Chloroplast phylogenomics and the dynamic evolution of Santalum (Santalaceae)

Xiaojin Liu
Research Institute of Tropical Forestry, Chinese Academy of Forestry
Daping Xu (gzfsrd@163.com)
Research Institute of Tropical Forestry, Chinese Academy of Forestry
Zhou Hong
Research Institute of Tropical Forestry, Chinese Academy of Forestry
Ningnan Zhang
Research Institute of Tropical Forestry, Chinese Academy of Forestry
Zhiyi Cui
Research Institute of Tropical Forestry, Chinese Academy of Forestry

Research Article

Keywords: Santalum, sandalwood, chloroplast genome, indels, SSR, inversion

Posted Date: February 8th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-154608/v1

Abstract

Background

Santalum (Santalaceae, sandalwood) is a hemiparasitic genus including approximately 15 extant species. It is known for its aromatic heartwood oil, which is used in incense and perfume. Demand for sandalwood-based products has led to drastic over-harvesting, and wild *Santalum* populations are now threatened. Knowledge of the phylogenetic relationships and genetic diversity will be critical for the conservation and proper management of this genus. Here, we sequenced the chloroplast genome of 11 *Santalum* species. The data were then used to investigate the chloroplast genome evolutionary dynamics and relationships and divergence time within *Santalum* and related species.

Results

The *Santalum* chloroplast genome contains the typical quadripartite structures, ranging from 143,291 to 144,263 bp. The chloroplast genome contains 124 genes. The whole set of *ndh* genes and the *infA* gene were found to lose their function. Between 17 and 31 SSRs were found in the *Santalum* chloroplast genome, and mononucleotide simple sequence repeats (SSRs) were the major type. The P-distance among the *Santalum* species was 0.0003 to 0.00828. Three mutation hotspot regions, 14 small inversions, and 460 indels events were discovered in the *Santalum* chloroplast genome. Our phylogenomic assessment provides improved resolution compared to past analyses. Our divergence time analysis shows that the crown age of *Santalum* was 8.46 Mya, the first divergence occurred around 6.97 Mya, and diversification was complete within approximately 1 Mya.

Conclusions

By sequencing the 12 chloroplast genomes of *Santalum*, we gain insight into the evolution of its chloroplast genomes. The chloroplast genome sequences had sufficient polymorphic information to elucidate the evolutionary history of *Santalum*.

Background

Sandalwood (*Santalum* L., Santalaceae) is known for its aromatic heartwood oil, which is used in incense and perfume [1]. *Santalum* comprises 15 extant species and about 14 varieties, which are widely distributed in India, Australia, and the Pacific Islands [2]. More than a quarter of *Santalum* species are distributed in the Hawaiian Islands, which is one of the distribution centers of the genus [2, 3]. All sandalwoods are hemiparasitic plants, taking a portion of their water and nutrients from the roots of host plants.

Because of its high value and the demand for the valuable sandalwood oil, sandalwood is one of the most heavily exploited plants throughout its range. The endemic species (*S. fernandezianum*) became extinct during the last century due to human exploitation [4]. *S. freycinetianum* var. *lanaiense* from the Hawaiian Islands, *S. insulare* var. *hendersonense* from Henderson Island, and *S. boninense* from Bonin Islands and *S. insulare* from the Cook Islands and French Polynesia [5] are now rare or threatened by extinction [6, 7]. Understanding the relationships of these species and genetic diversity is critical for conservation and proper management of this genus.

Various taxonomies have been proposed and there has been much debate about the relationships within *Santalum* [8, 9]. However, there remains confusion on some species, where the “distinction between taxa is often not clear-cut” [8], with morphological diversity within ranges and even within populations, for example, the red-flowered taxa, *S. haleakalae* and *S. freycinetianum* [10]. Another issue is that island populations with only weak morphology differences have been separated into different species. Meanwhile, morphologically similar populations on different islands were sometimes separated into distinct species [11].

Compared to using morphology alone, molecular data plus morphological data offer a better opportunity to discover cryptic species and to reveal evolutionary histories [12]. Several studies have used molecular data (e.g., the chloroplast genome marker *trnK* intron and nuclear DNA markers ITS and ETS) to infer the phylogenetic relationships of *Santalum* [2, 3, 10, 12, 13]. However, the phylogenetic resolution achieved in these studies has been insufficient to confidently determine the evolution history of *Santalum*. Therefore, sampling more genetic characters, such as complete chloroplast genomes, can provide better phylogenetic resolution to help address the relationships within *Santalum.*

Chloroplast genomes are usually inherited uniparentally, without recombination and, thus, effective expansion of the genetic information. Phylogenetic analyses based on whole chloroplast genome sequences have been widely used at different taxonomic levels [14–16]. Chloroplast genome data also provide effective genetic markers to resolve complex evolutionary histories [17, 18]. Also, a comparison of chloroplast sequences can help understand the evolutionary patterns of plants.

To better resolve the relationships within *Santalum* and to gain insight into the pattern of *Santalum* chloroplast genome evolution, we sequenced the chloroplast genome of 11 species of *Santalum.* Specifically, we attempted to (1) investigate the relationships within *Santalum*, (2) estimate the divergence time of *Santalum*, and (3) elucidate chloroplast genome evolution within *Santalum.*

Results

Structural characteristics of the Santalum chloroplast genome

The complete chloroplast genomes of *Santalum* species were assembled into circular molecules and contained the typical quadripartite structures (Fig. 1 and Supplemental Table S1). The *Santalum* chloroplast genomes ranged from 143,291 bp (*S. acuminatum*) to 144,263 bp (*S. boninense*) in length (Table 1), with LSCs (large single copies) of 82,944 bp (*S. acuminatum*) to 83,942 bp (*S. paniculatum*), IRs (inverted repeat) of 24,477 bp (*S. paniculatum*) to 24,511 bp (*S. album*), and SSCs (small single copy) of 11,237 (*S. leptocladum*) to 11,379 bp (*S. acuminatum*). The overall GC content was 38.0%.

Table 1
Characteristics of newly sequenced *Santalum* chloroplast genomes.

	Nucleatide length (bp)				Nmuber of genes				
Species	Total	LSC	IR	SSC	Protein	tRNA	rRNA	Total	Genbank Accession number
S. acuminatum	143291	82944	24484	11379	67	30	4	101	MW464925
S. album	144034	83793	24488	11265	67	30	4	101	MW464915
S. album	144101	83802	24511	11277	67	30	4	101	MW464922
S. boninense	144263	83912	24501	11349	67	30	4	101	MW464916
S. ellipticum	144250	83911	24495	11349	67	30	4	101	MW464917
S. ellipticum var. *littorale*	144255	83911	24498	11348	67	30	4	101	MW464920
S. freycinetianum var. *pyrularium*	143895	83582	24481	11351	67	30	4	101	MW464921
S. leptocladum	143801	83576	24494	11237	67	30	4	101	MW464918
S. sp.	143923	83603	24489	11342	67	30	4	101	MW464919
S. paniculatum	144239	83942	24477	11343	67	30	4	101	MW464914
S. spicatum	143638	83314	24495	11334	67	30	4	101	MW464924
S. yasi	144019	83736	24497	11289	67	30	4	101	MW464923

The *Santalum* chloroplast genomes had 72 protein-coding genes, 35 tRNA genes, eight rRNA genes, and nine pseudogenes. The whole set of *ndh* genes and the *infA* gene were found to have lost their function. The *ndhA* gene had complete loss of function, and the other *ndh* genes and *infA* were pseudogenizations. Sixteen genes have introns, with two (*ycf3* and *clpP*) harbor two introns.

Comparative analyses of the chloroplast genome

A total of 17–31 SSRs were found in the *Santalum* chloroplast genomes. Mono-, di-, tri -, tetra-, penta-, and hexanucleotide SSRs were all discovered (Fig. 2 and Supplemental Table S2). The majority of SSRs were mononucleotide repeats in all *Santalum* species, followed by tetranucleotide repeats. Tri- and tetranucleotide repeats were not found, and dinucleotide repeats were limited to one

occurrence in *S. acuminatum* and *S. album*. Most of the mononucleotide repeats were composed of A/T, with very little G/C. The LSC region contained the largest number of SSRs (184), with 39 identified in the SSC region and 66 in the IR region.

The chloroplast genomes were plotted using mVISTA and with *S. leptocladum* as the reference. The results revealed collineation, no rearrangement, and high sequence similarity across the chloroplast genome (Fig. 3). There were 2,352 variable sites in the 145,671-bp *Santalum* chloroplast genome alignment (Table 3). The overall nucleotide diversity (π) was 0.0036. The SSC exhibited the highest π value (0.00926), compared with the IR (0.00087) and LSC (0.00457) regions. The genetic p-distance and number of nucleotide substitutions among these ten *Santalum* species are given in Supplemental Table S3. The mean genetic distance was 0.00401, the lowest divergence (p-distance: 0.0003) was between *S. ellipticum* and *S. ellipticum* var. *littorale*, and the largest sequence divergence (p-distance: 0.00828) was between *S. spicatum* and *S. yasi*.

Table 3
Sequences divergence of *Santalum* chloroplast genomes.

Regions	Alignment length (bp)	Number of variable sites			Nucleotide polymorphism	
		Polymorphic	Singleton	Parismony informative	Nucleotide diversity	Haplptypes
LSC	84,949	1,704	1,149	549	0.00457	12
SSC	11,527	455	298	157	0.00926	12
IR	24,617	96	68	28	0.00087	11
Whole plastomes	145,671	2,352	1,582	764	0.00366	12

To identify the mutation hotspots in the chloroplast genome, the nucleotide diversity values are displayed in Fig. 4. The number of single nucleotide substitutions ranged from 0 to 46, and the π value ranged from 0 to 0.01485 within a 800-bp sliding window size. We defined the mutation hotspots with pi values > 0.012. There were three regions (*ccsA–trnL*, *ΨndhE–ΨndhG-rps15*, and *ycf1*), and those three regions were all located within the SSC region. Among these three regions, the *ccsA-trnL* had the highest nucleotide diversity values.

The most commonly employed loci used in plant phylogeny and DNA barcoding (e.g., *rbcL*, *matK*, *trnH-psbA*) were not selected in our study. We compared the sequence divergence of highly variable regions and the three conventional candidate chloroplast DNA barcodes (*matK*, *rbcL*, and *trnH-psbA*). Sequence variation values, such as genetic distance, nucleotide diversity, and the number of variable sites, are given in Supplemental Table S4. The three newly identified markers had higher genetic divergence and had more information sites than the three conventional candidate chloroplast DNA markers. The primers designed for the three variable markers are given in Supplemental Table S5.

Microstructural mutation variable

Among the chloroplast genomes of *Santalum* species, there were 460 indels in total, including 269 normal indels, 104 repeat-related indels, and 87 SSR-related indels. Most of the indels (77.17%, 355 times) were in the spacer regions, 57 indels were found in the intron regions (12.39%), 26 indels occurred in the pseudogene regions, and 22 indels in the exon regions. All SSR-related indels were located in non-coding regions. The length of normal indels ranged from 1 to 331 bp (Fig. 5), and 1-bp indels were the major type (37.92%). The longest normal indel occurred in the *ycf4-cemA* region, and was a deletion in *S. spicatum*. Repeat related indels ranged from 2 bp to 28 bp; the longest indel was located in *atpH-atpI* and was an insert in *S. boninense*, *S. paniculatum*, and *S. ellipticum* var. *littorale*. Most of the repeat-related indels were 4 to 6 bp long (71.15%). A total of 109 regions had indels: *ycf3-trnS* had 17 indels, followed by *trnL-rpl32* (15 indels), *rps16-trnQ* (14 indels), *atpH-atpI* (13 indels), *petA-psbJ* (12 indels), and *matK-rps16* (12 indels). For the coding regions, the *ycf1* gene had the most indels (9 indels).

Fourteen small inversions were identified in the *Santalum* chloroplast genome. All of the inversions and their inverted repeating franking sequences formed stem-loop structures. The inversions length was 2 to 33 bp, and the franking repeats were from 7 bp to 25 bp. There was no correlation between the length of inversion and the franking repeats sequences. Seven inversions occurred in the LSC regions; four were located in the SSC region, and three in the IR regions. All the inversions were located in the non-coding regions. Five inversions (in *ndhB*, *rpl33-rps18*, *rps15-ycf1*, *trnH-psbA*, and *trnM-atpE*) were specific to *S. acuminatum*. The inversion in *ndhD-*

psaC occurred in *S. spicatum*, while the inversions in *trnL-rpl32* and *petN-psbM* were specific to *S. album*. *S. album* had one sample with inversions at y*cf2-trnL* and *psaJ-rpl33*.

Phylogenetic inference

The **13CPG** dataset matrix included 150,415 nucleotide sites, of which 6,259 were variable sites. The second data matrix, **70g50s**, contained 66 protein-coding genes and four rRNA genes from 50 Santalales species. After excluding ambiguous regions and sites, this dataset contained 56,789 nucleotide sites, of which 13,458 (23.70%) were parsimony-informative sites.

The optimal partitioning scheme using the **70g50s** dataset identified under the Akaike Information Corrected Criterion (AICc) and using strict hierarchical clustering analysis in PartitionFinder (lnL= -263607.113888; AICc = 528592.787194) contained 57 partitions (Supplemental Table S6). The ML tree under the unpartitioned and the three partitioned schemes produced identical topologies (Fig. 6 and Supplemental Figures S1–S3). The ML tree inferred from the **13CPG** and **70g50s** datasets were similar to the phylogenetic relationships of *Santalum* species (Fig. 6).

According to the **70g50s** datasets, we inferred the phylogeny of Santalales. The ML tree showed that all the family was generally resolved and supported a monophyletic clade. *Erythropalum scandens* (Erythropalaceae) were selected as the outgroup, according to the results of Chen et al. and Guo et al. [19, 20]. Ximeniaceae was supported position as early diverging lineages. Loranthaceae and Schoepfiaceae formed a clade (BS = 100/PP = 1). Opiliaceae followed by Cervantesiacea were successive sisters to a clade comprising the remaining Santalales. Santalaceae was sister to Viscaceae plus Amphorogynaceae (BS = 100/PP = 1).

All *Santalum* species formed a monophyletic clade (BS = 100/PP = 1) and were sister to *Osyris wightiana* within Santalaceae. *Santalum* had a shortened branch on the phylogenetic tree, indicating low divergence among *Santalum* species. *S. spicatum* was the first diverging branch. *S. acuminatum* was sister to the remaining species, which formed two lineages. The first lineage included three species (*S. leptocladum*, *S. freycinetianum* var. *pyrularium*, and *S.* sp.) and the second lineage include three branches. Two samples of *S. album* were sister to the remaining species, and the relationships of the three branches were not clear (70g50s: BS = 48/BI = 0.53, 13CPG: BS = 49/BI = 0.72).

The estimated divergence time

Bayesian relaxed molecular clock analyses suggested that the crown age of the Santalales was 113.91 Mya (Fig. 7). The split between the Santalaceae and its closest relatives, Viscaceae and Amphorogynaceae, occurred 81.07 Mya (95% HPD: 71.71–96.27 Mya). The mean crown ages of Santalaceae, Viscaceae, and Amphorogynaceae were 38.44, 47.87, and 6.18 Mya, respectively. The crown age of *Santalum* was 8.46 Mya (95% HPD: 3.8–14.06 Mya) in the later Miocene. The first divergence occurred around 6.97 Mya (95% HPD: 3.03–12 Mya), followed by independent branch-splitting events within the two lineages at 3.02 Mya (95% HPD: 1.41–4.95 Mya). Diversification within the two lineages occurred over a short period of approximately 1 Mya.

Discussion

Santalum **chloroplast genome evolution and variation**

Our findings reveal that the *Santalum* chloroplast genomes have highly similar genome structures, genome sizes, and gene contents (Figs. 3 and 4). Our findings are similar to other chloroplast genome studies reporting that the single-copy regions and non-coding regions are more variable than IRs and coding regions [21, 22]. The variation in size relative to other angiosperm species was mainly due to some missing genes (Fig. 1).

All encoded NAD(P)H dehydrogenase complex (*Ndh*) genes in the *Santalum* chloroplast genome had functional or physical losses. The *ndh* genes are the earliest functional losses in the chloroplast genome of hemiparasites [23, 24]. All the *ndh* genes have losses in the Santalales hemiparasites [19, 20], suggesting that the chloroplast NDH pathway is not essential in these lineages, or this function has been transferred to the nuclear genome [24, 25]. Another degraded chloroplast gene in the *Santalum* chloroplast genome was *infA*; this mutation was detected in most Santalales hemiparasite [20]. The *infA* gene is a translation initiation factor of the translation initiation complex [26]. Loss and pseudogenization of *infA* has occurred in many hemiparasites and holoparasitic plants [24, 27]. This gene has also been independently lost multiple times among photoautotrophic plant lineages [28].

SSRs are important markers for population genetics and germplasm management [29], and chloroplast genome SSRs have been used to analyze the material from introgression [30]. A total of 17 to 31 SSRs were found in the *Santalum* chloroplast genome, and mononucleotide SSRs were the most frequent (Fig. 2). The number of SSR in the *Santalum* chloroplast genome was low compared to photoautotrophic plant lineages [31–34].

Microstructural mutation events are ubiquitous in chloroplast genome evolution, but have been little studied [31]. Indels and small inversions were analyzed in this study (Fig. 5 and Table 2). Based on Dong et al. [31], we classified the indels mutations into three categories: SSR-related indels, repeat-related indels, and normal indels. The normal indels were the most frequent in the *Santalum* chloroplast genome, and the size was also variable, ranging from 1 to 331 bp (Fig. 5). Slipped strand mispairing (SSM) has been suggested as the mechanism leading to most SSR-related indels [35, 36]. DNA recombination has also been proposed to cause repeat related indels [35, 36]. These different mechanisms might be responsible for the observed differences in indel length.

Table 2

The genomic location and length distributions of 14 inversions. Ssp: *S.* sp.; Sl: *S. leptocladum*; Sa-1: *S. album* LXJ-20-002; Sa-2: *S. album* LXJ-20-009; Sb: *S. boninense*; Se: *Santalum ellipticum*; Sp: *S. paniculatum*; Sel: *S. ellipticum* var. l*ittorale*; Sf: *S. freycinetianum* var. *pyrularium*; Sy: *S. yasi*; Ss: *S. spicatum*; Sau: *S. acuminatum*.

			Length (bp)		Inversions											
Region	Position	Location	loop	stem	*Ssp*	*Sl*	*Sa-1*	*Sa-2*	*Sb*	*Se*	*Sp*	*SeL*	*Sf*	*Sy*	*Ss*	*Sac*
LSC	petA-psbJ	spacer	6	25	no	no	yes	yes	yes	yes	no	yes	yes	yes	no	no
LSC	petN-psbM	spacer	12	19	no	no	yes	yes	no	no	no	no	no	no	no	no
LSC	psaJ-rpl33	spacer	14	18	no	yes	no	yes	yes	yes	yes	no	no	no	no	no
LSC	rpl33-rps18	spacer	29	12	no	no	no	no	no	no	no	no	no	no	no	yes
LSC	trnH-psbA	spacer	30	14	no	no	no	no	no	no	no	no	no	no	no	yes
LSC	trnK-rps16	spacer	2	16	no	no	no	no	no	no	no	no	no	yes	yes	no
LSC	trnM-atpE	spacer	18	14	no	no	no	no	no	no	no	no	no	no	no	yes
IR	ndhB	pseudo gene	2	8	no	no	no	no	no	no	no	no	no	yes	yes	yes
IR	ndhB	pseudo gene	8	7	no	no	no	no	no	no	no	no	no	no	no	yes
IR	ycf2-trnL	spacer	4	21	no	yes	no	yes	yes	yes	no	no	no	yes	no	yes
SSC	ndhD-psaC	spacer	30	7	no	no	no	no	no	no	no	no	no	no	yes	no
SSC	rps15-ycf1	spacer	16	19	no	no	no	no	no	no	no	no	no	no	no	yes
SSC	trnL-rpl32	spacer	33	14	no	no	yes	yes	no	no	no	no	no	no	no	no
SSC	trnN-trnR	spacer	4	8	no	no	no	no	yes	yes	yes	yes	no	no	no	yes

Mutation hotspot regions in the chloroplast genome have been identified in most plant lineages, and studies have identified those markers were more variable than universal chloroplast markers [22, 33, 37, 38]. We identified three variable regions (*ccsA–trnL*, *ΨndhE–ΨndhG-rps15*, and *ycf1*) compared with the Santalum chloroplast genome. The *ycf1* gene has been identified in several

lineages, such as *Dalbergia* [39], *Diospyros* [22], and *Quercus* [40]. Studies have shown that *ycf1* is phylogenetically useful [41] and is associated with a high success rate for DNA barcoding [42]. *ccsA–trnL* and *ΨndhE–ΨndhG-rps15* have been less widely used in the phylogeny and DNA barcoding.

Phylogenetic relationships of Santalum

Based on a few morphological characters, such as the floral tube color, placement of the ovary, the *Santalum* has been classified into three or four sections [8, 11]. However, the molecular data was not supported by the sectional classification of *Santalum*. For example, the *S. boninense* from Section *Santalum* was sister to *S. paniculatum* and *S. ellipticum* from Section *Hawaiiensia* (Fig. 6). *S. acuminatum* and *S. spicatum* were in the formerly recognized Australian genus *Eucarya.* The ITS, ETS, and GBSSI sequences (nuclear data) support the notion that these two species form a monophyletic group [2, 3], which conflicts with the chloroplast genome results (Fig. 6). Incomplete lineage sorting and chloroplast genome capture might account for this discordance [43, 44].

There were several auto- and allopolyploid species according to an estimate by measuring the C value [3]. Chloroplast genome data showed the maternal line of the allopolyploid species and were used to identify the maternal progenitor. The allopolyploid species of *S. ellipticum*, *S. paniculatum*, and *S. boninense* formed a clade, and there were no diploids species in this clade, although the maternal parents were unresolved. The progenitors might be extinct or not yet discovered. The biogeography suggests that *Santalum* island colonists tend to be polyploids [3].

Conclusions

The analyzed *Santalum* chloroplast genomes have a similar structure, gene number, and gene order. Diversification of the *Santalum* chloroplast genome is explained by the presence of mutation hotspots regions, small inversions, and the co-occurrence of different types of indel or single nucleotide polymorphisms. The phylogeny and divergence time analysis based on the complete chloroplast genome discovered that the *Santalum* species likely originated by radiation evolution, and most speciation events occurred less than 1 Mya. Owing to the incomplete sampling and the existing polyploids species in *Santalum*, further studies with extended sampling and additional nuclear markers will be necessary.

Methods

Plant materials and sequencing the chloroplast genomes

We collected 12 individual samples representing ten currently described *Santalum* species. Details of the 12 samples collected in this study are given in Supplemental Table S1. Specimens of these samples were preserved in the herbarium of Research Institute of Tropical Forestry, Chinese Academy of Forestry. Xiaojin Liu identified all samples. Leaf tissues were dried using silica gel for subsequent DNA extraction. These materials are from cultivated plants, and permission is not required to collect them.

DNA was extracted using the modified CTAB DNA extraction protocol [45]. We used the low-coverage whole-genome sequencing method to obtain the whole chloroplast genome. Paired-end libraries with 350-bp inserts were prepared and then sequenced in Illumina iSeq X-ten platform at Novogene. Each sample yielded about 4 Gb of data.

Chloroplast genome assembly and annotation

The raw reads were filtered using Trimmomatic v0.36 [46] to remove the adaptors, low-quality reads, and sites. The clean data were used to assemble the chloroplast genome using GetOrganelle [47]. The newly sequenced chloroplast genomes were annotated using Plann [48] using *Santalum album* (GenBank Accession number: MK675809) as the reference. The chloroplast genome maps were visualized using OGDRAW [49]. The complete chloroplast genomes were deposited in GenBank (MW464914 to MW464925).

Genome comparison

The mVISTA program was used to analyze the variation in the *Santalum* chloroplast genomes [50], using the chloroplast genome of *S. leptocladum* as the reference for sequence annotation. SSRs in the chloroplast genome were identified using the MISA software. The parameters implemented in MISA are as follows: repeat units ≥ 10 for mononucleotide, repeat units ≥ 6 for dinucleotides, repeat units ≥ 5 for trinucleotides, repeat units ≥ 4 for tetranucleotides, and repeat units ≥ 3 for pentanucleotides and hexanucleotides.

The whole *Santalum* chloroplast genome alignments were performed with MAFFT [51] and adjusted manually. We used the genetic P-distances and the number of SNPs (single nucleotide substitutions) to assess the divergence among the *Santalum* species. The P-distances and the number of SNPs were calculated using MEGA X [52]. To explore the mutation hotspots in the chloroplast genome, nucleotide diversity (π) was calculated using the software DnaSP v6 [53] by sliding window analysis with a window size of 800 bp and a step size of 100 bp.

Microstructural mutation events

Indels and small inversions were identified based on the aligned chloroplast genome sequence matrix, according to Dong et al. [31]. The indel types were divided into three categories: repeat-related indels, normal indels, and SSR-related indels. Inversions were first identified using the REPtuter program and then checked and confirmed by reexamining the sequence matrix. Inversions form a stem-loop structure, including the inversion sequences and inverted repeat at the opposite franking end [31].

Phylogenetic analyses

Phylogenetic analysis was conducted to elucidate the interspecific phylogenetic relationships within *Santalum* and the phylogenetic positions within Santalales. To make good use of the chloroplast genomes in phylogenetic analyses, we generated two datasets for phylogenetic inference. The first dataset (**13CPG**) was the 12 *Santalum* complete chloroplast genome sequences with *Osyris wightiana* used as an outgroup. The second dataset (**70g50s**) was 70 coding genes, including 12 *Santalum* samples and 38 Santalales species, including nine families.

Maximum likelihood (ML) and Bayesian inference (BI) methods were used to infer phylogenetic relationships. For the **13CPG** dataset, the best-fit model was found with ModelFinder [54]. For the **70g50s** dataset, we used the following partitioning schemes: (1) unpartitioned, (2) partitioned by genes, (3) partitioned with the rcluster algorithm (data pre-partitioned by locus), and (4) partitioned with the hcluster algorithm (data pre-partitioned by locus). All partitioning analyses were run in PartitionFinder 2 [55]. RAxML-NG [56] was run for the ML tree with 500 bootstrap replicates.

Mrbayes v3.2 [57] was used to perform the BI tree. For the **70g50s** dataset, we used the hcluster partitioning scheme for BI analyses because this scheme resulted in the highest log-likelihood in the ML analyses. The **13CPG** dataset used the GTR + G model (the best-fit model from ModelFinder) for BI analyses. The BI analysis was run with two independent chains and prior for 20 million generations with sampling every 1000 generations. The initial 25% of the sampled trees were discarded as burn-in. The stationarity was regarded as having been reached when the average standard deviation of split frequencies remained below 0.01.

Molecular clock dating

The **70g50s** dataset was used to estimate the divergence times of Santalales using five priors. The root age of the tree (crown age of Santalales) was set to 114 Mya (95% HPD: 112–116 Mya) according to the divergence time estimate of angiosperms [58]. The stem age of Loranthaceae was constrained to 72 Mya (95% HPD: 70.4–73.6 Mya) based on the fossil of *Cranwellia* [59, 60] and the result of Liu et al. [61] and the average age of the most recent common ancestor (MRCA) of Loranthaceae was set to 59 Mya (95% HPD: 57.4–60.6 Mya) [61]. The stem age of Viscaceae with 72 Mya (95% HPD: 70.4–73.6 Mya) according to Vidal-Russell and Nickrent [62]. The split age of *Santalum* and *Osyris* was set using the external calibration of 32–48 Mya, as estimated by the fossil-calibrated of Harbaugh and Baldwin [2] and Wikström et al. [58].

The divergence time of Santalales was performed in BEAST2 [63]. BEAST analyses were done using the uncorrelated lognormal relaxed molecular clock model.

The prior tree Yule model was selected, and the Markov Chain Monte Carlo (MCMC) tool was run for 400,000,000 generations with sampling every 10,000 generations. We conducted two separate MCMC runs and used Tracer 1.6 [64] to evaluate convergence and ensure sufficient and effective sample size for all parameters surpassing 200. A maximum credibility tree was then built using TreeAnnotator v2.4.7, with the initial 10% of trees discarded as burn-in.

Abbreviations

AICc: Akaike Information Corrected Criterion; BI: Bayesian analyses; CTAB: Cetyl trimethylammonium bromide; DnaSP: DNA sequences polymorphism; ETS: External transcribed spacer of ribosomal DNA; Gb: Giga base; GTR: General time reversible; HPD:

Highest posterior density; IR: Inverted repeat; ITS: Internal transcribed spacer of ribosomal DNA; LSC: Large single copy; ML: Maximum Likelihood; Mya: Million years ago; Ndh: Encoded NAD(P)H dehydrogenase complex; rRNA: ribosomal RNA; SNP: Single nucleotide polymorphism; SSC: Small single copy; SSR: Simple sequence repeats; tRNA: Transfer RNA; π: Nucleotide diversity.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

The 12 newly assembled chloroplast genomes were deposited in GenBank under the accession numbers of MW464914 to MW464925.

Competing Interests

The authors declare no conflicts of interest.

Funding

This study was supported by the Fundamental Research Funds for the Central Non-profit Research Institution of Chinese Academy of Forestry (CAFYBB2019QB003, CAFYBB2016QB010), the National Natural Science Foundation of China (Granted No. 31500512) and the National Key Research and Development Program of China (Granted No. 2016YFD060060503). The funding agencies had no role in the design of the experiment, analysis, and interpretation of data and in writing the manuscript.

Authors' contributions

XL, DX, and ZH conceived the ideas and designed methodology, XL,ZH,NZ, and ZC collected the data, XL analysed the data, XL wrote the manuscript, DX supervised the project. All authors contributed critically to the drafts and gave final approval for publication.

References

1. Merlin M, VanRavenswaay D: **The History of human impact on the genus Santalum in Hawai'i**. In: *In: Hamilton, Lawrence; Conrad, C Eugene, technical coordinators Proceedings of the Symposium on Sandalwood in the Pacific; April 9–11*, 1990; *Honolulu, Hawaii Gen Tech Rep PSW-GTR-122 Berkeley, CA: Pacific Southwest Research Station, Forest Service, US Department of Agriculture: p 46–60: 1990*. 6–60.
2. Harbaugh DT, Baldwin BG: **Phylogeny and biogeography of the sandalwoods (Santalum, Santalaceae): repeated dispersals throughout the Pacific**. *Am J Bot* 2007, **94**(6):1028–1040.
3. Harbaugh Danica T: **Polyploid and Hybrid Origins of Pacific Island Sandalwoods (Santalum, Santalaceae) Inferred from Low-Copy Nuclear and Flow Cytometry Data**. *Int J Plant Sci* 2008, **169**(5):677–685.
4. Stuessy TF, Marticorena C, Rodriguez R R, Crawford DJ, Silva O M: **Endemism in the vascular flora of the Juan Fernández Islands**. *Aliso: A Journal of Systematic and Evolutionary Botany* 1992, **13**(2):297–307.
5. Butaud J-F, Rives F, Verhaegen D, Bouvet J-M: **Phylogeography of Eastern Polynesian sandalwood (Santalum insulare), an endangered tree species from the Pacific: a study based on chloroplast microsatellites**. *J Biogeogr* 2005, **32**(10):1763–1774.
6. Maina SL, Pray LA, DeFilipps RA: **A historical note on the endangered Santalum boninensis (Santalaceae) of the Ogasawara Islands: early reports by Takasi Tuyama**. *Atoll Res Bull* 1988.
7. Waldren S, Florence J, Chepstow-Lusty AJ: **Rare and endemic vascular plants of the Pitcairn Islands, south-central Pacific Ocean: A conservation appraisal**. *Biol Conserv* 1995, **74**(2):83–98.
8. Stemmermann L: **Observations on the genus Santalum (Santalaceae) in Hawaii**. *Pacific Science* 1980, **34**:41–53.

9. Hewson H, George A: **Santalaceae**. In: *Flora of Australia*. vol. 22. Collingwood, Victoria, Australia: Australian Government Publishing Service: Canberra; 1984: 29–67.
10. Harbaugh DT, Oppenheimer HL, Wood KR, Wagner WL: **Taxonomic revision of the endangered Hawaiian red-flowered sandalwoods (Santalum) and discovery of an ancient hybrid species**. *Syst Bot* 2010, **35**(4):827–838.
11. Wagner WL, Herbst DR, Sohmer SH: **Manual of the flowering plants of Hawai**: University of Hawai'i Press; 1999.
12. Harbaugh DT: **A taxonomic revision of Australian northern sandalwood (Santalum lanceolatum, Santalaceae)**. *Aust Syst Bot* 2007, **20**(5):409–416.
13. Lichao J, Tuo H, Eleanor ED, Yonggang Z, Andrew JL, Yafang Y: **Applicability of chloroplast DNA barcodes for wood identification between Santalum album and its adulterants**. *Holzforschung* 2019, **73**(2):209–218.
14. Lloyd Evans D, Joshi SV, Wang J: **Whole chloroplast genome and gene locus phylogenies reveal the taxonomic placement and relationship of Tripidium (Panicoideae: Andropogoneae) to sugarcane**. *BMC Evol Biol* 2019, **19**(1):33.
15. Zhao D-N, Ren Y, Zhang J-Q: **Conservation and innovation: plastome evolution during rapid radiation of Rhodiola on the Qinghai-Tibetan Plateau**. *Mol Phylogenet Evol* 2019:106713.
16. Duan L, Harris AJ, Su C, Zhang Z-R, Arslan E, Ertuğrul K, Loc PK, Hayashi H, Wen J, Chen H-F: **Chloroplast Phylogenomics Reveals the Intercontinental Biogeographic History of the Liquorice Genus (Leguminosae: Glycyrrhiza)**. *Frontiers in Plant Science* 2020, **11**(793).
17. Mohamoud YA, Mathew LS, Torres MF, Younuskunju S, Krueger R, Suhre K, Malek JA: **Novel subpopulations in date palm (Phoenix dactylifera) identified by population-wide organellar genome sequencing**. *BMC Genomics* 2019, **20**(1):498.
18. Qiao J, Zhang X, Chen B, Huang F, Xu K, Huang Q, Huang Y, Hu Q, Wu X: **Comparison of the cytoplastic genomes by resequencing: insights into the genetic diversity and the phylogeny of the agriculturally important genus Brassica**. *BMC Genomics* 2020, **21**(1):480.
19. Chen X, Fang D, Wu C, Liu B, Liu Y, Sahu SK, Song B, Yang S, Yang T, Wei J *et al*: **Comparative Plastome Analysis of Root- and Stem-Feeding Parasites of Santalales Untangle the Footprints of Feeding Mode and Lifestyle Transitions**. *Genome Biol Evol* 2020, **12**(1):3663–3676.
20. Guo X, Liu C, Zhang G, Su W, Landis JB, Zhang X, Wang H, Ji Y: **The Complete Plastomes of Five Hemiparasitic Plants (Osyris wightiana, Pyrularia edulis, Santalum album, Viscum liquidambaricolum, and V. ovalifolium): Comparative and Evolutionary Analyses Within Santalales**. *Frontiers in Genetics* 2020, **11**(597).
21. Dong W, Xu C, Cheng T, Lin K, Zhou S: **Sequencing angiosperm plastid genomes made easy: A complete set of universal primers and a case study on the phylogeny of Saxifragales**. *Genome Biol Evol* 2013, **5**(5):989–997.
22. Li W, Liu Y, Yang Y, Xie X, Lu Y, Yang Z, Jin X, Dong W, Suo Z: **Interspecific chloroplast genome sequence diversity and genomic resources in *Diospyros*. *BMC Plant Biol* 2018, **18**(1):210.**
23. Wickett NJ, Zhang Y, Hansen SK, Roper JM, Kuehl JV, Plock SA, Wolf PG, dePamphilis CW, Boore JL, Goffinet B: **Functional gene losses occur with minimal size reduction in the plastid genome of the parasitic liverwort Aneura mirabilis**. *Mol Biol Evol* 2008, **25**(2):393–401.
24. Krause K: **Piecing together the puzzle of parasitic plant plastome evolution**. *Planta* 2011, **234**(4):647–656.
25. Lin CS, Chen JJ, Chiu CC, Hsiao HC, Yang CJ, Jin XH, Leebens-Mack J, de Pamphilis CW, Huang YT, Yang LH: **Concomitant loss of NDH complex-related genes within chloroplast and nuclear genomes in some orchids**. *The Plant Journal* 2017, **90**(5):994–1006.
26. Wicke S, Schneeweiss GM, Depamphilis CW, Muller KF, Quandt D: **The evolution of the plastid chromosome in land plants: gene content, gene order, gene function**. *Plant Mol Biol* 2011, **76**(3–5):273–297.
27. Roquet C, Coissac E, Cruaud C, Boleda M, Boyer F, Alberti A, Gielly L, Taberlet P, Thuiller W, Van Es J *et al*: **Understanding the evolution of holoparasitic plants: the complete plastid genome of the holoparasite Cytinus hypocistis (Cytinaceae)**. *Ann Bot* 2016.
28. Millen RS, Olmstead RG, Adams KL, Palmer JD, Lao NT, Heggie L, Kavanagh TA, Hibberd JM, Gray JC, Morden CW *et al*: **Many parallel losses of *infA* from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus**. *Plant Cell* 2001, **13**(3):645–658.

29. Li B, Lin F, Huang P, Guo W, Zheng Y: **Development of nuclear SSR and chloroplast genome markers in diverse Liriodendron chinense germplasm based on low-coverage whole genome sequencing**. *Biological Research* 2020, **53**(1):21.
30. Bryan GJ, McNicoll J, Ramsay G, Meyer RC, De Jong WS: **Polymorphic simple sequence repeat markers in chloroplast genomes of Solanaceous plants**. *Theoretical and Applied Genetics* 1999, **99**(5):859–867.
31. Dong W, Xu C, Wen J, Zhou S: **Evolutionary directions of single nucleotide substitutions and structural mutations in the chloroplast genomes of the family Calycanthaceae**. *BMC Evol Biol* 2020, **20**(1):96.
32. Kim G-B, Lim CE, Kim J-S, Kim K, Lee JH, Yu H-J, Mun J-H: **Comparative chloroplast genome analysis of Artemisia (Asteraceae) in East Asia: insights into evolutionary divergence and phylogenomic implications**. *BMC Genomics* 2020, **21**(1):415.
33. Wu Z, Liao R, Yang T, Dong X, Lan D, Qin R, Liu H: **Analysis of six chloroplast genomes provides insight into the evolution of Chrysosplenium (Saxifragaceae)**. *BMC Genomics* 2020, **21**(1):621.
34. Feng S, Zheng K, Jiao K, Cai Y, Chen C, Mao Y, Wang L, Zhan X, Ying Q, Wang H: **Complete chloroplast genomes of four Physalis species (Solanaceae): lights into genome structure, comparative analysis, and phylogenetic relationships**. *BMC Plant Biol* 2020, **20**(1):242.
35. Li YC, Korol AB, Fahima T, Beiles A, Nevo E: **Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review**. *Mol Ecol* 2002, **11**(12):2453–2465.
36. Kelchner SA: **The evolution of non-coding chloroplast DNA and its application in plant systematics**. *Ann Mo Bot Gard* 2000, **87**(4):482–498.
37. Han T, Li M, Li J, Lv H, Ren B, Chen J, Li W: **Comparison of chloroplast genomes of Gynura species: sequence variation, genome rearrangement and divergence studies**. *BMC Genomics* 2019, **20**(1):791.
38. Ren T, Li Z-X, Xie D-F, Gui L-J, Peng C, Wen J, He X-J: **Plastomes of eight Ligusticum species: characterization, genome evolution, and phylogenetic relationships**. *BMC Plant Biol* 2020, **20**(1):519.
39. Song Y, Zhang Y, Xu J, Li W, Li M: **Characterization of the complete chloroplast genome sequence of Dalbergia species and its phylogenetic implications**. *Sci Rep* 2019, **9**(1):20401.
40. Pang X, Liu H, Wu S, Yuan Y, Li H, Dong J, Liu Z, An C, Su Z, Li B: **Species Identification of Oaks (Quercus L., Fagaceae) from Gene to Genome**. *Int J Mol Sci* 2019, **20**(23).
41. Dastpak A, Osaloo SK, Maassoumi AA, Safar KN: **Molecular Phylogeny of Astragalus sect. Ammodendron (Fabaceae) Inferred from Chloroplast ycf1 Gene**. *Annales Botanici Fennici* 2018, **55**(1–3):75–82.
42. Dong W, Xu C, Li C, Sun J, Zuo Y, Shi S, Cheng T, Guo J, Zhou S: **ycf1, the most promising plastid DNA barcode of land plants**. *Sci Rep* 2015, **5**:8348.
43. Lee-Yaw JA, Grassa CJ, Joly S, Andrew RL, Rieseberg LH: **An evaluation of alternative explanations for widespread cytonuclear discordance in annual sunflowers (Helianthus)**. *New Phytol* 2019, **221**(1):515–526.
44. Wang M, Zhang L, Zhang Z, Li M, Wang D, Zhang X, Xi Z, Keefover-Ring K, Smart LB, DiFazio SP *et al*: **Phylogenomics of the genus Populus reveals extensive interspecific gene flow and balancing selection**. *New Phytol* 2020, **225**(3):1370–1382.
45. Li J, Wang S, Jing Y, Wang L, Zhou S: **A modified CTAB protocol for plant DNA extraction**. *Chin Bull Bot* 2013, **48**(1):72–78.
46. Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina sequence data**. *Bioinformatics* 2014, **30**(15):2114–2120.
47. Jin J-J, Yu W-B, Yang J-B, Song Y, dePamphilis CW, Yi T-S, Li D-Z: **GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes**. *Genome Biology* 2020, **21**(1):241.
48. Huang DI, Cronk QCB: **Plann: A command-line application for annotating plastome sequences**. *Applications in Plant Sciences* 2015, **3**(8):1500026.
49. Greiner S, Lehwark P, Bock R: **OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes**. *Nucleic Acids Res* 2019, **47**(W1):W59-W64.
50. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I: **VISTA: computational tools for comparative genomics**. *Nucleic Acids Res* 2004, **32**:W273-W279.
51. Katoh K, Standley DM: **MAFFT multiple sequence alignment software version 7: improvements in performance and usability**. *Mol Biol Evol* 2013, **30**(4):772–780.

52. Kumar S, Stecher G, Li M, Knyaz C, Tamura K: **MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms**. *Mol Biol Evol* 2018, **35**(6):1547–1549.
53. Rozas J, Ferrer-Mata A, Sanchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, Sanchez-Gracia A: **DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets**. *Mol Biol Evol* 2017, **34**(12):3299–3302.
54. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS: **ModelFinder: fast model selection for accurate phylogenetic estimates**. *Nat Methods* 2017, **14**(6):587–589.
55. Lanfear R, Frandsen PB, Wright AM, Senfeld T, Calcott B: **PartitionFinder 2: New Methods for Selecting Partitioned Models of Evolution for Molecular and Morphological Phylogenetic Analyses**. *Mol Biol Evol* 2016.
56. Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A: **RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference**. *Bioinformatics* 2019, **35**(21):4453–4455.
57. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP: **MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space**. *Syst Biol* 2012, **61**(3):539–542.
58. Wikström N, Savolainen V, Chase MW: **Evolution of the angiosperms**: **calibrating the family tree**. *Proceedings of the Royal Society of London Series B: Biological Sciences* 2001, **268**(1482):2211–2220.
59. Mildenhall DC: **Cranwellia costata n.sp. and Podosporites erugatus n.sp. from middle Pliocene (? early Pleistocene) sediments, South Island, New Zealand**. *J R Soc N Z* 1978, **8**(3):253–274.
60. Mildenhall DC: **New Zealand late Cretaceous and cenozoic plant biogeography: A contribution**. *Palaeogeography, Palaeoclimatology, Palaeoecology* 1980, **31**:197–233.
61. Liu B, Le CT, Barrett RL, Nickrent DL, Chen Z, Lu L, Vidal-Russell R: **Historical biogeography of Loranthaceae (Santalales): Diversification agrees with emergence of tropical forests and radiation of songbirds**. *Mol Phylogenet Evol* 2018, **124**:199–212.
62. Vidal-Russell R, Nickrent DL: **The first mistletoes: Origins of aerial parasitism in Santalales**. *Mol Phylogenet Evol* 2008, **47**(2):523–537.
63. Bouckaert R, Heled J, Kuhnert D, Vaughan T, Wu CH, Xie D, Suchard MA, Rambaut A, Drummond AJ: **BEAST 2: a software platform for Bayesian evolutionary analysis**. *PLoS Comp Biol* 2014, **10**(4):e1003537.
64. Rambaut A, Suchard M, Xie D, Drummond A: **Tracer v1**. **6**. In.; 2014: Available from http://beast.bio.ed.ac.uk/Tracer.

Figures

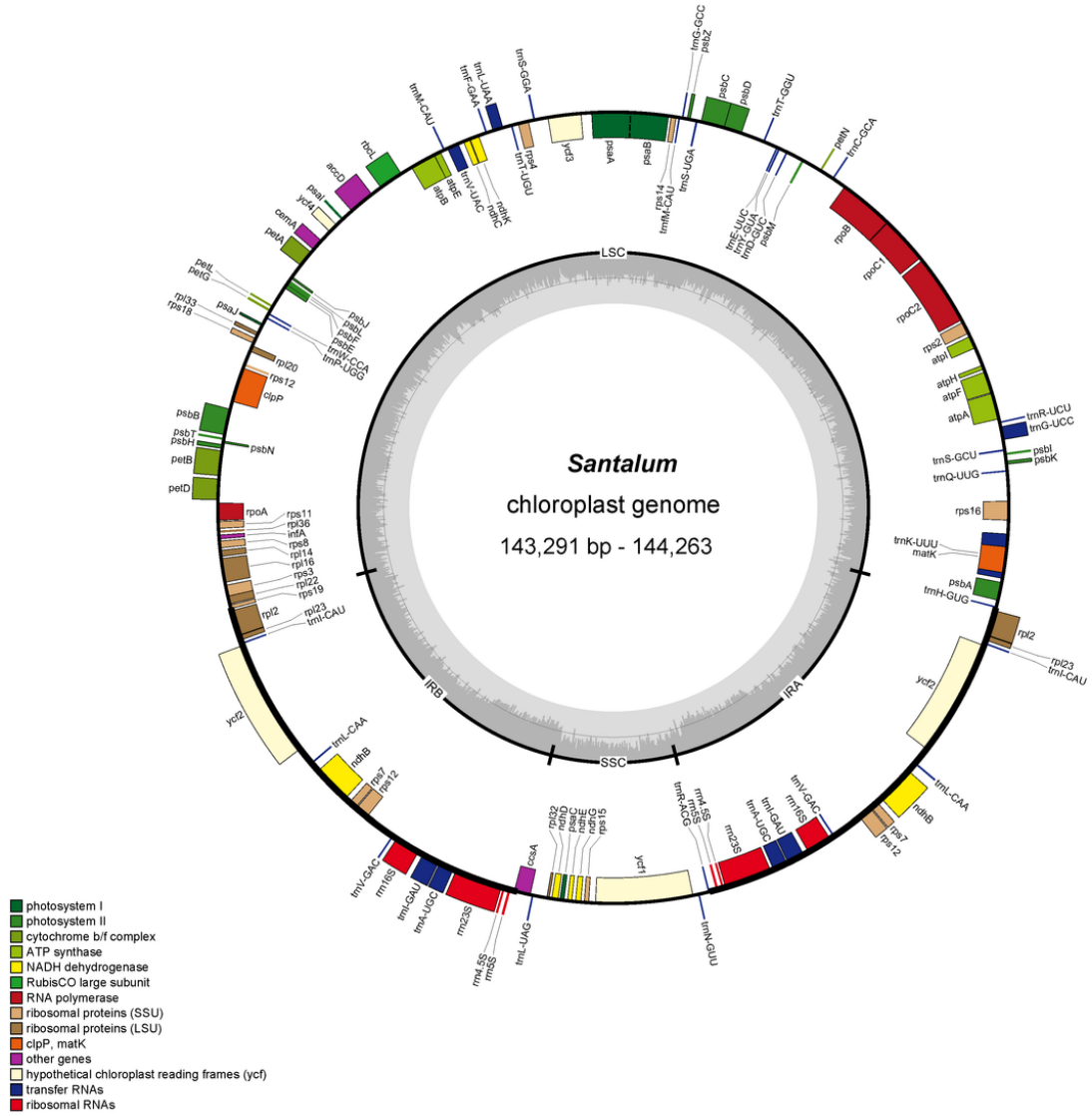

Figure 1

Chloroplast genome map of Santalum. Genes shown inside circle are transcribed counterclockwise, gene outside are transcribed clockwise. Difference functional groups of genes are showed in different color.

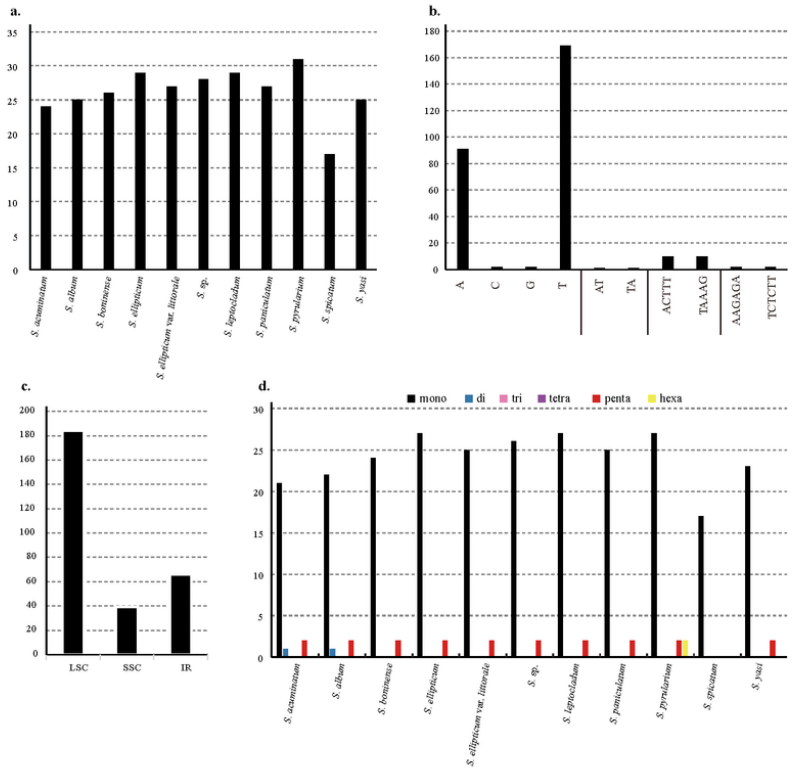

Figure 2

Frequency of SSRs in the Santalum chloroplast genomes. a. The number of SSRs detected in the different Santalum species; b. The number of SSR motifs in different repeat class types; c. The frequency of SSRs in LSC, IR, and SSC regions; d. The number of SSR types in different Santalum species.

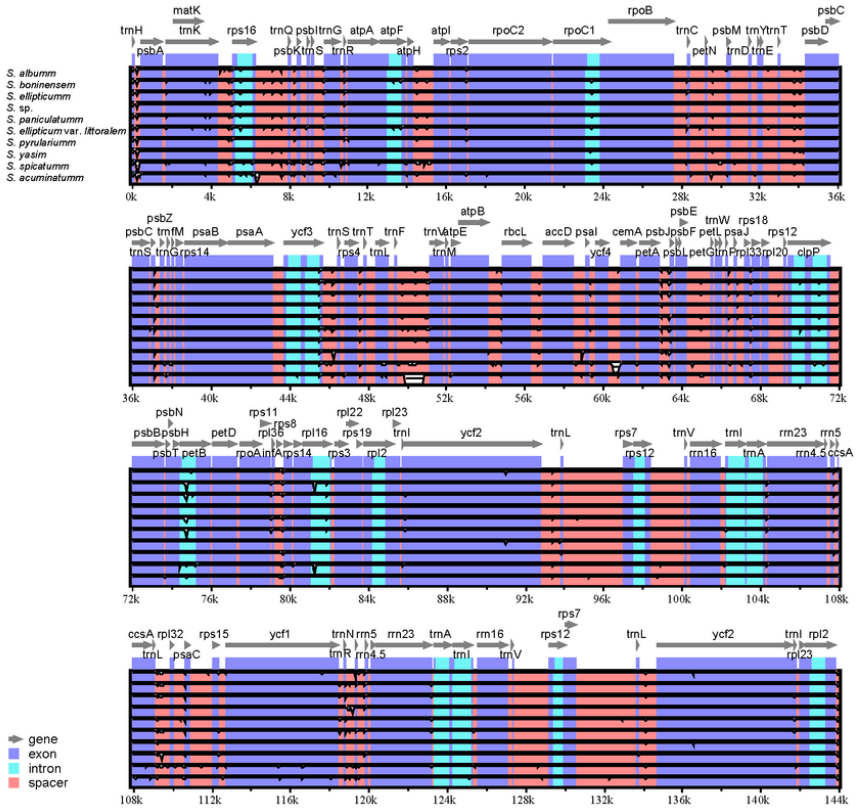

Figure 3

Visualization of genome alignment of the Santalum chloroplast genomes using S. leptocladum as reference by mVISTA. The x-axis showed the coordinate between the chloroplast genome.

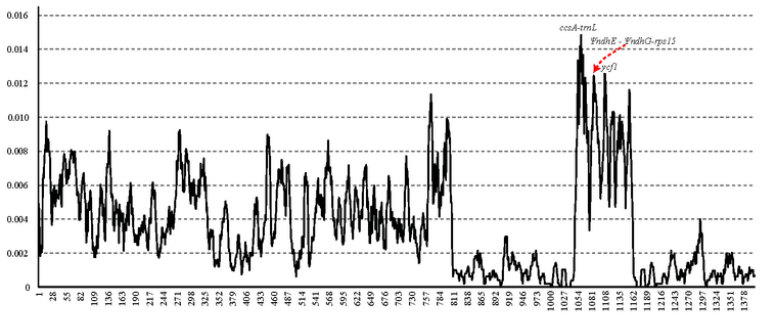

Figure 4

Sliding window test of nucleotide diversity (π) in the Santalum chloroplast genomes. The three mutation hotspot regions (π > 0.012) were annotated. The window size was set to 800 bp and the sliding windows size with 100 bp. X-axis, position of the midpoint of a window; Y-axis, π values of each window.

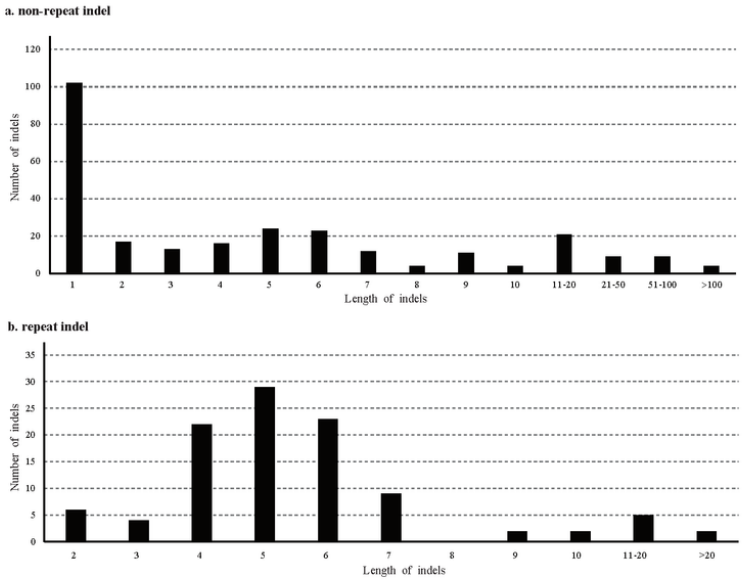

Figure 5

Number and size of indels in the Santalum chloroplast genomes. a. Non-repeat indels; b. Repeat indels.

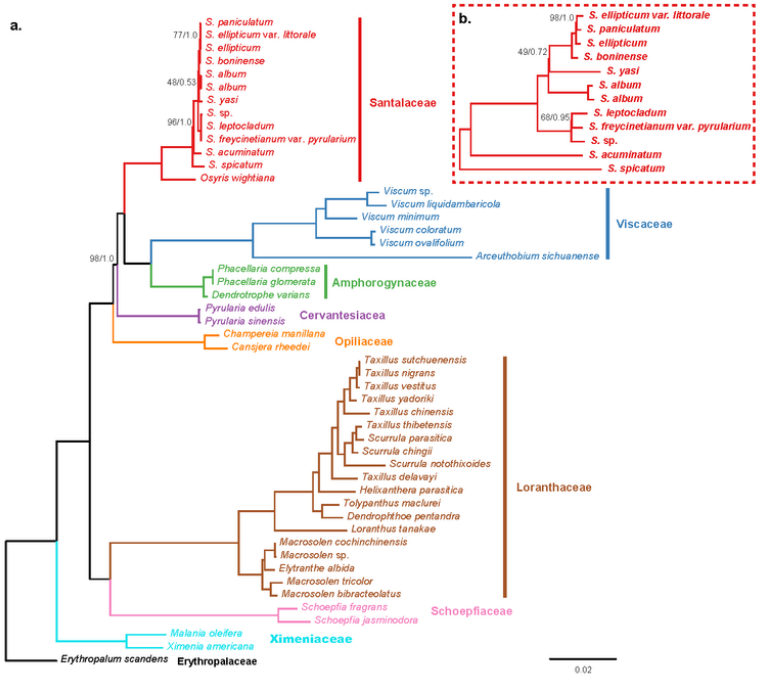

Figure 6

Phylogenetic trees of Santalales. a. ML tree with strict hierarchical clustering partitioning scheme using 70g50s dataset. ML bootstrap support value/Bayesian posterior probability presented at each node. b. ML tree with GTR+G model using 13CPG dataset. ML bootstrap support value/Bayesian posterior probability presented at each node. ML=100/BI=1.0 were not present.

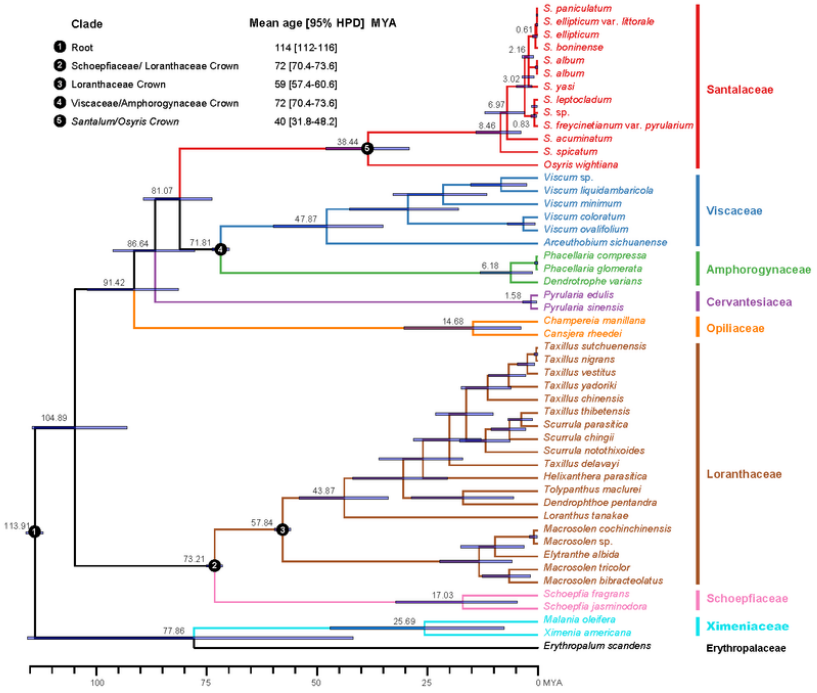

Figure 7

Divergence times of Santalales obtained from BEAST analysis based on 70g50s dataset. Mean divergence time of the nodes were shown next to the nodes while the blue bars correspond to the 95% highest posterior density (HPD). Black circles indicate the five calibration points.

Supplementary Files

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

- FigureS1.pdf
- FigureS2.pdf
- FigureS3.pdf
- TableS1Samplinginformation.xlsx
- TableS2SSRs.xlsx
- TableS3speciesdistance.xlsx
- TableS4markervariable.xlsx

- TableS5primers.xlsx
- TableS6partitionmodes.xlsx