

De novo assembled nuclear, chloroplast, and mitochondrial genomes show high intraspecific variation in the tropical rainforest species *Symphonia globulifera*

Sanna Olsson ^{1,*} Rocío Bautista,² M. Gonzalo Claros ^{3,4} Myriam Heuertz ⁵ Ivan Scotti ⁶

¹Instituto de Ciencias Forestales (ICIFOR-INIA), Consejo Superior de Investigaciones Científicas (CSIC), Madrid E-28040, Spain

²Plataforma Andaluza de Bioinformática, Universidad de Málaga, Málaga E-29590, Spain

³Instituto de Hortofruticultura Subtropical y Mediterránea (IHSM-UMA-CSIC), Málaga E-29010, Spain

⁴Department of Molecular Biology and Biochemistry, Universidad de Málaga, Málaga E-29071, Spain

⁵University. Bordeaux, INRAE, BIOGECO, Cestas F-33610, France

⁶INRAE, UR629 URFM, Ecologie des Forêts Méditerranéennes, Avignon Cedex 9 F-84914, France

*Corresponding author: Instituto de Ciencias Forestales (ICIFOR-INIA), Consejo Superior de Investigaciones Científicas (CSIC), Crta. La Coruña km 7.5, Madrid E-28040, Spain.

Email: sanna.olsson@inia.csic.es

Symphonia globulifera (Clusiaceae) has emerged as a model organism in tropical forest ecology and evolution due to its significant ecological role and complex biogeographical history. Originating from Africa, this species has independently colonized Caribbean, Central, and South America three times, becoming a key component of tropical ecosystems across these regions. Despite the ecological importance of *S. globulifera* and other tropical tree species, our understanding of their genomic architecture remains limited compared to temperate species. To bridge this gap, we present a comparative analysis of two de novo assembled nuclear genomes of *S. globulifera*—one from a South American individual and one from an African individual—and report newly assembled chloroplast and mitochondrial genomes. Initial assembly of the organelles was performed using GetOrganelle, and the results were compared with corresponding publicly available sequences from closely related *Garcinia* species. Our study introduces novel genomic resources, including an annotated nuclear draft genome based on Illumina short reads, the first chloroplast genome assembly for the genus, and a set of assembled mitochondrial gene sequences. Additionally, we provide a set of single-copy nuclear gene alignments identified by BUSCO as well as manually curated coding plastid genes, which will serve as valuable tools for future comparative analyses and phylogenetic studies. Our preliminary results based on chloroplast genes and limited sampling suggest that *Garcinia* might be nonmonophyletic. The detected differences in nuclear and organellar genomes reveal high intraspecific variation, emphasizing the importance of genome-wide sampling for understanding tropical tree evolution.

Keywords: tropical trees; genome assembly; genomic diversity; organelles; markers; phylogeny

Introduction

With hyperdiverse tropical ecosystems facing important threats to their conservation (Barlow et al. 2018; Ferrer-Paris et al. 2019) and eco-evolutionary processes, e.g., speciation and evolutionary change, being highly variable in tropical trees depending on population sizes and spatiotemporal environmental turnover (Allié et al. 2015; Pennington and Lavin 2016), it is important to study them in more detail. However, our understanding of eco-evolutionary processes in tropical tree species lags far behind that of temperate species (Holliday et al. 2017; De Kort et al. 2021). This is particularly true for the availability of genomic resources, essential for conservation genetics studies (Formenti et al. 2022; Heuertz et al. 2023; Theissinger et al. 2023).

Plant genome assemblies typically focus on the nuclear genome, which encodes the vast majority of functional genes (Heslop-Harrison and Schwarzacher 2011). However, chloroplast and mitochondrial genomes also play essential roles in

plant physiology and evolution. Chloroplasts are central to photosynthesis and are widely used in plant phylogenetics due to their relatively conserved structure and uniparental inheritance (Daniell et al. 2016). Mitochondria are involved in respiration and cellular energy production, and mitochondrial genes can provide complementary phylogenetic signals and insights into cytoplasmic inheritance (O’Conner and Li 2020). Compared to nuclear genomes, organellar genomes differ markedly in size, structure, and copy number: nuclear genomes are large and complex (typically hundreds of Mb to several Gb); chloroplast genomes are small (~120 to 160 kb), highly conserved, and present in hundreds to thousands of copies per cell; and mitochondrial genomes are highly variable in size and structure (~200 to 2,400 kb in plants) and often fragmented or rearranged (Sloan et al. 2012; Daniell 2016). Mitochondrial genomes are more difficult to assemble than chloroplasts due to their bigger size and high structural variability (Sloan 2013). These differences present specific

challenges and opportunities for genome assembly and comparative genomics.

Symphonia globulifera (Clusiaceae) has become a model organism in tropical forest ecology and evolution, due to its ecological importance and intriguing distribution pattern (Dick and Heuertz 2008; Torroba-Balmori et al. 2017; Schmitt et al. 2021). With an origin in Africa, the species has spread to the Caribbean, Central, and South America three independent times (Dick et al. 2003). It occurs in a variety of environments, and as a highly abundant species, it plays a pivotal role in the maintenance of forest ecosystems and serves as a resource for numerous human activities, including timber and traditional medicinal use (Oyen 2005; ter Steege et al. 2006).

Previous studies that used SNP or SSR markers in target regions have revealed high levels of intraspecific genetic diversity in *S. globulifera* (Dick and Heuertz 2008; Torroba-Balmori et al. 2017; Schmitt et al. 2021). However, genome-scale comparisons across continents are still lacking. While a draft nuclear genome was previously assembled from an African individual (Olsson et al. 2017), genomes from other parts of its range are missing. Similarly, although chloroplast or mitochondrial genomes have been assembled for several *Garcinia* species (Wee et al. 2023), no chloroplast or mitochondrial genomes are available for *Symphonia* or other Clusiaceae outside *Garcinia*. Chloroplast genomes available from the most closely related families within the same order, Malpighiales, are in Podostemaceae (Bedoya et al. 2019) and in *Hypericum perforatum* (Liu et al. 2023). The only available mitochondrial genome from Clusiaceae is that of *Garcinia mangostana* (Wee et al. 2022). The lack of these genomic resources limits our ability to understand both the evolutionary relationships within the family and the extent of genomic differentiation within the species.

The goals of our study are (i) to assemble and annotate the nuclear, chloroplast, and mitochondrial genomes of a South American individual of *S. globulifera*; (ii) to compare these assemblies with existing genomic resources from an African individual and closely related species; and (iii) to assess intraspecific genetic variation in the species. By providing novel, much-needed genomic resources for this tropical species, including an annotated nuclear draft genome, the first chloroplast assembly for the genus, and a set of single-copy nuclear genes and manually curated coding plastid gene sets, our study delivers resources for tropical plant genomics, useful for comparative analyses and phylogenies. The observed differences in gene content and sequence diversity between African and South American individuals highlight the substantial intraspecific variation present in the species and underscore the need for multiple reference genomes per species to represent the genetic complexity in tropical trees.

Material and methods

Sampling and genomic DNA sequencing

Leaves from the same *S. globulifera* individual that was used for transcriptomic sequencing by Brousseau et al. (2014) were collected in 2014. Whole genomic DNA was extracted from them using a CTAB protocol as described by Le Provost et al. (2003). Preparation of Illumina TruSeq DNA libraries and HiSeq2000 sequencing of the paired-end 100-bp libraries on the Illumina 1.8 sequencing platform was carried out by IGA Technology Services (Institute of Applied Genomics, Udine, Italy) in the same year. Qubit 2.0 Fluorometer (Invitrogen, Carlsbad, CA) was used for DNA and final library quantification. Quality of DNA was tested by Agilent 2100 Bioanalyzer High Sensitivity DNA assay (Agilent Technologies, Santa Clara, CA) prior to sequencing. Four different

libraries were prepared, and two of them were repeated due to low yield of reads.

Preprocessing

Quality of the reads was checked with FastQC before and after trimming. Trimmomatic v0.36 (Bolger et al. 2014) was used to quality filter and remove adapter sequences with the following parameters: ILLUMINACLIP:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:50. Exact duplicates were removed with the script `remove_dup_PE.py` (available at <https://github.com/Zhong-Lab-UCSD/MARIO>) before further analyses.

Removal of chloroplast and mitochondrial reads was performed by mapping with `bwa-mem` against Clusiaceae organelle genomes. Only the unmapped read-pairs containing both reads were kept for nuclear genomic work.

Chloroplast genome assembly of African individual

The reads from Olsson et al. (2017) were downloaded from <https://datadryad.org/stash/dataset/doi:10.5061/dryad.78ng1>. The recovery of chloroplast reads, both paired-end and mate-pair, was performed by SeqTrimNext (Falgueras et al. 2010). Using the chloroplast organelle genomes database of NCBI, the assembly was carried out using the A5 pipeline (Coil et al. 2015) and the SOAPdenovo (Xie et al. 2014) assembly. These protocols were selected because only long mate-pair reads were available for this individual, which limited the use of organelle-specific short-read assemblers such as GetOrganelle.

In the first method, the A5 pipeline used both paired and mate-pair reads as input with default parameters. This pipeline consists of five steps: read cleaning by Trimmomatic and error correction with SGA's k-mer module; assembly of contigs and scaffolds with the IDBA-UD algorithm; misassembly correction via read mapping; and a final round of stringent scaffolding.

The second method, SOAPdenovo, assembled fragments and paired-end sequencing from the libraries with variable insert sizes. Read overlaps were represented using de Bruijn graphs, with erroneous connections removed to generate the final graph. The results from the A5 pipeline were selected for downstream analyses due to longer scaffolds and higher N50 value.

The final scaffold derived from the A5 pipeline was mapped against the *Jatropha curcas* chloroplast genome (European Nucleotide Archive, ENA, accession number FJ695500.1, by Asif et al. (2010); see also Wee et al. (2023) using Bowtie2 v2.1 (Langmead et al. 2019) with default parameters. Mapped scaffolds were ordered and merged into one pseudochloroplast genome using Mauve (Darling et al. 2004). The assembly was annotated using GeSeq (Tillich et al. 2017, available at <https://chlorobox.mpimp-golm.mpg.de/geseq.html>; accessed on 18 August 2023). Four *Garcinia* species: *Garcinia gummi-gutta*, *G. mangostana*, *Garcinia oblongifolia*, and *Garcinia pedunculata* (see Table 1 for accession numbers) were used as references for the BLAST-like Alignment Tool (BLAT; Kent 2002) implemented in GeSeq.

Chloroplast genome assembly and annotation of American individual

GetOrganelle v1.7.7.0 (Jin et al. 2020) was used for chloroplast assembly, using all fastq reads as input. Parameters suggested for low-coverage assembly were used, i.e., -R 35, -k 13,21,33,45,65,85, `embplant_pt`, -w 60, and the chloroplast genome of *G. mangostana* (ENA accession number MZ823408, Wee et al. 2023) as seed. The assembly resulted in six scaffolds, of which five were selected as suggested by the program. The sixth scaffold was verified by

BLAST search not to belong to the assembly before discarding it as contamination. The assembly was visualized with Bandage (Wick et al. 2015). RegCloser (Cao et al. 2023) was used to try to close the gaps in the assembly, unsuccessfully.

Manual alignment of the assembled scaffolds of both *S. globulifera* individuals, along with *G. mangostana* var. Mesta (ENA accession number MZ823408), and four RefSeq *Garcinia* chloroplasts (ENA accession numbers NC_036341, NC_047250, NC_048983 and NC_050384, see also Wee et al. 2023) was performed in PhyDE v1.0 (Müller et al. 2005). This facilitated scaffold orientation, detection of potential chimeras, and estimation of missing data. The manually curated assembly was annotated using GeSeq, in the same manner as for the African individual, with the same four *Garcinia* species as references in the BLAT step. The plastome map was visualized using the Organellar Genome DRAW (OGDRAW v1.3.1) program with default parameters (Greiner et al. 2019). The annotated chloroplast genome was submitted to ENA with accession number OZ172679.1 (Table 1).

Although different assembly protocols were used for the two individuals due to the distinct sequencing technologies (mate-pair long reads vs Illumina short reads), both assemblies were aligned to the same reference sequences, manually curated, and annotated using the same pipeline and references to ensure consistency. This strategy minimized potential biases introduced by methodological differences in downstream comparisons.

Phylogenetic analysis of chloroplast genes

The gene regions (including exons and introns) of the 77 protein-coding genes were extracted from the GeSeq annotated assembly of the American *S. globulifera* chloroplast genome using the function GenBank Feature Extractor from Sequence Manipulation Suite (Stothard 2000). The same genes from the African *S. globulifera* chloroplast genome were retrieved from the manually curated alignment. The corresponding genes of all available *Garcinia* species, *Erythroxylum novogranatense*, and *J. curcas* (see Wee et al. 2023) were downloaded from GenBank (Table 1). The latter two were used as outgroups in the phylogenetic analysis. The individual gene alignments were concatenated with SequenceMatrix v1.8 (Vaidya et al. 2011), defining *J. curcas* as an outgroup. Sequences were aligned with Mafft (Katoh et al. 2002) using alignment option FFT-NS-I. The alignments were edited manually, applying CDS adjustments reported by Wee et al. 2023 (*atpF*, *cemA*, *ndhA*, *ndhD*, *ndhE*, *ndhK*, *petD*, *psbM*, *rpoC1*, *rps12*, *rps16*, *rps19*). Repeats were identified in the noncoding regions of *petD* and *clpP*, and the alignments were adjusted according to the criteria laid out by Kelchner (2000). The *rpl16* intron was deleted in the two outgroup species due to very high variability and uncertain homology assessment. The data matrix was analyzed by maximum likelihood (ML) after automatic model selection using ModelFinder (Kalyaanamoorthy et al. 2017) implemented in IQTree v1.6.12 (Nguyen et al. 2015) applying 1000 ultrafast bootstrap replicates (Hoang et al. 2018). Consensus topology and support values were drawn using TreeGraph2 (Stöver and Müller 2010).

Mitochondrial genome assembly and annotation

The mitochondrial genome of the American individual was assembled with GetOrganelle with default settings using the *G. mangostana* (OM759996) mitochondrial genome as seed. All FASTQ reads were used as input, as for the chloroplast genome assembly. The resulting scaffolds were aligned with the only available Clusiaceae mitochondrial assembly of *G. mangostana* (Wee et al. 2022) to retrieve the protein-coding genes.

Table 1. GenBank accession numbers of the full chloroplast genomes used in the phylogenetic study.

Species	Family	Accession	Size in bp
<i>Jatropha curcas</i>	Euphorbiaceae	NC_012224	163,856
<i>Erythroxylum novogranatense</i>	Erythroxylaceae	NC_030601	163,937
<i>Garcinia anomala</i>	Clusiaceae	MW582313	156,774
<i>Garcinia gummi-gutta</i>	Clusiaceae	NC_047250	156,202
<i>Garcinia mangostana</i> var. Mesta	Clusiaceae	MZ823408	156,580
<i>Garcinia mangostana</i> var. Thailand	Clusiaceae	NC_036341	158,179
<i>Garcinia oblongifolia</i>	Clusiaceae	NC_050384	156,577
<i>Garcinia pedunculata</i>	Clusiaceae	NC_048983	157,688
<i>Garcinia paucinervis</i>	Clusiaceae	MT501656	157,702
<i>Symphonia globulifera</i> (this study)	Clusiaceae	OZ172679	156,363

Nuclear genome characterization

The GenomeScope v2.0 web application (Ranallo-Benavidez et al. 2020, available at <http://genomescope.org/genomescope2.0>) was used to analyze genome characteristics and estimate the proportion of heterozygosity. It was also used to visualize the *k*-mer plot. The genome size was estimated by implementing *k*-mer counting methods. For the purposes of comparison, both Jellyfish v2.3.0 (Marçais and Kingsford 2011) and KMC v3.2.0 (Kokot et al. 2017) were used for *k*-mer count, setting the number of *k*-mers to 21. Max *k*-mer coverage was set in Jellyfish and GenomeScope to 1,000,000 instead of the default 10,000 as recommended to repetitive and highly heterozygous genomes.

Nuclear genome assembly and annotation

Platanus-allee v2.2.2 (Kajitani et al. 2019) was used for the genome assembly as a suitable assembler for highly heterozygous genomes, although optimally recommended to be used with mate-pair libraries and >80% genomic coverage. A metric of gene completeness of the assembly was determined with BUSCO v5.4.7 using both the eudicots_odb10 and embryophyta_odb10 databases (Manni et al. 2021). In addition, quality metrics were obtained with QUAST-LG (Mikheenko et al. 2018) and BUSCO v5.4.7 (Manni et al. 2021).

Functional annotation was done using the TOA (Taxonomy-oriented Annotation) pipeline (Mora-Márquez et al. 2021) in the mode of amino acid annotation. TOA is a transcriptome annotation platform with a focus on plants. Information from genomic databases were used sequentially in this order: Gymno PLAZA 1.0, Dicots PLAZA 4.0, Monocots PLAZA 4.0, NCBI RefSeq Plantmand, and NCBI Nucleotide Database (NT, selecting plant sequences), complemented with information from NCBI Gene, InterPro, and Gene Ontology databases as implemented in the pipeline. Taxonomic annotation with TOA detected 38 amino acids other than plants as first hit organisms. As the corresponding scaffolds were short (length between 47 and 281 bp), separate scaffolds, they were removed from the assembly before evidence-based prediction as likely contaminations.

The assembly was softmasked using the RepeatMasker v2.0.4 (Smit et al. 2013–2015) pipeline using the default NCBI search engine in RepeatModeler (Smit and Hubley 2008–2015). Simple reads were left unmasked, as they are frequently part of genes. The masked genome assembly was submitted to ENA with accession number GCA_965614385.1. Evidence-based gene prediction in

the masked genome assembly was performed using the BRAKER2 pipeline, applying amino acid data as reference.

Intraspecific genetic diversity

To estimate the number of differences between two geographically distant individuals, Transcriptome Ortholog Alignment Sequence Tools (TOAST; [Wcisel et al. 2020](#)) was used to retrieve the cds of orthologous single copy genes identified by BUSCO from both the African and American individual. The complete genes present in both individuals were concatenated with the Perl script seqCat.pl v1.0 ([Bininda-Emonds 2005](#)) into a 448,390 bp long concatenated matrix. The corresponding coding nucleotide sequences were extracted and aligned with MAFFT ([Katoh and Standley 2013](#)) with default settings and trimmed with trimAl v1.4.rev15 ([Capella-Gutiérrez et al. 2009](#)) using the option -strict. Alignment statistics for each orthogroup were obtained with SeqState ([Müller 2005](#)).

Results and discussion

Chloroplast genome assembly of African and American individuals

The de novo assembly of mate-pair (long) reads from the African individual using the A5 pipeline yielded 14 scaffolds. Aligning the scaffolds with the *Garcinia* chloroplast revealed that the assembled contigs did not cover the entire organelle. Therefore, a complete reconstruction of the chloroplast was not attempted. Instead, the coding sequences were retrieved for intraspecific comparison.

In contrast, the assembly based on Illumina reads of the American individual yielded a more complete plastome. De novo assembly using GetOrganelle resulted in five scaffolds ([Supplementary Fig. 13](#)). The average coverage of the chloroplast genome assembly with Illumina reads was 55x. Aligning against *Garcinia* sequences allowed us to reconstruct an almost complete 156,363 bp long circular plastome of *S. globulifera*, including a LSC (85,173 bp) and SSC (17,102 bp) separated by 27,044-bp-long inverted repeat (IR) regions ([Fig. 1](#)). However, six gaps remained in the repeat-rich intergenic regions of the assembly, indicated in the complete chloroplast assembly with 12 Ns (the exact number of Ns being unknown). The total number of missing characters was estimated to be 1,831 bp, based on the alignment with *Garcinia* species.

A total of 128 genes were annotated, of which 77 were unique protein-coding genes with six duplicated genes, 30 unique tRNAs (seven duplicated genes), and 4 rRNAs (four duplicated genes). All duplicated genes were located in the IR region. The chloroplast gene content and order were conserved in comparison with *Garcinia* species ([Wee et al. 2023](#)). Due to the availability of these closely related *Garcinia* chloroplast genomes, the rather modest amount of Illumina reads, together with long-read resources from another individual, allowed us to reconstruct an almost complete chloroplast genome of the species and to analyze intra-specific variation. Photo credit: Benoit Burban.

Phylogenetic relationships of *Symphonia* and *Garcinia* based on chloroplast genes

The alignment consisting of 77 protein-coding chloroplast genes of the two *S. globulifera* and seven *Garcinia* individuals, together with the two outgroup species (*J. curcas* and *E. novogranatense*), was 82,086 bp long. All genes were retrieved from all specimens, except for *rps16*, which was not present in the outgroup species ([Supplementary Table 1](#)). The highest number of indels was

observed in the *ycf3_exon2* region. The phylogenetic reconstruction shows a fully supported *Garcinia*–*Symphonia* clade ([Fig. 2](#)). The internal nodes are mostly fully supported, with two exceptions: neither the position of *Garcinia paucinervis* within the *Garcinia*–*Symphonia* clade (bootstrap support 55) nor the sister clade relationship of *S. globulifera* with *G. mangostana* var. Thailand and *G. pedunculata* (bootstrap support 53) is well supported.

According to phylogenetic studies ([Sweeney 2008](#), based on two nuclear markers; [Gaudeul et al. 2024](#), based on one nuclear and three chloroplast markers), *Garcinia* is sister genus to *Symphonia*. Our analyses suggest that *Garcinia* is potentially paraphyletic as *S. globulifera* forms a clade with *G. pedunculata* and *G. mangostana* var. Thailand. The nonmonophyly of *G. mangostana* lineages based on chloroplast sequences has been described by [Wee et al. \(2023\)](#). However, these results were based on a limited number of species, as is the case with our study. In addition, biological phenomena like incomplete lineage sorting, hybridization, limited marker resolution, or chloroplast capture could also explain why *S. globulifera* groups together with some *Garcinia* species. The recent study by [Gaudeul et al. \(2024\)](#) suggests a broad circumscription of *Garcinia* to include *Allanblackia* and some other previously described genera, but before knowing if *Symphonia* should be included in *Garcinia*, or any other taxonomic changes, a more thorough phylogenetic study is based on a comprehensive taxon selection and, ideally, including also nuclear single-copy genes. The chloroplast genome and manually curated chloroplast gene alignments will be an important resource to study phylogenetic relationships of *Symphonia* and *Garcinia*. The same applies to the nuclear single-copy genes identified in this study. These genes form the basis for designing and testing primers that can be used to amplify them in a larger number of species.

Mitochondrial genome assembly and annotation

The de novo assembly resulted in 14 scaffolds of which 8 were selected. The challenges of mitochondrial assembly are well known (as explained in the Introduction), and the short-read data alone did not yield a reasonable de novo assembly. However, when aligning the eight scaffolds against the *Garcinia* reference, all the same protein-coding genes as in *G. mangostana* were retrieved ([Supplementary Fig. 2](#)), which will be an additional resource for evolutionary studies. A comparison of these genes between the two species shows a very low level of differences ([Supplementary Table 2](#)).

Nuclear genome characterization

The sequencing of 12 libraries yielded a total of a bit more than one billion reads. The genome was estimated to have high heterozygosity (1.62 to 1.69%) and low coverage (×16.6). According to GenomeScope ([Vurture et al. 2017](#)), a heterozygosity rate above 1.5% is considered high. The genome size was estimated by Jellyfish to be between 1.25 and 1.26 Gbp, of which unique sequences 411 to 414 Mbp and read error rate 0.10%. KMC estimated a smaller genome size, 0.95–0.96 Gbp, and higher read error rate (0.14%). The first peak in GenomeScope profile was lower than the second one, which is probably due to low coverage and high heterozygosity ([Supplementary Fig. 3](#)). The profile also showed a lower peak at around 90x coverage, which was interpreted as high coverage organellar reads, as the removal of organellar reads was incomplete.

Nuclear genome assembly and annotation

The assembly had a N50 = 8,971 and yielded 72,765 contigs of which 8,037 were longer than 10,000 bp and 14 longer than

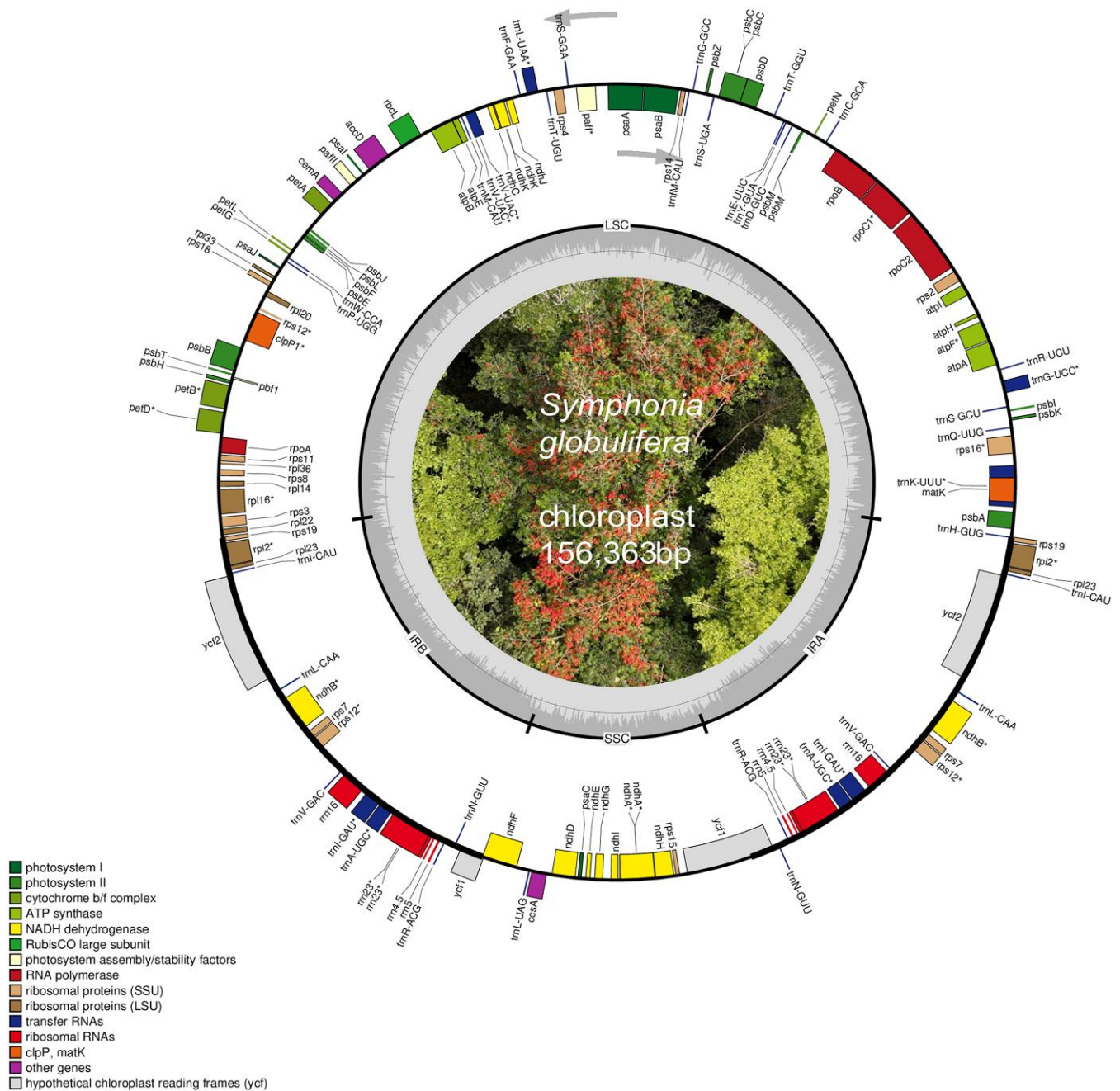


Fig. 1. The circular plastome of *S. globulifera* (American individual). Genes inside the circle are transcribed clockwise, while genes outside the circle are transcribed anti-clockwise, as indicated by the gray arrows. The gray bars inside the circle represent the GC content of the sequence. Asterisks (*) indicate genes containing intron(s).

50,000 bp (Supplementary Table 3). The number of long scaffolds and BUSCO completeness was slightly better than the existing draft genome of the species (Olsson et al. 2017, comparison in Fig. 3), based on the number of complete genes. There were 522 (32.3%) complete (compared to 465 in earlier genome) and 319 (19.8%) fragmented genes when compared against the embryophyta_odb10 database, while 753 (32.4%) complete (compared to 681 in earlier) and 288 (12.4%) fragmented genes when compared against the eudicots_odb10 database. Despite not meeting the optimal conditions for genome assembly with the chosen assembler, the produced genome assembly is slightly better than the previously available draft genome from an African individual.

The RepeatMasker pipeline masked 490,173 repeats totaling 166,387,956 bp. RepeatScout found 2,285 families. However, these results should be considered with caution as the N50 of the assembly was low (<10,000 bp) and consequently RepeatModeler may not perform well with short contigs (Smit 2008–2015). The BRAKER2 annotation pipeline resulted in 79,619 coding genes, which were made available on Zenodo (DOI: 10.5281/zenodo.13270663), together with the genome assembly and annotation file.

Intraspecific genetic diversity

A total of 1,131 complete BUSCO single copy genes were retrieved from the *S. globulifera* genomes, of which 256 were present in both the African and American individual. Genetic diversity was

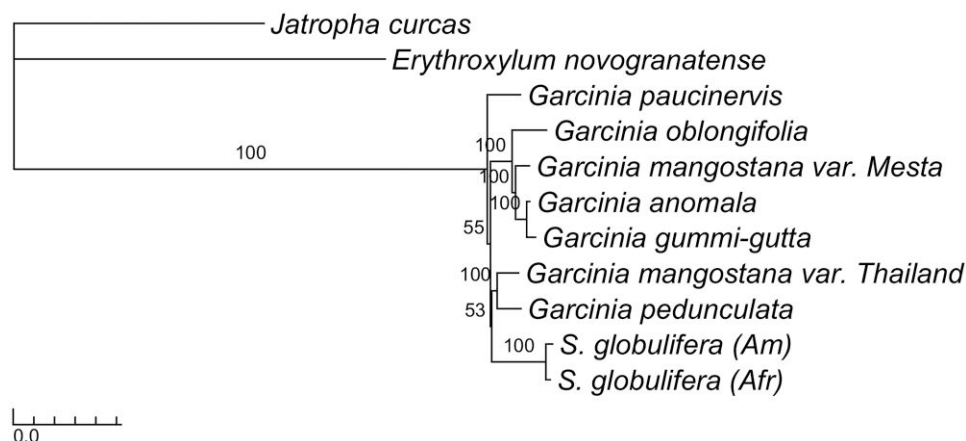


Fig. 2. Phylogenetic relationships of *S. globulifera* in relation to *Garcinia* based on 77 coding chloroplast genes. The support values labeled at each node indicate bootstrapped ML values. *J. curcas* and *E. novogranatense* served as outgroups.

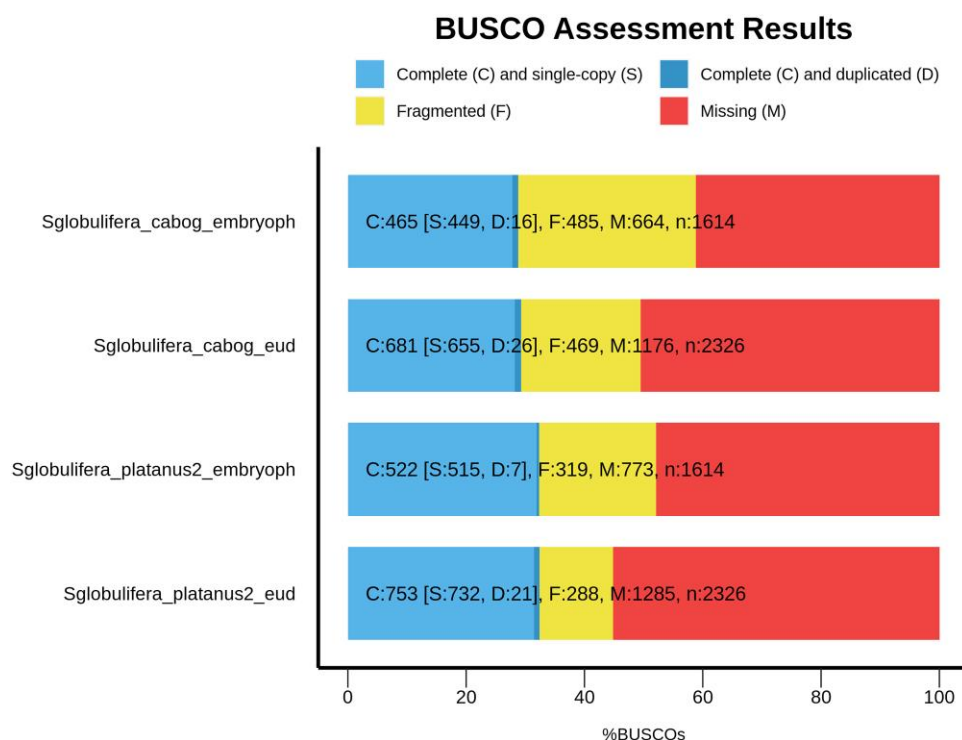


Fig. 3. Completeness of the assembled genome (*Sglobulifera_platanus2*) and for comparison, the existing genome assembly of an African individual (*Sglobulifera_cabog*). The databases used were the eudicots_odb10 (eud) and embryophyta_odb10 (embryoph).

measured in these shared 256 nuclear orthologous single-copy genes, which yielded a 448,390-bp-long concatenated alignment. Great variability in the length could be observed as the trimmed alignment (without introns present in only one of the specimens and without spurious sequences and poorly aligned regions) was only 161,212 bp long, of which 6.14% were variable.

It has become clear that one reference genome per species is not enough to capture the intraspecific genetic diversity in plants, as individuals of the same species often exhibit significant differences in the genome size, structural variants, and nucleotide-level polymorphisms such as SNPs and indels (Bayer et al. 2020). In addition to economically important species, genomic information from multiple individuals will be needed for tropical species to advance research on them, especially in species with high levels of intraspecific variation such as *S. globulifera*.

This high intraspecific genetic diversity has important implications for evolutionary biology and conservation. It suggests a rich genomic substrate for adaptation to diverse environments and may reflect complex demographic histories, including historical isolation, migration, or hybridization (Zuellig et al. 2014). From a conservation standpoint, recognizing and preserving such diversity is crucial for maintaining the species' evolutionary potential, especially under changing climate conditions.

Conclusions

This study provides the first comprehensive genome-wide comparison between African and South American individuals of *S. globulifera*, including de novo assembled nuclear, chloroplast, and mitochondrial genomes. The results reveal significant

intraspecific genomic variation, underscore the importance of genomic resources in tropical species, and point out the need for a comprehensive phylogenetic study of relationships between *Symphonia* and *Garcinia*. The new genomic resources, including annotated genes and curated alignments, will be instrumental for future studies in tropical phylogenetics, comparative genomics, and conservation biology. Together, these findings highlight the need to include multiple individuals in plant genome projects to fully capture the diversity and evolutionary history within widespread tropical species.

Data availability

The raw reads, nuclear genome assembly (GCA_965614385.1), and chloroplast assembly (OZ172679.1) are available from the ENA under project PRJEB62664. Nuclear coding sequences from the de novo genome assembly, the annotated genome assembly, and pairwise alignments of nuclear single-copy genes between African and American *S. globulifera* specimens are available from Zenodo under DOI 10.5281/zenodo.13270663 together with aligned chloroplast and mitochondrial coding genes.

Supplemental material available at [C3](#) online.

Acknowledgments

We acknowledge the Supercomputing Centre of Galicia (CESGA), SCBI (Supercomputing and Bioinformatics) Centre of the University of Malaga, as well as CSC—Finnish IT Center for Science for the allocation of computational resources. Three anonymous reviewers are acknowledged for helpful comments that significantly improved the original manuscript.

Funding

This study received funding from the French National Research Agency (ANR) under the project FLAG (ANR-12-ADAP-0007) and the "Investissement d'Avenir" grant managed by ANR (CEBA: ANR-10-LABX-25-01), the Spanish Ministry of Science and Innovation (MICINN) under the project AFFLORA (CGL2012-40129-C02-02), cofunded by the ERDF (European Regional Development Fund), as well as Plan Andaluz de Investigación, Desarrollo e Innovación under the grant P10-CVI-6075.

Conflicts of interest. None declared.

Literature cited

- Allié E et al. 2015. Pervasive local-scale tree-soil habitat association in a tropical forest community. *PLoS One*. 10:e0141488. <https://doi.org/10.1371/journal.pone.0141488>.
- Asif MH et al. 2010. Complete sequence and organisation of the *Jatropha curcas* (Euphorbiaceae) chloroplast genome. *Tree Genet Genomes*. 6:941–952. <https://doi.org/10.1007/s11295-010-0303-0>.
- Barlow J et al. 2018. The future of hyperdiverse tropical ecosystems. *Nature*. 559:517–526. <https://doi.org/10.1038/s41586-018-0301-1>.
- Bayer PE, Golicz AA, Scheben A, Batley J, Edwards D. 2020. Plant pan-genomes are the new reference. *Nat Plants*. 6:914–920. <https://doi.org/10.1038/s41477-020-0733-0>.
- Bedoya AM et al. 2019. Plastid genomes of five Species of riverweeds (Podostemaceae): structural organization and comparative analysis in Malpighiales. *Front Plant Sci*. 10:1035. <https://doi.org/10.3389/fpls.2019.01035>.
- Bininda-Emonds ORP. 2005. Seqcat.pl v1.0. Program distributed by the author, AG systematik und evolutions biologie. IBU—Fakultät V, Carl von Ossietzky Universität Oldenburg.
- Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 30:2114–2120. <https://doi.org/10.1093/bioinformatics/btu170>.
- Brousseau L, et al. 2014. High-throughput transcriptome sequencing and preliminary functional analysis in four Neotropical tree species. *BMC Genom*. 15:238.
- Cao S, Li M, Li LM. 2023. RegCloser: a robust regression approach to closing genome gaps. *BMC Bioinformatics*. 24:249. <https://doi.org/10.1186/s12859-023-05367-0>.
- Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. Trimal: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. 25:1972–1973. <https://doi.org/10.1093/bioinformatics/btp348>.
- Coil D, Jospin G, Darling AE. 2015. A5-miseq: an updated pipeline to assemble microbial genomes from Illumina MiSeq data. *Bioinformatics*. 31:587–589. <https://doi.org/10.1093/bioinformatics/btu661>.
- Daniell H, Lin C-S, Yu M, Chang W-J. 2016. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. *Genome Biol*. 17:134. <https://doi.org/10.1186/s13059-016-1004-2>.
- Darling AC, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res*. 14:1394–1403. <https://doi.org/10.1101/gr.2289704>.
- De Kort H et al. 2021. Life history, climate and biogeography interactively affect worldwide genetic diversity of plant and animal populations. *Nat Commun*. 12:516. <https://doi.org/10.1038/s41467-021-20958-2>.
- Dick CW, Abdul-Salim K, Bermingham E. 2003. Molecular systematic analysis reveals cryptic tertiary diversification diversification of a widespread tropical rain forest tree. *Am Nat*. 162:691–703. <https://doi.org/10.1086/379795>.
- Dick CW, Heuertz M. 2008. The complex biogeographic history of a widespread tropical tree species. *Evolution*. 62:2760–2774. <https://doi.org/10.1111/j.1558-5646.2008.00506.x>.
- Falgueras J et al. 2010. SeqTrim: a high-throughput pipeline for pre-processing any type of sequence read. *BMC Bioinformatics*. 11:38. <https://doi.org/10.1186/1471-2105-11-38>.
- Ferrer-Paris JR et al. 2019. An ecosystem risk assessment of temperate and tropical forests of the Americas with an outlook on future conservation strategies. *Conserv Lett*. 12:e12623. <https://doi.org/10.1111/conl.12623>.
- Formenti G et al. 2022. The era of reference genomes in conservation genomics. *Trends Ecol Evol*. 37:197–202. <https://doi.org/10.1016/j.tree.2021.11.008>.
- Gaudeul M, Sweeney P, Munzinger J. 2024. An updated infrageneric classification of the pantropical species-rich genus *Garcinia* L. (Clusiaceae) and some insights into the systematics of New Caledonian species, based on molecular and morphological evidence. *Phytokeys*. 239:73–105. <https://doi.org/10.3897/phytokeys.239.112563>.
- Greiner S, Lehwork P, Bock R. 2019. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Res*. 47:W59–W64. <https://doi.org/10.1093/nar/gkz238>.
- Heslop-Harrison JS, Schwarzacher T. 2011. Organization of the plant genome in chromosomes. *Plant J*. 66:18–33. <https://doi.org/10.1111/j.1365-3113.2011.04544.x>.
- Heuertz M et al. 2023. The application gap: genomics for biodiversity and ecosystem service management. *Biol Conserv*. 278:109883. <https://doi.org/10.1016/j.biocon.2022.109883>.

- Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBBoot2: improving the ultrafast bootstrap approximation. *Mol Biol Evol.* 35:518–522. <https://doi.org/10.1093/molbev/msx281>.
- Holliday JA et al. 2017. Advances in ecological genomics in forest trees and applications to genetic resources conservation and breeding. *Mol Ecol.* 26:706–717. <https://doi.org/10.1111/mec.13963>.
- Jin J-J et al. 2020. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biol.* 21:241. <https://doi.org/10.1186/s13059-020-02154-5>.
- Kajitani R et al. 2019. Platanus-alley is a de novo haplotype assembler enabling a comprehensive access to divergent heterozygous regions. *Nat Commun.* 10:1702. <https://doi.org/10.1038/s41467-019-09575-2>.
- Kalyanamoorthy S, Minh BQ, Wong T, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods.* 14:587–589. <https://doi.org/10.1038/nmeth.4285>.
- Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30:3059–3066. <https://doi.org/10.1093/nar/gkf436>.
- Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 30:772–780. <https://doi.org/10.1093/molbev/mst010>.
- Kelchner SA. 2000. The evolution of non-coding chloroplast DNA and its application in plant systematics. *Ann Mo Bot Gard.* 87:482–498. <https://doi.org/10.2307/2666142>.
- Kent WJ. 2002. BLAT—the BLAST-like alignment tool. *Genome Res.* 12:656–664. <https://doi.org/10.1101/gr.229202>.
- Kokot M, Dlugosz M, Deorowicz S. 2017. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics.* 33:2759–2761. <https://doi.org/10.1093/bioinformatics/btx304>.
- Langmead B, Wilks C, Antonescu V, Charles R. 2019. Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics.* 35:421–432. <https://doi.org/10.1093/bioinformatics/bty648>.
- Le Provost G, Paiva J, Pot D, Brach J, Plomion C. 2003. Seasonal variation in transcript accumulation in wood-forming tissues of maritime pine (*Pinus pinaster* Ait.) with emphasis on a cell wall glycine-rich protein. *Planta.* 217:820–830. <https://doi.org/10.1007/s00425-003-1051-2>.
- Liu X et al. 2023. Complete chloroplast genome of *Hypericum perforatum* and dynamic evolution in *Hypericum* (Hypericaceae). *Int J Mol Sci.* 24:16130. <https://doi.org/10.3390/ijms242216130>.
- Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol.* 38:4647–4654. <https://doi.org/10.1093/molbev/msab199>.
- Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics.* 27:764–770. <https://doi.org/10.1093/bioinformatics/btr011>.
- Mikheenko A, Pribelski A, Saveliev V, Antipov D, Gurevich A. 2018. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics.* 34:i142–i150. <https://doi.org/10.1093/bioinformatics/bty266>.
- Mora-Márquez F, Chano V, Vázquez-Poletti JL, López de Heredia U. 2021. TOA: a software package for automated functional annotation in non-model plant species. *Mol Ecol Resour.* 21:621–636. <https://doi.org/10.1111/1755-0998.13285>.
- Müller KF. 2005. SeqState primer design and sequence statistics for phylogenetic DNA data sets. *Appl Bioinformatics.* 4:65–69. <https://doi.org/10.2165/00822942-200504010-00008>.
- Müller KF, Quandt D, Müller J, Neinhuis C. 2005. PhyDE® 0.995: Phylogenetic Data Editor. <http://www.phyde.de>.
- Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol.* 32:268–274. <https://doi.org/10.1093/molbev/msu300>.
- O’Conner S, Li L. 2020. Mitochondrial fostering: the mitochondrial genome may play a role in plant orphan gene evolution. *Frontiers of Plant Science* 11, 600117. <https://doi.org/10.3389/fpls.2020.600117>.
- Olsson S et al. 2017. Development of genomic tools in a widespread tropical tree, *Symphonia globulifera* L.f.: a new low-coverage draft genome, SNP and SSR markers. *Mol Ecol Resour.* 17:614–630. <https://doi.org/10.1111/1755-0998.12605>.
- Oyen LPA. 2005. *Symphonia globulifera* [WWW Document]. (accessed 3.24.25). <https://prota.prota4u.org/protav8.asp?g=pe&p=Symphonia+globulifera>
- Pennington RT, Lavin M. 2016. The contrasting nature of woody plant species in different neotropical forest biomes reflects differences in ecological stability. *New Phytol.* 210:25–37. <https://doi.org/10.1111/nph.13724>.
- Ranallo-Benavidez TR, Jaron KS, Schatz MC. 2020. GenomeScope 2.0 and smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun.* 11:1432. <https://doi.org/10.1038/s41467-020-14998-3>.
- Schmitt S, Tysklind N, Hérault B, Heuert M. 2021. Topography drives microgeographic adaptations of closely related species in two tropical tree species complexes. *Mol Ecol.* 30:5080–5093. <https://doi.org/10.1111/mec.16116>.
- Sloan DB et al. 2012. Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. *PLoS Biol.* 10:e1001241. <https://doi.org/10.1371/journal.pbio.1001241>.
- Sloan DB. 2013. One ring to rule them all? Genome sequencing provides new insights into the ‘master circle’ model of plant mitochondrial DNA structure. *New Phytol.* 200:978–985. <https://doi.org/10.1111/nph.12395>.
- Smit AFA, Hubley R. 2008-2015. RepeatModeler Open-1.0. <http://www.repeatmasker.org>.
- Smit AFA, Hubley R, Green P. 2013-2015. RepeatMasker Open-4.0. <http://www.repeatmasker.org>.
- Stothard P. 2000. The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *Biotechniques.* 28:1102–1104. <https://doi.org/10.2144/00286ir01>.
- Stöver BC, Müller KF. 2010. TreeGraph 2: combining and visualizing evidence from different phylogenetic analyses. *BMC Bioinform.* 11:7. <https://doi.org/10.1186/1471-2105-11-7>.
- Sweeney PW. 2008. Phylogeny and floral diversity in the genus *Garcinia* (Clusiaceae) and relatives. *Int J Plant Sci.* 169:1288–1303. <https://doi.org/10.1086/591990>.
- Ter Steege H et al. 2006. Continental-scale patterns of canopy tree composition and function across Amazonia. *Nature.* 443:444–447. <https://doi.org/10.1038/nature05134>.
- Theissinger K et al. 2023. How can genomics help biodiversity conservation. *Trends Genet.* 39:545–559. <https://doi.org/10.1016/j.tig.2023.01.005>.
- Tillich M et al. 2017. Geseq—versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* 45:W6–W11. <https://doi.org/10.1093/nar/gkx391>.
- Torroba-Balmori P et al. 2017. Altitudinal gradients, biogeographic history and microhabitat adaptation affect fine-scale spatial genetic structure in African and neotropical populations of an

- ancient tropical tree species. PLoS One. 12:e0182515. <https://doi.org/10.1371/journal.pone.0182515>.
- Vaidya G, Lohman DJ, Meier R. 2011. SequenceMatrix: concatenation software for the fast assembly of multi-gene datasets with character set and codon information. Cladistics. 27:171–180. <https://doi.org/10.1111/j.1096-0031.2010.00329.x>.
- Vurtture GW, et al.. 2017. GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics. 33:2202–2204. <https://doi.org/10.1093/bioinformatics/btx153>.
- Wcisel DJ et al. 2020. Transcriptome ortholog alignment sequence tools (TOAST) for phylogenomic dataset assembly. BMC Evol Biol. 20:41. <https://doi.org/10.1186/s12862-020-01603-w>.
- Wee CC et al. 2022. Mitochondrial genome of *Garcinia mangostana* L. variety mesta. Sci Rep. 12:9480. <https://doi.org/10.3390/plants12040930>.
- Wee C-C, et al.. 2023. Plastomes of *Garcinia mangostana* L. and comparative analysis with other *Garcinia* species. Plants. 12:930. <https://doi.org/10.3390/plants12040930>.
- Wick RR, Schultz MB, Zobel J, Holt KE. 2015. Bandage: interactive visualization of de novo genome assemblies. Bioinformatics. 31:3350–3352. <https://doi.org/10.1093/bioinformatics/btv383>.
- Xie Y et al. 2014. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics. 30:1660–1666. <https://doi.org/10.1093/bioinformatics/btu077>.
- Zuellig MP, Kenney AM, Sweigart AL. 2014. Evolutionary genetics of plant adaptation: insights from new model systems. Curr Opin Plant Biol. 18:44–50. <https://doi.org/10.1016/j.pbi.2014.01.001>.

Editor: A. Sethuraman