Horticulture Research

Horticulture Research, 2024, **11**: uhae100

https://doi.org/10.1093/hr/uhae100

Article

HortGenome Search Engine, a universal genomic search engine for horticultural crops

Sen Wang[1,2,‡], Shangxiao Wei[1,2,‡], Yuling Deng[1,2], Shaoyuan Wu[1,2], Haixu Peng[1,2], You Qing[1,2], Xuyang Zhai[1,2], Shijie Zhou[1,2], Jinrong Li[1,2], Hua Li[1,2], Yijian Feng[1,2], Yating Yi[1,2], Rui Li[1,2], Hui Zhang[1,2], Yiding Wang[5], Renlong Zhang[5], Lu Ning[2,6], Yuncong Yao[1,*], Zhangjun Fei[3,4,*] and Yi Zheng[1,2,*]

[1]Beijing Key Laboratory for Agricultural Application and New Technique, College of Plant Science and Technology, Beijing University of Agriculture, Beijing 102206, China
[2]Bioinformatics Center, Beijing University of Agriculture, Beijing 102206, China
[3]Boyce Thompson Institute, Cornell University, Ithaca, NY 14853, USA
[4]USDA-ARS, Robert W. Holley Center for Agriculture and Health, Ithaca, NY 14853, USA
[5]College of Intelligent Science and Engineering, Beijing University of Agriculture, Beijing 102206, China
[6]Library, Beijing University of Agriculture, Beijing 102206, China
*Corresponding authors. E-mails: yaoyc_20@126.com; zf25@cornell.edu; yz@moilab.net
‡These authors contributed equally to the work.

Abstract

Horticultural crops comprising fruit, vegetable, ornamental, beverage, medicinal and aromatic plants play essential roles in food security and human health, as well as landscaping. With the advances of sequencing technologies, genomes for hundreds of horticultural crops have been deciphered in recent years, providing a basis for understanding gene functions and regulatory networks and for the improvement of horticultural crops. However, these valuable genomic data are scattered in warehouses with various complex searching and displaying strategies, which increases learning and usage costs and makes comparative and functional genomic analyses across different horticultural crops very challenging. To this end, we have developed a lightweight universal search engine, HortGenome Search Engine (HSE; http://hort.moilab.net), which allows for the querying of genes, functional annotations, protein domains, homologs, and other gene-related functional information of more than 500 horticultural crops. In addition, four commonly used tools, including 'BLAST', 'Batch Query', 'Enrichment analysis', and 'Synteny Viewer' have been developed for efficient mining and analysis of these genomic data.

Introduction

Horticultural crops comprise fruits, vegetables, floricultural and ornamental plants, as well as beverage, medicinal and aromatic plants, and have played critical roles in food supply, human health, and beautifying landscapes. With the growing human population, new demands are placed on the yield, quality, diversity, and nutritional value of horticultural crops. Decoding the genomes of horticultural crops not only provides an opportunity to investigate gene functions and regulatory networks [1, 2], but also serves as the cornerstone for functional and comparative genomics studies [3, 4] and paves a path to resolve complex quantitative trait loci (QTLs) of important horticultural traits [5]. Advanced genome editing technologies have been demonstrated in recent years to have a great potential for improving the quality and yield of horticultural crops [6], and reference genomes provide precise sequences for the application of genome editing technologies. Thus, genome sequencing plays a crucial role in horticultural crop improvement, and serves as an important foundation for understanding the history of crop domestication and evolution.

With the rapid advances of sequencing technologies, especially the PacBio HiFi long-read sequencing technology, various horticultural crop genomes have been deciphered, including those with high heterozygosity and polyploidy levels. To store, mine, and analyse the large-scale genomics data of horticultural crops, numerous databases have been developed, such as Sol Genomics Network (SGN), Genome Database for Rosaceae (GDR), Cucurbit Genomics Database (CuGenDB), among others [7–10]. These genomic databases integrate genomes, genes, and functional annotation information, as well as transcriptome and variome data, and implement widely used data mining and analysis tools such as BLAST, functional enrichment analysis, and genome browsers [11–14]. It is noteworthy that some plant genome databases extend the integration by including non-coding RNAs or other types of 'omics' data [15–17]. Additionally, some databases distinguish themselves by developing unique analytical tools tailored to specific research needs or data types, such as the 'Syntenic Gene @ Subgenome' and 'MicroSynteny' modules in BRAD [18], sRNA target prediction in SapBase [15], as well as CRISPR design and GWAS tools in CPBD [17]. Most of these databases manage genomic data for plants from a single family or species [19]. Therefore, the genomic resources of horticultural crops are scattered in different databases, and these databases exhibit different ways of presenting and utilizing results, resulting

Received: 1 December 2023; **Accepted:** 27 March 2024; **Published:** 8 April 2024; **Corrected and Typeset:** 1 June 2024

Figure 1. Search interface and result pages in HortGenome Search Engine. **a**–**e** Screenshots of the search interfaces. **f** Gene list of search results, including plant name, gene/transcript ID, genomic location and functional description.

in certain difficulties for users, especially in terms of searching tools that differ in complexity and functionality. This creates a learning curve for users seeking to search, browse, and conduct comparative analysis of genomic data across a broader range of plant species. In recent years, there has been an increasing focus on using search engines to explore the genetic makeup of plants [20]. This has proven to be an invaluable tool for researchers who are interested in studying plant genomics, functional genomics, and molecular assisted breeding. To this end, we have developed the HortGenome Search Engine (HSE; http://hort.moilab.net), a lightweight universal search engine for the genomic data of horticultural crops. Compared to other genomic databases, HSE stands out for its search engine-like interface that allows users to easily search genomic data without requiring prior knowledge. Currently, the searchable genomic data includes species information, gene sequences, comprehensive functional annotations, and homologous gene pairs. Currrently HSE contains data of 502 genome assemblies for horticultural crops covering fruit trees, vegetables, ornaments, and beverage plants, as well as model plant species, Arabidopsis and rice. In addition to

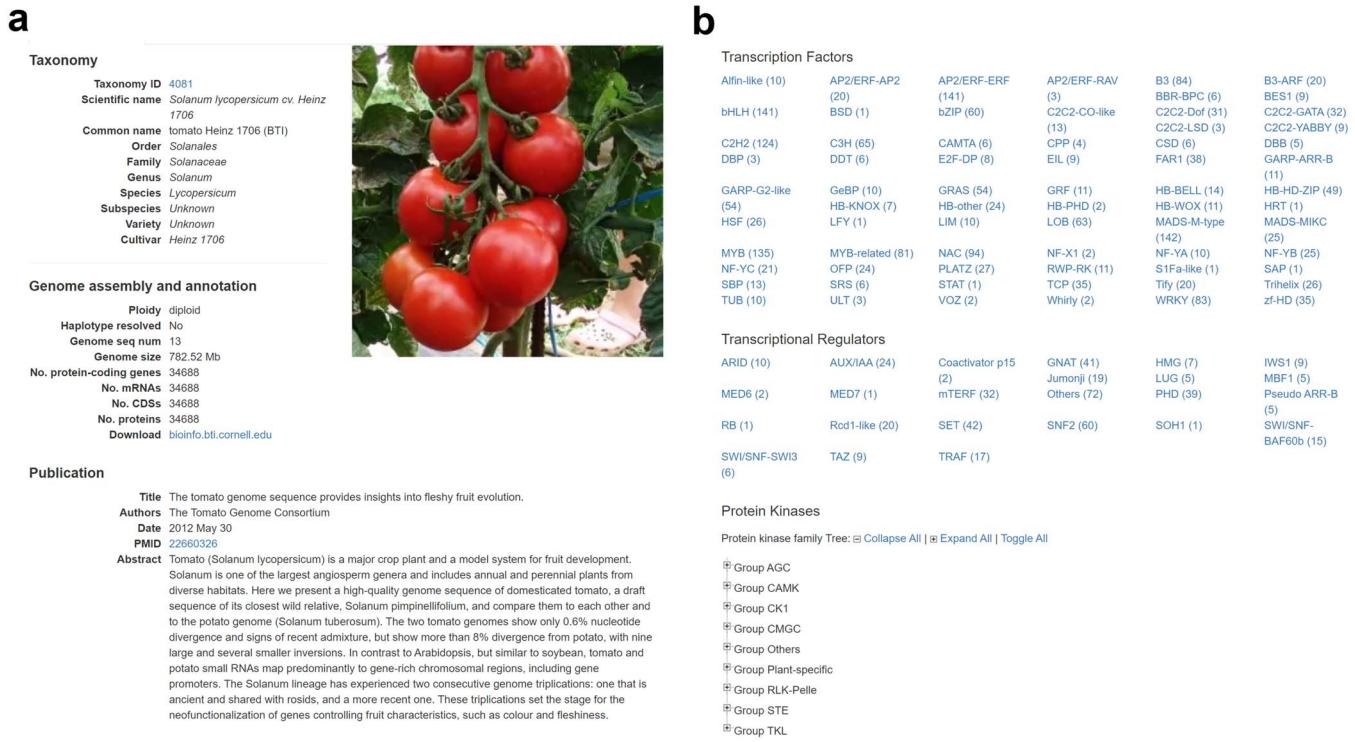

Figure 2. Genome page in HortGenome Search Engine **a** Screenshot of the genome page containing the genome information and picture of the plant. **b** Screenshot of the genome page containing transcription factors, transcriptional regulators, and protein kinases identified from the genome.

the searching function, several commonly used genomic data mining and analysis tools have been implemented in HSE, including 'BLAST', 'Batch Query', 'Enrichment analysis', and 'Synteny Viewer'.

Database contents and features

Preparation of genomic data

More than 1000 genome assemblies of nearly 800 plant species had been sequenced and published by the end of 2021 [21, 22]. Genomic data of horticultural crops, including the genome sequences, gene structure annotations in general feature format (GFF), and mRNA, coding (CDS) and protein sequences of protein-coding genes, were collected from plant genomics, comparative genomics, and plant family-specific databases, such as Phytozyme [23], Ensembl Plants [24], Genome Warehouse in National Genomics Data Center [25], SGN [7], GDR [8], and others. For some genome assemblies, only the genome sequences and GFF files are available; therefore, the corresponding mRNA, CDS, and protein sequences were extracted using the gffread program [26]. We further performed quality control on the collected genomic data to ensure the accuracy of the data to be included in the database. For example, genome assemblies that lack a GFF file or have an inaccurate GFF file in which the numbers of genes or gene IDs were inconsistent with the corresponding mRNA, CDS and protein sequence files were excluded. Finally, a total of 502 genome assemblies for horticultural crops, as well as the model plant species Arabidopsis and rice, were collected and included in the database (Table S1, see online supplementary material). Besides the genomic data, the taxonomy information, statistics of genome assemblies, associated publications, and images of the plant species have also been collected from the PlaBiPD database (https://www.plabipd.de/), published manuscripts, and other data sources, and included in the database.

Gene functional annotation

We used the pipeline described in our previous studies [9, 27] to generate comprehensive functional annotations for all protein-coding genes of the collected genome assemblies of horticulture plants. Briefly, protein sequences of the predicted genes were blasted against the GenBank non-redundant (nr), UniProt (TrEMBL and SwissProt), and Arabidopsis protein databases using DIAMOND [28] with an E-value cutoff of 1e-4. Based on the identified homologs from the UniProt and Arabidopsis protein databases, concise and informative functional descriptions were assigned to each gene using the AHRD program (https://github.com/groupschoof/AHRD). Protein sequences were further compared against the InterPro database using InterProScan [29] to identify functional protein domains. Transcription factors (TFs), transcriptional regulators (TRs), and protein kinases (PKs) were identified using the iTAK pipeline [30].

To generate GO and KEGG pathway annotations for functional enrichment analyses, protein sequences were compared against the EggNOG database using eggnog-mapper [31]. The assigned GO terms of genes/transcripts retrieved from the eggnog-mapper results were converted to the GO Annotation File (GAF) format. In the eggnog-mapper results, some non-plant KEGG pathways were assigned to plant genes/transcripts. For example, the tomato gene *Solyc09g008400*, which encodes a serine/threonine protein phosphatase 2A regulatory subunit protein, was assigned to map05165, the human papillomavirus infection pathway. These non-plant pathways were manually identified and removed from the eggnog-mapper results.

Synteny blocks and homologous gene pairs

Identifying synteny blocks and homologous gene pairs within or across genomes lays the groundwork for discovering and dating ancient genomic evolution events, as well as for inferring gene

Figure 3. Gene family page in HortGenome Search Engine. Screenshots of the list and download links **(a)**, locations on chromosomes **(b)**, structure **(c)**, and functional domains **(d)** of the tomato bZIP family genes.

functions [32]. Detection of synteny blocks among all the 502 genomes yield more than 120 000 pairwise genome comparisons. In addition, synteny blocks and gene pairs were also identified between any genome assemblies and their corresponding model plants (i.e., Arabidopsis for eudicot plants and rice for monocot plants). Briefly, the CDS of each genome were arranged in the

Figure 4. Gene feature page in HortGenome Search Engine. **a** Screenshot of the gene page containing basic information and gene structure. **b** Screenshot of the gene page containing gene, mRNA, CDS, and protein sequences. **c** Screenshot of the homolog genes and sequence alignments from the BLAST results. **d** Screenshot of the functional domains predicted from the protein sequence of the gene. **e** Screenshot of the GO terms assigned to the gene. **f** Screenshot of the gene page containing collinear gene pairs.

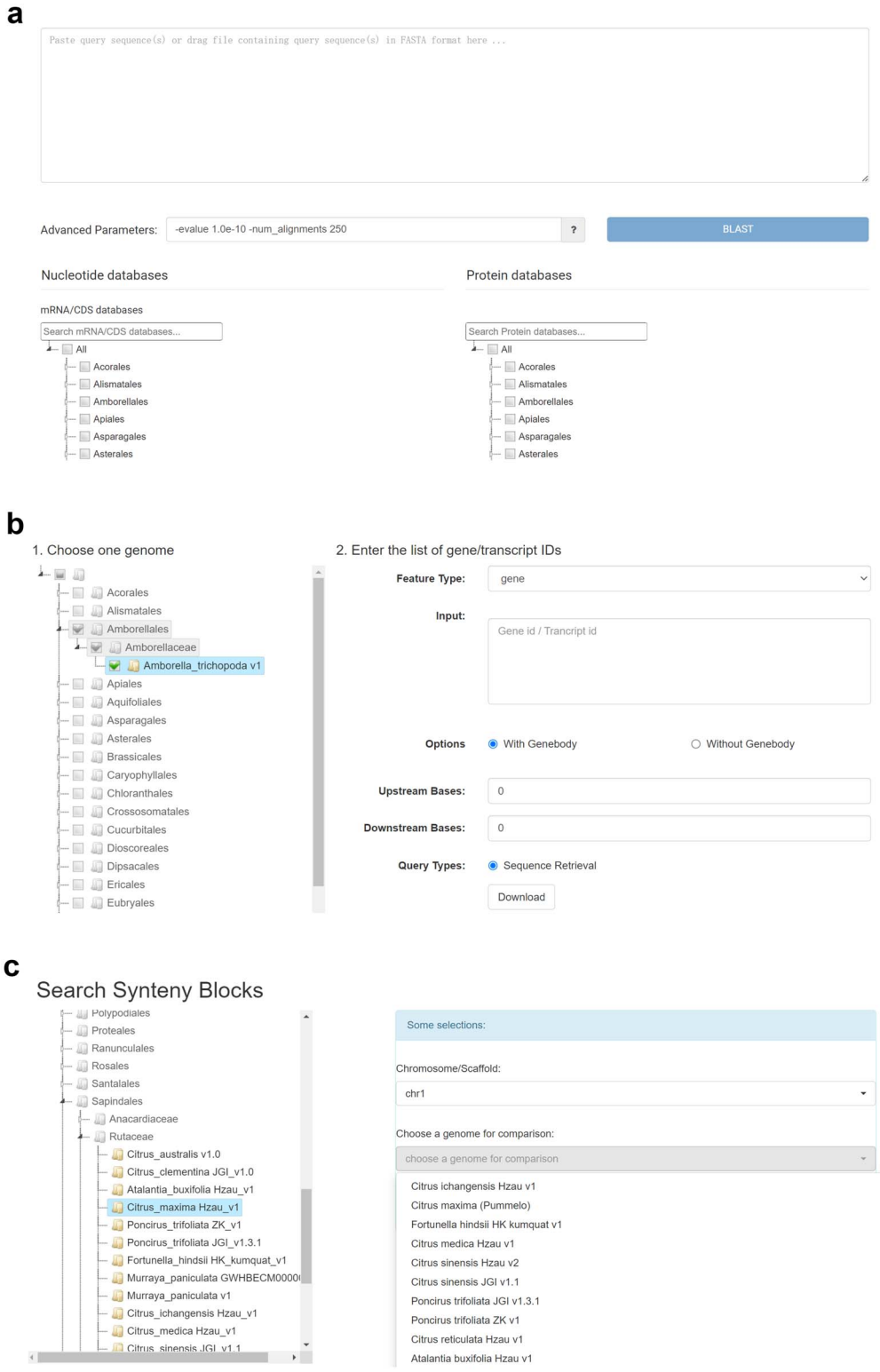

Figure 5. Query interfaces of data mining tools in HortGenome Search Engine. **a** Screenshot of the BLAST query page. **b** Search interface of 'Batch Query'. **c** Search interface of 'Synteny Viewer'.

order based on the GFF file, and then the CDS from different chromosomes, linkage groups, or scaffolds of the two compared genomes were aligned using the LASTZ program with default parameters. Syntenic blocks and homologous gene pairs were then identified using the python version of MCScanX [33], which implements a new BLAST filter to remove weak syntenic regions and tandem duplications [32]. In the end, more than 50 million syntenic blocks and about 1 billion homologous gene pairs were identified from pairwise genome comparisons and imported into the back-end database.

Data integration and indexing

Genome sequences, gene structures, and functional descriptions are imported into MongoDB, a popular NoSQL document database (https://www.mongodb.com/). Currently the database contains more than 40 million records of genes and transcripts from 502 genome assemblies. The top BLAST hits (homologs), GO terms, and InterPro domains assigned to each protein-coding gene have been imported into MongoDB, resulting in more than 150 million records in the database for searching. Indexing of gene/transcript IDs, functional descriptions, GO and Interpro terms, and TF/TR and protein kinase family names has been performed in the database, allowing for efficient search of large amounts of data. The interactive web interfaces have been developed using the Flask web framework and HTML.

Database functions

Search interface

To enhance user convenience in searching large-scale genomic data of horticultural crops, we have designed the search page to resemble popular search engines such as Google and Microsoft Bing. Multiple search methods have been streamlined into a single search box, thereby allowing users to search for genes of interest by entering various types of keywords and other related information, without requiring any prior experience or specialized training (Fig. 1a). Currently, the keywords could be the name of the species and gene, gene ID, the functional description of the gene, the family name of the transcription factor or protein kinase, or the GO or IPR ID. It is acknowledged that scientific names of crop species may be more challenging to enter accurately than common names. Additionally, it is often difficult for users to remember precise information, such as IDs for genes, GO, and IPR terms. To address this issue, we have implemented an auto-completion function for entering keywords. This feature prompts users with suggestions based on the information stored in the backend database after entering 2–3 characters, aiding in the accurate entry of information mentioned above. For example, when users search for tomato genetic information, they can use the common name 'tomato' or the Latin name *Solanum lycopersicum* for the query. When entering the first few characters, HSE will automatically prompt and complete the corresponding name for users to choose (Fig. 1b). After selecting species keywords, users can enter other keywords such as gene ID, gene name, gene functional description, etc (Fig. 1c–e).

The search returns a gene list with the corresponding species name, gene/transcript IDs, gene locations, and gene functional descriptions (Fig. 1f). The species name and gene/transcript IDs are linked to the corresponding genome page of the species and gene/transcript pages, respectively. In addition, if the user enters a keyword that combines the name of a species and the name of a specific TF/TR/PK family (Fig. 1e), the results will directly return to the corresponding gene family page of the species.

Genome page

The genome page displays basic information about the plant species and the genome assembly, and is comprised of three sections: taxonomy, genome assembly and annotation, and publication. The taxonomy section provides the scientific name, common name, and taxonomy information of the plant species, and the taxonomy ID is linked to the GenBank taxonomy database. The 'genome assembly and annotation' section shows the information about genome assembly size, the numbers of genome sequences, genes, mRNAs, CDS, and proteins, as well as the ploidy level information and the download link of the genome assembly. For the publication section, the title, authors, abstract, and publication date of the corresponding genome paper, which were automatically retrieved from PubMed according to the PubMed Identifier (PMID), are displayed (Fig. 2a).

On the genome page, an additional pagination is available to display the names and numbers of transcription factors, transcriptional regulators, and protein kinases identified for the selected genome (Fig. 2b). Clicking on a family name directs to the corresponding gene family page.

Gene family page

The gene family page displays homologous genes belonging to the same family, as well as gene location, structure, and functional domains. At present, only genes from the transcription factor, transcriptional regulator, and protein kinase families identified by iTAK [30] can be searched and displayed. For example, searching for the bZIP transcription factor of tomato will display all 60 bZIP genes identified in the genome. The page provides download links to retrieve the gene list, CDS, protein, and promoter sequences of these bZIP genes (Fig. 3a). The location of genes on chromosomes, gene structure, and protein functional domains are valuable information to study gene families. Therefore, the gene family page of HSE diaplays the images of gene location, structure, and protein functional domains for these homologous genes (Fig. 3b–d), which provide convenience for studying the function and evolution of the corresponding gene families.

Gene and transcript page

Each gene or transcript has a detailed feature page that contains all the related sequences and annotation information. The gene feature page forms different paginations based on the content types (Fig. 4). The overview pagination contains information about plant species, gene ID, location, strand, and functional description, as well as transcripts belonging to this gene. The gene structure is represented by its primary transcript and displayed using FeatureViewer [34] (Fig. 4a). The sequence pagination contains gene, mRNA (primary transcript), CDS, and protein sequences (Fig. 4b). In the BLAST pagination, it shows the top five homologs identified from the GenBank, UniProt, and TAIR databases, respectively. The BLAST hit accession IDs are linked to the corresponding databases, which allow users to access the expression, interaction, protein structure, and other information of the homologous genes from other databases. The detailed sequence alignment of the BLAST result is shown in a popup page when clicking the 'Show' link (Fig. 4c). The domain pagination lists the functional domains identified from the protein sequence of this gene (Fig. 4d). The gene ontology pagination lists the GO terms assigned to this gene and the GO IDs are linked to the AmiGO database, which provides details of the GO terms (Fig. 4e). The TF/TR/PK pagination shows the family name if the gene is identified as belonging to a specific TF/TR/PK family, which is linked to the corresponding gene family page. The syntelog pagination contains the collinear gene pairs and syntenic blocks related to this gene (Fig. 4f).

BLAST

We implemented the online BLAST tool, one of the most widely used tools in genome databases, using the SequenceServer [35]. In the query interface, the indexed genomes are organized in a hierarchical taxonomy display using jsTree (https://www.jstree.com/). The BLAST indexed databases are categorized into nucleotide

and protein databases. The nucleotide databases include the BLAST indexes for genome and mRNA/CDS sequences, and protein databases contain all indexes of protein sequences. With this interface, the BLAST search can be performed more flexibly (Fig. 5a). For example, by providing a DNA or protein sequence, the user can search against the sequences from a single plant species, or across the entire genus and family, or all plant species in the database. This provides a useful tool for studying gene function and evolution.

Batch Query

Genomic and functional genomic studies typically generate large lists of interesting genes, and retrieving nucleotide or protein sequences and functional annotations of these genes for downstream analyses is essential to understand the underlying biological processes. Similar to the online BLAST tool, a hierarchical taxonomy tree is provided in the 'Batch Query' interface for easily selecting the genome to be analysed. The query options will be changed dynamically according to the selected feature type. By selecting the 'gene' feature type, sequences containing exons, introns and the upstream and downstream sequences of a list of genes can be extracted (Fig. 5b). By selecting 'mRNA' or 'protein' feature type, in addition to extracting mRNA and protein sequences, the query also allows for retrieving functional descriptions, and family information for TFs, TRs, and PKs.

Enrichment analysis

Genomic and functional genomic analyses are capable of producing extensive lists of genes that are of interest. However, it is crucial to translate these lists into biologically relevant information to gain a deeper understanding of the underlying molecular mechanisms of the related biological processes. Enrichment analysis is a potent method that can be employed to identify classes of genes that are overrepresented in a list of genes. This approach enables the identification of highly dynamical biological processes or biochemical pathways under specific experimental conditions or developmental stages. In order to facilitate the enrichment analysis of gene and transcript data for hundreds of genomes, a hierarchical taxonomy tree has been constructed for the 'GO Enrichment Analysis' and 'KEGG Enrichment Analysis' tools, utilizing the same structure as that used in BLAST and 'Batch Query'. The 'GO Enrichment Analysis' tool has been implemented through the use of the Perl module GO::TermFinder, which employs the hypergeometric distribution test to determine enriched GO terms [36]. Similarly, the 'KEGG Enrichment Analysis' tool has been developed using KEGG pathways assigned to genes via eggnog-mapper, with enrichment significance calculated through the hypergeometric distribution test. The resulting enrichment analysis output page provides a list of enriched GO terms and KEGG pathway names, with links to the relevant GO and KEGG databases [37, 38]. Additionally, genes corresponding to each enriched GO term or KEGG pathway are included with links to relevant gene pages in HSE. Overall, GO and KEGG enrichment analyses are essential tools for the interpretation of genomic and functional genomic data, and their use is critical for advancing our understanding of complex biological systems.

Synteny Viewer

We have previously developed 'Synteny Viewer' as an extension module of Tripal to view genome synteny and homologous gene pairs between different cucurbit genomes [9]. The tool has been adopted by many genome databases, including Genome Database for Rosaceae (https://www.rosaceae.org) [8], ZEAMAP (http://zeamap.com) [39], etc. In HSE, the 'Synteny Viewer' has been re-implemented using Python/FLASK for managing the large amount of comparative genomic data generated from hundreds of plant genomes. To facilitate the search of a massive amount of synteny blocks and homologous gene pairs, the genome selection form is designed with genomes well organized through a hierarchical taxonomy tree. The chromosome/scaffold selection drop-down list and the compared genome drop-down list will be automatically updated according to the selected genome (Fig. 5c). The search result provides a circos plot that displays synteny blocks for query and compared chromosomes/scaffolds. Each synteny block is linked to a complete list of homologous gene pairs within the block, and each gene is linked to the detailed gene feature page mentioned above.

Using HSE to identify tomato TCP TFs

The TCP transcription factors are crucial regulatory proteins in plants that are involved in regulating plant morphology and structure by modulating pathways such as cell proliferation and hormone responses [40]. In tomato, a previous study reported the identification of 30 candidate TCP genes through BLAST searches against genes in the tomato genome (version SL2.40) and expressed sequence tags (ESTs) using Arabidopsis TCP proteins or TCP domains as queries [41]. Utilizing the 'Gene family page' in HSE grants immediate access to the 35 TCP transcription factors in the version SL4.0 of the tomato genome (Table S2, see online supplementary material). Of these TCP transcription factors, 29 align with those previously identified by Parapunova *et al.* [41]. The one TCP gene uniquely identified in Parapunova *et al.* [41] is obsolete in the newer version of the tomato genome (SL4.0). A detailed examination of the six TCP genes uniquely identified in HSE through the gene page confirms that these genes indeed possess the TCP functional domains, validating their classification as TCP transcription factors. Of these six TCP TFs, two are newly annotated in the version SL4.0 of the tomato genome, while the remaining four failed to be identified by Parapunova *et al.* [41]. Therefore, with the intuitive HSE search feature, users can quickly access the precise details of genes from specific families, which helps to streamline the research by minimizing redundant data analysis tasks. Additionally, the 'Gene family page' offers a wealth of essential information for transcription factor and protein kinase families, including details on genome positioning, functional domains, and gene structure illustrations, all of which serve as crucial resources for comprehensive gene family studies.

Conclusions and future directions

We have developed a universal search engine, HSE, that allows querying genes, functional annotations, and homologous gene pairs for hundreds of genomes of horticultural crops. More than 19 million genes with comprehensive functional annotations as well as ~50 million synteny blocks and 1 billion homologous gene pairs from 502 genome assemblies are stored in NoSQL document-oriented database for searching. It is worth mentioning that multiple indexes have been established on the document-oriented database to facilitate users to search genes in a more flexible way through a simple search box, which sets HSE apart from other plant genomic databases. Furthermore, several popular data mining tools of genomic databases have been implemented in HSE, including enrichment analysis of GO terms and KEGG pathways, 'Batch Query' for retrieving gene sequences and functional annotations, 'Synteny Viewer', and BLAST.

We will continue to collect genomic data of horticultural crops for HSE. HSE will be updated every six months or new horticultural genomes are available. In addition, users can submit genomes to HSE through the online genome collection forms in a timely manner (http://hort.moilab.net/genome/submit). In the future, we will expand the scope of data search to cover other omics data such as gene regulatory networks, gene expression, methylation, genotype, and phenotype. Furthermore, additional online data mining and visualization tools based on the horticultural crop genomes will be implemented in HSE.

Acknowledgements

We thank Bo Yuan at Lianzhi Technology Co., Ltd for assistance with computing acceleration. This work was supported by grants from the Beijing University of Agriculture (Start-up fund) to Y.Z., Young Teachers' Research and Innovation Capacity Enhancement Program QJKC2022044 and Beijing Municipal Education Commission Scientific Research Plan Project KM202310020010 to S.Wang. The computing power was supported by the Alibaba Cloud.

Author contributions

Z.F. and Y.Z. designed the project. S.Wei, Y.D., S.Wu, H.P., X.Z., S.Z., J.L., H.L., Y.F., Y.Yi, R.L., H.Z., Y.W., R.Z., L.N., and Y.Z. performed data collection. S.Wei, Y.Q., H.P., and S.Wang performed data analysis. S.Wei and Y.Z. wrote the code for database construction. Y.Yao, Z.F., and Y.Z. supervised the project and wrote the manuscript. All authors read and approved the final manuscript.

Data availability

All datasets have been made publicly available at http://hort.moilab.net/.

Conflict of interest statement

The authors declare that they have no conflict of interest.

Supplementary data

Supplementary data is available at *Horticulture Research* online.

References

1. Nurk S, Walenz BP, Rhie A. *et al.* HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res*. 2020;**30**:1291–305
2. Sun X, Jiao C, Schwaninger H. *et al.* Phased diploid genome assemblies and pan-genomes provide insights into the genetic history of apple domestication. *Nature Genet*. 2020;**52**: 1423–32
3. Song X, Liu Z, Wan H. *et al.* Editorial: comparative genomics and functional genomics analyses in plants. *Front Genet*. 2021;**12**: 618
4. Wang X, Gao L, Jiao C. *et al.* Genome of *Solanum pimpinellifolium* provides insights into structural variants during tomato breeding. *Nat Commun*. 2020;**11**:5817
5. Alonge M, Wang X, Benoit M. *et al.* Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell*. 2020;**182**:145–161.e23
6. Xu J, Hua K, Lang Z. Genome editing for horticultural crop improvement. *Hortic Res*. 2019;**6**:113
7. 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
8. Jung S, Lee T, Cheng CH. *et al.* 15 years of GDR: new data and functionality in the Genome Database for Rosaceae. *Nucleic Acids Res*. 2019;**47**:D1137–45
9. 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
10. Yu J, Wu S, Sun H. *et al.* CuGenDBv2: an updated database for cucurbit genomics. *Nucleic Acids Res*. 2023;**51**: D1457–64
11. Chen S, Sun M, Xu S. *et al.* The pear genomics database (PGDB): a comprehensive multi-omics research platform for Pyrus spp. *BMC Plant Biol*. 2023;**23**:430
12. Li Q, Qi J, Qin X. *et al.* CitGVD: a comprehensive database of citrus genomic variations. *Hort Res*. 2020;**7**:12
13. 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
14. Xu H, Yu Q, Shi Y. *et al.* PGD: Pineapple Genomics Database. *Hortic Res*. 2018;**5**:66
15. Li J, Chen C, Zeng Z. *et al.* SapBase (Sapinaceae Genomic DataBase): a central portal for functional and comparative genomics of Sapindaceae species. preprint: not peer–reviewed at *bioRxiv*. 2022; 2022.11.25.517904
16. Da L, Liu Y, Yang J. *et al.* AppleMDO: a multi-dimensional omics database for apple co-expression networks and chromatin states. *Front Plant Sci*. 2019;**10**:485905
17. Liu H, Wang X, Liu S. *et al.* Citrus Pan-Genome to Breeding Database (CPBD): a comprehensive genome database for citrus breeding. *Mol Plant*. 2022;**15**:1503–5
18. Chen H, Wang T, He X. *et al.* BRAD V3.0: an upgraded Brassicaceae database. *Nucleic Acids Res*. 2022;**50**:D1432–41
19. Chen F, Song Y, Li X. *et al.* Genome sequences of horticultural plants: past, present, and future. *Hortic Res*. 2019;**6**:112
20. Esch M, Chen J, Colmsee C. *et al.* LAILAPS: the plant science search engine. *Plant Cell Physiol*. 2015;**56**:e8
21. Marks RA, Hotaling S, Frandsen PB. *et al.* Representation and participation across 20 years of plant genome sequencing. *Nature plants*. 2021;**7**:1571–8
22. Sun Y, Shang L, Zhu QH. *et al.* Twenty years of plant genome sequencing: achievements and challenges. *Trends Plant Sci*. 2022;**27**:391–401
23. Goodstein DM, Shu S, Howson R. *et al.* Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res*. 2012;**40**:D1178–86
24. Bolser DM, Staines DM, Perry E. *et al.* Ensembl plants: integrating tools for visualizing, mining, and analyzing plant genomic data. *Methods Mol Biol*. 2017;**1533**:1–31
25. Chen M, Ma Y, Wu S. *et al.* Genome Warehouse: a public repository housing genome-scale data. *Genom Proteom Bioinform*. 2021;**19**:584–9
26. Trapnell C, Williams BA, Pertea G. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*. 2010;**28**:511–5
27. Yue J, Liu J, Tang W. *et al.* Kiwifruit Genome Database (KGD): a comprehensive resource for kiwifruit genomics. *Hortic Res*. 2020;**7**:117
28. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods*. 2015;**12**:59–60

29. Mitchell AL, Attwood TK, Babbitt PC. *et al*. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res*. 2019;**47**:D351–60
30. Zheng Y, Jiao C, Sun H. *et al*. iTAK: a program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. *Mol Plant*. 2016;**9**:1667–70
31. Huerta-Cepas J, Szklarczyk D, Heller D. *et al*. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res*. 2019;**47**:D309–14
32. Tang H, Bowers JE, Wang X. *et al*. Synteny and collinearity in plant genomes. *Science*. 2008;**320**:486–8
33. 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
34. Garcia L, Yachdav G, Martin MJ. FeatureViewer, a BioJS component for visualization of position-based annotations in protein sequences. *F1000Research*. 2014;**3**:47
35. Priyam A, Woodcroft BJ, Rai V. *et al*. Sequenceserver: a modern graphical user interface for custom BLAST databases. *Mol Biol Evol*. 2019;**36**:2922–4
36. Boyle EI, Weng S, Gollub J. *et al*. GO::TermFinder–open source software for accessing gene ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*. 2004;**20**:3710–5
37. Carbon S, Douglass E, Dunn N. *et al*. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res*. 2019;**47**:D330–8
38. Kanehisa M, Furumichi M, Sato Y. *et al*. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res*. 2021;**49**:D545–51
39. Gui S, Yang L, Li J. *et al*. ZEAMAP, a comprehensive database adapted to the maize multi-omics era. *iScience*. 2020;**23**:101241
40. Uberti Manassero NG, Viola IL, Welchen E. *et al*. TCP transcription factors: architectures of plant form. *Biomol Concepts*. 2013;**4**:111–27
41. Parapunova V, Busscher M, Busscher-Lange J. *et al*. Identification, cloning and characterization of the tomato TCP transcription factor family. *BMC Plant Biol*. 2014;**14**:157