A target capture approach for phylogenomic analyses at multiple evolutionary timescales in rosewoods (*Dalbergia* spp.) and the legume family (Fabaceae)

Simon Crameri[1] | Simone Fior[1] | Stefan Zoller[1,2] | Alex Widmer[1]

[1] Institute of Integrative Biology, ETH Zurich, Zürich, Switzerland

[2] Genetic Diversity Centre (GDC), ETH Zurich, Zürich, Switzerland[1]

Running title

Target capture in *Dalbergia* and legumes

Correspondence

Simon Crameri, Institute of Integrative Biology, ETH Zurich, Zürich, Switzerland. Email: sfcrameri@gmail.com

Alex Widmer, Institute of Integrative Biology, ETH Zurich, Zürich, Switzerland. Email: alex.widmer@usys.ethz.ch

[1] Current address: Orniplan AG, Wiedingstrasse 78, Zürich, Switzerland.

Funding information

This work was supported by ETH Zurich and a grant from the Rübel Foundation to AW. The funders had no role in study design, data collection and analysis, or preparation of the manuscript.

Abstract

Understanding the genetic changes associated with the evolution of biological diversity is of fundamental interest to molecular ecologists. The assessment of genetic variation at hundreds or thousands of unlinked genetic loci forms a sound basis to address questions ranging from micro- to macro-evolutionary timescales, and is now possible thanks to advances in sequencing technology. Major difficulties are associated with i) the lack of genomic resources for many taxa, especially from tropical biodiversity hotspots, ii) scaling the numbers of individuals analyzed and loci sequenced, and iii) building tools for reproducible bioinformatic analyses of such datasets. To address these challenges, we developed a set of target capture probes for phylogenomic studies of the highly diverse, pantropically distributed and economically significant rosewoods (*Dalbergia* spp.), explored the performance of an overlapping probe set for target capture across the legume family (Fabaceae), and built a general-purpose bioinformatics pipeline. Phylogenomic analyses of *Dalbergia* species from Madagascar yielded highly resolved and well supported hypotheses of evolutionary relationships. Population genomic analyses identified differences between closely related species and revealed the existence of a potentially new species, suggesting that the diversity of Malagasy *Dalbergia* species has been underestimated. Analyses at the family level corroborated previous findings by the recovery of monophyletic subfamilies and many well-known clades, as well as high levels of gene tree discordance, especially near the root of the family. The new genomic and bioinformatics resources will hopefully advance systematics and ecological genetics research in legumes, and promote conservation of the highly diverse and endangered *Dalbergia* rosewoods.

49 ***KEYWORDS***

50 *Dalbergia*, rosewood, Fabaceae, Leguminosae, target capture, phylogenomics

51

1 | INTRODUCTION

The question how biological diversity evolves is of fundamental interest in ecology and evolution, and addressing it benefits from integrative approaches (Cutter, 2013; Rissler, 2016). Investigating evolutionary processes acting at the level of populations or groups of spatially interconnected populations (metapopulations) within species typically falls within the fields of population genetics and phylogeography. By contrast, analyses of evolutionary relationships among species and patterns of diversification in higher taxonomic groups fall within the realm of phylogenetics. Though it has long been recognized that "the same ecological and evolutionary processes that cause lineage divergence can also drive speciation" (Rissler, 2016), research in these fields has traditionally relied on different conceptual approaches, analytical methods, and molecular markers, generating a false dichotomy between fields aiming to address the same underlying processes. Today, the conceptualization of common theory combined with advances in methodology leveraging on next-generation sequencing (NGS) data offer the opportunity to jointly study the processes that drive the evolution of biological diversity from micro- to macro-evolutionary timescales.

Target capture (Mamanova et al., 2010) provides an efficient approach to acquire molecular information across broad evolutionary timescales when genomic regions with varying level of diversity are included in the experimental design (Jones & Good, 2016). It requires the design of capture probes that target unique regions in the genome to prevent conflation of orthologs and paralogs, and are characterized by a conserved core for in-solution hybridization and more variable flanking regions expected to provide parsimony informative sites (Lemmon et al., 2012). Combined with high-throughput sequencing, this approach allows for the analysis of hundreds or thousands of orthologous loci in dozens to hundreds of

individuals at moderate per-sample costs, and therefore strikes a good balance between locus information content and scalability to high numbers of individuals, including museum specimens (de La Harpe et al., 2017; Brewer et al., 2019). Hence, target capture holds a great potential to bridge the divide between phylogenetics, phylogeography and population genetics (de La Harpe et al., 2017; Nicholls et al., 2015; Rissler, 2016) and has increasingly been applied at macro-evolutionary, phylogeographic and micro-evolutionary timescales in a wide range of animals (e.g., Faircloth et al., 2012; Lemmon et al., 2012; Prum et al., 2015) and plants (e.g., de La Harpe et al., 2018; Koenen et al., 2020a; Mandel et al., 2014).

A global probe set targeting 353 putatively single-copy protein-coding genes has recently been developed for flowering plants (Angiosperms353; Johnson et al., 2019). Recent studies in various plant families have shown that the Angiosperms353 probe set represents a cost-effective resource to resolve phylogenetic relationships at the level of plant orders (e.g., Thomas et al., 2021), families (e.g., Siniscalchi et al., 2021), or at the infrageneric level (e.g., Ottenlips et al., 2021). However, several comparisons revealed that micro-evolutionary relationships are often better resolved when targeting more loci using taxon-specific probe sets (e.g., Shah et al., 2021; Siniscalchi et al., 2021; Ufimov et al., 2021). The development of taxon-specific probe sets therefore remains valuable for detailed phylogenetic and population genetic analyses (Yardeni et al., 2021).

Beside challenges associated with the *de-novo* probe design, processing and analysis of high-throughput sequencing data often involves complex and computationally demanding calculations. Target capture data are often analyzed using the PHYLUCE (Faircloth, 2016) or HybPiper (Johnson et al., 2016) bioinformatic pipelines. PHYLUCE was developed for analysis of sequences flanking ultraconserved genomic elements and has mainly been used at macro-

98 evolutionary and phylogeographic timescales in animal systems, whereas HybPiper is
99 optimized for datasets derived from probes designed in exons using Hyb-Seq (Weitemier et al.,
100 2014). There is thus a need for existing tools to be expanded with pipelines that are applicable
101 at deep to shallow evolutionary timescales (de La Harpe et al., 2017), while being independent
102 from high-quality annotated genomes or transcriptomes.

103 *Dalbergia* L.f. (Fabaceae) is a pantropical and ecologically diverse plant genus with c.
104 270 currently accepted species (WCVP, 2021), some of which have been described relatively
105 recently (e.g, Adema et al., 2016; Lachenaud, 2016; Wilding et al., 2021a, 2021b). Numerous
106 arborescent species are a source of rosewood (Bosser & Rabevohitra, 2002; Prain, 1904), a
107 high-quality timber sought-after on the international market and cause of conservation
108 concern (Schuurman & Lowry, 2009; Waeber et al., 2019). National and international
109 regulations have been established, aiming at sustainable exploitation and revenues (Barrett
110 et al., 2013; CITES, 2020), but illegal logging and trade continues (UNODC, 2016b, 2020;
111 Vardeman & Runk, 2020). The effective implementation of regulations demands that species
112 are reliably recognized and that extant population sizes are estimated to assess the potential
113 threat status. Developing a comprehensive understanding of species diversity in *Dalbergia*
114 and their evolutionary history, as well as a thorough knowledge of the ecology and distribution
115 of many traded species, has been hampered by several factors. There is a shortage of
116 collections and experts focusing on this taxonomically challenging genus, and current
117 treatments heavily rely on leaves and flowers and/or fruits for identification (Bosser &
118 Rabevohitra, 2002; de Carvalho, 1997; Lachenaud, 2016), which are rarely encountered
119 together in the field. As a result, the taxonomy of the genus is in need of extensive revision

(Wilding et al., 2021a), which could be supported by phylogenomic analyses targeting the nuclear genome (Crameri, 2020).

Motivated by the need for genomic resources to inform a reliable taxonomy and foster conservation practice, we introduce a target capture approach for anchored phylogenomic analyses in *Dalbergia* (Dalbergia2396 set). This genus belongs to the third largest angiosperm family (Fabaceae, a.k.a. Leguminosae or legume family), which is subject to extensive research in areas such as systematics (LPWG, 2017), ecology (Sprent et al., 2017), evolution (Koenen et al., 2021), speciation and rapid radiations (Hughes & Eastwood, 2006), and contains many agricultural crops (Mousavi-Derazmahalleh et al., 2018; Zhuang et al., 2019). This motivated us to further explore the applicability of our approach for analyses across the entire legume family, which resulted in a second probe set (Fabaceae1005 set). Both probe sets represent a subset of 6,555 conserved target regions distributed across the nuclear genome, derived from a combination of divergent reference capture using five published legume genomes, and a *de novo* assembly of a *Dalbergia* transcriptome. We also introduce a dedicated bioinformatics pipeline named CAPTUREAL supporting the analysis of high-throughput target capture sequencing data, with special emphasis on streamlined applicability, parallelization, and graphical output for informed parameter choices. The pipeline is designed for general application to target capture datasets, modular, and therefore easily customizable. We demonstrate the application of our approach to resolve phylogenetic relationships in the economically important and conservation-relevant genus *Dalbergia*. We then explore the utility for phylogenomic analyses at much deeper timescales by analyzing target capture data of various legume subfamilies. Finally, we test the utility of this approach at a micro-

142 evolutionary scale, and assess genetic variation among individuals and populations of two
143 closely related *Dalbergia* species from Madagascar.
144

145 # 2 | MATERIALS AND METHODS

146 ## 2.1 | Design of target capture probes and reference sequences

147 We produced a transcriptome assembly of a cultivated individual of *Dalbergia*
148 *madagascariensis* subsp. *antongilensis* Bosser & R. Rabev., based on 63 million paired-end
149 sequencing reads generated on an Illumina® HiSeq™ 2000 platform. We performed *de novo*
150 assembly of the transcriptome using Trinity release 2012-01-25 (Grabherr et al., 2011),
151 resulting in 146,484 scaffolds, which were between 201 and 17,129 bp long, with a mean
152 length of 815 bp (see Supplementary Methods). We then pairwise aligned the *Dalbergia*
153 transcriptome with reference genomes of five legume species available in public databases to
154 generate a set of 12,049 probes from 6,555 conserved target regions (see Supplementary
155 Methods). This probe set was used for synthesis of hybridizing probes at myBaits® Custom
156 Target Capture Kits (Arbor Biosciences; https://arborbiosci.com).
157

158 ## 2.2 | Taxon sampling for target capture probes validation

159 We created three taxon sets with contrasting levels of evolutionary divergence, ranging from
160 subfamilies to species to populations. The subfamily set (Table S1) included five of the six
161 legume subfamilies, as recognized in the most recent treatment (LPWG, 2017), and comprised
162 104 individuals (110 samples, six replicates; 99 species including three outgroups). Three
163 species of *Polygala* Tourn. ex L. (Polygalaceae) were included as the outgroup for the
164 subfamily set. The species set (Table S2) included members of the closely related genera

Dalbergia (at least 19 species), *Machaerium* Pers. (three species) and *Ctenodon* Baill. sensu Cardoso et al. (2020) (two species) and comprised 60 individuals (63 samples, three replicates; at least 26 species including two outgroups). Two species of *Aeschynomene* L. sensu stricto (s.str.) sensu Cardoso et al. (2020) were included as the outgroup for the species set. The population set (Table S3, Figure S4) included 51 individuals in total, 29 attributed to *D. monticola* Bosser & R. Rabev. from four sampling locations, and 22 attributed to *D. orientalis* Bosser & R. Rabev. from eleven sampling locations.

2.3 | Library preparation, target capture and sequencing

We extracted total genomic DNA from silica gel dried leaf tissue (185 extractions) or museum specimens deposited at the Paris (P) herbarium (11 extractions) using the CTAB protocol (Doyle and Doyle, 1987) or the DNeasy® Plant Mini Kit (Qiagen). We quantified DNA using the QuantiFluor® dsDNA system for a Quantus™ fluorometer (Promega) and checked DNA integrity on 1.5% agarose gels for a subset of samples. We prepared genomic DNA libraries for each sample using the NEBNext® Ultra II DNA Library Prep Kit for Illumina® (New England Biolabs), following manufacturer's instructions. We individually indexed samples to be pooled within the same sequencing lane during the PCR enrichment step using NEBNext® Multiplex Oligos for Illumina® (single-indexed with E7335 and E7500 kits, or dual-indexed with E6440 kit, New England Biolabs). We performed in-solution hybridization and target enrichment using our 12,049 tiled RNA probes. We pooled up to six individually indexed libraries during the hybridization step using a stratified random assignment of libraries to hybridization reactions. Stratification aimed at optimizing the sequencing coverage across samples and consisted in avoiding pooling of close relatives of *Cajanus cajan* with more distantly related

188 samples, and of museum specimens with silica gel dried leaf material. We obtained short read
189 data by combining sequencing runs from an Illumina® MiSeq™ (2×300 bp paired-end
190 sequencing, 99 libraries) at the Genetic Diversity Centre (GDC) Zurich, an Illumina® HiSeq™
191 4000 (2×150 bp paired-end sequencing, 88 libraries) at the Functional Genomics Center Zurich
192 (FGCZ) or Fasteris SA (Plan-les-Ouates, Switzerland), and an Illumina® NovaSeq™ 6000 SP flow
193 cell (2×150 bp paired-end sequencing, 9 libraries) at the FGCZ. We repeated DNA extraction,
194 hybridization and target enrichment sequencing for nine individuals (replicates) to assess
195 reproducibility. One sample (*Hassold 565*) was represented in each taxon set, nine samples
196 were represented in both the species and population sets, and nineteen samples were
197 represented in both the subfamily and species sets.
198
199 **2.4 | CAPTUREAL bioinformatics pipeline**
200 The bioinformatic pipeline CAPTUREAL was developed for this project and is accessible on
201 Github (https://github.com/scrameri/CaptureAl) as a documented sequence of scripts. These
202 include bash and R scripts (R Core Team, 2020) to manage and visualize data with APE version
203 5.3 (Paradis & Schliep, 2018), DATA.TABLE version 1.12 (Dowle & Srinivasan, 2019), and TIDYVERSE
204 version 1.3.0 (Wickham et al., 2019). Where appropriate, computations are carried out for
205 multiple samples or target regions in parallel using GNU PARALLEL (Tange, 2011). The CAPTUREAL
206 pipeline streamlines the mapping of quality-trimmed reads to target regions, the exclusion of
207 loci targeting multi-copy genes and taxa with insufficient data coverage, and the alignment of
208 orthologous loci for downstream phylogenetic analyses. At various critical steps, the pipeline
209 outputs summary statistics and graphs that inform the user on the effects of specific filtering
210 parameters, allowing for informed parameter choices.

The pipeline is divided into seven steps to process quality-filtered reads. Steps 1 to 5 are always required, and 1) map the sequencing reads to target regions, 2) assemble mapped reads separately for each target region, 3) identify the most-likely orthologous contigs, 4) identify taxa and target regions with high capture sensitivity and specificity, and 5) create trimmed alignments of the kept taxa and target regions. Steps 6 and 7 are optional, and 6) combine physically neighboring and overlapping alignments to 7) generate longer and more representative reference sequences as starting points for reiteration of steps 1 to 5. Such reiteration can improve mapping success, and mitigates potential biases arising from the reference sequences used (Hahn et al., 2013).

In our analyses, we executed the pipeline separately and iteratively for different taxon sets. We first applied steps 1 to 5 to twelve representative samples each from the subfamily and species sets, followed by steps 6 and 7 to generate longer and taxon-specific reference sequences for target regions that can each be efficiently recovered in these taxon sets, and then reiterated steps 1 to 5 for all samples of the subfamily and species sets using the new reference sequences and more stringent target region filtering parameters (see Tables S1–S3 and Supplementary Methods for details). We also performed steps 6 and 7 after the second iteration of the species set analysis to produce reference sequences for analysis of the population set. Bioinformatic analyses were carried out on a multi-core LINUX server (GDC Zurich) or on the EULER scientific compute cluster (ETH Zurich). The sequence of executed commands and the chosen parameters are provided in Supplementary Methods.

2.4.1 | Step 1: Read mapping

Quality-filtered reads of each sample are mapped against the reference sequences (one sequence per target region) using the BWA-MEM algorithm (Li, 2013). The minimum alignment score and mapping quality can be adjusted as needed. Coverage statistics are computed using SAMTOOLS (Li & Durbin, 2009) and BEDTOOLS (Quinlan & Hall, 2010), and target regions are visually filtered for adequate coverage across samples using *filter.visual.coverages.R*, which allows to apply filtering thresholds that are informed by visualizations of coverage statistics (see Supplementary Methods). The main output of step 1 are BAM files a list of retained target regions.

2.4.2 | Step 2: Sequence assembly

Read pairs are extracted from quality-filtered reads when at least one read mapped to any of the retained target regions with the specified minimum mapping quality. Extracted reads are assembled separately for each sample and region using DIPSPADES (Safonova et al., 2015) based on haplocontigs generated by SPADES (Bankevich et al., 2012, see Supplementary Methods). The main output of step 2 are consensus contiguous sequences (contigs hereafter) for each sample and each target region.

2.4.3 | Step 3: Orthology assessment

Sequence assembly may yield multiple contigs per sample for some target regions, e.g., due to capture of several fragments of the same genomic region (e.g., in degraded museum specimens), or due to unspecific capture of paralogs (Johnson et al., 2016). The most likely orthologous contig(s) of each sample in each target region are determined using an exhaustive Smith-Waterman alignment (Smith & Waterman, 1981) between all contigs and the reference

sequences using EXONERATE (Slater & Birney, 2005). The best-matching contig is defined based on the EXONERATE alignment statistics as the most likely orthologous contig for each sample and target region, and further contigs that did not overlap with one another or the best-matching contig, but aligned with a sufficient alignment score to other parts of the target region are retained. These contigs likely represent fragments of the same region, and can therefore be combined with the best-matching contig to form a contiguous sequence (orthologous contig hereafter, see Supplementary Methods). The main output of step 3 is a single-sequence FASTA file with the putative orthologous contig for each sample and each target region.

2.4.4 | Step 4: Sample and region filtering

Successful target capture depends on whether sequence data can be collected for a high proportion of target regions (capture sensitivity, Jones & Good, 2016) in a high proportion of focal taxa, and whether the captured sequences are orthologs of the target regions (capture specificity). Target regions are visually filtered for high capture sensitivity and specificity across focal taxa using *filter.visual.assemblies.R*, which allows to apply filtering thresholds that are informed by visualizations of EXONERATE alignment statistics generated in step 3. These thresholds can be set globally to remove generally poorly sequenced samples or target regions, but they can also be set as the minimum fraction of samples required to pass a specified filtering threshold in order for a target region to be retained. Taxon groups can be defined, in which case the required capture sensitivity and specificity parameters need to be met in all considered taxon groups separately, thus preventing target regions from being

poorly represented in rare taxon groups (see Supplementary Methods). The main output of step 4 is a list of samples and a list of target regions to keep.

In our analyses, we defined the four subfamilies represented by multiple taxa as taxon groups in the subfamily set. In the species set we defined four taxon groups based on our preliminary phylogenetic results and phylogenetic relationships inferred by Hassold et al. (2016). These were subgroup (SG) 1 (species with large flowers and paniculate inflorescences), SG2 (species with large flowers and racemose inflorescences), SG3 (species with small flowers from East Madagascar), and SG4 (species with small flowers from West and North Madagascar).

2.4.5 | Step 5: Target region alignment and alignment trimming

A multi-sequence FASTA file is generated for all retained target regions, containing the respective orthologous contigs of all retained samples. Sequences are then aligned using MAFFT (Katoh & Standley, 2013), allowing for different alignment options. Alignments are trimmed at both ends until an alignment site shows nucleotides across a specified minimum fraction of aligned sequences, along with a specified maximum nucleotide diversity (i.e., the mean number of base differences between all sequence pairs). In addition, internal trimming is performed by only keeping sites with nucleotides in a specified minimum fraction of aligned sequences. Potential mis-assemblies or mis-alignments at contig ends are further resolved using a sliding window approach that identifies and masks sequences with large deviations from the alignment consensus (see Supplementary Methods). The main output of step 5 are potentially overlapping trimmed alignments for each kept target region.

2.4.6 | Step 6: Merging of overlapping alignments

Shorter but physically close target regions facilitate sequence assembly in lower-quality samples but can lead to overlaps in trimmed alignments of neighboring target regions. Such overlaps can be identified by aligning consensus sequences of target region alignments. Specifically, consensus sequences are generated by calling IUPAC ambiguity codes if a given minor allele frequency threshold is reached, or a gap if a given base frequency threshold is not reached. Local alignments between different consensus sequences are identified using BLAST+ version 2.7.1 (Camacho et al., 2009), and filtered for non-reciprocal hits between alignment ends of target regions located on the same linkage group. Orthologous contigs that are part of different, overlapping alignments are then aligned using MAFFT. The resulting merged alignments are then collapsed to represent different orthologous contigs of the same individual as a single sequence, a process that can be visually inspected if needed. Trimming is then applied as in step 5, and sets of two to several consecutively overlapping alignments are then each replaced by a single merged alignment if merging was successful (see Supplementary Methods). The main output of step 6 are non-overlapping trimmed alignments for each kept target region.

2.4.7 | Step 7: Generation of representative reference sequences

To mitigate potential biases arising from the reference sequences used, a new set of target region reference sequences can be generated based on the target region alignments generated in the two previous steps. For this purpose, a consensus sequence is generated for each alignment as in step 6, but separate consensus sequences can be generated for different specified taxon groups (see step 4). These sets of taxon group specific consensus sequences

are then aligned, and representative consensus sequences are generated as in step 6 (see Supplementary Methods). These taxon-specific reference sequences are the main output of step 7 and can be used to refine mapping, assembly and alignment by reiterating steps 1 to 5.

2.4.8 | Alignment assessment and filtering

We characterized all non-overlapping trimmed alignments for the number of gaps, gap ratio (i.e, the fraction of non-nucleotides in the alignment), total nucleotide diversity, average nucleotide diversity per site, and alignment length, as well as the number and proportion of segregating and parsimony informative sites. We then visually filtered alignments using *filter.visual.alignments.R*, which allows to apply filtering thresholds that are informed by visualizations of alignment statistics (see Supplementary Methods). We used the filtered alignments after the second iteration of step 6 for phylogenetic analyses.

2.5 | Phylogenetic analyses

We performed phylogenetic analyses with both the subfamily and species sets, using a supermatrix (concatenation) approach and a gene tree summary approach. For the supermatrix approach, we ran maximum likelihood searches on the concatenated alignments using RAxML version 8.2.11 (Stamatakis, 2014) with rapid bootstrap analysis and search for the best-scoring tree in the same run (-f a option), 100 bootstrap replicates, and the GTRCAT approximation of rate heterogeneity (see Supplementary Methods). For the gene tree summary approach, we ran RAxML jobs separately for each alignment using the same settings as for the supermatrix approach to generate gene trees. Following Zhang et al. (2018), we collapsed branches in gene trees if they had bootstrap support values below 10 using NEWICK

utilities (Junier & Zdobnov, 2010), and we performed species tree analyses with ASTRAL-III version 5.6.3 (Mirarab et al., 2014; Zhang et al., 2018) and standard parameters, except for full branch annotation (see Supplementary Methods). For the subfamily set, we additionally evaluated the quartet support for fifteen different subfamily topologies (i.e., all possible topologies with Caesalpinioideae, Dialioideae, Papilionoideae and (Cercidoideae, Detarioideae) as ingroups; Figure S2), using the tree scoring option in ASTRAL-III and a file with the assignment of taxa to subfamilies or the outgroup. All phylogenetic trees were displayed using GGTREE version 2.0.2 (Yu et al., 2016).

2.6 | Population genetic analyses

We carried out population genetic analyses for the population dataset. We mapped quality-filtered reads against the target region reference sequences that were representative of the species set after the second iteration using BWA-MEM. We verified efficient recovery of target regions by plotting heatmaps of coverage statistics, removed PCR duplicates using PICARD TOOLS version 2.21.3 (Broad Institute, 2019), and capped excessive coverage to 500 using *biostar154220.jar* (Lindenbaum, 2015). We then called variants using FREEBAYES version 1.1.0-3-g961e5f3 (Garrison & Marth, 2012) and standard parameters, except for a minimum alternate fraction of 0.05, a minimum repeat entropy of 1, and evaluation of only the four best alleles. Variants were filtered using VCFTOOLS version 0.1.15 (Danecek et al., 2011) and VCFLIB version 1.0.1 (Garrison, 2012), which was also used to decompose complex variants (see Supplementary Methods). We then used VCFR version 1.10.0 (Knaus & Grünwald, 2017) and ADEGENET version 2.1.1 (Jombart, 2008; Jombart & Ahmed, 2011) to generate *genind* and *genlight* objects that represented the SNP allele table with associated metadata such as

individual missingness, species identification, and sampling location. We used the SNP subset with zero missingness to conduct principal component analysis (PCA) based on the centered covariance matrix, as well as to calculate a neighbor-joining (NJ) tree (Saitou & Nei, 1987) on Nei's genetic distances, as implemented in POPPR version 2.8.1 (Kamvar et al., 2014). We also used the allele table to create a SNP subset for population clustering analysis using STRUCTURE version 2.3.4 (Pritchard et al., 2000). Specifically, we kept SNPs with genotype data in at least 95% of individuals, and we randomly sampled up to three SNPs per target region for linkage disequilibrium pruning and computational ease. STRUCTURE analyses were performed for one to ten demes (*K*), using 110,000 iterations, including a burn-in period of 10,000 iterations, with ten replicates per simulation (see Supplementary Methods). Replicate STRUCTURE results were aligned and visualized using CLUMPAK (Kopelman et al., 2015) and default settings.

3 | RESULTS

3.1 | Two probe sets for target capture across legumes and *Dalbergia*

We obtained 0.13 to 13.76 (median: 1.56) million raw read pairs per sample, of which we retained 86.55% to 99.34% (median: 93.82%) after quality trimming (Tables S1 – S3). In the first iteration applied to twelve representative samples, reads mapped to 6,519 or 6,287 of the 6,555 initial target regions in the subfamily or species set, respectively (step 1). Of these we retained 3,436 or 4,908 target regions, which showed adequate coverage across taxon groups. After assembly (step 2) and orthology assessment (step 3), 2,710 or 4,181 target regions passed the region specificity and sensitivity filters of lower stringency (step 4). Following alignment and trimming (step 5), overlapping portions in 207 or 377 regions were successfully merged, resulting in 2,468 or 3,736 non-overlapping trimmed alignments (step 6).

Longer and more representative consensus sequences were generated from these target regions (step 7) and used as references for mapping quality-trimmed reads of the complete taxon sets (step 1, see Tables S1 and S2). We retained 1,917 or 3,418 target regions with adequate coverage (Figures S5 and S6), of which 1,020 or 2,407 passed the specificity and sensitivity filters of higher stringency (step 4) after assembly. Merging of overlapping alignments in 15 or 11 regions yielded 1,005 (subfamily set) or 2,396 (species set) distinct alignments (step 5), of which 666 represented the same regions in both sets. The corresponding Fabaceae1005 and Dalbergia2396 probe sets, along with refined taxon-specific reference sequences are deposited on Dryad. Corresponding gene annotations in the *Cajanus cajan* genome are given in Tables S4 and S5. For phylogenetic analyses, we excluded 19 or 7 alignments with a gap ratio above 0.35 or 0.3 or a nucleotide diversity above 0.35 or 0.15, leaving 986 (subfamily set) or 2,389 (species set) alignments.

Quality-trimmed reads mapped to all 2,396 target regions in the population set (step 1) using reference sequences that were representative of the species set after the second iteration for mapping (Figure S7). Variant calling resulted in 203,916 raw variants and 116,500 filtered SNPs after decomposing complex variants, of which 60,204 (51.68%) were bi-allelic with no missing data and were used for PCA and NJ tree reconstruction. Random sampling of up to three SNPs per target region resulted in a subset of 5,042 SNPs for STRUCTURE analyses.

3.2 | Phylogenomic analyses across legumes

Phylogenetic analysis of 986 alignments recovered each of the five sampled subfamilies as monophyletic, and many well-established clades and relationships received ≥95% support

using both the gene tree summary method ASTRAL-III (Figure 1) and the supermatrix method (Figure S1). These included the subfamilies Cercidoideae and Detarioideae found to be sister taxa, the mimosoid clade within the recently re-circumscribed subfamily Caesalpinioideae (LPWG, 2017), as well as the Angylocalyceae-Dipterygeae-Amburaneae (ADA, Cardoso et al., 2012), Cladrastis (Wojciechowski, 2013) and Meso-Papilionoideae (Wojciechowski, 2013) clades within Papilionoideae. We also recovered the Sophoreae and Genisteae clades (Cardoso et al., 2013) within Genistoids sensu lato (s.l.) (Cardoso et al., 2012; Wojciechowski et al., 2004). Within the Dalbergioids s.l. (Wojciechowski et al., 2004), we recovered the Amorpheae clade (McMahon & Hufford, 2004) as sister to the rest of the group, which includes the Dalbergioids s.str. clade (Lavin et al., 2001), containing the *Adesmia*, *Pterocarpus* and *Dalbergia* subclades (Lavin et al., 2001), respectively. *Ctenodon brasilianus* (Poir.) D.B.O.S.Cardoso, P.L.R.Moraes & H.C.Lima and *C. nicaraguensis* (Oerst.) A.Delgado were found to be more closely related to *Machaerium* than to *Aeschynomene*. Within the Non-Protein-Amino-Acid-Accumulating (NPAAA) clade (Cardoso et al., 2012; Wojciechowski et al., 2004), we recovered the Millettioid s.l. clade (Wojciechowski et al., 2004), containing the genera *Indigofera* and *Millettia*, and the Phaseoleae s.l. (Vatanparast et al., 2018), as well as the Hologalegina (Wojciechowski, 2013) clade, including the Robinioids and the inverted-repeat-lacking clade (IRLC, Wojciechowski et al., 2004).

Other relationships among subfamilies remained unresolved using both phylogenetic methods (Figure 1, Figure S1). In particular, a clade comprising Caesalpinioideae, Cercidoideae, Detarioideae and Dialioideae as sister group to Papilionoideae was not supported in the supermatrix tree, and was recovered in only 47% of quartet trees. We evaluated quartet scores (i.e., the fraction of induced quartet trees) of fourteen further

topologies for relationships among sampled subfamilies (Figure S2) using the tree scoring option in ASTRAL-III in combination with a file that mapped taxa to subfamilies or to the outgroup. The subfamily topology presented in Figure 1 showed the highest normalized quartet score (38.40%). Two alternative topologies received a similar normalized quartet score of 38.36% (Figure S2) and involved a clade composed of Caesalpinioideae and Papilionoideae. Further contentious relationships between major groups concerned the three clades within Meso-Papilionoideae, where the clade formed by Dalbergioids s.l. and Genistoids s.l. was recovered only in 36% of quartet trees, and in relationships within Caesalpinioideae, Detarioideae, and Genisteae. All except one genus with multiple sampled accessions were recovered as monophyletic, the exception being *Cytisus*, which was paraphyletic with respect to *Lembotropis nigricans*. Pairs of replicates each grouped together (Figure 1, Figure S1).

3.3 | Phylogenomic analyses in *Dalbergia*

Phylogenetic analysis of 2,389 alignments recovered samples of *Dalbergia* as monophyletic with ≥95% support using both ASTRAL-III (Figure 2) and the supermatrix method (Figure S3). Within *Dalbergia*, we recovered two large and exclusively Malagasy clades, which we name Madagascar Supergroup I and II. All Malagasy species represented by multiple accessions were recovered as highly supported clades, with the exception of *D. normandii*. Four non-Malagasy *Dalbergia* specimens and *D. bracteolata* Baker were each found to represent separate lineages.

Within Supergroup I, one clade comprised samples of *Dalbergia chapelieri* s.l., while the remaining samples belonged to a sister group containing three monophyletic species and

462 a basal and paraphyletic *D. normandii*. Within Supergroup II, two clades contained species
463 distributed in the humid east of Madagascar, while the third contained species distributed in
464 the seasonally dry west and north of the island. Within *D. chapelieri* s.l. and *D. monticola*,
465 which were each represented by six individuals, we observed geographic structure, with
466 specimens from northeast and southeast Madagascar forming sister groups. Pairs of replicates
467 each grouped together (Figure 2, Figure S3).
468
469 **3.4 | Population genomic analyses**
470 Principal component analysis revealed three distinct clusters of individuals along principal
471 component (PC) 1 (explaining 27.58% of the total variation) and PC 2 (11.26%; Figure 3).
472 Individuals of *D. orientalis* separated along PC1, while individuals originally attributed to *D.*
473 *monticola* formed two distinct groups mainly along PC2. The unexpected smaller cluster (in
474 purple) comprised samples from a single broad sampling location in north-eastern
475 Madagascar (location 5, see Figure S4 and Table S3) where both *D. monticola* and *D. orientalis*
476 were also collected. The same three clusters were also recovered in STRUCTURE analyses (Figure
477 S8), where biologically meaningful clustering solutions were found for K = 2 (separating *D.*
478 *orientalis* from the rest) and K = 3 (further separating the unexpected smaller cluster; Figure
479 S9). Within *D. orientalis* and the larger cluster of presumed *D. monticola*, the NJ tree reflects
480 isolation by distance at a broad geographical scale, separating specimens from north-eastern
481 (locations 1 to 6), central-eastern (locations 7 and 8) and south-eastern Madagascar (locations
482 9 to 13; Figure 3). A similar geographic pattern was recovered by STRUCTURE assuming K = 5
483 (Figure S8), although that clustering solution received much lower support (Figure S9).
484 Clustering solutions assuming higher K did not recover additional meaningful structure. K = 7

485 showed an unrealistic probability by *K* of 1 (Figure S9B), which may be related to the presence
486 of 'ghost clusters' with near-zero admixture proportions.
487

488 4 | DISCUSSION

489 Understanding the diversity and diversification of species and evolutionary lineages requires
490 an integrative approach that links studies of micro-evolutionary processes to analyses of
491 macro-evolutionary relationships (de La Harpe et al., 2017). Genetic data form a preferable
492 source of information for investigations across broad evolutionary scales, as a large number
493 of loci distributed across the nuclear genome can represent the spectrum of evolutionary
494 rates at different scales of sample divergence. The present study introduces two overlapping
495 sets of target capture probes for phylogenomic studies at micro- to macro-evolutionary
496 timescales in rosewoods (Dalbergia2396 set) and across the legume family (Fabaceae1005
497 set), together with the flexible and modular bioinformatic pipeline CAPTUREAL, which
498 streamlines the processing of sequencing reads for phylogenomic and population genomic
499 analyses while visually informing users on the effect of critical parameter choices. We
500 demonstrated the utility of individual assemblies per target region to produce alignments of
501 hundreds of loci suitable for concatenation and multispecies coalescent approaches, which
502 confirmed phylogenomic conflicts at the root of the legume family, and provided an
503 unprecedented resolution of evolutionary relationships among lineages and species of the
504 taxonomically complex genus *Dalbergia*. Remapping of sequencing reads further made it
505 possible to identify thousands of informative sites amenable to population genomic analyses,
506 which revealed the existence of a potentially new cryptic *Dalbergia* species. Together, these
507 results illustrate that our newly developed probe sets are efficient tools for studies of species

508 diversity and diversification in rosewoods (*Dalbergia* spp.) and more broadly in the
509 economically important and highly diverse legume family.
510
511 **4.1 | Target capture probes**
512 The target capture probes presented here are part of a growing collection of genomic
513 resources for legume phylogenomics. Other probe sets for target capture in legumes have
514 been developed, focusing on different groups within the family. Our probes can be compared
515 to existing sets designed or validated at the level of legume species (Peng et al., 2017), genera
516 (e.g., de Sousa et al., 2014; Nicholls et al., 2015; Shavvon et al., 2017), or subfamilies (Koenen
517 et al., 2020a; Vatanparast et al., 2018) and across angiosperms (Johnson et al, 2019) to identify
518 overlaps in target regions for legume phylogenomics. For example, it would be interesting to
519 compare our probes with those of Vatanparast et al. (2018), who identified 423 target regions
520 based on 30 transcriptomes, of which 27 were sampled from the NPAAA clade, one from
521 Genistoids, and one each of the Caesalpinioideae and Cercidoideae subfamilies, limiting the
522 probe set validation to 25 species of the NPAAA clade. Capture of additional, less conserved
523 target regions across the legume family could be achieved by designing multiple probes for
524 hybridization in the same target region in different legume groups, as applied for studies
525 across angiosperms (Johnson et al., 2019). Such a probe design could profit from existing
526 legume probe sets but should rely on a stringent selection of targets that accounts for paralogs
527 (Vatanparast et al., 2018), which originated as a consequence of multiple whole-genome
528 duplication events in legumes (Egan & Vatanparast, 2019; Koenen et al., 2021).
529 In this study, we enriched DNA libraries from three taxon sets spanning micro-
530 evolutionary (populations) to macro-evolutionary (family) timescales, using a single set of

12,049 RNA probes targeting 6,555 genomic regions conserved across five Meso-Papilionoideae genomes and a *Dalbergia* transcriptome. We then identified 2,396 and 1,005 target regions with high capture specificity and sensitivity within the species-rich genus *Dalbergia* (Dalbergia2396 probe set) and more broadly across legumes (Fabaceae1005 probe set). We used our CAPTUREAL pipeline to refine phylogenomic and population genomic analyses using taxon-specific and longer reference sequences. This procedure has both benefits and drawbacks. An advantage is that different but overlapping probe sets amenable for efficient target capture in different focal groups can be identified, and that a single enriched DNA library can be included in multiple data sets spanning different evolutionary timescales. On the other hand, bioinformatic analyses took longer due to the iterative refinement, and only a portion of captured sequence data was ultimately used for phylogenomic or population genomic analyses in each focal group (see Tables S1 – S3). Higher costs per used sequence could be compensated by enriching DNA of up to six individuals in a single hybridization reaction, a strategy that has been used successfully in other studies (e.g., de La Harpe et al., 2018; Yardeni et al., 2021).

4.2 | CAPTUREAL bioinformatics pipeline

The CAPTUREAL pipeline starts with the mapping of quality-trimmed reads to conserved target regions identified during probe design, followed by assembly on a per-region basis, orthology assessment, and filtering for target regions with high capture sensitivity and specificity for downstream analyses. This approach differs from the PHYLUCE pipeline (Faircloth, 2016), where quality-trimmed reads are first assembled, and then matched to target regions. CAPTUREAL simplifies the assembly of reads specific to each locus, circumventing the

554 challenging task of *de novo* assembly of contigs from the large pool of sequencing reads
555 representative of thousands of loci (reviewed by Chaisson et al., 2015). Likewise, alignments
556 are conducted in clearly defined target regions in which overlap among individual contigs is
557 higher. However, assembly per region is more time-consuming and requires reference
558 sequences for the initial mapping step. This might introduce a reference bias when divergent
559 sequences are not mapped (Lunter & Goodson, 2011). We addressed this problem by
560 generating consensus sequences that are representative of a given taxon set and by limiting
561 analyses to target regions that can be efficiently recovered in all groups of that taxon set.
562 These set-specific reference sequences can then be used to iteratively refine mapping,
563 assembly, and target region filtering for any set of taxa. Our approach is conceptually similar
564 to the HybPiper pipeline (Johnson et al., 2016), which also employs a mapping-assembly
565 strategy, and uses depth of coverage and percent identity to the target region to choose
566 between multiple contigs, before it identifies intron/exon boundaries using target peptide
567 sequences and extracts coding sequences for alignment. While the HybPiper pipeline is
568 designed specifically for the Hyb-Seq approach (Johnson et al., 2016), in which exons are the
569 primary targets and flanking non-coding regions are used as supplementary data for analyses
570 at shallow evolutionary scales (Weitemier et al., 2014), CaptureAI is more general in scope and
571 neither requires nor leverages knowledge about intron/exon boundaries in the targeted
572 regions. It is therefore suitable for application in systems lacking high-quality annotated
573 reference genomes or transcriptomes. The main strengths of this pipeline are its modularity,
574 which allows for an iterative refinement of read mapping, assembly and alignment, its
575 flexibility given by user-defined parameters, the merging of alignments representing

physically overlapping target regions, and the visualization of key summary statistics and alignments along the workflow to inform the user on critical analysis parameters.

4.3 | Macro- and micro-evolutionary patters in *Dalbergia*

Dalbergia species endemic to Madagascar were recovered as two large, well-supported and fully resolved clades, each exclusively comprising Malagasy species. These two clades were previously identified on the basis of three chloroplast markers, but phylogenetic relationships within clades were not resolved, which exposed traditional DNA barcoding as insufficient for genetic discrimination between closely related *Dalbergia* species (Hassold et al., 2016; see Tables S2 and S3). Supergroup I and II are morphologically divergent and largely correspond to Group 1 and 2 reported by Bosser & Rabevohitra (2002). Supergroup I is characterized by a glabrous gynoecium with a long and slender style and relatively large flowers, while Supergroup II is characterized by a pubescent gynoecium with a short and squat style and relatively small flowers. The two supergroups are both more closely related to non-Malagasy taxa than to each other, suggesting at least two independent colonizations of Madagascar followed by species diversification. The only sampled Malagasy species not belonging to either of the two supergroups is *D. bracteolata*, which occurs on Madagascar as well as in mainland East Africa. A further species, which is endemic to Madagascar and morphologically divergent from Supergroups I and II (*D. xerophila* Bosser & R. Rabev.) was not included in this study.

Within Supergroup I, two well-supported subclades were resolved, which differ in their inflorescence structure. Within *Dalbergia chapelieri* s.l., a widely distributed species complex with paniculate inflorescences, northeastern and southeastern populations can be distinguished using the present data as well as chloroplast variation (Hassold et al., 2016). The

other subclade within Supergroup I contains species from eastern Madagascar with mostly racemose inflorescences, including a potentially new species, *Dalbergia* sp. 24. Collections belonging to this entity were previously believed to be conspecific with *D. maritima* var. *pubescens* (see Hassold et al., 2016) but show geographic (i.e., north-east vs. central-east), morphological (i.e., more numerous leaflets that are smaller, more oblong and less coriaceous) and genetic (Figure 2, Figure S3) differences compared to the type material (*Service Forestier 32824*). The type (collected in 1985) showed a slightly longer terminal branch compared to other samples in the concatenation tree (Figure S3) but clearly grouped with two recently collected conspecific samples from central-east Madagascar. The same subclade also contains material of two highly valued rosewood species, *D. occulta* and *D. normandii* (note that in Hassold et al. (2016), sterile material of *D. normandii* was erroneously identified as *D. madagascariensis*).

Supergroup II includes two clades distributed in the humid and sub-humid east and northwest of Madagascar, and a large third clade centered in the drier west and north of the island. Morphological synapomorphies characterizing these clades require further genetic and morphological analyses. The geographic separation in major eco-geographic regions of Madagascar suggests that climate regimes may have played a significant role in shaping the evolution of these groups, which thus constitute promising model systems to study processes of ecological divergence, along the same lines as studies that have investigated elements of the Malagasy fauna (Vences et al., 2009).

Our results revealed relationships among Supergroups I and II and non-Malagasy taxa that are incompatible with the plastid phylogeny of Hassold et al. (2016), in particular with regard to *Dalbergia melanoxylon* (Africa), *D. ecastaphyllum* (America and Africa), and *D.* cf.

oliveri (Asia). Incongruence between nuclear and plastid phylogenies is common at various evolutionary timescales in many plant groups (e.g., Lee et al., 2021; Pelser et al., 2010), and while the multispecies approach applied in this study is expected to return a phylogeny that reflects nuclear evolution accounting for incomplete lineage sorting, conflicts in gene tree topologies due to hybridization and chloroplast capture can further underlie the observed differences.

Our target capture approach demonstrated great potential to facilitate the resolution of several taxonomic conundrums within the genus, which likely resulted from few observable and diagnostic morphological characters, insufficient collection effort, and the difficulty of distinguishing between heritable and plastic trait variation within and among *Dalbergia* species (Lachenaud, 2016). The integration of highly informative museum specimens, including a nomenclatural type collected in 1985, enabled the accurate identification of recently collected but often sterile specimens, and was crucial in detecting misidentifications or potential taxonomic inadequacies (Buerki & Baker, 2016), as shown for *D. maritima* var. *pubescens* or *D. monticola.*

Population genomic analyses of 51 individuals readily separated the two closely related species *Dalbergia monticola* and *D. orientalis*, as well as a sympatric and syntopic but genetically differentiated entity, which could previously not be differentiated from the other two species based on three chloroplast markers (Hassold et al., 2016). The lack of admixture between *D. monticola* and this third cluster, the similarity in leaf characters, and the absence of known morphologically similar species occurring in the region, prompts us to hypothesize the latter to reflect a separate, yet undescribed cryptic species. Both *D. monticola* and *D. orientalis* are distributed from northeastern to south-eastern Madagascar, co-occur in various

localities, but differ in their predominant altitudinal distribution (Madagascar Catalogue, 2021). Population structure within both species was uncovered using our target capture approach and appears to be sufficient to distinguish specimens from the northeast (locations 1 to 6), central-east (locations 7 and 8), and southeast of the island (locations 9 to 13). These results indicate that genetic species identification and provenancing, at least to this broad geographic scale, may be feasible, which would have important implications for forensic timber identification and for tracing geographic hotspots of the illegal trade in these valuable timber species (UNODC, 2016a).

4.4 | Phylogenetic analyses across legumes

At the family level, 1,005 merged regions of the 6,555 targeted regions passed our stringent sensitivity and specificity filters, suggesting that many target regions were not efficiently captured across taxa. However, phylogenetic analysis of 986 nuclear target regions recovered multiple known clades within monophyletic subfamilies with strong bootstrap and quartet support, providing excellent resolution comparable to that obtained in the recent nuclear phylogenomic analysis of transcriptome and genome-wide data across legumes (Koenen et al., 2020b). As in that study, we found high support for Cercidoideae and Detarioideae as sister taxa, a relationship that was never inferred in analyses based on chloroplast genes (LPWG, 2017) or plastomes (Koenen et al., 2020b; Zhang et al., 2020). As shown in both studies, the other relationships among subfamilies are difficult to resolve. Our most supported subfamily topology (38.4% quartet support, Figure S2A) recovered the Papilionoideae as sister to a clade comprised of Caesalpinioideae, Dialioideae, and the Cercidoideae/Detarioideae clade, while Koenen et al. (2020b) demonstrated a successive divergence of the

Cercidoideae/Detarioideae clade, Dialioideae, Caesalpinioideae and Papilionoideae in all nuclear analyses. This alternative topology received almost equivalent overall quartet support (38.36%) in our analyses (Figure S2C), as did a third hypothesis in which Caesalpinioideae and Papilionoideae are sister to Dialioideae and the Cercidoideae/Detarioideae clade (Figure S2B). These nearly equally supported subfamily topologies can be explained by short deep internodes associated with conflicting bipartitions and are consistent with the idea of a nearly simultaneous evolutionary origin of all six legume subfamilies, causing incomplete lineage sorting (Koenen et al., 2020b). Taxon sampling may additionally contribute to the contentious deep-branching relationships. The monotypic Duparquetioideae subfamily could not be analyzed, and a portion of gene trees may suffer from long branch attraction between *Polygala* and Papilionoideae, which both exhibit markedly higher substitution rates compared to the other legume subfamilies (Koenen et al., 2020b). Additional outgroup taxa such as members of the Quillajaceae family could alleviate this problem, and permit a more accurate inference of subfamily relationships.

Substantial gene tree incongruence was also found with respect to the relationships among the three large clades within Meso-Papilionoidae. The sister relationship between Dalbergioids s.l. and Genistoids s.l. received only slightly higher quartet support than the two alternative hypotheses, which is consistent with previous results (Koenen et al., 2020b). Similarly, conflicting topologies affected most branches within Genisteae. By contrast, our analyses confirm that the genus *Aeschynomene* sensu RUDD (1955), which consisted of the former *A.* sect. *Aeschynomene* and *A.* sect. *Ochopodium* Vogel, is non-monophyletic (Ribeiro et al*.,* 2007). The recently re-established *Ctenodon* (= *A.* sect. *Ochopodium*, Cardoso et al., 2020) is sister to *Machaerium*, and these two genera form the sister group to *Dalbergia.*

4.5 | Conclusions and perspectives

The resources developed here for Fabaceae and in particular the genus *Dalbergia* bridge micro- and macro-evolutionary timescales and will hopefully facilitate community-driven efforts to advance legume genomics. Comprehensive sampling and sequencing by target capture of *Dalbergia* across its distribution range, and in particular from the hotspot of diversity in Madagascar, can yield valuable insights into the origin and diversification of the genus, thereby informing conservation policies and the taxonomic revision of Malagasy *Dalbergia*. The obtained sequence data will further serve to build a reference library for molecular identification of CITES-listed *Dalbergia* species, which would make a significant contribution toward the conservation of the valuable and endangered rosewoods.

ACKNOWLEDGEMENTS

We thank Sonja Hassold and the Missouri Botanical Garden (MBG) Madagascar team for organising and conducting field work, the Botanic Garden of the University of Zurich for the opportunity to sample legume plants from their living collection, and to MBG in St. Louis (USA) and the Paris herbarium (P) for access to their DNA bank and voucher specimens, respectively. We also thank Peter Phillipson and Nicholas Wilding for discussing sample determinations, *Dalbergia* taxonomy and Malagasy plant diversity. We are grateful to Claudia Michel for laboratory work and the Genetic Diversity Centre Zurich (GDC) for helpful support, in particular to Silvia Kobel for sequencing and Niklaus Zemp for bioinformatics. Finally, we thank Erik Koenen, Colin Hughes, Pete Lowry, Martin Fischer and Nicholas Wilding for their valuable inputs and comments on the manuscript.

AUTHOR CONTRIBUTIONS

SC and AW designed the study and collected samples. SZ assembled the draft *Dalbergia* transcriptome and designed target capture probes. SC and SZ analyzed data and wrote CAPTUREAL. SC, SF and AW wrote the manuscript with contributions from SZ.

DATA AVAILABILITY STATEMENT

Raw target capture sequencing reads generated for this study are deposited in the European Nucleotide Archive (ENA) at EMBL-EBI under accession number PRJEB41848 (https://www.ebi.ac.uk/ena/browser/view/PRJEB41848). Transcriptome sequencing reads as well as the draft *Dalbergia* transcriptome, sequences representing the initial 12,049 RNA probes and 6,555 target regions, the Fabaceae1005 and Dalbergia2396 probe sets, longer and taxon-specific reference sequences used for mapping, final alignments for the subfamily and species sets (all in FASTA format), and SNP data from the population set (VCF format) are available on Dryad (https://doi.org/10.5061/dryad.73n5tb2z7). The bioinformatics pipeline CAPTUREAL is available and further documented on Github (https://github.com/scrameri/CaptureAl). Because *Dalbergia* species are under threat from illegal exploitation, we have systematically refrained from making detailed distribution maps and precise geo-coordinates publicly available. Specimen records for collections from Madagascar are provided in the Catalogue of the Plants of Madagascar (Madagascar Catalogue, 2021), but with restricted public access to precise geo-coordinates (delivered on demand to bona fide researchers).

736

737 **ORCID**

738 Simon Crameri https://orcid.org/0000-0002-5516-1018

739 Simone Fior https://orcid.org/0000-0003-1173-1477

740 Alex Widmer https://orcid.org/0000-0001-8253-5137

741

References

Adema, F., Ohashi, H., & Sunarno, B. (2016). Notes on Malesian Fabaceae (Leguminosae-Papilionoideae) 17. The genus *Dalbergia*. *Blumea, 61*(3), 186–206. https://doi.org/10.3767/000651916X693905

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, *19*(5), 455–477. https://doi.org/10.1089/cmb.2012.0021

Barrett, M. A., Brown, J. L., & Yoder, A. D. (2013). Conservation: Protection for trade of precious rosewood. *Nature, 499*, 29. https://doi.org/10.1038/499029c

Bosser, J., & Rabevohitra, R. (2002). Tribe Dalbergieae. In D. J. Du Puy, J. N. Labat, R. Rabevohitra, J. F. Villiers, J. Bosser, & J. Moat (Eds.), *The Leguminosae of Madagascar* (pp. 321–361). Royal Botanical Gardens, Kew.

Brewer, G. E., Clarkson, J. J., Maurin, O., Zuntini, A. R., Barber, V., Bellot, S., Biggs, N., Cowan, R. S., Davies, N. M. J., Dodsworth, S., Edwards, S. L., Eiserhardt, W. L., Epitawalage, N., Frisby, S., Grall, A., Kersey, P. J., Pokorny, L., Leitch, I. J., Forest, F., & Baker, W. J. (2019). Factors affecting targeted sequencing of 353 nuclear genes from herbarium specimens spanning the diversity of Angiosperms. *Frontiers in Plant Science, 10*, 1102. https://doi.org/10.3389/fpls.2019.01102

Broad Institute (2019). *Picard Toolkit, version 2.21.3.* http://broadinstitute.github.io/picard

Buerki, S., & Baker, W. J. (2016). Collections-based research in the genomic era. *Biological Journal of the Linnean Society, 117*, 5–10. https://doi.org/10.1111/bij.12721

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics, 10*, 421. https://doi.org/10.1186/1471-2105-10-421

Cardoso, D. B. O. S., de Queiroz, L. P., Pennington, R. T., de Lima, H. C., Fonty, E., Wojciechowski, M. F., & Lavin, M. (2012). Revisiting the phylogeny of papilionoid legumes: New insights from comprehensively sampled early-branching lineages. *American Journal of Botany*, *99*(12), 1991–2013. https://doi.org/10.3732/ajb.1200380

Cardoso, D., Pennington, R. T., de Queiroz, L. P., Boatwright, J. S., Van Wyk, B. E., Wojciechowski, M. F., & Lavin, M. (2013). Reconstructing the deep-branching relationships of the papilionoid legumes. *South African Journal of Botany, 89*, 58–75. http://dx.doi.org/10.1016/j.sajb.2013.05.001

Cardoso, D. B. O. S., Mattos, C. M. J., Filardi, F., Delgado-Salinas, A., Lavin, M., de Moraes, P. L. R., Tapia-Pastrana, F., & de Lima, H. C. (2020). A molecular phylogeny of the pantropical papilionoid legume *Aeschynomene* supports reinstating the ecologically and morphologically coherent genus *Ctenodon*. *Neodiversity, 13*(1), 1–38. http://doi.org/10.13102/neod.131.1

Chaisson, M. J. P., Wilson, R. K., & Eichler, E. E. (2015). Genetic variation and the *de novo* assembly of human genomes. *Nature Reviews Genetics, 16*, 627–640. https://doi.org/10.1038/nrg3933

CITES (2020). *Appendices I, II and III*. Convention on International Trade in Endangered Species of Wild Fauna and Flora. Retrieved 27 January, 2021, from https://cites.org/eng/app/appendices.php

Crameri, S. (2020). *Phylogenomics, species discovery and integrative taxonomy in Dalbergia (Fabaceae) precious woods from Madagascar* [Doctoral thesis, ETH Zurich, Zurich, Switzerland. https://doi.org/10.3929/ethz-b-000487274

[dataset] Crameri, S., Fior, S., Zoller, S., & Widmer, A. (2022) Data from: A target capture approach for phylogenomic analyses at multiple evolutionary timescales in rosewoods (*Dalbergia* spp.) and the legume family (Fabaceae). European Nucleotide Archive (ENA). https://www.ebi.ac.uk/ena/browser/view/PRJEB41848

[dataset] Crameri, S., Fior, S., Zoller, S., & Widmer, A. (2022) Data from: A target capture approach for phylogenomic analyses at multiple evolutionary timescales in rosewoods (*Dalbergia* spp.) and the legume family (Fabaceae). Dryad Digital Repository. https://doi.org/10.5061/dryad.73n5tb2z7

Cutter, A. D. (2013). Integrating phylogenetics, phylogeography and population genetics through genomes and evolutionary theory. *Molecular Phylogenetics and Evolution, 69*(3), 1172–1185. https://doi.org/10.1016/j.ympev.2013.06.006

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G.T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. *Bioinformatics, 27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

de Carvalho, A.M. (1997). A synopsis of the genus *Dalbergia* (Fabaceae: Dalbergieae) in Brazil. *Brittonia, 49*, 87–109. https://doi.org/10.2307/2807701

de La Harpe, M., Paris, M., Karger, D. N., Rolland, J., Kessler, M., Salamin, N., & Lexer, C. (2017). Molecular ecology studies of species radiations: current research gaps, opportunities and challenges. *Molecular Ecology, 26*(10), 2608–2622. https://doi.org/10.1111/mec.14110

de La Harpe, M., Hess, J., Loiseau, O., Salamin, N., Lexer, C., & Paris, M. (2018). A dedicated target capture approach reveals variable genetic markers across micro- and macro-evolutionary time scales in palms. *Molecular Ecology Resources, 19*(1), 221–234. https://doi.org/10.1111/1755-0998.12945

de Sousa, F., Bertrand, Y. J. K., Nylinder, S., Oxelman, B., Eriksson, J. S., & Pfeil, B. E. (2014). Phylogenetic properties of 50 nuclear loci in Medicago (Leguminosae) generated using multiplexed sequence capture and next-generation sequencing. *PLoS ONE, 9*(10), e109704. https://doi.org/10.1371/journal.pone.0109704

Dowle, M., & Srinivasan, A. (2019). *data.table: Extension of `data.frame`. R package version 1.12.8*. https://CRAN.R-project.org/package=data.table

Doyle, J. J., & Doyle, J. L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin, 19*(1), 11–15.

Egan, A. N., & Vatanparast, M. (2019). Advances in legume research in the genomics era. *Australian Systematic Botany, 32*(6), 459–483. https://doi.org/10.1071/SB19019

Faircloth, B. C. (2016). PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics*, *32*(5), 786–788. https://doi.org/10.1093/bioinformatics/btv646

Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., & Glenn, T. C. (2012). Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology, 61*(5), 717–726. https://doi.org/10.1093/sysbio/sys004

831 Garrison, E. (2012). *Vcflib: A C++ library for parsing and manipulating VCF files*. Github.
832 https://github.com/ekg/vcflib
833 Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read
834 sequencing. *arXiv,* 1207.3907. https://arxiv.org/abs/1207.3907
835 Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X.,
836 Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A.,
837 Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., &
838 Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a
839 reference genome. *Nature Biotechnology, 29*, 644–652.
840 https://doi.org/10.1038/nbt.1883
841 Hahn, C., Bachmann, L., & Chevreux, B. (2013). Reconstructing mitochondrial genomes directly
842 from genomic next-generation sequencing reads – a baiting and iterative mapping
843 approach. *Nucleic Acids Research, 41*(13), e129. https://doi.org/10.1093/nar/gkt371
844 Hassold, S., Lowry II, P. P., Bauert, M. R., Razafintsalama, A., Ramamonjisoa, L., & Widmer, A.
845 (2016). DNA barcoding of Malagasy rosewoods: towards a molecular identification of
846 CITES-listed *Dalbergia* species. *PLoS ONE, 11*(6), e0157881.
847 https://doi.org/10.1371/journal.pone.0157881
848 Hughes, C. E., & Eastwood, R. (2006). Island radiation on a continental scale: Exceptional rates
849 of plant diversification after uplift of the Andes. *Proceedings of the National Academy*
850 *of Sciences of the United States of America, 103*(27), 10334–10339.
851 https://doi.org/10.1073/pnas.0601928103
852 Johnson, M. G., Gardner, E. M., Liu, Y., Medina, R., Goffinet, B., Shaw, A.J., Zerega, N. J. C., &
853 Wickett, N. J. (2016). HybPiper: Extracting coding sequence and introns for
854 phylogenetics from high-throughput sequencing reads using target enrichment.
855 *Applications in Plant Sciences, 4*(7), 1600016. https://doi.org/10.3732/apps.1600016
856 Johnson, M. G., Pokorny, L., Dodsworth, S., Botigué, L. R., Cowan, R. S., Devault, A., Eiserhardt,
857 W. L., Epitawalage, N., Forest, F., Kim, J. T., Leebens-Mack, J. H., Leitch, I. J., Maurin,
858 O., Soltis, D. E., Soltis, P. S., Wong, G. K.-S., Baker, W. J., & Wickett, N. J. (2019). A
859 universal probe set for targeted sequencing of 353 nuclear genes from any flowering
860 plant designed using k-medoids clustering. *Systematic Biology, 68*(4), 594–606.
861 https://doi.org/10.1093/sysbio/syy086
862 Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers.
863 *Bioinformatics, 24*(11), 1403–1405. https://doi.org/10.1093/bioinformatics/btn129
864 Jombart, T., & Ahmed, I. (2011). adegenet 1.3-1: new tools for the analysis of genome-wide
865 SNP data. *Bioinformatics, 27*(21), 3070–3071.
866 https://doi.org/10.1093/bioinformatics/btr521
867 Jones, M. R., & Good, J. M. (2016). Targeted capture in evolutionary and ecological genomics.
868 *Molecular Ecology, 25*(1), 185–202. https://doi.org/10.1111/mec.13304
869 Junier, T., & Zdobnov, E. M. (2010). The Newick utilities: high-throughput phylogenetic tree
870 processing in the UNIX shell. *Bioinformatics, 26*(13), 1669–1670.
871 https://onlinelibrary.wiley.com/doi/10.1111/mec.13304
872 Kamvar, Z. N., Tabima, J. F., & Grünwald, N. J. (2014). Poppr: an R package for genetic analysis
873 of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ, 2*, e281.
874 https://doi.org/10.7717/peerj.281

875 Katoh, K., & Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7:
876 improvements in performance and usability. *Molecular Biology and Evolution, 30*(4),
877 772–780. https://doi.org/10.1093/molbev/mst010
878 Knaus, B. J., & Grünwald, N. J. (2017). VCFR: a package to manipulate and visualize variant call
879 format data in R. *Molecular Ecology Resources, 17*(1), 44–53.
880 https://doi.org/10.1111/1755-0998.12549
881 Koenen, E. J. M., Kidner, C., Souza, E. R., Simon, M. F., Iganci, J. R., Nicholls, J. A., Brown, G. K.,
882 de Queiroz, L. P., Luckow, M., Lewis, G. P., Pennington, R. T., & Hughes, C. E. (2020a).
883 Hybrid capture of 964 nuclear genes resolves evolutionary relationships in the
884 mimosoid legumes and reveals the polytomous origins of a large pantropical radiation.
885 *American Journal of Botany, 107*(12), 1710–1735. https://doi.org/10.1002/ajb2.1568
886 Koenen, E. J. M., Ojeda, D. I., Steeves, R., Migliore, J., Bakker, F. T., Wieringa, J. J., Kidner, C.,
887 Hardy, O. J., Pennington, R. T., Bruneau, A., & Hughes, C. E. (2020b). Large-scale
888 genomic sequence data resolve the deepest divergences in the legume phylogeny and
889 support a near-simultaneous evolutionary origin of all six subfamilies. *New Phytologist,*
890 *225*(3), 1355–1369. https://doi.org/10.1111/nph.16290
891 Koenen, E. J. M., Ojeda, D. I., Bakker, F. T., Wieringa, J. J., Kidner, C., Hardy, O. J., Pennington,
892 R. T., Herendeen, P. S., Bruneau, A., & Hughes, C. E. (2021). The origin of the legumes
893 is a complex paleopolyploid phylogenomic tangle closely associated with the
894 Cretaceous-Paleogene (K-Pg) mass extinction event. *Systematic Biology, 70*(3), 508–
895 526. https://doi.org/10.1093/sysbio/syaa041
896 Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A., & Mayrose, I. (2015). Clumpak:
897 a program for identifying clustering modes and packaging population structure
898 inferences across *K*. *Molecular Ecology Resources, 15*(5), 1179–1191.
899 https://doi.org/10.1111/1755-0998.12387
900 Lachenaud, O. (2016). *Dalbergia* L. f., nom. cons. Pp. 101–153 in: *Flore du Gabon* (vol. 49), M.
901 S. M. Sosef, J. Florence, L. Ngok Banak, H. P. Bourobou Bourobou, P. Bissiengou et al.,
902 Margraf Publishers, Leiden.
903 Lavin, M., Pennington, R. T., Klitgaard, B. B., Sprent, J. I., de Lima, H. C., & Gasson, P. E. (2001).
904 The dalbergioid legumes (Fabaceae): delimitation of a pantropical monophyletic clade.
905 *American Journal of Botany, 88*(3), 503–533. https://doi.org/10.2307/2657116
906 Lee, A. K., Gilman, I. S., Srivastav, M., Lerner, A. D., Donoghue, M. J., & Clement, W. L. (2021).
907 Reconstructing Dipsacales phylogeny using Angiosperms353: issues and insights.
908 *American Journal of Botany, 108*(7), 1122–1142. https://doi.org/10.1002/ajb2.1695
909 Lemmon, A. R., Emme, S. A., & Lemmon, E. M. (2012). Anchored hybrid enrichment for
910 massively high-throughput phylogenomics. *Systematic Biology, 61*(5), 727–744.
911 https://doi.org/10.1093/sysbio/sys049
912 Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
913 *arXiv.* https://arxiv.org/abs/1303.3997
914 Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler
915 transform. *Bioinformatics, 25*(14), 1754–1760.
916 https://doi.org/10.1093/bioinformatics/btp324
917 Lindenbaum, P. (2015). *Jvarkit: java utilities for bioinformatics.* Github.
918 https://github.com/lindenb/jvarkit

LPWG (Legume Phylogeny Working Group) (2017). A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny. *Taxon, 66*(1), 44–77. https://doi.org/10.12705/661.3

Lunter, G., & Goodson, M. (2011). Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research, 21*, 936–939. https://doi.org/10.1101/gr.111120.110

Madagascar Catalogue (2021). *Catalogue of the Plants of Madagascar*. Missouri Botanical Garden, St. Louis, USA. Retrieved November 19, 2021, from http://www.tropicos.org/Project/Madagascar

Mamanova, L., Coffey, A. J., Scott, C. E., Kozarewa, I., Turner, E. H., Kumar, A., Howard, E., Shendure, J., & Turner, D. J. (2010). Target-enrichment strategies for next-generation sequencing. *Nature Methods, 7*, 111–118. https://doi.org/10.1038/nmeth.1419

Mandel, J. R., Dikow, R. B., Funk, V. A., Masalia, R. R., Staton, S. E., Kozik, A., Michelmore, R. W., Rieseberg, L. H., & Burke, J. M. (2014). A target enrichment method for gathering phylogenetic information from hundreds of loci: an example from the *Compositae*. *Applications in Plant Sciences, 2*(2), 1300085. https://doi.org/10.3732/apps.1300085

McMahon, M., & Hufford, L. (2004). Phylogeny of Amorpheae (Fabaceae: Papilionoideae). *American Journal of Botany, 91*(8), 1219–1230. https://doi.org/10.3732/ajb.91.8.1219

Mirarab, S., Reaz, R., Bayzid, M. S., Zimmermann, T., Swenson, M. S., & Warnow, T. (2014). ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*, *30*(17), i541–i548. https://doi.org/10.1093/bioinformatics/btu462

Mousavi-Derazmahalleh, M., Bayer, P. E., Hane, J. K., Valliyodan, B., Nguyen, H. T., Nelson, M. N., Erskine, W., Varshney, R. K., Papa, R., & Edwards, D. (2018). Adapting legume crops to climate change using genomic approaches. *Plant, Cell & Environment, 42*(1), 6–19. https://doi.org/10.1111/pce.13203

Nicholls, J. A., Pennington, R. T., Koenen, E. J. M., Hughes, C. E., Hearn, J., Bunnefeld, L., Dexter, K. G., Stone, G. N., & Kidner, C. A. (2015). Using targeted enrichment of nuclear genes to increase phylogenetic resolution in the neotropical rain forest genus *Inga* (Leguminosae: Mimosoideae). *Frontiers in Plant Science, 6*, 710. https://doi.org/10.3389/fpls.2015.00710

Ottenlips, M. V., Mansfield, D.H., Buerki, S., Feist, M. A. E., Downie, S. R., Dodsworth, S., Forest, F., Plunkett, G. M., & Smith, J. F. (2021). Resolving species boundaries in a recent radiation with the Angiosperms353 probe set: the *Lomatium packardiae*/*L. anomalum* clade of the *L. triternatum* (Apiaceae) complex. *American Journal of Botany, 108*(7), 1217–1233. https://doi.org/10.1002/ajb2.1676

Paradis, E., & Schliep, K. (2018). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics, 35*(3), 526–528. https://doi.org/10.1093/bioinformatics/bty633

Pelser, P. B., Kennedy, A. H., Tepe, E. J., Shidler, J. B., Nordenstam, B., Kadereit, J. W., & Watson, L. E. (2010). Patterns and causes of incongruence between plastid and nuclear Senecioneae (Asteraceae) phylogenies. American Journal of Botany *97*(5), 856–873. https://doi.org/10.3732/ajb.0900287

Peng, Z., Fan, W., Wang, L., Paudel, D., Leventini, D., Tillman, B. L., & Wang, J. (2017). Target enrichment sequencing in cultivated peanut (*Arachis hypogaea* L.) using probes

963 designed from transcript sequences. *Molecular Genetics and Genomics, 292*, 955–965.
964 https://doi.org/10.1007/s00438-017-1327-z
965 Prain, D. (1904). The Species of *Dalbergia* of South-Eastern Asia. *Annals of the Royal Botanic*
966 *Garden, Calcutta, 10*(1), 1–114.
967 Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using
968 multilocus genotype data. *Genetics, 155*(2), 945–959.
969 Prum, R. O., Berv, J. S., Dornburg, A., Field, D. J., Townsend, J. P., Lemmon, E. M., & Lemmon,
970 A. R. (2015). A comprehensive phylogeny of birds (Aves) using targeted next-
971 generation DNA sequencing. *Nature, 526*, 569–573.
972 https://doi.org/10.1038/nature15697
973 Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic
974 features. *Bioinformatics, 26*(6), 841–842.
975 https://doi.org/10.1093/bioinformatics/btq033
976 R Core Team (2020). R: A language and environment for statistical computing. R Foundation
977 for Statistical Computing, Vienna, Austria. https://www.R-project.org
978 Ribeiro, R. A., Lavin, M., Lemos-Filho, J. P., & Filho, C. (2007). The genus *Machaerium*
979 (Leguminosae) is more closely related to *Aeschynomene* Sect. *Ochopodium* than to
980 *Dalbergia*: Inferences from combined sequence data. *Systematic Botany, 32*(4), 762–
981 771. https://doi.org/10.1043/06-79.1
982 Rissler, L. J. (2016). Union of phylogeography and landscape genetics. *Proceedings of the*
983 *National Academy of Sciences of the United States of America, 113*(29), 8079–8086.
984 https://doi.org/10.1073/pnas.1601073113
985 Safonova, Y., Bankevich, A., & Pevzner, P. A. (2015). dipSPAdes: assembler for highly
986 polymorphic diploid genomes. *Journal of Computational Biology, 22*(6), 528–545.
987 https://doi.org/10.1089/cmb.2014.0153
988 Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing
989 phylogenetic trees. *Molecular Biology and Evolution, 4*(4), 406–425.
990 Schuurman, D., & Lowry II, P. P. (2009). The Madagascar rosewood massacre. *Madagascar*
991 *Conservation and Development, 4*, 98–102. https://doi.org/10.4314/mcd.v4i2.48649
992 Shah, T., Schneider, J. V., Zizka, G., Maurin, O., Baker, W., Forest, F., Brewer, G. E., Savolainen,
993 V., Darbyshire, I., & Larridon, I. (2021). Joining forces in *Ochnaceae* phylogenomics: a
994 tale of two targeted sequencing probe kits. *American Journal of Botany, 108*(7), 1201–
995 1216. https://doi.org/10.1002/ajb2.1682
996 Shavvon, R. S., Osaloo, S. K., Maassoumii, A. A., Moharrek, F., Erkul, S. K., Lemmon, A. R.,
997 Lemmon, E. M., Michalak, I., Muellner-Riehl, A. N., & Favre, A. (2017). Increasing
998 phylogenetic support for explosively radiating taxa: The promise of high-throughput
999 sequencing for *Oxytropis* (Fabaceae). *Journal of Systematics and Evolution, 55*(4), 385–
1000 404. https://doi.org/10.1111/jse.12269
1001 Siniscalchi, C. M., Hidalgo, O., Palazzesi, L., Pellicer, J., Pokorny, L., Maurin, O., Leitch, I. J.,
1002 Forest, F., Baker, W. J., & Mandel, J. R. (2021). Lineage-specific vs. universal: A
1003 comparison of the Compositae1061 and Angiosperms353 enrichment panels in the
1004 sunflower family. *Applications in Plant Sciences, 9*(7), e11422.
1005 https://doi.org/10.1002/aps3.11422
1006 Slater, G. S. C., & Birney, E. (2005). Automated generation of heuristics for biological sequence
1007 comparison. *BMC Bioinformatics, 6*, 31. https://doi.org/10.1186/1471-2105-6-31

Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular sequences. *Journal of Molecular Biology 147*, 195–197.

Sprent, J. I., Ardley, J., & James, E. K. (2017). Biogeography of nodulated legumes and their nitrogen-fixing symbionts. *New Phytologist, 215*(1), 40–56. https://doi.org/10.1111/nph.14474

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics, 30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Tange, O. (2011). GNU parallel: the command-line power tool. ;*Login:*, *36*(1), 42–47. https://www.usenix.org/system/files/login/articles/105438-Tange.pdf

Thomas, S. K., Liu, X., Du, Z. Y., Dong, Y., Cummings, A., Pokorny, L., Xiang, Q. Y., & Leebens-Mack, J. H. (2021). Comprehending Cornales: phylogenetic reconstruction of the order using the Angiosperms353 probe set. *American Journal of Botany, 108*(7), 1112–1121. https://doi.org/10.1002/ajb2.1696

Ufimov, R., Zeisek, V., Píšová, S., Baker, W.J., Fér, T., van Loo, M., Dobeš, C., & Schmickl, R. (2021). Relative performance of customized and universal probe sets in target enrichment: A case study in subtribe Malinae. *Applications in Plant Sciences, 9*(7), e11442. https://doi.org/10.1002/aps3.11442

UNODC (United Nations Office on Drugs and Crime) (2016a). Best practice guide for forensic timber identification. United Nations, New York. https://www.unodc.org/documents/Wildlife/Guide_Timber.pdf

UNODC (United Nations Office on Drugs and Crime) (2016b). World wildlife crime report: trafficking in protected species. United Nations, New York. https://www.unodc.org/documents/data-and-analysis/wildlife/World_Wildlife_Crime_Report_2016_final.pdf

UNODC (United Nations Office on Drugs and Crime) (2020). World wildlife crime report: trafficking in protected species. United Nations, New York. https://www.unodc.org/documents/data-and-analysis/wildlife/2020/World_Wildlife_Report_2020_9July.pdf

Vardeman, E., & Runk, J. V. (2020). Panama's illegal rosewood logging boom from *Dalbergia* retusa. *Global Ecology and Conservation, 23*, e01098. https://doi.org/10.1016/j.gecco.2020.e01098

Vatanparast, M., Powell, A., Doyle, J. J., & Egan, A. N. (2018). Targeting legume loci: A comparison of three methods for target enrichment bait design in Leguminosae phylogenomics. *Applications in Plant Sciences, 6*(3), e1036. https://doi.org/10.1002/aps3.1036

Vences, M., Wollenberg, K. C., Vieites, D. R., & Lees, D. C. (2009). Madagascar as a model region of species diversification. *Trends in Ecology and Evolution, 24*(8), 456–465. https://doi.org/10.1016/j.tree.2009.03.011

Waeber, P. O., Schuurman, D., Ramamonjisoa, B., Langrand, M., Barber, C. V., Innes, J. L., Lowry II, P. P., & Wilmé, L. (2019). Uplisting of Malagasy precious woods critical for their survival. *Biological Conservation, 235*, 89–92. https://doi.org/10.1016/j.biocon.2019.04.007

WCVP (2021). *World Checklist of Vascular Plants, version 2.0*. Facilitated by the Royal Botanic Gardens, Kew. Retrieved November 19, 2021, from http://wcvp.science.kew.org

Weitemier, K., Straub, S. C. K., Cronn, R. C., Fishbein, M., Schmickl, R., McDonnell, A., & Liston, A. (2014). Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. *Applications in Plant Sciences, 2*(9), 1400042. https://doi.org/10.3732/apps.1400042

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., & Yutani, H. (2019). Welcome to the Tidyverse. *The Journal of Open Source Software, 4*(43), 1686. https://doi.org/10.21105/joss.01686

Wilding, N., Phillipson, P. B., Andriambololonera, S., Bernard, R., Crameri, S., Rakotonirina, N., Rakotovao, C., Randrianaivo, R., Razakamalala, R., & Lowry II, P. P. (2021a). Taxonomic studies on Malagasy *Dalbergia* (Fabaceae). I. Two new species from northern Madagascar, and an emended description for *D. manongarivensis*. *Candollea, 76*(2), 237–249. https://doi.org/10.15553/c2021v762a4

Wilding, N., Phillipson, P. B., & Crameri, S. (2021b). Taxonomic studies on Malagasy *Dalbergia* (Fabaceae). II. A new name for *D. mollis* and the reinstatement of *D. chermezonii*. *Candollea, 76*(2), 251–257. https://doi.org/10.15553.c2021v762a5

Wojciechowski, M. F. (2013). Towards a new classification of Leguminosae: Naming clades using non-Linnaean phylogenetic nomenclature. *South African Journal of Botany, 89*, 85–93. http://dx.doi.org/10.1016/j.sajb.2013.06.017

Wojciechowski, M. F., Lavin, M., & Sanderson, M. J. (2004). A phylogeny of legumes (Leguminosae) based on analysis of the plastid *matK* gene resolves many well-supported subclades within the family. *American Journal of Botany, 91*(11), 1846–1862. https://doi.org/10.3732/ajb.91.11.1846

Yardeni, G., Viruel, J., Paris, M., Hess, J., Crego, C. G., de La Harpe, M., Rivera, N., Barfuss, M. H. J., Till, W., Guzmán-Jacob, V., Krömer, T., Lexer, C., Paun, O., & Leroy, T. (2021). Taxon-specific or universal? Using target capture to study the evolutionary history of rapid radiations. *Molecular Ecology Resources,* Early View, 4 October 2021. https://doi.org/10.1111/1755-0998.13523

Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T.-Y. (2016). GGTREE: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution, 8*(1), 28–36. https://doi.org/10.1111/2041-210X.12628

Zhang, C., Rabiee, M., Sayyari, E., & Mirarab, S. (2018). ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, *19*, 153. https://doi.org/10.1186/s12859-018-2129-y

Zhang, R., Wang, Y.-H., Jin, J.-J., Stull, G. W., Bruneau, A., Cardoso, D., de Queiroz, L. P., Moore, M. J., Zhang, S.-D., Chen, S.-Y., Wang, J., Li, D.-Z., & Yi, T.-S. (2020). Exploration of Plastid Phylogenomic Conflict Yields New Insights into the Deep Relationships of Leguminosae. *Systematic Biology, 69*(4), 613–622. https://doi.org/10.1093/sysbio/syaa013

Zhuang, W., Chen, H., Yang, M., Wang, J., Pandey, M.K., Zhang, C., Chang, W.-C., Zhang, L., Zhang, X., Tang, R., Garg, V., Wang, X., Tang, H., Chow, C.-N., Wang, J., Deng, Y., Wang, D., Khan, A. W., Yang, Q., Cai, T., Bajaj, P., Wu, K., ... & Varshney, R. K. (2019). The genome of cultivated peanut provides insight into legume karyotypes, polyploid

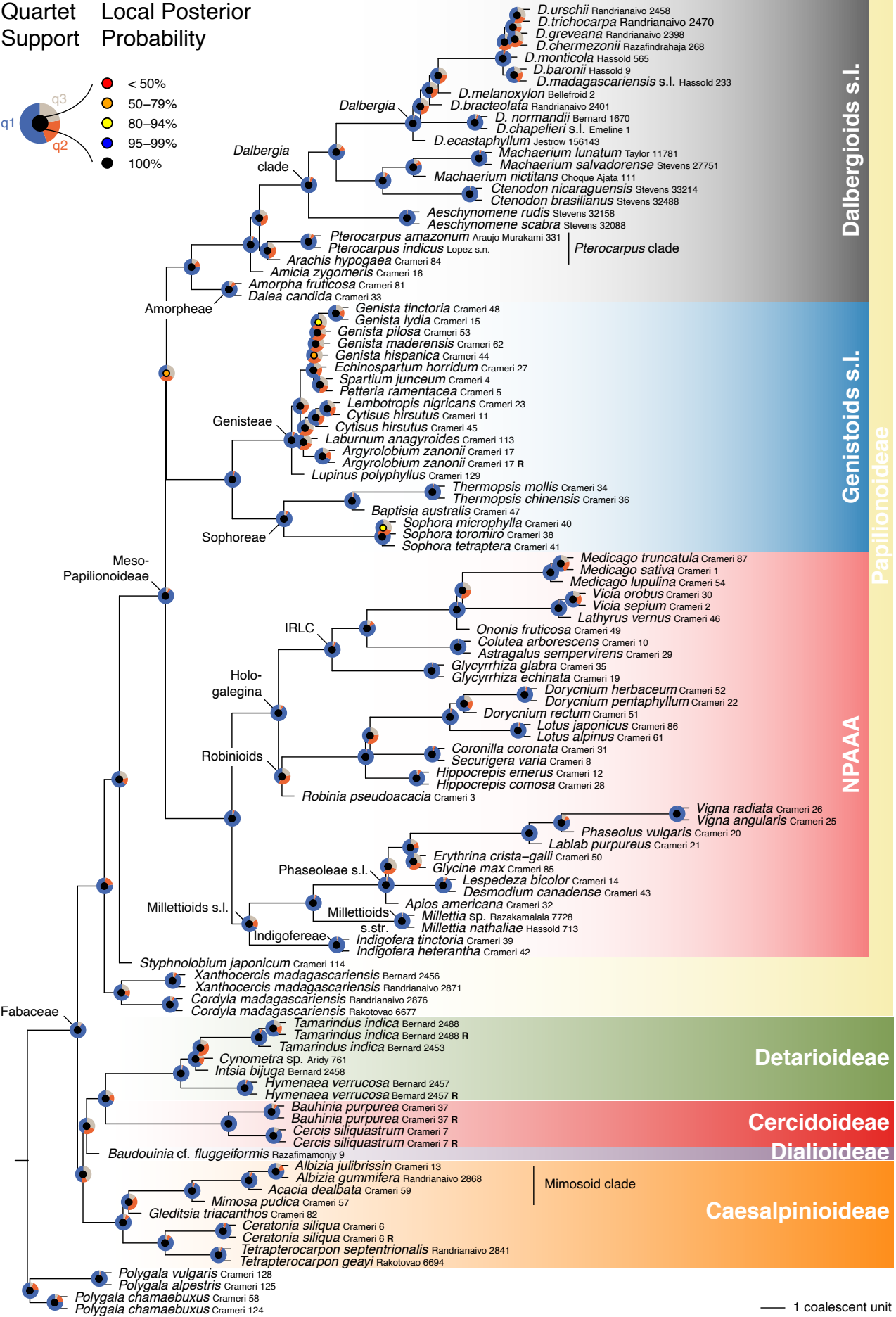
evolution and crop domestication. *Nature Genetics, 51*, 1–17. https://doi.org/10.1038/s41588-019-0402-2

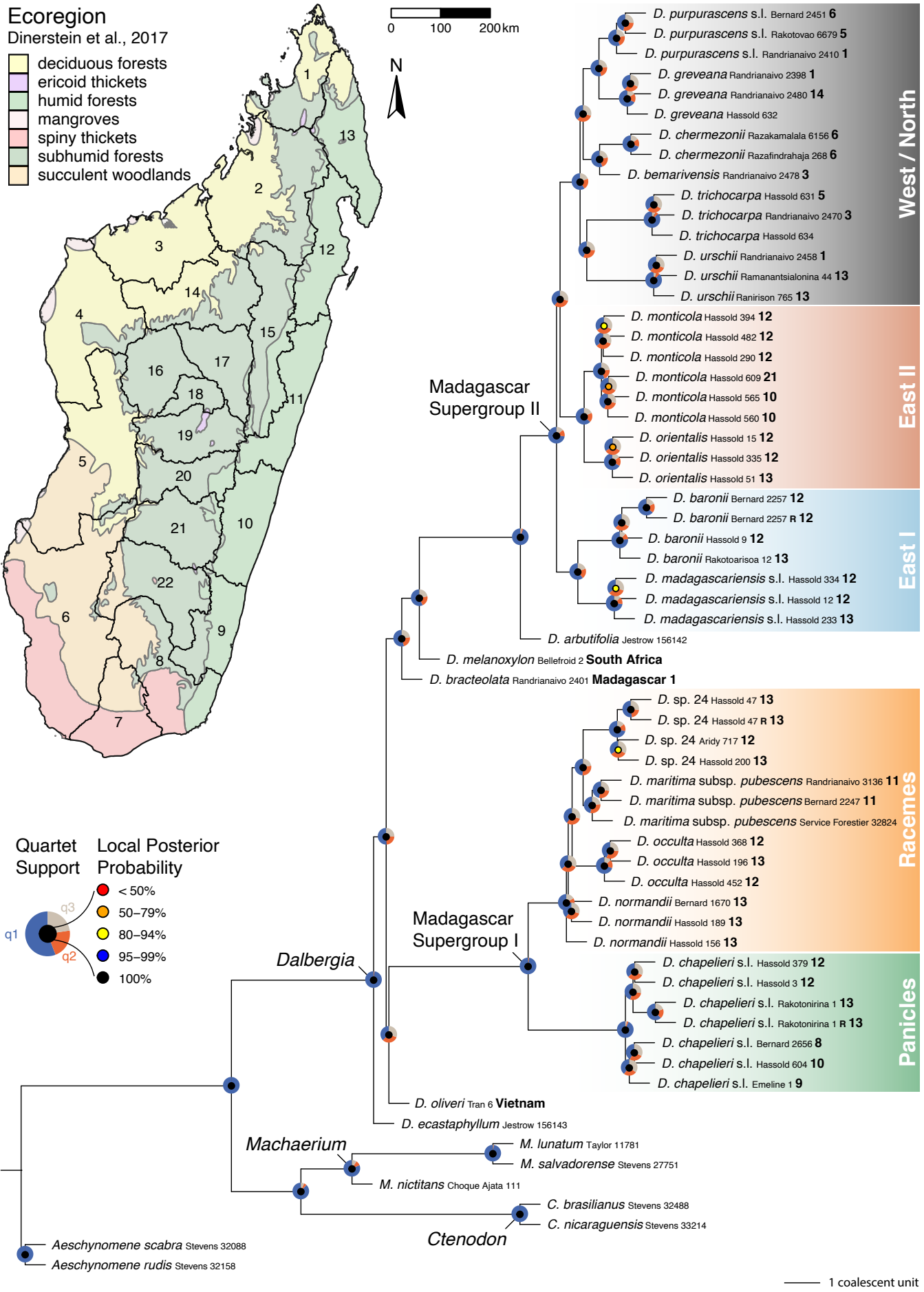
Figure Captions

FIGURE 1 Coalescent-based phylogeny of the subfamily set (n = 110) inferred using ASTRAL-III on 986 gene trees. Pie charts on each node denote the fraction of gene trees that are consistent with the shown topology (q1; blue), and with the first (q2; orange) and second (q3; gray) alternative topologies. Local posterior probabilities are shown as color-coded circles on each node (see inset legend). Replicate specimens are labelled with a bold 'R'. 860 gene trees (87.22%) had missing taxa. The overall normalized quartet score was 88.82%.

FIGURE 2 Coalescent-based phylogeny of the species set (n = 63) inferred using ASTRAL-III on 2,389 gene trees. Pie charts on each node denote the fraction of gene trees that are consistent with the shown topology (q1; blue), and with the first (q2; orange) and second (q3; gray) alternative topologies. Local posterior probabilities are shown as color-coded circles on each node (see inset legend). The geographic origins of accessions from Madagascar are indicated as bold numbers in the tree, which correspond to political regions of Madagascar, as well as to ecological regions (see inset map). Replicate specimens are labelled with a bold 'R'. 1,014 gene trees (42.44%) had missing taxa. The overall normalized quartet score was 85.42%.

FIGURE 3 PCA and NJ tree of the population set (n = 51) inferred from 60,204 biallelic SNPs with no missing data. Numbers adjacent to tree branches denote sampling locations as shown in Figure S4.


Quartet Support
Local Posterior Probability
q1
q2
q3
< 50%
50–79%
80–94%
95–99%
100%
D.urschii Randrianaivo 2458
D.trichocarpa Randrianaivo 2470
D.greveana Randrianaivo 2398
D.chermezonii Razafindrahaja 268
D.monticola Hassold 565
D.baronii Hassold 9
D.madagascariensis s.l. Hassold 233
D.melanoxylon Bellefroid 2
D.bracteolata Randrianaivo 2401
D. normandii Bernard 1670
D.chapelieri s.l. Emeline 1
D.ecastaphyllum Jestrow 156143
Dalbergia
Dalbergia clade
Machaerium lunatum Taylor 11781
Machaerium salvadorense Stevens 27751
Machaerium nictitans Choque Ajata 111
Ctenodon nicaraguensis Stevens 33214
Ctenodon brasilianus Stevens 32488
Aeschynomene rudis Stevens 32158
Aeschynomene scabra Stevens 32088
Pterocarpus amazonum Araujo Murakami 331
Pterocarpus indicus Lopez s.n.
Pterocarpus clade
Arachis hypogaea Crameri 84
Amicia zygomeris Crameri 16
Amorpha fruticosa Crameri 81
Dalea candida Crameri 33
Amorpheae
Dalbergioids s.l.
Genista tinctoria Crameri 48
Genista lydia Crameri 15
Genista pilosa Crameri 53
Genista maderensis Crameri 62
Genista hispanica Crameri 44
Echinospartum horridum Crameri 27
Spartium junceum Crameri 4
Petteria ramentacea Crameri 5
Lembotropis nigricans Crameri 23
Cytisus hirsutus Crameri 11
Cytisus hirsutus Crameri 45
Laburnum anagyroides Crameri 113
Argyrolobium zanonii Crameri 17
Argyrolobium zanonii Crameri 17 R
Lupinus polyphyllus Crameri 129
Genisteae
Thermopsis mollis Crameri 34
Thermopsis chinensis Crameri 36
Baptisia australis Crameri 47
Sophora microphylla Crameri 40
Sophora toromiro Crameri 38
Sophora tetraptera Crameri 41
Sophoreae
Genistoids s.l.
Papilionoideae
Meso-Papilionoideae
Medicago truncatula Crameri 87
Medicago sativa Crameri 1
Medicago lupulina Crameri 54
Vicia orobus Crameri 30
Vicia sepium Crameri 2
Lathyrus vernus Crameri 46
IRLC
Ononis fruticosa Crameri 49
Colutea arborescens Crameri 10
Astragalus sempervirens Crameri 29
Glycyrrhiza glabra Crameri 35
Glycyrrhiza echinata Crameri 19
Holo-galegina
Dorycnium herbaceum Crameri 52
Dorycnium pentaphyllum Crameri 22
Dorycnium rectum Crameri 51
Lotus japonicus Crameri 86
Lotus alpinus Crameri 61
Coronilla coronata Crameri 31
Securigera varia Crameri 8
Hippocrepis emerus Crameri 12
Hippocrepis comosa Crameri 28
Robinioids
Robinia pseudoacacia Crameri 3
NPAAA
Vigna radiata Crameri 26
Vigna angularis Crameri 25
Phaseolus vulgaris Crameri 20
Lablab purpureus Crameri 21
Erythrina crista–galli Crameri 50
Glycine max Crameri 85
Phaseoleae s.l.
Lespedeza bicolor Crameri 14
Desmodium canadense Crameri 43
Apios americana Crameri 32
Millettioids s.l.
Millettioids s.str.
Millettia sp. Razakamalala 7728
Millettia nathaliae Hassold 713
Indigofereae
Indigofera tinctoria Crameri 39
Indigofera heterantha Crameri 42
Styphnolobium japonicum Crameri 114
Xanthocercis madagascariensis Bernard 2456
Xanthocercis madagascariensis Randrianaivo 2871
Cordyla madagascariensis Randrianaivo 2876
Cordyla madagascariensis Rakotovao 6677
Fabaceae
Tamarindus indica Bernard 2488
Tamarindus indica Bernard 2488 R
Tamarindus indica Bernard 2453
Cynometra sp. Aridy 761
Intsia bijuga Bernard 2458
Hymenaea verrucosa Bernard 2457
Hymenaea verrucosa Bernard 2457 R
Detarioideae
Bauhinia purpurea Crameri 37
Bauhinia purpurea Crameri 37 R
Cercis siliquastrum Crameri 7
Cercis siliquastrum Crameri 7 R
Cercidoideae
Dialioideae
Baudouinia cf. fluggeiformis Razafimamonjy 9
Albizia julibrissin Crameri 13
Albizia gummifera Randrianaivo 2868
Acacia dealbata Crameri 59
Mimosa pudica Crameri 57
Gleditsia triacanthos Crameri 82
Ceratonia siliqua Crameri 6
Ceratonia siliqua Crameri 6 R
Tetrapterocarpon septentrionalis Randrianaivo 2841
Tetrapterocarpon geayi Rakotovao 6694
Mimosoid clade
Caesalpinioideae
Polygala vulgaris Crameri 128
Polygala alpestris Crameri 125
Polygala chamaebuxus Crameri 58
Polygala chamaebuxus Crameri 124
1 coalescent unit


Ecoregion
Dinerstein et al., 2017
deciduous forests
ericoid thickets
humid forests
mangroves
spiny thickets
subhumid forests
succulent woodlands
0
100
200 km
N
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
West / North
East II
East I
Racemes
Panicles
D. purpurascens s.l. Bernard 2451 6
D. purpurascens s.l. Rakotovao 6679 5
D. purpurascens s.l. Randrianaivo 2410 1
D. greveana Randrianaivo 2398 1
D. greveana Randrianaivo 2480 14
D. greveana Hassold 632
D. chermezonii Razakamalala 6156 6
D. chermezonii Razafindrahaja 268 6
D. bemarivensis Randrianaivo 2478 3
D. trichocarpa Hassold 631 5
D. trichocarpa Randrianaivo 2470 3
D. trichocarpa Hassold 634
D. urschii Randrianaivo 2458 1
D. urschii Ramanantsialonina 44 13
D. urschii Ranirison 765 13
D. monticola Hassold 394 12
D. monticola Hassold 482 12
D. monticola Hassold 290 12
D. monticola Hassold 609 21
D. monticola Hassold 565 10
D. monticola Hassold 560 10
D. orientalis Hassold 15 12
D. orientalis Hassold 335 12
D. orientalis Hassold 51 13
D. baronii Bernard 2257 12
D. baronii Bernard 2257 R 12
D. baronii Hassold 9 12
D. baronii Rakotoarisoa 12 13
D. madagascariensis s.l. Hassold 334 12
D. madagascariensis s.l. Hassold 12 12
D. madagascariensis s.l. Hassold 233 13
Madagascar Supergroup II
D. arbutifolia Jestrow 156142
D. melanoxylon Bellefroid 2 South Africa
D. bracteolata Randrianaivo 2401 Madagascar 1
D. sp. 24 Hassold 47 13
D. sp. 24 Hassold 47 R 13
D. sp. 24 Aridy 717 12
D. sp. 24 Hassold 200 13
D. maritima subsp. pubescens Randrianaivo 3136 11
D. maritima subsp. pubescens Bernard 2247 11
D. maritima subsp. pubescens Service Forestier 32824
D. occulta Hassold 368 12
D. occulta Hassold 196 13
D. occulta Hassold 452 12
D. normandii Bernard 1670 13
D. normandii Hassold 189 13
D. normandii Hassold 156 13
Madagascar Supergroup I
D. chapelieri s.l. Hassold 379 12
D. chapelieri s.l. Hassold 3 12
D. chapelieri s.l. Rakotonirina 1 13
D. chapelieri s.l. Rakotonirina 1 R 13
D. chapelieri s.l. Bernard 2656 8
D. chapelieri s.l. Hassold 604 10
D. chapelieri s.l. Emeline 1 9
Dalbergia
D. oliveri Tran 6 Vietnam
D. ecastaphyllum Jestrow 156143
Machaerium
M. lunatum Taylor 11781
M. salvadorense Stevens 27751
M. nictitans Choque Ajata 111
C. brasilianus Stevens 32488
C. nicaraguensis Stevens 33214
Ctenodon
Aeschynomene scabra Stevens 32088
Aeschynomene rudis Stevens 32158
Quartet Support
Local Posterior Probability
q1
q2
q3
< 50%
50–79%
80–94%
95–99%
100%
1 coalescent unit