A highly contiguous, scaffold-level nuclear genome assembly for the Fever tree (*Cinchona pubescens* Vahl) as a novel resource for research in the Rubiaceae

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

[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, 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

Abstract

Background: The Andean Fever tree (*Cinchona* L.; Rubiaceae) is the iconic source of bioactive quinine alkaloids, which have been vital to treating malaria for centuries. *C. pubescens* Vahl, in particular, has been an essential source of income for several countries within its native range in north-western South America. However, an absence of available genomic resources is essential for placing the *Cinchona* species within the tree of life and setting the foundation for exploring the evolution and biosynthesis of quinine alkaloids.

Findings: We address this gap by providing the first highly contiguous and annotated nuclear and organelle genome assemblies for *C. pubescens*. 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. Our nuclear genome assembly comprises 603 scaffolds comprising a total length of 904 Mb, and the completeness represents ~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 moderately complete, displaying 83% of the BUSCO gene set and a small fraction of genes (4.6%) classified as fragmented. Additionally, we report *C. pubescens* plastome with a length of ~157 Kb and a GC content of 37.74%. We demonstrate the utility of these novel genomic resources by placing *C. pubescens* in the Gentianales order using additional plastid and nuclear datasets.

Conclusions: Our study provides the first genomic resource for *C. pubescens*, thus opening new research avenues, including the provision of crucial genetic resources for analysis of alkaloid biosynthesis in the Fever tree.

Keywords: Oxford Nanopore Technologies, Rubiaceae, RNA-seq, whole-genome sequencing, *Cinchona*, quinine

Data Description

1.1 Background

The Andes biodiversity hotspot hosts over 28,000 species [1], of which 3,805 benefit humanity [2]; unfortunately, nuclear genomic resources are only available for a limited number of such diversity (179 spp – Genomes NCBI database accessed on 26 May 2022). The fever tree (*Cinchona* L., Rubiaceae) is a genus comprising 24 species native to the Eastern slopes of the Andes mountain range in South America ([3,4]; Fig. 1) and one of the most economically important genera in the family, second only to coffee [5]. The genus is widely known as the source of at least 35 quinine alkaloids (quinolines), which alleviate the fever symptoms associated with malaria [6]. As such, fever trees have played a crucial role in the economies and livelihoods of people worldwide for centuries [7,8].

Despite this genus's tremendous historical and economic importance, DNA sequence datasets for *Cinchona* are relatively meagre, limited to 252 DNA Sanger sequences available in the NCBI repository (accessed on May 17, 2021; [9]). 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, specifically of the Cinchonoideae subfamily where the Cinchoneae tribe is, are based on just a handful of nuclear (ITS) and plastid (*mat*K, *rcb*L, *rps*16, *trn*L-F) data sets. 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 [10,11] (including the genus *Cinchona*, which shows very unclear relationships). Furthermore, studies that examine the relationships between species of this genus are equally scarce [7,8]. A

recent genome-wide phylogenetic tree for the order Gentianales [12] provided strong support for *C. pubescens* as a 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 [13][14]. However, several species in the genus 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 [15,16]. Among them, *C. pubescens* or red Cinchona bark is now widely cultivated throughout the tropics, with some instances where the species has escaped cultivation and become invasive [17]. Extensive research on the structure, abundance, and chemical composition of quinine alkaloids in *Cinchona* has been conducted [18], revealing the further potential for novel drug discovery. However, the identity of the genes involved in the synthetic pathway of quinine alkaloids remains elusive.

Nuclear genome assemblies are critical to our understanding of the origin and domestication of useful plants and are a cornerstone resource for breeders [19–21]. 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 2*n*=34. The assemblies were generated using a combination of extensive long-read Nanopore (~218x) and short-read Illumina paired-end 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 [22] than in the three previously published genome assemblies in Rubiaceae, namely for *Chiococca alba [22]*, *Coffea canephora* [23], and *Coffea arabica [24]*. 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 [22,25]. 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 and RNA sequencing

We sampled leaves from a single *Cinchona pubescens* individual propagated vegetatively from a tree collected in Tanzania in 1977 and cultivated in the Temperate House of the Royal Botanic Gardens, Kew (RBG Kew), UK (Accession Number 1977-69; 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 assessing 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 the paired-end 150 bp libraries using the HiSeq X Ten chemistry. For transcriptome library preparations, total RNA was extracted from 1000 mg of frozen-ground leaf, young bract, mature bract, flower anthesis, flower bud (older), flower bud (young), leaf bud and young leaf 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 [26]. 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 using the 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. DNA libraries were prepared using the ONT Ligation Sequencing Kit (SQK-LSK109, Oxford Nanopore Technologies, UK). During sequencing on the PromethION platform, re-loads were performed when required. Though yield was slightly lower in sequencing for these re-loaded 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 [27], with modifications as described in Pellicer et al. [28]. Freshly collected tissue from the same individual sampled for DNA and RNA sequencing was measured together with *Oryza sativa* L. 'IR-36' as the calibration standard using general-purpose buffer" (GPB) [29] supplemented with 3% PVP-40 and β-mercaptoethanol [28]. 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 were analysed 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*) [30] and resulted in a 1C-value of 1.1 Gb. Additionally, using the full Illumina short-read dataset, we additionally implemented a k-mer counting method to

characterise the genome in Jellyfish v.2.2.10 [31] setting a kmer size of 21. We used GenomeScope to visualise the kmer plot [32]. However, we did not deem the kmer counting method sufficiently accurate for genome size estimation but reported the estimated genome-wide heterozygosity rates output by GenomeScope as 0.869-0.889%.

1.4 Short read data processing of the chloroplast genome assembly

Sequencing of the DNA Illumina library generated 428M paired-end reads, representing 128.4 Gb of raw data. RNA sequencing produced 385M paired-end reads, representing 115.5 Gb (Table 1). The quality of the raw reads was assessed using the FastQC software [33], and quality trimming was conducted using the software AdapterRemoval2 v.2.3.1 [34]. Here, bases with Phred score quality <30 and read lengths <50 bp were removed together with adapter sequences. The final short-read dataset was 131 Gb and contained 384,626,011 paired reads, which corresponds to an estimated 464.8x coverage (based on the genome size of 1.1 Gb/1C, see *Estimation of genome size*).

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 values are very similar to those reported for the *Coffea arabica* plastid genome, which is reported to be 155,189 bp in length and has a GC content of 37.4% [25].

We annotated the plastid genome assembly of *C. pubescens* in CHLOROBOX [35], which implements GeSeq [36], tRNAscan-SE v2.0.5 [37], and ARAGORN v1.2.38 [38]. 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 [39] (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 [40], where we used BWA [41] for alignment, specifying the backtrack algorithm, and filtering minimum quality equal to zero to maximise recovery. After PCR duplicate filtering, the coverage of unique hits was 7,960x.

1.5 Long-read nuclear genome assembly, quality assessment and ploidy levels

The quality and quantity of the PromethION sequencing output conducted across four flow cells were evaluated in NanoPlot v.1.82 [42] 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 [43] 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 [44],

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 of 2,783,363 bp, representing ~904 Mb (~82% of the genome size; Tab. 1). Lastly, a round of scaffold correction was implemented in RACON v.1.4.3 [45] 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 [46] 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 alignment file (SAM) 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 [47]; and ii) estimating the completeness of the genome as implemented in the software BUSCO v.5.2 and using the viridiplantae_odb10 [48]. 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 92.4% conserved eudicot genes, of which 87.5% were single copy, 4.9% duplicated, and 5.6% fragmented. The remaining BUSCO genes were labelled as missing (2%). Taken together, our results suggest that our nuclear genome assembly presents high contiguity and quality with high completeness.

Lastly, to evaluate the ploidy levels of *C. pubescens* through the newly assembled genome, we computed allele frequencies from reads mapped against two scaffolds, "utg 230" and "utg2" derived from our genome assembly, covering 9,568,509 bp (~106x coverage) and 14,628,764 bp (~103x coverage), respectively. The reads were obtained from the mapping procedure conducted to assess the quality of the corrected nuclear genome assembly (see above). We relied on the software ploidyNGS to compute allele frequencies [49], using the –g option (i.e. guess ploidy levels), a maximum read depth of 100 option (-d 100) and a maximum allele frequency of 0.95 (-m 0.95). Our analysis revealed that the genome of *C. pubescens* is diploid (Fig. 3) as inferred by the comparison of Kolmogorov-Smirnoff distances between the allele frequencies computed from our read mappings and those derived from simulated data [49].

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 [50] 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 [51] 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 [52], 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.5.12 and the representative plant set viridiplantae_odb10, which currently includes 72

species, of which 56 are angiosperms. Our assembled transcriptome captured 92.7% (394/425) of the BUSCO set as complete genes, of the remainder, 3.1% of the genes were fragmented, and 4.2% 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 [53] for a combination approach of *ab initio* and transcript evidence-based on RNA-seq data. As AUGUSTUS considered the transcripts' evidence as Expressed Sequence Tag (EST), we first generated hints from the transcriptome data by aligning the transcripts to the genome using BLAT v 3.5 [54]. Then, we set the hint parameters to rely on the hints and anchor the gene structure. We predicted 72,305 CDSs using the hints and tomato (*Solanum lycopersicum* L.) as the reference species. The completeness of the CDSs was estimated using BUSCO v.5.12 and vidiriplantae odb10 [48] as reference: 68.4% represented complete BUSCOs, 63.5% were single-copy, 26% were fragmented and 5.6% were missing. We summarized the metrics of the CDSs statistics with GenomeQC (Table 1) [55]

1.7 Nuclear and plastid phylogenomics of *Cinchona*

We verified the nuclear genome's phylogenetic placement using the reference sequences of the 353 low-copy nuclear genes that are conserved across angiosperms from the Plant and Fungal Trees of Life project [56]. Here, we sampled the gene sequences for the 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 [57] hosted by the Royal Botanic Gardens, Kew. To include the *C. pubescens* of this study in the analysis of the 353 low copy nuclear genes of selected Gentianales, we then retrieved these genes from our RNA-seq data using the pipeline HybPiper v.1.3.1 [58]. 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 [59]. The gene sequences produced by HybPiper

were aligned with the data for 19 selected Gentianales species using MAFFT v7.453 [60] and then they were concatenated into a supermatrix for phylogenomic analyses.

We implemented the maximum likelihood approach using RAxML-HPC V.8 [61] with a GTRGAMMA substitution model for each gene and a rapid bootstrap analysis with 500 replications. Then we filtered the bipartition trees that had >=20% support using Newick utilities [62]. The resulting trees were rooted using phyx v1.2.1 [63], 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 shows the support of the gene trees for the main topology, q2 shows the support for the first alternative topology, and q3 shows the support for the second alternative topology [64]. We incorporated these scores into the species tree with an R script [65]. All trees were visualized with FigTree v.1.4.4 [66].

In the nuclear phylogenomic tree resulting from the 353 low copy nuclear genes (Fig. 3), *C. pubescens* clusters within the Cinchonoideae, which is more closely related with the Ixoroideae group than with Rubioideae. Most nodes are 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 [67]. 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 show the tripartite structure [68]. 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 (ML) 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 [69].

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 (16 had 100% support and all but one node had 100% support). 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 [70,71], 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. Additionally, the ingroup showed that the Cinchonoideae and Ixoroideae subfamilies are sisters and form a clade while Rubioideae is placed as sister to this clade.

The placement of *C. pubescens* in the Cinchonoideae subfamily cluster using both the plastid and nuclear data presented in this study is consistent with previous taxonomic and phylogenetic studies [72] and gives support to the robustness of the assembled nuclear and plastid genomes. 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 [73,74].

2. Conclusion

Using a combination of extensive short and long-read DNA datasets, we deliver the first highly contiguous and robust nuclear and plastid genome assemblies for one of the historically most traded and economically important *Cinchona* species, *C. pubescens*. The

abundant genomic resources provided here 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 identify the genes involved in the biosynthetic pathways of quinine alkaloids synthesis, identify the underpinning genetic diversity of these genes both between and within species, and open doors on how the expression of these genes is regulated. Our nuclear scaffold-level and plastid genome assembly will enable future reference-guided assemblies, variant calling, and gene annotation to enhance functional analysis within the *Cinchona* genus, with the 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 sequence data, and nuclear and plastid assemblies are available at the NBCI repository, under the BioProject number PRJNA768351.

Competing interests

The authors declare that they have no competing interests.

Funding

NR, AA, CB, and NC 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 and the Royal Botanic Gardens, Kew.

Author contributions

IJL, OAPE, NR, MT and AA conceived the study. MN and OAPE collected plant tissue. OAPE, IJL, RFP, MT and AA generated datasets. OAPE, NC, CK and MT conducted in-silico analyses. NC and OAPE 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 and Claes Persson for taxonomic advice. Two anonymous reviewers and the associate editor provided constructive feedback to this manuscript.

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 [75].

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. **A**. The coalescent-based species tree estimation of the Gentianales order is inferred from low copy nuclear gene trees. Pie charts positioned on the nodes represent the percentage of the gene tree quartets that agree with the topology of the main species tree (blue) and the

other two alternative gene tree quartets (orange: second child [R], sister group [S] | first child [L], any other branch [O]; and grey: RO|LS). The genomic data of *Cinchona pubescens* terminal marked with "*" was newly produced in this study. **B**. Phylogenetic tree showing the relationships of twenty Gentianales species built from the whole plastid genome. Unless shown, numbers on the branches represent Likelihood Bootstrap Percentages of 100. The coloured boxes indicate the subfamily/family that sampled terminals belong to. Allele frequencies for the fourth most common alleles, derived from reads mapped against the scaffolds "utg230" (C) and "utg2" (D). Note that frequency distributions of the first and second most frequent alleles are either skewed towards 0 or 1, denoting homozygote variants, and 0.5, denoting heterozygote, diploid variants. (Inset): Frontal (left) and lateral view (right) of the sequenced specimen (ID 1977-69) of *Cinchona pubescens* (photos: O. A. Pérez-Escobar).

Table 1. Summary table for the Illumina WGS and RNA-Seq libraries.

Table 2. Summary assembly statistics for *C. pubescens* using SMARTdenovo and RACON.

Table 3. Annotation metrics summary statistics for the nuclear annotation of *C. pubescens*.

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. Pérez-Escobar OA, Zizka A, Bermúdez MA, Meseguer AS, Condamine FL, Hoorn C, et al.. The Andes through time: evolution and distribution of Andean floras. *Trends Plant Sci*. 27:364–782022;

2. Gori B, Ulian T, Bernal HY, Diazgranados M. Understanding the diversity and biogeography of Colombian edible plants. *Sci Rep*. 12:78352022;

3. Andersson L. A revision of the genus Cinchona (Rubiaceae-Cinchoneae). *Memoirs-New York botanical garden*. NYBG NEW YORK BOTANICAL GARDEN; 1998;

4. Maldonado C, Persson C, Alban J, Antonelli A, Rønsted N. Cinchona anderssonii (Rubiaceae), a new overlooked species from Bolivia. *Phytotaxa*. Magnolia Press; :203–82017;

5. Steere WC. The Cinchona-Bark Industry of South America. *Sci Mon*. American Association for the Advancement of Science; 61:114–261945;

6. Kacprzak KM. Chemistry and biology of Cinchona alkaloids. *Nat Products Bioprospect*. Springer-Verlag: Berlin; :605–412013;

7. Lee MR. Plants against malaria. Part 1: Cinchona or the Peruvian bark. *J R Coll Physicians Edinb*. 32:189–962002;

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

9. : Home - Nucleotide - NCBI. https://www.ncbi.nlm.nih.gov/nuccore Accessed 2021 May 17.

10. 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); 54:17–282005;

11. Manns U, Bremer B. Towards a better understanding of intertribal relationships and stable tribal delimitations within Cinchonoideae s.s. (Rubiaceae). *Mol Phylogenet Evol*. 56:21–392010;

12. Antonelli A, Clarkson JJ, Kainulainen K, Maurin O, Brewer GE, Davis AP, et al.. Settling a family feud: a high-level phylogenomic framework for the Gentianales based on 353 nuclear genes and partial plastomes. *Am J Bot*. 108:1143–652021;

13. Rusby HH. The Genus Cinchona in Bolivia. Bulletin of the Torrey Botanical Club.

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

15. Eyal S. The Fever Tree: from Malaria to Neurological Diseases. *Toxins* . 2018; doi: 10.3390/toxins10120491.

16. . IUCN Red List of threatened species. *Choice* . American Library Association; 43:43–2185 – 43–21852005;

17. 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; 140:297–3072007;

18. 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; p. 45–68.

19. Renner SS, Wu S, Pérez-Escobar OA, Silber MV, Fei Z, Chomicki G. A chromosome-level genome of a Kordofan melon illuminates the origin of domesticated watermelons. *Proc Natl Acad Sci U S A*. 2021; doi: 10.1073/pnas.2101486118.

20. Al-Mssallem IS, Hu S, Zhang X, Lin Q, Liu W, Tan J, et al.. Genome sequence of the date palm Phoenix dactylifera L. *Nat Commun*. 4:22742013;

21. Shukla VK, Doyon Y, Miller JC, DeKelver RC, Moehle EA, Worden SE, et al.. Precise genome modification in the crop species Zea mays using zinc-finger nucleases. *Nature*. 459:437–412009;

22. 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; doi: 10.1093/dnares/dsaa013.

23. 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*. 345:1181–42014;

24. 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*. 16:1756–662018;

25. 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; 4:2436–72019;

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

27. Doležel J, Kubaláková M, Suchánková P, Kovářová P, Bartoš J, Šimková H. Chromosome Analysis and Sorting. Flow Cytometry with Plant Cells.

28. 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*. 2222:325–612021;

29. 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; 100:8752007;

30. Bennett MD, Smith JB. Nuclear dna amounts in angiosperms. *Philos Trans R Soc Lond B Biol Sci*. 274:227–741976;

31. Marçais, Kingsford. JELLYFISH–fast, parallel k-mer counting for DNA. *Bioinformatics*.

32. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, et al.. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics*. 33:2202–42017;

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

34. Schubert M, Lindgreen S, Orlando L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res Notes*. 9:882016;

35. : MPI-MP CHLOROBOX - GeSeq. https://chlorobox.mpimp-golm.mpg.de/geseq.html Accessed 2021 May 18.

36. 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*. 45:W6–112017;

37. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*. 25:955–641997;

38. Laslett D, Canback B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res*. 32:11–62004;

39. Greiner S, Lehwark P, Bock R. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Res*. 47:W59–642019;

40. 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.

41. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 25:1754–602009;

42. De Coster W. NanoPlot. Github;

43. 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*. 27:722–362017;

44. 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;

45. Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res*. 27:737–462017;

46. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 34:3094–1002018;

47. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al.. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 25:2078–92009;

48. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 31:3210–22015;

49. Augusto Corrêa Dos Santos R, Goldman GH, Riaño-Pachón DM. ploidyNGS: visually exploring ploidy with Next Generation Sequencing data. *Bioinformatics*. 33:2575–62017;

50. 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*. 29:644–522011;

51. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al.. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 29:15–212013;

52. 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*. 31:5654–662003;

53. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res*. 34:W435–92006;

54. Kent WJ. BLAT—The BLAST-Like Alignment Tool. *Genome Res*. 12:656–642002;

55. Manchanda N, Portwood JL 2nd, Woodhouse MR, Seetharam AS, Lawrence-Dill CJ, Andorf CM, et al.. GenomeQC: a quality assessment tool for genome assemblies and gene structure annotations. *BMC Genomics*. 21:1932020;

56. 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; doi: 10.1093/sysbio/syab035.

57. : Index of /pub/paftol. http://sftp.kew.org/pub/paftol/ Accessed 2021 May 19.

58. 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; 4:16000162016;

59. Li H. seqtk Toolkit for processing sequences in FASTA/Q formats. *GitHub*. 767:692012;

60. Katoh K, Kuma K-I, Toh H, Miyata T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res*. 33:511–82005;

61. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 30:1312–32014;

62. Junier T, Zdobnov EM. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. *Bioinformatics*. 26:1669–702010;

63. Brown JW, Walker JF, Smith SA. Phyx: phylogenetic tools for unix. *Bioinformatics*. 33:1886–82017;

64. Zhang C, Rabiee M, Sayyari E, Mirarab S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*. 19:1532018;

65. Bellot S. scripts. Github;

66. Rambaut A. FigTree v1. 4.

67. Yi D-K, Kim K-J. Complete chloroplast genome sequences of important oilseed crop Sesamum indicum L. *PLoS One*. 7:e358722012;

68. 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*. 15:e02322952020;

69. 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; p. 1–8.

70. 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; doi: 10.3390/ijms19071962.

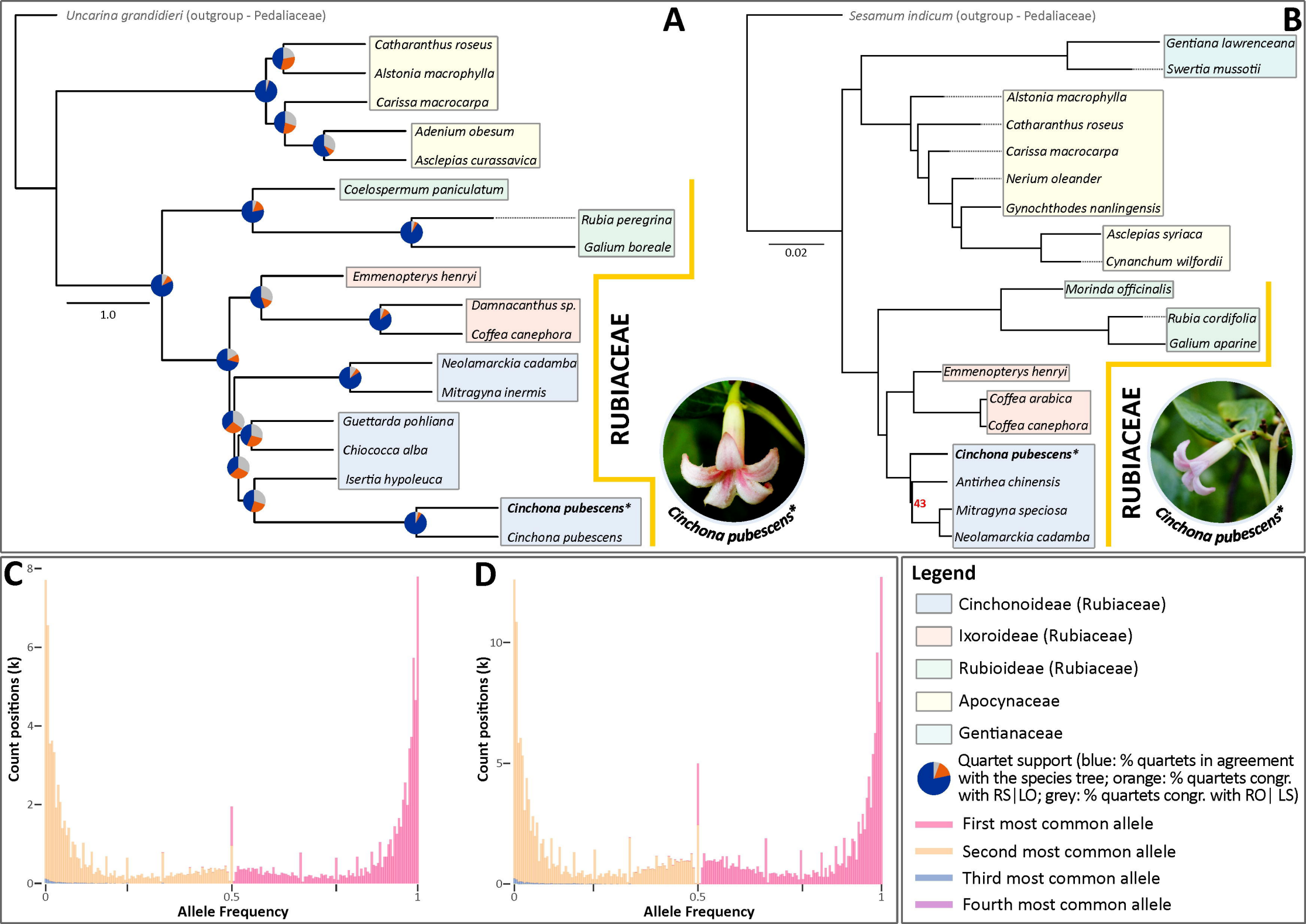
71. Zhang Y, Zhang J-W, Yang Y, Li X-N. Structural and Comparative Analysis of the Complete Chloroplast Genome of a Mangrove Plant: Scyphiphora hydrophyllacea Gaertn. f. and Related Rubiaceae Species. *For Trees Livelihoods*. Multidisciplinary Digital Publishing Institute; 10:10002019;

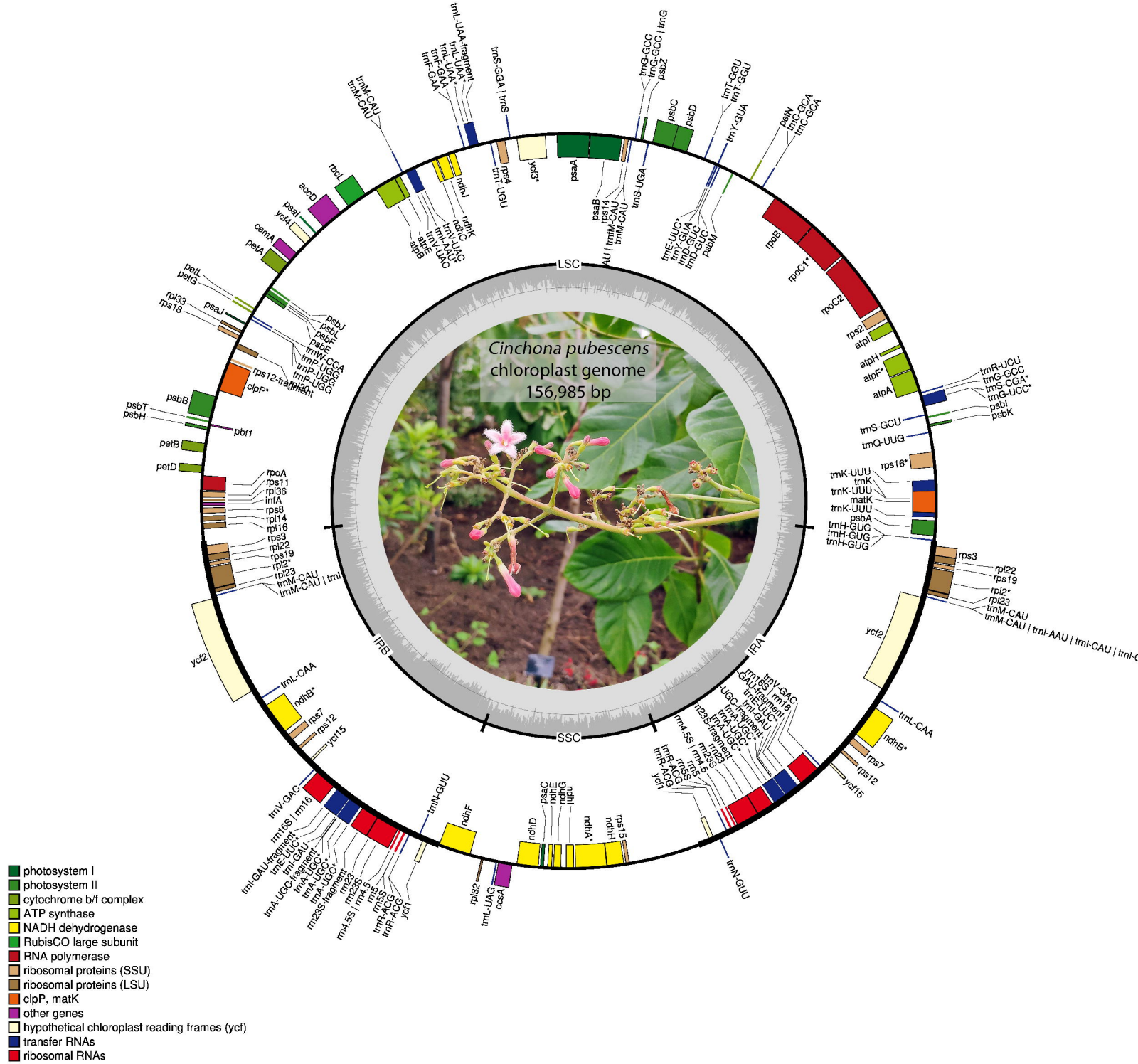
72. Robbrecht E, Manen J-F. The Major Evolutionary Lineages of the Coffee Family (Rubiaceae, Angiosperms). Combined Analysis (nDNA and cpDNA) to Infer the Position of Coptosapelta and Luculia, and Supertree Construction Based on rbcL, rps16, trnL-trnF and atpB-rbcL Data. A New Classification in Two Subfamilies, Cinchonoideae and Rubioideae. *Syst Geogr Plants*. National Botanic Garden of Belgium; 76:85–1452006;

73. LaPierre N, Egan R, Wang W, Wang Z. De novo Nanopore read quality improvement using deep learning. *BMC Bioinformatics*. 20:5522019;

74. 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*. 12:602021;

75. 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*. 24:973–842015;


A
Uncarina grandidieri (outgroup - Pedaliaceae)
Catharanthus roseus
Alstonia macrophylla
Carissa macrocarpa
Adenium obesum
Asclepias curassavica
Coelospermum paniculatum
Rubia peregrina
Galium boreale
Emmenopterys henryi
Damnacanthus sp.
Coffea canephora
Neolamarckia cadamba
Mitragyna inermis
Guettarda pohliana
Chiococca alba
Isertia hypoleuca
Cinchona pubescens*
Cinchona pubescens
1.0
RUBIACEAE
Cinchona pubescens*
B
Sesamum indicum (outgroup - Pedaliaceae)
Gentiana lawrenceana
Swertia mussotii
Alstonia macrophylla
Catharanthus roseus
Carissa macrocarpa
Nerium oleander
Gynochthodes nanlingensis
Asclepias syriaca
Cynanchum wilfordii
0.02
Morinda officinalis
Rubia cordifolia
Galium aparine
Emmenopterys henryi
Coffea arabica
Coffea canephora
Cinchona pubescens*
Antirhea chinensis
Mitragyna speciosa
Neolamarckia cadamba
43
RUBIACEAE
Cinchona pubescens*
C
Count positions (k)
Allele Frequency
D
Count positions (k)
Allele Frequency
Legend
Cinchonoideae (Rubiaceae)
Ixoroideae (Rubiaceae)
Rubioideae (Rubiaceae)
Apocynaceae
Gentianaceae
Quartet support (blue: % quartets in agreement with the species tree; orange: % quartets congr. with RS|LO; grey: % quartets congr. with RO| LS)
First most common allele
Second most common allele
Third most common allele
Fourth most common allele


Cinchona pubescens
chloroplast genome
156,985 bp
LSC
IRA
SSC
IRB
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

biosample_accession	Library_ID	Tissue	Source	Technology	Yield (Gb)	Total Number of reads (million)	%GC
SAMN22031859	CP9014LG	Leave	Genomic	Illumina HiSeq X	128.4	428	42
SAMN22031859	CP9014L	Leave	Transcriptomic	Illumina HiSeq X	38	101.6	42
SAMN22031859	CP9014YL	Young leaf	Transcriptomic	Illumina HiSeq X	32.6	87.1	46
SAMN22031859	CP9014LB	Leaf bud	Transcriptomic	Illumina HiSeq X	NA	NA	NA
SAMN22031859	CP9014Y	Young bract	Transcriptomic	Illumina HiSeq X	NA	NA	NA
SAMN22031859	CP9014B	Mature bract	Transcriptomic	Illumina HiSeq X	10.3	83.32	42
SAMN22031859	CP9014F	Flower in anthesis	Transcriptomic	Illumina HiSeq X	17.3	93.28	43
SAMN22031859	CP9014FBO	Flower bud - older	Transcriptomic	Illumina HiSeq X	18.6	127.9	43
SAMN22031859	CP9014FB	Flower bud - young	Transcriptomic	Illumina HiSeq X	17.9	103.8	43

	SMARTdenovo	RACON
Size_includeN	903037179	9.04E+08
Size_withoutN	903037179	9.04E+08
Seq_Num	603	603
Mean_Size	1497574	1499914
Median_Size	801066	802662
Longest_Seq	14628764	14747124
Shortest_Seq	33171	25882
GC_Content (%)	33.17	33.07
N50	2783363	2802128
L50	93	92
N90	682446	684435
Gap (%)	0	0

Number of gene models	72305
Minimum gene length	275
Maximum gene length	65857
Average gene length	4579.5
Number of exons	339698
Average number of exons per gene model	4.7
Average exon length	298.3
Number of transcripts	72305
Average number of transcripts per gene model	1
Number of gene models less than 200bp length	0