BMC Genomics

DATABASE Open Access

TriticeaeExpDB: a centralized transcriptomic resource for Triticeae research

Tingting Li[1,2†], Lihong Xiao[1†], Haitao Zhang[2], Yiting Su[2], Ruimin Li[3], Yihan Li[2*], Licao Cui[2*] and Xiaojun Nie[1*]

Abstract

The Triticeae tribe, encompassing staple crops like bread wheat (*Triticum aestivum*), barley (*Hordeum vulgare*), and rye (*Secale cereale*), underpins global food security due to its agronomic importance and genetic diversity. While RNA-seq has revolutionized transcriptome analysis by enabling insights into environmental adaptation, stress response and domestication signature, its potential is limited by data fragmentation, pipeline heterogeneity, and lack of standardization. To address these challenges, we present Triticeae Expression Database (TriticeaeExpDB: https://www.triticeaeexpdb.cn), a comprehensive platform integrating and uniformly reprocessing > 10,000 RNA-seq samples from 528 BioProjects across 17 Triticeae species. The database provides manually curated Triticeae datasets encompassing developmental series, abiotic/biotic stress responses, mutant lines, and experimental genetic population, establishing a high-quality repository for cross-species comparative genomics. Users can query expression profiles using gene identifiers or sequence homology, with results augmented by integrated functional annotations and cross-species ortholog mapping. Advanced toolkits support biological quality control, differential expression analysis, functional enrichment, co-expression network construction, and tissue-specificity evaluation, with interactive visualizations of hierarchical heatmaps, enrichment bubble plots, and network graphs. Custom workflows allow integration of private data with public repositories for comparative analyses. By bridging data accessibility with analytical scalability, TriticeaeExpDB accelerates gene discovery and molecular breeding in Triticeae, serving as both a centralized knowledgebase and an extensible research platform.

Keywords Triticeae, RNA-seq, Gene expression, Database, Toolkit

Background

The Triticeae tribe within the Poaceae family, comprising approximately 500 diploid and polyploid species distributed across 27 genera, represents one of the most economically important groups within the grass family [1]. This tribe includes staple cereal crops such as bread wheat (*Triticum aestivum*), barley (*Hordeum vulgare*), and rye (*Secale cereale*), which have been fundamental to human nutrition since the Neolithic Agricultural Revolution [2]. These crops are foundational to global food security, providing the primary source of carbohydrates and proteins worldwide. In developing countries, they contribute up to 80% of daily caloric intake, reflecting their heightened reliance on staple crops [3]. Beyond

†Tingting Li and Lihong Xiao contributed equally to this work.

*Correspondence:
Yihan Li
liyihan@jxau.edu.cn
Licao Cui
cuilicao@jxau.edu.cn
Xiaojun Nie
small@nwsuaf.edu.cn
[1]State Key Laboratory of Crop Stress Biology in Arid Areas, College of Agronomy, Northwest A&F University, Yangling, Shaanxi 712100, China
[2]College of Bioscience and Engineering, Jiangxi Agricultural University, Nanchang, Jiangxi 330045, China
[3]National Navel Orange Engineering Research rsity, Ganzhou, Jiangxi 341000, China

their agricultural significance, Triticeae species serve as valuable models for studying genetics, cytogenetics, and evolutionary biology due to their complex genomes, which include both diploid and polyploid members [4]. The genetic diversity within this tribe, particularly that conserved in wild relatives, constitutes an important resource for developing crops with enhanced resistance to climate-associated abiotic stresses, such as drought and salinity [5]. However, despite their importance, yield gains in these crops have stagnated in recent decades, underscoring the urgent need for innovative breeding strategies.

The Triticeae tribe poses significant challenges for genome assembly due to its exceptional genomic features, including expansive genome sizes, transposable element predominance, and recurrent polyploidization events (in the case of wheat) [6]. Early assembly efforts were hindered by these complexities; however, advances in next-generation sequencing technologies and innovative assembly algorithms have now enabled chromosome-scale genome reconstructions for multiple Triticeae species [7, 8]. These breakthroughs have spurred the rapid expansion of genomic resources, unlocking new opportunities for research on crop domestication, natural variation, and precision breeding [9]. Comparative genomic analyses of sequenced Triticeae members reveal distinct architectural features, such as extreme size variation and high repetitive content. For instance, diploid species like *Hordeum vulgare* and *Secale cereale* possess genomes ranging from 4.3 to 7.9 Gb, encoding over 40,000 annotated genes. The tetraploid *Triticum dicoccoides* genome spans 10.5 Gb with 65,012 protein-coding genes, while the hexaploid wheat genome reaches 14.5 Gb, of which ~ 80% consists of repetitive elements [10]. The availability of high-quality reference genomes has been complemented by large-scale transcriptomic datasets, facilitating systematic investigations into developmental regulation and stress-responsive gene networks. Integration of multi-omics data is now deepening our understanding of the genetic mechanisms governing domestication and environmental adaptation in this agronomically indispensable tribe [11].

Transcriptomic profiling has become a cornerstone of functional genomics, enabling hypothesis generation through genome-wide expression analysis [12]. High-throughput RNA sequencing (RNA-seq) has largely replaced microarray technology due to decreasing costs, standardized protocols, and enhanced resolution in transcriptional quantification [13]. This approach facilitates comprehensive gene expression profiling, including the detection of splice variants, structural annotations, and differential expression under diverse biological conditions [14]. In plant research, extensive RNA-seq datasets have been generated through studies on spatiotemporal expression patterns, developmental processes, and stress responses in key crops such as *Oryza sativa*, *Zea mays*, and *Triticum aestivum* [15]. Most sequencing initiatives now routinely submit raw data to international repositories, including the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA), EMBL-EBI's European Nucleotide Archive (ENA), and the China National Center for Bioinformation's Genome Sequence Archive (GSA). These archives collectively host vast volumes of plant RNA-seq data, representing a largely untapped resource for secondary analyses. Despite improved data accessibility, transforming raw RNA-seq data into biological insights remains computationally intensive. Processing pipelines demand specialized bioinformatics expertise and high-performance computing infrastructure, which are often unavailable to research groups lacking dedicated computational support [13]. These persistent technical barriers create an analytical bottleneck, limiting the ability of smaller laboratories to leverage public datasets for hypothesis-driven or comparative studies.

Despite extensive efforts to consolidate Triticeae transcriptomes through various RNA-seq repositories, a significant portion of these datasets remains underutilized. Current databases lack seamless integration with published RNA-seq data, particularly for wild relatives, limiting comprehensive expression profiling. Furthermore, existing resources host processed data from individual studies without standardization, rendering cross-project comparisons unreliable due to inconsistent bioinformatic pipelines and divergent reference genome versions. To maximize the utility of these RNA-seq datasets, a unified processing framework and a searchable, user-friendly database are imperative. To address this need, we developed the Triticeae Expression Database (TriticeaeExpDB, https://www.triticeaeexpdb.cn). This comprehensive platform integrates a total of 11,016 RNA-seq libraries spanning 17 Triticeae species, including 6078 from *Triticum aestivum*, 2962 from *Hordeum vulgare*, and 530 from *Hordeum spontaneum*. These datasets, manually curated from the SRA, ENA, DNA Data Bank of Japan (DDBJ), and China National Center for Bioinformation – National Genomics Data Center (CNCB-NGDC), cover diverse tissues, developmental stages, abiotic/biotic stresses, mutant lines, and experimental genetic populations. TriticeaeExpDB enables one-click retrieval of gene expression profiles, along with integrated functional annotations and cross-species orthologs within the Triticeae. Additionally, the platform provides a comprehensive suite of analytical tools, including sequence similarity searches, biological quality control, differential expression analysis, functional enrichment, co-expression network construction, and tissue-specificity index calculations, to support in-depth transcriptomic exploration.

As the first comprehensive expression atlas for Triticeae, TriticeaeExpDB accelerates fundamental research in Triticeae genome evolution and empowers translational studies by linking gene expression patterns to phenotypic outcomes. Its standardized datasets and integrated analytical tools bridge the gap between sequence variation and functional genomics, positioning the platform as an indispensable resource for crop improvement and evolutionary studies.

Construction and content

Reference genome acquisition and RNA-seq data collection

A total of 17 species from the Triticeae tribe were included in our study. Genomic sequence data was downloaded in FASTA format, while functional and structural annotations were obtained in GFF or GTF format. If multiple versions were available for the same species, we selected the most recent or favored assembly for download. The genomic information of six species, namely *Aegilops bicornis*, *Aegilops longissima*, *Aegilops searsii*, *Aegilops sharonensis*, *Aegilops speltoides*, and *Dasypyrum villosum* was obtained from CNCB-NGDC (https://ngdc.cncb.ac.cn/) [16]. The genomes of eight species, including *Aegilops tauschii*, *Hordeum vulgare*, *Secale cereale*, *Triticum aestivum*, *Triticum dicoccoides*, *Triticum durum*, *Triticum spelta* and *Triticum urartu* were downloaded the Ensembl Plants database (https://plants.ensembl.org/) [17]. *Triticum monococcum* was downloaded from the figshare platform (https://info.figshare.com/). *Hordeum spontaneum* was retrieved from the IPK database (https://barley-pangenome.ipk-gatersleben.de/). *Hordeum vulgare* var. *nudum* was obtained from the WheatOmics database (http://wheatomics.sdau.edu.cn/jbrowse.html) [18].

All RNA-seq samples were aggregated from the following mainstream databases: SRA, ENA, DDBJ databases and CNCB—NGDC (as of September 1, 2025). A total of 528 studies comprised of 11,016 RNA-seq experiments were collected for TriticeaeExpDB construction. These datasets comprehensively capture the expression profiles of Triticeae species, encompassing diverse subspecies/varieties/genotypes, developmental stages/tissues, biotic/abiotic stress conditions, as well as mutant lines and experimental genetic populations (*e.g.*, RIL, NIL, DH, and et al.) (Supplementary Material 1: Table S1 and Supplementary Material 1: Table S2). All samples were manually curated to label their detail information inferred from study description and related publications.

Quality control and gene expression estimation

Given that NCBI, ENA, and DDBJ share data through synchronized interoperability via the International Nucleotide Sequence Database Collaboration (INSDC), RNA-seq datasets were retrieved from the NCBI SRA database using the prefetch utility in SRAtoolkit v3.0.1 [19]. Subsequently, the downloaded SRA files were converted to FASTQ format using the parallel-fastq-dump tool v3.0.1 (https://github.com/rvalieris/parallel-fastq-dump). Additionally, transcriptome data from the CNCB-NGDC database were downloaded in FASTQ format via Hypertext Transfer Protocol (HTTP) protocol.

The quality control of raw reads was performed using Trimmomatic v0.40 [20] with the following parameter options: Minlen of 90, Trailing of 3, Leading of 3, and SlidingWindow of 4:5. Hisat v2.2.1 [21] was used to build the index for the genomic assembly and to align the RNA-seq reads onto the reference genome. SAM files were converted into BAM format and then the BAM files were sorted using the 'hbS' and 'sort' options of SAMtools v1.17 [22]. Stringtie v2.1.4 [23] was used for calculating the gene expression level for each gene. Transcripts Per Million (TPM) and Fragments Per Kilobase of exon per Million fragments mapped (FPKM) were obtained and can be chosen by users for custom analyses. Raw read count numbers were also obtained using the Python script prepDE.py release accompanying StringTie.

Database implementation and web interface

The main website of the database is hosted on a Tencent Cloud server with specifications of an 8-core CPU, 64 GB RAM, 500 GB storage, and a 10Mbps bandwidth (https://cloud.tencent.com/). The domain name (https://www.triticeaeexpdb.cn) was registered and linked to the server's IP address. The server operates on the Linux CentOS v7.9 operating system (http://www.centos.org). To facilitate platform development and web page creation, a Nginx v1.25.0 reverse proxy server architecture was configured.

The front-end web interface was developed using HTML5 (https://www.w3.org/html/), JavaScript (https://www.javascript.com/), CSS3 (http://www.w3.org), and Bootstrap5 (https://getbootstrap.com/). On the server side, the back-end was implemented using PHP (https://www.php.net/). MySQL v5.7.22 (https://www.mysql.com/) was utilized for storing, maintaining, and managing the gene expression. Custom PHP code was written to facilitate data retrieval from MySQL, which was then transferred to the front end for display.

Functional annotation and orthologous gene group identification

Pfam domains, Gene Ontology (GO) terms, and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways were assigned using EggNOG-mapper v2.1.12 (http://eggnog-mapper.embl.de/) [24] to enable functional term retrieval and enrichment analysis. Orthologous gene clusters were identified via OrthoFinder v2.5.5 [25], which utilized the longest protein sequences of genes and pairwise sequence similarities calculated by DIAMOND

v2.1.8 [26]. This workflow identified a total of 45,625 orthologous gene groups across 17 species, providing a framework for exploring gene expression conservation.

Integration of data mining tools

Biological quality control

Principal component analysis (PCA) was conducted using the PCA() function from the FactoMineR package v2.11 in R, allowing users to define the number of components extracted. Results are visualized as two-dimensional scatter plots generated with ggplot2 v3.5.2. For correlation analysis, expression matrices were first filtered to retain genes with non-zero standard deviation using the apply() function. Pairwise sample correlations (Pearson, Kendall, or Spearman) were then computed with the cor() function, and the resulting matrix was rendered as a hierarchical heatmap using the pheatmap package v1.0.13 to illustrate expression similarity patterns across samples.

Expression data retrieval and visualization

A custom PHP script was developed to automate expression value retrieval from the MySQL database. Heatmap generation utilized Plotly.js (https://plotly.com/) for dynamic visualization. The framework enables real-time exploration of multi-dimensional data through zooming, panning, and hover-triggered tooltips, with no reliance on external plugins.

Differential gene expression analysis

Differential expression analysis was implemented using EdgeR [27] and DESeq2 [28] integrated into the database. Significantly differentially expressed genes were identified under user-defined thresholds of FDR-adjusted p-value and log2-fold change, with results visualized as volcano plots generated via plotly.js.

Functional enrichment analysis

GO enrichment analysis was performed using clusterProfiler [29], an R/Bioconductor package for comparing biological themes among gene clusters. Enriched GO terms were visualized as bubble plots and gene-GO term networks using Plotly.js, with dark yellow and gray reflecting gene and GO terms, respectively.

Sequence homology searching

Sequence similarity searches were conducted using ViroBLAST [30], a standalone BLAST web server integrated into the database. A BLASTable database for 17 Triticeae species was constructed using makeblastdb, formatting coding sequences (CDS) and protein sequences into BLAST-compatible formats.

Co-expression network construction

Weighted gene co-expression network analysis was performed using the R package WGCNA [31]. Normalized expression data underwent preprocessing, including low-abundance gene filtering (mean FPKM < 1) and signed hybrid network construction with a soft-thresholding power $\beta = 12$. Module detection employed the dynamic tree cut algorithm (minimum module size = 30 genes). Co-expression modules were correlated with user-uploaded phenotypic data to prioritize functional modules. Results were visualized as cluster dendrogram, module-trait heatmaps, and module network graphs.

Tissue specificity analysis

Tissue specificity indices (τ) were calculated using TissueEnrich v1.26.0 [32] to identify tissue-specific genes from RNA-seq expression datasets. Gene expression data were structured as SummarizedExperiment objects for efficient data handling. Samples from multiple Bioprojects were consolidated and classified into nine tissue categories: Seedling and Whole Plant, Shoot System, Inflorescence and Spike, Seed and Endosperm, Root System, Floral Organs, Meristematic and Epidermal Tissues, Artificially Cultured Tissues, and Mixed Tissues. Mean FPKM values were computed for each tissue group, and tissue-specific genes were identified using the teGeneRetrieval function to evaluate cross-tissue expression patterns.

Genes were classified into six categories based on expression profiles: Not Expressed genes (FPKM < 1 in all tissues); Tissue Enriched genes (FPKM ≥ 1 in one tissue with ≥ fivefold higher expression than all others); Group Enriched genes (FPKM ≥ 1 in 1–3 tissues with ≥ fivefold higher expression than remaining tissues); Tissue Enhanced genes (FPKM ≥ 1 in one tissue with ≥ fivefold higher expression than the mean of other tissues); Expressed in All genes (FPKM ≥ 1 in all tissues); and Mixed genes (not fitting other categories).

Utility and discussion

Data integration, curation, and quality assessment in TriticeaeExpDB

To establish a comprehensive transcriptional atlas of Triticeae species, we integrated RNA-seq datasets from 528 studies encompassing 17 species, sourced from NCBI, EMBL, DDBJ, and NGDC, totaling 11,016 RNA-seq samples (Fig. 1 and Supplementary Material 1: Table S3). To minimize quantification biases, we reprocessed all data using a unified bioinformatics pipeline with the latest reference genomes. The dataset was dominated by three key species: *Triticum aestivum* (6078 samples, 55%), *Hordeum vulgare* (2962 samples, 27%), and *Hordeum spontaneum* (530 samples, 5%). Each study and associated samples were manually curated to ensure accurate

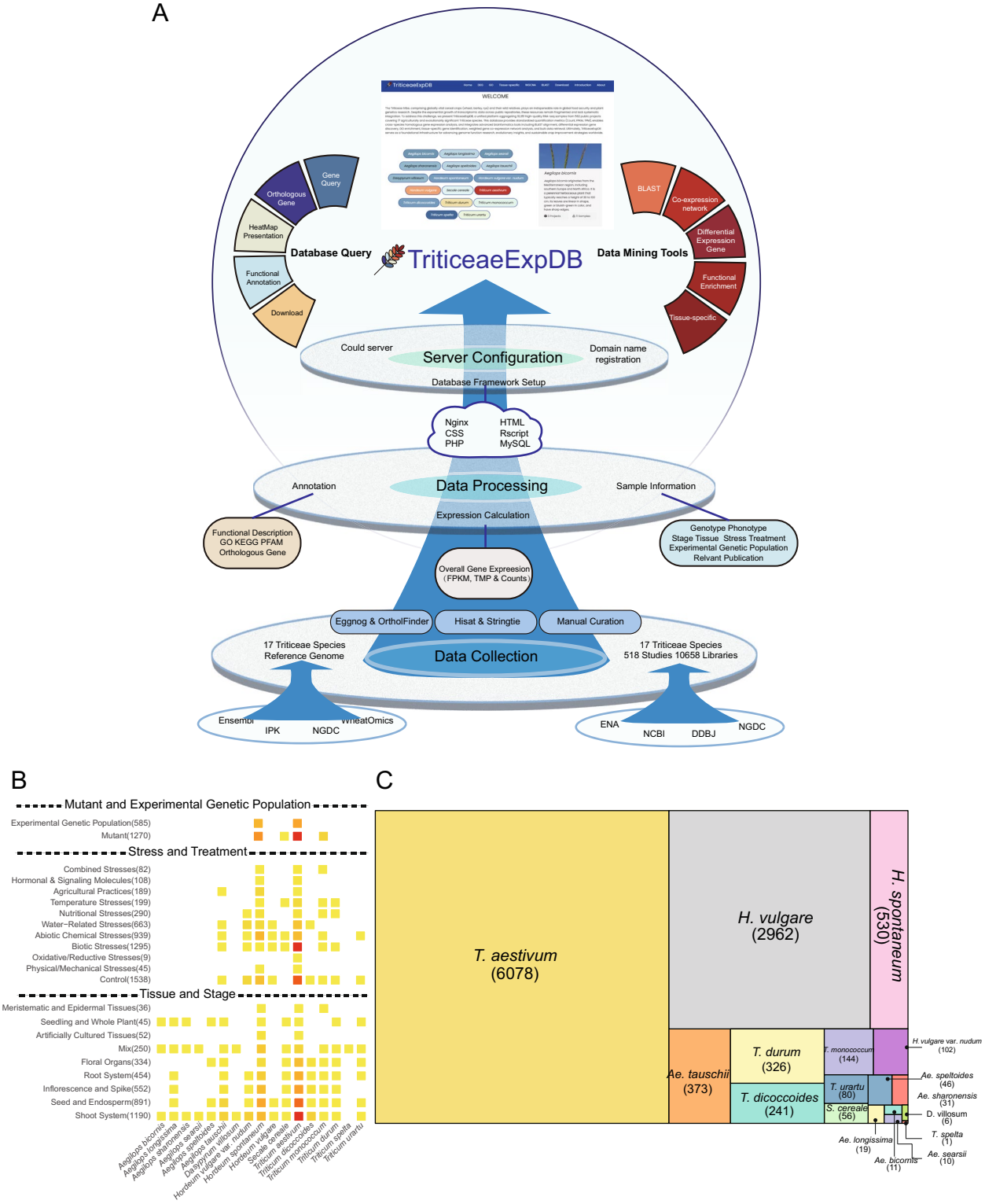

Fig. 1 Framework of TriticeaeExpDB and RNA-seq sample statistics. **A** Workflow for constructing TriticeaeExpDB. **B** Statistics of public RNA-seq libraries incorporated in TriticeaeExpDB. **C** Treemap representation of RNA-seq sample distribution across species

metadata annotation, including experimental conditions, tissue types, developmental stages, stress treatments, mutant lines, and experimental genetic populations. This meticulous annotation enables precise dataset retrieval and cross-study comparisons.

To validate data reliability, we assessed reproducibility using randomly selected samples from multiple studies. Alignment metrics confirmed high mapping efficiency, with most reads aligning concordantly to reference genomes. Intra-study reproducibility was robust, as evidenced by high mean correlation coefficients. Multivariate analyses, including PCA and hierarchical clustering, demonstrated tight grouping of biological replicates and clear separation between distinct conditions (Supplementary Material 2: Figure S1). Low-quality or inconsistent samples were systematically excluded. In summary, TriticeaeExpDB integrates rigorously curated transcriptomic data with standardized analytical workflows, ensuring reliability for comparative genomics and functional studies. This resource facilitates efficient reuse of public RNA-seq data, supporting advancements in Triticeae research.

Overview of TriticeaeExpDB

The TriticeaeExpDB integrates transcriptomic data spanning diverse biological contexts, including tissue-specific expression (*e.g.* grains, leaves, anthers, spikes, and et al.), stress responses (*e.g.* salt, drought, temperature extremes, fungal infections, and et al.), and genetic populations (*e.g.* RIL: Recombinant Inbred Lines, NIL: Near Isogenic Lines, DH: Doubled Haploid, and et al.) (Fig. 2). All datasets undergo rigorous curation, with raw sequencing data derived from controlled experiments and accompanied by comprehensive metadata annotations.

The platform features an intuitive interface supporting queries across 17 Triticeae species, accessible by both Latin and common names. Each species portal displays key statistics including BioProject counts, sample numbers, and reference genome versions. Gene queries support both single-gene or multi-gene searches and batch analysis via file uploads, with options to download complete datasets for offline analysis. BioProjects are systematically indexed using descriptive tags to enable efficient dataset discovery.

Gene information pages consolidate functional annotations from multiple sources, including GO terms, KEGG pathways, and Pfam domains. Project information sections provide detailed experimental metadata such as associated publications, project descriptions, and experimental designs. Expression data are available in TPM, FPKM, and raw count formats, with options to view either individual replicate values or biologically averaged expression levels while maintaining experimental group integrity.

Interactive heatmaps enable visualization of expression patterns, with downloadable PNG images and exportable expression matrices. Comprehensive sample metadata includes genotype or phenotype characteristics, developmental stages, tissue types, and treatment conditions. For cross-species comparisons, the database provides ortholog group IDs that facilitate examination of expression conservation and divergence across the 17 included Triticeae species.

Application toolkits

Beyond its core querying capabilities, TriticeaeExpDB integrates comprehensive analytical tools for biological quality control, differentially expressed genes (DEG) analysis, GO enrichment, co-expression network construction, sequence alignment, and tissue-specific expression evaluation, significantly enhancing research efficiency (Fig. 3).

For biological quality control, the platform provides an interactive module for PCA and sample correlation analysis. Users can upload their expression matrix to perform PCA, visualizing global expression patterns and batch effects through a 2D scatter plot. The accompanying results table details the principal component scores for each sample across multiple dimensions (*e.g.*, PC1 *vs.* PC2). Sample-to-sample expression consistency is further evaluated via a correlation heatmap, which uses a color gradient (blue to red) to represent pairwise Pearson correlation coefficients, allowing rapid identification of potential outliers and technical artifacts. Both the PCA scores and the correlation matrix are searchable, sortable, and exportable, providing a transparent foundation for downstream analysis.

The platform features a BLAST-based homology analysis module that enables sequence similarity searches against curated Triticeae coding sequences and protein databases. Users can submit nucleotide or amino acid sequences via FASTA format through either a text interface or file upload, with five optimized BLAST execution modes available for different research purposes. Key parameters include adjustable E-value thresholds, word sizes, and target sequence matching limits. Results are presented in an interactive table sorted by E-value, displaying alignment metrics such as bit scores, percent identity, query-subject overlap lengths, and hyperlinked NCBI accession numbers for direct metadata retrieval.

For comparative transcriptomic analysis, the platform incorporates established differential expression tools (DESeq2 and edgeR) with default stringent thresholds ($|\log2FC| \geq 2$; adjusted p-value < 0.01), while permitting user-defined parameter adjustments. Analytical outputs include interactive summary tables and volcano plots for visualizing significant expression changes across tissues, developmental stages, and stress conditions.

Fig. 2 Overview of the TriticeaeExpDB interface. **A** Database homepage. **B** Key features and analytical toolkits of TriticeaeExpDB. **C** Gene expression query interface, exemplified by a search for *Hordeum vulgare*. **D** Gene expression search results. **E** Functional annotations including GO terms and KEGG pathways for queried genes. **F** List of homologous genes

Fig. 3 Application toolkits integrated in TriticeaeExpDB. **A** Data quality control (**B**) Triticeae homologues. **C** Differential gene expression analysis. **D** Functional enrichment analysis. **E** Tissue-specific gene evaluation. **F** Gene co-expression network analysis

GO enrichment analysis is performed using clusterProfiler, allowing functional annotation of gene lists against customizable Triticeae reference genomes. The system provides flexible parameter settings for statistical thresholds (p-value and q-value), multiple testing corrections (BH, Bonferroni, or FDR), and ontology category filters. Results comprise detailed enrichment tables alongside visual representations, including fold change-associated bubble plots and gene-term interaction networks that elucidate functional relationships.

The platform implements WGCNA for identifying co-expressed gene networks, enabling users to submit custom expression matrices and trait data. Analysis outputs include: module-trait association heatmaps, hierarchical clustering dendrograms of expression patterns, and comprehensive network tables with pairwise Pearson correlation coefficients.

To assess tissue-specific gene expression patterns, we performed enrichment analysis using the TissueEnrich package in R. Genes were systematically classified into six distinct categories based on their expression profiles: Not Expressed, Tissue Enriched, Group Enriched, Tissue Enhanced, Expressed in All, and Mixed, with Tissue Enriched, Group Enriched, and Tissue Enhanced genes defined as tissue-specific. For user-submitted gene queries, TriticeaeExpDB dynamically retrieves and displays their expression patterns, identifies tissue-specific expression (where applicable), and visualizes expression levels across nine major tissue types in interactive bar plots. Example datasets and detailed analytical protocols are provided in the Help and FAQ sections to facilitate user exploration.

Case studies

TriticeaeExpDB serves as a valuable resource for in silico screening to identify genotypes, tissues or conditions exhibiting differential expression of target genes. As an example, we investigated *Btr1*, a key gene controlling rachis brittleness in barley, which, together with *Btr2*, determines the difference in spike fragility between wild and cultivated barley [33]. Using the *Btr1* reference sequence (GenBank ID: AKV61769.1), we performed a BLAST search against the barley whole-genome CDS dataset in TriticeaeExpDB, identifying a perfect match (*HvBtr1*: *HORVU.MOREX.r3.3HG0235600*). Subsequent analysis of RNA-seq data from 16 developmental stages and tissues (BioProject PRJEB14349) [34] revealed minimal expression (FPKM < 1) across all samples (Supplementary Material 2: Figure S2). Orthologs in wild barley and bread wheat also showed similarly low expression levels. These results demonstrate TriticeaeExpDB's ability to reliably detect low-abundance transcripts, suggesting that *Btr1* function may rely on precise spatial–temporal regulation rather than high transcript abundance, a hallmark of key regulatory genes with strong tissue (*e.g.*, rachis) and stage (*e.g.*, grain maturation) specificity.

In addition, we characterized the expression pattern of the wheat domestication gene *Q* (*TaQ-5A*: *TraesCS5A02G473800*), which influences critical traits such as tough rachis, free-threshing habit, spike morphology, 1000-grain weight, and plant height [35]. Expression analysis revealed broad, high expression levels across multiple spike, grain and leaf developmental stages (Supplementary Material 2: Figure S3). For instance, in spike, expression reached an FPKM of 67.55, while the average FPKM across samples in grain and leaf was 3.91 and 7.31, indicating that *Q* is widely expressed and likely functional at numerous developmental stages in wheat. Together, these contrasting examples, a low-expression gene with specific regulatory roles and a broadly expressed, high-abundance domestication gene, demonstrate the utility of TriticeaeExpDB for uncovering diverse expression patterns underlying important biological functions.

Additionally, TriticeaeExpDB facilitates the discovery of candidate genes associated with agronomic traits. We analyzed RNA-seq data from allohexaploid wheat under long-term salinity stress (BioProject PRJNA573996) [36], which included root and leaf tissues under control and stress conditions (three replicates each, totaling 12 samples). Differential expression analysis of root samples identified 628 upregulated and 1,587 downregulated genes (Fig. 4). GO enrichment analysis of upregulated genes revealed significant enrichment in 111 molecular function (MF), 181 biological process (BP), and 3 cellular component (CC) terms. WGCNA was employed to assess gene connectivity, and the top 10 most differentially expressed genes were selected for further evaluation. Integration with WGCNA network data and ortholog annotation identified six high-priority candidates. Notably, these genes exhibited strong upregulation (LogFC > 8, $P < 0.01$) in barley under salt stress (BioProject PRJNA489775), suggesting potential co-evolutionary conservation. These candidates warrant functional characterization to elucidate their roles in stress adaptation in Triticeae crops.

TriticeaeExpDB powers Triticeae multi-omics research through an integrated gene expression platform

The Triticeae tribe, a critical component of Poaceae, plays a pivotal role in global food security, comprising essential cereal crops such as *Triticum aestivum*, *Hordeum vulgare*, and *Secale cereale* [10]. Among these, polyploid wheat varieties have emerged as dominant cultivars in temperate regions, surpassing other major cereals like maize and rice in agricultural significance [37]. Recent progress in high-throughput sequencing and computational biology has facilitated the assembly of highly repetitive Triticeae genomes, while the widespread adoption

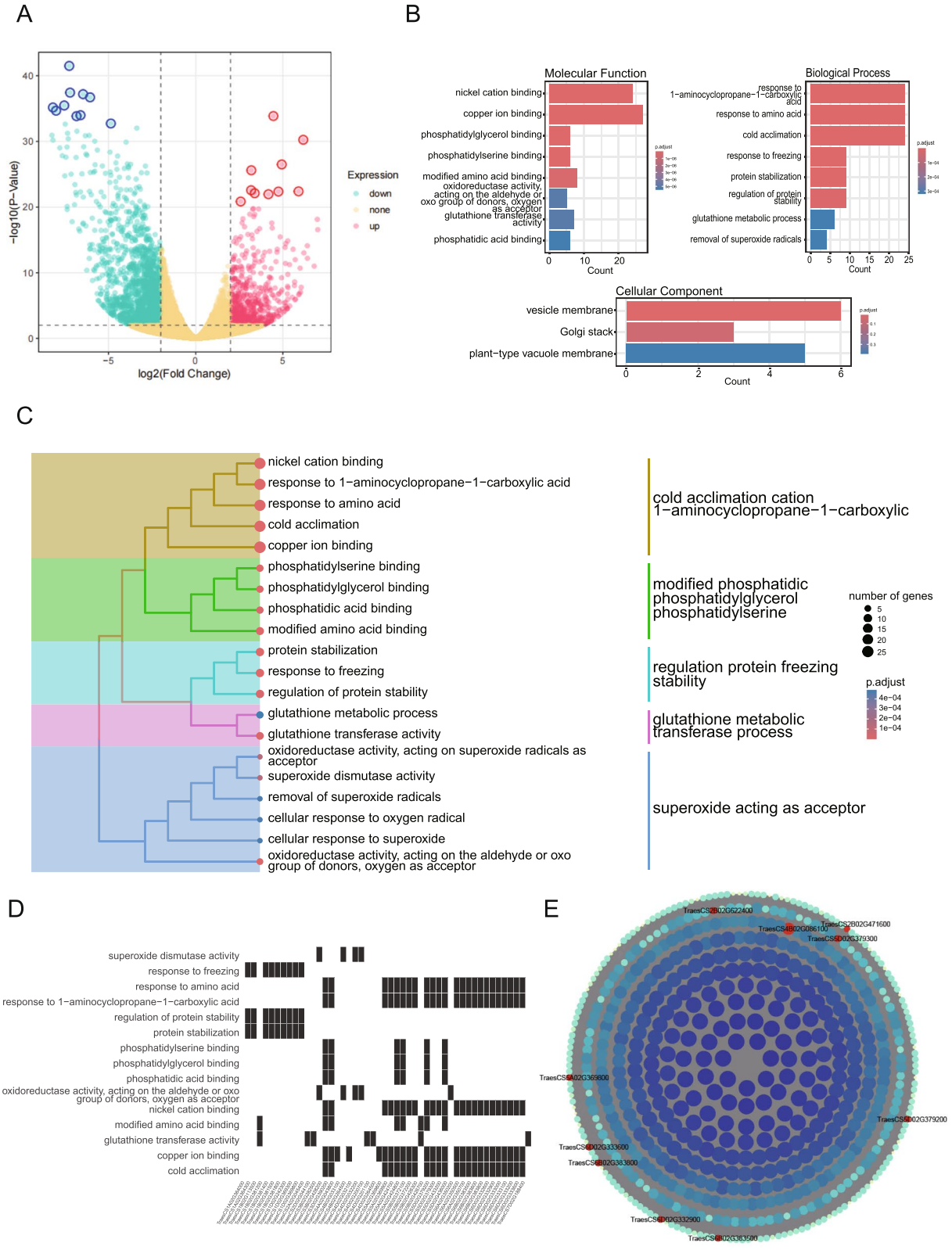

Fig. 4 Case study demonstrating data mining using TriticeaeExpDB (BioProject PRJNA573996 as example). **A** Volcano plot of differentially expressed genes. **B** GO enrichment analysis of upregulated genes. **C** Dendrogram topology constructed based on semantic similarity of enriched terms or a user-defined distance matrix. **D** Gene-pathway association matrix. **E** Constructed co-expression network

of cost-effective next-generation sequencing has yielded extensive transcriptomic datasets [38]. This rapid expansion of omics data underscores the need for integrated databases to enable systematic organization, analysis, and functional genomics exploration.

Several transcriptomic databases have been developed for Triticeae species, including WheatExp [39], WheatOmics [18] and BarleyExpDB [11]. While these resources provide valuable data, they are predominantly species-specific and lack integrated toolkits tailored for comparative transcriptomics across the entire Triticeae tribe. Moreover, comprehensive platforms that encompass wild relatives and understudied Triticeae species remain scarce. Existing databases also seldom combine orthology-based gene matching with uniformly processed expression profiles from multiple species, limiting their utility for investigating cross-species expression conservation.

TriticeaeExpDB addresses these limitations through three key advancements: quantitative scale, qualitative functionality, and scalable architecture. First, in data scale and species coverage, TriticeaeExpDB integrates over 10,000 uniformly reprocessed RNA-seq samples, an order of magnitude larger than most current resources, spanning 17 Triticeae species, including essential wild relatives. This extensive dataset, coupled with more than 45,000 pre-computed orthologous gene groups, establishes the dedicated platform for cross-Triticeae comparative transcriptomics.

Second, in analytical functionality, TriticeaeExpDB transcends the role of a mere data repository. Unlike broader multi-omics portals such as WheatOmics, our platform offers an integrated suite of workflows specifically optimized for Triticeae transcriptomics. Its core innovation lies in seamlessly linking ortholog-aware query systems with downstream analytical modules, including differential expression, co-expression networks, and tissue-specificity assessments. This end-to-end, species-aware pipeline directly tackles the challenges of analyzing polyploid genomes and enables reliable cross-species comparisons, a capability not cohesively implemented in existing platforms.

Beyond Triticeae-focused resources, general plant genomics databases such as GrainGenes (for temperate cereals and legumes) [40] and Gramene (a comparative genomics resource for grasses) [41] offer indispensable access to multi-omics data across diverse species. TriticeaeExpDB complements rather than competes with these established repositories. Whereas GrainGenes and Gramene serve as broad-spectrum portals for comparative genomics, TriticeaeExpDB is a deeply specialized platform dedicated to Triticeae transcriptomics. It distinguishes itself through uniform reprocessing of tribe-wide RNA-seq data, provision of pre-computed orthologs for cross-species queries, and integration of analytical toolkits optimized for polyploid and comparative expression analysis, features that address specific yet critical needs within the Triticeae research community. Thus, TriticeaeExpDB serves as a curated, expression-centric module that enhances the ecosystem anchored by these larger databases. We plan to implement bidirectional linking between TriticeaeExpDB and these foundational resources to improve user connectivity and data interoperability.

Third, its scalable and modular architecture ensures long-term utility and adaptability. Built on a cloud-based LNMP stack, the platform can process over 1,000 new RNA-seq samples per day, enabling timely updates. The design accommodates future integration of telomere-to-telomere genomes [42], upcoming barley Morex V4 assemblies [43], and pangenome references. Planned enhancements also include AI-powered tools for predictive breeding, positioning TriticeaeExpDB as a continuously evolving research platform rather than a static database.

Looking ahead, we envision TriticeaeExpDB serving as a central hub for multi-omics integration in Triticeae research. Future developments will prioritize the incorporation of emerging transcriptomic technologies, including full-length RNA-seq, single-cell RNA-seq, and spatial transcriptomics, to resolve splicing dynamics and tissue-specific expression at unprecedented resolution [44, 45]. Ultimately, we aim to connect genomic, transcriptomic, proteomic, and phenomic data to elucidate the regulatory networks underlying key agronomic traits. Beyond the Triticeae tribe, we plan to enhance interoperability with broader plant bioinformatics ecosystems, such as Gramene and GrainGenes, positioning TriticeaeExpDB as a specialized transcriptomic module within a connected data landscape.

To ensure that these ambitions are sustainably supported, we have implemented a concrete maintenance and governance plan. The platform benefits from sustained institutional funding and infrastructure, backed by our team's proven record of maintaining public databases for over five years. We commit to performing at least two major data updates annually, integrating newly published RNA-seq datasets, alongside continuous functional optimizations and security enhancements. All major updates will be permanently archived in public repositories to ensure the traceability of data associated with published research. User-focused communication is maintained through a dedicated contact email and a regularly updated news section. Collectively, these measures form a robust, transparent framework designed to sustain TriticeaeExpDB as a stable, evolving, and community-oriented resource for the long term.

By maintaining open-data principles and community-driven framework, TriticeaeExpDB not only redefines comparative transcriptomics but also serves as a replicable model for other polyploid crops facing climate adaptation challenges.

Conclusions

We introduce TriticeaeExpDB, a comprehensive expression atlas and analytical platform for Triticeae species, integrating a total of 11,016 high-quality RNA-seq samples from 17 species systematically curated from 528 BioProjects. All samples were uniformly processed using a standardized computational pipeline to ensure cross-comparability. The platform facilitates multidimensional queries of spatiotemporal expression patterns, supporting TPM, FPKM, and raw count metrics, alongside integrated analytical tools for comparative group analyses and hypothesis generation. As a centralized repository, TriticeaeExpDB enhances accessibility to Triticeae transcriptomic data, serving as both a discovery engine and a hypothesis-testing resource for the research community.

Abbreviations

BP	Biological Process
CC	Cellular Component
CDS	Coding Sequences
CNCB-NGDC	China National Center for Bioinformation - National Genomics Data Center
DDBJ	DNA Data Bank of Japan
DEG	Differentially Expressed Genes
DH	Doubled Haploid
ENA	European Nucleotide Archive
FPKM	Fragments Per Kilobase of exon per Million fragments mapped
GO	Gene Ontology
GSA	Genome Sequence Archive
HTTP	Hypertext Transfer Protocol
INSDC	International Nucleotide Sequence Database Collaboration
KEGG	Kyoto Encyclopedia of Genes and Genomes
MF	Molecular Function
NCBI	National Center for Biotechnology Information
NIL	Near Isogenic Lines
PCA	Principal component analysis
RIL	Recombinant Inbred Lines
RNA-seq	RNA sequencing
SRA	Sequence Read Archive
TPM	Transcripts Per Million
TriticeaeExpDB	Triticeae Expression Database

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-026-12630-0.

Supplementary Material 1.

Supplementary Material 2.

Acknowledgements

We would like to express our gratitude to the research groups that made their reference genomes publicly available. Due to space constraints, we apologize for not being able to cite all relevant papers in this manuscript. We are also grateful to the High-Performance Computing platform at Northwest A&F University for their assistance with data processing.

Authors' contributions

TL and LX contributed equally to this work. TL: Writing—original draft, Formal analysis, Visualization. LX: Writing—original draft, Formal analysis, Visualization. HZ: Writing—original draft, Data curation. YS: Writing—original draft, Formal analysis, Visualization. RL: Writing—original draft, Supervision. YL: Writing—original draft, Writing—review & editing, Supervision. LC: Writing—review & editing, Supervision. XN: Writing—review & editing, Supervision. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the Open Project Program of the State Key Laboratory for Crop Stress Resistance and High-Efficiency Production (Grant No. SKLCSRHPKF11), the Natural Science Foundation of Jiangxi Province (Grant No. 20232BAB205012), the Major Project on Agricultural Bio-breeding of China (Grant No. 2023ZD0402601), and the National Natural Science Foundation of China (Grants No. 32360157, 32372100, and U22A20457). The funding agencies had no role in the study design, data collection, analysis, interpretation, or manuscript preparation.

Data availability

The original contributions of this study are included in the article and supplementary materials. Detailed information regarding the expression profiles across the 17 Triticeae species can be downloaded directly from https://www.triticeaeexpdb.cn. For any further inquiries, please contact the corresponding author.

Declarations

Ethics approval and consent to participate

This study did not involve human participants, animal subjects, or any materials requiring specific permits/ethical approval.

Consent for publication

All authors have read and approved the manuscript for publication.

Competing interests

The authors declare no competing interests.

Received: 25 October 2025 / Accepted: 3 February 2026
Published online: 09 February 2026

References

1. Soreng RJ, Peterson PM, Romaschenko K, Davidse G, Zuloaga FO, Judziewicz EJ, et al. A worldwide phylogenetic classification of the Poaceae (Gramineae). J Syst Evol. 2015;53(2):117–37.
2. Sato K, Yoshida K, Takumi S. RNA-Seq-based DNA marker analysis of the genetics and molecular evolution of Triticeae species. Funct Integr Genomics. 2021;21(5–6):535–42.
3. Schnurbusch T. Wheat and barley biology: towards new frontiers. J Integr Plant Biol. 2019;61(3):198–203.
4. Al-Saghir M. Taxonomy and phylogeny in Triticeae: a historical review and current status. Adv Plants Agr Res. 2016;3:00108.
5. Nevo E, Chen G. Drought and salt tolerances in wild relatives for wheat and barley improvement. Plant Cell Environ. 2010;33(4):670–85.
6. Monat C, Padmarasu S, Lux T, Wicker T, Gundlach H, Himmelbach A, et al. TRITEX: chromosome-scale sequence assembly of Triticeae genomes with open-source tools. Genome Biol. 2019;20(1):284.
7. Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, et al. DNA sequencing at 40: past, present and future. Nature. 2017;550(7676):345–53.
8. Gao Z, Bian J, Lu F, Jiao Y, He H. Triticeae crop genome biology: an endless frontier. Front Plant Sci. 2023;14:1222681.
9. Sun Y, Shang L, Zhu QH, Fan L, Guo L. Twenty years of plant genome sequencing: achievements and challenges. Trends Plant Sci. 2022;27(4):391–401.

10. Wang X, Yan X, Hu Y, Qin L, Wang D, Jia J, et al. A recent burst of gene duplications in Triticeae. Plant Commun. 2022;3(2):100268.
11. Li T, Li Y, Shangguan H, Bian J, Luo R, Tian Y, et al. BarleyExpDB: an integrative gene expression database for barley. BMC Plant Biol. 2023;23(1):170.
12. Liu J, Lang K, Tan S, Jie W, Zhu Y, Huang S, et al. A web-based database server using 43,710 public RNA-seq samples for the analysis of gene expression and alternative splicing in livestock animals. BMC Genomics. 2022;23(1):706.
13. Zhang H, Zhang F, Yu Y, Feng L, Jia J, Liu B, et al. A comprehensive online database for exploring ~20,000 public Arabidopsis RNA-Seq libraries. Mol Plant. 2020;13(9):1231–3.
14. Liu J, Yin F, Lang K, Jie W, Tan S, Duan R, et al. MetazExp: a database for gene expression and alternative splicing profiles and their analyses based on 53 615 public RNA-seq samples in 72 metazoan species. Nucleic Acids Res. 2022;50(D1):D1046-d1054.
15. Yu Y, Zhang H, Long Y, Shu Y, Zhai J. Plant public RNA-seq database: a comprehensive online database for expression analysis of ~45 000 plant public RNA-Seq libraries. Plant Biotechnol J. 2022;20(5):806–8.
16. Database Resources of the National Genomics Data Center. China National Center for Bioinformation in 2025. Nucleic Acids Res. 2025;53(D1):D30-d44.
17. Dyer SC, Austine-Orimoloye O, Azov AG, Barba M, Barnes I, Barrera-Enriquez VP, et al. Ensembl 2025. Nucleic Acids Res. 2025;53(D1):D948-d957.
18. Ma S, Wang M, Wu J, Guo W, Chen Y, Li G, et al. WheatOmics: a platform combining multiple omics data to accelerate functional genomics studies in wheat. Mol Plant. 2021;14(12):1965–8.
19. Leinonen R, Sugawara H, Shumway M. The sequence read archive. Nucleic Acids Res. 2011;39(Database issue):D19-21.
20. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
21. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37(8):907–15.
22. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
23. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol. 2019;20(1):278.
24. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Mol Biol Evol. 2021;38(12):5825–9.
25. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20(1):238.
26. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12(1):59–60.
27. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40.
28. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
29. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS J Integr Biol. 2012;16(5):284–7.
30. Deng W, Nickle DC, Learn GH, Maust B, Mullins JI. ViroBLAST: a stand-alone BLAST web server for flexible queries of multiple databases and user's datasets. Bioinformatics. 2007;23(17):2334–6.
31. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559.
32. Jain A, Tuteja G. Tissueenrich: tissue-specific gene enrichment analysis. Bioinformatics. 2019;35(11):1966–7.
33. Pourkheirandish M, Hensel G, Kilian B, Senthil N, Chen G, Sameri M, et al. Evolution of the Grain Dispersal System in Barley. Cell. 2015;162(3):527–39.
34. Mascher M, Gundlach H, Himmelbach A, Beier S, Twardziok SO, Wicker T, et al. A chromosome conformation capture ordered sequence of the barley genome. Nature. 2017;544(7651):427–33.
35. Liu P, Xue S, Jia J, Zhao G, Liu J, Hu Y, et al. The wheat transcription factor Q functions in gibberellin biosynthesis and signaling and regulates height and spike length. Plant Cell. 2025;37(8):koaf183.
36. Bhanbhro N, Xiao B, Han L, Lu H, Wang H, Yang C. Adaptive strategy of allohexaploid wheat to long-term salinity stress. BMC Plant Biol. 2020;20(1):210.
37. Pont C, Leroy T, Seidel M, Tondelli A, Duchemin W, Armisen D, et al. Tracing the ancestry of modern bread wheats. Nat Genet. 2019;51(5):905–11.
38. Adamski NM, Borrill P, Brinton J, Harrington SA, Marchal C, Bentley AR, et al. A roadmap for gene functional characterisation in crops with large genomes: lessons from polyploid wheat. Elife. 2020;9:e55646.
39. Pearce S, Vazquez-Gross H, Herin SY, Hane D, Wang Y, Gu YQ, et al. WheatExp: an RNA-seq expression database for polyploid wheat. BMC Plant Biol. 2015;15:299.
40. Yao E, Blake VC, Sharma PK, Poretsky E, Michel S, Birkett CL, et al. GrainGenes: genetics, genomes, and pangenomes. Genetics. 2025;iyaf270.
41. Tello-Ruiz MK, Jaiswal P, Ware D. Gramene: a resource for comparative analysis of plants genomes and pathways. Methods Mol Biol. 2022;2443:101–31.
42. Navrátilová P, Toegelová H, Tulpová Z, Kuo YT, Stein N, Doležel J, et al. Prospects of telomere-to-telomere assembly in barley: analysis of sequence gaps in the MorexV3 reference genome. Plant Biotechnol J. 2022;20(7):1373–86.
43. Liu S, Li K, Dai X, Qin G, Lu D, Gao Z, et al. A telomere-to-telomere genome assembly coupled with multi-omic data provides insights into the evolution of hexaploid bread wheat. Nat Genet. 2025;57(4):1008–20.
44. Shaw R, Tian X, Xu J. Single-cell transcriptome analysis in plants: advances and challenges. Mol Plant. 2021;14(1):115–26.
45. Byrne A, Cole C, Volden R, Vollmers C. Realizing the potential of full-length transcriptome sequencing. Philos Trans R Soc Lond B Biol Sci. 2019;374(1786):20190097.

Publisher's Note