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

Simon Crameri1 | Simone Fior1 | Stefan Zoller1,2 | Alex Widmer1

1Institute of Integrative Biology, ETH Zurich, Zürich, Switzerland
2Genetic Diversity Centre (GDC), ETH Zurich, Zürich, Switzerland

Correspondence
Simon Crameri and Alex Widmer, Institute of Integrative Biology, ETH Zurich, Zürich, Switzerland.
Emails: sfcrameri@gmail.com (S.C.); alex.widmer@usys.ethz.ch (A.W.)

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.

Handling Editor: C. Alex Buerkle

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 macroevolutionary 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 analysed and loci sequenced; and (iii) building tools for reproducible bioinformatic analyses of such data sets. To address these challenges, we developed target capture probes for genomic 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 the general purpose bioinformatic pipeline CapturAl. Phylogenomic analyses of Malagasy *Dalbergia* species 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 bioinformatic resources, including the Fabaceae1005 and Dalbergia2396 probe sets, will hopefully advance systematics and ecological genetics research in legumes, and promote conservation of the highly diverse and endangered *Dalbergia* rosewoods.

Keywords
*Dalbergia*, Fabaceae, Leguminosae, phylogenomics, rosewood, target capture
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. Although 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 macroevolutionary 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 orthologues and paralogues, and are characterized by a preserved 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 (Brewer et al., 2019; de La Harpe et al., 2017). 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 macroevolutionary, phylogeographic and microevolutionary 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, Kidner, et al., 2020; 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 microevolutionary 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., 2022).

In addition to 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 analysed using the PHYLUCe (Faircloth, 2016) or HyPhy (Johnson et al., 2016) bioinformatic pipelines. PHYLUCe was developed for analysis of sequences flanking ultraconserved genomic elements and has mainly been used at macro-evolutionary and phylogeographic timescales in animal systems, whereas HyPhy is optimized for data sets derived from probes designed in exons using Hye-Seq (Weitemier et al., 2014).

There is thus a need for existing tools to be expanded with pipelines that are applicable at deep to shallow evolutionary timescales (de La Harpe et al., 2017), while being independent from high-quality annotated genomes or transcriptomes.

Dalbergia L.f. (Fabaceae) is a pantropical and ecologically diverse plant genus with c. 270 currently accepted species (WCVP, 2021), some of which have been described relatively recently (e.g., Adema et al., 2016; Lachenaud, 2016; Wilding, Phillipson, Andriambololonera, et al., 2021; Wilding, Phillipson, & Crameri, 2021). Numerous arborescent species are a source of rosewood (Bosser & Rabevohitra, 2002; Prain, 1904), a high-quality timber sought-after on the international market and cause of conservation concern (Schuurman & Lowry II, 2009; Waeger et al., 2019). National and international regulations have been established, aiming at sustainable exploitation and revenues (Barrett et al., 2013; CITES, 2020), but illegal logging and trade continues (UNODC, 2016b, 2020; Vardeman & Runk, 2020). The effective implementation of regulations demands that species are reliably recognized and that extant population sizes are estimated to assess the potential threat status. Developing a comprehensive understanding of species diversity in Dalbergia and their evolutionary history, as well as a thorough knowledge of the ecology and distribution of many traded species, has been hampered by several factors. There is a shortage of collections and experts focusing on this taxonomically challenging genus, and current treatments heavily rely on leaves and flowers and/or fruits for identification (Bosser & Rabevohitra, 2002; de Carvalho, 1997; Lachenaud, 2016), which are rarely encountered together in the field. As a result, the taxonomy of the genus is in need of extensive revision (Wilding, Phillipson, Andriambololonera, et al., 2021), 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 probe 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 probe set). Both probe sets represent a subset of 6555 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 bioinformatic 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 data sets, 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 analysing target capture data of various legume subfamilies. Finally, we test the utility of this approach at a microevolutionary scale, and assess genetic variation among individuals and populations of two closely related *Dalbergia* species from Madagascar.

2 | MATERIALS AND METHODS

2.1 | Design of target capture probes and reference sequences

We produced a transcriptome assembly of a cultivated individual of *Dalbergia madagascariensis* subsp. *antongilensis* Bosser & R. Rabev., based on 63 million paired-end sequencing reads generated on an Illumina HiSeq 2000 platform. We performed de novo assembly of the transcriptome using Trinity release 2012-01-25 (Grabherr et al., 2011), resulting in 146,484 scaffolds, which were between 201 and 17,129 bp long, with a mean length of 815 bp (see Appendix S1). We then pairwise aligned the *Dalbergia* transcriptome with reference genomes of five legume species available in public databases to generate a set of 12,049 probes extracted from the *Cajanus cajan* (L.) Millsp. v1.0 genome (https://www.ncbi.nlm.nih.gov/assembly/GCF_000340665.1) targeting 6555 conserved genomic regions (see Appendix S1). This probe set was used for synthesis of hybridizing probes at myBaits Custom Target Capture Kits (Arbor Biosciences; https://arborbiosci.com).

2.2 | Taxon sampling for target capture probes validation

We created three taxon sets with contrasting levels of evolutionary divergence, ranging from subfamilies to species to populations. The subfamily set (Table S1) included five of the six legume subfamilies, as recognized in the most recent treatment (LPWG, 2017), and comprised 104 individuals (110 samples, six replicates; 99 species including three outgroups). Three species of *Polygala* Tourn. ex L. (Polygalaceae) were included as the outgroup for the subfamily set. The species set (Table S2) included members of the closely related genera *Dalbergia* (at least 19 species), *Maachaerium* 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 11 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 & 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 the 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 samples, and of museum specimens with silica gel dried leaf material. We obtained short read data by combining sequencing runs from an Illumina MiSeq (2 × 300bp paired-end sequencing, 99 libraries) at the Genetic Diversity Centre (GDC) Zurich, an Illumina HiSeq 4000 (2 × 150bp paired-end sequencing, 88 libraries) at the Functional Genomics Center Zurich (FGCZ) or Fasteris SA (Plan-les-Ouates, Switzerland), and an Illumina NovaSeq 6000 SP flow cell (2 × 150bp paired-end sequencing, 9 libraries) at the FGCZ. We repeated DNA extraction, hybridization and target enrichment sequencing for nine individuals (replicates) to assess reproducibility. One sample (Hassold S65) was represented in each taxon set, nine samples were represented in both the species and population sets, and nineteen samples were represented in both the subfamily and species sets.

2.4 | CAPTUREAl bioinformatics pipeline

The bioinformatic pipeline CAPTUREAl was developed for this project and is accessible on GitHub (https://github.com/scrameri/CaptureAl).
As a documented sequence of scripts. These include bash and R scripts (R Core Team, 2022) to manage and visualize data with APE version 5.3 (Paradis & Schliep, 2018), data.table version 1.12 (Dowle & Srinivasan, 2019), and tidyverse version 1.3.0 (Wickham et al., 2019). Where appropriate, computations are carried out for multiple samples or target regions in parallel using GNU parallel (Tange, 2011). The CAPTUREAL pipeline streamlines the mapping of quality-trimmed reads to target regions, the exclusion of loci targeting multicopy genes and taxa with insufficient data coverage, the alignment of orthologous loci for downstream phylogenetic analyses, and the generation of longer and taxon-specific reference sequences for population genomic analyses. At various critical steps, the pipeline outputs summary statistics and graphs that inform the user on the effects of specific filtering parameters, allowing for informed parameter choices.

The pipeline is divided into seven steps to process quality-filtered reads (Figure 1). 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 neighbouring and overlapping alignments to (7) generate longer and more representative reference sequences as starting points for population genomic analyses and reiteration of steps 1 to 5. Such reiteration can improve mapping success, and can mitigate potential biases arising from the initial reference sequences used (Hahn et al., 2013). This can improve the assembly and alignment of target regions for phylogenetic analyses, and remapping Illumina reads to taxon-specific reference sequences can also improve variant calling and subsequent population genomic analyses. CAPTUREAL includes scripts to filter raw variants and to convert them to genind and genlight objects for various population genetic and genomic analyses in R, and to sample single nucleotide polymorphisms (SNPs) for ST prevented analyses (Pritchard et al., 2000) and similar programs.

In our analyses, we executed the pipeline separately and iteratively for different taxon sets. We first applied steps 1 to 5 to representative samples of each of the subfamily and species sets, followed by steps 6 and 7 to generate longer and taxon-specific reference sequences for target regions that were efficiently enriched in these taxon sets. We 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 Appendix S1). We also performed steps 6 and 7 after the second iteration of the species set analysis to produce reference sequences for population genomic analyses of the population set. Bioinformatic analyses were carried out on a multicore 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 Appendix S1.

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 filtered for adequate coverage across samples using filter.coverages.R, which allows to apply filtering thresholds that are informed by visualizations of coverage statistics (see Appendix S1). The main output of step 1 are BAM files, coverage statistics, and 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 downstream analyses by repeating Steps 1–5 or 1–7.
for each sample and target region using \texttt{dipSPAdes} (Safonova et al., 2015). In the process, \texttt{dipSPAdes} uses contiguous sequences generated by \texttt{SPAdes} (Bankevich et al., 2012) and seeks to collapse overlapping contiguous sequences representing alternate alleles of heterozygous individuals (haplocontigs) to a single sequence, and to extend collapsed sequences to fewer and longer consensus contiguous sequences (contigs hereafter; Safonova et al., 2015). The main output of step 2 is a FASTA file with zero to multiple contigs 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), due to specific capture of paralogues (Johnson et al., 2016), or haplocontigs that failed to be merged into a consensus contig by \texttt{dipSPAdes}. The most likely orthologous contig(s) of each sample and each target region are determined using the exhaustive Smith-Waterman alignment (Smith & Waterman, 1981) implemented in \texttt{exonerate} (Slater & Birney, 2005). For each sample and target region, the contig showing the highest \texttt{exonerate} alignment score to the target region reference sequence is assumed to be the most likely orthologous contig of that target region, if the normalized alignment score (defined as the raw \texttt{exonerate} alignment score divided by the target alignment length) meets a user-specified threshold. Any other contig that overlaps with the best-matching contig but shows a lower alignment score is assumed to be paralogous or an alternative haplocontig, and is discarded from downstream analyses. Further contigs that do not overlap with one another or the best-matching contig, but align to other parts of the target region with the required normalized alignment score, are retained. Such contigs were often observed in assemblies of degraded museum specimens and probably represent fragments of the same region. These fragments can therefore be prepended or appended to the best-matching contig using the directionality and number of gap-filling characters (hyphens) as indicated by the \texttt{exonerate} alignment statistics, to form a single contiguous sequence per sample and target region (orthologous contig hereafter, see Appendix S1). The main output of step 3 is a text file containing the \texttt{exonerate} alignment statistics, and 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 orthologues of the target regions (capture specificity). Target regions are filtered for high capture sensitivity and specificity across focal taxa using \texttt{filtervisual}. \texttt{assemblies.R}, which applies filtering thresholds that are set by the user as informed by visualizations of \texttt{exonerate} alignment statistics generated in step 3, before any contigs are selected or combined. Capture sensitivity thresholds can be set globally to remove generally poorly sequenced samples (i.e., minimum fraction of target regions with at least one contig) or target regions (i.e., minimum fraction of samples with at least one contig). Capture specificity thresholds can be set as the required fraction of samples belonging to a specified taxon group and passing a certain filtering threshold in order for a target region to be retained. If taxon groups are defined, the indicated capture sensitivity and specificity thresholds need to be met in each considered taxon group, thus preventing target regions from being systematically missing in taxon groups with few available samples (see Appendix S1). 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 phylogenomic 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). The second and more stringent iteration required at least 75% (subfamily set) or 85% (species set) of samples to have one or more contigs, as indicated by the \texttt{exonerate} alignment statistics generated in step 3, for a target region to be retained. Furthermore, we only retained target regions for which no more than two contigs had been assembled in at least 50% (subfamily set) or 70% (species set) of samples of each taxon group (see Appendix S1).

### 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 \texttt{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 Appendix S1). The main output of step 5 are trimmed alignments for each kept target region.

### 2.4.6 Step 6: Merging of overlapping alignments

This optional step aims at resolving potential overlaps between physically close alignments generated in step 5, which may
negatively affect downstream analyses. 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 across the alignment 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 nonreciprocal hits between alignment ends of target regions located on the same genomic scaffold or chromosome (identified from the name of the target region, if available). All orthologous contigs that are part of different, overlapping alignments are then written to a single FASTA file and aligned using MAFFT. The resulting alignments consist of more sequences than samples and need to be collapsed to represent neighbouring target regions of the same individual as a single sequence (supercontig), an automatic process that can be visually inspected. 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 Appendix S1). The main output of step 6 are nonoverlapping trimmed alignments for each kept target region.

2.4.7 | Step 7: Generation of representative reference sequences

In this optional step, the target region alignments generated in the two previous steps are used to produce longer and more representative target region reference sequences. This can mitigate potential shortcomings or biases arising from the reference sequences used in step 1. 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 Appendix S1). 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 nonoverlapping 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 filtered alignments using filter:visual.alignments.R, which allows filtering thresholds that are informed by visualizations of alignment statistics (see Appendix S1) to be applied. We used the filtered alignments after the second iteration of step 6 for phylogenomic analyses.

2.5 | Phylogenomic analyses

We performed phylogenomic 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 Appendix S1). 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 Appendix S1). For the subfamily set, we additionally evaluated the quartet support for 15 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 genomic analyses

We carried out population genomic analyses for the population set by relying on SNPs inferred from variant calling, which was based on mapped reads and revealed individual-level allelic variation. 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 HCard 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 Appendix S1). We then used vcnR 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 neighbour-joining
(step 7) and used as references for mapping quality-trimmed reads. Consensus sequences were generated from these target regions using blast alignments (step 6). Longer and more representative sequences were successfully merged, resulting in 2468 or 3736 nonoverlapping alignments (step 5), of which 726 represented the same regions in both sets. The corresponding tiled probe sequences (3273 for the Fabaceae1005 probe set and 6190 for the Dalbergia2396 probe set) extracted from the initial 12,049 probes are deposited on Dryad (https://doi.org/10.5061/dryad.73n5tb2z7) along with refined taxon-specific reference sequences. Corresponding gene annotations in the Cajanus cajan genome are given in Tables S6 and S7. For phylogenomic 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 2389 (species set) alignments.

Quality-trimmed reads mapped to all 2396 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 S9). 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 three SNPs per target region (or all SNPs if less than three SNPs remained) resulted in a subset of 7156 SNPs for Structure analyses.

2.7 | Comparison to the Angiosperms353 probe set

We used BLAST+ to search for correspondence between our two final sets of reference sequences (i.e., the consensus sequences of genomic regions targeted by the Fabaceae1005 and Dalbergia2396 probe sets, respectively) and 11,781 supercontigs representing 42 angiosperm taxa and 353 target regions of the Angiosperms353 probe set (Johnson et al., 2019). We performed reciprocal BLAST+ searches applying a maximum expect value (evalue) of 1E-04, a minimum of 100bp alignment length, and retaining only the best hits between the same pair of target regions (see Appendix S1).

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 12 representative samples, reads mapped to 6519 or 6287 of the 6555 initial target regions in the subfamily or species set, respectively (step 1). Of these we retained 3436 or 4908 target regions, which showed adequate coverage across taxon groups. After assembly (step 2) and orthology assessment (step 3), 2710 or 4181 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 2468 or 3736 nonoverlapping 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). In the second iteration, we retained 1917 or 3418 target regions with adequate coverage (Figures S5 and S6), of which 1020 or 2407 passed the specificity and sensitivity filters of higher stringency (step 4) after assembly (see Figures S7 and S8 for visualizations). Merging of overlapping alignments in 15 or 11 regions yielded 1005 (subfamily set) or 2396 (species set) distinct alignments (step 5), of which 726 represented the same regions in both sets. The corresponding tiled probe sequences (3273 for the Fabaceae1005 probe set and 6190 for the Dalbergia2396 probe set) extracted from the initial 12,049 probes are deposited on Dryad (https://doi.org/10.5061/dryad.73n5tb2z7) along with refined taxon-specific reference sequences. Corresponding gene annotations in the Cajanus cajan genome are given in Tables S6 and S7. For phylogenomic 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 2389 (species set) alignments.

Quality-trimmed reads mapped to all 2396 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 S9). 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 three SNPs per target region (or all SNPs if less than three SNPs remained) resulted in a subset of 7156 SNPs for Structure analyses.

3.2 | Phylogenomic analyses across legumes

Phylogenomic analysis of 986 alignments recovered each of the five sampled subfamilies as monophyletic, and many well-established clades and relationships received ≥95% support (blue and black node support dots) using both the gene tree summary method ASTRAL-III (Figure 2) 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-Dipterygaceae-Ambaranaceae (ADA, Cardoso et al., 2012), Cladrastis (Wojciechowski, 2013) and Meso-Papilionoideae (Wojciechowski, 2013) clades within Papilionoideae. We also recovered the Sophoreae and Genistae 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 (represented by a single accession of Amicia zygomeris DC.), Pterocarpus and Dalbergia subclades (Lavin et al., 2001), respectively. Ctenodon brasilianus (Poir.) D.B.O.S.Cardoso, P.L.R.Moras & H.C.Lima and C. nicaraguensis (Oerst.) A.Delgado were found to be more closely related to Machaerium than to Aeschynomene. Within the NPAAA we recovered the Millettiod 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 (Figures 2, S1). In particular, a clade comprising Caesalpinioideae, Cercidioideae, 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 14 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 2 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 parageneal with respect to Lembotropis nigricans. Pairs of replicates each grouped together (Figures 2, S1).

3.3 Phylogenetic analyses in Dalbergia

Phylogenomic analysis of 2389 alignments recovered samples of Dalbergia as monophyletic with ≥95% support (blue and black node support dots) using both ASTRAL-III (Figure 3) 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 chapeleri s.l., while the remaining samples belonged to a morphologically divergent sister group containing three monophyletic species and a basal and paraphyletic D. normandii. Within Supergroup II, two clades contained species distributed in the humid east of Madagascar, while the third contained species distributed in the seasonally dry west and north of the island. Within D. chapeleri s.l. and D. monticola, which were each represented by six individuals, we observed geographic structure, with specimens from northeast and southeast Madagascar forming sister groups. Pairs of replicates each grouped together (Figures 3, S3).

3.4 Population genomic analyses

Principal component analysis revealed three distinct clusters of individuals along principal component (PC) 1 (explaining 27.58% of the total variation) and PC 2 (11.26%; Figure 4a). Individuals of Dalbergia orientalis separated along PC1, while individuals originally attributed to D. monticola formed two distinct groups mainly along PC2. The unexpected smaller cluster (D. sp. B, which corresponds to D. sp. 17 in Crameri (2020); in purple in Figure 4) comprised samples from a single broad sampling location in north-eastern Madagascar (location 5, see Figure S4 and Table S3) where both D. monticola and D. orientalis were also collected. The same three clusters were also recovered in STRUCTURE analyses (Figures S3b–c and S11), where biologically meaningful clustering solutions were found for K = 2 (separating D. orientalis from the rest) and K = 3 (further separating the unexpected smaller cluster). Within D. orientalis and the larger cluster of true D. monticola, the NJ tree reflects isolation by distance at a broad geographical scale, separating specimens from north-eastern (locations 1 to 6), central-eastern (locations 7 and 8) and south-eastern Madagascar (locations 9 to 13; Figures 3a and S4). A similar geographic pattern was recovered by STRUCTURE assuming K = 5 and K = 7 (Figure 4c), although these clustering solutions received much lower support (Figure 4b). Clustering solutions assuming higher K did not recover significant additional structure (Figure S11).
3.5 Comparison to the Angiosperms353 probe set

Results of the BLAST+ analyses are shown in Tables S8 and S9. In summary, 34 (3.38%) of the Fabaceae1005 target regions and 75 (3.13%) of the Dalbergia2396 target regions were reciprocal best hits to 32 (9.07%) and 65 (18.41%) target regions of the Angiosperms353 set, respectively. In twelve cases, there were reciprocal hits between a single Angiosperms353 region and two of our regions. All these cases involved contiguous subregions in longer supercontigs of the Angiosperms353 set corresponding to two contiguous regions in the Cajanus cajan genome.

4 DISCUSSION

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

4.1 Target capture probes

The target capture probes presented here are part of a growing collection of genomic resources for legume phylogenomics. Other
probe sets for target capture in legumes have been developed, focusing on different groups within the family, and designed or validated at the level of legume species (Peng et al., 2017), genera (e.g., de Sousa et al., 2014; Nicholls et al., 2015; Shavvon et al., 2017), or above (Koenen, Kidner et al., 2020; Vatanparast et al., 2018), and across angiosperms (Johnson et al., 2019). Our reciprocal BLAST analyses revealed that at least 65 (18.41%) genomic regions targeted by the Angiosperms353 probe set are likely also targeted by the Dalbergia2396 probe set. The fraction of target regions overlapping with the Angiosperms353 set was nearly identical in the Fabaceae1005 set (3.38%) compared to the Dalbergia2396 set (3.13%). This suggests that a possibly higher degree of conservation in genomic regions targeted by the Fabaceae1005 set compared to the Dalbergia2396 set may refer to divergence of the Cajanus cajan genome on which our probe sequences were based, rather than a higher degree of conservation across angiosperms. In light of the considerable overlap between the Dalbergia2396 and Angiosperms353 probe sets, it would be interesting to extend such comparisons to identify overlaps and complementarity in probe sets specifically designed for legume phylogenomics. Capture of additional, less conserved target regions across the legume family could be achieved by designing multiple probes for hybridization in the same target region in different legume groups, as applied for studies across angiosperms (Johnson et al., 2019). Such a probe design could profit from existing legume probe sets but should rely on a stringent selection of targets that accounts for paralogues (Vatanparast et al., 2018), which originated as a consequence of multiple whole-genome duplication events in legumes (Egan & Vatanparast, 2019; Koenen et al., 2021).

In this study, we enriched DNA libraries from three taxon sets spanning microevolutionary (populations) to macroevolutionary (family) timescales, using a single set of 12,049 RNA probes targeting 6555 genomic regions conserved across five Meso-Papilionoideae genomes and a Dalbergia transcriptome. We then identified 2396 and 1005 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., 2022). To further reduce costs, future target enrichment experiments in Dalbergia could be focused on genomic regions targeted by the 6190 sequences of the Dalbergia2396 probe set.

### 4.2 | CaptureAL bioinformatic 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 challenging task of de novo assembly of contigs from the large pool of sequencing reads representative of thousands of loci (reviewed by Chaisson et al., 2015). Likewise, alignments are conducted in clearly defined target regions in which overlap among individual contigs is higher. However, assembly per region is more time-consuming and requires reference sequences for the initial mapping step. This might introduce a reference bias when divergent sequences are not mapped (Lunter & Goodson, 2011). We addressed this problem by generating consensus sequences that are representative of a given taxon set and by limiting analyses to target regions that can be efficiently recovered in all groups of that taxon set. These set-specific reference sequences can then be used to iteratively refine mapping, assembly, and target region filtering for any set of taxa. Our approach is conceptually similar to the HysPiPER pipeline (Johnson et al., 2016), which also employs a mapping-assembly strategy, and uses depth of coverage and percent identity to the target region to choose between multiple contigs, before it identifies intron/exon boundaries using target peptide sequences and extracts coding sequences for alignment. While the HysPiPER pipeline is designed specifically for the Hys-SEQ approach (Johnson et al., 2016), in which exons are the primary targets and flanking noncoding regions are used as supplementary data for analyses at shallow evolutionary scales (Weitemier et al., 2014), CaptureAL is more general in scope and neither requires nor leverages knowledge about intron/exon boundaries in the targeted regions. It is therefore suitable for application in systems lacking high-quality annotated reference genomes or transcriptomes. The main strengths of this pipeline are its modularity, which allows for an iterative refinement of read mapping, assembly, and alignment, its 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 microevolutionary patterns 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). Supergroups I and II are morphologically divergent and largely correspond to Group 1 and 2 reported by Bosser and 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 superfamilies 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 superfamilies 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 that have a strong tendency to produce racemose inflorescences. It includes a potentially new species, *Dalbergia* sp. A, which corresponds to *D. sp. 24* in Crameri (2020) and to which no provisional name has yet been assigned. Collections belonging to this entity were previously believed to be conspecific with *D. maritima* subsp. *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 3, Figure S3) differences compared to the type material (Service Forestier 32,824). 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 subhumid east and northwest of Madagascar, and a large third clade centred in the drier west and north of the island. The identification of morphological synapomorphies characterizing these clades require further genetic and morphological analyses. The geographic separation in major ecogeographic 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 of 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* (Americas 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 limited 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, enables 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* subsp. *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 presumably 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, 2022). 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 tracking geographic hotspots of the illegal trade in these valuable timber species (UNODC, 2016a).

### 4.4 | Phylogenomic analyses across legumes

At the family level, 1005 merged regions of the 6555 initially targeted regions passed our stringent sensitivity and specificity filters, suggesting that many target regions were not efficiently captured across taxa. However, phylogenomic 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, Ojeda, et al., 2020). 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, Ojeda, et al., 2020; 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, Ojeda, et al. (2020) 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, Ojeda, et al., 2020). Taxon sampling may additionally contribute to the contentious deep-branching relationships. The monotypic Duparquetioideae subfamily could not be analysed, 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, Ojeda, et al., 2020). 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-Papilionoideae. 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, Ojeda, et al., 2020). Similarly, conflicting topologies affected most branches within Genistaeae. By contrast, our analyses confirm that the genus Aeschynomene sensu Ruoo (1955), which consisted of the former A. sect. Aeschynomene and A. sect. Ochopodium Vogel, is nonmonophyletic (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 macroevolutionary 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.

### AUTHOR CONTRIBUTIONS

Simon Crameri and Alex Widmer designed the study and collected samples. Stefan Zoller assembled the draft Dalbergia transcriptome and designed target capture probes. Simon Crameri and Stefan Zoller analysed data and wrote CaptureAl. Simon Crameri, Simone Fior and Alex Widmer wrote the manuscript with contributions from Stefan Zoller.

### ACKNOWLEDGEMENTS

We thank the Missouri Botanical Garden (MBG) Madagascar team (especially Sylvie Andriambololonera, Roger Bernard, Nivo Rakotonirina, Charles Rakotovao, Richard Randrianaivo, and Richardson Razakamalala), the University of Antananarivo (especially Harisoa Ravaomanalina), Sonja Hassold and Hoa Thi Tran for organizing and conducting fieldwork, the Botanic Garden of the University of Zurich and the Fairchild Tropical Botanic Garden for the opportunity to sample legume plants from their living collection, and the herbaria of Zurich (ZT), Paris (P), Geneva (G) and MBG (MO) 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 very grateful to Claudia Michel for laboratory work and the Genetic Diversity Centre (GDC) at ETH Zurich 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. Open access funding provided by Eidgenössische Technische Hochschule Zurich.

### CONFLICT OF INTERESTS

The authors declare no conflict of interest.

### OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The raw sequencing data is available in the European Nucleotide Archive (ENA) at EMBL-EBI under accession number PRJEB41848 (https://www.ebi.ac.uk/ena/browser/view/PRJEB41848). The Dalbergia madagascariensis s.l. transcriptome, probe sequences for target capture, taxon-specific reference sequences, final alignments, and SNP data are available in Dryad (https://doi.org/10.5061/dryad.73n5tb2z7).
This article has earned an Open Materials badge for making publicly available the components of the research methodology needed to reproduce the reported procedure and analysis. The bioinformatic pipeline CaptureAI and R scripts for downstream population genomic analyses and visualizations are available and further documented on GitHub (https://github.com/scrameri/CaptureAI).

**DATA AVAILABILITY STATEMENT**

Raw target capture sequencing reads generated for this study have been deposited in the European Nucleotide Archive (ENA) at EMBL-EBI under accession number PRJEB41848 (https://www.ebi.ac.uk/ena/browser/view/PRJEB41848; Crameri et al., 2022a). Transcriptome sequencing reads as well as the draft Dalbergia transcriptome, sequences representing the initial 12,049 RNA probes and 6555 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 for population genetics and visualizations are available and further documented on GitHub (https://github.com/scrameri/CaptureAI). 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, 2022), but with restricted public access to precise geo-coordinates (delivered on demand to bona fide researchers).

**ORCID**

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

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

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

**REFERENCES**


Crameri, S., Fior, S., Zoller, S., & Widmer, A. (2022b). 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.73n5b2z7


Egan, A. N., & Vatanparast, M. (2019). Advances in legume research in

Faircloth, B. C. (2016). PHYLUCE is a software package for the analysis


**SUPPORTING INFORMATION**

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Crameri, S., Fior, S., Zoller, S., & Widmer, A. (2022). A target capture approach for phylogenomic analyses at multiple evolutionary timescales in rosewoods (*Dalbergia* spp.) and the legume family (Fabaceae). Molecular Ecology Resources, 22, 3087–3105. https://doi.org/10.1111/1755-0998.13666