

RESEARCH ARTICLE

Integrative utilization of genomic resources for improved phylogenetic resolution in Sonerileae (Melastomataceae)

Luo Chen^{1,2}  | Marie Claire Veranso-Libalah^{1,3}  | J. Peter Quakenbush⁴  |
 Ying Liu⁵  | Fabián A. Michelangeli⁶  | Diego F. Morales-Briones¹  |
 Gudrun Kadereit^{1,2} 

¹Prinzessin Therese von Bayern-Lehrstuhl für Systematik, Biodiversität & Evolution der Pflanzen, Ludwig-Maximilians-Universität München, Munich 80638, Germany

²Botanische Staatssammlung und Botanischer Garten München, Staatliche Naturwissenschaftliche Sammlungen Bayerns (SNSB), Munich 80638, Germany

³Senckenberg Research Institute and Natural History Museum, Frankfurt am Main 60325, Germany

⁴Department of Biology, Grand Valley State University, Allendale, MI 49401, USA

⁵School of Ecology, Sun Yat-sen University, Shenzhen, 518107, China

⁶Center for Biodiversity & Evolution, New York Botanical Garden, Bronx, NY 10458, USA

Correspondence

Luo Chen, Prinzessin Therese von Bayern-Lehrstuhl für Systematik, Biodiversität & Evolution der Pflanzen, Ludwig-Maximilians-Universität München, 80638 Munich, Germany.
 Email: Luo.Chen@lmu.de, chenluo0322@gmail.com

Funding information

Deutsche Forschungsgemeinschaft, Grant/Award Numbers: VE1291/1-1, KA1816/12-1, KA1816/12-2

Abstract

Premise: Advances in new next-generation sequencing (NGS) methods have revolutionized phylogenetics, yet challenges remain in effectively utilizing data from a wide range of sources. A well-resolved and broadly sampled phylogeny for Sonerileae, the second-largest tribe in Melastomataceae, is still lacking, hindering our understanding of its phylogenetic relationships and evolutionary history.

Methods: We combined target enrichment, genome skimming, RNA-Seq, and deep genome skimming (DGS) data to reconstruct the phylogenetic relationships within Sonerileae. We sampled 184 accessions representing 41 of the 47 genera in the tribe, including new sequences obtained with Hyb-Seq and DGS, and publicly available sequences. A newly developed reference comprising 5626 loci was employed to integrate data from different sources.

Results: Our analyses produced a well-supported, near-comprehensive phylogeny for the tribe, outperforming previous and current studies that used only target enrichment probe sets as the reference for loci extraction. For the long-unresolved Asian Sonerileae, our phylogeny is comparable to that of a study focused specifically on that group, where a draft genome was required to identify orthologs.

Conclusions: Our analyses demonstrate that using an inclusive reference file enables the integration of genomic data generated with different sequencing strategies. We provide an effective approach for combining heterogeneous data for phylogenomic analysis.

KEYWORDS

data integration, deep genome skimming, genome skimming, Melastomataceae, phylogenomics, RNA-Seq, Sonerileae, target enrichment

The increasing availability of next-generation sequencing (NGS) data has significantly changed the field of phylogenetics in the past 10 years (Guo et al., 2023). In response, a variety of sequencing strategies have been established for phylogenetic studies, including target enrichment, genome skimming, deep genome skimming (DGS), Hyb-Seq, and RNA-Seq. Among these, target enrichment, or target capture, is currently one of the most popular approaches for plant phylogenomics. It selects targeted genomic regions before sequencing, providing a cost-effective approach to obtain hundreds to thousands of loci

(Mandel et al., 2014). This method has proven effective across various taxonomic scales, from population-level studies (e.g., Villaverde et al., 2018; Jiménez-Mena et al., 2022) to higher level phylogenetics (e.g., Benítez-Villaseñor et al., 2023). A universal probe set targeting 353 nuclear loci across angiosperms, the Angiosperms353 kit (Johnson et al., 2018), exemplifies this approach and has been successfully used to resolve numerous flowering plant lineages, including species sampled from herbarium specimens over 200 years old (Zuntini et al., 2024).

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *American Journal of Botany* published by Wiley Periodicals LLC on behalf of Botanical Society of America.

Another commonly used sequencing strategy is genome skimming, which was originally developed as a low-coverage method to capture high-copy genomic regions (Straub et al., 2012). It has been widely applied in organellar genome sequencing due to the high copy numbers of plastid and mitochondrial genomes (Twyford and Ness, 2017; Trevisan et al., 2019). In addition to these high-copy regions, the potential to recover low-copy nuclear loci from genome skimming data was noted in the original study and has since been explored in subsequent studies (Straub et al., 2012; Blischak et al., 2014; Pezzini et al., 2023; Pouchon and Boluda, 2023).

More recently, Liu et al. (2021) demonstrated that with sufficient sequencing depth (at least 10× in their data set), numerous high-quality single-copy nuclear genes could also be assembled from genome skimming data. They termed this approach deep genome skimming, which enables the assembly of numerous single-copy nuclear genes and complete plastomes, mitochondrial genomes, and often the entire ribosomal DNA sequences without the need for target enrichment. This approach has since been applied to various lineages of organisms (e.g., Liu et al., 2021; Yang et al., 2023; Hu et al., 2024; Quattrini et al., 2024; Niu et al., 2025) and is particularly effective for groups with relatively small genomes (e.g., ~500 Mb, Liu et al., 2021; ~900 Mb, Yang et al., 2023; ~500–750 Mb, Quattrini et al., 2024), but can also be applied to relatively large genomes as long as a sufficient number of reads are generated (e.g., ~3 Gb in Hu et al., 2024; ~1.66–2.32 Gb in Niu et al., 2025).

Hyb-Seq is a sequencing strategy that combines target enrichment and genome skimming. It was initially proposed for sequencing targeted genomic regions using on-target reads and for recovering high-copy genomic regions through off-target reads from the enriched library (Weitemier et al., 2014), making it essentially interchangeable with target enrichment. The term can also refer to an approach that involves sequencing both enriched (target enrichment) and unenriched (genome skimming) libraries (Villaverde et al., 2018; Dodsworth et al., 2019) to improve recovery of high-copy regions of the genome; here, we use this definition.

RNA-Seq, or transcriptome sequencing, was originally developed for transcriptome profiling (Wang et al., 2009), but it also has great utility in phylogenetic studies (Oakley et al., 2012; Wickett et al., 2014; Leebens-Mack et al., 2019). When orthologs are accurately identified, the phylogenetic results are comparable to those obtained from whole-genome sequencing (Cheon et al., 2020). The challenge with RNA-Seq is that total RNA extraction requires fresh tissue, flash-frozen tissue stored at –80°C (Chomczynski and Sacchi, 2006), or tissue preserved in reagents such as RNALater (Yockteng et al., 2013). Although a few studies have successfully extracted RNA from silica-dried plant tissues (Chen et al., 2022; He et al., 2022; Ruiz-Vargas et al., 2024), this approach has not been widely tested across different groups and remains impractical for herbarium specimens.

Despite the potential to significantly improve data (re-)utilization and the general importance of integrative data use, no one has yet explored the compatibility of these different

types of genomic data when combined into a single phylogenomic inference or optimizing the strategy. A common strategy is to recover loci targeted for enrichment from RNA-Seq and DGS data (Bossert et al., 2019; Wang et al., 2024). However, this strategy uses only the subset of reads corresponding to targeted loci, leaving much of the remaining information in the RNA-Seq and DGS data underutilized. As more data are generated by different research groups using various sequencing strategies, it becomes increasingly necessary to identify an effective approach to integrating all NGS data sets available for a given group of organisms while maximizing information in the sequencing data.

Large, species-rich plant groups that have been studied for a long time by multiple research groups provide good examples of the challenges of data integration because the available sequence data sets are often heterogeneous. The Melastomataceae are one of the 10 largest families of flowering plants, comprising 5858 species across 173 genera in three subfamilies, one of which is further divided into 21 tribes (Ulloa Ulloa et al., 2022). Sonerileae (Figure 1) is the second-largest tribe within Melastomataceae, containing 47 genera and around 1080 species distributed throughout the tropics (Lin et al., 2022; Liu et al., 2022; Veranso-Libalah et al., 2023; Liu et al., 2024; Quakenbush et al., 2025; Zhuang et al., 2025). The generic delimitations within Sonerileae have long been unresolved because many of the morphological traits used are homoplasious and sampling for molecular phylogenetic studies has been limited (Zhou et al., 2019a, 2019b). A recent phylogenomic study (Zhou et al., 2022) on the Asian Sonerileae, which included samples of most recognized Asian genera and reconstructed their phylogeny using DGS data, identified 34 lineages within the Asian Sonerileae. Based on the results of this study, the genus *Perilimnastes* was reinstated (Liu et al., 2024), and *Neophoanthus* was newly described (Lin et al., 2022). The authors further expanded sampling for the *Cyphotheca-Plagiopetalum-Sporoxeia* clade and described three new genera, i.e., *Chiehchenii*, *Neophyllagathis*, and *Sporocyphoxeia*, based on phylogenetic and morphological evidence (Zhuang et al., 2025). In another phylogenomic study (Quakenbush et al., 2025) that focused on the systematics of fleshy-fruited Sonerileae using mostly target enrichment data, the fleshy-fruited Sonerileae were found to be nonmonophyletic. Within this group, *Pachycentria* and *Plethiandra* were recovered as nested within *Medinilla* and were consequently synonymized to maintain the monophyly of *Medinilla*. However, African and neotropical lineages remain underrepresented in both studies. A separate phylogenomic study (Maurin et al., 2021) using Angiosperms353 on the Myrtales included some African and neotropical Sonerilean genera, but the sampling was limited because the study focused on familial relationships. While the Kew Tree of Life Explorer (Baker et al., 2021) is gradually incorporating more species, the sampling for Sonerileae remains limited. Consequently, our understanding of the relationships among African and neotropical Sonerileae is still primarily based on a Sanger sequence-based phylogeny (Veranso-Libalah et al., 2023), in

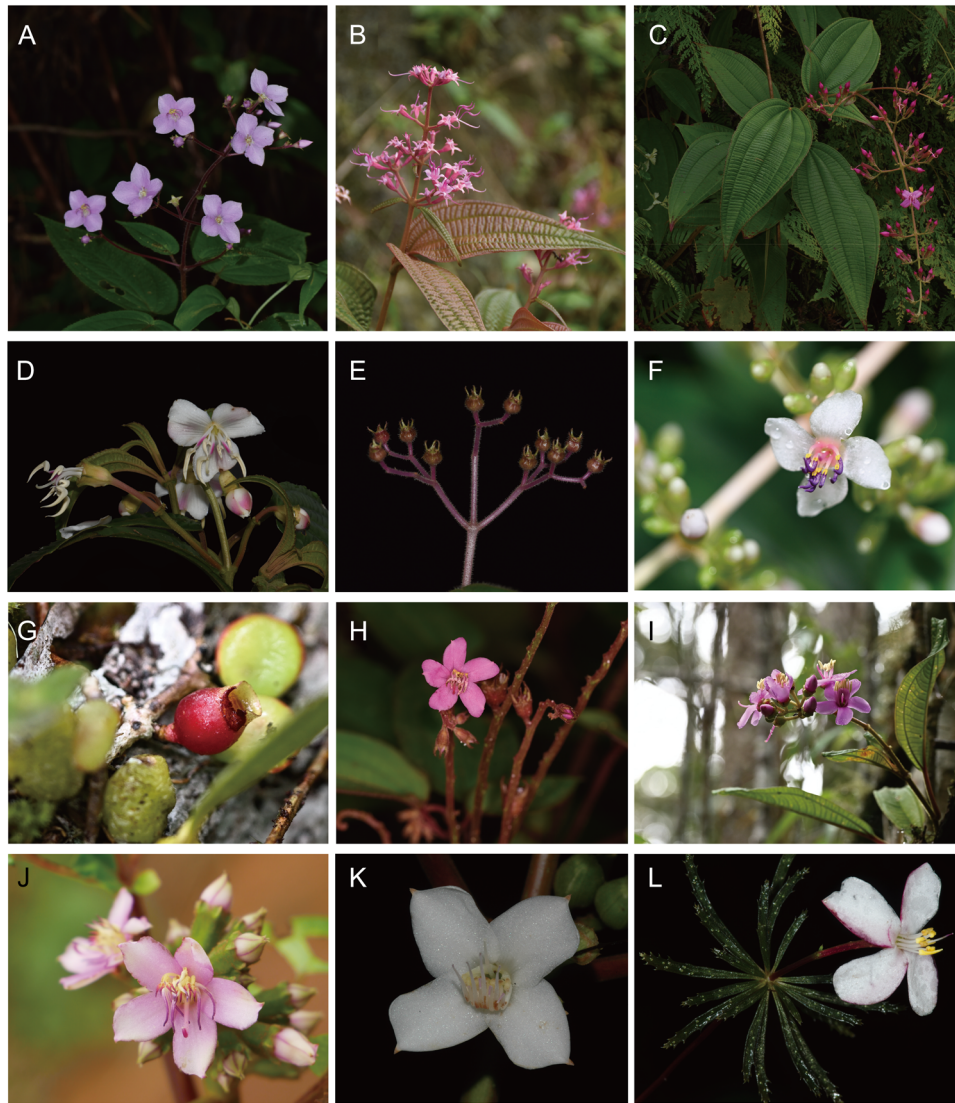


FIGURE 1 Representatives of the major clades of Sonerileae. (A) *Bredia hirsuta*, (B) *Blastus cavaleriei*, (C) *Oxyspora paniculata*, (D) *Cyphotheca montana*, (E) *Scorpiothyrus shangszeensis*, (F) *Medinilla pendula*, (G) *Medinilla sedifolia*, (H) *Mendelia mollis*, (I) *Gravesia laxiflora*, (J) *Dicellandra barteri*, (K) *Opisthocentra clidemioides*, (L) *Tryssophyton merumense*. Photo credits: A–E, Ying Liu; F–I, Luo Chen; J, Andreas Gröger; K, Mario Terra; L, John L. Clark.

which neotropical genera and the African endemic *Benna* were identified as successive sister lineages to the remaining Sonerileae, which is divided into an African-Malagasy endemic clade and a predominantly Asian clade. Based on this study, the genus *Mendelia* was described, and *Bourdaria* was reinstated.

A well-represented and well-supported phylogeny is crucial for understanding the relationships among the African, Asian, and neotropical Sonerileae and the evolutionary history of the tribe, including its biogeography and trait evolution. A large amount of sequence data for Sonerileae has already been published. Therefore, the aim of our current study is to mobilize and reuse these published data, extract loci, and integrate them into a new analysis, adding previously missing or underrepresented taxa. To reach this aim, we developed a new reference comprising 5626 loci for Sonerileae based on genomic and transcriptomic data. Using

this integrative reference, we reconstructed a near-comprehensive phylogeny for the tribe by combining previously and newly generated Hyb-Seq and DGS data with all available NGS data, including RNA-Seq, DGS, target enrichment from different probe sets, Hyb-Seq, and annotated genomes. A primary goal of our study was to assess how far this approach advances our understanding of the systematics and evolution of Sonerileae.

MATERIALS AND METHODS

Sampling and sequencing

We included 184 accessions, 179 of which belong to Sonerileae, representing 41 of the 47 genera, which include

four of the six neotropical genera, all nine afrotropical endemic genera, 27 of the 31 Asian/Oceanian endemic genera, and the only afrotropical-Asian disjunct genus within the tribe, *Medinilla*, for which we sampled all alliances recognized by Quakenbush et al. (2025). All 34 Asian lineages identified by Zhou et al. (2022), supported by both the single-copy ortholog (SCO) and genomic SNP data sets in that study, were included. Additionally, we selected four *Melastoma* and one *Osbeckia* species from the tribe Melastomateae as outgroups. To obtain a near-complete genus-wide phylogeny, we incorporated a variety of data types: Among the 184 accessions, 73 were newly sequenced, while sequence data for the remaining 111 accessions were obtained from public resources: RNA-Seq data for 15 accessions from the NCBI Sequence Read Archive (SRA) and China National GeneBank (CNGb), annotated genomes for two accessions from the National Genomics Data Center (NGDC), Angiosperms353 target enrichment data for 16 accessions from the European Nucleotide Archive (ENA), and DGS data for 78 accessions from the SRA. Details on the vouchers and accession numbers are provided in Appendix S1.

Of the 73 newly sequenced accessions, 70 were sequenced using Hyb-Seq and the remaining three using DGS. For the 70 accessions sequenced with Hyb-Seq, target enrichment data for 55 had previously been published (Quakenbush et al., 2025), to which genome skimming data were now added. In addition, complete Hyb-Seq data were newly generated for 15 accessions. DNA extraction, library preparation, target enrichment, and next-generation sequencing were performed as described by Quakenbush et al. (2025). The Melastomataceae probes (Jantzen et al., 2020) were used to capture 384 loci during target enrichment, while ~1 Gb of data were generated during genome skimming. For the DGS samples, DNA was extracted following Quakenbush et al. (2025). Library preparation and sequencing for these samples were performed at Novogene Europe (Cambridge, UK). Paired-end reads were generated on the Illumina NovaSeq X Plus system (San Diego, CA, USA), with a requested data yield of 6 GB per sample.

Nuclear reference construction

To effectively integrate the various data in this study, we developed a new nuclear reference sequence file using 16 RNA-Seq samples and the annotated genome for *Melastoma candidum* (Zhong et al., 2023) and *Barthea barthei* (Huang et al., 2024). These 18 samples included five species from Melastomateae, 12 species from Sonerileae, and one species from Miconieae. The pipeline described here (Figure 2) largely follows that of Yang and Smith (2014) and Morales-Briones et al. (2021) (https://bitbucket.org/yanglab/phylogenomic_dataset_construction). Raw reads of the RNA-Seq data were processed to correct random sequencing errors using Rcorrector v1.0.6 (Song and Florea, 2015), remove adapters and low-quality sequences with Trimmomatic v0.39 (Bolger et al., 2014), and filter organelle reads using Bowtie2 v2.4.5 (Langmead and Salzberg, 2012). De novo transcriptome

assembly was performed using Trinity v2.13.2 (Grabherr et al., 2011), then quality was assessed and low-quality transcripts were filtered out using Transrate v1.0.3 (Smith-Unna et al., 2016), and chimeric transcripts were removed using a blast-based method (Yang and Smith, 2013). Filtered transcripts were clustered into putative genes using Corset v1.09 (Davidson and Oshlack, 2014) with the longest transcript kept. Coding sequences (CDS) were identified using TransDecoder v5.3.0 (Haas et al., 2013). Before homology inference, transcript redundancy was further reduced using CD-HIT v4.8.1 (Fu et al., 2012). Homology was inferred by first performing all-by-all BLASTN v2.10.0+ (Camacho et al., 2009) on the 16 transcriptomes and two annotated genomes and then using MCL (van Dongen and Abreu-Goodger, 2012) to group transcripts into clusters, with the hit fraction cutoff set to 0.4 and inflation values to 1.4. The resulting 12,497 clusters were aligned using MAFFT v7.525 with the algorithm “localpair” (Katoh and Standley, 2013). The aligned clusters were trimmed using Phyutility v2.7.1 (Smith and Dunn, 2008) to discard columns with more than 90% missing data. Single homologous gene trees were inferred using IQ-Tree v2.3.5 (Minh et al., 2020), with ModelFinder (Kalyaanamoorthy et al., 2017) employed to select the best substitution model. The raw single trees were trimmed by masking monophyletic and paraphyletic tips that belong to the same species, retaining the tip with the most unambiguous characters in the trimmed alignment. Spurious long branches were then pruned with TreeShrink v1.3.9 (Mai and Mirarab, 2018). The MO (monophyletic outgroup) method (Yang and Smith, 2014) was employed to infer orthology. Gene trees without taxon duplication were retained as one-to-one orthologs if they had monophyletic outgroups. For trees with taxon duplication, only those with monophyletic, nonrepetitive outgroups were rerooted, and the orthologous subtree with the highest taxon number was retained. The resulting orthologs with fewer than 12 taxa were discarded. We also filtered gene trees if the Sonerileae or Melastomateae taxa did not form a clade. The corresponding fasta files of the final 5639 gene trees were written, combined, and checked with Hybpiper (Johnson et al., 2016) command `check_targetfile` and processed with command `fix_targetfile` to correct reading frames and filter sequences with low-complexity regions, resulting in 5626 retained loci. To assess the overlap among the newly developed 5626-locus reference, the Melastomataceae probe set, and the Angiosperms353 probe set (specifically a tailored mega353 file for Myrtales, see McLay et al. (2021) and Chen et al. (2023)), we built a local BLAST database from the new reference using the application `makeblastdb` (Camacho et al., 2009) and searched the sequences against the other two references using the application `blastn` with default settings.

Reads processing, assembly, and reference loci recovery

Reads processing, assembly, and reference loci recovery were conducted using Captus v1.2.0 (Ortiz et al., 2023 [preprint]).

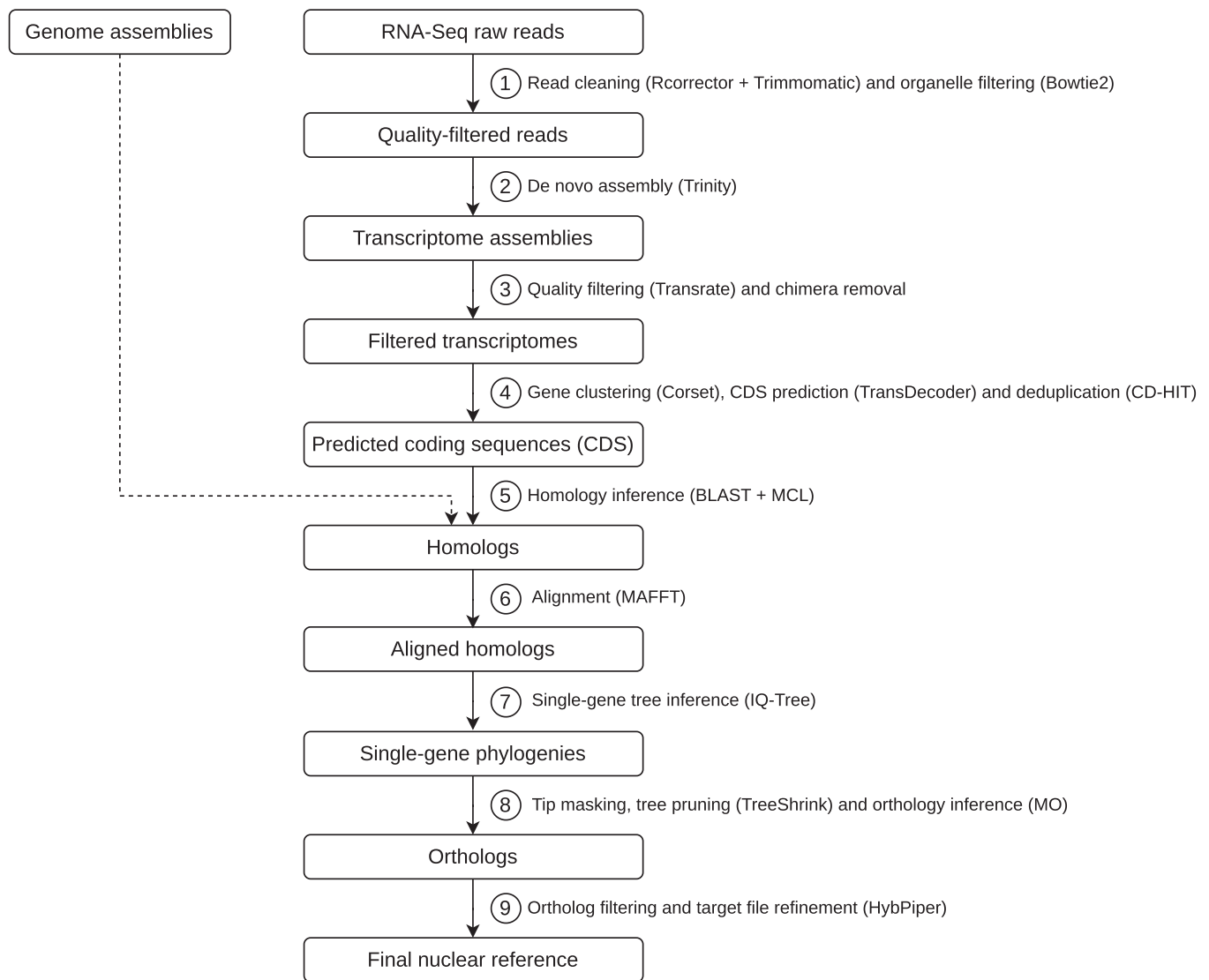


FIGURE 2 Workflow for constructing the nuclear reference used in this study.

Since the sequences from the RNA-Seq samples and annotated genomes used to construct the new reference were already available, these data were excluded from this step. Target enrichment and skimming data generated by Hyb-Seq for this study were first merged. Raw reads from the DGS, Angiosperms353 target enrichment, and Hyb-Seq data were trimmed using BBDuk from BBTools (Bushnell, 2022). For the Angiosperms353 data, trimming followed the default settings in Captus, whereas the DGS and Hyb-Seq data were trimmed with more stringent parameters (“--trimq 20 --maq 20”). The cleaned reads were then de novo assembled into contigs using MEGAHIT (Li et al., 2015) with default parameters. Reference loci were recovered from the assemblies using Scipio (Hatje et al., 2011) with minimum percentage identity (option --nuc_min_identity) set to 80 and coverage (option --nuc_min_coverage) to 50. The newly constructed reference was used to recover loci. The percentage of recovered sequence length relative to the reference

sequence was calculated in Captus and visualized using a custom Python script (ChatGPT [OpenAI, 2026] was used to assist with coding for this script. The generated code was critically reviewed, tested, and revised by the authors before use.). To maximize data use, loci in the Melastomataceae probe set but absent from the new reference were also used to recover loci.

To compare the performance of the new reference and the Melastomataceae probe set, we recovered the 396 loci of the Melastomataceae probe set (Jantzen et al., 2020; Dagallier and Michelangeli, 2024) across all 184 assemblies, including the RNA-Seq sequences.

Orthology inference

The recovered reference loci were retrieved using Captus for further analysis. For the new reference data set,

reference sequences from Sonerileae and Melastomateae were added to each cluster. This step was not necessary for the 396-locus data set. Each cluster was aligned and trimmed using MAFFT and Phyutility, following the same procedures as in the construction of the nuclear reference. Sequences shorter than 100 bp or with more than 90% missing sites were discarded. Single gene trees were inferred using IQ-Tree with the GTRGAMMA substitution model, and branch support was assessed using 1000 ultrafast bootstrap replicates. Monophyletic and paraphyletic tips were pruned, spurious long branches were trimmed, and orthology inference using the MO method followed the same approach as in nuclear reference construction. The five sampled Melastomateae species were set as outgroups, and only trees with at least 19 tips were retained. We calculated the occupancy of each taxon in the final orthologs for both the new reference and the 396-locus data set.

Phylogenetic analyses

The orthologs were realigned, trimmed, and gene trees were inferred using the same methodology as in the orthology inference. Species trees were then reconstructed using ASTRAL-IV v1.19.4.5 (Zhang et al., 2018; Zhang and Mirarab, 2022; Tabatabaee et al., 2023). The tree based on the new reference is referred to as ST1, and the tree based on the 396-locus data set is referred to as ST2. We also inferred a species tree (ST3) for the 16 RNA-Seq samples and two annotated genomes used for reference construction, based on the 5639 orthologs. The phylogenetic results were visualized using ggtree 3.10.0 (Yu et al., 2017) in R version 4.3.2 (R Core Team, 2023).

Assessing gene tree conflicts

The quartet-based method, quartet sampling (QS) (Pease et al., 2018), was used to assess the level of discordance among gene trees. For a given phylogeny, any internal branch can divide the tree into four sets of taxa. Quartet sampling repeatedly and randomly selects one taxon from each set and evaluates the likelihoods of all three possible topologies: one that is concordant and two that are discordant with the given phylogeny, based on the given sequence data. The species tree ST1 and an alignment generated by concatenating gene alignments used for gene tree inference with AMAS (Borowiec, 2016) were used as input. For each internal branch, 6000 replicates were performed, and the option --genetrees was used to enable partitioned analysis. Likelihood evaluations were carried out using RaxML (Stamatakis, 2014). Quartet concordance (QC), quartet differential (QD), and quartet informativeness (QI) scores were generated. The QS results were visualized using an R script developed by Liu et al. (2025).

RESULTS

New reference and sequence recovery

The newly developed reference comprises 5626 loci represented by a total of 76,834 sequences. The length of each locus ranges from 300 to 11,127 bp, with an average of 1419 bp and a median of 1275 bp. According to the BLAST results, 176 of these loci are putatively shared with the Angiosperms353 loci, and 221 are putatively the same as those in the Melastomataceae probe set. For details of the matching loci, see Appendix S2.

Of the 5626 target loci, 52–5625 loci were recovered per sample with Captus (Figure 3; Appendix S3). Sequence recovery differed among data sets: Angiosperms353 samples recovered 52 to 2353 loci (average 1082, median 1180), Hyb-seq samples recovered 74 to 5622 loci (average 5024, median 5578), and DGS samples recovered 3883 to 5625 loci (average 5586, median 5619).

Orthology inference

Both the newly developed reference and the Melastomataceae probe set were used for phylogenetic analysis. We obtained 5710 orthologs from the new reference data set, including 106 orthologs derived from the 175 non-overlapping loci from the Melastomataceae probe set, while 225 orthologs were retrieved from the 396-locus data set. We calculated the frequency of taxa in the final sets of orthologs. The representation across different sequencing strategies for the new reference data set was as follows (Appendix S4; Figure 4): Annotated genome samples had 2774 to 5459 loci represented (average 4117); RNA-Seq samples had 1972 to 5214 loci (average 3024, median 2456); Hyb-Seq samples had 62 to 4287 loci (average 2775, median 3241); Angiosperms353 samples had 10 to 1336 loci (average 483, median 359); and DGS samples had 1499 to 4703 loci (average 3738, median 3994). The representation across different sequencing strategies for the 396-locus data set is listed in Appendix S5 (see also Appendix S6).

Phylogenetic analyses

The species tree ST1 inferred from the new reference data set is well resolved with most nodes fully supported (LPP = 1) (Figure 5). The neotropical taxa (*Opisthocentra*, *Boyania colombiana*, *Boyania kenwurdackii* + *Phainantha*, *Tryssophyton*) are successive sisters to the rest of the tribe, followed by the African endemic *Benna*, and then *Kendrickia* (an Asian genus) plus the afrotrropical endemic clade. After these, the tree revealed a primarily Asian clade, with only one dispersal back to the Afro-tropics within the diverse genus *Medinilla*. Within the afrotrropical clade, *Bourdaria* is sister to all other genera, followed by a grade composed of *Gravesia*, *Dicellandra*,

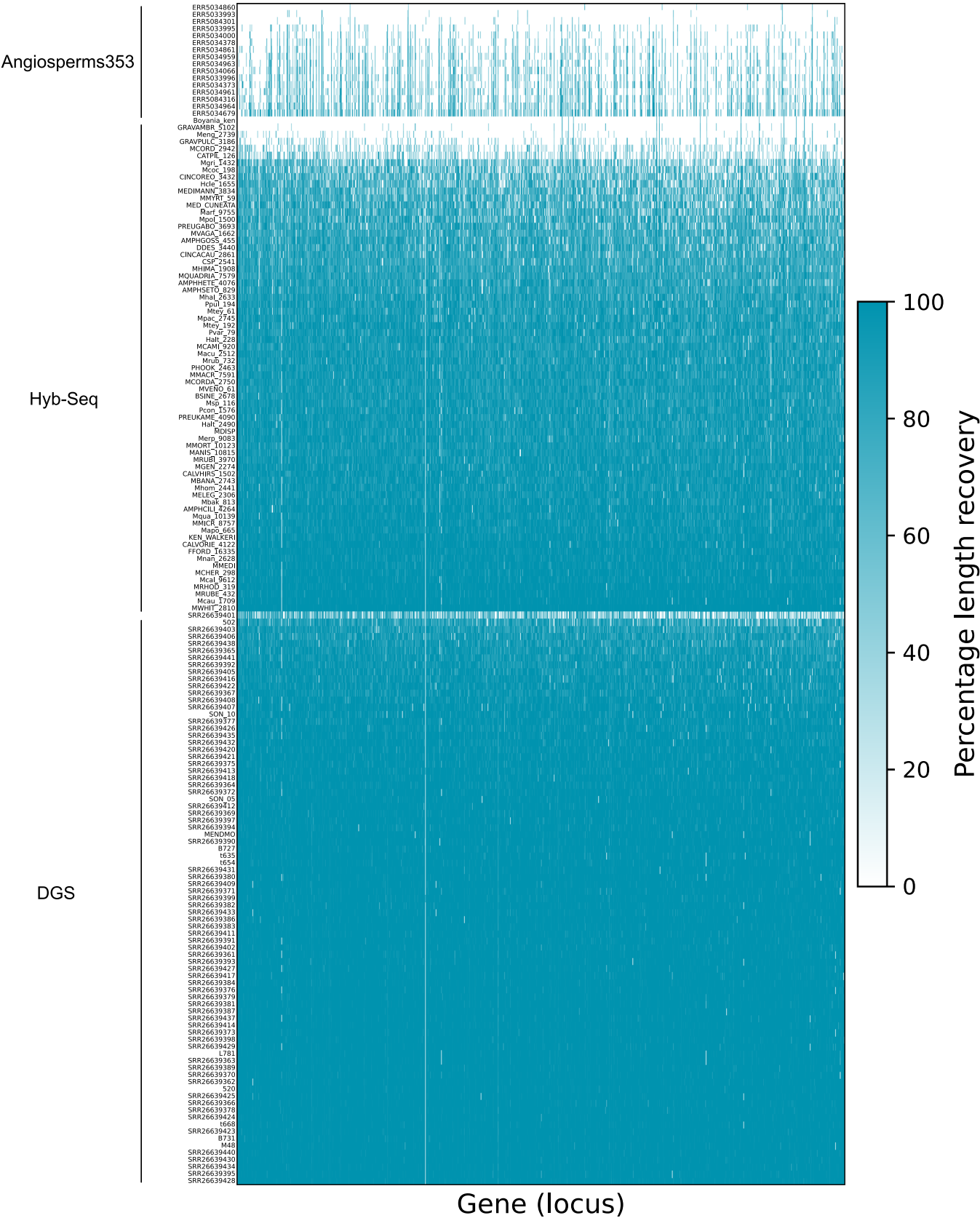


FIGURE 3 Heat map of percentage length recovery of new reference loci across samples sequenced with Angiosperms353 probe set, Hyb-Seq, and deep genome skimming (DGS).

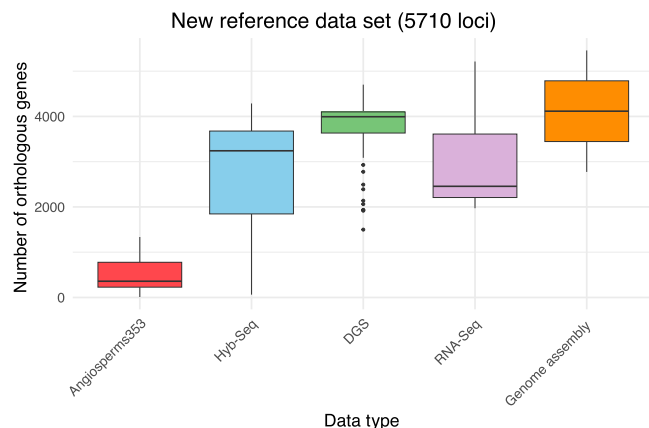


FIGURE 4 Distribution of orthologs per taxon by different sequencing strategies in the final 5710 orthologs from the new reference data set.

Calvoa, *Mendelia*, and *Preussiella*. *Amphiblemma* is sister to *Cinnabotrys* (Figure 5).

Within the primarily Asian clade, all 34 sampled lineages and nine clades (clades A–I) identified by Zhou et al. (2022) were recovered. The sampled taxa formed two major clades. The first corresponds to clade C, which includes *Medinilla*, *Tigridiopalma*, and *Nephoanthus* (formerly the *Phyllagathis prostrata* clade). The second major clade contains the remaining genera, starting with a lineage largely corresponding to clade A, now incorporating *Catanthera*. Clade A includes *Heteroblemma*, *Catanthera*, the *Phyllagathis* (thyrsoid) clade, the *Phyllagathis* (pentamerous) clade, and the *Driessenia* (isomorphic) clade. The next branching clade is clade B, comprising *Tashiroea*, *Poillannammia*, *Scorpiothyrsus*, and *Driessenia axantha*. Next is *Sonerila*, then clade D, which contains *Sarcopyramis* and *Barthea*. The next clade includes clade E (comprising *Anerincleistus*, *Phyllagathis scortechinii*, *Phyllagathis griffithii*, *Oxyspora bullata*, and *Phyllagathis tuberculata*-*P. hispida*), *Phyllagathis rotundifolia*, *Oxyspora paniculata*, and the *Oxyspora* (Bornean) clade. This clade is followed by clade F (which includes *Aschistanthera* and *Styrophyton*), *Perilimnastes* (formerly the *Phyllagathis* [raphides] clade), then clade G (consisting of *Sporoxeia*, *Sporocyphoxeia*, *Neophyllagathis*, *Cyphotheca*, *Chiehchenii*, and *Plagiopetalum*), then *Fordiophyton*, and finally clade I (which includes *Bredia*, *Blastus*, and clade H, containing *Kerriothyrsus*, *Phyllagathis erecta*, and *Fordiophyton breviscapum*).

The species tree ST2 (Appendix S7) inferred from the 396-locus data set receives less support ($LPP < 0.9$) for many nodes but is topologically similar to ST1, with some conflicting relationships. The first branching clade is also *Opisthocentra*, followed by *Boyania colombiana*, *Trysophyton*, *Boyania kenwurdackii*, and *Phainantha*, then *Benna*. Within the afrotropical clade, relationships are mostly identical to ST1, except one species of *Gravesia* is placed elsewhere, nested within the Asian clade A. Within the primarily Asian clade, *Kendrickia* and *Sonerila* are successively sister to the rest, and clade A is recovered as sister to

clade C with low support ($LPP = 0.54$). In contrast to ST1, clade D and *Styrophyton* are not resolved as monophyletic, with *Oxyspora* sp. ERR5084301 sequenced from PAFTOL failing to group with the other *Styrophyton* species. Additionally, some branches of ST2 are notably long, all of which are in the primarily Asian clade.

The species tree ST3 (Appendix S8), inferred using transcriptomic and genomic data, is topologically consistent with ST1.

Gene tree conflicts

The QS analysis revealed abundant conflicts among gene trees, with branches having low local posterior probabilities, often with high discordance (Appendix S9). In the backbone, notable areas of low or counter-support and high discordance include the Asian clade D ($LPP = 0.79$, $QC/QD/QI = -0.014/0.78/0.2$), and the clade comprising Asian clade E + *Phyllagathis rotundifolia* + *Oxyspora paniculata* + *Oxyspora* (Bornean) clade ($LPP = 0.22$, $QC/QD/QI = 0.007/0.96/0.25$). Within the identified major clades, many branches have high LPP but are weakly supported ($QC < 0.2$), particularly within the Asian clade.

DISCUSSION

An effective approach to integrating sequencing data from multiple sources

Efforts to integrate data from various sources, including target enrichment data, have primarily focused on recovering targeted loci from other data sets (Bossert et al., 2019; Wang et al., 2024). In this study, we compared using the Melastomataceae probe set alone for locus recovery with an approach that additionally used new reference sequences designed to capture all available data. The phylogenetic tree inferred from the Melastomataceae probe set loci was overall less well-supported compared to the phylogenetic tree based on the data set acquired by using the new reference loci (Figure 5; Appendix S7). Several well-established clades were not resolved as monophyletic in the former tree but were monophyletic in the latter.

Importantly, mapping data types onto the phylogeny showed that taxa sequenced with different strategies (target enrichment with Angiosperms353 probe set, Hyb-Seq, DGS, RNA-Seq, and annotated genomes) were interspersed across the tree rather than forming clades according to data type (Figure 5). This pattern supports the robustness of our integrative framework and suggests that heterogeneous genomic data can be reliably combined when appropriate reference construction and orthology inference are applied.

It has been suggested that increasing the number of gene trees (as has been done here) can improve the accuracy of species tree inference (Zhang et al., 2018), particularly under conditions of high incomplete lineage sorting (ILS)

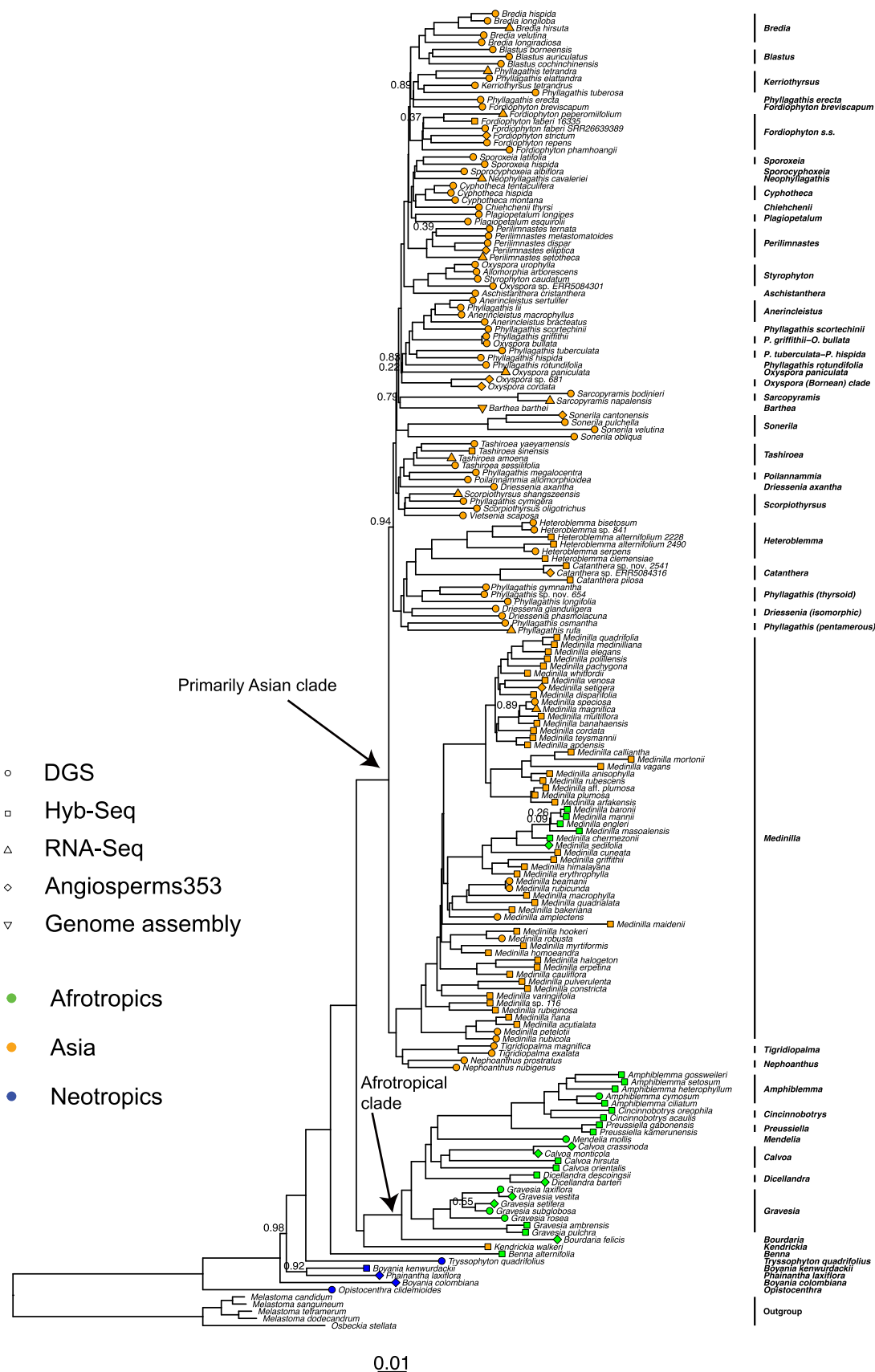


FIGURE 5 ASTRAL species tree ST1 inferred from 184 taxa and 5710 orthologs obtained from the new reference data set, showing the relationships within Sonerileae. Local posterior probabilities (LPP) are shown above the branches where LPP < 1.

(Molloy and Warnow, 2017). It has also been mathematically demonstrated that the number of true gene trees required by ASTRAL to accurately reconstruct the species tree is proportional to the inverse of the shortest branch length (Shekhar et al., 2018). Given the high level of ILS observed in the Asian Sonerileae in the previous study (Zhou et al., 2022), it is likely that the phylogenetic informativeness of the Melastomataceae probe set is insufficient for certain parts of the tree. This limitation may be further exacerbated by the low on-target rate of this probe set (Amarasinghe et al., 2021; Jantzen et al., 2022; Quakenbush et al., 2025), making the integrative approach employed here even more valuable.

The phylogenetic relationships within the Asian Sonerileae have long been unresolved, even when phylogenomic data were used (Baker et al., 2021; Maurin et al., 2021). In our study, most nodes of the Asian Sonerileae were resolved, consistent with the results of Zhou et al. (2022), in which 332 orthologs were obtained by sequencing a genome and several transcriptomes, followed by retrieving orthologs from genome resequencing data (DGS data) through read mapping. Our comparable results demonstrated that this more inclusive approach—which does not require genome sequencing and integrates diverse sequencing data sets, including target enrichment data—can produce accurate phylogenies while leveraging previously generated data rather than repeating the sequencing effort.

By including a large number of loci in the reference file, we increased the likelihood of maximizing the use of sequenced reads data. Nearly half of the Angiosperms353 loci (176 of 353) and more than half of the Melastomataceae probe set loci (221 out of 396) were found to match these orthologs (Appendix S2). The recovery was further improved by including the loci unique to the Melastomataceae probe set (i.e., those not overlapping with the new reference) in the phylogenetic analysis. For instance, *Gravesia setifera*, which was sequenced with the Angiosperms353 probe set, had not previously been grouped with other sampled *Gravesia* species (Baker et al., 2021; Maurin et al., 2021). In this study, represented by only two loci, *Gravesia* was placed within the primarily Asian clade in the phylogeny based on 396 loci; however, with 21 loci representing it, all sampled *Gravesia* species formed a monophyletic clade in the phylogeny using the new reference.

Notably, seven of 16 taxa sequenced using Angiosperms353 had more than 400 loci represented in the orthologs (Appendix S4), surpassing the number of targeted loci, suggesting that off-target reads can be effectively utilized (Allio et al., 2020; Costa et al., 2021; Reichelt et al., 2021; Lagou et al., 2024). Additionally, we demonstrated that Hyb-Seq and DGS can recover thousands of loci, with DGS performing slightly better (Hyb-Seq: average 2774, median 3241; DGS: average 3738, median 3994). The sizes of the sequenced Melastomataceae genomes range from 246.4 to 299.8 Mb (Hao et al., 2022; Zhong et al., 2023; Huang et al., 2024), making it easy to achieve high coverage with genome sequencing, which may explain the effectiveness of Hyb-Seq

and DGS in this study. Since DGS does not require target enrichment, it also reduces the time and costs associated with designing and developing probe sets, extensive lab work, and chemicals. The method employed in this study offers a promising approach for integrating data generated from different sequencing strategies.

Despite its advantages, the integrative approach presented here also has several potential limitations. First, although this approach performed well for Melastomataceae, its applicability to lineages with much larger genomes remains to be tested. Secondly, integrating heterogeneous data sets inevitably results in uneven locus recovery and missing data across taxa, which may affect resolution or support at some nodes. Finally, the performance of the approach depends on the taxonomic coverage of the reference file, which requires access to transcriptomic/genomic resources. Although large-scale genome sequencing projects such as 10KP (Cheng et al., 2018) and the Darwin Tree of Life (The Darwin Tree of Life Project Consortium, 2022) are rapidly increasing available resources, generating comprehensive coverage across many plant groups will take time.

Phylogenetic relationships within Sonerileae

By integrating various types of sequencing data, we generated a well-supported, near-comprehensive, genus-wide phylogeny for Sonerileae (Figure 5). This phylogeny is largely consistent with previous studies regarding the relationships among the African, Asian, and Neotropical clades (Baker et al., 2021; Zhou et al., 2022; Veranso-Libalah et al., 2023) but provides greater resolution and improved sampling within those clades. The neotropical lineages and the African endemic *Benna* are successively sister to the rest of the tribe, which is further divided into a predominantly afrotrropical clade and a primarily Asian clade.

The five sampled neotropical species from four genera appear as the first diverging lineages in the tribe. This pattern is similar to the results of Baker et al. (2021) and Veranso-Libalah et al. (2023), although the exact branching order of these lineages differs. *Boyania* is recovered as non-monophyletic, with *B. colombiana* sister to all Sonerileae except *Opisthocentra*, and *B. kenwurdackii* sister to *Phainantha laxiflora*. *Boyania* comprises three species and had already been shown to be nonmonophyletic based on Sanger data (Bacci et al., 2019). *Phainantha* has five recognized species, of which only two have been sequenced (one here), and they were resolved as sisters based on Sanger sequencing (Bacci et al., 2019; Wurdack and Michelangeli, 2019). *Boyania kenwurdackii* was represented by only 62 sequences and *P. laxiflora* by only 362. Incorporating more well-sequenced neotropical taxa, including the unsampled monotypic genera *Neblinanthra* and *Tateanthus*, may help clarify the relationships of these early-branching clades and assess whether gene flow contributed to the weak support.

The placement of the Sri Lankan endemic *Kendrickia* as sister to the afrotrropical clade, first identified by

Quakenbush et al. (2025), with *Kendrickia* represented by 255 loci, has been confirmed by our results with 3223 loci representing it. The internal relationships across different genera within the afrotropical clade are largely consistent with the Sanger-sequence-based phylogeny (Veranso-Libalah et al., 2023), with the exceptions of the *Amphiblemma-Cincinnobotrys-Preussiella* clade and the placement of *Bourdaria*. In our phylogeny, *Amphiblemma* is sister to *Cincinnobotrys*, and *Bourdaria* is sister to the rest of the afrotropical clade, whereas in the Sanger phylogeny, *Amphiblemma* is sister to *Preussiella*, and *Gravesia*, instead of *Bourdaria*, is recovered as sister to the rest of the Afrotropical clade. Our results support the taxonomic treatments by Veranso-Libalah et al. (2023), which separate *Mendelia* from *Amphiblemma* and *Bourdaria* from *Cincinnobotrys*.

Zhou et al. (2022) identified 34 lineages (potentially representing genera after taxonomic revision) and nine major clades within the remaining Asian Sonerileae. In this study, we sampled all of these lineages and also included new sequences for many fleshy-fruited taxa. All 34 sampled lineages and the nine major clades were successfully recovered, corroborating the findings of Zhou et al. (2022). However, there are four discordant placements between our species tree (Figure 5) and the Astral coalescent tree from Zhou et al. (2022). These discrepancies involve the backbone of the Asian Sonerileae, the placement of *Oxyspora paniculata* and *Driessenia axantha*, and the placement and composition of clade F. In our phylogeny, clade C is the first branching clade of the Asian Sonerileae, followed by clade A and clade B, then the rest of the Asian Sonerileae. By contrast, Zhou et al. (2022) placed clade A as the first branching clade, followed by a clade comprising clade B and C, and then the rest of Asian Sonerileae. The very short branches along the backbone suggest a history of rapid radiation. Our results are consistent with the species tree based on transcriptomic and genomic data (Appendix S8). In our phylogeny, *Oxyspora paniculata* is sister to a clade comprising clade E and *Phyllagathis rotundifolia*, whereas Zhou et al. (2022) placed it as sister to clade F. The LPP support is weak, and the QS analysis shows counter-support in both phylogenies (LPP = 0.83, QC/QD/QI = -0.016/0.77/0.26; LPP = 0.65, QC/QD/QI = -0.26/0.28/0.81). In our phylogeny, *Driessenia axantha* is nested within clade B, while in Zhou et al. (2022), it is placed as sister to clade B. In our analysis, the placement was well supported based on LPP but weakly supported based on the QC score (LPP = 1, QC/QD/QI = 0.015/0.95/0.25), whereas it was fully supported (LPP = 1, QC/QD/QI = 1/-/1) in the phylogeny of Zhou et al. (2022). Furthermore, clade F of Zhou et al. (2022) comprises *Aschistanthera*, *Styrophyton*, and the *Oxyspora* Bornean clade. In contrast, our phylogeny does not place the Bornean *Oxyspora* clade within clade F; instead, it is recovered as sister to a clade comprising Asian clade E, *Phyllagathis rotundifolia*, and *Oxyspora paniculata*. This relationship, however, is weakly supported (LPP = 0.22; QC/QD/QI = 0.007/0.96/0.25).

The placements of the fleshy-fruited taxa are consistent with those of Quakenbush et al. (2025). *Kendrickia* is sister to the afrotropical clade, and *Heteroblemma* and *Catanthera* form a clade that is sister to the *Phyllagathis* (thyrsoid) clade. The former genera *Pachycentria* and *Plethiandra* are nested within *Medinilla*. Within the highly diverse genus *Medinilla*, all 15 clades identified by Quakenbush et al. (2025) were recovered, with most internal relationships aligning closely. The exception is *Medinilla maidenii*, which in our phylogeny forms a clade with the *Medinilla* Western + Eastern Superclade with full support (LPP = 1), whereas Quakenbush et al. (2025) found it formed a clade with the *Medinilla* Western Superclade, albeit with low support (LPP = 0.65). Our QS results for the branch (QC/QD/QI = 0.00088/0.95/0.35) indicate rapid radiation, rather than gene flow. This case serves as an example of improving the accuracy of species tree inference by increasing the number of genes in the context of ILS.

CONCLUSIONS

In this study, we demonstrated the potential of integrating various types of NGS data for phylogenetic analysis by constructing an inclusive reference from publicly available genomic and transcriptomic data. Applying this approach to Sonerileae resulted in a well-represented and well-supported phylogeny. For Sonerileae, this phylogenetic framework provides a basis for future systematic work, including improving sampling for unsampled and undersampled genera (e.g., the neotropical lineages), investigating the causes of gene tree conflicts, and integrating molecular and morphological evidence. Such efforts will be important for resolving the remaining taxonomic and phylogenetic uncertainties and for understanding trait evolution and biogeographic history of the tribe. This approach shows promise for unifying heterogeneous NGS data generated through different sequencing strategies and can be broadly applied to other groups facing similar challenges.

AUTHOR CONTRIBUTIONS

L.C. conceptualized and designed the study, performed lab work and data analysis, and drafted the manuscript. M.C.V.L. conducted fieldwork, acquired funding, and provided supervision. J.P.Q. conducted fieldwork and lab work. Y.L. contributed key data. F.A.M. conducted fieldwork and provided supervision. D.F.M.B. assisted with data analysis. G.K. acquired funding and provided supervision. All authors contributed to revising the manuscript.

ACKNOWLEDGMENTS

We thank BR, BRLU, CAS, CSH, E, K, MJG, M, MO, NY, TNM, UPS, the Makiling Center for Mountain Ecosystems, Marie Selby Botanical Gardens, Munich Botanical Garden, Darin S. Penneys, Deepthi Yakandawala, Nina Peck, Kate Armstrong, and Maxim Nuraliev for providing samples. We thank Martina Silber for help with DNA extraction and

preparing plates for sequencing. We thank Monique Romero-Brito for sharing Hyb-Seq data of *Boyania kenwurdackii*. We thank Andreas Gröger, Mario Terra, and John L. Clark for photographs of live members of the Sonerileae. Luo Chen also thanks Dirk Metzler for his suggestions on this study. Finally, we thank the Editor-in-Chief, Sean W. Graham, Associate Editor Thais Vasconcelos, and two anonymous reviewers for their valuable comments and suggestions. This study was funded by the German Science Foundation (DFG), project numbers KA1816/12-1, VE1291/1-1, and KA1816/12-2. Open Access funding enabled and organized by Projekt DEAL.

DATA AVAILABILITY STATEMENT

The newly developed reference, final sequence alignments, gene trees, and species trees generated in this study, as well as scripts used to run analyses are available at Zenodo (<https://doi.org/10.5281/zenodo.19453003>).

ORCID

Luo Chen  <https://orcid.org/0000-0003-0727-9545>

Marie Claire Veranso-Libalah  <https://orcid.org/0000-0001-7847-1740>

J. Peter Quakenbush  <https://orcid.org/0000-0003-2378-5392>

Ying Liu  <https://orcid.org/0000-0003-0613-837X>

Fabián A. Michelangeli  <https://orcid.org/0000-0001-7348-143X>

Diego F. Morales-Briones  <https://orcid.org/0000-0003-1535-5739>

Gudrun Kadereit  <https://orcid.org/0000-0003-0094-8769>

REFERENCES

- Allio, R., A. Schomaker-Bastos, J. Romiguier, F. Prosdocimi, B. Nabholz, and F. Delsuc. 2020. MitoFinder: efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. *Molecular Ecology Resources* 20: 892-905.
- Amarasinghe, P., S. Joshi, N. Page, L. S. Wijedasa, M. Merello, H. Kathriarachchi, R. D. Stone, et al. 2021. Evolution and biogeography of *Memecylon*. *American Journal of Botany* 108: 628-646.
- Bacci, L. F., F. A. Michelangeli, and R. Goldenberg. 2019. Revisiting the classification of Melastomataceae: implications for habit and fruit evolution. *Botanical Journal of the Linnean Society* 190: 1-24.
- Baker, W. J., P. Bailey, V. Barber, A. Barker, S. Bellot, D. Bishop, L. R. Botigué, et al. 2021. A comprehensive phylogenomic platform for exploring the angiosperm tree of life. *Systematic Biology* 71: 301-319.
- Benítez-Villaseñor, A., C. Granados Mendoza, S. Wanke, M. Peñafiel Cevallos, M. E. Freire, E. M. Lemmon, A. R. Lemmon, and S. Magallón. 2023. The use of anchored hybrid enrichment data to resolve higher-level phylogenetic relationships: a proof-of-concept applied to Asterales (Eudicotyledoneae; Angiosperms). *Molecular Phylogenetics and Evolution* 181: 107714.
- Blischak, P. D., A. J. Wenzel, and A. D. Wolfe. 2014. Gene prediction and annotation in *Penstemon* (Plantaginaceae): a workflow for marker development from extremely low-coverage genome sequencing. *Applications in Plant Sciences* 2: 1400044.
- Bolger, A. M., M. Lohse, and B. Usadel. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 2114-2120.
- Borowiec, M. L. 2016. AMAS: A fast tool for alignment manipulation and computing of summary statistics. *PeerJ* 4: e1660.
- Bossert, S., E. A. Murray, E. A. B. Almeida, S. G. Brady, B. B. Blaimer, and B. N. Danforth. 2019. Combining transcriptomes and ultraconserved elements to illuminate the phylogeny of Apidae. *Molecular Phylogenetics and Evolution* 130: 121-131.
- Bushnell, B. 2022. BBMap: A fast, accurate, splice-aware aligner. BBTools official website: <https://bbmap.org> GitHub: <https://github.com/bbushnell/BBTools>
- Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10: 421.
- Chen, L., W.-T. Jin, X.-Q. Liu, and X.-Q. Wang. 2022. New insights into the phylogeny and evolution of Podocarpaceae inferred from transcriptomic data. *Molecular Phylogenetics and Evolution* 166: 107341.
- Chen, L., G. Kadereit, and M. C. Veranso-Libalah. 2023. Combining Angiosperms353 and Sanger data provides support for the reinstatement of the genus *Myrianthemum* (Melastomataceae). *Botanical Journal of the Linnean Society* 203: 123-133.
- Cheng, S., M. Melkonian, S. A. Smith, S. Brockington, J. M. Archibald, P.-M. Delaux, F.-W. Li, et al. 2018. 10KP: A phylodiverse genome sequencing plan. *GigaScience* 7: giy013.
- Cheon, S., J. Zhang, and C. Park. 2020. Is phylotranscriptomics as reliable as phylogenomics? *Molecular Biology and Evolution* 37: 3672-3683.
- Chomczynski, P., and N. Sacchi. 2006. The single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction: Twenty-something years on. *Nature Protocols* 1: 581-585.
- Costa, L., A. Marques, C. Buddenhagen, W. W. Thomas, B. Huettel, V. Schubert, S. Dodsworth, et al. 2021. Aiming off the target: recycling target capture sequencing reads for investigating repetitive DNA. *Annals of Botany* 128: 835-848.
- Dagallier, L.-P. M. J., and F. A. Michelangeli. 2024. An updated and extended version of the Melastomataceae probe set for target capture. *Applications in Plant Sciences* 12: e11564.
- Davidson, N. M., and A. Oshlack. 2014. Corset: enabling differential gene expression analysis for *de novo* assembled transcriptomes. *Genome Biology* 15: 410.
- Dodsworth, S., L. Pokorny, M. G. Johnson, J. T. Kim, O. Maurin, N. J. Wickett, F. Forest, and W. J. Baker. 2019. Hyb-Seq for flowering plant systematics. *Trends in Plant Science* 24: 887-891.
- Fu, L., B. Niu, Z. Zhu, S. Wu, and W. Li. 2012. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28: 3150-3152.
- Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* 29: 644-652.
- Guo, C., Y. Luo, L.-M. Gao, T.-S. Yi, H.-T. Li, J.-B. Yang, and D.-Z. Li. 2023. Phylogenomics and the flowering plant tree of life. *Journal of Integrative Plant Biology* 65: 299-323.
- Haas, B. J., A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, J. Bowden, M. B. Couger, et al. 2013. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols* 8: 1494-1512.
- Hao, Y., Y.-Z. Zhou, B. Chen, G.-Z. Chen, Z.-Y. Wen, D. Zhang, W.-H. Sun, et al. 2022. The *Melastoma dodecandrum* genome and the evolution of Myrtales. *Journal of Genetics and Genomics* 49: 120-131.
- Hatje, K., O. Keller, B. Hammesfahr, H. Pillmann, S. Waack, and M. Kollmar. 2011. Cross-species protein sequence and gene structure prediction with fine-tuned WebScipio 2.0 and Scipio. *BMC Research Notes* 4: 265.
- He, J., R. Lyu, Y. Luo, J. Xiao, L. Xie, J. Wen, W. Li, et al. 2022. A phylotranscriptome study using silica gel-dried leaf tissues produces an updated robust phylogeny of Ranunculaceae. *Molecular Phylogenetics and Evolution* 174: 107545.
- Hu, X.-Z., C. Guo, S.-Y. Qin, D.-Z. Li, and Z.-H. Guo. 2024. Deep genome skimming reveals the hybrid origin of *Pseudosasa gracilis* (Poaceae: Bambusoideae). *Plant Diversity* 46: 344-352.
- Huang, W., B. Xu, W. Guo, Z. Huang, Y. Li, and W. Wu. 2024. *De novo* genome assembly and population genomics of a shrub tree *Barthea*

- barthei* (Hance) Krass provide insights into the adaptive color variations. *Frontiers in Plant Science* 15: 1365686.
- Jantzen, J. R., P. Amarasinghe, R. A. Folk, M. Reginato, F. A. Michelangeli, D. E. Soltis, N. Cellinese, and P. S. Soltis. 2020. A two-tier bioinformatic pipeline to develop probes for target capture of nuclear loci with applications in Melastomataceae. *Applications in Plant Sciences* 8: e11345.
- Jantzen, J. R., P. J. F. Guimarães, L. C. Pederneiras, A. L. F. Oliveira, D. E. Soltis, and P. S. Soltis. 2022. Phylogenomic analysis of *Tibouchina* s.s. (Melastomataceae) highlights the evolutionary complexity of Neotropical savannas. *Botanical Journal of the Linnean Society* 199: 372–411.
- Jiménez-Mena, B., H. Flávio, R. Henriques, A. Manuzzi, M. Ramos, D. Meldrup, J. Edson, et al. 2022. Fishing for DNA? Designing baits for population genetics in target enrichment experiments: guidelines, considerations and the new tool supeRbait. *Molecular Ecology Resources* 22: 2105–2119.
- Johnson, M. G., E. M. Gardner, Y. Liu, R. Medina, B. Goffinet, A. J. Shaw, N. J. C. Zerega, and N. J. Wickett. 2016. HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Applications in Plant Sciences* 4: 1600016.
- Johnson, M. G., L. Pokorny, S. Dodsworth, L. R. Botigué, R. S. Cowan, A. Devault, W. L. Eiserhardt, et al. 2018. A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using *k*-medoids clustering. *Systematic Biology* 68: 594–606.
- Kalyanamoorthy, S., B. Q. Minh, T. K. F. Wong, A. von Haeseler, and L. S. Jermini. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* 14: 587–589.
- Katoh, K., and D. M. Standley. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* 30: 772–780.
- Lagou, L. J., G. Kadereit, and D. F. Morales-Briones. 2024. Phylogenomic analysis of target enrichment and transcriptome data uncovers rapid radiation and extensive hybridization in the slipper orchid genus *Cypripedium*. *Annals of Botany* 134: 1229–1250.
- Langmead, B., and S. L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9: 357–359.
- Leebens-Mack, J. H., M. S. Barker, E. J. Carpenter, M. K. Deyholos, M. A. Gitzendanner, S. W. Graham, I. Grosse, et al. 2019. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574: 679–685.
- Li, D., C.-M. Liu, R. Luo, K. Sadakane, and T.-W. Lam. 2015. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31: 1674–1676.
- Lin, C.-W., T.-C. Hsu, H. Luu, T.-Y. Yang, and S. Wei. 2022. *Nephoanthus* (Melastomataceae: Sonerileae), a new genus segregated from *Phyllagathis* s.l., with a new species from southern Vietnam. *Phytotaxa* 547: 66–76.
- Liu, B.-B., Z.-Y. Ma, C. Ren, R. G. J. Hodel, M. Sun, X.-Q. Liu, G.-N. Liu, et al. 2021. Capturing single-copy nuclear genes, organellar genomes, and nuclear ribosomal DNA from deep genome skimming data for plant phylogenetics: a case study in Vitaceae. *Journal of Systematics and Evolution* 59: 1124–1138.
- Liu, S.-Y., Y.-Y. Yang, Q. Tian, Z.-Y. Yang, S.-F. Li, P. J. Valdes, A. Farnsworth, et al. 2025. An integrative framework reveals widespread gene flow during the early radiation of oaks and relatives in Quercoideae (Fagaceae). *Journal of Integrative Plant Biology* 67: 1119–1141.
- Liu, Y., J.-H. Dai, Q.-Y. Zhuang, C.-Y. Zou, and K.-N. Ma. 2024. Resurrection of *Perilimnastes* (Sonerileae, Melastomataceae) with description of a new species *P. nana*. *PhytoKeys* 238: 11–31.
- Liu, Y., M. C. Veranso-Libalah, G. Kadereit, R.-C. Zhou, J. P. Quakenbush, C.-W. Lin, and J. S. Wai. 2022. Systematics of the tribe Sonerileae. In R. Goldenberg, F. A. Michelangeli, and F. Almeda [eds.], *Systematics, evolution, and ecology of Melastomataceae*, 321–343. Springer International, Cham, Switzerland.
- Mai, U., and S. Mirarab. 2018. TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees. *BMC Genomics* 19: 272.
- Mandel, J. R., R. B. Dikow, V. A. Funk, R. R. Masalia, S. E. Staton, A. Kozik, R. W. Michelmore, et al. 2014. A target enrichment method for gathering phylogenetic information from hundreds of loci: an example from the Compositae. *Applications in Plant Sciences* 2: 1300085.
- Maurin, O., A. Anest, S. Bellot, E. Biffin, G. Brewer, T. Charles-Dominique, R. S. Cowan, et al. 2021. A nuclear phylogenomic study of the angiosperm order Myrtales, exploring the potential and limitations of the universal Angiosperms353 probe set. *American Journal of Botany* 108: 1087–1111.
- McLay, T. G. B., J. L. Birch, B. F. Gunn, W. Ning, J. A. Tate, L. Nauheimer, E. M. Joyce, et al. 2021. New targets acquired: improving locus recovery from the Angiosperms353 probe set. *Applications in Plant Sciences* 9: e11420.
- Minh, B. Q., H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. von Haeseler, and R. Lanfear. 2020. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution* 37: 1530–1534.
- Molloy, E. K., and T. Warnow. 2017. To include or not to include: the impact of gene filtering on species tree estimation methods. *Systematic Biology* 67: 285–303.
- Morales-Briones, D. F., G. Kadereit, D. T. Tefarikis, M. J. Moore, S. A. Smith, S. F. Brockington, A. Timoneda, et al. 2021. Disentangling sources of gene tree discordance in phylogenomic data sets: testing ancient hybridizations in Amaranthaceae s.l. *Systematic Biology* 70: 219–235.
- Niu, Z.-Y., Z.-X. Zhang, Z.-Y. Cai, Y.-H. Tong, S.-J. Zeng, S. W. Graham, and N.-H. Xia. 2025. Phylogenomic and morphological evidence supports the reinstatement of the bamboo genus *Clavinodum* from *Oligostachyum* (Poaceae: Bambusoideae). *Molecular Phylogenetics and Evolution* 207: 108327.
- Oakley, T. H., J. M. Wolfe, A. R. Lindgren, and A. K. Zaharoff. 2012. Phylotranscriptomics to bring the understudied into the fold: monophyletic Ostracoda, fossil placement, and Pancrustacean phylogeny. *Molecular Biology and Evolution* 30: 215–233.
- Ortiz, E. M., A. Höwener, G. Shigita, M. Raza, O. Maurin, A. Zuntini, F. Forest, et al. 2023. A novel phylogenomics pipeline reveals complex pattern of reticulate evolution in Cucurbitales. *bioRxiv* <https://doi.org/10.1101/2023.10.27.564367> [preprint].
- Pease, J. B., J. W. Brown, J. F. Walker, C. E. Hinchliff, and S. A. Smith. 2018. Quartet sampling distinguishes lack of support from conflicting support in the green plant tree of life. *American Journal of Botany* 105: 385–403.
- Pezzini, F. F., G. Ferrari, L. L. Forrest, M. L. Hart, K. Nishii, and C. A. Kidner. 2023. Target capture and genome skimming for plant diversity studies. *Applications in Plant Sciences* 11: e11537.
- Pouchon, C., and C. G. Boluda. 2023. REFMAKER: Make your own reference to target nuclear loci in low coverage genome skimming libraries. Phylogenomic application in Sapotaceae. *Molecular Phylogenetics and Evolution* 186: 107826.
- Quakenbush, J. P., L. Chen, D. S. Penneys, T. J. Barkman, Y. Liu, D. Yakandawala, M. C. Veranso-Libalah, and G. Kadereit. 2025. Systematics of the fleshy-fruited Sonerileae (Melastomataceae). *Taxon* 74: 39–65.
- Quattrini, A. M., L. J. McCartin, E. E. Easton, J. Horowitz, H. H. Wirshing, H. Bowers, K. Mitchell, et al. 2024. Skimming genomes for systematics and DNA barcodes of corals. *Ecology and Evolution* 14: e11254.
- R Core Team. 2023. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Website <https://www.R-project.org/>.
- Reichelt, N., J. Wen, C. Pätzold, and M. S. Appelhaus. 2021. Target enrichment improves phylogenetic resolution in the genus *Zanthoxylum* (Rutaceae) and indicates both incomplete lineage sorting and hybridization events. *Annals of Botany* 128: 497–510.
- Ruiz-Vargas, N., K. Ramanauskas, A. S. Tyska, E. C. Bretz, M. T. S. Yeo, R. J. Mason-Gamer, and J. F. Walker. 2024. Transcriptome data from

- silica-preserved leaf tissue reveal gene flow patterns in a Caribbean bromeliad. *Annals of Botany* 133: 459-472.
- Shekhar, S., S. Roch, and S. Mirarab. 2018. Species tree estimation using ASTRAL: How many genes are enough? *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 15: 1738-1747.
- Smith-Unna, R., C. Boursnell, R. Patro, J. M. Hibberd, and S. Kelly. 2016. *TransRate*: reference-free quality assessment of *de-novo* transcriptome assemblies. *Genome Research* 26: 1134-1144.
- Smith, S. A., and C. W. Dunn. 2008. Phyutility: A phyloinformatics tool for trees, alignments and molecular data. *Bioinformatics* 24: 715-716.
- Song, L., and L. Florea. 2015. Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. *GigaScience* 4: 48.
- Stamatakis, A. 2014. RAXML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30: 1312-1313.
- Straub, S. C. K., M. Parks, K. Weitemier, M. Fishbein, R. C. Cronn, and A. Liston. 2012. Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. *American Journal of Botany* 99: 349-364.
- Tabatabaee, Y., C. Zhang, T. Warnow, and S. Mirarab. 2023. Phylogenomic branch length estimation using quartets. *Bioinformatics* 39: i185-i193.
- The Darwin Tree of Life Project Consortium. 2022. Sequence locally, think globally: The Darwin Tree of Life Project. *Proceedings of the National Academy of Sciences, USA* 119: e2115642118.
- Trevisan, B., D. M. C. Alcantara, D. J. Machado, F. P. L. Marques, and D. J. G. Lahr. 2019. Genome skimming is a low-cost and robust strategy to assemble complete mitochondrial genomes from ethanol preserved specimens in biodiversity studies. *PeerJ* 7: e7543.
- Twyford, A. D., and R. W. Ness. 2017. Strategies for complete plastid genome sequencing. *Molecular Ecology Resources* 17: 858-868.
- Ulloa Ulloa, C., F. Almeda, R. Goldenberg, G. Kadereit, F. A. Michelangeli, D. S. Penneys, R. D. Stone, and M. C. Veranso-Libalah. 2022. Melastomataceae: Global diversity, distribution, and endemism. In R. Goldenberg, F. A. Michelangeli, and F. Almeda [eds.], *Systematics, evolution, and ecology of Melastomataceae*, 3-28. Springer International, Cham, Switzerland.
- van Dongen, S., and C. Abreu-Goodger. 2012. Using MCL to extract clusters from networks. In J. van Helden, A. Toussaint, and D. Thiery [eds.], *Bacterial molecular networks: methods and protocols*, 281-295. Springer, NY, NY, USA.
- Veranso-Libalah, M. C., H. Mertes, R. D. Stone, L. Chen, T. Stévant, F. Almeda, X. M. van der Burgt, and G. Kadereit. 2023. Phylogeny and systematics of the tribe Sonerileae (Melastomataceae) in Africa: a revised taxonomic classification. *Journal of Systematics and Evolution* 61: 657-681.
- Villaverde, T., L. Pokorny, S. Olsson, M. Rincón-Barrado, M. G. Johnson, E. M. Gardner, N. J. Wickett, et al. 2018. Bridging the micro- and macroevolutionary levels in phylogenomics: Hyb-Seq solves relationships from populations to species and above. *New Phytologist* 220: 636-650.
- Wang, X., T. Xiong, Y. Wang, X. Zhang, and M. Sun. 2024. Integrating genomic sequencing resources: an innovative perspective on recycling with universal Angiosperms353 probe sets. *Horticulture Advances* 2: 4.
- Wang, Z., M. Gerstein, and M. Snyder. 2009. RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics* 10: 57-63.
- Weitemier, K., S. C. K. Straub, R. C. Cronn, M. Fishbein, R. Schmickl, A. McDonnell, and A. Liston. 2014. Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. *Applications in Plant Sciences* 2: 1400042.
- Wickett, N. J., S. Mirarab, N. Nguyen, T. Warnow, E. Carpenter, N. Matasci, S. Ayyampalayam, et al. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences, USA* 111: E4859-E4868.
- Wurdack, K. J., and F. A. Michelangeli. 2019. Systematics and relationships of *Tryssophyton* (Melastomataceae), with a second species from the Pakaraima Mountains of Guyana. *PhytoKeys* 136: 1-21.
- Yang, L.-H., X.-Z. Shi, F. Wen, and M. Kang. 2023. Phylogenomics reveals widespread hybridization and polyploidization in *Henckelia* (Gesneriaceae). *Annals of Botany* 131: 953-966.
- Yang, Y., and S. A. Smith. 2013. Optimizing *de novo* assembly of short-read RNA-seq data for phylogenomics. *BMC Genomics* 14: 328.
- Yang, Y., and S. A. Smith. 2014. Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: Improving accuracy and matrix occupancy for phylogenomics. *Molecular Biology and Evolution* 31: 3081-3092.
- Yockteng, R., A. M. R. Almeida, S. Yee, T. Andre, C. Hill, and C. D. Specht. 2013. A method for extracting high-quality RNA from diverse plants for next-generation sequencing and gene expression analyses. *Applications in Plant Sciences* 1: 1300070.
- Yu, G., D. K. Smith, H. Zhu, Y. Guan, and T. T.-Y. Lam. 2017. ggtree: An R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* 8: 28-36.
- Zhang, C., and S. Mirarab. 2022. Weighting by gene tree uncertainty improves accuracy of quartet-based species trees. *Molecular Biology and Evolution* 39: msac215.
- Zhang, C., M. Rabiee, E. Sayyari, and S. Mirarab. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* 19: 153.
- Zhong, Y., W. Wu, C. Sun, P. Zou, Y. Liu, S. Dai, and R. Zhou. 2023. Chromosomal-level genome assembly of *Melastoma candidum* provides insights into trichome evolution. *Frontiers in Plant Science* 14: 1126319.
- Zhou, Q.-J., J.-H. Dai, C.-W. Lin, W.-L. Ng, T. Van Do, J. S. Wai, F. A. Michelangeli, et al. 2022. Out of chaos: Phylogenomics of Asian Sonerileae. *Molecular Phylogenetics and Evolution* 175: 107581.
- Zhou, Q.-J., C.-W. Lin, J.-H. Dai, R.-C. Zhou, and Y. Liu. 2019a. Exploring the generic delimitation of *Phyllagathis* and *Bredia* (Melastomataceae): a combined nuclear and chloroplast DNA analysis. *Journal of Systematics and Evolution* 57: 256-267.
- Zhou, Q., C.-W. Lin, W. L. Ng, J. Dai, T. Denda, R. Zhou, and Y. Liu. 2019b. Analyses of plastome sequences improve phylogenetic resolution and provide new insight into the evolutionary history of Asian Sonerileae/Dissochaeteae. *Frontiers in Plant Science* 10: 1477.
- Zhuang, Q.-Y., J.-H. Dai, C.-Y. Zou, Y. Liu, Z.-Y. Yu, X. Jiang, Q.-J. Zhou, et al. 2025. Generic delimitation in the *Cyphotheca-Plagiopetalum-Sporoxeia* clade. *Molecular Phylogenetics and Evolution* 205: 108290.
- Zuntini, A. R., T. Carruthers, O. Maurin, P. C. Bailey, K. Leempoel, G. E. Brewer, N. Epitawalage, et al. 2024. Phylogenomics and the rise of the angiosperms. *Nature* 629: 843-850.

SUPPORTING INFORMATION

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

Appendix S1: Detailed list of taxa sampled in this study, including voucher information for newly generated data and accession numbers for public data.

Appendix S2: Matching loci between the newly developed reference comprising 5626 loci and the Melastomataceae probe set containing 396 loci, and the Angiosperms353 probe set.

Appendix S3: Summary of sequence recovery across samples.

Appendix S4: Frequency of taxa in the final orthologs for the new reference data set. Taxa sequenced using the same strategy are additionally separately listed.

Appendix S5: Frequency of taxa in the final orthologs for the 396-locus data set. Taxa sequenced using the same strategy are additionally separately listed.

Appendix S6: Distribution of orthologs per taxon by different sequencing methods in the final 225 orthologs from the 396-locus data set.

Appendix S7: ASTRAL species tree ST2 inferred from 184 taxa and 225 orthologs from the 396-locus data set, showing the relationships within Sonerileae. Local posterior probabilities (LPP) are shown above branches where $LPP < 1$.

Appendix S8: ASTRAL species tree ST3 inferred from 18 taxa and 5626 orthologs using only genomic and transcriptomic data, showing the relationships within Sonerileae. Local posterior probabilities (LPP) are shown above branches where $LPP < 1$.

Appendix S9: Gene tree conflicts measured with quartet sampling (QS) analysis. Each branch is annotated

with three QS scores: quartet concordance (QC), quartet differential (QD), and quartet informativeness (QI).

How to cite this article: Chen, L., M. C. Veranso-Libalah, J. P. Quakenbush, Y. Liu, F. A. Michelangeli, D. F. Morales-Briones, and G. Kadereit. 2026. Integrative utilization of genomic resources for improved phylogenetic resolution in Sonerileae (Melastomataceae). *American Journal of Botany* 113(6): e70216. <https://doi.org/10.1002/ajb2.70216>