Plant Biotechnology Journal
Open Access

Plant Biotechnology Journal (2024) **22**, pp. 1491–1503 doi: 10.1111/pbi.14281

The evolution history of an allotetraploid mangrove tree analysed with a new tool Allo4D

Yuan Wang[1], Yulong Li[1,2], Weihong Wu[1], Shao Shao[1], Qi Fang[1], Shaohua Xu[1,2], Zixiao Guo[1], Suhua Shi[1] and Ziwen He[1,*]

[1]*State Key Laboratory of Biocontrol and Guangdong Provincial Key Laboratory of Plant Resources, School of Life Sciences, Sun Yat-sen University, Guangzhou, Guangdong, China*

[2]*School of Ecology, Sun Yat-sen University, Guangzhou, Guangdong, China*

Received 8 June 2023;
revised 10 December 2023;
accepted 20 December 2023.
**Correspondence* (Tel +862084113677; fax +862084113677; email heziwen@mail.sysu.edu.cn

Keywords: mangrove, allotetraploid, subgenomes, adaptive evolution.

Summary

Mangrove species are broadly classified as true mangroves and mangrove associates. The latter are amphibious plants that can survive in the intertidal zone and reproduce naturally in terrestrial environments. Their widespread distribution and extensive adaptability make them ideal research materials for exploring adaptive evolution. In this study, we *de novo* assembled two genomes of mangrove associates (the allotetraploid *Barringtonia racemosa* ($2n = 4x = 52$) and diploid *Barringtonia asiatica* ($2n = 2x = 26$)) to investigate the role of allopolyploidy in the evolutionary history of mangrove species. We developed a new allotetraploid-dividing tool Allo4D to distinguish between allotetraploid scaffold-scale subgenomes and verified its accuracy and reliability using real and simulated data. According to the two subgenomes of allotetraploid *B. racemosa* divided using Allo4D, the allopolyploidization event was estimated to have occurred approximately one million years ago (Mya). We found that *B. racemosa*, *B. asiatica*, and *Diospyros lotus* shared a whole genome duplication (WGD) event during the K-Pg (Cretaceous-Paleozoic) period. K-Pg WGD and recent allopolyploidization events contributed to the speciation of *B. racemosa* and its adaptation to coastal habitats. We found that genes in the glucosinolates (GSLs) pathway, an essential pathway in response to various biotic and abiotic stresses, expanded rapidly in *B. racemosa* during polyploidization. In summary, this study provides a typical example of the adaptation of allopolyploid plants to extreme environmental conditions. The newly developed tool, Allo4D, can effectively divide allotetraploid subgenomes and explore the evolutionary history of polyploid plants, especially for species whose ancestors are unknown or extinct.

Introduction

Mangroves and mangrove associates grow within tropical and subtropical intertidal zones and face continuous environmental stress due to high salt, hypoxia, and UV radiation (Tomlinson, 2016). Although mangrove associates have not developed vivipary like some mangrove species, they have evolved several typical adaptive traits, including salt glands, well-developed aerial roots, and nondormant seeds (Quadros *et al.*, 2021). They grow from the intertidal zone to inland, a wider area than the true mangroves.

Whole genome duplication (WGD) or polyploidy refers to the doubling of chromosomes, including autopolyploidy (both copies are from the same parental species) and allopolyploidy (the copies are from different species). Polyploidy is an essential mechanism for speciation and genome evolution (de Peer *et al.*, 2017; Otto, 2007) and has been recognized as a powerful force for speciation and adaptive evolution (Adams and Wendel, 2005; de Peer *et al.*, 2017; Soltis and Soltis, 2016; Wood *et al.*, 2009). It frequently occurs in flowering plants, and recent evidence suggests that extant angiosperms experienced at least one round of WGD prior to differentiation (Jiao, 2018; Jiao *et al.*, 2011). The process of whole-genome duplication, chromosome fusion, deletion, and diploidization creates a wondrous cycle (Wendel, 2015). Plants undergo a series of genetic and epigenetic changes, such as DNA loss, transposon suppression, and homologous recombination during polyploidy, which leads to subfunctionalization, neofunctionalization, and novel regulatory interactions, providing an important genetic basis for species evolution (Doyle *et al.*, 2008). Previous studies have suggested that WGD played a crucial role in the colonization and diversification of Rhizophoreae ~70 Mya (Xu *et al.*, 2017) and that WGD contributed to the salt homeostasis of *Aegiceras corniculatum* ~35 Mya (Feng *et al.*, 2021). Similarly, environmental adaptation mechanisms (salt and waterlogging tolerance) are associated with polyploidization in *Hibiscus hamabo* (Wang *et al.*, 2022).

Related studies on mangroves, summarized in Additional file 1: Figure S1, focused on diploid and autopolyploid species. Therefore, the speciation and adaptive evolution of allopolyploid mangroves need to be further researched. In addition to the scarcity of polyploidy in mangroves, the assembly and division of allopolyploid genomes also pose challenges for this research, such as higher heterozygosity and a lack of progenitor species. Although there has been a proliferation of software for dividing polyploid subgenomes, such as DipHiC (Wu *et al.*, 2022), WGDI (Sun *et al.*, 2022), and SubPhaser (Jia *et al.*, 2022), there is still no suitable method to process scaffold-scale genome assembly.

The genus *Barringtonia* (Lecythidaceae), with approximately 40 species globally, is striking for its large, showy flowers and woody fruit capsules (Payens, 1967). They have a wide distribution range from the Pacific to the Marquesas Islands, mainly in the Malesian and Pacific regions. Many of these occur near rivers and lakes or in freshwater swamps and inundated areas (Prance, 2012). Among them, *B. racemosa* and its related species *B. asiatica* are the only mangroves and the most widespread species in this genus, extending from the Pacific to East Africa. They have typical characteristics of the genus *Barringtonia*, i.e., noticeable flowers and ovoid-cylindrical or pyramidal woody fruits (Additional file 1: Figure S2). Interestingly, we found *B. racemosa* is allotetraploid while *B. asiatica* is diploid. Therefore, they are excellent materials for studying the origin and evolution of allopolyploid mangroves.

In this study, we assembled high-quality genomes of *B. racemosa* and *B. asiatica* and developed a new tool to divide the subgenomes of allotetraploids for scaffold-scale genome assembly. We used comparative and phylogenetic genomic methods to solve the following scientific problems: (1) What is the speciation process of allotetraploid *B. racemosa*, and when did the two subgenomes diverge and fuse? (2) Have *B. racemosa* and *B. asiatica* experienced other recent WGD events, and have these polyploidization events contributed to intertidal adaptation? (3) How were the genes in the polyploid *B. racemosa* genome retained and lost after polyploidization, and whether this helped them adapt to the intertidal environment? Our study offers an essential tool for dividing other allotetraploid genomes into scaffold-scale genome assemblies. More importantly, this study provides a typical example of the origin and evolution of allopolyploid plants.

Results

Genome sequencing, assembly, and annotation

For *B. racemosa* and *B. asiatica* sequencing, we produced 151.43 Gb (~120× coverage) and 129.40 Gb (~164× coverage) of Pacbio long-read sequencing data, 92.61 Gb (~73× coverage) and 43.98 Gb (~56× coverage) of short reads of Illumina sequencing data (Additional file 2: Table S1), as a part of the world mangrove genomes sequencing project (He *et al.*, 2022). We also generated 135.48 Gb (~108× coverage) of Hi-C reads (Additional file 2: Table S1). The estimated genome sizes of *B. racemosa* and *B. asiatica* using the kmer spectrum method are consistent with the assembly sizes, 1.26 Gb and 0.79 Gb, respectively (Table 1). After mapping the short reads to the assembled genome, we calculated the heterozygosity of *B. racemosa* (0.134%) and *B. asiatica* (0.132%).

The final scaffold-scale genome assemblies contained 891 and 452 scaffolds of *B. racemosa* and *B. asiatica*, and the corresponding N50 lengths were 9.85 Mb and 9.73 Mb, respectively (Table 1). An assessment of the Benchmarking Universal Single Copy Orthologs (BUSCO) showed that the completeness of the two genomes was 98.8% and 95.2%, respectively (Additional file 2: Table S2). These results indicate that the genome assemblies are of sufficiently high quality for subsequent analyses. Based on the Hi-C data, we obtained a chromosome-scale genome assembly of *B. racemosa* with 26 pseudochromosomes that anchored 98% of the gene content (Figure S3). All strands of the chromosomes were adjusted in the same direction based on the collinearity between *B. racemosa* and *B. asiatica*.

Table 1 Genome assembly and annotation statistics for *B. racemosa* and *B. asiatica*

	B. racemosa	*B. asiatica*
Chromosome number	$2n = 4x = 52$	$2n = 2x = 26$
Estimated genome size (flow cytometry, Gb)	1.32	NA
Estimated genome size (kmer, Gb)	1.32	0.83
Assembled genome size (Gb)	1.26	0.79
Scaffold number	891	452
N50 length (Mb)	9.85	9.73
N50 count	40	25
Longest scaffold (Mb)	40.41	28.31
Number of scaffolds (≥1 Kb)	891	452
GC content (%)	37.10	37.14
Number of predicted genes	58 875	28 625
Average CDS length (bp)	1206.83	1190.05
Repeat sequences (%)	59.25	63.15

We predicted 58 875 and 28 625 genes in *B. racemosa* and *B. asiatica*, and the average lengths of their coding sequences were 1207 and 1190 bp, respectively (Table 1). Transposable element annotation showed 59.25% and 63.15% repetitive sequences, with total lengths of 748 519 825 bp and 497 276 499 bp, respectively. Long-terminal repeat retrotransposons accounted for the most significant proportion (46.61% and 52.54% in *B. racemosa* and *B. asiatica*, respectively, Additional file 2: Table S3).

In addition to doubling the chromosome number (*B. racemosa*, $2n = 52$; *B. asiatica*, $2n = 26$) (Kowal, 1989; Morawetz, 1986), the estimated genome size of *B. racemosa* (1.32 Gb) was much larger than that of *B. asiatica* (0.83 Gb) (Table 1). Each chromosome of *B. asiatica* corresponded to two chromosomes of *B. racemosa* (Figure 1a). These results suggest that *B. racemosa* is a tetraploid species. GenomeScope plot showed that 'aaab' genotype was estimated to be ~0 while 'aabb' genotype was estimated to be >7% (Figure 1b), which was determined as allotetraploid. The kmer spectra of *B. racemosa* show a prominent 2× peak based on the result of Tetmer (Figure 1c), clearly indicating allotetraploidy.

Development and testing of Allo4D

Distinguishing the subgenomes of *B. racemosa* is essential for subsequent analyses of its origin and evolution. However, in previous studies, the methods for identifying subgenomes were based only on chromosome-scale genome assembly and were mostly combined with Hi-C interaction signal results. This is unsuitable for our scaffold-scale genome assembly. In fact, these methods require higher sequencing depth and cost, such as 615× genome data of octoploid strawberry, 48× coverage of PacBio HiFi reads, and ~100× Hi-C data of polyploid *Echinochloa* (Edger *et al.*, 2019; Wu *et al.*, 2022). The recently published WGDI and SubPhaser pipelines were used for allopolyploid subgenome phasing and required chromosome-scale genome assembly (Jia *et al.*, 2022; Sun *et al.*, 2022). Therefore, developing a subgenome-distinguishing tool for scaffold-scale genome assembly is an urgent need.

We mainly applied the collinear and evolutionary relationships between allotetraploid and relatively diploid species to distinguish subgenomes. This pipeline has four major steps (Figure 2): (1) align the genome data and obtain a collinear relationship. Genes

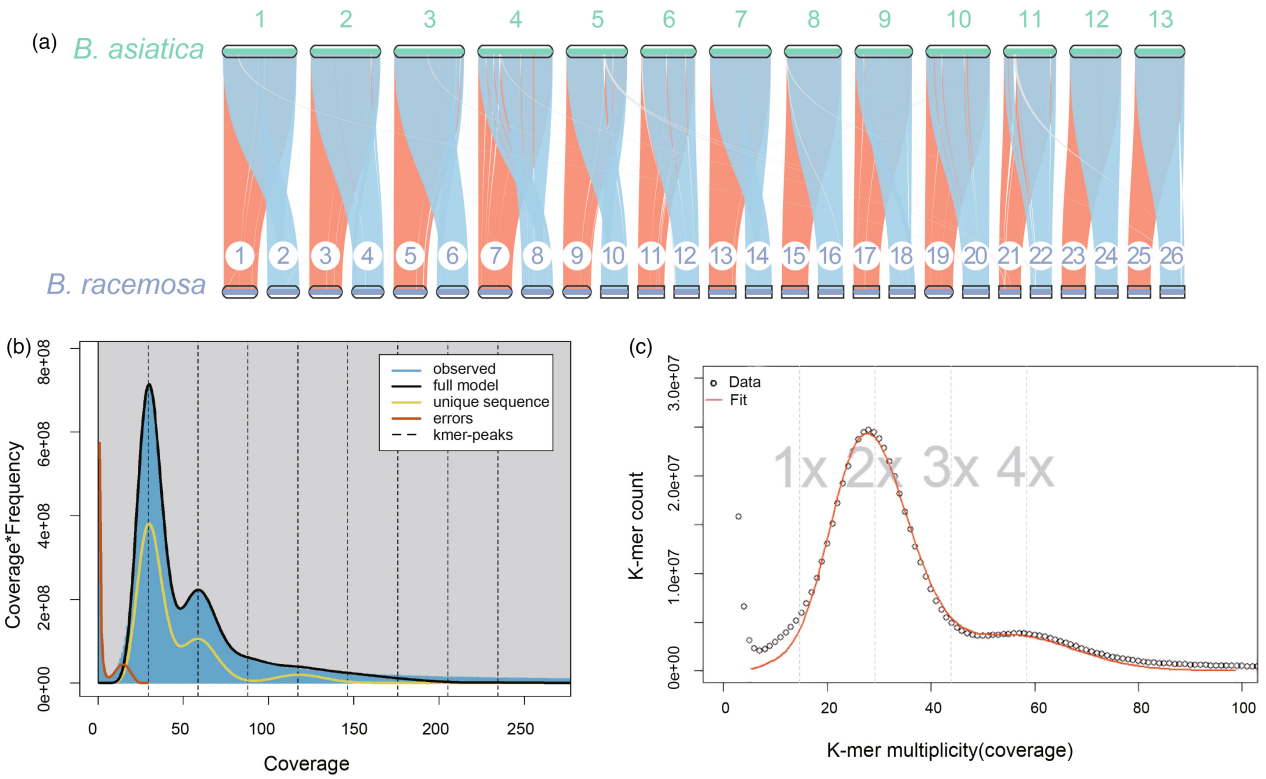


Figure 1 *Barringtonia racemosa* is an allotetraploid. (a) The collinearity between *B. racemosa* and *B. asiatica*. Red represents subgenome A, and blue represents subgenome B. (b) GenomeScope plots for *B. racemosa* (k = 19). The ratio of genotypes was calculated (aaaa: 86.9%, aaab: 0.001%, aabb: 7.44%, aabc: 0.001%, abcd: 5.67%). (c) Schematic of the shapes of the kmer spectra of *B. racemosa*, and the best-fit peak is 2×.

generated by ancient duplication events were removed during this process. (2) Select replicated gene pairs (two copies in *B. racemosa* and one copy in *B. asiatica*) and combine blocks into clusters according to the gene position in *B. asiatica*; however, only clusters with more than five collinear pairs were retained. (3) Construct phylogenetic trees within clusters and identify the optimal tree based on evolutionary relationships. We defined the gene clustered with diploidy as SubA gene, while the others were defined SubB gene. (4) Divide subgenomes using the SubA and SubB genes as markers. All of the Allo4D code was uploaded to GitHub (https://github.com/yuanw-18/Allo4D).

There were two evolutionary hypotheses regarding generating the phylogenetic trees in Step 3. Hypothesis 1 is one of the genes in the allotetraploid cluster with diploidy (Topology 1), whereas Hypothesis 2 is that the two copies within the allotetraploid cluster are together (Topology 2) (Figure 2d). We chose the optimal phylogenetic tree according to gene number and bootstrap support values, and only the genes from the optimal phylogenetic tree were used in the subsequent analysis.

We used the tool on real allotetraploid *Arabidopsis suecica* and simulated allotetraploid rice genome data to evaluate the accuracy and reliability of Allo4D, as detailed in the Methods section. Three crucial metrics (accuracy, precision, and recall) were used to assess pipeline performance. When the average genome length exceeded 3 Mb, the three metrics in both *Arabidopsis suecica* and rice genomes exceeded 90% (Additional file 1: Figures S4 and S5, Additional file 2: Tables S4 and S5). The performance of this pipeline fluctuated to a certain extent for the fragmented genome (below 3 Mb); however, each metric exceeded 75% (Additional file 1: Figures S4 and S5, Additional file 2: Tables S4 and S5). These results demonstrated the accuracy and reliability of our pipeline for distinguishing scaffold-scale allotetraploid subgenomes.

Division of the *B. racemosa* subgenome

We used the Allo4D tool to divide the subgenomes of *B. racemosa*. A total of 148 clusters supported Hypothesis 1, whereas only 13 clusters supported Hypothesis 2. In addition, the number of gene pairs in Topology 1 was significantly higher than in Topology 2, and trees with support values greater than 80% exceeded 90% in Topology 1, whereas it was less than 50% in Topology 2 (Additional file 1: Figure S6b,c). These results support Hypothesis 1, suggesting that the evolutionary relationship between *B. racemosa* and *B. asiatica* is consistent with Topology 1 and that *B. racemosa* originates from two ancestral diploid species. One of these is more closely related to *B. asiatica*, and the existing *B. racemosa* is derived from the hybridization between these two ancestral species. Therefore, we only used data from Topology 1 for subsequent analysis. Using Allo4D, 92.27% of the genes were divided into subgenomes A and B, accounting for 91.27% of the total genome (Additional file 2: Table S6, and Additional file 1: Figure S7). We also compared the subgenome phasing results between the scaffold and chromosome-scale genome assemblies (Additional file 2: Tables S7 and S8). All phased scaffolds were included in the phased chromosome results, and scaffold dividing results accounted for 92.17% of the chromosome dividing results, suggesting that scaffold-scale genome assembly can also distinguish subgenomes clearly.

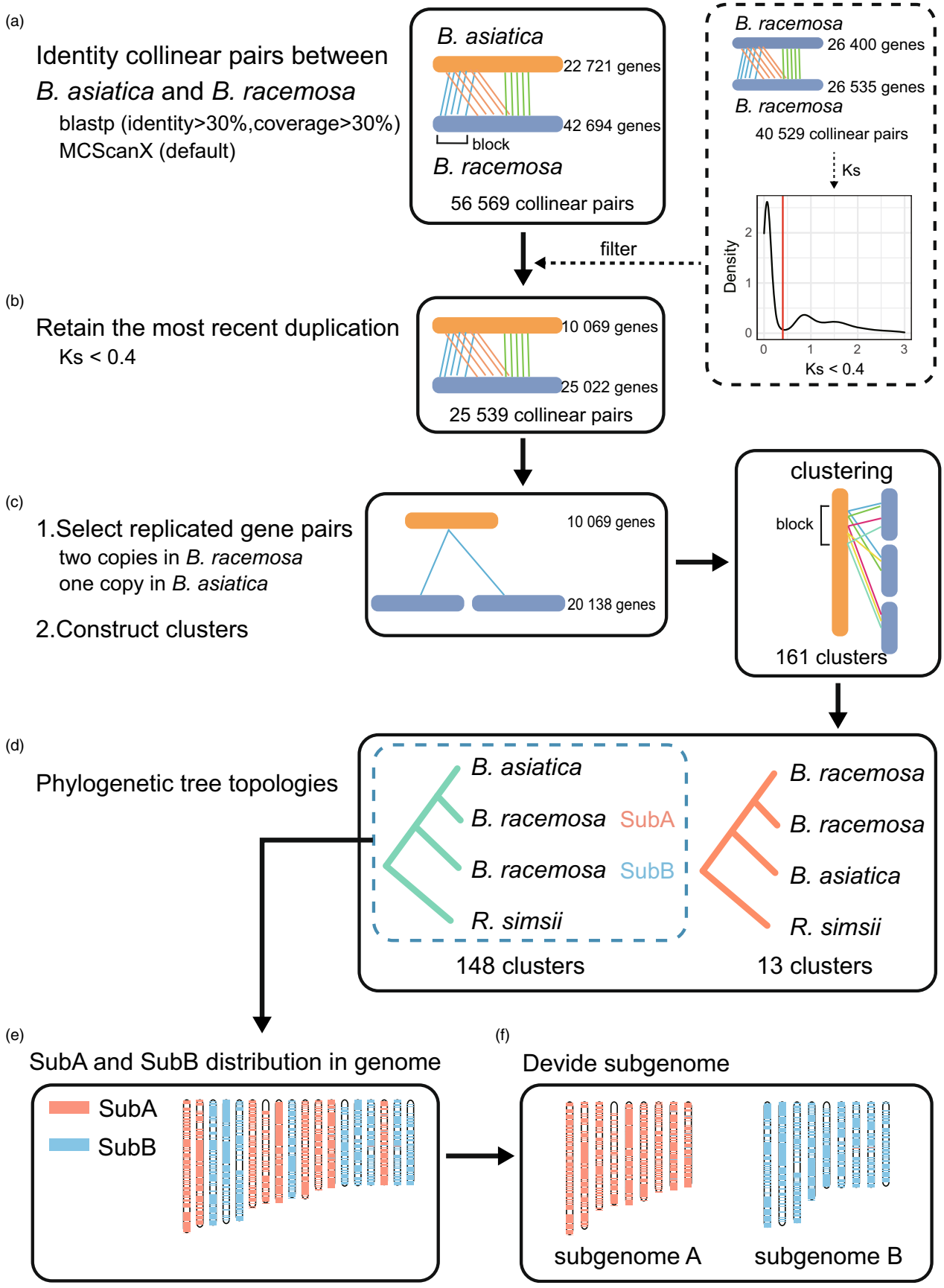


Figure 2 Pipeline of subgenome identification using Allo4D. (a) Identify collinear pairs between *B. asiatica* and *B. racemosa*. The orange bar represents *B. asiatica*, and the blue bar represents *B. racemosa*. (b) Retain the most recent duplication. (c) Select replicated gene pairs and construct clusters. (d) Analyse the phylogenetic tree in each cluster. (e) Distribution of SubA and SubB genes in *B. racemosa* genome. Red represents SubA genes, and blue represents SubB genes. (f) Division of *B. racemosa*.

Evolutionary history of *B. racemosa*

The phylogenetic relationships among *B. racemosa*, *B. asiatica*, and other related species were reconstructed to determine their phylogenetic position. We obtained 321 single-copy orthogroups from 11 species and constructed a phylogenetic tree using RAxML-NG with the GTR + F + R5 model. The subgenome A diverged from *B. asiatica* around 5.59 Mya, and subgenome B separated from *B. asiatica* at approximately 7.8 Mya (Figure 3a). We estimated that the allotetraploidization event of *B. racemosa* was 1 Mya based on the transposable element (TE) divergence rates between the two subgenomes (Additional file 1: Figure S8).

The results of the synteny analysis and Ks distributions suggest that *B. racemosa* has recently undergone two whole-genome duplication events, whereas *B. asiatica* has experienced one duplication event in the last few million years. We detected 2 : 1 syntenic relationships between *D. lotus* and *V. vinifera* and *B. asiatica* and *V. vinifera*, and a 4 : 1 syntenic relationship between *B. racemosa* and *V. vinifera* (Figure 3c). The Ks distributions also provide evidence for one shared WGD event among *D. lotus*, *B. asiatica*, and *B. racemosa* (Ks range = 0.65–1.05, time around K-Pg (Cretaceous-Paleogen boundary) (Akagi *et al.*, 2020)), and *B. racemosa* experienced one specific WGD which is the allopolyploidization process (Ks range = 0.03–0.10) (Figure 3d).

Our study indicated that WGD events were accompanied by extreme climates. It occurred approximately 66 million years ago at the K-Pg boundary with global cooling and darkness (Nichols and Johnson, 2008). Similarly, the allopolyploidization event (~1 Mya) underwent a global average surface temperature (GAST) and sea surface temperature (SST) reconstruction period. The global temperature gradually cooled until approximately one million years ago, after which cooling stalled (Martínez-Garcia *et al.*, 2010; Snyder, 2016). Whole-genome duplication events under climate-change conditions may play an important role in adaptation and speciation.

We found that the common ancestors of *B. racemosa* and *B. asiatica* diverged into species A (subgenome A ancestor) and

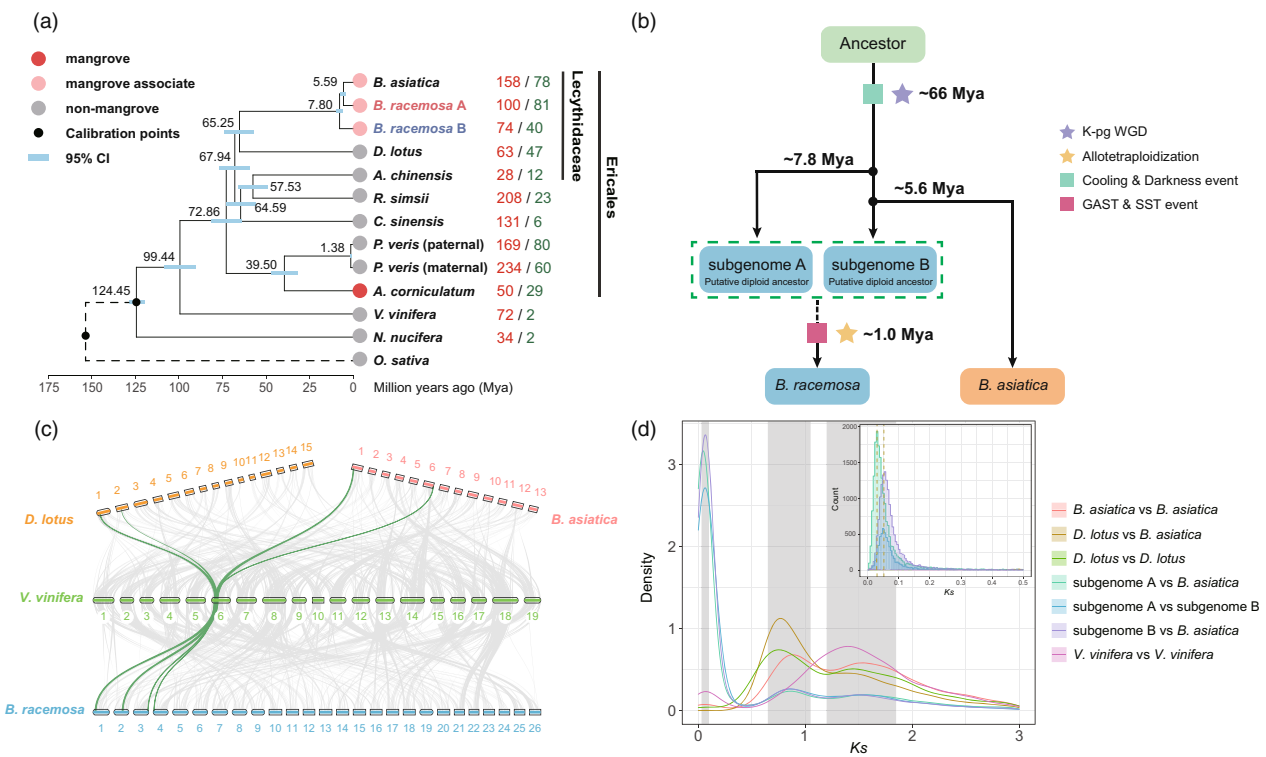
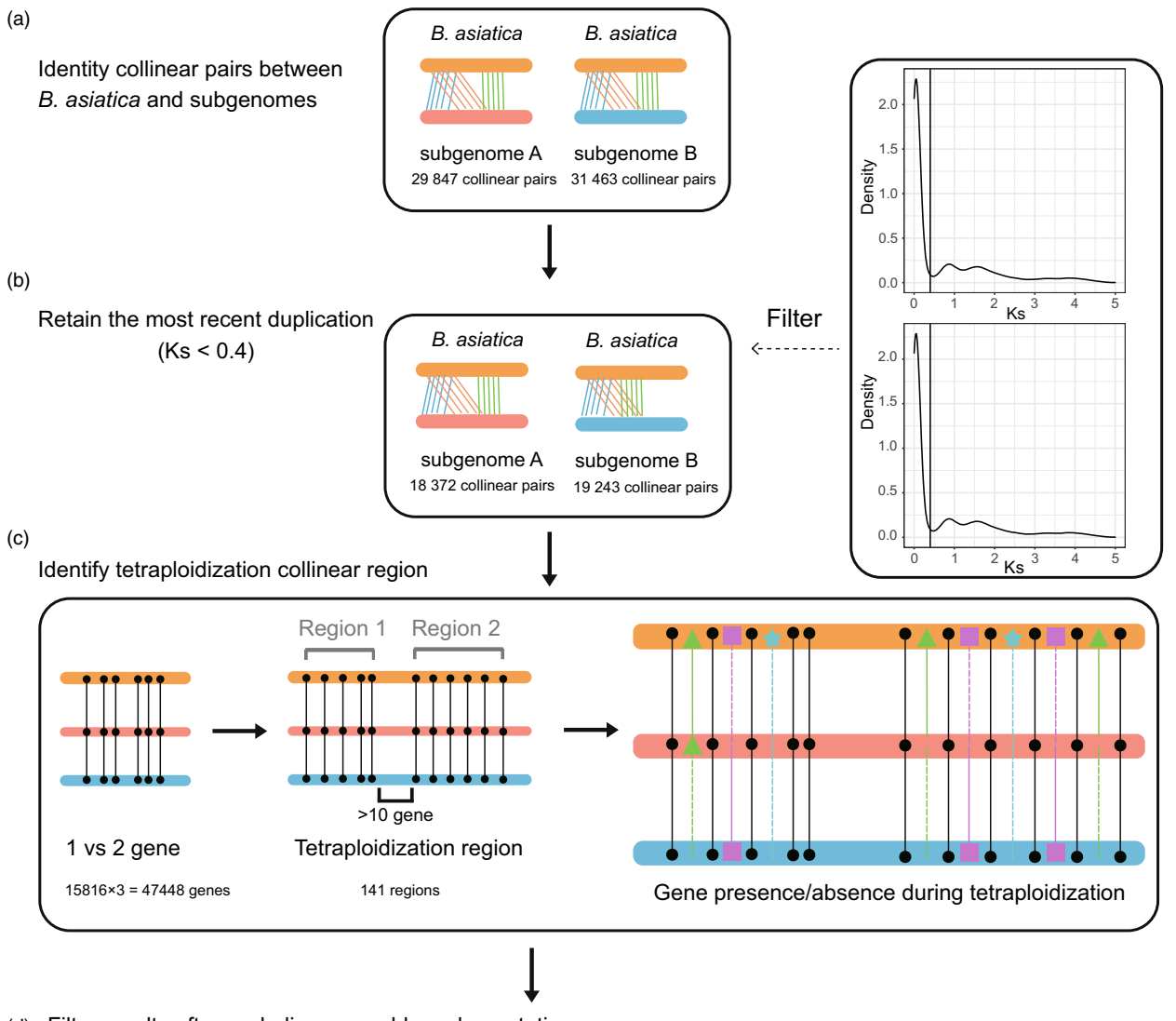


Figure 3 Genome evolutionary history. (a) Phylogenetic relationship among *B. racemosa*, *B. asiatica*, and other related species. Blue node bars represent 95% confidence intervals. The black nodes are two fossil calibration points. The circles of different colours indicate the type of species divided (dark red: mangrove; light red: mangrove associates; grey: non-mangrove). The numbers in red and green indicate the numbers of gene families that have expanded and contracted rapidly, respectively. (b) The evolutionary model between *B. asiatica* and *B. racemosa*. GAST means global average surface temperature reconstruction event; SST means sea surface temperature reconstruction event. (c) Syntenic relationships among *D. lotus*, *B. asiatica*, and *B. racemosa* relative to *V. vinifera*. One of the syntenic blocks among the four species is shown in green. (d) The distribution of synonymous substitution rate (Ks) of orthologues and paralogues among *D. lotus*, *B. asiatica*, *B. racemosa*, and *V. vinifera* is shown. The shades of grey represent three duplication events in *B. racemosa* (0.03–0.10; 0.65–1.05; and 1.20–1.85). The Ks details of *B. racemosa* are enlarged on the upper right.

species D (putative ancestor) ~7.8 million years ago. Species D was then separated into species B (subgenome B ancestor) and *B. asiatica* ~5.6 million years ago. The allotetraploid *B. racemosa* was produced by hybridizing species A and B around one million years ago (Figure 3b).

Genome reconstruction in *B. racemosa*

We investigated gene retention and loss in *B. racemosa* to explore the relationship between patterns of genome reconstruction after allopolyploidization and environmental adaptation. The presence and absence of genes in two subgenomes of *B. racemosa* were determined using the *B. asiatica* genome as a reference (Figure 4). There were 18 372 and 19 243 collinear pairs in the subgenomes A and B, respectively. We identified 141 allotetraploidization regions, including four patterns of presence and absence (two copies in *B. racemosa* were retained or lost, and one of the genes was lost in subgenome A or subgenome B). After excluding assembly and annotation errors, the percentage of two



(d) Filter results after excluding assembly and annotation errors

Group	Number	Percentage
●	15749	55.02%
■	205	0.72%
▲	303	1.06%
★	734	2.56%

- ● Gene presence both in subgenome A & subgenome B
- ■ Gene absence in subgenome A
- ▲ Gene absence in subgenome B
- ★ Gene absence in *B. racemosa*

Figure 4 Pipeline identifying present and absent genes. (a) Identify collinear pairs between *B. asiatica* and the two subgenomes. The orange bar represents *B. asiatica*, and the red and blue bars represent the two subgenomes. (b) Retain the most recent duplication. (c) Identify present and absent genes. (d) The statistics of the finalized present and absent genes after filtering.

homologous genes both present in *B. racemosa* was the highest (55.02%), whereas only 2.56% of the homologous genes were absent. The homologous genes absent in subgenome A accounted for 0.72%, while that accounted 1.06% in subgenome B.

The selection pressure for these gene pairs was calculated to determine the evolutionary rates of the present and absent genes. We found that positive selected genes (Ka/Ks >1, *P*-value <0.05) were primarily associated with resistance (Additional file 2: Table S9). Surprisingly, the homologous genes that were only present in subgenome A were enriched in transferase activity, response to stimulus, and organelle organization pathway. In contrast, homologous genes that were only present in subgenome B were enriched in ATPase activity, hormone biosynthetic processes, and lipid metabolic pathways. In addition, genes present in both subgenomes were primarily enriched in the nucleocytoplasmic fusion and reproduction-related pathways (Additional file 1: Figure S9).

Expansion of glucosinolate-related genes in *B. racemosa*

Expanding gene families and genes retained by whole-genome duplication play essential roles in the adaptive evolution of plants, such as cold, heat, and salt stress (Feng *et al.*, 2023; Wu *et al.*, 2020). Enrichment analysis of the rapidly expanding gene families in *B. racemosa* and *B. asiatica* revealed that many genes were enriched in the glucosinolates (GSLs) pathway. Other genes were enriched in salt stress, response to external stimuli, and ion transport pathways (Figure 5a,b). GSLs are essential for plant defence and responses to biotic and abiotic stressors. It also has been the focus of recent studies due to its positive tumour treatment effects (Aghajanzadeh *et al.*, 2018; del Carmen Martínez-Ballesta *et al.*, 2013; Justen and Fritz, 2013; Sønderby *et al.*, 2010; Soundararajan and Kim, 2018; Vig *et al.*, 2009). Enrichment analysis of the genes retained by WGD in *B. racemosa* and *B. asiatica* showed that these genes were enriched in insect resistance, energy metabolism, and ion transport pathways (Additional file 1: Figure S10), which is consistent with the function of GSLs.

A. thaliana GSLs-related genes were searched in the other eight species of Ericales, and the number of related genes was counted (Additional file 2: Table S10). In the core structure pathway, the number of *SOT* (*SOT16*, *SOT17*, and *SOT18*) genes in *B. racemosa* was higher than that in other species (Figure 5c). The *SOT* gene family is divided into eight groups and 21 genes in *A. thaliana* (Khan *et al.*, 2008), which are involved in insect response, stress resistance, and plant growth. However, only *SOT16*, *SOT17*, and *SOT18* play essential roles in the core structure pathway; therefore, the *SOT* genes mentioned below refer to these three genes. *SOT* genes encode a desulphoglucosinolate sulphotransferase involved in the final step of glucosinolate core structure biosynthesis, which promotes the substitution of SO_4^{2-} and accelerates the accumulation of glucosinolates.

We found 24 and 18 *SOT* genes in *B. racemosa* and *B. asiatica*, respectively, and 9 of these genes arose from a recent tetraploid. There are 21 and 18 *SOT* genes produced by tandem duplication in *B. racemosa* and *B. asiatica* (Additional file 1: Figure S11), and the domains of these *SOT* genes remained intact (Additional file 1: Figure S12). The massive expansion of *SOT* genes enables *B. racemosa* and *B. asiatica* to adapt to complex coastal woodlands and resist insects and fungi.

Discussion

We developed a tool (Allo4D) to divide allotetraploid subgenomes based on a scaffold-scale genome assembly in this study. It was successfully tested using simulated rice and real *Arabidopsis suecica* data. Our approach compensates for previous technical limitations that can only divide subgenomes at the genome level. Compared with similar software and methods, it requires less data, significantly reducing sequencing costs.

When we tested the alignment method with WGDI and MCScan, the collinear genes extracted using WGDI produced more consistent subgenome typing results than MCScan. WGDI slightly increased the number of SubA (from 27 304 to 27 566) and SubB genes (from 27 020 to 27 590) in *B. racemosa*. Therefore, for poorly assembled genomes, we recommend that users use WGDI for upstream alignment analysis and then perform Allo4D genome typing.

We found two recent whole genome duplication events in *B. racemosa* (~66 Mya and ~1 Mya), which accompanied the K-Pg boundary and GAST and SST reconstruction events, respectively. WGD events are closely associated with the timing of dramatic climatic and environmental changes (Wu *et al.*, 2020). This also allows polyploids to have stronger evolutionary potential, including plant speciation, diversification, genome evolution, and adaptive selection (Zhang *et al.*, 2019). Two recent WGD events in *B. racemosa* promoted the expansion of genes involved in salt tolerance, responses to external stimuli, and insect and fungal resistance, which helped it adapt to extreme environments. Similarly, the ~35 Mya WGDs in *Aegiceras corniculatum* enabled it to respond to high-salt stimuli (Feng *et al.*, 2021). Previous studies on mangrove genomes indicate that the impact of whole-genome duplication on the adaptive evolution of mangroves is universal (Xu *et al.*, 2017). Although significant progress has been made in plant WGD events, our understanding of the mechanisms underlying the relationship between polyploidy and extreme environmental adaptation is still limited.

After whole genome duplication, the genomes were reconstructed, with some genes absent and some retained. In a recent study of *Cucumis*, the majority of the genes in progenitors were preserved in the offspring genomes, and most genomic changes emerged immediately after interspecific hybridization (Yu *et al.*, 2021). Similarly, only 2.56% of genes were absent in *B. racemosa*. After allopolyploidization, the two subgenomes of *B. racemosa* preferentially retained specific genes with different functions, suggesting that biased fractionation (gene loss occurring unevenly between duplicated genomic regions) occurred in the allopolyploid.

It is vital that genes present in both subgenomes are related to reproductive development, indicating a 'smart genome reconstruction strategy', enabling it to cope with nucleocytoplasmic incompatibility and organelle remodelling caused by polyploidy. In addition, most positive selected genes were related to resistance and defence, providing a stronger evolutionary potential for *B. racemosa*.

Experimental procedures

Plant material

We collected fresh and healthy leaves of *B. racemosa* and *B. asiatica* from Dong Zhai Gang National Nature Reserve, Haikou, China, and Pasir Ris Park, Singapore, for sequencing. Plant material was immediately frozen in liquid nitrogen and stored at −80 °C until total genomic DNA and RNA extraction. High-quality genomic DNA and RNA were extracted using the cetyltrimethylammonium bromide (CTAB) method (Doyle and Doyle, 1987).

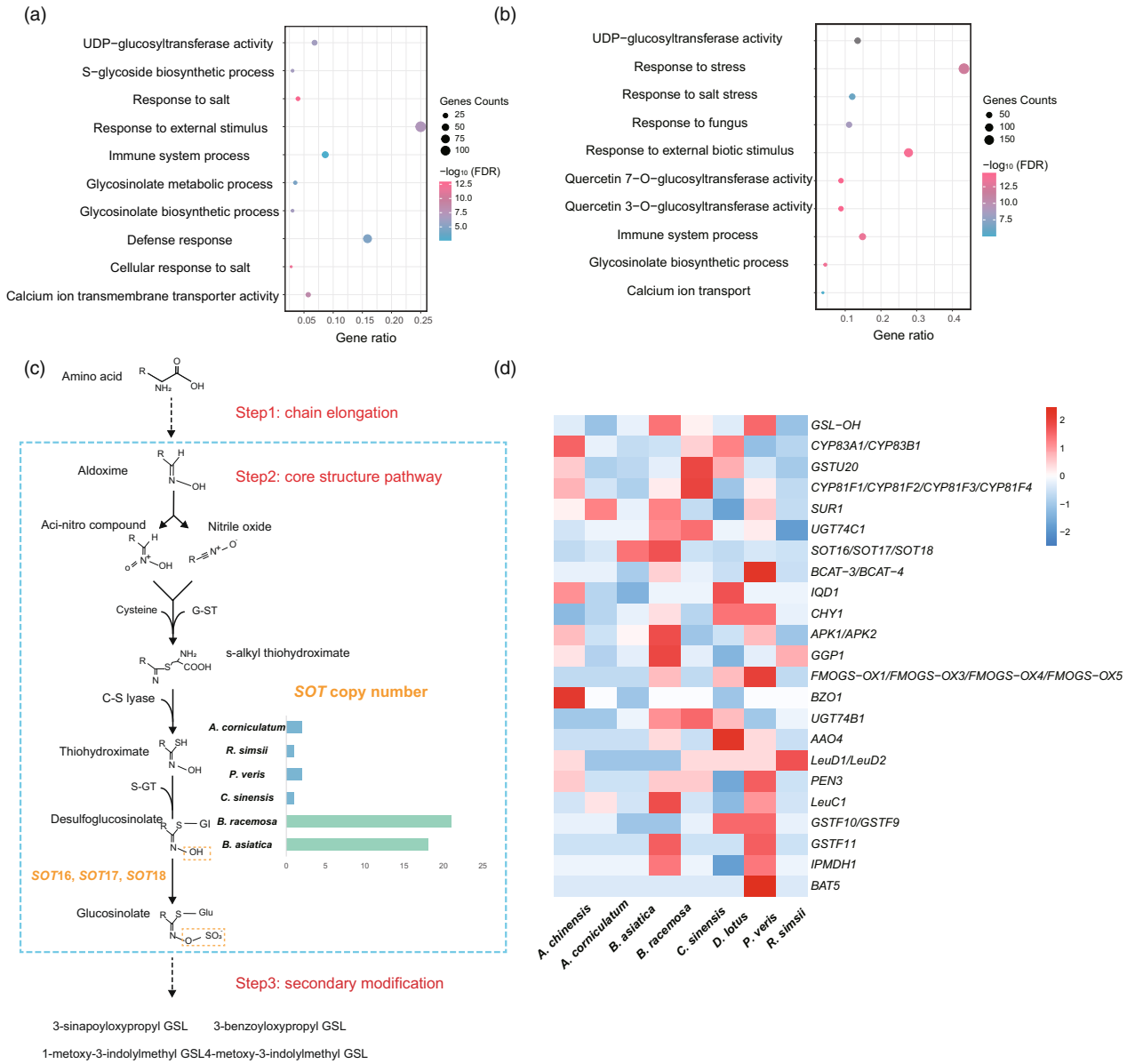


Figure 5 Rapid expansion gene families and evolution of glucosinolates (GSLs) biosynthesis pathways in *B. racemosa* and *B. asiatica*. The significantly enriched GO terms of rapid expansion gene families of *B. racemosa* (a) and *B. asiatica* (b). (c) The partial biosynthesis pathway of GSLs. There are three main steps in this pathway (steps 1 and 3 are not shown in the figure, and step 2 is surrounded by a blue dotted box). The bar graph on the right shows the number of *SOT* genes in each species. G-ST, glutathione-S-transferase; S-GT, S-glucosyltransferase; SOT, sulphotransferase. (d) The heat map of GSLs-related gene copy number. The intensity of the heatmap represents the column-scaled copy number (*Z-score*).

Library construction, sequencing, and assembly

To construct short-insert libraries, we prepared DNA using TIANamp Plant DNA Kits and then sheared it with a sonication device (500 bp). The DNA was subjected to paired-end (150 bp) sequencing on an Illumina X-TEN platform for genomic library construction. RNA-seq libraries were built using the NEBNext Ultra RNA Library Prep Kit to facilitate the prediction of protein-coding genes.

For PacBio library construction, genomic DNA was sheared to ~20 kb, and short fragments (<7 kb) were filtered using BluePipin. The filtered DNA was converted into a proprietary SMRTbell library using the PacBio DNA Template Preparation Kit. Single-molecule real-time (SMRT) long-read sequencing was performed using the PacBio Sequel II sequencing platform (v3.0).

Hi-C libraries were prepared according to Padmarasu *et al.* (2019). The samples were cross-linked with 3% formaldehyde at 4 °C for 30 min, and quenched with 0.375 M glycine for 5 min. After cell lysis, endogenous nucleases were deactivated using 0.3% SDS. Chromatin DNA was digested with 100 U MboI (NEB), biotin-labelled and ligated with 50 U T4 DNA ligase (NEB). Following reverse cross-linking, DNA was extracted using a QIAamp DNA Mini Kit (Qiagen). Purified DNA fragments (300–500 bp) were subjected to blunt-end repair, A-tailing, and adaptor addition, followed by biotin-streptavidin-mediated pull-down and PCR amplification. Finally, the Hi-C libraries were

quantified and sequenced using the MGI-seq platform (BGI, China).

We assembled the *de novo* genomes of *B. racemosa* and *B. asiatica* based on subreads from PacBio sequencing using FALCON (v2.2.4) and wtdbg2 (v2.5) with optimized parameters (Chin *et al.*, 2016; Ruan and Li, 2020). The raw contig genomes were polished using Quiver (SMRT Analysis v2.3.0) with default parameters (Chin *et al.*, 2013). Finally, several rounds of iterative error correction were performed on the raw genomes using Pilon (v1.22) to improve the primary assembly accuracy.

Clean Hi-C reads were evaluated and qualified using HiC-Pro (Servant *et al.*, 2015). Subsequently, the organization and orientation of the genomic groups were determined using a 3D-DNA pipeline (Dudchenko *et al.*, 2017). The Hi-C interaction patterns were visualized using HiCPlotter (Akdemir and Chin, 2015). The chromosomes of *B. racemosa* were compiled as chromosomes 01–26 based on their homologous relationship with *B. asiatica*. To visualize collinearity between *B. racemosa* and *B. asiatica*, the scaffold-scale genome assembly of *B. asiatica* was anchored to the chromosome-scale genome assembly of *B. racemosa* using RagTag (Alonge *et al.*, 2022).

Genome annotation

Protein-coding genes were predicted using three methods: ab initio, homology-based, and transcriptome-based. For ab initio prediction, we used Augustus (v2.5.5) (Stanke *et al.*, 2006) and GeneMark (v4.32) (Besemer *et al.*, 2001) to predict the coding genes. For homology-based prediction, exonerate (v2.2.0) (Slater and Birney, 2005) was used to generate gene structures using homologous proteins from sequenced related plant species in Ericales and model plant species. For the transcriptome-based forecast, RNA-seq data were mapped against genomes using TopHat (v2.1.1) (Trapnell *et al.*, 2009), and Cufflinks (v2.2.1) (Trapnell *et al.*, 2012) was used to generate a final transcriptome assembly. Finally, all gene models from the three methods were integrated using EvidenceModeler (EVM) to produce non-redundant and consensus gene sets (Haas *et al.*, 2008).

Repetitive sequences annotation

The Extensive *de novo* TE Annotator (EDTA) (v2.0.1) with default parameters (Ou *et al.*, 2019) was used to predict whole-genome TE and evaluate LTR insertion times to obtain long terminal repeat (LTR) retrotransposons, inverted terminal repeat (TIR) transposons, miniature inverted transposable elements (MITEs), and Helitrons.

Genome quality assessment

GCE (v1.0.3) (Liu *et al.*, 2013) was used to estimate the genome size using filtered Illumina shorts reads (kmer = 21, -M 10000). We aligned the Illumina short reads to the assembled genome using BWA (v0.7.17) (Li, 2013; Li and Durbin, 2009). GATK (v4.2.0.0) was used to calculate heterozygosity (SNP: MQ < 30.0, FS > 30.0, SOR >4.0; Indel: MQ < 35.0, FS > 30.0, SOR >4.0) (DePristo *et al.*, 2011; der Auwera and O'Connor, 2020; McKenna *et al.*, 2010). We also evaluated the completeness of the assembled genome using BUSCO (v3.0.2) based on the eudicotyledons_odb10 database (Seppey *et al.*, 2019).

Allo4D pipeline design and test

We developed a new tool (Allo4D) to distinguish allopolyploid subgenomes (Figure 2). First, we analysed the inter- and intragenomic collinearity between allotetraploid and relative diploid species using MCScanX with default parameters. We also used the alignment method in WGDI to test for collinear genes. Subsequently, the Ks values of the gene pairs across syntenic blocks were calculated with Yang-Nielsen (YN) model in the PAML software package (v4.9) (Wang *et al.*, 2010). After filtering the collinear genes between the two genomes generated by ancient duplication events according to Ks values (synonymous substitutions per site), we selected duplicate collinear pairs (one *B. asiatica* gene and two *B. racemosa* genes). Clusters were constructed according to the block number of *B. asiatica* (only clusters with more than five collinear pairs were retained). Phylogenetic trees were constructed for each set using IQ-TREE (v2.0.3) (Nguyen *et al.*, 2015) with one outgroup (e.g., *Rhododendron simsii* was used in our study) (Yang *et al.*, 2020). We selected only the clusters with the most significant number of tree topologies (green tree in Figure 2d for subsequent analysis). We defined the gene clustered with diploidy as the SubA gene, while the other was the SubB gene. The SubA and SubB genes were marked on the allotetraploid genome, and subgenomes A and B were divided according to these markers. If the number of SubA genes exceeded 90% of the SubB genes on one scaffold, the scaffold was classified as subgenome A. Conversely, it was classified as subgenome B when the number of SubB genes exceeded 90% of the SubA genes on one scaffold. We used Allo4D to divide the subgenomes in the chromosome-scale genome assembly of *B. racemosa* and compared the phasing results between the chromosome-scale and scaffold-scale genome assemblies.

The pipeline was applied to simulate rice and real *Arabidopsis suecica* allotetraploid genomes to further test the reliability. In addition, the average allotetraploid genome length was cleaved from 0.5 Mb to 10 Mb to accommodate the genome at the scaffold level. We obtained four genomes (*Oryza sativa indica*, *Oryza sativa japonica*, *Oryza punctata*, and *Leersia perrieri*) from RiceRelativesGD (http://ibi.zju.edu.cn/ricerelativesgd/) (Mao *et al.*, 2019) for simulated rice data. The genomes of *Oryza sativa indica* and *Oryza punctata* were combined as a putative allotetraploid and *Oryza sativa japonica* as a relative diploid. To test the actual allotetraploid genome data, we used the Allo4D pipeline to distinguish the subgenomes of *Arabidopsis suecica* (Jiang *et al.*, 2021), using the relatively diploid *Arabidopsis thaliana* to help divide. Three metrics—accuracy, precision, and recall— were used to assess the performance of our pipeline, and their definitions are as follows:

1. Accuracy (%) = Correctly classified genome length/Total genome length
2. Precision: subgenome A Precision (%) = Correctly classified length of subgenome A/(Correctly classified length of subgenome A + length of subgenome B misclassified as subgenome A)
   subgenome B Precision (%) = Correctly classified length of subgenome B/(Correctly classified length of subgenome B + length of subgenome A misclassified as subgenome B)
3. Recall: subgenome A Recall (%) = Correctly classified length of subgenome A/(Correctly classified length of subgenome A + length of subgenome A misclassified as subgenome B + unclassified length of subgenome A)
   subgenome B Recall (%) = Correctly classified length of subgenome B/(Correctly classified length of subgenome B + length of subgenome B misclassified as subgenome A + unclassified length of subgenome B).

Ploidy and subgenome identification of *Barringtonia racemosa*

In addition to the evidence regarding chromosome number and genome size, we analysed the collinearity between *B. racemosa* and *B. asiatica* to identify ploidy. The Python version of MCScanX (v1.2.7) (Tang *et al*., 2008) was used to identify the synteny blocks of the two genomes. Synteny blocks were identified based on homologous gene pairs between species using the jcvi.compara.catalog module, and low-gene-density blocks were removed using the jcvi.compara.synteny module (--minspan = 30). We used the jcvi.graphics.karyotype module to visualize high-quality synteny blocks using default parameters.

The genomic kmer spectrum of *B. racemosa* was analysed to determine whether the genome was heterologous or homologous. First, raw Illumina short reads were filtered using fastp (v0.23.4) (Chen *et al*., 2018), which was used to remove adapters and sequences with low base quality using default parameters. We then used Jellyfish (v2.3.0) (Marçais and Kingsford, 2011) to compute a histogram of the kmer frequencies (kmer length = 21 bp). Genome size and ploidy were estimated using GenomeScope 2.0 (-k 21, -p 4) (Ranallo-Benavidez *et al*., 2020). We also used Tetmer (Becher *et al*., 2020) to determine ploidy using default parameter settings.

Comparative phylogenetic and gene family analyses

OrthoFinder (v2.4.0) (Emms and Kelly, 2019) was used to identify the orthologous genes in *B. racemosa*, *B. asiatica*, and nine other representative plant species (Additional file 2: Table S11). Ultimately, 321 single-copy homologous were identified. These protein sequences were aligned using MAFFT (v7.464) (Katoh and Standley, 2013) with default parameters and then converted to codon alignments using PAL2NAL (v14) (Suyama *et al*., 2006). We used Gblocks (v0.91b) (Castresana, 2000) to trim the alignments (-t = c -b4 = 5 -b5 = a) and merged the sequences of the same species. We used ModelFinder in IQ-TREE (Nguyen *et al*., 2015) to predict the best-fit substitution structure model and constructed a phylogenetic tree using RAxML-NG (v1.0.1) (Kozlov *et al*., 2019). MCMCTEE in PAML (v4.9) (Yang, 2007) was used to estimate divergence time based on two fossils (1. the root node of eudicots and monocots was constrained to between 125 and 247 Mya; 2. the common ancestor of eudicots was placed at 119.6–128.63 Mya) (Morris *et al*., 2018). The results were visualized using the R package ggtree (Yu *et al*., 2017).

CAFE (v5.0) (Mendes *et al*., 2021) was used in gene family analysis after removing the large families with more than 100 gene copies in one or more species. Gene families that expanded and contracted were identified, and then we obtained the rapid expansion gene family (the gene count expanded significantly, *P*-value <0.01).

Divergent time estimation and whole genome duplication events

We also estimated the allotetraploidization time of *B. racemosa* (merge time between two subgenomes) using a method described for common carp (Xu *et al*., 2019). First, we used Nucmer (v4.0.0) (Delcher *et al*., 2003) to align subgenomes A and B (alignment length > 1000, alignment identity >90) to obtain a matching region. RepeatModeler (v2.0.1) (http://www.repeatmasker.org/) was used to create a repeat sequences database of the matching regions, and TEs were identified in the subgenomes using RepeatMasker (v4.1.1) (https://www.repeatmasker.org/). We used calcDivergenceFromAlign.pl (a script in RepeatMasker) to calculate the substitutions of TEs, and divergence was obtained by comparing the TEs between *B. racemosa* and the repeat sequences database obtained above. Finally, the ggplot2 (Wickham, 2016) R package was used to show divergence values using a Generalized Additive Model (GAM) to fit the curves. The merger and divergence time points between the two subgenomes corresponded to TE divergence rates of 2.4% and 18.5% in subgenomes A and B, respectively. The estimated divergent time of *B. racemosa* and *B. asiatica* was 7.8 Mya, and therefore, the allotetraploidization time is: 7.8 Mya × 2.4/18.5 = 1.01 Mya.

We performed collinearity analysis and Ks value calculations to determine the rate of evolution and whole-genome duplication events. The synteny blocks of *B. racemosa*, *B. asiatica*, *D. lotus*, and *V. vinifera* were identified using the same method (Figure 2a). We also used MCScanX (Tang *et al*., 2008) to verify these results and to determine collinear blocks after searching for homologous genes using BLASTP (v2.9.0) (*e*value <1*e*-10). The Ks values between these collinear pairs were calculated using KaKs_Calculator (v2.0) with the Yang-Nielsen (YN) model (Wang *et al*., 2010).

Genome presence and absence identification

To study the reconstruction pattern of *B. racemosa* genome after allopolyploidy, we identified the genes present and absent in *B. racemosa* with home scripts using *B. asiatica* genome as a reference (Figure 4). First, we analysed collinearity between the subgenomes of *B. racemosa* and *B. asiatica* and then filtered collinear genes generated by ancient duplication events using the same method as in subgenomes division. We also used one vs. two gene pairs identified as tetraploidized collinear regions. The other original genes within every region of *B. asiatica* were inserted, and then we obtained the present and absent genes in *B. racemosa* genome using *B. asiatica* as a reference. In addition, absent genes were retained only when they were not in the result of BLASTP (*e*value <1*e*-10) between the two genomes.

We also used short-reads mapping and genome alignment methods to eliminate falsely identified loss of genes due to assembly and annotation errors. Short reads of *B. racemosa* were aligned to its genome using BWA (v0.7.17) (Li and Durbin, 2009), and BEDTools (v2.29.2) (Quinlan and Hall, 2010) were used to calculate the coverage around the deletion region in the subgenomes of *B. racemosa*. If the coverage reached 100%, the genome assembly was considered continuous and complete. To eliminate annotation errors, we aligned the genomes of *B. racemosa* and *B. asiatica* with Lastz (v1.04.18) (Harris, 2007), using axtChain (Kent *et al*., 2003), RepeatFiller (Osipova *et al*., 2019), and chainCleaner (Suarez *et al*., 2017) in GenomeAlignmentTools (https://github.com/hillerlab/GenomeAlignmentTools) to construct a clean chain. Finally, we used TOGA (v1.1.5) to determine whether the orthologous genes of *B. asiatica* were present in the subgenomes of *B. racemosa* (Kirilenko *et al*., 2023).

We also calculated the selective pressure between these homologues, which had four conditions: two copies in *B. racemosa* were retained or lost, and one of the genes was lost in subgenome A or subgenome B. We then identified positive selected genes (Ka/Ks >1, *P*-value <0.05, Fisher's exact test) and identified their functions in TAIR (https://www.arabidopsis.org/) using BLASTP. Finally, we performed Gene Ontology (GO) enrichment (*P*-value <0.05) for these genes using clusterProfiler (v4.0) (Yu *et al*., 2012).

Functional protein annotation and enrichment analysis

Protein domains were identified by mapping to the InterProScan and Pfam databases using InterProScan (v5.51-85) (Jones *et al*., 2014) and HMMER (v3.3.2) (Finn *et al*., 2011). For functional protein annotation of *B. racemosa* and *B. asiatica*, the annotated genes were aligned to proteins in the NR (https://www.ncbi.nlm.nih.gov/), SwissProt (https://www.uniprot.org/), and KOG databases (Koonin *et al*., 2004). Annotations were also assigned to the Gene Ontology (http://geneontology.org/) metabolic pathways to obtain more functional information.

We carried out GO enrichment (*P*-value <0.05) for the rapid expansion family and duplicated genes retained by the recent WGD of *B. racemosa* and *B. asiatica*. The R package clusterProfiler (Yu *et al*., 2012) was used for this analysis, and the R package ggplot2 (Wickham, 2016) was used to visualize the results.

Identification of the critical glucosinolates-related genes

We searched for candidate genes of glucosinolates (GSLs) in all eight Ericales genomes (*B. racemosa*, *B. asiatica*, *D. lotus*, *A. chinensis*, *R. simsii*, *C. sinensis*, *P. veris*, and *A. corniculatum*) against *Arabidopsis thaliana* (Athaliana_Araport11, https://phytozome-next.jgi.doe.gov) using the result of OrthoFinder above. The essential GSLs-related genes in *A. thaliana* were used for subsequent analysis (Song *et al*., 2021). We predicted their conserved domains and gene structures to confirm the accuracy of these genes (https://www.ncbi.nlm.nih.gov/Structure/cdd/).

Acknowledgements

We thank Dr. Li He, Dr. Xingtan Zhang, Dr. Zhenyang Liao, Dr. Xianliang Song, Dr. Liyuan Wang, Dr. Hannes Becher, Dr. Ghillean Prance, Dr. Xiao Feng for their insightful comments. This project was supported by the National Natural Science Foundation of China (32170230, 31971540, 31830005); the Guangdong Basic and Applied Basic Research Foundation (2023B1515020083); and the Innovation Group Project of Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai) (311021006).

Conflict of interest statement

The authors declare that they have no competing interests.

Author contributions

Z.H. designed and conceptualized the study. Y.W., S.Shao, S.X., Z.G., S.Shi, and Z.H. collected materials. Y.W., W.W., and Q.F. performed the experiments. Y.W., Y.L. W.W., and Z.H. performed the data analysis. Y.W. and Z.H. wrote the manuscript. All authors read and approved the final manuscript.

Code availability statement

All codes used in the project were submitted to Github (https://github.com/yuanw-18/code).

Data availability statement

The genome assembly and annotation data have been deposited in the Genome Warehouse (GWH), at the National Genomics Data Center (NGDC) (https://ngdc.cncb.ac.cn/gwh/), under project accession number PRJCA020101. The scaffold-scale genome assembly (v1.0) was deposited under the accession numbers GWHEQTY00000000 (*B. asiatica*) and GWHEQTZ00000000 (*B. racemosa*). The chromosome-scale genome assembly (v2.0) of *B. racemosa* was deposited under the accession number GWHDUDD00000000. All the raw sequencing data of Pacbio, RNAseq, and Hi-C have been deposited in the Genome Sequence Archive (GSA), which is available at https://ngdc.cncb.ac.cn/gsa/, under the accession number CRA013727.

References

Adams, K.L. and Wendel, J.F. (2005) Polyploidy and genome evolution in plants. *Curr. Opin. Plant Biol.* **8**, 135–141.

Aghajanzadeh, T.A., Reich, M., Kopriva, S. and De Kok, L.J. (2018) Impact of chloride (NaCl, KCl) and sulphate (Na_2SO_4, K_2SO_4) salinity on glucosinolate metabolism in Brassica rapa. *J. Agron. Crop Sci.* **204**, 137–146.

Akagi, T., Shirasawa, K., Nagasaki, H., Hirakawa, H., Tao, R., Comai, L. and Henry, I.M. (2020) The persimmon genome reveals clues to the evolution of a lineage-specific sex determination system in plants. *PLoS Genet.* **16**, e1008566.

Akdemir, K.C. and Chin, L. (2015) HiCPlotter integrates genomic data with interaction matrices. *Genome Biol.* **16**, 1–8.

Alonge, M., Lebeigle, L., Kirsche, M., Jenike, K., Ou, S., Aganezov, S. *et al*. (2022) Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol.* **23**, 1–19.

der Auwera, G.A. and O'Connor, B.D. (2020) *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. Sebastopol: O'Reilly Media.

Becher, H., Brown, M.R., Powell, G., Metherell, C., Riddiford, N.J. and Twyford, A.D. (2020) Maintenance of species differences in closely related tetraploid parasitic Euphrasia (Orobanchaceae) on an isolated Island. *Plant Commun.* **1**, 100105.

Besemer, J., Lomsadze, A. and Borodovsky, M. (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.* **29**, 2607–2618.

del Carmen Martínez-Ballesta, M., Moreno, D.A. and Carvajal, M. (2013) The physiological importance of glucosinolates on plant response to abiotic stress in Brassica. *Int. J. Mol. Sci.* **14**, 11607–11625.

Castresana, J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540–552.

Chen, S., Zhou, Y., Chen, Y. and Gu, J. (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890.

Chin, C.-S., Alexander, D.H., Marks, P., Klammer, A.A., Drake, J., Heiner, C., Clum, A. *et al*. (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569.

Chin, C.-S., Peluso, P., Sedlazeck, F.J., Nattestad, M., Concepcion, G.T., Clum, A., Dunn, C. *et al*. (2016) Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054.

Delcher, A.L., Salzberg, S.L. and Phillippy, A.M. (2003) Using MUMmer to identify similar regions in large sequence sets. *Curr. Protoc. Bioinform.* **00**, 10–13.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A. *et al*. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498.

Doyle, J.J. and Doyle, J.L. (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* **19**, 11–15.

Doyle, J.J., Flagel, L.E., Paterson, A.H., Rapp, R.A., Soltis, D.E., Soltis, P.S. and Wendel, J.F. (2008) Evolutionary genetics of genome merger and doubling in plants. *Annu. Rev. Genet.* **42**, 443–461.

Dudchenko, O., Batra, S.S., Omer, A.D., Nyquist, S.K., Hoeger, M., Durand, N.C., Shamim, M.S. *et al*. (2017) De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95.

Edger, P.P., Poorten, T.J., VanBuren, R., Hardigan, M.A., Colle, M., McKain, M.R., Smith, R.D. *et al*. (2019) Origin and evolution of the octoploid strawberry genome. *Nat. Genet.* **51**, 541–547.

Emms, D.M. and Kelly, S. (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 1–14.

Feng, X., Li, G., Xu, S., Wu, W., Chen, Q., Shao, S., Liu, M. *et al*. (2021) Genomic insights into molecular adaptation to intertidal environments in the mangrove *Aegiceras corniculatum*. *New Phytol.* **231**, 2346–2358.

Feng, X., Li, G., Wu, W., Lyu, H., Wang, J., Liu, C., Zhong, C. *et al*. (2023) Expansion and adaptive evolution of the WRKY transcription factor family in Avicennia mangrove trees. *Mar. Life Sci. Technol.* **5**, 155–168.

Finn, R.D., Clements, J. and Eddy, S.R. (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29–W37.

Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O. *et al*. (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* **9**, 1–22.

Harris, R.S. (2007) *Improved Pairwise Alignment of Genomic DNA*. Pennsylvania: The Pennsylvania State University.

He, Z., Feng, X., Chen, Q., Li, L., Li, S., Han, K., Guo, Z. *et al*. (2022) Evolution of coastal forests based on a full set of mangrove genomes. *Nat. Ecol. Evol.* **6**, 1–12.

Jia, K.-H., Wang, Z.-X., Wang, L., Li, G.-Y., Zhang, W., Wang, X.-L., Xu, F.J. *et al*. (2022) SubPhaser: a robust allopolyploid subgenome phasing method based on subgenome-specific k-mers. *New Phytol.* **235**, 801–809.

Jiang, X., Song, Q., Ye, W. and Chen, Z.J. (2021) Concerted genomic and epigenomic changes accompany stabilization of Arabidopsis allopolyploids. *Nat. Ecol. Evol.* **5**, 1382–1393.

Jiao, Y. (2018) Double the genome, double the fun: genome duplications in angiosperms. *Mol. Plant* **11**, 357–358.

Jiao, Y., Wickett, N.J., Ayyampalayam, S., Chanderbali, A.S., Landherr, L., Ralph, P.E., Tomsho, L.P. *et al*. (2011) Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**, 97–100.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H. *et al*. (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240.

Justen, V.L. and Fritz, V.A. (2013) Temperature-induced glucosinolate accumulation is associated with expression of BrMYB transcription factors. *HortScience* **48**, 47–52.

Katoh, K. and Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780.

Kent, W.J., Baertsch, R., Hinrichs, A., Miller, W. and Haussler, D. (2003) Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci. USA* **100**, 11484–11489.

Khan, N.A., Singh, S. and Umar, S. (2008) *Sulfur Assimilation and Abiotic Stress in Plants*. Heidelberg: Springer.

Kirilenko, B.M., Munegowda, C., Osipova, E., Jebb, D., Sharma, V., Blumer, M., Morales, A.E. *et al*. (2023) Integrating gene annotation with orthology inference at scale. *Science* **380**, eabn3107.

Koonin, E.V., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Krylov, D.M., Makarova, K.S., Mazumder, R. *et al*. (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol.* **5**, 1–28.

Kowal, R.R. (1989) Chromosome numbers of Asteranthos and the putatively related Lecythidaceae. *Brittonia* **41**, 131–135.

Kozlov, A.M., Darriba, D., Flouri, T., Morel, B. and Stamatakis, A. (2019) RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453–4455.

Li, H. (2013) *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. arXiv Prepr. arXiv1303.3997.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760.

Liu, B., Shi, Y., Yuan, J., Hu, X., Zhang, H., Li, N. *et al*. (2013) *Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects*. arXiv Prepr. arXiv1308.2012.

Mao, L., Chen, M., Chu, Q., Jia, L., Sultana, M.H., Wu, D., Kong, X. *et al*. (2019) RiceRelativesGD: a genomic database of rice relatives for rice research. *Database* **2019**, baz110.

Marçais, G. and Kingsford, C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770.

Martínez-Garcia, A., Rosell-Melé, A., McClymont, E.L., Gersonde, R. and Haug, G.H. (2010) Subpolar link to the emergence of the modern equatorial Pacific cold tongue. *Science (80-.).* **328**, 1550–1553.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K. *et al*. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303.

Mendes, F.K., Vanderpool, D., Fulton, B. and Hahn, M.W. (2021) CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics* **36**, 5516–5518.

Morawetz, W. (1986) Remarks on karyological differentiation patterns in tropical woody plants. *Plant Syst. Evol.* **152**, 49–100.

Morris, J.L., Puttick, M.N., Clark, J.W., Edwards, D., Kenrick, P., Pressel, S., Wellman, C.H. *et al*. (2018) The timescale of early land plant evolution. *Proc. Natl. Acad. Sci. USA* **115**, E2274–E2283.

Nguyen, L.-T., Schmidt, H.A., Von Haeseler, A. and Minh, B.Q. (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274.

Nichols, D.J. and Johnson, K.R. (2008) *Plants and the KT Boundary*. Cambridge, UK: Cambridge University Press.

Osipova, E., Hecker, N. and Hiller, M. (2019) RepeatFiller newly identifies megabases of aligning repetitive sequences and improves annotations of conserved non-exonic elements. *Gigascience* **8**, giz132.

Otto, S.P. (2007) The evolutionary consequences of polyploidy. *Cell* **131**, 452–462.

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J.R.A., Hellinga, A.J., Lugo, C.S.B. *et al*. (2019) Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 1–18.

Padmarasu, S., Himmelbach, A., Mascher, M. and Stein, N. (2019) In situ Hi-C for plants: an improved method to detect long-range chromatin interactions. *Methods Mol. Biol.* **1993**, 441–472.

Payens, J. (1967) A monograph of the genus Barringtonia. *Blumea* **15**, 157–263.

de Peer, Y., Mizrachi, E. and Marchal, K. (2017) The evolutionary significance of polyploidy. *Nat. Rev. Genet.* **18**, 411–424.

Prance, G.T. (2012) A revision of Barringtonia (Lecythidaceae). *Allertonia* **12**, 1–164.

Quadros, A.F., Helfer, V., Nordhaus, I., Reuter, H. and Zimmer, M. (2021) Functional traits of terrestrial plants in the intertidal: a review on mangrove trees. *Biol. Bull.* **241**, 123–139.

Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842.

Ranallo-Benavidez, T.R., Jaron, K.S. and Schatz, M.C. (2020) GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1–10.

Ruan, J. and Li, H. (2020) Fast and accurate long-read assembly with wtdbg2. *Nat. Methods* **17**, 155–158.

Seppey, M., Manni, M. and Zdobnov, E.M. (2019) BUSCO: assessing genome assembly and annotation completeness. *Methods Mol. Biol.* **1962**, 227–245.

Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.-J., Vert, J.-P., Heard, E. *et al*. (2015) HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 1–11.

Slater, G.S.C. and Birney, E. (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinform.* **6**, 1–11.

Snyder, C.W. (2016) Evolution of global temperature over the past two million years. *Nature* **538**, 226–228.

Soltis, P.S. and Soltis, D.E. (2016) Ancient WGD events as drivers of key innovations in angiosperms. *Curr. Opin. Plant Biol.* **30**, 159–165.

Sønderby, I.E., Geu-Flores, F. and Halkier, B.A. (2010) Biosynthesis of glucosinolates-gene discovery and beyond. *Trends Plant Sci.* **15**, 283–290.

Song, X., Wei, Y., Xiao, D., Gong, K., Sun, P., Ren, Y., Yuan, J. *et al*. (2021) Brassica carinata genome characterization clarifies U's triangle model of evolution and polyploidy in Brassica. *Plant Physiol.* **186**, 388–406.

Soundararajan, P. and Kim, J.S. (2018) Anti-carcinogenic glucosinolates in cruciferous vegetables and their antagonistic effects on prevention of cancers. *Molecules* **23**, 2983.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S. and Morgenstern, B. (2006) AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439.

Suarez, H.G., Langer, B.E., Ladde, P. and Hiller, M. (2017) chainCleaner improves genome alignment specificity and sensitivity. *Bioinformatics* **33**, 1596–1603.

Sun, P., Jiao, B., Yang, Y., Shan, L., Li, T., Li, X., Xi, Z. *et al*. (2022) WGDI: A user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes. *Mol. Plant* **15**, 1841–1851.

Suyama, M., Torrents, D. and Bork, P. (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609–W612.

Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M. and Paterson, A.H. (2008) Synteny and collinearity in plant genomes. *Science* **320**, 486–488.

Tomlinson, P.B. (2016) *The Botany of Mangroves*. New York: Cambridge University Press.

Trapnell, C., Pachter, L. and Salzberg, S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H. *et al*. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578.

Vig, A.P., Rampal, G., Thind, T.S. and Arora, S. (2009) Bio-protective effects of glucosinolates-a review. *LWT-Food Sci. Technol.* **42**, 1561–1572.

Wang, D., Zhang, Y., Zhang, Z., Zhu, J. and Yu, J. (2010) KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics, Proteom. Bioinform.* **8**, 77–80.

Wang, Z., Xue, J.-Y., Hu, S.-Y., Zhang, F., Yu, R., Chen, D., Van de Peer, Y. *et al*. (2022) The genome of *Hibiscus hamabo* reveals its adaptation to saline and waterlogged habitat. *Hortic. Res.* **9**, uhac067.

Wendel, J.F. (2015) The wondrous cycles of polyploidy in plants. *Am J Bot* **102**, 1753–1756.

Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag.

Wood, T.E., Takebayashi, N., Barker, M.S., Mayrose, I., Greenspoon, P.B. and Rieseberg, L.H. (2009) The frequency of polyploid speciation in vascular plants. *Proc. Natl. Acad. Sci.* **106**, 13875–13879.

Wu, S., Han, B. and Jiao, Y. (2020) Genetic contribution of paleopolyploidy to adaptive evolution in angiosperms. *Mol. Plant* **13**, 59–71.

Wu, D., Shen, E., Jiang, B., Feng, Y., Tang, W., Lao, S., Jia, L. *et al*. (2022) Genomic insights into the evolution of Echinochloa species as weed and orphan crop. *Nat. Commun.* **13**, 1–16.

Xu, S., He, Z., Zhang, Z., Guo, Z., Guo, W., Lyu, H., Li, J. *et al*. (2017) The origin, diversification and adaptation of a major mangrove clade (Rhizophoreae) revealed by whole-genome sequencing. *Natl. Sci. Rev.* **4**, 721–734.

Xu, P., Xu, J., Liu, G., Chen, L., Zhou, Z., Peng, W., Jiang, Y. *et al*. (2019) The allotetraploid origin and asymmetrical genome evolution of the common carp Cyprinus carpio. *Nat. Commun.* **10**, 1–11.

Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591.

Yang, F.-S., Nie, S., Liu, H., Shi, T.-L., Tian, X.-C., Zhou, S.-S., Bao, Y.-T. *et al*. (2020) Chromosome-level genome assembly of a parent species of widely cultivated azaleas. *Nat. Commun.* **11**, 1–13.

Yu, G., Wang, L.-G., Han, Y. and He, Q.-Y. (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. *Omi. J. Integr. Biol.* **16**, 284–287.

Yu, G., Smith, D.K., Zhu, H., Guan, Y. and Lam, T.T.-Y. (2017) ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36.

Yu, X., Wang, P., Li, J., Zhao, Q., Ji, C., Zhu, Z., Zhai, Y. *et al*. (2021) Whole-genome sequence of synthesized allopolyploids in *Cucumis* reveals insights into the genome evolution of allopolyploidization. *Adv. Sci.* **8**, 2004222.

Zhang, K., Wang, X. and Cheng, F. (2019) Plant polyploidy: origin, evolution, and its influence on crop domestication. *Hortic. Plant J.* **5**, 231–239.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure S1 Summary of current information on WGDs in mangroves.
Figure S2 Photos of sequencing species.
Figure S3 The genome-wide analysis of chromatin interactions in the *Barringtonica racemosa* genome based on Hi-C data.
Figure S4 The statistics for dividing subgenomes in *Arabidopsis suecica* using Allo4D.
Figure S5 The statistics for dividing subgenomes in simulated rice data using Allo4D.
Figure S6 The phylogenetic tree in each cluster.
Figure S7 The distribution of SubA and SubB genes in *B. racemosa* genome.
Figure S8 The distribution of sequence divergence rates of transposable elements (TEs) as percentages of subgenome sizes of *B. racemosa*.
Figure S9 Enriched GO functional categories of present and absent genes in *B. racemosa* ($p < 0.05$).
Figure S10 GO enrichment analysis for WGD retained genes in *B. racemosa* and *B. asiatica*.
Figure S11 Distribution of *SOT* genes in *B. racemosa* and *B. asiatica*.
Figure S12 The conserved domains of *SOT* genes.
Table S1 Library and sequence data information of *B. racemosa* and *B. asiatica* genomes.
Table S2 The BUSCO result of *B. racemosa* and *B. asiatica* genomes.
Table S3 Summary of the annotated TEs in the genome.
Table S4 Summary of the Allo4D statistics in *Arabidopsis thaliana*.
Table S5 Summary of the Allo4D statistics in stimulated rice.
Table S6 Statistics of *B. racemosa* subgenome division results.
Table S7 Subgenome genome phase results based on the chromosome-scale and scaffold-scale genome assemblies.
Table S8 Subgenome genome phase statistics between the chromosome-scale and scaffold-scale genome assemblies.
Table S9 Statistics of the positive selected absent and present genes.
Table S10 The number of genes involved in glucosinolate metabolism pathways.
Table S11 Genomic data used for phylogenomic and gene family analyses.

