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,*], 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, China 102206

[2]Bioinformatics Center, Beijing University of Agriculture, Beijing, China 102206

[3]Boyce Thompson Institute, Cornell University, Ithaca, NY 14853

[4]USDA-ARS, Robert W. Holley Center for Agriculture and Health, Ithaca, NY, 14853

[5]College of Intelligent Science and Engineering, Beijing University of Agriculture, Beijing, China, 102206

[6]Library, Beijing University of Agriculture, Beijing, China, 102206

[#]These authors contributed equally to the work.

Sen Wang, wangsen@moilab.net

Shangxiao Wei, wsx22@moilab.net

Yuling Deng, jiapengyou@moilab.net

Shaoyuan Wu, wushaoyuan@moilab.net

Haixu Peng, holdianmiss@moilab.net

You Qing, qingyou@moilab.net

Xuyang Zhai, 0weixiao1@moilab.net

Shijie Zhou, zsj@moilab.net

Jinrong Li, ljr22@moilab.net

Hua Li, lh22@moilab.net

Yijian Feng, fyj22@moilab.net

Yating Yi, yyt22@moilab.net

Rui Li, lr22@moilab.net

Hui Zhang, 18076158874@moilab.net

Yiding Wang, wyd1021244674@gmail.com

Renlong Zhang, zrl@bua.edu.cn

Lu Ning, ninglu@bua.edu.cn

*Correspondence:

Yuncong Yao (yaoyc_20@126.com), Phone number: +8613910831039, Fax: +8601080769125;

Zhangjun Fei (zf25@cornell.edu), Phone number: +16072543234, Fax: +16072541234;

Yi Zheng (yz@moilab.net), Phone number: +8618535137613, Fax: +8601080769125;

Running title (50 characters)

Searching Horticultural Genomic Data

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 querying genes, functional annotations, protein domains, homologs, and other gene-related functional information of more than 400 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 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 analyze 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]. Most of these databases manage genomic data for plants from a single-family or species[11]. Therefore, the genomic resources of horticultural crops are scattered in different databases, and these databases exhibit different ways of presenting and utilizing results, resulting 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[12]. 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, it 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. The HortGenome Search Engine contains data of 434 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 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 have been sequenced and published by the end of 2021[13,14]. 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[15], Ensembl Plants[16], Genome Warehouse in National Genomics Data Center[17], 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[18]. 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 434 genome assemblies for horticultural crops, as well as the model plant species Arabidopsis and rice, were collected and included in the database (Table S1). 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[9,19] 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[20] 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[21] to identify functional protein domains. Transcription factors (TFs), transcriptional regulators (TRs), and protein kinases (PKs) were identified using the iTAK pipeline[22].

To generate GO and KEGG pathway annotations for functional enrichment analyses, protein sequences were compared against the EggNOG database using eggnog-mapper[23]. 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 functions[24]. Detection of synteny blocks among all the 434 genomes would yield more than 90,000 pairwise genome comparisons, which is time-consuming and computationally not feasible. Therefore, in our study syntenic blocks and homologous gene pairs were identified only between any two genome assemblies from species within the same family, and within each genome assembly. 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 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[25], which implements a new BLAST filter to remove weak syntenic regions and tandem duplications[24]. In the end, a total of 1,832,351 syntenic blocks and 413 million homologous gene pairs were identified from 6,994 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 34 million records of genes and transcripts from 434 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 126 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 (Figure 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 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, the HSE will automatically prompt and complete the corresponding name for users to choose (Figure 1B). After selecting species keywords, users can enter other keywords such as gene ID, gene name, gene functional description, etc (Figure 1C-E).

The search returns a gene list with the corresponding species name, gene/transcript IDs, gene locations, and gene functional descriptions (Figure 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 user enters a keyword that combines the name of a species and the name of a specific TF/TR/PK family (Figure 1E), the results will directly return to the corresponding gene family page of the species (Figure 3).

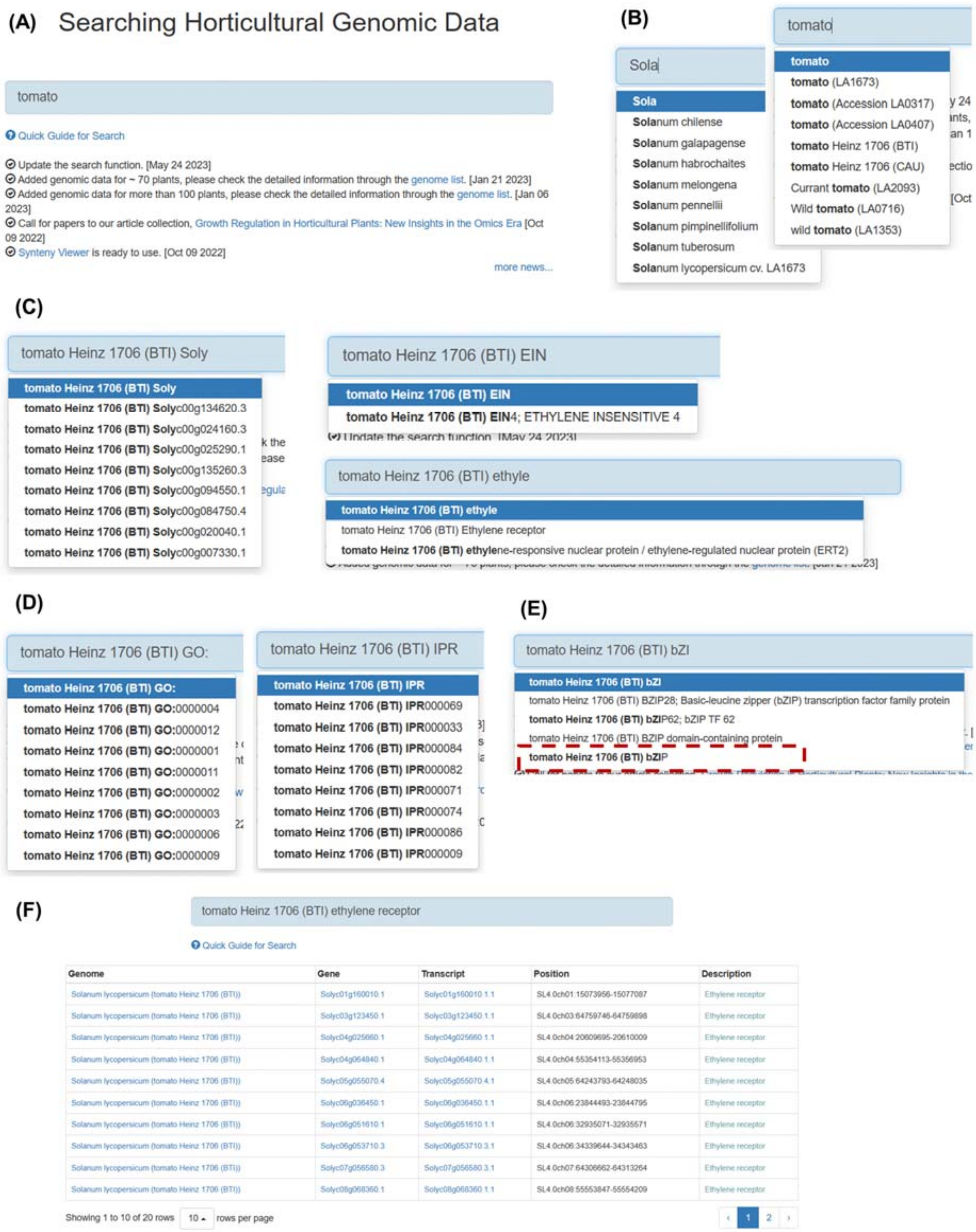


Figure 1. Search interface and result pages in HortGenome Search Engine. (A-E) Screenshots of the search interfaces. **(F)** Gene list of search results, including plant name, gene/transcript ID, genomic location and functional description.

Genome page display

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 (Figure 2A).

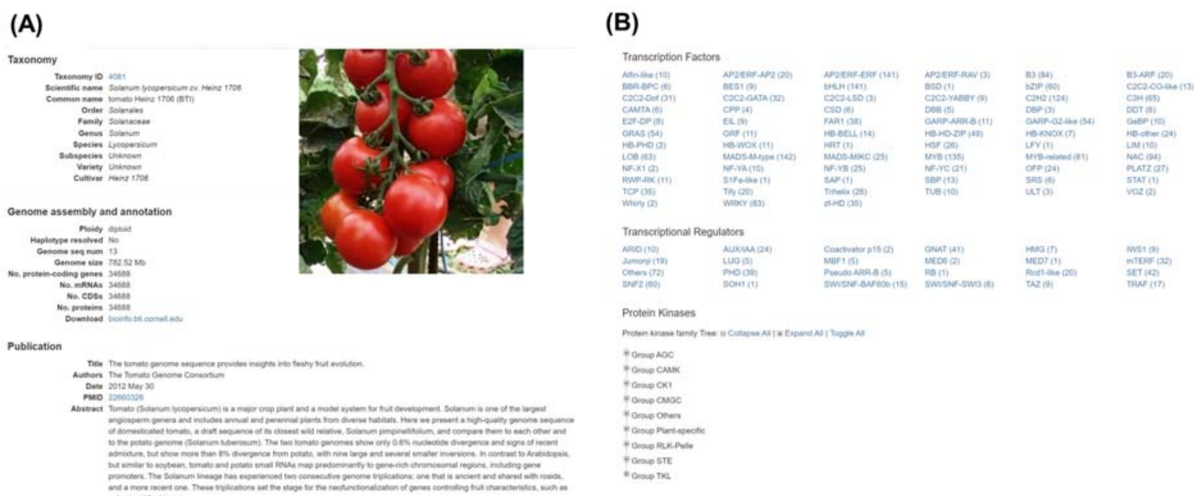


Figure 2. Genome page in HortGenome Search Engine. (A) Screenshot of the genome page containing the genome information and picture of the plant. **(B)** Screenshot of the genome page containing transcription factors, transcriptional regulators, and protein kinases identified from the genome.

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 (Figure 2B). Clicking on a family name directs to the corresponding gene family page.

Gene family page display

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.

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[22] 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 gene list, CDS, protein, and promoter sequences of these bZIP genes (Figure 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 (Figure 3B-D), which provide convenience for studying the function and evolution of the corresponding gene families.

Gene and transcript page display

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 (Figure 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 [26] (Figure 4A). The sequence pagination contains gene, mRNA (primary transcript), CDS, and protein sequences (Figure 4B). In the BLAST pagination, it shows the top 5 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 (Figure 4C). The domain pagination lists the functional domains identified from the protein sequence of this gene (Figure 4D). The gene ontology pagination lists the GO terms assigned to this gene and

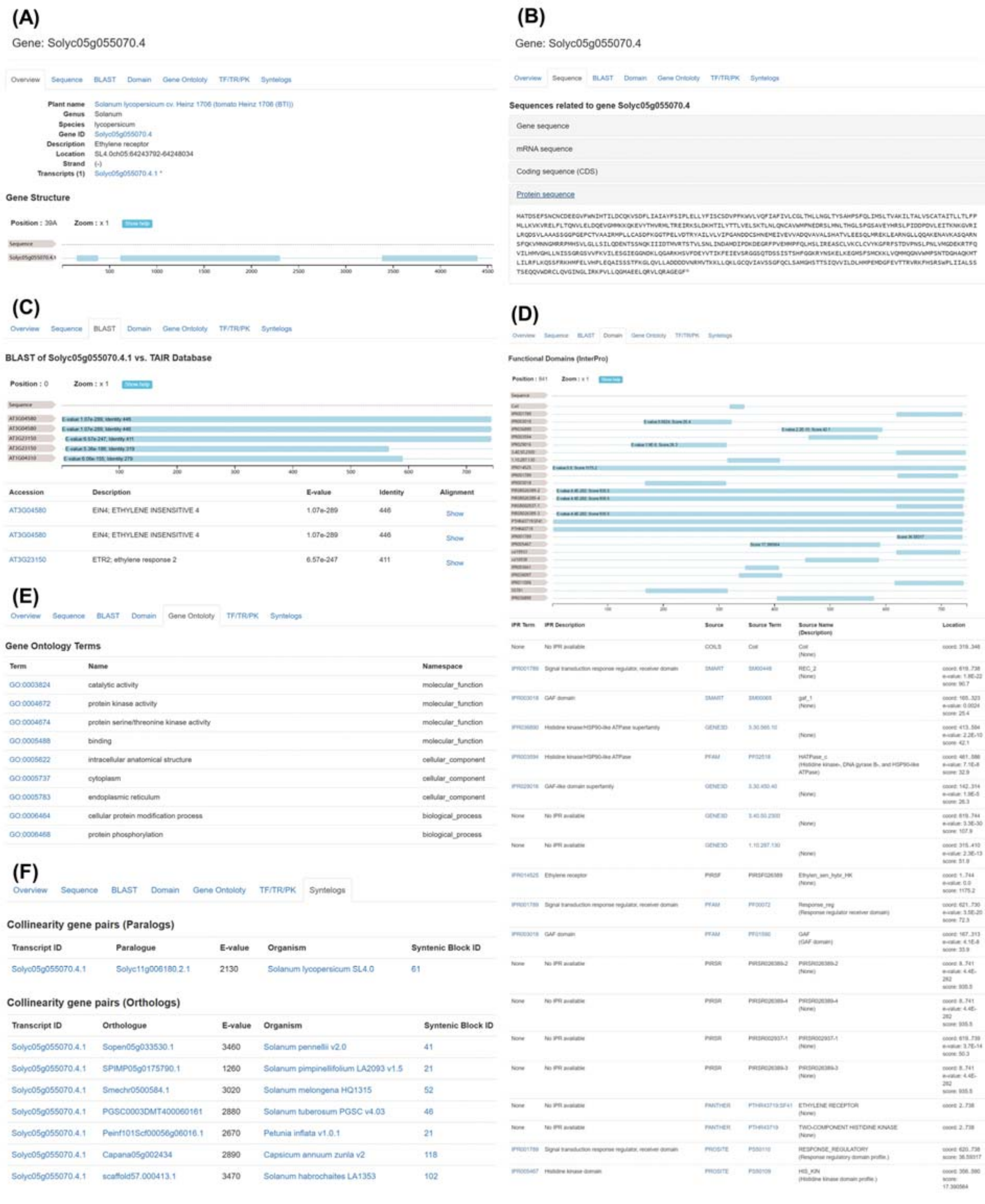


Figure 4. Gene feature page in HortGenome Search Engine. (A) Screenshot of the gene page

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.

the GO IDs are linked to the AmiGO database which provides details of the GO terms (Figure 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 (Figure 4F).

BLAST

We implemented the online BLAST tool, one of the most widely used tools in genome databases, using the SequenceServer[27]. 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 (Figure 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 analyzed. 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 (Figure 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[28]. 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[29,30]. 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.

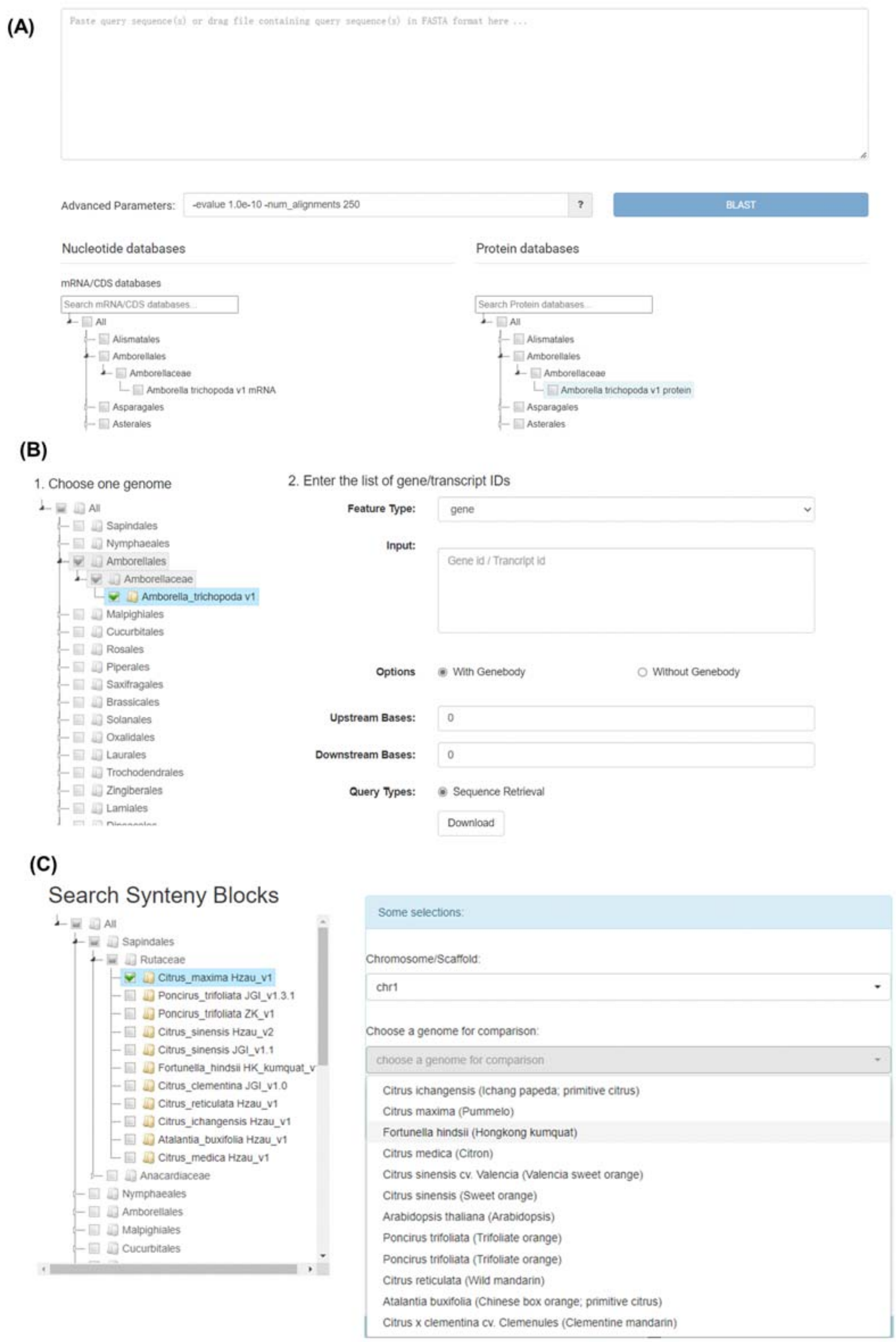


Figure 5. Query interfaces of data mining tools in HortGenome Search Engine. (A) Screenshot of the BLAST query page. **(B)** Search interface of 'Batch Query'. **(C)** Search interface of 'Synteny Viewer'.

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)[31], 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 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 (Figure 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.

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 16 million genes with comprehensive functional annotations as well as 1,832,351 synteny blocks and 413 million homologous gene pairs from 434 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 by contacting us. In the future, we will expand the scope of data search to cover other omics data such as gene regulatory networks, gene expression, 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.W. The computing power was supported by the Alibaba Cloud.

Contributions

Z. Fei and Y. Zheng designed the project. S. Wei, Y. Deng, S. Wu, H. Peng, X. Zhai, S. Zhou, J. Li, H. Li, Y. Feng, Y. Yi, R. Li, H. Zhang, Y. Wang, R. Zhang, L. Ning, and Y. Zheng performed data collection. S. Wei, Y. Qing, H. Peng, and S. Wang performed data analysis. S. Wei and Y. Zheng wrote the code for database construction. Y. Yao, Z. Fei, and Y. Zheng supervised the project and wrote the manuscript. All authors read and approved the final manuscript.

Data availability statement

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

Conflict of interest

The authors declare that they have no conflict of interest.

Supplementary data

Supplementary data is available at Horticulture Research online.

Reference:

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**: 1291–1305.

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. *Nat Genet 2020 5212* 2020; **52**: 1423–1432.

3 Song X, Liu Z, Wan H, Chen W, Zhou R, Duan W. 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–D1041.

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–D1145.

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–D1136.

10 Yu J, Wu S, Sun H *et al.* CuGenDBv2: an updated database for cucurbit genomics. *Nucleic Acids Res* 2023; **51**: D1457–D1464.

11 Chen F, Song Y, Li X *et al.* Genome sequences of horticultural plants: past, present, and future. *Hortic Res* 2019; **6**: 112.

12 Esch M, Chen J, Colmsee C *et al.* LAILAPS: the plant science search engine. *Plant Cell Physiol* 2015; **56**: e8.

13 Marks RA, Hotaling S, Frandsen PB, VanBuren R. Representation and participation across 20 years of plant genome sequencing. *Nat Plants* 2021; **7**: 1571–1578.

14 Sun Y, Shang L, Zhu QH, Fan L, Guo L. Twenty years of plant genome sequencing: achievements and challenges. *Trends Plant Sci* 2022; **27**: 391–401.

15 Goodstein DM, Shu S, Howson R *et al.* Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* 2012; **40**: D1178-D1186.

16 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.

17 Chen M, Ma Y, Wu S *et al.* Genome Warehouse: A public repository housing genome-scale data. *Genomics Proteomics Bioinformatics* 2021; **19**: 584–589.

18 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**: 511–515.

19 Yue J, Liu J, Tang W *et al.* Kiwifruit Genome Database (KGD): a comprehensive resource for kiwifruit genomics. *Hortic Res* 2020; **7**: 117.

20 Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015; **12**: 59–60.

21 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–D360.

22 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**: 1667–1670.

23 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–D314.

24 Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH. Synteny and collinearity in plant genomes. *Science* 2008; **320**: 486–488.

25 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.

26 Garcia L, Yachdav G, Martin MJ. FeatureViewer, a BioJS component for visualization of position-based annotations in protein sequences. *F1000Research* 2014; **3:** 47.

27 Priyam A, Woodcroft BJ, Rai V *et al.* Sequenceserver: A modern graphical user interface for custom BLAST databases. *Mol Biol Evol* 2019; **36**: 2922–2924.

28 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**: 3710–3715.

29 Carbon S, Douglass E, Dunn N *et al*. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res* 2019; **47**: D330–D338.

30 Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res* 2021; **49**: D545–D551.

31 Gui S, Yang L, Li J *et al.* ZEAMAP, a comprehensive database adapted to the maize multi-omics era. *iScience* 2020; **23**: 101241.

FIGURES LEGENDS

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.

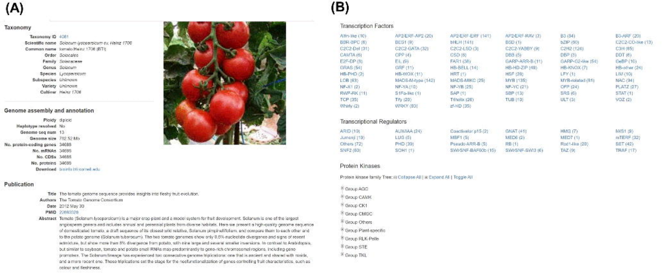
Figure 2. Genome page in HortGenome Search Engine. (A) Screenshot of the genome page containing the genome information and picture of the plant. **(B)** Screenshot of the genome page containing transcription factors, transcriptional regulators, and protein kinases identified from the genome.

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.

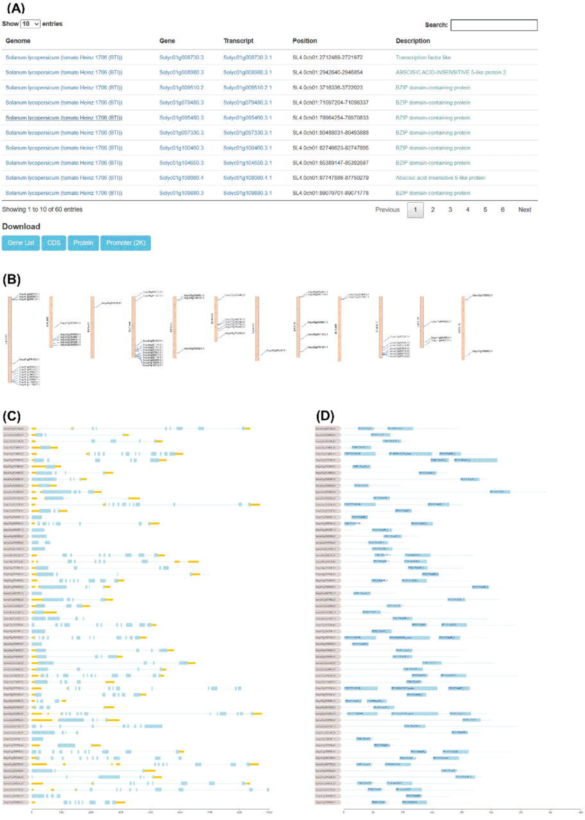
Figure 4. Gene feature page in HortGenome Search Engine. (A) Screenshot of the gene page containing basic information and gene structure. **(B)** Screenshot of the gene page containing gene, mRNA, CDS, and protein sequences. **(C)** Screenshot of the homolog genes and sequence alignments from the BLAST 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’.

(A)
Searching Horticultural Genomic Data
tomato
Quick Guide for Search
Update the search function. [May 24 2023]
Added genomic data for ~ 70 plants, please check the detailed information through the genome list. [Jan 21 2023]
Added genomic data for more than 100 plants, please check the detailed information through the genome list. [Jan 06 2023]
Call for papers to our article collection, Growth Regulation in Horticultural Plants: New Insights in the Omics Era [Oct 09 2022]
Synteny Viewer is ready to use. [Oct 09 2022]
more news...
(B)
Sola
Sola
Solanum chilense
Solanum galapagense
Solanum habrochaites
Solanum melongena
Solanum pennellii
Solanum pimpinellifolium
Solanum tuberosum
Solanum lycopersicum cv. LA1673
tomato
tomato
tomato (LA1673)
tomato (Accession LA0317)
tomato (Accession LA0407)
tomato Heinz 1706 (BTI)
tomato Heinz 1706 (CAU)
Currant tomato (LA2093)
Wild tomato (LA0716)
wild tomato (LA1353)
(C)
tomato Heinz 1706 (BTI) Soly
tomato Heinz 1706 (BTI) Soly
tomato Heinz 1706 (BTI) Solyc00g134620.3
tomato Heinz 1706 (BTI) Solyc00g024160.3
tomato Heinz 1706 (BTI) Solyc00g025290.1
tomato Heinz 1706 (BTI) Solyc00g135260.3
tomato Heinz 1706 (BTI) Solyc00g094550.1
tomato Heinz 1706 (BTI) Solyc00g084750.4
tomato Heinz 1706 (BTI) Solyc00g020040.1
tomato Heinz 1706 (BTI) Solyc00g007330.1
tomato Heinz 1706 (BTI) EIN
tomato Heinz 1706 (BTI) EIN
tomato Heinz 1706 (BTI) EIN4; ETHYLENE INSENSITIVE 4
tomato Heinz 1706 (BTI) ethyle
tomato Heinz 1706 (BTI) ethyle
tomato Heinz 1706 (BTI) Ethylene receptor
tomato Heinz 1706 (BTI) ethylene-responsive nuclear protein / ethylene-regulated nuclear protein (ERT2)
(D)
tomato Heinz 1706 (BTI) GO:
tomato Heinz 1706 (BTI) GO:
tomato Heinz 1706 (BTI) GO:0000004
tomato Heinz 1706 (BTI) GO:0000012
tomato Heinz 1706 (BTI) GO:0000001
tomato Heinz 1706 (BTI) GO:0000011
tomato Heinz 1706 (BTI) GO:0000002
tomato Heinz 1706 (BTI) GO:0000003
tomato Heinz 1706 (BTI) GO:0000006
tomato Heinz 1706 (BTI) GO:0000009
tomato Heinz 1706 (BTI) IPR
tomato Heinz 1706 (BTI) IPR
tomato Heinz 1706 (BTI) IPR000069
tomato Heinz 1706 (BTI) IPR000033
tomato Heinz 1706 (BTI) IPR000084
tomato Heinz 1706 (BTI) IPR000082
tomato Heinz 1706 (BTI) IPR000071
tomato Heinz 1706 (BTI) IPR000074
tomato Heinz 1706 (BTI) IPR000086
tomato Heinz 1706 (BTI) IPR000009
(E)
tomato Heinz 1706 (BTI) bZI
tomato Heinz 1706 (BTI) bZI
tomato Heinz 1706 (BTI) BZIP28; Basic-leucine zipper (bZIP) transcription factor family protein
tomato Heinz 1706 (BTI) bZIP62; bZIP TF 62
tomato Heinz 1706 (BTI) BZIP domain-containing protein
tomato Heinz 1706 (BTI) bZIP
(F)
tomato Heinz 1706 (BTI) ethylene receptor
Quick Guide for Search
Genome | Gene | Transcript | Position | Description
Solanum lycopersicum (tomato Heinz 1706 (BTI)) | Solyc01g160010.1 | Solyc01g160010.1.1 | SL4.0ch01:15073956-15077087 | Ethylene receptor
Solanum lycopersicum (tomato Heinz 1706 (BTI)) | Solyc03g123450.1 | Solyc03g123450.1.1 | SL4.0ch03:64759746-64759898 | Ethylene receptor
Solanum lycopersicum (tomato Heinz 1706 (BTI)) | Solyc04g025660.1 | Solyc04g025660.1.1 | SL4.0ch04:20609695-20610009 | Ethylene receptor
Solanum lycopersicum (tomato Heinz 1706 (BTI)) | Solyc04g064840.1 | Solyc04g064840.1.1 | SL4.0ch04:55354113-55356953 | Ethylene receptor
Solanum lycopersicum (tomato Heinz 1706 (BTI)) | Solyc05g055070.4 | Solyc05g055070.4.1 | SL4.0ch05:64243793-64248035 | Ethylene receptor
Solanum lycopersicum (tomato Heinz 1706 (BTI)) | Solyc06g036450.1 | Solyc06g036450.1.1 | SL4.0ch06:23844493-23844795 | Ethylene receptor
Solanum lycopersicum (tomato Heinz 1706 (BTI)) | Solyc06g051610.1 | Solyc06g051610.1.1 | SL4.0ch06:32935071-32935571 | Ethylene receptor
Solanum lycopersicum (tomato Heinz 1706 (BTI)) | Solyc06g053710.3 | Solyc06g053710.3.1 | SL4.0ch06:34339644-34343463 | Ethylene receptor
Solanum lycopersicum (tomato Heinz 1706 (BTI)) | Solyc07g056580.3 | Solyc07g056580.3.1 | SL4.0ch07:64306662-64313264 | Ethylene receptor
Solanum lycopersicum (tomato Heinz 1706 (BTI)) | Solyc08g068360.1 | Solyc08g068360.1.1 | SL4.0ch08:55553847-55554209 | Ethylene receptor
Showing 1 to 10 of 20 rows
10
rows per page
1
2



(A)
Taxonomy
Taxonomy ID 4081
Scientific name Solanum lycopersicum cv. Heinz 1706
Common name tomato Heinz 1706 (BTI)
Order Solanales
Family Solanaceae
Genus Solanum
Species Lycopersicum
Subspecies Unknown
Variety Unknown
Cultivar Heinz 1706
Genome assembly and annotation
Ploidy diploid
Haplotype resolved No
Genome seq num 13
Genome size 782.52 Mb
No. protein-coding genes 34688
No. mRNAs 34688
No. CDSs 34688
No. proteins 34688
Download bioinfo.bti.cornell.edu
Publication
Title The tomato genome sequence provides insights into fleshy fruit evolution.
Authors The Tomato Genome Consortium
Date 2012 May 30
PMID 22660326
Abstract Tomato (Solanum lycopersicum) is a major crop plant and a model system for fruit development. Solanum is one of the largest angiosperm genera and includes annual and perennial plants from diverse habitats. Here we present a high-quality genome sequence of domesticated tomato, a draft sequence of its closest wild relative, Solanum pimpinellifolium, and compare them to each other and to the potato genome (Solanum tuberosum). The two tomato genomes show only 0.6% nucleotide divergence and signs of recent admixture, but show more than 8% divergence from potato, with nine large and several smaller inversions. In contrast to Arabidopsis, but similar to soybean, tomato and potato small RNAs map predominantly to gene-rich chromosomal regions, including gene promoters. The Solanum lineage has experienced two consecutive genome triplications: one that is ancient and shared with rosids, and a more recent one. These triplications set the stage for the neofunctionalization of genes controlling fruit characteristics, such as colour and fleshiness.
(B)
Transcription Factors
Alfin-like (10)
AP2/ERF-AP2 (20)
AP2/ERF-ERF (141)
AP2/ERF-RAV (3)
B3 (84)
B3-ARF (20)
BBR-BPC (6)
BES1 (9)
bHLH (141)
BSD (1)
bZIP (80)
C2C2-CO-like (13)
C2C2-Dof (31)
C2C2-GATA (32)
C2C2-LSD (3)
C2C2-YABBY (9)
C2H2 (124)
C3H (85)
CAMTA (6)
CPP (4)
CSD (8)
DBB (5)
DBP (3)
DDT (6)
E2F-DP (8)
EIL (9)
FAR1 (38)
GARP-ARR-B (11)
GARP-G2-like (54)
GeBP (10)
GRAS (54)
GRF (11)
HB-BELL (14)
HB-HD-ZIP (48)
HB-KNOX (7)
HB-other (24)
HB-PHD (2)
HB-WOX (11)
HRT (1)
HSF (28)
LFY (1)
LIM (10)
LOB (63)
MADS-M-type (142)
MADS-MIKC (25)
MYB (135)
MYB-related (81)
NAC (94)
NF-X1 (2)
NF-YA (10)
NF-YB (25)
NF-YC (21)
OFP (24)
PLATZ (27)
RWP-RK (11)
S1Fa-like (1)
SAP (1)
SBP (13)
SRS (6)
STAT (1)
TCP (35)
Tify (20)
Trihelix (26)
TUB (10)
ULT (3)
VOZ (2)
Whirly (2)
WRKY (83)
zf-HD (35)
Transcriptional Regulators
ARID (10)
AUX/IAA (24)
Coactivator p15 (2)
GNAT (41)
HMG (7)
IWS1 (9)
Jumonji (19)
LUG (5)
MBF1 (5)
MED6 (2)
MED7 (1)
mTERF (32)
Others (72)
PHD (39)
Pseudo ARR-B (5)
RB (1)
Rcd1-like (20)
SET (42)
SNF2 (60)
SOH1 (1)
SWI/SNF-BAF60b (15)
SWI/SNF-SWI3 (6)
TAZ (9)
TRAF (17)
Protein Kinases
Protein kinase family Tree: Collapse All | Expand All | Toggle All
Group AGC
Group CAMK
Group CK1
Group CMGC
Group Others
Group Plant-specific
Group RLK-Pelle
Group STE
Group TKL



(A)

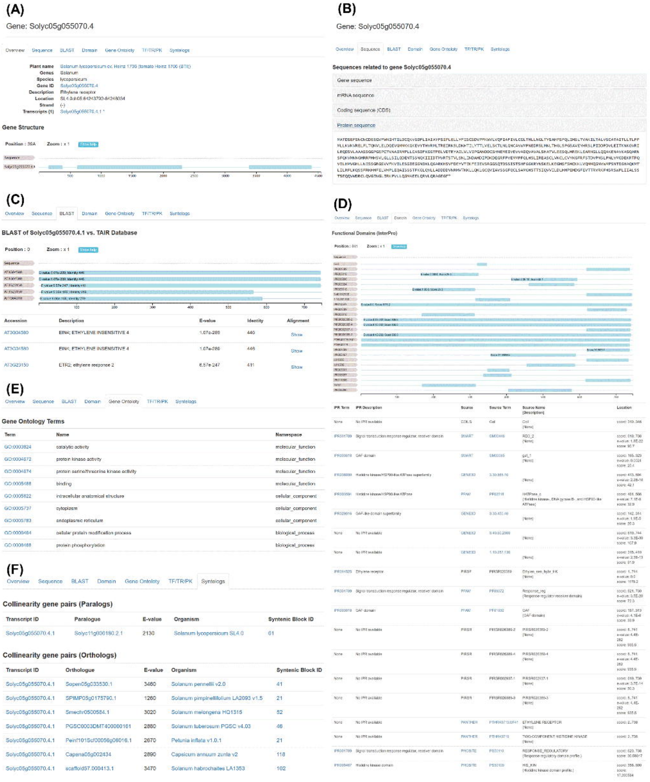
Show 10 entries   Search:

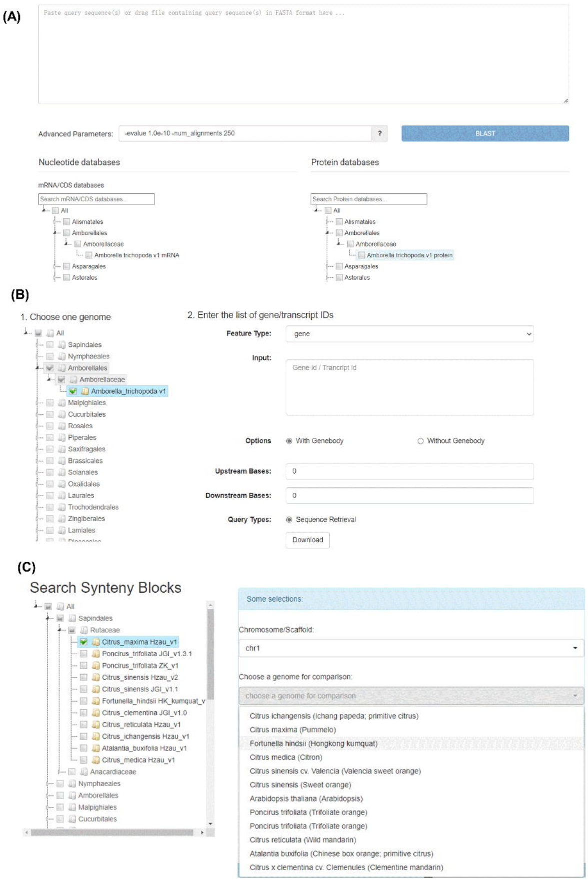
Genome	Gene	Transcript	Position	Description
Solanum lycopersicum (tomato Heinz 1706 (BTI))	Solyc01g008730.3	Solyc01g008730.3.1	SL4.0ch01:2712489-2721972	Transcription factor like
Solanum lycopersicum (tomato Heinz 1706 (BTI))	Solyc01g008980.3	Solyc01g008980.3.1	SL4.0ch01:2942640-2946854	ABSCISIC ACID-INSENSITIVE 5-like protein 2
Solanum lycopersicum (tomato Heinz 1706 (BTI))	Solyc01g009510.2	Solyc01g009510.2.1	SL4.0ch01:3716336-3722023	BZIP domain-containing protein
Solanum lycopersicum (tomato Heinz 1706 (BTI))	Solyc01g079480.3	Solyc01g079480.3.1	SL4.0ch01:71097204-71098337	BZIP domain-containing protein
Solanum lycopersicum (tomato Heinz 1706 (BTI))	Solyc01g095460.3	Solyc01g095460.3.1	SL4.0ch01:78964254-78970833	BZIP domain-containing protein
Solanum lycopersicum (tomato Heinz 1706 (BTI))	Solyc01g097330.3	Solyc01g097330.3.1	SL4.0ch01:80488031-80493888	BZIP domain-containing protein
Solanum lycopersicum (tomato Heinz 1706 (BTI))	Solyc01g100460.3	Solyc01g100460.3.1	SL4.0ch01:82746623-82747895	BZIP domain-containing protein
Solanum lycopersicum (tomato Heinz 1706 (BTI))	Solyc01g104650.3	Solyc01g104650.3.1	SL4.0ch01:85389147-85392687	BZIP domain-containing protein
Solanum lycopersicum (tomato Heinz 1706 (BTI))	Solyc01g108080.4	Solyc01g108080.4.1	SL4.0ch01:87747886-87750279	Abscisic acid insensitive 5 like protein
Solanum lycopersicum (tomato Heinz 1706 (BTI))	Solyc01g109880.3	Solyc01g109880.3.1	SL4.0ch01:89070701-89071778	BZIP domain-containing protein

Showing 1 to 10 of 60 entries   Previous 1 2 3 4 5 6 Next

Download

Gene List CDS Protein Promoter (2K)


(A)
Gene: Solyc05g055070.4
Overview Sequence BLAST Domain Gene Ontology TF/TR/PK Syntelogs
Plant name Solanum lycopersicum cv. Heinz 1706 (tomato Heinz 1706 (BTI))
Genus Solanum
Species lycopersicum
Gene ID Solyc05g055070.4
Description Ethylene receptor
Strand (-)
Gene Structure
(B)
Gene: Solyc05g055070.4
Sequences related to gene Solyc05g055070.4
Gene sequence
mRNA sequence
Coding sequence (CDS)
Protein sequence
(C)
BLAST of Solyc05g055070.4.1 vs. TAIR Database
Accession Description E-value Identity Alignment
AT3G04580 EIN4; ETHYLENE INSENSITIVE 4 1.07e-289 446 Show
AT3G04580 EIN4; ETHYLENE INSENSITIVE 4 1.07e-289 446 Show
AT3G23150 ETR2; ethylene response 2 6.57e-247 411 Show
(D)
Functional Domains (InterPro)
IPR Term IPR Description Source Source Term Source Name (Description) Location
(E)
Gene Ontology Terms
Term Name Namespace
GO:0003824 catalytic activity molecular_function
GO:0004672 protein kinase activity molecular_function
GO:0004674 protein serine/threonine kinase activity molecular_function
GO:0005488 binding molecular_function
GO:0005622 intracellular anatomical structure cellular_component
GO:0005737 cytoplasm cellular_component
GO:0005783 endoplasmic reticulum cellular_component
GO:0006464 cellular protein modification process biological_process
GO:0006468 protein phosphorylation biological_process
(F)
Collinearity gene pairs (Paralogs)
Transcript ID Paralogue E-value Organism Syntenic Block ID
Solyc05g055070.4.1 Solyc11g006180.2.1 2130 Solanum lycopersicum SL4.0 61
Collinearity gene pairs (Orthologs)
Transcript ID Orthologue E-value Organism Syntenic Block ID
Solyc05g055070.4.1 Sopen05g033530.1 3460 Solanum pennellii v2.0 41
Solyc05g055070.4.1 SPIMP05g0175790.1 1260 Solanum pimpinellifolium LA2093 v1.5 21
Solyc05g055070.4.1 Smechr0500584.1 3020 Solanum melongena HQ1315 52
Solyc05g055070.4.1 PGSC0003DMT400060161 2880 Solanum tuberosum PGSC v4.03 46
Solyc05g055070.4.1 Peinf101Scf00056g06016.1 2670 Petunia inflata v1.0.1 21
Solyc05g055070.4.1 Capana05g002434 2890 Capsicum annuum zunla v2 118
Solyc05g055070.4.1 scaffold57.000413.1 3470 Solanum habrochaites LA1353 102


(A)
Paste query sequence(s) or drag file containing query sequence(s) in FASTA format here ...
Advanced Parameters:
-evalue 1.0e-10 -num_alignments 250
?
BLAST
Nucleotide databases
Protein databases
mRNA/CDS databases
Search mRNA/CDS databases
All
Alismatales
Amborellales
Amborellaceae
Amborella trichopoda v1 mRNA
Asparagales
Asterales
Search Protein databases
All
Alismatales
Amborellales
Amborellaceae
Amborella trichopoda v1 protein
Asparagales
Asterales
(B)
1. Choose one genome
All
Sapindales
Nymphaeales
Amborellales
Amborellaceae
Amborella_trichopoda v1
Malpighiales
Cucurbitales
Rosales
Piperales
Saxifragales
Brassicales
Solanales
Oxalidales
Laurales
Trochodendrales
Zingiberales
Lamiales
2. Enter the list of gene/transcript IDs
Feature Type:
gene
Input:
Gene id / Trancript id
Options
With Genebody
Without Genebody
Upstream Bases:
0
Downstream Bases:
0
Query Types:
Sequence Retrieval
Download
(C)
Search Synteny Blocks
All
Sapindales
Rutaceae
Citrus_maxima Hzau_v1
Poncirus_trifoliata JGI_v1.3.1
Poncirus_trifoliata ZK_v1
Citrus_sinensis Hzau_v2
Citrus_sinensis JGI_v1.1
Fortunella_hindsii HK_kumquat_v
Citrus_clementina JGI_v1.0
Citrus_reticulata Hzau_v1
Citrus_ichangensis Hzau_v1
Atalantia_buxifolia Hzau_v1
Citrus_medica Hzau_v1
Anacardiaceae
Nymphaeales
Amborellales
Malpighiales
Cucurbitales
Some selections:
Chromosome/Scaffold:
chr1
Choose a genome for comparison:
choose a genome for comparison
Citrus ichangensis (Ichang papeda; primitive citrus)
Citrus maxima (Pummelo)
Fortunella hindsii (Hongkong kumquat)
Citrus medica (Citron)
Citrus sinensis cv. Valencia (Valencia sweet orange)
Citrus sinensis (Sweet orange)
Arabidopsis thaliana (Arabidopsis)
Poncirus trifoliata (Trifoliate orange)
Poncirus trifoliata (Trifoliate orange)
Citrus reticulata (Wild mandarin)
Atalantia buxifolia (Chinese box orange; primitive citrus)
Citrus x clementina cv. Clemenules (Clementine mandarin)