Computational and Structural Biotechnology Journal 21 (2023) 3327–3338

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Database Article

SoybeanGDB: A comprehensive genomic and bioinformatic platform for soybean genetics and genomics

Haoran Li [a,1], Tiantian Chen [a,1], Lihua Jia [a,1], Zhizhan Wang [a], Jiaming Li [a], Yazhou Wang [a], Mengjia Fu [a], Mingming Chen [a], Yuping Wang [a], Fangfang Huang [a], Yingru Jiang [a], Tao Li [a], Zhengfu Zhou [b], Yang Li [a,*], Wen Yao [a,*], Yihan Wang [a,*]

[a] National Key Laboratory of Wheat and Maize Crop Science, College of Life Sciences, Henan Agricultural University, Zhengzhou 450002, China
[b] Henan Academy of Crop Molecular Breeding, Henan Academy of Agricultural Sciences, Zhengzhou 450002, China

ARTICLE INFO

Article history:
Received 1 January 2023
Received in revised form 9 June 2023
Accepted 10 June 2023
Available online 12 June 2023

Keywords:
Soybean
Bioinformatic platform
Genome database
Genomic variation
Zhonghuang 13

ABSTRACT

Soybean (*Glycine max* (L.) Merr.) is a globally significant crop, widely cultivated for oilseed production and animal feeds. In recent years, the rapid growth of multi-omics data from thousands of soybean accessions has provided unprecedented opportunities for researchers to explore genomes, genetic variations, and gene functions. To facilitate the utilization of these abundant data for soybean breeding and genetic improvement, the SoybeanGDB database (https://venyao.xyz/SoybeanGDB/) was developed as a comprehensive platform. SoybeanGDB integrates high-quality de novo assemblies of 39 soybean genomes and genomic variations among thousands of soybean accessions. Genomic information and variations in user-specified genomic regions can be searched and downloaded from SoybeanGDB, in a user-friendly manner. To facilitate research on genetic resources and elucidate the biological significance of genes, SoybeanGDB also incorporates a variety of bioinformatics analysis modules with graphical interfaces, such as linkage disequilibrium analysis, nucleotide diversity analysis, allele frequency analysis, gene expression analysis, primer design, gene set enrichment analysis, etc. In summary, SoybeanGDB is an essential and valuable resource that provides an open and free platform to accelerate global soybean research.

1. Introduction

Soybean, one of the world's most essential crops, is widely used in diverse food products and serves as an excellent source of protein for animal feeds (https://www.fao.org/). Achieving high yield and enhancing the qualities of oil and protein are the major goals of soybean breeding and genetic improvement, which require systematic investigation of genomic variations associated with agronomical traits. Leveraging high-throughput and precise genotyping and phenotyping data obtained from experimental or natural/wild populations represents an efficient strategy to uncover the genetic mechanisms underlying diverse phenotype variations in various organisms [1–3]. With the rapid advancement of sequencing technology, a large number of soybean accessions have been sequenced to elucidate the genetic basis of various agronomic traits and facilitate molecular breeding in soybean [4–7]. In different types of soybean biological studies, genomic variations such as SNPs, InDels, structural variations (or genomic rearrangements), and copy number variations have been employed to identify candidate variations/genes for in-depth functional analysis in soybean.

With the explosive growth of biological data generated by high-throughput sequencing and other biotechnologies, a comprehensive genome database has become an indispensable tool for studying various organisms [8–10]. Currently, three soybean-related genome databases are available, including SoyBase [11,12], SoyKB (Soybean knowledge base) [13], and WildsoyDB [14]. SoyBase is the leading soybean database based on the genome of *Glycine max* L. cv. Williams 82, which was the first reference genome for soybean and has greatly contributed to functional genomic studies of soybean [15]. SoyBase encompasses genome sequences and gene annotations of Williams 82, along with diverse datasets such as gene expression data, epigenetic data, molecular markers, and phenotypic traits

* Corresponding authors.
E-mail addresses: liyang@henau.edu.cn (Y. Li), yaowen@henau.edu.cn (W. Yao), yihanwang@vip.163.com (Y. Wang).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.csbj.2023.06.012

[11,12]. SoyKB, also based on the Williams 82 genome, is a comprehensive knowledge database developed for functional genomics research and molecular breeding of soybean [13]. It provides various analysis tools for data visualization and functional genomic studies. WildsoyDB is another integrated online platform that aggregates four soybean genomes, including *Glycine max* Williams 82 (a2v1 and a4v1), *Glycine max* Zhonghuang 13 (v2) and *Glycine soja* W05 (v1). It offers various functional modules for analyzing genomic resources [14].

Over the past few years, a golden reference genome for *Glycine max* L. cv. Zhonghuang 13, a widely cultivated soybean variety in China, has been reported [16,17]. Following this, a pan-genome analysis was conducted based on the genome sequence of Zhonghuang 13 and high-quality de novo assemblies of 26 other representative soybeans [4]. Furthermore, the high-quality genomes of *Glycine max* Lee, *Glycine max* Hwangkeum, *Glycine soja* PI483463, *Glycine soja* W05 and six perennial *Glycine* species have enriched the genomic resources available for soybean research [18–21]. Moreover, the re-sequencing data of diverse soybean accessions, which is another valuable resource for mining candidate variations or genes, is consistently increasing [4,7]. Despite the enormous amount of newly sequenced data reported in recent years, including high-quality genome assemblies of soybean and genomic variations among thousands of soybean accessions [4,7,18–21], it has not been fully integrated into any database such as SoyBase, SoyKB or WildsoyDB [11–14]. To address this issue, we have developed an integrative genomic database, SoybeanGDB. It aggregates comprehensive information from 39 high-quality de novo assembled soybean genomes, as well as 15,446,616 SNPs and 4136,231 InDels identified using re-sequencing data from 2898 soybean accessions and 7869,806 SNPs among 481 accessions. SoybeanGDB incorporates diverse functional modules that facilitate genomic and genetic studies. Users can search, browse, analyze, and download details related to coding/non-coding regions, SNPs and InDels. SoybeanGDB also hosts various versatile analytical tools such as JBrowse 2, expression analysis, BLAST, Primer-design, and gene set enrichment analysis, serving the soybean community.

2. Materials and methods

2.1. Data collection

A comprehensive collection of 39 high-quality de novo genome assemblies of soybean, previously reported in various studies, has been integrated into SoybeanGDB (Table S1) [4,11,17–22]. Out of these, the genomes of 28 soybean accessions consisting of three wild soybeans, nine landraces and 16 improved cultivars, were downloaded from the Genome Warehouse (https://ngdc.cncb.ac.cn/gwh/) [23]. The genomes of *Glycine max* Williams 82 and six perennial *Glycine* species were obtained from SoyBase (https://www.soybase.org/) [11,21]. In addition, the genomes of *Glycine max* Lee and *Glycine soja* PI483463 were collected from Phytozome (https://phytozome.jgi.doe.gov) [19]. The genomes of *Glycine soja* W05 and Hwangkeum were obtained from http://www.wildsoydb.org/ and https://www.ncbi.nlm.nih.gov/nuccore/, respectively [18,20].

SNPs and InDels among 2898 soybean accessions, which consisted of 1747 improved soybean cultivars, 1048 soybean landraces, and 103 *Glycine soja* accessions, were downloaded from the Genome Variation Map database (https://ngdc.cncb.ac.cn/gvm/home) [24]. Moreover, SNPs from 481 soybean accessions, comprising 429 cultivated varieties and 52 *Glycine soja* accessions, were obtained from https://data.nal.usda.gov/search/type/dataset [7].

The gene expression levels (FPKM) of *Glycine max* Zhonghuang 13 (55,443 genes across 27 samples from various tissues and stages), an experimental line A81–356022 (38,177 genes from seven tissues and seven stages in seed development), *Glycine soja* W05 (21,402 genes from trifoliate and primary leaves, and roots of young seedlings), and 102 soybean accessions (56,044 genes form 18-day-old leaves) were collected from previous studies conducted by Shen et al. [17], Severin et al. [25], Qi et al. [26] and Li et al. [27], respectively. These details can be found in the “Accessions” and “Data source” pages of the SoybeanGDB database.

2.2. Data processing

2.2.1. Processing of soybean genomic data

Based on the genome annotation file in GFF3 format, the sequences of genes, CDSs, cDNAs, and proteins from the 39 soybean genomes were extracted using in-house R scripts. Subsequently, the extracted sequences were stored as R data files in SoybeanGDB.

Transposable elements (TEs) in the 39 soybean genomes were annotated using two tools: Extensive de-novo TE Annotator (EDTA) v2.0.0 and RepeatMasker 4.1.0 [28,29]. EDTA was employed with the following parameters “–overwrite 1 –sensitive 1 22 –anno 1 –evaluate 1 –threads 50”. This allowed the construction of a filtered non-redundant TE library based on the whole genome sequence of Zhonghuang 13. Next, RepeatMasker 4.1.0 was used to identify transposons in each of the 39 soybean genomes by utilizing the filtered non-redundant TE library.

Transcription factors and transcriptional regulators in the 39 soybean genomes were identified and classified using the prediction tool iTAK v1.6 with the parameters “perl iTAK.pl -a 10 protein.fasta” [30].

2.2.2. Identification of syntenic and rearranged regions

To define syntenic regions and identify structural variations, including duplicated region, inverted region, translocated region, tandem repeat, deletions, and insertions, between two soybean genomes, we employed SyRI to compare each of the 37 genomes with either Zhonghuang 13 or Williams 82 reference genome. The parameters of SyRI were “python syri -c bamfile -r fastafile1 -q fastafile2 -k -F B” [31]. In this process, minimap2 was used to align the chromosome-level sequences of the 37 genomes to the reference genomes Zhonghuang 13 or Williams 82, with the parameter “-ax asm5 –eqx -t 64” [32]. The alignment results for each pair of genomes were then submitted to SyRI, which identified the syntenic and rearranged regions with the parameter “-k -F S”.

2.2.3. Processing of SNPs and InDels

The SNP data among 2898 accessions were downloaded from the Genome Variation Map database [24]. For each soybean accession, the raw datasets stored SNP likelihoods across all variation sites. To process the raw SNPs among the 2898 soybean accessions, we followed the criteria reported in a previous study with slight modifications [33]. First, low-quality SNP sites were removed if the total number of soybean accessions with a heterozygote genotype or homozygous minor genotype was < 5. After that, only biallelic SNPs were retained. Then, SNP sites with a missing rate $> 5\%$ and a read depth ≤ 500 or $\geq 50{,}000$ were removed. Further, the raw SNP data with character genotype matrices of “A, C, G and T” were converted into integer sparse matrices of “0” and “1” using in-house R scripts. To facilitate efficient loading of the SNP data, SNP sites in each 50-Mb non-overlapping genomic region of Zhonghuang 13 genome were stored as an R data file. Subsequently, all the SNPs were annotated using SnpEff (http://pcingola.github.io/SnpEff/) based on the genomic structure of Zhonghuang 13 [34].

The InDels underwent the same processing procedure as SNPs, with the exception that multiallelic InDels were retained. To facilitate fast retrieval using Tabix, the resulting InDel data files were compressed using bgzip [35].

2.2.4. Identification of orthologous gene

Based on the longest protein sequences annotated for each gene in the 39 soybean genomes, orthologous genes were identified using OrthoFinder with the parameter "-f" [36].

2.2.5. GO and KEGG analysis

Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) annotations were conducted for all protein-coding genes in each soybean genome using eggNOG 5.0 with the default settings [37]. clusterprofiler 3.14 was utilized for GO and KEGG enrichment analysis [38].

2.3. Construction of the SoybeanGDB database

SoybeanGDB was built based on R/Shiny with the help of various other R packages (such as shinyWidgets, Shinycssloaders, Shinydashboard, Shinysky, data.table, ggplot2, IRanges, S4Vectors, Biostrings, and corrplot) (Table S2), which had been successfully applied in previous studies [33,39–41]. The structure of SoybeanGDB consisted of two primary R scripts: ui.R and server.R, along with several auxiliary R scripts. The ui.R defined the graphical interface of SoybeanGDB and collected user inputs that were then conveyed to the server side. The server.R conducted calculations on the server side and displayed the output on the graphical interface of SoybeanGDB. The R/Shiny framework allowed for effective data organization, user-friendly interface, and data visualization in SoybeanGDB. To ensure compatibility and accessibility, all functionalities of SoybeanGDB were extensively tested on five popular Internet browsers (Google Chrome, Apple Safari, Firefox, 360 Browser, and Microsoft Edge) across different Operating Systems (Windows, Linux, and MacOS).

2.4. Deployment of the SoybeanGDB database on a Linux web server

SoybeanGDB was deployed on a Linux web server running CentOS release 7.9.2009 as the operating system. The Linux web server was equipped with eight Intel processors (Intel(R) Xeon(R) Platinum 8255 C CPU @ 2.50 GHz) and 32 GB of random access memory (RAM).

3. Results

SoybeanGDB is a freely accessible public database that offers versatile functionalities for conducting functional genomics studies, including genome/transcriptome search, SNPs/InDels search and various tools (Fig. 1). To enhance the usability of SoybeanGDB, details and hyperlinks for the 39 genomes and two variation datasets from the 2898 soybean accessions and 481 accessions are provided under the "Accessions" menu. Additionally, a comprehensive tutorial of SoybeanGDB is available under the "Help" menu.

3.1. Searching and browsing 39 high-quality de novo assembled soybean genomes for protein-coding genes and other genomic features

A total of 39 high-quality genomes were incorporated into SoybeanGDB, including five wild soybeans, nine landraces, 19 improved cultivars, and six perennial *Glycine* species (Table S1). Six functionalities were developed and implemented within the "Genomes" menu of SoybeanGDB, enabling users to retrieve genes, transposable elements, and other genome features from any one of the 39 soybean genomes (Fig. 2).

Firstly, users can query any one of the 39 soybean genomes using a single gene ID, a single genomic region, or multiple gene IDs. The main panel of the output page displays gene(s) or genomic region in details, including gene structure, gene annotation, sequences (gene, CDS, cDNA, and protein), and transposable elements (Fig. 2A, B). SoybeanGDB contains a total of 47,142,361 annotated transposable elements across the 39 soybean genomes (Table S3). The download buttons in the uppermost panel of the output page allow users to export the results in plain text, table, or graph format. Furthermore, the "Genomic distribution of location" page allows users to visualize the chromosomal distribution of user-input genes from any one of the 39 soybean genomes using the circlize R package (Fig. 2C) [42].

A submenu is available in SoybeanGDB for searching and analyzing transcription factors (TFs) and transcriptional regulators (TRs) in any one of the 39 genomes (Fig. 2D). A total of 139,364 TFs from 69 families and 29,578 TRs from 24 families were identified in the 39 genomes and further classified using iTAK (Table S4) [30]. Among them, *Glycine syndetika* and Zhonghuang 13 had the lowest number of TFs (2486) and TRs (579), while *Glycine dolichocarpa* possessed the highest number of TFs (4705) and TRs (1361) (Table S4). One possible reason for the largest number of TFs and TRs in *Glycine dolichocarpa* might be that it is an allotetraploid (2 n = 4x =80) resulting from the hybridization between *Glycine syndetika* (2 n = 40) and *Glycine tomentella* D3 (2 n = 40). To facilitate easy access and output, details on TFs or TRs of interest, including description information, sequences, and annotations, are provided.

Synteny is critical for exploring the evolutionary relationships among various genomes or genes. To facilitate the investigation of evolutionary patterns between different soybean accessions or gene families, we utilized SyRI to identify syntenic blocks between Zhonghuang 13 (or Williams 82) and the remaining 37 soybean genomes[31]. A summary of the syntenic regions is provided in Table S5. A user-friendly interface is provided for searching, browsing, and downloading the syntenic regions between any two soybean genomes (Fig. 2E). Protein-coding genes located within a syntenic region can be easily retrieved from SoybeanGDB.

In addition to identifying syntenic regions, SyRI facilitate the detection of genomic structural variations, including duplicated regions, inverted regions, translocated regions, tandem repeats, deletions, and insertions, between two soybean genomes (Table S5). To enhance the retrieval of these structural variations between Zhonghuang 13 (or Williams 82) and the other 37 soybean genomes, a submenu was implemented (Fig. 2F). Furthermore, the output page provides information on protein-coding gene(s) associated with a structural variation, allowing users to download the data for downstream analysis.

3.2. Browsing 39 high-quality de novo assembled soybean genomes using JBrowse 2

Except for searching by gene IDs or genomic locations, the 39 high-quality genomes in SoybeanGDB can be explored with the help of JBrowse 2. JBrowse 2 is a full-featured and interactive genome browser widely utilized for the rapid visualization and exploration of large-scale genomic datasets [43]. For each of the 39 genomes, diverse tracks were deployed to facilitate browsing of genomic features, including genome sequences, protein-coding gene models, GO/KEGG annotations of protein-coding genes, and transposable elements (Fig. 3A-C). Researchers can selectively choose one or multiple tracks to view the genomic features of interest. Upon selecting an element within a track, the corresponding information is displayed in the right panel of the genome browser. To enable comparative analysis between Zhonghuang 13 and the other 38 genomes, SoybeanGDB incorporated the Gbrowse_synteny tool, which allows all syntenic regions with their corresponding annotations to be easily displayed in parallel (Fig. 3A). For the genome of Zhonghuang 13, additional tracks are available for browsing SNPs and InDels among 2898 soybean accessions, allowing researchers to explore their effects on gene structures (Fig. 3D).

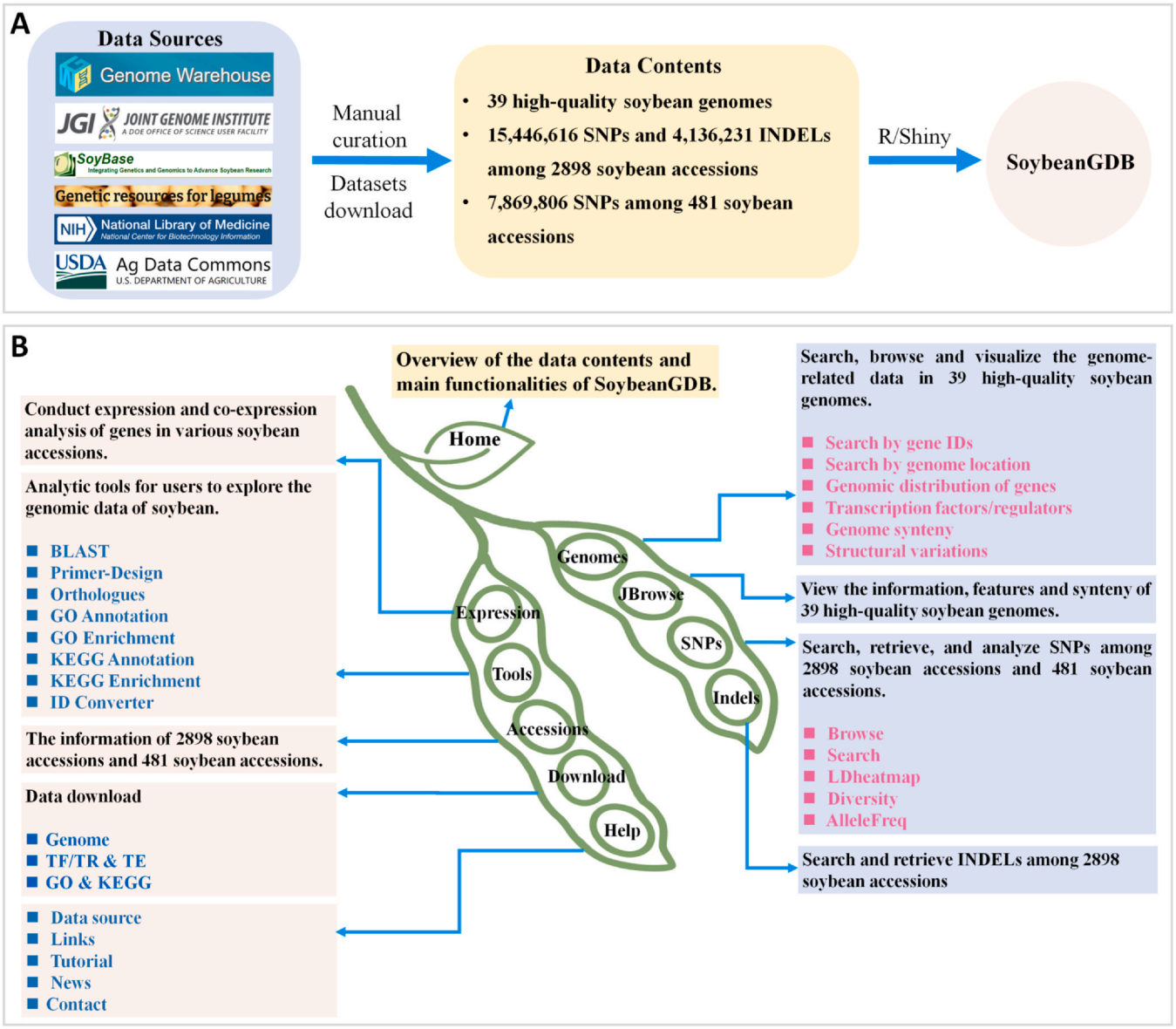

Fig. 1. Overview of the SoybeanGDB database. (A) The workflow for data collection and processing to build the SoybeanGDB database. (B) The framework and major functions of SoybeanGDB.

3.3. Searching, browsing, and analyzing SNPs among thousands of soybean accessions

SNPs represent the most prevalent form of genetic variation associated with complex phenotypes. To facilitate the exploration of phenotypic variations in soybean, a total of 15,446,616 high-quality SNPs among 2898 soybean accessions mapped in Zhonghuang 13, as well as 7,869,806 SNPs among 481 soybean accessions mapped in Williams 82, were integrated into the SoybeanGDB database (Table S6; Table S7). Among the 2898 soybean accessions, intergenic regions accounted for approximately 84.15% of the SNPs, while introns of genes accounted for 5.63% of the SNPs. Similarly, in the 481 accessions, intergenic regions accounted for about 81.45% of the SNPs, while introns accounted for 9.88% of the SNPs. To aid users in exploring the impact of SNPs on soybean genetic regulations, the "SNPs" menu provides two datasets and five useful tools for users (Fig. 4).

(i) Browse: SNPs sites within a gene or genomic region of interest can be visualized in the main panel of the SNP browser. These SNPs sites are represented by inverted triangles in various colors (Fig. 4A). The position of each inverted triangle is determined by the X-axis (chromosome positions of SNP sites) and the Y-axis (a random value). To construct the SNP browser, we employed the ggplot2 and plotly R packages [44,45]. The browsed results can be further filtered by selecting a subset of soybean accessions or by setting the mutation effect of SNPs. The final results can be output as either a plain text file or a static PDF file.

(ii) Search: in this interface, a table summarizing the genotype across SNP sites in a user-input gene or genomic region among the selected soybean accessions is shown in the main panel (Fig. 4B). The results, which comprise essential information about SNP sites, genotypes across the retrieved SNP sites, and gene annotations, can be downloaded as TXT files for further analysis using other tools.

(iii) LDheatmap: in genetic studies of agronomic traits, linkage disequilibrium (LD) between adjacent SNP sites is commonly visualized as a heatmap to define LD blocks. In this submenu, the LDheatmap R package is used to visualize pairwise linkage

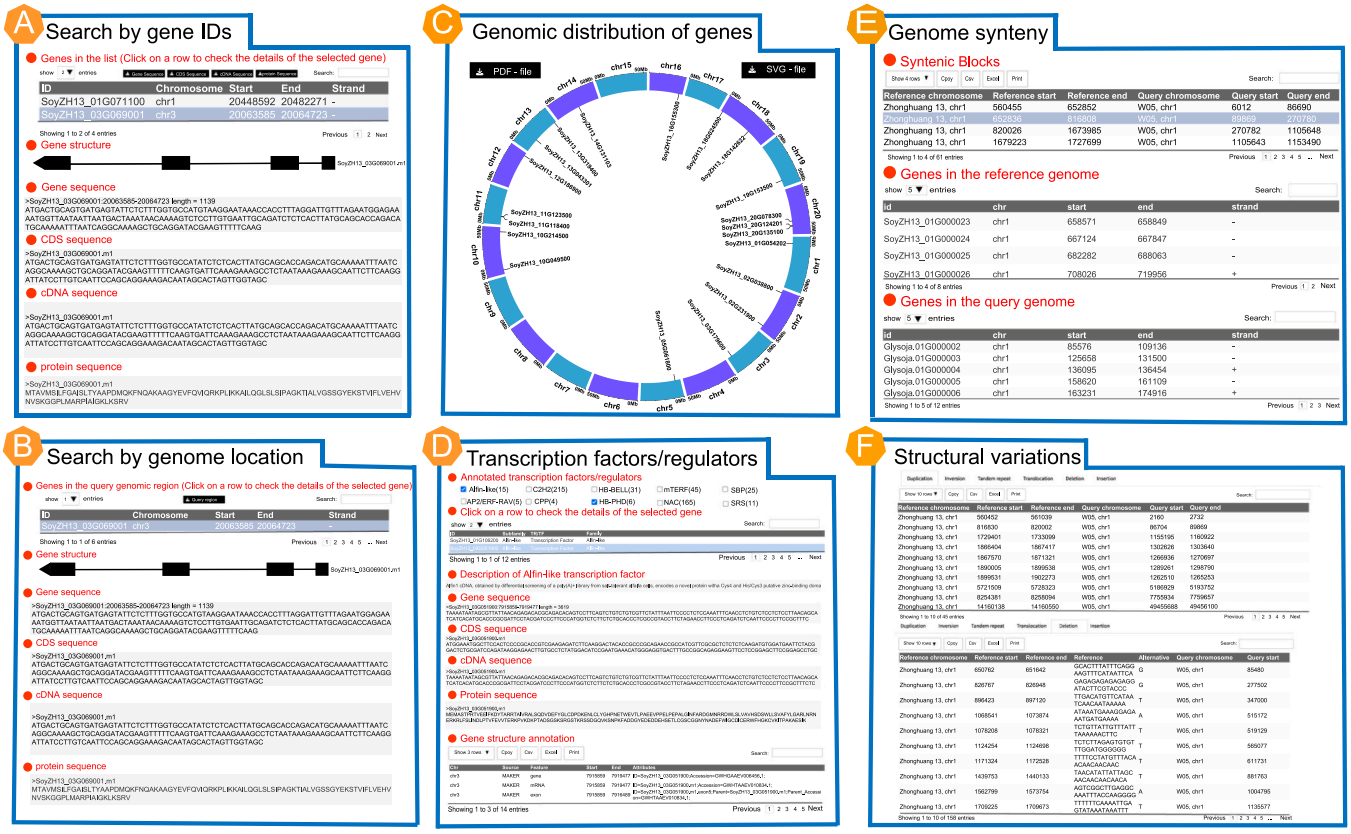

Fig. 2. Browse and search genomic features of 39 high-quality de novo assembled soybean genomes. (A) Search any one of the 39 soybean genomes by one or multiple gene IDs. (B) Search any one of the 39 soybean genomes by a single genomic region. (C) Visualize the distribution of user-input genes in any one of the 39 soybean genomes. (D) Browse the transcription factors/transcriptional regulators identified in any one of the 39 soybean genomes. (E) Browse syntenic regions between two soybean genomes. (F) Browse structural variations identified between two soybean genomes.

disequilibrium between SNP sites [46]. The linkage disequilibrium between pairwise SNP sites in a user-input gene or genomic region is calculated and displayed as a triangular correlation heatmap (Fig. 4C). Moreover, the SNP sites used in this analysis can be screened by the mutation effect of SNPs or the soybean accessions. Several options with graphical interfaces are provided to customize the appearance of the heatmap, including figure flipping and color setting. The resulting heatmap can be downloaded as a PDF or SVG file.

(iv) Diversity: nucleotide diversity among populations of different ecotypes serves as an important indicator of artificial selection during domestication. The pegas R package was employed to calculate the nucleotide diversity between different soybean ecotypes from the 2898 accessions or 481 accessions [47]. The user-input genomic region is divided into non-overlapping windows with 10 SNPs, and the sequence diversity for the chosen soybean ecotypes is then calculated (Fig. 4D). Finally, the results are visualized as line charts using the ggplot2 R package. Export options include images (PDF or SVG files) or plain texts (TXT file).

(v) AlleleFreq: this interface provides functionality to calculate and display the allele frequency of user-input SNP sites across different soybean ecotypes, using either 2898 accessions or 481 accessions (Fig. 4E). Each input SNP site should be represented by a 10-digit integer (e.g., 0133024709), where the first two digits denote the chromosome ID and the remaining eight digits indicate the genomic position of each SNP site in the genome of Zhonghuang 13. The allele frequency of user-input SNP sites will be visualized as pie charts using this module of SoybeanGDB, demonstrating the selection of favorable alleles in different soybean subgroups. The calculated allele frequency can be downloaded as a plain text file or as an image in PDF or SVG format.

3.4. Searching and retrieving InDels among 2898 soybean accessions

Insertions and deletions (InDels) are another significant form of genomic variation that has important implications for gene functions. A total of 4,136,231 high-quality InDels were identified among 2898 soybean accessions mapped in Zhonghuang 13 and deposited in the SoybeanGDB database (Table S6). On the “INDELs” page, users can efficiently search for InDels among selected soybean accessions using a gene ID or a genomic region (Fig. 4F). All InDels located in the user-input genomic region will be presented in a table in the main panel of the output page. Furthermore, the results can be downloaded in CSV, Excel, and TXT formats for downstream analysis.

3.5. Gene expression analysis

The expression data of Zhonghuang 13, A81–356022, W05 and 102 soybean accessions, obtained from previous studies, were integrated into SoybeanGDB to investigate gene expression during soybean development [17,25–27]. For Zhonghuang 13, a total of 27 samples from cotyledons, roots, stems, leaves, flowers, pods, pod& seeds and seeds at various developmental stages were included. A81–356022 contributed 14 samples from roots, nodules, leaves, flowers, pods, pod-shells at two developmental stages, and seeds at seven developmental stages. Additionally, three tissues from W05 and leaves from 102 accessions were also included. The expression levels (FPKM) of user-input genes can be retrieved as a table and visualized as a heatmap (Fig. 5A).

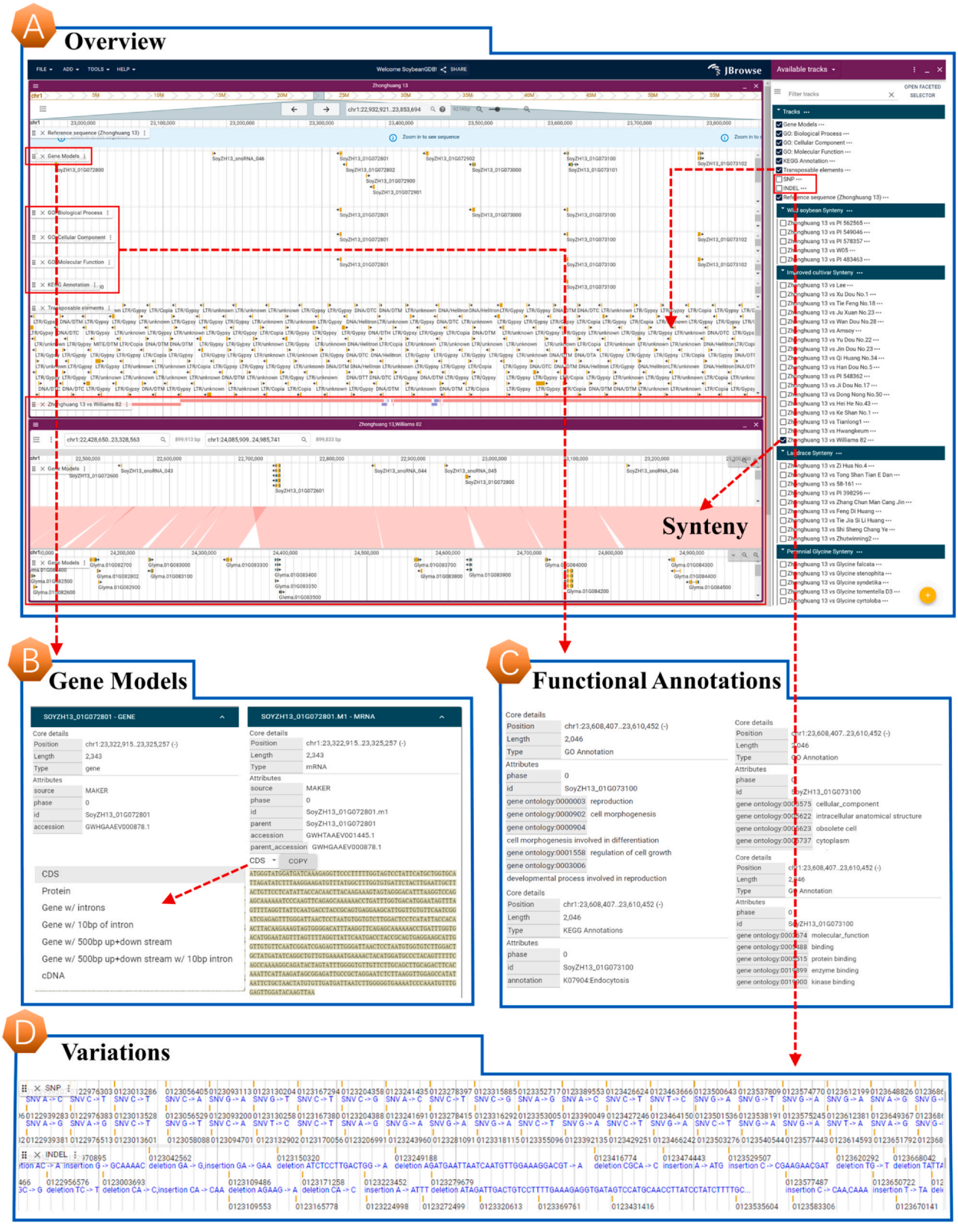

Fig. 3. The JBrowse module of SoybeanGDB. (A) JBrowse 2 visualization of chr1:22932921–23696811 in the Zhonghuang 13 genome. Six tracks are displayed, including one for protein-coding genes (B), three for GO annotations (C), one for KEGG annotations, one for transposable elements, and one for synteny. (D) Two tracks for SNPs and InDels in the Zhonghuang 13 genome are hidden.

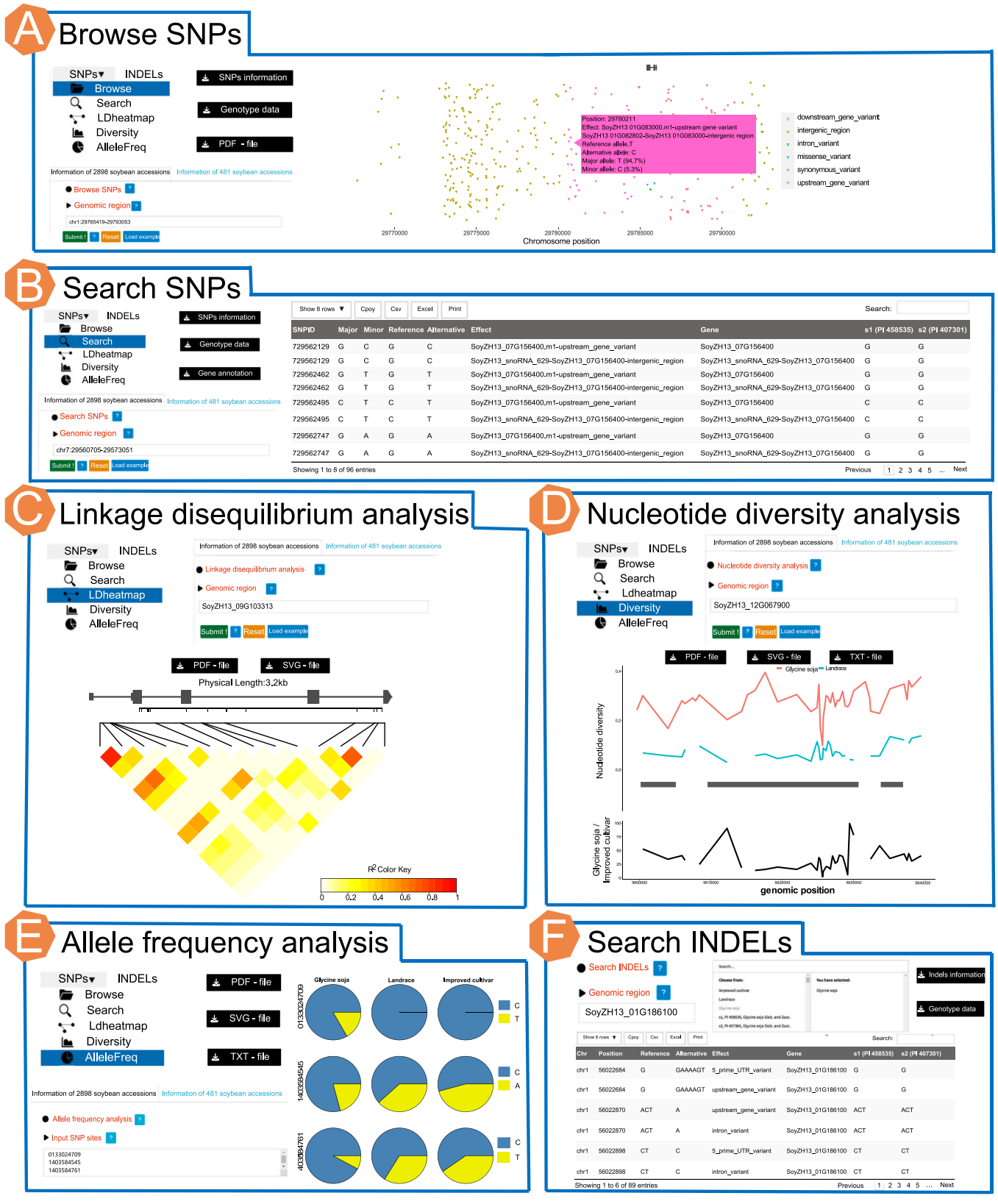

Fig. 4. Functional modules for the analysis of SNPs and InDels among thousands of soybean accessions. (A) Browse and visualize SNPs in a user-input genomic region. (B) Search for SNPs in a user-input genomic region. (C) Linkage disequilibrium analysis of SNPs. (D) Nucleotide diversity analysis of SNPs. (E) Allele frequency analysis of SNPs. (F) Search InDels in a user-input genomic region.

Co-expression analysis is commonly used to disclose potential gene functions in various organisms, as transcription-associated genes are often functionally related [48,49]. To enhance the biological understanding of gene functions in soybean, SoybeanGDB provides a functionality to calculate the correlation coefficient of expression levels using a list of user-input genes. The correlation

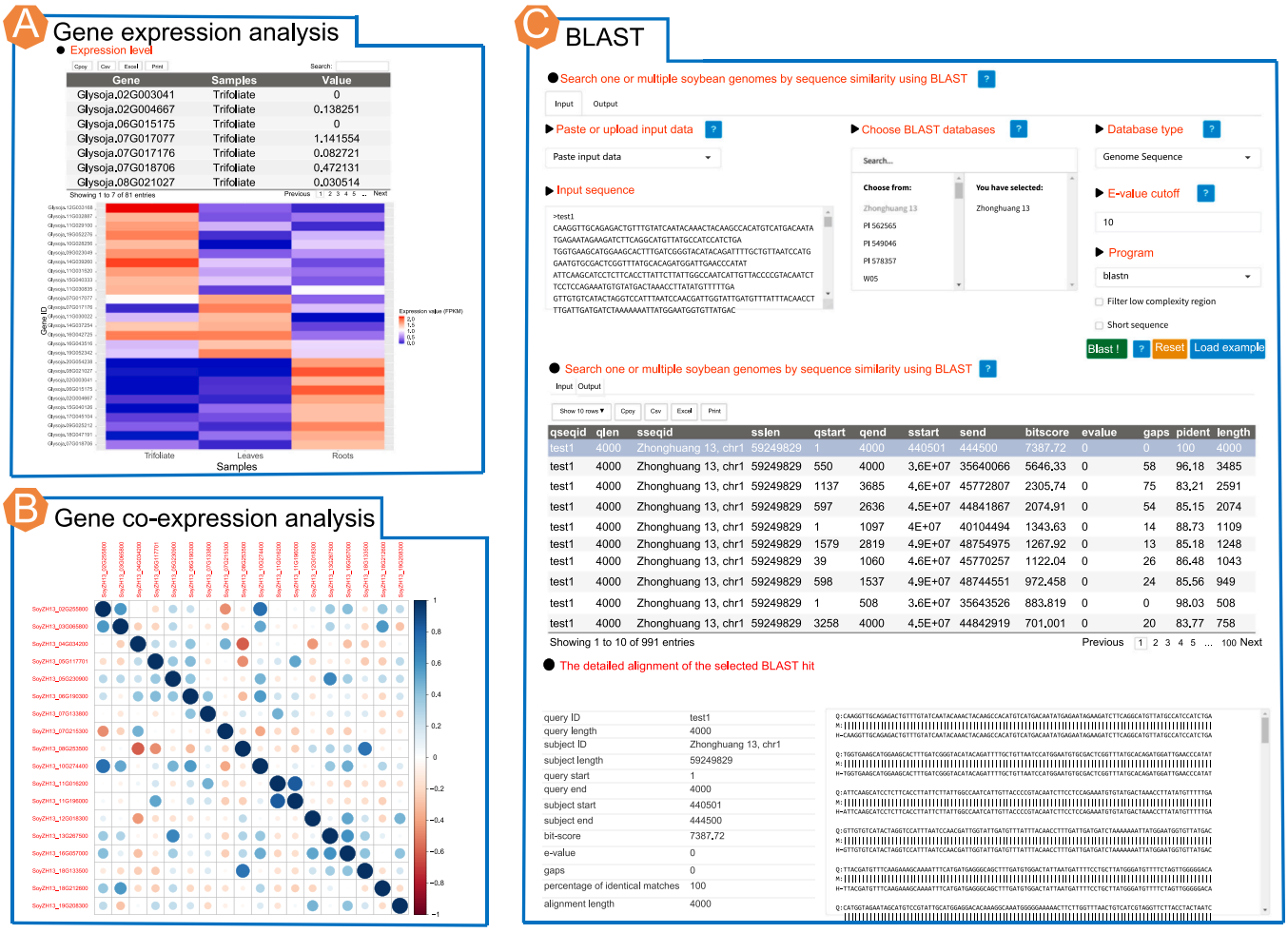

Fig. 5. Tools for gene expression analysis and BLAST alignment implemented in SoybeanGDB. (A) Browse and visualize the expression level of genes in different tissues and stages of Zhonghuang 13. (B) Calculate and visualize the expression correlations between user-input genes of Zhonghuang 13. (C) Search one or multiple soybean genomes by sequence similarity utilizing BLAST.

coefficients can be visualized as a table or a heatmap (Fig. 5B), and downloaded for further analysis.

3.6. Tools for functional genomic studies of soybean

SoybeanGDB also provides a suite of popular bioinformatics tools (Fig. 5C, Fig. 6 A-D).

3.6.1. BLAST

BLAST was implemented in SoybeanGDB as a graphical interface for searching homologous genes or sequences in one or more of the 39 high-quality de novo assembled soybean genomes (Fig. 5C) [50]. In the Input panel, users can conveniently submit query sequences (e.g., genome sequence, gene sequence, protein sequence, CDS sequence) in FASTA format. These input sequences can be utilized to search for homologs using blastn, tblastn, or tblastx. The BLAST databases for all 39 soybean genomes were constructed in SoybeanGDB. One or multiple query sequences can be submitted for each search. The BLAST results can be viewed in the Output panel and downloaded for further analysis.

3.6.2. Primer design

This module aims to assist users in designing PCR primers that flank one or multiple SNPs or InDels. The primers can be used as molecular markers to genotype one or multiple soybean accessions at the targeting SNPs or InDels (Fig. 6A). Primer3 was utilized as the backend engine [51]. The graphical interface of this module provides various options to set parameters of Primer3. The main panel displays five pairs of designed primers in a table, which can be easily downloaded. Moreover, the positions of the most suitable candidate primers, as well as the SNPs and InDels sites on the templates, are shown below the table.

3.6.3. Orthologous gene search

This tool was designed to facilitate the fast and accurate retrieval of orthologous genes among the 39 high-quality soybean genomes identified by OrthoFinder (Fig. 6B) [36]. Given a single or multiple user-input gene ID(s) from any one of the 39 genomes, the corresponding orthologs in the remaining 38 genomes are displayed in a table or a heatmap. The results can be easily downloaded as a CSV or Excel file.

3.6.4. Gene set annotation and enrichment analysis

GO and KEGG are two major annotation systems widely used for the investigation of gene functions in protein-coding genes [52,53]. All protein-coding genes in the 39 soybean genomes underwent annotation using the EggNOG database, which provided GO terms and KEGG pathways [37]. As a result, a total of 462,505 GO terms were associated with 1,005,057 genes, and 5237 KEGG pathways were identified for 472,983 genes in SoybeanGDB (Table S8). In this module, users can enter a meaningful gene set derived from biological experiments and bioinformatics analysis. For a given gene set,

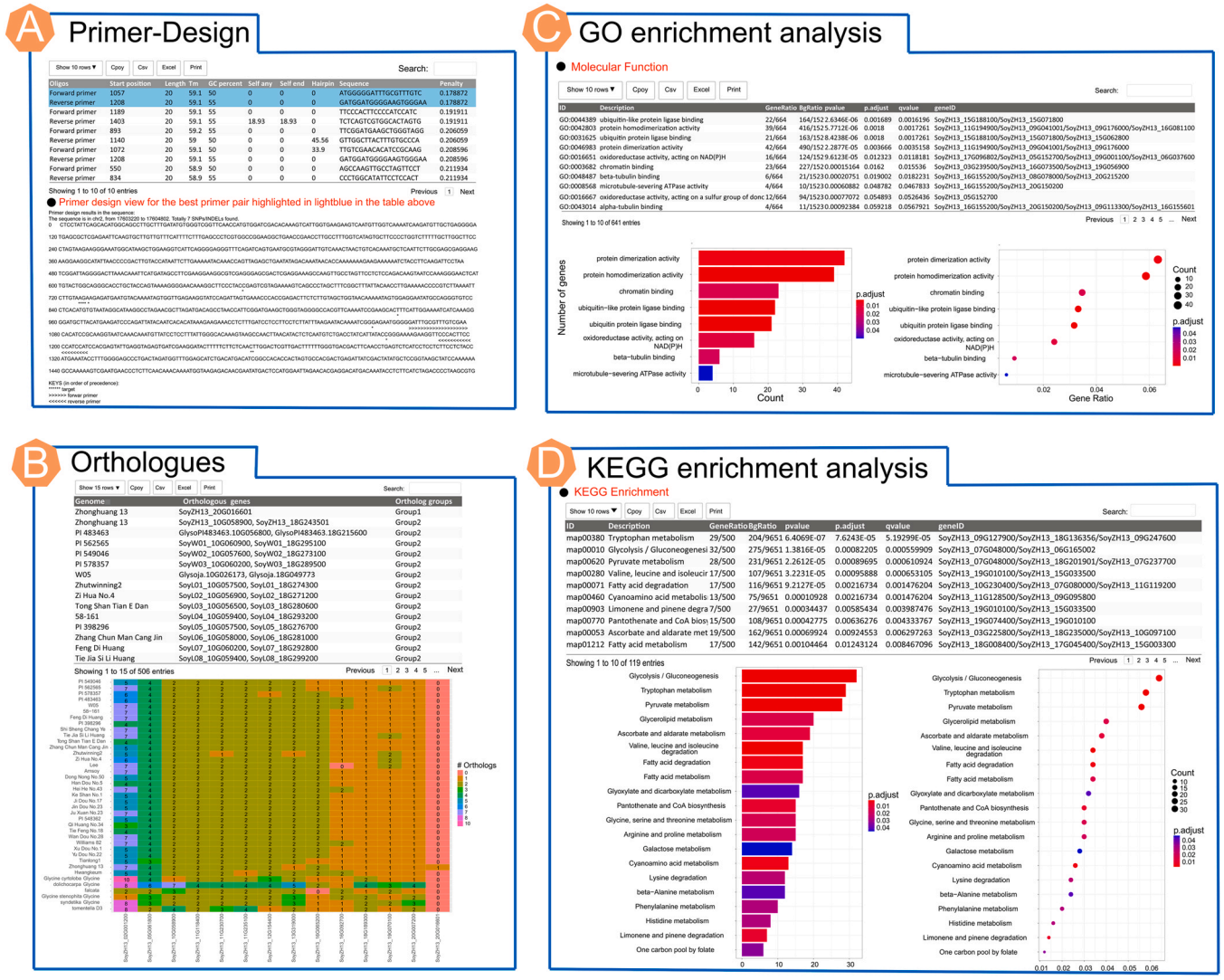

Fig. 6. Bioinformatics tools implemented in SoybeanGDB. (A) Design primers based on the genome of Zhonghuang 13 targeting SNPs and InDels in a user-input genomic region or gene locus. (B) Search and retrieve orthologous genes among the 39 soybean genomes. (C) GO enrichment analysis of a list of input genes in any of the 39 soybean genomes. (D) KEGG enrichment analysis of a list of input genes in any of the 39 soybean genomes.

the GO and KEGG annotations can be easily retrieved from SoybeanGDB in the form of tables and visualized as bar plots. In addition, we provide functional modules that allow users to perform GO or KEGG enrichment analysis on a user-input gene set from any one of the 39 soybean genomes, utilizing clusterProfiler (Fig. 6C,D) [38]. For visualization, only the 30 largest GO terms/KEGG pathways and the top 30 significant enrichment results are displayed in the bar plot.

3.7. Data download

All processed data, including genomic sequences, CDS/cDNA/protein sequences, gene annotations, transposable elements, transcription factors/transcription regulators, as well as GO/KEGG annotations for any one of the 39 soybean genomes, can be downloaded from the download page of SoybeanGDB. These data are stored in a compressed gzip file format, to accelerate downloading.

3.8. A case study of the SoybeanGDB database

We used *GmPRR3b* (*SoyZH13_12G067700*) as an example to illustrate the functionalities of SoybeanGDB. *GmPRR3b*, encoding a PSEUDO-RESPONSE-REGULATOR protein, has been reported to play a regulatory role in soybean flowering and maturity across different geographical areas during domestication [54]. Firstly, we investigated the nucleotide diversity in the genomic region of *GmPRR3b* among soybean accessions of different ecotypes using the “SNPs -> Diversity” module of SoybeanGDB. By inputting the gene ID “SoyZH13_12G067700″ and setting relevant parameters, we observed a significant reduction in the nucleotide diversity in the genomic region of *GmPRR3b* among the 1048 landraces compared to wild soybeans, which approached zero among the 1747 improved cultivars (Fig. 7A). A specific SNP site (chr12, 1205587110) in GmPRR3b was found to undergo a change from C to T, resulting in the production of a truncated protein that serves as a key variant influencing the functional divergence of *GmPRR3b* [54]. By using the “SNPs -> AlleleFreq” module of SoybeanGDB and inputting the SNP site(s) (e.g., 1205587110) from *GmPRR3b*, we observed a significant decrease in the ratio of T during the domestication and improvement process of soybean, with the ratio being 91.89% in wild soybeans, 15.17% in landraces, and 2.40% in improved cultivars (Fig. 7B). These findings provided strong evidence supporting the selection of *GmPRR3b* during the domestication and improvement of soybean, aligning with a previous study [54]. Next, we investigated the

Fig. 7. An example to demonstrate the application of SoybeanGDB. (A) Nucleotide diversity of *GmPRR3b* among soybean accessions of different ecotypes (wild soybean, landrace, and improved cultivar) calculated based on SNPs. (B) The spectrum of allele frequencies at six SNP sites of *GmPRR3b* in different ecotypes. The red box indicates the key SNP site. (C) Expression pattern of *GmPRR3b* at different tissues and developmental stages. (D) Orthologous genes of *GmPRR3b* identified in the 39 soybean genomes. I–IV represent wild soybeans, landraces, improved cultivars, and perennial *Glycine* species, respectively.

expression pattern of *GmPRR3b* across 27 samples using the "Expression -> Gene expression analysis" module of SoybeanGDB. By inputting the gene ID "*SoyZH13_12G067700*", we found that the expression level of *GmPRR3b* peaked during the flower bud differentiation stage (flo-1) and subsequently reduced to a stable level during the flower development process (flo-2 to flo-5) (Fig. 7C). Finally, the orthologous genes of *GmPRR3b* in the 39 soybean accessions were searched using the "Tools -> Orthologues" module by inputting the gene ID "*SoyZH13_12G067700*". Two orthologous genes of *GmPRR3b* were identified in 33 soybean accessions, including five wild soybeans, eight landraces, nineteen improved cultivars, and one perennial *Glycine* species (Fig. 7D). Only a single copy of *GmPRR3b* was identified in the genomes of *Glycine syndetika* and a soybean landrace (Zhutwinning2). Due to tandem duplication, three or more orthologous genes of *GmPRR3b* were found in the other four genomes of perennial *Glycine* species. The above cases demonstrate the feature-rich capabilities of SoybeanGDB and its easy-to-use analytic tools.

4. Discussion and conclusion

With the rapid advancements in genomics studies and sequencing technologies, numerous biological databases and web applications have been developed in the field of plant research [8,9,55]. Soybean, which originated in China and encompasses over 60,000 accessions, is a key crop utilized worldwide for vegetable oil and protein feed production [4]. Recent advances in soybean research have demonstrated the vital importance of large-scale omics data from massive soybean accessions in gene function analysis, population genetic analysis, and breeding of new varieties [4–7]. To this end, we constructed the SoybeanGDB database using R/Shiny [39]. Different from the datasets in existing soybean databases [11–14], SoybeanGDB currently hosts the largest number of genomic data and genomic variations, while also providing a variety of easy-to-use bioinformatics analytic tools as graphical interfaces.

SoybeanGDB is particularly valuable for researchers without a background in bioinformatics. Firstly, users can easily and efficiently

search and browse genomic information using the “Genomes” and “JBrowse” models. The gene expression and co-expression analysis of gene(s) can be obtained using the “Expression” module. Moreover, SoybeanGDB provides significant genetic variation resources including SNPs and Indels, which can be used to identify candidate variations/genes and develop molecular markers for soybean breeding. For example, users can mine soybean accessions that possess beneficial alleles of candidate variations/genes and utilize these accessions as donors in breeding, using the “SNPs”, “INDELs” and “Tools -> Orthologues” models. Users can further utilize the “Primer-Design” module to design molecular markers for genotyping multiple soybean accessions in molecular breeding.

In summary, SoybeanGDB was developed by integrating invaluable soybean data, which included high-quality de novo genome assemblies of 39 soybean accessions, and high-quality genomic variations. These variations comprised 15,446,616 SNPs and 4,136,231 InDels among 2898 soybean accessions mapped to the Zhonghuang 13 genome, as well as 7,869,806 SNPs among 481 soybean accessions mapped to the Williams 82 genome. To facilitate the utilization of the collected data, SoybeanGDB incorporated a variety of easy-to-use bioinformatics analytic tools presented as graphical interfaces. In addition, we plan to regularly update the SoybeanGDB database by incorporating the newest available resources and developing new tools, to meet the ever-growing needs of the soybean community. We hope that SoybeanGDB will make a significant contribution to global soybean research.

CRediT authorship contribution statement

W.Y., Y.L. and Yihan Wang conceived and designed this study; H.L., W.Y., Y.L., Yihan Wang, T.C., L.J., Z.W., J.L., Yazhou Wang, M.F., M.C., Yuping Wang, F.H., Y.J., T.L. and Z.Z. collected and analyzed the data; H.L., W.Y., Y.L., Yihan Wang, T.C. and L.J. constructed the database with help from Z.W., J.L., Yazhou Wang, M.F., M.C., Yuping Wang, T.C., F.H., Y.J., T.L. and Z.Z.; Y.L., W.Y. and Yihan Wang wrote the paper.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (31900451; 32101745), the research start-up fund to topnotch talents of Henan Agricultural University (30500581), the research start-up fund to young talents of Henan Agricultural University (30500941), the Scientific and Technological Research Project of Henan Province (212102110243), and the Natural Science Foundation of Henan Province (232300420011).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.06.012.

References

[1] Guo Z, Liu X, Zhang B, et al. Genetic analyses of lodging resistance and yield provide insights into post-Green-Revolution breeding in rice. Plant Biotechnol J 2021;19(4):814–29.
[2] Li J, Yuan D, Wang P, et al. Cotton pan-genome retrieves the lost sequences and genes during domestication and selection. Genome Biol 2021;22(1):119.
[3] Qin P, Lu H, Du H, et al. Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. Cell 2021;184(13):3542–58. e16.
[4] Liu Y, Du H, Li P, et al. Pan-genome of wild and cultivated soybeans. Cell 2020;182(1):162–76. e13.
[5] Zhou Z, Jiang Y, Wang Z, et al. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat Biotechnol 2015;33(4):408–14.
[6] Kou K, Yang H, Li H, et al. A functionally divergent SOC1 homolog improves soybean yield and latitudinal adaptation. Curr Biol 2022;32(8):1728–42.
[7] Valliyodan B, Brown AV, Wang J, et al. Genetic variation among 481 diverse soybean accessions, inferred from genomic re-sequencing. Sci Data 2021;8(1):50.
[8] Portwood II JL, Woodhouse MR, Cannon EK, et al. MaizeGDB 2018: the maize multi-genome genetics and genomics database. Nucleic Acids Res 2019;47(D1):D1146–54.
[9] Zhao H, Li J, Yang L, et al. An inferred functional impact map of genetic variants in rice. Mol Plant 2021;14(9):1584–99.
[10] 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.
[11] Brown AV, Conners SI, Huang W, et al. A new decade and new data at SoyBase, the USDA-ARS soybean genetics and genomics database. Nucleic Acids Res 2021;49(D1):D1496–501.
[12] Grant D, Nelson RT, Cannon SB, et al. SoyBase, the USDA-ARS soybean genetics and genomics database. Nucleic Acids Res 2010;38(suppl_1):D843–6.
[13] Joshi T, Fitzpatrick MR, Chen S, et al. Soybean knowledge base (SoyKB): a web resource for integration of soybean translational genomics and molecular breeding. Nucleic Acids Res 2014;42(D1):D1245–52.
[14] Xiao Z, Wang Q, Li M-W, et al. Wildsoydb DataHub: a platform for accessing soybean multiomic datasets across multiple reference genomes. Plant Physiol 2022;190(4):2099–102.
[15] Schmutz J, Cannon SB, Schlueter J, et al. Genome sequence of the palaeopolyploid soybean. Nature 2010;463(7278):178–83.
[16] Shen Y, Liu J, Geng H, et al. De novo assembly of a Chinese soybean genome. Sci China Life Sci 2018;61(8):871–84.
[17] Shen Y, Du H, Liu Y, et al. Update soybean Zhonghuang 13 genome to a golden reference. Sci China Life Sci 2019;62(9):1257–60.
[18] Xie M, Chung CY-L, Li M-W, et al. A reference-grade wild soybean genome. Nat Commun 2019;10(1):1216.
[19] Valliyodan B, Cannon SB, Bayer PE, et al. Construction and comparison of three reference-quality genome assemblies for soybean. Plant J 2019;100(5):1066–82.
[20] Kim M-S, Lee T, Baek J, et al. Genome assembly of the popular Korean soybean cultivar Hwangkeum. G3 Genes Genomes Genet 2021;11(10). jkab272.
[21] Zhuang Y, Wang X, Li X, et al. Phylogenomics of the genus *Glycine* sheds light on polyploid evolution and life-strategy transition. Nat Plants 2022;8(3):233–44.
[22] Jia J, Ji R, Li Z, et al. Soybean DICER-LIKE2 regulates seed coat color via production of primary 22-nucleotide small interfering RNAs from long inverted repeats. Plant Cell 2020;32(12):3662–73.
[23] Chen M, Ma Y, Wu S, et al. Genome warehouse: a public repository housing genome-scale data. Genom Proteomics Bioinformatics 2021;19(4):584–9.
[24] Li C, Tian D, Tang B, et al. Genome variation map: a worldwide collection of genome variations across multiple species. Nucleic Acids Res 2021;49(D1):D1186–91.
[25] Severin AJ, Woody JL, Bolon Y-T, et al. RNA-Seq atlas of *Glycine max*: a guide to the soybean transcriptome. BMC Plant Biol 2010;10(1):160.
[26] Qi X, Li M-W, Xie M, et al. Identification of a novel salt tolerance gene in wild soybean by whole-genome sequencing. Nat Commun 2014;5(1):4340.
[27] Li D, Liu Q, Schnable PS. TWAS results are complementary to and less affected by linkage disequilibrium than GWAS. Plant Physiol 2021;186(4):1800–11.
[28] Ou S, Su W, Liao Y, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol 2019;20(1):275.
[29] Smit A., Hubley, R., Green, P. RepeatMasker Open-4.1.0. 2013–2015. Available from: wwwrepeatmaskerorg. 2019.
[30] Zheng Y, Jiao C, Sun H, et al. iTAK: A program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. Mol Plant 2016;9(12):1667–70.
[31] Goel M, Sun H, Jiao W-B, et al. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. Genome Biol 2019;20(1):277.
[32] Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 2018;34(18):3094–100.
[33] Zhou W, Wang L, Zheng W, et al. MaizeSNPDB: a comprehensive database for efficient retrieve and analysis of SNPs among 1210 maize lines. Comp Struct Biotechnol J 2019;17:1377–83.
[34] Cingolani P, Platts A, Wang LL, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms. SnpEff Fly 2012;6(2):80–92.
[35] Li H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics 2011;27(5):718–9.
[36] Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol 2019;20(1):238.
[37] Huerta-Cepas J, Szklarczyk D, Heller D, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res 2019;47(D1):D309–14.
[38] Yu G, Wang L-G, Han Y, et al. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 2012;16(5):284–7.
[39] Jia L, Yao W, Jiang Y, et al. Development of interactive biological web applications with R/Shiny. Brief Bioinform 2022;23(1). bbab415.
[40] Jia L, Li Y, Huang F, et al. LIRBase: a comprehensive database of long inverted repeats in eukaryotic genomes. Nucleic Acids Res 2022;50(D1):D174–82.

[41] Yao W, Huang F, Zhang X, et al. ECOGEMS: efficient compression and retrieve of SNP data of 2058 rice accessions with integer sparse matrices. Bioinformatics 2019;35(20):4181–3.
[42] Gu Z, Gu L, Eils R, et al. circlize implements and enhances circular visualization in R. Bioinformatics 2014;30(19):2811–2.
[43] Cain S, Haw R, Bridge C, et al. JBrowse 2: An extensible open-source platform for modern genome analysis. Cancer Res 2022;82(12_Supplement):6400.
[44] Inc. P.T. Collaborative data science. Montréal: Plotly Technologies Inc. 2015.
[45] Wickham H. Ggplot2: Elegant Graphics for Data Analysis. New York: Springer; 2009. p. 212.
[46] Shin J-H, Blay S, McNeney B, et al. LDheatmap: an R function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. J Stat Softw 2006;16(3):9.
[47] Paradis E. pegas: an R package for population genetics with an integrated–modular approach. Bioinformatics 2010;26(3):419–20.
[48] Movahedi S, Van Bel M, Heyndrickx KS, et al. Comparative co-expression analysis in plant biology. Plant Cell Environ 2012;35(10):1787–98.
[49] van Dam S, Võsa U, van der Graaf A, et al. Gene co-expression analysis for functional classification and gene–disease predictions. Brief Bioinform 2018;19(4):575–92.
[50] Ye J, McGinnis S, Madden TL. BLAST: improvements for better sequence analysis. Nucleic Acids Res 2006;34(suppl_2):W6–9.
[51] Untergasser A, Cutcutache I, Koressaar T, et al. Primer3-new capabilities and interfaces. Nucleic Acids Res 2012;40(15):e115.
[52] Ogata H, Goto S, Sato K, et al. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 1999;27(1):29–34.
[53] The Gene Ontology Consortium. The gene ontology resource: 20 years and still GOing strong. Nucleic Acids Res 2019;47(D1):D330–8.
[54] Lu S, Dong L, Fang C, et al. Stepwise selection on homeologous PRR genes controlling flowering and maturity during soybean domestication. Nat Genet 2020;52(4):428–36.
[55] Mansueto L, Fuentes RR, Borja FN, et al. Rice SNP-seek database update: new SNPs, indels, and queries. Nucleic Acids Res 2017;45(D1):D1075–81.