Agriculture Communications 2 (2024) 100041

Contents lists available at ScienceDirect

Agriculture Communications

journal homepage: www.sciencedirect.com/journal/agriculture-communications

The genomic database of fruits: A comprehensive fruit information database for comparative and functional genomic studies

Jingyi Liu [a,1], Chenchen Huang [a,1], Dingsheng Xing [a], Shujing Cui [a], Yanhong Huang [a], Can Wang [a], Ruohan Qi [a], Zhuo Liu [a], Rong Zhou [b], Xiao Ma [a,c,**], Xiaoming Song [a,*]

[a] School of Life Sciences/Library, North China University of Science and Technology, Tangshan 063210, China
[b] Department of Food Science, Aarhus University, Aarhus 8200, Denmark
[c] College of Horticultural Science & Technology, Hebei Normal University of Science & Technology, Qinhuangdao 066600, China

HIGHLIGHTS

- The Genomic Database of Fruits is the first large-scale collection of fruit genomic data and bioinformatics analysis results.
- Based on whole-genome analyses of 44 fruit species, we constructed a comprehensive, user-friendly fruit database called "The Genomic Database of Fruits" (http://tgdf.bio2db.com/).

ARTICLE INFO

Keywords:
The genomic databases of fruits
Fruit
Functional gene
Gene annotation
CRISPR
Tool

ABSTRACT

Fruit has an important role in human nutrition and health; therefore, the systematic study of fruit genomic data is essential. The Genomic Database of Fruits (TGDF, http://tgdf.bio2db.com/), established through whole-genome analyses of 44 fruit species, is a comprehensive, user-friendly fruit database. TGDF contains a wealth of functional genes, including 11,350 flowering genes, 3161 auxin signaling genes, 2164 anthocyanin synthesis genes, 1464 abscisic acid (ABA) synthesis genes, 10,931 cell division and expansion genes, 1786 starch synthesis genes, 294 fruit size genes, and 6311 sugar transporter genes. Additionally, TGDF contains 1,433,368 CRISPR guide sequences from various fruit genomes, along with information on homologous genes and duplication types for the 44 fruit species. TGDF contains 6,417,060 gene annotations sourced from TrEMBL, SwissProt, Nr, and Gene Ontology databases, along with tools such as Sequence Fetch, BLAST, Synteny, and JBrowse for bioinformatics analyses. Transcriptomic data were also collected and collated from fruits, including details on instruments, tissues, or growth stages. This comprehensive, user-friendly resource is the first collection of fruit genomic data. Users can easily download genomic sequences, gene annotations, and bioinformatics analysis results from TGDF, which will be updated continually. We anticipate that TGDF will become a primary resource for fruit comparative and functional genomic studies.

1. Introduction

Fruit is rich in vitamins, carbohydrates, dietary fiber, minerals, carotenoids, and other nutritional sources, and plays an indispensable role in human nutrition and health [1–3]. Over the past decade, advances in sequencing technologies and reductions in sequencing costs have led to the sequencing of numerous plant genomes. In particular, the emergence of third-generation sequencing technologies has substantially increased plant genomic data quantity and quality [4]. Grape (*Vitis vinifera*), the first crop to be sequenced [5], marked the beginning of the sequencing of fruit genomes and the realization that genome sizes vary greatly across fruit species. For instance, the genome size of Gansu peach (*Prunus kansuensis*), sequenced in 2022, is 198 Mb, while the genome size of grape (*V. vinifera*), sequenced in 2007, is 1.03 GB [5,6], with the genome sizes of other fruit species falling within this interval.

* Corresponding author.
** Corresponding author.
E-mail addresses: maxiaoxiaos@sina.com (X. Ma), songxm@ncst.edu.cn (X. Song).
[1] The authors contributed equally to this work.

https://doi.org/10.1016/j.agrcom.2024.100041
Received 25 December 2023; Received in revised form 7 March 2024; Accepted 28 March 2024
Available online 24 May 2024

The analysis of fruit genomes offers important resources for comparative genomics, functional genomics, and molecular biology studies. However, most fruit genome sequences lack gene sets and annotations. Currently, the databases of several fruits are available, including the Citrus Genome Variation Database (http://citgvd.cric.cn) [7], Citrus Pangenomic Breeding Database (http://citrus.hzau.edu.cn/) [8], Persimmon Genome Database (http://www.persimmongenome.cn), Kiwifruit Genome Database (http://kiwifruitgenome.org/) [9], Grape Genome Database (http://www.grapegenomics.com/) [10], Bayberry Database (http://www.bayberrybase.cn/) [11], Pineapple Genome Database (http://pineapple.angiosperms.org/pineapple/html/index.html) [12], Pear Genome Project (http://peargenome.njau.edu.cn/) [13], and Blueberry Genome Database (http://bioinformatics.towson.edu/BBGD/) [14]. In addition, family- and genus-level genome databases, such as the Rosaceae Genome Database (https://www.rosaceae.org) [15] and the Gourd Genomics Database (http://cucurbitgenomics.org/) [16], are also available.

Despite the existence of multiple fruit genome tools, a comprehensive, user-friendly fruit genome resource has not been available until now, which has hindered comparative and functional genomic studies. Therefore, we established "The Fruit Genome Database" (TGDF), a large-scale database for sharing fruit genome sequences, along with gene annotation information and transcriptomic data. TGDF, an open-access resource, sources genomic sequences primarily from the National Center of Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov) and Ensembl Plants (http://plants.ensembl.org) [17,18]. The aim of this endeavor is to provide a centralized platform for fruit genome data. In this study, we present an overview of its interface and related datasets, which will enable researchers to carry out comparative and functional genomic studies of various fruit species at the whole-genome level.

2. Materials and methods

2.1. Retrieval of fruit genome and gene annotation resources

We obtained genome sequence files, GFF sequence files, coding sequence (CDS) sequence files, and protein sequence files of 44 fruit species using different public databases: the National Center for Biotechnology Information Database (NCBI, https://www.ncbi.nlm.nih.gov/), Ensembl Plants (http://plants.ensembl.org), National Genomics Data Center (https://bigd.big.ac.cn/gwh) [19], Rosaceae Genome Database (https://www.rosaceae.org), and Gourd Genome Database (http://cucurbitgenomics.org/). In addition, several single genome databases were searched, such as the Kiwifruit Genome Database (http://kiwifruitgenome.org/), Pineapple Genomics Database (http://pineapple.angiosperms.org/pineapple/html/index.html), and Citrus Genome Database (Supplementary Table 1). We selected the most recent version or the most commonly used version of each genome as the study object. Genomic information for each fruit species, including genome size, gene number, sequencing information, genome database, and species classification, was collected. To eliminate redundant sequences, alternatively spliced sequences were removed using custom Perl scripts.

2.2. Gene function annotation

Swiss-Prot, TrEMBL [20], Gene Ontology (http://geneontology.org/) [21], and non-redundant protein sequence database (https://www.ncbi.nlm.nih.gov) were used to annotate the genes of the 44 fruit species, and the gene annotations were deposited into TGDF.

2.3. Functional gene identification

The anthocyanin synthesis genes, auxin signaling genes, flowering genes, abscisic acid (ABA) signaling genes, cell division and expansion genes, and starch synthesis genes of each target fruit species were searched and identified using BLASTp (E-value $<$ 1e-5; identity $>$60%; score $>$150). To identify candidate genes, we analyzed the sequences of auxin signaling genes, anthocyanin synthesis genes, flowering genes, and other genes from 44 fruit species, comparing them with sequences from other species. Given that the candidate genes are highly homologous to the functional genes of model species such as *Arabidopsis thaliana*, it is likely that these genes share similar functions across different species. After the candidate genes were screened, a domain analysis was performed to verify whether the candidate genes had the target domains.

2.4. CRISPR-Cas9 target design

CasFinder was used to design Cas9 targets for CRISPR [22], while RepeatMasker was used to screen repetitive genomic sequences in each species [23]. Subsequently, an index was created for each genome using Bowtie [24], and CasFinder pipeline scripts (CasValue_v2.pl and CasFinder.pl) were applied to CRISPR guide sequences [22]. Internal Perl scripts were then used to filter candidate sequences, yielding specific sequences for each gene.

2.5. Homologous gene identification

OrthoFinder (v2.0) was used to identify homologous genes. BLASTp (E value $<$ 1e-5) was used to determine protein sequence similarities between different species. Subsequently, cluster analysis was performed using the MCL algorithm (I $>$ 1.5), and trees were constructed using gene families from all species.

2.6. Collinearity and gene duplication types

Collinearity blocks within and between genomes were analyzed using MCScanX with default parameters [25]. Initially, BLASTp alignment was used to identify homologous sequences between and across genomes (E-value $<$ 1e-5). Subsequently, collinearity blocks were analyzed by comparing chromosome, gene, start position, and stop position columns in giff files with BLASTp results, and collinearity was detected using TBTools [26]. Finally, the duplication type was predicted using MCScanX (duplicate_gene_classifier) [25].

2.7. Database construction

TDGF was established using Django framework and MySQL database, after incorporating several programming languages such as HTML, CSS, JavaScript, R, and Python [27], and an interactive interface was established for the easy retrieval of information [28]. Diagrams are generated using Echarts, an open-source visualization library based on JavaScript, while data are stored and managed through server storage functions, with access rights set up to safeguard against data breaches. Additionally, Python and Perl scripts were used to process the genomic data, and various bioinformatics tools were used to explore the biological significance of each genomic dataset.

3. Results

3.1. Overview of fruit research

We collected genomic data from 44 fruit crops, including 1 monocotyledon and 43 dicotyledons (Supplementary Fig. 1a). Most dicots belonged to the Rosaceae family, with a total of 20 species. Most fruit crops were sequenced between 2017 and 2019, representing 54.55% of all fruit genomes (Fig. 1a; Supplementary Fig. 1b). We also collected related genomic information for these crops (Supplementary Table 1).

3.2. Overview of the homepage

We conducted a comprehensive bioinformatics analysis of the genomic data of these fruit species, which included identifying major

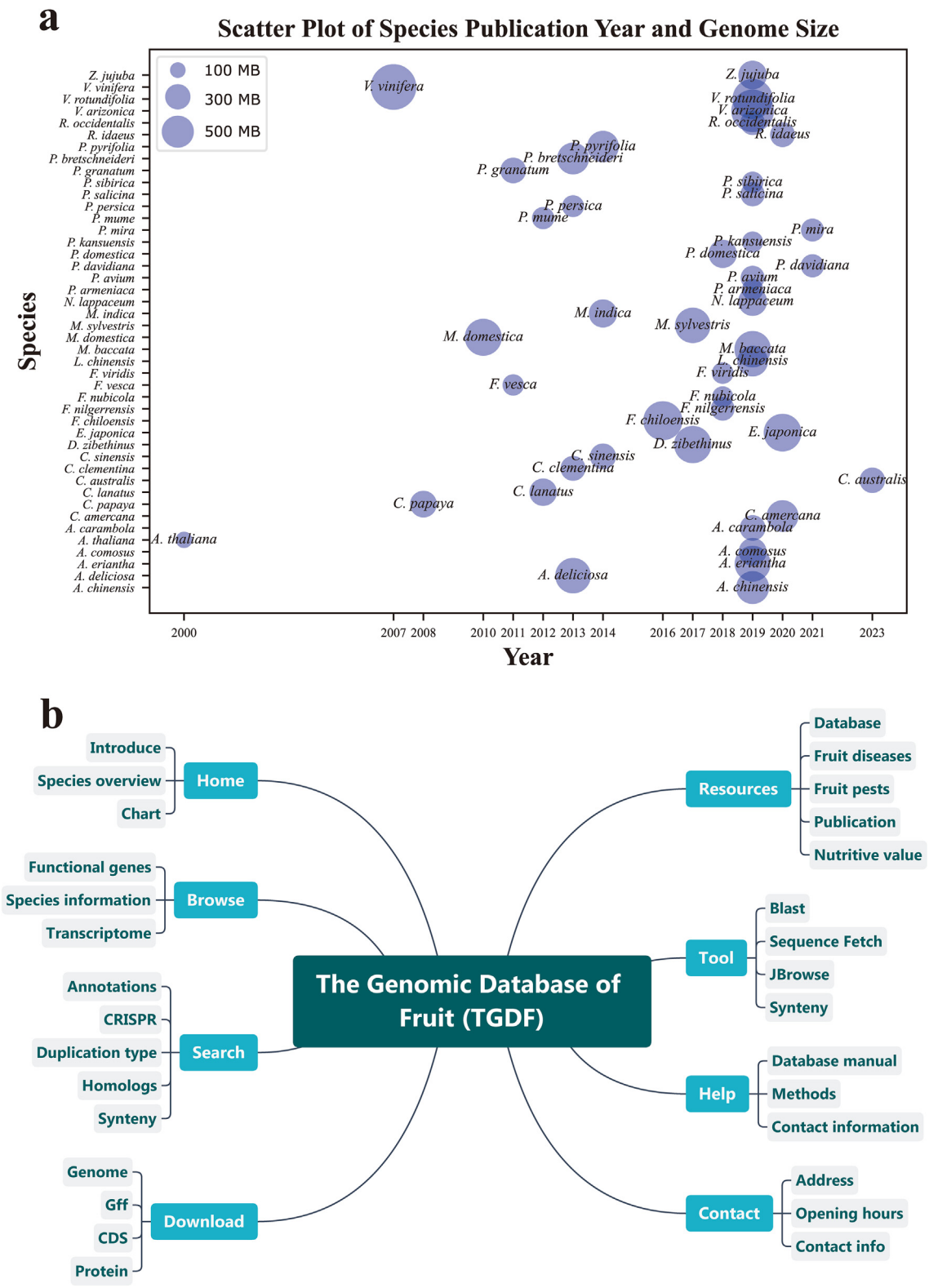

Fig. 1. Overview of fruit genome sequencing and TGDF database creation. (a) Species information and major genome sequencing indicators (publication year, assembled genome size, and sequencing technology) of the 44 fruit crops from the years 2007–2023. (b) Architecture of 'The Genomic Database of Fruits' (TGDF) database.

functional genes, annotating genes, designing specific guide sequences for CRISPR, analyzing homologous genes, and characterizing repetitive types. Functional genes of interest included those involved in auxin signaling, anthocyanin synthesis, and flowering.

Finally, we established TGDF to help users access these genome resources, along with querying the bioinformatics analysis results (Fig. 1b). All genomic information is stored in the database's backend, and users can easily access this information through the database's frontend. Here,

we provide descriptions of the homepage, browsing interface, search function, download feature, resource repository, tools, and contact page (Fig. 1b; Fig. 2).

3.3. Home interface

The TGDF homepage consists of three sections: brief introduction, species overview, and chart section. The species information is organized by family and species, and each fruit crop is associated with its Latin name, acronym, common name, species classification, genome size, chromosome number, and species picture. This provision helps users to understand the different types of fruits. Additionally, we provide trees and diagrams of the important functional genes in each fruit species. Histograms display the number of auxin signaling, anthocyanin synthesis, and flowering genes for each fruit species, facilitating the comparison of functional genes across different species.

3.4. Search interface

In TGDF, we provide gene annotations, CRISPR guide sequences, homologues, and repetitive types of the 44 fruit species based on four protein databases (UniProt knowledgebases [SwissProt, TrEMBL], Nr, and GO), we annotated 51.19% (*Fragaria nubicola*) to 99.98% (*Prunus persica*) of all genes in each species (Fig. 3; Supplementary Table 2), and these annotations are available for download. In addition, we included the external links to the genomes of the 44 fruit species, which are accessible through UniProt, NCBI, and the comparative genomics platform.

To facilitate gene editing studies in various fruit species, we designed 1,433,368 CRISPR guide sequences for all genes and stored them in TGDF. The success rate for different species ranged from 25.20% (*P. kansuensis*) to 98.90% (*Actinidia eriantha*) (Fig. 3; Supplementary Table 3).

To clarify the genetic evolutionary relationships, we used OrthoFinder to analyze the homologous genes. Among the 44 fruit species, there were 58,616 orthogroups, with 21,412 species-specific groups (Supplementary Tables 4 and 5). We also displayed these groups in TGDF, along with the phylogenetic tree of each group. Species trees were constructed using homologous genes to reveal the evolutionary relationships among these fruit species.

To explore gene duplication or loss after whole genome repeat (WGD) events, we conducted a collinearity analysis of amino acid sequences within and between two species. For the 44 fruit species, we detected 5 duplication types in each gene: dispersed, singleton, proximal, tandem, and whole genome replication (WGD)/segment (Fig. 3; Supplementary Table 6). WGD/segment duplication was predominant in several species that experienced WGD events, accounting for the highest proportion of duplication types in *Prunus domestica* (83.98%), followed by *V. vinifera* (68.81%) and *Actinidia chinensis* (67.07%).

3.5. Browser interface

The functional genes identified from the whole genomes of 44 fruit crops, including genes related to flowering, anthocyanin synthesis, auxin signaling, ABA signaling, cell division and expansion, and starch synthesis, are available in TGDF (Fig. 4; Supplementary Table 7). Species information of related fruit species is also provided.

In TGDF, users can access species information for all fruits, organized by family and species, including taxonomic identifiers, common names, Wikipedia links, and species overviews, which surpasses the content available in other fruit crop interfaces. In addition to the genomic data, we collected the transcriptomic data of the 44 fruit species. Our interface provides users with login numbers, sample information, instruments used, tissue details, and table-version information, all of which facilitate searches and downloads. Users can also download tables, which provide valuable information on gene expression profiles in various fruit species.

3.6. Download interface

TGDF includes genome-related datasets, including genome sequences, GFF files, coding sequences [CDS], and protein sequences, from 44 fruit species. These genome datasets are accessible through our interface and can be easily downloaded for comparative genomic analysis of fruit crops.

3.7. Tools interface

TGDF is equipped with various tools, such as BLAST, Sequence Fetch, Synteny, and JBrowse, which assist users in performing comparative and functional genomic analyses. Sequence Fetch enables rapid retrieval of target sequence information, offering users the convenience of browsing, searching, and downloading relevant information, thus expediating bioinformatic analysis.

In TGDF, we created a user-friendly interface with BLAST using CDS and protein sequences from the 44 fruit species. Users can seamlessly perform sequence alignments by either uploading a file in Fasta format or copying the sequence directly into the designated box.

Synteny offers two collinearity analysis tools: Toolkit (MCscanX) and Python MCscan. With Python MCscan users can upload BED and CDS files online for collinearity analysis. Configuration files ("layout" and "seqids") can be used in consultation with the Python MCscan manual. For the database schema, users select 2 or 4 species to carry out collinearity analysis. For MCscanX, users upload BLAST or GFF files for collinearity analysis. Different configuration files yield four collinearity types: point plotters, circular plotters, bar plotters, and dual synchronous plotters.

Additionally, we incorporated the JBrowse tool to display fruit crop characteristics and gene sequences. This tool allows users to examine the details of the selected genes on the corresponding chromosome sequence.

3.8. Resources and contact information

TGDF provides comprehensive information on fruit crops, common fruit diseases and their treatments, relevant publications, fruit nutritional values, and much more. Users can easily access these resources and quickly gain insights into their research status. In addition, contact information, such as email addresses, is provided so that users can easily contact the authors.

4. Discussion

In recent years, as the genomes of more fruit species have been sequenced, specific databases, such as PGD for pineapple [12], KGD for kiwifruit [9], CitGVD for citrus [7], and DGD for persimmon, have emerged. While these databases offer valuable information for the study of fruit genomes, they typically only contain one or a few closely related species. On the other hand, TGDF integrates the information from these databases and contains bioinformatics analysis results of genomes from 44 fruit species. Therefore, this integration offers several advantages. Firstly, TGDF is the only platform containing genomic data from 44 fruit species. Secondly, we provide CRISPR guide sequences, identify functional genes, and annotate genes for fruit crops. Therefore, these aspects not only address current challenges in plant research but also strengthen comparative and functional genomic studies in fruit crops.

Additionally, TGDF provides information on gene duplication types across the 44 fruit species. We also identified homologous genes within or between the genomes, constructing species trees and storing the information in TGDF. Finally, TGDF contains resources for the major fruits, helping users to understand the current status of research and to obtain information on target genes for further study.

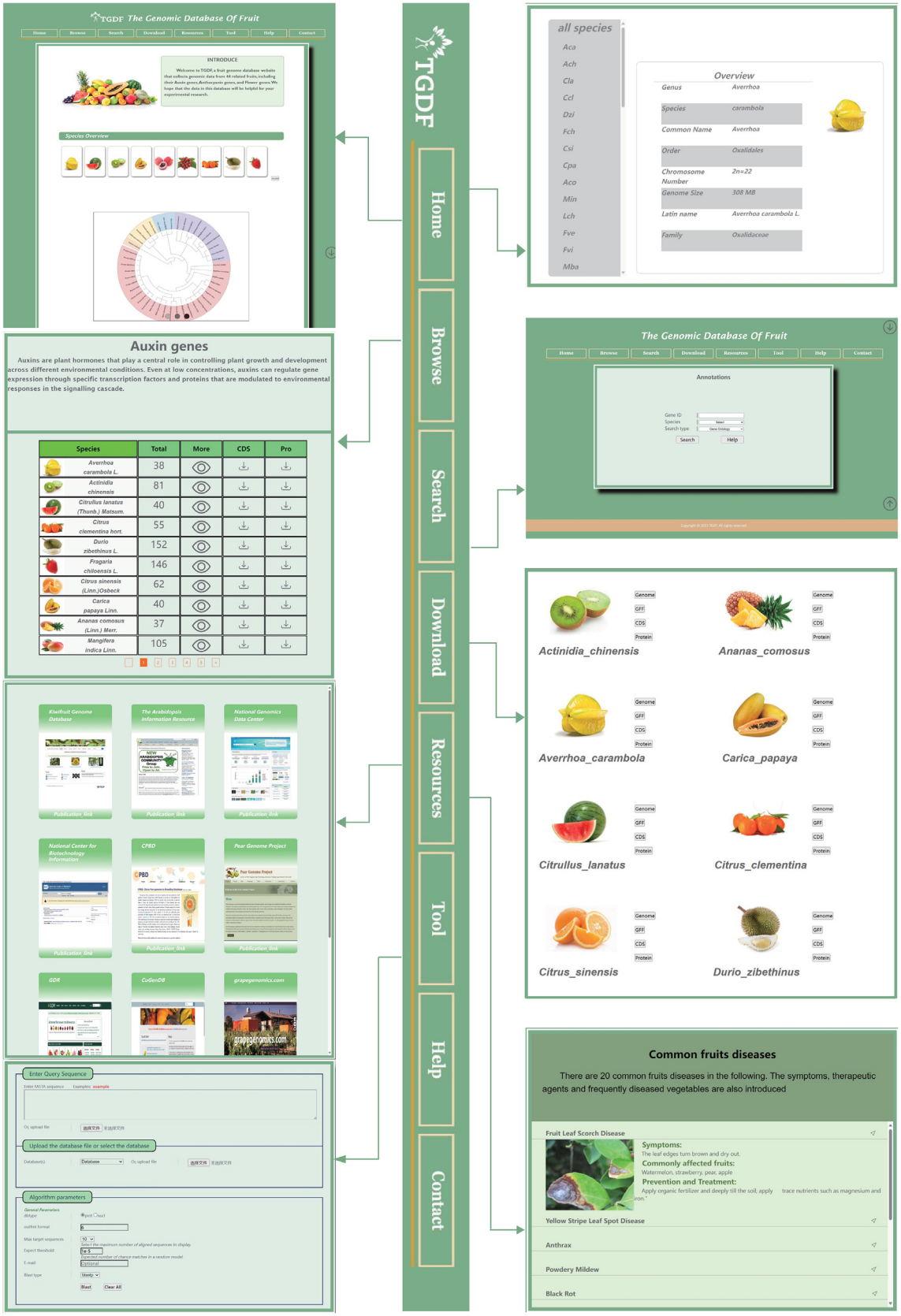

Fig. 2. Overview of TGDF showing the homepage and other features, including home, browse, resources, search, tool, download, help, contact information interfaces.

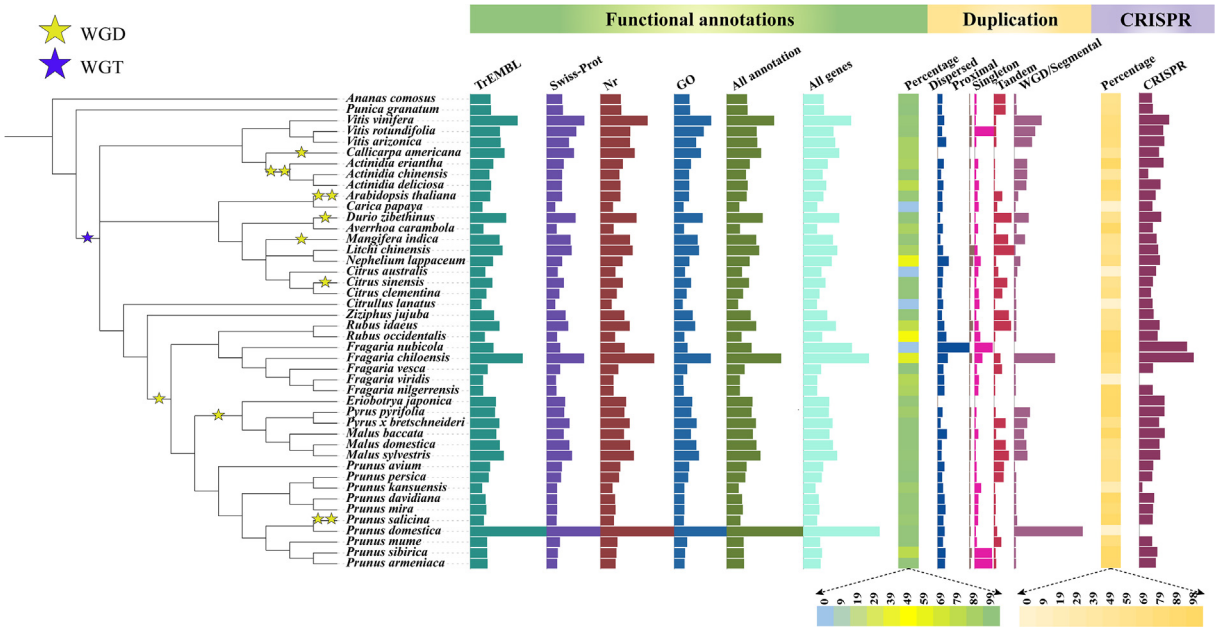

Fig. 3. Bar plots of the number of gene functional annotations, gene duplication types, and CRISPR guide sequences in the 44 fruit species. Whole-genome duplication (WGD) and whole-genome triplication (WGT) events are indicated by yellow and blue, respectively. The specific values are presented in Supplementary Tables 2, 3, and 6.

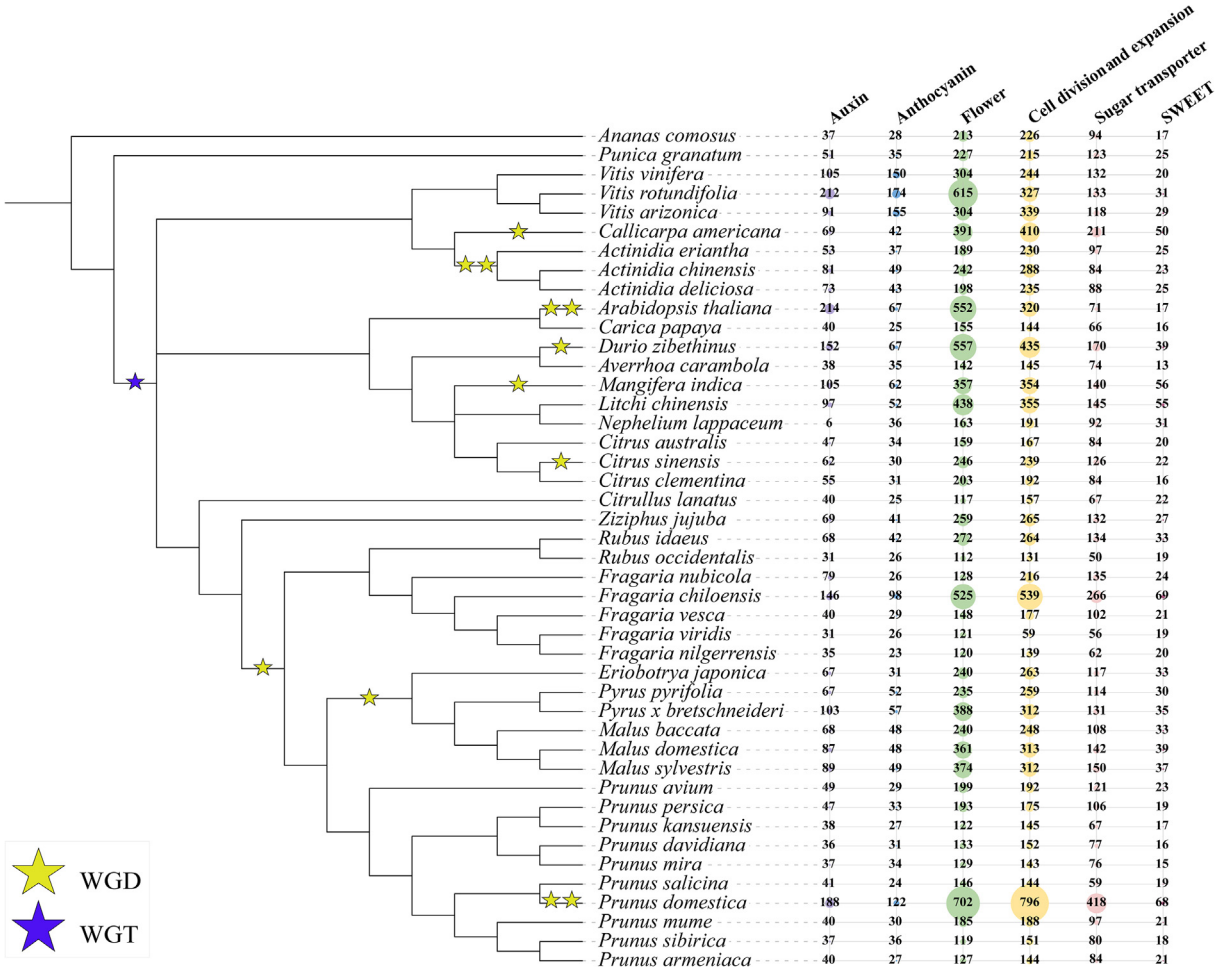

Fig. 4. Plot of the number of members of several functional gene families, including auxin signaling, anthocyanin synthesis, flowering, cell division and expansion, sugar transporter, and sweet genes in the 44 fruit crops. The numbers for each gene family were transformed by $\log_2$. The size of the circle represents the size of the value, and the specific values are presented in Supplementary Table 7.

5. Conclusion

We constructed The Genomic Database of Fruits (TGDF, http://tgdf.bio2db.com/), a user-friendly platform, by analyzing the genomes of 44 fruit species. TGDF contains functional genes, CRISPR guide sequences, and bioinformatics analysis results from 44 fruit species. We anticipate that TGDF will emerge as an important resource for comparative and functional genomic studies in fruit crops in the future.

Declaration of interests

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

Acknowledgments

This work was supported by the Natural Science Foundation for National Natural Science Foundation of China (32172583), Distinguished Young Scholar of Hebei (C2022209010), Natural Science Foundation of Hebei (C2021209005), and Key Lab. of Nucleic Research, Tangshan (2022TS003b).

Authors' contributions

X.S. conceived the project and initiated the research. X.S., X.M., and J.L. supervised and managed the research. X.S., J.L., C.W, X.M., and R.Q. generated and collected the data. Bioinformatics analysis and database construction were led by X.S., J.L., C.H., D.X., S.C., Y.H., and Z.L. The manuscript was organized, written, and revised by X.S., J.L., X.M., and R.Z. All authors read and revised the manuscript.

Supplementary information

Supplementary information to this article can be found online at https://doi.org/10.1016/j.agrcom.2024.100041.

References

[1] Ma X, Chang Y, Li F, Yang J, Ye L, Zhou T, et al. CsABF3-activated CsSUT1 pathway is implicated in pre-harvest water deficit inducing sucrose accumulation in citrus fruit. Hortic Plant J 2024;10(1):103–14.
[2] Wang W, Yu J, Du M, Wang J, Hu D. Basic helix-loop-helix (bHLH) transcription factor MdbHLH3 negatively affects the storage performance of postharvest apple fruit. Hortic Plant J 2022;8(6):700–12.
[3] Chen H, Ji H, Zhu S, Zhu K, Ye J, Deng X. Carotenoid and transcriptome profiles of a novel citrus cultivar 'Jinlegan' reveal mechanisms of yellowish fruit formation. Hortic Adv 2023;1(1):5.
[4] Mei Y, Jing D, Tang S, Chen X, Chen H, Duanmu H, et al. InsectBase 2.0: a comprehensive gene resource for insects. Nucleic Acids Res 2021;50(D1):D1040–5.
[5] Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 2007;449(7161):463–7.
[6] Cao K, Peng Z, Zhao X, Li Y, Liu K, Arus P, et al. Chromosome-level genome assemblies of four wild peach species provide insights into genome evolution and genetic basis of stress resistance. BMC Biol 2022;20(1):139.
[7] Li Q, Qi J, Qin X, Dou W, Lei T, Hu A, et al. CitGVD: a comprehensive database of citrus genomic variations. Hortic Res 2020;7:12.
[8] Liu H, Wang X, Liu S, Huang Y, Guo YX, Xie WZ, et al. Citrus Pan-Genome to Breeding Database (CPBD): a comprehensive genome database for citrus breeding. Mol Plant 2022;15(10):1503–5.
[9] Yue J, Liu J, Tang W, Wu YQ, Tang X, Li W, et al. Kiwifruit Genome Database (KGD): a comprehensive resource for kiwifruit genomics. Hortic Res 2020;7:117.
[10] Minio A, Massonnet M, Figueroa-Balderas R, Vondras AM, Blanco-Ulate B, Cantu D. Iso-seq allows genome-independent transcriptome profiling of grape berry. Development. G3 (Bethesda) 2019;9(3):755–67.
[11] Ren H, He Y, Qi X, Zheng X, Zhang S, Yu Z, et al. The bayberry database: a multiomic database for Myrica rubra, an important fruit tree with medicinal value. BMC Plant Biol 2021;21(1):452.
[12] Xu H, Yu Q, Shi Y, Hua X, Tang H, Yang L, et al. PGD: pineapple genomics database. Hortic Res 2018;5:66.
[13] Wu J, Li L-T, Li M, Khan MA, Li X-G, Chen H, et al. High-density genetic linkage map construction and identification of fruit-related QTLs in pear using SNP and SSR markers. J Exp Bot 2014;65(20):5771–81.
[14] Alkharouf NW, Dhanaraj AL, Naik D, Overall C, Matthews BF, Rowland LJ. BBGD: an online database for blueberry genomic data. BMC Plant Biol 2007;7:5.
[15] Jung S, Lee T, Cheng CH, Buble K, Zheng P, Yu J, et al. 15 years of GDR: new data and functionality in the genome database for Rosaceae. Nucleic Acids Res 2019;47(D1):D1137–d1145.
[16] Zheng Y, Wu S, Bai Y, Sun H, Jiao C, Guo S, et al. Cucurbit Genomics Database (CuGenDB): a central portal for comparative and functional genomics of cucurbit crops. Nucleic Acids Res 2019;47(D1):D1128–d1136.
[17] Bolser DM, Staines DM, Perry E, Kersey PJ. Ensembl plants: integrating tools for visualizing, mining, and analyzing plant genomic data. Methods Mol Biol 2017;1533:1–31.
[18] Sayers EW, Beck J, Bolton EE, Bourexis D, Brister JR, Canese K, et al. Database resources of the national center for Biotechnology information. Nucleic Acids Res 2021;49(D1):D10–d17.
[19] National Genomics Data Center Members and Partners. Database resources of the national genomics data center in 2020. Nucleic Acids Res 2020;48(D1):D24–d33.
[20] UniProt Consortium. UniProt: the universal protein knowledge base in 2021. Nucleic Acids Res 2021;49(D1):D480–d489.
[21] Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res 2021;49(D1):D325–d334.
[22] Aach J, Mali P, Church GM. CasFinder: flexible algorithm for identifying specific Cas9 targets in genomes. 2014.
[23] Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics 2009;Chapter 4:4.10.11–14.10.14.
[24] Giannoulatou E, Park SH, Humphreys DT, Ho JW. Verification and validation of bioinformatics software without a gold standard: a case study of BWA and Bowtie. BMC Bioinf 2014;15 Suppl 16(Suppl 16):S15.
[25] Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res 2012;40(7):e49.
[26] Chen C, Chen H, Zhang Y, Thomas HR, Frank MH, He Y, et al. TBtools: an integrative toolkit developed for interactive analyses of big biological data. Mol Plant 2020;13(8):1194–202.
[27] Feng S, Liu Z, Chen H, Li N, Yu T, Zhou R, et al. PHGD: an integrative and user-friendly database for plant hormone-related genes. iMeta 2024;3(1):e164.
[28] Yu T, Ma X, Liu Z, Feng X, Wang Z, Ren J, et al. TVIR: a comprehensive vegetable information resource database for comparative and functional genomic studies. Hortic Res 2022;9:uhac213.