https://doi.org/10.1186/s12870-025-06397-x

**RESEARCH** **Open Access**

# Comparative genomics profiling of *Citrus* species reveals the diversity and disease responsiveness of the GLP pangenes family

Muhammad Tahir ul Qamar[1,2†], Kinza Fatima[3†], Muhammad Junaid Rao[4†], Qian Tang[1†], Muhammad Sadaqat[5], Baopeng Ding[6,7], Ling-Ling Chen[1,2] and Xi-Tong Zhu[1,2*]

## Abstract

*Citrus* is an important nutritional fruit globally; however, its yield is affected by various stresses. This study presents the draft pangenome of *Citrus,* developed using 11 species to examine their genetic diversity and identify members of the germin-like proteins (GLPs) gene family involved in disease responsiveness. The developed sequence-based pangenome contains 954 Mb sequence and 74,755 genes. The comparative genomics analysis revealed the presence-absence variations (PAVs) among the *Citrus* genomes and species-specific protein-coding genes. Gene-based pangenome analysis revealed 4,936 new genes missing in the reference genome and highlighted the core and shell genes with putative functions in stress regulation. The pangenome-wide identification of GLP gene family members indicated the intraspecies diversity among the members across 11 genomes by analyzing their gene structure, motifs, and chromosomal distribution patterns. The synteny and evolutionary constraints analyses of Citrus GLPs provide detailed evidence of their evolutionary conservation and divergence. Further, the interaction, functional enrichment, and promoter analysis revealed their involvement in abiotic-, biotic-stress, signaling, and development-related pathways. The expression patterns of *C. sinensis* GLPs were studied in Huanglongbing (HLB) and Citrus canker disease. Several genes including *CsGLPs1-2* and *CsGLPs8-4* showed changes in expression patterns under both disease conditions. The qRT-PCR analysis revealed that these two genes were highly expressed in leaves infected with HLB disease across seven HLB-tolerant and susceptible citrus species. This *Citrus* pangenome and pangenes family study offers a comprehensive resource and new insights into the structural and functional diversity, identifying candidate genes that are important for future research to understand the stress-responsive mechanisms in *Citrus*.

**Keywords** *Citrus*, Citrus canker, Expression profiles, Genetic diversity, Huanglongbing, Pangenome, Pangenome-wide analysis, Pangenes, Presence-absence variations (PAVs)

[†]Muhammad Tahir ul Qamar, Kinza Fatima, Muhammad Junaid Rao and Qian Tang contributed equally to this work.

*Correspondence:
Xi-Tong Zhu
xtzhu@gxu.edu.cn
Full list of author information is available at the end of the article

## Introduction

*Citrus Spp.*, belonging to the *Rutaceae* family and comprising a diverse group of flowering plants, have long been regarded as vital crops worldwide due to their economic significance and nutritional value. Citrus fruit crops are cultivated for their succulent fruits and juices, characterized by their aroma and taste. They are not only a valuable source of fresh fruits and juices but also serve as essential components of the global food industry, contributing to employment and trade in many regions [1]. However, despite their economic importance, citrus plants face many challenges, including diseases caused by pathogens such as bacteria, fungi, and viruses. Huanglongbing (HLB), a disease commonly known as citrus greening disease, is caused by a bacterium *Candidatus* Liberibacter asiaticus (*Ca*Las). The symptoms of this disease include smaller and green with misshaped fruits, asymmetrical and upright yellow leaves, brownish-black smaller aborted seeds, thin canopies, distorted foliage, and tree decline. This leads to overall decline in fruit growth and juice value as once the plant is infected, it ultimately leads to its death [2, 3]. Another disease, called citrus canker is caused by bacterium *Xanthomonas citri* pv *citri* (*Xcc*). This causes necrotic lesions on fruit, leaves, stems, and twigs. The infected fruits and trees develop yellow and watery borders leading to defoliation and death of the tree over time [4]. These diseases pose a significant threat to citrus production and necessitate a better understanding of the mechanisms underlying disease resistance in these plants [5]. Earlier studies have shown that the bulk of cultivated citrus is mainly contributed by three species belonging to the *Citrus* genus, *Citrus medica* (citron), *C. reticulata* (mandarin), and *C. maxima* (pomelo). Later, another species *C. ichangensis* (papeda) was also proposed as the fourth foundational species [6].

The genetic basis of disease resistance in plants has been a subject of intense research in recent years. Various families of genes have been identified and characterized, each playing a specific role in the intricate network of plant defense mechanisms. One such gene family, the GLP family, has garnered considerable attention due to its involvement in plant defense responses against various pathogens. The *GLPs* constitute diverse and functionally important gene family members in various plant species. *GLPs* have been implicated in numerous physiological processes, such as plant defense responses, oxidative stress tolerance, cell wall modification, ROS scavenging, and signal transduction pathways [5, 7]. GLP gene family has already been identified and studied for their roles in many plants including *Arabidopsis thaliana* [8], *Oryza sativa* [8], *Zea mays* [5], *Arachis hypogaea* [9], *Triticum aestivum* [10], *Solanum tuberosum* [11], *Cucumis sativus* [12], *Cucumis melo* [13], and *Camellia sinensis* [14].

Pangenome studies in *Setaria italica* [15], *Sorghum bicolor* [16], *Brassica oleracea* [17], Soybeans [18], potato [19], and tomato [20] indicated that structural variations have crucial roles in genetic improvement. Previous studies have reported insights into *Citrus* diversity and gene families involved in abiotic and biotic stresses [6, 21]. Pangenomes are used to study the entire genetic content as well as genomic variations in a species under study [22]. Pangenome graphs provide a compact representation of the entire genome, capturing their sequence similarities, and all types of variations among them [23]. Sequence-based pangenome captures the sequences at the genic as well as non-genic level. Whereas, the gene-based pangenome captures the entire set of genes as well as orthologous gene families among the members of the same species [24]. Pangene is a gene model or allele found in a similar genomic location in multiple members of the same species or closely related species. A pangene set is a genus- or species-level equivalent to several terms that are commonly used to describe genes that correspond at a deeper taxonomic level: "gene family", "orthogroup", or "ortholog set". The common use of pangene is to capture presence-absence variation (PAV) at the gene level [25, 26].

In this study, we performed pangenome analysis of 11 *Citrus* species including *Atalantia buxifolia* (Chinese box orange) [27], *C. ichangensis* (Ichang papeda) [28], *C. reticulata* (Mandarin) [29], *Fortunella hindsii* (Kumquat) [30], *C. sinensis* (Sweet orange) [31], *C. grandis* or *C. maxima* (Pummelo) [28], *C. medica* (Citron) [28], *C. clementina* (clementines) [32], *C. unshiu* Marc (satsumas) [33], *Murraya paniculata* (Orange jasmine) [34], and *Poncirus trifoliata* (Trifoliate orange) [35]. We report a full spectrum of genetic variations and diversity across 11 *Citrus* species. By incorporating the functional characterization of gene and gene families, we provide a comprehensive resource that could serve as a foundation for future studies on genetic improvement against stresses. Further, we also incorporated the identification studies of GLP gene family, reporting members across the 11 *Citrus* species. The comprehensive structural and functional evaluation of these identified GLP members indicated their potential roles in plant development and resilience against disease stress. The gene expression profiles unveiled the genes responsive to Huanglongbing (HLB) and citrus canker disease. Overall, this study provides comprehensive insights into the variations among *Citrus* genomes at both the sequence and gene family levels and further uncovers their roles in conferring resistance against environmental stresses.

## Methods

### Sequence-based pangenome analysis

The *Citrus* pangenome was constructed using publicly available 11 genomes (Table S1). Based on the genome size, GC ratio, and nucleotide identity, these genomes were selected and proceeded for the construction of a sequence-based pangenome. Initially, the high-quality reference genome *C. grandis* and the other genomes were analyzed using the Nucmer program from the MUMmer v4.0.0beta2 [36]. *C. grandis* was selected as the reference genome and other query genomes were compared iteratively based on their evolutionary order [37] (*M.paniculata* > *A. buxifolia* > *P.trifoliata* > *C. ichangensis* > *C. reticulata* > *F.hindsii* > *C. sinensis* > *C. medica* > *C. clementina* > *C. unshiu*) [37]. The ppsPCP [38] was run using default parameters (–coverage 0.9, –sim_pav 0.95, –sim_gene 0.8) and multiple threads on a high computing facility of Huazhong Agricultural University, Wuhan 430070, China to construct the sequence-based pangenome. An in-house Perl script was used to retrieve PAVs-flanking sequences (100 bp) from the feature coordinates in the GFF3 files of reference and query genomes, which were further used to determine the position of PAVs on the reference genome chromosomes using BLASTn [39]. Perl script was also used to filter PAVs accuracy. Based on the flanking sequence's locations on query and reference genomes, the PAVs were categorized into high-confidence, average-confidence, and low-confidence PAVs. Low confidence PAVs were split into two cases (Low case 1 and Low case 2) based on mapping situations (Figure S1). PAVs were screened and added to develop a draft pangenome. Overlapped genes with PAVs were also harvested and an annotation file was prepared. After categorization, PAVs were placed on the chromosomes of the reference *C. grandis* genome. To assess the quality of the developed *Citrus* pangenome, all protein-coding genes were extracted from all 10 *Citrus* genomes and used as query sequences to be mapped against the reference genome (*C. grandis*) and pangenome respectively using BLAT [40] with 80% similarity. Repbase [41], RepeatModeler [42], and RepeatMasker [43] were used to identify and annotate Transposable elements (TEs). In-house Perl scripts were used to investigate the densities of genes, GC contents, and TEs across the pangenome which were further used to draw the circle graphs by using Circos software [44]. All the densities were counted using a 100-kb window. To evaluate the quality of the developed *Citrus* pangenome and its suitability to be used as a reference to assemble *Citrus* species sequenced by NGS approaches, *C. reticulata* paired-end sequencing data was mapped to the developed pangenome using BWA v0.7.17 [45] with default parameters.

### Genes and genes-family-based pangenome analysis of *Citrus*

A pangenome based on annotated protein-coding genes was also constructed. All assemblies used were annotated by the ab initio gene prediction method [46] except *C. unshiu* Marc which was annotated by using MAKER-P [47]. Related genes across 11 *Citrus* assemblies were clustered by grouping gene models using GET_HOMOLOGUES-EST [48] with a minimum alignment coverage of 90%. This algorithm takes CDS sequences as input and generates BLASTN [39] hits to drive Markov clustering of CDS sequences. The resulting pangenes clusters were reduced to non-redundant gene clusters by collapsing clusters with ≥ 95% identity and ≥ 90% alignment coverage. These pangenes clusters were validated by performing silhouette analysis with k = 10. These clusters were further taken to compute average nucleotide identity (ANI) matrices [49] to summarize the overall similarity among clusters.

Subsequently, a single gene from each cluster was selected as the pangene representative and further analyzed using GET_HOMOLOGUES-EST. The resulting clusters were classified into core (gene clusters which contained all 11 species), soft-core (gene clusters contained 10–11 species), shell (gene clusters contained 2–9 species), and cloud (gene clusters contained only 1 variety) subsets. Shell genes were further divided into reference genes and non-reference genes. Gene ontology analyses were performed for non-reference genes using eggNOG-mapper v2 [50] and AgriGO v2.0 [51]. Curves describing the pangenome and the core genome size were fitted in R using the nls (nonlinear least squares) function from package stats. Points used in regression corresponded to all the possible combinations of genomes. The combinations of genomes were obtained according to the following formula: 10!/(n!(10—n)!), n = [1,10], and the pangenome size was modelled using the power law regression y = AxB + C. The core genome size was modelled using exponential regression y = AeBx + C. The model was fitted using means [51]. Protein sequences from 11 species were compared using all-by-all DIAMOND [52] followed by Orthofinder [53] to cluster them into orthologous gene families, which were then used to estimate the *Citrus* pangenome size. The gene families shared by all *Citrus* species constitute the core gene sets, while those shared by fewer than 11 *Citrus* species constitute the shell gene sets.

Orthologous gene families were used to construct a species tree. The families were first mapped with MAFFT v7.407 [54] and the phylogenetic tree was constructed using FastTree v2 [55]. The divergence time of 11 *Citrus* species was estimated by r8s v1.81 [56] and CAFE v2.2 [57] was used to detect gene family expansion and contraction.

### Functional characterization of the pangenome

All the pangenome genes were split into two groups corresponding to core and shell genes which were compared for gene length, exon number, coding sequences, upstream TEs, and nonsynonymous substitutions per nonsynonymous site (dN) to the number of synonymous substitutions per synonymous site (dS) ratio (dN/dS ratio). The groups were compared using Mann–Whitney *U*-test as implemented in R function Wilcox.test (two-tailed test).

For dN/dS analysis, the protein sequences were aligned with each gene cluster using Clustal Omega v1.2.4 [58] and the multiple sequence alignment was performed using PAL2NAL v14 [59]. The dN/dS ratio was calculated by SNAP v2.1.1 [60]. For TEs, Repbase [41], RepeatModeler [61], and RepeatMasker [43] were used to identify and annotate Transposable elements (TEs). Next, in-house Perl and R scripts were used to measure fraction coverage and upstream distance of TEs. The core and shell pangenes were functionally annotated using eggNOG-mapper v2 36 and eggNOG 5.0 database [62], AgriGO v2.0 [63], and Blast2GO [64]. In eggNOG-mapper v2, all the proteins of 11 species were first re-annotated by searching them in the eggNOG 5.0 database. After annotation, GO terms were retrieved for each gene, and run_GOseq.pl script of Trinity v2.6.6 [65] was used to perform GO enrichment analysis for core and shell genes. In AgriGO v2.0, *C. sinensis* was used as background reference.

The shell genes were searched for their enrichment network using ClueGO [66] with *A. thaliana* as a reference. All predicted networks were analyzed using CluePedia [67] and visualized using Cytoscape [68]. Next, GeneMANIA [69] was used to functionally group the promising shell genes network for improved biological interpretation.

### Pangenome-wide identification of *Citrus* GLPs

The 32 *A. thaliana* protein sequences were obtained from the Ensemble Plants database (https://plants.ensembl.org/index.html) [70] and used as queries to conduct the BLASTp search against 11 *Citrus* genomes. The candidate *Citrus* GLPs were searched for Cupin_1 domain (Pfam ID: PF00190) to identify GLP gene family sequences. The physiochemical properties were predicted by using the ExPASy ProtParam tool (https://web.expasy.org/protparam/) [71].

### *Phylogenetics, conserved motifs, and gene structure analysis of Citrus GLPs*

The identified GLP protein sequences from 11 *Citrus* and 13 other plant species (*A. thaliana* [8], *O. Sativa* [8], *A. hypogaea* [9], *A. duranensis* [9], *A. ipaensis* [9], *Mangifera indica*, *Malus domestica*, *Pyrus bretschneideri*, *Psidium Guajava*, *Prunus persica*, *Solanum lysopercicum*, *S. tubersom* [72], and *Z. Mays* [5]) (Table S9) were aligned using ClustalW [73]. A maximum likelihood tree with 1000 bootstraps was constructed using IQTREE Web Server [74] and visualized using the Interactive Tree of Life server, iTOL (https://itol.embl.de/) [75].

The conserved motifs in 11 *Citrus* GLPs were searched using the MEME tool (https://meme-suite.org/meme/) [76]. The gff3 files were used to construct gene structure using TBtools [77].

### *Chromosomal distribution, comparative syntenic, Ka/Ks rate, and gene duplication analysis*

The chromosomal locations of each *Citrus GLPs* were retrieved from the gff/gff3 files and mapped to chromosomes by using the gene location visualization tool of the TBtools software. The synteny relationships of orthologous GLP genes were explored among three chromosomal-level Citrus species including *C. sinensis*, *C. grandis*, and *P. trifoliata*. The MCScanX tool [78] was employed with an e-value $< 1.0e^{-5}$ to analyze and visualize synteny among these genomes.

The duplicated genes were analyzed at 70% criteria. DnaSP v.6 software [79] was used to predict the Ka/Ks values of duplicated genes. the divergence time for the duplicated gene pairs was also calculated by using the formula $t = Ks/2\lambda \times 10^{-6}$, where the λ value for dicots, $1.5 \times 10^{-8}$, calculates the time of duplication in million-year units [80].

### *miRNA, PPI, and GO enrichment analysis*

The coding sequences were used to identify the putative miRNAs targeting the *CsGLP* genes using the miRNAsong database [81]. The PPIs among *C. sinensis* GLP proteins were predicted using STRING database [82] and visualized using Cytoscape [68]. The GO enrichment analysis was performed using the PANNZER database [83].

### *Cis-regulatory elements prediction and expression analysis of Citrus GLPs*

The 2-kb upstream regions of *C. sinensis GLPs* were used to search for *cis*-regulatory elements using PlantCARE online tool [84].

The expression pattern of *C. sinensis GLPs* was analyzed in two disease conditions: before and post Huanglongbing (HLB) disease-infected leaves of *C. sinensis* plant (BioProject: PRJNA797721) and leaves infected with citrus canker disease (BioProject: PRJNA836261). After evaluating the quality FastQC [85], the high-quality paired-clean reads were mapped to the indexed genome using HISAT [86]. The raw counts for each gene family member were quantified using StringTie [87]. and the heatmaps were generated using the fragments per kilobase of transcript per million mapped reads (FPKM) values.

### RNA extraction and quantitative Real-Time PCR (qRT-PCR) analysis

The seeds of seven citrus species (*M.paniculata*, *A. buxifolia*, *C. ichangensis*, *C. medica*, *C. maxima*, *C. sinensis*, and *C. reticulata*) were obtained from the Citrus Genomics and Genetics Improvement Laboratory, College of Horticulture and Forestry Sciences (CHFS), Huazhong Agricultural University, Wuhan, China. These seven citrus species were grown under controlled environmental conditions in a growth chamber (having 60 ± 3% humidity, 27 ± 2 °C temperature, and 8000 LUX light intensity) with recommended fertilizer and water treatment. After six months, six plants from each citrus species were selected for HLB inoculation. The HLB pathogen *Candidatus* Liberibacter asiaticus (*C*Las) was introduced using *Diaphorina citri* (Asian citrus psyllid) vectors that had been reared on HLB-infected trees for over four weeks. Approximately 50 psyllids were transferred to each experimental plant following previously established protocols [88]. Three plants from each species were maintained as healthy controls without *C*Las exposure. One month following psyllid introduction, DNA was extracted from the leaf tissue of all experimental plants and analyzed for HLB infection utilizing a nested PCR method. Subsequently, three HLB-infected plants from each species were selected for quantitative PCR analysis. The nested PCR results using corresponding primer sequences which are detailed in the Supplementary nested PCR figure and primer file (Note S2, Figure S48). The HLB-infected and control leaves of seven species were collected for RNA extraction.

Total RNA was extracted using Zomanbio (Cat no. ZP401-2) total RNA-pure reagent (Lot#200F12F), and 1 µg (µg) of total RNA was used for the complementary DNA (cDNA) synthesis. cDNA was synthesized using the Zomanbio (M-MLV, ZR102-3) reverse transcriptase kit (Beijing, ZOMAN Biotechnology Co., Ltd.) according to the manufacturer's instructions. The cDNA concentration was measured using Qubit fluorometer (Invitrogen, China) and normalized to 100 ng/µL before qRT-PCR [88, 89]. For qRT-PCR, ChamQ universal master mix SYBR (Vazyme, Q711-02) and a LongGene (Model: q2000b) fluorescence quantitative PCR instrument (Langji Scientific instrument Co., Ltd.; Hangzhou, China) were used. The expression levels of candidate genes (*CsGLP1-2* and *CsGLP8-4*) were quantified using the $2^{-\Delta\Delta CT}$ method [90]. qRT-PCR was performed using gene-specific primers detailed in (Table S10). Actin gene from Citrus was used as an internal reference (control) (Table S10). Statistix 8.1 (Tallahassee Florida, United States) statistical software was used for analyzing all qRT-PCR data and the Excel program was used to generate graphs.

## Results

### Sequence-based pangenome of *Citrus*

A total of 11 *Citrus* genomes were used to construct the pangenome. A summary of each *Citrus* genome assembly is given in Table S1. Compared to other genomes, *C. grandis* has the highest quality genome with a genome size of 344.88 Mb, 30,123 genes, 45.85% TEs, and 31% GC content, thus, it was chosen as the reference genome. The GC content in selected *Citrus* verities varied from 30 to 35% and the mean identity observed among *Citrus* species was 95.00% making them good candidates to be used for pangenome construction (Figure S2).

At first, the pangenome was developed to get an initial estimate of *Citrus* pangenome size at the DNA sequence level. The genomes were compared, and PAVs were confirmed and categorized. Further, the filtered PAVs were compared to the query *C. grandis* genome, and their boundaries were corrected to extract full-length genes. After boundary correction, a total of 43,693; 52,748; 22,301; 14,233; 12,035; 11,628; 3,615; 19,791; 3,305, and 10,404 PAVs were screened from *M. paniculata*, *A. buxifolia*, *P. trifoliata*, *C. ichangensis*, *C. reticulata*, *F. hindsii*, *C. sinensis*, *C. medica*, *C. clementina,* and *C. unshiu* Marc genomes respectively and added into developing draft pangenome (Figure S3-4). The highest number of PAVs were contributed by *A. buxifolia* (52,748), while *C. clementina* contributed the lowest number of PAVs (33,05). The longest PAV sequence was extracted from *F. hindsii* which was 1,029,660 bp long. The average length of the PAV sequence added to the pangenome was 3,503 bp, much larger than the 100 bp minimum length cutoff (Table S2).

After developing a sequence-based draft pangenome, overlapped genes with PAVs were also harvested and an annotation file was prepared. In total 12,400; 9,775; 4,381; 2,907; 2,145; 4,592; 3,156; 3,209; 378 and 1,689 genes were contributed by *M. paniculata*, *A. buxifolia*, *P. trifoliata*, *C. ichangensis*, *C. reticulata*, *F. hindsii*, *C. sinensis*, *C. medica*, *C. clementina*, and *C. unshiu* Marc respectively

(Figure S5). This sequence-based *Citrus* pangenome was 954 Mb and contained 74,755 genes, containing 620 Mb more sequence and 44,632 more genes than the reference genome *C. grandis* (334 Mb genome / 30,123 genes) (Figure S6).

### *PAVs categorization and distribution on the reference genome*

The identified PAVs were categorized into high, average, low case 1, low case 2, N_flanked, and N_mapped PAVs, and their total numbers were estimated for every *Citrus* variety. *P. trifoliata* possesses the highest number of high-confidence category PAVs, while *M. paniculata* followed by *A. buxifolia* possesses the highest numbers of average confidence category PAVs (Figure S7).

After the classification, the distribution of PAVs on the reference *C. grandis* genome was analyzed. PAVs were found distributed all along the chromosomes of the reference genome. However, chr2 and chr5 showed some variation, where most of the PAVs were enriched on the arms of chromosomes. A few PAVs were also found to be distributed on chrUn, indicating that these unknown sequences may have some useful biological functions (Fig. 1B).

### *Comparative analysis between reference and the pangenome*

In case of mapping against the reference genome, the mapping rate was between 82 to 98%. The maximum number of genes (98%) from *C. clementina* and a minimum number of genes (82%) from *M. paniculata* were mapped to the reference genome (Table S3).

However, in case of pangenome, the mapping rate was found ~ 100%. Maximum number of genes (99.99%) from *C. clementina* and minimum genes (99.84%) from *A. buxifolia* were mapped to the pangenome. The mapping rate of every species was more than 99% (Table S4). These results show a better quality, mapping rate, and completeness of the developed *Citrus* pangenome (Figure S8). Thus, it can be used as an alternative way to develop more complete genome assemblies.

### *Characterization and quality testing of pangenome*

The distributions of genes, TEs, and GC contents in the pangenome were studied. Some regions (e.g., portion shared by chr3 and *C. clementina*) had very low densities of genes and TEs, whereas other regions (chr5, chr8, chr9, and *M. paniculata*) had high densities. These results were consistent with the previous results of PAVs and gene screening. However, GC% varied among different regions. The developed *Citrus* pangenome had 34% GC contents and 56% TEs in total (Fig. 1C).

NGS paired-end sequencing data of *C. reticulata* genome was mapped against the developed pangenome to affirm whether it can be used as a reference to assemble new genomes with reference-guided assembly approaches [91]. *C. reticulata* genome was assembled using *C. grandis* as a reference and the mapping percentage was 90%. However, while using the developed *Citrus* pangenome as a reference, a mapping rate of 99.89% was achieved which endorses the quality and usage of this pangenome as a reference.

## Genes and genes-family-based pangenome analysis of *Citrus*

### *Gene-based pangenome analysis*

The gene-based pangenome was developed using a single representative from each gene cluster that resulted in 59,261 pangenes clusters. These clusters were further categorized into core, soft-core, shell, and cloud clusters (Fig. 1D, Table S5). Further analysis focused on high-confidence pangenome consisting of 26,092 core and shell pangenes. The shell pangenes were analyzed to check the number of genes belonging to the reference genome and the new genes. A total of 4,936 new genes were found that were missing in the reference genome and these non-reference genes were termed accessory genes (Fig. 1E, S9).

To further investigate the biological functions of accessory genes, their ratios from each variety and their GO were examined. It was found that most of the accessory genes (56%) belonged to only 3 species (*M. paniculata, A. buxifolia,* and *C. ichangensis*). Most of the accessory genes were found to be involved in the metabolic processes including; organic substance metabolic process (GO:0071704), cellular metabolic process (GO:0044237), nitrogen compound metabolic process (GO:0006807), cellular macromolecule metabolic process (GO:0044260), cellular protein metabolic process (GO:0044267), and RNA metabolic process (GO:0016070) (Fig. 2E).

### *Pangenome modelling*

Stepwise addition of *Citrus* species in independent clustering runs showed an increase in the total number of PAVs and genes in both core and shell regions of the pangenome (Fig. 2A-B). The changes within the total gene set of the pangenome and the genes shared among all 11 genomes (core genes) were studied. The pangenes increased as additional genomes were added one by one. However, the core genes decreased with the addition of *Citrus* genomes (Fig. 2C-D). These results demonstrate

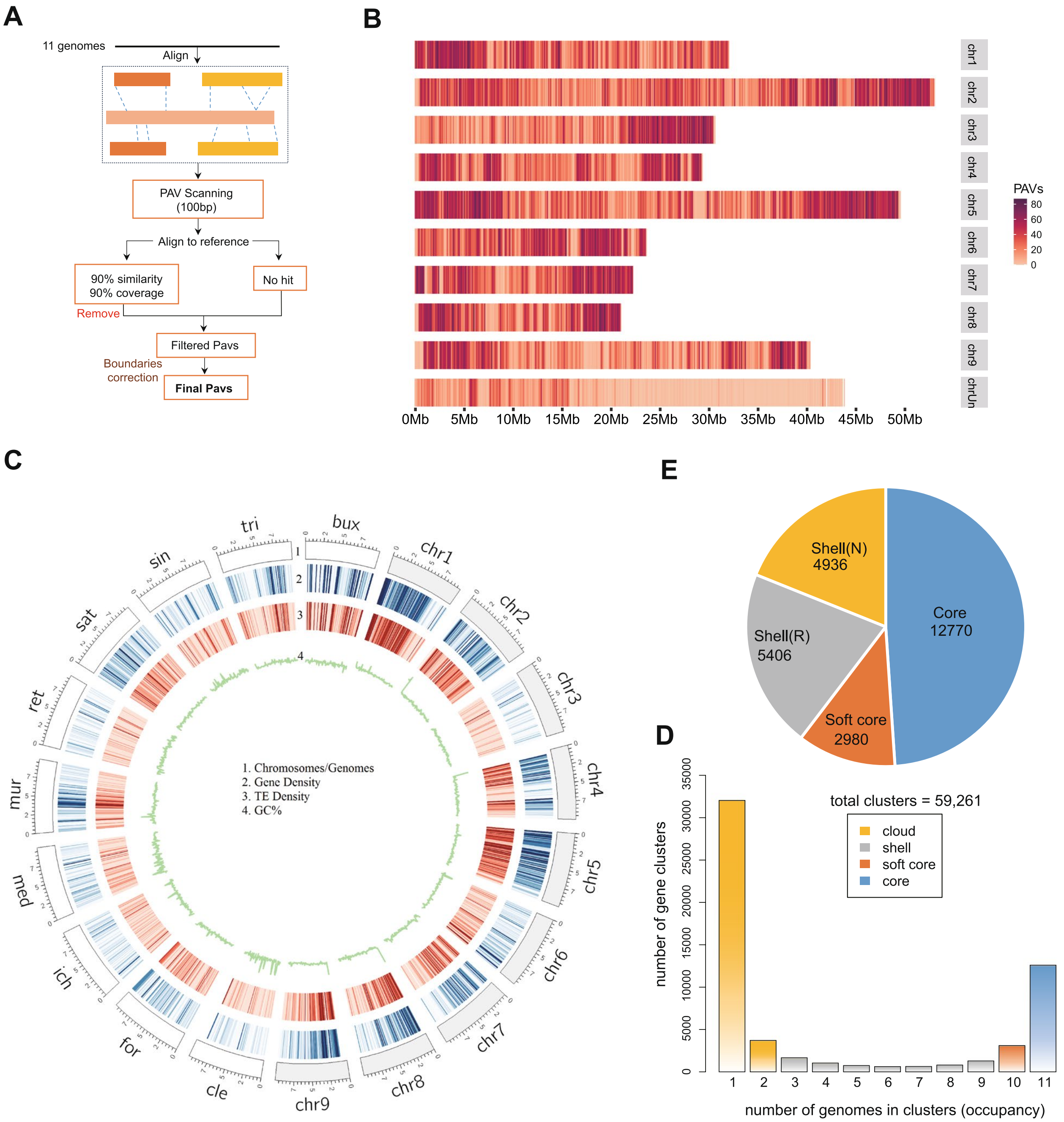

**Fig. 1** **A** *Citrus* pangenome construction. **B** The distribution of PAVs on reference genome *C. grandis*. **C** Characteristics of the *Citrus* pangenome. The distribution of the genes, TEs, and GC contents across the pangenome is shown in the circos plot. The calculations to show the distribution were based on a 100 Kb window. The "chr1-9" represents *C. grandis* (reference) chromosomes from 1 to 9, while "cle", "for", "ich", "med", "mur", "ret", "sat", "sin", "tri" and "bux" represent *C. clementina, F. hindsii, C. ichangensis, C. medica, M. paniculata, C. reticulata, C. unshiu, C. sinensis, P. trifoliata,* and *A. buxifolia*, respectively. **D** Gene-based comprehensive *Citrus* pangenome including cloud region. **E** Core and shell regions of gene-based pangenome. Distribution of pangenes into core, soft-core, shell reference, and shell non-reference categories. "R" represents the genes present in the reference genome, while "N" represents accessory genes completely missing in the reference genome

that the gene-based pangenome within the context of the current set of *Citrus* species, is closed. Further addition of sequences within this set of *Citrus* species will likely result in very negligible gene discovery. However, more genes can be discovered if the quality of these genomes improves further.

### *Orthologous clustering, phylogeny, and divergence time inference of Citrus pangene families*

Pangenome analysis based on orthologues gene families was performed to gain insights into pangene families. The results showed that 62.64% of gene families were species-specific and present only in one variety. This could be the possible reason for the presence of significant phenotypic variations among *Citrus* species. Among the 11 species, *C. unshiu* showed the highest number (11,322) of species-specific gene families. The remaining 35.75% orthologous gene families were conserved across all 11 *Citrus* genomes. However, only 0.02% of orthologous gene families were found to be shared between more than 1 variety (Fig. 3A).

The phylogenetic tree based on the pangenes families from 11 *Citrus* species showed that *M. paniculata* is the oldest variety and diverged about 25 million years ago compared to the other *Citrus* species. *A. buxifolia* and *P. trifoliata* were also found to be more divergent than the other species (Fig. 3B). These results are consistent with the initial sequence-based and gene-based pangenomes results, where it was identified that *M. paniculata* is the most outlier variety. The expansion and contraction analysis of pangene families showed that different species possessed different numbers of expanded or contracted gene families. It was found that *M. paniculata* has 818 expanded gene families but 9,723 contracted gene families, which is the largest number of contracted gene families among all 11 species which might be linked to phenotypic and biological variation between *M. paniculata* and other *Citrus* species.

## Functional characterization of the pangenome

### *Comparative analysis between core and shell regions*

Core and shell regions of the *Citrus* pangenome were compared to get insights into pangenes variations. Gene and CDS lengths of core genes were shorter than the shell genes. Similarly, core genes had fewer exons per gene than shell genes (Fig. 3C-E). Moreover, shell genes had higher dN/dS ratios compared to the core genes ($p < 2.2e\text{-}^{16-}$, Welch two-sample *t*-test) (Fig. 3F). This overall higher frequency of non-synonymous and synonymous substitutions of shell genes suggested reduced functional constraint and a high evolutionary rate of shell genes. TEs from each *Citrus* genome were annotated and analyzed to get insights into their potential roles in the evolution of shell genes. It was found that TEs were significantly closer and covered a bigger fraction of the upstream region of the shell genes compared to the core genes ($p < 2.2\text{E} - 16$, Wilcoxon signed-rank test). For core genes, more than 81% had no TEs coverage within 1 kb upstream regions, whereas, only 72% of shell genes had no TEs coverage within 1 kb upstream regions (Fig. 3G). Around 27% of shell genes possess the closest TEs within 1 kb upstream compared with only 21% of core genes (Fig. 3H).

### *GO enrichment and network analysis of core and shell genes*

To functionally characterize the core ad shell genes, their GO enrichment analysis was performed in terms of biological process (BP), cellular component (CC), and molecular function (MF).

The core genes were enriched for essential BPs including, cellular process (GO:0009987), cellular metabolic process (GO:0044237), biological regulation (GO:0065007), response to stimulus (GO:0048584), regulation of biological process (GO:0050789), and regulation of cellular process (GO:0050794) (Figure S10A). The shell genes were found to be enriched in BPs beneficial for defense such as response to stress (GO:0006950), defense response (GO:0006952), signaling (GO:0023052), and immune system process (GO:0002376) (Figure S10B).

In terms of CCs, the core pangenes were found enriched in the cell part (GO:0044464), intracellular (GO:0005622), intracellular part (GO:0044424), organelle (GO:0043226), cytoplasm (GO:0005737), and nucleus (GO:0005634) (Figure S11A). While, shell pangenes were found enriched in the membrane (GO:0016020), cell periphery (GO:0071944), plasma membrane (GO:0005886), and cell junction (GO:0030054) (Figure S11B).

In MFs category, the core pangenes were found enriched in protein binding (GO:0005515), transcription regulator activity (GO:0140110), DNA binding (GO:0003677), and ion binding (GO:0043167) (Figure S12A). While, shell pangenes were found enriched in RNA binding (GO:0003723), nuclease activity (GO:0004518), lyase activity (GO:0016829), and terpene synthase activity (GO:0010333) (Figure S12B).

Network analysis of shell genes revealed their enrichment defense and signal transduction pathways. The "defense response" network was further studied in detail. A total of 184 unique *Citrus* shell genes were found enriched in this network. This network contained

(See figure on next page.)

**Fig. 2** **A** Number of PAVs contributed by each variety to gradually developing clusters of pangenome. **B** The number of genes contributed by each variety to gradually developing clusters of pangenome. **C** Modelling of the pangenome by adding one genome to the pool at a time. **D** Modelling of the core genome by adding one genome to the pool at a time. **E** GO enrichment of accessory genes in biological processes. **F** The Presence absence variation (PAVs) among the *GLPs* identified among 11 *Citrus* genomes. 15 members were classified as core genes. 42 members were classified as shell genes. One member of *C. grandis* was classified as the shell or unique gene

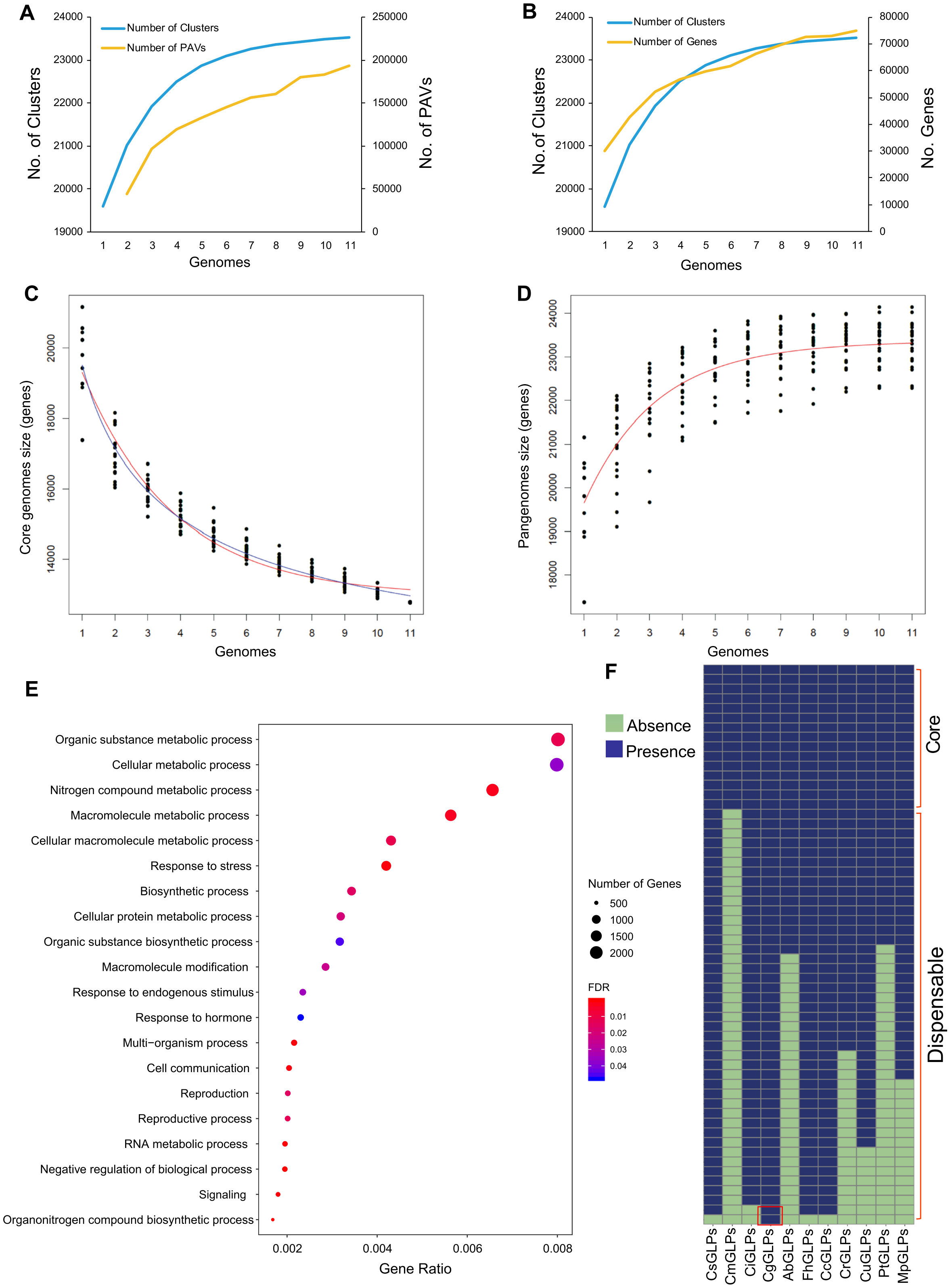

**Fig. 2** (See legend on previous page.)

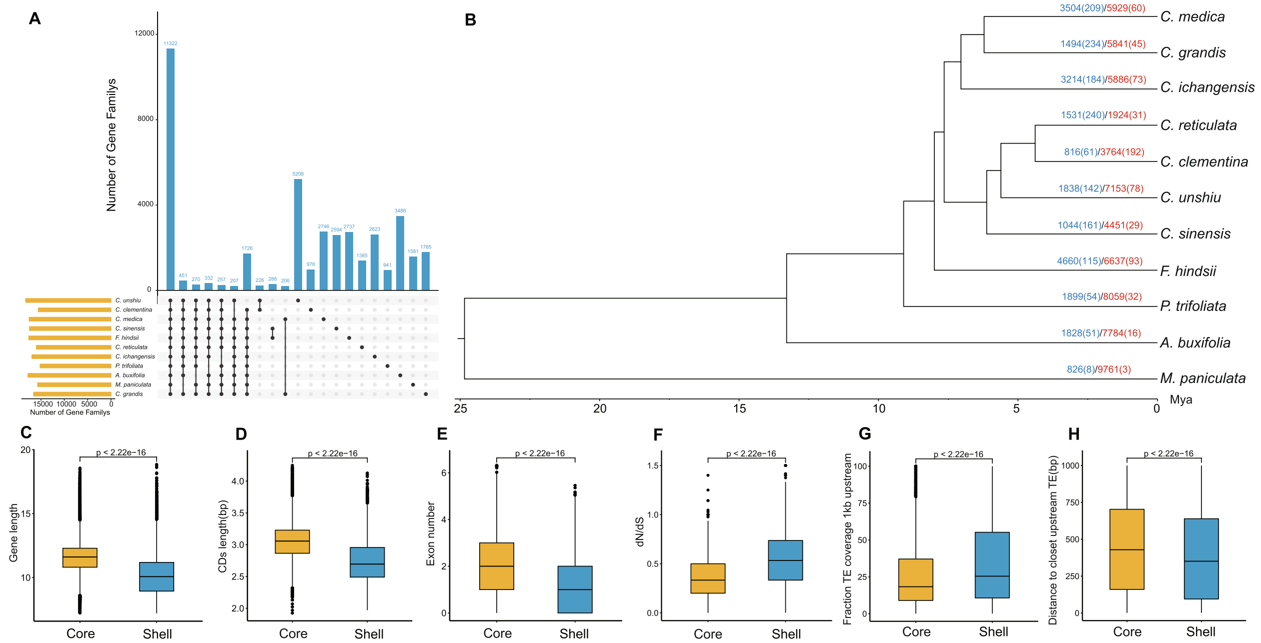

**Fig. 3** **A** The number of gene families shared by different combinations of *Citrus* species. **B** Phylogenetic tree with expansion and contraction information of gene families in different *Citrus* species. Mya: million years ago; Blue font: the number of expanded gene family; Blue font in brackets: rapidly expanded gene family; Red font: the number of contracted gene family; Red font in brackets: rapidly contracted gene family. Comparison of core and shell genes. **C** genes length, **D** CDS length, **E** exon number, **F** nonsynonymous/synonymous substitution ratio, **G** fraction of TE coverage of 1 Kb upstream, and (**H**) Distance to closest upstream TE

five sub-clusters: response to external biotic stimulus, response to bacterium, defense response to bacterium, response to other organisms, and defense response to other organisms (Figure S13).

Moreover, the gene cluster involved in "response to external biotic stimulus" was also analyzed. Total 14 genes; BTB/POZ domain-containing protein NPY2 isoform X2 (Phytozome ID: orange1.1g007080m), autophagy-related protein 8C (Phytozome ID: orange1.1g033424m), inhibitor of trypsin and hageman factor-like (Phytozome ID: orange1.1g033055m), ubiquitin-like protein (Phytozome ID: orange1.1g034423m), autophagy-related protein 8i (Phytozome ID: orange1.1g033214m), glu S.griseus protease inhibitor-like (Phytozome ID: orange1.1g035202m), phytochrome A (Phytozome ID: orange1.1g001235m), autophagy-related protein 8e (Phytozome ID: orange1.1g033384m), autophagy-related protein 8f (Phytozome ID: orange1.1g033114m), unknown (Phytozome ID: orange1.1g039274m), autophagy-related protein 8C (Phytozome ID: orange1.1g033381m), phytochrome E (Phytozome ID: orange1.1g001210), root phototropism protein 2 (Phytozome ID: morange1.1g007944m), and phytochrome C (Phytozome ID: orange1.1g001215m) were found enriched in this cluster. The co-expression network was also generated between these genes and their homologs in *A. thaliana*. This revealed a complex interaction network with other 20 genes. This network cluster also contained members of the plant phytochrome (Phy) family: A, B, C, D, and E (Figure S14).

## Pangenome-wide identification of *Citrus* GLPs

A total of 57 genes from *C. sinensis* (CsGLPs); 15 from *C. medica* (CmGLPs); 56 from *C. ichangensis* (CiGLPs); 58 from *C. grandis* (CgGLPs); 30 from *A. buxifolia* (AbGLPs); 57 from *F. Hindsii* (FhGLPs) and *C. clementina* (CcGLPs); 40 from *C. reticulata* (CrGLPs); 50 from *C. unshiu* Marc (CuGLPs); 29 from *P. trifoliata* (PtGLPs); and 43 genes from *M. paniculata* (MpGLPs) genome were identified. All these identified members were confirmed for the presence of the Cupin domain. Each member was named according to their phylogenetic relationships (Table S6). Among these identified members, 15 were shared by all genomes. 42 members had varied numbers in different varieties species and were termed dispensable genes. Further, one unique gene belonging to *C. grandis* genome was identified (Fig. 2F).

The *Citrus* GLP proteins had pI values ranging from 4–10, indicating both acidic and basic behaviors. The II values of most of these proteins were less than 40, suggesting their stability in the test tube. However, a few proteins had II values above 40, indicating instability in the test tube. The high AI values of all the *Citrus* GLP proteins suggested that these proteins are thermally stable. Additionally, most of these GLP proteins were found to be hydrophobic based on their GRAVY values. However, some proteins showed deviant behavior; their negative GRAVY values indicated that these proteins are hydrophilic (Table S6, Note S1).

### *Phylogenetic relationships of Citrus GLPs*

A phylogenetic tree among *Citrus* and other species GLPs was constructed to analyze the evolutionary relationships. The phylogenetic tree was clustered into six groups: 1–6. Group 6 comprised the largest clade containing members from all 11 *Citrus Spp*. Members of this group shared homology with *O. sativa* GLPs (OsGLPs). Group 2 comprised the smallest clade and contained no member from any of the identified GLPs from 11 *Citrus Spp*. Members of this group, the GLPs from *Z. mays* (ZmGLPs) shared homology with OsGLPs. Additionally, *C. medica* contained members only in Groups 4 and 6. Similarly, *P. trifliata* GLPs members were found only in Groups 4,5, and 6 and contained no members in Groups 1,2, and 3 (Fig. 4).

### *Conserved motifs and gene structure analysis of Citrus GLPs*

The conserved motifs and gene structure of *Citrus* GLPs were analyzed to gain insights into their evolutionary patterns. Members belonging to clades 1, 3–6 shared a similar pattern of motif conservation across all *Citrus Spp*. However, fewer members of clade 6 had different conserved motifs, but their conservation pattern was the same across all species. The gene structure showed that the exons and introns patterns were conserved among

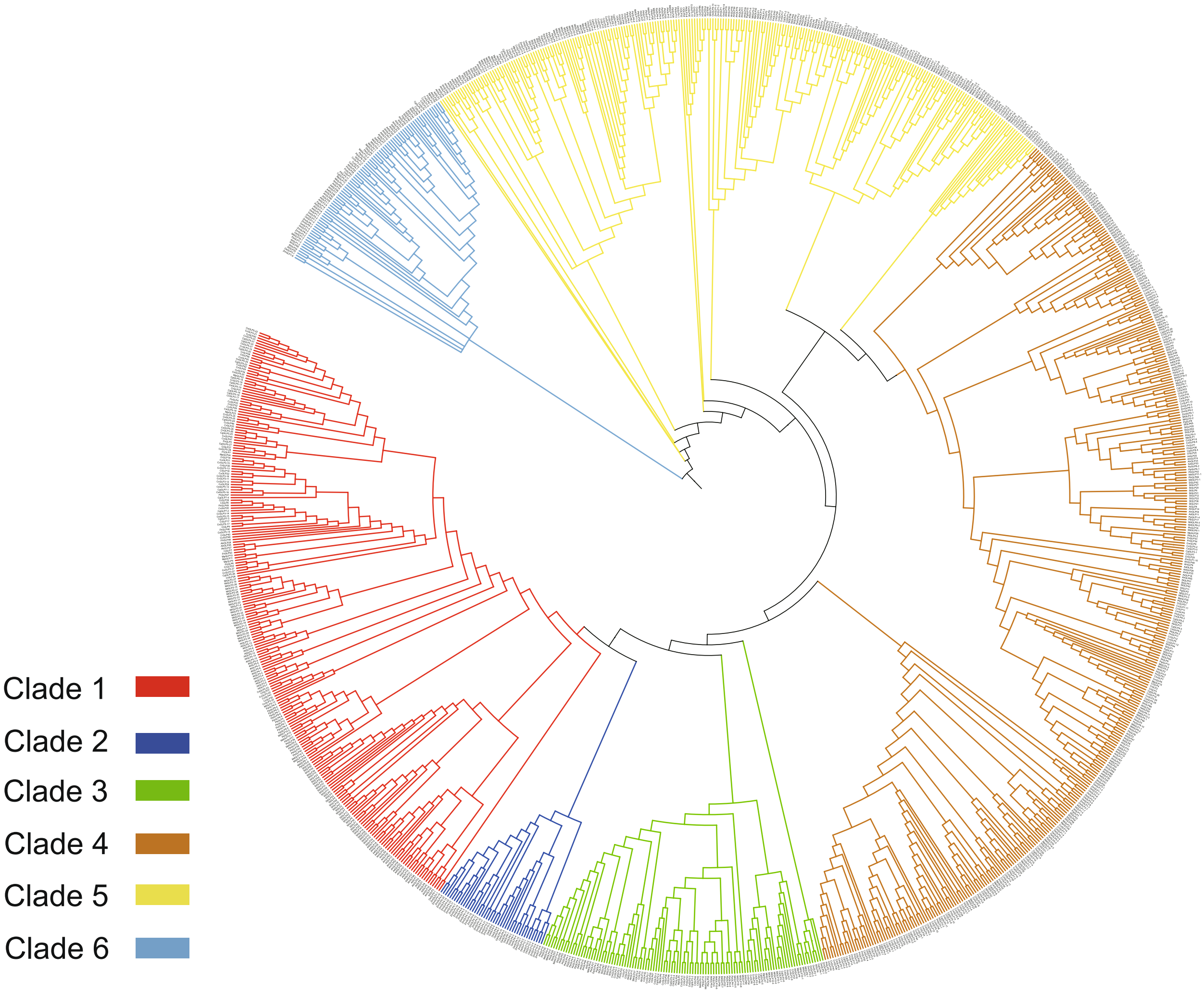

**Fig. 4** A maximum-likelihood (ML) phylogenetic tree with 1000 bootstrap replicates was generated using GLP members from 11 *Citrus* and 13 other plant species. Different colors specify different clades

the members of the same phylogenetic clade, however, it varied across species. Most of the members contained one exon and no intron (Figure S15-S25).

### *Chromosomal mapping, comparative syntenic, and gene duplication analysis*

The chromosomal location of genes was identified to evaluate the distribution pattern of *Citrus GLPs.* The *GLP* members from *C. sinensis*, *C. grandis*, and *P. trifoliata* were distributed on nine chromosomes in each genome and showed conservation in distribution patterns. The highest number of conserved genes were present on Chr5. However, *CsGLP10* was present on another unknown scaffold. The *GLP* members from the rest of the *Citrus* genomes were randomly distributed on different scaffolds (Figure S26-36).

According to a synteny analysis, *C. sinensis* and the other two inherited plant species, including *C. grandis* and *P. trifoliata*, all have substantial orthologs of the *GLP* genes (Fig. 5A, Table S11). Two *CsGLPs* (*CsGLP1-1* and *CsGLP1-2*) on chr1 showed syntenic relationships, respectively, with two *CgGLPs* (*CgGLP1-2* and *CgGLP1-3*) on chr1 and two *PtGLPs* (*PtGLP7-3* and *PtGLP7-2*) on chr7. On chr2, one *CsGLP* (*CsGLP2-6*) displayed synteny with one *CgGLP* (*CgGLP2-1*) on chr2, while no syntenic associations were observed in *P. trifoliata*. On chr3, two *CsGLP* genes (*CsGLP3-1* and *CsGLP3-3*) exhibited synteny with two *CgGLPs* (*CgGLP3-2* and *CgGLP3-3*) on chr3 and one *PtGLP* (*PtGLP5-2*) on chr5. On chr4, only one *CsGLP* (*CsGLP3-1*) have synteny correlation with two *CgGLPs* (*CgGLP5-14* and *CgGLP5-30*) on chr5 and one *PtGLP* (*PtGLP3-11*) on chr3. On chr5, 12 *CsGLPs* (*CsGLP5-1, CsGLP5-2, CsGLP5-4, CsGLP5-9, CsGLP5-12, CsGLP5-13, CsGLP5-18, CsGLP5-21, CsGLP5-22, CsGLP5-23, CsGLP5-24,* and *CsGLP5-29*) demonstrate syntenic associations with 11 *CgGLPs* (*CgGLP5-1, CgGLP5-10, CgGLP5-13, CgGLP5-16, CgGLP5-18, CgGLP5-20, CgGLP5-23, CgGLP5-24, CgGLP5-28, CgGLP7-1,* and *CgGLP9-2*) on chr5, chr7, and chr9 as well as eight *PtGLPs* (*PtGLP3-1, PtGLP3-3, PtGLP3-5, PtGLP3-6, PtGLP3-7, PtGLP3-8, PtGLP3-9,* and *PtGLP9-2*) on chr3 and chr9. On chr6, two *CsGLPs* (CsGLP6-1 and CsGLP6-2) have syntenic connections with two *CgGLPs* (*CgGLP5-14* and *CgGLP5-30*) on chr5, as well as two *PtGLPs* (*PtGLP3-11* and *PtGLP6-1*) on chr3 and chr6. Notably, *CsGLP6-1* is linked to both *CgGLP5-14* and *CgGLP5-30*, and *PtGLP3-11*, while *CsGLP6-2* exhibits connections with *PtGLP6-1*. For chr7, four *CsGLPs* (*CsGLP7-1, CsGLP7-2, CsGLP7-3,* and *CsGLP7-4*) display synteny with three *CgGLPs* (*CgGLP5-15, CgGLP7-5*, and *CgGLP7-6*) on chr5 and chr7, as well as four *PtGLPs* (*PtGLP4-1, PtGLP4-2, PtGLP10,* and *PtGLP10*) on chr4 and chrUn. On chr8, three *CsGLPs* (*CsGLP8-1, CsGLP8-3,* and *CsGLP8-4*) genes have syntenic correlations with three CgGLPs (*CgGLP8-1, CgGLP8-3,* and *CgGLP8-4)* on chr8 and three *PtGLPs* (*PtGLP8-1, PtGLP8-2,* and *PtGLP8-3*) on chr8. On chr9, two *CsGLPs* (*CsGLP9-1* and *CsGLP9-2*) exhibit synteny with four *CgGLPs* (*CgGLP5-28, CgGLP5-29, CgGLP9-1,* and *CgGLP9-2*) on chr5 and chr9, as well as four *PtGLPs* (*PtGLP3-9, PtGLP3-10, PtGLP9-1,* and *PtGLP9-2*) on chr3 and chr9. Lastly, on chrUn, only one *CsGLP10* exhibit synteny with two *CgGLPs* (*CgGLP5-29* and *CgGLP9-1*) on chr5 and chr9, as well as two *PtGLPs* (*PtGLP3*−10 and *PtGLP9-1*) on chr3 and chr9. Overall, this comprehensive synteny analysis highlights the evolutionary conservation and divergence of GLP genes across *C. sinensis, C. grandis, and P. trifoliata*, illustrating their importance in genomic studies.

The gene duplication analysis showed that *C. sinensis* contains 63 pairs of duplicated genes. Most of these resulted from tandem duplication while 13 pairs were found to be segmentally duplicated. 32 pairs of *CiGLPs* were found, of these 2 pairs resulted from tandem duplication while the rest showed segmental duplication. *C. grandis*, *F. Hindsii*, and *M. paniculata* contained 87, 53, and 14 pairs of duplicated genes originating from both segmental and tandem duplication. *C. clementina* showed the highest number of duplicated gene pairs, 103; and only 5 of these were segmentally duplicated. *A. buxifolia* showed the minimum number of duplicated gene pairs, five; out of these only one was segmentally duplicated. *C. reticulata*, *C. unshiu*, and *P. trifoliata* contained 13, 50, and 11 pairs of duplicated genes originating from both segmental and tandem duplication. These para-homologous genes' Ka/Ks ratio across 11 genomes ranged from 0.07 to 4.95, showing both positive and negative selection events. The time of divergence of these duplicated gene pairs was

(See figure on next page.)

**Fig. 5** **A** Synteny analysis of GLP genes among *C. sinensis*, *C.grandis*, and, *P.trifoliata* genomes. The collinear blocks among these species' genomes are shown with grey lines, while syntenic pairs of GLP genes are highlighted with pink. The chromosome number is indicated above each respective chromosome of individual genomes. **B** predicted miRNAs potentially targeting the *ScGLPs* and the target sites. **C** Predicted biological processes (BP), cellular components (CC), and molecular functions (MF) associated with *C. sinensis* GLPs. **D** Interactions among CsGLPs and other homologous proteins

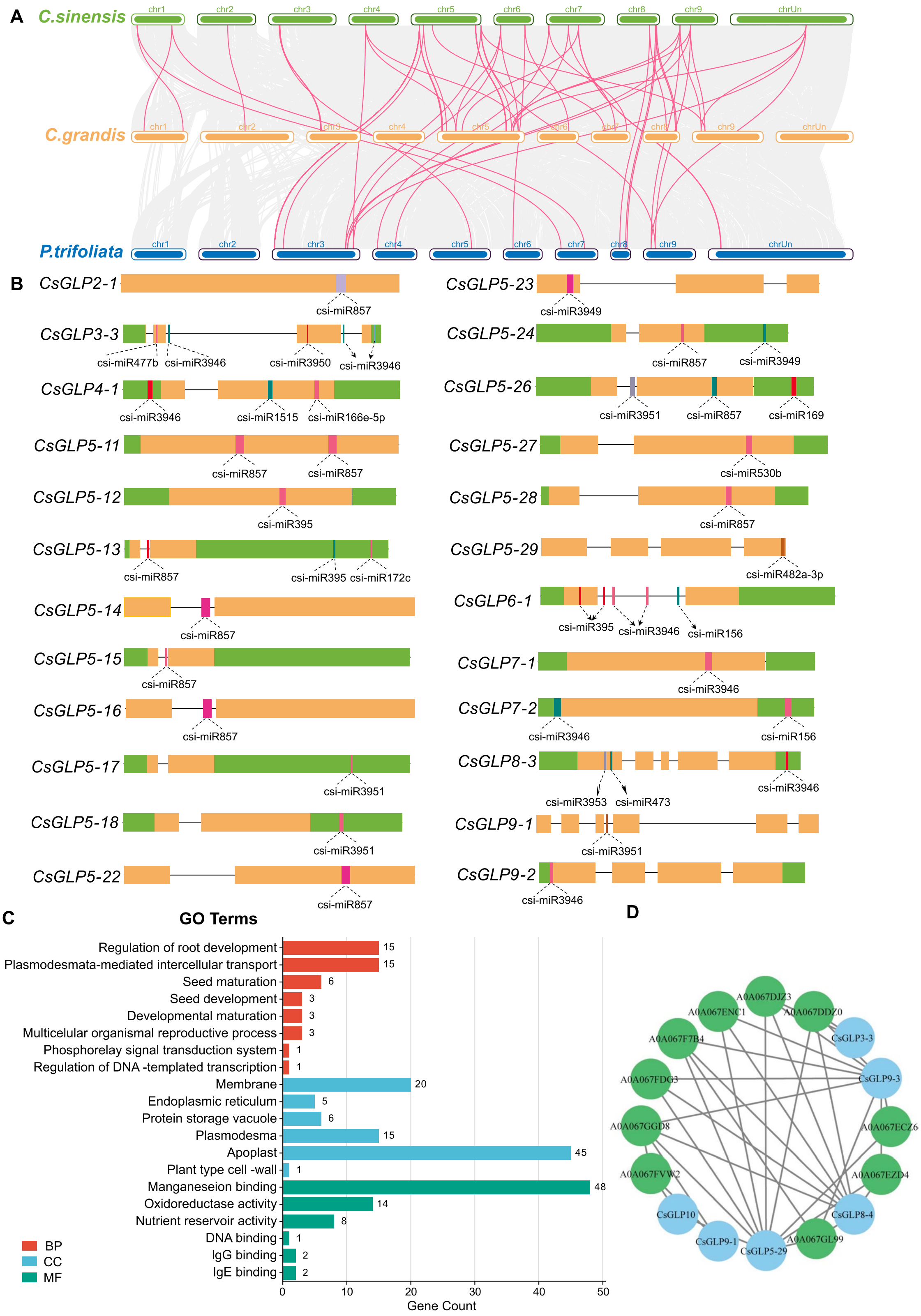

**Fig. 5** (See legend on previous page.)

between 0.4 and 154.81 million years ago (MYA), suggesting a significant period of evolutionary divergence (Table S7).

### *miRNA prediction, PPI, and GO enrichment analysis*

The miRNAs targeting the *CsGLPs* were predicted to gain insights into miRNA-mediated post-transcriptional stress regulation of these genes. 24 *CsGLPs* were targeted by miRNAs belonging to 16 different families (Fig. 5B, Table S8). However, further studies are required to determine their involvement in biological roles of *CsGLPs*. A PPI network of CsGLPs was generated to check their functional relativity. The CsGLP3-3, CsGLP5-29, CsGLP8-4, CsGLP9-1, CsGLP9-3, and CsGLP10 interacted with other homologous proteins from *C. sinensis*, which showed their potential roles in growth, development, and stress regulation mechanisms (Fig. 5D).

Based on GO analysis, the *CsGLPs* were found to be involved in various BPs including root and seed development; and transcription regulation. CCs revealed their localization in the membrane, vacuole, cell wall, etc. The MFs associated with these genes included ion binding, DNA binding, and nutrient reservoir activity (Fig. 5C). All these terms indicated their involvement in growth and stress responsiveness mechanisms.

### *Cis-regulatory element analysis of Citrus GLPs*

The *Citrus* GLPs were analyzed for the *cis*-regulatory elements present in their promoter region. All the identified *GLPs* from *Citrus* genomes contained several *cis*-elements associated with development, hormone, light, and stress responsiveness. Five elements related to development were CAT-box, MBSI, circadian, HD-Zip 1, and o2-site. Elements associated with hormone responsiveness included P-box, TGA-element, ABRE, CGTCA-motif, and TCA-element. Light-related elements included G-box, GT1-motif, GATA-motif, and Box 4. Moreover, GC-motif, LTR, TC-rich repeats, and MBS were the elements related to stress responsiveness (Figure S37-47).

### *Expression profiling of CsGLPs*

The RNA-seq data of *C. sinensis* was analyzed to identify the responsiveness of *GLP* genes to HLB and citrus canker disease conditions. For HLB disease the expression was analyzed in before and post-infection conditions in leaves of *C. sinensis* plant. Few *CsGLP* genes showed a change in their expression patterns. *CsGLPs1-2* showed a higher expression in pre- and post-HLB infection conditions. Similarly, *CsGLPs2-1*, *CsGLPs3-3*, *CsGLPs4-1*, *CsGLPs5*-20, CsGLPs6*−1*, *CsGLPs6-2*, *CsGLPs8-1*, *CsGLPs8-3,* and *CsGLPs8-4* also showed a change in their expression patterns (Fig. 6A).

The expression data of *C. sinensis* leaves infected with *Xanthomonas citri* (XC infection) and *X. aurantifolii* (XA infection) was analyzed at different stages of diseased conditions. *CsGLP* genes including *CsGLPs1-2, CsGLPs1-3*¸ *CsGLPs4-1, CsGLPs7-4, CsGLPs6-1, CsGLPs6-2, CsGLPs8-1, CsGLPs8-2,* and *CsGLPs8-4* showed a change in their expression patterns (Fig. 6B).

### *Expression validation of CsGLPs through qRT-PCR*

To further validate the expression levels of *CsGLPs* under HLB disease conditions, qRT-PCR was performed on leaf samples of both HLB tolerant (*M. paniculata*, *A. buxifolia,* and *C. ichangensis*) and susceptible (*C. medica, C. maxima, C. sinensis,* and, *C. reticulata*) citrus varieties. The selected species were chosen for their diverse genetic backgrounds, significant commercial value, and distinct phenotypic traits, ensuring they serve as a comprehensive representation of the broader citrus germplasm. The results revealed a significant upregulation of *CsGLPs1-2* (Cs1gpb016570) in response to HLB infection across all seven tested citrus varieties. Notably, the highest expression levels were observed in the HLB-tolerant varieties, specifically *M. paniculata*, *A. buxifolia,* and *C. ichangensis*. This upregulation pattern was consistently observed in susceptible species as well (Fig. 7A). Furthermore, a parallel expression trend was noted for *CsGLPs8-4* (Cs8gpb018800) across the same set of citrus species (Fig. 7B). *CsGLPs8-4* showed higher expression levels in HLB-infected leaves as compared to the samples under controlled conditions. These findings collectively suggest that *CsGLPs* are highly responsive to HLB pathogenicity, indicating their potential role in the molecular mechanisms underlying the citrus plant's response to HLB disease.

## Discussion

In this study, we developed a comprehensive draft pangenome using 11 publicly available *Citrus* genomes which provided insights into the intraspecies diversity and stress responsiveness of citrus plants. The selected genomes were prioritized due to the availability of high-quality sequencing data generated using advanced platforms such as PacBio, Illumina, and Sanger, ensuring reliability and accuracy for downstream analyses. Five species: *C. grandis*, *F.hindsii*, *M. paniculata*, *P.trifoliata,* and *C. unshiu* Marc were sequenced by PacBio long-reads sequencing platform [92]. Other five species: *A. buxifolia*, *C. ichangensis*, *C. reticulata*, *C. sinensis*, and *C. medica* were sequenced by Illumina short-reads sequencing platforms [93]. One variety, *C. clementina* was sequenced by Sanger sequencing protocol [32]. All the *Citrus* genomes except *C. unshiu* Marc were annotated by combined Ab initio gene prediction evidence

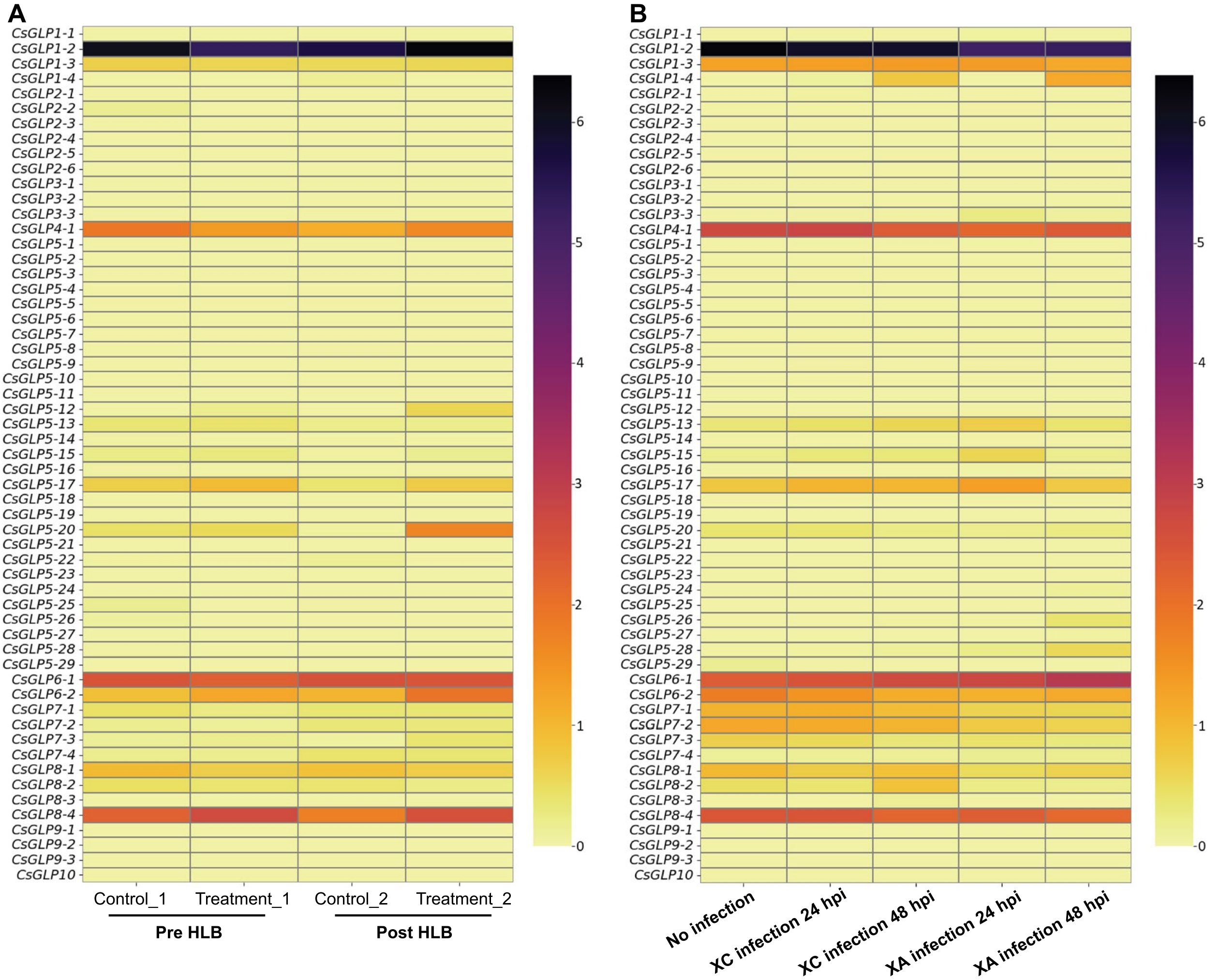

**Fig. 6** Heatmap showing the expression patterns of *CsGLPs*, (**A**) before and post condition of leaves infected with HLB disease (**B**) leaves infected with citrus canker disease; *X. citri* (XC) and *X. aurantifolii* (XA)

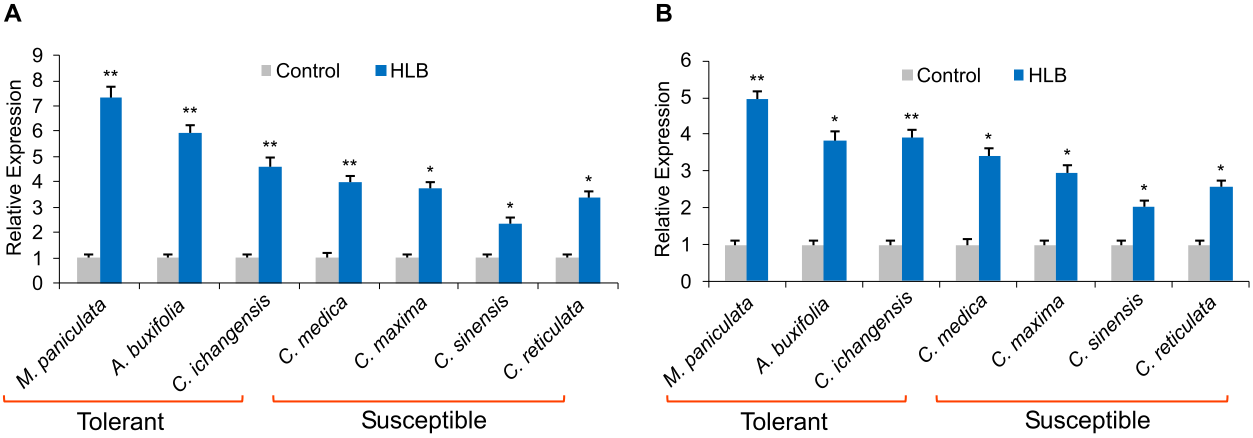

**Fig. 7** Expression analysis of leaves infected from HLB, across seven species. **A** *CsGLPs1-2* (Cs1g_pb016570), (**B**): *CsGLPs8-4* (Cs8g_pb018800). Each graph column signifies the mean of 3 replications and the student's t-test was used to compare the control and HLB-infected samples at * $p < 0.05$ and ** $p < 0.01$

[46], homology sequences analysis, and RNA sequencing (RNA-seq) analysis. Whereas, *C. unshiu* Marc was annotated using gene modeling approach through MAKER-P [47]. Thus, in the case of *Citrus* species, more than 80% of the sequence has been assembled and annotated. *C. medica* has the biggest assembled genome of 368.46 Mb with 32,579 genes. The sequence identity among these genomes was found 95% on average, however, *M.paniculata* shared fewer identical sequences than other cultivars. Based on sequence identity, GC content, gene number, and TEs, *C. grandis* was used as a reference genome in the pangenome construction to capture the genetic diversity among the *Citrus* genomes. All these assembled and annotated genomes were successfully used in this study to identify the SVs, particularly the PAVs contributing to *Citrus* genomic diversity, and to identify the candidate genes contributing to HLB and citrus canker disease responsiveness.

We identified 3,305–52,748 PAVs contributed by the *Citrus* genomes. The major portion of these PAVs was contributed by the two species: *A. buxifolia* and *M.paniculata*. However, when mapped against the pangenome, a better mapping rate was observed. *C. clementina, C. unshiu, M. paniculata,* and *C. sinensis* showed 99.99%, 99.98%, 99.97%, and 99.96% genes mapped, respectively. These results are consistent with the previous studies on pangenomes providing better resources to capture and study genetic diversity by representing more than 90% of the estimated genome size [21, 94]. In our study, the developed pangenome had 56% TEs. Another study reporting the super-pangenome of four species of cultivated *Citrus* by producing chromosome-scale de novo assemblies showed each genome composed of ~ 50% TEs [95]. The current study showed 59,261 pangenes clusters containing 12,770 core genes, 2,980 soft-core, and 10,342 shell genes. Droc et al., implemented super-pangenome of four species of cultivated citrus and showed 32,962 pangenes consisting of 25,291 core genes, 2,431 as soft-core, and 5,171 shell or dispensable genes [95]. These results show a comparable trend of number of pangenes with the current study as core genes contribute the highest number of pangenes, followed by shell and then soft-core genes. However, the number of pangenes varies as the number of genomes used to construct the pangenome. Moreover, the modelling of the pangenome suggested a closed pangenome with a finite number of genes as observed in other studies [17, 96, 97]. Similarly, Huang et al., also reported an increase in gene family with the addition of *Citrus* genomes, finally reaching a plateau [6]. However, the pangenome analysis done using 10 Citrus accessions reported by Gao et al., showed that gene number also increased with the addition of genomes but did not reach the closeness [21]. The orthologous gene family analysis showed that 35.75% of gene families were core which is consistent with a previous report on 18 *Citrus* species pangenome, where 35.8% of the gene families were found to be core [6]. Similarly, private pangene families observed were 62.64% which is comparable to the 64.6% private gene families reported by Gao et al., [21]. Moreover, shell genes also contained fewer exons per gene than the core genes consistent with previous reports concerning genes displaying PAV [98, 99]. In citrus super-pangenome study, out of these 32,962 orthologous pangenes, *C. medica* contributed the highest number of diagnostic PAVs which provides insights into their evolutionary pressures and species-specific traits [95]. The pangene families analyzed using 18 species belonging to *Citrus* and *Citrus*-related genera in previous study showed 333 expanded and 401 contracted gene families [6]. However, our study reports 818 expanded and 9,723 contracted gene families which is the largest number of contracted gene families among all 11 *Citrus* species. Such a large number of contracted gene families might be responsible for phenotypic and biological variation between *M. paniculate* and other *Citrus* species.

The phylogenetic tree among 11 *Citrus* species showed that *M. paniculata*, *A. buxifolia,* and *C. trifoliata* were more divergent compared with other species. While, *C. medica*, *C. grandis*, and *C. ichangensis* had the highest number of expanded gene families. These results are consistent with the initial sequence-based and gene-based pangenomes results, where it was identified that *M.paniculata* is the most outlier variety and also supported by previous studies *C. ichangensis and C. trifoliata* were found among early diverging *Citrus* species [6]. These results are also consistent with the initial sequence-based and gene-based pangenomes results, where it was identified that *M.paniculata* is the most outlier variety and such a large number of contracted gene families might be responsible for phenotypic and biological variation between *M.paniculata* and other *Citrus* species. The GO and network analysis of the pangenes revealed their roles mainly in stress response, defense response, and immune system responsiveness. Members of phytochrome superfamily have also been observed during network clustering which has significant roles in bacteria, algae, fungi, and plants [100–102]. The presence of Phy family member proteins in the defense cluster shows that they are not only involved in the light activation process but also possibly involved in defense processes in *Citrus*. These findings are reinforced by the results shown by the study on super-pangenome of citrus where the presence of PRR genes in four citrus species show their role in adaptive stress responsiveness to environmental conditions [95].The results from GO revealed that significant number of accessory genes (4,936) with potential biological functions were missing from the *C. grandis* genome. These could be the candidate genes responsible for genetic and phenotypic variation among different *Citrus* species. Thus, emphasizing the use of pangenome instead of a single reference genome.

The pangenome-wide analysis of GLP gene family in 11 *Citrus* species provided a comprehensive overview of intraspecies genomic diversity and its responsiveness in disease pathogenicity. The number of identified members fluctuated with the size of the genome, however, the overall number was consistent with previously identified members [8,5]. Previously, this gene family has been reported to be involved in stress-responsiveness such as heat and salt [72]; disease resistance [103], and powdery mildew stress [10] which suggests the contribution of GLP gene family members in stress responsiveness.

The phylogenetic tree showed that this gene family members are divided into six clades based on homology [8], however, *C. medica* and *P. trifoliata* showed deviant behavior. *C. medica* contained members only in 4 and 6 clades, while, *P. trifoliata* contained members only in 4, 5, and 6 clades. The phylogenetic tree constructed using pan gene family clusters also showed *P. trifoliata* as the divergent species. Some motifs were highly conserved across the *Citrus Spp.*, suggesting their crucial roles in essential biological processes. Other motifs displayed variations or specificity within certain *Citrus Spp.*, implying functional divergence and adaptations. Intraspecies conserved motifs analysis showed the similarity and conservation in intraspecies proteomes. Conserved motifs observed in *C. grandis* and *C. sinensis* were large in size and exhibited a linear distribution pattern throughout the gene family like motifs identified and characterized in *C. melo* [13] and *T. aestivum* [10]. The exons were observed to be relatively conserved in terms of their lengths and positions within the gene structure, suggesting conserved functional domains or regions. In contrast, the introns varied significantly in length and sequence composition, highlighting the potential for intron divergence and evolution. Chromosomal mapping showed the uneven distribution of *GLP* genes on chromosomes and scaffolds of *Citrus Spp.* All these identified *GLP* members exhibited both positive and purifying selection.

Syntenic relationships indicate the preservation of gene order despite speciation, aligning with findings in other plant families where key gene families exhibit strong synteny. In *Brassica rapa*, synteny analysis identified syntenic gene pairs with other Brassicaceae species, highlighting the preservation of gene order and function within this family [104]. Similarly, a genome-wide study in peanut (*A. hypogaea*) identified 84 GLP genes, with synteny analysis revealing segmental duplications as a key mechanism for GLP gene family expansion [9]. In potato (*S. tuberosum*), 70 GLP genes were identified, and synteny analysis indicated that tandem duplications contributed to the expansion of this gene family [72]. These findings align with our synteny analysis in Citrus species, which demonstrated significant conservation of GLP genes among *C. sinensis, C. grandis, and P. trifoliata*, suggesting that *GLP* genes play fundamental roles maintained through evolutionary pressures. However, the observed divergences, such as the absence of certain syntenic associations in *P. trifoliata*, may reflect species-specific adaptations or genomic rearrangements [105]. In conclusion, synteny analyses across different plant species, including Citrus, *B. rapa*, peanut, and potato, reveal that while GLP gene families are generally conserved, they also exhibit species-specific expansions and rearrangements. These patterns underscore the dynamic nature of plant genomes and the role of *GLP* genes in plant development and stress responses. *Cis*-regulatory elements are the key players in regulating the stress-responsive activities of *GLP* genes, thus regulating gene expression [10]. The results of *cis*-regulatory elements predicted main functions in which *GLPs* are involved including hormone responsiveness, stress responsiveness, growth and development responsiveness, and light responsiveness. These results are consistent with the ones identified using pangenome for functional characterization. Further, the GO and PPI analysis also indicated the roles of these genes in stress regulation activities. miRNAs belonging to multiple families also targeted *CsGLPs*, thus, controlling their expression levels as previously reported in potato [72]. However, further studies are needed to validate the miRNAs-mediated gene expression.

Previously, reports have been carried out to understand the genetic basis of HLB disease in *Citrus* accessions [21, 106, 107]. Similarly, various studies have been carried out to identify the expression of genes against the citrus bacterial canker studies [108, 109]. In this study, the RNA-seq expression analysis of publically available data was carried to understand the involvement of *CsGLP* genes in HLB and citrus canker disease condition. This study identifies the differential expression of certain *GLP* genes including *CsGLP1-2, CsGLP2-1*, *CsGLP3-3*, *CsGLP4-1, CsGLP6-1*, *CsGLP6-2*, *CsGLP8-3,* and *CsGLP8-4* in disease resistance against HLB and citrus canker infection. Further, the expression levels of the candidate *CsGLP1-2 and CsGLP8-4* genes were upregulated in the tolerant and susceptible *Citrus* species infected with HLB disease. The differential expression patterns observed between susceptible and tolerant citrus species underscore the importance of *CsGLPs* in the plant's defense strategy, potentially offering insights into breeding programs aimed at enhancing HLB resistance in Citrus crops.

Overall, this study has provided a comprehensive resource including sequence, gene, and gene family-based pangenome. The pangenome provided several PAVs and additional genes that could contribute to various phenotypic and genotypic variations among *Citrus* genomes. Additionally, the GLP gene family analysis also revealed the intraspecies diversity among *Citrus* genomes and identified putative *CsGLP* genes responsible for HLB tolerance in *Citrus* accessions.

## Conclusion

Our study represents a significant advancement in our understanding of the genomic diversity and stress responsiveness within *Citrus* species by developing a detailed *Citrus* pangenome using 11 genomes. The PAVs identified across the genomes reveal critical insights into the structural and functional diversity among *Citrus* species, with important implications for breeding programs aiming to improve disease resistance and crop resilience. The gene-based pangenome analysis provided a detailed classification of core and shell genes, revealing that shell genes, which exhibit higher evolutionary rates and reduced functional constraints, are enriched in stress-responsive functions, particularly in defense mechanisms and immune responses. The GLP pangenes family analyses further evaluated the intraspecific diversity across species at the gene family level. Similar to PAVs obtained through pangenome, the GLP gene family members also showed structural diversity across 11 genomes including core, dispensable, and one unique gene. Further, functional analysis showed the putative involvement of *Citrus GLPs* in stress and development-related mechanisms. Transcriptome analysis showed the difference in expression of certain *CsGLPs* in HLB and citrus canker disease pathogenicity. The two genes *CsGLPs1-2* and *CsGLPs8-4* showed an elevated expression in both susceptible and tolerant citrus species under HLB disease conditions. qRT-PCR results also confirmed the higher expression of *CsGLPs1-2* and *CsGLPs8-4* across seven *Citrus* species infected with HLB. The differential expression of these genes in response to disease stress across multiple species suggests their potential as molecular markers for breeding programs targeting disease resistance. Overall, this study lays the foundation for leveraging pangenomics in *Citrus* to address challenges related to biotic and abiotic stresses, ultimately contributing to the sustainability and productivity of *Citrus* crops.

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12870-025-06397-x.

Supplementary Material 1.

Supplementary Material 2.

**Acknowledgements**

The authors are thankful to the Bioinformatics Computing Platform of the National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, China, for providing the computational resources for this study.

**Authors' contributions**

M.T.Q., K.F., M.J.R., and Q.T. contributed equally to the analysis and experiments of the study and wrote the first draft of the manuscript. M.S. assisted in the data acquisition and interpretation. B.D. provided technical expertise and support for the experiments. L.L.C. and X.T.Z. conceptualized and supervised the project and secured funding. All authors reviewed, revised, and approved the final version of the manuscript.

**Funding**

This work was supported by the Guangxi Science and Technology Major Program (AA23062085), Guangxi Natural Science Foundation (2024GXNSFGA010003), and the National Natural Science Foundation of China (32270712).

**Data availability**

The data presented in this study are available within the article or in its supplementary material. All scripts utilized in this study are publicly accessible on GitHub at the following link: https://github.com/tahirulqamar/Citrus_Pangenes.

## Declarations

**Ethics approval and consent to participate**

Not applicable.

**Consent for publication**

Not applicable.

**Competing interests**

The authors declare no competing interests.

**Author details**

[1]State Key Laboratory for Conservation and Utilization of Subtropical Agro-Bioresources, College of Life Science and Technology, Guangxi University, Nanning, Guangxi 530004, China. [2]National Key Laboratory of Crop Genetic Improvement, College of Informatics, Huazhong Agricultural University, Wuhan 430070, China. [3]College of Natural & Agricultural Sciences, University of California, Riverside, CA 92521, USA. [4]State Key Laboratory of Subtropical Silviculture, College of Forestry and Biotechnology, Zhejiang A&F University, Hangzhou, Zhejiang 311300, China. [5]UMR CNRS 6553 Ecosystèmes, Biodiversité, Evolution (ECOBIO), Université de Rennes 1, Rennes, France. [6]Engineering Research Center of Coal-Based Ecological Carbon Sequestration Technology of the Ministry of Education and Key Laboratory of National Forest and Grass Administration for the Application of Graphene in Forestry, Shanxi Datong University, Datong, Shanxi 037009, People's Republic of China. [7]College of Horticulture, Shanxi Agricultural University, Taigu, Shanxi 030801, People's Republic of China.

Received: 22 November 2024 Accepted: 13 March 2025
Published online: 27 March 2025

## References

1. Liu Y, Heying E, Tanumihardjo SA. History, global distribution, and nutritional importance of citrus fruits. Compr Rev Food Sci Food Saf. 2012;11:530–45.
2. Tiwari T, et al. Genetic and physiological characteristics of CsNPR3 edited citrus and their impact on HLB tolerance. Front Genome Ed. 2024;6:1485529.
3. Mubeen M, et al. Innovative strategies for characterizing and managing huanglongbing in citrus. World J Microbiol Biotechnol. 2024;40:342.
4. Waqif H, et al. Algal macromolecular mediated synthesis of nanoparticles for their application against citrus canker for food security. Int J Biol Macromol. 2024;263:130259.
5. Mao L, et al. ZmGLP1, a germin-like protein from maize, plays an important role in the regulation of pathogen resistance. Int J Mol Sci. 2022;23(22):14316.
6. Huang Y, et al. Pangenome analysis provides insight into the evolution of the orange subfamily and a key gene for citric acid accumulation in citrus fruits. Nat Genet. 2023;55:1964.
7. Dunwell JM, Gibbings JG, Mahmood T, Saqlan Naqvi SM. Germin and germin-like proteins: Evolution, structure, and function. CRC Crit Rev Plant Sci. 2008;27:342–75.
8. Li L, Xu X, Chen C, Shen Z. Genome-wide characterization and expression analysis of the Germin-like protein family in rice and Arabidopsis. Int J Mol Sci. 2016;17:1622.
9. Yang Q, et al. Genome-wide identification of germin-like proteins in peanut (Arachis hypogea L.) and expression analysis under different abiotic stresses. Front Plant Sci. 2023;13:1–26.
10. Yuan B, et al. Genome-Wide Identification and Characterization of Germin and Germin-Like Proteins (GLPs) and Their Response Under Powdery Mildew Stress in Wheat (Triticum aestivum L.). Plant Mol Biol Report. 2021;39:821–32.
11. Gangadhar BH, et al. Enhanced thermo-tolerance in transgenic potato (Solanum tuberosum L.) overexpressing hydrogen peroxide-producing germin-like protein (GLP). Genomics. 2021;113:3224–34.
12. Liao L, Hu Z, Liu S, Yang Y, Zhou Y. Characterization of germin-like proteins (Glps) and their expression in response to abiotic and biotic stresses in cucumber. Horticulturae. 2021;7:412.
13. Zhang Z, et al. Genome-wide identification, characterization, and expression analysis related to low-temperature stress of the cmglp gene family in Cucumis melo L. Int J Mol Sci. 2022;23:8190.
14. Fu JY, et al. Identification and functional analysis of germin-like protein Gene family in tea plant (*Camellia sinensis*). Sci Hortic (Amsterdam). 2018;234:166–75.
15. He Q, et al. A graph-based genome and pan-genome variation of the model plant *Setaria*. Nat Genet. 2023;55:1232–42.
16. Ruperao P, et al. *Sorghum* pan-genome explores the functional utility for genomic-assisted breeding to accelerate the genetic gain. Front Plant Sci. 2021;12:1–17.
17. Golicz AA, et al. The pangenome of an agronomically important crop plant *Brassica oleracea*. Nat Commun. 2016;7: 13390.
18. Liu Y, et al. Pan-genome of wild and cultivated soybeans. Cell. 2020;182:162-176.e13.
19. Guo Y. Pangenome and the diversity of potato species. Nat Food. 2023;4:638.
20. Gao L, et al. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. Nat Genet. 2019;51:1044–51.
21. Gao Y, et al. Citrus genomic resources unravel putative genetic determinants of Huanglongbing pathogenicity. iScience. 2023;26:106024.
22. Jayakodi M, Schreiber M, Stein N, Mascher M. Building pan-genome infrastructures for crop plants and their use in association genetics. DNA Res. 2021;28:dsaa030.
23. Garrison E, et al. Building pangenome graphs. Nat Methods. 2024;21:1–5.
24. Tahir ul Qamar M, Zhu X, Khan MS, Xing F, Chen LL. Pan-genome: a promising resource for noncoding RNA discovery in plants. Plant Genome. 2020;13:e20046.
25. Cannon SB, Lee H-O, Weeks NT, Berendzen J. Pandagma: a tool for identifying pan-gene sets and gene families at desired evolutionary depths and accommodating whole genome duplications. Bioinformatics. 2024;btae526:btae526. https://doi.org/10.1093/bioinformatics/btae526.
26. Contreras-Moreira B, et al. GET_PANGENES: calling pangenes from plant genome alignments confirms presence-absence variation. Genome Biol. 2023;24:223.
27. Cheng C, Shi X, Wu J, Zhang Y, Lü P. Genome-scale computational identification and characterization of utr introns in *Atalantia buxifolia*. Horticulturae. 2021;7:1–13.
28. Wang X, et al. Genomic analyses of primitive, wild and cultivated citrus provide insights into asexual reproduction. Nat Publ Gr. 2017;49(5):765–72.
29. Wang L, et al. Genome of wild mandarin and domestication history of mandarin. Molecular Plant. 2018;11:1024–37.
30. Zhu C, et al. Genome sequencing and CRISPR / Cas9 gene editing of an early flowering Mini-Citrus (Fortunella hindsii). Plant Biotechnol J. 2019;17(11):2199–210.
31. Xu Q, et al. The draft genome of sweet orange (*Citrus sinensis*). Nat Genet. 2013;45:59–66.
32. Wu GA, et al. Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. Nat Biotechnol. 2014;32:656–62.
33. Shimizu T, et al. Draft sequencing of the heterozygous diploid genome of Satsuma (Citrus unshiu Marc.) using a hybrid assembly approach. Front Genet. 2017;8:180.
34. Liu H, et al. Citrus Pan-Genome to Breeding Database (CPBD): a comprehensive genome database for citrus breeding. Mol Plant. 2022;15:1503–5.
35. Peng Z, et al. A chromosome-scale reference genome of trifoliate orange (*Poncirus trifoliata*) provides insights into disease resistance, cold tolerance and genome evolution in Citrus. Plant J. 2020;104:1215–32.
36. Marçais G, et al. MUMmer4: a fast and versatile genome alignment system. PLOS Comput Biol. 2018;14:e1005944.
37. Wu GA, et al. Genomics of the origin and evolution of Citrus. Nature. 2018;554:311–6.
38. Tahir Ul Qamar M, Zhu X, Xing F, Chen LL. ppsPCP: a plant presence/absence variants scanner and pan-genome construction pipeline. Bioinformatics. 2019;35:4156–8.
39. Camacho C, et al. BLAST + : architecture and applications. BMC Bioinformatics. 2009;9:1–9.
40. Kent WJ. BLAT—the BLAST-like alignment tool. Genome Res. 2002;12:656–64.
41. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015;6:11.
42. Flynn JM, et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proceedings of the Natl Acad Sci. 2020;117:9451–7.
43. Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinforma. 2009;25:4.10.1-4.10.14.
44. Krzywinski M, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009;19:1639–45.
45. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.
46. Stanke M, et al. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 2006;34:W435–9.
47. Campbell MS, et al. MAKER-P: A tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant Physiol. 2014;164:513–24.
48. Contreras-Moreira B, et al. Analysis of plant pan-genomes and transcriptomes with GET_HOMOLOGUES-EST, a clustering solution for sequences of the same species. Front Plant Sci. 2017;8:238135.
49. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20:53–65.
50. Huerta-Cepas J, et al. Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. Mol Biol Evol. 2017;34:2115–22.
51. Du Z, Zhou X, Ling Y, Zhang Z, Su Z. agriGO: a GO analysis toolkit for the agricultural community. Nucleic Acids Res. 2010;38:W64–70.
52. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12:59–60.

53. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20:1–14.
54. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. Nucleic Acids Res. 2002;30:3059–66.
55. Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5:e9490.
56. Sanderson MJ. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics. 2003;19:301–2.
57. De Bie T, Cristianini N, Demuth JP, Hahn MW. CAFE: a computational tool for the study of gene family evolution. Bioinformatics. 2006;22:1269–71.
58. Sievers F, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011;7:539.
59. Suyama M, Torrents D, Bork P, Delbru M. PAL2NAL : robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 2006;34:609–12.
60. Alicai T, et al. Cassava brown streak virus has a rapidly evolving genome: implications for virus speciation, variability, diagnosis and host resistance. Sci Rep. 2016;6:36164.
61. Smit AFA, Hubley R, Green P. RepeatMasker. 1996.
62. Huerta-Cepas J, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 2019;47:D309–14.
63. Tian T, et al. agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. Nucleic Acids Res. 2017;45:W122–9.
64. Conesa A, et al. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005;21:3674–6.
65. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Dawn Thompson DA, IAmit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Nir Hacohen, Hacohen N, Rhind N, Federica di Palma, Bruce WN, Friedman AR. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat. Biotechnol. 2013;29:644–652.
66. Bindea G, et al. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics. 2009;25:1091–3.
67. Bindea G, Galon J, Mlecnik B. CluePedia Cytoscape plugin: pathway insights using integrated experimental and in silico data. Bioinformatics. 2013;29:661–3.
68. Shannon P, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498–504.
69. Montojo J, et al. GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. Bioinformatics. 2010;26:2927–8.
70. Bolser D, et al. Ensembl plants: integrating tools for visualizing, mining, and analyzing plant genomic data. in Plant genomics databases 1–31. Springer; 2017.
71. Gasteiger E, et al. Protein identification and analysis tools on the ExPASy server. proteomics Protoc. Handb. 2005:571–607.
72. Zaynab M, et al. Genome-wide identification and expression profiling of germin-like proteins reveal their role in regulating abiotic stress response in potato. Front Plant Sci. 2022;12:1–19.
73. Thompson JD, Gibson TJ, Higgins DG. Multiple Sequence Alignment Using ClustalW and ClustalX. Curr Protoc Bioinforma. 2003;00:1–22.
74. Trifinopoulos J, Nguyen LT, von Haeseler A, Minh BQ. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. Nucleic Acids Res. 2016;44:W232–5.
75. Letunic I, Bork P. Interactive tree of life (iTOL) v5: An online tool for phylogenetic tree display and annotation. Nucleic Acids Res. 2021;49:W293–6.
76. Bailey TL, et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 2009;37:W202–8.
77. Chen C, et al. TBtools, a toolkit for biologists integrating various biological data handling tools with a user-friendly interface. Mol Plant. 2018;13(8):1194–202.
78. Wang Y, et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 2012;40:e49.
79. Rozas J, et al. DnaSP 6: DNA sequence polymorphism analysis of large data sets. Mol Biol Evol. 2017;34:3299–302.
80. Fatima K, et al. Integrated omics and machine learning-assisted profiling of cysteine-rich-receptor-like kinases from three peanut spp. revealed their role in multiple stresses. Front Genet. 2023;14:1252020.
81. Barta T, Peskova L, Hampl A. MiRNAsong: A web-based tool for generation and testing of miRNA sponge constructs in silico. Sci Rep. 2016;6:1–8.
82. von Mering C, et al. STRING: A database of predicted functional associations between proteins. Nucleic Acids Res. 2003;31:258–61.
83. Törönen P, Medlar A, Holm L. PANNZER2: a rapid functional annotation web server. Nucleic Acids Res. 2018;46:W84–8.
84. Rombauts S, Déhais P, Van Montagu M, Rouzé P. PlantCARE, a plant cis-acting regulatory element database. Nucleic Acids Res. 1999;27:295–6.
85. Wingett SW, Andrews S. FastQ Screen: a tool for multi-genome mapping and quality control. F1000Research. 2018;7:1338.
86. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37:907–15.
87. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc. 2016;11:1650–67.
88. Lee JA, et al. Asymptomatic spread of huanglongbing and implications for disease control. Proc Natl Acad Sci. 2015;112:7605–10.
89. Li W, Hartung JS, Levy L. Quantitative real-time PCR for detection and identification of Candidatus Liberibacter species associated with citrus huanglongbing. J Microbiol Methods. 2006;66:104–15.
90. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2−ΔΔCT method. Methods. 2001;25:402–8.
91. Schneeberger K, et al. Reference-guided assembly of four diverse Arabidopsis thaliana genomes. Proc Natl Acad Sci U S A. 2011;108:10249–54.
92. Rhoads A, Au KF. PacBio sequencing and its applications. Genomics Proteomics Bioinformatics. 2015;13:278–89.
93. Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol. 2008;26:1135–45.
94. Wang K, et al. Duck pan-genome reveals two transposon insertions caused bodyweight enlarging and white plumage phenotype formation during evolution. iMeta. 2024;3:e154.
95. Droc G. et al. A super-pangenome for cultivated citrus reveals evolutive features during the allopatric phase of their reticulate evolution. bioRxiv. 2024. https://doi.org/10.1101/2024.10.17.618847.
96. Hirsch CN, et al. Insights into the maize pan-genome and pan-transcriptome. Plant Cell. 2014;26:121–35.
97. Li Y, et al. De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nat Biotechnol. 2014;32:1045–52.
98. Bush SJ, et al. Presence-absence variation in A. thaliana is primarily associated with genomic signatures consistent with relaxed selective constraints. Mol Biol Evol. 2014;31:59–69.
99. Schatz MC, et al. Whole genome de novo assemblies of three divergent strains of rice, *Oryza sativa*, document novel gene space of aus and indica. Genome Biol. 2014;15:1–16.
100. Rockwell NC, Su Y-S, Lagarias JC. Phytochrome structure and signaling mechanisms. Annu Rev Plant Biol. 2006;57:837–58.
101. Qiu X, Sun G, Liu F, Hu W. Functions of plant phytochrome signaling pathways in adaptation to diverse stresses. Int J Mol Sci. 2023;24:1320.
102. Xiang S, Wu S, Jing Y, Chen L, Yu D. Phytochrome B regulates jasmonic acid-mediated defense response against *Botrytis cinerea* in *Arabidopsis*. Plant Divers. 2022;44:109–15.
103. Lu M, Han Y-P, Gao J-G, Wang X-J, Li W-B. Identification and analysis of the germin-like gene family in soybean. BMC Genomics. 2010;11: 620.
104. Cheng F, Wu J, Fang L, Wang X. Syntenic gene analysis between Brassica rapa and other Brassicaceae species. Front Plant Sci. 2012;3:198.
105. Ollitrault P, et al. Comparative genetic mapping and a consensus interspecific genetic map reveal strong synteny and collinearity within the Citrus genus. Front Plant Sci. 2024;15:1475965.
106. Gottwald TR, Bassanezi RB, Paulo S. Citrus Huanglongbing: The Pathogen and Its Impact Plant Health Progress Plant Health Progress. Plant Heal Prog. 2007;1993:36.
107. Ma W, et al. Citrus Huanglongbing is a pathogen-triggered immune disease that can be mitigated with antioxidants and gibberellin. Nat Commun. 2022;1:11–13.
108. Li Q, et al. Genomewide analysis of the CIII peroxidase family in sweet orange (Citrus sinensis) and expression profiles induced by Xanthomonas citri subsp. citri and hormones. J Genet. 2020;99:10.

109. Huang X, et al. Genome-wide identification and characterization of the sweet orange (*Citrus sinensis*) basic helix-loop-helix (bHLH) family reveals a role for CsbHLH085 as a regulator of citrus bacterial canker resistance. Int J Biol Macromol. 2024;267:131442.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.