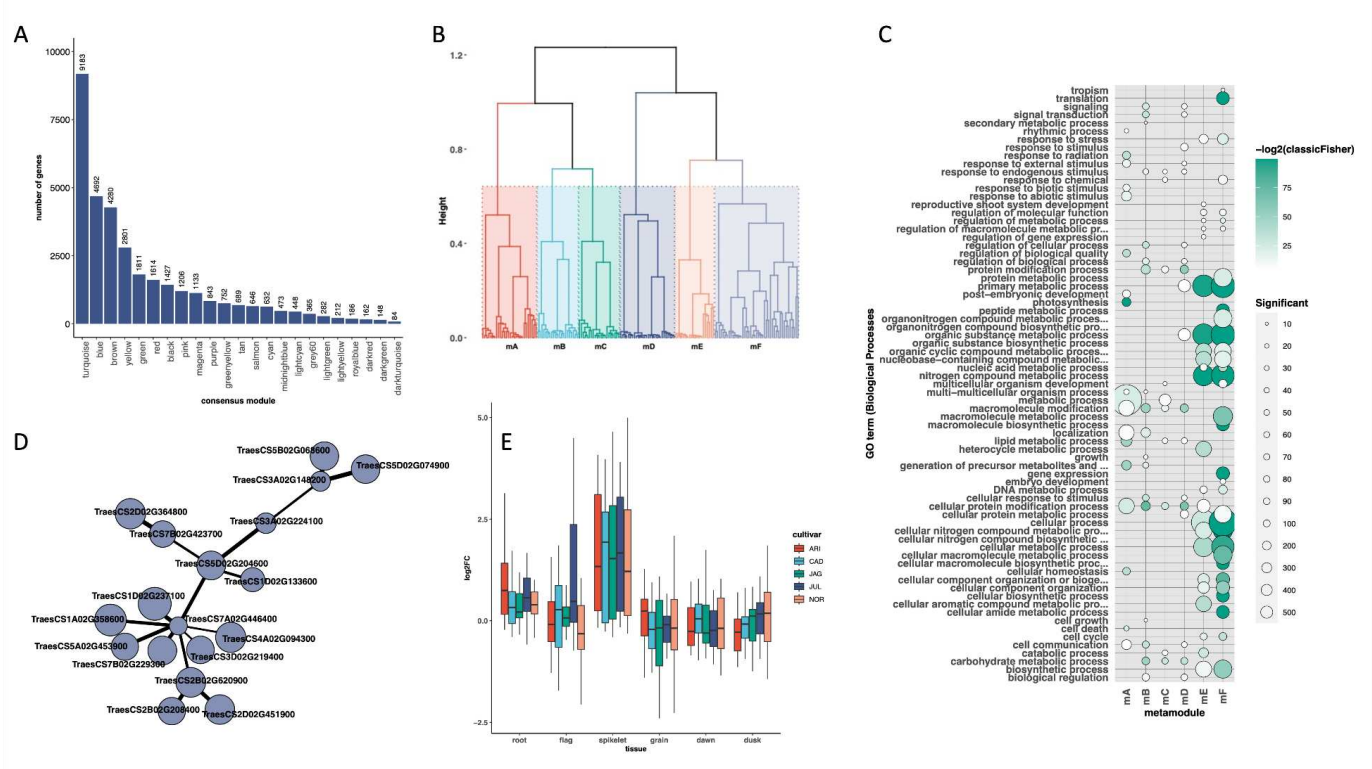

Figure 3. Components of the wheat consensus network with functional annotation and UK cultivar Claire-specific differences A. Number of genes in each module from the seven cultivar consensus network. **B.** Hierarchical clustering of 23 consensus module eigengenes identifying six metamodules. **C.** GO-slim terms of biological processes associated with genes in 6 metamodules. Only terms with p<0.05 and >10 significant genes are shown. Bubble colour indicates the -log$_2$*p*-value significance from Fisher's exact test and size indicates the frequency of the GO-slim term in the underlying EBI Gene Ontology Annotation database (larger bubbles indicate more general terms). **D**. Claire-specific network of 18 genes with divergent expression patterns compared to all other cultivars. Node size is scaled to the log$_2$ average expression +1 of each gene in spikelet tissues and edge width reflects the weight of the connection between nodes. **E.** Log$_2$ fold change in expression across all tissues, between Claire and remaining cultivars of 18 Claire-specific genes.

Conserved patterns of co-expression

In addition to the nine cultivars for which we generated *de novo* annotations, we also collected expression data from the same tissues and conditions for the remaining five cultivars included in the wheat pan-genome. To exploit these additional transcriptomic datasets in the absence of *de novo* annotations, we employed the consensus network function of WGCNA[17] which uses expression counts from a single reference (CSv1.1) to enable the identification of gene expression modules that represent biological pathways conserved across all cultivars.

To explore how regulatory networks are conserved across tissues and cultivars of the pan-transcriptome, we constructed our consensus co-expression network spanning seven cultivars;

four with chromosome pseudomolecule assemblies (Arina*LrFor*, Julius, Jagger, Norin 61) and three with scaffold assemblies; two UK derived (Cadenza, Claire) and one key CIMMYT (Weebill) cultivar. Together these seven cultivars represent a wide range of the genetic variation observed in the pan-genome[4].

To identify sets of genes with highly correlated expression patterns across cultivars we used high-confidence genes expressed at greater than five normalised counts in at least three samples, with less than two standard deviations in normalised counts, between sample replicates. The resulting 48,337 genes from each cultivar (338,359 genes in total) were used to construct the consensus network, with modules ranging in size from 84 to 9,183 genes (**Figure 3A**). This consensus network accounted for 70% (34,069 genes) of the genes from each cultivar in our original dataset. These modules comprised co-expressed genes that exhibited highly similar expression patterns across all six developmental stages in all seven cultivars. A comparison of the consensus module eigengenes (MEs) for these 23 modules demonstrated the conservation of expression patterns between cultivars (**Extended Data Figure 5**). We used hierarchical clustering to further collapse the 23 consensus modules into six consensus metamodules (**Figure 3B**); comprising 32,936 genes from each cultivar. GO-slim term analysis (**Figure 3C**) and distinct transcription factor superfamily membership (**Extended Data Figure 6**) indicated that each metamodule could be associated with distinct biological processes (**Supp. Table 5**).

We defined a threshold for classifying inter-consensus module relationships and used this to make pairwise comparisons of the ME for each consensus module and identify modules with divergent or similar patterns of expression. We then used these module relationships to compare how the 30-let triads were split across our consensus network. Within the genes used to build our consensus network, we identified 6,867 of the 10,521 complete triad sets (50.8%). 3,640 (53.0%) of these triads were assigned to modules within the network, with the remaining 3,227 (47%) triads having at least one member present in the unclustered set of genes. This set of genes that could not be fitted into the consensus network will contain genes with low variance or low expression across tissues, and genes that do not show the same pattern of expression across all seven cultivars. Of the 3,640 triads within our network 3,548 (96%) belonged to either the same or similar expression modules, reflecting the conservation of expression between the A, B and D sub-genomes. 2,431 of these triads (66.8%) belonged to the stable category of 30-lets defined previously through comparison of individual triad expression balance across all tissues and 9 cultivars. The identification of co-expression modules containing similarly expressed triad members reveals additionally conserved genes, tightly connected to these stable triads. Using the consensus network, we also observed 146 of the 3,640 triads (4%) where sub-genome members were split across divergent modules. These triads were significantly enriched for GO terms associated with the regulation of signal transduction and DNA metabolic process (**Supp. Table 6**).

Using a consensus network approach to identify cultivar-specific co-expression patterns

Whilst our consensus network enables robust biological inferences through the identification of conserved gene sets across all cultivar members of the network, we also report a set of 14,268 unclustered genes that cannot be fitted to the consensus modules. These 14,268 genes may have an expression profile that correlates with a consensus ME, but as this pattern of expression is not conserved across all seven cultivars, these genes will not be placed within the consensus network. Within our unclustered gene set we identified 1,753 such genes where the pattern of expression for a single cultivar was closely correlated to a consensus ME displaying an expression profile divergent to the module containing the same gene in the remaining six cultivars. This enabled us to identify sets of co-expressed cultivar-specific genes (**Supp. Table 7**). We visualised 12 of these cultivar-specific network fragments using igraph[18] including a set of 18 linked genes with increased expression in spike tissue in Claire compared to other cultivars (**Figure 3D & 3E**). The most highly connected gene in this subnetwork (TraesCS7A02G446400) is a transducin/WD40 repeat protein. These proteins are key regulators of both plant developmental and stress processes, and are known to participate in histone modification, transcriptional regulation and signal transduction[19]. Additional genes in this cluster, are annotated as protein phosphatases, an eRF1 transcription factor, a calcium-binding EF-hand domain-containing protein and a polyadenylation specificity factor. We hypothesise that the cultivar-specific expression pattern observed in Claire linked to increased expression in spike tissues could be the result of cultivar-specific regulation of a developmental or stress response.

Co-expression network analysis using *de novo* gene models

Our consensus network approach, using a common reference, enabled us to identify high confidence, conserved expression modules and identify cultivar-specific co-expressed gene sets. However, the use of a common reference meant that we were unable to assess the *de novo* contribution of each genome to the consensus network. Of the seven cultivars within the consensus network, Arina*LrFor*, Jagger, Julius and Norin 61 each have corresponding *de novo* annotations. We used these *de novo* gene models to identify a total of 4,682 *de novo* annotated genes without a corresponding CSv1.1 orthologue and used expression counts from these *de novo* gene models to build a *de novo* co-expression network of 13 modules (3,975 genes, **Supp. Table 8**). Each of these 13 modules could be closely correlated with at least one consensus module from the consensus network (**Supp. Table 8**), indicating that our *de novo* modules were not exhibiting patterns of expression distinct from those previously identified in the seven-cultivar consensus network. One of these *de novo* derived modules was significantly enriched (p<0.000003) for Jagger *de novo* gene models with increased expression in flag leaf, spike and root tissues (**Extended Data Figure 7**). The three most significant GO terms enriched within this module of 50 genes indicated a role in transcriptional regulation (GO:0065007, GO:0031323, GO:0050789). 15 of these 50 genes were also annotated as transcription factors with FHY3/FAR1 DNA binding domains (**Supp. Table 8**). These domains are known to be involved in phytochrome signalling in *Arabidopsis*[20] and in wheat are hypothesised to contribute to the regulation of *Ppd-B1a* and *PhyC* known to control photoperiodic sensitivity to flowering[21].

Our work demonstrates the strength of a consensus network approach in identifying potentially biologically conserved pathways between cultivars where *de novo* annotations are not available

for each member of the network. Using this method we were able to reveal cultivar-specific co-expressed genes for several cultivars including Claire and Weebill, for which we do not currently have *de novo* gene models. In addition, further extending our co-expression analysis to include the *de novo* gene models of four of the chromosome level assemblies revealed additional *de novo* co-expressed modules exhibiting cultivar-specificity, such as the *de novo* module enriched for Jagger genes, that would not have been captured in the consensus network.

Developing *de novo* annotations for all 14 of the cultivars within the wheat pan-genome[17] will be invaluable in uncovering the complete regulatory network landscape of the wheat pan-transcriptome. Associating these co-expression profiles with the core, shell and cloud components of the wheat pan-genome will enable us to explore how structural rearrangements and introgressions across the wheat genome perturb these regulatory networks.

A case study: Uncovering variation in the prolamin super-family 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. Along with HMW-glutenins they are also potential triggers for various immune reactions in a subset of the human population. As a case study, we investigated both the qualitative and quantitative differences in the 687 genes from the prolamin superfamily and HMW-glutenins across the newly generated wheat pan-genome and pan-transcriptome data. We observed clear expression differences both for individual developmental stages and also between wheat cultivars for many genes from the prolamin superfamily highlighting spatiotemporal variation in expression profile (**Figure 4A**).

Comparison of reference grain allergens identified in the Chinese Spring reference genome (IWGSC v1.1) and across the pan-genome cultivars[22,23,24] with the expression patterns of potentially immune reactive gene products indicated differences in the major allergens and antigens (glutenins and gliadins). SY Mattis and LongReach Lancer showed lower gene expression levels in alpha and gamma gliadins with gene set enrichment analysis of gene families highlighting gamma gliadins are primarily enriched in the downregulated genes **(Extended Data Figure 8, Supp. Table 9&10)**.

Detailed analysis of celiac disease (CD) related epitopes encoded in the gliadin and glutenin genes in the pan-genome revealed variability in their expression patterns. We found lower expression of HLA-DQ epitope containing genes in SY Mattis and LongReach Lancer and higher values in Cadenza and Jagger. Cultivar-specific analysis showed that Arina*LrFor* and SY Mattis contained lower alpha gliadin DQ epitope expressions due to significant differences in the expression activities of the three sub-genomes which might be affected by differences in the related transcription factor gene expression profiles (**Figure 4B, 4C, Supp. Table 11&12, Extended Data Figure 9**). While sub-genome specific expression patterns of gamma gliadin DQ epitopes did not reveal significant variation, the expression of alpha-gliadin genes with DQ epitopes originating from the A genome was lower in SY Mattis and LongReach Lancer, while the highly immunogenic D genome alpha gliadin epitope expression levels were lower in the

cultivar Arina*LrFor* (**Extended Data Figure 10A**). 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**).

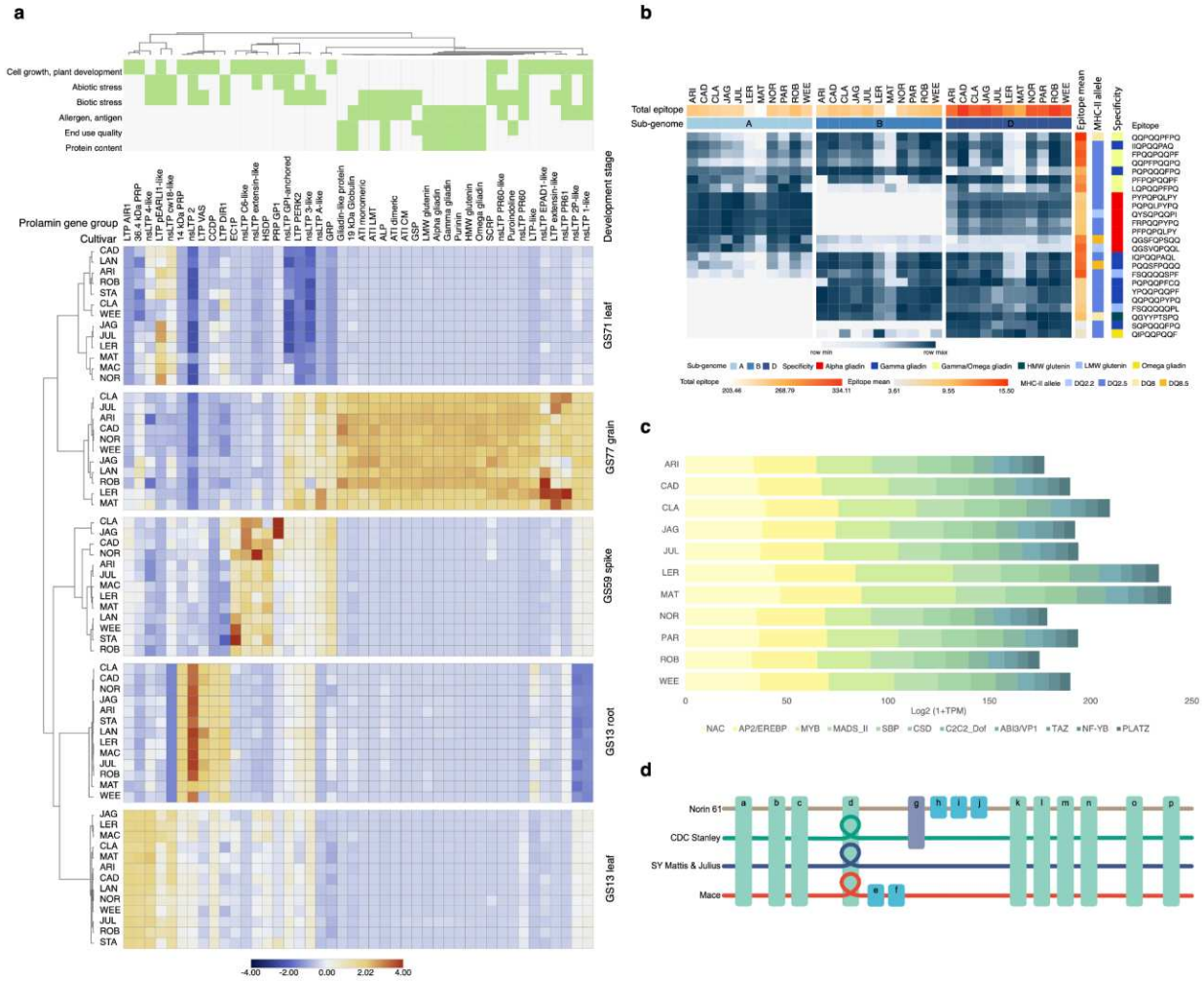

Figure 4. **Gene and expression variation in the prolamin family across the wheat pan-cultivars. A**. Prolamin superfamily gene expression across cultivars; **B**. Celiac disease epitope expression across cultivars. Epitope expression profiles were calculated as sum of gene expression profiles with the highlighted HLA-DQ epitopes for each sub-genome. **C**. Relative proportion of cumulative expression profiles of transcription factor families showing strong co-expression pattern (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 regulators of storage protein gene expression. **D**. Representation of the variation graph for the region of 6D containing the alpha gliadin locus (**Extended Data Figure 10B)**. 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 (blue: in common to all 4 cultivars, orange: 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 (**Extended Data Figure 11**).

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 (**Extended Data Figure 10B**). 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 (**Extended Data Figure 10B, Supp. Table 13**). 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 Arina*LrFor*, Norin 61 and Mace, respectively (**Extended Data Figure 10B, Supp. Table 13**).

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, Extended Data Figure 11**). 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, and 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 (**Extended Data Figure 10B**). 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. Comparison of promoter profiles indicates differences in the expression regulation when epitope-poor and epitope-rich gene copy variants of the same chromosome 6D locus are compared. 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[25].

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[26], often associated with adaptation to biotic or abiotic stress[13]. 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 an 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 stress responses in the developing grain and the photoperiodic control of flowering. 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.

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

Plant Materials and Growth Conditions

The 14 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 (Norin 61, Julius, Jagger, Arina*LrFor*, Robigus, Claire and SY Mattis), seedlings were transferred in 40-well trays (7 days after sowing) to a vernalisation CER running at 6 °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 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 of 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.

Alignment and Gene Expression Analysis

Samples were aligned to the IWGSC Chinese Spring RefSeq 1.1 reference genome, using HISAT2 v2.0.4, acting as a common reference to allow inclusion of UK cultivars and comparison with *de novo* annotations, and normalised counts were generated using DESeq2 with the RefSeq 1.1 gene set[27,16,28].

GO term analysis

Functional enrichment of differentially expressed genes for biological processes was performed using the gene ontology enrichment analysis package, topGO[29] 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[30].

Tissue-Specific Index

The specificity of gene expression to developmental stages was determined using the tissue-specific index [31]. 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 conditions 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[32].

Sub-genome 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).

Co-expression Analysis

Network construction

The WGCNA R package[17] (R version 3.6.0) was used to build co-expression networks for seven cultivars (Arina*LrFor*, Cadenza, Claire, Jagger, Julius, Norin 61 and Weebill) 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 matrix for the seven selected cultivars containing DESeq2 normalised counts aligned to 102,443 high confidence CSv1.1 genes was filtered and genes, where the sum of counts across all samples was greater than 5 in at least 3 samples, were retained (92,976 genes). To reduce background noise we removed genes where the expression of any replicate was >2σ from the mean expression of that sample set. The resulting 48,337 genes were submitted to WGCNA to construct a signed hybrid consensus network using the blockwiseConsensusModules () function. A soft power threshold of 18 was used, together with the following parameters; minModuleSize = 30, corType = bicor, maxPOutliers = 0.05, mergeCutHeight = 0.3, minKMEtoStay = 0.2, maxBlockSize = 46,000. Eigengenes were then extracted for each module, per cultivar from the resulting consensus network.

Defining thresholds 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 consensus module comparison, using the Pearson correlation distance. We used the maximum distance of each of these pairwise comparisons, for each module and calculated the median of these maxima. Next, we investigated the proportion of 1,663 triads (from 30-lets) identified as split across the 23 previously defined consensus 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 75% 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[33] to determine the optimal number of clusters for the 7 cultivar 23 consensus module eigengenes. The resulting Dunn index[34] and silhouette width[35] indicated that the optimum number of clusters for our ME dataset was 6. We calculated the pairwise Pearson correlation coefficients for all our 161 cultivar consensus ME (cor()) and converted this to a dissimilarity matrix (as.dist()). We used hierarchical clustering of this dissimilarity matrix (hclust()) to define consensus metamodules. As the JAG magenta consensus module fell into a different metamodule to the remaining 6 cultivars we omitted the magenta consensus module from our metamodule construction and downstream enrichment analysis. We used the moduleEigengenes() function to compute the ME of each metamodule and carried out GOterm and TF superfamily enrichment analysis.

Transcription factor superfamily enrichment

Genes annotated as members of transcription factor (TF) superfamilies[5] were identified in each metamodule and the frequency of each TF superfamily compared to the frequency observed in the 32,936 genes used to construct metamodules. TF families were classed as either significantly under or overrepresented in each module using Fisher's exact test ($p <= 0.05$).

Identifying cultivar-specific expression patterns

We used the consensuskME() function in WGCNA to determine the maximum eigengene-based connectivity (kME) of each gene within the unclustered gene set of 14,268 genes, to the consensus ME for each cultivar. Those genes with a positive association greater than 0.7 to a consensus ME were retained. To reveal putative networks specific to a single cultivar we identified genes from the 13,708 genes that demonstrated associations greater than 0.7 kME for at least one cultivar, per gene, where a minimum of 4 out of the 7 cultivars exhibited >0.7 kME and that in pairwise comparison to all other cultivars, for the specific cultivar being assessed, the gene was assigned to a divergent module.

Connectivity within each set of genes demonstrating single cultivar-specificity was determined using the R package igraph[18]. Using the graph adjacency() function, graph adjacencies were created for each specific cultivar set based on the Pearson correlation distances between genes in a pairwise fashion. These directed graphs were simplified to remove multiple edges and loops, and filtered to retain only those connections with an absolute Pearson correlation > 0.8. The mst() function using the prim algorithm was used to create a minimum spanning tree and the resulting subgraphs were visualised using plot() with isolated nodes excluded.

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[36] (v2.7.8a) and further assembled into transcripts by StringTie[37] (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[38] (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[39] (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 [40] (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[41] (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[42]). Augustus was run using the wheat model.

A consolidated set of gene models was selected using Mikado[43], 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[44] (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[45] (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

724 genes. The evidence-based part of the annotation pipeline is deposited at
725 https://github.com/PGSB-HMGU/plant.annot.
726
727 **Consolidation**
728 Pairwise whole genome alignments were generated using lastz[46]. The resulting alignments were
729 stitched together into a single whole genome alignment using TBA/multiz[47]. The MAF output
730 was converted into HAL format using maf2hal[48].
731 *De novo* gene annotation from one cultivar was lifted over to all other cultivars using the whole
732 genome alignment and the halLiftover tool, whereas only full-length gene models were kept.
733 Missing gene models in one cultivar were identified using bedtools[49].
734
735 **Tandem array detection**
736 Tandem arrays were identified using the tandem discovery model from the JCVI package[50].
737 Expression bias was calculated using a modified method described previously[5]. Here we used
738 normalised read counts instead of TPM values and a cut-off of 0.8. The following categories
739 were assigned: only1 for tandems with only one gene expressed and no expression data for
740 second gene; expressed1 for tandems in which only one gene is expressed under all RNASeq
741 conditions; variable where expression can shift between array members depending on the
742 condition; balanced, where both array members are equally expressed. noExpr states that no
743 expression data was available.
744
745 **Orthogroup analysis**
746 The longest isoforms from high-confidence genes were used as input for Orthofinder[51].
747 Orthofinder was run using standard parameters. We used the UpSetR in the R package
748 (http://gehlenborglab.org/research/projects/upsetr/) to analyse and visualise how many
749 orthogroups are shared between the cultivars or are unique to a single species. GENESPACE[9]
750 was used to derive and visualise syntenic relationships between all chromosomes and sub-
751 genomes.
752
753 **Analysis of Canadian-specific genes**
754 Taking each genome in turn as a reference, kmers of length 51 were identified from genic
755 regions using the annotation for that reference. These kmers were used to search the genomes
756 of the other cultivars and a coverage score was computed[52] between each gene in the
757 reference and every other genome. The coverage score (a value between 0 and 1) can be used
758 as a proxy for sequence similarity/difference between genes in different cultivars where values
759 closer to 0 indicate greater difference and values closer to 1 indicate similarity. Coverage scores
760 for genes along chromosomes were plotted using the seaborn visualisation library[53] in Jupyter
761 notebook. Coverage scores were also visualised as heatmaps with coverage scores close to 0
762 represented as dark bands.
763
764 **Comparative analysis of immune reactive regions in the wheat pan-genome**
765 *Reference allergen identification and chromosome 6D comparison*
766 Reference allergens in the wheat pan-genome were filtered using blastn algorithm against the
767 identified sequences in the IWGSC v1 gene annotation v1.1[22]. To identify unannotated gliadin

and glutenin gene models and to compare the potential immune reactivity of the wheat cultivars, known coeliac diseases 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 for the detailed comparison of 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[22].

Promoter motif enrichment analysis

1000 bp 5'-end non-coding sequences were extracted from the chromosome 6D loci and used for motif enrichment analysis in MEME-SEA[54]. The JASPAR core plant 2022 motif collection was used as a background database.

Epitope expression analysis

Epitope expression values were calculated using the TPM gene expression values of genes where the reverse translated consensus epitope sequence was detected multiplied by the number of epitopes in each sequence. The obtained values were summed for each epitope type as well as summed for epitope types at genome levels.

Gene co-expression analysis

TPM>1 log2 transformed TPM gene expression data were used to create a grain co-expression network using 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.

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[55], 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 pan-graph construction using PGGB[56] 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 pan-graph

Using ODGI toolkit[57] 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 **Supp. Table 14** extracts 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 sub pan-graph using odgi view (**Supp. Table 15).**

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 BBS/E/T/000PR9816 (NC1 - Supporting EI's ISPs and the UK Community with Genomics and Single Cell Analysis) for data generation and BB/CCG1720/1 for the physical HPC infrastructure and data centre delivered via the NBI Research Computing group. 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 and URPP Evolution in Action. 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

856 investigated introgressions and copy number variation. RRP performed co-expression analysis,
857 constructed regulatory networks and gliadin pan-graph, and investigated cultivar-specific
858 expression. JW performed genome kmer comparison. AJ performed an analysis of allergen and
859 prolamin diversity, MG supervised the prolamin analysis. KS investigated Norin-specific gene
860 expression patterns. GN imported and managed data for EnsemblPlants. BW, TL, RRP, JW, CU,
861 MS & AH contributed to preparing the manuscript. PB, WH, BC, CU, NS, SK, JP, KM & CP
862 provided intellectual input. Syngenta provided funding for data generation. The 10+ Wheat
863 Genome Project, EL, AB, AH and MS for conceptualisation of the project.
864
865 **Ethics declarations**
866 The authors declare no competing interests.
867
868 **References**

869 1. Braun, H.-J., Rajaram, S. & Ginkel, M. van. CIMMYT's approach to breeding for wide
870 adaptation. *Euphytica* **92**, 175–183 (1996).
871 2. Erenstein, O. *et al.* Wheat Improvement, Food Security in a Changing Climate. 47–66 (2022)
872 doi:10.1007/978-3-030-90673-3_4.
873 3. Levy, A. A. & Feldman, M. Evolution and origin of bread wheat. *Plant Cell* **34**, 2549–2567
874 (2022).
875 4. Walkowiak, S. *et al.* Multiple wheat genomes reveal global variation in modern breeding.
876 *Nature* **588**, 277–283 (2020).
877 5. Ramirez-Gonzalez, R. H. *et al.* The transcriptional landscape of polyploid wheat. *Science (New*
878 *York, NY)* **361**, eaar6089 (2018).
879 6. Jin, M. *et al.* Maize pan-transcriptome provides novel insights into genome complexity and
880 quantitative trait variation. *Scientific Reports* **6**, 18936 (2016).
881 7. Thind, A. K. *et al.* Chromosome-scale comparative sequence analysis unravels molecular
882 mechanisms of genome dynamics between two wheat cultivars. *Genome Biol.* **19**, 104 (2018).
883 8. Khan, A. W. *et al.* Super-Pangenome by Integrating the Wild Side of a Species for Accelerated
884 Crop Improvement. *Trends Plant Sci.* **25**, 148–158 (2020).
885 9. Lovell, J. T. *et al.* GENESPACE tracks regions of interest and gene copy number variation
886 across multiple genomes. *eLife* **11**, e78526 (2022).
887 10. He, F. *et al.* Genomic variants affecting homoeologous gene expression dosage contribute to
888 agronomic trait variation in allopolyploid wheat. *Nat. Commun.* **13**, 826 (2022).
889 11. Leister, D. Tandem and segmental gene duplication and recombination in the evolution of
890 plant disease resistance genes. *Trends Genet.* **20**, 116–122 (2004).
891 12. Hanada, K., Zou, C., Lehti-Shiu, M. D., Shinozaki, K. & Shiu, S.-H. Importance of Lineage-
892 Specific Expansion of Plant Tandem Duplicates in the Adaptive Response to Environmental
893 Stimuli . *Plant Physiol.* **148**, 993–1003 (2008).
894 13. Bayer, P. E., Golicz, A. A., Scheben, A., Batley, J. & Edwards, D. Plant pan-genomes are the
895 new reference. *Nat. Plants* **6**, 914–920 (2020).

14. 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).
15. Gordon, S. P. *et al.* Extensive gene content variation in the Brachypodium distachyon pan-genome correlates with population structure. *Nat. Commun.* **8**, 2184 (2017).
16. 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).
17. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform.* **9**, 559 (2008).
18. Cśardi, G. & Nepusz, T. The igraph software package for complex network research. in.
19. Hu, R. *et al.* Genome-wide identification and analysis of WD40 proteins in wheat (Triticum aestivum L.). *BMC Genom.* **19**, 803 (2018).
20. Li, J., Li, G., Wang, H. & Deng, X. W. Phytochrome Signaling Mechanisms. *Arab. Book* **2011**, e0148 (2011).
21. Kiseleva, A. A., Potokina, E. K. & Salina, E. A. Features of Ppd-B1 expression regulation and their impact on the flowering time of wheat near-isogenic lines. *BMC Plant Biol.* **17**, 172 (2017).
22. Juhász, A. *et al.* Genome mapping of seed-borne allergens and immunoresponsive proteins in wheat. *Sci. Adv.* **4**, eaar8602 (2018).
23. 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).
24. 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).
25. Liao, W.-W. *et al.* A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
26. Cheng, H. *et al.* Frequent intra- and inter-species introgression shapes the landscape of genetic variation in bread wheat. *Genome Biol.* **20**, 136 (2019).
27. 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).
28. 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).
29. Rahnenfuhrer, A. A. topGO: Enrichment Analysis for Gene Ontology. R package version 2.46.0. (2021).
30. 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).
31. Yanai, I. *et al.* Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650–659 (2005).
32. Kryuchkova-Mostacci, N. & Robinson-Rechavi, M. A benchmark of gene expression tissue-specificity metrics. *Brief. Bioinform.* **18**, 205–214 (2017).

934 33. Brock, G., Pihur, V., Datta, S. & Datta, S. clValid: An R Package for Cluster Validation. *Journal*
935 *of Statistical Software* **25**, 1–22 (2008).
936 34. Dunn†, J. C. Well-Separated Clusters and Optimal Fuzzy Partitions. *J. Cybern.* **4**, 95–104
937 (1974).
938 35. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster
939 analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
940 36. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
941 37. Kovaka, S. *et al.* Transcriptome assembly from long-read RNA-seq alignments with
942 StringTie2. *Genome Biol.* **20**, 278 (2019).
943 38. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA
944 and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
945 39. Gremme, G., Brendel, V., Sparks, M. E. & Kurtz, S. Engineering a software tool for gene
946 structure prediction in higher organisms. *Inf. Softw. Technol.* **47**, 965–978 (2005).
947 40. Ghosh, S. & Chan, C.-K. K. Plant Bioinformatics, Methods and Protocols. *Methods Mol. Biol.*
948 **1374**, 339–361 (2015).
949 41. Hoff, K. J. & Stanke, M. Predicting Genes in Single Genomes with AUGUSTUS. *Curr. Protoc.*
950 *Bioinform.* **65**, e57 (2019).
951 42. Nachtweide, S. & Stanke, M. Gene Prediction, Methods and Protocols. *Methods Mol. Biol.*
952 **1962**, 139–160 (2019).
953 43. Venturini, L., Caim, S., Kaithakottil, G. G., Mapleson, D. L. & Swarbreck, D. Leveraging
954 multiple transcriptome assembly methods for improved gene structure annotation. *GigaScience*
955 **7**, giy093- (2018).
956 44. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search
957 tool. *Journal of molecular biology* **215**, 403–410 (1990).
958 45. Seppey, M., Manni, M. & Zdobnov, E. M. Gene Prediction, Methods and Protocols. *Methods*
959 *Mol. Biol.* **1962**, 227–245 (2019).
960 46. Kiełbasa, S. M., Wan, R., Sato, K., Horton, P. & Frith, M. C. Adaptive seeds tame genomic
961 sequence comparison. *Genome Res.* **21**, 487–493 (2011).
962 47. Blanchette, M. *et al.* Aligning Multiple Genomic Sequences With the Threaded Blockset
963 Aligner. *Genome Res.* **14**, 708–715 (2004).
964 48. Hickey, G., Paten, B., Earl, D., Zerbino, D. & Haussler, D. HAL: a hierarchical format for
965 storing and analyzing multiple genome alignments. *Bioinformatics* **29**, 1341–1342 (2013).
966 49. Quinlan, A. R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr. Protoc.*
967 *Bioinform.* **47**, 11.12.1-11.12.34 (2014).
968 50. Tang, H. *et al.* Synteny and Collinearity in Plant Genomes. *Science* **320**, 486–488 (2008).
969 51. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome
970 comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157
971 (2015).

52. 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.

53. Waskom, M. seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).

54. 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.

55. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).

56. Garrison, E. *et al.* Building pangenome graphs. *Biorxiv Prepr Serv Biology* (2023) doi:10.1101/2023.04.05.535718.

57. Guarracino, A., Heumos, S., Nahnsen, S., Prins, P. & Garrison, E. ODGI: understanding pangenome graphs. *Bioinformatics* **38**, btac308- (2022).