Online Resource 1

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

Genetic Resources and Crop Evolution

A. Menacho, C. L. Saldaña, C. Mendoza, W. Salazar, C. Machicado, C. I. Arbizu

Corresponding author: Carlos I. Arbizu, Facultad de Ingeniería y Ciencias Agrarias, Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas (UNTRM), Chachapoyas, Amazonas, Peru; carlos.arbizu@untrm.edu.pe

Supplementary methods and results. This file gives the settings, intermediate results and checks summarized in the main text; the related main-text sections are given in parentheses after each heading.

S1. Sequencing and read processing (Materials and methods: Read processing and genome size estimation; Results, first paragraph)

Adapters were detected automatically for paired-end data (--detect_adapter_for_pe). Both read ends were trimmed with a sliding window (--cut_front and --cut_tail; default 4-bp window and mean quality threshold of 20), and reads shorter than 30 bp were discarded (--length_required 30). The default read-level quality filter of fastp was also applied: reads with >40% of bases below Phred 15 or with >5 N bases were discarded. Sequencing statistics are summarized in Table ESM1.

Table ESM1. *Sequencing and read-filtering statistics.*

Statistic	Raw reads	After filtering
Reads	647,554,078 (323,777,039 pairs)	633,106,646 (97.77%)
Read length (bp)	150	146 (mean)
Total bases (Gbp)	97.13	92.62
GC content (%)	37.42	—
Mean Phred quality per position	—	>37 at all positions
Bases ≥Q30 (%)	—	94.39
Duplication rate (%)	10.97	—
Insert size of overlapping pairs	peak at 250 bp; ~97% of pairs overlapping	—
Nominal coverage	—	~185× (499.1–503.2 Mb estimate)

S2. Genome survey (Results: Genome survey)

GenomeScope 2.0 estimated a haploid genome size of 491.4–498.0 Mb and 1.06–1.18% heterozygosity at k = 17 (kcov = 70.6×), and 499.1–503.2 Mb and 0.97–1.04% at k = 21 (kcov = 67×); the homozygous peak lies at twice kcov. Repeat content decreased with k, from 215.7–218.6 Mb at k = 17 to 114.8–115.7 Mb at k = 21, so the two k values differed about twofold in repeat content despite similar size estimates. At k = 31 the model converged on a different solution (578.0–587.3 Mb; heterozygosity 0.10–0.21%; model fit 83.46–95.04%) whose kcov (53×) was lower than the ~62× expected from the k = 21 fit [$kcov_{31} \approx kcov_{21} \times (L - 31 + 1)/(L - 21 + 1)$, with L = 146 bp]; this underestimate alone accounts for the larger size (587.3 Mb × 53/61.7 ≈ 505 Mb). No chromosome count for *C. spruceanum* is recorded in the Chromosome Counts Database (Rice et al. 2015); the only count recorded for the genus in that database, 2n = 34 in *C. candidissimum* (Vahl) DC. (Kiehn 1995b), was interpreted as diploid with x = 17 (Kiehn 1995a), but ploidy was not verified independently for *C. spruceanum*.

S3. ABySS assembly and IUPAC codes (Materials and methods: De novo genome assembly)

The k values 71, 79, 87 and 95 were evaluated in steps of 8, and k = 99 was added close to the upper range commonly used for 150-bp reads. IUPAC ambiguity codes were introduced almost exclusively at the scaffolding stage of ABySS, where alternative sequence paths are merged into consensus sequences: the k = 87 assembly contained 0 codes in unitigs, 130 in contigs and 71,155 in scaffolds, and the final draft assembly 70,258.

S4. Contiguity statistics (Materials and methods: Assembly quality assessment)

During k optimization QUAST was run with --large and --eukaryote, under which statistics are computed for sequences ≥3 kb. Scaffold statistics of the final draft assembly were computed on all sequences (--min-contig 1000). Contig statistics were computed with a custom Python script on all contigs obtained by splitting scaffolds at runs of ≥10 N, without a minimum length, because the --split-scaffolds option of QUAST filters by length only the last fragment of each scaffold.

S5. Read depth and the short-sequence fraction (Materials and methods: Assembly quality assessment; Results: Length filtering, refined contamination screening and final draft assembly)

Because end-to-end alignment requires the whole read to fall within a sequence, Bowtie2 underestimates depth in sequences whose length approaches the read length. After correcting for the number of possible read start positions, depth still increased with length, from ~51× (200–499 bp) to ~75× (500–999 bp), ~100× (1–3 kb) and ~131× (≥3 kb). The relatively high depth of sequences <200 bp (32.2×) probably reflects shorter, quality-trimmed reads mapping to collapsed repeats. The fraction removed by length filtering contained 38–50% of the assembly-specific (erroneous) k-mers, which explains the small increase in QV after filtering. Sequences between 1 and 3 kb contained genuine single-copy sequence, whereas sequences shorter than 1 kb were dominated by low-depth and allelic content.

S6. Organellar sequences (Materials and methods: Contamination screening and organellar filtering; Results: Contamination screening and nuclear genome definition)

The asm20 preset of minimap2 was chosen to accommodate the divergence expected between genera, and *Coffea arabica* was used as a mitochondrial reference because it belongs to the same subfamily as *Calycophyllum*, Dialypetalanthoideae (formerly Ixoroideae; Razafimandimbison and Rydin 2024). The 85,876 organellar sequences totalled 12.16 Mb (mean length 142 bp; length-weighted mean depth 120×, underestimated for short sequences); 81,832 (11.55 Mb) aligned best to the plastome and 4,044 (0.60 Mb) to the mitochondrial references. The plastid-derived fraction, ~75 times the size of the plastome, most likely consists of error- or variant-containing fragments arising from the very high coverage of the plastid genome. Because 98.7% of these sequences were shorter than 200 bp and only 13 were ≥3 kb (0.10 Mb), the removal of nuclear insertions of organellar DNA (NUPTs/NUMTs) is unlikely to have affected the nuclear assembly substantially; moreover, the ≥90% alignment-coverage criterion retains long nuclear sequences that contain only short organellar insertions.

S7. Refined contamination screen (Materials and methods and Results: Length filtering, refined contamination screening and final draft assembly)

Sequences were removed when (R1) all hits were to non-plant cellular organisms and GC was ≥40%; (R2) all hits were viral and GC was ≥45%; (R3) they had no hit and GC was ≥50%; (R4) they had at least one plant hit but most hits were bacterial and GC was ≥55%; or (R5) they had no hit, depth was <20× and GC was ≥40%. Sequences with exclusively non-plant hits but plant-like GC (<40%) were retained. The BUSCO datasets used for validation were OrthoDB v12 releases of 22 May 2026. The fungal sequences (~153×) were separated from the plant fraction (GC 34.9%; ~135×) by GC content and taxonomy rather than by depth. Of the 33 eudicotyledons BUSCO matches located in removed sequences (Online Resource 2, Table S5), only nine involved genes that lost their complete status. The GC content of the reads (37.4%) was higher than that of the final assembly (34.2%), partly because the reads also include organellar and microbial sequences. The final assembly was written with a fixed line width of 80 characters.

S8. Repeat annotation (Materials and methods: Repeat annotation)

RepeatMasker v4.1.5 was run with the species-specific library (-lib), soft-masking (-xsmall), GFF output (-gff) and alignment output (-a). The library (1,428 families) comprised 186 LTR retrotransposon, 59 DNA transposon, 41 LINE and 9 rolling-circle families, 11 RNA-derived families (8 rRNA, 2 snRNA and 1 tRNA) and 1,122 unclassified families (78.6%). Unclassified elements accounted for 69.4% of the interspersed repeat content.

S9. SSR comparison (Materials and methods: SSR identification and comparative analysis; Results: SSR abundance, motif composition and comparison with related Rubiaceae genomes)

The chromosome-level *C. canephora* assembly GCA_036785865.1 (scaffold N50 = 50.1 Mb; Salojärvi et al. 2024) was used for SSR comparisons, whereas GCA_900059795.1 was used for annotation transfer because it is the *C.*

canephora assembly with a public gene annotation. The higher SSR density of *C. spruceanum* was driven by mononucleotide (344.6 vs. 251.9 and 191.2 per Mb in *C. canephora* and *E. henryi*) and dinucleotide repeats (80.2 vs. 55.5 and 57.8 per Mb), whereas tetra-, penta- and hexanucleotide repeats were rare in all three genomes (≤4.4 per Mb).

S10. Gene annotation (Materials and methods: Structural annotation of protein-coding genes; Functional annotation. Results: Structural and functional annotation of the *C. spruceanum* proteome)

The reference annotations comprised 25,574 (*C. canephora*), 33,619 (*C. eugenioides*) and 52,343 genes (*C. arabica*, an allotetraploid; Salojärvi et al. 2024). The *C. eugenioides* and *C. arabica* annotations correspond to NCBI Annotation Release 100. The protein-mode BUSCO analysis used v6.1.0, the version installed at the time of annotation; genome-mode analyses used v6.0.0 with the same dataset. Genes were flagged as carrying transposable-element domains when eggNOG-mapper assigned any of the following Pfam domains: RVT_1–3, rve, Retrotrans_gag and related gag domains, RVP_2, Asp_protease_2, UBN2, MULE, Dimer_Tnp_hAT, DDE_Tnp, Transposase, Transpos_assoc, PMD and zf-BED; the most frequent were reverse transcriptase (RVT_1, RVT_3), retrotransposon gag (Retrotrans_gag) and integrase (rve) domains. DIAMOND returned Swiss-Prot hits for 38,040 transcripts (76.5%), of which 36,056 (72.5%) met the high-confidence criteria.