Online Resource 2

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 Tables S1–S9

Table S1 Comparison of genome survey statistics across three k-mer sizes (k = 17, 21, and 31) evaluated with GenomeScope v2.0

Parameter	k = 17	k = 21	k = 31
Homozygosity (%)	98.82–98.94	98.96–99.03	99.79–99.90
Heterozygosity (%)	1.06–1.18	0.97–1.04	0.10–0.21
Haploid genome size (Mb)	491.4–498.0	499.1–503.2	578.0–587.3
Repeat content (Mb)	215.7–218.6	114.8–115.7	62.8–63.8
Unique sequence (Mb)	275.7–279.4	384.4–387.5	515.2–523.5
Model fit (%)	69.81–96.11	80.31–96.40	83.46–95.04
k-mer coverage, kcov (×)	70.6	67	53
Read error rate (%)	0.55	0.59	0.47

k-mers were counted with Jellyfish v2.3.1 on quality-filtered paired-end reads; values are the ranges reported by GenomeScope 2.0 (rounded), and kcov is the single best-fit value. k = 21 was retained (see Results: Genome survey). At k = 31 the fitted model did not reproduce the bimodal spectrum; its kcov (53×) was below the ~62× expected from the k = 21 fit for the shorter effective read length per k-mer [$kcov_{31} \approx kcov_{21} \times (L - 31 + 1)/(L - 21 + 1)$, L = 146 bp], which alone accounts for its larger size estimate (587.3 Mb × 53/61.7 ≈ 505 Mb).

Table S2 K-mer size optimization for the ABySS assembly: contiguity and BUSCO gene-space completeness (eudicotyledons_odb12) across the five k values evaluated

k	N50 (bp)	Longest sequence (bp)	Complete (%)	Single-copy (%)	Duplicated (%)	Fragmented (%)	Missing (%)
71	7,891	147,760	60.6	53.0	7.6	27.5	11.9
79	11,221	159,109	77.3	62.8	14.5	17.6	5.1
87	13,584	502,507	84.2	65.2	19.1	11.6	4.1
95	16,193	2,256,758	87.4	63.3	24.1	9.3	3.3
99	17,270	3,393,101	88.2	61.5	26.7	9.0	2.8

Contiguity metrics (N50, longest sequence) were calculated with QUAST v5.3.0 on sequences ≥3 kb (--large option). k = 87 was selected for the final assembly as the value with the highest proportion of complete single-copy BUSCOs. These assemblies had not yet been screened for contamination: the longest sequence at k = 87 was later identified as bacterial, and those at k = 95 and 99 are likely microbial.

Table S3 Distribution of sequence lengths in the nuclear assembly of *Calycophyllum spruceanum* after first-pass decontamination, and mean read depth per length class

Length class (bp)	Sequences	Sequences (%)	Total length (bp)	Length (%)	Mean read depth (×)
<200	429,385	44.38	64,885,820	7.98	32.2
200–499	366,531	37.89	105,220,909	12.94	24.4
500–999	81,368	8.41	55,752,432	6.86	58.4
1,000–2,999	40,033	4.14	69,054,609	8.49	91.5
≥3,000	50,143	5.18	518,361,065	63.74	129.2
Total	**967,460**	**100.00**	**813,274,835**	**100.00**	–

Mean read depth is the length-weighted mean of per-sequence depth obtained with samtools coverage after mapping the trimmed reads with Bowtie2 (end-to-end mode) to the assembly before decontamination; it is underestimated in sequences close to the read length (150 bp). Sequences ≥1,000 bp (90,176 sequences; 587,415,674 bp) were subsequently subjected to the refined contamination screen (Table S5).

Table S4 Effect of minimum sequence length thresholds on the nuclear assembly of *Calycophyllum spruceanum* after first-pass decontamination

Minimum length	Sequences	Total length (Mb)	N50 (bp)	Complete BUSCOs (%)	Duplicated BUSCOs (%)	QV	k-mer completeness (%)	Homozygous-peak k-mers (%)
None	967,460	813.27	6,746	84.2	19.1	70.44	93.52	99.8
≥500 bp	171,544	643.17	10,220	84.2	19.1	71.58	91.64	98.9
≥1,000 bp	**90,176**	**587.42**	**11,650**	**84.2**	**19.1**	**71.60**	**89.44**	**97.9**
≥3,000 bp	50,143	518.36	13,585	83.9	19.0	71.57	84.34	94.2

N50 was calculated with SeqKit v2.10.1 on all sequences of each set. BUSCO v6.0.0 was run in genome mode against eudicotyledons_odb12 (n = 2,805). QV and k-mer completeness were calculated with Merqury v1.3 (k = 21). Homozygous-peak k-mers: percentage of read k-mers with multiplicity 130–199× (a window extending from the homozygous peak, ~134× or twice kcov, to ~1.5 times the peak, chosen to exclude heterozygous k-mers at ~67× and repeated k-mers; it also includes k-mers of the fungal sequences, at ~132×, which were still present in these sets) that are present in the assembly. Values correspond to the sets before the refined contamination screen (see Materials and methods: Length filtering, refined contamination screening and final draft assembly), which subsequently removed 80.49 Mb of microbial sequences from the ≥1,000 bp set. The selected threshold (≥1,000 bp) is shown in bold.

Table S5 Refined taxon-annotated GC–coverage contamination screen applied to the ≥1,000 bp set: rules and sequences removed

Rule	Criterion	Assigned group	Sequences	Length (bp)	GC (%)	Mean depth (×)	BUSCOs[a]
R1	Only non-plant cellular hits; GC ≥40%	Bacteria	12,974	50,127,263	67.6	35.2	25
		Fungi	341	23,208,322	50.0	152.7	8
		Metazoa	1,988	3,193,970	45.7	14.0	0
		Other eukaryotes	13	130,686	66.7	35.7	0
R2	Only viral hits; GC ≥45%	Viruses	3	5,611	59.5	14.6	0
R3	No hit; GC ≥50%	–	1,397	1,810,401	62.0	16.6	0
R4	≥1 plant hit, majority bacterial; GC ≥55%	Bacteria	180	766,815	68.7	44.2	0
R5	No hit; depth <20×; GC ≥40%	–	899	1,245,911	45.0	13.6	0
Total removed			**17,795**	**80,488,979**			**33**
Reference	Viridiplantae hits (retained)	–	35,286	360,823,660	34.9	135.0	–

Taxonomic assignment from DIAMOND v2.1.23 blastx hits against NCBI nr (e-value ≤1 × 10^{-25}); GC content excludes N and ambiguity codes; depth from Bowtie2 mapping (see Materials and methods: Assembly quality assessment). [a]Eudicotyledons_odb12 BUSCO matches (complete, duplicated or fragmented) located in the removed sequences. Sequences with exclusively non-plant hits but plant-like GC (<40%; 477 sequences, 1.66 Mb), including putative endogenous viral elements (~170×, GC ~36%), were retained.

Table S6 BUSCO completeness of the sequences removed as bacterial and fungal during refined contamination screening

Removed set	Lineage dataset (n)	C (%)	S (%)	D (%)	F (%)	M (%)
Bacterial (13,154 seq.; 50.89 Mb)	bacteria_odb12 (116)	100.0	2.6	97.4	0.0	0.0
	alphaproteobacteria_odb12 (426)	100.0	0.5	99.5	0.0	0.0
	methylobacteriaceae_odb12 (1,239)	99.9	5.6	94.3	0.0	0.1
Fungal (341 seq.; 23.21 Mb)	dothideomycetes_odb12 (4,643)	84.7	84.4	0.3	1.8	13.5

BUSCO v6.0.0 in genome mode; OrthoDB v12 datasets released on 22 May 2026. C, complete; S, single-copy; D, duplicated; F, fragmented; M, missing. The bacterial set comprises the sequences removed by rules R1 (Bacteria) and R4. High duplication in bacterial datasets indicates several co-assembled bacterial genomes; the near absence of duplication in the fungal dataset is consistent with a single haploid genome.

Table S7 BUSCO gene-space completeness of the *Calycophyllum spruceanum* nuclear assembly and final draft assembly (sequences ≥1,000 bp after refined contamination screening), evaluated against three lineage datasets of increasing taxonomic specificity

	viridiplantae_odb12 (n = 822)		embryophyta_odb12 (n = 2,026)		eudicotyledons_odb12 (n = 2,805)	
BUSCO metric	**Nuclear**	**Final**	**Nuclear**	**Final**	**Nuclear**	**Final**
Complete (%)	89.7	83.8	83.4	82.6	84.2	83.9
Single-copy (%)	69.7	65.8	65.8	65.4	65.2	64.9
Duplicated (%)	20.0	18.0	17.5	17.2	19.1	19.0
Fragmented (%)	9.6	15.0	14.4	14.9	11.6	11.7
Missing (%)	0.7	1.2	2.3	2.5	4.1	4.4

BUSCO v6.0.0 was run in genome mode using miniprot. “Nuclear” refers to the assembly after first-pass decontamination, before length filtering; “Final” refers to the final draft assembly (sequences ≥1,000 bp after refined contamination screening). eudicotyledons_odb12 was used as the primary dataset throughout the main text, as the most specific dataset available for *Calycophyllum* (Rubiaceae). The decrease in viridiplantae_odb12 completeness between the Nuclear and Final assemblies reflects BUSCOs previously completed on microbial sequences (see Results: Length filtering, refined contamination screening and final draft assembly): of the 48 viridiplantae BUSCOs that lost complete status, 44 became fragmented and 4 missing, and all had been scored as complete on sequences later removed as microbial.

Table S8 Summary statistics of structural annotation metrics for the *Calycophyllum spruceanum* final draft assembly (GeMoMa v1.9), corresponding to Fig. 3

Metric	n	Mean	Median	Min–Max
Gene length (bp)	36,690	2,626.06	1,853	42–40,634
Transcript CDS total length (bp)	49,702	1,193.69	1,026	42–15,327
CDS features per mRNA	49,702	4.88	3	1–60
Predicted protein length (aa)	49,702	396.90	341	13–5,108

Gene length is calculated per gene feature; CDS total length, CDS count, and protein length are calculated per transcript (mRNA). 507 genes (1.4%) encoded proteins shorter than 50 aa; no protein contained internal stop codons.

Table S9 Twenty most represented KEGG pathways in the *Calycophyllum spruceanum* predicted proteome

KEGG ID	Pathway	Transcripts	% of total
map01100	Metabolic pathways	5,072	10.20
map01110	Biosynthesis of secondary metabolites	3,218	6.47
map04626	Plant–pathogen interaction	933	1.88
map04620	Toll-like receptor signaling pathway[a]	877	1.76
map04016	MAPK signaling pathway – plant	751	1.51
map04075	Plant hormone signal transduction	729	1.47
map00940	Phenylpropanoid biosynthesis	717	1.44
map00500	Starch and sucrose metabolism	582	1.17
map03010	Ribosome	532	1.07
map01200	Carbon metabolism	481	0.97
map04144	Endocytosis	452	0.91
map04141	Protein processing in endoplasmic reticulum	452	0.91
map03040	Spliceosome	396	0.80
map00040	Pentose and glucuronate interconversions	371	0.75
map01230	Biosynthesis of amino acids	357	0.72
map00520	Amino sugar and nucleotide sugar metabolism	346	0.70
map00460	Cyanoamino acid metabolism	326	0.66
map00190	Oxidative phosphorylation	324	0.65
map00380	Tryptophan metabolism	319	0.64
map04120	Ubiquitin mediated proteolysis	298	0.60

Transcripts assigned by eggNOG-mapper v2.1.15 to each KEGG reference pathway. Only pathways defined by KEGG for *A. thaliana* (list retrieved on 16 October 2026) and outside the Human Diseases category (map05xxx) are shown. Of the 396 reference maps assigned, 260 are not defined for *A. thaliana* or belong to Human Diseases and were excluded. Percentages are relative to the 49,702 predicted transcripts. [a]Listed by KEGG for *A. thaliana*; these assignments most likely reflect KO terms of conserved signalling components (e.g., ROP GTPases and MAP kinases) shared with this animal immune pathway, rather than a Toll-like receptor system.