12-1-2022

TVIR: a comprehensive vegetable information resource database for comparative and functional genomic studies

Tong Yu

Xiao Ma

Zhuo Liu

Xuehuan Feng

Zhiyuan Wang

See next page for additional authors

Authors

Tong Yu, Xiao Ma, Zhuo Liu, Xuehuan Feng, Zhiyuan Wang, Jun Ren, Rui Cao, Yingchao Zhang, Fulei Nie, and Xiaoming Song

Horticulture Research

Horticulture Research, 2022, **9**: uhac213

https://doi.org/10.1093/hr/uhac213

Article

TVIR: a comprehensive vegetable information resource database for comparative and functional genomic studies

Tong Yu[1,†], Xiao Ma[1,†], Zhuo Liu[1,†], Xuehuan Feng[2,†], Zhiyuan Wang[1], Jun Ren[3], Rui Cao[1], Yingchao Zhang[1], Fulei Nie[1], and Xiaoming Song[1,2,4,*]

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

*Corresponding author: E-mail: songxm@ncst.edu.cn
†Equal contribution.

Abstract

Vegetables are an indispensable part of the daily diet of humans. Therefore, it is vital to systematically study the genomic data of vegetables and build a platform for data sharing and analysis. In this study, a comprehensive platform for vegetables with a user-friendly Web interface—The Vegetable Information Resource (TVIR, http://tvir.bio2db.com)—was built based on the genomes of 59 vegetables. TVIR database contains numerous important functional genes, including 5215 auxin genes, 2437 anthocyanin genes, 15 002 flowering genes, 79 830 resistance genes, and 2639 glucosinolate genes of 59 vegetables. In addition, 2597 N6-methyladenosine (m6A) genes were identified, including 513 writers, 1058 erasers, and 1026 readers. A total of 2 101 501 specific clustered regularly interspaced short palindromic repeat (CRISPR) guide sequences and 17 377 miRNAs were detected and deposited in TVIR database. Information on gene synteny, duplication, and orthologs is also provided for 59 vegetable species. TVIR database contains 2 346 850 gene annotations by the Swiss-Prot, TrEMBL, Gene Ontology (GO), Pfam, and Non-redundant (Nr) databases. Synteny, Primer Design, Blast, and JBrowse tools are provided to facilitate users in conducting comparative genomic analyses. This is the first large-scale collection of vegetable genomic data and bioinformatic analysis. All genome and gene sequences, annotations, and bioinformatic results can be easily downloaded from TVIR. Furthermore, transcriptome data of 98 vegetables have been collected and collated, and can be searched by species, tissues, or different growth stages. TVIR is expected to become a key hub for vegetable research globally. The database will be updated with newly assembled vegetable genomes and comparative genomic studies in the future.

Introduction

Vegetables are necessary for human health as they provide rich sources of dietary fiber, minerals, vitamins, and other nutrients [1–4]. The decreased cost of sequencing has led to numerous vegetable genomes being released in the past decade. In particular, the amount and quality of vegetable genomes have markedly increased due to the emergence of third-generation sequencing in recent years [5]. The first vegetable genome to be sequenced was that of *Cucumis sativus* (cucumber) in 2009, and since then the genomes of several vegetable species have been sequenced [6]. For example, two high-quality and chromosome-level genomes of celery and coriander were assembled in our laboratory [7, 8]. Genome sizes vary significantly among different vegetable species. The assembled genome size of garlic (*Allium sativum*) was 16.24 Gb, which was similar that of onion (*Allium cepa*) (14.90 Gb) [9, 10]. However, the genome size of *Neoporphyra haitanensis* (Laver), a lower plant from the phylum Rhodophyta, was only 49.67 Mb. Moreover, pan-genome studies of several vegetable species have been reported recently, including *Brassica rapa*, *Brassica oleracea*, *Brassica napus*, *Raphanus sativus*, *Solanum lycopersicum*, and *Solanum melongena* [11–18].

The availability of these vegetable genomes provides valuable resources for comparative and functional genomics studies, such as the evolution, domestication, and molecular mechanisms of vegetables. The genome sequences of most major vegetable crops have been deposited in public databases, such as the National Center for Biotechnology Information (NCBI) database [19]. However, most genome sequences have not been updated and often lack gene sets and annotation. Currently, several organism-specific databases of vegetables have been constructed, such as the Onion Genome Sequencing Project (https://www.oniongenome.wur.nl) [10], Celery Genome Database (CGD) (http://celerydb.bio2db.com) [7], Coriander Genome Database (http://cgdb.bio2db.com) [8, 20, 21], Radish Genome Database (RadishGD) (http://radish-genome.org/) [22], Pepper Genome Platform (PGP) (http://peppergenome.snu.ac.kr), Eggplant Genome Database (http://www.eggplant-hq.cn/Eggplant/home/index) [23], Lettuce Genome Resource (https://lgr.genomecenter.ucdavis.edu)

Received: 8 June 2022; **Accepted:** 14 September 2022; **Published:** 19 September 2022; **Corrected and Typeset:** 1 December 2022

[24], *Beta vulgaris* Resource (https://bvseq.boku.ac.at/index.shtml) [25, 26], Melonomics (http://melonomics.cragenomica.es) [27], Asparagus Genome Project (http://asparagus.uga.edu/tripal/) [28], Sweet Potato Genome database (http://public-genomes-ngs.molgen.mpg.de/sweetpotato/) [29], White Lupin Genome database (https://www.whitelupin.fr/download.html) [30], and the pan-genome information resource for *B. napus* (BnPIR) (http://cbi.hzau.edu.cn/bnapus/) [31]. In addition, some databases have been constructed for species at the genus or family level, such as the *Brassica* Database (BRAD) (http://brassicadb.cn) [32], The Brassicaceae Genome Resource (TBGR) (http://www.tbgr.org.cn) [33], Solanaceae Genomics Network (SGN) (https://solgenomics.net/) [34], and CuGenDB for Cucurbitaceae (http://cucurbitgenomics.org) [35].

Despite these individual resources, an integrated, comprehensive, and up-to-date database of gene resources for the vegetable community is lacking. Therefore, to make all the vegetable genome sequences, annotated data, and transcriptome information accessible to the vegetable research community, a comprehensive gene resource platform for vegetables— The Vegetable Information Resource (TVIR)—was built in the current study. The purpose of the TVIR platform is to provide a repository that can be used for comparative and functional genomics studies of vegetable species at the whole-genome level. An overview of the TVIR database interfaces, including Browse, Search, Charts, Tools, Download, and Resources, is provided in this report. The database will promote vegetable research by providing several convenient tools and rich omics data resources.

Results

Overview of vegetable research

According to statistics, the genomes of 65 vegetable species have been sequenced to date (Fig. 1a, Supplementary Data Table S1). In terms of taxonomic distribution, these species comprise 57 eudicots, 6 monocots, 1 basal angiosperm, and 1 Rhodophyta (Supplementary Data Fig. S1a). Among the eudicots, the highest number of species was in the order Cucurbitales (14), followed by Fabales (12), Brassicales (9), and Solanales (9), suggesting their importance to humans. The first vegetable crop to be fully sequenced and assembled was cucumber, released in 2009. Most other vegetable crops have been sequenced in the past 5 years (2017–21), accounting for 67.69% (44) of all assembled vegetable genomes (Fig. 1a, Supplementary Data Fig. S1b).

Related information resources for public genomes from 65 vegetable crops were collected (Supplementary Data Table S1). Due to incomplete genome or gene annotation data for six species, only the remaining 59 species were selected for subsequent analysis.

Architecture of TVIR database

Systematic analyses were performed on these collected genomic data, such as gene annotation, orthologous genes, transcription factors (TFs), CRISPR guide sequences, m6A, miRNA, synteny, duplication type, and detection of the main functional genes. Several important functional genes were identified for inclusion in TVIR, including auxin, anthocyanin, flowering, resistance, and glucosinolate genes. Finally, TVIR database was constructed to facilitate access to and employment of these genomic resources and bioinformatic analysis results (Fig. 1b). All the genomic resources were stored in the back-end tables of MySQL, which could be easily accessed through the front-end web application. A detailed description of TVIR database, including Search, Browse, Tools, Download, Resources, and Help interfaces (Figs 1b and 2) is provided below.

Search

In this section, users can search the gene annotation, CRISPR guide sequences, duplication type, homologs, and synteny for 59 vegetable species. Based on the five protein databases (Gene Ontology, Nr, Pfam, Swiss-Prot, and TrEMBL), 73.02% (*Pisum sativum*) to 99.88% (*Benincasa hispida*) of all genes were annotated in each species (Fig. 3, Supplementary Data Table S2). All annotations can be searched according to the gene identifier in TVIR.

To facilitate gene-editing research on vegetable species, CRISPR guide sequences for all genes were designed and deposited in TVIR. In total, 2 101 501 specific CRISPR guide sequences were designed in all examined species, and the success rate of guide sequence design in all species ranged from 52.95% (*Manihot esculenta*) to 99.73% (*Cucurbita argyrosperma*) (Fig. 3, Supplementary Data Table S3).

To explore gene evolution in each vegetable species, homologous gene analysis was performed using the OrthoFinder program. There were 78 256 orthogroups detected among the 59 vegetable species, of which 30 388 groups were species-specific (Supplementary Data Tables S4 and S5). In addition to displaying these orthogroups and downloading sequences from TIVR database, phylogenetic trees reconstructed using the genes in each group were also illustrated. The species tree was reconstructed using orthologous genes to uncover the phylogenetic relationships of these vegetable species.

To clarify the situation of gene duplication or loss after whole-genome duplication (WGD) or whole-genome triplication (WGT) events in vegetable species, syntenic analyses were conducted using amino acid sequences within species or between any two species. The syntenic genes and corresponding figures were then integrated into TVIR database. Furthermore, five duplication types, including singleton, proximal, tandem, dispersed, and WGD/segmental, were detected for each gene in the 59 vegetable species (Fig. 3, Supplementary Data Table S6), and WGD/segmental duplication was dominant in several species that underwent WGD or WGT events, especially for recent genome duplication events. For example, WGD/segmental repeat accounted for the highest proportion of duplication type in *Euryale ferox* (66.57%), followed by *Brassica juncea* (65.32%) and *B. napus* (60.99%) (Fig. 3, Supplementary Data Table S6). This is because *E. ferox* has experienced ancient WGD and recent WGT events [36]. *B. juncea* and *B. napus* not only experienced ancient WGD and recent WGT events, but also tetraploids were formed by diploid hybridization [37, 38]. The proportion of WGD/segmental repeat type in four Cucurbitaceae species (*C. argyrosperma*, *Cucurbita maxima*, *Cucurbita moschata*, and *Cucurbita pepo*) was also >50% due to their recent WGD event [39–41]. Similarly, the proportion of WGD/segmental repeat type was >50% in *B. rapa*, *Lupinus albus*, and *Zingiber officinale* because they underwent a recent WGT event or two rounds of recent WGD events [30, 42–44].

Browse

The Browse section provides species information and sequences of functional genes, TFs, m6A, transcriptome, and miRNAs of related vegetable species. For user convenience, the information in the Browse interface is presented according to the family and species. Comprehensive information for each species is provided, including taxonomy ID, common name, classification, chromosome number, genome size, and species pictures. These

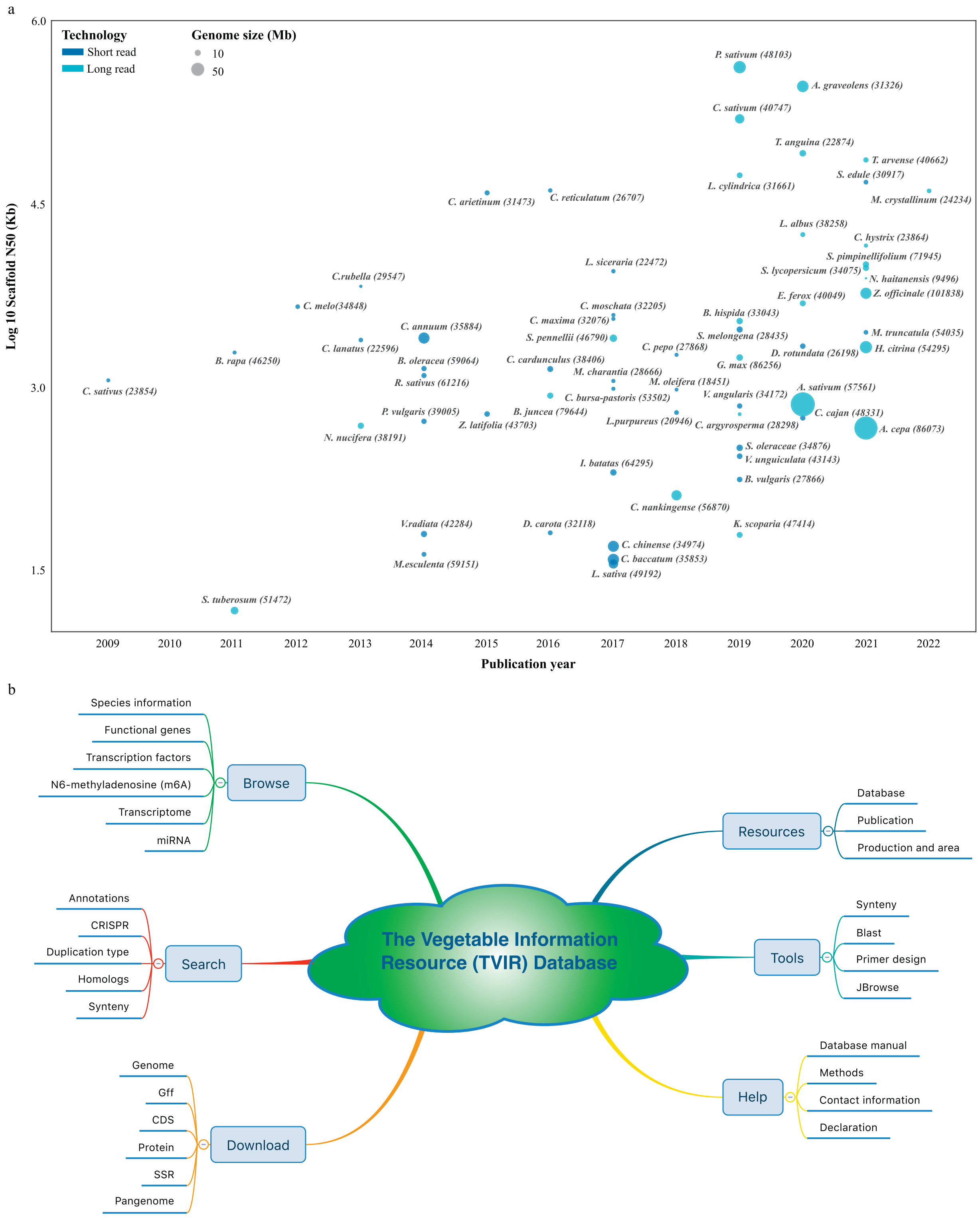
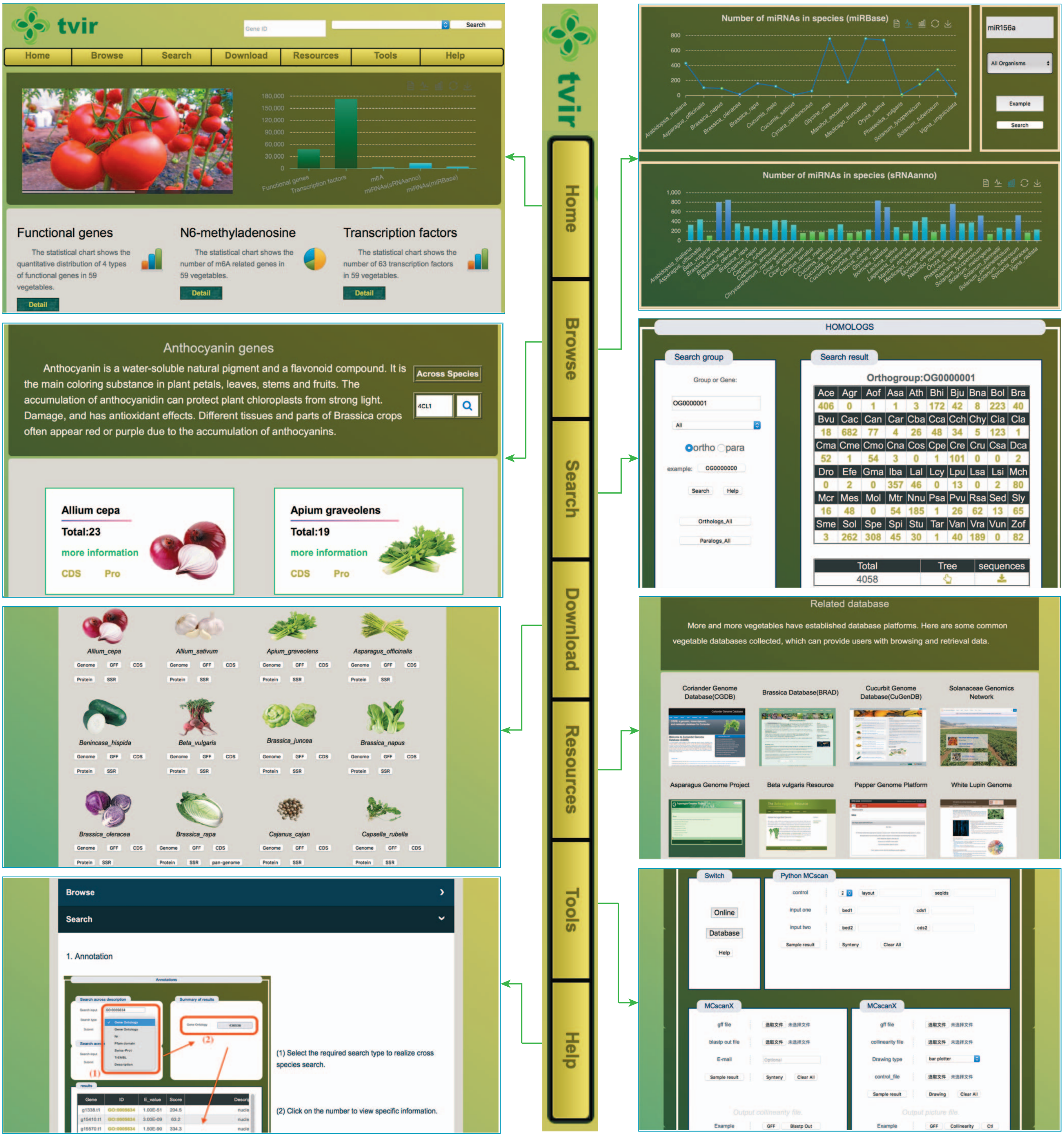


Figure 1. Overview of vegetable genome sequencing and TVIR database. **a** Species information and major genome sequencing indicators (publication year, assembled genome size, gene number, scaffold N50, and sequencing technology) of 65 vegetables from the years 2009 to 2022. **b** Architecture of TVIR database.

resources can help users rapidly understand the related vegetable species.

TFs have critical roles in various stresses as well as plant growth and development [45, 46]. In total, 172 493 TFs from 63 families were identified in all examined species (Fig. 4a, Supplementary Data Table S7). The four TF families with the largest number of genes were myeloblastosis (MYB) (18797), nucleotide binding sites (NBSs) (15053), APETALA2/ethylene-responsive element binding factors (AP2/ERF) (11992), and basic helix–loop–helix (bHLH) (10609). Compared with *Arabidopsis thaliana*, the NF-X1, LEAFY (LFY), and growth-regulating factor (GRF) gene families were significantly expanded, while the glabrous-



Functional genes

The statistical chart shows the quantitative distribution of 4 types of functional genes in 59 vegetables.

N6-methyladenosine

The statistical chart shows the number of m6A related genes in 59 vegetables.

Transcription factors

The statistical chart shows the number of 63 transcription factors in 59 vegetables.

Anthocyanin genes

Anthocyanin is a water-soluble natural pigment and a flavonoid compound. It is the main coloring substance in plant petals, leaves, stems and fruits. The accumulation of anthocyanidin can protect plant chloroplasts from strong light. Damage, and has antioxidant effects. Different tissues and parts of Brassica crops often appear red or purple due to the accumulation of anthocyanins.

Allium cepa
Total:23
more information
CDS Pro

Apium graveolens
Total:19
more information
CDS Pro

Orthogroup:OG0000001

Ace	Agr	Aof	Asa	Ath	Bhi	Bju	Bna	Bol	Bra
406	0	1	1	3	172	42	8	223	40
Bvu	Cac	Can	Car	Cba	Cca	Cch	Chy	Cia	Cla
18	682	77	4	26	48	34	5	123	1
Cma	Cme	Cmo	Cna	Cos	Cpe	Cre	Cru	Csa	Dca
52	1	54	3	0	1	101	0	0	2
Dro	Efe	Gma	Iba	Lal	Lcy	Lpu	Lsa	Lsi	Mch
0	2	0	357	46	0	13	0	2	80
Mcr	Mes	Mol	Mtr	Nnu	Psa	Pvu	Rsa	Sed	Sly
16	48	0	54	185	1	26	62	13	65
Sme	Sol	Spe	Spi	Stu	Tar	Van	Vra	Vun	Zof
3	262	308	45	30	1	40	189	0	82

Total	Tree	sequences
4058	☝	⤓

More and more vegetables have established database platforms. Here are some common vegetable databases collected, which can provide users with browsing and retrieval data.

(1) Select the required search type to realize cross species search.

(2) Click on the number to view specific information.

Figure 2. Overview of TVIR database with main interfaces and internal features, including Home, Browse, Resources, Search, Tools, Download, and Help interfaces.

enhancer-binding protein (GeBP), basic leucine zipper 2 (bZIP-2), signal transducer and activator of transcription (STAT), and NOZZLE/SPOROCYTELESS (NZZ/SPL) gene families were contracted in most examined species (Fig. 4b, Supplementary Data Table S8). Almost all TF families were expanded in *Z. officinale* and *Glycine max*, which underwent one additional recent WGD event and two WGD events, respectively (Fig. 4b) [43, 44, 47].

Notably, one common WGT event occurred in *Brassica* and *R. sativus*, and most TF families were significantly expanded in *R. sativus* but not in *B. rapa* and *B. oleracea* [42, 48, 49]. This might be due to most genes being lost after WGT events, consistent with previous reports [42, 49]. In the two tetraploid species of the genus *Brassica* (*B. napus* and *B. juncea*), most TF families were expanded because they were formed by the crossing of two diploid species of the genus *Brassica* [37, 38] (Fig. 4b).

Several important agronomic-related functional gene families were identified, including 5215 auxin genes, 2437 anthocyanin genes, 15 002 flowering genes, 79 830 resistance genes [563 receptor-like proteins (RLPs), 64 214 receptor-like kinases (RLKs), and 15 053 NBSs], edna and 2639 glucosinolate genes in the

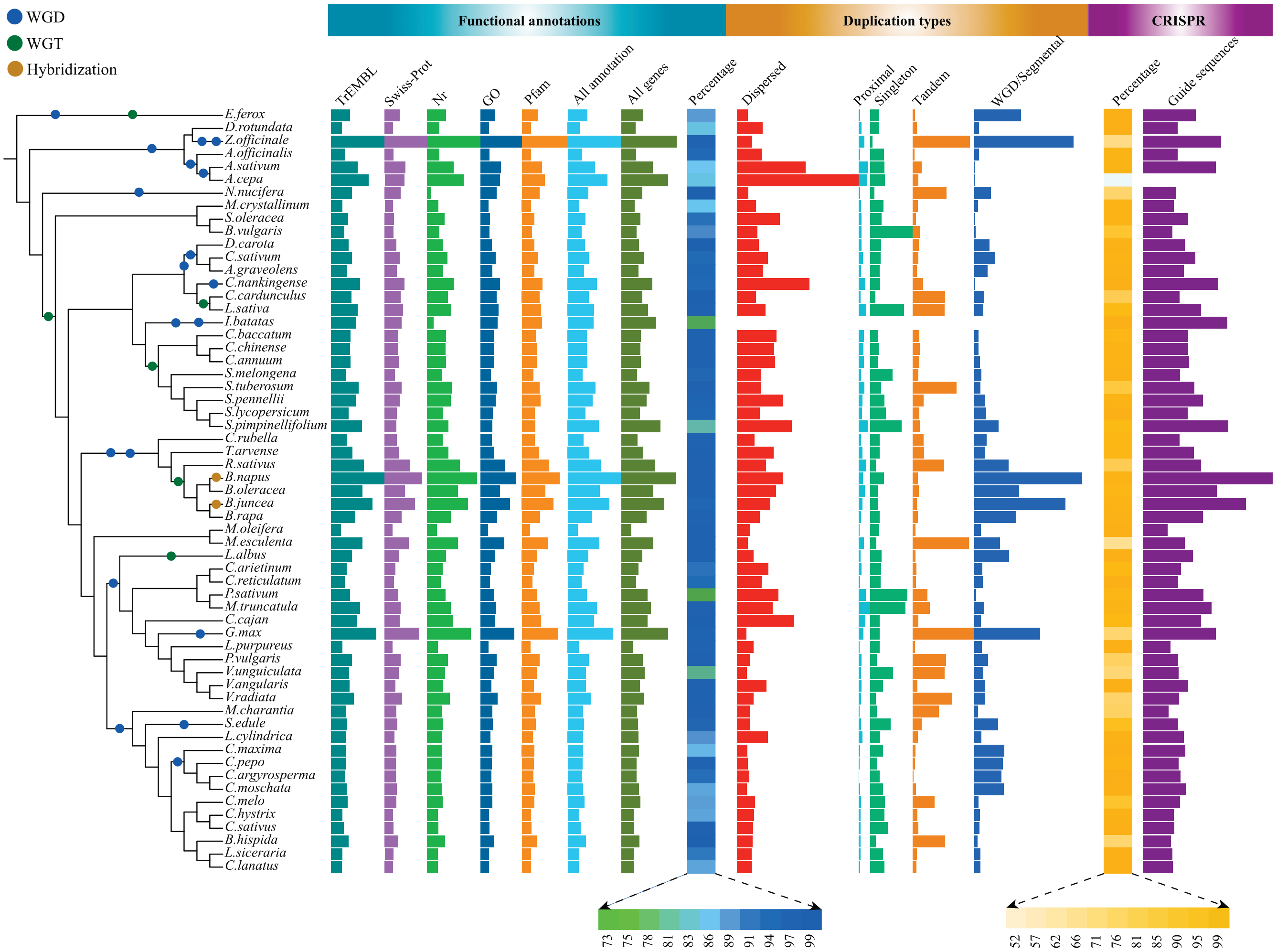


Figure 3. Bar plots of the number of gene functional annotations, gene duplication types, and CRISPR guide sequences in 59 vegetable species. WGD, WGT, and genome hybridization events are indicated by blue, green, and orange circles, respectively. The specific values can be obtained from Supplementary Data Tables S2, S3, and S6.

genomes of 59 vegetable species (Fig. 5, Supplementary Data Table S9). All of these functional genes play essential roles in the breeding of vegetable species and research on them. Users can quickly search and download related genes directly from TIVR database. For example, all 4-coumarate:CoA ligase 1 (*4CL1*) genes in all species can be searched at the anthocyanin gene browsing interface. Furthermore, the *4CL1* gene list is provided and queried sequences are available for users to download at the bottom of the page. Users can employ these datasets to conduct further comparative analysis among cross-species.

One of the most important RNA modifications in plants is m6A methylation. The function and characterization of m6A are a major focus in current plant research. The genes involved in m6A modification are highly conserved across different plants [50]. Therefore, the m6A genes in vegetable species were identified and deposited in TVIR database. Moreover, the m6A genes were classified into writers, erasers, and readers according to their functions. A total of 513 writers, 1058 erasers, and 1026 readers were identified in the 59 vegetable species (Fig. 5, Supplementary Data Table S9). The writers were further divided into 262 methyltransferase-A with 70-kDa subunit (MTA70), 105 Hakai, 111 Wilms' tumor 1-associated protein (Wtap), and 35 Virilizer genes. These m6A gene resources in TVIR database will provide valuable guidance for the genetic improvement of vegetable crops by epitranscriptome manipulation.

To facilitate the study of vegetable miRNAs, information on miRNAs was collected from the sRNAanno and miRBase databases (Supplementary Data Table S10). In the sRNAanno database 13 454 miRNAs were identified from 37 vegetables and *A. thaliana*, while in the miRBase database 3913 miRNAs were identified from 15 vegetables and *A. thaliana*. In TVIR database the bar or line charts clearly display the number of miRNAs in each species, making it easier for users to perform comparisons among different vegetable species. Furthermore, the hairpin sequences, mature sequences, and target genes of miRNAs were predicted and can be downloaded from TVIR database. Moreover, the structure of each miRNA is illustrated and shown in TVIR. Publications related to each miRNA and links to the NCBI database are also provided. The multi-select dropdown menu allows users to select species to search miRNAs according to their needs. Users can also select all organisms from the dropdown menu, which will then show the search for a certain miRNA in all species.

In addition to vegetable genome data, the transcriptome data of 98 vegetables were collected and collated. Users can easily and quickly select the corresponding transcriptome data according to the Latin name of the species, the plant tissue, or the settings of different growth and development periods. In the retrieved

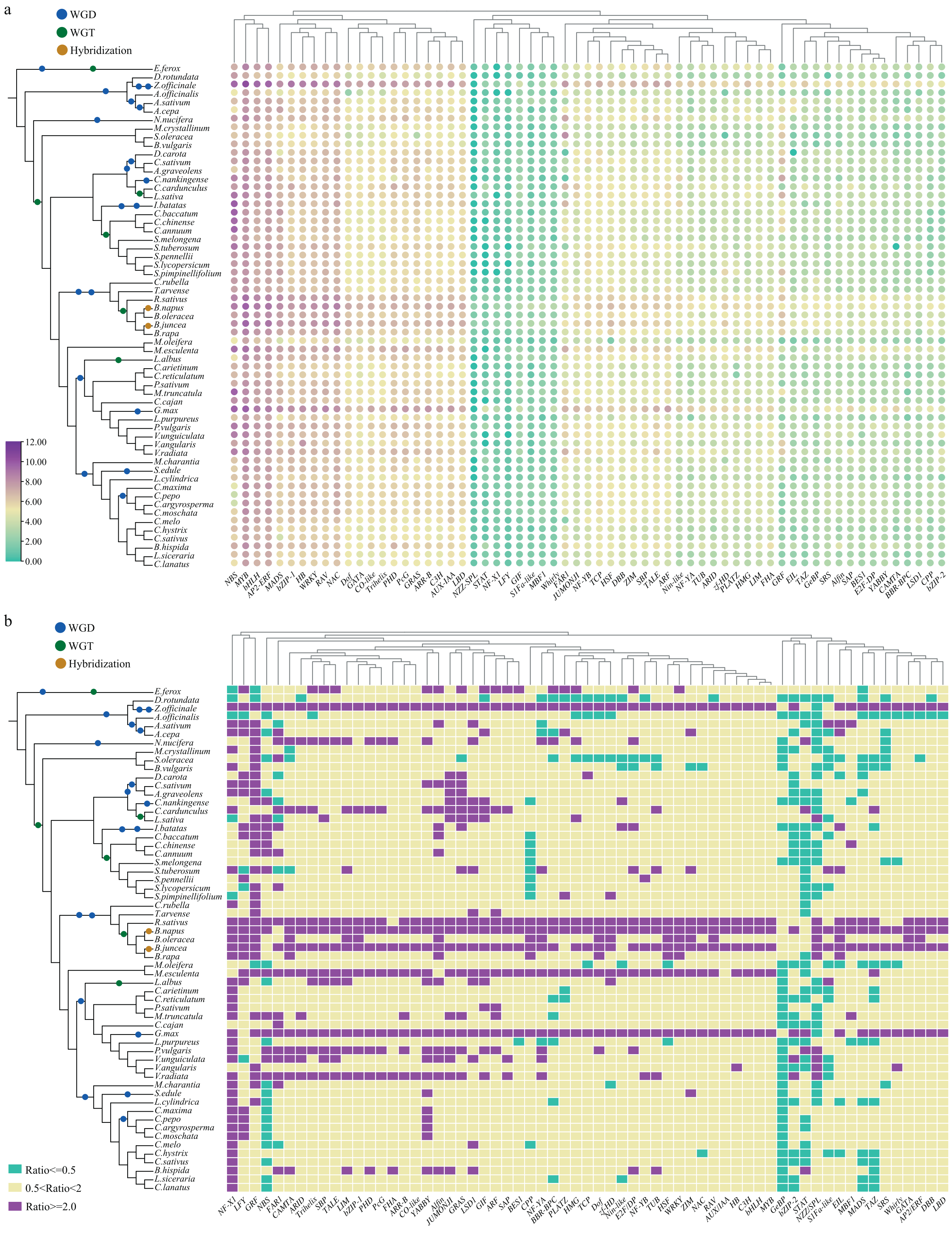


Figure 4. Heat map of TF family analysis in 59 vegetable crops. **a** Circle plot showing the number of members of each TF family in each species. The number of each TF was transformed by $\log_2$. **b** The ratio of number of members of each TF family compared with *A. thaliana* in each vegetable species. Purple indicates a ratio of ≥ 2, green indicates a ratio of ≤ 0.5, and yellow indicates 0.5 < ratio < 2. The specific values of ratios can be obtained from Supplementary Data Table S8.

information table, users are provided with accession numbers and links to transcriptome data, sample information, sequencing library, and author information. Moreover, the information table can be downloaded, providing useful data resources for the retrieval, collection, and analysis of gene expression of vegetable species.

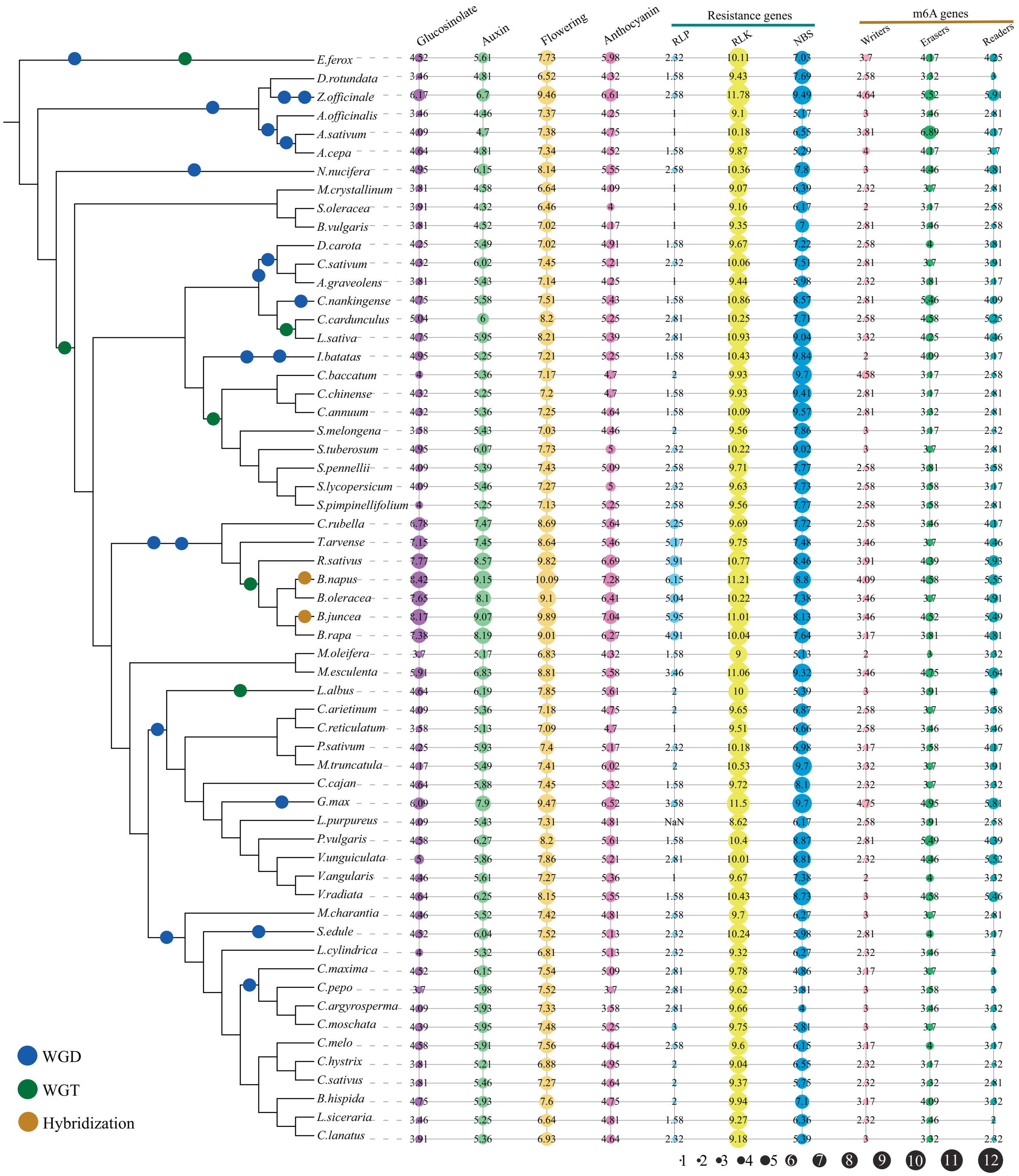


Figure 5. Plot of the number of members of several functional gene families, including auxin, anthocyanin, flowering, glucosinolate, resistance, and m6A genes, in the 59 vegetable crops. The numbers for each gene family were transformed by $\log_2$. Circle size represents the size of the value, and specific values can be obtained from Supplementary Data Table S9.

Tools

Four popular tools—Synteny, Blast, Primer Design, and JBrowse—are included in TVIR database to help users conduct comparative and functional genomic analyses. Two syntenic tools, the Multiple Collinearity Scan toolkit (MCScanX) and Python MCscan, are provided and users can conduct syntenic analyses among species with these tools. Both Python MCscan and MCScanX have online and database modes in TVIR database.

For Python MCscan, users can upload bed and coding sequence (CDS) files for syntenic analysis in the online mode. The configuration files 'layout' and 'seqids' can be created according to the manual of the Python MCscan program. Depending on the user's needs, this tool can perform a collinear analysis of two to four species at a time. For database mode, users only need to select two to four species in TVIR database, then the syntenic diagram is rapidly illustrated.

For MCScanX, users can upload Blast and general feature format (gff) files for syntenic analysis in the online mode. The database also provides a visualization using the results of syntenic analysis, including gff and collinear files. Four syntenic types, including dot plotter, circle plotter, bar plotter, and dual synteny plotter, can be obtained by selecting different configuration files. For the database mode, users only need to select two species among the 59 vegetable species, and the corresponding syntenic diagram is rapidly displayed according to the selected drawing type.

The Blast tool was included in TVIR database to facilitate sequence alignment. A user-friendly interface was created and Blast databases were constructed using CDS and protein sequences of the 59 vegetable species. All users can easily perform sequence alignment by uploading a Fasta format file or directly copying sequences to the frame. A Primer Design tool was deposited in TVIR to help researchers design primers for genes in the 59 vegetable species by entering the related gene accession number. Moreover, users can also design primers for their gene sequences with Fasta format. In addition, a JBrowse tool was built to show the genomic sequences and features of vegetable genes. This tool allows users to check the detailed information for selected genes on the corresponding chromosome or scaffold sequences.

Download

The genome sequences, gff, CDS, and protein sequences of each vegetable species can be obtained from TVIR database. Moreover, the pan-genomes of seven species, including *B. rapa*, *C. sativus*, *L. albus*, *R. sativus*, *S. lycopersicum*, *S. melongena*, and *Solanum tuberosum*, were also provided. Finally, the simple sequence repeat (SSR) markers were identified from all the coding genes of each species, and they can also be downloaded from TVIR database. All these genomic datasets and resources facilitate rapid and easy comparative genomic analyses of vegetable crops.

Resources and Help

The Resources section of TVIR provides information on the database, publication, yield, and planting area of related vegetables. Through the Resources page, users can quickly understand the research value and status of major vegetables, and rapidly obtain relevant vegetable data resources. In the Help section, a detailed manual for each interface of the TVIR database is supplied. Furthermore, e-mail addresses and phone numbers are provided to enable users to easily contact us.

Discussion

In this study, 65 vegetable species with sequenced genomes were collected, and comparative and functional genomics studies were subsequently performed on the 59 species that had well-annotated genomes. As increasing numbers of vegetable genomes have been sequenced, several species-specific vegetable databases have been built, such as RadishGD for radish [22], CGD for celery [7], CGDB for coriander [8, 20, 21], and EGD for eggplant [23]. Moreover, several databases were constructed for species in the same genera or family, such as BRAD for *Brassica* [32], TBGR for Brassicaceae [33], SGN for Solanaceae (https://solgenomics.net/) [34], and CuGenDB for Cucurbitaceae (http://cucurbitgenomics.org) [35].

Although the above databases provide rich resources for vegetable research, they only contain one or several closely related species. Compared with these databases, TVIR integrates most resources of these websites and provides systematic analysis results using 59 vegetable genomes; consequently, TVIR has a number of advantages over existing databases. Firstly, TVIR database contains comprehensive genomic resources from public genomes of 59 vegetable species, making it the first large-scale collection of vegetable genomic data. Secondly, TVIR contains lots of important functional genes (flowering, resistance, anthocyanin, glucosinolate, and auxin genes), m6A, guide sequences of CRISPR, synteny, and orthologs, which were detected by using a series of bioinformatics analyses. Thirdly, TVIR provides the gene annotation information on the 59 vegetable species from five annotation databases. Finally, TVIR includes the Blast, Primer Design, Synteny, and JBrowse tools, which help users easily conduct comparative genomic analyses in vegetable species. TVIR aims to provide researchers with related information for a queried gene, including sequences, domains, annotation, homologs, synteny, and gene family. This ‘data consumer’-oriented function could help people studying vegetables to obtain an overview of queried genes, and this will be particularly useful when designing experiments to verify gene function. Therefore, several user-oriented software and web servers are provided in TVIR to facilitate gene analysis for vegetable molecular biologists.

In conclusion, TVIR will facilitate both comparative and functional genomic studies in vegetables, and potentially plant species. Researchers can easily retrieve and download the target functional genes for cross-species comparative study. Researchers will be encouraged to submit their genome and transcriptome sequences to TVIR, and technical assistance will be assigned for uploading and handling their sequences. TVIR database will be continuously improved and updated with new omics sequences and comparative genomic tools.

Materials and methods

Retrieval of genome and gene annotation resources

Genome sequences, gff files, CDS, and protein sequences of each vegetable species were collected from several major public or private databases. For example, most genome datasets were downloaded from NCBI (https://www.ncbi.nlm.nih.gov), JGI (https://phytozome-next.jgi.doe.gov), CuGenDB (http://cucurbitgenomics.org/), BRAD (http://brassicadb.cn) [32], TBGR (http://www.tbgr.org.cn) [33], the Solanaceae Genomics Network (https://solgenomics.net/), and other related databases for single species (Supplementary Data Table S1). Spliced genes were deleted to avoid redundant sequences using a custom Perl script. A total of 65 vegetable species was retrieved for genome sequencing. Genomic information for these species was collated, including their classification, genome size, gene number, scaffold N50, chromosome number, publication status, sequencing information, and genome access databases (Supplementary Data Table S1).

Functional annotation of genes

Gene annotations of 59 vegetable species were conducted using five protein databases, including Swiss-Prot and TrEMBL of the UniProt knowledgebase (https://www.uniprot.org) [51], Pfam (v34.0) (http://pfam.xfam.org) [52], Gene Ontology (GO, http://geneontology.org) [53], and the non-redundant protein sequence database (Nr, https://www.ncbi.nlm.nih.gov). All gene annotations are displayed in TVIR database. SSR markers were developed according to our previous reports [54–56].

Identification of orthologous, paralogous, and xenologous genes

OrthoFinder (v2.0) was used to identify orthologs, paralogs, and xenologs [57]. First, the Blastp program was used to obtain the similarity relationships among the protein sequences in different species (E-value <1e−5). Cluster analysis was then performed using the MCL algorithm (inflation value >1.5), and gene and species trees were built using each gene family of all species.

Detection of collinearity and duplication types

Collinearity analysis was performed by the MCScanX software with default parameters [58]. First, Blastp was used to search for potential homologous genes (E-value $<10^{-5}$) among different species. Then, the collinearity was detected according to the gff files and Blast results. Finally, collinear relationships were illustrated using TBtools [59]. The duplicate_gene_classifier program from MCScanX was used to predict the duplication types [58].

Identification of transcription factors and functional genes

The 63 TF gene families were detected, using the Pfam database, from the protein sequences of 59 vegetable species (E-value <1e−5) [52]. The 295 flowering genes of *Arabidopsis* were collected from previous studies and the FLOR-ID database [60–62]. The 73 glucosinolate genes of *Arabidopsis* were collected from previous reports [49, 63, 64], and the 41 anthocyanin genes and 151 auxin genes of *Arabidopsis* were obtained from the BRAD website [32]. Homologous flowering, glucosinolate, anthocyanin, and auxin genes in other vegetable species were identified by Blastp (E-value <1e−5; identity >60%; score >150) with manual checking. Furthermore, the identified candidate genes were further verified by their domains, and related genes with domains were annotated in TVIR database.

Detection of resistance genes

Three kinds of resistance (R) genes were predominantly identified NBS, RLK, and RLP genes [65, 66]. The NBS genes of each species were detected by the accession number PF00931 in the Pfam annotation with an E-value <1e−5. RLK genes were obtained from the Pfam annotation by the keyword 'kinase'; in total, 15 gene families were assigned to the RLK genes. The 56 RLP genes of *Arabidopsis* were downloaded from the BRAD website [32], and homologous genes in other vegetable species were detected by the Blastp program (E-value <1e−5; identity >60%; score >150).

m6A identification

The m6A genes comprised three groups, including writers, readers [IYT521-B homology (YTH)], and erasers [Alkylation repair protein-B (AlkB)] [50]. Writers had four gene families, including MTA70, WTAP, HAKAI, and VIRILIZER, which were identified by the accession numbers PF05063, PF17098, PF18408, and PF15912, respectively. The YTH genes of readers were detected by the accession number PF04146, and AlkB genes of erasers were detected by the accession number PF13532 of the Pfam annotation.

miRNA collection and target gene identification

The mature, hairpin sequences, and gff files of miRNAs were downloaded from sRNAanno and miRBase (Release 22.1) [67, 68]. The structure of each miRNA was created using the ViennaRNA package (v2.5.0) with slight modifications in batch [69]. The target genes of each miRNA were predicted using the TargetFinder program [70]. All the above datasets were sorted by in-house Perl scripts so they could be displayed in TVIR.

Cas9 target sequence design for CRISPR

The CasFinder pipeline was used to design the Cas9 target sites for CRISPR [62]. Firstly, the repetitive genome sequences were screened using the RepeatMasker program for each species [71]. The index was then created for each genome by the Bowtie program [72]. Finally, the scripts CasValue_v2.pl and CasFinder.pl from the CasFinder pipeline were adopted to design the guide sequences for the CRISPR study [62]. The candidate sequence was filtered by in-house Perl scripts to obtain the specific sequence for each gene.

Database construction

Based on the Django framework, the TVIR database was built with MySQL database management according to our previous reports [33, 73] and using several programming languages, such as Python, JavaScript, HyperText Markup Language (HTML), and Cascading Style Sheets (CSS). Several interactive Web interfaces were constructed to help users conveniently search TVIR and obtain the required information. Echarts was used to show the charts, and is an open-source visualization tool integrated into JavaScript. Vegetable genome and gene sequences were processed by Python or Perl scripts, and some bioinformatics tools were employed to perform comparative genomic analyses.

Acknowledgements

This work was supported by the Natural Science Foundation for Distinguished Young Scholar of Hebei Province (C2022209010), the National Natural Science Foundation of China (32172583, 31801856), the Natural Science Foundation of Hebei (C2021209005), the Educational and teaching reform research and practice project of North China University of Science and Technology (L21106), and the China Postdoctoral Science Foundation (2020M673188, 2021T140097).

Author contributions

X.S. conceived the project and was responsible for project initiation. X.S. and T.Y. supervised and managed the project and research. The data collection and bioinformatics analyses were led by X.S., T.Y., X.M., Z.L., Z.W., and X.F. The database construction was led by X.S., T.Y., Z.L., and F.N. The manuscript was organized, written, and revised by X.S., T.Y., X.F., X.M., R.C., Y.Z., and J.R. All authors read and approved the manuscript.

Data availability

All materials and related data in this study are provided in the TVIR database and supplementary files. Other datasets are available upon request to the corresponding author.

Conflict of interest

The authors declare no competing interests.

Supplementary data

Supplementary data is available at *Horticulture Research* online.

References

1. Chen F, Song Y, Li X *et al.* Genome sequences of horticultural plants: past, present, and future. *Hortic Res*. 2019;**6**:112.
2. Weng Y. Inaugural editorial: vegetable research. *Veg Res*. 2021;**1**:1.
3. Pei Q, Yu T, Wu T *et al.* Comprehensive identification and analyses of the Hsf gene family in the whole-genome of three Apiaceae species. *Hortic Plant J*. 2021;**7**:457–68.
4. Pei Q, Li N, Bai Y *et al.* Comparative analysis of the *TCP* gene family in celery, coriander and carrot (family Apiaceae). *Vegetable Res*. 2021;**1**:5.
5. Mei Y, Jing D, Tang S *et al.* InsectBase 2.0: a comprehensive gene resource for insects. *Nucleic Acids Res*. 2021;**50**:D1040–5.
6. Huang S, Li R, Zhang Z *et al.* The genome of the cucumber, *Cucumis sativus* L. *Nat Genet*. 2009;**41**:1275–81.
7. Song X, Sun P, Yuan J *et al.* The celery genome sequence reveals sequential paleo-polyploidizations, karyotype evolution and resistance gene reduction in Apiales. *Plant Biotechnol J*. 2021;**19**:731–44.
8. Song X, Wang J, Li N *et al.* Deciphering the high-quality genome sequence of coriander that causes controversial feelings. *Plant Biotechnol J*. 2020;**18**:1444–56.
9. Sun X, Zhu S, Li N *et al.* A chromosome-level genome assembly of garlic (*Allium sativum*) provides insights into genome evolution and allicin biosynthesis. *Mol Plant*. 2020;**13**:1328–39.
10. Finkers R, van Kaauwen M, Ament K *et al.* Insights from the first genome assembly of onion (*Allium cepa*). *G3 Genes|Genomes|Genetics*. 2021;**11**:jkab243.
11. Golicz AA, Bayer PE, Barker GC *et al.* The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nat Commun*. 2016;**7**:13390.
12. He Z, Ji R, Havlickova L *et al.* Genome structural evolution in *Brassica* crops. *Nat Plants*. 2021;**7**:757–65.
13. Cai X, Chang L, Zhang T *et al.* Impacts of allopolyploidization and structural variation on intraspecific diversification in *Brassica rapa*. *Genome Biol*. 2021;**22**:166.
14. Song JM, Guan Z, Hu J *et al.* Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of *Brassica napus*. *Nat Plants*. 2020;**6**:34–45.
15. Bayer PE, Scheben A, Golicz AA *et al.* Modelling of gene loss propensity in the pangenomes of three *Brassica* species suggests different mechanisms between polyploids and diploids. *Plant Biotechnol J*. 2021;**19**:2488–500.
16. Barchi L, Rabanus-Wallace MT, Prohens J *et al.* Improved genome assembly and pan-genome provide key insights into eggplant domestication and breeding. *Plant J*. 2021;**107**:579–96.
17. Gao L, Gonda I, Sun H *et al.* The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat Genet*. 2019;**51**:1044–51.
18. Zhang X, Liu T, Wang J *et al.* Pan-genome of *Raphanus* highlights genetic variation and introgression among domesticated, wild, and weedy radishes. *Mol Plant*. 2021;**14**:2032–55.
19. Sayers EW, Beck J, Bolton EE *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 2021;**49**:D10–7.
20. Song X, Nie F, Chen W *et al.* Coriander Genomics Database: a genomic, transcriptomic, and metabolic database for coriander. *Hortic Res*. 2020;**7**:55.
21. Wu T, Feng S, Yang Q *et al.* Integration of the metabolome and transcriptome reveals the metabolites and genes related to nutritional and medicinal value in *Coriandrum sativum*. *J Integr Agric*. 2021;**20**:1807–18.
22. Yu HJ, Baek S, Lee YJ *et al.* The radish genome database (RadishGD): an integrated information resource for radish genomics. *Database (Oxford)*. 2019;**2019**:baz009.
23. Wei Q, Wang J, Wang W *et al.* A high-quality chromosome-level genome assembly reveals genetics for important traits in eggplant. *Hortic Res*. 2020;**7**:153.
24. Reyes-Chin-Wo S, Wang Z, Yang X *et al.* Genome assembly with in vitro proximity ligation data and whole-genome triplication in lettuce. *Nat Commun*. 2017;**8**:14953.
25. Rodriguez Del Rio A, Minoche AE, Zwickl NF *et al.* Genomes of the wild beets *Beta patula* and *Beta vulgaris* ssp. *maritima*. *Plant J*. 2019;**99**:1242–53.
26. Lehner R, Blazek L, Minoche AE *et al.* Assembly and characterization of the genome of chard (*Beta vulgaris* ssp. *vulgaris* var. *cicla*). *J Biotechnol*. 2021;**333**:67–76.
27. Garcia-Mas J, Benjak A, Sansevarino W. The genome of melon (*Cucumis melo* L.). *Proc Natl Acad Sci USA*. 2012;**109**:11872–7.
28. Harkess A, Zhou J, Xu C *et al.* The asparagus genome sheds light on the origin and evolution of a young Y chromosome. *Nat Commun*. 2017;**8**:1279.
29. Yang J, Moeinzadeh MH, Kuhl H *et al.* Haplotype-resolved sweet potato genome traces back its hexaploidization history. *Nature Plants*. 2017;**3**:696–703.
30. Hufnagel B, Marques A, Soriano A *et al.* High-quality genome sequence of white lupin provides insight into soil exploration and seed quality. *Nat Commun*. 2020;**11**:492.
31. Liu D, Yu L, Wei L *et al.* BnTIR: an online transcriptome platform for exploring RNA-seq libraries for oil crop *Brassica napus*. *Plant Biotechnol J*. 2021;**19**:1895–7.
32. Chen H, Wang T, He X *et al.* BRAD V3.0: an upgraded Brassicaceae database. *Nucleic Acids Res*. 2022;**50**:D1432–41.
33. Liu Z, Li N, Yu T *et al.* The Brassicaceae genome resource (TBGR): a comprehensive genome platform for Brassicaceae plants. *Plant Physiol*. 2022;**190**:226–37.
34. Fernandez-Pozo N, Menda N, Edwards JD *et al.* The Sol Genomics Network (SGN)—from genotype to phenotype to breeding. *Nucleic Acids Res*. 2015;**43**:D1036–41.
35. Zheng Y, Wu S, Bai Y *et al.* Cucurbit Genomics Database (CuGenDB): a central portal for comparative and functional genomics of cucurbit crops. *Nucleic Acids Res*. 2019;**47**:D1128–36.
36. Yang Y, Sun P, Lv L *et al.* Prickly waterlily and rigid hornwort genomes shed light on early angiosperm evolution. *Nat Plants*. 2020;**6**:215–22.
37. Yang J, Liu D, Wang X *et al.* The genome sequence of allopolyploid *Brassica juncea* and analysis of differential homoeolog gene expression influencing selection. *Nat Genet*. 2016;**48**:1225–32.
38. Chalhoub B, Denoeud F, Liu S *et al.* Plant genetics. Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. *Science*. 2014;**345**:950–3.
39. Barrera-Redondo J, Ibarra-Laclette E, Vázquez-Lobo A *et al.* The genome of *Cucurbita argyrosperma* (silver-seed gourd) reveals faster rates of protein-coding gene and long noncoding RNA turnover and neofunctionalization within *Cucurbita*. *Mol Plant*. 2019;**12**:506–20.
40. Sun H, Wu S, Zhang G *et al.* Karyotype stability and unbiased fractionation in the paleo-allotetraploid *Cucurbita* genomes. *Mol Plant*. 2017;**10**:1293–306.
41. Montero-Pau J, Blanca J, Bombarely A *et al.* De novo assembly of the zucchini genome reveals a whole-genome duplication associated with the origin of the *Cucurbita* genus. *Plant Biotechnol J*. 2018;**16**:1161–71.
42. Wang X, Wang H, Wang J *et al.* The genome of the mesopolyploid crop species *Brassica rapa*. *Nat Genet*. 2011;**43**:1035–9.

43. Cheng SP, Jia KH, Liu H *et al.* Haplotype-resolved genome assembly and allele-specific gene expression in cultivated ginger. *Hortic Res*. 2021;**8**:188.
44. Li HL, Wu L, Dong Z *et al.* Haplotype-resolved genome of diploid ginger (*Zingiber officinale*) and its unique gingerol biosynthetic pathway. *Hortic Res*. 2021;**8**:189.
45. Song X, Li Y, Hou X. Genome-wide analysis of the AP2/ERF transcription factor superfamily in Chinese cabbage (*Brassica rapa* ssp. *pekinensis*). *BMC Genomics*. 2013;**14**:573.
46. Song X, Nie F, Chen W *et al.* Coriander genomics database: a genomic, transcriptomic, and metabolic database for coriander. *Hortic Res*. 2020;**7**:55.
47. Moharana KC, Venancio TM. Polyploidization events shaped the transcription factor repertoires in legumes (Fabaceae). *Plant J*. 2020;**103**:726–41.
48. Shirasawa K, Hirakawa H, Fukino N *et al.* Genome sequence and analysis of a Japanese radish (*Raphanus sativus*) cultivar named 'Sakurajima Daikon' possessing giant root. *DNA Res*. 2020;**27**:dsaa010.
49. Song X, Wei Y, Xiao D *et al. Brassica carinata* genome characterization clarifies U's triangle model of evolution and polyploidy in *Brassica*. *Plant Physiol*. 2021;**186**:388–406.
50. Yue H, Nie X, Yan Z *et al.* N6-methyladenosine regulatory machinery in plants: composition, function and evolution. *Plant Biotechnol J*. 2019;**17**:1194–208.
51. UniProt C. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res*. 2021;**49**:D480–9.
52. Mistry J, Chuguransky S, Williams L *et al.* Pfam: the protein families database in 2021. *Nucleic Acids Res*. 2021;**49**:D412–9.
53. Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res*. 2021;**49**:D325–34.
54. Song X, Yang Q, Bai Y *et al.* Comprehensive analysis of SSRs and database construction using all complete gene-coding sequences in major horticultural and representative plants. *Hortic Res*. 2021;**8**:122.
55. Song X, Ge T, Li Y *et al.* Genome-wide identification of SSR and SNP markers from the non-heading Chinese cabbage for comparative genomic analyses. *BMC Genomics*. 2015;**16**:328.
56. Song X, Li N, Guo Y *et al.* Comprehensive identification and characterization of simple sequence repeats based on the whole-genome sequences of 14 forest and fruit trees. *For Res*. 2021;**1**: 1–10.
57. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol*. 2019;**20**:238.
58. Wang Y, Tang H, Debarry JD *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res*. 2012;**40**:e49.
59. Chen C, Chen H, Zhang Y *et al.* TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Mol Plant*. 2020;**13**:1194–202.
60. Li H, Fan Y, Yu J *et al.* 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*. 2018;**19**:E3632.
61. Bouche F, Lobet G, Tocquin P *et al.* FLOR-ID: an interactive database of flowering-time gene networks in *Arabidopsis thaliana*. *Nucleic Acids Res*. 2016;**44**:D1167–71.
62. Aach J, Mali P, Church GM. CasFinder: flexible algorithm for identifying specific Cas9 targets in genomes. *bioRxiv*. 2014;005074.
63. Wang H, Wu J, Sun S *et al.* Glucosinolate biosynthetic genes in *Brassica rapa*. *Gene*. 2011;**487**:135–42.
64. Cheng F, Wu J, Wang X. Genome triplication drove the diversification of *Brassica* plants. *Hortic Res*. 2014;**1**:14024.
65. Li P, Quan X, Jia G *et al.* RGAugury: a pipeline for genome-wide prediction of resistance gene analogs (RGAs) in plants. *BMC Genomics*. 2016;**17**:852.
66. Sekhwal MK, Li P, Lam I *et al.* Disease resistance gene analogs (RGAs) in plants. *Int J Mol Sci*. 2015;**16**:19248–90.
67. Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. *Nucleic Acids Res*. 2019;**47**:D155–62.
68. Chen C, Li J, Feng J *et al.* sRNAanno—a database repository of uniformly annotated small RNAs in plants. *Hortic Res*. 2021;**8**:45.
69. Lorenz R, Bernhart SH, Siederdissen HZ *et al.* ViennaRNA package 2.0. *Algorithms Mol Biol*. 2011;**6**:26.
70. Kielbasa SM, Bluthgen N, Fahling M *et al.* Targetfinder.org: a resource for systematic discovery of transcription factor target genes. *Nucleic Acids Res*. 2010;**38**:W233–8.
71. Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics*. 2009; Chapter 4, Unit 4.10;25.
72. Giannoulatou E, Park SH, Humphreys DT *et al.* Verification and validation of bioinformatics software without a gold standard: a case study of BWA and bowtie. *Bioinformatics*. 2014;**15**:S15.
73. Yu T, Bai Y, Liu Z *et al.* Large-scale analyses of heat shock transcription factors and database construction based on whole-genome genes in horticultural and representative plants. *Hortic Res*. 2022;**9**:uhac035.