Nucleic Acids Research, 2025, **53**, D1623–D1632
https://doi.org/10.1093/nar/gkae1040
Advance access publication date: 11 November 2024
Database issue

SoIR: a comprehensive Solanaceae information resource for comparative and functional genomic study

Zhuo Liu [1,†], Shaoqin Shen [1,†], Chunjin Li [1,†], Chenhao Zhang [1,†], Xiang Chen [2,†], Yanhong Fu [1], Tong Yu [1], Rong Zhou [3,4], Dongxu Liu [2], Qing-Yong Yang [2,*] and Xiaoming Song [1,*]

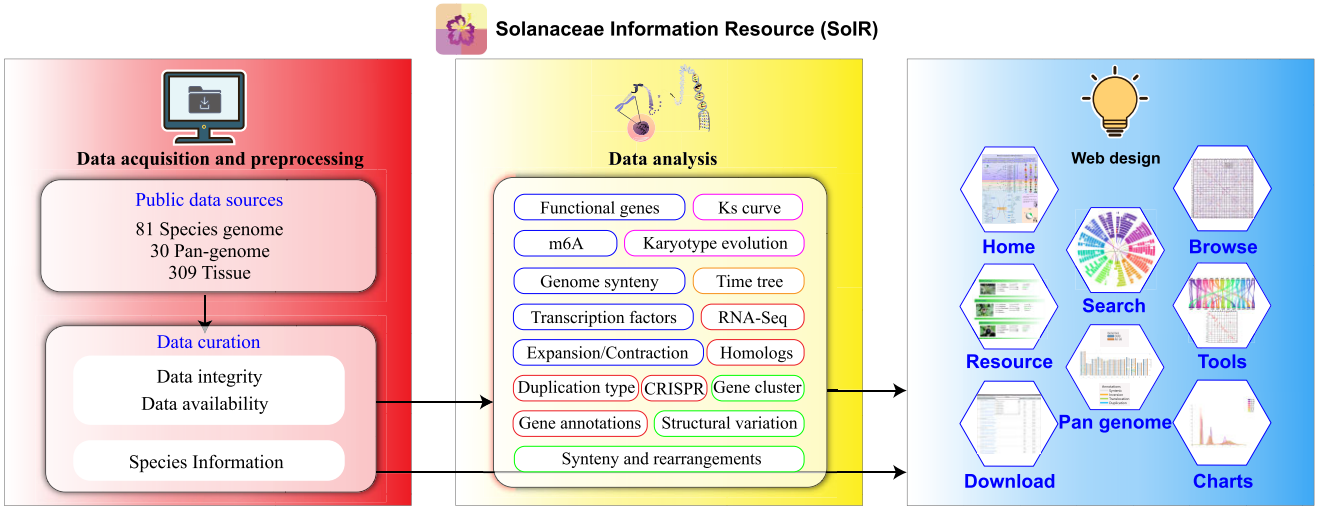
[1]School of Life Sciences/School of Basic Medical Sciences/Key Laboratory for Quality of Salt Alkali Resistant TCM of Hebei Administration of TCM, North China University of Science and Technology, Tangshan, Hebei 063210, China
[2]National Key Laboratory of Crop Genetic Improvement, Hubei Engineering Technology Research Center of Agricultural Big Data, Hubei Hongshan Laboratory, Huazhong Agricultural University, Wuhan 430070, China
[3]College of Horticulture, Nanjing Agricultural University, Nanjing 210095, China
[4]Department of Food Science, Aarhus University, Aarhus 8200, Denmark

[*]To whom correspondence should be addressed. Tel: +86 315 8805607; Fax: +86 315 8805607; Email: songxm@ncst.edu.cn
Correspondence may also be addressed to Qing-Yong Yang. Tel: +86 27 87280877; Fax: +86 27 87280877; Email: yqy@mail.hzau.edu.cn
[†]The first five authors should be regarded as Joint First Authors.

Abstract

The Solanaceae family, which includes economically important crops such as tomatoes, potatoes and peppers, has experienced a rapid expansion in genomic data due to advancements in sequencing technologies. However, existing databases are limited by incomplete species representation, a lack of comprehensive comparative genomic tools and the absence of systematic pan-genomic analyses. To address these gaps, we developed the Solanaceae Information Resource (SoIR, https://soir.bio2db.com), a comprehensive genomics database for the Solanaceae family. SoIR integrates genomic data from 81 species and transcriptomic data from 41 species, encompassing a total of 3 908 408 gene annotations derived from Gene Ontology, nonredundant protein, Pfam, Swiss-Prot and TrEMBL databases. The resource also includes 3 437 115 CRISPR guide sequences, 212 395 transcription factors and 19 086 genes associated with methylation modification. In addition to species-specific analyses, SoIR provides extensive bioinformatics tools for investigating gene family evolution, phylogenetic relationships and karyotype reconstruction across 25 fully sequenced genomes. With advanced tools such as Blast, Synteny and Sequence Alignment, the platform provides users with interactive and intuitive visualizations for conducting cross-species comparative genomics. As the first comprehensive pan-genomic resource for the entire Solanaceae family, SoIR facilitates in-depth cross-species analysis, supporting global research initiatives in plant evolution, functional genomics and crop improvement.

Graphical abstract



Received: August 9, 2024. **Revised:** September 27, 2024. **Editorial Decision:** October 12, 2024. **Accepted:** October 19, 2024

Introduction

The Solanaceae family, with ∼100 genera and 2700–3000 species, is one of the most economically and biologically significant angiosperm families. The family includes widely cultivated and consumed crops such as potato (*Solanum tuberosum*), tomato (*Solanum lycopersicum*), eggplant (*Solanum melongena*), pepper (*Capsicum annuum*) and tobacco (*Nicotiana tabacum*), all of which are essential for global agriculture and human nutrition (1,2). Beyond their agricultural importance, species such as tomato, potato and tobacco serve as crucial model organisms in biological research.

The genomic study of Solanaceae began with major milestones such as the publication of the first potato genome by the Potato Genome Sequencing Alliance in 2011 (3), followed by the first tomato genome by the Tomato Genome Alliance in 2012 (4). These achievements marked the beginning of a new era of genomics within the Solanaceae family. Since then, advancements in sequencing technologies have driven significant progress in Solanaceae genomics, expanding genome and transcriptome studies to a variety of species, including peppers (5–7), tomatoes (4,8), potatoes (2,3,9) and eggplants (10). To date, genomic data have been assembled for 81 Solanaceae species across 13 genera (Supplementary Table S1), providing a rich foundation for further research.

These extensive genomic resources enable comparative and functional genomics investigation into key aspects of Solanaceae biology, such as evolution pathways, domestication processes and underlying molecular mechanisms. Several comprehensive multi-omics databases have been established based on these genomic datasets, including the Solanaceae Genomics Network (SGN, http://solgenomics.net) (11), Tomato Functional Genomics Database (TFGD, http://ted.bti.cornell.edu) (12), Eggplant Genome Database (EggplantGD, http://eggplant.kazusa.or.jp) (13), Pepper Genomics Database (PepperGD, https://peppergenome.snu.ac.kr) (6) and Potato Genomics Resource (Spud DB, http://spuddb.uga.edu) (14). Among these, SGN stands out for its widespread use and recognition within the scientific community.

Despite the progress, the rapid development of omics technologies, particularly in the realm of pan-genomics, has created a growing need to systematically organize and analyze the vast and ever-expanding body of data. To meet this challenge, we developed the Solanaceae Information Resource (SoIR), a comprehensive genome resource platform. SoIR aims to serve as a central repository for comparative genomics, functional genomics, variation analysis and evolutionary research within the Solanaceae family. By providing robust tools and resources, SoIR is designed to facilitate and advance cutting-edge research in the Solanaceae family.

Materials and methods

Data sources and processing

Genome sequences, gene annotations in general feature format (GFF), coding sequences (CDS) and protein sequences for various Solanaceae species were collected from several major public databases. The majority of genomes were obtained from the National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov), the National Genomics Data Center (https://ngdc.cncb.ac.cn), the Potato Database (https://solomics.agis.org.cn/potato/) (1), the Vegetable Information Resource (http://tvir.bio2db.com) (15) and the Sol Genomics Network (SGN) (https://solgenomics.net) (11). Additional species-specific data were sourced from other specialized databases (Supplementary Table S1). In total, genomic data for 81 Solanaceae species were compiled, with detailed information on taxonomy, sequencing information, references and related databases. For species with multiple genome versions, the most recent, high-quality or widely used versions were selected for further bioinformatics analysis.

Genomic analysis

Four databases were used to annotate 81 Solanaceae species: the Pfam database (16), the UniProt database (17), the Nonredundant Protein Sequence database (https://www.ncbi.nlm.nih.gov) and the Gene Ontology (GO) database (18). The CasFinder pipeline was used for designing CRISPR–Cas9 target sites (19). Orthologous and paralogous sequences were identified using OrthoFinder (20), and gene trees and species trees were constructed with FastTree (21).

Collinearity analysis was conducted using MCScanX, based on BLASTP results and GFF annotations (22,23). The resulting collinearity blocks and point diagrams were saved in JSON format, and visualizations were created using the D3 library. MCScanX's duplicate_gene_classifier function was used to categorize gene duplicates.

Evolution analysis

Gene families across the 81 species were identified using OrthoFinder (20). These gene families were then classified into single-copy and multi-copy groups using the MCL graph clustering algorithm. The parsed MCL output was fed into CAFÉ (v4.2.1) to analyze gene family contraction and expansion (24). Homologous sequences were aligned using MUSCLE (25), and protein alignments were converted to codon alignments using PAL2NAL (26). The nonsynonymous (K_a) and synonymous (K_s) substitution rates were calculated using the yn00 module of PAML (27). K_s density distributions were determined with three modules of WGDI: Kspeak (kp), PeaksFit (pf) and KsFigures (kf) (28).

Phylogenetic trees were constructed using maximum likelihood models implemented in RaxML (29). Divergence times were estimated based on single-copy gene families and species trees, using the r8s method (30). Time calibration points were extracted from the Timetree database (31). To infer the evolutionary trajectory of Solanaceae species, we analyzed homologous gene dot plots, tracing chromosomal evolution from ancestral to present configurations based on previous published studies (32,33).

Transcription factor and candidate functional gene identification

Transcription factor (TF) families were identified using the PlantTFDB database (34,35). *Arabidopsis* anthocyanin-related genes were drawn from previous studies, amounting to 41 genes in total (36,37). For the identification of homologous flowering and anthocyanin genes in other species, BLASTP was used with stringent criteria (*E*-value <1e−5; identity >60%; score >150), followed by manual verification to ensure accuracy (36,38). Hormone-related genes were identified by employing a combination of BLAST and Pfam methods, following established protocols (36,39). Resistance gene

analogs were detected using the pipeline of RGAugury (40). Additionally, m^6A, including categories of writers, readers and erasers (41), 5-methyladenosine genes and histone H3 genes were identified using the Pfam database, adhering to previous reports (15,42).

Pan-genome analysis

Pan-genome analysis was conducted using protein sequences from 81 Solanaceae species. Gene families were inferred using OrthoFinder (20). Following established methodologies (43,44), pan-genome analysis classified gene clusters into core, soft-core, dispensable and specific clusters. Core clusters were defined as those present across all 81 genomes, while soft-core clusters were those present in 79–80 species. Dispensable clusters were defined as present in 2–78 species, and specific clusters as unique to a single species. The number of protein-coding genes in both the pan-genome and core genome was estimated using PanGP (v1.0.1) with a random sampling algorithm, selecting a sample size of 2000 and conducting 80 replicates. Pairwise genome alignments between potato and tomato were conducted using the nucmer program from the MUMmer package (v4.0.0beta2) (45). Structural variant (SV) detection was performed using SVMU (v0.4-alpha) (46) to generate copy number variants (CNVs), insertions and deletions. For SV calling, minimap2 (v2.21-r1071) (47) was used to generate paired genome alignments, which were then processed by SyRI (v1.2) (48) to identify and classify SVs, including insertions, deletions, single nucleotide polymorphisms (SNPs), inversions and translocations.

Transcriptome analysis

Publicly available transcriptome data were retrieved from the NCBI and NGDC databases. Initial quality control of raw sequencing data was conducted using fastp software (49) to ensure data quality. Adapter sequences were trimmed using Trimmomatic (v0.36) (50), after which the filtered reads were aligned to the reference genome using hisat2 (v2.2.1) (51). Gene expression quantification was performed using the run-featurecounts.R script (52). Finally, expression levels for all genes were normalized and summarized based on FPKM (fragments per kilobase million) and TPM (transcripts per million) values (53), with all data merged into a comprehensive expression matrix.

Database construction

The SoIR database was constructed following established methodologies (38,42,54). Specifically, the database architecture was built with the Django framework, paired with MySQL for efficient database management and storage of genome-related datasets. The SoIR platform was developed using HTML, CSS, JavaScript and Python (55). Data visualization, including charts and graphs, was performed using ECharts and the D3 library. An interactive web interface was designed to allow users to easily access and retrieve relevant information from the SoIR database (56). The interaction between the frontend and backend was handled using Python, JavaScript and HTML, ensuring smooth extraction, processing and presentation of data from the MySQL database to the user.

Results

Overview of the main interfaces of the SoIR database

To facilitate comprehensive comparative analysis of Solanaceae genomes, we collected genomic data from 81 Solanaceae species (Supplementary Table S1). Systematic bioinformatics analyses were performed on these datasets, including gene annotation, homology, duplication types, CRISPR guide sequences, pan-genome analysis, homologous gene identification, TF families and m^6A gene information. Additional analyses included gene family identification, divergence time estimation, homologous collinearity and karyotype evolution analysis across 25 fully sequenced Solanaceae species. Furthermore, transcriptome analyses were conducted on 41 Solanaceae species with available data. The SoIR database was developed to provide user-friendly access to these extensive genomic resources and bioinformatics results. The database interface features multiple sections, including Home, Browse, Pan-genome, Search, Charts, Resources, Download, Tools, Help and Contact (Figure 1 and Supplementary Figure S1).

Home interface

The Home interface provides a comprehensive overview of gene family evolution across 36 species, including 35 Solanaceae species and 1 grape species (*Vitis vinifera*) (Supplementary Figure S2A and B, and Supplementary Table S2). A total of 49 437 gene families were identified, with *Physalis grisea* exhibiting the highest number of gene families (22 617), followed by *Physalis pruinosa* (22 578) and *Solanum commersonii* (22 407) (Supplementary Table S3). Among these, 6789 gene families were conserved across all 36 species (Supplementary Figure S2B and Supplementary Table S3). Notably, *Solanum laciniatum* had 954 unique gene families, more than any other Solanaceae species.

Gene family contraction and expansion analyses were also performed (Supplementary Figure S2C and Supplementary Table S3), with *Solanum retroflexum* (18 123) showing the highest number of expanded gene families, followed by *S. laciniatum* (16 472) and *Solanum candolleanum* (15 784). The greatest contraction was observed in *S. laciniatum* (16 803). Phylogenetic and divergence time analysis estimated that Solanaceae and grape diverged around 118 million years ago (Mya), while *Nicotiana* diverged from other Solanaceae species ∼23.9 Mya (Figure 1A).

A species phylogenetic tree was prominently displayed on the SoIR homepage, with interactive elements enabling users to access the expansion/contraction modules within the Browse section by hovering over or clicking branch nodes (Figure 2A). Species names and images are hyperlinked to species-specific search results pages, providing detailed information and analysis results (Supplementary Figure S1). A site structure diagram was also added to the homepage for easy navigation.

Search interface

The Search interface allows users to access genomic data, including gene annotations, homologous genes, synteny, duplication types, CRISPR guide sequences and transcriptome data (Supplementary Figure S1). A total of 3 908 408 genes from 81 Solanaceae species were annotated using

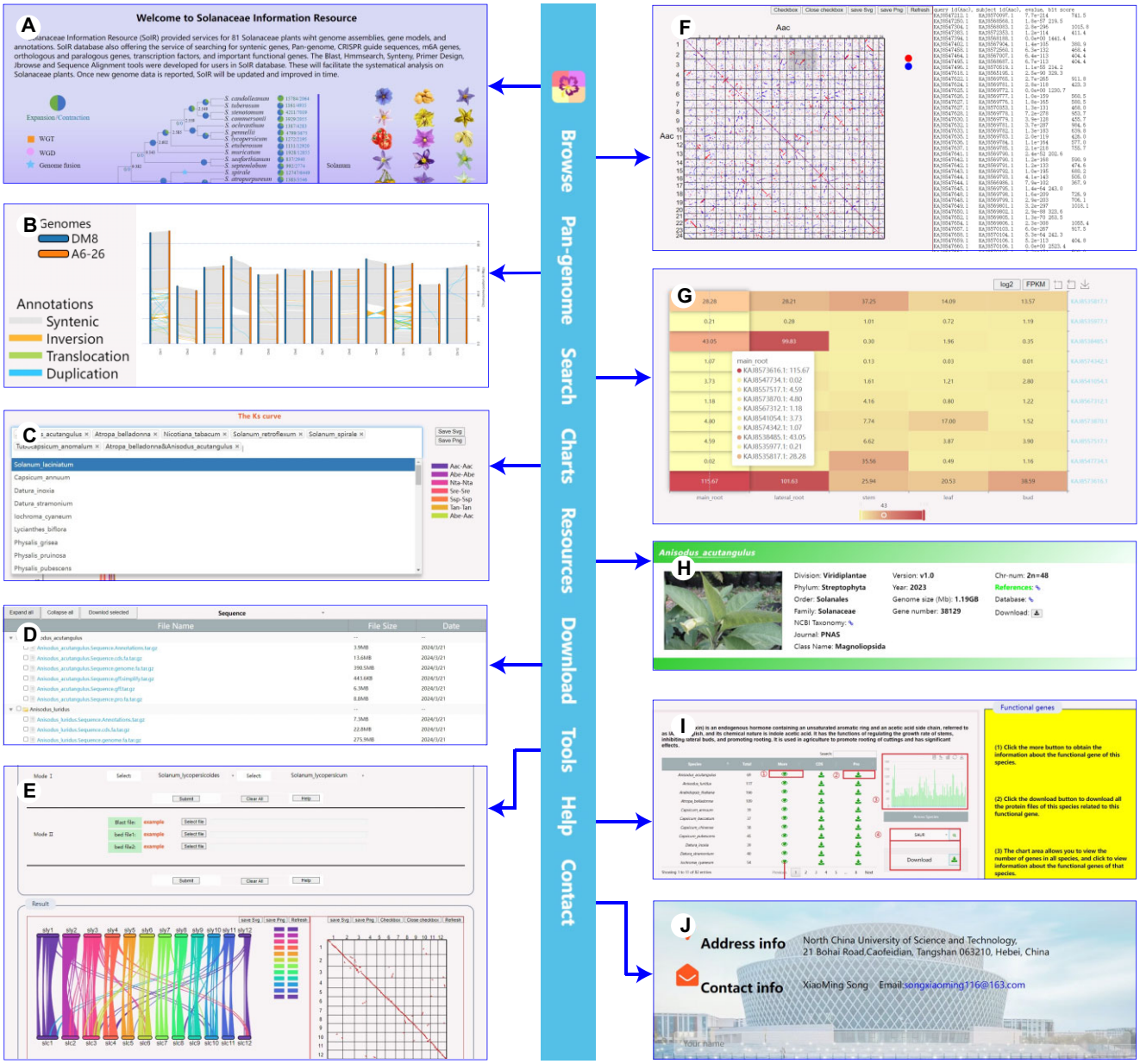


Figure 1. Overview of the main interfaces and internal features of the SolR database. (**A**) Home interface. (**B**) Pan-genome interface. (**C**) Charts interface. (**D**) Download interface. (**E**) Tools interface. (**F**) Browse interface. (**G**) Search interface. (**H**) Resources interface. (**I**) Help interface. (**J**) Contact interface.

five protein databases, with annotation completeness varying between 85.34% in *Solanum pimpinellifolium* and 99.98% for *Solanum lycopersicum* (Supplementary Figure S3 and Supplementary Table S4). Users can search annotations by gene ID or by functional identifiers such as gene symbols (e.g. flowering locus) or login number (e.g. GO: 0035448 and Pfam: PF04874) for cross-species queries.

Homologous genes were identified using OrthoFinder, yielding 68 820 orthorhombic groups, with 12 454 being species-specific (Supplementary Tables S5 and S6). Gene phylogenetic trees were constructed for each group, allowing users to select specific orthogroup (OG) numbers and species for detailed analysis. Genome collinearity analysis was performed to investigate gene duplication and loss patterns across Solanaceae species. Homologous regions were globally aligned, showing a gene ratio of ~3:1 in most Solanaceae species compared to grape. Syntenic gene blocks were visualized using a circular collinearity diagram, with options for users to view syntenic gene clusters within 10, 20 or 30 neighboring genes (Figure 2B). Retention of homologous genomic regions varied across species, with *Solanum pennellii* exhibiting the highest gene retention rate (23.6%) and *Iochroma cyaneum* the lowest (14.1%) (Supplementary Tables S7 and S8).

Five types of gene duplications (dispersed, proximal, singleton, tandem and whole-genome duplication) were detected across all 81 species (Supplementary Figure S3 and Supplementary Table S9). Additionally, 3 437 115 CRISPR guide sequences were generated for all genes in the dataset (Supplementary Figure S3 and Supplementary Table S10). Gene expression was calculated for 42 Solanaceae species in 11 different tissues under normal conditions (Supplementary Table S11). Following the selection of species,

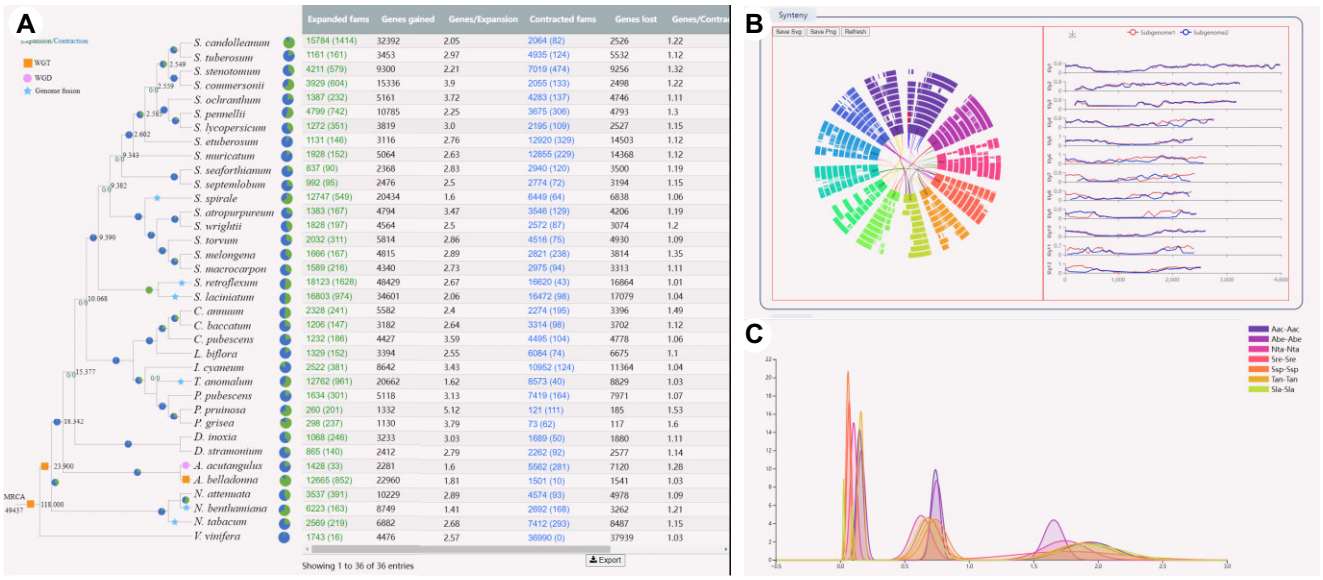


Figure 2. Overview of the Solanaceae Evolution interface in the SoIR database. (**A**) Expansion/contraction of gene families and divergence time estimation. (**B**) Global comparison of homologous regions in the genomes of Solanaceae species. (**C**) K_s curve for collinear gene pairs in intergenomic and intragenomic blocks.

the tissue of the species and 10 sample genes were shown in the SoIR database (Figure 1G).

Browse interface

The Browse interface provides user with interactive access to key genomic analysis results, such as genome synteny, TFs and functional gene families. The 'Genome synteny' tool allows users to view syntenic dot plots and diagrams for both intraspecies and interspecies comparisons at the chromosome level (Figure 1F). A total of 212 395 TFs from 58 families were detected in Solanaceae species (Supplementary Table S12 and Supplementary Figure S4A). Gene families such as SAP, HB-other, NF-YB and FAR1 showed significant expansion compared to *Arabidopsis*, while STAT, CPP, RAV, GeBP and NZZ-SPL families were reduced in most Solanaceae species (Supplementary Figure S4B and Supplementary Table S13). Notably, *Atropa belladonna* displayed substantial TF expansion following a recent whole-genome triplication (WGT) event, whereas *Aegiphila acutangulus* experienced a whole-genome duplication (WGD) but retained only a portion of the expanded TFs, likely due to subsequent gene loss (Supplementary Figure S4B).

In addition, phytohormone genes (765 376), flowering genes (17 907), anthocyanin genes (3429) and resistance genes (146 132) were predicted across the 81 Solanaceae species (Supplementary Figure S5 and Supplementary Table S14). Users can browse or download these genes by searching for specific terms, such as 'FLC' for flowering genes, and perform cross-species comparative analysis. The database also includes m^6A (3022), histone H3 (14 151) and 5-methyladenosine (1913) genes (Supplementary Table S14 and Supplementary Figure S5), contributing valuable resources for genetic studies in Solanaceae species.

Pan-genome interface

A family-wide pan-genome analysis was conducted across 81 Solanaceae species, identifying 69 580 gene clusters that include 3 888 681 genes (Supplementary Figure S6A–C). This analysis revealed 4555 core clusters, containing 1 112 519 genes, representing 6.55% of clusters and 28.61% of total genes. In addition, 6006 soft-core clusters contained 1 051 384 genes (27.04%), while 47 091 dispensable clusters included 1 679 094 genes (43.18%). Specific clusters unique to individual species totaled 11 928, covering 45 684 genes (1.17%). Further pan-genomic analysis was conducted for seven potato species and 23 tomato species (Supplementary Table S15). For potatoes, 35 120 gene clusters were identified, comprising 251 032 genes. These clusters included 15 390 core clusters, 8984 soft-core clusters, 10 049 dispensable clusters and 697 specific clusters. In tomatoes, 88 351 gene clusters were identified, comprising 678 539 genes, including 886 core clusters, 867 soft-core clusters, 82 873 dispensable clusters and 3725 specific clusters. All gene cluster data are accessible in the SoIR database, allowing users to select and view data for specific species.

Comparative analyses of structural variations were conducted by aligning six potato genomes against the *S. tuberosum* (DM8) reference genome. We identified between 2.1 and 3.01 million SNPs, along with 158 635 insertions, 153 732 deletions, 29 586 CNVs, 2171 translocations and 623 inversions. Similarly, for 22 tomato genomes aligned to the *S. lycopersicum* (SL4) reference, we detected between 0.06 and 3.4 million SNPs, 86 525 insertions, 61 470 deletions, 15 894 CNVs, 597 translocations and 843 inversions.

We established a comprehensive SV landscape by plotting all assembly variations and their collinearity with the reference genome. This landscape illustrates both collinearity and rearrangements among the six potato genomes in relation to the *S. tuberosum* (DM8) reference genome, as well as the relationship between the 22 tomato genomes and the *S. lycopersicum* (SL4) reference genome. Additionally, the collinearity and rearrangements are dynamically displayed in the database (Figures 1B and 3A), allowing users to select samples and examine variations. By clicking on specific regions of mutation collinearity and rearrangement, users can view the affected genes (Figure 3B and C).

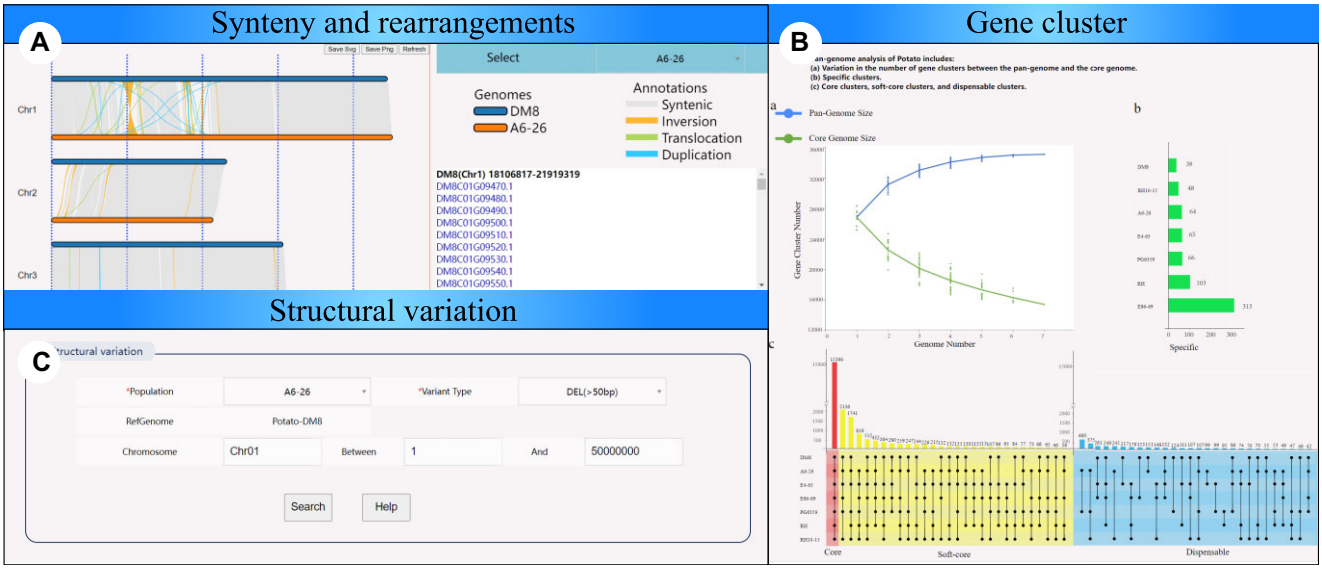


Figure 3. Overview of the Pan-genome interface in the SoIR database. (**A**) Distribution of SVs in two representative Solanaceae species. (**B**) Visualization of gene clusters within the database. (**C**) Interface for searching various types of genetic variations.

Charts interface

To understand Solanaceae genome evolution, we conducted analyses of K_s values and chromosomal karyotype changes. The Charts interface provides interactive visualizations of the K_s curve and karyotype evolution (Figure 1C). We explored genome evolution by analyzing the distribution of average synonymous substitution levels (K_s) both within and between species (Supplementary Figure S7 and Supplementary Table S16). Most Solanaceae species experienced two WGT events: the γ event, common to eudicots, and a Solanaceae-specific WGT. Additional polyploidy events occurred in *A. acutangulus* and *A. belladonna* following the common WGT event (57,58). Species divergence was assessed based on K_s peak values (Supplementary Figure S7I and J), adjusted for shared evolutionary events. The peak for the γ event was centered at 1.23, while the Solanaceae-specific WGT peak was at 0.41 (Supplementary Figure S7E–H). The database allows users to explore K_s peak maps within and between species for comparative analysis (Figure 2C).

The ancestral karyotype of eudicot species is denoted by chromosomes G1–G7, with chromosomes designated as A1–A7, B1–B7 and C1–C7 post-γ event. Homology relationships among most Solanaceae species are depicted in dot plots (Supplementary Figure S8A–C) with a 1:1 ratio. The evolution from tomato's 12 chromosomes to the 23 ancestral chromosomes of Solanaceae following WGT events is illustrated (Supplementary Figure S9). The ancestral Solanaceae karyotype prior to the WGT event was inferred using the coffee (GCA_900059795.1) reference genome. Comparison with coffee revealed a 1:3 homologous collinearity relationship (Supplementary Figure S10). We identified 15 ancestral chromosomes for Solanaceae and reconstructed their evolutionary trajectory from ancient to modern chromosomes (Supplementary Figure S11A). The transition from the 21 ancestral chromosomes of core eudicots to the 15 chromosomes of Solanaceae involved four end-to-end joining (EEJ) and two nested chromosome fusions (NCFs) (Supplementary Figure S11B). Following the WGT event, Solanaceae genomes ranged from 15 to 45 chromosomes, with observations of 20 EEJs, 2 NCFs and multiple chromosome crossovers.

Download interface

The Download interface provides users with access to datasets for 81 Solanaceae species, including genome assemblies, GFF files, CDS and protein sequences, along with their corresponding annotation data (Figure 1D). Special GFF files include chromosome information, gene start and end positions, and positive and negative strand details, enabling users to conduct collinearity and related analyses. Users have the option to download datasets in bulk based on their selections.

Tools interface

To support genomic analysis, the SoIR database features six widely used tools: Blast, Synteny, Primer Design, JBrowse, Hmmsearch and Sequence Alignment. The Blast tool enables sequence alignment using nucleotide and protein sequences from the 81 Solanaceae species. Users can perform similarity searches by either entering sequences directly or uploading files in FASTA format. The collinearity tool facilitates collinearity analysis, generating dot plots and synteny diagrams to illustrate collinearity relationships. It offers two operational modes: Mode I allows users to generate images directly by selecting species, while Mode II processes input files for visualization (Figure 1E).

The Primer Design tool assists users in designing primers by allowing FASTA sequence uploads. For Hmmsearch, users can upload sequences or enter Pfam domain ID for structural predictions. JBrowse provides a platform for exploring genomic sequences and annotations, featuring interactive gene ID functionalities that deliver detailed information. Finally, the Sequence Alignment tool enables microsynteny visualization.

Resource, Help and Contact interfaces

The Resource interface offers detailed information on each Solanaceae species, including genome size, chromosome num-

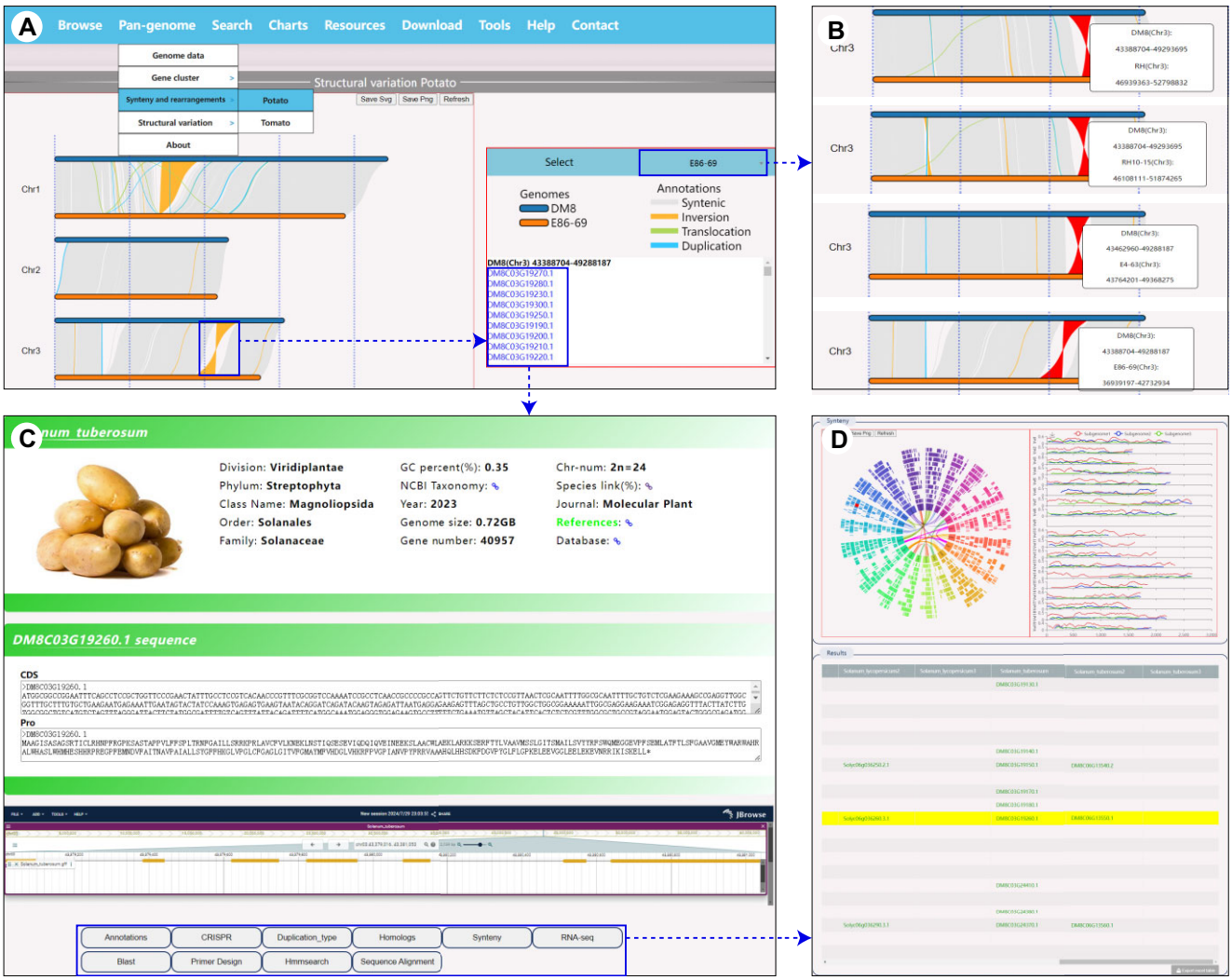


Figure 4. Case study panels from the SoIR database. (**A**) Structural variation display interface for potato. (**B**) Inversion events observed in four potato samples. (**C**) Detailed information on the gene '*DM8C03G19260.1*' located near the inversion region. (**D**) Specific display page for the gene '*DM8C03G19260.1*'.

ber and links to related databases (Figure 1H). It also provides comprehensive links to external Solanaceae databases. The Help interface includes a detailed manual for navigating the SoIR database (Figure 1I), while the Contact interface provides an email address for user inquiries (Figure 1J).

Case study

To illustrate the application of the SoIR database, we present a case study on the identification of structural variations. Inversions serve as significant catalysts for local adaptation and diversification by protecting inverted sequences from recombination. Our study identified structural variations in seven potato varieties, documenting a total of 623 inversions, with 70 of these exceeding 1 Mb in size. Initial comparisons using the potato structural variation interface (Figure 4A) revealed that the DM8 variety exhibited inversions at approximately DMchr03:43.39–49.29Mb relative to four other potato samples (E4-63, E86-69, RH, RH10-5) (Figure 4B), consistent with previous reports in *Nature* (1). By selecting the inverted region, users can access detailed gene information within this segment (Figure 4A), allowing for in-depth analysis of gene functions, as shown in Figure 4. Clicking on a gene initiates a search in the database, leading to a dedicated results page (Figure 4C). Notably, this inversion region contained 520 genes associated with carotenoid content in tubers (1). Our database identified the gene '*DM8C03G19260.1*' in the DM8 genome, located near the inversion breakpoint, which was homologous to the '*Soltu.DM.03G018410*' gene in the DM6.1 version, also linked to carotenoid content (1). The database provides comprehensive information on this gene, including its sequence, structure, functionality and associated TFs (Figure 4C). It further details gene functional annotations, CRISPR applications, synteny genes, RNA sequencing data, homologs and gene duplication types (Figure 4D). Moreover, users can easily access online analytical tools with a single click, where gene sequences are pre-populated for further analysis.

Discussion

Several databases have been developed for individual of Solanaceae species, such as SGN (11), TFGD (12), EggplantGD (13), PepperGD (6) and Spud DB (14). While these databases serve as valuable resources, they typically focus on a limited number of closely related species. For instance, SGN

Table 1. Comparison of SoIR with other accessible Solanaceae databases

Database	Species	Annotation types	Gene synteny	Gene features	Transcriptome	Pan-genome	Karyotype	Gene homology	Transcription factors	Functional genes
SoIR	81	6	√	√	√	√	√	√	√	√
SGN	16	1	–	√	√	–	–	–	–	–
TFGD	1	3	–	–	√	–	–	–	–	–
EggplantGD	1	–	–	–	–	–	–	–	–	–
Spud DB	3	1	–	√	–	–	–	–	–	–
PotatoHub	1	–	–	–	–	–	–	–	–	–
TGSol	3	1	–	√	–	–	–	–	–	–

is a widely recognized Solanaceae database, providing genetic maps and markers for key species. However, despite their breath, existing databases exhibit several limitations:

1. *Limited data integration*: Current resources offer restricted integration of genomic data across Solanaceae species, with many newly sequenced genomes absent.
2. *Lack of comparative analysis*: There is insufficient provision for comparative analysis among Solanaceae species, especially regarding functional gene similarities and phylogenetic relationships.
3. *Absence of pan-genomic analysis*: Pan-genomic analyses across Solanaceae species are largely absent from existing databases.
4. *Non-intuitive data display*: Data presentation is often cumbersome, with inadequate dynamic visualizations that hinder exploration.
5. *Suboptimal interface interaction*: User interaction between database interfaces tends to be inefficient and fragmented.

In contrast, the SoIR addresses these problems by incorporating genomic data from 81 Solanaceae species and offering comprehensive bioinformatics analyses (Table 1). SoIR provides unique insights into ancestral karyotypes and chromosome evolution, a feature lacking in other databases. By reconstructing ancestral chromosomes, SoIR sheds light on the impact of polyploidization events on genome diversity (59). While ancestral karyotypes have been reconstructed for other plant families such as Apiaceae, Cucurbitaceae and Asteraceae (32,60,61), Solanaceae's ancestral karyotypes have not been thoroughly investigated. By selecting well-assembled genomes at the chromosome levels, we inferred ancestral chromosomes and reconstructed evolutionary trajectories, thereby bridging gaps in our understanding of Solanaceae chromosome karyotype evolution.

SoIR represents the first comprehensive pan-genomic resource for the Solanaceae family, providing systematic and in-depth analyses across all sequenced species. Additionally, it supports detailed gene analysis, including methylation, CRISPR sequences and various key functional genes from 81 Solanaceae species, with results stored within the database. These resources are critical for advancing plant research and functional genomics (62,63).

The SoIR integrates a wide array of genomic resources, such as Blast, Synteny, Primer Design and Hmmsearch tools, facilitating extensive comparative analyses. With enhanced JavaScript functionality, users can interact with dynamic SVG visualizations, exploring species trees and gene retention data effectively. Advanced mapping technologies, such as D3.js, are employed to optimize information-rich visualizations such as dot maps, synteny diagrams and collinear circle maps, enabling users to examine gene pairs, gene lists and detailed functional information interactively.

In conclusion, the SoIR database offers a robust platform for mining genomic and transcriptomic data from 81 Solanaceae species, with transcriptome data available for 41 species. It provides distinct advantages over existing databases and addresses many of their limitations. Furthermore, we plan to implement routine system updates every 6 months, focusing on expanding the genomic datasets and refining analytical tools to enhance both user experience and database functionality.

Data availability

All materials and data related to this study are available in the SoIR database (https://soir.bio2db.com) and in the supplementary files. The synteny display tool is accessible on GitHub (https://github.com/songxm-ncst/SoIR-Synteny_HTML) and Zenodo (https://doi.org/10.6084/m9.figshare.27231963.v1), which can be accessed at https://zenodo.org/records/13942996.

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

Author contributions: X.S. conceived the project and was responsible for the initiation of the project. X.S., Q.-Y.Y. and Z.L. supervised and managed the project and research. Data generation and collection were performed by X.S., Z.L., S.S. and C.L. Bioinformatics analysis was led by X.S., Z.L., S.S., C.L., C.Z., Y.F., X.C. and T.Y. The manuscript was organized, written and revised by X.S., Q.-Y.Y., Z.L., S.S., D.L. and R.Z. All the authors read and revised the manuscript.

Funding

Natural Science Fund for Distinguished Young Scholars of Hebei Province [C2022209010]; National Key Research and Development Program of China [2023YFF1002000]; National Natural Science Foundation of China [32172583]; S&T Program of Hebei [23372505D]; Hebei Natural Science Foundation [H2023209084].

Conflict of interest statement

None declared.

References

1. Tang,D., Jia,Y., Zhang,J., Li,H., Cheng,L., Wang,P., Bao,Z., Liu,Z., Feng,S., Zhu,X., *et al.* (2022) Genome evolution and diversity of wild and cultivated potatoes. *Nature*, **606**, 535–541.
2. Yang,X., Zhang,L., Guo,X., Xu,J., Zhang,K., Yang,Y., Yang,Y., Jian,Y., Dong,D., Huang,S., *et al.* (2023) The gap-free potato genome assembly reveals large tandem gene clusters of agronomical importance in highly repeated genomic regions. *Mol. Plant*, **16**, 314–317.
3. Xu,X., Pan,S., Cheng,S., Zhang,B., Mu,D., Ni,P., Zhang,G., Yang,S., Li,R., Wang,J., *et al.* (2011) Genome sequence and analysis of the tuber crop potato. *Nature*, **475**, 189–195.
4. Sato,S., Tabata,S., Hirakawa,H., Asamizu,E., Shirasawa,K., Isobe,S., Kaneko,T., Nakamura,Y., Shibata,D., Aoki,K., *et al.* (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, **485**, 635–641.
5. Qin,C., Yu,C.S., Shen,Y.O., Fang,X.D., Chen,L., Min,J.M., Cheng,J.W., Zhao,S.C., Xu,M., Luo,Y., *et al.* (2014) Whole-genome sequencing of cultivated and wild peppers provides insights into *Capsicum* domestication and specialization. *Proc. Natl Acad. Sci. U.S.A.*, **111**, 5135–5140.
6. Liu,F., Zhao,J., Sun,H., Xiong,C., Sun,X., Wang,X., Wang,Z., Jarret,R., Wang,J., Tang,B., *et al.* (2023) Genomes of cultivated and wild *Capsicum* species provide insights into pepper domestication and population differentiation. *Nat. Commun.*, **14**, 5487.
7. Chen,W., Wang,X., Sun,J., Wang,X., Zhu,Z., Ayhan,D.H., Yi,S., Yan,M., Zhang,L., Meng,T., *et al.* (2024) Two telomere-to-telomere gapless genomes reveal insights into *Capsicum* evolution and capsaicinoid biosynthesis. *Nat. Commun.*, **15**, 4295.
8. Su,X., Wang,B., Geng,X., Du,Y., Yang,Q., Liang,B., Meng,G., Gao,Q., Yang,W., Zhu,Y., *et al.* (2021) A high-continuity and annotated tomato reference genome. *BMC Genomics*, **22**, 898.
9. Aversano,R., Contaldi,F., Ercolano,M.R., Grosso,V., Iorizzo,M., Tatino,F., Xumerle,L., Dal Molin,A., Avanzato,C., Ferrarini,A., *et al.* (2015) The *Solanum commersonii* genome sequence provides insights into adaptation to stress conditions and genome evolution of wild potato relatives. *Plant Cell*, **27**, 954–968.
10. Wei,Q., Wang,J., Wang,W., Hu,T., Hu,H. and Bao,C. (2020) A high-quality chromosome-level genome assembly reveals genetics for important traits in eggplant. *Hortic. Res.*, **7**, 153.
11. Fernandez-Pozo,N., Menda,N., Edwards,J.D., Saha,S., Tecle,I.Y., Strickler,S.R., Bombarely,A., Fisher-York,T., Pujar,A., Foerster,H., *et al.* (2015) The Sol Genomics Network (SGN)—from genotype to phenotype to breeding. *Nucleic Acids Res.*, **43**, D1036–D1041.
12. Fei,Z., Joung,J.G., Tang,X., Zheng,Y., Huang,M., Lee,J.M., McQuinn,R., Tieman,D.M., Alba,R., Klee,H.J., *et al.* (2011) Tomato Functional Genomics Database: a comprehensive resource and analysis package for tomato functional genomics. *Nucleic Acids Res.*, **39**, D1156–D1163.
13. Hirakawa,H., Shirasawa,K., Miyatake,K., Nunome,T., Negoro,S., Ohyama,A., Yamaguchi,H., Sato,S., Isobe,S., Tabata,S., *et al.* (2014) Draft genome sequence of eggplant (*Solanum melongena* L.): the representative solanum species indigenous to the old world. *DNA Res.*, **21**, 649–660.
14. Hirsch,C.D., Hamilton,J.P., Childs,K.L., Cepela,J., Crisovan,E., Vaillancourt,B., Hirsch,C.N., Habermann,M., Neal,B. and Buell,C.R. (2014) Spud DB: a resource for mining sequences, genotypes, and phenotypes to accelerate potato breeding. *Plant Genome*, **7**, https://doi.org/10.3835/plantgenome2013.12.0042.
15. Yu,T., Ma,X., Liu,Z., Feng,X., Wang,Z., Ren,J., Cao,R., Zhang,Y., Nie,F. and Song,X. (2022) TVIR: a comprehensive vegetable information resource database for comparative and functional genomic studies. *Hortic. Res.*, **9**, uhac213.
16. Mistry,J., Chuguransky,S., Williams,L., Qureshi,M., Salazar,G.A., Sonnhammer,E.L.L., Tosatto,S.C.E., Paladin,L., Raj,S., Richardson,L.J., *et al.* (2021) Pfam: the protein families database in 2021. *Nucleic Acids Res.*, **49**, D412–D419.
17. Bateman,A., Martin,M.J., Orchard,S., Magrane,M., Ahmad,S., Alpi,E., Bowler-Barnett,E.H., Britto,R., Cukura,A., Denny,P., *et al.* (2023) UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.*, **51**, D523–D531.
18. Carbon,S., Douglass,E., Good,B.M., Unni,D.R. and Elser,J. (2021) The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.*, **49**, D325–D334.
19. Aach,J., Mali,P. and Church,G. (2014) CasFinder: flexible algorithm for identifying specific Cas9 targets in genomes through model selection and model averaging. bioRxiv doi: https://doi.org/10.1101/005074, 12 May 2014, preprint: not peer reviewed.
20. Emms,D.M. and Kelly,S. (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.*, **20**, 238.
21. Price,M.N., Dehal,P.S. and Arkin,A.P. (2010) FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, 10.
22. Wang,Y., Tang,H., Debarry,J.D., Tan,X., Li,J., Wang,X., Lee,T.H., Jin,H., Marler,B., Guo,H., *et al.* (2012) MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.*, **40**, e49.
23. Liu,Z., Fu,Y., Wang,H., Zhang,Y., Han,J., Wang,Y., Shen,S., Li,C., Jiang,M., Yang,X., *et al.* (2023) The high-quality sequencing of the *Brassica rapa* 'XiangQingCai' genome and exploration of genome evolution and genes related to volatile aroma. *Hortic. Res.*, **10**, uhad187.
24. De Bie,T., Cristianini,N., Demuth,J.P. and Hahn,M.W. (2006) CAFE: a computational tool for the study of gene family evolution. *Bioinformatics*, **22**, 1269–1271.
25. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
26. Suyama,M., Torrents,D. and Bork,P. (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.*, **34**, W609–W612.
27. Yang,Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.*, **24**, 1586–1591.
28. Sun,P., Jiao,B., Yang,Y., Shan,L., Li,T., Li,X., Xi,Z., Wang,X. and Liu,J. (2022) WGDI: a user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes. *Mol. Plant*, **15**, 1841–1851.
29. Stamatakis,A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.
30. Sanderson,M.J. (2003) r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics*, **19**, 301–302.
31. Kumar,S., Stecher,G., Suleski,M. and Hedges,S.B. (2017) TimeTree: a resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol.*, **34**, 1812–1819.
32. Song,X., Sun,P., Yuan,J., Gong,K., Li,N., Meng,F., Zhang,Z., Li,X., Hu,J., Wang,J., *et al.* (2021) The celery genome sequence reveals sequential paleo-polyploidizations, karyotype evolution and resistance gene reduction in Apiales. *Plant Biotechnol. J.*, **19**, 731–744.
33. Shen,S., Li,N., Wang,Y., Zhou,R., Sun,P., Lin,H., Chen,W., Yu,T., Liu,Z., Wang,Z., *et al.* (2022) High-quality ice plant reference genome analysis provides insights into genome evolution and allows exploration of genes involved in the transition from C3 to CAM pathways. *Plant Biotechnol. J.*, **20**, 2107–2122.
34. Zhang,Y., Zhang,Y., Li,B., Tan,X., Zhu,C., Wu,T., Feng,S., Yang,Q., Shen,S., Yu,T., *et al.* (2022) Polyploidy events shaped the expansion of transcription factors in Cucurbitaceae and exploitation of genes for tendril development. *Hortic. Plant J.*, **8**, 562–574.
35. Jin,J., Tian,F., Yang,D.C., Meng,Y.Q., Kong,L., Luo,J. and Gao,G. (2017) PlantTFDB 4.0: toward a central hub for transcription

factors and regulatory interactions in plants. *Nucleic Acids Res.*, **45**, D1040–D1045.
36. Feng,S., Liu,Z., Chen,H., Li,N., Yu,T., Zhou,R., Nie,F., Guo,D., Ma,X. and Song,X. (2024) PHGD: an integrative and user-friendly database for plant hormone-related genes. *iMeta*, **3**, e164.
37. Chen,H., Wang,T., He,X., Cai,X., Lin,R., Liang,J., Wu,J., King,G. and Wang,X. (2022) BRAD V3.0: an upgraded Brassicaceae database. *Nucleic Acids Res.*, **50**, D1432–D1441.
38. Wu,T., Liu,Z., Yu,T., Zhou,R., Yang,Q., Cao,R., Nie,F., Ma,X., Bai,Y. and Song,X. (2024) Flowering genes identification, network analysis, and database construction for 837 plants. *Hortic. Res.*, **11**, uhae013.
39. Feng,S., Li,N., Chen,H., Liu,Z., Li,C., Zhou,R., Zhang,Y., Cao,R., Ma,X. and Song,X. (2024) Large-scale analysis of the ARF and Aux/IAA gene families in 406 horticultural and other plants. *Mol. Hortic.*, **4**, 13.
40. Li,P., Quan,X., Jia,G., Xiao,J., Cloutier,S. and You,F.M. (2016) RGAugury: a pipeline for genome-wide prediction of resistance gene analogs (RGAs) in plants. *BMC Genomics*, **17**, 852.
41. Yue,H., Nie,X., Yan,Z. and Weining,S. (2019) N^6-Methyladenosine regulatory machinery in plants: composition, function and evolution. *Plant Biotechnol. J.*, **17**, 1194–1208.
42. Liu,Z., Li,N., Yu,T., Wang,Z., Wang,J., Ren,J., He,J., Huang,Y., Shi,K., Yang,Q., *et al.* (2022) The Brassicaceae Genome Resource (TBGR): a comprehensive genome platform for Brassicaceae plants. *Plant Physiol.*, **190**, 226–237.
43. Kang,M., Wu,H., Liu,H., Liu,W., Zhu,M., Han,Y., Liu,W., Chen,C., Song,Y., Tan,L., *et al.* (2023) The pan-genome and local adaptation of *Arabidopsis thaliana*. *Nat. Commun.*, **14**, 6259.
44. Wang,T., Duan,S., Xu,C., Wang,Y., Zhang,X., Xu,X., Chen,L., Han,Z. and Wu,T. (2023) Pan-genome analysis of 13 *Malus* accessions reveals structural and sequence variations associated with fruit traits. *Nat. Commun.*, **14**, 7377.
45. Marçais,G., Delcher,A.L., Phillippy,A.M., Coston,R., Salzberg,S.L. and Zimin,A. (2018) MUMmer4: a fast and versatile genome alignment system. *PLoS Comput. Biol.*, **14**, e1005944.
46. Chakraborty,M., Emerson,J.J., Macdonald,S.J. and Long,A.D. (2019) Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. *Nat. Commun.*, **10**, 4872.
47. Li,H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–3100.
48. Goel,M., Sun,H., Jiao,W.B. and Schneeberger,K. (2019) SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.*, **20**, 277.
49. Chen,S., Zhou,Y., Chen,Y. and Gu,J. (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, **34**, i884–i890.
50. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
51. Kim,D., Paggi,J.M., Park,C., Bennett,C. and Salzberg,S.L. (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.*, **37**, 907–915.
52. Liao,Y., Smyth,G.K. and Shi,W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.
53. Li,N., Wang,Y., Zheng,R. and Song,X. (2022) Research progress on biological functions of lncRNAs in major vegetable crops. *Veg. Res.*, **2**, 14.
54. Meng,F., Tang,Q., Chu,T., Li,X., Lin,Y., Song,X. and Chen,W. (2022) TCMPG: an integrative database for traditional Chinese medicine plant genomes. *Hortic. Res.*, **9**, uhac060.
55. Yu,T., Bai,Y., Liu,Z., Wang,Z., Yang,Q., Wu,T., Feng,S., Zhang,Y., Shen,S., Li,Q., *et al.* (2022) Large-scale analyses of heat shock transcription factors and database construction based on whole-genome genes in horticultural and representative plants. *Hortic. Res.*, **9**, uhac035.
56. Yang,P., Yuan,Y., Yan,C., Jia,Y., You,Q., Da,L., Lou,A., Lv,B., Zhang,Z. and Liu,Y. (2024) AlliumDB: a central portal for comparative and functional genomics in *Allium*. *Hortic. Res.*, **11**, uhad285.
57. Wang,Y.J., Tain,T., Yu,J.Y., Li,J., Xu,B., Chen,J., D'Auria,J.C., Huang,J.P. and Huang,S.X. (2023) Genomic and structural basis for evolution of tropane alkaloid biosynthesis. *Proc. Natl Acad. Sci. U.S.A.*, **120**, e2302448120.
58. Zhang,F., Qiu,F., Zeng,J., Xu,Z., Tang,Y., Zhao,T., Gou,Y., Su,F., Wang,S., Sun,X., *et al.* (2023) Revealing evolution of tropane alkaloid biosynthesis by analyzing two genomes in the Solanaceae family. *Nat. Commun.*, **14**, 1446.
59. Murat,F., Armero,A., Pont,C., Klopp,C. and Salse,J. (2017) Reconstructing the genome of the most recent common ancestor of flowering plants. *Nat. Genet.*, **49**, 490–496.
60. Wang,J., Yuan,M., Feng,Y., Zhang,Y., Bao,S., Hao,Y., Ding,Y., Gao,X., Yu,Z., Xu,Q., *et al.* (2022) A common whole-genome paleotetraploidization in Cucurbitales. *Plant Physiol.*, **190**, 2430–2448.
61. Kong,X., Zhang,Y., Wang,Z., Bao,S., Feng,Y., Wang,J., Yu,Z., Long,F., Xiao,Z., Hao,Y., *et al.* (2023) Two-step model of paleohexaploidy, ancestral genome reshuffling and plasticity of heat shock response in Asteraceae. *Hortic. Res.*, **10**, uhad073.
62. Vu,T.V., Nguyen,N.T., Kim,J., Das,S., Lee,J. and Kim,J.Y. (2022) The obstacles and potential solution clues of prime editing applications in tomato. *Biodes. Res.*, **2022**, 0001.
63. Cao,H., Wu,T., Shi,L., Yang,L. and Zhang,C. (2023) Alternative splicing control of light and temperature stress responses and its prospects in vegetable crops. *Veg. Res.*, **3**, 17.

Received: August 9, 2024. **Revised:** September 27, 2024. **Editorial Decision:** October 12, 2024. **Accepted:** October 19, 2024