



The reference genome sequence of the scarlet follicle, *Sterculia lanceolata*, reveals a paleo-polyploidization and its impact on fruit quality and fruit dehiscence

Youtao Hu¹ · Youpeng Zhang¹ · Hongbin Zhang¹ · Jiahao Zhang¹ · Guilian Guo¹ · Kejing Yang² · Bing-Yan Shao² · Jia-Yu Xue² · Robert Henry³ · Wenquan Wang¹ · Fei Chen¹

Received: 8 September 2025 / Revised: 1 December 2025 / Accepted: 30 January 2026
© The Author(s) 2026

Abstract

Sterculia lanceolata, a tree species of the Malvaceae family with notable ornamental and medicinal value, has long been constrained in genetic research and breeding applications due to the lack of genomic resources. In this study, we report for the first time a high-quality, chromosome-level genome assembly of this species, aimed at elucidating its evolutionary history and the genetic basis of key traits. We constructed the genome using PacBio HiFi sequencing and further assembled it into 20 pseudochromosomes with the aid of Hi-C technology, yielding a final genome assembly size of 602.8 Mb with a contig N50 of 29.3 Mb and a BUSCO completeness of 98.7%. The assembly includes the identification of 20 pseudochromosomes and the annotation of 35,873 protein-coding genes, with an annotation rate of 96.4%. By integrating genomic data from other Malvaceae species, we analyzed the karyotype evolution of *S. lanceolata* and revealed the basal ploidy level of the family. Comparative genomic analyses uncovered significant syntenic relationships and whole-genome duplication (WGD) events among Malvaceae species, thereby clarifying the trajectory of karyotype evolution. Moreover, the study identified key regulatory gene families associated with fruit dehiscence (homologs of *SHP1/2*, *FUL*, *IND*, and *ALC*) that have undergone extensive expansion in *S. lanceolata* as a consequence of ancient polyploidy events. The reference genome provided in this study not only serves as a critical resource for evolutionary research in Malvaceae but also establishes a foundational framework for molecular breeding, genetic improvement, and conservation of *S. lanceolata* and related species.

Keywords *Sterculia lanceolata* · Malvaceae · Whole-genome duplication (WGD) · Karyotype evolution · Seed development · Fruit dehiscence · Comparative genomics

✉ Robert Henry
robert.henry@uq.edu.au

✉ Wenquan Wang
wangwenquan@itbb.org.cn

✉ Fei Chen
feichen@hainanu.edu.cn

¹ National Key Laboratory for Tropical Crop Breeding, College of breeding and multiplication (Sanya Institute of Breeding and Multiplication), Hainan University, Sanya 572025, China

² College of Horticulture, Bioinformatics Center, Academy for Advanced Interdisciplinary Studies, Nanjing Agricultural University, Nanjing 210095, China

³ Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, St Lucia, Brisbane, Australia

Introduction

Sterculia lanceolata, an important member of the genus *Sterculia* within the family Malvaceae, is a striking subtropical evergreen tree (Ma et al. 2016). It is renowned for its distinctive economic value and long-standing traditional medicinal uses (Prasad et al. 2022), with a broad geographic distribution primarily across subtropical regions of Asia, spanning southern China, Vietnam, Thailand, India, and beyond (Bawri et al. 2024). The species exhibits remarkable adaptability to the humid and variable climatic conditions characteristic of these regions. In addition to serving as a major regional source of timber, its roots, leaves, and seeds have been widely utilized in traditional medicine (Van Sam et al. 2008). They have been commonly applied to treat

symptoms such as pain, swelling, and bruising caused by trauma or blood stasis. Previous studies have shown that plants of the genus *Sterculia* are generally rich in flavonoids and their derivatives, compounds that have been demonstrated to possess diverse biological activities, including antibacterial, anti-inflammatory, antioxidant, and anticancer properties (Wang et al. 2024). Within forest ecosystems, *S. lanceolata* also plays an essential role in providing both food sources and habitats for a variety of organisms. Despite its pronounced ecological and economic importance, our understanding of its genetics remains extremely limited. To date, the NCBI public database contains only a single transcriptome dataset derived from its leaves (PRJNA435648), and no complete nuclear genome has yet been reported for any member of the genus *Sterculia* (Eum et al. 2019). This knowledge gap has posed a significant barrier to uncovering the complex molecular mechanisms underlying its physiological characteristics, medicinal value, and ecological adaptability (Ganie et al. 2015).

This knowledge gap highlights the urgency of conducting in-depth genetic studies on *S. lanceolata*, as such research is essential to deciphering the genetic codes underlying its remarkable adaptability and economically important traits. The Malvaceae family also encompasses globally significant crops such as cotton (*Gossypium raimondii*) and durian (*Durio zibethinus*), and important ornamental plant species such as *Hibiscus* (Cvetković et al. 2021). In addition, recent chromosome-level genomes and metabolomic analyses of tropical woody and liana species such as *Ficus altissima* and *F. hirta*, the invasive vine *Merremia boissiana* and the ornamental tree *Spathodea campanulata* have highlighted how high-quality genomes can illuminate evolutionary divergence, specialized metabolite biosynthesis and ecological adaptation in tropical plants (Zhang et al. 2025; Wang et al. 2025; Guo et al. 2025). Although the genomes of multiple cotton and durian species have been decoded, providing profound insights into key traits such as fiber development and flavor compound biosynthesis, investigations into the evolutionary origins and genetic foundations of its important sister genus, *Sterculia*, remain in their infancy. More importantly, for certain lineage-specific biological traits within Malvaceae—such as the molecular mechanisms regulating precise fruit dehiscence following maturation—there is still a lack of in-depth cross-species comparative research (Pabón-Mora et al. 2014).

Therefore, this study aims to construct a high-quality, chromosome-level reference genome of *S. lanceolata*. By conducting an in-depth analysis of its genomic background and evolutionary history, we seek to identify the key gene families that determine its unique fruit morphology. Compared with existing studies on the genus *Sterculia*, which have been largely limited to morphological descriptions and

transcriptome sequencing (Eom and Na 2019), this research provides, for the first time, a complete nuclear genome assembly, representing a major advancement. This will enable a deeper understanding of the molecular mechanisms underlying its ecological adaptability and valuable economic traits.

In addition, our findings bear direct significance for research across the entire Malvaceae family. Given the phylogenetic position of *S. lanceolata* within Malvaceae, our analysis of whole-genome duplication events and identification of key gene family expansions provide a valuable resource for comparative genomics. Specifically, the identification of critical genes associated with fruit dehiscence in *S. lanceolata* may offer novel genetic targets for the improvement of other related crops. Building on these insights, our study not only deepens the understanding of this economically important forest species but also provides essential scientific evidence for the evolution, genetic improvement, and conservation of Malvaceae.

Materials and methods

Experimental materials and sequencing technologies

Samples for sequencing of *S. lanceolata* were collected in the vicinity of the Innovation Research and Education Valley teaching area, Yazhou District, Sanya City, Hainan Province, China. DNA was extracted from young leaves, while samples for transcriptome analysis were obtained from various tissues of *S. lanceolata*, including roots, stems, functional leaves, and seeds. Samples were immediately flash-frozen in liquid nitrogen after collection and stored at -80°C in an ultra-low temperature freezer for subsequent nucleic acid extraction. Genomic DNA was randomly sheared, and short-read libraries of *S. lanceolata* were constructed following the TruSeq DNA sample preparation protocol. These libraries were then subjected to paired-end sequencing (150 bp read length) on the NovaSeq 6000 platform. In addition, a PacBio HiFi library was constructed. The library was prepared for sequencing using the PacBio Binding Kit (Wenger et al. 2019), in which primers and polymerases were ligated to the library. The final reaction products were purified with AMPure PB Beads and subsequently sequenced on the Revio sequencing system. Furthermore, Hi-C libraries were sequenced on the Illumina NovaSeq platform with paired-end reads of 150 bp. All transcriptome libraries were sequenced on the Illumina platform (Quail et al. 2008), generating paired-end reads of 150 bp. All sequencing data were processed using fastp v0.23.4 (Chen et al. 2018) to filter out low-quality reads and obtain clean data for downstream analyses.

Genome size estimation based on k-mer analysis

A k-mer analysis was performed for genome survey. To obtain clean reads, the raw data were filtered by removing low-quality reads, short reads, adapter sequences, and polyG tails. Subsequently, KMC v3.2.4 (Kokot et al. 2017) was employed to conduct k-mer frequency distribution analysis with $k=19$. Genome size, heterozygosity, and repeat content were estimated using GenomeScope v2.0 (Ranallo-Benavidez et al. 2020). The genome size of *S. lanceolata* was estimated to be 465.5 Mb, with a heterozygosity of 0.321%.

Chromosome-level genome assembly

First, the raw HiFi data were converted into FASTQ format using samtools v1.21 (Li et al. 2009). To obtain clean HiFi and Hi-C data, fastp v0.23.4 (Chen et al. 2018) was applied to filter out low-quality reads, short reads, adapter sequences, and polyG tails from the raw datasets. The filtered HiFi data were then assembled into contigs using hifiasm 0.19.9 (Cheng et al. 2021) with default parameters. For chromosome-scale scaffolding, Hi-C data processing was performed using the ALLHiC v0.9.13 pipeline (Zhang et al. 2019). Clean Hi-C reads were first aligned to the assembled contigs to generate a Hi-C interaction matrix, and ALLHiC was then used to cluster contigs into linkage groups and to order and orient them within each group according to Hi-C contact frequency, thereby building chromosome-scale scaffolds. The scaffold organization was further refined with 3D de novo Assembly (3D-DNA, v180114) (Dudchenko et al. 2017), and potential mis-joins were manually inspected and corrected based on the Hi-C contact maps visualized in Juicebox v1.11.08 (Durand et al. 2016). To assess assembly quality, we conducted a comprehensive evaluation using BUSCO v5.1 (Seppey et al. 2019) with the eudicots_odb10 database to examine the presence and completeness of single-copy orthologs. Ultimately, we obtained a high-quality, well-assembled *S. lanceolata* genome sequence, providing a solid foundation for future functional studies.

Gene prediction and annotation

We utilized RepeatModeler v2.0.3 (Flynn et al. 2020) to construct a *de novo* repeat library for the classification of repetitive sequences, and employed RepeatMasker v4.1.2 (Chen 2004) to identify them. In total, 376.7 Mb of repetitive elements were identified, constituting 59.6% of the *S. lanceolata* genome. All transcriptomic data were aligned to the genome using HISAT2 v2.2.1 (Kim et al. 2019). Subsequently, BRAKER v3.0.3 (Gabriel et al. 2024) was used to automatically train a species-specific parameter model and annotate gene structures, leveraging alignment evidence

from both transcriptomic data and proteins. For functional annotation, the predicted protein-coding genes were aligned against the eggNOG database (Hernández-Plaza et al. 2023). This pipeline assigns orthology-based functional information, including Gene Ontology (GO) terms, KEGG Orthology (KO) identifiers, and functional categories to each query protein, providing a comprehensive functional description for the *S. lanceolata* gene set.

Assembly and annotation of the chloroplast and mitochondrion genome

To assemble the chloroplast genome, we used the GetOrganelle v1.7.7.1 (Jin et al. 2020) software to assemble the Illumina paired-end sequencing data, specifying the $-k$ option with values set to 21, 45, 65, 85, and 105, respectively. For mitochondrial genome assembly, we utilized the Oatk v1.0 (Zhou et al. 2025) toolkit for *de novo* assembly from PacBio HiFi sequencing reads, invoking the syncasm module within Oatk to filter specific k-mer sequences from the raw HiFi reads. We applied $-k$ 1001 to set the minimum k-mer length and used $-c$ 150 to set the minimum coverage threshold.

The resulting FASTA file was used as the input for annotation on the OGDRAW platform (Lohse et al. 2007) (<https://chlorobox.mpimp-golm.mpg.de/OGDraw.html>). Following the platform's guidelines, we annotated the organellar genomes and chose the GenBank (GB) format for the output. To visualize the annotated genome, we accessed the "Upload" section of the OGDRAW online tool and submitted the GB file containing the *S. lanceolata* genome sequence. After selecting the appropriate options, we submitted the file for processing. OGDRAW then generated and displayed the annotated organellar genome in a graphical format, thus providing a clear representation of the organellar genetic structure.

Synteny and whole-genome duplication analysis

We selected the protein sequence data of *H. littoralis*, *F. kwangsiensis*, *F. hainanensis*, and *S. lanceolata* as the basis for our collinearity analysis. To ensure the reliability of the results, we performed both intra- and inter-genomic self- and cross-comparisons using the BLASTP v2.12.0 tool, with an E-value threshold set to $1e-5$. Subsequently, we utilized MCScanX v1.0.0 (Wang et al. 2012) to identify high-confidence collinear blocks. For the visualization of collinearity information, we used JCVI v1.5.1 (Tang et al. 2024), which effectively displays inter-chromosomal collinear regions. For the analysis of whole-genome duplication (WGD) events, we applied WGD v1.1.054 (Schranz et al. 2012) to calculate the WGD incidence rate and used the

WGDI v0.6.5 (Sun et al. 2022) software to plot the *Ks* value distribution, thereby estimating the timing and frequency of WGD events.

Polyploidy and karyotype analysis

We pre-processed the protein data for *S. lanceolata*, *D. zibethinus*, *G. raimondii*, and *H. schizopetalus* using BLAST v2.9.0–2 (Altschul et al. 1997), selecting the output file format “-outfmt 6” and “-num_alignment 2”. Subsequently, we constructed the ancestral Malvaceae karyotype (AMK) using WGDI v0.6.5 (Sun et al. 2022) with the “-km” option. Based on the protein sequence information, we inferred the modern karyotypes of the studied species. To quantify the fission and fusion events that occurred in the evolutionary history of these species, we enumerated all homologous segments and calculated the corresponding fission and fusion counts.

Identification and functional enrichment analysis of WGD genes

To identify genes in the *S. lanceolata* genome derived from whole-genome duplication (WGD), we first performed an all-versus-all BLASTP v2.12.0 (Lavigne et al. 2008) search of all protein sequences, setting an E-value threshold of 1e-5. Using these alignment results and gene location data, we detected collinear blocks and classified gene duplication types with the MCScanX v1.0.0 (Wang et al. 2012) toolkit. Subsequently, we conducted a Gene Ontology (GO) enrichment analysis on the identified set of WGD genes. We used the web-based tool agriGO v2.0 (Tian et al. 2017) (<https://systemsbiology.cau.edu.cn/agriGOv2/>) for the enrichment analysis. agriGO v2.0 performs gene enrichment using the Singular Enrichment Analysis (SEA) method, with the SEA parameters set as follows: Fisher’s exact test, Yekutieli correction (FDR under dependency), a significance level of 0.05, a minimum of 5 mapped genes, and analysis of the complete GO and molecular function. GO enrichment results were visualized as color-coded graphs and heatmap-like layouts using DataColor, which provides distinctive color mapping to highlight complex biological data relationships (He et al. 2024).

Phylogenetic analysis and divergence time estimation

We collected protein data for 39 species from public databases such as TropiCODB (Dai et al. 2025) (<https://bioinformatics.hainanu.edu.cn/tropdb/>) and PLAZA 5.0 (Van Bel et al. 2022) (<https://bioinformatics.psb.ugent.be/plaza/version>

[s/plaza_v5_dicots/](https://plaza_v5_dicots/)), focusing on plants from the Malvaceae family and the model plant *Arabidopsis thaliana*. Using OrthoFinder v2.5.5 (Emms and Kelly 2019), a tree-based method, we identified high-quality orthogroups among these species. The OrthoFinder output was optimized with TRIMAL (Capella-Gutiérrez et al. 2009), applying the -gt 0.6 option to remove short sequences within each orthogroup and using a -cons 60 threshold to retain only the most conserved sites. Subsequently, we performed a Maximum Likelihood (ML) analysis on the concatenated protein sequences of these species using the parallel version of RAXML v8.2.13 (Stamatakis 2014) (raxmlHPC-PTHREADS). This analysis aimed to infer the phylogenetic relationships in the context of *S. lanceolata*, thereby providing a clearer understanding of its evolutionary position relative to other plant species.

To estimate the divergence times among closely related species, including *H. littoralis*, *F. kwangsiensis*, *F. hainanensis*, *T. cacao*, *D. zibethinus*, *B. ceiba*, *G. hirsutum*, and *S. lanceolata*, we utilized the protein sequences of single-copy genes from Malvaceae plants. The divergence time estimation was conducted using the mcmctree v4.9 (Puttick 2019) tool in the PAML package. Following the analysis, the estimated divergence times were integrated into a phylogenetic tree reconstructed with IQTREE v2.4.0 (Minh et al. 2020), thereby providing a comprehensive understanding of the evolutionary relationships among the species.

Identification and phylogeny of key fruit dehiscence genes

To elucidate the molecular regulatory mechanism of follicle dehiscence along the ventral suture in *S. lanceolata*, we first selected five known key regulatory genes from the model plant *Arabidopsis thaliana*: SHP1 (AT3G58780), SHP2 (AT2G42830), FUL (AT5G60910), IND (AT4G00120), and ALC (AT5G67110). The proteome data used in this study included those of *S. lanceolata*, *Gossypium hirsutum* (upland cotton), *Durio zibethinus* (durian), and *Arabidopsis thaliana*. Based on the conserved protein domain features of the target genes in *A. thaliana*, we downloaded the Hidden Markov Model (HMM) file for the MADS-box domain from the Pfam database v35.0 (Mistry et al. 2021). Using the hmmsearch program within the HMMER package v3.3.2 (Finn et al. 2011) with default parameters, a search was conducted against the combined proteomes of the four aforementioned species.

To elucidate the evolutionary relationships among the members of each gene family, we constructed phylogenetic trees for each family separately. First, a multiple sequence alignment of the homologous protein sequences for each family was performed using MAFFT v7.525 (Katoh and

Standley 2013). Subsequently, based on the alignment results, phylogenetic trees were constructed using FastTree v2.1.11 (Price et al. 2010) with the WAG amino acid substitution model and the approximate maximum-likelihood method. Branch reliability was assessed based on Shimodaira-Hasegawa-like (SH-like) local support values. Finally, the resulting tree files were imported into the iTOL v6 (Letunic and Bork 2024) online platform for visualization and annotation.

Results

Chromosome-level genome assembly

To gain a deeper understanding of the genetic characteristics and evolutionary history of *S. lanceolata*, we employed advanced sequencing technologies to successfully construct a high-quality, chromosome-level reference genome. The final assembly, corresponding to the primary (collapsed) haploid assembly generated by hifiasm, has a total size of 602.8 Mb with a contig N50 of 29.3 Mb, indicating excellent continuity. Using high-throughput chromosome conformation capture (Hi-C) technology, the vast majority of sequences were successfully anchored to 20 pseudochromosomes, representing the $2n=40$ diploid genome (Fig. 1a). The Hi-C interaction heatmap clearly displayed strong diagonal signals and very weak inter-chromosomal interactions, strongly supporting the high accuracy and completeness of our genome assembly at the chromosome level (Fig. 1a). To evaluate the overall quality of the genome, k-mer analysis was performed. The k-mer analysis indicated that the *S. lanceolata* genome has low heterozygosity (0.321%) (Fig. 1b). Such low heterozygosity reduces allelic divergence and makes accurate separation of haplotypes more challenging, but it facilitates the construction of a highly contiguous primary assembly, which provides a solid foundation for subsequent high-quality genome assembly, annotation, and downstream analyses.

Genome completeness is a core metric for evaluating its quality. We assessed the completeness of the genome using BUSCO v5.1 (Seppey et al. 2019). The results showed that the BUSCO completeness of the genome assembly reached 98.7% (Fig. 1c), indicating that nearly all core conserved genes were fully represented. This high score demonstrates that the *S. lanceolata* genome we obtained is a high-quality, highly complete reference genome, sufficient to support downstream analyses such as comparative genomics, evolutionary studies, and functional gene mining. In addition, we successfully assembled the complete chloroplast (160,400 bp) and mitochondrial (161,707 bp)

genomes (Fig. 1e and f), providing additional layers of genetic information for species identification and phylogenetic research.

Repeat sequence characterization and genome annotation

Following repeat annotation using both *de novo* prediction and homology-based approaches, we found that *S. lanceolata* contains a substantial proportion of repetitive sequences, accounting for 59.59% of the genome. Among these, 42.46% are retrotransposons and 3.59% are DNA transposons, with long terminal repeats (LTRs) comprising 40.94% of the genome. The overall genomic GC content reached 32.71% (Fig. 1d). By integrating *de novo* prediction, homology-based alignment, and transcriptome annotation, we identified a total of 35,873 protein-coding genes, with a BUSCO v5.1 completeness of 96.4% for the genome (Table 1). Notably, the number of genes in *S. lanceolata* exceeds that of closely related Malvaceae subfamily species, including *H. littoralis* (29,154 genes) (He et al. 2022) and *S. hainanensis* (34,388 genes) (Dong et al. 2025), suggesting that *S. lanceolata* may encode a broader repertoire of stress-resistance and developmental genes. This provides a valuable genetic resource for horticultural breeding research, with the potential to introduce novel genes to enhance resistance and growth-related traits.

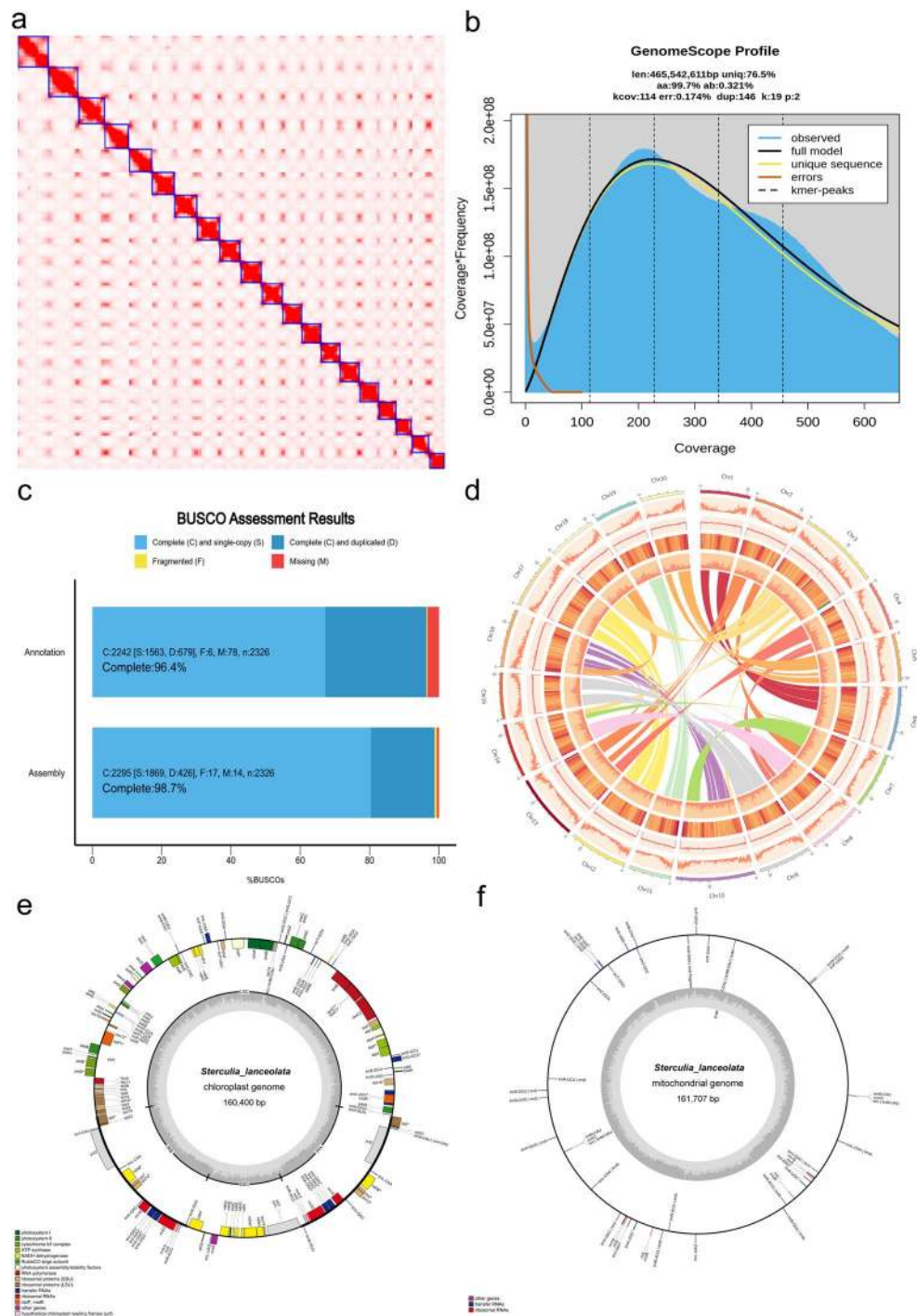
Ancient Whole-Genome duplications shaping evolutionary dynamics

Whole-genome duplication (WGD) or polyploidy events are key drivers of plant evolution, providing abundant genetic material that facilitates adaptation and species diversification. To investigate the evolutionary history of *S. lanceolata* and related species, we constructed a phylogenetic tree encompassing multiple subfamilies within Malvaceae (Fig. 2d). The tree clearly places *S. lanceolata* within the subfamily Sterculioideae and reveals well-defined phylogenetic relationships with other Malvaceae species.

Notably, we identified several ancient polyploidy events at key nodes of the phylogenetic tree. For example, on the branch of Helicteroideae, which includes durian (*Durio zibethinus*), we detected a pronounced whole-genome triplication (WGT) event (Fig. 2d). This finding is consistent with previous studies and helps explain the large genome size and complex genetic background of durian.

To elucidate the polyploidy history and chromosomal evolution of Malvaceae plants, we conducted a comparative genomics analysis of *S. lanceolata* and its closely related species. Within the *S. lanceolata* genome, the distribution

Fig. 1 Genomic features of *S. lanceolata*. **(a)** Hi-C heatmap of the genome assembly. **(b)** Distribution of 19-bp short reads. **(c)** Assessment of completeness for the genome assembly and gene set annotation. **(d)** Genomic features of *S. lanceolata* (chromosome karyotype; gene density; GC content per Mb; repeat content per Mb; long terminal repeat (LTR) distribution; identification of syntenic regions among different chromosomes). **(e)** Chloroplast genome of *S. lanceolata*. **(f)** Mitochondrial genome of *S. lanceolata*



of synonymous substitution rates (K_s) for paralogous gene pairs exhibits a pronounced peak at approximately 1.6, providing clear evidence for an ancient whole-genome duplication (WGD) event (Fig. 2a).

More importantly, analysis of K_s distributions of paralogous gene pairs clarified the polyploidization history of *S. lanceolata*. The K_s distribution shows a distinct peak at ~ 0.3 in *S. lanceolata* (Fig. 2a, red solid line). In combination with the WGD signals annotated on the species

phylogeny (Fig. 2d), this pattern indicates a lineage-specific whole-genome duplication in *S. lanceolata*. In addition, an older K_s peak at approximately 1.5–2.0 is shared among several Malvaceae species and corresponds to ancient polyploidy events at the family or deeper level. Inter-species synteny analysis (Fig. 2b) shows conserved collinear blocks and lineage-specific rearrangements among these genomes, which are consistent with the inferred WGD history.

Table 1 Genome features of *S. lanceolata*

Genome	<i>S. lanceolata</i>
Ploidy	2n=40
Assembled genome size (Mb)	602.8
Genomic heterozygosity (%)	0.321
Contig N50 (Mb)	29.3
Number of contigs	188
Number of pseudochromosomes	20
Repeat sequence content (%)	59.59
GC content (%)	32.71
Number of gene models	35,873
Genome BUSCOs (%)	98.7
Gene BUSCO (%)	96.4

Chromosomal rearrangements shaped the unique 20 chromosomes of *S. lanceolata*

Our reconstruction of karyotype evolution (Fig. 3) indicates that a common ancestor of Malvaceae possessed a karyotype with 11 chromosomes ($n=11$). Following this ancestral stage, a major divergence occurred. The lineage leading to Malvoideae (including *Gossypium raimondii* and *Hibiscus schizopetalus*) underwent an ancient whole-genome pentaplication (WGM) event. In contrast, the lineage leading to *S. lanceolata* within Sterculioideae experienced an independent whole-genome duplication (WGD) event. Following this WGD, the *S. lanceolata* genome underwent extensive chromosomal rearrangements, including reciprocal translocations (RTA), end-to-end joining (EEJ), fissions, and losses, ultimately stabilizing into its current karyotype of 20 chromosome pairs ($n=20$, $2n=40$).

For comparison, its close relative durian (*Durio zibethinus*, subfamily Helicteroideae) appears to have undergone another independent whole-genome triplication (WGT) event, followed by a series of complex fissions, nested chromosome fusions (NCFs), and other rearrangements, resulting in its modern karyotype of $2n=56$. *Gossypium raimondii* has maintained a relatively stable karyotype ($2n=26$) with few subsequent changes, whereas *Hibiscus schizopetalus* experienced at least two additional WGD events, followed by extensive chromosomal fissions and fusions, culminating in its current karyotype of $2n=42$. These diverse evolutionary trajectories vividly illustrate the substantial differences in genome structural evolution among lineages following WGD events, and it is precisely these differences that have shaped the rich and varied biological characteristics of Malvaceae species.

Dynamic changes in seed nutrient composition and gene accumulation driven by polyploidy

The seeds of *S. lanceolata* have attracted considerable attention due to their rich nutritional content and unique flavor.

We conducted a systematic study of the morphology and nutritional composition of *S. lanceolata* seeds during their development from immature to mature stages (Fig. 4). Morphologically, both the fruit (follicle) and seeds undergo significant color changes throughout development (Figs. 4e–g).

The results provided compelling evidence: among the WGD-retained genes, GO terms related to “Seed development” and “Starch metabolic process” were significantly enriched (Fig. 5a). This finding directly indicates that the WGD event in *S. lanceolata* did not randomly increase gene copy numbers, but rather led to a directed, large-scale expansion of gene repertoires associated with the synthesis and accumulation of seed nutrients. These expanded gene sets provide abundant raw material for the complex regulatory networks governing seed metabolism.

The metabolite profiling results provided direct evidence supporting this conclusion. We found that, as the seeds matured, the contents of several key mineral elements, including magnesium (Mg), phosphorus (P), and potassium (K), increased significantly (Fig. 5b), which is critical for energy storage in seeds and subsequent germination. Total starch content peaked during the mid-developmental stage and slightly declined at maturity, while amylose content also exhibited dynamic changes throughout seed development (Fig. 5f and g). Interestingly, we observed that flavonoid content was very high during the early stages of seed development but decreased sharply upon maturation (Fig. 5h). Flavonoids are generally associated with plant defense and bitterness, and their reduction is considered an evolutionary strategy to enhance seed palatability, attract animals for consumption, and facilitate dispersal (Cipollini and Levey 1997).

Expansion and mechanistic study of gene families related to fruit dehiscence

To validate this model and investigate the unique dehiscence mechanism in *S. lanceolata*, we conducted a systematic phylogenetic analysis of these key gene families (Fig. 6a–d). In the analyses of the *SHP1/2*, *IND*, *FUL*, and *ALC* gene families, we observed that, compared with the model plant *Arabidopsis thaliana*, *S. lanceolata* (denoted as “Slan” in the figures) exhibits a notable expansion in gene copy numbers across most of these families (Figs. 6a–c). For instance, in the phylogenetic tree of the *SHP1/2* gene family, multiple “Slan” genes cluster together, forming species-specific gene clusters (Fig. 6a). In addition, Fig. 6e provides a schematic summary of the regulatory relationships among *FUL*, *SHP1/2*, *IND* and *ALC* in the fruit valves and dehiscence zone.

Fruit dehiscence is a critical developmental process governed by a finely tuned gene regulatory network that ensures

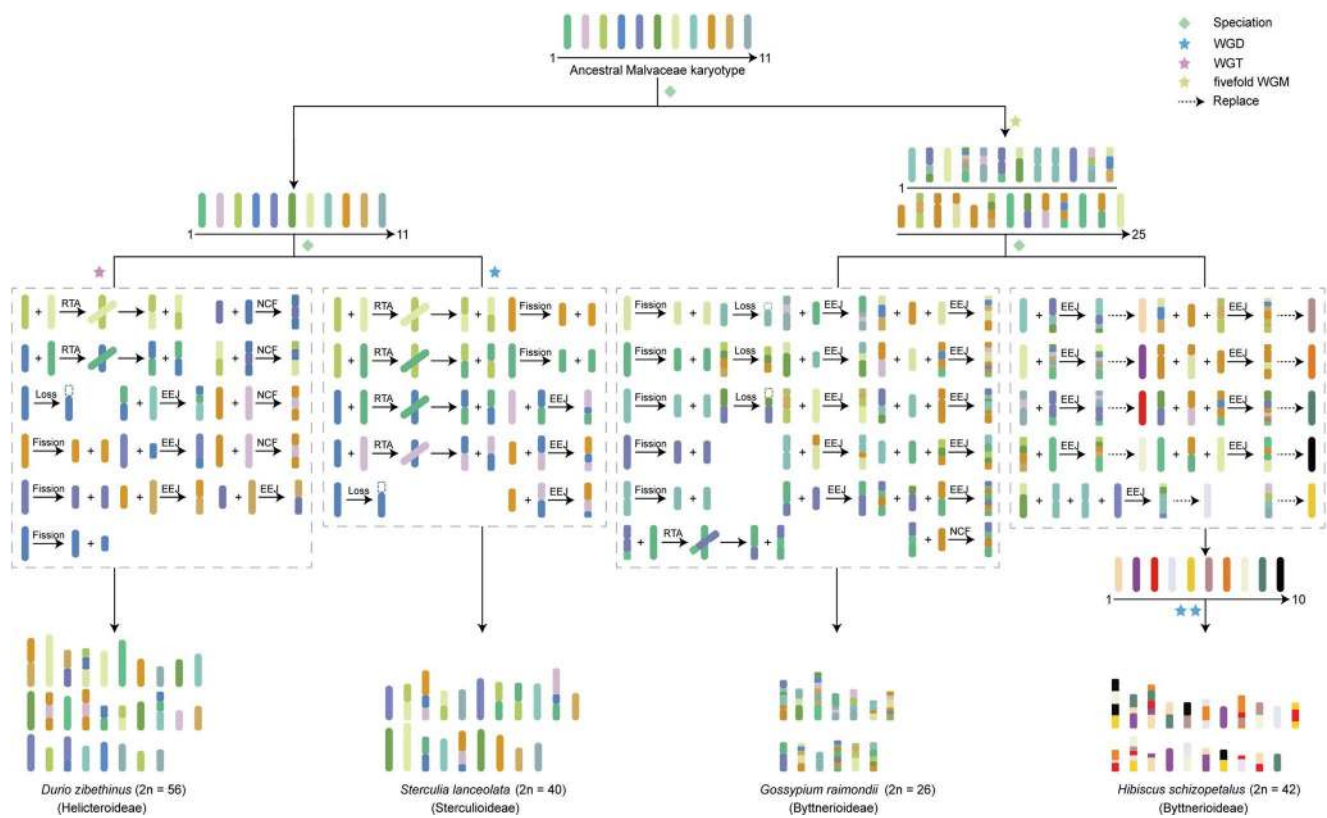
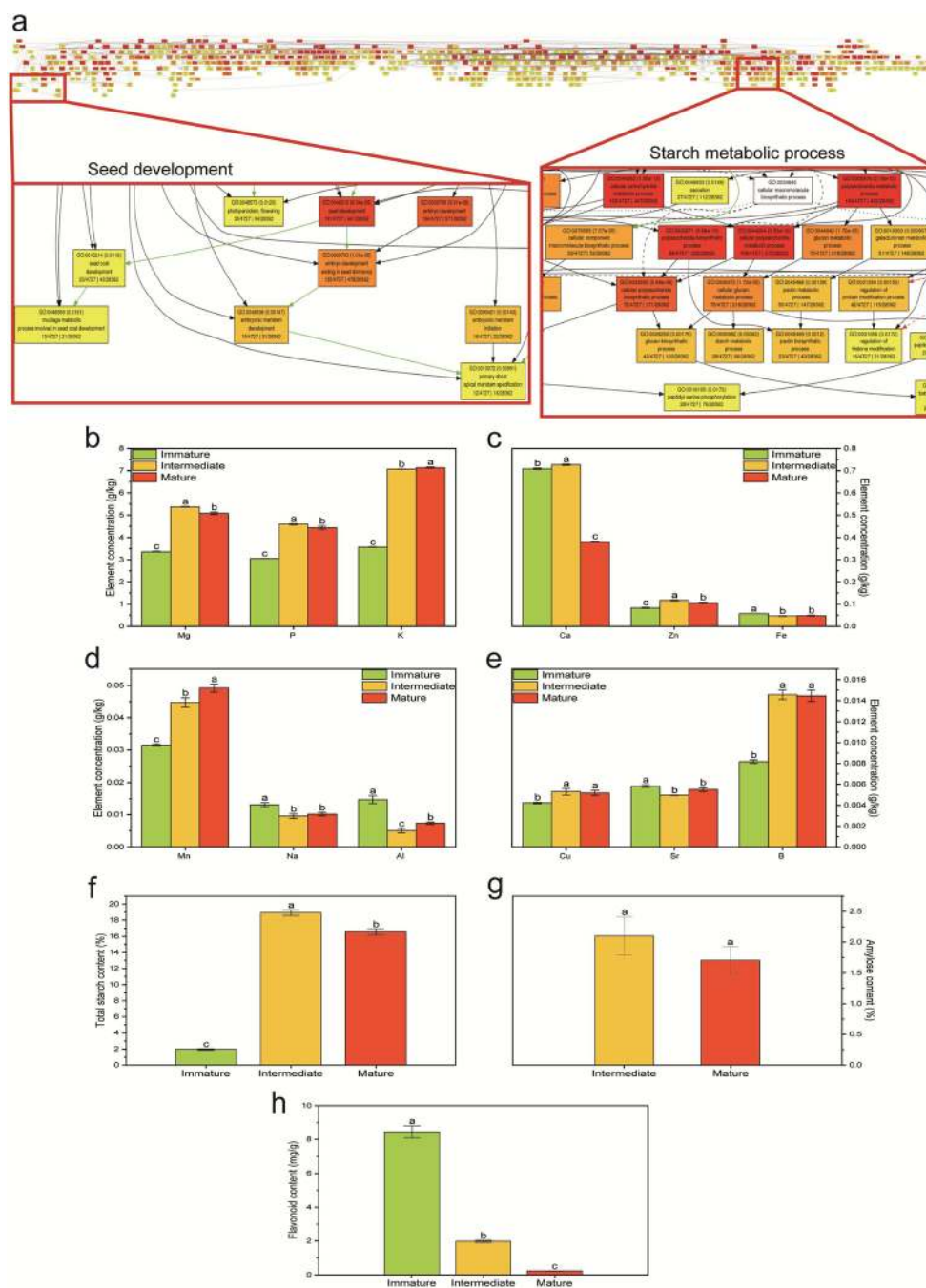


Fig. 3 Karyotype analysis of *S. lanceolata* and three other Malvaceae species

Fig. 4 Morphological features of *S. lanceolata*. (a) Morphology of a mature plant. (b–d) Follicles (fruits) at different developmental stages. (e–g) Seeds at different developmental stages. (h–j) Longitudinal sections of seeds at various developmental stages



Fig. 5 GO functional annotation and physiological–biochemical changes during seed development in *S. lanceolata*. **(a)** GO functional annotation of *S. lanceolata*, with seed development and starch metabolism highlighted within the red box. **(b–e)** Changes in mineral element contents in seeds at different developmental stages. **(f)** Changes in total starch content in seeds at various developmental stages. **(g)** Changes in amylose content in seeds at different developmental stages. **(h)** Changes in flavonoid content in seeds at different developmental stages

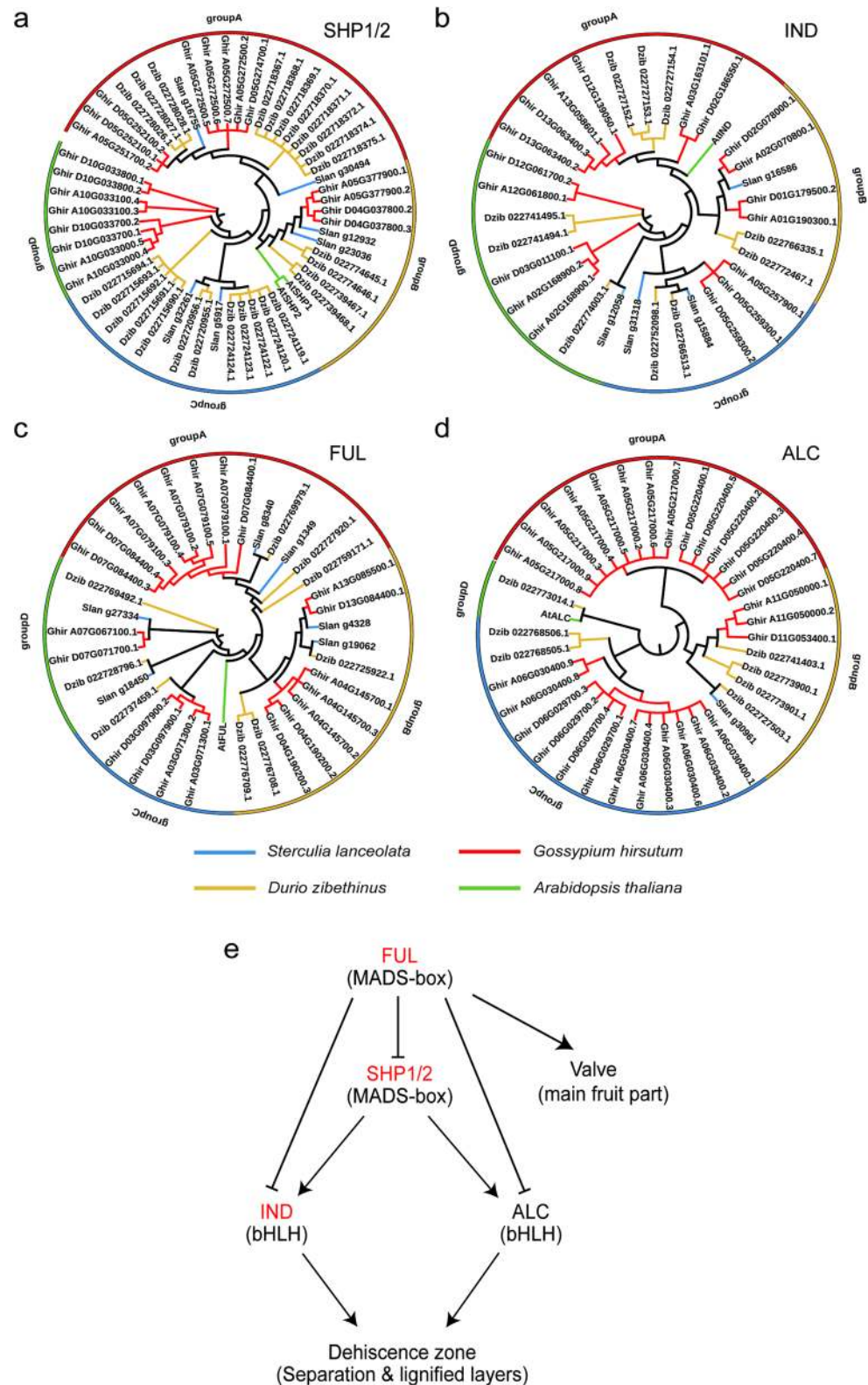


lignified layer, working in concert with *ALC* to complete the formation of the separation layer, whereas loss of *ALC* typically results in dehiscence defects. In contrast, the *MADS*-box gene *FUL* plays a dominant role in valve growth and development, restricting differentiation of the dehiscence zone by antagonizing the *SHP1/2–IND/ALC* pathway. Through the activation by *SHP1/2* and the antagonistic action of *FUL*, this regulatory network collectively establishes the spatial boundary between the valve and valve margin, thereby enabling controlled dehiscence of mature fruits along a predetermined interface.

Discussion

This study reports the first high-quality, chromosome-level reference genome for *S. lanceolata*, filling a critical gap in the genomics of the *Sterculia* genus within the Malvaceae family. This achievement not only establishes a solid genetic foundation for dissecting the species' unique economic and medicinal traits but also provides a new and crucial phylogenetic node for resolving the complex evolutionary history of the entire Malvaceae family. The assembled genome exhibits excellent continuity (Contig N50=29.3 Mb) and

Fig. 6 Phylogenetic relationships and regulatory pathway model of fruit dehiscence genes in *S. lanceolata*. (a–d) Maximum likelihood phylogenetic trees of the *SHP1/2*, *IND*, *FUL*, and *ALC* gene families. (e) Simplified model of the regulatory pathway controlling fruit dehiscence in *Arabidopsis thaliana*



high completeness (BUSCO=98.7%), with a quality sufficient to support sophisticated comparative genomics and functional gene discovery. High-quality, chromosome-level assemblies of this type are increasingly recognized as

indispensable resources for resolving plant genome structure, evolutionary history and trait-associated loci (Zhou et al. 2022). Furthermore, the genome's low heterozygosity offers critical insights into the species' reproductive biology,

population history, and provides a crucial baseline for future conservation genetics. This research marks a pivotal transition for the study of the *Sterculia* genus, advancing it from traditional morphology and limited transcriptomics into the whole-genome era.

The whole-genome duplication (WGD) events revealed in our study are a core driving force in the evolution of *S. lanceolata* and the entire Malvaceae family. Through *Ks* distribution analysis, we clearly identified two pivotal polyploidization events: an ancient WGD ($Ks \approx 1.6$) shared among Malvaceae species and a more recent, lineage-specific WGD ($Ks \approx 0.3$) within the *S. lanceolata* lineage. This recent WGD event is key to understanding the unique biological traits of *S. lanceolata*. Polyploidization provides abundant raw genetic material, enabling organisms to adapt to environmental changes and evolve novel functions, and ancient waves of polyploidization in flowering plants have been closely linked to facilitated adaptation to environmental stress (Wu et al. 2020; Zhang et al. 2020).

Compared to the whole-genome triplication (WGT) experienced by durian (*Durio zibethinus*) (Teh et al. 2017) and the pentaploidy (WGM) event in the cotton lineage (Wendel and Cronn 2002), the WGD in *S. lanceolata* demonstrates the diversity and complexity of polyploidization history within Malvaceae. Following a WGD, the genome typically undergoes drastic rearrangement and a process of diploidization. Our karyotype analysis reconstructed the evolutionary trajectory of *S. lanceolata* from a putative ancestral Malvaceae karyotype of $n=11$. This pathway involved a WGD event (theoretically forming $n=22$), followed by a series of complex chromosomal rearrangements including reciprocal translocations (RTA), fissions, and losses, which ultimately stabilized to form its characteristic $n=20$ ($2n=40$) karyotype. This model not only perfectly explains the origin of its modern chromosome number but also vividly illustrates the universal mechanisms of dynamic genome evolution post-polyploidization (Soltis et al. 2015).

One of the primary breakthroughs of this study is the direct link established between a macro-evolutionary WGD event and the species' key economic traits. We innovatively identified the set of genes that originated from the recent WGD and were subsequently retained, then performed a functional enrichment analysis on this specific gene set. The analysis revealed significant enrichment in the categories of 'seed development' and 'starch metabolic process.'

This finding provides strong evidence that gene retention post-WGD was non-random, instead favoring the selective expansion and preservation of genes related to nutrient accumulation in seeds. We then corroborated these genetic findings with metabolomic data, which demonstrated that as seeds mature, key mineral elements and total starch content increase significantly, while flavonoid compounds

associated with astringency decrease sharply. Therefore, the observed changes in flavonoid content in *S. lanceolata* seeds likely reflect an evolutionary compromise balancing the dual demands of "self-protection" and "attracting dispersers" during co-evolution. This forms a complete evidentiary chain—from a genomic evolutionary event to functional gene set expansion, and ultimately to the formation of a desirable economic trait—profoundly uncovering the evolutionary underpinnings of the high nutritional value of *S. lanceolata* seeds. This research paradigm, which tightly integrates WGD with economic traits, remains relatively rare in plant genomics and is analogous both to research on the expansion of starch-related genes in potato (Van Harsseelaar et al. 2017) and to work showing that an ancient WGD contributed to the diversification of flavor-related genes in the tea plant (*Camellia sinensis*) (Wang et al. 2021).

Similarly, the WGD event provides clues for elucidating another key biological trait in *S. lanceolata*: fruit dehiscence. The ability of a fruit to dehisce at the appropriate time is crucial for plant seed dispersal and propagation (Balanzà et al. 2016). In the classical Arabidopsis model, *FUL* is primarily expressed in the fruit valves and suppresses the formation of the dehiscence zone, whereas *SHP1/2*, *IND* and *ALC* act together to promote differentiation and lignification of the dehiscence zone, ultimately leading to fruit opening. Our study revealed that the core gene families regulating fruit dehiscence, including *SHP1/2*, *FUL*, *IND*, and *ALC*, all exhibit significant copy number expansion in the *S. lanceolata* genome. It is plausible that this WGD-driven increase in gene dosage led to a more complex and finely-tuned regulatory network. For instance, the expanded gene copies, potentially through subfunctionalization or neofunctionalization, may have enabled more precise control over the dehiscence zone and valve margins, thereby ensuring that the follicle dehisces efficiently and accurately along the ventral suture upon maturation (Lü et al. 2025). This discovery not only clarifies the developmental mechanisms in *S. lanceolata* but also offers potential homologous gene targets for the genetic improvement of other crops, such as preventing yield loss from pod shattering in rapeseed (*Brassica napus*) (Raman et al. 2014).

Conclusions

This study reports the first successful construction of a high-quality, chromosome-level reference genome for *S. lanceolata*, revealing its genomic structure, gene composition, and evolutionary features. Through in-depth comparative genomic analysis, we elucidated that *S. lanceolata* underwent an independent, recent whole-genome duplication (WGD) event and have reconstructed in detail its

evolutionary trajectory, which led to the formation of its current $n=20$ karyotype via large-scale chromosomal rearrangements post-WGD.

Crucially, this study establishes a direct link between this macro-scale genomic evolutionary event and the species' key agronomic traits. We demonstrate that the WGD-driven expansion of gene families directly resulted in the large-scale duplication and retention of key genes associated with 'seed development,' 'starch metabolism,' and 'pod dehiscence.' These changes at the genetic level are ultimately manifested at the phenotypic level as the rich accumulation of seed nutrients and a precise fruit dehiscence mechanism.

In conclusion, this research not only provides an invaluable resource for evolutionary and comparative genomics of the Malvaceae family but also establishes a solid molecular foundation for the genetic improvement, germplasm conservation, and exploitation of the medicinal value of *S. lanceolata*.

Acknowledgements We gratefully acknowledge the valuable comments and suggestions provided by the editor and anonymous reviewers.

Author contributions The authors confirm their contributions to the manuscript as follows: Conceptualization and project supervision: Wang W, Chen F, R. Henry; Plant material collection: Hu Y, Guo G, Zhang Y; Data analysis: Hu Y, Zhang Y, Guo G, Zhang H, Zhang J, Shao B, Xue J; Draft manuscript preparation: Hu Y, Zhang J, Yang K. All authors reviewed the results and approved the final version of the manuscript.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. This study was supported by the Startup Funding for the Double First-Class Discipline of Crop Science at Hainan University (RZ2100003362). Fei Chen is also supported by a grant from Hainan University (XTCX2022NYB04).

Data availability The raw genomic sequencing data and transcriptomic reads have been deposited in the National Genomics Data Center under accession number PRJCA028606. The genome assembly, annotation files, and protein sequences are available at FigShare (<https://doi.org/10.6084/m9.figshare.30632720>).

Declarations

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25(17):3389–3402. <https://doi.org/10.1093/nar/25.17.3389>
- Balanà V, Roig-Villanova I, Di Marzo M, Masiero S, Colombo L (2016) Seed abscission and fruit dehiscence required for seed dispersal rely on similar genetic networks. *Development* 143(18):3372–3381. <https://doi.org/10.1242/dev.135202>
- Bawri A, Baro D, Sarma K, Teron R, Tanti B (2024) *Sterculia lanceolata* Cav. var. *Lanceolata* (Malvaceae): an addition to flora of India. *Vegetos* 37(6):2642–2646. <https://doi.org/10.1007/s42535-023-00745-8>
- Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T (2009) TrimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25(15):1972–1973. <https://doi.org/10.1093/bioinformatics/btp348>
- Cheng H, Concepcion GT, Feng X, Zhang H, Li H (2021) Haplotype-resolved de Novo assembly using phased assembly graphs with hifiasm. *Nat Methods* 18(2):170–175. <https://doi.org/10.1038/s41592-020-01056-5>
- Chen N (2004) Using repeat masker to identify repetitive elements in genomic sequences. *Curr Protocols Bioinf* 5(1):41011–41014. <https://doi.org/10.1002/0471250953.bi0410s25>
- Chen S, Zhou Y, Chen Y, Gu J (2018) Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34:i884–i890. <https://doi.org/10.1093/bioinformatics/bty560>
- Cipollini ML, Levey DJ (1997) Secondary metabolites of fleshy vertebrate-dispersed fruits: adaptive hypotheses and implications for seed dispersal. *Am Nat* 150(3):346–372. <https://doi.org/10.1086/286069>
- Cvetković T, Areces-Berazain F, Hinsinger DD, Thomas DC, Wieringa JJ, Ganesan SK et al (2021) Phylogenomics resolves deep subfamilial relationships in malvaceae Sl. G3(7):jkab136 <https://doi.org/10.1093/g3journal/jkab136>
- Dai W, He S, Luo Y, Zhao C, Yang Z, Zhang J et al (2025) Tropi-CODB: A multi-omics resource for supporting biodesign in tropical crops. *BioDesign Res* 7(1):100003. <https://doi.org/10.1016/j.bidere.2025.100003>
- Dong Z, Jin S, Fan R, Sun P, Shao L, Zhao T et al (2025) High-quality genome of *firmitaria hainanensis* provides insights into the evolution of malvaceae subfamilies and the mechanism of their wood density formation. *J Genet Genomics* 52(6):812–825. <https://doi.org/10.1016/j.jgg.2024.12.009>
- Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC et al (2017) De Novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* 356(6333):92–95. <https://doi.org/10.1126/science.aal3327>
- Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES et al (2016) Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst* 3(1):99–101. <https://doi.org/10.1016/j.cels.2015.07.012>
- Emms DM, Kelly S (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* 20(1):238. <https://doi.org/10.1186/s13059-019-1832-y>
- Eom SH, Na J-K (2019) Leaf transcriptome data of two tropical medicinal plants: *sterculia lanceolata* and *Clausena excavata*. *Data Brief* 25:104297. <https://doi.org/10.1016/j.dib.2019.104297>
- Eum SM, Kim S-Y, Hong JS, Roy NS, Choi S, Paik J et al (2019) Transcriptome analysis and development of SSR markers of ethnobotanical plant *sterculia lanceolata*. *Tree Genet Genomes* 15(3):37. <https://doi.org/10.1007/s11295-019-1348-3>

- Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* 39(suppl2):W29–W37. <https://doi.org/10.1093/nar/gkr367>
- Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C et al (2020) RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci* 117(17):9451–9457. <https://doi.org/10.1073/pnas.1921046117>
- Gabriel L, Brūna T, Hoff KJ, Ebel M, Lomsadze A, Borodovsky M et al (2024) BRAKER3: fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Res* 34(5):769–777. <https://doi.org/10.1101/gr.278090.123>
- Ganie SH, Upadhyay P, Das S, Sharma MP (2015) Authentication of medicinal plants by DNA markers. *Plant Gene* 4:83–99. <https://doi.org/10.1016/j.plgene.2015.10.002>
- Guo G, García-Caparrós P, Zhang J, Zhang J, Li Y, He S et al (2025) Chromosomal reference genome of *Merremia boissiana*: unveiling the secrets of the tropical rainforest's killer plant. *Trop Plants* 4(1). <https://doi.org/10.48130/tp-0025-0007>
- Hernández-Plaza A, Szklarczyk D, Botas J, Cantalapiedra CP, Giner-Lamia J, Mende DR et al (2023) EggNOG 6.0: enabling comparative genomics across 12 535 organisms. *Nucleic Acids Res* 51(D1):D389–D394. <https://doi.org/10.1093/nar/gkac1022>
- He S, Dong W, Chen J, Zhang J, Lin W, Yang S et al (2024) DataColor: unveiling biological data relationships through distinctive color mapping. *Hortic Res* 11(2):uhad273. <https://doi.org/10.1093/hr/uhad273>
- He Z, Feng X, Chen Q, Li L, Li S, Han K et al (2022) Evolution of coastal forests based on a full set of Mangrove genomes. *Nat Ecol Evol* 6(6):738–749. <https://doi.org/10.1038/s41559-022-01744-9>
- Jin J-J, Yu W-B, Yang J-B, Song Y, DePamphilis CW, Yi T-S et al (2020) GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biol* 21(1):241. <https://doi.org/10.1186/s13059-020-02154-5>
- Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 30(4):772–780. <https://doi.org/10.1093/molbev/mst010>
- Kim D, Paggi JM, Park C, Bennett C, Salzberg SL (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* 37(8):907–915. <https://doi.org/10.1038/s41587-019-0201-4>
- Kokot M, Długosz M, Deorowicz S (2017) KMC 3: counting and manipulating k-mer statistics. *Bioinformatics* 33(17):2759–2761. <https://doi.org/10.1093/bioinformatics/btx304>
- Lavigne R, Seto D, Mahadevan P, Ackermann H-W, Kropinski AM (2008) Unifying classical and molecular taxonomic classification: analysis of the Podoviridae using BLASTP-based tools. *Res Microbiol* 159(5):406–414. <https://doi.org/10.1016/j.resmic.2008.03.005>
- Letunic I, Bork P (2024) Interactive tree of life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res* 52(W1):W78–W82. <https://doi.org/10.1093/nar/gkac268>
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N et al (2009) The sequence alignment/map format and samtools. *Bioinformatics* 25(16):2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
- Lohse M, Drechsel O, Bock R (2007) OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. *Curr Genet* 52(5):267–274. <https://doi.org/10.1007/s00294-007-0161-y>
- Lü T-F, Gao F, Su T-B, Dong Y (2025) Developmental genetics of fruit diversity in brassicaceae. *Curr Opin Plant Biol* 85:102707. <https://doi.org/10.1016/j.pbi.2025.102707>
- Ma C, Xiong Y, Li L, Guo D (2016) Root and leaf decomposition become decoupled over time: implications for below- and above-ground relationships. *Funct Ecol* 30(7):1239–1246. <https://doi.org/10.1111/1365-2435.12619>
- Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A et al (2020) IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol* 37(5):1530–1534. <https://doi.org/10.1093/molbev/msaa015>
- Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer EL et al (2021) Pfam: the protein families database in 2021. *Nucleic Acids Res* 49(D1):D412–D419. <https://doi.org/10.1093/nar/gkaa913>
- Pabón-Mora N, Wong GK-S, Ambrose BA (2014) Evolution of fruit development genes in flowering plants. *Front Plant Sci* 5:300. <https://doi.org/10.3389/fpls.2014.00300>
- Prasad N, Thombare N, Sharma SC, Kumar S (2022) Production, processing, properties and applications of Karaya (*Sterculia species*) gum. *Ind Crops Prod* 177:114467. <https://doi.org/10.1016/j.indcrop.2021.114467>
- Price MN, Dehal PS, Arkin AP (2010) FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5(3):e9490. <https://doi.org/10.1371/journal.pone.0009490>
- Puttick MN (2019) MCMCtreeR: functions to prepare MCMCtree analyses and visualize posterior ages on trees. *Bioinformatics* 35(24):5321–5322. <https://doi.org/10.1093/bioinformatics/btz554>
- Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R et al (2008) A large genome center's improvements to the illumina sequencing system. *Nat Methods* 5(12):1005–1010. <https://doi.org/10.1038/nmeth.1270>
- Raman H, Raman R, Kilian A, Detering F, Carling J, Coombes N et al (2014) Genome-wide delineation of natural variation for pod shatter resistance in brassica napus. *PLoS ONE* 9(7):e101673. <https://doi.org/10.1371/journal.pone.0101673>
- Ranallo-Benavidez TR, Jaron KS, Schatz MC (2020) GenomeScope 2.0 and smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun* 11(1):1432. <https://doi.org/10.1038/s41467-020-14998-3>
- Schranz ME, Mohammadin S, Edger PP (2012) Ancient whole genome duplications, novelty and diversification: the WGD radiation Lag-Time model. *Curr Opin Plant Biol* 15(2):147–153. <https://doi.org/10.1016/j.pbi.2012.03.011>
- Sepey M, Manni M, Zdobnov EM (2019) BUSCO: assessing genome assembly and annotation completeness. *Gene prediction: methods and protocols*. Springer, pp 227–245. <https://doi.org/10.1093/bioinformatics/btv351>
- Soltis PS, Marchant DB, Van de Peer Y, Soltis DE (2015) Polyploidy and genome evolution in plants. *Curr Opin Genet Dev* 35:119–125. <https://doi.org/10.1016/j.gde.2015.11.003>
- Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313. <https://doi.org/10.1093/bioinformatics/btu033>
- Sun P, Jiao B, Yang Y, Shan L, Li T, Li X et al (2022) WGDI: a user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes. *Mol Plant* 15(12):1841–1851. <https://doi.org/10.1016/j.molp.2022.10.018>
- Tang H, Krishnakumar V, Zeng X, Xu Z, Taranto A, Lomas JS et al (2024) JCVI: A versatile toolkit for comparative genomics analysis. *Imeta* 3(4):e211. <https://doi.org/10.1002/imt2.211>
- Teh BT, Lim K, Yong CH, Ng CCY, Rao SR, Rajasegaran V et al (2017) The draft genome of tropical fruit Durian (*Durio zibethinus*). *Nat Genet* 49(11):1633–1641. <https://doi.org/10.1038/ng.3972>
- Tian T, Liu Y, Yan H, You Q, Yi X, Du Z et al (2017) AgriGO v2. 0: a GO analysis toolkit for the agricultural community, 2017 update.

- Nucleic Acids Res 45(W1):W122–W129. <https://doi.org/10.1093/nar/gkx382>
- Van Bel M, Silvestri F, Weitz EM, Kreft L, Botzki A, Coppens F et al (2022) PLAZA 5.0: extending the scope and power of comparative and functional genomics in plants. *Nucleic Acids Res* 50(D1):D1468–D1474. <https://doi.org/10.1093/nar/gkab1024>
- Van Harselaar JK, Lorenz J, Senning M, Sonnewald U, Sonnewald S (2017) Genome-wide analysis of starch metabolism genes in potato (*Solanum tuberosum* L.). *BMC Genomics* 18(1):37. <https://doi.org/10.1186/s12864-016-3381-z>
- Van Sam H, Baas P, Keßler P (2008) Traditional medicinal plants in Ben En National park, Vietnam. *Blumea: Biodivers Evol Biogeogr Plants* 53(3):569–601. <https://doi.org/10.3767/000651908X607521>
- Wang L, Yan YC, Hao YM, Zhuo XX, Duan HT, Zhang Q et al (2024) A review on the phytochemistry and Pharmacology of the genus *sterculia*. *Chem Biodivers* 21(8):e202400963. <https://doi.org/10.1002/cbdv.202400963>
- Wang S, Guo G, Zhang J, Wang W, Chen F (2025) Chromosome-level genome of *Spathodea campanulata* revealed genetic characteristics of Malvidin and Cyanidin accumulation in the corolla. *Trop Plants* 4(1). <https://doi.org/10.48130/tp-0025-0004>
- Wang Y, Chen F, Ma Y, Zhang T, Sun P, Lan M et al (2021) An ancient whole-genome duplication event and its contribution to flavor compounds in the tea plant (*Camellia sinensis*). *Hortic Res* 8. <https://doi.org/10.1038/s41438-021-00613-z>
- Wang Y, Tang H, DeBarry JD, Tan X, Li J, Wang X et al (2012) MCS-canX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res* 40(7):e49–e49. <https://doi.org/10.1093/nar/gkr1293>
- Wendel JF, Cronn RC (2002) Polyploidy and the evolutionary history of cotton. *Adv Agron* 87:139–186
- Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT et al (2019) Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* 37(10):1155–1162. <https://doi.org/10.1038/s41587-019-0217-9>
- Wu S, Han B, Jiao Y (2020) Genetic contribution of paleopolyploidy to adaptive evolution in angiosperms. *Mol Plant* 13(1):59–71. <https://doi.org/10.1016/j.molp.2019.10.012>
- Zhang J, Li Z, Liang Y, He S, Guo G, Yang S et al (2024) Diploid wax Apple (*Syzygium samarangense*) genome identified NAC genes regulating fruit development. *Hortic Res* 11(3):uhae025. <https://doi.org/10.1093/hr/uhae025>
- Zhang J, Wang D, Guo G, Wang S, Zhang J, Liang Y et al (2025) Genomic and metabolomic analysis of *ficus altissima* and *ficus hirta*: evolutionary divergence and medicinal implications. *Trop Plants* 4(1). <https://doi.org/10.48130/tp-0025-0014>
- Zhang L, Wu S, Chang X, Wang X, Zhao Y, Xia Y et al (2020) The ancient wave of polyploidization events in flowering plants and their facilitated adaptation to environmental stress. *Plant Cell Environ* 43(12):2847–2856. <https://doi.org/10.1111/pce.13898>
- Zhang X, Zhang S, Zhao Q, Ming R, Tang H (2019) Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat Plants* 5(8):833–845. <https://doi.org/10.1038/s41477-019-0487-8>
- Zhou C, Brown M, Blaxter M, Consortium DT, o. LP, McCarthy SA, Durbin R (2025) Oatk: a de Novo assembly tool for complex plant organelle genomes. *Genome Biol* 26(1):235. <https://doi.org/10.1186/s13059-025-03676-6>
- Zhou Y, Zhang J, Xiong X, Cheng Z-M, Chen F (2022) De Novo assembly of plant complete genomes. *Trop Plants* 1(1):1–8. <https://doi.org/10.48130/TP-2022-0007>

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