https://doi.org/10.1093/plphys/kiac266 PLANT PHYSIOLOGY 2022: 190: 226–237

Plant Physiology®

Breakthrough Tech

The Brassicaceae genome resource (TBGR): A comprehensive genome platform for Brassicaceae plants

Zhuo Liu [1], Nan Li,[1] Tong Yu [1], Zhiyuan Wang [1], Jiaqi Wang [1], Jun Ren [2], Jinghua He [1], Yini Huang,[1] Keqian Shi,[1] Qihang Yang,[1] Tong Wu,[1] Hao Lin [3] and Xiaoming Song [1,3,4,*,†]

1 School of Life Sciences, North China University of Science and Technology, Tangshan, Hebei 063210, China
2 Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing 100081, China
3 School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China
4 Food Science and Technology Department, University of Nebraska-Lincoln, Lincoln, Nebraska 68588, USA

*Author for correspondence: songxm@ncst.edu.cn
†Senior author

X.S. conceived the project and was responsible for the project initiation. X.S. and Z.L. supervised and managed the project and research. The data collection, bioinformatics analyses, and database construction were led by X.S., Z.L., N.L., T.Y., Z.W., J.W., J.H., Y.H., K.S., Q.Y., and T.W. The manuscript was organized, written and revised by X.S., Z.L., N.L., J.R. and H.L. All authors read and approved the manuscript.
The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (https://academic.oup.com/plphys/pages/general-instructions) is Xiaoming Song (songxm@ncst.edu.cn).

Abstract

The Brassicaceae is an important plant family. We built a user-friendly, web-based, comparative, and functional genomic database, The Brassicaceae Genome Resource (TBGR, http://www.tbgr.org.cn), based on 82 released genomes from 27 Brassicaceae species. The TBGR database contains a large number of important functional genes, including 4,096 glucosinolate genes, 6,625 auxin genes, 13,805 flowering genes, 36,632 resistance genes, 1,939 anthocyanin genes, and 1,231 m6A genes. A total of 1,174,049 specific guide sequences for clustered regularly interspaced short palindromic repeats and 5,856,479 transposable elements were detected in Brassicaceae. TBGR also provides information on synteny, duplication, and orthologs for 27 Brassicaceae species. The TBGR database contains 1,183,851 gene annotations obtained using the TrEMBL, Swiss-Prot, Nr, GO, and Pfam databases. The BLAST, Synteny, Primer Design, Seq_fetch, and JBrowse tools are provided to help users perform comparative genomic analyses. All the genome assemblies, gene models, annotations, and bioinformatics results can be easily downloaded from the TBGR database. We plan to improve and continuously update the database with newly assembled genomes and comparative genomic studies. We expect the TBGR database to become a key resource for the study of the Brassicaceae.

Introduction

Comprising approximately 4,000 species and 338 genera, Brassicaceae is a large family in the order Brassicales (Al-Shehbaz et al., 2006; Walden et al., 2020). Brassicaceae contains many important vegetable, oilseed, and feed crop species (Cheng et al., 2014; Song et al., 2021a, 2021b). Arabidopsis (*Arabidopsis thaliana*), an important model organism in plant biology, is also within this family. Brassicaceae provides an excellent system for studying genome evolution and polyploidy.

The "U's triangle" model consists of six widely cultivated Brassicaceae species, including three diploid species (Chinese

Received March 22, 2022. Accepted May 3, 2022. Advance access publication June 7, 2022

cabbage [*Brassica rapa*, AA, 2n = 2x = 20], Cabbage [*Brassica oleracea*, CC, 2n = 2x = 18], and Black mustard [*Brassica nigra*, BB, 2n = 2x = 16]) and three tetraploid species (mustard [*Brassica juncea*, AABB, 2n = 4x = 36], Rapeseed [*Brassica napus*, AACC, 2n = 4x = 38], and Ethiopian mustard [*Brassica carinata*, BBCC, 2n = 4x = 34]) (Nagaharu, 1935; Song et al., 2021a). Because of the ubiquity of *Brassica* "U's triangle" species as models for the study of polyploidization and genome hybridization, much progress has been made in comparative and functional genomic research on these *Brassica* species in recent years, and this has generated large amounts of omics data.

Since the genome of *A. thaliana* was sequenced in 2000, the genomes of several Brassicaceae species have been sequenced (Wang et al., 2011; Chalhoub et al., 2014; Liu et al., 2014; Parkin et al., 2014; Yang et al., 2016). Recently, we presented a high-quality and chromosome-level genome sequence of *B. carinata* (Song et al., 2021a). With continuously declining sequencing costs and improvements in bioinformatics analysis technology, the pan-genomes of several species of Brassicaceae, such as *B. rapa*, *B. oleracea*, and *B. napus*, have been analyzed; genus-wide pan-genome studies have also been conducted (Golicz et al., 2016; Song et al., 2020a; Bayer et al., 2021; Cai et al., 2021; He et al., 2021). Most genomes of Brassicaceae species have already been assembled and released, but more work is needed to mine these genomic datasets.

We built the "The Brassicaceae Genome Resource" (TBGR) database to make all genome sequences and annotated data of Brassicaceae accessible to the Brassicaceae research community. The purpose of the TBGR database is to provide a repository that can be used for comparative and functional genomics analyses of Brassicaceae species at the whole-genome scale. Here, we present an overview of the interfaces of the TBGR database, including the Browse, Charts, Search, Tools, Resources, and Download interfaces that we designed to help users analyze TBGR data. This database provides a convenient and useful tool that will promote Brassicaceae research.

Results

Overview of the main interfaces of the TBGR database

We collected genomic information resources from 82 public genomes of 27 Brassicaceae species (Supplemental Table S1). We then conducted a systematic bioinformatics analysis of these data, including gene annotation, synteny, duplication type, clustered regularly interspaced short palindromic repeats (CRISPR) guide sequences, transposable elements (TEs), homologous genes, transcription factors (TFs), N6-methyladenosine (m6A), and identification of the main functional genes. We identified many important functional genes in the Brassicaceae family, including glucosinolate, auxin, flowering, resistance, and anthocyanin genes. Finally, we built the TBGR database to help users easily query, compare, and download these genome resources and results of these bioinformatics analyses. Using these available datasets and related bioinformatics tools, the genome information was stored in backend tables of MySQL, which can be easily accessed by the frontend web application. Here, we provide a detailed description of the interfaces of the TBGR database, including the Browse, Charts, Search, Download, Tools, Resources, Help, and Contact interfaces (Figure 1).

Search interface

In this section, we detail the search function that can be used to obtain the gene annotation, synteny, guide sequences of CRISPR, duplication type, and homologous genes for 27 Brassicaceae species (Figure 2). We conducted the functional annotation for all genes of these species. Based on the four databases (Pfam, UniProt knowledgebase [Swiss-Prot, TrEMBL], nonredundant protein sequence database [Nr], and Gene Ontology [GO] database), 80.35% London rocket (*Sisymbrium irio*) to 99.98% Saltwater cress (*Eutrema salsugineum*) of the genes were annotated in different species (Supplemental Table S2). All of these annotations can be searched in the TBGR database using the gene id of each species. Furthermore, users can enter the gene symbol (e.g. *FLOWERING LOCUS C* [*FLC*]) or the accession number (e.g. GO:0005634 for GO database and PF07145 for Pfam database) of each functional database to realize the cross-species search function. Moreover, we added the external links for the genes of 27 representative genomes of Brassicaceae species. The external links included UniProt, National Center for Biotechnology Information (NCBI), and Comparative Genomics platform. In addition, we also linked the *A. thaliana* genes to The Arabidopsis Information Resource and *B. napus* genes to *Brassica napus* pan-genome information resource (BnPIR).

To clarify the evolutionary relationships of genes in Brassicaceae, we conducted a homologous gene analysis using OrthoFinder software. A total of 55,714 orthogroups were identified among 27 species, among which 13,226 groups belonged to the species-specific orthogroups (Supplemental Tables S3 and S4). In addition to displaying these groups and downloading sequences on our database, we also displayed the evolutionary trees for the genes of each group. Based on the orthologous genes, the phylogenetic tree of these species was constructed to uncover the evolutionary relationship of these Brassicaceae species (Figure 3).

To explore patterns of gene duplication or loss after whole-genome duplication events in Brassicaceae species, we performed genome syntenic analyses. The syntenic genes between any two or more species of Brassicaceae are provided in the TBGR database. We also provided an option to identify syntenic genes within a range (Flanking 10, 20, 30 genes) around a gene of interest. Furthermore, the five duplication types (dispersed, proximal, singleton, tandem, and whole-genome duplication [WGD]/segmental) were identified for each gene of 27 species (Figure 3; Supplemental Table S5). We found that the WGD/segmental duplication was dominant in the six Brassica U's triangle species, Camelina (*Camelina sativa*), and Maca (*Lepidium meyenii*). This phenomenon was

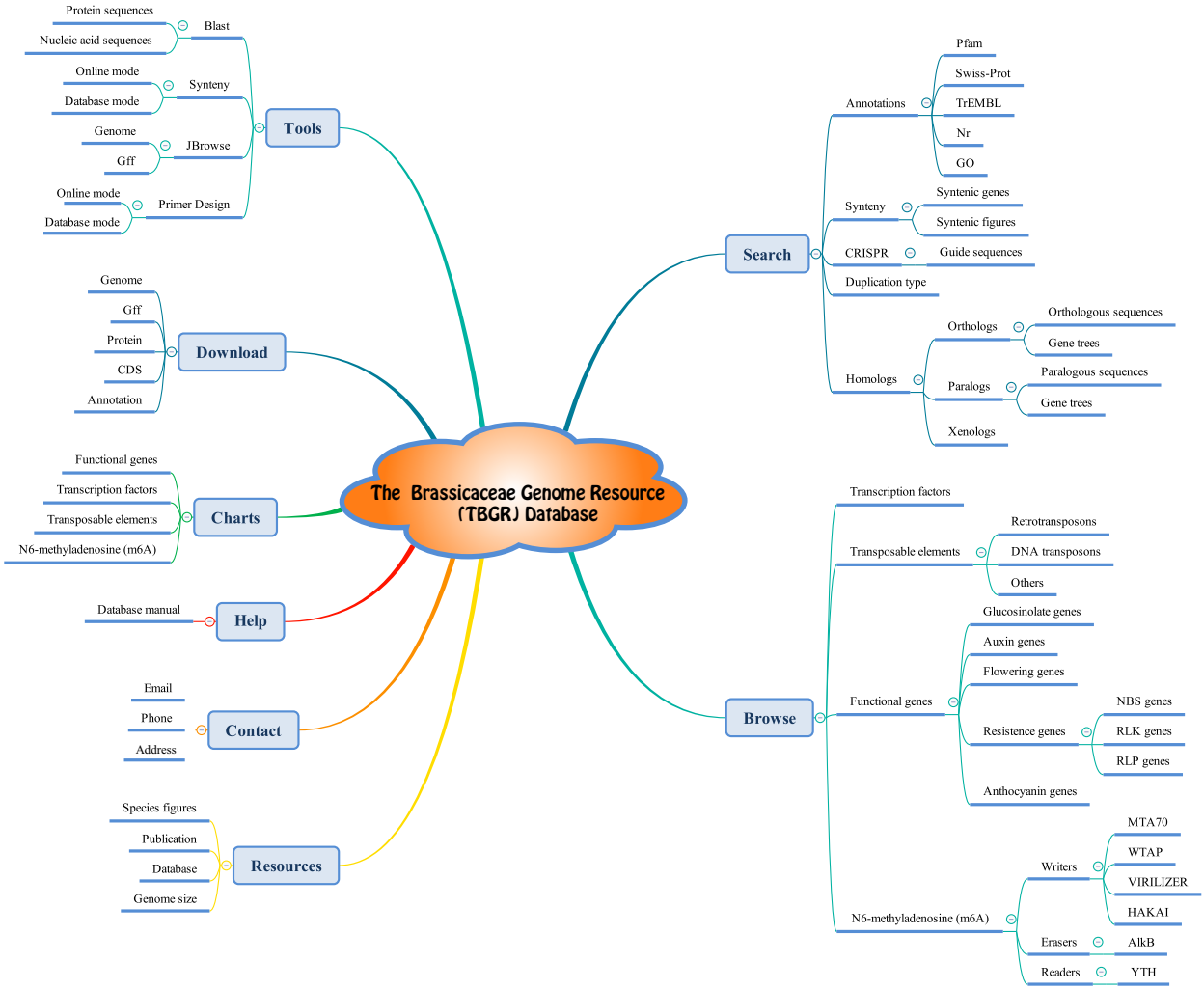

Figure 1 The architecture of "The Brassicaceae Genome Resource" (TBGR) database. The different colors represent the main interfaces of TBGR. Gff, general feature format; RLK, receptor like kinases; RLP, receptor like protein; MTA70, 70 kDa subunit of methyltransferase-A; AlkB, Alkylation repair protein-B; YTH, IYT521-B homology.

due to the fact that each Brassica species and *C. sativa* had one additional whole-genome triplication (WGT) event compared to other Brassicaceae species, and *L. meyenii* also had two additional WGD events (Kagale et al., 2014; Jeong et al., 2016; Zhang et al., 2016; Song et al., 2021a).

To facilitate CRISPR research on Brassicaceae species, we designed guide sequences for all genes of these species and provided the search function in our database (Figure 2). A total of 1,174,049 specific guide sequences were designed for genes of all species, and the success rate of guide sequence design ranges from 79.46% (*Schrenkiella parvula*) to 99.76% (*Capsella grandiflora*) in different species (Figure 3; Supplemental Table S6). Furthermore, the off-target sequences for each guide sequence were identified and deposited in our database.

Browse interfaces

We identified the TEs, TFs, m6A, and other important functional genes from the whole genomes of 27 Brassicaceae species (Figure 2). For user convenience, comprehensive information for all of these TEs and genes are provided according to family and species in the Browse interface.

Large numbers of TEs are present in plant genomes, and their genomic positions can change, which can lead to the appearance of new mutations, rewiring of the gene regulatory network, and even enlarging genome size (Lee and Kim, 2014; Naville et al., 2019; Nishihara, 2020). We annotated the TEs of the genomes of 27 Brassicaceae species and performed a detailed classification (Figure 3; Supplemental Figure S2). A total of 5,856,479 TEs were detected in all species, which were further divided into 4,155,956 retrotransposons, 1,012,983 DNA transposons, and 687,540 other TEs (Supplemental Table S7). Furthermore, we found that there was obviously a positive correlation between the expansion of the transposon and the genome size ($r = 0.79$) (Supplemental Figure S3). Therefore, these identified transposons provide rich resources for future research on the expansion of the genome.

TFs play an important role in plant development and stress responses, and they regulate downstream genes by binding to specific DNA sequences (Song et al., 2013, 2020a,

Figure 2 Overview of the main interfaces and internal features of the TBGR database including the Home, Browse, Charts, Resources, Search, Download, Tools, Help, and Contact interfaces.

2020b). A total of 85,220 TFs from 63 families were detected in all species examined (Figure 4A; Supplemental Table S8). The four families with the largest number of genes were Myeloblastosis_DNA-bind (MYB) (9,604), APETALA2/Ethylene-Responsive Factor (AP2/ERF) (6,087), nucleotide-binding site (5,746), and basic helix–loop–helix (5,331), which were markedly higher than the number of genes in other families. Comparing with *A. thaliana*, we found that nuclear TF, X-box binding 1 (NF-X1), and growth-regulating factor gene families were obviously expanded, while the signal transducer and activator of transcription gene family numbers were contracted in most species (Figure 4B; Supplemental Table S9). Interestingly, we found that compared to other Brassicaceae species, nearly all TF families were expanded in *C. sativa* and *L. meyenii* due to their additional WGT event and two WGD events, respectively (Kagale et al., 2014; Zhang et al., 2016) (Figure 4B). However, although an additional WGT event occurred in *Brassica* and *R. sativus*, most TF families did not expand substantially in the four diploid species (*B. nigra*, *B. rapa*, *B. oleracea*, and *R. sativus*). This result indicated that most genes might have been lost after genome duplication, a finding that was also consistent with previous reports (Wang et al., 2011; Song et al., 2021a). Among three tetraploid *Brassica* species (*B. napus*, *B. juncea*, and *B. carinata*), almost all TF families were expanded since they come from three diploid *Brassica* species that hybridized with each other (Chalhoub et al., 2014; Yang et al., 2016; Song et al., 2021a) (Figure 4B).

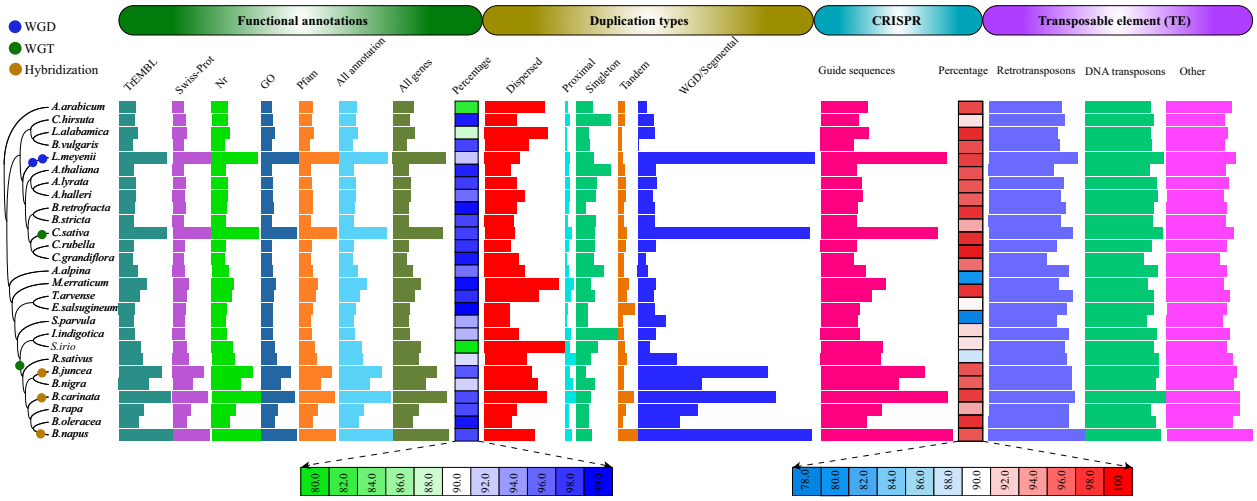

Figure 3 Bar plots of the number of functional annotations, duplication types, guide sequences of CRISPR, and TEs in 27 Brassicaceae species. The WGD and WGT indicate the WGD and WGT, respectively. Nonredundant protein sequence database (Nr); GO. The specific data used in this figure were from Supplemental Tables S2 and S5–S7.

Furthermore, we identified 4,096 glucosinolate genes, 6,625 auxin genes, 13,805 flowering genes, 1,939 anthocyanin genes, and 36,632 resistance genes in the genomes of 27 Brassicaceae species (Figure 5; Supplemental Tables S10 and S11). The resistance genes were further divided into 14 families according to their domains (Supplemental Table S11). These genes have played an important role in the breeding and research of Brassicaceae. Users can query and download related genes directly from our website. For example, in the flowering gene browsing interface, the user can find *FLC* genes in all species by entering "*FLC*" in the input box of the lower right corner. Then, the *FLC* gene list and queried sequence data are provided for users to download at the bottom of the result page, so that users can use these data for further cross-species comparative analysis.

The m6A methylation is one of the most important types of RNA modification. The characterization of plant m6A and its function is a major focus of current plant research. A previous study indicates that the evolutionary relationships among genes related to m6A modification are highly conserved across plants (Yue et al., 2019). Therefore, we explored the *m6A* genes in Brassicaceae species and provided them in the TBGR database. We detected 268 m6A writers, 419 erasers, and 544 readers in 27 Brassicaceae species (Figure 5; Supplemental Table S10). The 268 m6A writers were further divided into 113 with 70 kDa subunit of methyltransferase-A (MTA70), 42 Wilms' tumor 1-associating protein (WTAP), 28 Virilizer, and 85 Hakai genes. These rich *m6A* gene resources in the TBGR database will contribute to the genetic improvement of Brassicaceae species through epitranscriptome manipulation in the future.

Charts interface

This interface provides several interactive plots to view TFs, m6A, TEs, and important functional genes of 27 Brassicaceae species (Figure 2). The number of glucosinolate, auxin, flowering, anthocyanin, and m6A genes in each species is shown in a histogram that permits intuitive comparison of the number of genes of each type among different species. The bar charts and line charts show the number of genes in each TF family and each type of TEs in each species, which makes it easier for users to make comparisons among species. The multi-select dropdown menu allows users to select species to view according to their needs. These charts can also be easily downloaded using the link in the upper right corner of this interface.

Download interface

The genome-related datasets (general feature format [Gff], genome, coding sequence [CDS], and protein sequences) of 82 genomes from 27 Brassicaceae species can be downloaded from this interface (Figure 2). There are multiple versions of the genome or pan-genome of several species, such as *B. rapa*, *B. napus*, and *B. oleracea*. In addition, comprehensive information for these genomes is provided, including the genome version, genome size, gene number, chromosome number, scaffold N50, sequencer type, publication information (date and journal), reference, database, and their links for each species. All of the genomic datasets and resources can be easily downloaded and used for comparative genomic analyses. In the Download interface, we provide the main bioinformatics analysis pipeline used in this study. If users want to perform related analysis on other versions of the genome, they can also perform similar analysis by referring to our methods or pipelines of this interface in TBGR database.

Tools interface

Four popular tools (Basic Local Alignment Search Tool [BLAST], Synteny, Primer Design, JBrowse) are embedded in the TBGR database to help users perform genomic analyses (Figure 2).

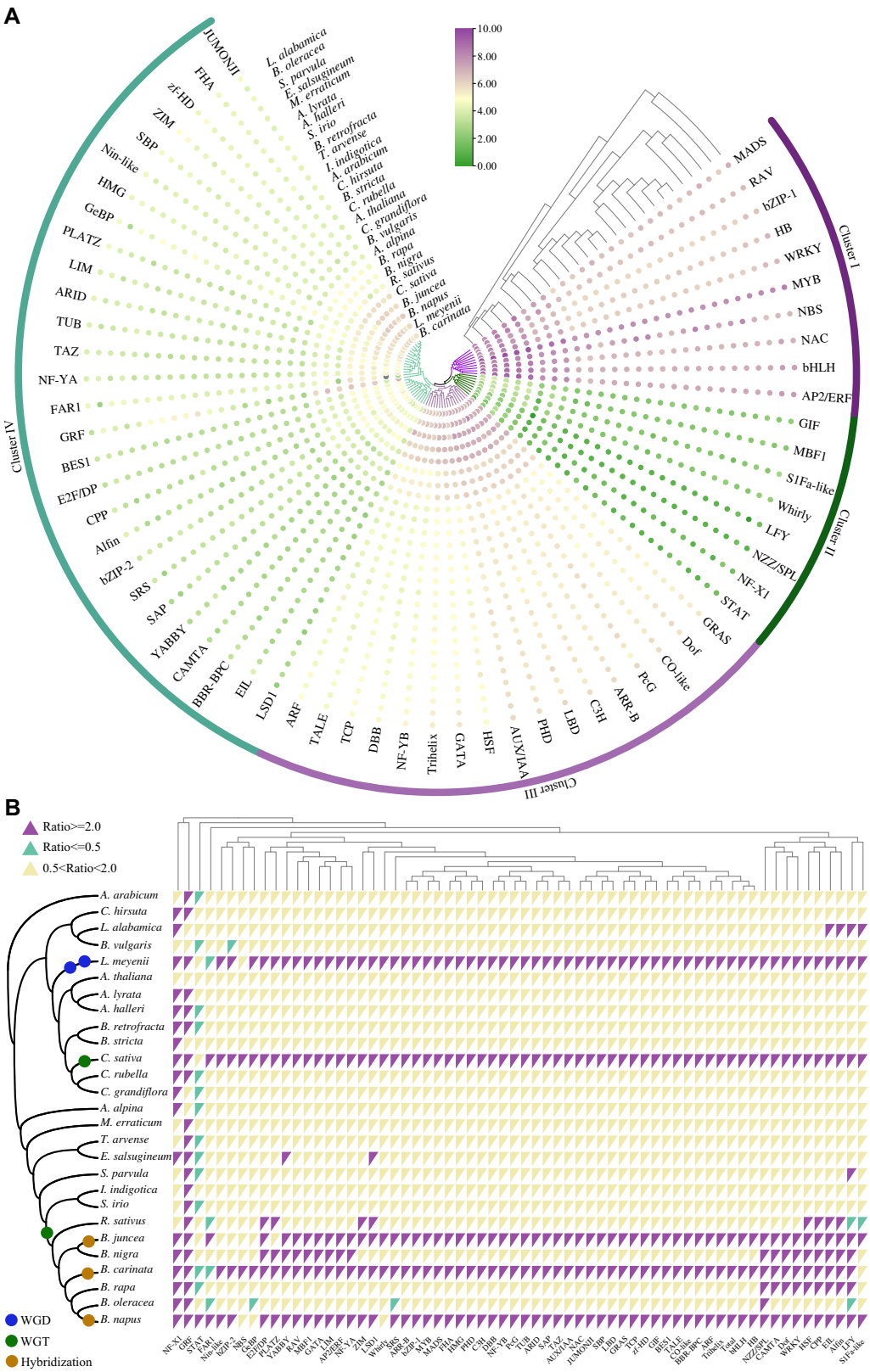

Figure 4 Analysis of TF families in the genomes of 27 Brassicaceae species. A, Circle plot showing the number of members of each TF family. The number for each TF was $\log_2$ transformed. The four clusters (I–VI) were obtained by cluster analysis according to the number of members of each TF family in the 27 species. B, The ratio of transcription factor family number in each species compared with that of *A. thaliana*.

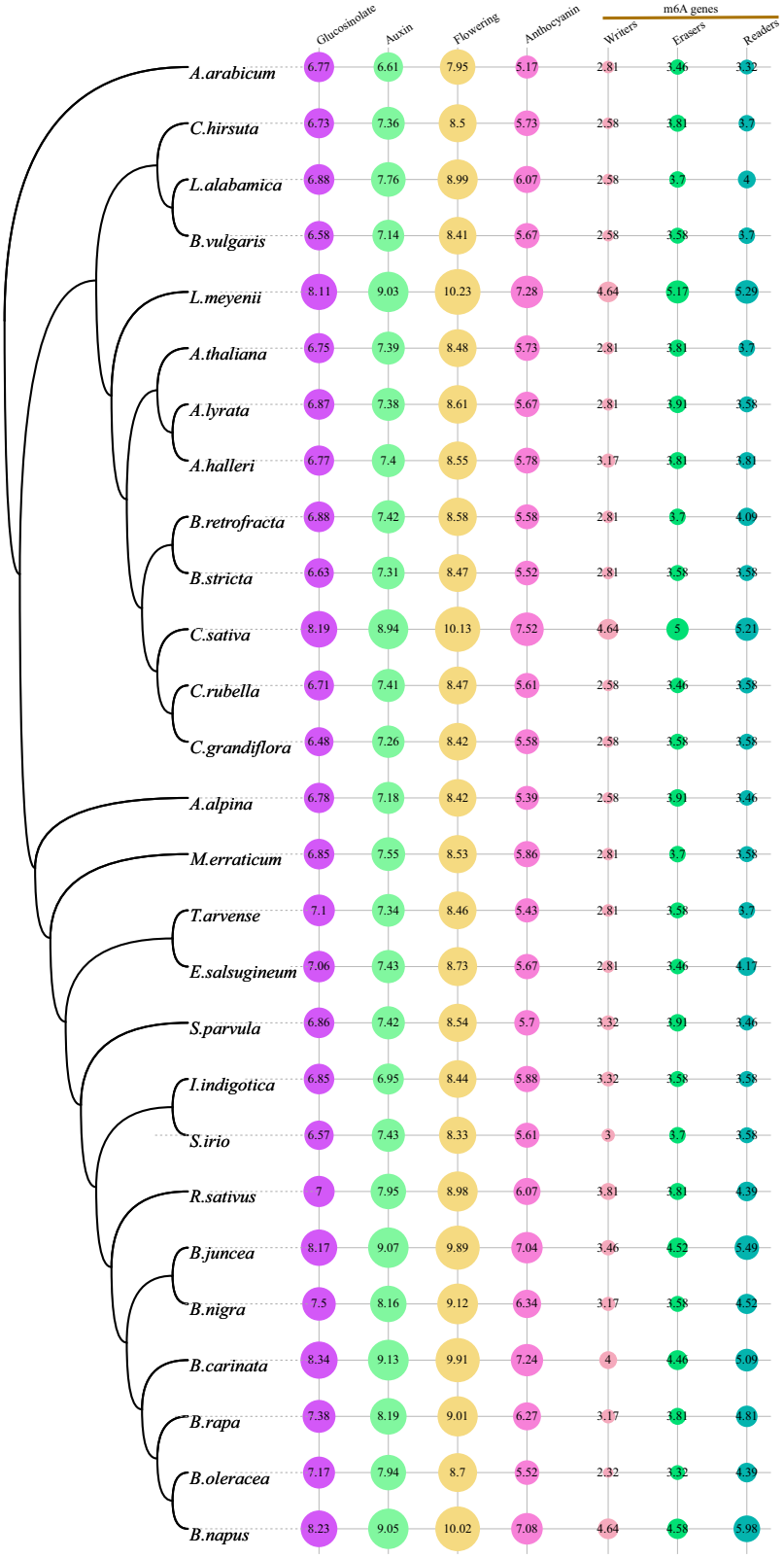

Figure 5 Plot of the number of functional gene families (glucosinolate, auxin, flowering, anthocyanin, and m6A genes) in the 27 Brassicaceae species. The numbers for each functional gene family were $\log_2$ transformed. The size of the dot represents the $\log_2$-transformed gene number, and the color represents different types of genes.

The BLAST tool is used in the TBGR database to help users conduct sequence alignment. We provide a user-friendly graphic interface and built the BLAST database using the nucleotide sequences and protein sequences of 27 Brassicaceae species. Users can easily perform similarity analysis by copying a sequence to the frame or uploading a file in FASTA format.

We developed a syntenic tool (Synteny) to help users perform collinearity analysis within or between species. The tool was divided into online and database modes. In the online mode, users can upload Gff and BLAST files for collinearity analysis. The website also provides a visualization function for the results of collinearity analysis. By adjusting the configuration file, four forms of display effects can be achieved, including bar plotter, circle plotter, dot plotter, and dual synteny plotter. For the database mode, users only need to select two Brassicaceae species among the 27 species, and they can quickly display the collinearity diagram.

A tool (Primer Design) was developed to help users design the primers for the genes of 27 Brassicaceae species by entering the gene accession under the Database mode. In addition, the users can design the primers by uploading their own nucleic acid sequences in FASTA format in online mode.

A genome browser (JBrowse) was constructed to display the genomic data and features of Brassicaceae genes. Users can query the genomic sequences of each chromosome or scaffold, which enables users to view the detailed information for selected genes.

Resource, help, and contact interfaces

In the Resource interface, the figures, genome size, and links to relevant studies and the database of each Brassicaceae species are provided (Figure 2). In addition, we have collected common databases of Brassicaceae species and provided links in the Resource interface to facilitate access to these resources. In the Help interface, users are provided with a detailed manual of the TBGR database and the abbreviations of gene symbols and species names. In the Contact interface, users are provided our emails and phone numbers.

Discussion

Several databases of single species of Brassicaceae have been built, such as genomic resources for Alpine rock-cress (*Arabis alpina*) (http://www.arabis-alpina.org/) (Jiao et al., 2017), the radish genome database (RadishGD) (http://radish-genome.org/) (Yu et al., 2019), Pennycress Genomics for field penny-cress. (*Thlaspi arvense*) (http://pennycress.umn.edu), the Brassica Genomics Database for *B. carinata* (http://brassicadb.bio2db.com) (Song et al., 2021a), and the BnPIR (http://cbi.hzau.edu.cn/bnapus/) (Liu et al., 2021). In addition, some databases have been constructed for multiple *Brassica* species, such as the Brassica database (BRAD, http://brassicadb.cn) (Chen et al., 2022), the Brassica.info database (https://www.brassica.info), the Crucifer Genome Initiative (http://cruciferseq.ca), the Brassica Genome Database (BrGDB, http://www.plantgdb.org/BrGDB/), and Brassica Genome (http://www.brassicagenome.net). These databases provide rich resources for Brassica studies, and their main focus is on genome data dissemination and visualization. Most of the analysis data in the BRAD database focuses mainly on one species (*B. rapa*); however we performed bioinformatics analyses of the genomic data for 27 representative genomes of Brassicaceae species and displayed them in our database. For example, BRAD provides users only the ability to search for TF gene families and several important functional gene families of *B. rapa*. Compared with BRAD, our database contains TF gene families and several important functional gene families of 27 representative genomes of Brassicaceae species.

Moreover, we added more analyses and data below, which are not available in other public resources database of Brassicaceae. First, we performed the m6A, CRISPR guide sequences, and TEs analyses of 27 representative genomes of Brassicaceae species, and finally deposited these results in our database. These three research contents are not only hotspots of current research in plants, but also very important for comparative and functional genomics studies of Brassicaceae species. Second, our synteny tool in TBGR database not only provides collinearity analysis and shows collinearity diagram among Brassicaceae species in the database, but also provides user upload function. Users can upload relevant data for collinearity analysis for any other species, which facilitates genome collinearity analysis between non-Brassicaceae and Brassicaceae species. Third, our database contains duplication type information of each gene of 27 representative genomes of Brassicaceae species. Furthermore, we also performed the orthologous and paralogous gene identification within or between 27 representative genomes of Brassicaceae species. Then, the phylogentic trees were constructed based on these orthologous and paralogous gene families, and shown in the TBGR database. Finally, comprehensive information for 82 genomes of 27 Brassicaceae species are provided in the Download interface, including the species classification, genome information, publication information, and database for each species. All of these information can help users quickly understand the 82 genomic information of various versions (Supplemental Table S1), and users can choose to use them according to their needs. However, this information does not exist in other databases of Brassica species.

Compared with the existing platforms, the TBGR database integrates most of the resources of these websites, including a systematic analysis of these genomic data. Therefore, the TBGR database can help users mine data from the genomes of all Brassicaceae species, and it has specific features that facilitate comparison among these existing databases. First, the TBGR database contains comprehensive genomic information from 82 public genomes of 27 Brassicaceae species, so it provides a wealth of resources for genomics research in this group. Second, based on the bioinformatic analysis of

these genomic data, we obtained a large number of important functional genes (glucosinolate, auxin, flowering, resistance, and anthocyanin genes), m6A, CRISPR guide sequences, TEs, and other related data resources. We then integrated all of these data into our website, which provides rich resources for researchers of Brassicaceae and other plants. Third, this database provides information on the synteny and orthologs between any two Brassicaceae species. Fourth, gene annotation information of the 27 Brassicaceae species from four annotation databases is provided. Finally, the Blast, Synteny, Primer Design, and JBrowse tools were built into the TBGR database, which helps users easily perform comparative genomic analyses of Brassicaceae species.

In conclusion, the TBGR database will facilitate both comparative genomic and functional genomic studies in plants, especially for Brassicaceae species. Users can easily retrieve and download the target functional genes for research for cross-species comparative analysis. In the future, we will continuously improve and update the newly assembled genomes and comparative genomic tools in the TBGR database.

Materials and methods

Collection of genome resources

Genome-related datasets such as GFF files, genome sequences, CDSs, and protein sequences of each Brassicaceae species were collected from several public databases. Most datasets were obtained from the NCBI (https://www.ncbi.nlm.nih.gov), Phytozome (https://phytozome-next.jgi.doe.gov), the BRAD (http://brassicadb.cn), the BnPIR (http://cbi.hzau.edu.cn/bnapus/), and other related databases (Supplemental Table S1). Alternatively, spliced genes were removed using a custom Perl script to prevent redundant sequences from being included in subsequent analyses. We obtained 82 genomes of 27 Brassicaceae species and provided detailed information on these plants, such as their classification, sequencing information, references, and related databases (Supplemental Table S1). For species with multiple versions of the genome, we choose the version that is of high quality, the latest or commonly used by researchers. The representative genomes for further bioinformatics analysis were marked with red color in Supplemental Table S1. On TBGR database, we added a list of selected representative genomes in the Download interface, and marked the background color as yellow. For another interface of the TBGR database, we also added the corresponding genome version information after the species name.

TE detection

RepeatMasker (v4.1.1), RepeatModeler (v2.0.1), and HelitronScanner (v1.0) were used for transposon prediction (Tempel, 2012; Xiong et al., 2014) (Supplemental Figure S1). RepeatMasker was used to compare the model species Zebrafish (*Danio rerio*) and *A. thaliana* in the Repbase database (Bao et al., 2015). RepeatModeler and RepeatMasker were used to predict TEs *de novo*. Finally, HelitronScanner was used to predict helitron type transposons (Xiong et al., 2014). The pearson correlation coefficient between the expansion of the transposon and the genome size was calculated using the "cor" function package of R (https://www.r-project.org).

Gene functional annotation

The gene annotations of 27 Brassicaceae species were performed using four databases, including the Pfam database (v34.0) (http://pfam.xfam.org) (Mistry et al., 2021), UniProt knowledgebase (Swiss-Prot, TrEMBL) (https://www.uniprot.org) (UniProt C, 2021), nonredundant protein sequence database (Nr) (https://www.ncbi.nlm.nih.gov), and GO database (http://geneontology.org) (Gene Ontology C, 2021). Furthermore, we obtained the level of GO annotation using the goatools (Klopfenstein et al., 2018). All the annotation information was sorted into tables in batches using the Perl program for display in the TBGR database.

Identification of orthologous, paralogous, and xenologous genes

Orthologs, paralogs, and xenologs were identified using OrthoFinder (v2.0) (Emms and Kelly, 2019). First, the similarity relationships between the protein sequences of all species were based on Blastp similarity scores (*E*-value $<$1e-5). Cluster analysis was conducted using the MCL graph clustering algorithm (Inflation value $>$1.5), and single-copy and multi-copy gene families were obtained. Gene trees and species trees were constructed using each gene family across all species using the FastTree software (Price et al., 2009).

Synteny and duplication type detection

The Multiple Collinearity Scan toolkit (MCScanX) was used to conduct gene collinearity analysis with default parameters (Wang et al., 2012). First, BLASTP was used to search for potential anchors (*E*-value $<10^{-5}$; top five matches) between each gene in multiple genomes. Collinearity analysis was then performed using the BLAST results and GFF files. Finally, collinear relationships within one species or between two species were drawn using TBtools (Chen et al., 2020). The program (duplicate_gene_classifier) in MCScanX was used to infer types of duplicated genes (Wang et al., 2012).

Functional gene identification

The Pfam database was used to identify the main 63 TF gene families from all the protein sequences of 27 species (*E*-value $<$1e-5) (Mistry et al., 2021). A total of 73 Arabidopsis glucosinolate genes were collected from previous studies (Wang et al., 2011; Cheng et al., 2014; Song et al., 2021a); 295 Arabidopsis flowering genes were collected from the FLOR-ID database and previous studies (Aach et al., 2014; Cheng et al., 2014; Bouche et al., 2016; Li et al., 2018; Song et al., 2021a); and 151 Arabidopsis auxin genes and 41 Arabidopsis anthocyanin genes were downloaded from the BRAD database (Chen et al., 2022). Functional-related genes in other Brassicaceae species were identified using BLASTP (*E*-value $<$1e-5; score $>$150) and checked

manually. The resistance gene analogs were detected in 27 Brassicaceae species using the pipeline of RGAugury (Li et al., 2016).

m6A analysis

The m6A genes were mainly divided into three groups: writers, erasers (Alkylation repair protein-B [AlkB]), and readers (IYT521-B homology [YTH]) (Yue et al., 2019). Writers contained four gene families, 70 kDa subunit of methyltransferase-A (MTA70), WTAP, Virilizer, and Hakai, which were identified using the numbers "PF05063", "PF17098", "PF15912", and "PF18408", respectively. The AlkB gene family of erasers was identified using the number "PF13532", and the YTH gene family of readers was identified using the number "PF04146" with the *E*-value $<$1e-5.

CRISPR–Cas9 target sequence identification

The CRISPR–Cas9 target sites were detected by the CasFinder system (Aach et al., 2014). First, RepeatMasker was used to shield repetitive sequences in each genome (Tarailo-Graovac and Chen, 2009). Bowtie was then used to generate the index for each genome with default parameters (Giannoulatou et al., 2014). The two scripts CasFinder.pl and CasValue_v2.pl in CasFinder were used to identify the guide sequences for each gene with default parameters except –*x* = 5 (Aach et al., 2014). Finally, the candidate guide sequences were filtered using in-house Perl scripts to obtain the specific guide sequences for each gene. The off-target sequences for each guide sequence were detected using CasOT program (Xiao et al., 2014).

Database construction

The TBGR database was constructed based on the Django framework and MySQL database management. The TBGR database contains several databases that store processed genome-related datasets in MySQL. The TBGR database was written in programming languages, including HTML, CSS, JavaScript, and Python. The charts were generated by Echarts, which is an open-source visualization library implemented in JavaScript. The interactive Web interface was built to enable users to conveniently access the TBGR database and obtain relevant information. Python, JavaScript, and HTML were used to transmit query requirements and extract datasets rapidly from the MySQL database. The collected genomic datasets were processed using Python and Perl scripts, and several bioinformatics tools were used to explore the biological meaning of the genomic datasets.

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure S1. The flowchart for TE identification for 27 Brassicaceae species used in this study.

Supplemental Figure S2. The classification of TEs in the genomes of 27 Brassicaceae species.

Supplemental Figure S3. The correlation analysis of TE and genome size in 27 Brassicaceae species.

Supplemental Table S1. The detailed information of 82 genomes from 27 Brassicaceae species.

Supplemental Table S2. The statistics of gene functional annotation of 27 Brassicaceae species compiled from four databases.

Supplemental Table S3. The overall statistics of orthologous genes in all 27 Brassicaceae species.

Supplemental Table S4. The statistics of orthologous genes in each species of 27 Brassicaceae.

Supplemental Table S5. The statistics of orthologous genes in each species of 27 Brassicaceae. WGD, whole genome duplication.

Supplemental Table S6. The statistics of guide sequence number for CRISPR studies in 27 Brassicaceae species.

Supplemental Table S7. The number of different types of TE in the genome of 27 Brassicaceae species.

Supplemental Table S8. The number of genes in each TF family for all 27 Brassicaceae species.

Supplemental Table S9. The ratio of the number of genes in each TF family for each of the 27 Brassicaceae species compared with *A. thaliana*.

Supplemental Table S10. The number of important functional genes related to glucosinolate, auxin, flowering, anthocyanin, and *m6A* genes in 27 Brassicaceae species.

Supplemental Table S11. The number of resistance genes for each classification in 27 Brassicaceae species.

Funding

This work was supported by the National Natural Science Foundation of China (32172583, 31801856), Natural Science Foundation of Hebei (C2021209005), Innovation and entrepreneurship training program for college students of North China University of Science and Technology (Grant No. X2020009), and the China Postdoctoral Science Foundation (2020M673188, 2021T140097).

Conflict of interest statement. The authors declare no competing interests.

References

Aach J, Mali P, Church GM (2014) CasFinder: Flexible algorithm for identifying specific Cas9 targets in genomes. bioRxiv, doi: 10.1101/005074

Al-Shehbaz IA, Beilstein MA, Kellogg EA (2006) Systematics and phylogeny of the Brassicaceae (Cruciferae): an overview. Plant Syst Evol **259**: 89–120

Bao W, Kojima KK, Kohany O (2015) Repbase update, a database of repetitive elements in eukaryotic genomes. Mob DNA **6**: 11

Bayer PE, Scheben A, Golicz AA, Yuan Y, Faure S, Lee H, Chawla HS, Anderson R, Bancroft I, Raman H, et al. (2021) Modelling of gene loss propensity in the pangenomes of three Brassica species suggests different mechanisms between polyploids and diploids. Plant Biotechnol J **19**: 2488–2500

Bouche F, Lobet G, Tocquin P, Perilleux C (2016) FLOR-ID: an interactive database of flowering-time gene networks in *Arabidopsis thaliana*. Nucleic Acids Res **44**: D1167–D1171

Cai X, Chang L, Zhang T, Chen H, Zhang L, Lin R, Liang J, Wu J, Freeling M, Wang X. (2021) Impacts of allopolyploidization and

structural variation on intraspecific diversification in *Brassica rapa*. Genome Biol **22**: 166

Chalhoub B, Denoeud F, Liu S, Parkin IA, Tang H, Wang X, Chiquet J, Belcram H, Tong C, Samans B, et al. (2014) Plant genetics. Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. Science **345**: 950–953

Chen C, Chen H, Zhang Y, Thomas HR, Frank MH, He Y, Xia R (2020) TBtools: an integrative toolkit developed for interactive analyses of big biological data. Mol Plant **13**: 1194–1202

Chen H, Wang T, He X, Cai X, Lin R, Liang J, Wu J, King G, Wang X (2022) BRAD V3.0: an upgraded Brassicaceae database. Nucleic Acids Res **50**: D1432–D1441

Cheng F, Wu J, Wang X (2014) Genome triplication drove the diversification of *Brassica* plants. Hortic Res **1**: 14024

Emms DM, Kelly S (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol **20**: 238

Gene Ontology C (2021) The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res **49**: D325–D334

Giannoulatou E, Park SH, Humphreys DT, Ho JW (2014) Verification and validation of bioinformatics software without a gold standard: a case study of BWA and Bowtie. BMC Bioinformatics **15(Suppl 16)**: S15

Golicz AA, Bayer PE, Barker GC, Edger PP, Kim H, Martinez PA, Chan CK, Severn-Ellis A, McCombie WR, Parkin IA, et al. (2016) The pangenome of an agronomically important crop plant *Brassica oleracea*. Nat Commun **7**: 13390

He Z, Ji R, Havlickova L, Wang L, Li Y, Lee HT, Song J, Koh C, Yang J, Zang M, et al. (2021) Genome structural evolution in *Brassica* crops. Nat Plants **7**: 757–765

Jeong YM, Kim N, Ahn BO, Oh M, Chung WH, Chung H, Jeong S, Lim K-B, Hwang Y-J, Kim G-B, et al. (2016) Elucidating the triplicated ancestral genome structure of radish based on chromosome-level comparison with the *Brassica* genomes. Theor Appl Genet **129**: 1357–1372

Jiao WB, Accinelli GG, Hartwig B, Kiefer C, Baker D, Severing E, Willing E-M, Piednoel M, Woetzel S, Madrid W, et al. (2017) Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. Genome Res **27**: 778–786

Kagale S, Koh C, Nixon J, Bollina V, Clarke WE, Tuteja R, Spillane C, Robinson SJ, Links MG, Clarke C, et al. (2014) The emerging biofuel crop *Camelina sativa* retains a highly undifferentiated hexaploid genome structure. Nat Commun **5**: 3706

Klopfenstein DV, Zhang L, Pedersen BS, Ramirez F, Warwick Vesztrocy A, Naldi A, et al. (2018) GOATOOLS: a python library for gene ontology analyses. Sci Rep **8**: 10872

Lee SI, Kim NS (2014) Transposable elements and genome size variations in plants. Genomics Inform **12**: 87–97

Li H, Fan Y, Yu J, Chai L, Zhang J, Jiang J, Cui C, Zheng B, Jiang L, Lu K (2018) Genome-wide identification of flowering-time genes in *Brassica* species and reveals a correlation between selective pressure and expression patterns of vernalization-pathway genes in *Brassica napus*. Int J Mol Sci **19**: 3632

Li P, Quan X, Jia G, Xiao J, Cloutier S, You FM (2016) RGAugury: a pipeline for genome-wide prediction of resistance gene analogs (RGAs) in plants. BMC Genomics **17**: 852

Liu S, Liu Y, Yang X, Tong C, Edwards D, Parkin IAP, Zhao M, Ma J, Yu J, Huang S, et al. (2014) The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. Nat Commun **5**: 3930

Liu D, Yu L, Wei L, Yu P, Wang J, Zhao H, Zhang Y, Zhang S, Yang Z, Chen G, et al. (2021) BnTIR: an online transcriptome platform for exploring RNA-seq libraries for oil crop *Brassica napus*. Plant Biotechnol J **19**: 1895–1897

Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, et al. (2021) Pfam: The protein families database in 2021. Nucleic Acids Res **49**: D412–D419

Nagaharu U (1935) Genome analysis in *Brassica* with special reference to the experimental formation of *B. napus* and peculiar mode of fertilization. Jpn J Bot **7**: 389–452

Naville M, Henriet S, Warren I, Sumic S, Reeve M, Volff JN, Chourrout D (2019) Massive changes of genome size driven by expansions of non-autonomous transposable elements. Curr Biol **29**: 1161–1168 e1166

Nishihara H (2020) Transposable elements as genetic accelerators of evolution: contribution to genome size, gene regulatory network rewiring and morphological innovation. Genes Genet Syst **94**: 269–281

Parkin IAP, Koh C, Tang H, Robinson SJ, Kagale S, Clarke WE, Town CD, Nixon J, Krishnakumar V, Bidwell SL, et al. (2014) Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid *Brassica oleracea*. Genome Biol **15**: R77

Price MN, Dehal PS, Arkin AP (2009) FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol **26**: 1641–1650

Song JM, Guan ZL, Hu JL, Guo CC, Yang ZQ, Wang S, Liu DX, Wang B, Lu SP, Zhou R, et al. (2020a) Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of *Brassica napus*. Nat Plants **6**: 34–45

Song X, Li Y, Hou X (2013) Genome-wide analysis of the AP2/ERF transcription factor superfamily in Chinese cabbage (*Brassica rapa* ssp. *pekinensis*). BMC Genomics **14**: 573

Song X, Nie F, Chen W, Ma X, Gong K, Yang Q, Wang J, Li N, Sun P, Pei Q, et al. (2020b) Coriander genomics database: a genomic, transcriptomic, and metabolic database for coriander. Hortic Res **7**: 1–10

Song X, Wei Y, Xiao D, Gong K, Sun P, Ren Y, Yuan J, Wu T, Yang Q, Li X, et al. (2021a) *Brassica carinata* genome characterization clarifies U's triangle model of evolution and polyploidy in *Brassica*. Plant Physiol **186**: 388–406

Song X, Yang Q, Bai Y, Gong K, Wu T, Yu T, Pei Q, Duan W, Huang Z, Wang Z, et al. (2021b) Comprehensive analysis of SSRs and database construction using all complete gene-coding sequences in major horticultural and representative plants. Hortic Res **8**: 122

Tarailo-Graovac M, Chen N (2009) Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics Chapter 4(Unit 4): 10

Tempel S (2012) Using and understanding RepeatMasker. Methods Mol Biol **859**: 29–51

UniProt C (2021) UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res **49**: D480–D489

Walden N, German DA, Wolf EM, Kiefer M, Rigault P, Huang XC, Kiefer C, Schmickl R, Franzke A, et al. (2020) Nested whole-genome duplications coincide with diversification and high morphological disparity in Brassicaceae. Nat Commun **11**: 3795

Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, Lee T-H, Jin H, Marler B, Guo H (2012) MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res **40**: e49

Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, Bai Y, Mun J-H, Bancroft I, Cheng F, et al. (2011) The genome of the mesopolyploid crop species *Brassica rapa*. Nat Genet **43**: 1035–1039

Wang H, Wu J, Sun S, Liu B, Cheng F, Sun R, Wang X (2011) Glucosinolate biosynthetic genes in *Brassica rapa*. Gene **487**: 135–142

Xiao A, Cheng Z, Kong L, Zhu Z, Lin S, Gao G, Zhang B (2014) CasOT: a genome-wide Cas9/gRNA off-target searching tool. Bioinformatics **30**: 1180–1182

Xiong W, He L, Lai J, Dooner HK, Du C (2014) HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. Proc Natl Acad Sci U S A **111**: 10263–10268

Yang J, Liu D, Wang X, Ji C, Cheng F, Liu B, Hu Z, Chen S, Pental D, Ju Y, et al. (2016) The genome sequence of allopolyploid *Brassica juncea* and analysis of differential homoeolog gene expression influencing selection. Nat Genet **48**: 1225–1232

Yu HJ, Baek S, Lee YJ, Cho A, Mun JH (2019) The radish genome database (RadishGD): an integrated information resource for radish genomics. Database (Oxford) **2019**: baz009

Yue H, Nie X, Yan Z, Weining S (2019) N6-methyladenosine regulatory machinery in plants: composition, function and evolution. Plant Biotechnol J **17**: 1194–1208

Zhang J, Tian Y, Yan L, Zhang G, Wang X, Zeng Y, Zhang J, Ma X, Tan Y, Long N, et al. (2016) Genome of plant Maca (*Lepidium meyenii*) illuminates genomic basis for high-altitude adaptation in the central Andes. Mol Plant **9**: 1066–1077