

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

Sen Wang<sup>1,2,#</sup>, Shangxiao Wei<sup>1,2,#</sup>, Yuling Deng<sup>1,2</sup>, Shaoyuan Wu<sup>1,2</sup>, Haixu Peng<sup>1,2</sup>, You Qing<sup>1,2</sup>, Xuyang Zhai<sup>1,2</sup>, Shijie Zhou<sup>1,2</sup>, Jinrong Li<sup>1,2</sup>, Hua Li<sup>1,2</sup>, Yijian Feng<sup>1,2</sup>, Yating Yi<sup>1,2</sup>, Rui Li<sup>1,2</sup>, Hui Zhang<sup>1,2</sup>, Yiding Wang<sup>5</sup>, Renlong Zhang<sup>5</sup>, Lu Ning<sup>2,6</sup>, YunCong Yao<sup>1,\*</sup>, Zhangjun Fei<sup>3,4,\*</sup>, Yi Zheng<sup>1,2,\*</sup>

<sup>1</sup>Beijing Key Laboratory for Agricultural Application and New Technique, College of Plant Science and Technology, Beijing University of Agriculture, Beijing, China 102206

<sup>2</sup>Bioinformatics Center, Beijing University of Agriculture, Beijing, China 102206

<sup>3</sup>Boyce Thompson Institute, Cornell University, Ithaca, NY 14853

<sup>4</sup>USDA-ARS, Robert W. Holley Center for Agriculture and Health, Ithaca, NY, 14853

<sup>5</sup>College of Intelligent Science and Engineering, Beijing University of Agriculture, Beijing, China, 102206

<sup>6</sup>Library, Beijing University of Agriculture, Beijing, China, 102206

<sup>#</sup>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

33 Renlong Zhang, [zrl@bua.edu.cn](mailto:zrl@bua.edu.cn)

34 Lu Ning, [ninglu@bua.edu.cn](mailto:ninglu@bua.edu.cn)

35

36 \*Correspondence:

37 Yuncong Yao ([yaoyc\\_20@126.com](mailto:yaoyc_20@126.com)), Phone number: +8613910831039, Fax: +8601080769125;

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

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

40

# **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.moiab.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<sup>1,2</sup>, but also serves as the cornerstone for functional and comparative genomics studies<sup>3,4</sup> and paves a path to resolve complex QTLs of important horticultural traits<sup>5</sup>. Advanced genome editing technologies have been demonstrated in recent years to have a great potential for improving the quality and yield of horticultural crops<sup>6</sup>, 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<sup>7-10</sup>. Most of these databases manage genomic data for plants from a single-family or species<sup>11</sup>. 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<sup>12</sup>. 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<sup>13,14</sup>. 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 Phytozome<sup>15</sup>, Ensembl Plants<sup>16</sup>, Genome Warehouse in National Genomics Data Center<sup>17</sup>, SGN<sup>7</sup>, GDR<sup>8</sup> 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<sup>18</sup>. 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<sup>9,19</sup> 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<sup>20</sup> 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<sup>21</sup> to identify functional protein domains. Transcription factors (TFs), transcriptional regulators (TRs), and protein kinases (PKs) were identified using the iTAK pipeline<sup>22</sup>.

To generate GO and KEGG pathway annotations for functional enrichment analyses, protein sequences were compared against the EggNOG database using eggno-mapper<sup>23</sup>. The assigned GO terms of genes/transcripts retrieved from the eggno-mapper results were converted to the GO Annotation File (GAF) format. In the eggno-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 eggno-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<sup>24</sup>. 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<sup>25</sup>, which implements a new BLAST filter to remove weak syntenic regions and tandem duplications<sup>24</sup>. 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).

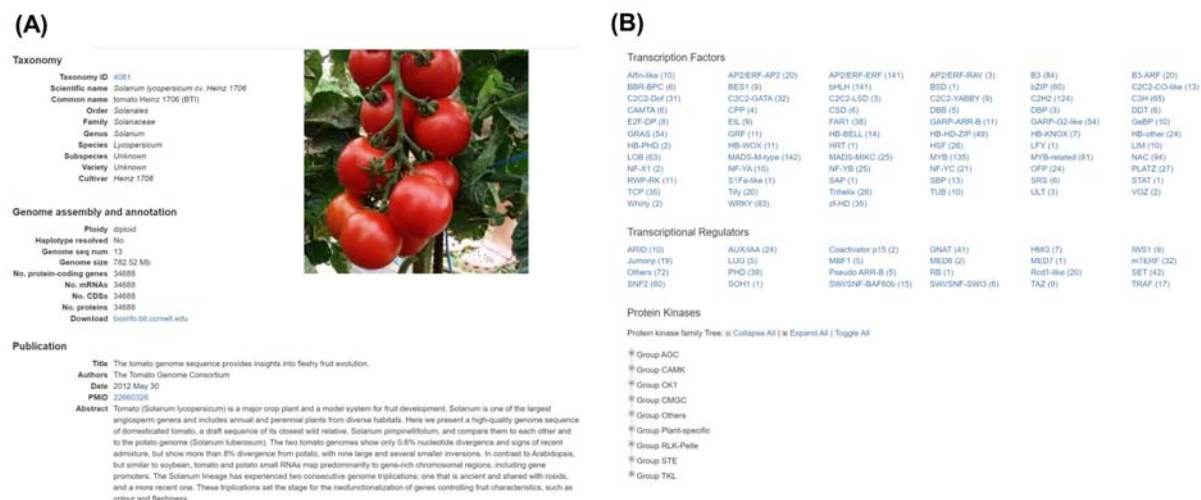
The search returns a gene list with the corresponding species name, gene/transcript IDs, gene locations, and gene functional descriptions (Figure 1F). The species name and gene/transcript IDs are linked to the corresponding genome page of the species and gene/transcript pages, respectively. In addition, 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

(A)

Show 10 entries

Search:

Genome	Gene	Transcript	Position	Description
Solanum lycopersicum (tomato Heinz 1706 (BT))	Solyc01g008730.3	Solyc01g008730.3.1	SL4.0ch01.2712469-2721972	Transcription factor like
Solanum lycopersicum (tomato Heinz 1706 (BT))	Solyc01g008980.3	Solyc01g008980.3.1	SL4.0ch01.2942640-2946854	ABSCISIC ACID-INSENSITIVE 5-like protein 2
Solanum lycopersicum (tomato Heinz 1706 (BT))	Solyc01g009510.2	Solyc01g009510.2.1	SL4.0ch01.3716336-3722023	BZIP domain-containing protein
Solanum lycopersicum (tomato Heinz 1706 (BT))	Solyc01g079490.3	Solyc01g079490.3.1	SL4.0ch01.71097204-71098337	BZIP domain-containing protein
Solanum lycopersicum (tomato Heinz 1706 (BT))	Solyc01g095460.3	Solyc01g095460.3.1	SL4.0ch01.78964254-78970833	BZIP domain-containing protein
Solanum lycopersicum (tomato Heinz 1706 (BT))	Solyc01g097330.3	Solyc01g097330.3.1	SL4.0ch01.80488031-80493888	BZIP domain-containing protein
Solanum lycopersicum (tomato Heinz 1706 (BT))	Solyc01g100460.3	Solyc01g100460.3.1	SL4.0ch01.82746523-82747895	BZIP domain-containing protein
Solanum lycopersicum (tomato Heinz 1706 (BT))	Solyc01g104650.3	Solyc01g104650.3.1	SL4.0ch01.85389147-85392687	BZIP domain-containing protein
Solanum lycopersicum (tomato Heinz 1706 (BT))	Solyc01g108080.4	Solyc01g108080.4.1	SL4.0ch01.87747886-87750279	Abcisic acid-insensitive 5-like protein
Solanum lycopersicum (tomato Heinz 1706 (BT))	Solyc01g109890.3	Solyc01g109890.3.1	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)

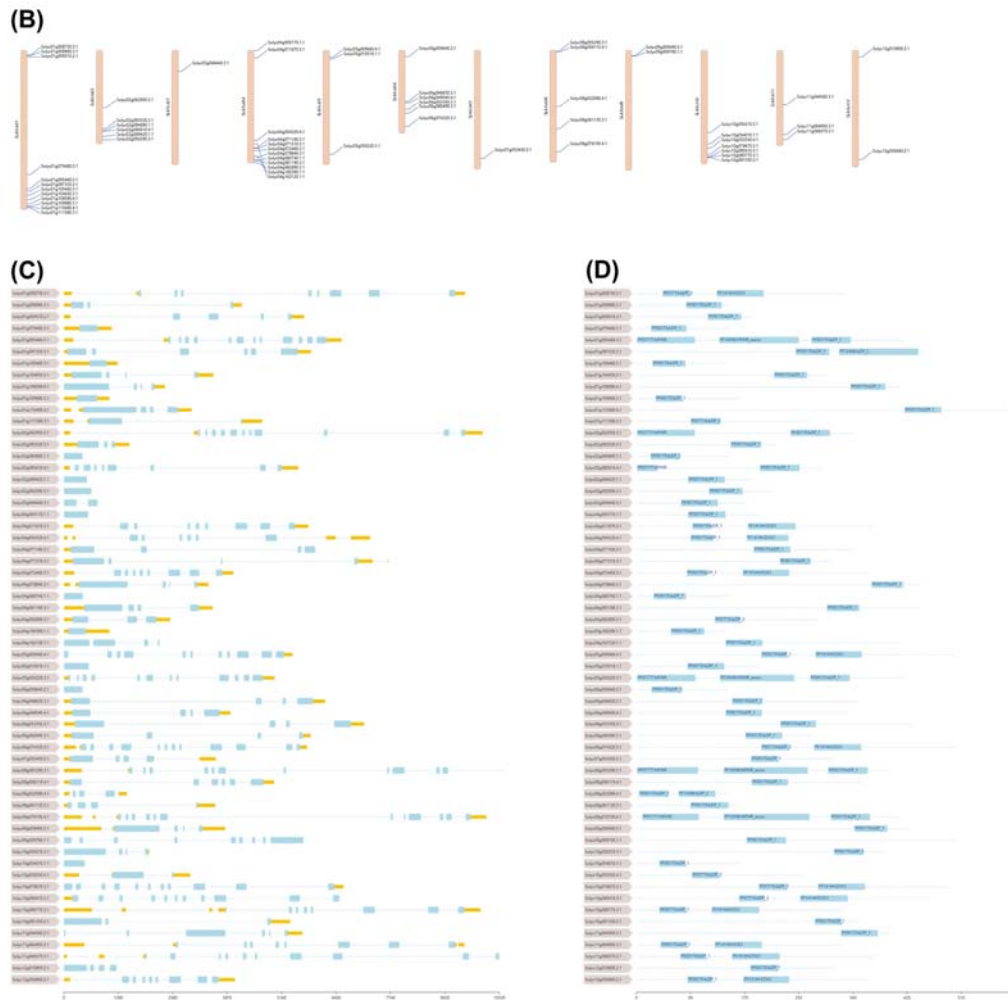


Figure 3. Gene family page in HortGenome Search Engine. Screenshots of the list and

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<sup>22</sup> 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 displays 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<sup>26</sup> (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



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<sup>27</sup>. 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<sup>28</sup>. Similarly, the ‘KEGG Enrichment Analysis’ tool has been developed using KEGG pathways assigned to genes via eggno-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<sup>29,30</sup>. 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.

**(A)**

Paste query sequence(s) or drag file containing query sequence(s) in FASTA format here ...

Advanced Parameters:  ?

Nucleotide databases

mRNA/CDS databases

Search mRNA/CDS databases...

- All
- Alismatales
- Amborellales
  - Amborellaceae
    - Amborella trichopoda v1 mRNA
- Asparagales
- Asterales

Protein databases

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
- Piperiales
- Saxifragales
- Brassicales
- Solanales
- Oxalidales
- Laurales
- Trochodendrales
- Zingiberales
- Lamiales
- Phageliales

2. Enter the list of gene/transcript IDs

Feature Type:

Input:

Options: ☒ With Genebody ☐ Without Genebody

Upstream Bases:

Downstream Bases:

Query Types: ☒ Sequence Retrieval

**(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:

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 (Trifoliolate orange)
- Poncirus trifoliata (Trifoliolate orange)
- Citrus reticulata (Wild mandarin)
- Atalantia buxifolia (Chinese box orange; primitive citrus)
- Citrus x clementina cv. Clemenules (Clementine mandarin)

## **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<sup>9</sup>. The tool has been adopted by many genome databases, including Genome Database for Rosaceae (<https://www.rosaceae.org>)<sup>8</sup>, ZEAMAP (<http://zeamap.com>)<sup>31</sup>, 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/>

366     **Conflict of interest**

367     The authors declare that they have no conflict of interest.

368

369     **Supplementary data**

370     Supplementary data is available at Horticulture Research online.

371

# 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 (LA1673)
- tomato (Accession LA0317)
- tomato (Accession LA0407)
- tomato Heinz 1706 (BTI)
- tomato Heinz 1706 (CAU)
- Current tomato (LA2093)
- Wild tomato (LA0716)
- wild tomato (LA1353)

## (C)

tomato Heinz 1706 (BTI) Soly

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) IPR

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))	Soly01G190070.1	Soly01G190070.1.1	SL4-00R1:1907399-19073997	ethylene receptor
Solanum lycopersicum (tomato Heinz 1706 (BTI))	Soly03G123450.1	Soly03G123450.1.1	SL4-00R3:5476849-5476866	ethylene receptor
Solanum lycopersicum (tomato Heinz 1706 (BTI))	Soly04G16680.1	Soly04G16680.1.1	SL4-00R4:2053959-2053959	ethylene receptor
Solanum lycopersicum (tomato Heinz 1706 (BTI))	Soly04G16430.1	Soly04G16430.1.1	SL4-00R4:5535411-5535555	ethylene receptor
Solanum lycopersicum (tomato Heinz 1706 (BTI))	Soly05G155070.4	Soly05G155070.4.1	SL4-00R5:6424793-6424835	ethylene receptor
Solanum lycopersicum (tomato Heinz 1706 (BTI))	Soly06G136450.1	Soly06G136450.1.1	SL4-00R6:2304493-2304479	ethylene receptor
Solanum lycopersicum (tomato Heinz 1706 (BTI))	Soly06G1610.1	Soly06G1610.1.1	SL4-00R6:3205071-3205071	ethylene receptor
Solanum lycopersicum (tomato Heinz 1706 (BTI))	Soly06G15710.3	Soly06G15710.3.1	SL4-00R6:3432844-3434583	ethylene receptor
Solanum lycopersicum (tomato Heinz 1706 (BTI))	Soly07G165580.3	Soly07G165580.3.1	SL4-00R7:5430692-5431254	ethylene receptor
Solanum lycopersicum (tomato Heinz 1706 (BTI))	Soly08G16830.1	Soly08G16830.1.1	SL4-00R8:5553847-5554259	ethylene receptor

Showing 1 to 10 of 20 rows 10 rows per page

(A)

Taxonomy

Taxonomy ID	4281
Scientific name	<i>Solanum lycopersicum</i> ex. Heinz (708)
Common name	tomato Heinz 1706 (571)
Order	Solanales
Family	Solanaceae
Genus	Solanum
Species	<i>Lycopersicon</i>
Subspecies	Unknown
Variety	Unknown
Cultivar	Heinz (708)



Genome assembly and annotation

Project	d1000
Platform/technology	N/A
Genome size (Mb)	12
Genome size (Mb)	710 (5) (10)
No. protein-coding genes	34,681
No. mRNAs	34,681
No. CDSs	34,681
No. proteins	34,681
Download	<a href="#">GenBank</a> <a href="#">NCBI</a> <a href="#">Ensembl</a>

Publication

Title	The tomato genome sequence provides insights into fleshy fruit evolution.
Authors	The Tomato Genome Consortium
Date	2012 May 10
PMID	22653594
Abstract	Tomato ( <i>Solanum lycopersicum</i> ) is a major crop plant and a model system for fruit development. <i>Solanum</i> 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, <i>Solanum pimpinellifolium</i> , and compare them to each other and to the potato genome ( <i>Solanum tuberosum</i> ). The two tomato genomes show only 0.5% nucleotide divergence and signs of recent admixture, but show more than 8% divergence from potato, with one 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 <i>Solanum</i> lineage has experienced two consecutive genome duplications: one that is ancient and shared with apple, and a more recent one. These duplications set the stage for the neofunctionalization of genes controlling fruit characteristics, such as colour and fleshiness.

(B)

Transcription Factors

AP2-ERF (10)	AP2/ERF-AP2 (20)	AP2/ERF-ERF (140)	AP2/ERF-MY (2)	ERF (34)	ERF-AP2 (20)
BHLH (141)	BRN (8)	BHLH (141)	ERF (1)	ERF (34)	ERF-ERF (10)
CHC2-ERF (30)	CHC2-ERF (30)	CHC2-ERF (30)	CHC2-ERF (30)	ERF (34)	ERF-ERF (10)
ERF (34)	ERF (34)	ERF (34)	ERF (34)	ERF (34)	ERF-ERF (10)
ERF-ERF (10)	ERF-ERF (10)	ERF-ERF (10)	ERF-ERF (10)	ERF-ERF (10)	ERF-ERF (10)
ERF-ERF (10)	ERF-ERF (10)	ERF-ERF (10)	ERF-ERF (10)	ERF-ERF (10)	ERF-ERF (10)
ERF-ERF (10)	ERF-ERF (10)	ERF-ERF (10)	ERF-ERF (10)	ERF-ERF (10)	ERF-ERF (10)
ERF-ERF (10)	ERF-ERF (10)	ERF-ERF (10)	ERF-ERF (10)	ERF-ERF (10)	ERF-ERF (10)
ERF-ERF (10)	ERF-ERF (10)	ERF-ERF (10)	ERF-ERF (10)	ERF-ERF (10)	ERF-ERF (10)
ERF-ERF (10)	ERF-ERF (10)	ERF-ERF (10)	ERF-ERF (10)	ERF-ERF (10)	ERF-ERF (10)

Transcriptional Regulators

ARF (10)	ARF (10)	ARF (10)	ARF (10)	ARF (10)	ARF (10)
ARF (10)	ARF (10)	ARF (10)	ARF (10)	ARF (10)	ARF (10)
ARF (10)	ARF (10)	ARF (10)	ARF (10)	ARF (10)	ARF (10)
ARF (10)	ARF (10)	ARF (10)	ARF (10)	ARF (10)	ARF (10)
ARF (10)	ARF (10)	ARF (10)	ARF (10)	ARF (10)	ARF (10)
ARF (10)	ARF (10)	ARF (10)	ARF (10)	ARF (10)	ARF (10)
ARF (10)	ARF (10)	ARF (10)	ARF (10)	ARF (10)	ARF (10)
ARF (10)	ARF (10)	ARF (10)	ARF (10)	ARF (10)	ARF (10)
ARF (10)	ARF (10)	ARF (10)	ARF (10)	ARF (10)	ARF (10)
ARF (10)	ARF (10)	ARF (10)	ARF (10)	ARF (10)	ARF (10)

Protein Kinases

Protein Kinase family Two (10) (10) (10) (10) (10) (10)

- Group A00
- Group A01
- Group A02
- Group A03
- Group A04
- Group A05
- Group A06
- Group A07
- Group A08
- Group A09
- Group A10
- Group A11
- Group A12
- Group A13
- Group A14
- Group A15
- Group A16
- Group A17
- Group A18
- Group A19
- Group A20
- Group A21
- Group A22
- Group A23
- Group A24
- Group A25
- Group A26
- Group A27
- Group A28
- Group A29
- Group A30
- Group A31
- Group A32
- Group A33
- Group A34
- Group A35
- Group A36
- Group A37
- Group A38
- Group A39
- Group A40
- Group A41
- Group A42
- Group A43
- Group A44
- Group A45
- Group A46
- Group A47
- Group A48
- Group A49
- Group A50
- Group A51
- Group A52
- Group A53
- Group A54
- Group A55
- Group A56
- Group A57
- Group A58
- Group A59
- Group A60
- Group A61
- Group A62
- Group A63
- Group A64
- Group A65
- Group A66
- Group A67
- Group A68
- Group A69
- Group A70
- Group A71
- Group A72
- Group A73
- Group A74
- Group A75
- Group A76
- Group A77
- Group A78
- Group A79
- Group A80
- Group A81
- Group A82
- Group A83
- Group A84
- Group A85
- Group A86
- Group A87
- Group A88
- Group A89
- Group A90
- Group A91
- Group A92
- Group A93
- Group A94
- Group A95
- Group A96
- Group A97
- Group A98
- Group A99
- Group A100
- Group A101
- Group A102
- Group A103
- Group A104
- Group A105
- Group A106
- Group A107
- Group A108
- Group A109
- Group A110
- Group A111
- Group A112
- Group A113
- Group A114
- Group A115
- Group A116
- Group A117
- Group A118
- Group A119
- Group A120
- Group A121
- Group A122
- Group A123
- Group A124
- Group A125
- Group A126
- Group A127
- Group A128
- Group A129
- Group A130
- Group A131
- Group A132
- Group A133
- Group A134
- Group A135
- Group A136
- Group A137
- Group A138
- Group A139
- Group A140
- Group A141
- Group A142
- Group A143
- Group A144
- Group A145
- Group A146
- Group A147
- Group A148
- Group A149
- Group A150
- Group A151
- Group A152
- Group A153
- Group A154
- Group A155
- Group A156
- Group A157
- Group A158
- Group A159
- Group A160
- Group A161
- Group A162
- Group A163
- Group A164
- Group A165
- Group A166
- Group A167
- Group A168
- Group A169
- Group A170
- Group A171
- Group A172
- Group A173
- Group A174
- Group A175
- Group A176
- Group A177
- Group A178
- Group A179
- Group A180
- Group A181
- Group A182
- Group A183
- Group A184
- Group A185
- Group A186
- Group A187
- Group A188
- Group A189
- Group A190
- Group A191
- Group A192
- Group A193
- Group A194
- Group A195
- Group A196
- Group A197
- Group A198
- Group A199
- Group A200
- Group A201
- Group A202
- Group A203
- Group A204
- Group A205
- Group A206
- Group A207
- Group A208
- Group A209
- Group A210
- Group A211
- Group A212
- Group A213
- Group A214
- Group A215
- Group A216
- Group A217
- Group A218
- Group A219
- Group A220
- Group A221
- Group A222
- Group A223
- Group A224
- Group A225
- Group A226
- Group A227
- Group A228
- Group A229
- Group A230
- Group A231
- Group A232
- Group A233
- Group A234
- Group A235
- Group A236
- Group A237
- Group A238
- Group A239
- Group A240
- Group A241
- Group A242
- Group A243
- Group A244
- Group A245
- Group A246
- Group A247
- Group A248
- Group A249
- Group A250
- Group A251
- Group A252
- Group A253
- Group A254
- Group A255
- Group A256
- Group A257
- Group A258
- Group A259
- Group A260
- Group A261
- Group A262
- Group A263
- Group A264
- Group A265
- Group A266
- Group A267
- Group A268
- Group A269
- Group A270
- Group A271
- Group A272
- Group A273
- Group A274
- Group A275
- Group A276
- Group A277
- Group A278
- Group A279
- Group A280
- Group A281
- Group A282
- Group A283
- Group A284
- Group A285
- Group A286
- Group A287
- Group A288
- Group A289
- Group A290
- Group A291
- Group A292
- Group A293
- Group A294
- Group A295
- Group A296
- Group A297
- Group A298
- Group A299
- Group A300
- Group A301
- Group A302
- Group A303
- Group A304
- Group A305
- Group A306
- Group A307
- Group A308
- Group A309
- Group A310
- Group A311
- Group A312
- Group A313
- Group A314
- Group A315
- Group A316
- Group A317
- Group A318
- Group A319
- Group A320
- Group A321
- Group A322
- Group A323
- Group A324
- Group A325
- Group A326
- Group A327
- Group A328
- Group A329
- Group A330
- Group A331
- Group A332
- Group A333
- Group A334
- Group A335
- Group A336
- Group A337
- Group A338
- Group A339
- Group A340
- Group A341
- Group A342
- Group A343
- Group A344
- Group A345
- Group A346
- Group A347
- Group A348
- Group A349
- Group A350
- Group A351
- Group A352
- Group A353
- Group A354
- Group A355
- Group A356
- Group A357
- Group A358
- Group A359
- Group A360
- Group A361
- Group A362
- Group A363
- Group A364
- Group A365
- Group A366
- Group A367
- Group A368
- Group A369
- Group A370
- Group A371
- Group A372
- Group A373
- Group A374
- Group A375
- Group A376
- Group A377
- Group A378
- Group A379
- Group A380
- Group A381
- Group A382
- Group A383
- Group A384
- Group A385
- Group A386
- Group A387
- Group A388
- Group A389
- Group A390
- Group A391
- Group A392
- Group A393
- Group A394
- Group A395
- Group A396
- Group A397
- Group A398
- Group A399
- Group A400
- Group A401
- Group A402
- Group A403
- Group A404
- Group A405
- Group A406
- Group A407
- Group A408
- Group A409
- Group A410
- Group A411
- Group A412
- Group A413
- Group A414
- Group A415
- Group A416
- Group A417
- Group A418
- Group A419
- Group A420
- Group A421
- Group A422
- Group A423
- Group A424
- Group A425
- Group A426
- Group A427
- Group A428
- Group A429
- Group A430
- Group A431
- Group A432
- Group A433
- Group A434
- Group A435
- Group A436
- Group A437
- Group A438
- Group A439
- Group A440
- Group A441
- Group A442
- Group A443
- Group A444
- Group A445
- Group A446
- Group A447
- Group A448
- Group A449
- Group A450
- Group A451
- Group A452
- Group A453
- Group A454
- Group A455
- Group A456
- Group A457
- Group A458
- Group A459
- Group A460
- Group A461
- Group A462
- Group A463
- Group A464
- Group A465
- Group A466
- Group A467
- Group A468
- Group A469
- Group A470
- Group A471
- Group A472
- Group A473
- Group A474
- Group A475
- Group A476
- Group A477
- Group A478
- Group A479
- Group A480
- Group A481
- Group A482
- Group A483
- Group A484
- Group A485
- Group A486
- Group A487
- Group A488
- Group A489
- Group A490
- Group A491
- Group A492
- Group A493
- Group A494
- Group A495
- Group A496
- Group A497
- Group A498
- Group A499
- Group A500
- Group A501
- Group A502
- Group A503
- Group A504
- Group A505
- Group A506
- Group A507
- Group A508
- Group A509
- Group A510
- Group A511
- Group A512
- Group A513
- Group A514
- Group A515
- Group A516
- Group A517
- Group A518
- Group A519
- Group A520
- Group A521
- Group A522
- Group A523
- Group A524
- Group A525
- Group A526
- Group A527
- Group A528
- Group A529
- Group A530
- Group A531
- Group A532
- Group A533
- Group A534
- Group A535
- Group A536
- Group A537
- Group A538
- Group A539
- Group A540
- Group A541
- Group A542
- Group A543
- Group A544
- Group A545
- Group A546
- Group A547
- Group A548
- Group A549
- Group A550
- Group A551
- Group A552
- Group A553
- Group A554
- Group A555
- Group A556
- Group A557
- Group A558
- Group A559
- Group A560
- Group A561
- Group A562
- Group A563
- Group A564
- Group A565
- Group A566
- Group A567
- Group A568
- Group A569
- Group A570
- Group A571
- Group A572
- Group A573
- Group A574
- Group A575
- Group A576
- Group A577
- Group A578
- Group A579
- Group A580
- Group A581
- Group A582
- Group A583
- Group A584
- Group A585
- Group A586
- Group A587
- Group A588
- Group A589
- Group A590
- Group A591
- Group A592
- Group A593
- Group A594
- Group A595
- Group A596
- Group A597
- Group A598
- Group A599
- Group A600
- Group A601
- Group A602
- Group A603
- Group A604
- Group A605
- Group A606
- Group A607
- Group A608
- Group A609
- Group A610
- Group A611
- Group A612
- Group A613
- Group A614
- Group A615
- Group A616
- Group A617
- Group A618
- Group A619
- Group A620
- Group A621
- Group A622
- Group A623
- Group A624
- Group A625
- Group A626
- Group A627
- Group A628
- Group A629
- Group A630
- Group A631
- Group A632
- Group A633
- Group A634
- Group A635
- Group A636
- Group A637
- Group A638
- Group A639
- Group A640
- Group A641
- Group A642
- Group A643
- Group A644
- Group A645
- Group A646
- Group A647
- Group A648
- Group A649
- Group A650
- Group A651
- Group A652
- Group A653
- Group A654
- Group A655
- Group A656
- Group A657
- Group A658
- Group A659
- Group A660
- Group A661
- Group A662
- Group A663
- Group A664
- Group A665
- Group A666
- Group A667
- Group A668
- Group A669
- Group A670
- Group A671
- Group A672
- Group A673
- Group A674
- Group A675
- Group A676
- Group A677
- Group A678
- Group A679
- Group A680
- Group A681
- Group A682
- Group A683
- Group A684
- Group A685
- Group A686
- Group A687
- Group A688
- Group A689
- Group A690
- Group A691
- Group A692
- Group A693
- Group A694
- Group A695
- Group A696
- Group A697
- Group A698
- Group A699
- Group A700
- Group A701
- Group A702
- Group A703
- Group A704
- Group A705
- Group A706
- Group A707
- Group A708
- Group A709
- Group A710
- Group A711
- Group A712
- Group A713
- Group A714
- Group A715
- Group A716
- Group A717
- Group A718
- Group A719
- Group A720
- Group A721
- Group A722
- Group A723
- Group A724
- Group A725
- Group A726
- Group A727
- Group A728
- Group A729
- Group A730
- Group A731
- Group A732
- Group A733
- Group A734
- Group A735
- Group A736
- Group A737
- Group A738
- Group A739
- Group A740
- Group A741
- Group A742
- Group A743
- Group A744
- Group A745
- Group A746
- Group A747
- Group A748
- Group A749
- Group A750
- Group A751
- Group A752
- Group A753
- Group A754
- Group A755
- Group A756
- Group A757
- Group A758
- Group A759
- Group A760
- Group A761
- Group A762
- Group A763
- Group A764
- Group A765
- Group A766
- Group A767
- Group A768
- Group A769
- Group A770
- Group A771
- Group A772
- Group A773
- Group A774
- Group A775
- Group A776
- Group A777
- Group A778
- Group A779
- Group A780
- Group A781
- Group A782
- Group A783
- Group A784
- Group A785
- Group A786
- Group A787
- Group A788
- Group A789
- Group A790
- Group A791
- Group A792
- Group A793
- Group A794
- Group A795
- Group A796
- Group A797
- Group A798
- Group A799
- Group A800
- Group A801
- Group A802
- Group A803
- Group A804
- Group A805
- Group A806
- Group A807
- Group A808
- Group A809
- Group A810
- Group A811
- Group A812
- Group A813
- Group A814
- Group A815
- Group A816
- Group A817
- Group A818
- Group A819
- Group A820
- Group A821
- Group A822
- Group A823
- Group A824
- Group A825
- Group A826
- Group A827
- Group A828
- Group A829
- Group A830
- Group A831
- Group A832
- Group A833
- Group A834
- Group A835
- Group A836
- Group A837
- Group A838
- Group A839
- Group A840
- Group A841
- Group A842
- Group A843
- Group A844
- Group A845
- Group A846
- Group A847
- Group A848
- Group A849
- Group A850
- Group A851
- Group A852
- Group A853
- Group A854
- Group A855
- Group A856
- Group A857
- Group A858
- Group A859
- Group A860
- Group A861
- Group A862
- Group A863
- Group A864
- Group A865
- Group A866
- Group A867
- Group A868
- Group A869
- Group A870
- Group A871
- Group A872
- Group A873
- Group A874
- Group A875
- Group A876
- Group A877
- Group A878
- Group A879
- Group A880
- Group A881
- Group A882
- Group A883
- Group A884
- Group A885
- Group A886
- Group A887
- Group A888
- Group A889
- Group A890
- Group A891
- Group A892
- Group A893
- Group A894
- Group A895
- Group A896
- Group A897
- Group A898
- Group A899
- Group A900
- Group A901
- Group A902
- Group A903
- Group A904
- Group A905
- Group A906
- Group A907
- Group A908
- Group A909
- Group A910
- Group A911
- Group A912
- Group A913
- Group A914
- Group A915
- Group A916
- Group A917
- Group A918
- Group A919
- Group A920
- Group A921
- Group A922
- Group A923
- Group A924
- Group A925
- Group A926
- Group A927
- Group A928
- Group A929
- Group A930
- Group A931
- Group A932
- Group A933
- Group A934
- Group A935
- Group A936
- Group A937
- Group A938
- Group A939
- Group A940
- Group A941
- Group A942
- Group A943
- Group A944
- Group A945
- Group A946
- Group A947
- Group A948
- Group A949
- Group A950
- Group A951
- Group A952
- Group A953
- Group A954
- Group A955
- Group A956
- Group A957
- Group A958
- Group A959
- Group A960
- Group A961
- Group A962
- Group A963
- Group A964
- Group A965
- Group A966
- Group A967
- Group A96

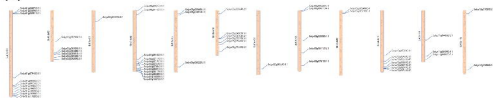
**(A)**

Show 10 entries

Search: 

Genome	Gene	Transcript	Position	Description
<i>Solanum lycopersicum</i> (tomato Heinz 1706 (ST1))	Solyd1g008730.3	Solyd1g008730.3.1	SL4 chr01:2712485-2721972	Transcription factor like
<i>Solanum lycopersicum</i> (tomato Heinz 1706 (ST1))	Solyd1g008980.3	Solyd1g008980.3.1	SL4 chr01:2942645-2949504	ARSCISIC ACID-INSENSITIVE 5-like protein 2
<i>Solanum lycopersicum</i> (tomato Heinz 1706 (ST1))	Solyd1g00910.2	Solyd1g00910.2.1	SL4 chr01:3716336-3722023	BZIP domain-containing protein
<i>Solanum lycopersicum</i> (tomato Heinz 1706 (ST1))	Solyd1g079480.3	Solyd1g079480.3.1	SL4 chr01:71067204-71096337	BZIP domain-containing protein
<i>Solanum lycopersicum</i> (tomato Heinz 1706 (ST1))	Solyd1g085480.3	Solyd1g085480.3.1	SL4 chr01:78964254-78970833	BZIP domain-containing protein
<i>Solanum lycopersicum</i> (tomato Heinz 1706 (ST1))	Solyd1g097330.3	Solyd1g097330.3.1	SL4 chr01:8048601-80493889	BZIP domain-containing protein
<i>Solanum lycopersicum</i> (tomato Heinz 1706 (ST1))	Solyd1g100400.3	Solyd1g100400.3.1	SL4 chr01:82746623-82747895	BZIP domain-containing protein
<i>Solanum lycopersicum</i> (tomato Heinz 1706 (ST1))	Solyd1g104650.3	Solyd1g104650.3.1	SL4 chr01:85389147-85392887	BZIP domain-containing protein
<i>Solanum lycopersicum</i> (tomato Heinz 1706 (ST1))	Solyd1g108680.4	Solyd1g108680.4.1	SL4 chr01:8747688-87750279	Acidic acid interactive 5-like protein
<i>Solanum lycopersicum</i> (tomato Heinz 1706 (ST1))	Solyd1g109880.3	Solyd1g109880.3.1	SL4 chr01:88070751-89071776	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 \(2K\)](#)
**(B)****(C)****(D)**

(A)

Gene: Solyc05g055070.4



### Gene Structure



(C)



BLAST of Solyc05g055070.4.1 vs. TAIR Database



(E)



Gene Ontology Terms

Term	Gene	Namespace
GO:0002024	cathepsin activity	molecular_function
GO:0004072	protein kinase activity	molecular_function
GO:0004274	protein serine/threonine kinase activity	molecular_function
GO:0051885	kinning	molecular_function
GO:0039222	intracellular anatomical structure	cellular_component
GO:0005737	cytoplasm	cellular_component
GO:0005783	endoplasmic reticulum	cellular_component
GO:0009084	cellular protein modification process	biological_process
GO:0009185	protein of endoplasmic	biological_process

(F)



## Collinearity gene pairs (Paralogs)

Transcript ID	Paralogue	E-value	Organism	Syntenic Block ID
Solyc5g0655070.4.1	Solyc11g006160.2.1	2130	<i>Solanum lycopersicum</i> SL4.0	61

## Collinearity gene pairs (Orthologs)

Accession ID	Ortholog	E-value	Organism	Symbolic Block
Syrc009560599.A.1	SPERP05G33530.1	1480	<i>Solanum pennellii</i> v2.0	41
Syrc009560599.A.1	SPMP05G173790.1	3260	<i>Solanum pimpinellifolium</i> LA2953 v1.5	21
Syrc009560599.A.1	Swed01G056884.1	2020	<i>Solanum melongena</i> HD1516	52
Syrc009560599.A.1	P05G03G3DMT400050121	3880	<i>Solanum tuberosum</i> P05G v4.83	48
Syrc009560599.A.1	Pter1715G0005609016.1	2870	<i>Petunia inflata</i> v1.0.1	21
Syrc009560599.A.1	Capon05G002354	2890	<i>Capparis americana</i> Zurich v2	118
Syrc009560599.A.1	CA19G197.000413.1	3470	<i>Solanum habrochloides</i> LA1553	192

(B)

Gene: Solyc05g055070.4



## Sequences related to gene Scylc05g055070.4



(D)



Functional Domains (InterPro)



Year	Project Description	Start	End	Source Name (Description)	Location
2016	No PM available	2016.0	2016.0	CH	1000 750 240
2017-2018	Significant increases in project engagement, moderate climate	2017.0	2018.0	ND2.0	1000 110 700 Project 1: 100.0 www.9.7
2019-2020	SAF decline	2019.0	2020.0	SAF.0	1000 100 320
2021-2022	Stable to increase (2021) then a 40% drop in 2022	2021.0	2022.0	2021.0	1000 100 320
2023-2024	Stable to increase (2023) then a 40% drop	2023.0	2024.0	2023.0	1000 100 320
2025-2026	SAF decline to approximately	2025.0	2026.0	2025.0	1000 100 320
None	No PM available	2016.0	2016.0	CH	1000 750 240
None	No PM available	2016.0	2016.0	CH	1000 750 240
2018-2019	Climate change	2018.0	2019.0	ND2.0	1000 110 700 Project 1: 100.0 www.9.7

[illegible]

**(A)**

Paste query sequence(s) or drag file containing query sequence(s) in FASTA format here ...

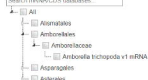
Advanced Parameters:  ?

BLAST

#### Nucleotide databases

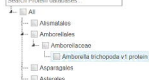
##### mRNA/CDS databases

Search mRNA/CDS databases.

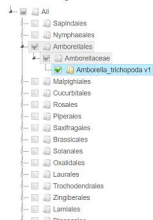


#### Protein databases

Search Protein databases.

**(B)**

#### 1. Choose one genome



#### 2. Enter the list of gene/transcript IDs

Feature Type:

gene

Input:

Gene ID / Transcript ID

Options

☒ With Genebody

☐ Without Genebody

Upstream Bases:

0

Downstream Bases:

0

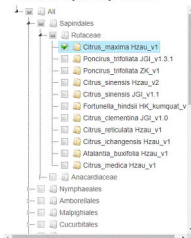
Query Type:

☒ Sequence Retrieval

Download

**(C)**

#### Search Synteny Blocks



Some selections:

Chromosome/Scaffold:

chr1

Choose a genome for comparison:

choose a genome for comparison

Citrus ichangensis (Ichang papaya; 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 (Trifoliolate orange)

Poncirus trifoliata (Trifoliolate orange)

Citrus reticulata (Wild mandarin)

Atalantia buxifolia (Chinese box orange; primitive citrus)

Citrus x clementina cv. Clementines (Clementine mandarin)