Genome assembly of *Chiococca alba* uncovers key enzymes involved in the biosynthesis of unusual terpenoids

Kin H. Lau\(^{1,\dagger}\), Wajid Waheed Bhat\(^{2,\ddagger}\), John P. Hamilton\(^{1}\), Joshua C. Wood\(^{1}\), Brieannie Vaillancourt\(^{1}\), Krystle Wieger-Rininger\(^{1}\), Linsey Newton\(^{1}\), Britta Hamberger\(^{2}\), Daniel Holmes\(^{3}\), Bjoern Hamberger\(^{2,4,*}\), and C. Robin Buell\(^{1,4,5,*}\)

\(^{1}\)Department of Plant Biology, \(^{2}\)Department of Biochemistry and Molecular Biology, \(^{3}\)Department of Chemistry, \(^{4}\)MSU AgBioResearch, and \(^{5}\)Plant Resilience Institute, Michigan State University, East Lansing, MI 48824, USA

*To whom correspondence should be addressed. Tel. (517) 353-5597. Email: buell@msu.edu (C.R.B.); Tel. (517) 884-6964. hamberge@msu.edu (B.H.)*

\(^{\dagger}\)Present address: Bioinformatics and Biostatistics Core, Van Andel Institute, Grand Rapids, MI 49503, USA.

\(^{\ddagger}\)These authors contributed equally to this work.

Received 22 March 2020; Accepted 29 June 2020

Abstract

*Chiococca alba* (L.) Hitchc. (snowberry), a member of the Rubiaceae, has been used as a folk remedy for a range of health issues including inflammation and rheumatism and produces a wealth of specialized metabolites including terpenes, alkaloids, and flavonoids. We generated a 558 Mb draft genome assembly for snowberry which encodes 28,707 high-confidence genes. Comparative analyses with other angiosperm genomes revealed enrichment in snowberry of lineage-specific genes involved in specialized metabolism. Synteny between snowberry and *Coffea canephora* Pierre ex A. Froehner (coffee) was evident, including the chromosomal region encoding caffeine biosynthesis in coffee, albeit syntelogs of N-methyltransferase were absent in snowberry. A total of 27 putative terpene synthase genes were identified, including 10 that encode diterpene synthases. Functional validation of a subset of putative terpene synthases revealed that combinations of diterpene synthases yielded access to products of both general and specialized metabolism. Specifically, we identified plausible intermediates in the biosynthesis of merilactone and ribenone, structurally unique antimicrobial diterpene natural products. Access to the *C. alba* genome will enable additional characterization of biosynthetic pathways responsible for health-promoting compounds in this medicinal species.

Key words: *Chiococca alba*, genome, 10× linked reads, alkaloid, terpene synthase

1. Introduction

Species in the Rubiaceae family produce a wide range of specialized metabolites including alkaloids, terpenoids, and flavonoids.\(^{1-6}\) The most well-known compounds include the stimulatory alkaloid caffeine from *Coffea* species and the anti-malarial alkaloid quinine from *Cinchona*, a compound on the World Health Organization’s List of...
Essential Medicines. Lesser known is the medicinal shrub, Chiococca alba (L.) Hitchc. (Fig. 1A; also known as snowberry, ‘cabinca’, West Indian milkberry, and other regional names), that occurs naturally in tropical regions of North and South America. In Brazilian traditional medicine, C. alba is used to treat a wide variety of ailments. Its roots are used as a diuretic, an anti-viral, an anti-inflammatory, and as a treatment for rheumatism, hysteria, and snakebites. Although the use of the leaves is less frequent, leaves have been used to treat asthma, headaches, and diarrhea. Contemporary research has validated the use of C. alba as a folk medicine by confirming its anti-inflammatory and anti-microbial properties. Tests with the DNA-repair deficient Saccharomyces cerevisiae mutant, RS321, suggests that C. alba may have anti-cancer activity.

A large number of plant specialized metabolites have been isolated from C. alba (Fig. 1B) including lignans, coumarins, ketoalcohols, triterpenes, iridoids, quinoline alkaloids, flavonoids, and saponins. Chiococca alba is also a rich source of diterpenes, including merilactone and ribenone. Merilactone is a structurally unique C19 nor-diterpene found only in C. alba. It is potentially synthesized from a diterpene scaffold, followed by ring cleavage, carbon loss and lactonization. Ribenone has an unusual heteroatom-containing ring with demonstrated activity against Leishmania and potential anti-cancer activity. Chiococca alba also makes a number of kaurene type diterpenes like 1-hydroxy-18-nor-kaur-4,16-dien-3-one; 15-hydroxy-kaur-16-en-3-one, kaur-16-en-19-ol; kaurenoic acid; ent-17-hydroxy-16α-kauran-3-one and merilactone. Although considerable work has been done on the phytochemical and pharmacological aspects of C. alba, genes involved in the biosynthesis of merilactone, ribenone, and other ent-kaureno-derived C. alba diterpenes have not been identified.

Terpene compounds are prevalent throughout the Plant Kingdom and are classified based on the number of isoprene units in the scaffold; diterpenes have four isoprene units and are typically derived from geranylgeranyl diphasphate (GGPP) via diterpene synthases (diTPSs) yielding the diterpene skeleton, which can then undergo further modification by cytochromes P450, acyl transferases, or other enzymes. Bioactive diterpenes from C. alba fall into the labdane-related superfamily, whose biosynthesis is characterized by initial bicyclization of GGPP by a Class II diTPS from the TPS-c subfamily into a bicyclic prenyl diphasphate intermediate [e.g. copalyl/ labdadienyl diphasphate (CPP)]. The resulting intermediate undergoes further cyclization and/or rearrangement catalysed by Class I diTPS from the TPS-e subfamily that acts to remove the diphasphate moiety and can form additional rings, double bonds, or hydroxyl groups.

Despite the diverse biochemical profiles of Rubiaceae species and their medicinal importance, genome assembly efforts within the Rubiaceae family have been limited beyond those with Coffea. Here, we present a de novo draft-quality genome assembly for C. alba, which has a basic chromosome number between 12 and 14, consistent with a 2n = 2x ploidy level. We assembled 10 Genomes linked reads using the Supernova assembler, an approach that has recently been implemented in several systems including a soybean relative, a pepper F1 hybrid, wild perch fish, and human. This assembly and other genomic analyses facilitated the identification and functional characterization of five diTPSs (CaTPS1-5) involved in the biosynthesis of C. alba diterpenes. Access to the C. alba genome sequence along with validated terpene biosynthetic pathway genes will enable improved understanding and discovery of specialized metabolites in this medicinal plant species.

2. Materials and methods

2.1. Plant materials and growth conditions

Mature C. alba plants were procured from Sweet Bay Nursery (Parrish, FL, USA) and grown in a greenhouse under ambient photoperiod and 24°C day/17°C night temperatures. Nicotiana benthamiana plants were grown in a plant growth room under 16-h light (24°C) and 8-h dark (17°C) regime.

2.2. DNA and RNA library preparation and sequencing

Genomic DNA was isolated from young leaves of a mature plant following a modified cetyl trimethylammonium bromide method. Illumina-compatible whole-genome sequencing (WGS) libraries were made from young leaf and sequenced on an Illumina HiSeq 4000 (Illumina, San Diego, CA, USA) in paired-end mode to 150 nt. Young leaf mate-pair libraries were prepared using the Nextera Mate Pair Library Prep Kit (Illumina, San Diego, CA, USA) and grown in a greenhouse under ambient photoperiod and 24°C day/17°C night temperatures. Nicotiana benthamiana plants were grown in a plant growth room under 16-h light (24°C) and 8-h dark (17°C) regime. Total RNA was isolated from mature root and mature leaf as described previously. Mature leaf and root RNA-Seq libraries were constructed by first purifying mRNA using the Dynabeads™ mRNA

![Figure 1. Chiococca alba. (A) Chiococca alba (snowberry) with fruit (left) and flower (right). (B) Chemical diversity of C. alba terpenoids. Representative monoterpenes, seco-iridoid glucoside Alboside I, triterpene Chiococcasaponin IV, diterpene kaurane-type ent-kaurenoic acid, nor-kaurane-type 1-hydroxy-18-nor-kaur-4,16-dien-3-one, ent-manoyl oxide-type Ribenone, plausibly nor-seco-pimarane-type Merilactone.](image-url)
DIRECT™ Purification Kit (Thermo Fisher Scientific, Waltham, MA, USA) followed by the KAPA Stranded RNA-Seq Kit (Roche, Basel, Switzerland) with NEBNext® Multiplex Oligos for Illumina® (New England Biolabs, Ipswich, MA, USA). RNA-Seq libraries were sequenced on either an Illumina HiSeq2500, HiSeq 4000, or NextSeq500 in paired-end mode to 150 nt. Total RNA was extracted from leaf and root of three independent mature C. alba plants using the Spectrum™ Plant Total RNA Kit (Sigma-Aldrich, St Louis, MO, USA). Residual DNA contamination was removed using the TURBO DNA-free™ kit (Thermo Fisher Scientific, Waltham, MA, USA). RNA-Seq libraries were prepared using the Illumina TrueSeq Stranded mRNA Library Preparation Kit with IDT for Illumina Unique Dual Index adapters. All libraries were sequenced on an Illumina HiSeq 4000 in single end mode to 50 nt. All RNA-Seq libraries generated in this study are listed in Supplementary Table S2.

2.3. Genome size estimation and k-mer analysis of the C. alba genome

Genome size of C. alba was determined using flow cytometry at the Benaroya Research Institute at Virginia Mason in Seattle, WA, USA. Paired-end whole-genome shotgun sequencing reads for the ALLPATHS-LG assembly (RUB_AI, RUB_AI_02 and RUB_AZ) were error-corrected using ALLPATHS-LG (v52488) and the counts of canonical k-mers in the sequencing reads was generated using Jellyfish (v2.2.0). Analysis of the k-mer counts was performed using GenomeScope (http://qb.cshl.edu/genomescope/).

2.4. Genome assembly and scaffold filtering

For the ALLPATHS-LG assembly (hereafter referred to as the AP assembly), paired-end and mate-pair reads were trimmed using Cutadapt (v1.11) with a quality cut-off of 15 and minimum read length of 25 nt. Mate-pair library reads were further processed with NextClip (v1.3.1) and only categories A, B, and C (R1 and/or R2 contained the junction adapter) were retained for assembly. ALLPATHS-LG (v52488) was run with default parameters with three WGS libraries with total reads equivalent to 86× coverage and two mate-pair libraries, with total reads equivalent to 30× coverage.

For the 10× Genomics assembly (hereafter referred to as 10×), Supernova (v2.0.0) was run with 310 M 150 nt reads from a single 10× library, equivalent to 70× raw coverage and 48× effective coverage after accounting for duplicated reads, as calculated by Supernova. FASTA files were extracted from the raw assembly using the ‘mksoput’ function (Supernova v2.0.1) with the ‘psedoap2’ style and minimum scaffold size of 1 kb; pseudohap 1 was chosen for downstream analysis. Redundant scaffolds were removed using the ‘redundancy reduction’ module of Redundans, which performs all-versus-all self-assembly of scaffolds using LAST. Redundans was run multiple times to identify optimal identity and coverage cut-offs, which were specified using ‘-identity’ and ‘-overlap’, respectively.

Mean scaffold read depth values were calculated from alignments of the 10× library to the different genome assemblies using BWA-MEM (bwa v0.7.12) with the ‘-M’ option followed by removal of duplicate reads using MarkDuplicates (picardTools v2.9.2; http://broadinstitute.github.io/picard/); the total read bases aligned to each scaffold were calculated using SAMtools (v1.4) bedcov and divided by the length of each scaffold minus gaps. Identity and coverage cut-offs that best balanced collapsing split haplotypes without collapsing paralogous regions (see Results) were chosen for the final filtered 10× assembly. To complement the mean scaffold read depth analysis, self-alignments of the genomes were also used to visualize the decrease in inter-scaffold hits as redundant scaffolds were removed; these genome alignments were performed using nuclmer (MUMmer v3.2.3) with the parameters ‘-maxmatch’ and ‘-minmatch 200’, followed by filtering for alignment length and identity using the ‘delta-filter’ function.

Scaffolds in the filtered 10× assembly were queried against the NCBI nt database (downloaded 1 May 2018) using blastn (BLAST+ v2.6.0) with default parameters except ‘-max_target_seqs’ was set to 100000. Despite using a relatively lenient filter of E-value < 4e-40 and Query Coverage Per Subject > 10%, only one non-Viridiplantae scaffold which was the PhiX sequencing control (NC_001422.1) was detected. To remove chloroplast sequences, scaffolds were queried against Rubiaceae chloroplast genomes downloaded from NCBI (Supplementary Table S3) and a nucleotide BLAST search was performed using filters of ‘Query Coverage Per Subject’ > 97, ‘Query Coverage Per HSP’ > 50 and identity > 97.

2.5. Genome assembly quality assessment

Standard sequence content and contiguity metrics were obtained using the ‘assembleathon_stats.pl’ from Assemblathon 2 (45) using gaps of 2 or more Ns to delineate contigs within scaffolds. BUSCO (v3.02.2b) was run using the ‘embryophytaodb9’ database (1440 BUSCO groups) in genome mode for the AP assembly, the unfiltered 10× assembly and the filtered 10× assembly, and in transcript mode for the longest ORF transcript of each gene annotated in the filtered 10× assembly.

Genomic WGS DNA reads were aligned as described above. RNA-Seq reads were aligned to the genome assemblies using HISAT2 (v2.1.0). Alignment metrics were obtained using SAMtools flagstat and Picard CollectAlignmentSummaryMetrics (picardTools v2.9.2; http://broadinstitute.github.io/picard/).

2.6. Gene annotation

A C. alba custom repeat library (CRL) was created with RepeatModeler (v1.0.8; http://repeatmodeler.org) using the unfiltered 10× assembly. The CRL was searched against a curated library of plant protein-coding genes and sequences, removing matches with ProtExcluder (v1.1) and then combined with the RepBase (v20150807) Viridiplantae repeats to create a final CRL. The filtered 10× assembly was then assembled with RepeatMasker (v4.0.6; http://repeatmasker.org) using the CRL with the ‘-s’, ‘-olow’, and ‘-no_is’ options, resulting in 260.89 Mb (47%) of masked sequence.

For subsequent gene annotation steps requiring RNA-Seq align-ments, only paired-end RNA-Seq libraries were used. Read alignments to the filtered 10× assembly were obtained using TopHat2 (v2.1.1) with the parameters ‘-min-intron-length 20’ and ‘-max-intron-length 20000’ in stranded mode, merging the alignments for the root libraries. Genome-guided transcript assemblies were constructed with these alignments using Trinity (v2.3.2) with the parameters ‘genome_guided_max_intron 10000’ and ‘--SS_lib_type RF’, filtering away transcripts shorter than 500 bp.

Ab initio gene prediction was performed by training AUGUSTUS (v3.2.2) on the soft-masked assembly using the leaf RNA-Seq alignments. Initial gene predictions were then generated using AUGUSTUS and the hard-masked assembly. The initial gene predictions were then refined using PASA2 (v2.0.2) utilizing the genome-guided transcript assemblies as evidence. High-confidence gene models were defined as transcripts with a best PFAM (v31) hit with sequence E-value < 1e-5 and domain E-value ≤ 1e-3 as identified using HMMER (v3.1b2) or having Fragments Per Kilobase of
transcript per Million mapped reads (FPKMs) > 0 in either the root or the leaf dataset as calculated using Cufflinks2 (v2.2.1)44 with the parameters ‘–multi-read-correct’ and ‘–max-intron-length 20000’ in stranded mode. Transcripts with a significant (based on above thresholds) best PFAM hit that was a transposable element-related domain were also removed.

Expression abundances from replicated leaf and root RNA-Seq libraries were generated to examine terpene synthase (TPS) genes involved in diterpene biosynthesis. The longest predicted peptide isoforms of each gene were filtered using Cutadapt (v1.18)34 with a quality cut-off score of 20 and a minimum read length of 30 nt. The RNA-Seq reads were then aligned to the filtered 10x assembly using HISAT2 (v2.1.0)45 in stranded mode with a minimum intron length of 20 and a maximum intron length of 60,000. FPKMs were generated using Cufflinks (v2.2.1)54 and SAMTools (v1.9)40 with the parameters ‘–multi-read-correct’, ‘–min-intron-length 10’, and ‘–max-intron-length 60000’ in stranded mode.

2.7. Comparative genome analyses
To identify homologous genes for syntenic analysis, BLASTP (BLAST+ v2.6.0)52 was run using the longest peptide isoform for each gene, comparing C. alba to itself, Coffea canephora to itself and C. alba versus C. canephora in both directions. BLAST hits were filtered for E-value ≤ 1e-5 and only the top five hits (based on bit score) for each query sequence were retained. These BLAST results were used to identify syntic blocks using MCScanX v2017032233 with default parameters and visualized using SynVisio (https://synviso.github.io). The longest predicted peptide isoforms of C. alba, C. canephora,24 Vitis vinifera (grapevine),39 Arabidopsis thaliana,37 and Amborella trichopoda28 (grapevine, Arabidopsis, and Amborella data were downloaded from Phytozome v12)39 were analysed using Orthofinder v2.2.750 with default settings except ‘-M msa’. Overlapping orthogroups membership was visualized using UpSet v1.3.3.61 Gene ontology (GO) terms were annotated using InterProScan v5.28.67.0.52 GO term enrichment for genes in C. alba-specific orthogroups (including orthogroups containing only one C. alba gene) was tested using Fisher’s exact tests implemented in topGO v2.34.0.44. Peptide sequences were aligned using Clustal Omega (v1.2.1) and a maximum likelihood tree was generated using MEGA version X.64 The Poisson correction parameter and pairwise deletion of gaps were applied. The reliability of branching was assessed by the bootstrap re-sampling method using 1000 bootstrap replications.

2.8. Identification and cloning of candidate genes involved in C. alba diterpene biosynthesis
Initial efforts to identify TPS genes in C. alba utilized de novo transcript assemblies generated by trimming RNA-Seq reads using Cutadapt (v1.11)34 with a quality cut-off of 30 and minimum read length of 30 nt and running Trinity (v2.3.2)49 with ‘–normalize_-max_read_cov 50’ and ‘–SS_lib_type RF’, retaining only transcripts that were >500 bp. To identify terpene synthase-related genes in C. alba in the initial de novo transcript assemblies and the C. alba genome, we utilized reference TPS peptides (mono-, sesqui-, and di-TPS) curated from multiple plant species by Terzemy35 and a list of P450s known to be involved in terpene synthesis (Supplementary Table S4). Reference terpene synthase genes were compared with predicted peptides in C. alba using BLASTP (BLAST+ v2.6.0.0) and the results were filtered for E-value ≤ 1e-10, subject coverage ≥ 80%, and identity ≥ 40%. To identify the precise genome coordinates for the TPS de novo Trinity transcripts in the C. alba genome, GMAP (v20160401)66 was used to align the longest isoform from each ‘gene’ as determined by Trinity to the genome assemblies; ‘–no-chimeras’ was set to avoid chimeric alignments, ‘-trimendexons 0’ was used to prevent trimming of alignment ends, and coverage and identity cut-offs were set using ‘–min-trimmed-coverage’ and ‘–min-identity’ respectively. Candidate genes were amplified from single-stranded cDNA generated from total root RNA using the RevertAid First Strand cDNA Synthesis Kit (Thermo Fisher Scientific, Waltham, MA, USA) with oligo(dT) primers. Putative TPS cDNAs were cloned into pJET1.2 vector (Thermo Fisher, USA). All primers used are listed in Supplementary Table S5.

2.9. Functional characterization of C. alba TPS candidates in N. benthamiana
For functional characterization of putative TPS genes in N. benthamiana, full-length coding sequences (CDS) of sequence-verified TPS genes were cloned into the pEAQ-HT vector57 (kindly provided by Prof. G. Lomonossoff, John Innes Centre, UK) using In-Fusion6 HD Cloning Plus (Takara Bio, CA, USA). The N. benthamiana system is considered to be highly efficient for characterization of TPSs largely due to native codon usage, presence of subcellular compartment for localization of diTPS, synthesis of the common precursor GGPP, and flexibility of using different combinations of TPSs as well as the ease of upscaling the infiltration process for production of milligram scale terpenoid compounds for structural elucidations.22,68–70 Thus, TPSs were transiently co-expressed in N. benthamiana together with the suppressor of gene-silencing p19,71 Coleus forskohlii (syn. Plectranthus barbatus) enzymes 1-deoxy-d-xylulose 5-phosphate synthase (CfDXS) and GGPP synthase (CfGGPPS), according to a previously established protocol72; coexpression of the upstream genes, CfDXS and CfGGPPS substantially increase the level of the substrate GGPP.22

Five to eight days post-infiltration, metabolites of three leaf disks (3-cm diameter, 1 disk per leaf) were extracted in 1 ml n-hexane with 1 mg/l 1-ecosanol as internal standard at room temperature overnight in an orbital shaker at 200 rpm. Plant material was collected by centrifugation and the organic phase transferred to GC vials for analysis. As controls, p19, CfDXS, and CfGGPPS were co-expressed with and without individual candidate genes. The functionally characterized reference diTPSs ZmAM2 (Zea mays),73 NmTPS2 (Nepeta mussinii),22 TWTPS21 (Tripterygium wilfordii),69 ZmSKL4 (Z. mays),74 EPTPS3,68 as well as CfTPS2 and CfTPS3 (C. forskohlii)70 were included for comparison.

2.10. Analytical procedures
All gas chromatography–mass spectrometry (GC-MS) analyses were performed on an Agilent 7890A GC with an Agilent VF-5ms column (30 m × 0.25 μm × 0.25 μm, with 10 mm EZ-Guard) and an Agilent 5975C detector. The inlet was set to 275°C (30 m) and 280°C (30 m), hold 3 min. For GC-MS-based root and leaf metabolomics, 500 mg of finely ground fresh roots and leaves were extracted for 24 h in 5 ml n-
hexane. The extracts were concentrated to 1 ml under N₂ and ana-
lysed by GC-MS using Method B as described above.

For ultra-high-performance liquid chromatography/mass spec-
trometry (UHPLC/MS) metabolomic analysis, 500 mg of fresh leaf 
and root tissues were finely ground and extracted with 10 ml 80% 
methanol by incubating in the dark at room temperature for 24 h. 
The extracts were concentrated down to 1 ml by blowing in N₂ gas. 
A 10 µl volume of each extract was subsequently analysed using a 
15 min gradient elution method on an Acquity BE C18 UHPLC 
column (2.1 x 100 mm, 1.7 µm, Waters) as previously described.65

2.11. Compound purification and NMR

For the production of diterpene compound sufficient for structural 
analysis by nuclear magnetic resonance (NMR), we used a large-
scale N. benthamiana agroinfiltration system as previously de-
described.68 Briefly, 25 N. benthamiana plants were vacuum co-
filtrated with a combination of CaTPS1/CaTPS5 and CfGGPPS/ 
CfDXS. After 5 days, 250 g (fresh weight) of agroinfiltrated leaves 
were finely ground using a blender and subjected to two rounds of 
night extractions in 500 ml hexane. The extract was dried on a 
rotary evaporator. The diterpene compound was purified from the 
resin using silica gel flash column chromatography with a mobile 
phase of 5% ethyl-acetate in hexane.

NMR experiments were performed on an Agilent DDR2 spec-
trometer operating at 499.90 MHz equipped with a PFG OneNMR 
Probe or on a Varian Inova spectrometer operating at 599.73 MHz 
-equipped with an indirect PFG HCN probe. Experiments were per-
formed without spinning and at ambient temperature unless other-
wise noted. The following parameters were used for 2D NMR 
experiments: NOESY: 16 transients, 200 increments, 500 ms mixing 
time; gCOSY: 8 transients, 256 increments; gHSQC: 8 transients, 
128 increments; gHMBC: 8 transients, 256 increments. Standard 
vendor supplied processing was performed on all datasets including 
zero filling, linear prediction, baseline correction, and apodization. 
CDCl₃ peaks were referenced to 7.26 and 77.00 ppm for ¹H and ¹³C 
spectra, respectively.

Optical rotation was measured with a Perkin Elmer Polarimeter 
341 (Überlingen, Germany) using a microcell (pathlength 100 mm, 
volume 1 ml). Measurements were carried out at 20°C, wavelength 
589 nm.

3. Results and discussion

3.1. Genome assembly and removal of redundant scaffolds

The haploid genome size of C. alba determined by flow cytometry 
was 567 Mb. k-mer analyses predicted a smaller genome size varying 
from 480 to 537 Mb with heterozygosity rates ranging from 0.37% 
to 0.49% depending on the k-mer size (Supplementary Fig. S1). 
Two genome assembly approaches were performed: a traditional 
approach using paired-end and mate-pair read assembly with the 
ALLPATHS-LG assembler31 and a newer method using a 10× 
Genomics linked read assembly using the Supernova assembly 
algorithm.36

As expected, the 10× assembly outperformed the AP assembly 
across all basic contiguity metrics, achieving a scaffold N50 of 
1.75 Mb versus 103 kb in the AP assembly, and a longest scaffold 
length of 10.1 Mb versus 727 kb (Table 1 and Supplementary Table 
S6). The 10× assembly had a total assembly size of 656 Mb com-
pared with 534 Mb in the AP assembly, whereas flow cytometry 

| Table 1. Assembly and annotation metrics of the C. alba genome |
|----------------|----------------|
| Metric         | Number         |
| Number of scaffolds          | 3,518          |
| Total size of scaffolds       | 557,689,917    |
| Longest scaffold             | 10,139,299     |
| Shortest scaffold            | 5,000          |
| Mean scaffold size           | 158,525        |
| Median scaffold size         | 10,194         |
| N50 scaffold length          | 2,354,703      |
| L50 scaffold count           | 63             |
| No. high-confidence genes    | 28,707         |
| No. high-confidence gene models | 43,217       |

estimated the genome as 567 Mb, suggestive that the 10× assembly 
has retained ‘split haplotypes’. We tested this by performing 
self-alignments of the assemblies (Supplementary Fig. S2). The 10× 
assembly had visibly more self-homology compared with the AP 
assembly, with duplicated sequences more prevalent in shorter 
scaffolds. Assessment of mean scaffold read depth revealed a bimodal 
distribution in the 10× assembly, with the more prominent peak 
collapsed haplotypes) at roughly double the read depth of the sec-
tary peak (split haplotypes), further confirming the presence of 
split haplotypes in the 10× assembly.

Sequence redundancy was also observed in a Supernova assembly 
of perch, in which 66.19 Mb of scaffolds was removed from an initial 
assembly of 1,024.4 Mb using filter thresholds of ≥ 99% identity 
and ≥ 95% coverage in alignments with other scaffolds in the assem-
ly.34 Thus, we implemented a similar approach for C. alba, by first 
removing scaffolds smaller than 5 kb then self-aligning the 10× 
assembly and filtering out redundant scaffolds. In addition, we added 
read depth analyses to inform the selection of identity and coverage 
cut-offs (Supplementary Fig. S3). Thresholds of 80% identity and 
95% coverage provided a balance between reducing the split haplo-
types and limiting the collapsing of paralogous regions (evidenced by 
read depth higher than the more prominent peak). From this assem-
ly, we also removed three scaffolds composed of only n totalling 
9,000 bp, an 8,200 bp scaffold corresponding to the PhiX sequencing 
control, and three chloroplast scaffolds totalling 138,637 bp; subse-
quent references to the ‘filtered 10× assembly’ refer to this assembly 
unless otherwise noted.

3.2. Assessing genome assembly completeness

Aligning WGS and RNA-Seq reads indicated comparably high levels 
of completeness in the AP assembly and the filtered 10× assembly 
with alignment rates of ~98% for the WGS libraries and between 
92% and 95% for the RNA-Seq libraries when aligned to either as-
seembly (Supplementary Tables S7 and S8). However, rates of prop-
ely paired read alignments emphasized the superior contiguity in the 
10× assembly, at ~98% for the WGS libraries and 89% for the 10× 
library (Supplementary Table S7).

To assess if informative scaffolds were inadvertently removed by 
efforts to filter out redundant scaffolds in the 10× assembly, read 
alignment rates were also obtained for the pre-filtered assembly. 
Comparison of the raw 10× assembly to the filtered 10× assembly, 
yielded a reduction of alignment rates by 0.6–1%. Conversely, high-
quality alignments (MAPQ ≥ 20, as defined by PicardTools) were 
increased by 0.8–0.9% for the WGS libraries, presumably due to the 
removal of highly similar, redundant target sequences in the filtered
10× assembly. Interestingly, high-quality alignments of the 10× library were decreased by 1% in the filtered 10× library. This decrease in high-quality alignments may be connected with a 2.4% decrease in properly paired reads compared with a 0.05–1.1% decrease for the WGS libraries; we speculate that the larger insert size of the 10× library may make paired alignments more haplotype-specific. Overall, alignment rates of RNA-Seq reads were reduced by 1.5–2.8%, but high-quality alignment rates decreased by only 0.4–1.1%. Consistent with the hypothesis that the scaffold filtering had minimal impact on the gene content of the 10× assembly, alignment of de novo assembled transcripts were a mere ~1% lower in the filtered 10× assembly, and missing BUSCOs increased by 0.7% (Supplementary Table S9). As to genic content, 92.2% complete, 2.0% fragmented, and 5.8% missing BUSCO orthologues support that the 10× assembly is robust (Supplementary Table S9). Overall, the metrics for the filtered 10× assembly indicate high quality and completeness for a draft genome assembly.

3.3. Genome annotation and orthologous relationships with other plant species

The filtered 10× assembly was annotated using an ab initio gene finder trained with RNA-Seq alignments followed by refinement of gene models using genome-guided transcript assemblies as evidence. A total of 34,878 genes encoded 49,586 gene models were annotated (Supplementary Table S10). The mean gene length and mean CDS length were 3.4 and 1.2 kb, respectively, with a mean of 5.55 exons per gene model. Filtering of the high-confidence gene set based on expression and protein domain signatures (see Materials and methods) led to a set of 28,707 high-confidence genes encoding 43,217 gene models. BUSCO analysis on the annotated genes raised the score to 93.2% complete, 2.6% fragmented, and 4.2% missing (Supplementary Table S9).

We identified syntenic blocks between the snowberry scaffolds and the C. canephora genome; a total of 750 blocks with 21,697 syntenic gene pairs were identified (Fig. 2A). To explore conservation of the caffeine N-methyltransferase (NMT) cluster in caffeine, we searched for C. alba regions with synteny to the coffee NMT genes. C. alba scaffolds 327 and 265 harboured syntenic blocks of 24 and 11 genes, respectively (Fig 2B and Supplementary Table S11) in which syntelogs of the NMT genes are absent. In fact, C. alba NMT orthologues predicted by OrthoFinder2, g2037.t1 and g34221.t1, are on separate scaffolds (343 and 560, respectively). Instead of Chromosome 9 of C. canephora, the longest syntenic blocks for scaffolds 343 and 560 were both matches to Chromosome 8 (Supplementary Table S11). Together, these results suggest that the caffeine NMT cluster is absent in C. alba reflecting the unique evolution of caffeine synthesis in the Coffea genus within the Rubiaceae family.76

In an orthogroup analysis of C. alba, C. canephora, V. vinifera, A. thaliana, and Amborella, 66% (9,318/14,362) of orthogroups containing at least two orthologues/paralogues were conserved across all five species (Supplementary Fig. S4). Consistent with evolutionary relatedness, the next highest intersections were between the two Rubiaceae species, C. alba and C. canephora, and between all four non-Amborella species at 888 and 797 orthogroups, respectively. GO term analysis of paralogous groups specific to C. alba, totalling 4,765 genes, highlighted genes that may be involved in C. alba-specific plant architecture and specialized metabolism, with enrichment in terms including ‘oxidation-reduction process’, ‘amino metabolic process’, ‘cellulose microfibril organization’, ‘quinone binding’, and ‘iron ion binding’ (Supplementary Tables S12 and S13). Additionally, there was an enrichment of ‘viral capsid’ and ‘viral process’ genes, which had homology to geminivirus protein domains. Extensive collinearity of two scaffolds harbouring these genes compared with the C. canephora genome indicated that these are viral integration events and not due to a contaminated sample (Supplementary Fig. S5).

Among orthogroups containing C. alba TPS genes, gene counts were similar between C. alba and C. canephora both when summed across orthogroups based on TPS type and within individual orthogroups (Fig. 3). Mild gene family expansions were observed in C. canephora in a monoTPS orthogroup (OG0003358) and a sesquiTPS orthogroup (OG0006883), with four coffee paralogues to one C. alba orthologue in both cases. For both orthogroups, the C. canephora expansion was driven by tandem duplication (Supplementary Table S14). The most remarkable result among all species was a dramatic expansion in a sesquiTPS orthogroup (OG0000030) in V. vinifera, with 45 genes compared with 11 in both C. canephora and C. alba, the next highest species. Interestingly, previous studies in grapevine identified distinct sesquiterpene profiles in different cultivars that are speculated to be caused by TPS variants77; expansion in the OG0000030 sesquiTPS orthogroup may have facilitated this functional diversity.

3.4. Terpene synthases in C. alba

To gain a deeper understanding into the biosynthesis of bioactive diterpenoids in C. alba, we mined its genome for TPSs using the previously described phylogeny-guided gene discovery strategy.69,78–81 We identified 27 putative TPS genes, of which, ten were classified as diTPSs (Fig. 4 and Supplementary Table S15). Of the identified diTPS genes, four (CaTPS1, 2, 6, and 7) were predicted as Class II diTPSs (TPS-c subfamily) based on the presence of characteristic DxD signature motif.82 Conversely, CaTPS3, 4, 5, 8, and 9 were classified as Class I diTPSs (TPS-e subfamily) as they featured the conserved DxDD and NDXX2(S/T)X3E motifs known to be relevant for catalysis in Class I diTPSs. CaTPS10 was predicted to be a member of TPS-f sub-family, members of which are involved in the biosynthesis of linear chain diterpenes including geranylannolool (TPS4; Fig. 5).83 Of the ten diTPSs identified in the C. alba, only the Class II diTPSs, CaTPS1, and CaTPS2 and the Class I diTPSs, CaTPS3, CaTPS4, and CaTPS5 could be retrieved as full-length cDNA for further functional analysis based on presence of a start and stop codon and alignments to other known, validated TPSs. To determine the enzymatic activity of these diTPSs, we conducted in vivo combinatorial assays using the transient Agrobacterium-mediated co-expression in N. benthamiana.22,68

3.5. Functional characterization of C. alba diTPSs

Combinatorial expression of Classes II and I diTPSs has been widely used to determine the identity and stereochemistry of enzyme products.22,68,81 Taking advantage of this modular pairwise activity of Classes II and I diTPSs in angiosperm labdane biosynthesis, we tested the C. alba diTPSs in combination with functionally characterized Classes II and I reference diTPSs. To investigate the function of CaTPS1 and CaTPS2, we transiently expressed them in combination with the Class I diTPS N. mussini ent-kaurene synthase which converts ent-CPP to ent-kaurenene (1),22,68 C. forskohlii miltiradiene synthase (CFTPS3), catalysing cyclization of (+)-CPP into miltiradiene,70 and Salvia sclarea sclareol synthase (ScSS), a promiscuous enzyme converting ent-CPP to ent-
manool, (+)-CPP to (+)-manool or labdenol diphosphate (LPP) to sclareol. As a reference and for product identification, we also expressed a suite of known Class II diTPS, (+)-CPS (CfTPS1), Z. mays ent-CPS (ZmAN2), T. wilfordii ent-8-LPP (TwTPS21), and C. forskohlii (+)-8-LPP (CfTPS2). This comparison enabled verification of the stereo-selectivity of CaTPS1 and CaTPS2.

CaTPS1 yielded a diterpene with identical retention time and mass spectrum to the product of ZmAN2, illustrating that the primary product of CaTPS1 is CPP [(1), Fig. 5]. The stereochemistry of CaTPS1 was investigated by co-expression of CaTPS1 with NmTPS2, CfTPS3 or SsSS. Combination of CaTPS1/NmTPS2 yielded a single major peak (Supplementary Fig. S6a) that matched the ZmAN2/NmTPS2 product (Supplementary Fig. S6b), ent-16-kaurene (3), establishing CaTPS1 as ent-CPP synthase (Supplementary Fig. S6). CfTPS3 is a Class I diTPS specific for diphosphate substrates in normal (+) configuration. The combination of CaTPS1 co-expressed with CfTPS3 did not yield a diterpene product. When co-expressed with SsSS, which has broad substrate specificity, the activity of CaTPS1 yielded manool in ent-configuration (Supplementary Fig. S7). The product was identified by comparison with the authentic standard afforded by the reference diTPSs ent-CPP synthase (ZmAN2) and SsSS. These results establish CaTPS1 as ent-CPP synthase.

The product was identified by comparison with the authentic standard afforded by the reference diTPSs ent-CPP synthase (ZmAN2) and SsSS. Ent-CPP synthases are involved in the biosynthesis of gibberellins and ent-kaurene-derived specialized metabolites.

Co-expression of CaTPS2 and NmTPS2 generated a product with a retention time and mass spectra matching to that of the reference combination TwTPS21/EpTPS1 product, (13R)-ent-manoyl oxide (4) (Supplementary Fig. S8c and d). To support the stereochemistry, we compared against the combination of CfTPS2/CfTPS3 yielding the stereoisomer, (13R)-manoyl oxide in normal configuration (Supplementary Fig. S8e). The retention time of the CfTPS2/CfTPS3 was distinct, indicating that CaTPS2 affords stereoselectively (5R, 8S, 9S, 10S)-labda-13-en-8-ol diphosphate [ent-8-LPP (2)] (Fig. 5 and Supplementary Fig. S8).
Class I diTPSs, CaTPS3-5 were tested in combination with CaTPS1, CaTPS2, and four reference Class II enzymes. Substrates accepted by each enzyme and the products are given in Supplementary Figs S6–S8. When combined with either of the 13R-CPP synthases, (CaTPS1 or ZmAN2), CaTPS3 and CaTPS4 resulted in the formation of 13R-kaurene (3) as identified by comparison to the reference combination of ZmAN2/NmTPS22 (Fig. 5 and Supplementary Fig. S6c–f), supporting a function of CaTPS3 and CaTPS4 as 13R-kaurene synthases. The diTPSs CaTPS3 and CaTPS4 share over 95% identity and their genes seem to have evolved as a result of local tandem duplication as they are directly adjacent in the snowberry genome (Supplementary Table S15).

Expression of CaTPS3 and CaTPS4 is partially overlapping. CaTPS3 is equally expressed in both leaf and root tissues (mean FPKM leaf 6.2 vs. root 5.8) whereas mean FPKM values for CaTPS4 were higher in root compared with leaf tissue (root 8.5 to leaf 6.3; Fig. 5). Since C. alba produces a suite of 13R-kaurene-derived specialized metabolites in the roots (Fig. 5), it is possible that both diTPS contribute to their biosynthesis. Since C. alba produces a suite of 13R-kaurene-derived specialized metabolites in the roots (Fig. 5), it is possible that both diTPS contribute to their biosynthesis, as well as the formation of gibberellin phytohormones, sharing the same diterpene scaffold.

3.6. CaTPS2 yields access to the precursor of ribenone

The combinatorial assay of CaTPS2 (ent-8-LPP synthase) with CaTPS3 and CaTPS4 yielded (13R)-ent-manoyl oxide (4) identified by comparison with the reference product formed by TwTPS21/EpTPS1 (Supplementary Fig. S8a and b). The configuration and stereochemistry of (13R)-ent-manoyl oxide is consistent with ribenone [3-keto-(13R)-ent-manoyl oxide, (7)]—a major C. alba diterpene, indicating that the biosynthetic route for ribenone may involve the combination of CaTPS2 and either CaTPS3 or CaTPS4. Accumulation of (13R)-ent-manoyl oxide (4) in C. alba, or Rubiaceae, has not been reported. We analysed root and leaf extracts of C. alba by GC-MS and confirmed the presence of (13R)-ent-manoyl oxide (4) in the root tissues alone (Supplementary Fig. S8f), consistent with the presence of ribenone and biosynthesis proceeding through ent-16-kaurene (3; Supplementary Fig. S9). CaTPS2 was highly expressed in leaves and moderately in root tissues of C. alba (Fig. 5) indicating that the encoded diTPS CaTPS2 may also supply ent-8-LPP to yet unidentified diterpenes in C. alba, beyond the biosynthesis of ribenone in C. alba roots through CaTPS4.

Figure 3. Gene counts of orthogroups containing C. alba genes predicted to be involved in terpene synthesis. Orthogroups were classified as mono-, sesqui-, or diTPSs by collapsing the TPS annotation for C. alba gene members. OG0000211 contained both sesqui- and mono-TPS genes.
3.7. CaTPS5 is a Class I diTPS catalysing formation of epi-dolabradiene

CaTPS5 only showed substantial activity with ent-CPP as substrate. When co-expressed with CaTPS1 or the reference enzyme ZmAN2, CaTPS5 converted ent-CPP into a single product (5) with a fragmentation pattern that matched retention time and mass spectrum of dolabradiene, product of a recently identified TPS (ZmSKL4) from *Z. mays* (Fig. 5 and Supplementary S10b). Dolabradiene is known to occur in two configurations, the 13(S)-dolabradiene and its 13(R)-epi-dolabradiene stereoisomer, which have been reported across different families including Araucariaceae, Euphorbiaceae, Rubiaceae, Cupressaceae, and liverworts.\(^{86-89}\) To determine the configuration at C-13, we scaled up the *N. benthamiana* transient expression. Hexane extract from about 250 g (fresh weight) *N. benthamiana* leaves infiltrated with constructs carrying CaTPS1, CaTPS5, and CjDXS/CjGGPPS enabled us to purify the product by silica chromatography. Chemical shift assignments were determined through analysis of \(^1H, \quad ^{13}C, \quad gHSQC, \quad gHMBC, \quad gCOSY,\) and NOESY NMR data (Supplementary Table S16 and Figs S11–S17). Relative stereochemistry was determined by inspection of NOESY data, energy minimized 3D structural models, and \(^13C\) chemical shift data. Key to relative stereochemical assignment were nuclear Overhauser effect (NOE) correlations between protons on C19 and C20 (1.07 and 0.75 ppm, respectively), establishing a *cis* configuration for those methyl groups. The relative stereochemistry of the vinyl group (C15/16) and methyl (C17) was determined by the presence of a NOE cross-peak between the vinylic proton (5.80 ppm, H15) and the methyl (0.75 ppm, H20; Supplementary Fig S17). Further evidence of this configuration is a strong NOE interaction of H15 and H16a, b (larger cross-peak on H16a) with H14b (1.35 ppm). The large (12.7 Hz) coupling constant to H8 is consistent with an axial position of H14b and H8, and consequently a position

![Phylogenetic tree of TPSs](image-url)
of H14b on the same face as C20 due to large NOE correlation to H15. Finally, the carbon chemical shifts of the vinyl and methyl groups (151.5/108.5 and 23.03 for C15/C16 and C17, respectively) are consistent with an axial methyl group and equatorial vinyl, and the stereochemical configuration as 13-epi-dolabradiene.90 Polarimetric analysis yielded the optical rotation of [α]D +86.25°/C14 (c. 0.0016, dimethyl sulfoxide [DMSO]), in consonance with the earlier reported value of [α]D +86,88 while the optical rotation for dolabradiene was reported as [α]D/C0 70.88 This configuration is consistent with the earlier described stereochemistry of merilactone.13

4. Summary and conclusions

We generated a high-quality draft genome sequence of the medicinal plant, C. alba, using 10× Genomics linked read technology. Annotation of the genome revealed 27,707 high-confidence genes with robust synteny with the caffeine-producing coffee genome. We were able to annotate 27 TPSs and identify genes that encode for plausible intermediates in the biosynthesis of the structurally unique antimicrobial diterpene natural products, merilactone and ribenone. As the second species in the Rubiaceae with a genome sequence, the C. alba genome provides a new resource to identify genes involved in specialized metabolism in a family with rich chemical diversity.

Data availability

Raw read sequences for the genome and transcriptome analyses are available in the National Center for Biotechnology Information under BioProject ID PRJNA543280. The genome assembly, annotation, gff, expression abundances, orthologous group membership as well as the GC-MS, LC-MS, and NMR datasets are available in the Dryad Digital Repository (https://doi.org/10.5061/dryad.00000000r).

Acknowledgements

The authors acknowledge the support from the facilities at Michigan State University including the Genomics as well as the Mass Spectrometry and Metabolomics Core, and the Max T. Rogers NMR Facility. We thank Philip Zerbe (University of California, Davis) for the ZmSKL4 clone and Reuben Peters (Iowa State University) for the ZmAN2 clone. The authors are thankful to Luis Manuel Peña Rodríguez (Unidad de Biotecnología, Centro de Investigación Científica de Yucatán, Mérida, México) for sharing C. alba diterpene standards.

Accession numbers

CaTPS1 (MK922246), CaTPS2 (MK922247), CaTPS3 (MK922248), CaTPS4 (MK922249), and CaTPS5 (MK922250).

Funding

Funding for this work was provided by a grant from the Michigan State University Strategic Partnership Grant Programme to C.R.B. and B.H.

Conflict of interest

B.H. and W.W.B. have filed a patent (PCT/US2019/044887) including diterpene synthases described in this study.

Supplementary data

Supplementary data are available at DNARES online.
