Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Genome assembly and functional annotation of capirona (Calycophyllum spruceanum): genomic insights into an Amazonian timber tree

Amelia Menacho
Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas (UNTRM)
Carla L. Saldaña
Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas (UNTRM)
Cristian Mendoza
Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas (UNTRM)
Wilian Salazar
Instituto Nacional de Innovación Agraria
Claudia Machicado
Universidad Peruana Cayetano Heredia
Carlos I. Arbizu
carlos.arbizu@untrm.edu.pe

Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas (UNTRM)

Research Article

Keywords: Capirona, Forest genetic resources, Genome survey, Microsatellites, Peruvian Amazon, Short-read sequencing

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11579457/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Calycophyllum spruceanum (Benth.) Hook.f. ex K.Schum. (Rubiaceae) is an Amazonian timber tree whose bark is used in traditional medicine, but genomic resources for the species are scarce. We report the first draft nuclear genome of the species, assembled from 323.8 million Illumina read pairs of a single tree (~ 185× coverage after trimming). k-mer analysis estimated a haploid genome size of 499–503 Mb and a heterozygosity of approximately 1%. After removal of organellar and microbial sequences, the final draft assembly spans 506.9 Mb in 72,381 scaffolds (N50 = 11.5 kb), with a GC content of 34.2%. It contains 83.9% of conserved single-copy eudicot genes (BUSCO) as complete copies; 19% of these genes are duplicated, consistent with retained allelic copies. Repeats cover 40.3% of the assembly, being mostly unclassified elements and LTR retrotransposons, with Gypsy elements about three times more abundant than Copia. We identified 266,057 simple sequence repeats (SSRs; 529.0 per Mb), a higher density than in *Coffea canephora* Pierre ex A.Froehner and *Emmenopterys henryi* Oliv., with AT-rich motifs predominating. Homology-based annotation from three *Coffea* genomes predicted 36,690 protein-coding genes, 97.1% of which received at least one homology-based functional annotation. This draft is the first nuclear reference for the genus *Calycophyllum* and a source of candidate SSR loci for the conservation and breeding of capirona genetic resources.

Introduction

Calycophyllum spruceanum (Benth.) Hook.f. ex K.Schum., known as capirona in Peru, is a canopy tree of the coffee family (Rubiaceae, subfamily Dialypetalanthoideae, formerly Ixoroideae; Razafimandimbison and Rydin 2024) native to the Amazon basin, from Bolivia, Peru and Colombia to the lower Amazon in Brazil (Tauchen et al. 2011). It is mainly a riverine species (Russell et al. 1999) and it is characterized by its dense and durable wood, which is used as lumber and building material (Tauchen et al. 2011). Its bark, shed every year, is used by Amazonian peoples such as the Shipibo-Conibo and Ese Eja to treat skin infections and wounds, which gave the species the name “tree of youth” (Peixoto et al. 2018).

Capirona has also been included in on-farm tree improvement in the Peruvian Amazon, where farmers selected mother trees for provenance and progeny trials (Weber et al. 2009). These trials revealed significant variation among provenances and especially among families in growth and wood density, with wood density more heritable than height or diameter (Sotelo Montes et al. 2006). Molecular knowledge of this variation, however, remains limited. Amplified Fragment Length Polymorphism (AFLP) markers showed that most genetic variation lies within populations along Peruvian tributaries of the Amazon (Russell et al. 1999). Fingerprints of the internal transcribed spacer (ITS) suggested that morphological diversity exceeds genetic diversity (Tauchen et al. 2011), and Random Amplified Polymorphic DNA (RAPD) markers separated four genetic groups in the eastern Peruvian Amazon (Saldaña et al. 2021). The complete plastome is also available (Saldaña et al. 2022). Target-enrichment sequencing with the Angiosperms353 nuclear gene panel has also been performed for one individual of *C. spruceanum* (Baker et al. 2022; ENA accession ERX7192628). None of these resources provides a genome-wide set of codominant nuclear markers for estimating heterozygosity, gene flow or parentage, and no nuclear genome assembly has been reported for the genus.

This gap reflects a broader bias in plant genomics. By January 2021, genome assemblies were available for 798 land plant species, about three quarters of them produced by teams in China, the USA and Europe, and only nine led by South American institutions (Marks et al. 2021). Well-curated genomes of Neotropical forest trees are still rare; the firstone was that of *Handroanthus impetiginosus* (Mart. ex DC.) Mattos (Bignoniaceae; Silva-Junior et al. 2018). Within Rubiaceae, nuclear genomes have been published mainly for crops and medicinal species, such as coffee

(Denoeud et al. 2014; Salojärvi et al. 2024), *Gardenia jasminoides* J.Ellis (Xu et al. 2020), kratom (*Mitragyna speciosa* Korth.; Brose et al. 2021; Pootakham et al. 2022) and the fever tree *Cinchona pubescens* Vahl (Allasi Canales et al. 2022). None of them is an Amazonian timber tree.

Illumina short reads alone cannot produce chromosome-level assemblies, but they provide a low-cost first reference. Short-read genome surveys and draft assemblies have been used to estimate genome features and to obtain SSR markers for genetic resources such as *Gardenia jasminoides* and a Peruvian landrace of *Capsicum chinense* Jacq. (Wang et al. 2021b; Estrada et al. 2024). Such drafts face three challenges: repeats fragment the assembly (Jiao and Schneeberger 2017), heterozygosity can split alleles into separate sequences (Kajitani et al. 2014), and contaminating sequences are common in public eukaryotic assemblies (Steinegger and Salzberg 2020).

Here we report the first draft nuclear genome of *C. spruceanum*, assembled from Illumina paired-end reads of the same tree used to assemble its plastome (Saldaña et al. 2022). Our objectives were to (i) estimate genome size and heterozygosity from k-mers; (ii) assemble the nuclear genome and screen it for organellar and microbial sequences; (iii) characterize its repetitive DNA and SSRs and compare SSRs with those of *Coffea canephora* Pierre ex A.Froehner and *Emmenopterys henryi* Oliv.; and (iv) predict and functionally annotate its protein-coding genes by homology. These resources provide a basis for developing codominant markers and for future genomic studies of capirona genetic resources.

Materials and methods

Sample collection and DNA extraction

A single tree of *C. spruceanum* growing at the San Bernardo Experimental Agricultural Station (EEA San Bernardo) of the National Institute for Agrarian Innovation (INIA), Tambopata, Madre de Dios, Peru, was sampled; the same individual was used for chloroplast genome assembly (Saldaña et al. 2022). A voucher specimen (USM 324323) was deposited in the Herbarium San Marcos (USM), Universidad Nacional Mayor de San Marcos. Genomic DNA was isolated from fresh leaves following a modified CTAB procedure (Doyle and Doyle 1987), its integrity was checked by electrophoresis on 1% agarose, and its concentration was measured fluorometrically (Qubit™ 4, Invitrogen). A Nextera XT library (Illumina) was sequenced on an Illumina HiSeq 2500 platform as 2 × 150 bp paired-end reads.

Read processing and genome size estimation

Read quality was inspected with FastQC v0.12.1 (Andrews 2014). Adapters and low-quality bases were removed with fastp v0.26.0 (Chen et al. 2018) using automatic paired-end adapter detection, sliding-window trimming of both read ends (4-bp window, mean Phred ≥ 20) and a minimum read length of 30 bp (full settings in Online Resource 1). Canonical k-mers were counted with Jellyfish v2.3.1 (Marçais and Kingsford 2011) at k = 17, 21 and 31, and the frequency histograms were fitted with GenomeScope 2.0 (Ranallo-Benavidez et al. 2020) under a diploid model (p = 2). The k value yielding the most consistent profile was used for comparisons with the assembly.

De novo genome assembly

The trimmed reads were assembled with ABySS v2.3.10 (Jackman et al. 2017) in Bloom filter mode (B = 50G; q = 15; other parameters at default values) at k = 71, 79, 87, 95 and 99, and the scaffolds produced by abyss-pe were used in all analyses. The k value yielding the highest proportion of complete single-copy BUSCOs (see Assembly quality assessment) was retained.

Assembly quality assessment

Contiguity was assessed with QUAST v5.3.0 (Gurevich et al. 2013), on sequences ≥ 3 kb (--large) during k optimization and on all scaffolds of the final assembly; contig statistics were computed after splitting scaffolds at runs of ≥ 10 N (Online Resource 1). Gene-space completeness was estimated with BUSCO v6.0.0 (Manni et al. 2021; Tegenfeldt et al. 2025) in genome mode with miniprot (Li 2023) against eudicotyledons_odb12 (n = 2,805), the most specific dataset available for Rubiaceae; viridiplantae_odb12 and embryophyta_odb12 were also evaluated. Base accuracy (QV) and k-mer completeness were estimated with Merqury v1.3 (Rhie et al. 2020) from a Meryl v1.4.1 database (k = 21) of the trimmed reads, and Merqury copy-number spectra were used to follow the fate of heterozygous and homozygous-peak read k-mers. Read depth per sequence was obtained by mapping the trimmed reads to the k = 87 assembly with Bowtie2 v2.4.5 (end-to-end mode; Langmead and Salzberg 2012) and SAMtools v1.21 coverage (Li et al. 2009).

Contamination screening and organellar filtering

Vector sequences were detected with BLASTn v2.16.0 (Altschul et al. 1990) against NCBI UniVec (accessed 21 August 2024; identity ≥ 90%, e-value ≤ 0.01). All sequences were searched against NCBI nr (downloaded 9 July 2026; Sayers et al. 2025) with DIAMOND v2.1.23 in blastx mode (--sensitive; e-value ≤ 1×10^{-25}; --max-target-seqs 5; Buchfink et al. 2021), and hits were assigned to lineages with NCBI Taxonomy. In a first pass, sequences without Viridiplantae hits and with mean depth < 5× were removed (Laetsch and Blaxter 2017). Organellar sequences were identified with minimap2 v2.28 (-x asm20; Li 2018) against the *C. spruceanum* plastome (OK326865.1) and the mitochondrial genomes of two Rubiaceae, *Scyphiphora hydrophylacea* C.F.Gaertn. (NC_057654.1) and *Coffea arabica* L. (OL789880.1–OL789881.1); sequences aligned over ≥ 90% of their length were removed.

Length filtering, refined contamination screening and final draft assembly

Minimum lengths of 500, 1,000 and 3,000 bp were applied with SeqKit v2.10.1 (Shen et al. 2024), and the filtered sets were compared by BUSCO completeness, QV, k-mer completeness and copy-number spectra; the threshold retained was the one that removed most short, low-depth sequences without loss of gene-space completeness or of homozygous-peak read k-mers. The length-filtered set was then re-screened by combining the taxonomic assignment of all DIAMOND hits with GC content and read depth (Laetsch and Blaxter 2017). Thresholds were derived from the sequences with Viridiplantae hits, 96.1% of whose length had 30–40% GC, and five rules (Online Resource 1) removed sequences with non-plant, viral, predominantly bacterial or no hits combined with GC content above the plant range. Sequences removed as bacterial or fungal were analysed with BUSCO against bacteria_odb12, alphaproteobacteria_odb12, methylobacteriaceae_odb12 and dothideomycetes_odb12 to test whether they represented co-assembled microbial genomes. The remaining sequences, renamed with simplified identifiers, constitute the final draft assembly.

Repeat annotation

A species-specific repeat library was built de novo with RepeatModeler v2.0.7 (Flynn et al. 2020), including structure-based LTR retrotransposon discovery (-LTRStruct; LTRharvest, Ellinghaus et al. 2008; LTR_retriever, Ou and Jiang 2018) and without Dfam or Repbase entries, and the final draft assembly was soft-masked with RepeatMasker v4.1.5 (Smit et al. 2013) using this library.

SSR identification and comparative analysis

SSRs were identified with MISA v2.1 (Beier et al. 2017) with minimum repeat numbers of 10, 6, 5, 5, 5 and 5 for mono- to hexanucleotide motifs; SSRs ≤ 100 bp apart were reported as compound. Each motif was grouped with its circular permutations and reverse complement into a canonical class (e.g., AG, GA, CT and TC), and density was expressed per Mb of unambiguous (A, C, G or T) sequence. The same settings were applied to *E. henryi* (GCA_051912405.1), the closest relative of *C. spruceanum* among the taxa sampled in plastid phylogenies (Saldaña et al. 2022), and to the chromosome-level assembly of *C. canephora* (GCA_036785865.1; Salojärvi et al. 2024), a further representative of the subfamily Dialypetalanthoideae.

Structural annotation of protein-coding genes

Protein-coding genes were predicted by homology with GeMoMa v1.9 (Keilwagen et al. 2018) on the unmasked final draft assembly, because masking can hide exons that overlap repeats. The 70,258 IUPAC ambiguity codes, not accepted by GeMoMa, were replaced by N without changing coordinates. Gene models were transferred from the three annotated *Coffea* genomes, the closest relatives with public annotations (Dialypetalanthoideae) including *C. canephora* (GCA_900059795.1; Denoeud et al. 2014), *C. eugenioides* S.Moore (GCF_003713205.1) and *C. arabica* (GCF_003713225.1). Homology searches used MMseqs2 v18 (Steinegger and Söding 2017) with ReAlign scoring, and the predictions were combined with the GeMoMa Annotation Filter, retaining models with start and stop codons and a relative score ≥ 0.75 per amino acid. Proteins were extracted with gffread v0.12.9 (Pertea and Pertea 2020), and the longest protein of each gene was evaluated with BUSCO v6.1.0 in protein mode against eudicotyledons_odb12.

Functional annotation

Predicted proteins were searched against Swiss-Prot (UniProt Consortium 2025) with DIAMOND v2.1.23 in blastp mode (e-value ≤ 1×10^{-5}; best hit), and hits with identity > 30% and e-value < 1×10^{-10} were considered high-confidence. Orthology-based annotation was performed with eggNOG-mapper v2.1.15 (Cantalapiedra et al. 2021) against eggNOG v5.0 (Huerta-Cepas et al. 2019), with non-electronic GO evidence and automatic taxonomic scope, yielding GO terms (Gene Ontology Consortium 2023), KEGG orthologs and pathways (Kanehisa et al. 2025), Pfam domains (Mistry et al. 2021) and COG categories (Galperin et al. 2021). Pathway summaries were restricted to KEGG maps defined for *Arabidopsis thaliana* (L.) Heynh. (KEGG pathway list retrieved on 16 October 2026), excluding the Human Diseases category. Assignments to pathways mainly characterized in animals may reflect shared KEGG orthologs and were not interpreted as evidence for the corresponding animal signaling systems. Genes with Pfam domains typical of transposable elements (listed in Online Resource 1) were flagged.

Results

Sequencing yielded 323.8 million read pairs (97.13 Gbp; GC 37.42%), of which 97.77% of the reads (92.62 Gbp; ~185× the estimated genome size) were retained after filtering, with 94.39% of bases ≥Q30 (Online Resource 1).

Genome survey

At k = 17 and 21, GenomeScope fitted a diploid profile with distinct heterozygous and homozygous peaks (Online Resource 2, Table S1). At k = 21, the value retained, the haploid genome size was estimated at 499.1–503.2 Mb, with 0.97–1.04% heterozygosity, 114.8–115.7 Mb of repetitive sequence, a k-mer coverage of 67× and a model fit of 80.3–96.4%. The size estimated at k = 17 was similar (491.4–498.0 Mb), whereas at k = 31 the model did not reproduce the bimodal spectrum and was not used (Online Resource 2, Table S1).

De novo assembly and k-mer size optimization

Contiguity and completeness of the ABySS assemblies increased with k, from N50 = 7,891 bp and 60.6% complete BUSCOs at k = 71 to N50 = 17,270 bp and 88.2% at k = 99, but duplicated BUSCOs also increased, from 7.6% to 26.7% (Online Resource 2, Table S2), as expected when longer k-mers resolve alleles into separate paths in heterozygous genomes (Kajitani et al. 2014). k = 87 gave the highest proportion of complete single-copy BUSCOs (65.2%; 84.2% complete, 19.1% duplicated) and it was retained; this assembly comprised 1,096,826 sequences (834.7 Mb).

Contamination screening and nuclear genome definition

The first-pass screen identified 2 vector sequences, 43,492 low-depth sequences without plant hits and 85,876 organellar sequences (12.16 Mb; 81,832 plastid and 4,044 mitochondrial), which together amounted to 129,366 distinct sequences (Fig. 1). Organellar sequences were short (mean 142 bp; 98.7% <200 bp; only 13 ≥ 3 kb). The resulting nuclear assembly comprised 967,460 sequences (813.3 Mb) and retained 84.2% complete BUSCOs.

Length filtering, refined contamination screening and final draft assembly

The nuclear assembly was 1.6 times larger than the GenomeScope estimate, largely because of short sequences including a total of 917,317 sequences shorter than 3 kb (94.8% of the total) that accounted for 294.9 Mb. In addition, the mean read depth increased with sequence length from 24.4× (200–499 bp) to 129.2× (≥ 3 kb; Online Resource 2, Table S3). After removing sequences shorter than 500, 1,000 or 3,000 bp, BUSCO completeness remained almost unchanged (84.2%, 84.2% and 83.9%, respectively), and QV increased slightly (from 70.44 to 71.57–71.60), whereas k-mer completeness fell from 93.52% to 91.64%, 89.44% and 84.34%, respectively (Online Resource 2, Table S4). Of the read k-mers in the homozygous peak (multiplicity 130–199×), 99.8% were present in the unfiltered assembly, 98.9% and 97.9% remained after removing sequences shorter than 500 and 1,000 bp, respectively, and only 94.2% at ≥ 3,000 bp. Therefore, the ≥ 1,000 bp threshold was adopted.

The refined screening removed 17,795 sequences (80.49 Mb; 13.7%) of the ≥ 1,000 bp set that had escaped the first pass because they were sequenced at high depth (Online Resource 2, Table S5). Bacterial sequences (13,154 sequences totalling 50.89 Mb, with a mean GC content of ~ 68% and a mean read depth of ~ 35×) had best hits to phyllosphere Alphaproteobacteria (*Methylobacterium*, *Sphingomonas*, *Aureimonas* and *Methylorubrum*) and contained complete bacteria_odb12, alphaproteobacteria_odb12 and methylobacteriaceae_odb12 BUSCO sets (99.9–100%, 94.3–99.5% duplicated), indicating several co-assembled bacterial genomes. Fungal sequences (341 sequences totalling 23.21 Mb, with a mean GC content of 50.0% and a mean read depth of ~ 153×) had best hits to *Microthyrium microscopicum* (Dothideomycetes) and contained 84.7% of dothideomycetes_odb12 BUSCOs, almost all single-copy (Online Resource 2, Table S6); whereas Metazoa sequences (1,988 sequences totalling 3.19 Mb) matched mites. The removed sequences included the longest scaffolds of the length-filtered set (up to 502.5 kb) and accounted for 8% of the read k-mers.

The final draft assembly comprised 72,381 scaffolds representing 506.9 Mb (scaffold N50 = 11,462 bp; longest scaffold = 135,591 bp) and 101,385 contigs (contig N50 = 8,765 bp), with a GC content of 34.2% (Table 1; Fig. 1). It contained 83.9% complete BUSCOs (64.9% single-copy, 19.0% duplicated; 11.7% fragmented and 4.4% missing), its QV was 71.23 and its k-mer completeness 81.45% (~ 88.5% of the non-microbial read k-mers), and its size was within ~ 1% of the GenomeScope estimate. The nine complete BUSCOs lost with the refined screen, and the decrease of viridiplantae_odb12 completeness from 89.7% to 83.8% (Online Resource 2, Table S7), corresponded to conserved genes that had been scored as complete on microbial copies.

Table 1
Assembly and quality statistics of the final draft nuclear genome of *Calycophyllum spruceanum* (ABySS, k = 87; sequences ≥ 1,000 bp after first-pass and refined contamination screening).

Statistic	**Scaffolds**	**Contigs**
Contiguity		
Number of sequences	72,381	101,385
Total length (bp)	506,926,695	503,059,775
Longest sequence (bp)	135,591	109,076
N50 (bp)	11,462	8,765
L50	12,306	16,374
N90 (bp)	3,103	2,457
L90	45,380	57,887
auN (bp)	15,989	12,009
Sequences ≥ 10 kb	14,847	13,141
Sequences ≥ 50 kb	304	121
GC content (%)	34.23	34.23
N's per 100 kbp	777.91	15.21
Gene-space completeness (BUSCO)		
Complete (%)	83.9	–
Single-copy (%)	64.9	–
Duplicated (%)	19.0	–
Fragmented (%)	11.7	–
Missing (%)	4.4	–
Base accuracy (Merqury)		
Quality value (QV)	71.23	–
k-mer completeness (%)	81.45	–

Contiguity statistics were computed with QUAST v5.3.0 on both all scaffolds and contigs obtained by splitting scaffolds at runs of ≥ 10 N; N's per 100 kbp include 70,258 IUPAC codes. BUSCO v6.0.0 (eudicotyledons_odb12, n = 2,805) and Merqury v1.3 (k = 21) were run on the scaffolds. Merqury uses the reads that built the assembly, so QV mainly reflects errors at scaffold joins; k-mer completeness includes the k-mers of removed microbial sequences (8.0%).

Repetitive DNA content

RepeatModeler produced 1,428 repeat families, 78.6% of these were unclassified. RepeatMasker masked 204.2 Mb (40.29%) of the final draft assembly (Table 2), almost entirely interspersed repeats (38.58%), of which unclassified elements made up 26.78% of the assembly. Among classified repeats, LTR retrotransposons predominated (9.49%), with Gypsy/DIRS1 elements (6.84%) about three times more abundant than Ty1/Copia (2.22%); LINEs (1.14%, almost all L1/CIN4) and DNA transposons (1.17%, mainly hobo-Activator and MULE-MuDR) were minor components, and no SINEs were detected.

Table 2
Repetitive DNA content of the *Calycophyllum spruceanum* final draft assembly, annotated with RepeatMasker v4.1.5 using a species-specific de novo repeat library (1,428 families) generated with RepeatModeler v2.0.7

Repetitive element	**Number of elements**	**Length (bp)**	**% of assembly**
Retroelements	67,605	53,884,753	10.63
SINEs	0	0	0.00
LINEs	10,461	5,763,668	1.14
LTR elements	57,144	48,121,085	9.49
BEL/Pao	436	172,404	0.03
Ty1/Copia	11,817	11,276,128	2.22
Gypsy/DIRS1	43,061	34,675,661	6.84
DNA transposons	10,838	5,945,399	1.17
hobo-Activator	4,542	2,430,167	0.48
MULE-MuDR	3,776	2,196,394	0.43
Tourist/Harbinger	1,160	500,632	0.10
Tc1-IS630-Pogo	290	81,313	0.02
Unclassified	427,355	135,742,864	26.78
Total interspersed repeats	–	**195,573,016**	**38.58**
Rolling-circles	1,851	1,896,523	0.37
Small RNA	2,728	301,781	0.06
Simple repeats	149,652	5,220,334	1.03
Low complexity	25,858	1,228,584	0.24
Total masked sequence	–	**204,220,238**	**40.29**

Percentages are relative to the final draft assembly (506,926,695 bp). The *de novo* library (1,428 families: 186 LTR, 59 DNA transposon, 41 LINE, 9 rolling-circle, 11 RNA-derived and 1,122 unclassified) contains no Dfam or Repbase entries. Subclasses do not add up to class totals, and rolling-circle elements are not included in total interspersed repeats.

SSR abundance, motif composition and comparison with related Rubiaceae genomes

MISA identified 266,057 SSRs in the final draft assembly (529.0 per Mb; Table 3). Mononucleotide repeats predominated (65.1%), followed by dinucleotide (15.2%), compound (13.9%) and trinucleotide repeats (4.3%). Motifs were strongly AT-biased: A/T accounted for 93.1% of mononucleotide and AT/AT for 66.6% of dinucleotide repeats, and AAT/ATT (38.0%) and AAG/CTT (23.1%) were the most frequent trinucleotide classes. With the same settings, SSR density was lower in *C. canephora* (375.7 per Mb) and *E. henryi* (300.5 per Mb) than in *C. spruceanum*, mainly because of fewer mono- and dinucleotide repeats, whereas motif composition was similar across the three genomes (GC-rich classes made up 22.6%, 22.0% and 24.8% of trinucleotide repeats in *C. spruceanum*, *C. canephora* and *E. henryi*, respectively; Fig. 2).

Table 3
SSR types and most frequent canonical motif classes in the *Calycophyllum spruceanum* final draft assembly

SSR type	SSRs	SSRs per Mb	Classes	1st class	2nd class	3rd class
Mononucleotide	173,330	344.6	2	A/T (93.1)	C/G (6.9)	–
Dinucleotide	40,338	80.2	4	AT/AT (66.6)	AG/CT (22.6)	AC/GT (10.8)
Trinucleotide	11,545	23.0	10	AAT/ATT (38.0)	AAG/CTT (23.1)	ATC/GAT (9.9)
Tetranucleotide	2,234	4.4	28	AAAT/ATTT (52.4)	AAAG/CTTT (17.1)	AATT/AATT (5.2)
Pentanucleotide	875	1.7	68	AAAAG/CTTTT (28.8)	AAAAT/ATTTT (28.3)	AAAAC/GTTTT (5.7)
Hexanucleotide	674	1.3	146	AAAAAG/CTTTTT (16.5)	AAAAAT/ATTTTT (10.5)	AAAAGG/CCTTTT (3.0)
Compound	37,061	73.7	–	–	–	–
Total	**266,057**	**529.0**				

MISA v2.1 settings: 1–10 2–6 3–5 4–5 5–5 6 – 5; SSRs ≤ 100 bp apart were counted as compound. Canonical classes pool a motif with its circular permutations and reverse complement. Density is per Mb of A, C, G or T sequence (502.98 Mb). Classes, number of canonical classes detected; values in parentheses are percentages of each SSR type.

Structural and functional annotation of the *C. spruceanum* proteome

GeMoMa predicted 36,690 protein-coding genes and 49,702 transcripts (72.4 genes per Mb), with a median gene length of 1,853 bp, a median of 3 CDS per transcript and a median protein length of 341 residues (Fig. 3; Online Resource 2, Table S8). A total of 7,482 genes (20.4%) had more than one isoform, transferred from the reference annotations, and no protein contained internal stop codons. The longest protein of each gene recovered 76.4% complete BUSCOs (59.5% single-copy, 16.9% duplicated; 5.1% fragmented and 18.5% missing), below the genome-mode value (83.9%), as expected when models split across scaffold ends are discarded by the start- and stop-codon filter.

Overall, 35,632 genes (97.1%) received at least one homology-based functional annotation (Table 4). High-confidence Swiss-Prot hits were obtained for 24,864 genes (67.8%), and eggNOG-mapper assigned 35,631 genes (97.1%) to orthologous groups, and 87.1% to Pfam domains, 89.8% to COG categories, 43.0% to KO terms and 44.3% to GO terms. The five annotation types frequently co-occurred: 9,377 genes (25.6%) had all five and 73.1% had at

least three, whereas 1,380 (3.8%) were supported by a single type and 2,983 (8.1%) by none, 1,925 of which had only a low-confidence Swiss-Prot hit or an eggNOG orthologous group (Fig. 4a). Function unknown (S; 29.6% of genes) was the most frequent COG category, followed by signal transduction (T; 9.1%), transcription (K; 7.3%), posttranslational modification and chaperones (O; 7.1%) and replication, recombination and repair (L; 6.5%) (Fig. 4b). Among KEGG pathways defined for *A. thaliana*, metabolic pathways (map01100; 10.2% of transcripts) and biosynthesis of secondary metabolites (map01110; 6.5%) were the most represented, followed by plant–pathogen interaction, plant MAPK signaling, plant hormone signal transduction and phenylpropanoid biosynthesis (1.4–1.9% each; Online Resource 2, Table S9). Pfam domains typical of transposable elements occurred in 2,813 genes (7.7%), most likely transferred from transposon-derived genes of the *Coffea* references. After excluding these latter, the final balance was 33,877 genes.

Table 4
Functional annotation statistics of the *Calycophyllum spruceanum* predicted proteome (DIAMOND v2.1.23 against Swiss-Prot and eggNOG-mapper v2.1.15)

Functional category	**Transcripts**	**%**	**Genes**	**%**
Total predicted	49,702	100.0	36,690	100.0
Swiss-Prot hit (DIAMOND, e-value ≤ 1×10^{-5})	38,040	76.5	26,491	72.2
High-confidence Swiss-Prot hit	36,056	72.5	24,864	67.8
Assigned to eggNOG orthologous groups	48,591	97.8	35,631	97.1
With PFAM protein domains	44,479	89.5	31,955	87.1
With COG functional category	45,729	92.0	32,963	89.8
With KEGG orthology (KO) terms	22,921	46.1	15,761	43.0
With KEGG pathway	14,537	29.2	10,050	27.4
With Gene Ontology (GO) annotations	22,914	46.1	16,241	44.3
With at least one annotation (DIAMOND or eggNOG)	48,592	97.8	35,632	97.1
With transposable-element Pfam domains	3,203	6.4	2,813	7.7

Percentages are relative to 49,702 transcripts and 36,690 genes; a gene is counted when at least one transcript is annotated. High-confidence Swiss-Prot hits: identity > 30% and e-value < 1×10^{-10}. KEGG pathway counts include all reference maps. Transposable-element Pfam domains are listed in Online Resource 1; their count is an upper bound, because PMD and zf-BED domains also occur in domesticated genes.

Discussion

This study provides the first nuclear genome assembly for *Calycophyllum*, a genus of Neotropical timber trees for which genomic resources have so far consisted mainly of the plastome (Saldaña et al. 2022), Angiosperms353 target-enrichment data from one individual (Baker et al. 2022) and diversity studies based on dominant markers (Russell et al. 1999; Saldaña et al. 2021). The haploid genome size estimated for *C. spruceanum* (499.1–503.2 Mb) lies at the lower end of the range reported for sequenced Rubiaceae. It is close to that of *G. jasminoides* (550.6 Mb by flow cytometry; Xu et al. 2020) and smaller than those of *C. canephora* (~ 710 Mb estimated; Denoeud et al. 2014), *M. speciosa* (~ 727 Mb; Pootakham et al. 2022) and *C. pubescens* (~ 1.1 Gb; Allasi Canales et al. 2022). Its

heterozygosity (0.97–1.04%) is within the range reported for other family members (0.87–0.89% in *C. pubescens* and 0.68–2.2% in *G. jasminoides)* (Xu et al. 2020; Wang et al. 2021b; Allasi Canales et al. 2022). These estimates assume a diploid genome. No chromosome count is available for *C. spruceanum*; the only count recorded for the genus in the Chromosome Counts Database, 2n = 34 in *C. candidissimum* (Kiehn 1995b; Rice et al. 2015), was interpreted as diploid with x = 17, as in *Cinchona* (Kiehn 1995a). A diploid model is therefore plausible, but flow cytometry and a chromosome count of *C. spruceanum* are needed to confirm both genome size and ploidy.

Although based on a single paired-end library, the final draft assembly recovered 83.9% of eudicotyledons BUSCOs as complete, with scaffold and contig N50 values of 11.5 and 8.8 kb. As expected without long-insert libraries, contiguity was well below that of short-read assemblies supported by mate-pair data, such as that of *H. impetiginosus* (N50 of 81.3 kb; Silva-Junior et al. 2018). Fragmentation arises because repeats are the main source of breaks in short-read assemblies (Jiao and Schneeberger 2017) and the inserts of our library were short (peak at 250 bp; Online Resource 1). Duplicated BUSCOs increased with k and reached 19.0% in the final assembly, consistent with part of the heterozygous sequence having been assembled as separate allelic copies (Kajitani et al. 2014), although genuine paralogs may also contribute. We did not remove them with haplotig-purging tools such as purge_dups, which identify duplicates from the read-depth profile of long reads mapped to a primary assembly (Guan et al. 2020). In a fragmented short-read assembly, depth is underestimated in short sequences (Online Resource 1) and inflated in collapsed repeats (Wang et al. 2021a), so depth-based purging could discard genuine sequences. For the same reason, the agreement between the assembly size and the GenomeScope estimate (~ 1%) should not be taken as evidence of a fully collapsed haploid representation: retained alleles are probably offset by collapsed repeats.

Quality control also removed a high-depth microbial fraction (80.5 Mb; 13.7% of the length-filtered set); epiphytic microorganisms on the leaf surface were most likely extracted together with the plant DNA. Most of this fraction corresponded to Alphaproteobacteria such as *Methylobacterium* and *Sphingomonas*, which dominate bacterial communities on leaves (Delmotte et al. 2009; Vorholt 2012), and to a fungus related to *Microthyrium microscopicum*, a member of a family of leaf-associated fungi (Wu et al. 2011). These sequences had passed the depth-based first-pass screen and, before their removal, had inflated viridiplantae_odb12 completeness through microbial copies of conserved genes (Online Resource 2, Table S7), as can occur in contaminated assemblies (Challis et al. 2020). The organellar sequences removed (12.2 Mb) were mostly shorter than 200 bp, and nuclear insertions of organellar DNA usually represent less than 0.1% of plant nuclear genomes (Zhang et al. 2020), so their removal is unlikely to have substantially affected the nuclear gene space, although some genuine nuclear insertions may also have been removed.

Repeats covered 40.3% of the final assembly, less than the ~ 50–54% reported for *C. canephora*, *G. jasminoides* and *M. speciosa* (Denoeud et al. 2014; Xu et al. 2020; Pootakham et al. 2022). This difference is probably methodological rather than biological, because sequences shorter than 1 kb, which are enriched in repeats, were excluded, and repeats are collapsed in short-read assemblies (Wang et al. 2021a). The predominance of Gypsy over Ty1/Copia elements agrees with that reported for *M. speciosa* (8.48% vs. 6.36%; Pootakham et al. 2022). The large unclassified fraction (69.4% of interspersed repeats) is typical of *de novo* libraries of non-model trees; in *H. impetiginosus*, about 71% of repeat consensus sequences also remained unclassified (Silva-Junior et al. 2018).

The 266,057 SSRs identified constitute the first genome-wide microsatellite catalogue for *C. spruceanum*. Previous molecular studies of the species relied on dominant AFLP and RAPD markers and on ITS fingerprints (Russell et al. 1999; Tauchen et al. 2011; Saldaña et al. 2021). Codominant SSRs, in contrast, allow estimation of heterozygosity, gene flow and parentage, which are needed to manage the variation in growth and wood density documented in

provenance and progeny trials of capirona (Sotelo Montes et al. 2006; Weber et al. 2009). Mononucleotide repeats, the most abundant class, are poorly suited as markers because of homopolymer errors, so di- and trinucleotide loci (e.g., AAT/ATT and AAG/CTT) are the most promising candidates for primer design and validation in natural populations. The higher SSR density of *C. spruceanum* compared with *C. canephora* and *Emmenopterys henryi*, driven by mono- and dinucleotide repeats, should be confirmed with a more contiguous assembly; the AT-rich motif composition shared by the three genomes appears more robust.

Homology-based transfer from three *Coffea* annotations predicted 36,690 genes, within the range reported for other Rubiaceae (25,574 in *C. canephora*, 35,967 in *G. jasminoides* and 39,708 protein-coding genes in *M. speciosa*; Denoeud et al. 2014; Xu et al. 2020; Pootakham et al. 2022). The total is probably overestimated by retained allelic copies, reflected in the 16.9% of duplicated BUSCOs in the proteome, and by up to 2,813 genes with transposable-element domains inherited from the reference annotations; gene models of this kind were excluded from the high-confidence gene set of *M. speciosa* (Brose et al. 2021). Homology transfer also has blind spots. It cannot recover lineage-specific genes, it offers no transcript support for the 7,482 genes with alternative isoforms, and it discards models split across scaffold ends, which explains why the proteome (76.4% complete BUSCOs) is less complete than the genome (83.9%). In the argan genome, GeMoMa models served as hints for AUGUSTUS and were then refined with full-length transcripts (Rupp et al. 2024); RNA-seq data will be needed to refine the capirona models in a similar way. Even so, 97.1% of the genes received at least one homology-based functional annotation and 67.8% had high-confidence Swiss-Prot hits, so the gene set already provides a functional reference for the species.

Overall, this draft provides a fragmented reference for *C. spruceanum* that is nonetheless largely complete in gene space. Long-read sequencing and chromatin-conformation scaffolding (Jiao and Schneeberger 2017), together with RNA-seq, flow cytometry and a chromosome count, are the next steps toward a chromosome-level assembly. Population resequencing anchored to this draft and validated SSR panels could support the selection and conservation of capirona genetic resources in the Peruvian Amazon.

Conclusions

We assembled the first draft nuclear genome of *C. spruceanum* from Illumina paired-end reads alone. Although fragmented, the 506.9 Mb assembly recovers 83.9% of conserved eudicot genes and therefore captures most of the gene space. The assembly provides 266,057 SSRs, from which di- and trinucleotide loci can be selected for primer design and validation, and 36,690 homology-based gene models with functional annotations, which together form a first nuclear reference for *Calycophyllum*. Long reads, chromatin-conformation scaffolding and RNA-seq will be needed to obtain a chromosome-level assembly and an evidence-based annotation, and to support population genomic studies of capirona genetic resources.

Declarations

Declarations

Competing interests

The authors have no competing interests to declare that are relevant to the content of this article.

Funding

This research was funded by the project “Creación del servicio de agricultura de precisión en los Departamentos de Lambayeque, Huancavelica, Ucayali y San Martín 4 Departamentos” of the Ministry of Agrarian Development and Irrigation (MIDAGRI) of the Peruvian Government with grant number CUI 2449640. C.I.A. and A.M. are supported by project “Aplicación de la fenómica y genómica de alto rendimiento para iniciar el mejoramiento genético moderno de lupuna (*Ceiba* spp.), especie forestal con potencial maderable” with grant number PE 501087792-2024-PROCIENCIA.

Author Contribution

A.M.: data analysis, methodology, software, validation, writing-original draft preparation, review and editing. C.L.S.: data analysis, validation, review and editing. C.Me.: data analysis, validation, review and editing. W.S.: conceptualization, supervision, review and editing. C.Ma.: investigation, supervision, review and editing. C.I.A.: conceptualization, methodology, investigation, supervision, review and editing.

Acknowledgement

We thank the High-Performance Computing Cluster of the Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas (UNTRM–DATA SCIENCE) for providing computational resources for data analysis. We also gratefully acknowledge Fredy Velayarce and Jairo Vallejos from the Information Technology Office of UNTRM for their valuable technical assistance and support with the HPC server. In addition, we thank Eric Rodriguez, Maria Angélica Puyo and Cristina Aybar (INIA) for supporting the logistic activities in our laboratory.

Data Availability

The whole-genome sequencing reads generated in this study are available in the NCBI Sequence Read Archive under accession (in process) (BioProject PRJNA1547248; BioSample SAMN64310517). The final draft nuclear assembly has been deposited in GenBank under accession JCFKHX000000000; the version described in this paper is version JCFKHX010000000. The plastome sequence of the same individual is available in GenBank under accession OK326865.1 (Saldaña et al. 2022).

References

1. Allasi Canales N, Pérez-Escobar OA, Powell RF et al (2022) A highly contiguous, scaffold-level nuclear genome assembly for the fever tree (*Cinchona pubescens* Vahl) as a novel resource for Rubiaceae research. GigaByte. https://doi.org/10.46471/gigabyte.71
2. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2
3. Andrews S (2014) FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
4. Baker WJ, Bailey P, Barber V et al (2022) A comprehensive phylogenomic platform for exploring the angiosperm tree of life. Syst Biol 71:301–319. https://doi.org/10.1093/sysbio/syab035
5. Beier S, Thiel T, Münch T, Scholz U, Mascher M (2017) MISA-web: a web server for microsatellite prediction. Bioinformatics 33:2583–2585. https://doi.org/10.1093/bioinformatics/btx198

6. Brose J, Lau KH, Dang TTT et al (2021) The *Mitragyna speciosa* (Kratom) genome: a resource for data-mining potent pharmaceuticals that impact human health. G3 11:jkab058. https://doi.org/10.1093/g3journal/jkab058
7. Buchfink B, Reuter K, Drost HG (2021) Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods 18:366–368. https://doi.org/10.1038/s41592-021-01101-x
8. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J (2021) eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Mol Biol Evol 38:5825–5829. https://doi.org/10.1093/molbev/msab293
9. Challis R, Richards E, Rajan J, Cochrane G, Blaxter M (2020) BlobToolKit – interactive quality assessment of genome assemblies. G3 10:1361–1374. https://doi.org/10.1534/g3.119.400908
10. Chen S, Zhou Y, Chen Y, Gu J (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34:i884–i890. https://doi.org/10.1093/bioinformatics/bty560
11. Delmotte N, Knief C, Chaffron S et al (2009) Community proteogenomics reveals insights into the physiology of phyllosphere bacteria. Proc Natl Acad Sci USA 106:16428–16433. https://doi.org/10.1073/pnas.0905240106
12. Denoeud F, Carretero-Paulet L, Dereeper A et al (2014) The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. Science 345:1181–1184. https://doi.org/10.1126/science.1255274
13. Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull 19:11–15
14. Ellinghaus D, Kurtz S, Willhoeft U (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics 9:18. https://doi.org/10.1186/1471-2105-9-18
15. Estrada R, Tantalean JFC, Saldaña CL, Romero Y, Vilcara E, Arbizu CI (2024) Draft genome and SSR data mining of a Peruvian landrace of *Capsicum chinense*, the arnaucho chili pepper. Genet Resour Crop Evol 71:4017–4025. https://doi.org/10.1007/s10722-024-01941-4
16. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF (2020) RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci USA 117:9451–9457. https://doi.org/10.1073/pnas.1921046117
17. Galperin MY, Wolf YI, Makarova KS, Vera Alvarez R, Landsman D, Koonin EV (2021) COG database update: focus on microbial diversity, model organisms, and widespread pathogens. Nucleic Acids Res 49:D274–D281. https://doi.org/10.1093/nar/gkaa1018
18. Gene Ontology Consortium (2023) The Gene Ontology knowledgebase in 2023. Genetics 224:iyad031. https://doi.org/10.1093/genetics/iyad031
19. Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R (2020) Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics 36:2896–2898. https://doi.org/10.1093/bioinformatics/btaa025
20. Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: quality assessment tool for genome assemblies. Bioinformatics 29:1072–1075. https://doi.org/10.1093/bioinformatics/btt086
21. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ, von Mering C, Bork P (2019) eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res 47:D309–D314. https://doi.org/10.1093/nar/gky1085
22. Jackman SD, Vandervalk BP, Mohamadi H, Chu J, Yeo S, Hammond SA, Jahesh G, Khan H, Coombe L, Warren RL, Birol I (2017) ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. Genome Res 27:768–777. https://doi.org/10.1101/gr.214346.116

23. Jiao WB, Schneeberger K (2017) The impact of third generation genomic technologies on plant genome assembly. Curr Opin Plant Biol 36:64–70. https://doi.org/10.1016/j.pbi.2017.02.002
24. Kajitani R, Toshimoto K, Noguchi H et al (2014) Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res 24:1384–1395. https://doi.org/10.1101/gr.170720.113
25. Kanehisa M, Furumichi M, Sato Y, Matsuura Y, Ishiguro-Watanabe M (2025) KEGG: biological systems database as a model of the real world. Nucleic Acids Res 53:D672–D677. https://doi.org/10.1093/nar/gkae909
26. Keilwagen J, Hartung F, Paulini M, Twardziok SO, Grau J (2018) Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. BMC Bioinformatics 19:189. https://doi.org/10.1186/s12859-018-2203-5
27. Kiehn M (1995a) Chromosome survey of the Rubiaceae. Ann Missouri Bot Gard 82:398–408. https://doi.org/10.2307/2399890
28. Kiehn M (1995b) Chromosomenzählungen an Rubiaceae aus Costa Rica. Carinthia II Sonderheft 53:75–77
29. Laetsch DR, Blaxter ML (2017) BlobTools: interrogation of genome assemblies. F1000Research 6:1287. https://doi.org/10.12688/f1000research.12232.1
30. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. https://doi.org/10.1038/nmeth.1923
31. Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100. https://doi.org/10.1093/bioinformatics/bty191
32. Li H (2023) Protein-to-genome alignment with miniprot. Bioinformatics 39:btad014. https://doi.org/10.1093/bioinformatics/btad014
33. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352
34. 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
35. 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
36. Marks RA, Hotaling S, Frandsen PB, VanBuren R (2021) Representation and participation across 20 years of plant genome sequencing. Nat Plants 7:1571–1578. https://doi.org/10.1038/s41477-021-01031-8
37. Mistry J, Chuguransky S, Williams L et al (2021) Pfam: the protein families database in 2021. Nucleic Acids Res 49:D412–D419. https://doi.org/10.1093/nar/gkaa913
38. Ou S, Jiang N (2018) LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. Plant Physiol 176:1410–1422. https://doi.org/10.1104/pp.17.01310
39. Peixoto H, Roxo M, Koolen H, da Silva F, Silva E, Braun MS, Wang X, Wink M (2018) *Calycophyllum spruceanum* (Benth.), the Amazonian "tree of youth" prolongs longevity and enhances stress resistance in *Caenorhabditis elegans*. Molecules 23:534. https://doi.org/10.3390/molecules23030534
40. Pertea G, Pertea M (2020) GFF utilities: GffRead and GffCompare. F1000Research 9:304. https://doi.org/10.12688/f1000research.23297.2
41. Pootakham W, Yoocha T, Jomchai N et al (2022) A chromosome-scale genome assembly of *Mitragyna speciosa* (kratom) and the assessment of its genetic diversity in Thailand. Biology 11:1492.

https://doi.org/10.3390/biology11101492

42. 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
43. Razafimandimbison SG, Rydin C (2024) Phylogeny and classification of the coffee family (Rubiaceae, Gentianales): overview and outlook. Taxon 73:673–717. https://doi.org/10.1002/tax.13167
44. Rhie A, Walenz BP, Koren S, Phillippy AM (2020) Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol 21:245. https://doi.org/10.1186/s13059-020-02134-9
45. Rice A, Glick L, Abadi S, Einhorn M, Kopelman NM, Salman-Minkov A, Mayzel J, Chay O, Mayrose I (2015) The Chromosome Counts Database (CCDB) – a community resource of plant chromosome numbers. New Phytol 206:19–26. https://doi.org/10.1111/nph.13191
46. Rupp O, Roessner C, Lederer-Ponzer N, Wollenweber TE, Becker A, Lamaoui M (2024) Genome of *Argania spinosa* L.: insights into oil production and the tocopherol biosynthesis pathway. Genet Resour Crop Evol 71:4027–4042. https://doi.org/10.1007/s10722-024-01931-6
47. Russell JR, Weber JC, Booth A, Powell W, Sotelo Montes C, Dawson IK (1999) Genetic variation of *Calycophyllum spruceanum* in the Peruvian Amazon Basin, revealed by amplified fragment length polymorphism (AFLP) analysis. Mol Ecol 8:199–204. https://doi.org/10.1046/j.1365-294x.1999.00551.x
48. Saldaña CL, Cancan JD, Cruz W, Correa MY, Ramos M, Cuellar E, Arbizu CI (2021) Genetic diversity and population structure of capirona (*Calycophyllum spruceanum* Benth.) from the Peruvian Amazon revealed by RAPD markers. Forests 12:1125. https://doi.org/10.3390/f12081125
49. Saldaña CL, Rodriguez-Grados P, Chávez-Galarza JC, Feijoo S, Guerrero-Abad JC, Vásquez HV, Maicelo JL, Jhoncon JH, Arbizu CI (2022) Unlocking the complete chloroplast genome of a native tree species from the Amazon basin, capirona (*Calycophyllum spruceanum*, Rubiaceae), and its comparative analysis with other Ixoroideae species. Genes 13:113. https://doi.org/10.3390/genes13010113
50. Salojärvi J, Rambani A, Yu Z et al (2024) The genome and population genomics of allopolyploid *Coffea arabica* reveal the diversification history of modern coffee cultivars. Nat Genet 56:721–731. https://doi.org/10.1038/s41588-024-01695-w
51. Sayers EW, Beck J, Bolton EE et al (2025) Database resources of the National Center for Biotechnology Information in 2025. Nucleic Acids Res 53:D20–D29. https://doi.org/10.1093/nar/gkae979
52. Shen W, Sipos B, Zhao L (2024) SeqKit2: a Swiss army knife for sequence and alignment processing. iMeta 3:e191. https://doi.org/10.1002/imt2.191
53. Silva-Junior OB, Grattapaglia D, Novaes E, Collevatti RG (2018) Genome assembly of the Pink Ipê (*Handroanthus impetiginosus*, Bignoniaceae), a highly valued, ecologically keystone Neotropical timber forest tree. GigaScience 7:gix125. https://doi.org/10.1093/gigascience/gix125
54. Smit AFA, Hubley R, Green P (2013) RepeatMasker Open-4.0. http://www.repeatmasker.org
55. Sotelo Montes C, Hernández RE, Beaulieu J, Weber JC (2006) Genetic variation and correlations between growth and wood density of *Calycophyllum spruceanum* at an early age in the Peruvian Amazon. Silvae Genet 55:217–228. https://doi.org/10.1515/sg-2006-0029
56. Steinegger M, Salzberg SL (2020) Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank. Genome Biol 21:115. https://doi.org/10.1186/s13059-020-02023-1
57. Steinegger M, Söding J (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol 35:1026–1028. https://doi.org/10.1038/nbt.3988

58. Tauchen J, Lojka B, Hlásná-Čepková P, Svobodová E, Dvořáková Z, Rollo A (2011) Morphological and genetic diversity of *Calycophyllum spruceanum* (Benth) K. Schum (Rubiaceae) in Peruvian Amazon. Agric Trop Subtrop 44:212–218
59. Tegenfeldt F, Kuznetsov D, Manni M, Berkeley M, Zdobnov EM, Kriventseva EV (2025) OrthoDB and BUSCO update: annotation of orthologs with wider sampling of genomes. Nucleic Acids Res 53:D516–D522. https://doi.org/10.1093/nar/gkae987
60. UniProt Consortium (2025) UniProt: the Universal Protein Knowledgebase in 2025. Nucleic Acids Res 53:D609–D617. https://doi.org/10.1093/nar/gkae1010
61. Vorholt JA (2012) Microbial life in the phyllosphere. Nat Rev Microbiol 10:828–840. https://doi.org/10.1038/nrmicro2910
62. Wang P, Meng F, Moore BM, Shiu SH (2021a) Impact of short-read sequencing on the misassembly of a plant genome. BMC Genomics 22:99. https://doi.org/10.1186/s12864-021-07397-5
63. Wang W, Shao F, Deng X et al (2021b) Genome surveying reveals the complete chloroplast genome and nuclear genomic features of the crocin-producing plant *Gardenia jasminoides* Ellis. Genet Resour Crop Evol 68:1165–1180. https://doi.org/10.1007/s10722-020-01056-6
64. Weber JC, Sotelo Montes C, Ugarte J, Simons T (2009) Phenotypic selection of *Calycophyllum spruceanum* on farms in the Peruvian Amazon: evaluating a low-intensity selection strategy. Silvae Genet 58:172–179. https://doi.org/10.1515/sg-2009-0023
65. Wu HX, Schoch CL, Boonmee S, Bahkali AH, Chomnunti P, Hyde KD (2011) A reappraisal of Microthyriaceae. Fungal Divers 51:189–248
66. Xu Z, Pu X, Gao R et al (2020) Tandem gene duplications drive divergent evolution of caffeine and crocin biosynthetic pathways in plants. BMC Biol 18:63. https://doi.org/10.1186/s12915-020-00795-3
67. Zhang GJ, Dong R, Lan LN, Li SF, Gao WJ, Niu HX (2020) Nuclear integrants of organellar DNA contribute to genome structure and evolution in plants. Int J Mol Sci 21:707. https://doi.org/10.3390/ijms21030707

Figures

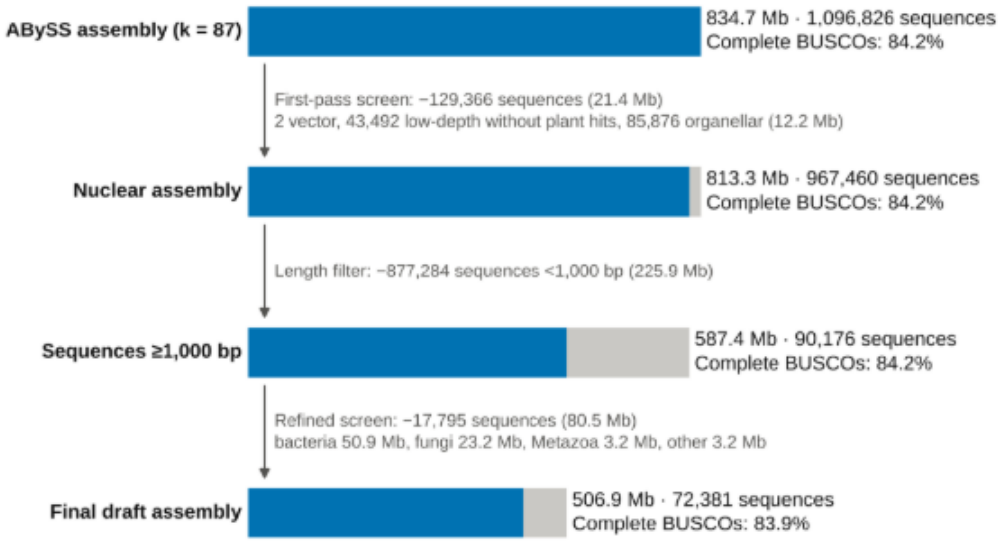


Figure 1

Filtering steps from the ABySS assembly (k = 87) to the final draft assembly of *Calycophyllum spruceanum*. Bars show total length (dark bars, sequence retained; light segments, length removed at the preceding step); labels give the number of sequences and the proportion of complete eudicotyledons_odb12 BUSCOs at each stage, and arrows the sequences removed at each step

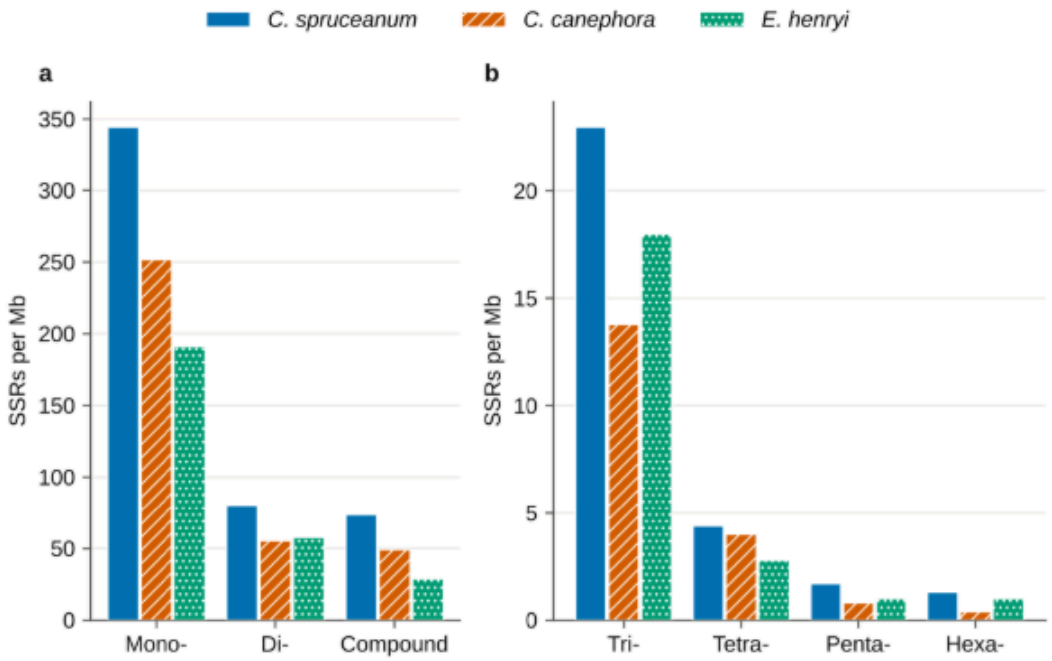


Figure 2

Density of SSRs in three Rubiaceae genomes. SSRs per Mb of unambiguous sequence (A, C, G or T) by repeat type, identified with MISA v2.1 (1-10 2-6 3-5 4-5 5-5 6-5) in the *C. spruceanum*final draft assembly and in the genomes of *C. canephora*(GCA_036785865.1) and *E. henryi* (GCA_051912405.1), with 266,057, 252,559 and 243,652 SSRs, respectively. (a) Mononucleotide (Mono-), dinucleotide (Di-) and compound SSRs. (b) Tri- to hexanucleotide SSRs, shown on a separate scale

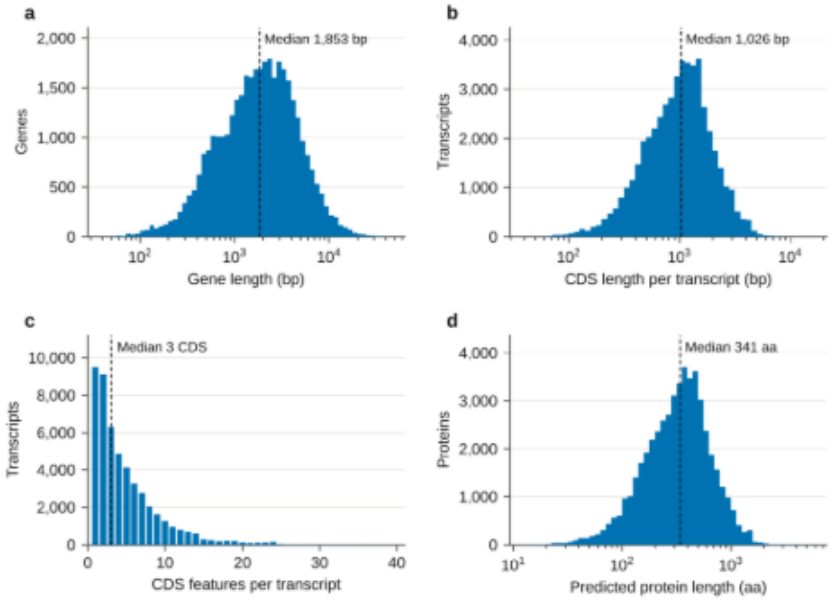


Figure 3

Gene structure characteristics of the protein-coding genes predicted in the *Calycophyllum spruceanum* final draft assembly. (a) Gene length. (b) Total coding sequence (CDS) length per transcript. (c) Number of CDS features per transcript. (d) Predicted protein length. Gene, CDS and protein lengths are shown on a logarithmic x-axis; dashed lines mark the medians. Structural annotation was generated with GeMoMa v1.9 using *Coffea canephora* (GCA_900059795.1), *C. eugenioides*(GCF_003713205.1) and *C. arabica* (GCF_003713225.1) as reference genomes

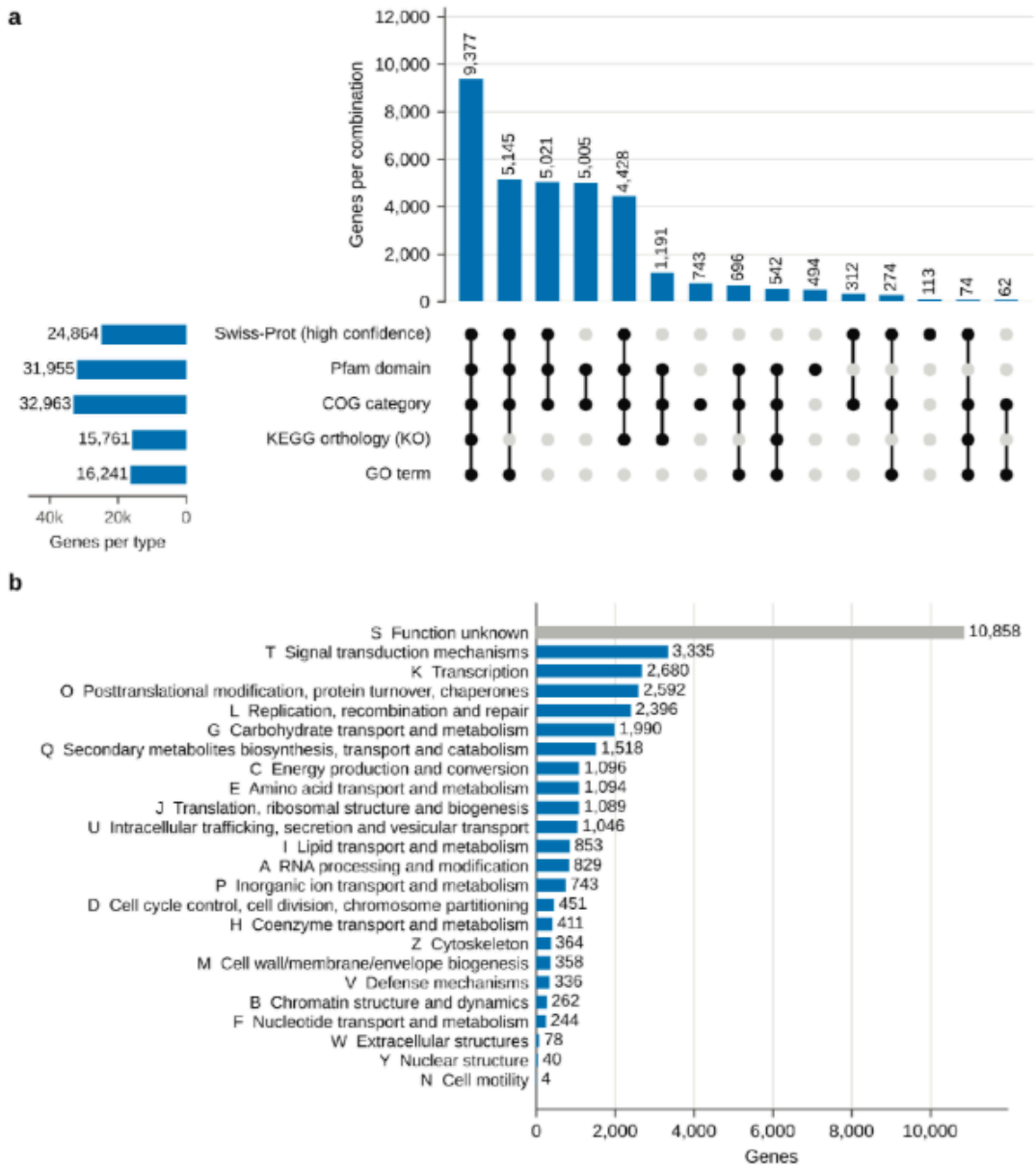


Figure 4

Functional annotation of the 36,690 predicted genes of *Calycophyllum spruceanum*. (a) Genes in the 15 most frequent combinations of five annotation types (high-confidence Swiss-Prot hit, Pfam domain, COG category, KEGG orthology term and GO term); filled dots mark the types present, and the bars on the left give the number of genes with each type. A further 2,983 genes had none of these types. (b) Genes in each COG functional category (Function unknown, S, shown as a lighter bar); a gene is counted in every category assigned to any of its transcripts

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM1.pdf
- ESM2.pdf