



OPEN

A comparative analysis of complete chloroplast genomes of seven *Ocotea* species (Lauraceae) confirms low sequence divergence within the *Ocotea* complex

Dimitrij Trofimov^{1,2}✉, Daniel Cadar³, Jonas Schmidt-Chanasit³, Pedro Luís Rodrigues de Moraes⁴ & Jens G. Rohwer¹✉

The genus *Ocotea* (Lauraceae) includes about 450 species, of which about 90% are Neotropical, while the rest is from Macaronesia, Africa and Madagascar. In this study we present the first complete chloroplast genome sequences of seven *Ocotea* species, six Neotropical and one from Macaronesia. Genome sizes range from 152,630 (*O. porosa*) to 152,685 bp (*O. aciphylla*). All seven plastomes contain a total of 131 (114 unique) genes, among which 87 (80 unique) encode proteins. The order of genes (if present) is the same in all Lauraceae examined so far. Two hypervariable loci were found in the LSC region (*psbA-trnH*, *ycf2*), three in the SSC region (*ycf1*, *ndhH*, *trnL(UAG)-ndhF*). The pairwise cp genomic alignment between the taxa showed that the LSC and SSC regions are more variable compared to the IR regions. The protein coding regions comprise 25,503–25,520 codons in the *Ocotea* plastomes examined. The most frequent amino acids encoded in the plastomes were leucine, isoleucine, and serine. SSRs were found to be more frequent in the two dioecious Neotropical *Ocotea* species than in the four bisexual species and the gynodioecious species examined (87 vs. 75–84 SSRs). A preliminary phylogenetic analysis based on 69 complete plastomes of Lauraceae species shows the seven *Ocotea* species as sister group to *Cinnamomum* sensu lato. Sequence divergence among the *Ocotea* species appears to be much lower than among species of the most closely related, likewise species-rich genera *Cinnamomum*, *Lindera* and *Litsea*.

The Lauraceae are among the most frequent woody plant families in moist tropical areas and include about 55 genera with 2500–3500 species^{1–3}. The genus *Ocotea* Aubl., in its current circumscription, is the largest genus among the Neotropical Lauraceae, consisting of about 400–450 recognized species^{2–8}. The number of Paleotropical *Ocotea* species is far smaller. The majority (34 spp.) is endemic to Madagascar, four are found in Continental Africa, three on Mauritius, one on Réunion Island, and one on the Comoro islands. *Ocotea foetens* (Aiton) Baill. is endemic to Macaronesia^{3,7}.

Most of the molecular phylogenetic studies in Lauraceae published so far focused on other genera or on the major evolutionary lineages in the Lauraceae and included only a relatively small number of *Ocotea* species^{9–11}. Nevertheless, they suggested that *Ocotea* was paraphyletic with respect to most other New World genera, viz. *Aniba* Aubl., *Damburneya* Raf., *Dicypellium* Nees & Mart., *Endlicheria* Nees, *Kubitzkia* van der Werff, *Licaria* Aubl., *Nectandra* Rol. ex Rottb., *Paraia* Rohwer, H.G. Richt. & van der Werff, *Pleurothyrium* Nees, *Rhodostemonodaphne* Rohwer & Kubitzki, *Umbellularia* (Nees) Nuttall and *Urbanodendron* Mez. A recent study based on RAD-seq data¹² added *Phyllostemonodaphne* Kosterm. to this list. These genera, plus presumably *Gamanthera* van der Werff and *Povedadaphne* W.C. Burger, which have not been studied yet, are collectively referred to as

¹Institute of Plant Science and Microbiology, Universität Hamburg, Ohnhorststraße 18, 22609 Hamburg, Germany. ²Institute of Ecology and Evolution, Universität Jena, Philosophenweg 16, 07743 Jena, Germany. ³Bernhard Nocht Institute for Tropical Medicine, Bernhard-Nocht-Straße 74, 20359 Hamburg, Germany. ⁴Departamento de Botânica, Instituto de Biociências, Universidade Estadual Paulista “Júlio de Mesquita Filho”, Av. 24 A 1515, Bela Vista, Rio Claro, Caixa Postal 199, São Paulo CEP 13506-900, Brazil. ✉email: dimitrij.trofimov@uni-jena.de; jens.rohwer@uni-hamburg.de

the *Ocotea* complex¹⁰ or Supraocotea¹², a group of about 950 species. A higher number of *Ocotea* species than in previous studies, plus representative species of other genera of the *Ocotea* complex, were studied by Trofimov et al.² and Trofimov and Rohwer³, with similar results. Using sequences of the nuclear internal transcribed spacer (ITS) and one of the most informative parts of the chloroplast genome, the *psbA-trnH* spacer, they separated two genera from *Ocotea* s. lat., namely *Mespilodaphne* Nees & Mart. and *Kuloa* Trofimov & Rohwer. In addition, several of the morphological groups described by Rohwer⁴ were confirmed as monophyletic in these studies. Resolution and/or support values at the lower nodes within the *Ocotea* complex, however, remained poor. Most of the other established chloroplast markers tested in the research group of the senior author [JGR] (*atpB-rbcL*, *matK*, *ndhF-rpl32*, *psbK-psbI*, *rbcL*, *rpl16*, *rpb2*, *rpl3-trnL*, *rpl32-trnL*, *rpoB*, *rpoC1*, *trnG-trnS*, *trnL-trnF*, and *trnT-trnL*) turned out to be less informative in molecular analyses of the *Ocotea* complex, or problematic because of too many single nucleotide repeats. Therefore, no significant improvement was to be expected from sequencing individual chloroplast markers any more. Sequencing of entire chloroplast (cp) genomes, on the other hand, is expected to yield a higher number of informative characters, which will probably lead to better support for the lower nodes within the *Ocotea* complex. The present study of selected *Ocotea* plastomes is intended as a first step towards this goal. The most recent phylogeny by Penagos Zuluaga et al.¹² based on RAD-seq data is fully resolved at the lower nodes, with strong bootstrap support for all of the basal and most of the more distal nodes, so that it will provide an ideal basis for comparison with our and future cp genome data.

The chloroplast (cp) genome is a circular molecule ranging in size from 107 to 218 kb. It shows a characteristic quadripartite structure with a pair of inverted repeats (IRs) separating a large single copy (LSC) and a small single copy (SSC) region^{13,14}. The typical angiosperm cp genome consists of 120–130 genes, coding mainly for RNAs and photosynthesis-related genes¹⁵.

Up to the present, the plastomes of Lauraceae were studied mainly in Asian species of *Actinodaphne* Nees, *Alseodaphne* Nees, *Beilschmiedia* Nees, *Cryptocarya* R. Br., *Caryodaphnopsis* Airy Shaw, *Cassytha* L., *Cinnamomum* Schaeff., *Dehaasia* Blume, *Endiandra* R. Br., *Eusideroxylon* Teijsm. & Binn., *Iteadaphne* Blume, *Laurus* L., *Lindera* Thunb., *Litsea* Lam., *Machilus* Nees, *Neocinnamomum* H. Liu, *Neolitsea* (Benth. & Hook. f.) Merr., *Nothaphoebe* Blume, *Parasassafras* D.G. Long, *Phoebe* Nees, *Sassafras* J. Presl and *Syndiclis* Hook. f.^{16–33}. Neotropical species were poorly represented in previous studies of the cp genome. Only *Nectandra angustifolia* (Schrad.) Nees & Mart. (but see below) and *Persea americana* Mill. have been studied so far, plus the North American *P. borbonica* (L.) Spreng.^{24,25,28}. These studies considerably improved support values among the major phylogenetic lineages in the Asian Lauraceae, especially among *Cassytha*, *Caryodaphnopsis* and *Neocinnamomum*. In other plant groups, such as the genus *Quercus* L., Poaceae-Arundinarieae and Rosaceae, they allowed resolving phylogenetic relationships on different levels^{34–36}.

In this study, we sequenced and analyzed the complete chloroplast genomes of six Neotropical and the only Macaronesian *Ocotea* species using Illumina high-throughput sequencing technology. This is the first study of this kind in this species-rich and ecologically important group. We describe the structure of the plastomes examined, amino acid percentage of protein-coding genes, content of Simple Sequence Repeats (SSRs), relative synonymous codon usage for protein coding nucleotides and variability values in the *Ocotea* plastomes, and compare them to 85 plastomes of Lauraceae. In addition, we performed a preliminary phylogenetic analysis to show the positions of the seven *Ocotea* species among 62 plastomes of Core Lauraceae in the sense of Rohwer and Rudolph³⁷ examined so far, i.e. Cinnamomeae, Laureae and Perseeae. The *Ocotea* complex forms the largest clade within the Cinnamomeae, with the Laureae and the Perseeae as consecutive sister groups. However, this paper does not have a phylogenetic focus but rather provides basic data on chloroplast genomes that may be used in future phylogenetic studies.

Results

Organization of the plastomes of *Ocotea*. The chloroplast genome sequences of the seven *Ocotea* species range from 152,630 bp in *O. porosa* (Nees & Mart.) Barroso to 152,685 bp in *O. aciphylla* (Nees & Mart.) Mez (Table 1). The plastomes show the typical quadripartite structure of chloroplast genomes. Two inverted repeat (IR) regions (20,009–20,015 bp) are separated by a large single copy (LSC) region (93,815–93,859 bp) and a small single copy (SSC) region (18,775–18,818 bp) (Fig. 1, Table 1). All seven *Ocotea* plastomes contain a total of 131 genes (114 unique), among which 87 (80 unique) encode proteins (Table 2). The order of genes (if present) is the same in all Lauraceae so far examined. Fourteen genes have one intron (*atpF*, *ndhA*, *ndhB*, *rpl2*, *rpl16*, *rpoC1*, *rps12*, *rps16*, *trnA-UGC*, *trnG-UCC*, *trnI-GAU*, *trnK-UUU*, *trnL-UAA*, and *trnV-UAC*), and two (*clpP* and *pafI*) have two introns (Table 2, Supplementary Table S1). The total GC content in plastomes is same in all *Ocotea* species examined (39.2%; Table 1). Contents of nucleotides in the LSC, IR and SSC of the plastomes were similar in all species of *Ocotea* examined (Supplementary Table S2). About 30.3–30.4%, 27.3–28.3%, and 32.9% were detected for A; 19.3–19.4%, 21.1–23.4%, and 21.1% for C; 18.6%, 21.0–23.4%, and 18.1% for G; and 31.6–31.7%, 27.2–28.3%, and 33.1% for T, respectively. The GC content in the IR regions was higher than in the LSC and SSC regions (44.4%, vs. 37.9–38.0% and 33.9–34.0%, respectively).

Determination of the most variable regions. The nucleotide diversity (Π) values within 600 bp across the seven *Ocotea* plastomes vary from 0 to 0.015, with a mean value of 0.001 (Fig. 2a). Four variable loci with $\Pi \geq 0.006$ were found in the LSC region (*psbA-trnH*, $\Pi = 0.007$; *ycf2*, $\Pi = 0.006$) and in the SSC region (*ycf1*, $\Pi = 0.008$; *ndhH*, $\Pi = 0.008$; *trnL(UAG)-ndhF*, $\Pi = 0.015$). At the family level, sequence divergence was calculated using published chloroplast genomes of *Alseodaphne*, *Cinnamomum*, *Laurus*, *Lindera*, *Litsea*, *Machilus*, *Neolitsea*, *Parasassafras*, *Persea*, *Phoebe*, and *Sassafras* (see “Materials and methods” section). Unfortunately, the sequence of *Nectandra angustifolia* (marked as “unverified” in GenBank) had to be excluded because it differs so strongly from those of all other Core Lauraceae that large parts of it could not be readily aligned. The Π values

	<i>O. aciphylla</i>	<i>O. daphnifolia</i>	<i>O. foetens</i>	<i>O. guianensis</i>	<i>O. odorifera</i>	<i>O. porosa</i>	<i>O. tabacifolia</i>
Raw reads no	2,290,010	2,161,228	5,898,677	8,358,607	4,373,153	1,985,942	2,103,610
Clean reads no.	2,287,092	2,143,358	5,896,116	8,355,240	4,370,826	1,984,720	2,102,190
Gene no. (unique)	131 (114)	131 (114)	131 (114)	131 (114)	131 (114)	131 (114)	131 (114)
Protein coding genes no. (unique)	87 (80)	87 (80)	87 (80)	87 (80)	87 (80)	87 (80)	87 (80)
tRNA genes no. (unique)	36 (30)	36 (30)	36 (30)	36 (30)	36 (30)	36 (30)	36 (30)
rRNA genes no. (unique)	8 (4)	8 (4)	8 (4)	8 (4)	8 (4)	8 (4)	8 (4)
cp genome length, bp	152,685	152,635	152,656	152,656	152,646	152,630	152,652
Large single copy (LSC) length, bp	93,849	93,834	93,859	93,858	93,841	93,815	93,851
Inverted repeat (IR) length, bp	20,009	20,009	20,011	20,009	20,009	20,015	20,009
Small single copy (SSC) length, bp	18,818	18,783	18,775	18,780	18,787	18,785	18,783
Protein coding (CDS) length, bp	76,577	76,591	76,591	76,567	76,591	76,559	76,576
Total GC content (%)	39.2	39.2	39.2	39.2	39.2	39.2	39.2

Table 1. Summary of seven complete plastomes of *Ocotea* species.

among the 69 plastomes vary from 0 to 0.022, with a mean value of 0.0045 (Fig. 2b). Variable loci with $P_i > 0.01$ were identified in the LSC region (*rps16-trnQ*, $P_i = 0.01$; *rpoB-psbD*, $P_i = 0.01$; *trnT-trnL*, $P_i = 0.01$; *rpl23-ycf2*, $P_i = 0.014$) and in the SSC region (*ycf1*, $P_i = 0.019$; *trnL(UAG)-ycf1*, $P_i = 0.022$). The open reading frames *ycf1* and *ycf2* are located in one of the IR regions (IRb), at the border of the SSC and the LSC region, respectively.

Comparative analysis of plastomes. A comparison of the LSC, IR and SSC junction positions in the *Ocotea* plastomes is shown in Fig. 3. The *ycf1* gene crosses the boundary between the IRb (1408 bp) and the SSC (4163 bp) regions. The *ycf2* gene is found in the boundary between the LSC (3852 bp) and the IRb (3162 bp) regions. Fragments (pseudogenes) of *ycf1* (1408 bp) and *ycf2* (3162 bp) are located in the IRa region. The distances between the *ndhF* gene and the *ycf1* fragment and between the *ycf2*-fragment and the *trnH* gene are 21 bp and 27 bp, respectively. The pairwise cp genomic alignment between six *Ocotea* species and *O. aciphylla* as reference showed very high similarity in all sequences (Fig. 4). The LSC and SSC regions were more variable in comparison with the IR regions. The noncoding regions showed a relatively higher mutation rate than protein-coding regions in the *Ocotea* plastomes examined.

Codon usage analysis. The count of codons in the plastoms examined here were 25,503–25,520 with an average number of about 25,514 (Supplementary Tables S3, S4). The effective number of codons (ENC), Codon Bias Index (CBI) as well the Scaled Chi-square (SCh2) were very similar in all *Ocotea* plastomes (56.59–56.62; 0.15; 0.073–0.074, respectively) (Supplementary Table S4). The GC content at coding positions is about 39.1% in the examined *Ocotea* plastomes. The GC contents at second and at third codon positions were also very similar (35.5–35.6%; 39.2% respectively). All possible codon types are used for each amino acid. The most frequent amino acids encoded in the *Ocotea* plastomes are leucine (Leu; 11.76–11.83%), isoleucine (Ile; 8.05%–8.11%), and serine (Ser; 7.93–8.01%) (Fig. 5). The amino acids arginine (Arg), glycine (Gly), lysine (Lys), phenylalanine (Phe), and valine (Val) account for 5.02–5.95% each. Least represented in the chloroplast genomes examined were cysteine (Cys; 1.81–1.87%) and tryptophan (Trp; 1.94–1.95%). The relative synonymous codon usage (RSCU) was greater than 1.0 in 31 codons (Supplementary Table S3). The count of preferred codons ending with A/U or G/C were 25 and six, respectively. The frequency of different codons coding for the same amino acid was almost the same in all *Ocotea* species examined. The Macaronesian *Ocotea foetens* presented slightly higher frequencies for arginine, cysteine, serine, histidine (His), and tyrosine (Tyr) in comparison with the Neotropical *Ocotea* species, whereas the contents of alanine (Ala), isoleucine and leucine were slightly lower.

Simple sequence repeats (SSRs) analysis. The seven chloroplast genomes examined showed a total 586 SSRs with a repeat length of one to six bp (Fig. 6a, Supplementary Table S5). These SSRs were mainly mononucleotide repeats (433 SSRs = 74%) of A or T (417), less frequently C or G (16). In addition, there were 65 dinucleotide repeats (11%), 21 tri- (4%), 55 tetra- (9%), five penta- (1%), and seven hexanucleotide repeats (1%). The numbers of SSRs observed in the different *Ocotea* species were relatively similar. In each plastome we identified 77–89 SSRs, incl. 57–67 mono-, nine or ten di-, three tri-, seven or eight tetra-, zero, one or two penta-, and zero, one or three hexanucleotide repeats (Fig. 6b). The SSRs were identified mainly in the LSC region (62–70 SSRs; Supplementary Table S5), compared to one or two and 11–17 SSRs in the IR and SSC regions, respectively.

Phylogenetic analysis of Lauraceae plastomes. The data matrix consisted of 160,629 characters, among which 5266 were variable but parsimony-uninformative, and 4631 were parsimony-informative. However, only 168 characters were parsimony-informative among the seven *Ocotea* species in this analysis. Most clades in the Maximum Likelihood analysis received 100% bootstrap support (ML-BS, Fig. 7, Supplementary

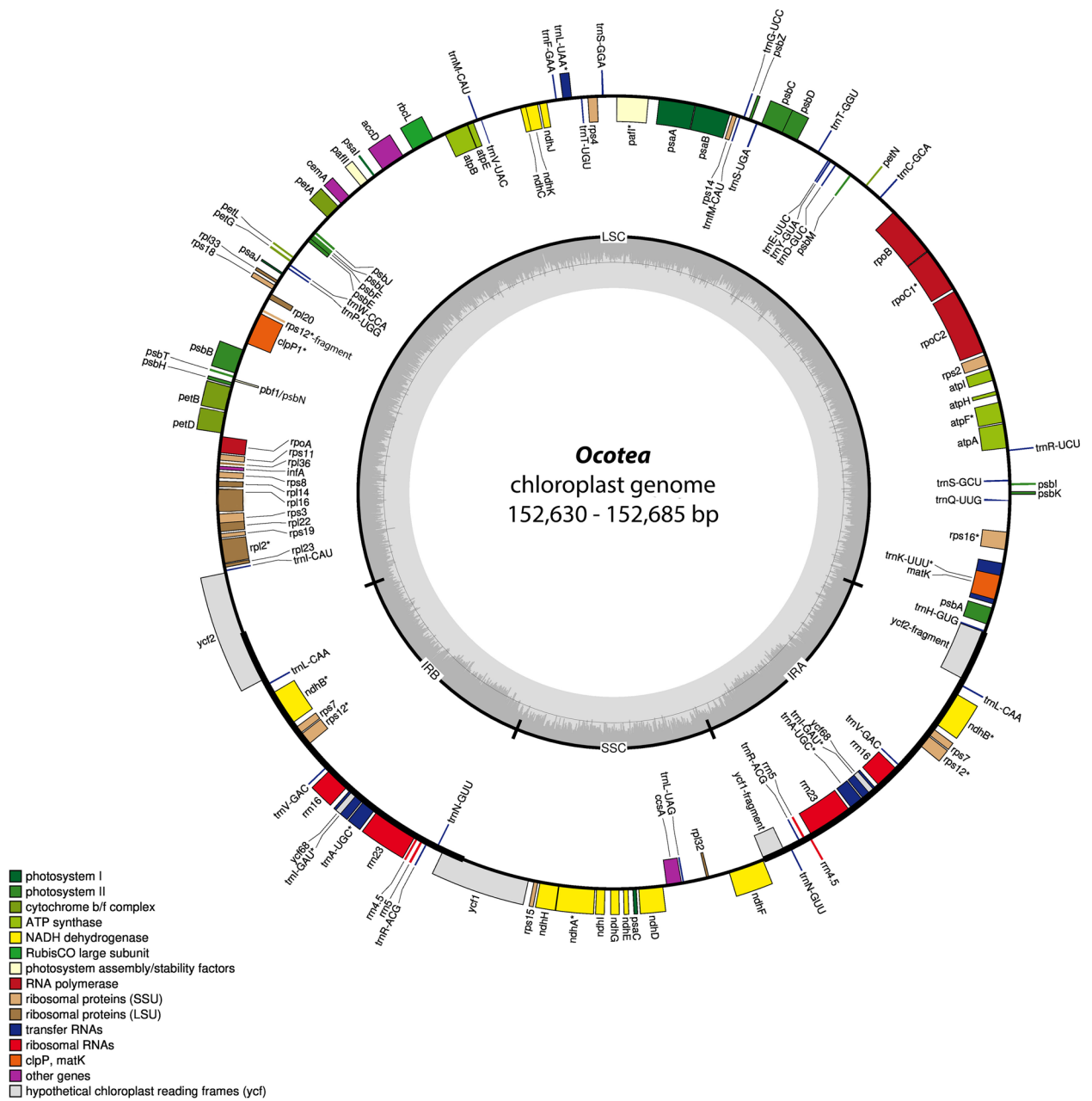


Figure 1. Gene map of *Ocotea* species (*O. aciphylla*, *O. daphnifolia*, *O. foetens*, *O. guianensis*, *O. odorifera*, *O. porosa*, and *O. tabacifolia*) chloroplast genomes. The genes shown on the inside and the outside of the outer circle are transcribed in clockwise and counterclockwise direction, respectively. The coloured bars denote gene functional groups. The dark gray and light gray shading within the inner circle correspond to percentage GC and AT content, respectively. *IR* inverted repeat, *LSC* large single copy, *SSC* small single copy.

Fig. S1). With the Perseeae defined as the outgroup, Laureae and Cinnamomeae are shown as sister clades in the ingroup. Among the Cinnamomeae, species of *Cinnamomum*, with *Sassafras* nested among them, form the sister group to the seven *Ocotea* species examined here. The Macaronesian *Ocotea foetens* is shown as sister taxon to the six Neotropical species. Among these, the two dioecious species, *Ocotea guianensis* Aubl. and *O. tabacifolia* (Meisn.) Rohwer, form the sister group to the remaining species, which are bisexual or gynodioecious [*O. daphnifolia* (Meisn.) Mez]. The latter clade, however, is barely supported (57% ML-BS). *Ocotea aciphylla* (Nees & Mart.) Mez appears as sister taxon to the remaining species, and among these *O. porosa* (Nees & Mart.) Barroso is shown as sister taxon to *O. daphnifolia* and *O. odorifera* (Vell.) Rohwer.

Category	Group of genes	Name of genes
Photosynthesis related genes	Rubisco	<i>rbcl</i>
	Photosystem I	<i>psaA, psaB, psaC, psaI, psaJ</i>
	Assembly/stability of photosystem I	<i>pafl**</i> , <i>paflI</i>
	Photosystem II	<i>psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, inhA/psbZ</i>
	ATP synthase	<i>atpA, atpB, atpE, atpF*</i> , <i>atpH, atpI</i>
	Cytochrome b/f complex	<i>petA, petB, petD, petG, petL, petN</i>
	Cytochrome c synthesis	<i>ccsA</i>
	NADH-dehydrogenase	<i>ndhA*</i> , <i>ndhB*</i> (2x), <i>ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK</i>
Transcription and translation related genes	Subunit of RNA polymerase	<i>rpoA, rpoB, rpoC1*</i> , <i>rpoC2</i>
	Ribosomal protein small subunit	<i>rps2, rps3, rps4, rps7</i> (2x), <i>rps8, rps11, rps12*</i> (2x), <i>rps12-fragment, rps14, rps15, rps16*</i> , <i>rps18, rps19</i>
	Ribosomal protein large subunit	<i>rpl2*</i> , <i>rpl14, rpl16*</i> , <i>rpl20, rpl22, rpl23, rpl32, rpl33, rpl36</i>
	Translation initiation factor	<i>infA</i>
RNA genes	Ribosomal RNAs	<i>rrn4.5</i> (2x), <i>rrn5</i> (2x), <i>rrn16</i> (2x), <i>rrn23</i> (2x)
	Transfer RNAs	<i>trnA-UGC*</i> (2x), <i>trnC-GCA, trnD-GUC, trnE-UUC, trnF-GAA, trnG-GCC, trnG-UCC*</i> , <i>trnH-GUG, trnI-CAU, trnI-GAU*</i> (2x), <i>trnK-UUU*</i> , <i>trnL-CAA</i> (2x), <i>trnL-UAA*</i> , <i>trnL-UAG, trnM-CAU, trnM-CAU, trnN-GUU</i> (2x), <i>trnP-UGG, trnQ-UUG, trnR-ACG</i> (2x), <i>trnR-UCU, trnS-GCU, trnS-GGA, trnS-UGA, trnT-GGU, trnT-UGU, trnV-GAC</i> (2x), <i>trnV-UAC*</i> , <i>trnW-CCA, trnY-GUA</i>
Miscellaneous group	Maturase	<i>matK</i>
	Envelope membrane protein	<i>cemA</i>
	Subunit of Acetyl-CoA-Carboxylase	<i>accD</i>
	Proteolysis	<i>clpP**</i>
Genes of unknown function	Conserved open reading frames	<i>ycf1, ycf1-fragment, ycf2, ycf2-fragment</i>
Putative pseudogenes		<i>ycf68</i> (2x)

Table 2. Genes encoded by seven *Ocotea* plastomes. (2x)—Gene represented by two copies. *Gene containing one intron. **Gene containing two introns.

Discussion

The genome sizes of the *Ocotea* species examined in this study are similar to those of other Core Lauraceae^{16–33}, as well as *Caryodaphnopsis* and *Neocinnamomum* species²⁵ (Supplementary Table S6). The genomes of Cryptocaryae (*Beilschmiedia*, *Cryptocarya*, *Endiandra*, *Eusideroxylon* and *Syndiclis*) are more than 5000 bp larger^{25,28}. The cp genome of the hemiparasitic *Cassytha*, on the contrary, is ca. 40,000 bp smaller than those of the Core Lauraceae²⁵. *Cassytha* has lost not only its functional *ndh* genes, like many hemiparasitic plants³⁸, but also an entire inverted repeat region.

The seven *Ocotea* chloroplast genomes show some length variation in all of their parts (LSC, IRs, SSC). Consistently smaller length variation was found among the species of *Alseodaphne* (3 spp.), *Endiandra* (4 spp.), *Neocinnamomum* (2 spp.), *Phoebe* (3 spp.) and *Syndiclis* (2 spp.) so far examined^{25–28}. Larger variation was found among the species of *Beilschmiedia* (6 spp.), *Caryodaphnopsis* (3 spp.), *Cinnamomum* (7 spp.), *Litsea* (14 spp.) and *Persea* (3 spp.)^{16,17,24,25,28–30}. However, the intrageneric differences are expected to increase, also in *Ocotea*, with increased number of species examined. It is therefore too early to make statements about the relative length variability in different clades. Surprisingly large intrageneric differences of about 6000 bp in total chloroplast, LSC and IR regions, and about 400 bp in the SSC region were observed among *Caryodaphnopsis* species, *C. henryi*, *C. malipoensis* and *C. tonkinensis*^{25,28}.

A total of 131 (114 unique) genes were identified in the *Ocotea* species examined. For most Lauraceae examined so far (species of *Actinodaphne*, *Alseodaphne*, *Beilschmiedia*, *Caryodaphnopsis*, *Cinnamomum*, *Cryptocarya*, *Eusideroxylon*, *Lindera*, *Machilus*, *Nectandra*, *Neocinnamomum*, *Neolitsea*, *Persea*, *Phoebe* and *Sassafras*), the number of genes was indicated as 128–130 (113 unique)^{23–27,33}. Lower numbers (127 total/112 unique) were reported for some species of *Actinodaphne*, *Cinnamomum*, *Lindera*, *Litsea*, and *Neolitsea*^{17,29,30}, but only 107 genes (total and unique) in two *Cassytha* species²⁵.

A total of 87 protein coding genes were identified in the *Ocotea* species examined. The counts of total protein coding genes in Lauraceae in previous studies ranged from 73 genes in *Cassytha* species via 79 in *Cinnamomum camphora* to 86 genes in *Caryodaphnopsis henryi* Airy Shaw^{16,25,30}. Consistently 85 protein coding genes have been reported for the genera of the early divergent Cryptocaryae (*Beilschmiedia*, *Cryptocarya* and *Eusideroxylon*), as far as they have been examined. Among the remaining Lauraceae, the most frequent count is 84³⁰. Lower numbers have been reported for *Cinnamomum micranthum* and *C. kanehirae* (83)²⁹, *Litsea glutinosa* (83)¹⁷, *Cinnamomum camphora* (79)¹⁶ and two *Cassytha* species (73)²⁵. The differences among the counts of genes in the Lauraceae species, except the hemiparasitic *Cassytha*, may be due to different annotation of genes. Particularly the *rpl22* gene has not been annotated in most of the earlier studies^{17,19,23,24,26,27,29,30,33,39–41}.

The *psbA-trnH*, *ycf1*, *ycf2*, *ndhH* and *trnL(UAG)-ndhF* regions were identified as hypervariable loci ($P_i \geq 0.006$) at the species level among the *Ocotea* species examined here. Seven hypervariable regions ($P_i > 0.014$), *ihbA-trnG*,

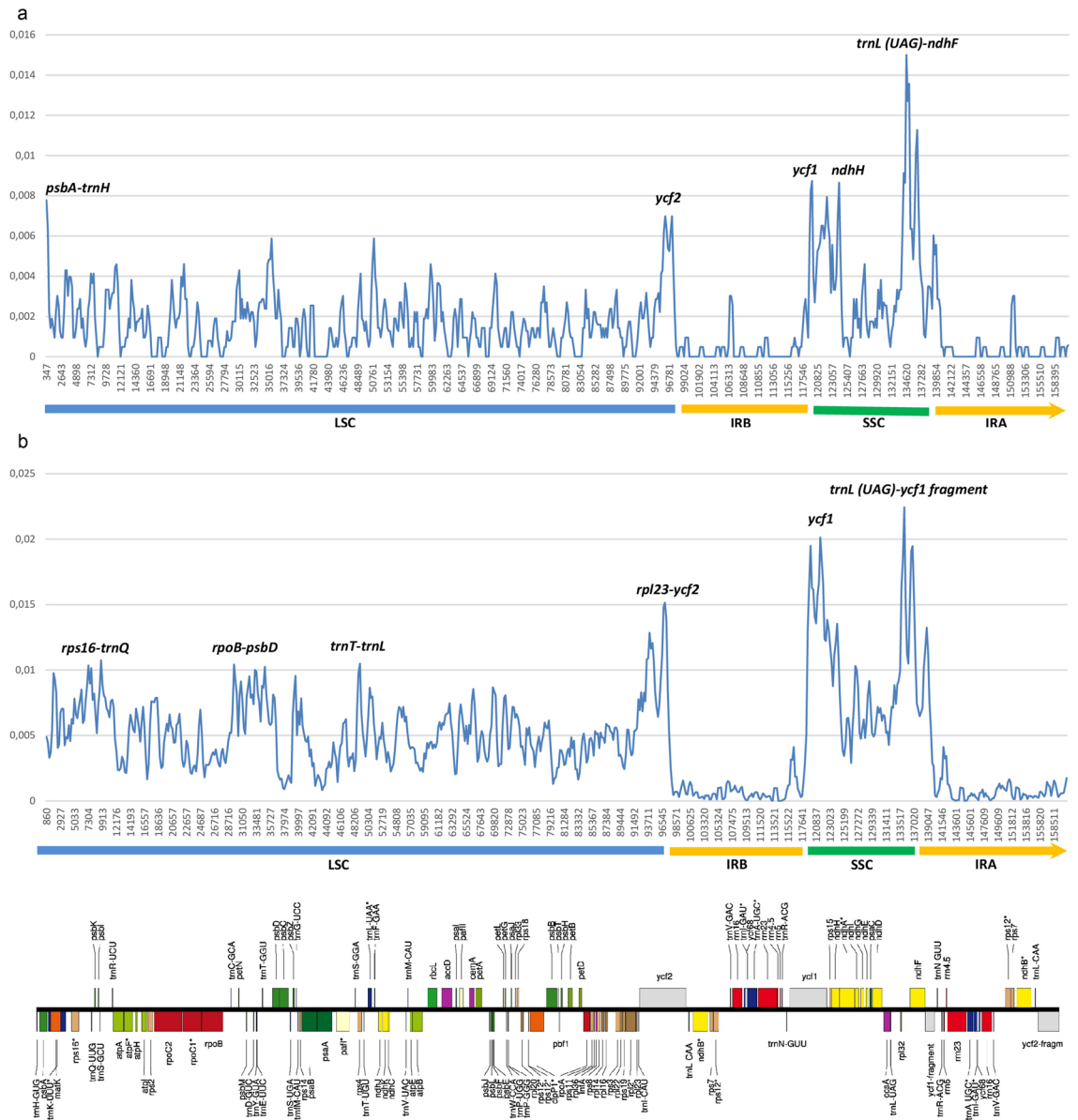


Figure 2. Comparison of the nucleotide variability (P_i) values, (a) among the seven *Ocotlea* plastomes and (b) among 69 plastomes of Lauraceae. The linear gene map of *Ocotlea* species is shown below.

ndhA, *ndhF-rpl32*, *psbK-psbI*, *rps16*, *trnS-trnG* and *ycf1* were identified in *Lindera* species³³. The *psbA-trnH*, *ycf2* and *ndhH* regions are not among the most variable regions in these species. *Alseodaphne* species show six hypervariable loci ($P_i > 0.006$), *accD-psaI*, *ndhF-rpl32*, *rps19-rpl3*, *rpl32-trnL*, *trnG-UCC*, and *ycf1*²⁶. Seven hypervariable loci ($P_i > 0.008$), *clpP*, *ndhF-rpl32*, *rpl32-trpL*, *rps8-rpl14*, *trnQ-psbI*, *ycf1*, and *ycf2*, were identified in *Machilus* species²³. At the family level, we identified additional hypervariable regions ($P_i \geq 0.01$) among 69 Core Lauraceae species, viz. *rpoB-psbD* and *trnT-trnL*. Zhao et al.³³ detected only *ndhF-rpl32* and *ycf1* as hypervariable loci ($P_i > 0.014$) among the Core Lauraceae. By comparing the *Ocotlea* plastomes using the mVISTA program^{42,43}, we confirmed that the IR regions are more conservative than the LSC and SSC regions. The LSC and SSC regions comprise more noncoding regions with higher mutation rates. The protein-coding sequences, including 80 genes, were longer in the *Ocotlea* plastomes than in the plastome of *Cinnamomum camphora* (76,509–76,560 bp = 25,503–25,520 codons vs. 63,654 bp = 21,218 codons)¹⁶.

In our study and in Chen et al.¹⁶ the codons coding for leucine and for cysteine were the most and the least frequent, respectively. 11.76–11.83% of the codons in *Ocotlea* and 10.87% in *Cinnamomum camphora* are coding for leucine, whereas only 1.81–1.87% or 1.25%, respectively, are coding for cysteine. Like in *C. camphora*, preferred codons in *Ocotlea* are more frequently ending in A/U than in G/C (27 vs. two in *C. camphora*, 25 vs. six in *Ocotlea*).

Simple sequence repeats (SSRs) are widely distributed in chloroplast genomes of Lauraceae. Chen et al.¹⁶ detected 81, 82, 83–88, and 86 SSRs in *Litsea*, *Machilus*, *Cinnamomum* and *Persea* species, respectively. In this study, we found more SSRs in the two Neotropical dioecious *Ocotlea* species examined, *O. guianensis* and *O.*

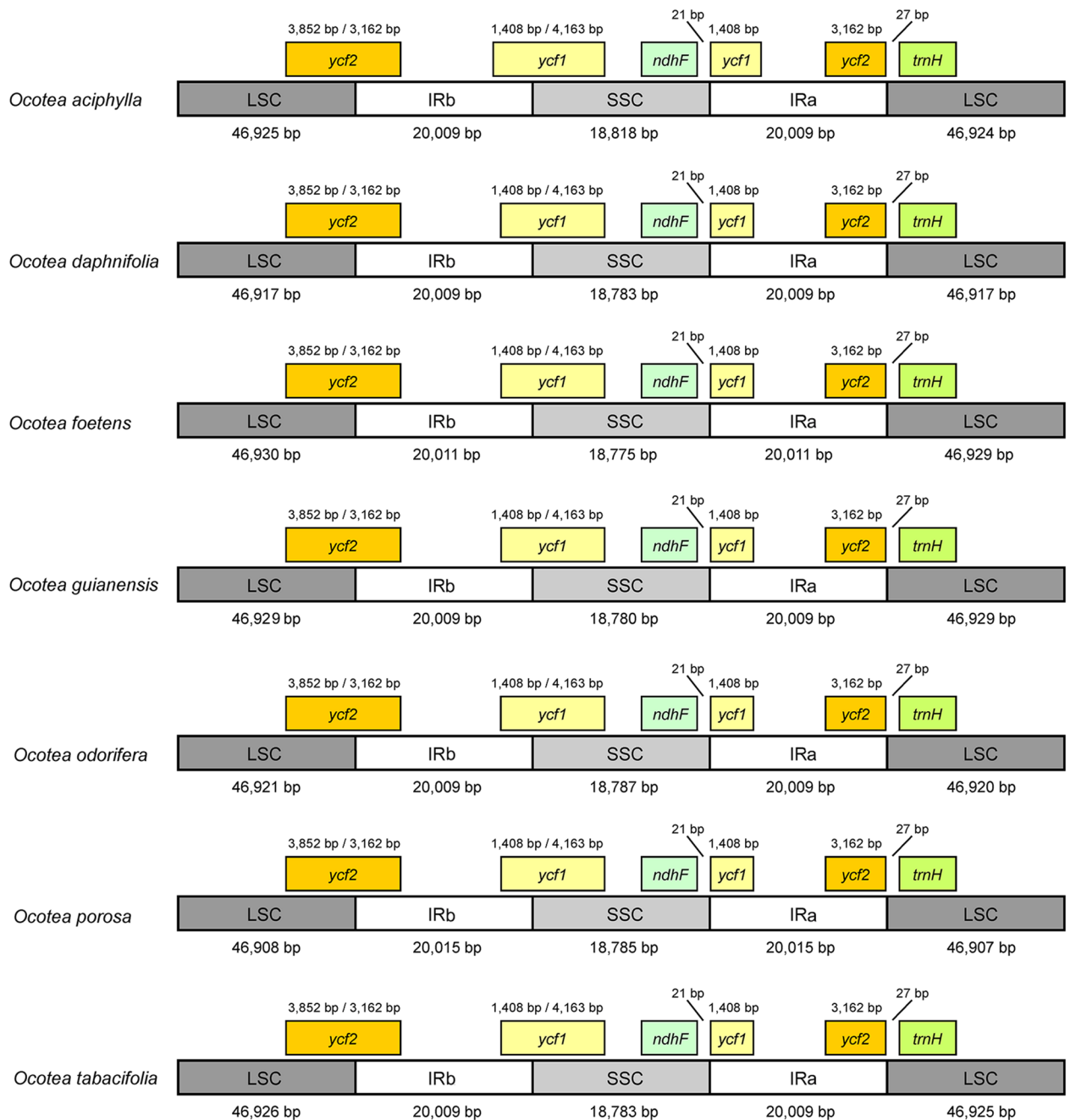


Figure 3. Comparison of LSC, IR, and SSC junction positions among seven *Ocotea* chloroplast genomes.

tabacifolia, than in the other four Neotropical species (87 vs. 75–82 SSRs), which are bisexual or gynodioecious (*O. daphnifolia*). The plastome of the Macaronesian *Ocotea foetens* contains 84 SSRs. It remains to be checked if the number of SSRs is indeed correlated with larger clades within the *Ocotea* complex. Mononucleotide SSRs are very predominant in the chloroplast sequences of Lauraceae. The counts of mononucleotide SSRs varied from 54 to 65 in *Litsea*, *Machilus*, *Cinnamomum* and *Persea* species¹⁶. In *Ocotea*, we detected 57–67 mononucleotide SSRs. Among the Neotropical *Ocotea* species, we found the highest numbers in the two dioecious species (66–67, vs. 57–61 in the other four species). The Macaronesian *Ocotea foetens* showed 63 mononucleotide SSRs. Hexanucleotide repeats were rare in all Lauraceae species examined so far. No hexanucleotide SSRs were found in *Ocotea daphnifolia* and *O. odorifera*. However, we detected three hexanucleotide repeats in *Ocotea porosa*, instead of only one in most other Lauraceae. The numbers of SSRs in the LSC, SSC and IR regions were similar for all Lauraceae species studied. Chen et al.¹⁶ detected 63, 16 and four SSRs in the LSC, SSC and IR regions of *Cinnamomum camphora* vs. 62–70, 11–17 and one or two SSRs in the *Ocotea* species in our study.

As expected, addition of the seven *Ocotea* species does not change the result of the phylogenetic analysis significantly compared to previous cp genome studies^{28,33}. The topology among the major clades, Cinnamomeae, Laureae and the Perseeae, is the same in all studies. As expected, the seven *Ocotea* species form a monophyletic

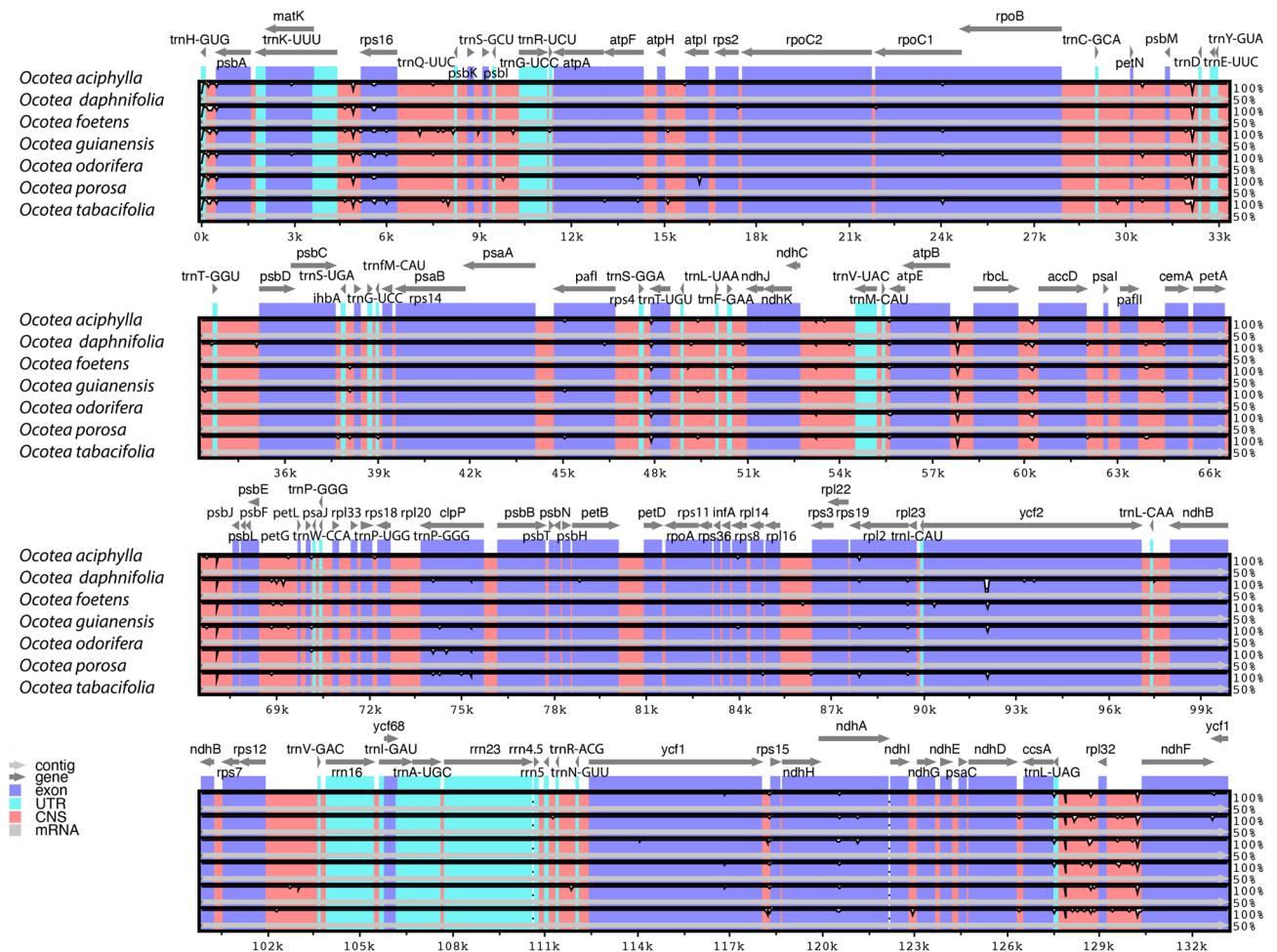


Figure 4. Visualization of *Ocottea* chloroplast genomes using mVISTA program with *O. aciphylla* as reference. CNS conserved non-coding sequence; UTR untranslated region.

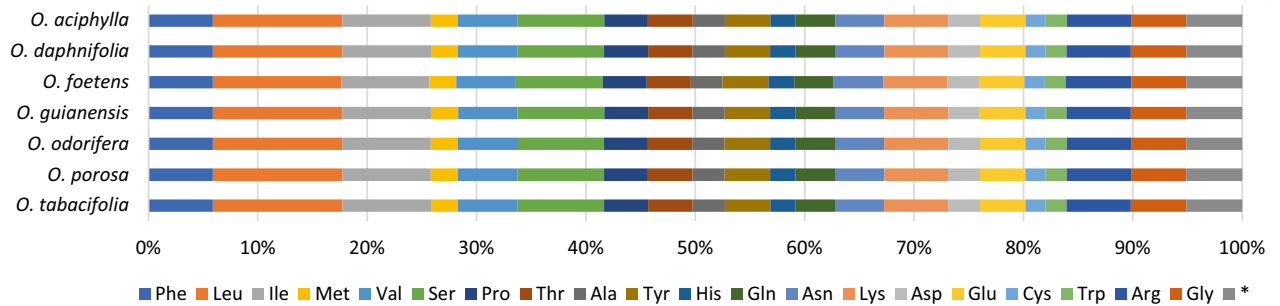


Figure 5. Amino acid percentage of protein-coding genes in seven *Ocottea* plastomes.

group that is sister to *Cinnamomum* s.lat., i.e., including *Sassafras*. It is unfortunate that the cp genome of the taxon recorded as ‘UNVERIFIED *Nectandra angustifolia*’ in GenBank (MF939340) is so divergent from all other Core Lauraceae that large parts of it could not even be aligned. Based on the results of earlier studies^{2,3,9,10,12}, *Nectandra* was expected to be nested in *Ocottea*, as sister taxon to the dioecious clade. Not only because of its aberrant sequence it is questionable if the species listed as *N. angustifolia* in the study of Song et al.²⁵ has been determined correctly. The real *N. angustifolia* is known from the type collection from Bahia only, so that it appears unlikely that it was cultivated in Sulawesi. Apart from *Nectandra angustifolia*, no complete plastomes have been sequenced so far in any of the genera that are usually found nested among the *Ocottea* species (*Aniba*, *Damburneya*, *Dicypellium*, *Endlicheria*, *Kubitzkia*, *Licaria*, *Mespilodaphne*, *Nectandra*, *Paraia*, *Pleurothyrium*, *Rhodostemonodaphne*, *Umbellularia* and *Urbanodendron*). The number of *Ocottea* species examined here is still too small to reach any definite conclusions about their phylogeny. There are, however, two differences compared to the recent study by Penagos et al.¹². In their study, the Old World *Ocottea* species (the clade named Palaeocottea) form the sister group to a clade named Praelicaria, which is represented by *Ocottea aciphylla*, *O. odorifera* and *O.*

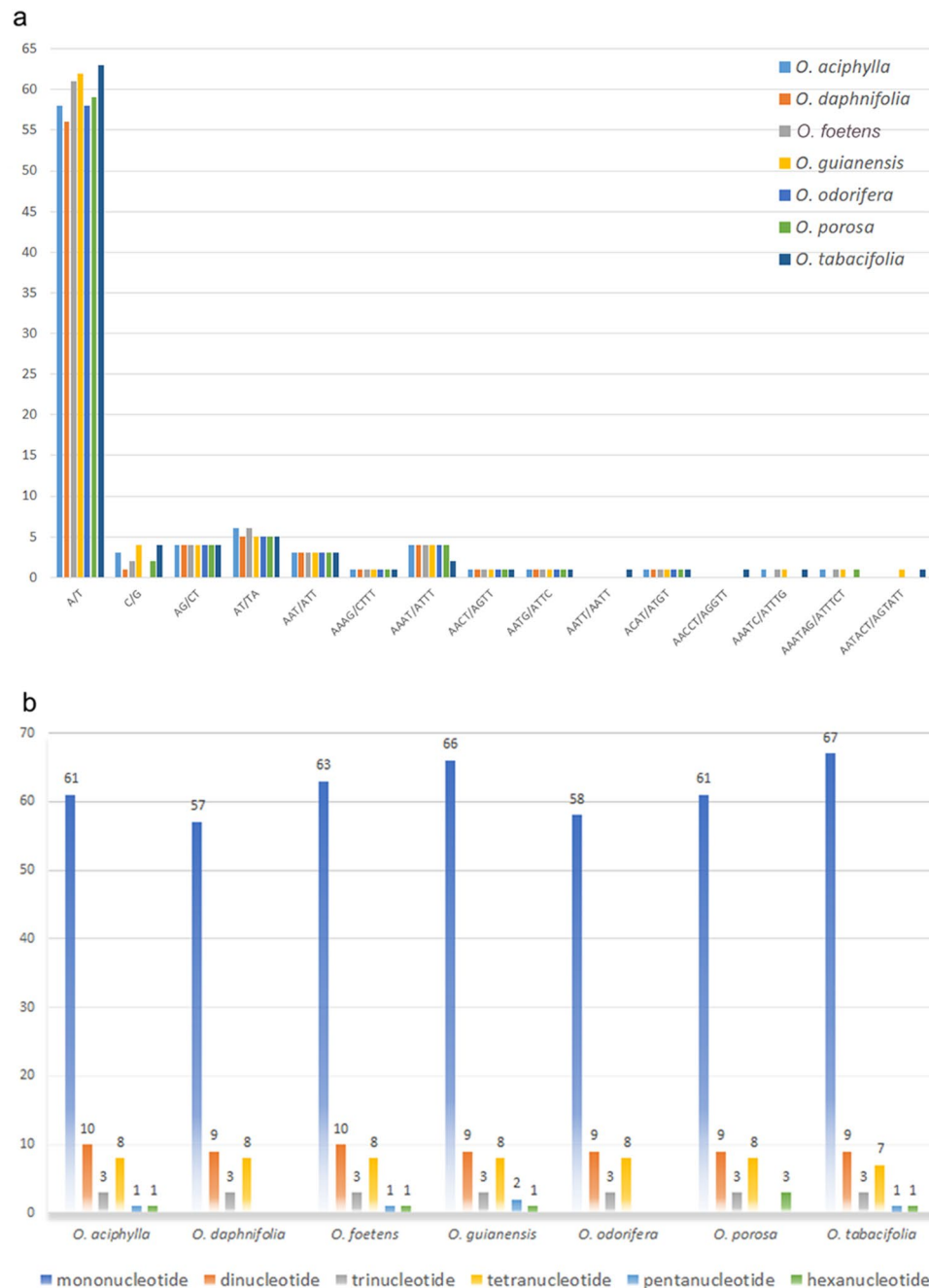


Figure 6. Simple sequence repeats (SSRs) in seven chloroplast genomes of *Ocotea*. (a) Counts of nucleotide repeats; (b) counts of mono-, di-, tri-, tetra-, penta- and hexanucleotides.

porosa in our study. *Ocotea daphnifolia*, which is nested among the Praelicaria taxa in our result, is a member of the *O. minarum* group and as such a member of the Pluriocotea clade in the study by Penagos et al.¹². In their result, the Pluriocotea clade is the sister group to a clade consisting of the dioecious taxa (Dioecotea, represented by *O. guianensis* and *O. tabacifolia* in our study), the *O. helicterifolia* group and the genera *Nectandra*, *Pleurothyrium* and *Damburneya*, which are not represented in our study. It needs to be checked if these differences persist when further cp genomes become available. As expected, entire plastomes have the potential to increase resolution and support values among the clades of the *Ocotea* complex. Our phylogeny is fully resolved, and not only the *Ocotea* complex receives 100% bootstrap support, but also four of the five nodes within it. There is still one node that is scarcely supported, but that may change with denser taxon sampling.

Sequence divergence among the seven *Ocotea* species is rather low, compared to the most closely related, likewise species-rich genera *Cinnamomum*, *Lindera* and *Litsea*. Even though we selected *Ocotea* species from widely divergent clades, there were only 168 parsimony-informative characters among them in the entire chloroplast genomes. If we arbitrarily select the first seven species of *Cinnamomum*, *Lindera* or *Litsea* from our data matrix, these numbers are 414, 423 or 410, respectively. This confirms the results of the tests of individual established

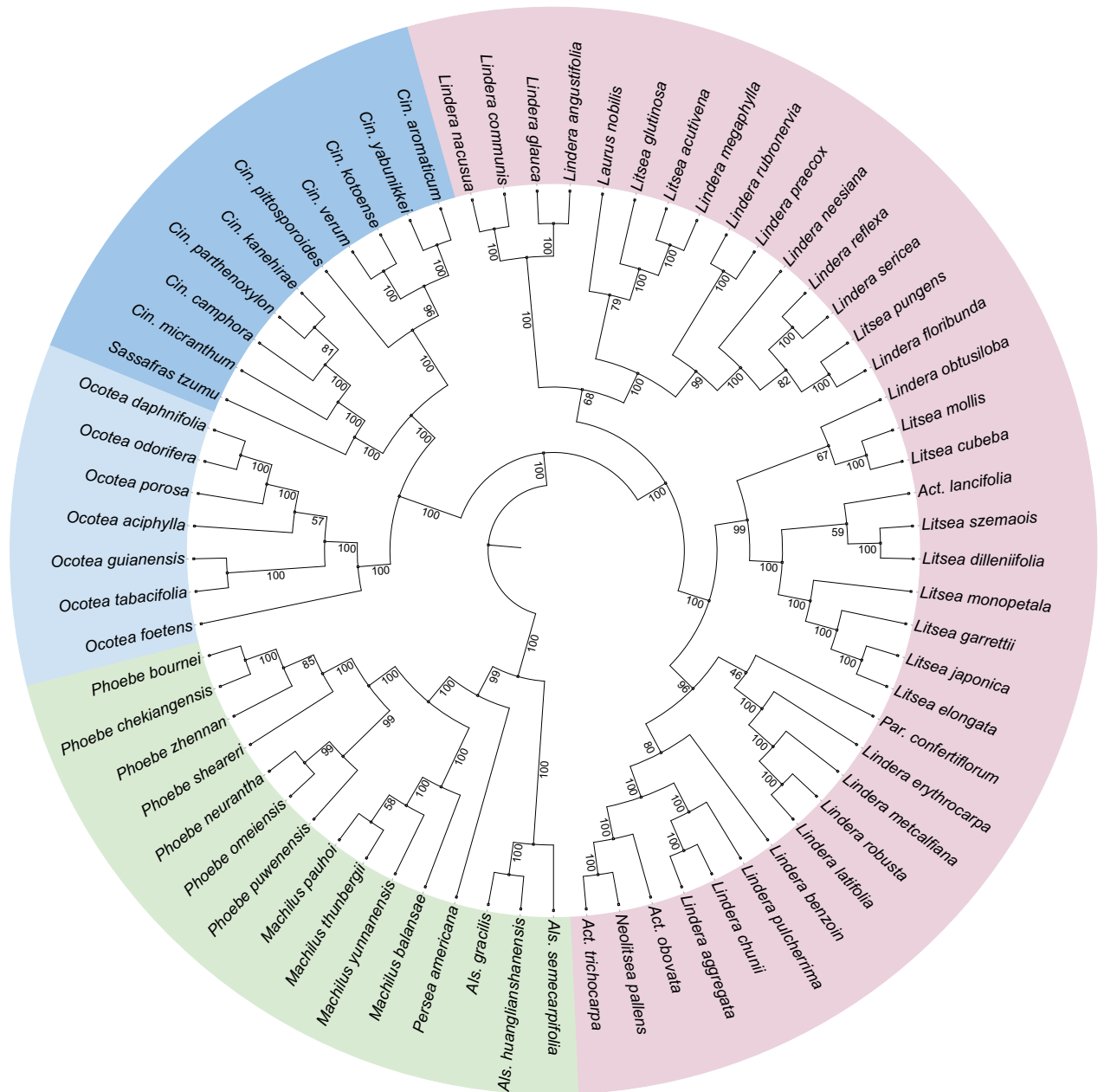


Figure 7. Maximum Likelihood phylogeny including 69 complete chloroplast genome sequences of Core Lauraceae. Color codes: blue—Cinnamomeae, incl. pale blue for *Ocotea* spp.; pink—Laureae; green—Perseeae. Maximum likelihood bootstrap support values (ML-BS) are shown next to the branches.

chloroplast markers mentioned in the introduction, and may point to a rather recent diversification of the *Ocotea* complex, as was first suggested by Chanderali et al.¹⁰. However, a much larger number of sequences will be required for a molecular clock analysis of this group.

Materials and methods

Plant materials. Silica-gel dried leaf material of seven *Ocotea* species, *O. aciphylla*, *O. daphnifolia*, *O. foetens*, *O. guianensis*, *O. odorifera*, *O. porosa*, and *O. tabacifolia*, was used for the present analysis (Supplementary Table S7). According to previous analyses^{2,3,12}, these species belong to different clades within the genus, except *Ocotea odorifera* and *O. porosa* from the *O. indecora* group. The plant material was collected in accordance with the relevant institutional, national, and international guidelines and legislation. PLRM obtained the collecting permits for the material collected in Brazil 2011. *Ocotea foetens* was collected in the Botanical Garden of Berlin, with permission of the curator G. Parolly, from a tree of unknown origin that had been growing in the garden for decades. Voucher specimens are deposited in the herbarium Rioclarense (HRCB) at the Universidade Estadual Paulista, Rio Claro (Brazil), the herbarium Hamburgense (HBG) at the University of Hamburg (Germany), and the garden herbarium of the Botanical Garden and Botanical Museum Berlin (Germany).

DNA preparation and chloroplast sequencing. DNA was isolated with the innuPREP Plant DNA Kit (Analytik Jena, Germany) according to the manufacturer's protocol, with modifications^{9,37}. DNA libraries were built using the QIAseq FX DNA Library Kit (Qiagen, Germany) and 120 ng of each DNA. Normalized samples were pooled and sequenced using the 300-cycles (2 × 150 bp paired-end) MiSeq reagent kit v3 (Illumina, San Diego, CA) on a MiSeq platform at the NGS Core Facility at the Bernhard Nocht Institute for Tropical Medicine, Hamburg, Germany. The generated raw reads were first checked qualitatively, with Phred quality score < 20 trimmed and filtered to remove polyclonal and low quality reads (< 55 bases long) using CLC workbench v. 20.0.1 (Qiagen).

Plastomes assembly and annotation. Analyses of genome sequence and genomic organization were performed using Geneious Prime 2021.0.3⁴⁴. The generated contigs of *Ocotea foetens* were assembled de novo and annotated using the plastomes of *Cinnamomum camphora* (GenBank accession number MH050970) and *Persea americana* (NC_031189) for comparison. The contigs of the remaining taxa were assembled and annotated using the chloroplast genome of *O. foetens* as a reference. The contigs were inspected visually for any signs of erroneous assembly. In a few cases, doubtful regions were verified by Sanger sequencing (methods described earlier^{2,3,9,11}). The circular plastome maps of *Ocotea* were drawn using OGDRAW v1.2^{29,33–38,44–47}.

Determination of the most variable regions of plastomes. The chloroplast genomes of seven *Ocotea* species and 63 other Lauraceae were downloaded from the NCBI GenBank (Supplementary Table S8). All 70 sequences were aligned using MAFFT v7⁴⁸ with default parameters. Visual inspection of the alignment showed that large parts of the sequence of *Nectandra angustifolia* could not be aligned with confidence, so that this species had to be removed. Ten small inversions (5–39 base pairs), bordered by long palindromic sequences, were identified and reversed, because earlier analyses had shown that the orientation of such hairpin loops varies even within a single population. In the final alignment, these inversions correspond to positions 272–276, 480–487, 29,786–29,810, 66,770–66,809, 69,509–69,522, 70,298–70,314, 117,696–117,701 (in *Parasassafra* only), 126,376–126,385, 132,449–132,505 and 140,474–140,479 (in *Parasassafra* only). Also a few additional minor adjustments of the alignment were made manually during inspection of the sequences, mostly in regions of SSRs. DnaSP v6⁴⁹ was used for calculating the nucleotide variability values (Pi) within the plastomes. The sliding window length was set to 600 bp, and the step size was set to 200 bp. Microsoft Excel⁵⁰ was used to plot the Pi values. These data were used to identify hypervariable regions among the seven *Ocotea* plastomes examined as well as among the sequences retrieved from the NCBI GenBank (Supplementary Tables S7, S8).

Comparative analysis of *Ocotea* plastomes. A comparison of the LSC, IR and SSC junction positions in the *Ocotea* plastomes was carried out in Geneious Prime 2021.0.3⁴⁴. The mVISTA program in Shuffle-LAGAN mode^{42,43} was used for the visualization of the differences in the seven *Ocotea* chloroplast genomes.

Codon usage and SSRs analyses. The protein-coding genes of *Ocotea* plastomes were extracted using the program Geneious Prime 2021.0.3⁴⁴. The sequences were aligned using MAFFT v7⁴⁸. Codon usage frequency, Codon Bias, and G + C content were calculated using the program DnaSP v6.

The SSR motifs were scanned using MISA v2.1⁵¹. The minimum thresholds were set to 10 repetitions for mononucleotide SSRs, five repeat units for dinucleotide SSRs, four repetitions for trinucleotide SSRs and three repetitions for tetra-, penta- and hexanucleotide SSRs. The maximum length of interruption between two SSRs was chosen as 100 bp.

Phylogenetic analysis of Lauraceae plastomes. The data matrix that had been prepared for the determination of the most variable regions was analyzed using maximum likelihood analyses (ML) in MEGA 10.2.5⁵², with the following parameters: nrep = 500, Tamura-Nei model, uniform rates among sites and Nearest-Neighbor-Interchange (NNI). The chloroplast genomes of the Perseeae (*Alseodaphne* spp., *Machilus* spp., *Persea americana*, and *Phoebe* spp.) were used as outgroup.

Data availability

The complete cp genome sequences of the seven *Ocotea* species have been submitted to the NCBI GenBank.

Received: 16 July 2021; Accepted: 28 December 2021

Published online: 21 January 2022

References

- Rohwer, J. G. Lauraceae. In *The Families and Genera of Vascular Plants* Vol. 2 (eds Kubitzki, K. et al.) (Springer, 1993).
- Trofimov, D., Moraes, P. L. R. & Rohwer, J. G. Towards a phylogenetic classification of the *Ocotea* complex (Lauraceae)—Classification principles and reinstatement of *Mespilodaphne*. *Bot. J. Linn. Soc.* **190**, 25–50. <https://doi.org/10.1093/botlinnean/boz010> (2019).
- Trofimov, D. & Rohwer, J. G. Towards a phylogenetic classification of the *Ocotea* complex (Lauraceae)—An analysis with emphasis on the Old World taxa and description of the new genus *Kuloa*. *Bot. J. Linn. Soc.* **192**, 510–535. <https://doi.org/10.1093/botlinnean/boz088> (2020).
- Rohwer, J. G. Prodrum einer Monographie der Gattung *Ocotea* Aubl. (Lauraceae), sensu lato. *Mitt. Inst. Allg. Bot. Hamburg* **1986**(20), 3–278 (1986).
- van der Werff, H. Studies in Malagasy Lauraceae II: New taxa. *Novon* **6**, 463–475. <https://doi.org/10.2307/3392057> (1996).
- van der Werff, H. A synopsis of *Ocotea* (Lauraceae) in Central America and Southern Mexico. *Ann. Missouri Bot. Gard.* **89**, 429–451. <https://doi.org/10.2307/3298602> (2002).

7. van der Werff, H. A revision of the genus *Ocotea* Aubl. (Lauraceae) in Madagascar and the Comoro Islands. *Adansonia* **35**, 235–279. <https://doi.org/10.5252/a2013n2a5> (2013).
8. van der Werff, H. Studies in Andean *Ocotea* (Lauraceae) IV Species with unisexual flowers and densely pubescent leaves, or with erect pubescence or domatia, occurring above 1000 m in altitude. *Novon* **25**, 343–393. <https://doi.org/10.3417/2016021> (2017).
9. Trofimov, D., Rudolph, B. & Rohwer, J. G. Phylogenetic study of the genus *Nectandra* (Lauraceae), and reinstatement of *Damburneya*. *Taxon* **65**(5), 980–996. <https://doi.org/10.12705/655.3> (2016).
10. Chandlerbali, A. S., van der Werff, H. & Renner, S. S. Phylogeny and historical biogeography of Lauraceae: Evidence from the chloroplast and nuclear genomes. *Ann. Missouri Bot. Gard.* **88**, 104–134. <https://doi.org/10.2307/2666133> (2001).
11. Rohde, R. *et al.* Neither *Phoebe* nor *Cinnamomum*—The tetrasporangiate species of *Aiouea* (Lauraceae). *Taxon* **66**(5), 1085–1111. <https://doi.org/10.12705/665.6> (2017).
12. Penagos Zuluaga, J. C. *et al.* Resolved phylogenetic relationships in the *Ocotea* complex (Supraocotea) facilitate phylogenetic classification and studies of character evolution. *Am. J. Bot.* **108**(4), 1–16. <https://doi.org/10.1002/ajb2.1632> (2021).
13. Palmer, J. D. Comparative organization of chloroplast genomes. *Annu. Rev. Genet.* **19**, 325–354. <https://doi.org/10.1146/annurev.ge.19.120185.001545> (1985).
14. Chumley, T. W. *et al.* The complete chloroplast genome sequence of *Pelargonium x hortorum*: Organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. *Mol. Biol. Evol.* **23**(11), 2175–2190. <https://doi.org/10.1093/molbev/msl089> (2006).
15. Ruhlman, T. A. & Jansen, R. K. The plastid genomes of flowering plants. *Methods Mol. Biol.* **1132**, 3–38. https://doi.org/10.1007/978-1-62703-995-6_1 (2014).
16. Chen, C. *et al.* The complete chloroplast genome of *Cinnamomum camphora* and its comparison with related Lauraceae species. *PeerJ* **5**, e3820. <https://doi.org/10.7717/peerj.38202017> (2017).
17. Hinsinger, D. D. & Strijk, J. S. Toward phylogenomics of Lauraceae: The complete chloroplast genome sequence of *Litsea glutinosa* (Lauraceae), an invasive tree species on Indian and Pacific Ocean islands. *Plant Gene* **9**, 71–79. <https://doi.org/10.1016/j.plgene.2016.08.002> (2017).
18. Jo, S., Kim, Y. K., Cheon, S. H., Fan, Q. & Kim, K. J. Characterization of 20 complete plastomes from the tribe Laureae (Lauraceae) and distribution of small inversions. *PLoS ONE* **14**(11), e0224622. <https://doi.org/10.1371/journal.pone.0224622> (2019).
19. Liao, Q., Ye, T. & Song, Y. Complete chloroplast genome sequence of a subtropical tree, *Parasassafras confertiflorum* (Lauraceae). *Mitochondrial DNA B* **3**(2), 1216–1217. <https://doi.org/10.1080/23802359.2018.1532331> (2018).
20. Wang, Q. *et al.* The complete chloroplast genome sequence of *Litsea cubeba*. *Mitochondrial DNA B* **5**(3), 2193–2194. <https://doi.org/10.1080/23802359.2020.1768961> (2020).
21. Liao, Q., Ye, T. & Song, Y. Complete chloroplast genome sequence of a subtropical tree, *Parasassafras confertiflorum* (Lauraceae [sic!]). *Mitochondrial DNA B* **3**(2), 1216–1217. <https://doi.org/10.1080/23802359.2018.1532331> (2018).
22. Qiu, Q., Yang, D., Xu, L., Xu, Y. & Wang, Y. The complete chloroplast genome sequence of *Litsea garrettii*. *Mitochondrial DNA B Resour.* **5**(1), 1105–1106. <https://doi.org/10.1080/23802359.2020.1768961> (2020).
23. Song, Y. *et al.* Comparative analysis of complete chloroplast genome sequences of two tropical trees *Machilus yunnanensis* and *Machilus balansae* in the family Lauraceae. *Front. Plant Sci.* **6**, 662. <https://doi.org/10.3389/fpls.2015.00662> (2015).
24. Song, Y., Yao, X., Tan, Y., Gan, Y. & Corlett, R. T. Complete chloroplast genome sequence of the avocado: Gene organization, comparative analysis, and phylogenetic relationships with other Lauraceae. *Can. J. For. Res.* **46**, 1293–1301. <https://doi.org/10.1139/cjfr-2016-0199> (2016).
25. Song, Y. *et al.* Evolutionary comparisons of the chloroplast genome in Lauraceae and insights into loss events in the Magnoliids. *Gen. Biol. Evol.* **9**(9), 2354–2364. <https://doi.org/10.1093/gbe/evx180> (2017).
26. Song, Y., Yao, X., Liu, B., Tan, Y. & Corlett, R. T. Complete plastid genome sequences of three tropical *Alseodaphne* trees in the family Lauraceae. *Holzforchung* **72**, 337–345. <https://doi.org/10.1515/hf-2017-0065> (2018).
27. Song, Y. *et al.* Comparative analysis of complete chloroplast genome sequences of two subtropical trees, *Phoebe sheareri* and *Phoebe omeiensis* (Lauraceae). *Tree Genet. Genomes* **13**, 120. <https://doi.org/10.1007/s11295-017-1196-y> (2017).
28. Song, Y. *et al.* Plastid phylogenomics improve phylogenetic resolution in the Lauraceae. *J. Syst. Evol.* **58**(4), 423–439. <https://doi.org/10.1111/jse.12536> (2020).
29. Wu, C. C., Chu, F. H., Ho, C. K., Sung, C. H. & Chang, S. H. Comparative analysis of the complete chloroplast genomic sequence and chemical components of *Cinnamomum micranthum* and *Cinnamomum kanehirae*. *Holzforchung* **71**(3), 189–197. <https://doi.org/10.1515/hf-2016-0133> (2017).
30. Xiao, T. W. *et al.* Conflicting phylogenetic signals in plastomes of the tribe Laureae (Lauraceae). *PeerJ* **8**, e10155. <https://doi.org/10.7717/peerj.10155> (2020).
31. Yuan, X., Li, Y. & Wang, Y. The complete chloroplast genome sequence of *Cinnamomum kotoense*. *Mitochondrial DNA B Resour.* **5**(1), 331–332. <https://doi.org/10.1080/23802359.2019.1703604> (2019).
32. Zhang, J., Li, Y. & Wang, Y. The complete chloroplast genome sequence of *Phoebe puwenensis*. *Mitochondrial DNA B Resour.* **5**(1), 218–219. <https://doi.org/10.1080/23802359.2019.1699469> (2019).
33. Zhao, M.-L. *et al.* Comparative chloroplast genomics and phylogenetics of nine *Lindera* species (Lauraceae). *Sci. Rep.* **8**, 8844. <https://doi.org/10.1038/s41598-018-27090-0> (2018).
34. Ma, P. F., Zhang, Y. X., Zeng, C. X., Guo, Z. H. & Li, D. Z. Chloroplast phylogenomic analyses resolve deep-level relationships of an intractable bamboo Tribe Arundinarieae (Poaceae). *Syst. Biol.* **63**(6), 933–950. <https://doi.org/10.1093/sysbio/syu054> (2014).
35. Yang, Y. C. *et al.* Comparative analysis of the complete chloroplast genomes of five *Quercus* species. *Front. Plant Sci.* **7**, 959. <https://doi.org/10.3389/fpls.2016.00959> (2016).
36. Zhang, S. D. *et al.* Diversification of Rosaceae since the late cretaceous based on plastid phylogenomics. *New Phytol.* **214**, 1355–1367. <https://doi.org/10.1111/nph.14461> (2017).
37. Rohwer, J. G. & Rudolph, B. Jumping genera: the phylogenetic positions of *Cassytha*, *Hypodaphnis*, and *Neocinnamomum* (Lauraceae) based on different analyses of trnK intron sequences. *Ann. Missouri Bot. Gard.* **92**(2), 153–178 (2005).
38. Shin, H. W. & Lee, N. S. Correction: Understanding plastome evolution in Hemiparasitic Santalales: Complete chloroplast genomes of three species, *Dendrotrophe varians*, *Helixanthera parasitica*, and *Macrosolen cochinchinensis*. *PLoS ONE* **13**(10), e0205616. <https://doi.org/10.1371/journal.pone.0205616> (2018).
39. Rabah, S. O. *et al.* Plastome Sequencing of ten nonmodel crop species uncovers a large insertion of mitochondrial DNA in cashew. *Plant Genome*. <https://doi.org/10.3835/plantgenome2017.03.0020> (2017).
40. Li, Y., Xu, W., Zou, W., Jiang, D. & Liu, X. Complete chloroplast genome sequences of two endangered *Phoebe* (Lauraceae) species. *Bot. Stud.* **58**(1), 37. <https://doi.org/10.1186/s40529-017-0192-8> (2017).
41. Liu, D., Liu, D., Li, M. & Chen, S. The complete chloroplast genome of *Phoebe zhennan*. *Mitochondrial DNA B* **4**(1), 1564–1565. <https://doi.org/10.1080/23802359.2019.1601525> (2019).
42. Brudno, M. *et al.* Comparative sequencing program. LAGAN and multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* **13**(4), 721–731. <https://doi.org/10.1101/gr.926603> (2003).
43. Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M. & Dubchak, I. VISTA: Computational tools for comparative genomics. *Nucl. Acids Res.* **32**, W273–W279. <https://doi.org/10.1093/nar/gkh458> (2004).
44. Kearsse, M. *et al.* Geneious basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649. <https://doi.org/10.1093/bioinformatics/bts199> (2012).

45. Tillich, M. *et al.* GeSeq—Versatile and accurate annotation of organelle genomes. *Nucl. Acids Res.* **45**, W6–W11. <https://doi.org/10.1093/nar/gkx391> (2017).
46. Kent, W. J. BLAT—The BLAST-like alignment tool. *Genome Res.* **12**(4), 656–664. <https://doi.org/10.1101/gr.229202> (2002).
47. Lohse, M., Drechsel, O. & Bock, R. OrganellarGenomeDRAW (OGDRAW)—A tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. *Curr. Genet.* **52**, 267–274. <https://doi.org/10.1007/s00294-007-0161-y> (2007).
48. Katoh, K., Rozewicki, J. & Yamada, K. D. MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. *Brief. Bioinform.* **20**(4), 1160–1166. <https://doi.org/10.1093/bib/bbx108> (2019).
49. Rozas, J. *et al.* DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Mol. Biol. Evol.* **34**(12), 3299–3302. <https://doi.org/10.1093/molbev/msx248> (2017).
50. Microsoft Excel. Microsoft Corporation (2018). Microsoft Excel.
51. Beier, S., Thiel, T., Münch, T., Scholz, U. & Mascher, M. MISA-web: A web server for microsatellite prediction. *Bioinformatics* **33**, 2583–2585. <https://doi.org/10.1093/bioinformatics/btx198> (2017).
52. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549. <https://doi.org/10.1093/molbev/msy096> (2018).

Acknowledgements

We thank the curator of HBG, Matthias Schultz, for allowing DNA extraction from some specimens. We are also thankful Stefan Wanke and Matthias Jost (both TU Dresden) for instructions on the de novo method. The collection of plant material was supported by a PROP-G-UNESP (Internacionalização dos Programas de Pós-Graduação, Edital 02/2011) grant to one of us (PLRM).

Author contributions

D.T., J.G.R., D.C. and J.S.-C. designed the experiment; D.T. and D.C. performed the experiment; P.L.R.d.M. collected the samples; D.T. analysed the results; D.T. and J.G.R. wrote the manuscript; all authors revised the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-021-04635-4>.

Correspondence and requests for materials should be addressed to D.T. or J.G.R.

Reprints and permissions information is available at www.nature.com/reprints.

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



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

© The Author(s) 2022