Calophyllaceae plastomes, their structure and insights in relationships within the clusioids

Rafaela Jorge Trad1,2, Fernanda Nunes Cabral3, Volker Bittrich5, Saura Rodrigues da Silva4 & Maria do Carmo Estanislau do Amaral1

A complete chloroplast genome is not yet available for numerous species of plants. Among the groups that lack plastome information is the clusioid clade (Malpighiales), which includes five families: Bonnetiaceae, Calophyllaceae, Clusiaceae, Hypericaceae, and Podostemaceae. With around 2200 species, it has few published plastomes and most of them are from Podostemaceae. Here we assembled and compared six plastomes from members of the clusioids: five from Calophyllaceae (newly sequenced) and one from Clusiaceae. Putative regions for evolutionary studies were identified and the newly assembled chloroplasts were analyzed with other available chloroplasts for the group, focusing on Calophyllaceae. Our results mostly agree with recent studies which found a general conserved structure, except for the two Podostemaceae species that have a large inversion (trnK-UUU–rbcL) and lack one intron from ycf3. Within Calophyllaceae we observed a longer LSC and reduced IRs in Mahurea exstipulata, resulting in some genic rearrangement, and a short inversion (psbJ–psbE) in Kielmeyera coriacea. Phylogenetic analyses recovered the clusioids and the five families as monophyletic and revealed that conflicts in relationships reported in the literature for the group agree with nodes concentrating uninformative or conflicting gene trees. Our study brings new insights about clusioid plastome architecture and its evolution.

Since the advance of next-generation sequencing complete plastomes or chloroplast coding sequences have been the most used source of information to explore phylogenetic relationships among plants at diverse scales. This is due to their high number of copies in a cell (1000 to 10,000), their much smaller size when compared to nuclear genomes (ca. 120 genes vs. thousands of genes), and the availability of numerous tools to assemble chloroplast genomes (e.g., NOVOPlasty, GetOrganelle, Geneious, MITOBim). Thus, it is usually easier to obtain the DNA from the chloroplast when compared to the nucleus. In addition, sequencing the chloroplast is cost advantageous, and the generated data require less computational power to be processed and analyzed. Despite the existence of a canonical chloroplast structure for land plants, which includes two unique regions (the large single copy-LSC, and the small single copy-SSC), and one duplicated region (the inverted repeat-IR), much variation has been already found at different taxonomic levels. Finally, most of the chloroplast genes are functional, thus usually conserved. These reasons make plastid DNA extremely attractive for high throughput sequencing. Furthermore, the possibility of using the chloroplast genome from a phylogenetically distant taxon as a reference for the assembly accelerates data processing. However, it is important to mention that the analysis of nuclear genes is becoming much easier.

Although the chloroplast genome is generally imagined as being a quadripartite circular structure, there is evidence that these circular molecules represent only a small fraction of the total chloroplast DNA. Indeed, chloroplast chromosomes are hypothesized to consist of branched and complex multigenomic forms. Even though phylogenetic conflict within plastid data has been known for over two decades, it was largely neglected until recently, when the traditional approach of treating the plastome as a single “supergene” has been questioned, creating some debate (see Doyle 2021). There is a growing body of evidence showing that plastid regions contribute differentially to phylogenetic support and that several of the plastome genes actually are uninformative. Finally, heteroplasmacy was recently quantified and may be more common than previously.

1Department of Plant Biology, Biology Institute, University of Campinas (UNICAMP), CP 6109, Campinas, SP 13083-970, Brazil. 2Macroeconomy Lab @ J3-166, Institute of Biological Sciences – ICB, Federal University of Minas Gerais (UFMG), Belo Horizonte, Campinas, MG 31270-901, Brazil. 3Departamento de Ciências e Linguagens, Instituto Federal de Minas Gerais – Campus Bambuí, Bambuí, MG 38900-000, Brazil. 4Department of Technology, UNESP - São Paulo State University, Campus Jaboticabal, Jaboticabal, SP 14884-900, Brazil. 5Volker Bittrich is an independent scientist, Campinas, Brazil. 6Email: rafajt@hotmail.com
expected. This phenomenon allows the occurrence of heteroplasmic recombination, which is a potential source of gene tree conflict. Besides heteroplasmic recombination, heterotachy, stochastic and systematic errors and horizontal gene transfer have been invoked as potential causes of gene tree conflict. These findings are already changing how we see plastomes and will probably influence future works using plastid data.

The highly diverse order Malpighiales includes 36 recognized families in its most recent circumscription and represents ca. 7.8% of eudicot diversity. Intrafamilial relationships remained poorly understood until Xi and collaborators (2012) presented a phylogeny based on 82 plastid, six mitochondrial, and three nuclear genes and suggested relationships between families in the order, albeit some were only poorly supported. Indeed, recently Cai and collaborators (2021) described what they called “the perfect storm” —a combination of incomplete lineage sorting, introgression, and gene tree estimation error that may occur simultaneously during periods of rapid radiation, making phylogenetic inference challenging—and showed the difficulties in resolving relationships especially along the backbone of Malpighiales. Within the order, the clusioid clade includes five families (Bonnetiaceae, Calophyllaceae, Clusiaceae, Hypericaceae and Podostemaceae), 94 genera and ca. 2200 species. Relationships within the clusioids had remained unclear—indeed, the very existence of the group was questionable—until somewhat over a decade ago. For example, Wawra (1886) included some of the taxa in the Ternstroemiaceae, today included in the Theaceae (Ericales), and Cronquist (1981) separated them in different orders and subclasses, while where Podostemaceae were to be placed was almost literally anybody’s guess. On the other hand, the families Hypericaceae and Calophyllaceae have frequently been included in Clusiaceae s.l. (sometimes called Guttiferae).

The relationships between the families included in the clusioid clade were explored in more depth by Ruhiel and collaborators (2011, 2013). The authors recognized two major clades: (Clusiaceae + Bonnetiaceae) and (Calophyllaceae (Hypericaceae + Podostemaceae)). Xi and collaborators (2012) recovered the five clusioid families as monophyletic and in a well-supported clade (85% bootstrap/0.91 posterior probability). Recently, Cai and collaborators (2021) also recovered the clusioid families in a well-supported clade using 423 single-copy nuclear loci, but internal relationships were very different—(Hypericaceae (Clusiaceae (Calophyllaceae + Bonnetiaceae)))—although poorly supported; Podostemaceae were not included. The conflicting relationships within the clusioids may result from the above-mentioned factors, i.e., “the perfect storm,” but also indicate that nuclear and plastid data have conflicting evolutionary stories, since most of Xi et al. data came from plastid genes. The use of chloroplast genomic data to investigate intergeneric relationships and/or genome within the clusioids is recent. The studies have focused mainly on Clusiaceae and Podostemaceae. Jin and collaborators (2020) included Calophyllaceae, but only one plastome, i.e., that of Mesua ferra L.

The family Calophyllaceae is pantropical and includes 14 genera and ca. 460 species. Within the Calophyllaceae two tribes are currently recognized: Endodesiumi, including two monotypic African genera (Endodesmis Benth. and Lebrunia Staner), and Calophyllaceae, including the remaining 12 genera (Calophyllum L., Caraipa Aubl., Clusia Planch. & Triana, Haplocratera Benth., Kayea Wall., Kielmeyera Mart. & Zucc., Mammea L., Mahurea Aubl., Marilia Sw., Mesua L., Neotuae Maguire, Pocoeulnzon Bodd.), which are distributed throughout the tropics. In both Ruhfel et al. and Cabral et al. Endodemia was well-supported as sister to the remaining genera of the family (Lebrunia was not included in either study). Relationships among the genera in Calophyllaceae are not completely understood, e.g., the position of Calophyllum + Mesua clade may be (1) sister to the clade Kayea + Pocoeulnzon, (2) sister to a clade that includes Mammea, Kayea and Pocoeulnzon, or (3) sister to a clade including the neotropical genera Caraipa, Clusia, Haplocarla, Kielmeyera, Mahurea, and Marilia.

Complete plastome information, including its annotation, is still not available for many species. Thus, the present work aims to reduce this knowledge gap for the family Calophyllaceae by presenting four complete and one nearly complete newly assembled and annotated plastid genomes from four different genera in the family (Calophyllum, Caraipa, Kielmeyera and Mahurea). One of the species Calophyllum brasiliense Cambess., Caraipa heterocarpa Ducke, Kielmeyera appariciana Siddi, K. coriacea Mart. & Zucc., and Mahurea exstipulata Benth. We also assembled and annotated the first plastid genome from Clusia L., the largest genus in the Clusiaceae. Our data were analyzed in the context of the clusioid clade, and we hope to improve our understanding of the evolution of the group.

**Results**

**Chloroplast structure.** The newly assembled plastomes presented a quadripartite structure with one LSC, one SSC and two IRs (Fig. 1). The total length ranged from 149,535 bp in Mahurea exstipulata to 160,253 bp in Calophyllum brasiliense, with mean depth coverage ranging from 52.0 to 474.7 reads. LSC, SSC and IR sizes and their respective GC content are presented in Table 1. Among the newly assembled plastomes M. exstipulata showed the most distinct plastome with the largest LSC (98,042 bp) and the smallest IRs (16,553 bp), respectively 9,923 to 12,143 bp longer and 8,966 to 10,781 bp shorter respectively than those of the other species. All species presented an SSC similar in size ranging from 17,464 bp in Caraipa heterocarpa to 160,253 bp. The total length ranged from 149,535 bp in Mahurea exstipulata to 160,253 bp in Calophyllum brasiliense, with mean depth coverage ranging from 52.0 to 474.7 reads. LSC, SSC and IR sizes and their respective GC content are presented in Table 1. Among the newly assembled plastomes M. exstipulata showed the most distinct plastome with the largest LSC (98,042 bp) and the smallest IRs (16,553 bp), respectively 9,923 to 12,143 bp longer and 8,966 to 10,781 bp shorter respectively than those of the other species. All species presented an SSC similar in size ranging from 17,464 bp in Calophyllum brasiliense to 19,102 bp in Clusia pananaparina (Aubl.) Choisy. The GC content was similar among all species (Table 1).

Five plastomes have 87 protein-coding genes, 37 transfer RNA (tRNA) and eight ribosomal (rRNA), totaling 132 genes. However, M. exstipulata had lost two tRNA genes and one of the copies of the rpl22 and ycf2 genes and both copies of the ycf15 gene, thus having a total of 125 genes in its plastome. A list of all genes is found in Table 2. Only clpP gene has two introns. The clusioids have lost one of the two ycf3 introns, thus this gene has two exons and one intron as in Jin et al. Among the duplicated genes in the IRs, there are seven tRNAs, eight rDNAs (seven in C. heterocarpa in which the ycf15 is pseudogenized, and four in M. exstipulata due to its shorter IR), and four rRNAs (Fig. 1). Overall, the gene content is similar within the clusioids, the two Podostemaceae (Marathrum foeniculaceum Bonpl. and Tristicha trifaria (Bory ex Willd.) Spreng.) being the most distinct.
Figure 1. Circular map of the five complete clusioid plastomes. (A) *Caraipa heterocarpa*, (B) *Kielmeyera appariciana*, (C) *K. coriacea*, (D) *Mahurea exstipulata*, (E) *Clusia panapanari*. The genes represented outside the circle are transcribed counterclockwise and those inside the outer circle are transcribed clockwise. Genes are colored according to their functional groups following the legend. The inner gray graphs indicate the GC content across the plastome.
Few differences in the limits of the four regions of the plastomes were observed (Fig. 2). The LSC-IRb limit is flanked by *rpl22* on the LSC side and by *rpl2* on the IRb side, with *rps19* spanning the junction in all species but *M. exstipulata*. This species lacks *rps19* in this region and has *ycf2* flanking the junction on the LSC side and *trnL* on the IRb side. The IRb-SSC junction always has the *ycf1*-fragment on IRb and the *ndhF* on SSC, while

**Table 1.** GenBank accession numbers and comparison of chloroplast genome size and GC content across three different regions (LSC, SSC, and IR) for 12 clusioid species. LSC large single copy, SSC small single copy, IR inverted repeat.

| Species                  | Family              | GenBank       | Coverage (mean) | Total reads | Mapped reads | LSC bp | GC % | SSC bp | GC % | IR bp | GC % | Full plastome bp | GC % |
|--------------------------|---------------------|---------------|-----------------|-------------|--------------|--------|------|--------|------|--------|------|------------------|------|
| Bonnetia paniculata      | Bonnetiaceae        | MK955182      | –               | –           | –            | 84,024 | 33.9 | 18,140 | 29.9 | 27,309 | 41.9 | 156,782           | 36.2 |
| Calophyllum brasiliense  | Calophyllaceae       | MW853786      | 52.0            | 8,922,582   | 61,660       | 88,119 | 34.2 | 17,464 | 30.6 | 27,334 | 42.2 | 160,253           | 36.5 |
| Caraipa heterocarpa      | Calophyllaceae       | MW853787      | 44.1            | 8,486,444   | 49,148       | 86,990 | 34.2 | 18,260 | 30.7 | 26,353 | 42.7 | 157,956           | 36.6 |
| Kielmeyera appariciana   | Calophyllaceae       | MW853788      | 447.3           | 8,424,602   | 504,031      | 87,648 | 34.2 | 18,300 | 30.5 | 25,519 | 42.7 | 158,986           | 36.6 |
| Kielmeyera coriacea      | Calophyllaceae       | MW853789      | 474.7           | 11,077,404  | 540,562      | 88,263 | 34.3 | 18,219 | 30.6 | 26,490 | 42.7 | 159,470           | 36.6 |
| Mahomma extipulata       | Calophyllaceae       | MW853790      | 298.2           | 7,486,882   | 333,152      | 98,042 | 34.6 | 18,395 | 30.7 | 16,553 | 45.4 | 149,535           | 36.5 |
| Mesua ferrea             | Calophyllaceae       | MK995181      | –               | –           | –            | 89,207 | 33.6 | 19,102 | 29.5 | 26,702 | 42.3 | 158,405           | 36.0 |
| Clusia panapanari        | Clusiaceae           | SRR7518735    | 78.6            | 2,452,164   | 89,207       | 85,899 | 33.6 | 19,102 | 29.5 | 26,702 | 42.3 | 158,405           | 36.0 |
| Cratoxylum cochinchinense| Clusiaceae           | MK995180      | –               | –           | –            | 85,640 | 34.0 | 18,892 | 29.9 | 26,272 | 42.1 | 157,076           | 36.2 |
| Garcinia gymnus-gutta    | Calophyllaceae       | NC_047250     | –               | –           | –            | 84,998 | 33.5 | 17,088 | 30.3 | 27,058 | 42.1 | 156,202           | 36.2 |
| Marathrum foetidesculent| Podostemaceae        | MK995178      | –               | –           | –            | 79,506 | 32.2 | 12,262 | 28.0 | 19,916 | 43.0 | 131,600           | 35.1 |
| Tristicha trifaria       | Podostemaceae        | MK995179      | –               | –           | –            | 79,002 | 33.7 | 12,717 | 30.7 | 19,623 | 43.5 | 130,967           | 36.3 |

**Table 2.** List of genes annotated in the six assembled plastomes: *Calophyllum brasiliense*, *Caraipa heterocarpa*: *Clusia panapanari*, *Kielmeyera appariciana*, *K. coriacea*, and *Mahomma extipulata*. *Genes with introns. Transpliced gene. *Present only in *M. exstipulata*. **Absent in *C. brasiliense*. ***Only three copies in *M. exstipulata.***
Long and simple sequence repeats (SSRs) and sequence polymorphism. REPuter$^{44}$ identified between 20 long repeats in *Tristicha trifaria* and 50 in *Cratoxylum cochinchinense* (Lour.) Blume most of them to be forward and palindromic repeats; complementary repeats were found only in *Garcinia gumi-gutta* (L.) N.Robson, which has two of them (Table 3). We found a repeat flanking rbcL in *Marathrum foeniculaceum* and a repeat flanking trnK in *T. trifaria*, each gene being at one end of the 50 kb inversion reported for the family. No repeats were found flanking the *psbI*–*psbE* inversion in *K. coriacea*, although they are often associated with genomic rearrangements$^{57}$. In almost all clusioid representatives, most long repeats were distributed throughout the LSC, the only exceptions were *B. paniculata* with 15/32 repeats in the LSC and *C. cochinchinense* with 21/50. The number of long repeats in the SSC ranged from one in *T. trifaria* to eight in *M. ferrea*, and in the IR it ranged from three in *M. foeniculaceum* to 31 in *C. cochinchinense* (Supplementary Figure S1). The presence of repeats in these genes was already noted in other studies$^{45-47}$. MISA$^{48,49}$ identified between 297 SSRs in *Caraipa heterocarpa* and 403 in *M. foeniculaceum*, most were mononucleotide A or T repeats. The two Podostemaceae species had fewer dinucleotide SSRs than all other species with 25 in *T. trifaria* and 28 in *M. foeniculaceum*. Tri- and tetranucleotide SSRs were found in all species. Pentanucleotide SSRs were found in almost all species (9/12), and hexanucleotide SSRs were found only in five species (Table 3).

Sequence polymorphism analyses indicated that the ten longest regions are equally distributed (four in IR, three in LSC and three in SSC) and that eight of them are CDSs. Nine of the ten regions with more segregating sites and more estimated mutations were the same; at least eight of them are CDSs, and six of them are in the LSC. All the ten regions with highest nucleotide diversity were intergenic spacers and seven out of ten are in the LSC. Overall, LSC and SSC had higher variability (nucleotide diversity between 2.9 and 3.5-fold, respectively) when compared to the IRs (Supplementary Table S2).

**Phylogenetic analyses.** Our 59 sequences complete alignment prior to the removal of poorly aligned regions on Gblocks$^{50}$ (nogb) has 77,015 bp, including 34,595 distinct patterns, 21,558 parsimony-informative sites, and 21.16% of gaps/missing data. The alignment processed on Gblocks$^{50}$ (gb) has 66,671 bp, including 30,582 distinct patterns, 19,548 parsimony-informative, and 12.94% gaps/missing data. We used three different methods (maximum-likelihood—ML, Bayesian inference—BI, and multispecies coalescent—MC) and eight different datasets (CU: 82 protein-coding genes concatenated unpartitioned—"supergene approach", CP: 82 protein-coding genes partitioned with individual evolutionary models, 82: a single consensus gene tree per locus used as input, 2050: 25 consensus gene trees per locus used as input; each of these four datasets have two versions: one without removal of poorly aligned regions—nogb, and one after removal using Gblocks$^{50}$—gb) in our analyses.

The trees we obtained exhibited the same topology for most relationships. In all analyses the clusioids were recovered in a strongly supported monophyletic clade with all the five families also being monophyletic.
Figure 3. Progressive Mauve alignment showing synteny and rearrangements for twelve clusoid plastomes. Brownish-orange bars represent the psbJ–psbE inversion in *Kielmeyera coriacea* and light blue bars represent the *trnK*–rbcL inversion in the two Podostemaceae species (*Marathrum foeniculaceum* and *Tristicha trifaria*).
Calophyllaceae was recovered sister to a clade composed by Podostemaceae and Hypericaceae. The position of Bonnetiaceae and Clusiaceae changed in different analyses and datasets. Within Calophyllaceae most relationships were stable with the two Kielmeyera species grouped in a clade, Caraipa heterocarpa being sister to this clade; Mahurea exstipulata was recovered sister to Kielmeyera + Caraipa clade, and Mesua ferrea and Calophyllum formed another clade. The position of Mammea americana was not stable and will be discussed below. Within Podostemaceae, Tristicha trifaria was recovered sister to the clade M. foeniculaceum + Podostemum ceratophyllum. In Hypericaceae, Cratoxylum cochinchinense was recovered sister to the clade Viaia ferruginea + Hypericum; relationships within Hypericum were not stable. In Clusiaceae, the two Clusia species included in the study grouped in a clade which was recovered sister to the clade including the two Garcinia species. The two Bonnetiaceae genera, represented by B. paniculata and Poinariart sp., were recovered in a clade. All stable branches mentioned above received 100% (ultrafast bootstrap—UB)/1 (posterior probability—PP) support values.

The three relationships that changed based on the analyses type and on the dataset were: (1) the position of the Bonnetiaceae and Clusiaceae, (2) the positions of Mammea (Calophyllaceae), and (3) the relationship between Hypericum fraseri Steud. and H. kalmianum L. (Hypericaceae). These relationships are represented in the network (Fig. 4). Conflicting relationships from 1 and 2 are summarized in Fig. 5. Filtering the poorly aligned regions, using partitioned schema, and allowing different evolutionary models usually increased support values in analyses and—importantly—changed the topology. For conflicting relationships 1, ML and BI unpartitioned analysis after removal of poorly aligned regions (CUgb), all ML and BI partitioned analyses (CPgb and CPnogb) and most MC species trees (82gb, 2050gb, 2050nogb) recovered Clusiaceae as sister to Bonnetiaceae (Fig. 5-IIA). Only ML and BI unpartitioned analysis without removal of poorly aligned regions (CUnogb), and MC species tree using a single gene tree per gene as input without removal of poorly aligned regions (82nogb) recovered Bonnetiaceae as sister to the remaining clusioid families (Fig. 5-IIIB). Support values for this relationship were generally lower to moderate. For conflicting relationship 2, all ML and BI consensus trees regardless of partitioning scheme or removal of poorly aligned regions (CUgb, CUnogb, CPgb, CPnogb) recovered Mammea as sister to the remaining Calophyllaceae (Fig. 5-IIIC). All MC species trees (82gb, 82nogb, 2050gb, 2050nogb) recovered Mammea as sister to the clade Mesua + Calophyllum, and this clade as sister to a clade including the remaining Calophyllaceae (Fig. 5-IIID). In all ML and BI analyses, support values for this relationship were moderate to high whereas they were low in MC analyses. Relationship 3 will not be discussed since few samples from the genus Hypericum L. were included here (3/370 species). PhyParts gene tree discordance analyses using MC trees after removal of poorly aligned regions (82gb, 2050gb) as input showed that the two of three branches with lower support in our analyses are the ones where there is more gene conflict (Fig. 5-I). Particularly for taxa in Calophyllaceae, around 50% of the genes were not informative or resulted in conflicting topologies. The discussion regarding gene tree discordance will be based on MC species trees generated using the datasets with poorly aligned regions removed (82gb and 2050gb).

**Discussion**

Five out of the six newly assembled plastomes (including the almost complete plastome of Calophyllum brasiliense) presented here showed a conserved structure with total size, general organization in a quadripartite structure, IR, LSC and SSC sizes, number of genes and GC content in agreement with the values for an ordinary angiosperm plastome. However, as is the case in several groups, our study focusing on Calophyllaceae revealed that there is variation of plastome structure for the family. Mahurea exstipulata chloroplast had only a single copy of the usually duplicated genes rpl2, rpl23 and ycf2, which moved from the IRs to the LSC, resulting in the contraction of former and the expansion of the latter and in distinct IR/single-copy regions junctions. The expansion or contraction of IRs is a common type of rearrangement in plastomes that was already documented for the clusioids in Podostemaceae and for some other groups like Geraniaceae, Acanthaceae and the genus Passiflora. However, this phenomenon was not observed in Calophyllaceae until the present study. Although
Podostemaceae species and *M. exstipulata* share IR contractions, their plastomes are quite distinct and these contractions represent independent evolutionary events. In *M. exstipulata* the IR contraction resulted from the loss of one copy of the genes *rpl2, rpl23* and *ycf2* and it was associated with a large expansion of the LSC; it is between 9,258 and 19,040 bp longer when compared to the other clusioids. The loss of a copy of the same three genes is a rare event and has been reported only in *Strobilanthes cusia* (Nees) Kuntze (Acanthaceae, Lamiales)53, and the loss of *rpl2* and *rpl23* is known from *Cuscuta reflexa* Roxb. (Convolvulaceae, Solanales) 54 and *Lonicera japonica* Thunb. (Caprifoliaceae, Dipsacales)55 On the other hand, in Podostemaceae the IR contraction resulted from both gene (*ycf1* and *ycf2*) and intron (*rps16*) loss and pseudogenization of *accD, rpl22,* and *clpP*, all associated with a reduction also in the single copy regions, which resulted in a smaller chloroplast as a whole. Among the newly sequenced plastomes, all species but *M. exstipulata* had the junctions between IR and LSC and SSC regions similar to those of a canonical angiosperm plastome56. The differences observed in the limits of the four main parts of the plastome correspond to one of the sources of variation in the plastome structure and have been described from different taxonomic levels, e.g., within Calophyllaceae and between the clusioids or within Ceridoidae legumes56. The usually slower evolutionary rates of genes IRs when compared to single copy regions is widely reported in the literature14,57 and was also confirmed in our study. Regarding genes and intergenic spacers, the latter were the most variable and should be considered in future evolutionary studies. In agreement with Walker and collaborators (2019)25, the *rpoC2* was the gene with more segregating sites.

Regarding other structural differences noted in the clusioid chloroplasts, the *trnK-UUU–rbcL* inversion was first noted for Podostemaceae by Bedoya et al. (2019)39 and Jin et al. (2020)40. Although Jin et al. (2020)40 also reported this inversion for *Cratoxylum cochinchinense* (Hypericaceae) and suggested it could be a synapomorphy of the clade composed by these two families, it was observed only in Podostemaceae in our study. We also registered the first inversion within Calophyllaceae (*psbJ–psbE*) in *Kielmeyera coriacea*. Repetitive regions are known to flank inversion breakpoints. In our data, it was observed only in Podostemaceae (Supplementary Table 1). The *ycf3* intron loss has not been reported for any other angiosperm except the clusioids to our knowledge and could be a synapomorphy of the clade but further investigation of the chloroplast structure of the closely related families is necessary to confirm this hypothesis. We speculated that the intron 2 was lost from *ycf3*, since in Petersen et al. (2011)38 the loss of intron 1 from that gene in tobacco resulted in phenotypically mutant plants and prevented the splicing of intron 2.

Phylogenetic relationships within the clusioid families are frequently reported as challenging and historically they have changed quite a lot. Savolainen et al. (2000)39, Chase et al. (2002)40 and Gustafsson et al. (2002)35 all used *rbcL* to infer relationships and recovered Podostemaceae (P) sister to Hypericaceae (H) and Calophyllaceae (Ca) sister to Clusiaceae s.s. (Cl) with Bonnetiaceae (B) as sister to the Podostemaceae + Hypericaceae clade (Savolainen et al. 2000, Chase et al. 2002)39,40 or as sister to a clade including all the other four families.
Figure 5. Relationships within the clusioid families and summary of conflicts. On the top, (I) Gene tree discordance within the clusioids represented in the coalescent-based species tree. Pie charts summarize the proportion of conflicting (red and green), concordant (blue) and non-informative (gray) genes for each branch. Numbers above branches indicate concordant genes at that node, and below conflicting genes. On the right the tree generated with one tree for each of the 82 genes (82), and on the left the tree generated with 25 independent replicates for each of the 82 genes (2050), both after removal of poorly aligned sequences (gb). On the bottom, (II) Summary of the main conflicting relationships recovered using maximum-likelihood (values in blue), Bayesian inference (values in red), and multispecies coalescent (values in black) and eight different datasets (CU: 82 protein-coding genes concatenated unpartitioned, CP: 82 protein-coding genes partitioned with individual evolutionary models, 82: a single consensus gene tree per locus used as input, 2050: 25 consensus gene trees from independent runs per locus used as input; each of these four datasets have two versions: one without removal of poorly aligned regions–nogb, and one after removal using Gblocks 50–gb). On the top of each tree there are support values for the branch highlighted in red. The two upper trees (A and B) represent alternative topologies for the position of Bonnetiaceae and Clusiaceae families. The two bottom trees (C and D) represent alternative topologies for the position of Mammea americana (Calophyllaceae). Support values are represented by ultrafast bootstrap (UB) or posterior probability (PP), and the respective dataset (CUgb, CUnogb, CPgb, CPnogb, 82gb, 82nogb, 2050gb, 2050nogb) are in parenthesis. A. Bonnetiaceae sister to Clusiaceae: 76% (UB, CUgb)/58% (UB, CPnogb)/66% (UB, CPgb)/0.76 (PP, CUgb)/0.58 (PP, CPnogb)/0.66 (PP, CPgb). B. Bonnetiaceae sister to the remaining clusioid families: 50% (UB, CUgb)/0.37 (UB, CUnogb).
proposed by Engler64 and mainly supported by the lack of the schizogenic latex or resin ducts in Bonnetiaceae. 82 gene trees (Fig. 5-I). Interestingly, a position of Bonnetiaceae as sister to the rest of the clusioid clade was present study. Ruhfel et al. (2011, 2016)35,41, recovered spp. (Hypericeae)). Within Calophyllaceae, only representatives from the tribe Calophylleae were included in the and gene flow. Therefore, for a better comprehension of phylogenetic relationships, more analytical refinement et al. (2016)41, with . Within Calophyllaceae, only representatives from the tribe Calophylleae were included in the sampling of the clade so far (194 species included), recovered the same topology of Wurdack and Davis (2009)62 as did Xi et al. (2012)33 using genes from the three genomic compartments and Jin et al. (2020)66 using 82 protein-coding plastid genes. More recently, Cai et al. (2021)34 recovered a distinct topology using a 423 nuclear gene matrix: ((Ca + B) + (H + P) + CI)-Podostemaceae were not included, however, there is much incongruence between gene trees and the species tree. Our results mostly agree with the topology from Wurdack and Davis (2009)62, but also reveal much gene tree disagreement, particularly for two relationships: (1) Clusiaceae + Bonnetiaceae that is supported by only 340 of the total 2050 gene trees and 18 of the total 82 gene trees, and (2) Calophyllaceae + (Hypericaceae + Podostemaceae) is supported by 500 of the total 2050 gene trees and 21 of the total 82 gene trees (Fig. 5-I). Interestingly, a position of Bonnetiaceae as sister to the rest of the clusioid clade was proposed by Engler64 and mainly supported by the lack of the schizogenic latex or resin ducts in Bonnetiaceae. However, this position was not recovered in any recent molecular phylogeny until the current study and in Baker et al.65, but in the latter Podostemaceae was not with the other clusioids.

Relationships within Bonnetiaceae, Clusiaceae, and Podostemaceae are largely in agreement with other studies30, Clusia and Garcinia L. being the only genera with more than one representative and recovered as monophyletic in all our analyses. Within Hypericaceae, our results recovered the same relationship as found by Ruhfel et al. (2016)34, with Cratoxylum cochinchinense (Cratoxyleae) + (Visnia ferruginea Kunth. (Visniaeae) + Hypericum spp. (Hypericaceae)). Within Calophyllaceae, only representatives from the tribe Calophylleae were included in the present study. Ruhfel et al. (2011, 2016)34,41, recovered Mammea in a clade with Kayea and Poeclionereae alone, and this clade as sister to a large clade including Calophyllum and Mesua. Recently, Cabral et al. (2021)62 evaluated relationships between Calophylleae genera. The authors recovered Mammea in a clade with Poeclionereae alone, and this clade as sister to a large clade including the genera Calophyllum, Caraipa, Chusia, Haploclathra, Kielmeyera, Marila, Mahurea, and Mesua, although with low support. Our analyses recovered both positions for Mammea. Mammea was recovered as sister to the clade Calophyllum + Mesua, as in Ruhfel et al. (2011, 2016)34,41, in all four coalescent schema, i.e., 82gb, 82nogb, 2050gb and 2050nogb, but all with low support (PP 0.37, 0.44, 0.76 and 0.22, respectively) (Fig. 5-II). The high level of uninformative gene trees (uninformative gene trees/total number of gene trees: 1195/2050, 13/82) associated with a considerable number of discordant trees (discordant trees/total number of trees: 677/2050, 60/82) may help to explain the low support observed (Fig. 5-I). In all four ML and BI concatenated analyses (CUgb, CUfgb, CPgb) Mammea was recovered as sister to a large clade including the other genera, as in Cabral et al. (2021)42, with high support (100% UB/1.0 PP), and the relationship between the clade (Mesua + Calophyllum) and the clade with the remaining genera was moderately to strongly supported (UB/PP: CUfgb: 100%/1.0, CUfgb: 100%/1.0, CPfgb: 87%/0.92) (Fig. 5-III).

The conflicts reported here show that partitioning the dataset, filtering for poorly aligned regions and the method of inference can impact the topology of the tree. The use of MSC, considering each plastid gene individually, was recently recommended to infer phylogenies with plastid data25; this approach breaks some assumptions of the model (see25 for a discussion). However, despite scarce, there is evidence of recombination in chloroplasts (see25,88), raising a question that deserves further exploration. Thus, it would be interesting to see studies tested for non-recombinant units in the plastid prior to phylogenetic inference. There is a lot to be learned about organellar genomes, including a deeper investigation of its multibranch structure.

Plastid data can be a good starting point to investigate phylogenetic relationships, particularly in large and neglected genera such as Clusia and Calophyllum, with around 300–400 and 200 species, respectively. Whenever possible, it should be combined with data from other genomic compartments. Obtaining genomic DNA from herbarium specimens to target nuclear regions is still challenging, although it may become more common with the development of probe sets that can be used in different groups and of standardized protocols46. In this scenario, chloroplast DNA data availability will also increase since this information is being recovered with nuclear DNA in target enrichment capture due to the presence of plastids in high copy number in a cell. Furthermore, an initial exploration may help to point out groups where more sampling is needed or where relationships will need a different source of data to be clarified.

The disagreements found in this study have been reported since the first molecular phylogenies for the clusioids appeared. We now need to add data from other genomic compartments (i.e., nucleus and mitochondria). However, what Cai et al. (2021)34 have shown with nuclear data is that reconstructing phylogenies in Malpighiales, to which the clusioids belong, represents a huge challenge due to incomplete lineage sorting, tree error estimation and gene flow. Therefore, for a better comprehension of phylogenetic relationships, more analytical refinement is required, such as testing different partitioned schema.

**Methods**

**Taxon Sampling, DNA extraction and sequencing.** Silica dried leaves of Kielmeyera appariciana, K. coriacea, Caraipa heterocarpa and Mahurea exstipulata were sampled during the development of R. J. Trad and F. N. Cabral theses and vouchers have been deposited in the University of Campinas herbarium (UEC). All the plants included in the study were collected with ICMBio fieldwork permits (n. 23954–4 and 41896–3 to R.J.T. and n. 33308 to E.N.C.), and INPA permit to collect in the Ducke Reserve for E.N.C. Collections and experimental research on the plants present in the study complied with international, national, and/or institutional guidelines. Kielmeyera specimens (Trad 192—K. appariciana and Trad 401—K. coriacea) were identified by Rafaela.
Chlorobox (https://chlorobox.mpimp-golm.mpg.de/). Potential pseudogenes were defined by Blast following v. 2.3.2 in Geneious 9.12 to check for structural differences such as inversions or rearrangements. IR boundaries false rearrangements. A total of 12 plastomes were aligned using progressive Mauve algorithm in Mauve Plugin quantified on NanoDrop Spectrophotometer (Thermo Scientific, Waltham, MA, USA) or on Qubit 2.0 (Invitrogen, Carlsbad, CA, USA). Genomic libraries for whole-genome were prepared and sequenced 2 × 150 bp on Illumina NextSeq 500 Mid-Output by Genohub (Austin, TX, USA). All adapters were removed by the company.

**Plastome structure and identity.** The gene content was compared between the five Calophyllaceae and the Clusiaceae newly assembled plastomes (Table 2). For general plastome structure and all phylogenetic analyses, all the annotated and checked plastomes available for the clusioid clade on GenBank were included: Bonnetia paniculata Spruce ex. Benth. (Bonnetiaceae), Messua ferrea L. (Calophyllaceae) Garcinia gummi-gutta (Clusiaceae), Cratoxylum cochinichense (Hypericaceae), Marathrum foeniculaceum (Podostemaceae) and Tristique trifaria (Podostemaceae). All the GenBank accession numbers are found in Table 1. We manually adjusted the beginning of each plastome on Geneious 9.12. Beginning in the same region is important to avoid Mauve infer false rearrangements. A total of 12 plastomes were aligned using progressive Mauve algorithm in Mauve Plugin v. 2.3.2 in Geneious 9.12 to check for structural differences such as inversions or rearrangements. IR boundaries were evaluated on IRScope online.77

**Repetitive regions and polymorphism.** Since reorganizations can be associated with small dispersed repeats, REPutter44 was used to identify direct, complement, palindromic, and reverse repeats with the following parameters: minimal size of 30 pb and Hamming distance of 3. For Calophyllum brasiiliense and for Kielmeyera coriacea the Ns and IUPAC ambiguous bases had to be manually removed from the respective fasta file in order to run REPutter.44 And MISA48,49 was used to identify simple sequence repeats (SSR) with a minimum number of 7, 4, 4, 3, 3, and 3, for mono-, di-, tri-, tetra-, penta- and hexanucleotide repeats, respectively. The parameters followed Silva et al. (2019)78.

To evaluate sequence polymorphism, protein-coding and intergenic regions were extracted with parseGenbank.pl script from MitoY47 or on Geneious 9.12 and aligned on MAFFT v. 7.30880 with the ‘adjustdirection’ option added to the default parameters. Only intergenic regions longer than 50 bp were included and Azara serrata Ruiz & Pav. (Salicaceae) was included in the alignments as an outgroup. Alignments were used as input on DnaSP 641 to calculate variation between regions. We compared total number of sites, of analyzed positions (NetSites), of segregating sites, of conserved sites, the estimated number of mutations, parsimony informative sites (PIS), proportion of PIS, nucleotide diversity and average number of substitutions per site. The proportion of PIS was calculated as the PIS/NetSites × 100 and gives an estimate about the absolute informativeness of the region. This information, when combined with the length of the region, may be helpful for markers design.

**Phylogenetic analysis.** For the phylogenetic analyses we assembled a dataset with the 82-protein-coding genes that includes two Bonnetiaceae species (two genera), seven Calophyllaceae species (six genera), four Clusiaceae species (two genera), five Hypericaceae species (three genera) and three Podostemaceae species (three genera). Additionally, data for other 36 species within Malpighiales, Averrhoa carambola L. (Oxalidales) and Elaeocodendron orientale Jacq. (Celastrales) were also included as outgroups. A list with all the GenBank accession numbers for additional species is included in the Supplementary Material S1. To assure poorly aligned regions were not interfering in tree topology, these regions were removed with Gblocks50 with the following parameters: -t = d -b5 = a -n = y -e = gb1 -d = y. Therefore, the following analyses were conducted both for the original dataset (nogb) and for the dataset after gblocks (gb). Maximum likelihood inference was conducted based on the concatenated approach for the unpartitioned (CUngb and CUgb) and partitioned datasets (CPnogb and CPgb) in IQ-tree 2.2.5. Models were selected based on Bayesian information criterion (BIC) and support was assessed.
through 1000 ultrafast bootstrap resampling. For both the nogb and gb alignments, a total of 25 independent runs were performed for each gene, totaling 2050 consensus gene trees which were included in a single file to generate the 2050 datasets, i.e., 2050nogb and 2050gb. Bayesian Inference for the unpartitioned (CUngenb and CUGgb) and partitioned datasets (CUnggb and CPGgb) was conducted in MrBayes v.3.2.7 with the best-fit model in accordance with BIC assessed in IQ-tree26,83. It was run for 30 million generations sampled every 1000 generations, using two runs and four chains, until the average standard deviation of split frequencies became less than 0.01, beginning with random trees. The initial trees were discarded after reaching stationarity (~25%). Also, a coalescent-based species tree estimation was done in Astral-III84. We conducted the analysis with one tree per gene as input (82gb and 82nogb) and with 25 trees per gene as input (2050gb and 2050nogb). All gene trees used as input in Astral-III84 were rooted and had branches with support lower than 10% collapsed in Newick Utilities85 through the functions nw_reroot and nw_ed. Rooting the trees and collapsing branches with low support are known to improve the performance of summary methods86.

To explore tree conflict within the clusoid families the splits.nex file generated in IQ-tree22 for both partitioned and unpartitioned analysis was visualized as a network in SplitsTree4 v. 4.16.287. Additionally, tree conflict was explored through PhyParts88 and pie charts were plotted on the species phylogeny using the PieCharts python script developed by M. Johnson88. PhyParts calculates the number of concordant gene trees, of the top alternative bipartition, of other conflicting topologies, and of uninformative genes for all branches in the tree. To root the trees and remove ultrafast bootstrap and branch length values the files for both PhyParts88 and PieCharts were prepared using Newick Utilities85 through the functions nw_reroot and nw_topo.

Data availability
The complete plastome sequences of Calophyllum brasiliense, Caraipa heterocarpa, Kielmeyera apparciana, K. carioidea, and Matura extispilata have been submitted to GenBank under the accession numbers MW853786MW853790.

Received: 10 April 2021; Accepted: 21 September 2021
Published online: 20 October 2021

References
1. Firetti, E. et al. Complete chloroplast genome sequences contribute to plant species delimitation: a case study of the Anemophaengae species complex. *Am. J. Bot.* 104, 1493–1509 (2017).
2. Thode, V. A. & Lohmann, L. G. Comparative chloroplast genomics at low taxonomic levels: a case study using *Amphilophium* (Bignonieae, Bignonieae). *Front. Plant Sci.* 10, 796. https://doi.org/10.3389/fpls.2019.00796 (2019).
3. Liu, B.-B., Campbell, C. S., Hong, D.-Y. & Wen, J. Phylogenetic relationships and chloroplast capture in the *Amphilophium-Mala-comes-Peraphyllum* clade (Maeleae, Rosaceae): Evidence from chloroplast genome and nuclear ribosomal DNA data using genome skimming. *Mol. Phylogenetics Evol.* 147, 106784 (2020).
4. Zhang, D. et al. Plastome sequence helps to resolve deep-level relationships of Populus in the family Salicaceae. *Front. Plant Sci.* 10, 26. https://doi.org/10.3389/fpls.2019.00005 (2019).
5. Ruful, B. R., Gitzendanner, M. A., Solis, P. S., Solis, D. E. & Burleigh, J. G. From algae to angiosperms—inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. *BMC Evol. Biol.* 14, 23. https://doi.org/10.1186/s12862-014-0023-4 (2014).
6. Marinho, L. C. et al. Plastomes resolve generic limits within tribe Clusiae (Clusiaceae) and reveal the new genus Arakawia. *Mol. Phylogen. Evol.* 134, 142–151 (2019).
7. Ravi, V., Khurana, J. P., Tyagi, A. K. & Khurana, P. An update on chloroplast genomes. *Plant Syst. Evol.* 271, 101–122 (2008).
8. Daniell, H., Lin, C.-S., Yu, M. & Chang, W.-J. Chloroplast genome diversity, evolution, and applications in genetic engineering. *Genome Biol.* 17, 134. https://doi.org/10.1186/s13059-016-1004-2 (2016).
9. Kellogg, E. A. & Bennett, J. L. The evolution of nuclear genome structure in seed plants. *Am. J. Bot.* 91, 1709–1725 (2004).
10. Dierckxsen, N., Mardulp, Y. & Smits, G. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Res.* 45, e18 (2017).
11. Jin, K. et al. GetOrganell: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biol.* 21, 241. https://doi.org/10.1186/s41598-020-66024-7 (2020).
12. Kearse, M. et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequenced data. *Bioinformatics* 28, 1647–1649 (2012).
13. Hahn, C., Bachmann, L. & Chevreux, B. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. *Nucleic Acids Res.* 41, e129. https://doi.org/10.1093/nar/gkt377 (2013).
14. Mower, J. P. & Vickrey, T. L. Structural diversity among plastid genomes of land plants in *Advances in Botanical Research* (ed. Chau, S. M. J. R.) 263–297 (Academic Press, 2018).
15. Dugas, D. V. et al. Mimosoid legume plastome evolution: IR expansion, tandem repeat expansions, and accelerated rate of evolution in clpP. *Syst. Biol.* 58, 16958. https://doi.org/10.1093/sysbio/syv16958 (2015).
16. Baker, W. J. et al. Exploring Angiosperm353: an open, community toolkit for collaborative phylogenomic research on flowering plants. *Am. J. Bot.* 108, 1059–1065 (2021).
17. Wicke, S., Schneebeck, G. M., dePamphilis, C. W., Müller, K. F. & Quandt, D. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. *Plant Mol. Biol.* 76, 273–297 (2011).
18. Bendich, A. J. Circular chloroplast chromosomes: the grand illusion. *Plant Cell* 16, 1661–1666 (2004).
19. Oldenburg, D. J. & Bendich, A. J. Most chloroplast DNA of maize seedlings in linear molecules with defined ends and branched forms. *J. Mol. Biol.* 335, 953–970 (2004).
20. Erickson, P. & Oxelman, B. Reticulate or tree-like chloroplast DNA evolution in Sileneae (Caryophyllaceae)? *Mol. Phylogen. Evol.* 48, 313–325 (2008).
21. Vogl, C. et al. Probabilistic analysis indicates discordant gene trees in chloroplast evolution. *J. Mol. Evol.* 56, 330–340 (2003).
22. Gonçalves, D. J. P., Simpson, B. R., Ortiz, E. M., Shimizu, G. H. & Jansen, R. K. Incongruence between species tree and gene trees and phylogenetic signal variation in plastid genes. *Mol. Phylogen. Evol.* 138, 219–232 (2019).
23. Gonçalves, D. J. P., Jansen, R. K., Ruiz-ru, T. A. & Mandel, J. R. Under the rug: abandoning persistent misconceptions that obfuscate organelle evolution. *Mol. Phylogen. Evol.* 151, 106903. https://doi.org/10.1016/j.ympev.2020.106903 (2020).
24. Doyle, J. D. Defining coalescent genes: theory meets practice in organelle phylogenomics. *Syst. Biol.* 6, 1–14 (2021).
25. Walker, J. E., Walker-Hale, N., Vargas, O. M., Larson, D. A. & Stull, G. W. Characterizing gene tree conflict in plastome-inferred phylogenies. *PeerJ* 7, e7747. https://doi.org/10.7717/peerj.7747 (2019).
67. Doyle, J. J. & Doyle, J. L. A rapid DNA isolation procedure from small quantities of fresh leaf tissue. *Phytochem. Bull.* 19, 11–15 (1997).
68. Caddah, M. K. Estudos Taxonômicos sobre o complexo Kielmeyera coriacea Mart. & Zucc. (Clusiaceae). MSc Dissertation, University of Campinas, SP, Brazil (2009).
69. Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: interactive visualisation of de novo genome assemblies. *Bioinformatics* 31, 3350–3352 (2015).
70. Langmead, B. & Salzberg, S. Fast gapped-read alignment with bowtie 2. *Nat. Methods* 9, 357–359 (2012).
71. Kajitani, R. et al. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads”. *Genome Res.* 24, 1384–1395 (2014).
72. Walker, J. F., Jansen, R. K., Zanis, M. J. & Emery, N. C. Sources of inversion variation in the small single copy (SSC) region of chloroplast genomes. *Am. J. Bot.* 102, 1751–1752 (2015).
73. Tillich, M. et al. GeSeq-verbatim and accurate annotation of organelle genomes. *Nucleic Acids Res.* 45, W6–W11. https://doi.org/10.1093/nar/gkx391 (2017).
74. Laslett, D. & Canback, B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.* 32, 11–16 (2019).
75. Silva, S. R. et al. Comparative genomic analysis of *Genlisea* (corkscrew plants–Lentibulariaceae) chloroplast genomes reveals an increasing loss of the ndh genes. *PLoS ONE* 13, e0190321. https://doi.org/10.1371/journal.pone.0190321 (2018).
76. Lobse, M., Drechsel, O., Kahlau, S. & Bock, R. OrganellarGenomeDRAW—a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Res.* 41, W575–W581. https://doi.org/10.1093/nar/gkt289 (2013).
77. Amiryousefi, A., Hyvönen, J. & Poczai, P. IRscope: an online program to visualize the junction sites of chloroplast genomes. *Bioinformatics* 34, 3030–3031 (2018).
78. Silva, S. R. et al. Intraspécific Variation within the *Utricularia amethystina* species morphotypes based on chloroplast genomes. *Int. J. Mol. Sci.* 20, 6130. https://doi.org/10.3390/ijms20246130 (2019).
79. Alverson, et al. Insights into the evolution of mitochondrial genome size from complete sequences of *Citrus lanatus* and *Cacarbita pepo* (Cucurbitaceae). *Mol. Biol. Evol.* 27, 1436–1448 (2010).
80. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780 (2013).
81. Rozas, J. et al. DnaSP 6: DNA sequence polymorphism analysis of large datasets. *Mol. Biol. Evol.* 34, 3299–3302 (2017).
82. Minh, B. Q. et al. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* 37, 1530–1534 (2020).
83. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587–589 (2017).
84. Zhang, C., Rabiee, M., Sayyari, E. & Mirarab, S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinf.* 19, 153 (2018).
85. Junier, T. & Zdobnov, E. M. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. *Bioinformatics* 26, 1669–1670 (2010).
86. Mirarab, S., Bayzid, M. S. & Warnow, T. Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. *Syst. Biol.* 65, 366–380 (2016).
87. Huson, D. H. & Bryant, D. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* 23, 254–267 (2006).
88. Johnson, M. G. PhyParts PieCharts python script. GitHub: https://github.com/mossmann/MITPythonNotebooks/blob/master/phypartspiecharts.py (2019).

**Acknowledgements**

We would like to thank Prof. Dr. Peter F. Stevens for the important discussions and suggestions on the manuscript. Profa. Dra. Anete Pereira de Souza for allowing us to use the LAGM (CRMEG – Unicamp) cluster, the Fundação de Apoio à Pesquisa do Estado de São Paulo (Auxílio à Pesquisa FAPESP 2012/51781-0 to M.C.E.A., R.J.T. and F.N.C., and a Doctoral Scholarship to R.J.T. n. 1124/13, and the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES – code 001) for a Doctoral Scholarship to R.J.T. and F.N.C., and a Doctoral Internship Fellowship to R.J.T. n. 205856/2014-4), the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq: PROTAX n. 562299/2010-6 and Universal Research Grant n. 482988/2013-4 to M.C.E.A., and a Potsdoctoral Fellowship 2018/02285-6 to S.R.S.), the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES – code 001) for a Doctoral Scholarship to R.J.T.

**Author contributions**

R.J.T., M.C.E.A and V.B. designed the study, R.J.T. and F.N.C. collected the data. R.J.T. and S.R.S. conceived and conducted the analyses. R.J.T., S.R.S. and V.B. discussed the results. R.J.T. wrote the first version of the manuscript with input from all authors. All authors critically reviewed the manuscript.

**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/s41598-021-99178-z.

**Correspondence** and requests for materials should be addressed to R.J.T.

**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.
