

Genome assembly provides insights into the genome evolution of *Baccaurea ramiflora* Lour.

Lour. Jianjian Huang

Hanshan Normal University

Jie Chen

Guangdong Ocean University

Min Shi

Hanshan Normal University

Jiaqi Zheng

Hanshan Normal University

Ming Chen

Hanshan Normal University

Linjun Wu

Hanshan Normal University

Hui Zhu

Hanshan Normal University

Yuzhong Zheng

Hanshan Normal University

Qinghan Wu

Hanshan Normal University

Fengnian Wu

`fengnian.wu@hstc.edu.cn`

Hanshan Normal University

Article

Keywords: *Baccaurea ramiflora*, Whole-genome sequencing, de novo assembly, Genome evolution

Posted Date: August 17th, 2023

DOI: <https://doi.org/10.21203/rs.3.rs-3094729/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on February 28th, 2024. See the published version at <https://doi.org/10.1038/s41598-024-55498-4>.

Abstract

Baccaurea ramiflora Lour., an evergreen tree of the *Baccaurea* genus of the Phyllanthaceae family, is mainly distributed in South Asia, Southeast Asia, and southern China, including southern Yunnan Province. It is a wild or semicultivated tree species of ornamental, edible, and medicinal value, with important development potential. Here, we report the whole-genome sequencing of *B. ramiflora* through the combination of PacBio SMRT and Illumina HiSeq 2500 sequencing techniques. The size of the assembled genome is 975.8 Mb, with a contig N50 of 509.33 kb and a longest contig of 7.74 Mb. Highly repetitive sequences make up 73.47% of the genome, of which 52.1% are long terminal repeat–retrotransposon sequences. A total of 29,172 protein-coding genes are predicted, of which 25,980 (89.06%) have been annotated, while 3,452 noncoding RNAs are found. Comparative genomic analysis indicates that *B. ramiflora* is closely related to the family Euphorbiaceae, being sister groups to each other that diverged approximately 59.9 Mya. During the evolutionary process, *B. ramiflora* has 278 candidate genes under positive selection. The synonymous substitution rate and collinearity analysis show that *B. ramiflora* only underwent one ancient genome-wide triploidization event and no recent genome-wide duplication event. This high-quality *B. ramiflora* genome provides a valuable resource for basic research and tree improvement programs of the Phyllanthaceae family.

Introduction

Baccaurea ramiflora Lour., an evergreen tree of the *Baccaurea* genus of the Phyllanthaceae family of the Malpighiales order (APG IV system) (Fig. 1), is distributed in Southeast Asia and South Asia, is mainly cultivated in South China, including in southern Yunnan Province^{1,2}, and is occasionally cultivated in Nepal, India, Myanmar, Bangladesh, Indochina, Thailand, and Malaysia³. It is an underutilized tropical wild fruit tree species with unique garden and landscape development value. It is a typical tropical plant with old stems that bear flowers, with a beautiful tree shape, several fruit colors (green, white, light yellow, orange, pink, purple, etc.), and different fruit shapes (spherical, long oval, olive, etc.). When fruiting, the fruits grow densely on old branches, making the tree an ideal choice for garden landscaping for its appealing fruits and branches⁴.

The pulp of *B. ramiflora*, sweet and sour, with a tender texture, can be consumed fresh or processed in the forms of fruit juice, preserved fruit, jam, and fruit wine⁵. It is rich in nutrients, with an edible rate of 49.2%, consisting of water (84.7%), carbohydrate (86.14%), fats (0.06%), crude protein (5.43%), fibers (0.29%), vitamin C (1.57 mg/100 g), titratable acid (1.99%), and total sugar (11.87%)^{6,7,8,9}. Generally, the pulp yield per plant of an adult *B. ramiflora* tree can reach 100–150 kg, so planting *B. ramiflora* can bring many economic benefits to farmers⁹. However, *B. ramiflora* has not yet been commercially planted and utilized⁷, although some landrace trees have been planted around the houses of local farmers. *B. ramiflora* of different origins can have vastly different flowering and fruiting times; the flowering and fruiting periods of *B. ramiflora* trees growing in Hainan and Southeast Asia areas are approximately 45 days earlier than

those in Chinese mainland, and the fruits mature between late May and early August, similar to those of lychee, suggesting that *B. ramiflora* has the potential to make up for the lack of seasonal fruit⁵.

B. ramiflora also has high medicinal and health value^{4,10}. Its leaves, fruits, stems, bark and seeds are the ingredients of many Chinese herbal medicines that have long been used to treat jaundice, constipation, indigestion, cellulitis, inflammation, and rheumatoid arthritis, as well as being an antidote for snake venom^{2,11,12,13,14}. The chemical and medicinal ingredients of *B. ramiflora* have been investigated extensively, and more than 30 compounds that have a host of biological activities such as analgesic, anthelmintic, antioxidant, antidiarrheal, anti-inflammatory, cytotoxic, hemolytic, hypoglycemic, hypolipidemic, insecticidal, neuropharmacological, platelet, antifungal, antibacterial, and other medicinal values have been isolated and characterized^{15,16,17,18,19,20}.

Although *B. ramiflora* is rich in germplasm resources, with diverse fruit phenotypes and varying fruit quality, artificially developed elite varieties have not been reported, and cultivated trees are mostly landraces without a unified evaluation standard. Due to the lack of investigations of the germplasm resources and genomics of *B. ramiflora*, elite strains of *B. ramiflora* have not been developed, and there are no large-scale commercial plantations of the tree^{5,7}. Genomic studies of *B. ramiflora* are important for understanding its germplasm resources but have not been conducted yet, which has limited the tree improvement work and thus the tree promotion and production. As a result, *B. ramiflora* breeding is still in its infancy. A complete *B. ramiflora* genome would be of great significance for the study of the diversity of *B. ramiflora* germplasm resources and the mining of functional genes related to agronomic traits. Here, we report the sequencing and assembly of *B. ramiflora* and the analysis of its gene family evolution, whole-genome duplication events, and collinearity. As a reference genome, this *B. ramiflora* genome can provide molecular data for research on the Phyllostachys family.

Results

Sequencing and assembly. The sequencing and assembly of the *B. ramiflora* genome were performed based on long-read PacBio SMRT sequencing in combination with short-read Illumina HiSeq2500 sequencing for error correction. Through approximately 60 Gb of next-generation sequencing data, we conducted a survey analysis of the genome with a K-mer of 19 and estimated that its genome size was approximately 973 Mb, with a heterozygosity of 0.634% and a repeat sequence ratio of 57.5% (Fig. S1). The Illumina HiSeq2500 sequencing generated approximately 68.6 Gb of raw reads, with a sequencing depth of 60×, while the PacBio SMRT sequencing generated 114.14 Gb of raw reads, with a sequencing depth of 100×. The *B. ramiflora* genome was assembled through different assembly methods (Table S1), including CANU + dereundancy, WTDBG (Canu error correction + WTDBG assembly), WTDBG + dereundancy, and Flye + dereundancy. The CANU + dereundancy method showed the lowest properly paired mapping rate (69.59%) in the next-gen sequencing data, and WTDBG, Flye, and WTDBG + dereundancy showed obvious abnormal secondary peaks in the GC depth graph (inconsistent with the GC content of the main peak, with an average GC depth of below 20×). Considering that the assembly

difficulty may be related to the complexity of the species itself, based on the positions of the abnormal peaks in the above assembly results, we proportionally filtered the original subreads, in which subreads with a GC content in the range of 40–45%/45–50% were filtered out, and the GC content of the filtered subread sequences essentially became normally distributed. The filtered subreads were assembled using the Canu assembler, after which the preliminary assembly did not show any abnormal GC peak, and the genome size after dereplication was close to the survey evaluation size, with a ContigN50 of approximately 300 kb. To further improve the assembly, after filtering out the subreads highly homologous to mitochondrial and chloroplast sequences, the remaining sequences were assembled using the Canu + deheterozygosity method, which achieved the best assembly, with a normal distribution of GC content and no abnormal peaks.

The size of the final assembly of *B. ramiflora* genome was approximately 975.8 Mb, which was slightly larger than the genome size estimated based on K-mer. This was likely because the high heterozygosity of the genome led two sister chromatids to be assembled together, resulting in a mixed genome greater than the actual size. The assembly contained 3,346 contigs, with the following statistics: GC content: 35.31%; N50: 509.33 kb; Contig Min: 1.63 kb; N90: 118.07Kb; Contig Max: 7.74 Mb. For plant genomes with high heterozygosity, an assembly with a contig N50 of 509.33 kb is considered good. Compared with the assembled genomes of other plant species of family Euphorbiaceae, i.e., *Ricinus communis* (genome size: 350.5 Mb; Contig N50 21.4 kb), *Manihot esculenta* (genome size: 693.6 Mb; Contig N50 3.3 Mb), *Hevea brasiliensis* (genome size: 1.5 Gb; Contig N50 152.7 kb), and *Jatropha curcas* (genome size: 266.8 Mb, Contig N50 134.3 kb) (data source: NCBI database), the genome of *B. ramiflora* is large, being smaller only than that of *H. brasiliensis*, with a Contig N50 only smaller than that of *M. esculenta*.

Genomic evaluation. To verify the accuracy and completeness of the assembly, we first evaluated the data mapping ratio of the DNA library, which accounted for 98.86% of the total. When we used the RNA libraries of three different tissues of *B. ramiflora* for alignment, the alignment rate was 88.38% for roots, 89.73% for stems, and 90.43% for leaves (Table S2). Finally, we evaluated the *B. ramiflora* genome assembly using the benchmarking universal single-copy orthologs (BUSCO)²¹ and found 2,121 BUSCOs, of which 97.5% were single-copy BUSCOs (86.7% were both unique and complete, and 10.8% were multicopy and complete), 0.8% were BUSCOs with incomplete coverage, and 1.7% were missing BUSCOs, indicating that most of the single-copy genes in the genome of *B. ramiflora* are included without repeats or overassembly. The genome assembly is high-quality and has high coverage.

Gene annotation. The information of assembled repeated sequences of *B. ramiflora* is summarized in Table S3. The assembled repeated sequences accounted for 73.47% (716.93 Mb) of the total sequences, higher than the K-mer estimation (57.5%), indicating the complexity of the *B. ramiflora* genome. Long terminal repeat (LTR)-retrotransposons are a class of transposon with the highest proportion of repeat sequences in plant genomes and the main cause of genome expansion, so high-quality sequences of LTR-retrotransposons are crucial to a genome assembly^{22,23}. Sequences of LTR-retrotransposons in the *B. ramiflora* genome, which accounted for 52.1% of the total sequence, were mainly of the types Copia (89.67 Mb, 9.19% of the genome) and Gypsy (281.22 Mb, 28.82% of the genome), and the LTR-

retrotransposons of other types were 137.53 Mb in length (14.09% of the genome), indicating that the *B. ramiflora* genome contains many unique LTR-retrotransposons. Tandem repeats were very rare, only accounting for 0.09% of the total. Unknown repeat sequences accounted for 15.91% of the total, indicating that the genome contains many new genes and unique genes. Statistical results showed that the size of the *B. ramiflora* genome was mainly determined by the LTR-retrotransposons with their repetitive sequences, which accounted for 70.91% of the total repetitive sequences and 52.1% of the total genome. There were still many unknown repeat sequences in the *B. ramiflora* genome. A total of 3,452 noncoding RNAs were found in the genome, which were mainly snRNA, rRNA, and tRNA sequences and numbered 1981, 674, and 526 respectively.

We performed a *de novo*, homology-based search and RNA-Seq to predict genes from repeat-masked *B. ramiflora* genome sequences. Table 1 shows the final prediction results. A total of 29,172 protein-coding genes, 127.25 Mb in total length, were predicted for the *B. ramiflora* genome, with an average size of 4.4 kb per gene, generating 32,692 transcripts (52.18 Mb in total length, 1.60 kb in size on average), each gene generating an average of 1.1 transcripts. There were a total of 185,435 exons, including 178,416 coding exons (281 bp on average), each transcript containing 5.7 exons on average; there were a total of 152,743 introns (629 bp on average). The total length of coding sequences (CDS) was 40.83 Mb (1,248 bp in size on average). Compared with some species of the Euphorbiaceae family, such as *R. communis* (number of predicted genes: 26,049), *M. esculenta* (number of predicted genes: 34,312), *H. brasiliensis* (number of predicted genes: 42,686), and *J. curcas* (number of predicted genes: 22,718) (data source: NCBI database), *B. ramiflora* has a medium number of genes, which is likely related to its high proportion of repeat sequences throughout its genome.

Table 1
Gene model prediction statistics in *B. ramiflora*

Type	Value
Number of genes	29,172
Total genic length (bp)	127,254,215
Mean gene length (bp)	4,362
Number of transcripts	32,692
Transcripts per gene	1.1
Total transcript length (bp)	52,174,963
Mean transcript length (bp)	1,595
Number of exons	185,435
Exons per transcript	5.7
Exon length (bp)	281
Number of coding exons	178,416
Number of introns	152,743
Mean intron length (bp)	629
coding region Total CDS length (bp)	40,826,528
Average length of coding regions (mean CDS length) (bp)	1,248

We next compared the gene function annotation results with those of five major databases (NR, Swiss-Prot, eggNOG, GO, and KEGG) and found that 25,980 genes (89.06%) had been annotated and 3,192 genes (10.94%) had not, indicating that there were still many unknown genes in the *B. ramiflora* genome, which suggested that *B. ramiflora* may have a highly unique evolutionary process. Some 25,962 (89.00%), 19,939 (68.355%), 24,699 (84.67%), 17,621 (60.40%), 8,724 (9.91%) coding genes had been annotated in the NR, SwissProt, eggNOG, GO, and KEGG databases, respectively. Fewer genes of *B. ramiflora* were found in the KEGG database, likely because *B. ramiflora* is a wild tree species containing many genes of unknown function.

Phylogenetic evolution analysis. To further understand the evolutionary position of *B. ramiflora*, we performed gene homology analysis on *B. ramiflora* and five related angiosperms (*R. communis*, *M. esculenta*, *H. brasiliensis*, *J. curcas*, and *Populus euphratica*) using *Arabidopsis thaliana* as the outgroup, the seven species containing 206,360 protein sequences, including 27,886, 18,174, 26,645, 30,653, 20,267, 28,266, and 22,938 sequences, respectively, with 3,283, 304, 747, 1,024, 542, 875, and 3,081 unique protein sequences and 6,267, 8,440, 5,836, 4,836, 8,214, 4,407, and 6,863 single-copy gene sequences. The result of gene family cluster analysis of five species of the original Euphorbiaceae (*B.*

ramiflora, *R. communis*, *M. esculenta*, *H. brasiliensis*, and *J. curcas*) is shown in Fig. 2. The common gene families contained a total of 11,667 genes, while *B. ramiflora*-specific gene families contained 1,150 genes, the most among the five species, suggesting that *B. ramiflora* has acquired new gene families/genes in the evolutionary process. In *B. ramiflora*-specific gene families, 102 GO enrichment categories, mainly in catalytic activity, transferase activity, kinase activity, hydrolase activity, and nucleic acid metabolic process, were found. Thirteen enriched KEGG pathways, mainly in three aspects, i.e., genetic information (mismatch repair, homologous recombination, basal transcription factors, DNA replication, RNA degradation, nucleotide excision repair, and spliceosome), metabolic pathways (monoterpenoid biosynthesis, isoquinoline alkaloid biosynthesis, glutathione metabolism, tyrosine metabolism, alanine, aspartate, and glutamate metabolism), and plant-pathogen interaction, were found.

We performed clustering analysis on the gene families of *B. ramiflora*, *R. communis*, *M. esculenta*, *H. brasiliensis*, *J. curcas*, *Populus euphratica*, and *A. thaliana* using the OrthoMCL method, in which 1,739 single-copy homologous genes were used to generate the phylogenetic tree. As shown in Fig. 3, *B. ramiflora* clustered into two sister groups with four species of the Euphorbiaceae family, which is consistent with the APG IV system, in which *B. ramiflora* belongs to the *Phyllanthus* genus. We estimated that the divergence between *B. ramiflora* and Euphorbiaceae occurred 59.9 Mya (in the range of 50.4–75.2 Mya), and *B. ramiflora* and *P. euphratica* diverged from Euphorbiaceae 66.6 Mya (in the range of 53.4–75.2 Mya), while *A. thaliana* diverged from the other species 108.0 Mya (in the range of 107–109 Mya).

Gene family expansion and contraction analysis. During the evolutionary process, 173 gene families of *B. ramiflora* expanded, mainly in metabolic pathways, including linoleic acid metabolism, monoterpenoid biosynthesis, alpha-linolenic acid metabolism, sesquiterpenoid and triterpenoid biosynthesis, glyoxylate and dicarboxylate metabolism, peroxisome, diterpenoid biosynthesis, tropane, piperidine and pyridine alkaloid biosynthesis, and anthocyanin biosynthesis; 22 gene families contracted, mainly in kinase activity, transferase activity, phosphorylation, metabolic process, and catalytic activity. Compared with four species of Euphorbiaceae, *B. ramiflora* showed nine collapsed gene families and no expanded gene families, indicating that *B. ramiflora* is closely related to the Euphorbiaceae family.

A total of 278 candidate positive selection genes ($P < 0.05$) were screened through the likelihood ratio test and then subjected to functional enrichment analysis, which registered them to 160 GO pathway categories and four KEGG pathway categories (Fig. 4).

Whole-genome duplication and collinearity analysis. Whole-genome duplication (WGD) events are of great significance in plant evolution and speciation. Accompanied by the doubling of genes, they provide new genes for plant evolution, accelerate the generation of new genes and the expansion of gene families, and improve the plant's ability to adapt to environmental changes, thereby promoting plant evolution. The K_s and transversions at four-fold degenerate sites (4DTv) distribution maps of the homologous genes of *B. ramiflora*, *P. euphratica*, and *M. esculenta* (Fig. 5) showed that the *B. ramiflora* genome had one peak, corresponding to K_s and 4DTv values of 2.4 and 0.6, respectively, which were also

present in both *M.* and *P. euphratica*, suggesting a genome-wide triplication event (γ) that occurred before the eudicot differentiation. No recent genomic duplication signal was detected in the *B. ramiflora* genome.

The *Vitis vinifera* genome contains chromosomal relics of its angiosperm ancestors and has been commonly adopted as a reference genome for analyzing the collinearity of angiosperm species. Through collinearity alignment of the *B. ramiflora* and *V. vinifera* genomes, we found 13,698 pairs of homologous genes. As shown in the dot plot (Fig. 6), the two genomes had many collinear fragments. However, long fragments at the same position hardly matched, suggesting that after their evolutionary divergence, no WGD events have occurred, which is consistent with the K_s and 4DTv values. We found that some fragments showed ambiguous matches at scattered spots in other intervals, which we speculated are relics left by the ancient genome-wide triplication γ event in the *B. ramiflora* genome, supporting the consistent collinearity between *B. ramiflora* and *V. vinifera*.

Discussion

Sequencing evaluation of the *B. ramiflora* genome. *B. ramiflora* is a dioecious plant with high heterozygosity and a complex genome, which makes the whole-genome sequencing and assembly of *B. ramiflora* very challenging. To sequence complex genomes with high heterozygosity and abundant repetitive sequences, it is often difficult to meet the sequencing requirements by relying solely on next-generation sequencing techniques. Third-generation sequencing uses single-molecule real-time sequencing techniques, which has the advantages of no PCR amplification and ultralong reads²⁴. Therefore, in this study, we used the Pac Bio SMRT technique in combination with next-generation sequencing data to correct and assemble the sequences of the complex genome of *B. ramiflora*²⁵.

Similar to whole-genome sequencing of birch (*Betula platyphylla*)²⁶, we found that the long-fragment genomic DNA extracted from *B. ramiflora* had some characteristics that make third-generation sequencing advantageous, such as poor DNA integrity and low concentration, yield, and purity (e.g., viscous samples). The leaves of *B. ramiflora* contain many secondary metabolites, including polyphenols, lactones, and sterols^{19,27,28}, which interfere with the extraction of long DNA fragments, and commercial DNA extraction kits do not meet the quality requirements of DNA extraction. In this study, we obtained high-quality long DNA fragments by first obtaining a large amount of crude DNA with the modified cetyltrimethylammonium bromide (CTAB) method and then further purifying it with the FC-121-4001 kit. Therefore, for woody plants with high heterozygosity, the improved CTAB method combined with a DNA extraction kit can be used for DNA extraction. Even so, the Reads N50 was still low in this study. DNA extraction of woody plants with highly complex genome needs to be further optimized to ensure higher DNA integrity.

***B. ramiflora* genome assembly.** According to their heterozygosity and proportion of repetitive sequences, genomes can be divided into highly heterozygous (heterozygosity > 0.8%), mildly heterozygous (heterozygosity > 0.5%), and highly repetitive (repetitive sequences > 50%)²⁹. The K-mer analysis showed

that the heterozygosity of the *B. ramiflora* genome is 0.634%, and repeat sequences account for 57.5% of the genome, making it a genome with moderate heterozygosity and a high proportion of repeats. It is difficult to assemble complex genomes with high heterozygosity. Therefore, we adopted the Canu + dereundancy method to filter the original subreads with abnormal peaks, which mainly filters the subreads with a GC content in the range of 40–45%/45–50%, and further filtered out subreads highly homologous to mitochondrial and chloroplast sequences, then assembled the remaining sequences using the Canu + deheterozygosity method to achieve the best assembly, which showed a normally distributed GC content with no abnormal peaks. Compared with the whole-genome assembly of *B. platyphylla* with high heterozygosity²⁶, which has a contig N50 of approximately 300 kb, indicating good assembly, the assembled *B. ramiflora* genome, with a high heterogeneity, had an N50 of 509.33 kb, indicating the assembling in this study was very effective. Based on the verification of previous next-generation resequencing and the assembly results of two third-generation libraries, we speculate that the low Contig N50 was caused by the complexity of the *B. ramiflora* genome. The size of the assembled genome was similar to that predicted from the K-mer, indicating that the genome was not overexpanded or collapsed and was thus well assembled.

Genome structure of *B. ramiflora*. Coding genes were predicted using the ab initio prediction, homology annotation, and RNA-seq annotation methods. Functional annotations were performed by aligning the genes with those in five public databases, and 29,172 protein-coding genes were predicted, of which 25,980 have been annotated, accounting for 89.06% of the predicted genes. Plant genomes vary significantly; the amplification of transposons and polyploidy events in genomes increase the number of repetitive sequences as well as the genome complexity^{23,30}. Transposons, especially LTR-retrotransposons that are widespread in plant genomes, drive genome expansion³¹. Repetitive sequences account for 73.47% of the *B. ramiflora* genome sequence, indicating that the *B. ramiflora* genome is highly repetitive. LTR-retrotransposons account for 52.1% of the genome sequence, and the Gypsy-type and Copia-type transposons account for 55.32% and 17.64%, respectively of the repetitive sequences, having the greatest impact on the repeat sequences of the *B. ramiflora* genome. In addition, unknown repeat sequences account for 15.91% of the genome, indicating that *B. ramiflora* has many unique conserved sequences, which is likely related to its evolution and may be of great significance for the expression of specific traits of *B. ramiflora*.

Comparative genomics of *B. ramiflora*. The phylogenetic tree of gene families of seven species, i.e., *B. ramiflora*, *R. communis*, *M. esculenta*, *H. brasiliensis*, *J. curcas*, *P. euphratica*, and *A. thaliana*, shows that *B. ramiflora* clusters into two groups with species of the *Euphorbiaceae* family, which is consistent with previous phylogenetic studies. *B. ramiflora* diverged from *Euphorbiaceae* species 59.9 Mya, while *P. euphratica* diverged from *Euphorbiaceae* species 66.6 Mya, indicating that *B. ramiflora* and *P. euphratica* are closely related to *Euphorbiaceae*, being sister groups to each other. A total of 278 genes have undergone positive selection, indicating that *B. ramiflora* has acquired many new genes in adapting to its environment, which according to the results of collinearity analysis resulted from the relics of the ancient whole-genome triplication γ event. The collinearity between *B. ramiflora* and *V. vinifera* suggests that the

two genomes have consistent collinearity, and after divergence from *V. vinifera*, no further WGD events occurred in *B. ramiflora*. In terms of WGD events and collinearity, *B. ramiflora* and *V. vinifera* have similar genomes and only experienced an ancient genome-wide triploidization event after the emergence of angiosperms, with no recent WGD events. The *B. ramiflora* genome is about twice the size of the *V. vinifera* genome, which is mainly caused by its large number of LTR-retrotransposons.

Materials and methods

Statement. The current study complies with relevant institution, nation, and international guidelines and legislation for experimental research and filed studies on plant, including the collection of plant materials.

Plant material. The material for this study was collected from Mountain xiexie in zhanjiang, Guangdong Province (21°36'11.23"N,110°21'16.45"E) and currently planted in the greenhouse of Hanshan Normal University in Chaozhou, Guangdong Province (23°39'45.44"N,116°37'46.09"E) and stored in liquid nitrogen. Crude genomic DNA was extracted from the young leaves using a modified CTAB method, in which the vortex step was replaced by prolonging the incubation time, and then further purified using the Truseq Nano DNA Sample Preparation Kit (Illumina, cat. number: FC-121-4001).

Sequencing and assessment of genome size. For short-read sequencing, the 450-bp PE library was constructed using the NEBNext Ultra DNA Library Prep Kit and sequenced on the Illumina HiSeq 2500 platform. All raw Illumina sequencing reads were trimmed and filtered using Trimmomatic v0.33 to remove adapters, reads with > 3% N, and low-quality reads (those with > 50% bases with a quality score of $Q < 3$)³². For long-read sequencing, the PacBio SMRT bell library was prepared using the Template Prep Kit, then sequenced under the diffusion loading mode. Genome size was estimated through 19-bp k-mer frequency analysis³³.

Genome assembly and evaluation. *De novo* assembly was performed using Falcon v0.3.0 (<https://github.com/PacificBiosciences/FALCON-integrate>) (length_cutoff = 5000; ength_cutoff_pr = 12000), Canuv1.6(<https://github.com/marbl/canu/releases>) (genomeSize = 1000 m; minReadLength = 2000), and HGAP3 (<https://github.com/PacificBiosciences/smrtmake>) (genomeSize = 1000 m; minReadLength = 2000; other default parameters) on an SGE cluster with multiple computer nodes, each consisting of 120 central processing units and 500 gigabytes of memory. The optimized configuration for the assembly was achieved using the HGAP pipeline, which then served as a reference. Finally, the Illumina data were aligned with the assembled results of the corrected third-gen data using Bwa mem v0.7.12, which was followed by correction with Pilon, ultimately generating the error-corrected assembly result of the third-generation and next-generation sequencing data³³.

The assembly was evaluated through three methods, data mapping ratio evaluation, GC-Depth evaluation, and BUSCO assembly integrity evaluation. The assembly integrity assessment adopted the BUSCO method (<http://busco.ezlab.org/>)³⁴, the intactness of the whole-genome sequencing assembly of

B. ramiflora was assessed, and the completeness of the assembled genome was assessed by aligning the assembled genes of *B. ramiflora* with the conserved sequences of closely related species.

Repeat sequence prediction. The predicted repeats in the genome were masked using RepeatMasker (v4-0-6)³⁵ with the following parameters: -s -nolow -norna-gff -engine ncbi -parallel 20. The transposable elements such as miniature inverted-repeat transposable elements and LTRs were identified through the structure prediction method. Known repeat sequences of *B. ramiflora* were searched for in the RepBase library. Finally, other repeats in *B. ramiflora* genome were collected using the de novo prediction method³⁶. The repeat sequences obtained through the above three methods were integrated to construct the unique repeat sequence library of *B. ramiflora*. The LTR-retrotransposon elements were identified by LTR_FINDER1.0.6³⁷ to construct the de novo repeat library, and their genomic locations were determined using RepeatMasker (v4-0-6).

Genome prediction and annotation. Protein-coding genes were annotated using a comprehensive strategy combining ab initio gene prediction, homology-based gene prediction, and Iso-Seq read-based prediction. The cDNA sequences of proteins were predicted based on homologous genes of five closely related species, *M. esculenta*, *J. curcas*, *H. brasiliensis*, *R. communis*, and *P. alba*, using GeMoMa-1.6.1. First, the genes of *B. ramiflora* were predicted de novo using GlimmerHMM, Augustus, Snap, and GeneMarkES^{38,39}. The RNA-Seq reads were aligned to the scaffold sequences using HISAT2 v2.0.4⁴⁰. The aligned reads were assembled using Cufflinks v2.2.1. Open reading frames were predicted using PASA v2.0.1⁴¹, of which those encoding 100–1000 AA with two or more CDSs that allowed full-length protein sequence alignment against reference sequences were selected. Gene structure was predicted with Augustus v3.0.3³⁸ combined with the RNA-Seq data. Specifically, parameter training was performed on the training set, and then intron hints (i.e., the positional information of predicted introns) were obtained according to the alignment result of RNA-Seq reads and Scaffold (TopHat v2.0.10), based on which, the gene structure was predicted. The gene structure was also predicted using SNAP and GlimmerHMM³⁹. Specifically, first, parameters were trained with the training set, and then the gene structure was predicted on the scaffold with masked repeats sequences. Second, intron hints were obtained using GeneMark-ET v4.5⁴² combined with Augustus v3.0.3, based on which the gene structure of the scaffold with masked repeat sequences was predicted. Finally, the above gene prediction results were integrated using EVM⁴³, and the untranslated region and alternative splicing of coding gene were predicted using PASA.

The gene annotation information was obtained by aligning with NR, Swiss-Prot, and eggNOGv4.5 using DIAMOND⁴⁴. The GO annotation was performed using Blast2GO, so the predicted genes were matched to their respective GO numbers. The KEGG annotation was performed using KOBAS 2.0⁴⁵ to obtain functional annotation of gene products. By individually aligning the genes to each of the five major databases, the corresponding gene function annotation results were obtained, and then the gene function annotation results were integrated to draw Venn diagrams to generate more accurate gene function annotation information.

Contract and expansion of *B. ramiflora* gene families and phylogenetic analysis. Homologous gene clustering analysis was conducted on *B. ramiflora* and five closely related species with well-annotated genomes, i.e., *R. communis*, *M. esculenta*, *H. brasiliensis*, and *J. curcas*, and *P. euphratica*, taking *A. thaliana* as the outgroup species and applying the OrthoMCL method⁴⁶. Specifically, the protein sequences of the above seven species were selected, of which the transcripts with the longest alternatively spliced coding regions were retained, while those coding for shorter than 50 amino acids were excluded, and then the protein-coding sequences were determined with all-vs.-all BLAST alignment, with the e value set to 1e-5, to calculate the similarity between the sequences. The clustering analysis was performed using the Markov clustering algorithm. The clustering result of protein families was shown in a Venn diagram. Unique gene families of *B. ramiflora* were obtained through protein family cluster analysis. The expansion and contract of gene families of the above-described seven species in the evolutionary process were analyzed in Cafe software.

Based on the clustering results of protein families, phylogenetic analysis of *B. ramiflora* was performed. Specifically, single-copy gene families of the seven species were selected as reference markers, and supergenes were constructed through 4DTv. Multiple-sequence alignment on the seven species was performed using Mafft v7.294⁴⁷, and the poorly matched regions were filtered using Gblocks v.0.91b⁴⁸ to select the most suitable base substitution model to obtain sequence files with better alignment results. The confidence of each branch was tested using Rax ML v8.0.19⁴⁹, the GTRGAMMA model in Maximum Likelihood (ML)⁵⁰ and 1000 bootstrap runs before the phylogenetic analysis. The divergence time was estimated based on phylogenetic analysis, by extracting the fourfold degenerate sites through the CDS alignment of single-copy gene family sequences and then generating complete sequences according to the order of species and genes. The divergence time was estimated using the MCMCtree program of PAML v4.6 software⁵¹, with the following parameters: burn in = 10,000; sample number = 100,000; sample frequency = 2; clock = 2. In reference to the known differentiation times of other species in the database of the Time Tree (<http://www.timetree.org/>)⁵², the divergence time of *B. ramiflora* was estimated.

Whole-genome duplication event and collinearity analysis. The protein sequences of homologous genes of *B. ramiflora*, *P. euphratica*, and *M. esculenta* were pairwise aligned, and the ratio of nonsynonymous substitution rate (K_a) over the synonymous substitution rate (K_s) ($\omega = K_a/K_s$) was calculated using the KaKs_Calculator v.2.0 software⁵³, and the 4D sites and 4DTv values of homologous genes were analyzed to obtain the genes with more than 20 sites. The divergence events between *B. ramiflora* and other species and their respective WGD events were predicted based on the 4DTv values. The collinear regions within *B. ramiflora* genome and the collinearity between *B. ramiflora* and *V. vinifera* were detected by MCScan v0.8⁵⁴ with the rule that each collinear region contain at least 20 homologous genes in the same order, based on which the collinearity map was generated. The multiple CDSs of single-copy gene families of the seven species were used for alignment, and the branch site-specific model in the codeml tool of PAML was used to detect whether each gene family had been positively selected in the subbranch of *B. ramiflora*, based on the ω value of the coding sequence (if $\omega > 1$, then the gene family has undergone positive selection), which was then tested based on the likelihood ratio of the two hypotheses.

Declarations

Data availability

Genome data has been uploaded to NCBI and will not be published until the paper is published. Please reference PRJNA845134 in your publication.

Acknowledgments

We would like to thank editor and reviewers who revised this manuscript providing insightful comments. This research was supported by the projects of Forestry Science and Technology Innovation Project of Guangdong province (2018KJCX023), Guangdong Provincial Key Laboratory of Functional Substances in Medicinal Edible Resources and Healthcare Products 2021B1212040015 and school-level projects of Hanshan Normal University (QD202122, QD202123).

Author Contributions

Jianjian Huang: Data curation, Formal analysis, Validation, Investigation, Visualization, Methodology, Writing - Original draft, Writing -Review & Editing, Funding acquisition. Jie Chen: Investigation, Methodology, Visualization. Min Shi: Data curation, Formal analysis, Methodology. Jiaqi Zheng: Data curation, Formal analysis, Methodology. Ming Chen: Data curation, Formal analysis, Methodology. Linjun Wu: Data curation, Formal analysis, Methodology. Hui Zhu: Investigation, Methodology. Yuzhong Zheng: Investigation, Methodology. Qinghan Wu: Investigation, Methodology. Fengnian Wu: Supervision, Conceptualization, Writing - Original draft, Writing - Review & Editing.

Conflict of interest

The authors declare that they have no conflict of interest.

Supplementary data

The online version contains supplementary material available at....

References

1. Li, P.T., 1994. Euphorbiaceae. Flora of China. 44(1), 131.
2. Goyal, A.K., Mishra, T., Sen, A., 2013. Antioxidant profiling of Latkan (*B. ramiflora* Lour.) wine. Indian J. Biotechnol. 12, 137–139.
3. Bhowmick, N., 2011. Some lesser known minor fruit crops of northern parts of West Bengal. Acta. Hortic. 2011, 890, 61–63.
4. Wang, H.J., Xing, Y.Q., Lin, S., Luo, Z.W., Gu, Y.F., 2013. Research and application of *Baccaurea ramiflora* fruit resources. Mod. Agr. Sci. Tech. 2013, 21, 122–123.

5. Chen, J., Wu, F.N., Wang, H.C., Guo, C.C., Zhang, W.G., Luo, P.S., Zhou, J., Hao, W.W., Yang, G.Y., Huang, J.J., 2023. Identification of key taste components in *Baccaurea ramiflora* Lour. fruit using non-targeted metabolomics. Food Sci. Hum. Well. 2023, 12, 94–101.
6. Hu, J.X., Xiao, C.F., Zheng, L.L., 2003. *Baccaurea ramiflora*: a wild fruit tree. South Chin. Fruits 2003, 32(4), 49.
7. Luo, P.S., Zhou, J., Chen, M.K., Xu, J., Lu, Y.C., Huang, L.J., Deng, X., 2014. Investigation on the germplasm resources and selection of elite individual plants of *Baccaurea ramiflora* in Guangxi. South Chin. Fruits 2014, 43(6), 82–86.
8. Li, W.X., Wei, C.Z., Kong, F.N., Lu, Y.C., Luo, P.S., Wei, Y., Yang, Z.Q., Zhou, J., 2015. Determination of soluble saccharide contents of *Baccaurea ramiflora* Lour. using anthrone colorimetric method. Chin. Hortic. Abstr. 2015, 12, 7–8.
9. Pandey, Y., Upadhyay, S., Bhatt, S.S., Sharma, L., Chanbisana, C., 2018. Nutritional compositions of *Baccaurea sapida* and *Eleaocarpus sikkimnesis* of Sikkim Himalaya. Int. J. Curr. Microbiol. App. Sci. 2018, 7(2), 2101–2106.
10. Inta, A., Trisonthi, P., Trisonthi, C., 2013. Analysis of traditional knowledge in medicinal plants used by Yuan in Thailand. J. Ethnopharmacol 2013, 149(1), 344–351.
11. Lin, Y.F., Yi, Z., Zhao, C.H., 2003. Color Atlas of Dai Medicine in China. Kunming: Yunnan Nationality Press, 2003, 380.
12. Rahim, Z.B., Rahman, M.M., Saha, D., Hosen, S.M.Z., Paul, S., Kader, S., 2012. Ethnomedicinal plants used against jaundice in Bangladesh and its economical prospects. Bull Pharm. Res. 2012, 2(2), 91–105.
13. Kalita, D., Saikia, J., Mukherjee, A.K., Doley, R., 2014. An ethnomedicinal survey of traditionally used medicinal plants for the treatment of snakebite in Moriga on district of Assam India. Int. J. Med. Aroma. Plants 2014, 4(2), 97–106.
14. Saha, S., Gouda, T.S., Srinivas, S.V., 2017. Preliminary phytochemical analysis and oral acute toxicity study of the leaves of *Baccaurea ramiflora* and *Microcos paniculata*. Saudi. J. Med. Pharm. Sci. 2017, 3(6), 444–449.
15. Usha, T., Middha, S.K., Bhattacharya, M., Lokesh, P., Goyal, A., 2014. Rosmarinic acid a new polyphenol from *Baccaurea ramiflora* Lour. leaf: a probable compound for its anti-inflammatory activity. Antioxidants 2014, 3(4), 830–842.
16. Mann, S., Sharma, A., Biswas, S., Gupta, R.K., 2015. Identification and molecular docking analysis of active ingredients with medicinal properties from edible *Baccaurea sapida*. Bioinformation 2015, 11(9), 437–443.
17. Pan, Z.H., Ning, D.S., Huang, S.S., Wu, Y.F., Ding, T., Luo, L., 2015. A new picrotoxane sesquiterpene from the berries of *Baccaurea ramiflora* with antifungal activity against *Colletotrichum gloeosporioides*. Nat. Prod. Res. 2015, 29(14), 1323–1327.
18. Saha, M.R., Dey, P., Chaudhuri, T.K., Goyal, A.K., Sen, A., 2016. Assessment of haemolytic cytotoxic and free radical scavenging activities of an underutilized fruit *Baccaurea ramiflora* Lour. (Roxb.)

- Muell. Indian J. Exp. Biol. 2016, 54(2), 115–125.
19. Usha, T., Pradhan, S., Goyal, A.K., Dhivya S., Middha, S.K., 2017. Molecular simulation-based combinatorial modeling and antioxidant activities of Zingiberaceae family rhizomes. *Pharmacogn. Mag.* 2017, 13(Suppl 3), S715.
 20. Nesa, M.L., Karim, S.S., Api, K., Sarker, M.M.R., Islam, M.M., Kabir, A., Sarker, M.K., Nahar, K., Asadujjaman, M., Munir, M.S., 2018. Screening of *Baccaurea ramiflora* (Lour.) extracts for cytotoxic analgesic anti-inflammatory neuropharmacological and antidiarrheal activities. *BMC. Complem. Altern. M.* 2018, 18(1), 35.
 21. Daccord, N., Celton, J.M., Linsmith, G., Becker, C., Choisne, N., Schijlen, E., Geest, H., Bianco, L., Micheletti, D., Velasco, R., Pierro, E.A.D., Gouzy, J., Rees, D.J.G., Guérif, P., Muranty, H., Durel, C.E., Laurens, F., Lespinasse, Y., Gaillard, S., Aubourg, S., Quesneville, H., Weigel, D., Weg, E.V.D., Troggio, M., Bucher, E., 2017. High-quality *de novo* assembly of the apple genome and methylome dynamics of early fruit development. *Nat. Genet.* 2017, 49, 1099.
 22. Ouyang, S., Buell, R., 2004. The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. *Nucleic Acids Res.* 2004, 32(90001), 360.
 23. Xu, L., Zhang, Y., Su, Y., Liu, L., Yang, J., Li, ZC., 2010. Structure and evolution of full-length LTR retrotransposons in rice genome. *Plant Syst. Evol.* 2010, 287(1), 19–28.
 24. Liu, Y.H., Wang, L., Yu, L., 2015. Principle and application of single-molecule real-time sequencing technology. *Hereditas* 2015, 37(3), 259–268.
 25. Yang, Y., Du, X.J., Liang, B., Guo, J.D., Cheng, X.Z., Wang, S., 2015. Development of third-generation sequencing technologies and related bioinformatics. *Food Res. Develop.* 2015, 36(10), 143–147.
 26. Wang, S., 2019. Whole genome sequencing and analysis of *Baccaurea platyphylla* (*B. platyphylla*). Harbin: Northeast Forestry University, 2019.
 27. Xu, J., Lin, Q., Liang, Z.Y., Deng, S.M., Zhong, C.X., 2007. Study on chemical constituents of essential oils from *Baccaurea ramiflora* Lour. fruit, leaf and root. *Food Sci.* 2007, 28(11), 439–442.
 28. Ning, D.S., Wu, Y.F., Lv, S.H., 2014. Chemical constituents of the stem and leaves of *Baccaurea ramiflora*. *Guihaia* 2014, 34(2), 160–162.
 29. Wu, Y.T.N., Du, H.Y., Li, F.D., 2014. Advances in whole genome sequencing of *Eucommia ulmoides*. Beijing: Social Science Literature Publishing House, 2014, 3.
 30. Moghe, G.D., Shiu, S.H., 2014. The causes and molecular consequences of polyploidy in flowering plants. *Ann. NY. Acad. Sci.* 2014, 1320(1), 16–34.
 31. Storz, G., 2002. An expanding universe of noncoding RNAs. *Science* 2002, 296(5571), 1260–1263.
 32. Vurture, G.W., Sedlazeck, F.J., Nattestad, M., Underwood, C.J., Fang, H., Gurtowski, J., Schatz, M.C., 2017. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* 2017, 33, 2202–2204.
 33. Koren, S., Rhie, A., Walenz, B.P., Dilthey, A.T., Bickhart, D.M., Kingan, S.B., Hiendleder, S., Williams, J.L., Smith, T.P.L., Phillippy, A.M., 2018. De novo assembly of haplotype-resolved genomes with trio

- binning. *Nat. Biotechnol.* 2018, 36(12), 1174–1182.
34. Seppey, M., Manni, M., Zdobnov, E.M., 2019. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods Mol. Biol.* 2019, 1962, 227–245.
35. Tarailo-Graovac, M., Chen, N., 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* 2009, 4(10), 1–14.
36. Price, A.L., Jones, N.C., Pevzner, P.A., 2005. *De novo* identification of repeat families in large genomes. *Bioinformatics* 2005, 21(1), 351–358.
37. Xu, Z., Wang, H., 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* 2007, 35, W265-268.
38. Mario, S., Rasmus, S., Stephan, W., Burkhard, M., 2004. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.* 2004, 32(2), 309–312.
39. Majoros, W.H., Pertea, M., Salzberg, S.L., 2004. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* 2004, 20(16), 2878–2879.
40. Kim, D., Langmead, B., Salzberg, S.L., 2015. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* 2015, 12(4), 357–360.
41. Roberts, A., Pimentel, H., Trapnell, C., Pachter, L., 2011. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* 2011, 27(17), 2325–2329.
42. Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y. O., Borodovsky, M., 2008. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Res.* 2008, 18(12), 1979–1990.
43. Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R., Wortman, J.R., 2008. Automated eukaryotic gene structure annotation using Evidence Modeler and the Program to Assemble Spliced Alignments. *Genome Biol.* 2008, 9(1), 7–16.
44. Buchfink, B., Xie, C., Huson, D.H., 2015. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 2015, 12(1), 59–60.
45. Xie, C., Mao, X.Z., Huang, J.J., Ding, Y., Wu, J.M., Dong, S., Kong, L., Gao, G., Li, C.Y., Wei, L.P., 2011. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids. Res.* 2011, 39(2), 316–322.
46. Li, L., Stoeckert, C.J., Roos, D.S., 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 2003, 13(9), 2178–2189.
47. Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 2013, 30(4), 772–780.
48. Castresana, J., 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* 2000, 17(4), 540–552.
49. Stamatakis, A., 2006. RAX ML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 2006, 22(21), 2688–2690.

50. Ronquist, F., Huelsenbeck, J.P., 2003. Mr Bayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 2003, 19(12), 1572–1574.
51. Yang, Z., 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 2007, 24(8), 1586–1591.
52. Hedges, S.B., Dudley, J., Kumar, S., 2006. Time Tree: a public knowledge-base of divergence times among organisms. *Bioinformatics* 2006, 22(23), 2971–2972.
53. Zhang, Z., Li, J., Zhao, X.Q., Wang, J., Wong, K.S., Yu, J., 2006. Ka Ks_Calculator: calculating Ka and Ks through model selection and model averaging. *Genome Proteom. Bioinf.* 2006, 4(4), 259.
54. Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M., Paterson, A.H., 2008. Synteny and collinearity in plant genomes. *Science* 2008, 320(5875), 486–488.

Figures



Figure 1

Ripening fruit of *B. ramiflora*.

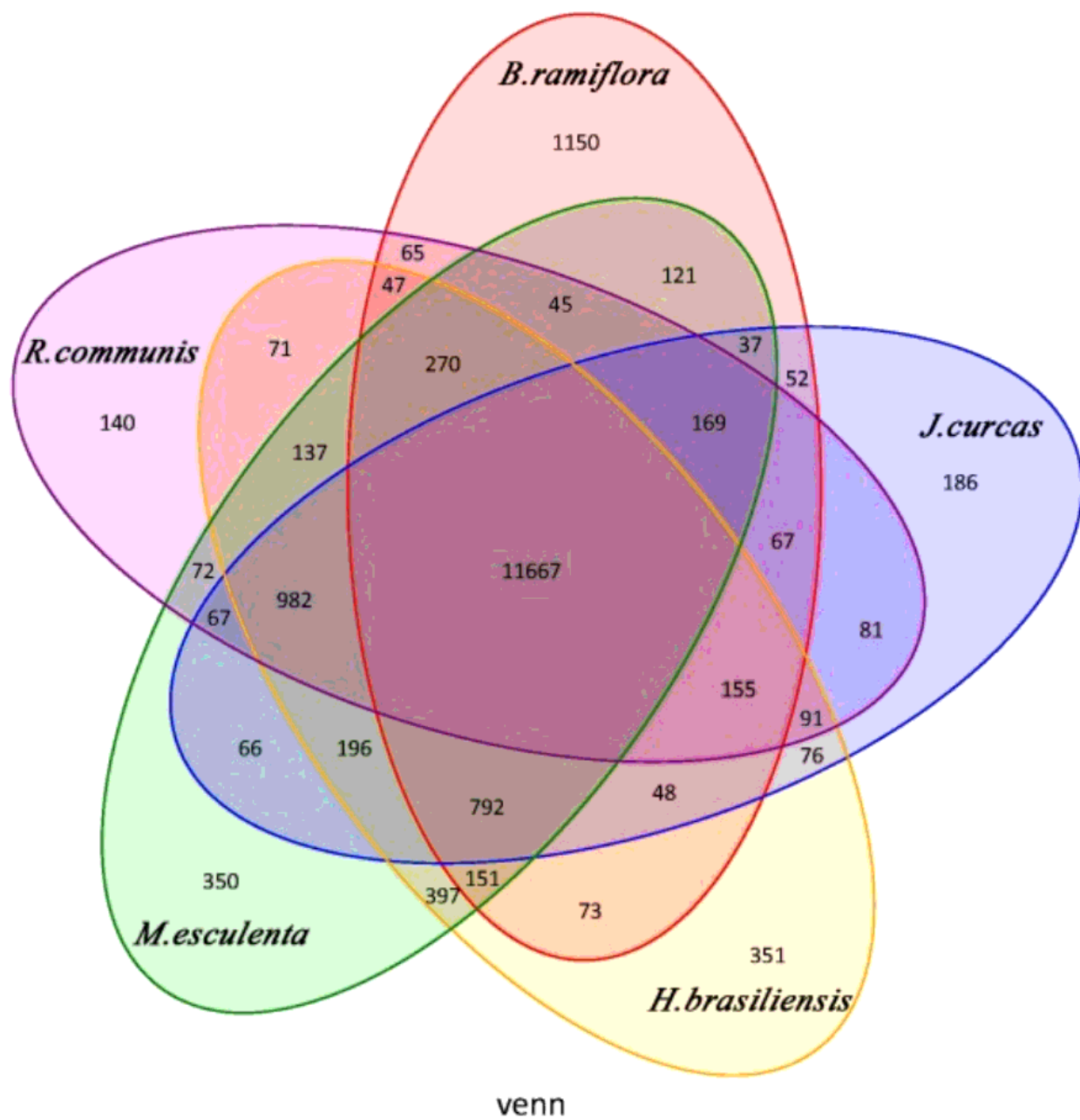


Figure 2

Venn diagrams of the gene families of five species of the original Euphorbiaceae.

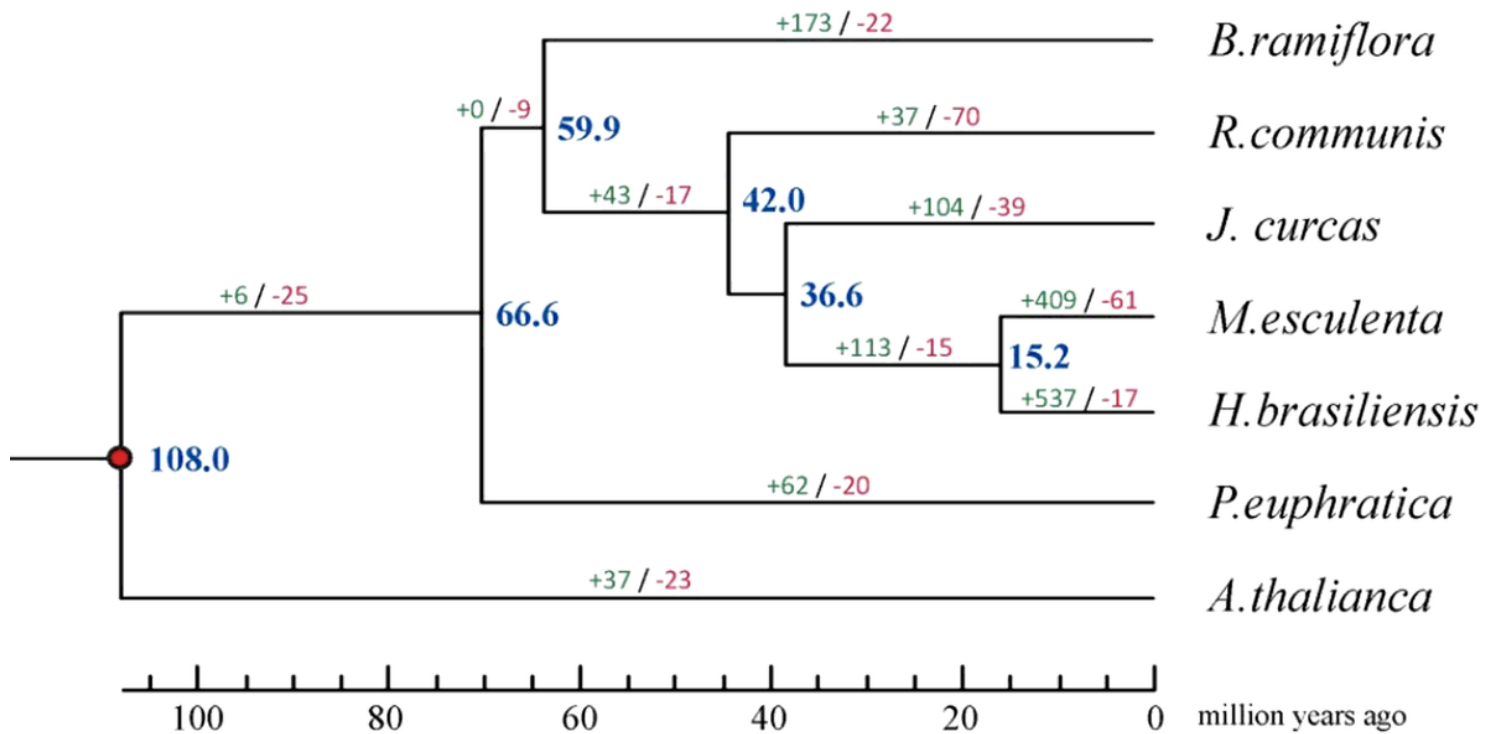


Figure 3

Phylogenetic tree construction of *B. ramiflora* based on single copy homologous sequences.

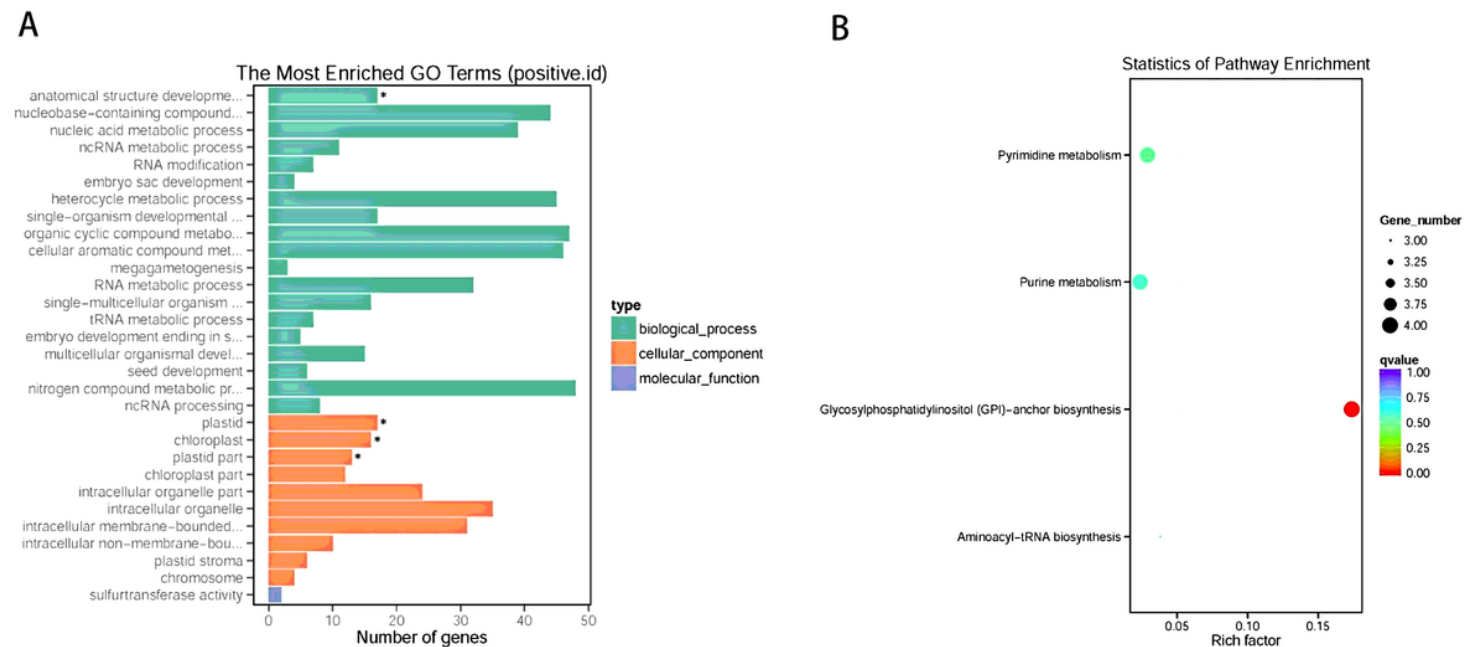


Figure 4

Positive selection analysis: GO enrichment (A) and KEGG (B) enrichment pathways of positive selection candidate genes.

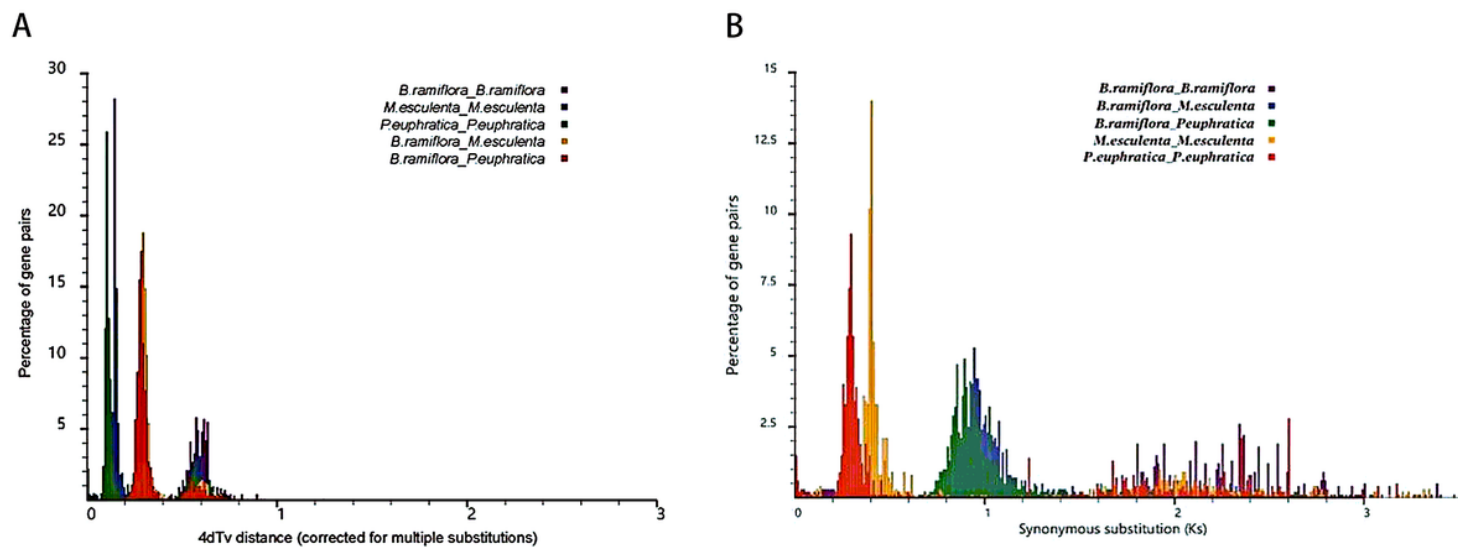


Figure 5

Genome-wide replication events: Pairwise K_s (A) and 4dTv (B) values of all paralogous gene pairs within *B. ramiflora*.

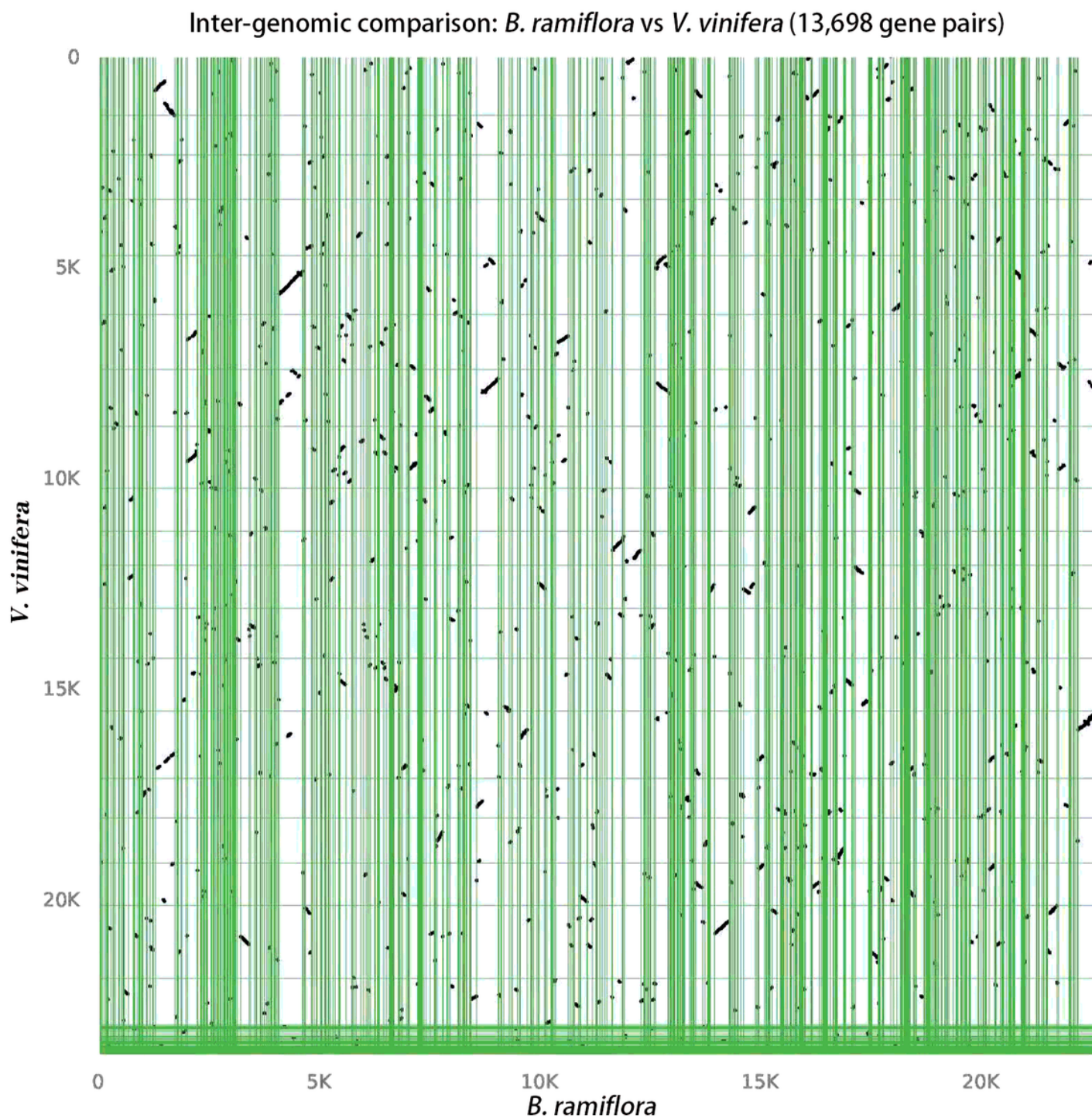


Figure 6

Intergenomic syntentic dot plot within *B. ramiflora* vs *V. vinifera*.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementarymaterial.docx](#)