Building a novel nuclear-organelle genomic framework for the fever tree (*Cinchona pubescens* Vahl) through short and long-read DNA data assemblies

Nataly Allasi Canales[1,2]*, Oscar A. Pérez-Escobar[2]*, Robyn F. Powell[2], Mats Töpel[3], Catherine Kidner[4], Mark Nesbitt[2], Carla Maldonado[5], Christopher J. Barnes[6], Nina Rønsted[1,7], Ilia J. Leitch[2+], Alexandre Antonelli[2,5,9,10+]

[1]Natural History Museum of Denmark, University of Copenhagen, Denmark.

[2]Royal Botanic Gardens, Kew, London, UK.

[3]University of Gothenburg, Department of Marine Sciences, Sweden.

[4]Royal Botanic Garden Edinburgh, UK

[5]Herbario Nacional de Bolivia, Instituto de Ecología, Universidad Mayor de San Andrés, La Paz, Bolivia

[6]The GLOBE Institute, University of Copenhagen, Denmark

[7]National Tropical Botanical Garden, Kalaheo, Hawaii, USA

[9]Gothenburg Global Biodiversity Centre, Sweden.

[10]Department of Plant Sciences, University of Oxford, Oxford, UK.

*These authors contributed equally to the study

+Joint senior authors

Corresponding authors: o.perez-escobar@kew.org (OAPE), or i.leitch@kew.org (IJL)

Abstract

Background: The Andean fever tree (*Cinchona* L.; Rubiaceae) is the iconic source of bioactive quinine alkaloids which have been key to treating malaria for centuries. In particular, *C. pubescens* Vahl has been an important source of income for several countries in its native range in north-western South America. However, the genomic resources required to place *Cinchona* species in the tree of life and to explore the evolution and biosynthesis of alkaloids are meagre. **Findings:** Using a combination of ~120 Gb of long sequencing reads derived from the Oxford Nanopore PromethION platform and 142 Gb of short read Illumina data, we address this gap by providing the first highly contiguous nuclear and organelle genome assemblies and their corresponding annotations. Our nuclear genome assembly consists of 603 scaffolds comprising a total length of 903 Mb, representing ~85% of the genome size (1.1 Gb/1C). This draft genome sequence was complemented by annotating 72,305 CDSs using a combination of *de novo* and reference-based transcriptome assemblies. Completeness analysis revealed that our assembly is highly complete, displaying 83% of the BUSCO gene set, and a small fraction of genes (4.6%) classified as fragmented. We demonstrate the utility of these novel genomic resources by placing *C. pubescens* in the Gentianales order using plastid and nuclear datasets.

Conclusions: Our study provides the first genomic resource for *C. pubescens* thus opening new research avenues, including the unravelling of the gene toolkit for alkaloid biosynthesis in the fever tree.

Keywords: Oxford Nanopore, Rubiaceae, transcriptomics, whole genome sequencing.

Data Description

1.1 Background

The fever tree (*Cinchona* L., Rubiaceae) is a genus of 24 species native to the Eastern slopes of the Andes mountain range in South America ([1,2]; Fig. 1) and perhaps one of the most economically important genera in the family, second only to coffee [3]. The genus is widely known as the source of a group of at least 35 quinine alkaloids collectively called quinolines that have been shown to be key to ameliorating the fever and chill episodes associated with malaria [4]. As such, the fever trees have played a crucial role in the economies and livelihoods of people worldwide for centuries [5,6].

Despite the tremendous historical and economic importance of *Cinchona*, DNA sequence datasets for *Cinchona* are rather meagre, limited to 252 DNA Sanger sequences available in the NCBI repository (accessed on May 17, 2021; [7]). More importantly, no nuclear and organellar reference genomes exist for any species of the genus. As such, important fundamental and applied questions – such as the mode and tempo of evolution of the fever tree, or the genetic pathways responsible for quinine alkaloid production – remain elusive. Previous phylogenetic studies of the Rubiaceae family, and more specifically of the Cinchonoideae subfamily where the Cinchoneae tribe is found, are based on just a handful of nuclear (ITS) and plastids data sets (matK, rcbL, rps16, trnL-F). They show an unresolved polytomy between the tribes and the seven genera of the Cinchoneae tribe that have so far been included in more specific studies [8,9] (including the genus *Cinchona*, which shows very unclear relationships). Furthermore, studies are lacking of the relationships within species of this genus [7,8]. A recent genome-wide phylogenetic tree for the order Gentianales (Antonelli et al., in press) provided strong support for *C. pubescens* as sister to *Isertia hypoleuca*, but the sampling was exclusively at the genus level and therefore did not include any other species of *Cinchona* nor other genera in tribe *Cinchoneae.*

The production of alkaloids is highest in *C. calisaya*, also known as yellow bark [10][11]. However, several species of the genus *Cinchona* have historically been harvested to provide sources of quinine alkaloids, one of the most traded natural products, resulting in significant reductions in their natural ranges and population size [12,13]. Nevertheless, among them, *C. pubescens* or red Cinchona bark is now widely cultivated throughout the New and Old-World tropics, with some instances where the species has escaped cultivation and become invasive [14]. Extensive research on the structure, abundance, and chemical composition of quinine alkaloids in the *Cinchona* genus have been conducted [15], revealing further potential in novel drug discovery. However, the identity of the genes involved in the synthetic pathway of quinine alkaloids remain elusive.

Here, we present the first high quality draft nuclear and plastid genomes of *C. pubescens*, which is characterised by having a genome size of 1.1 Gb (1C, this study) and a chromosome number of 2n=34. The assemblies were generated using a combination of extensive long-read Nanopore (~218x) and short-read Illumina paired-read datasets (~300x) jointly with state-of-the-art genome assemblers, resulting in a reference genome for which contiguity and quality are comparable to, or even higher [16] than in the three previously published genome assemblies in Rubiaceae, such as *Chiococca alba [16]*, *Coffea canephora* [17], and *Coffea arabica [18]*. The plastid genome from short-reads of *C. pubescens* had a length of 156,985 bp and a GC content of 37.74%, very similar to other Rubiaceae plastid genomes [16,19]. Lastly, we demonstrate the utility and reliability of our resources by constructing nuclear and plastid phylogenomic frameworks of *C. pubescens*.

1.2 Sampling and genomic DNA/RNA sequencing

We sampled leaves from a single *Cinchona pubescens* individual (1977-69, propagated vegetatively from a tree collected in Tanzania in 1977) cultivated in the Temperate House of the Royal Botanic Gardens, Kew (RBG Kew), UK (a voucher was also prepared which is deposited in the RBG Kew herbarium (**K**)). DNA was extracted from fresh tissue using two different protocols to produce paired-end Illumina and native Nanopore libraries. For Illumina DNA library preparations, we used 1000 mg of starting material that was first frozen with liquid nitrogen and subsequently ground in a mortar. The Qiagen DNeasy (Qiagen, Denmark) plant kit was used to extract DNA from the ground tissue, following the manufacturer's protocol. We built the libraries using the Illumina TruSeq PCR-free library (NEX, Ipswich, MA, USA) following the manufacturer's protocol, by first quantifying the DNA quantity and quality using a Nanodrop fluorometer (Thermo Scientific, Denmark) and then fragmenting oligonucleotide strands through ultrasonic oscillation using a Covaris ME220 (Massachusetts, USA) device to yield fragments with an average length of 350 bp. Then we sequenced on a lane the paired-end 150 bp libraries using the HiSeq X Ten chemistry. Total RNA was extracted from 1000 mg of frozen-ground leaf, bract, and flower tissue using the TRIZoL reagent (Thermo Fisher Scientific, Denmark) following the manufacturer's protocol. Illumina library preparation and sequencing were conducted by Genewiz GmbH (Leipzig, Germany).

The nanopore sequencing data were generated and base called as part of Oxford Nanopore's London Calling 2019 conference [20]. For Nanopore library preparation, 1000 mg of leaf tissue was frozen and ground with a mortar and pestle. The lysis was carried with Carlson lysis buffer (100 mM Tris-HCl, pH 9.5, 2% CTAB, 1.4 M NaCl, 1% PEG 8000, 20 mM EDTA) supplemented with β-mercaptoethanol. The sample was extracted with chloroform and precipitated with isopropanol. Finally, it was purified with QIAGEN Blood and Cell Culture DNA Maxi Kit (Qiagen, UK). Size selection was performed using the Circulomics Short Read Eliminator kit (Circulomics, MD, USA) to deplete fragments below 10 kb. Then libraries were prepared using the ONT Ligation Sequencing Kit (SQK-LSK109, Oxford Nanopore Technologies, UK). During sequencing on the PromethION library, re-loads were performed when required. Though yield was slightly lower in sequencing for these samples (over 50 Gb in 24 hours), the read N50 was over 48 kb (up from 28 kb without size selection).

1.3 Estimation of genome size

To accurately determine the genome size of *C. pubescens*, we followed the one-step flow cytometry procedure [21], with modifications as described in Pellicer et al. [22]. Freshly collected tissue from the same individual sampled for DNA/RNA sequencing was measured together with *Oryza sativa* L. ‘IR-36’ as the calibration standard using general purpose buffer” (GPB) [23] supplemented with 3% PVP-40 and β-mercaptoethanol [22]. The samples were analysed on a Partec Cyflow SL3 flow cytometer (Partec GmbH, Münster, Germany) fitted with a 100 mW green solid-state laser (532 nm, Cobolt Samba, Solna, Sweden). Three replicates were prepared and the output histograms analyzed using the FlowMax software v.2.4 (Partec GmbH, Münster, Germany). The 1C-value of *C. pubescens* was calculated as: (Mean peak position of *C. pubescens*/Mean peak position of *O. sativa*) × 0.49 Gb(= 1C value of *O. sativa*) [24] and resulted in a 1C-value of 1.1 Gb.

1.4 Short read data processing of the chloroplast genome assembly

Sequencing of the DNA Illumina library generated 428M paired reads, representing 310Gb of raw data. RNA sequencing produced 385M paired reads, representing 142.6 Gb. The quality of the raw reads was assessed using the FastQC software [25], and quality trimming was conducted using the software AdapterRemoval2 v.2.3.1 [26]. Here, bases with phred score quality <30 and read lengths <50 bp were removed together with adapter sequences. The final short-read dataset was of 131 Gb and contained 384,626,011 paired reads, which corresponds to 464.8x coverage (1.2 Gbases, see *DNA content assessment* section).

The plastid genome of *C. pubescens* was assembled using only short reads, as there were some discrepancies using the hybrid dataset. The toolkit GetOrganelle v.1.7.5, was used with the parameters suggested for assembling plastid genomes in Embryophyta (i.e., parameters -*R* 15, -*k* 21,45,65,85,105, -*F* embplant_pt). GetOrganelle produced a single linear representation of the *C. pubescens* plastid genome, with a length of 156,985 bp (Fig. 2) and a GC content of 37.74%. These are very similar to those reported for the *Coffea arabica* plastid genome, which is reported to be 155,189 bp in length and have a GC content of 37.4% [19].

We annotated the plastid genome assembly of *C. pubescens* in CHLOROBOX [27], which implements GeSeq [28], tRNAscan-SE v2.0.5 [29], and ARAGORN v1.2.38 [30]. CHLOROBOX annotations indicated that the *C. pubescens* genome has the typical angiosperm quadripartite structure, i.e., Inverted Repeat (IRa and IRb) (each 27,502 bp long), the Small Single Copy (SSC) region (18,051 bp), and the Large Single Copy (LSC) region (83,930 bp). We predicted 128 genes, of which 34-37 were tRNA (tRNAScan-SE and ARAGORN, respectively), 81 CDSs, and four ribosomal RNAs (rRNAs). The junction between SSC-IRa and LSC-IRa contains the *ycf1* (pseudogene) and *rps3* (gene), respectively. Similarly, the junction between IRb-SSC and LSC-IRb contains the *ycf1* (pseudogene) and *rsp3* (gene), respectively (Supp. Fig. 1). The final structural features of the *C. pubescens* plastid genome were generated using OGDRAW v. 1.3.1 [31] (Fig. 2) and edited manually. Finally, the quality of the plastid genome assembly was estimated by mapping the Illumina DNA short reads to the newly assembled genome using the bam pipeline in Paleomix [32], where we used BWA [33] for alignment, specifying the backtrack algorithm, and filtering minimum quality equal to zero. The overall mapping rate was 7.34%, prior to PCR duplicate filtering, and the coverage of unique hits was 7,960x.

1.5 Long-read nuclear genome assembly and quality assessment

The quality and quantity of the PromethION sequencing output conducted across four flow cells was evaluated in NanoPlot v.1.82 [34] independently for each flow cell, using as input the sequencing summary report produced by Guppy v3.0.3. Overall, the average read length, phred score quality and N50 following base calling with Guppy v3.0.3, and the High Accuracy model reached values of ~19,000 bp, 9, and ~46,000 bp, for mean read length, mean read quality, and read length N50, respectively (Supp. Tab. 1). A total of 13,252,640 quality-passed reads were produced, representing ~262 Gb and providing a theoretical genome coverage of ~218x. To assemble the raw Nanopore reads into scaffolds, we first corrected and trimmed the quality-passed reads using the software CANU v.1.9 [35] in correction and trimming mode with the following parameters: *genomeSize* = 1.1g, -*nanopore-raw*. This step generated a total of 1,265,511 reads, representing c. 89 Gb, or a theoretical genome coverage of 74x. Next, the corrected/trimmed reads were used as input into SMARTdenovo v.1.0 [36], using the following parameters: -*c* 1 (generate consensus mode), -*k* 16 (k-mer length) and -*J* 5000 (minimum read length). This step produced an assembly composed of 603 scaffolds with an N50 = 2,783,363 bp, representing ~903 Mb (~84% of the genome size; Tab. 1). Lastly, a round of scaffold correction was implemented in RACON v.1.4.3 [37] using as input the corrected Nanopore reads generated by CANU and an alignment SAM file produced by mapping the trimmed DNA Illumina reads against the assembly produced by SMARTdenovo. The alignment file was produced by Minimap2 v.2.18 [38] using the “accurate genomic read mapping” settings designed to map short read Illumina data (flag -*ax*). RACON was executed using an error threshold of 0.3 (-*e* flag), a quality threshold of 10 (-*q*), and a window length of 500 (-*w*). The corrected assembly differed little compared with the raw assembly produced by SMARTdenovo (Tab. 1).

We followed a two-pronged approach to assess the quality of our corrected nuclear genome assembly by: i) evaluating the proportion of Illumina reads that mapped against our new genome assembly using as input the SAM file generated by Minimap2 and computing coverage and mean depth values per scaffold, as implemented in the function *view* (flag –F 260) of the software Samtools v1.12 [39]; and ii) estimating the completeness of the genome as implemented in the software BUSCO v.5.12 and using the eudicots_odb10 [40]. A total of 827,098,761 reads were mapped against the corrected genome assembly, representing 99% of the trimmed reads used as input (241,498,983). Mean coverage and read depth ranged from 26-48x. The genome completeness analysis recovered a total of 87.6% conserved eudicot genes, of which 77% were single copy, 6% duplicated, and 4.6% fragmented. The remaining BUSCO genes were labelled as missing (12.4%). Taken together, our results suggest that our nuclear genome assembly presents high contiguity and quality with a moderate completeness.

1.6 Transcriptome assembly, candidate gene annotation, and quality assessment

To produce a comprehensive database of assembled transcripts, we generated reference-based and *de-novo* assemblies with the Trinity toolkit v. 2.8 [41] using the trimmed RNA-seq data. The reference-based assembly was conducted using as input the aligned RNA-seq trimmed reads against our new reference genome as produced by aligner STAR v.2.9 [42] with default settings, and a maximum intron length of 57,000 as estimated for *Arabidopsis thaliana* (flag --*genome_guided_max_intron*). The *de-novo* transcriptome assemblies were also produced using the default settings of Trinity and the trimmed RNA-seq reads as input. A comprehensive database of *de-novo* and reference-based assembled transcriptomes was compiled with the software PASA v.2.0.2 [43], using the following parameters: --*min_per_ID* 95, and --*min_per_aligned* 30.

To assess the completeness of the *de-novo* transcriptome assembly, we used BUSCO v. 3.0.2 and the representative plant set viridiplantae_odb10, which currently includes 72 species, of which 56 are angiosperms. Our assembled transcriptome captured 72.5% (312/430) of the BUSCO set as complete genes, of the remainder, 19.3% of the genes were fragmented, and 8.2% were missing.

We predicted the structure and identity of the genes in the nuclear genome using the comprehensive transcriptome assembly compiled with PASA. For this purpose, we used AUGUSTUS v3.3.3 [44] for a combination approach of *ab initio* and transcript evidence-based from RNA-seq data. As we considered the transcripts as EST, we first generated hints from the transcriptome data by aligning the transcripts to the genome using BLAT v 3.5 [45]. Then, we set the hint parameters to rely on the hints and anchor the gene structure. Finally, we predicted the genes using the hints and tomato (*Solanum lycopersicum* L.) as reference species.

The structural annotation of the nuclear genome was performed with AUGUSTUS, in a combined *ab initio* and evidence-based approach. As evidence, we used the assembled transcriptome contigs, which are very similar in size to ESTs and cover more than one exon. Thus, we incorporated EST alignments to the nuclear genome as an extrinsic source. First, we prepared the hints from ESTs by aligning the ESTs against the genome using BLAT and filtering out the short alignments. Then, we generated the hints from the EST alignments using blat2hints.pl. As we were using the ESTs as an anchor for the structural annotation, we changed the hint parameters, so AUGUSTUS used the hints as evidence. Finally, using tomato as a reference species, we predicted 72,305 CDSs.

1.7 Nuclear and plastid phylogenomics of *Cinchona*

To verify the nuclear genome, we inferred nuclear phylogenetic relationships using the reference sequences of 353 low copy nuclear genes that are conserved across angiosperms from the Plant and Fungal Trees of Life project [46]. Here, we sampled gene sequences of 18 taxa from the Gentianales, which included another *C. pubescens* from that study (Supp. Tab. 2) that are publicly available in the Tree of Life Explorer [47] hosted by the Royal Botanic Gardens, Kew. To include this study *C. pubescens* in the sampling of 353 low copy nuclear genes of selected Gentianales, we then retrieved these genes from the RNA-seq data utilized to produce transcriptome assemblies using the pipeline HybPiper v.1.3.1 [48]. Given the abundance of RNA-seq read data, to render the gene retrieval tractable, as input for HybPiper we used a subsample of the trimmed read data, as implemented in the software seqtk [49]. The gene sequences produced by HybPiper were aligned with the data for 19 selected Gentianales species using MAFFT v7.453 [50] and then they were concatenated into a supermatrix for phylogenomic analyses.

We implemented the maximum likelihood approach using RAxML-HPC V.8 [51] with a GTRGAMMA substitution model for each gene and a rapid bootstrap analysis with 500 replications. Then we filtered the bipartition trees that only had LBS>=20 using Newick utilities [52]. The resulting trees were rooted using phyx v1.2.1 [53], setting *Uncarina grandidi*eri (Baill.) Stapf (Lamiales) as the root. To estimate the species tree from the gene trees, we used the coalescent approach with ASTRAL 5.6.1 to calculate the quartet scores, which is the number of quartet trees present in the gene trees that are also present in the species tree. Q1 which shows the support of the gene trees for the main topology, q2 which supports the first alternative topology, and q3 showing the support for the second alternative topology [54]. We incorporated these scores into the species tree with an R script [55]. All trees were visualized with FigTree v.1.4.4 [56].

In the nuclear phylogenomic tree resulting from the 353 low copy nuclear genes (Fig. 3), *C. pubescens* clusters within the Cinchonoideae, which appears more closely related with the Ixoroideae group than with Rubioideae. Additionally, most of the nodes were also highly supported by quartet scores, showing that a large proportion of the gene trees agreed with the species tree.

For the plastid phylogeny, we used *Sesamum indicum* L. as an outgroup from the Lamiids cluster [57]. We performed maximum likelihood using the complete plastid genomes of the 20 species available to date in the Gentianales. All the plastid genomes we analysed had the classic quadripartite genomic structure, although some Rubiaceae species also show the tripartite structure [58]. We aligned the 20 Gentianales (Supp. Tab. 3) plastid genomes with MAFFT v7.427 using the default parameter settings to perform the multiple sequence alignments. Then we estimated the phylogenetic tree with the maximum likelihood approach using the GTRCAT model RAxML-HPC v.8. We conducted heuristic searches with 1000 bootstrap replicates (rapid bootstrapping and search for the best-scoring ML tree). Both analyses were performed on the Cipres Science Gateway [59].

As with the nuclear tree, the plastid trees were also clustered at the subfamily level, recovering the Cinchonoideae, Ixoroideae, and Rubioideae as natural groups, alongside the two species belonging to Pedilaceae used as outgroups for the phylogenetic analysis. For the plastid data, the vast majority of nodes were strongly supported (14 nodes had LBS=100), all but one node had a BS of 100%. However, we found *Gynochthodes nanlingensis* (Y.Z.Ruan) Razafim. & B.Bremer (Rubioideae) to cluster with other Apocynaceae species. While the same result has previously been reported in other studies [60,61], it seems to be due to an erroneous DNA sequence attributed to *G. nanlingensis* or a misidentification of the voucher, so it is recommended this is thoroughly checked before making any further statements in this regard. Additionally, the ingroup showed that the Cinchonoideae and Ixoroideae subfamilies are sisters while Rubioideae is placed as sister to this clade. The placement of *C. pubescens* in the Cinchonoideae subfamily cluster using both plastid and nuclear data presented in this study is consistent with previous taxonomic and phylogenetic studies [62] which gives support to the robustness to the assembled genome. As potential future work the Nanopore sequencing data could be re-base called using the latest algorithms from Oxford Nanopore to take advantage of recent developments in this area over the last few years which has seen continuous improvement in raw-read accuracy [63,64].

2. Conclusion

Using a combination of extensive short and long read DNA datasets, we deliver the first highly contiguous and robust nuclear-plastid genome assemblies for one of the historically most traded and economically important *Cinchona* species, *C. pubescens*. As the third species of the Rubiaceae family with a nuclear genome sequenced in great detail, the abundant genomic resources provided here will open up new research avenues to disentangle the evolutionary history of the Fever tree. In the short-term, these genomic tools will significantly help to unravel the genes involved in the biosynthetic pathways of quinine alkaloids synthesis, identifying the underpinning genetic diversity of these genes both between and within species, and shed light on how the expression of these genes is regulated. It is hoped that the nuclear and plastid genome presented here will become the foundation for the Fever tree's genomic data and databases. Our nuclear scaffolds and plastid genome assembly will enable future reference-guided assemblies, variant calling, and gene annotation to enhance functional analysis within the *Cinchona* genus, with potential to further explore the quinine alkaloid biosynthetic pathway in depth and hence enhance its potential for finding new medicinal leads to treat malaria.

Data availability

The genome sequences data, nuclear and plastid assemblies are available at the NBCI repository, under the BioProject number PRJAXXXXX.

Competing interests

The authors declare that they have no competing interests.

Funding

NR, AA, CB, and NC and received funding from H2020 MSCA-ITN-ETN Plant.ID, a European Union's Horizon 2020 research and innovation programme under grant agreement No 765000. OAPE acknowledges financial support from the Swiss Orchid Foundation and the Lady Sainsbury Fellowship at the Royal Botanic Gardens, Kew. PromethION sequencing (flow cells and consumables) were provided by Oxford Nanopore Technologies. AA and NR acknowledge funding from the SciLifeLab 2015 Biodiversity Program. AA is further funded by the Swedish Research Council, the Swedish Foundation for Strategic Research and the Royal Botanic Gardens, Kew.

Author contributions

IJL, OAPE, NR, CB, MT and AA conceived the study. MN, VC and OAPE collected plant tissue. OAPE, IJL, RFP, MT and AA generated datasets. OAPE, NC, CK and MT conducted in-silico analyses. OAPE and NC wrote the manuscript, with contributions from all co-authors.

Acknowledgements

We thank Jonathan Pugh, Vania Costa, and Simon Mayes for support and assistance during Nanopore sequencing preparation.

Legends

Figure 1. A. *Cinchona* trees in the Andean cloud forest. B. The *C. pubescens* specimen studied in this work (CP9014) growing in the Temperate House at the Royal Botanic Gardens, Kew,

UK. C. Inflorescence of *C. pubescens*, CP9014. D. *Cinchona* barks from the Economic Botany Collection, Kew, UK. E. Distribution map of the *Cinchona* genus across the American continent shown in blue dots, modified from Maldonado et al. 2015 [65].

Figure 2. Annotated *C. pubescens* plastome. Genes displayed on the inside of the circle are transcribed clockwise, while genes positioned on the outside are transcribed counter clockwise.

Figure 3. The coalescent-based species tree estimation of the Gentianales order without distances, inferred using low copy nuclear gene trees. Pie charts positioned on the nodes represent the percentage of the gene trees that agree with the topology of the main species tree (red) and the other two alternative topologies (cyan and grey). The “Cinchona_pubescens” specimen is from the Plant and Fungal Trees of Life project [46] and “C_pubescens_WGS” is from this study.

Figure 4. Phylogenetic tree showing the relationships of twenty Gentianales species built from the whole plastid genome. Numbers on the branches are bootstrap percentages of ML estimations. The coloured boxes depict the subfamily level that groups the species analysed.

Table 1. Summary assembling statistics for *C. pubescens* using SMARTdenovo and RACON.

Additional files

Supplementary table 1. Summary statistics of the Nanopore reads.

Supplementary table 2. Overview of the samples from the Tree of Life Explorer (Royal Botanic Gardens, Kew) that were used in the phylogenetic analysis to construct the coalescent tree.

Supplementary table 3. Sample overview of the specimens and their accession numbers used to infer the phylogenetic tree built using plastid data.

Supplementary figure 1. Plastid genome visualization of junctions IRb-SSC (*ycf1*) and LSC-IRb (*rsp3*).

References

1. Andersson L. A revision of the genus Cinchona (Rubiaceae-Cinchoneae). Memoirs of the New York Botanical Garden; 1998;80: 1-75.

2. Maldonado C, Persson C, Alban J, Antonelli A, Rønsted N. *Cinchona anderssonii* (Rubiaceae), a new overlooked species from Bolivia. Phytotaxa. Magnolia Press; 2017;203–8.

3. Steere WC. The Cinchona-Bark Industry of South America. Sci Mon. American Association for the Advancement of Science; 1945;61:114–26.

4. Kacprzak KM. Chemistry and biology of *Cinchona* alkaloids. Nat Products Bioprospect. Springer-Verlag: Berlin; 2013;605–41.

5. Lee MR. Plants against malaria. Part 1: *Cinchona* or the Peruvian bark. J R Coll Physicians Edinb. 2002;32:189–96.

6. Walker K, Nesbitt M. Just the Tonic: A Natural History of Tonic Water. Kew Publishing; 2019.

7. Home - Nucleotide - NCBI [Internet]. [cited 2021 May 17]. Available from: https://www.ncbi.nlm.nih.gov/nuccore

8. Andersson L, Antonelli A. Phylogeny of the Tribe Cinchoneae (Rubiaceae), Its Position in Cinchonoideae, and Description of a New Genus, *Ciliosemina*. Taxon. International Association for Plant Taxonomy (IAPT); 2005;54:17–28.

9. Manns U, Bremer B. Towards a better understanding of intertribal relationships and stable tribal delimitations within Cinchonoideae s.s. (Rubiaceae). Mol Phylogenet Evol. 2010;56:21–39.

10. Rusby HH. The Genus *Cinchona* in Bolivia. Bulletin of the Torrey Botanical Club. 1931. p. 523.

11. Council OFE. European pharmacopoeia. v. 1. Strasbourg, France: Council of Europe. 2016;

12. Eyal S. The Fever Tree: from Malaria to Neurological Diseases. Toxins. 2018;10.

13. IUCN Red List of threatened species. Choice . American Library Association; 2005;43:43–2185 – 43–2185.

14. Jäger H, Tye A, Kowarik I. Tree invasion in naturally treeless environments: Impacts of quinine (*Cinchona pubescens*) trees on native vegetation in Galápagos. Biol Conserv. Elsevier; 2007;140:297–307.

15. Sullivan DJ. *Cinchona* Alkaloids: Quinine and Quinidine. In: Staines HM, Krishna S, editors. Treatment and Prevention of Malaria: Antimalarial Drug Chemistry, Action and Use. Basel: Springer Basel; 2012. p. 45–68.

16. Lau KH, Bhat WW, Hamilton JP, Wood JC, Vaillancourt B, Wiegert-Rininger K, et al. Genome assembly of *Chiococca alba* uncovers key enzymes involved in the biosynthesis of unusual terpenoids. DNA Res. Oxford Academic; 2020;27.

17. Denoeud F, Carretero-Paulet L, Dereeper A, Droc G, Guyot R, Pietrella M, et al. The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. Science. 2014;345:1181–4.

18. Tran HTM, Ramaraj T, Furtado A, Lee LS, Henry RJ. Use of a draft genome of coffee (*Coffea arabica*) to identify SNPs associated with caffeine content. Plant Biotechnol J. 2018;16:1756–66.

19. Park J, Kim Y, Xi H, Heo K-I. The complete chloroplast genome of coffee tree, *Coffea arabica* L. “Blue Mountain” (Rubiaceae). Mitochondrial DNA Part B. Taylor & Francis; 2019;4:2436–7.

20. London Calling: Live first-time sequencing of the Fever Tree - the plant that some say has saved millions of lives from malaria [Internet]. [cited 2021 Jul 8]. Available from: https://nanoporetech.com/about-us/news/london-calling-live-first-time-sequencing-fever-tree-plant-some-say-has-saved

21. Doležel J, Kubaláková M, Suchánková P, Kovářová P, Bartoš J, Šimková H. Chromosome Analysis and Sorting. Flow Cytometry with Plant Cells. 2007. p. 373–403.

22. Pellicer J, Powell RF, Leitch IJ. The Application of Flow Cytometry for Estimating Genome Size, Ploidy Level Endopolyploidy, and Reproductive Modes in Plants. Methods Mol Biol. 2021;2222:325–61.

23. Loureiro J, Rodriguez E, Doležel J, Santos C. Two New Nuclear Isolation Buffers for Plant DNA Flow Cytometry: A Test with 37 Species. Ann Bot. Oxford University Press; 2007;100:875.

24. Bennett MD, Smith JB. Nuclear DNA amounts in angiosperms. Philos Trans R Soc Lond B Biol Sci. 1976;274:227–74.

25. Andrews S, Others. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom; 2010.

26. Schubert M, Lindgreen S, Orlando L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res Notes. 2016;9:88.

27. MPI-MP CHLOROBOX - GeSeq [Internet]. [cited 2021 May 18]. Available from: https://chlorobox.mpimp-golm.mpg.de/geseq.html

28. Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, et al. GeSeq - versatile and accurate annotation of organelle genomes. Nucleic Acids Res. 2017;45:W6–11.

29. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997;25:955–64.

30. Laslett D, Canback B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. 2004;32:11–6.

31. Greiner S, Lehwark P, Bock R. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. Nucleic Acids Res. 2019;47:W59–64.

32. Schubert M, Ermini L, Sarkissian CD, Jónsson H, Ginolhac A, Schaefer R, et al. Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. Nature Protocols. 2014. p. 1056–82.

33. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.

34. De Coster W. NanoPlot [Internet]. Github; [cited 2021 May 18]. Available from: https://github.com/wdecoster/NanoPlot

35. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27:722–36.

36. Ruan J. SMARTdenovo: Ultra-fast *de novo* assembler using long noisy reads. Github Available at: https://github com/ruanjue/smartdenovo [Accessed January 10, 2019]. 2018;

37. Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 2017;27:737–46.

38. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.

39. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

40. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2.

41. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29:644–52.

42. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.

43. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, et al. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 2003;31:5654–66.

44. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 2006;34:W435–9.

45. Kent WJ. BLAT—The BLAST-Like Alignment Tool. Genome Res. 2002;12:656–64.

46. Baker WJ, Bailey P, Barber V, Barker A, Bellot S, Bishop D, et al. A Comprehensive Phylogenomic Platform for Exploring the Angiosperm Tree of Life. Syst Biol. 2021

47. Index of /pub/paftol [Internet]. [cited 2021 May 19]. Available from: http://sftp.kew.org/pub/paftol/

48. Johnson MG, Gardner EM, Liu Y, Medina R, Goffinet B, Shaw AJ, et al. HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. Appl Plant Sci. Wiley; 2016;4:1600016.

49. Li H. seqtk Toolkit for processing sequences in FASTA/Q formats. GitHub. 2012;767:69.

50. Katoh K, Kuma K-I, Toh H, Miyata T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 2005;33:511–8.

51. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312–3.

52. Junier T, Zdobnov EM. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. Bioinformatics. 2010;26:1669–70.

53. Brown JW, Walker JF, Smith SA. Phyx: phylogenetic tools for unix. Bioinformatics. 2017;33:1886–8.

54. Zhang C, Rabiee M, Sayyari E, Mirarab S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics. 2018;19:153.

55. Bellot S. scripts [Internet]. Github; [cited 2021 Jun 24]. Available from: https://github.com/sidonieB/scripts

56. Rambaut A. FigTree v1. 4. 2012. [cited 2021 Jun 24]. Available from: https://github.com/rambaut/figtree/releases

57. Yi D-K, Kim K-J. Complete chloroplast genome sequences of important oilseed crop *Sesamum indicum* L. PLoS One. 2012;7:e35872.

58. Ly SN, Garavito A, De Block P, Asselman P, Guyeux C, Charr J-C, et al. Chloroplast genomes of Rubiaceae: Comparative genomics and molecular phylogeny in subfamily Ixoroideae. PLoS One. 2020;15:e0232295.

59. Miller MA, Pfeiffer W, Schwartz T. The CIPRES science gateway: a community resource for phylogenetic analyses. Proceedings of the 2011 TeraGrid Conference: Extreme Digital Discovery. New York, NY, USA: Association for Computing Machinery; 2011. p. 1–8.

60. Zhou T, Wang J, Jia Y, Li W, Xu F, Wang X. Comparative Chloroplast Genome Analyses of Species in *Gentiana* section *Cruciata* (Gentianaceae) and the Development of Authentication Markers. Int J Mol Sci. 2018;19.

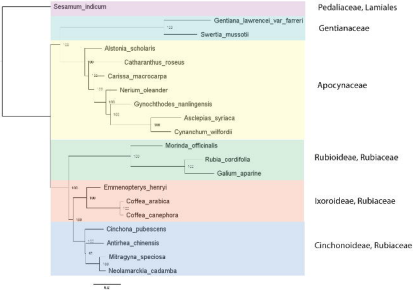
61. Zhang Y, Zhang J-W, Yang Y, Li X-N. Structural and Comparative Analysis of the Complete Chloroplast Genome of a Mangrove Plant: *Scyphiphora hydrophyllacea* Gaertn. f. and Related Rubiaceae Species. For Trees Livelihoods. Multidisciplinary Digital Publishing Institute; 2019;10:1000.

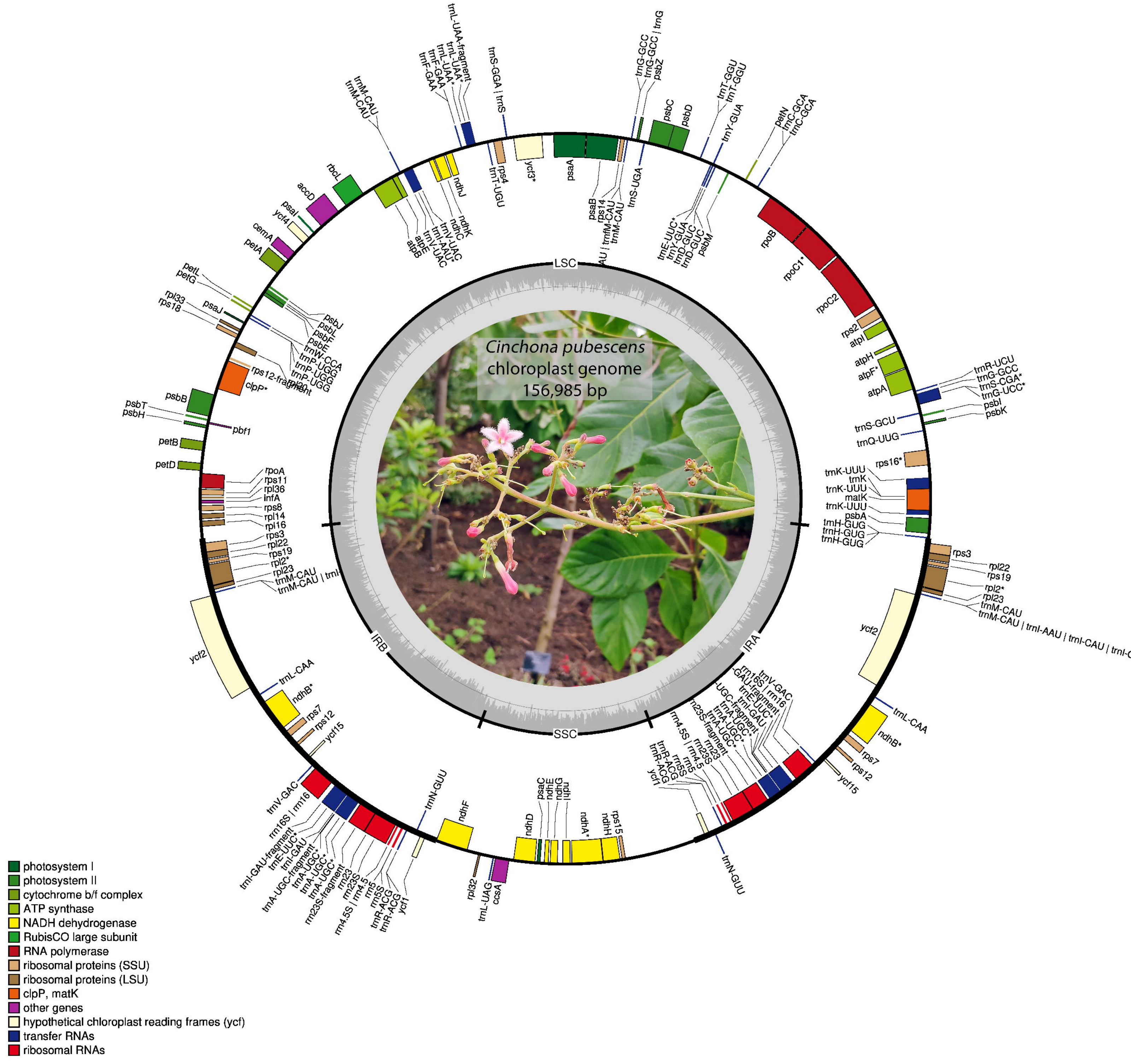
62. Robbrecht E, Manen J-F. The Major Evolutionary Lineages of the Coffee Family (Rubiaceae, Angiosperms). Combined Analysis (nDNA and cpDNA) to Infer the Position of Coptosapelta and Luculia, and Supertree Construction Based on rbcL, rps16, trnL-trnF and atpB-rbcL Data. A New Classification in Two Subfamilies, Cinchonoideae and Rubioideae. Syst Geogr Plants. National Botanic Garden of Belgium; 2006;76:85–145.

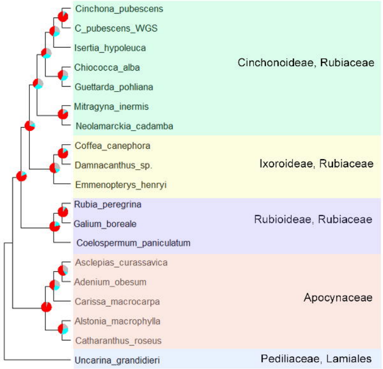
63. LaPierre N, Egan R, Wang W, Wang Z. *De novo* Nanopore read quality improvement using deep learning. BMC Bioinformatics. 2019;20:552.

64. Chen Y, Nie F, Xie S-Q, Zheng Y-F, Dai Q, Bray T, et al. Efficient assembly of nanopore reads via highly accurate and intact error correction. Nat Commun. 2021;12:60.

65. Maldonado C, Molina CI, Zizka A, Persson C, Taylor CM, Albán J, et al. Estimating species diversity and distribution in the era of Big Data: to what extent can we trust public databases? Glob Ecol Biogeogr. 2015;24:973–84.


Sesamum_indicum
Pedaliaceae, Lamiales
Gentiana_lawrencei_var_farreri
100
Swertia_mussotii
Gentianaceae
100
Alstonia_scholaris
100
Catharanthus_roseus
100
Carissa_macrocarpa
100
Nerium_oleander
100
Gynochthodes_nanlingensis
100
Asclepias_syriaca
100
Cynanchum_wilfordii
Apocynaceae
Morinda_officinalis
100
Rubia_cordifolia
100
Galium_aparine
Rubioideae, Rubiaceae
100
Emmenopterys_henryi
100
Coffea_arabica
100
Coffea_canephora
Ixoroideae, Rubiaceae
100
Cinchona_pubescens
100
Antirhea_chinensis
47
Mitragyna_speciosa
100
Neolamarckia_cadamba
Cinchonoideae, Rubiaceae


Cinchona pubescens
chloroplast genome
156,985 bp
LSC
SSC
IRA
IRB
trnL-UAA-fragment
trnL-UAA*
trnL-UAA
trnF-GAA
trnF-GAA
trnM-CAU
trnM-CAU
trnS-GGA | trnS
trnT-UGU
rps4
ycf3*
psaA
psaB
rps14
AU | trnfM-CAU
trnM-CAU
trnS-UGA
trnG-GCC
trnG-GCC | trnG
psbZ
psbC
psbD
trnT-GGU
trnT-GGU
trnY-GUA
trnE-UUC*
trnY-GUA
trnD-GUC
trnD-GUC
psbM
petN
trnC-GCA
trnC-GCA
rpoB
rpoC1*
rpoC2
rps2
atpI
atpH
atpF*
atpA
trnR-UCU
trnG-GCC
trnS-CGA*
trnG-UCC*
psbI
psbK
trnS-GCU
trnQ-UUG
rps16*
trnK-UUU
trnK
trnK-UUU
matK
trnK-UUU
psbA
trnH-GUG
trnH-GUG
trnH-GUG
rps3
rpl22
rps19
rpl2*
rpl23
trnM-CAU
trnM-CAU | trnI-AAU | trnI-CAU | trnI-C
ycf2
trnL-CAA
ndhB*
rps7
rps12
ycf15
trnV-GAC
rrn16S | rrn16
-GAU-fragment
trnE-UUC*
trnI-GAU
-UGC-fragment
trnA-UGC*
trnA-UGC*
rrn23S-fragment
rrn23S
rrn23
rrn4.5S | rrn4.5
rrn5S
rrn5
trnR-ACG
trnR-ACG
ycf1
trnN-GUU
rps15
ndhH
ndhA*
ndhI
ndhG
ndhE
psaC
ndhD
ccsA
trnL-UAG
rpl32
ndhF
trnN-GUU
ycf1
trnR-ACG
trnR-ACG
rrn5
rrn5S
rrn4.5S | rrn4.5
rrn23S-fragment
rrn23S
rrn23
trnA-UGC*
trnA-UGC*
trnA-UGC-fragment
trnI-GAU
trnE-UUC*
trnI-GAU-fragment
rrn16S | rrn16
trnV-GAC
ycf15
rps12
rps7
ndhB*
trnL-CAA
ycf2
trnM-CAU | trnI
trnM-CAU
rpl23
rpl2*
rps19
rpl22
rps3
rpl16
rpl14
rps8
infA
rpl36
rps11
rpoA
petD
petB
pbf1
psbH
psbT
psbB
clpP*
rps12-fragment
rpl20
trnP-UGG
trnP-UGG
trnW-CCA
psbE
psbF
psbL
psbJ
rps18
rpl33
psaJ
petG
petL
petA
cemA
ycf4
psaI
accD
rbcL
atpB
atpE
trnV-UAC
trnV-UAC*
trnM-CAU
ndhC
ndhK
ndhJ
🟩 photosystem I
🟩 photosystem II
🟩 cytochrome b/f complex
🟩 ATP synthase
🟨 NADH dehydrogenase
🟩 RubisCO large subunit
🟥 RNA polymerase
🟫 ribosomal proteins (SSU)
🟫 ribosomal proteins (LSU)
🟧 clpP, matK
🟪 other genes
⬜ hypothetical chloroplast reading frames (ycf)
🟦 transfer RNAs
🟥 ribosomal RNAs


Cinchona_pubescens
C_pubescens_WGS
Isertia_hypoleuca
Chiococca_alba
Guettarda_pohliana
Mitragyna_inermis
Neolamarckia_cadamba
Cinchonoideae, Rubiaceae
Coffea_canephora
Damnacanthus_sp.
Emmenopterys_henryi
Ixoroideae, Rubiaceae
Rubia_peregrina
Galium_boreale
Coelospermum_paniculatum
Rubioideae, Rubiaceae
Asclepias_curassavica
Adenium_obesum
Carissa_macrocarpa
Alstonia_macrophylla
Catharanthus_roseus
Apocynaceae
Uncarina_grandidieri
Pediliaceae, Lamiales