DATABASE

WheatCENet: A Database for Comparative Co-expression Network Analysis of Allohexaploid Wheat and Its Progenitors

Zhongqiu Li¹, Yiheng Hu²,³, Xuelian Ma¹, Lingling Da¹, Jiajie She¹, Yue Liu¹, Xin Yi², Yaxin Cao¹, Wenyong Xu¹, Yuannian Jiao²,³,* Zhen Su¹,*

¹State Key Laboratory of Plant Physiology and Biochemistry, College of Biological Sciences, China Agricultural University, Beijing 100193, China
²State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China
³University of Chinese Academy of Sciences, Beijing 100049, China

Received 12 August 2021; revised 16 March 2022; accepted 8 May 2022
Available online 1 June 2022
Handled by Long Mao

KEYWORDS
Co-expression network; Species comparison; Diploid and polyploid wheat; Functional annotation

Abstract Genetic and epigenetic changes after polyploidization events could result in variable gene expression and modified regulatory networks. Here, using large-scale transcriptome data, we constructed co-expression networks for diploid, tetraploid, and hexaploid wheat species, and built a platform for comparing co-expression networks of allohexaploid wheat and its progenitors, named WheatCENet. WheatCENet is a platform for searching and comparing specific functional co-expression networks, as well as identifying the related functions of the genes clustered therein. Functional annotations like pathways, gene families, protein–protein interactions, microRNAs (miRNAs), and several lines of epigenome data are integrated into this platform, and Gene Ontology (GO) annotation, gene set enrichment analysis (GSEA), motif identification, and other useful tools are also included. Using WheatCENet, we found that the network of WHEAT ABERRANT PANICLE ORGANIZATION 1 (WAPO1) has more co-expressed genes related to spike development in hexaploid wheat than its progenitors. We also found a novel motif of CCWWWWWGG (CArG) specifically in the promoter region of WAPO-A1, suggesting that neofunctionalization of the WAPO-A1 gene affects spikelet development in hexaploid wheat. WheatCENet is useful for investigating co-expression networks and conducting other analyses, and thus facilitates

* Corresponding authors.
E-mail: zhensu@cau.edu.cn (Su Z), jiaoyn@ibcas.ac.cn (Jiao Y).

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2022.04.007
1672-0229 © 2023 The Authors. Published by Elsevier B.V. and Science Press on behalf of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
Introduction

Wheat, in the Poaceae family, is the most widely grown food crop worldwide, providing an important source of nutrients for millions of people. Global food demand is increasing rapidly, with a 60%–70% increase in food production required by 2050 [1]. Basic research and optimization of wheat breeding are necessary to meet this demand. However, the bread wheat genome is large and complex (16 Gb) [2], and thus wheat research lags behind that of rice and maize. This lag is mainly because bread wheat is a recent allohexaploid, formed via two consecutive allopolyploidization events. The diploid wheat *Triticum urartu* (AA) and a yet unknown *Aegilops* species formed tetraploid wheat *Triticum dicoccoides* (AABB) around 0.5 million years ago (MYA). Subsequently, tetraploid wheat and *Aegilops tauschii* (DD) hybridized to form the hexaploid wheat *Triticum aestivium* (AABBDD) around 0.01 MYA [3,4]. As wheat is a primary food crop, wheat scientists have worked to sequence the genome of hexaploid wheat and its progenitors, and the completion of genome sequencing has laid a solid foundation for studying the functional, comparative, and evolutionary genomics of wheat [5,6].

Currently, public databases for wheat can be divided into four types: genome, transcriptome, proteome, and others. Genome databases, like Wheat@URGI portal [7], GrainGenes [8], CerealsDB [9], Wheat-SnpHub-Portal [10], WheatGmap [11], and Triticeae-Genetribe [12], provide genomic or genetic data and other useful tools. Transcriptome databases, including expVIP [13] and Wheat eFP Browser (https://bar.utoronto.ca/efp_wheat/cgi-bin/efpWeb.cgi), are usually used to provide the expression patterns of homeologs. Transcriptome-based co-expression networks like WheatNet have been constructed by DNA microarray datasets with an early genome assembly version [14]. Ketnetminer [15] and WheatOmics [16] only contain a hexaploid wheat network, not including wheat progenitors. Wheat Proteome provides searchable organ and developmental stage proteomic data [17]. Other databases like Triticeae Toolbox (T3) have phenotype and genotype data for barley, wheat, and oat [18]. The wheat microRNA Portal has integrated the abiotic stress response microRNAs (miRNAs) in wheat [19]. Despite the accumulation of large-scale RNA-seq data and the improved and high-quality wheat genomic sequence, there are still gaps in meeting the demands for co-expression analysis within allohexaploid wheat and its progenitors.

Thus, we developed the co-expression network comparison database WheatCENet for allohexaploid wheat and its progenitors, including four global networks (*T. aestivum, T. dicoccoides, T. urartu, and Ae. tauschii*) and two *T. aestivum* conditional networks (tissue-specific and stress-treated). WheatCENet integrates functional and epigenome sequencing data, and includes useful tools like gene set enrichment analysis (GSEA), Gene Ontology (GO) analysis, and motif analysis, which will help bench scientists easily pick key candidate genes for functional studies. In addition, this database will provide genomic scientists with a useful source for deciphering key molecular modules during the formation, evolution, and domestication of wheat.

Construction and content

Data sources and processing

For genome data, *T. aestivum* (AABBDD) was based on the International Wheat Genome Sequencing Consortium (IWGSC) Chinese Spring v1.0 genome assembly and v1.1 annotation (URGI) [2]; *T. dicoccoides* (AABB) was based on the Zavitan WEW_v1.0 genome assembly and annotation [20]: *T. urartu* (AA) was based on the Institute of Genetics and Developmental Biology, Chinese Academy of Sciences (IGDB) genome assembly and annotation [21]; and *Ae. tauschii* (DD) was based on the Chinese Academy of Agricultural Sciences (CAAS) genome assembly and annotation [22].

For transcriptome data, we used most of the available RNA-seq datasets in addition to the common tissues of leaf, root, and grain in the four studied species for robust constructed co-expression networks [23,24]. Finally, we collected 425 transcriptomic datasets (112 for *T. aestivum*, 153 for *T. dicoccoides*, 90 for *T. urartu*, and 70 for *Ae. tauschii*) from the NCBI Sequence Read Archive (SRA) (for detailed sample information, see Table S1). Quality control was conducted based on FastQC software (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), and Trimomatic software [25] was used to remove adapter sequences and low-sequencing-quality bases. The remaining sequence data (112 for *T. aestivum*, 153 for *T. dicoccoides*, 76 for *T. urartu*, and 42 for *Ae. tauschii*) were mapped to corresponding genomes, and then the fragments per kilobase of transcript per million mapped reads (FPKM) values of all protein-coding genes were calculated from each sample with parameter settings --star and --estimate-rspd using RSEM [26]. Those samples with a mapping rate < 50% were filtered out. Then, the R package “pheatmap” (https://github.com/raivokolde/pheatmap) was used to perform cluster analysis on all datasets, and the outlier samples were excluded (Figures S1 and S2). According to the boxplot of the reading score distribution, the quality of the remaining samples was acceptable (Figure S3). The mapping results of these data are listed in Table S2.

For epigenome data, all 25 *T. aestivum* epigenome datasets, including H3K4me3, H3K4me1, H3K36me3, H3K27me3, H3K9me2, H3K9ac, H3K27ac, DNase-seq, and CENH3, were downloaded from public platforms, including SRA. Quality control was conducted based on trim_galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) software with parameter settings q = 25 and stringency = 3. The sequence reads were mapped to the *T. aestivum* accession Chinese Spring v1.0 reference genome with the maximal exact matches (MEM) algorithm and default parameters by BWA software [27]. The enriched regions were called by MACS software [28] with the nomodel parameter. The details and mapping results of these data are listed in Table S3.

For collinear ortholog pairs, we first computed the top 5 Basic Local Alignment Search Tool (BLAST) results between two species using protein sequences based on the rank of bit score that met an E-value threshold of $1 \times 10^{-5}$, which was suggested by MCScanX tools [29]. Then the top 5 BLAST results with general feature format (gff) files were used to establish collinear ortholog pairs in *T. urartu*, *Ae. tauschii*, *T. dicoccoides*, and *T. aestivum* through MCScanX tools with default parameters.

**Co-expression network construction and network comparison**

We used the calculated FPKM values for constructing the four global networks and two *T. aestivum* conditional (tissue-specific and stress-treated) networks based on the Pearson correlation coefficient (PCC) and mutual rank (MR) algorithm [30]; every network covered at least 84.70% of genes (Table 1). We used the PCC and MR to measure the co-expression relationship between genes. First, based on the PCC distribution diagram of all gene pairs, the thresholds for negative and positive correlation were the values of the lowest 5% PCC values (~0.25) and highest 5% PCC values (0.5) in the *T. aestivum* global network, respectively (Figure S4). Second, we used the MR method to exclude poor co-expression gene pairs, because MR has been used successfully for this purpose in several plants such as *Arabidopsis*, maize, and bamboo [31]. Furthermore, the GO terms of biological processes related to multiple genes in the interval [4,20] were used to evaluate the accuracy of the co-expression network through the receiver operating characteristic (ROC) curve [32]. Finally, to compare the *T. aestivum* global networks and conditional networks or other species' networks while considering the coverage and connectivity of each network, we set the MR of all networks to 30.

Further, in order to verify whether the networks we built are comparable among allohexaploid wheat and its progenitors, we collected genes from literature, such as the light-dependent chlorophyll accumulation gene *TaCHLH* [33], the peroxidase gene *TaPOD1* [34], the storage protein activator gene *TaSPA*, and the genes involved in the photosynthesis pathway (*LHCA2*, *LHCA3*, *LHCb7*, and *psaK*) (Table S4). We queried the expression values of these genes in the leaf, root, and grain of the four species. The results suggest that the expression trends of these genes in samples with different developmental stages or stress treatments are still consistent and comparative among four species for analyzing the network (Figure S5).

**Functional module identification**

We used the CFinder software to identify modules containing more densely connected genes by combining positive and negative gene pairs together [35] (Table 1). Then we used the gene set annotations, like gene families, GO terms, and metabolic pathways (detailed information is in the Functional annotation and gene annotation section), to predict the functions of modules. Non-significant entries were then filtered out using Fisher’s tests and the multiple test correction method “Benjamini-Yekutieli” [false discovery rate (FDR), as referred to in the PlantGSAD [36]. As a result, 1867 functional modules in AABBDD, 625 functional modules in AABB, 524 functional modules in AA, and 851 functional modules in DD were...
obtained. These modules may be related to important agronomic traits.

Usage of co-expression network tools

In WheatCENet, a network search tool for one gene or a gene list was provided for the four global and two conditional (tissue-specific and stress-treated) co-expression networks, which were visualized using Cytoscape [37] (Figure 1A). Furthermore, we established collinear ortholog pairs in AA, DD, AABB, and AABBDD using MCScanX [29] tools. We used collinear gene pairs to determine the correlations between diploid and polyploid wheat. With the relationship of homologous genes, three types of co-expression networks between...
polyploid wheat and its progenitors can be compared: genes of a single species can be compared between the global network and tissue-specific network or stress-treatment network in the network comparison tool (Figure 1B); genes of two species, such as AA vs. AABB, DD vs. AABBDD, and AABB vs. AABBDD, can be compared in the network comparison tool; and genes of three species (DD, AABB, and AABBDD) can be compared in the ortholog network comparison tool. For instance, users can submit one gene of interest in our ortholog network comparison tool, and choose two or three species for the orthologous pair of this gene. Then all the orthologous gene pairs are highlighted and linked to each other with brown lines in the network to exhibit the conservation and diversification of the regulatory network during wheat evolution (Figure 1C). For all genes in the network, the annotations and relationships of genes in a network are listed in the tables. GSEA, GO analysis, and motif analysis tools are used to find the potential functions and regulations in the promoter regions of genes in the network (Figure 1D), as well as the expression profile and the distribution of genes on chromosomes of the network (Figure 1E). Moreover, the gene expression profiles and cis-elements of the homeologous sub-networks can be compared to find the similarities and differences of the network.

Functional annotation and gene annotation

For functional annotation (Figure 2A; Table 2), there are five types of data in WheatCENet, which can be searched by users to predict gene function. On the one hand, integrating these data is of great significance for genome-level gene annotation; on the other hand, these functional annotations can be used as background gene sets for data mining, such as enrichment analysis for network or functional modules.

Ontology has become a critical tool for enabling cross-species comparative analyses and increasing data sharing and reusability [38]. Thus, we collected 7391 GO, 444 Plant Ontology (PO), and 276 Trait Ontology (TO) annotation entries of AABBDD from URGI (https://wheat-urgi.versailles.inra.fr/Data), 1814 GO annotation entries of AA from MBKBASE (https://www.mbkbase.org/Tu/), and 7612 and 9716 GO annotation entries of DD and AABB separately annotated by Blast2GO software [39].

Metabolic pathways are responsible for the biosynthesis of complex metabolites, having an impact on the growth and development of plants or aiding plants in responding to biotic and abiotic stresses [40]. We collected Plant Reactome pathways of all species from Gramene [41]; PlantCyc pathways of AA, DD, and AABBDD were integrated from PlantCyc [42], while AABB was predicted by orthologs. For the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (https://www.kegg.jp/kegg/), DD came from KEGG using the ID conversion; AABBDD and AABB KEGG pathways were annotated by the KEGG tool KAAS; and AA came from MBKBASE.

Protein domains are important parts of proteins, and many domains either have specific functions or contribute to the function of their proteins in a specific way [43]. Protein domains were annotated by PfamScan tools (https://www.ebi.ac.uk/Tools/pfa/pfamscan/) based on a hidden Markov model [44]. We ultimately identified 4021 functional domains with 83,690 genes in AABBDD; 4052 functional domains with 49,301 genes in AABB; 3675 functional domains with 26,636 genes in AA; and 3930 functional domains with 26,275 genes in DD. Users can submit a gene list to WheatCENet, and then the protein domains and detailed information can be extracted to proceed to the downstream analysis (like adding protein domains when constructing evolutionary trees).

Gene families, such as those of transcription factors (TFs)/transcriptional regulators (TRs), protein kinases (PKs), carbohydrate-active enzymes, and ubiquitin, play important biological roles. We used iTAK software (http://bioinfo.bti.cornell.edu/cgi-bin/itak/index.cgi) to identify TF and PK families. We obtained a hidden Markov model from UUCD (https://uucd.biocuckoo.org/) to identify ubiquitin families. InParanoid (https://inparanoid.sbc.su.se/cgi-bin/index.cgi) software and Pfam domains were used to predict the carbohydrate-active enzyme families. So we annotated 20 ubiquitin families with 6641 genes, 90 TF/TR families with 7913 genes, 125 PK families with 5217 genes, and 94 carbohydrate-active enzyme families with 4290 genes in AABBDD; 20 ubiquitin families with 2667 genes, 91 TF/TR families with 4622 genes, 126 PK families with 3126 genes, and 93 carbohydrate-active enzyme families with 2749 genes in AABB; 20 ubiquitin families with 1503 genes, 90 TF/TR families with 2325 genes, 120 PK families with 1623 genes, and 91 carbohydrate-active enzyme families with 1326 genes in AA; and 19 ubiquitin families with 2068 genes, 91 TF/TR families with 2337 genes, 122 PK families with 1183 genes, and 92 carbohydrate-active enzyme families with 1456 genes in DD. By comparing the protein sequences of hexaploid AABBDD and its progenitors with wheat protein sequences of cytochrome P450 (CYP450) genes (https://drnelson.uthsc.edu/cytochromeP450.html), we further identified their P450 domains using PfamScan (https://www.ebi.ac.uk/Tools/pfa/pfamscan/). Finally, 1564, 823, 462, and 489 full-length CYP450 genes were found in AABBDD, AABB, AA, and DD, respectively (the detailed information and related analysis about CYP450 genes are shown in File S1).

miRNAs function in plant developmental plasticity, abiotic/biotic responses, and symbiotic/parasitic interactions [45]. By integrating miRNAs data including mature miRNAs and precursor miRNAs, 9626 miRNAs in AABBDD and 173 miRNAs in DD were collected from miRbase [46] and IWGSC, while 1702 miRNAs in AA were collected from the literature [47]. Then, all miRNA sequences were mapped to the corresponding genome by GMAP (http://research.pub. gene.com/gmap/), and the mature targets were predicted by psRNATarget (https://plantgrn.noble.org/psRNATarget/ or https://www.zhaolab.org/psRNATarget/). We also predicted protein–protein interaction (PPI) pairs in all species, with the experimentally validated PPI pairs in Arabidopsis predicted by InParanoid [48]. There were 149,844 PPI pairs with 12,972 genes in AABBDD, 53,169 PPI pairs with 9452 genes in AABB, 16,726 PPI pairs with 4818 genes in AA, and 16,226 PPI pairs with 5105 genes in DD (Table 2).

For gene annotation, gene search results included all known and predicted information. Taking WHEAT ABRANT PANICLE ORGANIZATION 1 (WAPO-A1) as an example, the gene detail page includes basic information: the gene locus
Figure 2  Description of functional annotation and supported tools in the database
A. Five types of data are included in functional annotation. Gene families include ubiquitin, TF/TR, PK, CYP450, and carbohydrate-active enzyme gene families. Ontology (such as GO, PO, and TO), protein domain, miRNA, and metabolic pathways (including PlantCyc, KEGG, and Plant Reactome), can be also browsed. B. Tools, like GSEA, GO analysis, motif analysis, BLAST, ID conversion, extraction of sequence/FPKM, and UCSC genome browser (including epigenome data in T. aestivum (AABBDD) and RNA-seq samples in T. dicoccoides (AABB) (not shown in the picture)), are supported in WheatCENet. miRNA, microRNA; KEGG, Kyoto Encyclopedia of Genes and Genomes; TF, transcription factor; TR, transcription regulator; PK, protein kinase; CYP450, cytochrome P450; BLAST, Basic Local Alignment Search Tool; PO, Plant Ontology; TO, Trait Ontology.
is TraesCS7A02G481600; it is a 1457-bp gene with two exons located on chromosome 7A that encodes an F-box-like protein, related to spikelet number per spike (SNS); the best orthologous gene can be linked to the detailed function in Arabidopsis thaliana and Oryza sativa; and the gene sequence, coding sequence, and protein sequence can also be downloaded. In addition, the co-expression network is shown, including orthologous genes and networks (global network, tissue-specific network, and stress-specific network) linked to the corresponding functional interface in diploid and polyploid wheat. Heuristic information, such as F-box ubiquitin family and GO terms related to flower development and regulation of circadian rhythm, is displayed. Moreover, the protein domain module shows an F-box-like functional domain with alignment start and end information; the expression pattern module shows the gene expression profiling in samples (WAPO-A1 is specifically expressed in spike); and the gene sequence, codon sequence, and protein sequence can also be downloaded. In addition, the co-expression network is shown, including orthologous genes and networks (global network, tissue-specific network, and stress-specific network) linked to the corresponding functional interface in diploid and polyploid wheat. Heuristic information, such as F-box ubiquitin family and GO terms related to flower development and regulation of circadian rhythm, is displayed. Moreover, the protein domain module shows an F-box-like functional domain with alignment start and end information; the expression pattern module shows the gene expression profiling in samples (WAPO-A1 is specifically expressed in spike); and the predicted function module shows the possible functional clues of WAPO-A1 identified by CFinder. For histone modification, the University of California at Santa Cruz (UCSC) [49] genome browser shows obvious peaks for H3K4me3 and H3K27me3, which are related to spikelet or flower development (Figure 3).

### Supported analyses and tools

There are three category analysis tools, including GSEA, GO analysis, and cis-element enrichment. GSEA was based on the data analysis processing of PlantGSAD [36]. Here, we used functional data information as gene sets. The annotation entries with FDR < 0.05 were used and displayed. GO analysis was based on agriGOv2 [50] data processing. Cis-element analysis (such as the functions of sequence scan, gene name scan, and custom scan) can identify significantly enriched motifs in the promoter region of one gene and thus predict possible functions. We also provided BLAST (DNA and protein), ID conversion (in different genome versions), and Sequence (gene ID or chromosome position) and FPKM extraction tools for users to conveniently obtain the information of a gene (Figure 2B).

### Case study: function analysis of the known gene WAPO1 using WheatCENet

Polyploidy (i.e., whole-genome duplication) is an important evolutionary feature in the plant kingdom, particularly in flowering plants [51], after which individual genes may experience nonfunctionalization, neofunctionalization, or subfunctionalization [52]. For example, WAPO1 is an orthologue of rice gene ABERRANT PANICLE ORGANIZATION 1 (APO1) and Arabidopsis gene UNUSUAL FLORAL ORGANS (UFO). UFO acts synergistically with floral meristem identity factor LEAFY (LFY) and restricts the expression of the class B floral organ identity genes in Arabidopsis [53]. The interactions between the orthologs of LFY and UFO have also been demonstrated in rice, petunia, Antirrhinum majus, and pea, suggesting that LFY and UFO are conserved among species [53,54]. EVERGREEN (EVE) encodes a WOX homologous domain protein, which is only expressed in the initial lateral IM and participates in the activation of the UFO homologous gene DOUBLE TOP (DOT) in petunia. The EVE ortholog of Compound Inflorescence (S) and UFO ortholog ANANTHA (AN) have a similar effect on inflorescence meristems in tomatoes and related nightshades [53]. In wheat, WAPO-A1 showed a highly significant association with total spikelet number (TSN) [55–58]. There are three genes [TraesCS7A02G481600 (WAPO-A1), TraesCS7B02G384000 (WAPO-B1), and TraesCS7D02G468700 (WAPO-D1)] in AABBD, two genes [TRIDC7AG067090 (not expressed) and TRIDC7BG06970] in ABB, no gene in AA, and one gene (evm.TU.Contig391.11) in DD. We searched the co-expression network of WAPO1 using WheatCENet (Figure 4), and found that spike development-related gene wheat FLO/LFY (WFL) [59],

### Table 2 Functional annotation collected from public databases or annotated by public software

<table>
<thead>
<tr>
<th>Classification</th>
<th>Category (source/method)</th>
<th>AABBDD module count (gene count)</th>
<th>AABB module count (gene count)</th>
<th>AA module count (gene count)</th>
<th>DD module count (gene count)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO</td>
<td>GO (IWGSC/Blast2GO/MBKBASE)</td>
<td>7391 (85,762)</td>
<td>9716 (49,242)</td>
<td>1814 (21,216)</td>
<td>7612 (24,143)</td>
</tr>
<tr>
<td>TO</td>
<td>TO (IWGSC/orthologous relationship)</td>
<td>276 (1000)</td>
<td>254 (528)</td>
<td>232 (218)</td>
<td>262 (308)</td>
</tr>
<tr>
<td>PO</td>
<td>PO (IWGSC)</td>
<td>444 (63,687)</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Pfam domain</td>
<td>Function domain</td>
<td>4021 (83,690)</td>
<td>4052 (49,301)</td>
<td>3675 (26,636)</td>
<td>3930 (26,275)</td>
</tr>
<tr>
<td>Gene family</td>
<td>Ubiquitin (UUCD)</td>
<td>20 (6641)</td>
<td>20 (2667)</td>
<td>20 (1503)</td>
<td>19 (2068)</td>
</tr>
<tr>
<td>TF/TR (tTAK)</td>
<td>TF/TR (tTAK)</td>
<td>90 (7913)</td>
<td>91 (4622)</td>
<td>90 (2325)</td>
<td>91 (2337)</td>
</tr>
<tr>
<td>PK (tTAK)</td>
<td>PK (tTAK)</td>
<td>125 (5217)</td>
<td>126 (3126)</td>
<td>120 (1623)</td>
<td>122 (1183)</td>
</tr>
<tr>
<td>CYP450 (BLAST + Pfam)</td>
<td>CYP450 (BLAST + Pfam)</td>
<td>264 (1564)</td>
<td>242 (823)</td>
<td>211 (462)</td>
<td>232 (489)</td>
</tr>
<tr>
<td>Carbohydrate-active enzyme (ortholog)</td>
<td>Carbohydrate-active enzyme (ortholog)</td>
<td>94 (4290)</td>
<td>93 (2749)</td>
<td>91 (1326)</td>
<td>92 (1456)</td>
</tr>
<tr>
<td>Metabolic pathway</td>
<td>Plant Reactome (Gramene)</td>
<td>204 (3476)</td>
<td>204 (2063)</td>
<td>192 (654)</td>
<td>202 (976)</td>
</tr>
<tr>
<td>PlantCyc (PlantCyc/ortholog)</td>
<td>PlantCyc (PlantCyc/ortholog)</td>
<td>493 (7100)</td>
<td>494 (4209)</td>
<td>426 (2057)</td>
<td>434 (2228)</td>
</tr>
<tr>
<td>KEGG (KEGG/KAAS/MBKBASE)</td>
<td>KEGG (KEGG/KAAS/MBKBASE)</td>
<td>139 (15,528)</td>
<td>139 (9312)</td>
<td>112 (2246)</td>
<td>140 (4692)</td>
</tr>
<tr>
<td>miRNA</td>
<td>miRNA target (miRbase/IWGSC/literature)</td>
<td>9626</td>
<td>–</td>
<td>1702</td>
<td>173</td>
</tr>
<tr>
<td>PPI</td>
<td>PPI (InParanoid)</td>
<td>149,844 (12,972)</td>
<td>53,169 (9452)</td>
<td>16,726 (4818)</td>
<td>16,226 (5105)</td>
</tr>
</tbody>
</table>

Note: “–” indicates that the data were not available. GO, Gene Ontology; TO, Trait Ontology; PO, Plant Ontology; miRNA, microRNA; PPI, protein–protein interaction; TF, transcription factor; TR, transcription regulator; PK, protein kinase; CYP450, cytochrome P450; KEGG, Kyoto Encyclopedia of Genes and Genomes.
TaSPL14 [60], TaSPL17 [61], and WUSCHEL-related homeo-box TaWOX4/TaWUS [62,63], were directly co-expressed with WAPO1 in AABBDD, but were not found in DD and AABB co-expression networks, even in the top 300 co-expressed genes. WAPO-A1 also seems to be co-expressed with more genes related to spike development than WAPO-B1 and WAPO-D1. WAPO-A1b in the H2 haplotype (present in the Chinese Spring IWGSC v1.0 genome) was associated with a higher SNS than WAPO-A1c&d in the H3 haplotype (present in WAPO-B1 and WAPO-D1) and WAPO-A1a in the H2 haplotype [57]. WAPO-B1 also has alleles conferring higher spikelet number at chromosome 7B [58]. Taken together, in addition to the effect of natural gene variation on gene expression, the co-expressed genes may also regulate spike development together with WAPO1 in wheat.

GO analysis of the WAPO1 co-expression network in AABBDD showed that these genes were related to reproductive shoot system development and post-embryonic development (Figure S6A). GSEA results for the network (Figure S6B) also indicated that the network of WAPO1 corresponded to floral transition, in which some gene sets were significantly enriched with reproductive meristem phase change, miR156 miRNA_target_network (miR156 could regulate TaSPL14 and TaSPL17 [60,64]), and squamosal promoter binding protein (SBP) TF. After exploring the regulatory regions through motif analysis, we found a novel motif of CCAWWWWWWGG (CArG) specifically identified in the promoter region of the WAPO-A1 gene in AABBDD but not in AABB. We also found that the CArG motif appeared in the genes co-expressed with WAPO1 (Figure 4).
Previous studies revealed that the promoter of \textit{WAPO1} ortholog contains predicted binding sites for the TFs of MADS-box and SBP-like genes [65]. The CArG motif is bound by MADS-box TFs that mainly participate in regulating flowering and floral/spikelet development [66–68]. Flower-specific TFs were confirmed to function in removing the H3K27me3 surrounding flower-specific regulatory elements in \textit{Arabidopsis thaliana} [69,70]. In wheat, the enriched CArG-box motifs were found in the spikelet-reduced H3K27me3 peaks [71]. We also found that \textit{WAPO1} is affected by H3K27me3 modification in different developmental stages using data from http://bioinfo.sibs.ac.cn/dynamic_epigenome/. H3K27me3 modification was lower in spikelet I at the booting stage than spikelet II at the flowering and seedling.
stages, but the expression was the opposite. The newly gained CArG motif might have a significant role in the evolution of the WAPO-A1 gene for functioning in the development of spikelets in hexaploid wheat.

**Discussion and conclusion**

With the latest developments in sequencing technologies and assembly methods, many high-quality sequenced genomes of wheat have been produced [72]. The large amounts of data generated require a platform for experimenters to search for genes of interest. Many wheat databases have thus been developed, such as URGI and GrainGenes, which are data repositories of *T. aestivum* and its relatives, and provide tools like BLAST and JBrowse [7,8]. CerealsDB and Wheat-SnpHub-Portal mainly focus on visualizing single nucleotide polymorphism (SNP) data [9,10]; WheatGmap is committed to the analysis of whole-genome sequencing (WGS) and whole-exome sequencing (WES) data of *T. aestivum* [11]; and Triticeae-GeneTribe provides homologous gene relationships among 12 Triticeae species and three out-groups (rice, maize, and *Arabidopsis*) [12]. None of these databases provide a comprehensive search engine for single genes, such as the gene annotation of WAPO-A1 (Figure 3) provided in WheatCENet. Unlike expression profiling databases such as expVIP and Wheat eFP Browser (https://bar.utoronto.ca/efp_wheat/cgi-bin/efpWeb.cgi), which only provide expression data of *T. aestivum* [13] WheatCENet integrates RNA-seq data of *T. aestivum*, *T. dicoccoides*, *T. urartu*, and *Ae. tauschii* from public platforms, which is useful for studying expression patterns of genes like WAPO1. Existing co-expression networks like WheatNet [14], Ketnetminer [15], and Wheatomics [16] only provide DNA microarray datasets with an early genome assembly version or just include the co-expression networks of hexaploid wheat.

The WheatCENet database aims to provide an online service platform for comparative analysis of gene functions from a multidimensional network across diploid and polyploid wheat species. WheatCENet also includes comprehensive functional annotations (e.g., gene family, GO, miRNA, and metabolic pathways) to predict gene functions. Moreover, WheatCENet includes online tools like GSEA, GO, module, and motif analysis to determine the possible functions of gene sets, and BLAST and ID conversion allow for gene ID conversion between different genome versions. By using WheatCENet, wheat researchers can quickly find key information about the desired gene, predict biological processes (es) the gene may participate in, and study the evolutionary history of the gene in wheat of different ploidies.

For genes that have been cloned but whose functional analysis is not very accurate in wheat, WheatCENet can help to further analyze the function. For example, we found that *Arabidopsis UFO* and wheat WAPO1 have the same 15th amino acid, phenylalanine (F) [55], so we also searched the motif of *UFO*, and identified a variant of the CArG motif in the promoter. Using our UCSC genome browser, we also found that H3K27me3 and H3K4me3 modifications have a peak in the gene body of WAPO1. However, H3K27me3 modification has obvious differences in different tissues and developmental stages, while H3K4me3 does not. Spikelet-reduced H3K27me3 peaks carrying the enriched CArG-box motifs have been found in wheat [71]. Taken together, gaining the CArG motif may affect WAPO-A1 functions in regulating the total number of spikelets in AABBDD, though this requires further experimental verification. Gene analysis methods like that used for WAPO1 can also be applied to other cloned genes, to find the evolutionary differences in wheat of different ploidies.

WheatCENet has more possibilities for improvement, and we will continue to update it in the future. For example, RNA-seq samples from different growth stages, various stress treatments, and more tissue types can be integrated to build a more robust co-expression network. Concurrently, the 10+ Wheat Genomes Project has provided 15 assemblies for different wheat lines from global breeding programs. These wheat accessions and tetraploid durum wheat could be used to analyze co-expression networks and modules with the increase of corresponding RNA-seq samples, so as to more closely link the network with variation and evolution. Epigenomic data like DNase-seq, ChIP-seq, ATAC-seq, MNase-seq, MeDIP-seq, and BS-seq, which can be used to find peaks with gene expression, can be integrated to clarify the complex relationship between gene expression and chromatin structure. These new additions to WheatCENet will help to mine gene function and breeding in wheat.

By constructing and comparing networks in diploid and tetraploid wheat progenitors, we can dissect the origin and evolution of co-expression networks to better understand the underlying genetic basis for various agronomically important traits of bread wheat. The new WheatCENet platform could facilitate bench scientists identifying key candidate genes for functional studies, and provide genomic scientists a reliable source to decipher key molecular modules during the formation, evolution, and domestication of wheat.

**Data availability**


**Competing interests**

The authors have declared no competing interests.

**CRediT authorship contribution statement**

Zhongqiu Li: Data curation, Formal analysis, Visualization, Writing – original draft. Yiheng Hu: Data curation, Formal analysis, Writing – review & editing. Xuelian Ma: Visualization, Formal analysis, Writing – review & editing. Lingling Da: Visualization, Formal analysis, Jiajie She: Software, Visualization, Yue Liu: Software, Formal analysis, Xin Yi: Visualization, Yaxin Cao: Software, Wenyong Xu: Conceptualization, Methodology, Writing – review & editing, Yuannian Jiao: Conceptualization, Methodology, Writing – review & editing. Zhen Su: Conceptualization, Methodology, Supervision, Project administration, Writing – review & editing. All authors have read and approved the final manuscript.
Acknowledgments
This work was supported by grants from the National Natural Science Foundation of China (Grant Nos. 31970629 and 31771467 to ZS, and 31870209 to YJ).

Supplementary material
Supplementary data to this study can be found online at https://doi.org/10.1016/j.gpb.2022.04.007.

ORCID

ORCID 0000-0003-1355-7353 (Zhongqiu Li)
ORCID 0000-0003-4404-7241 (Yiheng Hu)
ORCID 0000-0002-0774-8439 (Xuelian Ma)
ORCID 0000-0001-6572-0066 (Lingling Da)
ORCID 0000-0002-3286-6411 (Xuelian Ma)
ORCID 0000-0003-4404-7241 (Yiheng Hu)
ORCID 0000-0003-4286-6411 (Xin Yi)
ORCID 0000-0002-2249-3406 (Yaxin Cao)
ORCID 0000-0002-7925-2244 (Wenying Xu)
ORCID 0000-0002-8987-2782 (Yuannian Jiao)
ORCID 0000-0001-6579-3847 (Zhen Su)

References


Arabidopsis flowers and embryo developmental genes are repressed in seedlings
