






INVITED SPECIAL ARTICLE

For the Special Issue: Exploring Angiosperms353: a Universal Toolkit for Flowering Plant Phylogenomics

Joining forces in Ochnaceae phylogenomics: a tale of two targeted sequencing probe kits

Toral Shah^{1,2,6} , Julio V. Schneider³, Georg Zizka^{3,4}, Olivier Maurin¹ , William Baker¹ , Félix Forest¹ , Grace E. Brewer¹, Vincent Savolainen² ,
Iain Darbyshire¹ , and Isabel Larridon^{1,5} 

Manuscript received 29 September 2020; revision accepted 23 February 2021.

¹ Royal Botanic Gardens, Kew, Richmond, Surrey TW9 3AE, UK

² Department of Life Sciences, Imperial College, Silwood Park Campus, Ascot, Berks SL5 7PY, UK

³ Department of Botany and Molecular Evolution, Senckenberg Research Institute and Natural History Museum Frankfurt, Senckenberganlage 25, Frankfurt am Main D-60325, Germany

⁴ Institute of Ecology, Evolution and Diversity, Goethe University, Max-von-Laue-Str. 13, Frankfurt am Main 60438, Germany

⁵ Systematic and Evolutionary Botany Lab, Department of Biology, Ghent University, K.L., Ledeganckstraat 35, Gent 9000, Belgium

⁶ Author for correspondence (e-mail: T.Shah@kew.org)

Citation: Shah, T., J. V. Schneider, G. Zizka, O. Maurin, W. Baker, F. Forest, G. E. Brewer, I. Darbyshire, and I. Larridon. 2021. Joining forces in Ochnaceae phylogenomics: a tale of two targeted sequencing probe kits. *American Journal of Botany* 108(7): 1201–1216.

doi:10.1002/ajb2.1682

PREMISE: Both universal and family-specific targeted sequencing probe kits are becoming widely used for reconstruction of phylogenetic relationships in angiosperms. Within the pantropical Ochnaceae, we show that with careful data filtering, universal kits are equally as capable in resolving intergeneric relationships as custom probe kits. Furthermore, we show the strength in combining data from both kits to mitigate bias and provide a more robust result to resolve evolutionary relationships.

METHODS: We sampled 23 Ochnaceae genera and used targeted sequencing with two probe kits, the universal Angiosperms353 kit and a family-specific kit. We used maximum likelihood inference with a concatenated matrix of loci and multispecies-coalescence approaches to infer relationships in the family. We explored phylogenetic informativeness and the impact of missing data on resolution and tree support.

RESULTS: For the Angiosperms353 data set, the concatenation approach provided results more congruent with those of the Ochnaceae-specific data set. Filtering missing data was most impactful on the Angiosperms353 data set, with a relaxed threshold being the optimum scenario. The Ochnaceae-specific data set resolved consistent topologies using both inference methods, and no major improvements were obtained after data filtering. Merging of data obtained with the two kits resulted in a well-supported phylogenetic tree.

CONCLUSIONS: The Angiosperms353 data set improved upon data filtering, and missing data played an important role in phylogenetic reconstruction. The Angiosperms353 data set resolved the phylogenetic backbone of Ochnaceae as equally well as the family specific data set. All analyses indicated that both *Sauvagesia* L. and *Campylospermum* Tiegh. as currently circumscribed are polyphyletic and require revised delimitation.

KEY WORDS coalescence; custom; maximum likelihood; missing data; phylogenetic informativeness; probe kit; universal.

The use of molecular data to generate phylogenetic inferences to support plant classification has been constantly increasing over the last 30 years (Chase et al., 1993; McCormack et al., 2012; Xi et al., 2013; Wickett et al., 2014). In the last decade, there has been a rise

in the number of methods available for molecular data collection and inference of the angiosperm Tree of Life. Since the introduction of Sanger sequencing and the continuously improved application of the polymerase chain reaction (PCR) to amplify DNA, earlier

classifications largely based on plant morphology have been tested and improved (Soltis et al., 2012; Barrett et al., 2016).

Recently, developments in high-throughput sequencing have allowed cost-effective, genome-scale sequencing for a wider range of plant groups (Lemmon and Lemmon, 2013; McKain et al., 2018; Dodsworth et al., 2019). Through the selection and enrichment of specific regions of the genome (i.e., target enrichment or targeted sequencing; Turner et al., 2009; Mamanova et al., 2010), hundreds of low-copy nuclear genes can be captured using specific probes (also known as “baits”; Grover et al., 2012; Couvreur et al., 2019). This technique is rapidly emerging as a powerful tool for plant phylogenomics (Summerer, 2009; Mamanova et al., 2010; Lemmon, et al., 2012), with growing potential application across angiosperms at different taxonomic levels (Lemmon et al., 2012; Villaverde et al., 2018; Brewer et al., 2019; Soto Gomez et al., 2019; Murphy et al., 2020). Targeted sequencing encompasses a number of methodologies, such as anchored phylogenomics (Lemmon et al., 2012), exon capture (Mamanova et al., 2010), and Hyb-Seq (Weitemier et al., 2014). Once the genomic libraries are enriched for the regions of interest they are then sequenced on high-throughput sequencing platforms (Gnrirke et al., 2009; Grover et al., 2012; Fragoso-Martínez et al., 2017; Léveillé-Bourret et al., 2018; Johnson et al., 2019).

The wide availability of transcriptome data, particularly through the One Thousand Plants project (OneKP; Matasci et al., 2014), has facilitated relatively easy and affordable probe design for many plant groups (McKain et al., 2018). It is argued that taxon-specific probes are advantageous in obtaining higher gene recovery, with more variable loci appropriate to the target group, thus more confidently able to resolve phylogenetic relationships (Kadlec et al., 2017). Recently, there has been a surge in the development of taxon-specific probe kits for groups such as Compositae (Mandel et al., 2014), *Erica* Tourn. ex L. (Ericaceae) (Kadlec et al., 2017), Fabaceae (Leguminosae) (Vatanparast et al., 2018), *Buddleja* Houst. ex L. (Scrophulariaceae) (Chau et al., 2018), *Dioscorea* Plum. ex L. (Dioscoreaceae) (Soto Gomez et al., 2019), and Cyperaceae (Villaverde et al., 2020). Although custom-designed probe kits may yield more genes than universal probe kits, they require access to genomic resources and expertise on the particular plant group (Dodsworth et al., 2019). The development of universal probe kits brings huge possibilities for angiosperm phylogenetics, and their full potential is yet to be understood (Buddenhagen et al., 2016 [Preprint]; Johnson et al., 2019; Larridon et al., 2020). Examples of universal kits include the Angiosperms I kit for anchored phylogenomics (Buddenhagen et al., 2016 [Preprint]; Léveillé-Bourret et al., 2018), the Genealogy of Flagellate Plants kit (GoFlag; Breinholt et al., 2020 [Preprint]), and the Angiosperms353 probe kit (Johnson et al., 2019). Both the Angiosperms353 and GoFlag probe kits were designed using OneKP transcriptome data, focusing on a set of putatively single-copy nuclear loci that are orthologous across green plants. The Angiosperms353 kit includes 353 of these loci and was designed using *k*-medoids clustering (Bauckhage, 2015). The kit includes up to 15 variants for each of the 353 genes, making it widely applicable in up to 95% of angiosperm species and suitable for studies at both deep and shallow phylogenetic levels (Johnson et al., 2019; Dodsworth et al., 2019).

When reconstructing phylogenetic relationships, the increase in sampled loci and taxon sampling is known to improve the accuracy of phylogenetic reconstruction (Pollock et al., 2002; Zwiclk and Hillis, 2002; Hedtke et al., 2006; Crawley and Hilu, 2012; Alvizu et al., 2018; Jantzen et al., 2019). While this increase in data applies

to genomic data sets (Lee et al., 2011; McCormack et al., 2012), it also raises questions about phylogenetic informativeness and the quantity of missing data. Several studies have shown that phylogenetic reconstruction is less affected by missing data in data sets with hundreds or thousands of genes (Thomson and Shaffer, 2010; Jiang et al., 2014; Xi et al., 2016). Conversely, Sayyari et al. (2017) showed that missing data can have a negative impact on gene and species tree inference. Thus, the balance between increasing the number of genes and increasing missing data is still poorly known (Jiang et al., 2014; Hosner et al., 2016; Huang and Lacey Knowles, 2016; Streicher et al., 2016). Moreover, the uncertainties of how missing data impacts various phylogenetic inference methods, particularly coalescent-based approaches require more work (Crawley and Hilu, 2012; Roure et al., 2013; Hosner et al., 2016).

To explore this trade-off between more genes and the inclusion of missing data, investigation on phylogenetic informativeness of individual genes profiles could provide a quantitative measure of the power of each gene, allowing us to differentiate between phylogenetic noise and true biological relationships (Townsend, 2007). Straub et al. (2014) defined “phylogenetic noise” as any character within the sequence data that may obscure or contradict the true gene genealogy. Phylogenetic reconstruction may be impacted by genes and sites with high phylogenetic noise and large amounts of missing data (Townsend et al., 2012; Xi et al., 2016). Here, we investigated the impacts of missing data on phylogenetic informativeness when resolving phylogenetic relationships using the family Ochnaceae and two probe kits, a family-specific kit and a universal probe kit.

The pantropical family Ochnaceae consists of 32 genera and ca. 550 species (Christenhusz et al., 2017; POWO, 2019). Since APG III (2009), the family has been expanded to include Medusagynaceae and Quiinaeaceae, which form a distinct clade within the expanded Ochnaceae (Davis et al., 2005; Korotkova et al., 2009; Xi et al., 2012; Angiosperm Phylogeny Group, 2016). The family comprises three subfamilies: Medusagynoideae Reveal with one species occurring in the Seychelles, the neotropical Quiinoideae Luerss. with four genera (Schneider et al., 2002, 2006; Schneider and Zizka, 2017), and the pantropical Ochnoideae. The latter subfamily is further divided into four tribes: Testuleae J.V.Schneid., Luxemburgieae Horan., Sauvagesieae Ging. ex DC. and Ochneae Bartl. The tribe Ochneae is still further divided into three subtribes: Elvasiinae J.V.Schneid., Lophirinae J.V.Schneid., and Ochninae Kanis ex J.V.Schneid.

Relationships within Ochnaceae have been the focus of recent Sanger sequencing (Bissengou, 2014; Schneider et al., 2014) and phylogenomic studies (Schneider et al., 2020). Overall, these studies confirmed the monophyly of the wider family and presented a new infrafamilial classification. Despite many similarities, these studies have revealed some nodes requiring further study. First is the polyphyly of two genera: Schneider et al. (2014) found *Sauvagesia* to be polyphyletic with *Sauvagesia serrata* (Korth.) Sastre isolated from the rest of the genus, whilst Bissengou (2014) revealed polyphyly of the genus *Campylospermum*. In the latter case, species from West/Central Africa grouped sister to *Idertia* Farron, whilst species from East Africa and Madagascar were sister to *Rhabdophyllum* Tiegh. The second taxonomic concern is the relationships amongst genera of the subtribe Ochninae. In Schneider et al. (2014), *Campylospermum* was resolved as sister to the rest of Ochninae, and *Brackenridgea* A.Gray, *Ochna* L., *Idertia*, and *Rhabdophyllum* formed a clade. Their results directly contradict with morphological relationships observed by

Farron (1963) and Sosef (2008). In contrast, Bissengou (2014) resolved *Brackenridgea* as sister to *Ochna*, with the *Brackenridgea*-*Ochna* clade sister to *Campylopermum*, *Ouratea* Aubl., and *Rhabdophyllum*.

Schneider et al. (2020) used a custom bait kit that was designed to resolve relationships in Ochnaceae based on a sole ingroup transcriptome from Ochnaceae. Although most of the phylogenetic backbone of the family was resolved in that study, the bias toward Ochnaceae in terms of the number of successfully captured genes might have been one reason for the few remaining unclear relationships that still require additional data. Therefore, we were interested in the performance of a more universal bait kit that is supposed to contain more conserved genes that might be helpful in resolving these recalcitrant nodes. Our study aimed to investigate the power of two targeted sequencing kits, the universal Angiosperms353 kit (targeting 353 low- or single-copy nuclear genes with ca. 0.26 Mbp; Johnson et al., 2019), and an Ochnaceae-specific kit (targeting 275 low or single copy nuclear loci with ca. 0.66 Mbp; Schneider et al., 2020). The Angiosperms353 probe kit was designed using transcriptome data from more than 600 angiosperms and consists of 353 single-copy genes useful for phylogenetic studies in all angiosperm lineages (Johnson et al., 2019). In light of the rapid advances of high-throughput sequencing techniques, we explored the performance of both probe kits by investigating gene recovery, phylogenetic informativeness, and topological resolution. With hundreds of genes, the two targeted sequencing kits provide an excellent framework for exploring the impact of missing data using different filtering strategies when trying to resolve recalcitrant nodes leading to taxonomic uncertainties in the family Ochnaceae.

MATERIALS AND METHODS

Taxon sampling

Samples prepared for enrichment with the Angiosperms353 probe kit consisted of 31 accessions from 23 genera of a total of 32 accepted genera. Sampling for enrichment with the Ochnaceae-specific kit consisted of 26 accessions from 23 genera. Voucher information for both data sets is listed in Appendix 1. The number of accessions, summarized in Appendix S1, shows that although sampling slightly differs between the kits, the accessions selected for each probe kit equally cover the breadth of the family. The number of genera sampled is representative for each subfamily, tribe and subtribe, where possible the same species were sampled. All specimen information is provided in Appendix S2.

Sequences for outgroup taxa were obtained from the OneKP project and were chosen to represent closely related families within the order Malpighiales (Stevens, 2001 onward), i.e., *Rhizophora mangle* L. (Rhizophoraceae), *Mammea americana* L. (Calophyllaceae), *Hypericum perforatum* L. (Hypericaceae), and *Garcinia oblongifolia* Champ. ex Benth. and *G. livingstonei* T.Anderson (Clusiaceae).

DNA extraction, library preparation, hybridization and sequencing

Angiosperms353—Molecular work for the accessions enriched with the Angiosperms353 kit was conducted at the Sackler Phylogenomic Laboratory, within the Jodrell Laboratory at Royal Botanic Gardens, Kew (London, UK). The closest taxon to the

Ochnaceae used in the design of the Angiosperms353 kit is *Ochna serrulata* (Hochst.) Walp. Genomic DNA was extracted from leaf tissue obtained from herbarium specimens using a modified cetyltrimethylammonium bromide (CTAB) approach, with chloroform–isoamyl alcohol (Sevag) and precipitation in isopropanol at -20°C (Doyle and Doyle, 1987). Some accessions were sourced from the Kew DNA Bank (<https://dnabank.science.kew.org/>) (Appendix S2).

The samples were purified with Agencourt AMPure XP Beads (Beckman Coulter, Indianapolis, IN, USA) following the manufacturer's protocol. All DNA extracts were quantified with a Quantus Fluorometer (Promega, Madison, WI, USA) and run on a 1% agarose gel to assess their average fragment size. Samples with low concentration (not visible on a 1% agarose gel) were assessed on an Agilent Technologies 4200 TapeStation System (Santa Clara, CA, USA). DNA extracts with average fragment size above 350 bp were sonicated using a Covaris M220 Focused-ultrasonicator (Covaris, Woburn, MA, USA) following the manufacturer's protocol to obtain an average fragment size of 350 bp. Dual-indexed libraries for Illumina sequencing were prepared using the NEBNext Ultra II DNA Library Prep Kit and the NEBNext Multiplex Oligos for Illumina (Dual Index Primers 1 and 2; New England BioLabs, Ipswich, MA, USA) following the manufacturer's protocol, but using half the recommended volumes. Briefly, we used 200 ng (or minimum 50 ng) of the fragmented DNA for the end-preparation reaction. Following the adapter ligation and size-selection, the DNA fragments were amplified using eight cycles of PCR. The libraries were quantified using a Quantus Fluorometer, and fragment size was assessed with TapeStation using High Sensitivity D1000 ScreenTapes. The final library size including the adapters was ca. 500 bp on average. Samples with similar library concentration and fragment size were pooled and enriched with the Angiosperms353 probes, following the MyBaits manual (Johnson et al., 2019). The hybridization was performed for 24 h at 65°C , followed by 12 cycles of PCR. Final products were again run on the TapeStation to assess the fragment size so they could be pooled equimolarly for sequencing. After multiplexing library pools, sequencing was performed on an Illumina HiSeqX instrument (San Diego, CA, USA) at Macrogen (Seoul, South Korea) producing 2×150 -bp paired-end reads.

Ochnaceae-specific—The molecular work for samples enriched with the Ochnaceae-specific kit was conducted at the Senckenberg Research Institute in Germany and followed the protocol of Schneider et al. (2020). The data for the accessions enriched with the Ochnaceae-specific kit represent a subsampling of the data set of Schneider et al. (2020) (Appendix S2). The development of the custom bait kit is outlined in detail by Schneider et al. (2020). Briefly, the target loci were selected by using transcriptome data from *Ochna serrulata* and the clusoid sister group. Low-copy loci in the ingroup data set were identified by an all against all comparison of the ingroup transcriptome sequences using BLASTn analysis. Sequences with no or poor matches to any other sequences in the data set were assumed to be of low- or single-copy. Pairwise BLASTn analysis was used to identify homologous sequences in the outgroup transcriptome data. Finally, those ingroup sequences were selected that had a blast match to a gene in the *Ricinus communis* L. (Euphorbiaceae) genome (version 0.1; Chan et al. (2010); obtained from JGI Phytozome (2016)) with three or more introns. In a last step, mitochondrial and plastid sequences were identified and removed from further analysis. The custom bait kit was produced for the 275 randomly selected nuclear targets (ca. 0.66 Mbp total) with

a 3.75× tiling density at MYcroarray (now Arbor Biosciences, Ann Arbor, MI, USA).

For DNA extraction, up to 20 mg silica-dried or herbarium leaf material was ground to a fine powder using a Bead Ruptor 24 (Omni International, Kennesaw, GA, USA) with steel beads. For DNA extraction, we used a modified CTAB protocol or, for some samples, the innuPrep Plant DNA Kit-KFFLX (Analytik Jena, Thuringia, Germany). DNA purity was checked with a DS 11 spectrophotometer (DeNovix, Wilmington, DE, USA), and yield was measured with Qubit 1.0 (ThermoFisher, Waltham, MA, USA). Fragment size ranges of the DNA samples were checked on 1% agarose gels or with 2200TapeStation with a Genomic Tape (Agilent). For extracted DNA, fragmentation was carried out using a Bioruptor UCD 300 Next Gen sonicator (Diagenode, Liège, Belgium). For highly degraded DNA, sonication was omitted. Library preparation was done with the NEBNext Ultra II DNA Library Preparation kit (New England Biolabs) following the manufacturer's manual. Samples were further grouped based on DNA yield (sets of 50, 100, 250, and 500 ng DNA input) and phylogenetic relationship (e.g., species of the same genus). After adaptor ligation, the libraries were cleaned using Ampure XP beads (Beckman Coulter, Brea, CA, USA). This cleanup included a fragment size selection for samples with DNA input >50 ng to achieve a mean fragment size of approximately 400 bp. In the case of lower DNA input and highly degraded DNA, size selection was usually omitted. All libraries were enriched using 5–14 PCR cycles, depending on DNA input. PCR enrichment also served for multiplexing with the NEBNext Multiplex Oligos for Illumina (Dual Index Primer Set 1, New England Biolabs). After a final clean-up with Ampure XP beads, libraries were quantified using Qubit 1.0. Sequence capture was conducted using the Ochnaceae-specific MyBaits kit (now Arbor Biosciences) and the manufacturer's protocol, version 3.0. Equimolar amounts of up to six libraries were pooled in a single hybridization reaction. For highly degraded DNA samples, hybridization was performed in single reactions with the baits diluted 1:2. Hybridization was run for 16–21 h at a temperature of 65°C. For some Sauvagesieae, Quiinoideae, *Medusagyne*, a hybridization temperature of 60°C was chosen following Li et al. (2013). By reducing the stringency, we aimed at capturing more divergent loci, too. Post-capture libraries were enriched during 10–14 PCR cycles using the “reamp” primers (see Meyer and Kircher 2010; annealing temperature: 65°C) and a HiFi polymerase (KAPA HiFi HotStart Ready Mix, KAPA Biosystems, Wilmington, DE, USA). The enriched post-capture libraries were cleaned with 1.1× Ampure XP beads, eluted in 10 mM Tris-HCl, 0.05% Tween 20, pH 8.0, and quantified with Qubit 1.0. The pooled post-capture libraries were further pooled (up to 96 dual indexed samples) at equimolar amounts for sequencing on an Illumina HiSeq 2500 with 150-bp paired-end reads at Macrogen (Seoul, South Korea).

Contig assembly and multiple sequence alignment—The following bioinformatic methods were conducted for both data sets.

FastQC v. 0.11.7 (Andrews, 2010) was used to assess the quality of Illumina raw reads from the bait-enriched samples. The raw sequencing reads were then trimmed with Trimmomatic v.0.36 (Bolger et al., 2014) using the settings LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:36 to remove adapter sequences and portions of low quality. The HybPiper pipeline v.3 (Johnson et al., 2016) was used with default settings to process the quality-checked reads and recover the coding sequences for each locus. Outgroup sequences from the OneKP project (Wickett et al., 2014) were

added to each data set. Paired reads of samples enriched with the Angiosperms353 baits and the Ochnaceae baits were mapped to targets using BLASTx option (Altschul et al., 1990) and their respective amino acid target file. The sequences obtained from the BLASTx option were used for subsequent analysis because it was found to recover longer sequences. Mapped reads were then assembled into contigs with SPAdes v3.13.1 (Bankevich et al., 2012), and the `retrieve_sequences.py` script from the HybPiper suite was used with the `.aa` flag to produce outputs of a single sequence per gene, which is selected using length, similarity, and coverage. HybPiper flags potential paralogs when multiple contigs are discovered mapping well to a single reference sequence. All loci flagged as potential paralogs were removed from downstream analyses. Subsequent analyses were performed using exon-only data. Sequence recovery for both data sets is listed in Appendix S2. The percentage of gene recovery was calculated using the sum of the captured length per genes per individual divided by the sum of the mean length of all loci.

MAFFT v. 7.305b (Katoh et al., 2002) was used to align individual genes using the `-auto` flag. AMAS (Borowiec, 2016) was used to produce summary statistics for each alignment, evaluating the amount of missing data and the number of parsimony informative sites (Appendices S3 and S4).

Phylogenetic inference

Both assembled data sets were individually analyzed using the following approaches. An additional data set was generated by combining the genes from both probe kits. The two target files were tested for gene overlap using BLASTx. Duplicate genes (7 genes) were removed, and all other recovered genes from both data sets were combined resulting in 620 individual loci. Where two species were available for a genus, the species with higher gene recovery from its respective probe kit was selected to represent the genus.

Multispecies-coalescent (MSC) approach—The aligned exons were then used to infer individual maximum likelihood gene trees with IQTREE v.2.0 (Nguyen et al., 2015) with 1000 ultrafast bootstraps using the `-bb` option. Species trees were then inferred from the gene trees using ASTRAL-III v5.5.11 (Zhang et al., 2018) with the `-t 2` option providing full annotation outputs, including quartet support to allow visualization of the main topology, and first and second alternative as pie charts on the phylogenetic tree reconstruction.

Concatenation approach—An additional analysis was performed by concatenating exon alignments using AMAS for all loci. A species tree was generated from the concatenated exon alignments using IQTREE v.2.0, and then two measures of genealogical concordance were also calculated for each data set; gene concordance factor (gCF) and site concordance factor (sCF) using the options `-gcf` and `-scf` in IQTREE v.2.0 (Nguyen et al., 2015). The gCF and sCF values represent the percentage of gene trees containing that branch, and the number of alignment sites supporting that branch, respectively.

Phylogenetic informativeness and missing data

The phylogenetic informativeness of recovered loci was investigated using PhyDesign (López-Giráldez and Townsend, 2011; website: <http://phydesign.townsend.yale.edu/>). The program enables a quantitative measure of the strength of a gene in resolving topological relationships and branching order within a phylogenetic tree

(Townsend, 2007). For the analysis, a phylogenetic tree from the concatenation approach was made ultrametric using R v. 3.6.3 (R Core Team, 2020) with the `force.ultrametric` function in the package `phytools` v. 0.6-44 (Revell, 2012) and uploaded to `PhyDesign` along with a partitioned alignment of all genes. The rates of evolution were calculated using the `Rate4Site` algorithmic tool for identifying functional genes in proteins (Pupko et al., 2002). An initial analysis was conducted to identify genes with high substitution rates or with high phylogenetic noise. The presence of sharp “phantom” peaks close to present time suggests there are loci that may be containing more phylogenetic noise accounting for errors in the phylogenetic inference. Genes that were particularly “noisy” were profiled by `PhyDesign` with sharp ghost bandings close to time 0.

The impact of missing data was investigated at three thresholds, that is sites (single sequence positions) with more than or equal to 70%, 50%, and 30% missing data were omitted from the alignment using a custom python script. The filtered alignment was submitted to `PhyDesign` again to test for phylogenetic informativeness and phylogenetic noise. Subsequently, each filtered alignment was used for phylogenetic inference with the two methods outlined above.

RESULTS

Sequence recovery and data quality

The accessions enriched with the Ochnaceae-specific kit yielded higher gene recovery with an average of 62.4% capture success compared to the Angiosperms353 kit with an average of 40.9% capture success (Fig. 1A; Appendix S2). Enrichment efficiency was higher for the Ochnaceae-specific kit than the Angiosperms353 kit with 52% and 8%, respectively, accounting for number of genes recovered. Table 1 shows accessions enriched with the Ochnaceae-specific kit, which

recovered an average of 238 genes of a possible 275 genes, whereas the accessions enriched with the Angiosperms353 kit recovered an average of 224 genes of 353 genes. Heat maps (Appendix S5) showed higher gene recovery for the Ochnaceae-specific kit, with a higher number of genes and longer length per gene. When considering taxonomic bias, it is also notable that *Ochna* and closely related genera on average have higher gene recovery in the Ochnaceae-specific data set, which is consistent with the development of the probes of the Ochnaceae-specific kit from the *Ochna serrulata* transcriptome. However, more distantly related genera also showed high gene recovery. The Angiosperms353 kit recovered genes randomly across the family, and as expected due to its universal applicability also higher recovery of the outgroup taxa compared to the Ochnaceae-specific kit.

Summary statistics for the alignments obtained with each probe kit are provided in Table 1 and Appendices S3 and S4. Before data filtering, the Ochnaceae-specific data set had a higher proportion for both parsimony informative sites (PIS) and variable sites (VS), 27.6% and 57.9% respectively, compared to the Angiosperms353 data set with 17.9% PIS and 39.3% VS. The Angiosperms353 data set showed a positive correlation between the alignment length and the number of PIS (Fig. 1B), with an average alignment length of 319 bp. The Ochnaceae-specific data set showed a huge variability in PIS, irrespective of the alignment length, which was 540 bp on average. Additionally, the percentage of missing data showed the Angiosperms353 data set on average had slightly higher proportions of missing data with 41.5% compared to the Ochnaceae-specific data set with 36.3%.

Impact of missing data

Missing data had varying impacts on the phylogenetic inference based on the data sets obtained with the two probe kits. For the

data sets obtained with both probe kits, the proportion of VS steadily increased when removing more sites with missing data. The total number of loci was significantly reduced upon removal of missing data for the Angiosperms353 probe kit but was inconsequential for the Ochnaceae-specific kit. When considering the nodes of interest corresponding to the hitherto unclear phylogenetic relationships, including the polyphyly of *Sauvagesia* and *Campylospermum*, the relationships amongst genera of Ochninae, and the relationship of Medusagynoideae with the rest of the family, Fig. 2 shows the support of each node of interest retrieved under various scenarios. Incongruence between the two inference methods was highest for the Angiosperms353 data set; for this data set, the concatenation approach generated more consistent results than the MSC approach under varying filtering strategies and was more consistent with the topology obtained with the Ochnaceae-specific kit using both approaches. The MSC

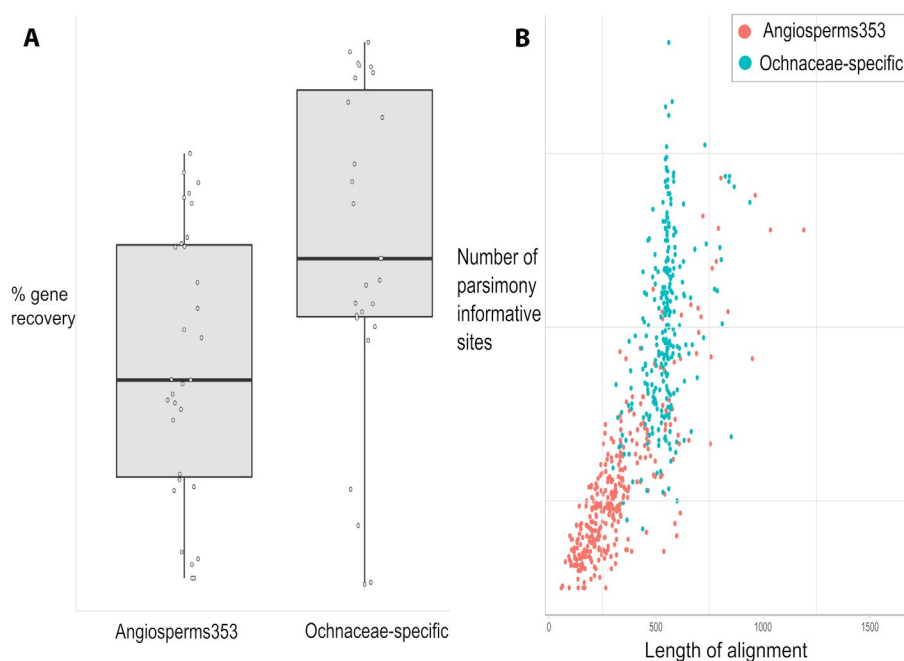


FIGURE 1. (A) Box plot showing gene recovery success per probe kit with the respective enriched samples. (B) Scatter plot of the relationship between length of alignment and the number of parsimony informative sites for each locus and probe kit.

TABLE 1. Summary statistics for alignments of all recovered loci showing values for number of genes, length of alignments, proportion of variable and parsimony informative sites, and percentage of missing data for each probe kit.

	Angiosperms353 kit				Ochnaceae-specific kit				Combined kits total
	Unfiltered	70%	50%	30%	Unfiltered	70%	50%	30%	—
Filtering threshold	Unfiltered	70%	50%	30%	Unfiltered	70%	50%	30%	—
Number of genes in kit	353				275				628
Average number of genes recovered	224				238				—
Number of genes retained in data set	350	211	211	210	275	275	275	274	620
Length of alignment (bp)	111,766	49,008	39,226	27,909	149,192	125,313	100,610	77,061	256,614
Mean length of alignment per locus	319				543				
Range length of alignment per locus (min-max) (bp)	58 – 1587				300 – 939				
Proportion of variable sites (%)	39.3	44.1	43.5	43.1	57.9	63.6	64.0	63.3	48.7
Proportion of parsimony informative sites (%)	17.9	22.6	22.6	22.7	27.6	32.0	34.8	35.7	21.6
Mean percentage of missing data per locus (%)	41.5	26.5	18.7	11.35	36.3	27.3	17.9	11.5	40.6

approach for the Angiosperms353 probe kit generated conflicting topological results under different filtering thresholds with poorer support. Taxonomic relationships for the concatenation method were congruent under all the filtering strategies for the Angiosperms353 data set, although the relationship of *Ochna* as sister to *Brackenridgea*, *Campylospermum* clade A, and *Idertia* received poor support. The MSC inference was conflicting when resolving the sister relationship to *Ochna*. For the Ochnaceae-specific data set, there was no difference for the nodes of interest in support or topology under any filtering scenarios, and strong support was retrieved for all nodes under both approaches. The increased number of genes through combining the data sets obtained with both kits resulted in a phylogenetic hypothesis with strong support for all nodes consistent with the topology obtained with the Ochnaceae-specific data set.

Phylogenetic relationships in Ochnaceae

In the subsequent sections, we refer to the Angiosperms353 data set with a relaxed threshold of 70% and the unfiltered Ochnaceae-specific data set for the description of the phylogenetic relationships because they provide the most robust topologies with strong support.

The topologies resulting from the two phylogenetic reconstruction approaches are largely congruent between the two data sets (Fig. 3A, B). Both data sets resolved Quinoideae and Medusagynoideae as sister to each other with 100% BS. Both subfamilies together were resolved as sister to Ochnoideae with 100% BS and gene concordance greater than 50%. Furthermore, the concatenation approach resulted in well-resolved topologies at the tribal and subtribal levels as currently circumscribed.

Probe Kit	Angiosperms353								Ochnaceae-specific								Combined	
	Concatenated-Supermatrix				Multispecies-coalescent (MSC)				Concatenated-Supermatrix				Multispecies-coalescent (MSC)				Concat.	MSC
Alignment Threshold of missing data removed	Unfiltered	≥70%	≥50%	≥30%	Unfiltered	≥70%	≥50%	≥30%	Unfiltered	≥70%	≥50%	≥30%	Unfiltered	≥70%	≥50%	≥30%	Unfiltered	
Medusagynoideae sister to Quinoideae	Green	Green	Green	Green	Gray	Yellow	Gray	Yellow	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green
Medusagynoideae sister to rest of Ochnaceae	Gray	Gray	Gray	Gray	Green	Green	Green	Green	Gray	Gray	Gray	Gray	Gray	Gray	Gray	Gray	Gray	Gray
Polyphyly of <i>Sauvagesia</i>	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green
Polyphyly of <i>Campylospermum</i>	Yellow	Green	Green	Green	Green	Yellow	Yellow	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green
<i>Ochna</i> sister to <i>Idertia</i> only	Gray	Gray	Gray	Gray	Yellow	Gray	Gray	Yellow	Gray	Gray	Gray	Gray	Gray	Gray	Gray	Gray	Gray	Gray
<i>Ochna</i> sister to <i>Brackenridgea</i> , <i>Campylospermum</i> I	Gray	Gray	Gray	Gray	Gray	Gray	Gray	Gray	Gray	Gray	Gray	Gray	Gray	Gray	Gray	Gray	Gray	Gray
<i>Ochna</i> sister to <i>Brackenridgea</i> , <i>Campylospermum</i> I, <i>Idertia</i> , <i>Rhabdophyllum</i> , <i>Campylospermum</i> II	Gray	Gray	Gray	Gray	Gray	Gray	Gray	Gray	Gray	Gray	Gray	Gray	Gray	Gray	Gray	Gray	Gray	Gray
<i>Ochna</i> sister to <i>Brackenridgea</i> , <i>Campylospermum</i> I, <i>Idertia</i>	Green	Yellow	Yellow	Yellow	Gray	Yellow	Gray	Gray	Green	Green	Green	Green	Green	Green	Green	Green	Green	Green

FIGURE 2. Color chart with support values. Green: nodes resolved with strong support, ≥70% BS or ≥0.75 LPP; yellow: nodes with poor support, <70% BS or <0.75 LPP, gray: unretrieved topology for taxonomic relationships and nodes of interest.

The topological MSC results for both data sets (Fig. 3C, D) were congruent with those of the concatenation approach. The data differed slightly in support, where Quiinoideae was inferred as sister to Medusagynoideae with strong support (local posterior probability [0.99 LPP]) for the Ochnaceae-specific data set, but poor support (0.45 LPP) for the Angiosperms353 data set. Both data sets inferred the two subfamilies as sister to Ochnoideae with strong support. The topology resulting from the Angiosperms353 data set resolved all tribes and subtribes as currently circumscribed with mixed LPP support. The Ochnaceae-specific data set largely placed all tribes and subtribes as previously circumscribed by Schneider et al. (2014), with the exception of *Philacra auriculata* Dwyer, which was placed here in the tribe Ochnaeae with moderate support (0.7 LPP), as sister to the rest of Ochninae. This unusual placement of the genus was not resolved by removing sites of missing data. Backbone nodes received stronger support in both reconstruction methods with the Angiosperms353 compared to the Ochnaceae-specific kit. For the concatenation approach, the tribe Ochnaeae as sister to Sauvagesieae was resolved with 98% BS for the Angiosperms353 kit compared to 76% BS for the Ochnaceae-specific kit.

Relationships within the subtribe Ochninae

The relationships of genera within subtribe Ochninae in all data sets generated through both methods, showed equally high BS and LPP support for most branches (Figs. 2, 3). The sister relationship to *Ochna* was congruent amongst all inferences, i.e., resolving a monophyletic *Ochna* as sister to a clade comprising *Brackenridgea*, *Campylospermum* (Clade B) and *Idertia*. The Ochnaceae-specific data set resolved this relationship with strong support for the concatenation and MSC approach (BS 100%, 0.99 LPP), respectively (Fig. 3B, D). The Angiosperms353 data set, inferred the same topology although with poor support for both approaches (concatenation, BS 49%; MSC, LPP 0.57; Fig. 3A, C). All analyses placed *Ouratea* as sister to the remainder of the subtribe. The Angiosperms353 coalescence-based inference is the only analysis placing *Campylospermum reticulatum* (P.Beauv.) Farron on a separate branch as sister to the rest of Ochninae but with poor support of LPP 0.51 (Fig. 3C).

Corroborating polyphyletic genera

All our analyses confirm the polyphyly of *Sauvagesia*. Both data sets and both methods resolve the genus as two unrelated clades with high support and strong gene tree agreement. *Sauvagesia serrata* is resolved as sister to *Schuermansia henningsii* K.Schum., while *Sauvagesia erecta* L. is sister to *Adenarake muriculata* Maguire & Wurdack. Both clades are placed in tribe Sauvagesieae (Fig. 3). The polyphyly of *Campylospermum* is also confirmed by all data sets, with both lineages of *Campylospermum* placed in subtribe Ochninae. All trees (Fig. 3) resolved *Campylospermum elongatum* (Oliv.) Tiegh. in a clade with *Brackenridgea* and *Idertia*, sister to *Ochna* with strong support with the exception of the Angiosperms353 concatenation result in which it is poorly supported (BS 49%). The second lineage, *Campylospermum reticulatum*, is placed in a separate clade as sister to *Rhabdophyllum calophyllum* (Hook.f.) Tiegh. in all data sets and approaches except the MSC approach based on the Angiosperms353 data set where *Campylospermum reticulatum* was found as sister to the remainder of subtribe Ochninae with low LPP of 0.51.

Phylogenetic informativeness and data filtering

Maximum net phylogenetic informativeness (NPI) ranged from 11.06 to 1041.7 for the unfiltered Angiosperms353 data set, with most genes reaching their peak informativeness between time interval 0.06 and 0.12 (Fig. 4). Most genes reached their peak before the divergence of tribes and subtribes; however, several genes show ghost bandings (sharp peaks) close to time 0. After removing sites with missing data at 70% or more, phylogenetic noise was reduced, significantly removing the ghost bandings toward the present, with maximum NPI ranging from 11.0 to 541.7 in the filtered data sets. The relationship of Medusagynoideae as sister to Quiinoideae, together sister to Ochnoideae obtained in the unfiltered data set incurred no change in support or topology after removing missing data. Likewise, the genus *Sauvagesia* remained polyphyletic with strong support even after filtering of missing data. The genus *Campylospermum* also remained polyphyletic, but with increased support from BS 67% to 100% after removing missing data.

Phylogenetic informativeness was investigated in the Ochnaceae-specific data set as well to evaluate the impact of missing data before and after data filtering at 70% (Fig. 5A, B). Maximum NPI ranged from 63.2 to 483.5 before data filtering. Most genes reached a peak NPI between time interval 0.08 and 0.16. Filtering the data at a threshold of 70% did not impact the maximum NPI. Ghost peaks for the Ochnaceae-specific data set were more significant without data filtering and were completely omitted after removing missing data. Phylogenetic relationships were not impacted by removing missing data. Only one node of interest, highlighting the polyphyly of *Sauvagesia* received a decrease in BS from 100% to 73% (Fig. 5).

A combined tree

Recovered genes from both data sets were combined to form a larger data set. The proportion of missing data and PIS of the combined data set was 40.6% and 21.6%, respectively, similar to that of the unfiltered data set of the Angiosperms353 probe kit of 41.5% and 17.9%, respectively (Table 1). Phylogenetic inference using a combined data set under both methods resolved strong support for all nodes of interest (Fig. 2), with congruent topologies to that of the Ochnaceae-specific data set under all filtering strategies, and to the Angiosperms353 data set under 70% filtering threshold for both inference methods (Fig. 6).

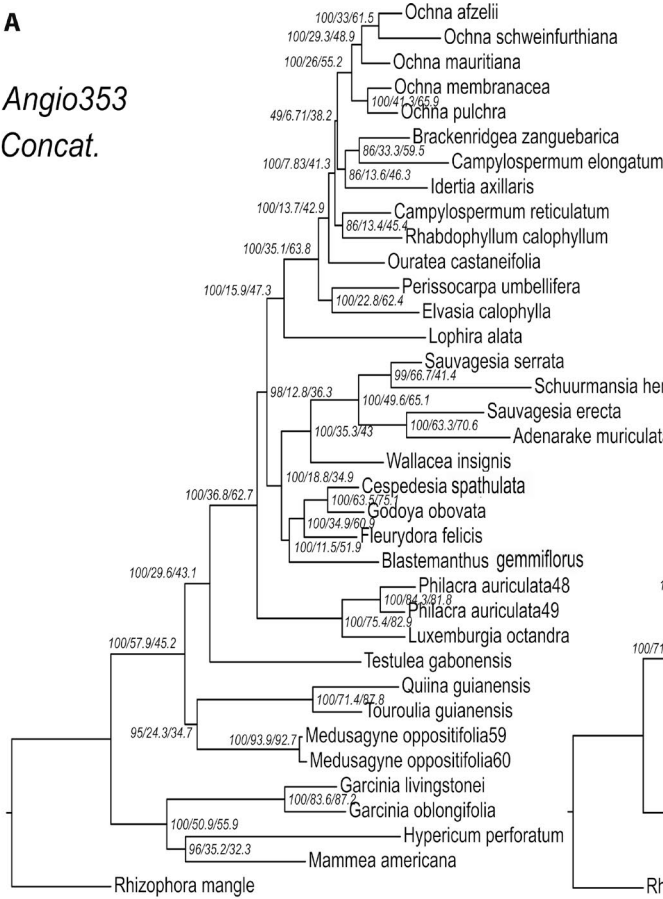
DISCUSSION

Angiosperms353 vs. Ochnaceae-specific probe kit

The Ochnaceae-specific kit recovered a substantially greater number of loci for *Ochna* and the other genera of subtribe Ochninae than the Angiosperms353 kit; this difference was smaller for more distantly related genera (Appendix S5). This result was expected due to the Ochnaceae-specific kit being developed using a transcriptome of *Ochna serrulata*, which should essentially share more loci with closely related genera. Although the results suggest there is some taxonomic bias with the Ochnaceae-specific kit toward *Ochna* and closely related genera, some more distantly related genera do have moderately high gene recovery, perhaps pertaining to sample quality rather than taxonomic affinity (Schneider et al., 2020). Efficiency of family-specific probe kits can be enhanced by

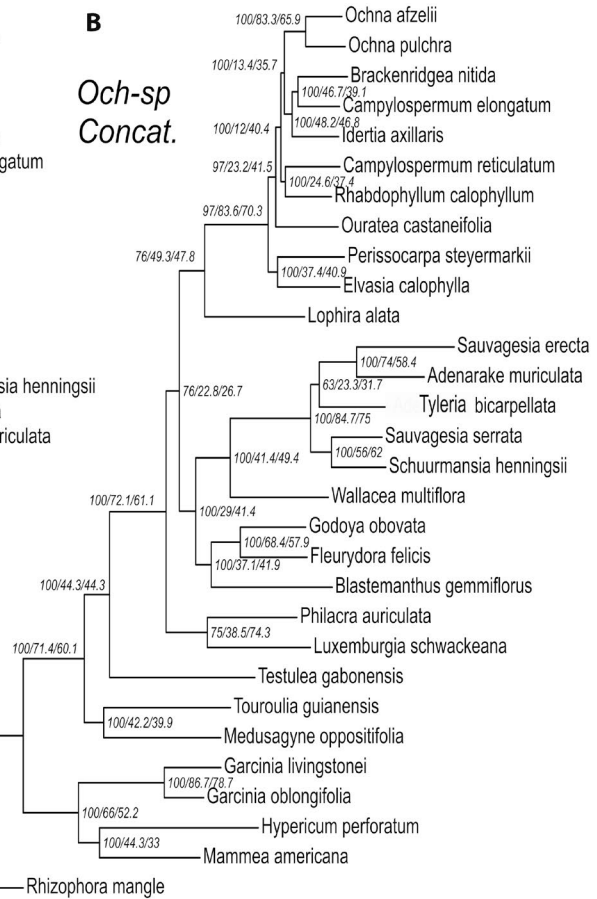
A

Angio353
Concat.



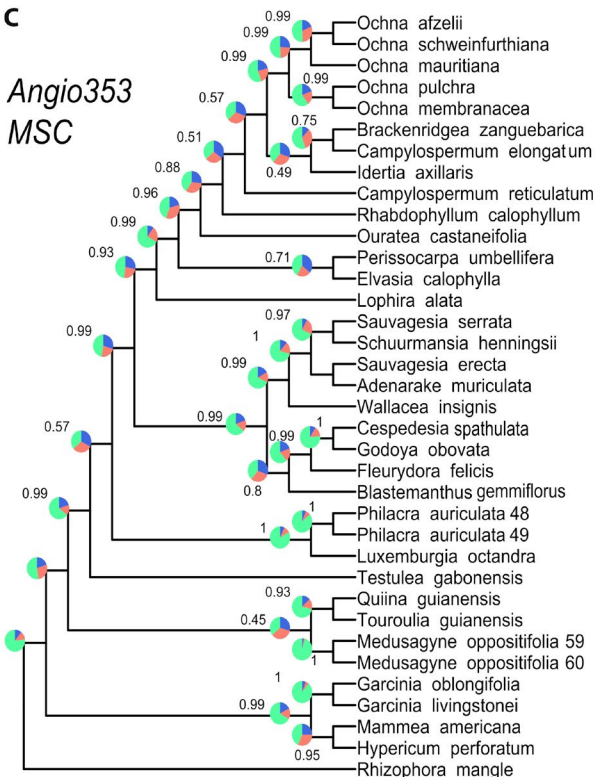
B

Och-sp
Concat.



C

Angio353
MSC



D

Och-sp
MSC

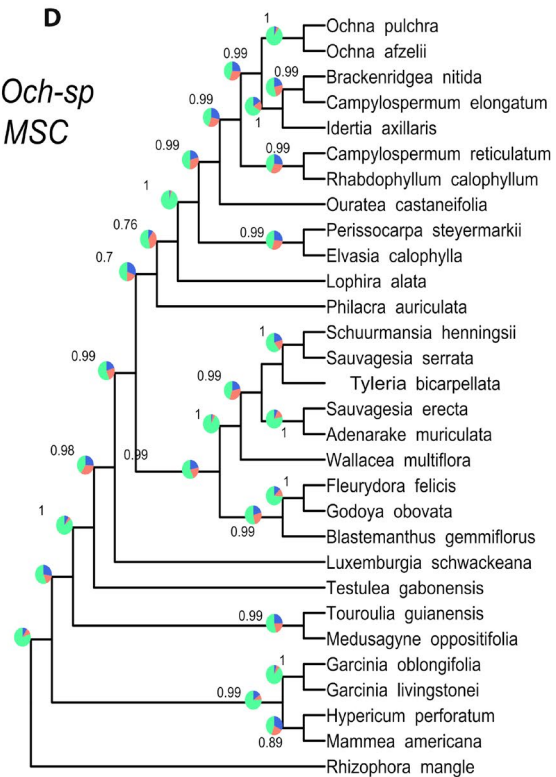


FIGURE 3. Phylogenetic reconstructions of Ochnaceae using concatenation (Concat.) data with IQTREE (A, B) and multispecies coalescence (MSC) approaches with ASTRAL-III (C, D) with the Angiosperms353 data with missing data filtered at a 70% threshold (A, C) and the unfiltered Ochnaceae-specific data (B, D). Values at nodes in A and B are bootstrap support derived from IQTREE, gene concordance factor (gCF), and site concordance factor (sCF). Pie charts indicate quartet support (green = number of gene trees that support this topology, blue = number of gene trees that support an alternative topology, red = number of gene trees that are non-informative). Numbers next to species epithets indicate multiple accessions for the same species.

expanding the genomic resources used in the development of the kit. Including transcriptome data spanning all the family's major clades would reduce taxonomic bias and promote gene recovery more widely shared across different clades. On the other hand, the Angiosperms353 kit was able to recover loci more randomly across the family, also an expected outcome since the kit was developed using orthologous genes from across angiosperms (Johnson et al., 2019). Outgroup taxa were also far better recovered using the Angiosperms353 kit, which would be expected given that the probes of the Ochnaceae kit are specific for this family. This result was consistent with the findings of Chau et al. (2018) who also compared targeted sequencing kits and found that the general locus set was more successful with outgroup taxa than the specific locus set.

We acknowledge the limitation of sampling differences between the each probe kit in our study. Additional factors not investigated in the present study include sample age, drying conditions, specimen preservation, and taxonomic affinity, which could also account for differences in recovery success (Brewer et al., 2019). Despite the limitations arising from using samples of different quantity and quality, the samples selected for each data set consisted of comparable representatives for each major clade in the family, including the same genera, and where possible the same species. Thus, we are confident that the observed differences in gene recovery adequately reflect the differences between the two kits. We can compare the capability of each probe kit in resolving some recalcitrant nodes in Ochnaceae by testing the informativeness of each kit and the general effects of missing data.

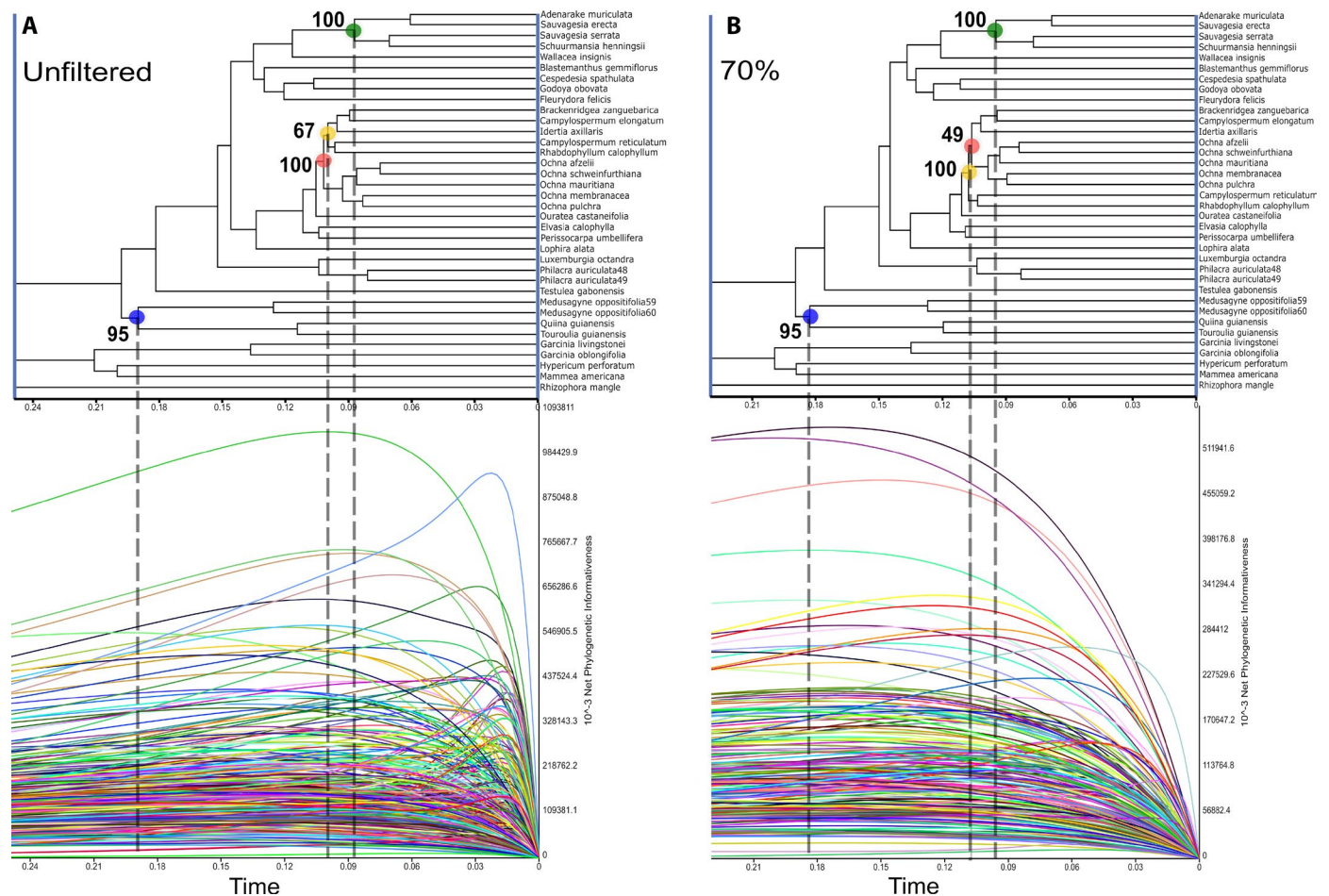


FIGURE 4. Net phylogenetic informativeness of the Angiosperms353 data through an arbitrary time scale with colored lines representing individual gene profiles: (A) profiles from the unfiltered data set, (B) profiles from the filtered data set where the sites with missing data $\geq 70\%$ were removed from all loci. Nodes of interest are highlighted by colored dots, indicating changes in support or topology after data filtering. Blue and green dots show no change. Yellow indicates increase in support, and red indicates a decrease in support and change in topology.

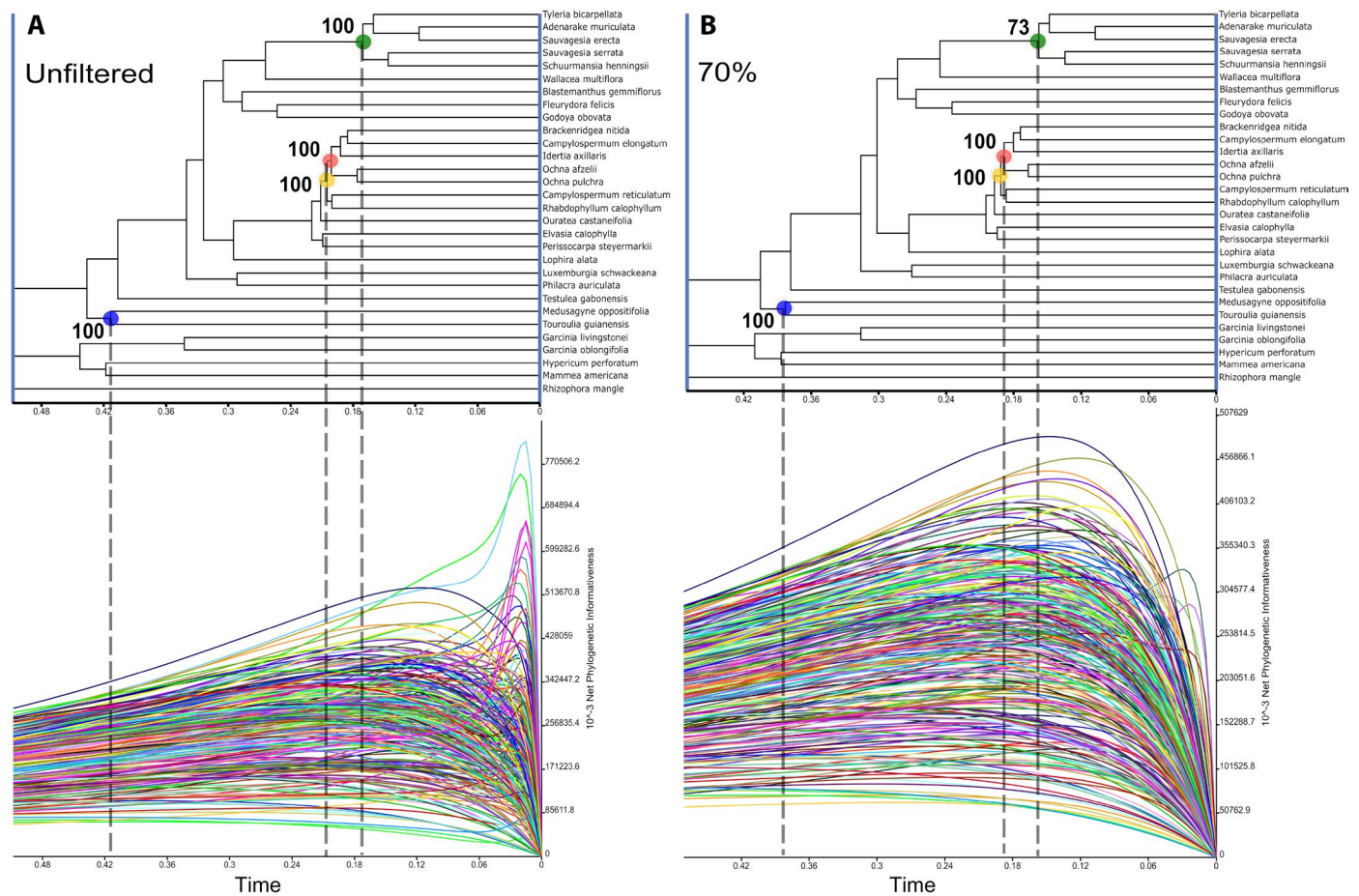


FIGURE 5. Net phylogenetic informativeness of the Ochnaceae-specific data through an arbitrary time scale with colored lines representing individual gene profiles: (A) profiles from the unfiltered data set, (B) profiles from the filtered data set where the sites with missing data $\geq 70\%$ were removed from all loci. Nodes of interest are highlighted by colored dots, indicating changes in support or topology after data filtering. Only the green dot shows decrease in support after data filtering. All other nodes of interest remained unchanged.

We found the Ochnaceae-specific kit recovered a higher number of genes and yielded a higher proportion of PIS and VS than the Angiosperms353 kit did (Fig. 1, Table 1). This outcome was anticipated because universal kits are likely to hold more conserved loci across wider phylogenetic distances. A recent study by Larridon et al. (2020) similarly found that the recovery of loci with the Cyperaceae-specific kit was higher than with the Angiosperms353 kit. On the other hand, they found the total number of PIS and VS was higher for the Cyperaceae-specific kit, although differed from our study when accounting for the size of the data set, as the relative proportions of PIS were similar between the kits. In this study, the unfiltered data set of the Ochnaceae-specific kit resulted in a more strongly supported phylogenetic tree compared to unfiltered data from the Angiosperms353 kit. These results are consistent with that of Kadlec et al. (2017) who found their custom-designed kit for the genus *Erica* Tourn. ex L. outperformed the universal approach; however, unlike their conclusions that a “made to measure” approach is superior to a universal one, we argue that a custom probe kit and universal kit are as effective in resolving phylogenetic relationships. With careful data filtering, the universal Angiosperms353 kit can effectively resolve phylogenetic relationships, many of which are congruent with results obtained with a family-specific kit. This finding again, agrees with the conclusions of Chau et al. (2018), showing that universal kits

can be as effective as taxon-specific kits regarding phylogenetic informativeness. The Angiosperms353 kit has the added advantage of being applicable across angiosperms, aiding outgroup inclusion and broadening applicability of the data beyond lineage-specific studies. Furthermore, we show that when both kits are used in combination, the increased amount of data improves phylogenetic resolution.

Phylogenetic informativeness and missing data

To our knowledge, no previous work has compared the impact of missing data on the amount of “noisy” loci variability between a universal and custom probe kit. In our study, as the two probe kits had minimal gene overlap (only seven genes), and a difference in overall gene recovery, we wanted to explore the phylogenetic informativeness of each kit independently. The presence of sharp “phantom” peaks close to present time in both probe kits suggests there are genes that may be responsible for phylogenetic noise more than others, potentially accounting for some unexpected relationships observed in the phylogenetic inference. Interestingly, the Ochnaceae-specific data showed more exaggerated “phantom” peaks than the Angiosperms353 data did, suggesting a higher prevalence of phylogenetic noise in the former. The Ochnaceae-specific kit obtained peak NPI close to present time, compared with deeper

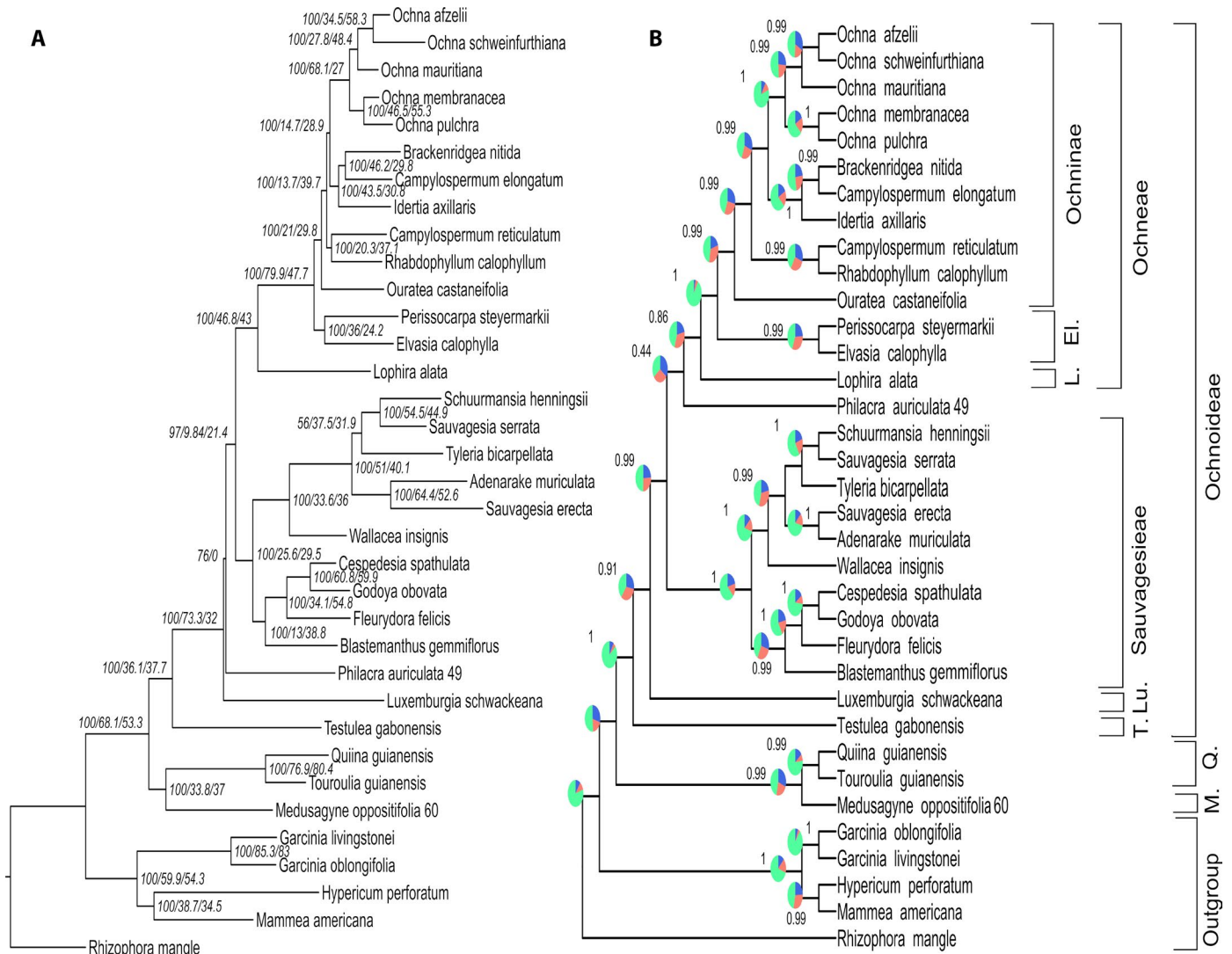


FIGURE 6. Phylogenetic reconstruction for Ochnaceae using the combined data set from both probe kits. (A) Concatenation approach with maximum likelihood-based inference with three values at each node; bootstrap support derived from IQTREE, gene concordance factor (gCF) and site concordance factor (sCF). (B) Multispecies coalescence approach with ASTRAL-III with pie charts indicating quartet support (green = number of gene trees that support this topology, blue = number of gene trees that support an alternative topology, red = number of gene trees that are non-informative). Numbers next to species epithets indicate multiple accessions for the same species. The ranks and suprageneric names are presented here; El. = Elvasiinae; L. = Lophirinae; Lu. = Luxemburgieae; M. = Medusagynoideae; T. = Testuleae; Q. = Quiinoideae.

nodes in the Angiosperms353 kit probe kit, potentially caused by the saturation of substitutions (Goremykin et al., 2010; Fragoso-Martínez et al., 2017). Despite having more exaggerated “phantom” peaks, the inference obtained with the Ochnaceae-specific kit was not impacted by removing sites of missing data, and there was negligible difference on the tree topologies and support. This outcome may be because even after removing missing data at all thresholds, the data set retained a significantly high number of loci with a high proportion of PIS (Table 1), therefore showing that missing data were not influencing the efficiency of the kit. The only oddity was the placement of *Philacra* within Ochneae in the coalescent topology. This unexpected placement is inconsistent with morphological data and in conflict with all other topologies, which place this genus in Luxemburgieae, suggesting that it is likely an artefact of low gene recovery for this taxon.

The removal of “noisy” loci improves phylogenetic inference. Additional ways to improve resolution in the phylogenetic tree and reduce phylogenetic noise may be to increase taxon sampling. Expanded taxonomic sampling can modify the alignment and in turn the branch lengths of the inferred phylogenetic tree. As a result, adding new taxa allows for different estimates of substitution rates of sites in the alignment that were poorly estimated before (López-Giráldez and Townsend, 2011; Fragoso-Martínez et al., 2017). However, it is important to also add taxa that represent nodes deeper than the clades and relationships of interest (López-Giráldez and Townsend, 2011).

The Angiosperms353 data set only had slightly more missing data than the Ochnaceae-specific data set (Table 1), and the removal of missing data at thresholds of 30%, 50% and 70%, respectively, showed varying effects on phylogenetic informativeness and

inference with both data sets. The Angiosperms353 data was more significantly influenced by filtering of missing data. For this data set, the coalescent-based approach was less consistent amongst different filtering thresholds, such as, for example, the relationship of Medusagynoideae relative to the other subfamilies and relationships in the subtribe Ochninae (Fig. 2). One reason for this may be that maximum phylogenetic informativeness is achieved at nodes more recent than the subfamily divergence, leaving that node with lower phylogenetic signal and more sensitive to phylogenetic noise (Townsend and Leuenberger, 2011). Furthermore, deep divergences such as the relationships of subfamilies are more difficult to resolve (King and Rokas, 2017) and potentially more sensitive to effects arising from missing data. This inconsistency in the coalescent approach aligns with the results of Hosner et al. (2016) who compared effects of missing data amongst different inference methods. They found that gene tree inference produced unexpected results with taxa placed in unusual positions. Gene tree conflict may explain the inconsistency in the placement of Medusagynoideae. A potential solution to this problem could be to use the most informative genes for gene tree estimation and species tree inference (Hosner et al., 2016). In general, it is less common to have a conflicting topology derived through coalescence, which theoretically should be able to better accommodate different gene histories (Heled and Drummond, 2010; Wu, 2012; Xi et al., 2014). A majority of studies have focused on missing data using a concatenation approach (Crawley and Hilu, 2012; Roure et al., 2013; Jiang et al., 2014; Hosner et al., 2016); however, in an age of genomic data sets, the use of MSC phylogenetic reconstruction is growing. Therefore, the discrepancy from the MSC result with the Angiosperms353 data provides an important insight into understanding the impact of missing data on MSC phylogenetic reconstruction and whether including or excluding more genes with missing data will impact overall the phylogenetic relationships.

Generally, for the Angiosperms353 data, removing missing data at a relaxed threshold of 70% increased support in both inference methods, particularly for the nodes of interest. The increase in support and the reduction of “phantom” peaks after removing sites of missing data shows that the sites where missing data were primarily found, were indeed responsible for increased phylogenetic noise. We chose to remove sites with missing data instead of removing entire genes to preserve as much information as possible. In some cases, entire genes were removed in the Angiosperms353 data set due to extremely low recovery in the initial alignment, with more missing data than the allocated thresholds. We found that filtering at more stringent thresholds of 30% and 50% generally reduced support across the trees. This finding suggests that despite removing missing data, there may be a trade-off between losing valuable phylogenetic information versus reducing phylogenetic noise.

Toward a robust phylogenetic framework

Conflicting evolutionary histories within Ochnaceae in recent studies and persistence of some uncertain relationships even with large taxon sampling and huge amounts of DNA sequence data has stimulated our investigation into the use of different probe kits in resolving recalcitrant clades and how species tree reconstruction methods and the impact of missing data influence phylogenetic resolution. Adding more genes such as those obtained with the Angiosperms353 kit (i.e., added to the family-specific data set of

Schneider et al. [2020]) is likely to improve phylogenetic resolution, although there is evidence that increasing data can also result in diminishing returns (Hosner et al., 2016). In Ochnaceae, the novel Angiosperms353 data set was primarily valuable in corroborating relationships obtained by other studies. Topologies from the different data sets were similar to the results of Schneider et al. (2020). Although our study is less taxon-rich, an important finding was the resolution of the subfamilial relationships with maximum support. In earlier studies, these relationships remained unclear or received low support (Xi et al., 2012; Schneider et al., 2014, 2020). Both kits resolved congruent relationships within Sauvagesieae. For example, *Blastemanthus* Planch. was strongly resolved as sister to a clade with *Fleurydora* A.Chev., *Cespedesia* Goudot, and *Godoya* Ruiz. & Pav., contrasting with the results of Schneider et al. (2014) who retrieved a polytomy for this clade, but agreeing with those of Schneider et al. (2020). Furthermore, in the present study, *Wallacea* Spruce ex Benth. & Hook.f. is placed as sister to the clade uniting *Adenarake* Maguire & Wurdack, *Sauvagesia*, and *Schuermansia* Blume, different to previous work placing *Fleurydora* as sister to *Wallacea* (Schneider et al., 2014). Additionally, in agreement with the results of Schneider et al. (2014, 2020) and Bissengou (2014), all phylogenetic reconstructions resolved both *Sauvagesia* and *Campylosporum* as polyphyletic. Our results show that *Sauvagesia serrata* (= *Neckia serrata*) from Asia is separated from core *Sauvagesia*, which is mainly neotropical and resolved as sister to *Tyleria* Gleason. Similarly, the polyphyly of *Campylosporum* is also supported by all the topologies. The west/central African species *Campylosporum elongatum* (Clade B of Schneider et al., 2020) is placed sister to *Brackenridgea*, which together are sister to *Idertia*. It is separate from *Campylosporum reticulatum* (Clade A of Schneider et al., 2020) which is resolved as sister to *Rhabdophyllum*. Although there is some support from morphology for the separation of the two clades of *Campylosporum* (Bissengou, 2014; Schneider et al., 2020), a more detailed assessment with a broader taxon sampling is required before a formal re-circumscription of this genus.

Overall, backbone nodes resolved by both approaches were better supported with the Angiosperms353 kit compared to the Ochnaceae-specific kit. On the other hand, more recent nodes in the subtribe Ochninae were more strongly supported resolved using the MSC approach with the Ochnaceae-specific kit. Thus, with the aim of finding a consensus topology, whilst mitigating as much taxonomic bias arising from each probe kit, we merged the data from both probe kits. Thereby, we obtained a strongly supported topology that was consistent with the results from the Ochnaceae-specific data set. It is well known from several studies (Pollock et al., 2002; Hedtke et al., 2006; Jantzen et al., 2019) that an increase in taxon sampling can greatly improve phylogenetic resolution. Fewer studies have shown that an increase in the number of characters whilst maintaining some missing data is equally advantageous in producing strongly supported topologies under different inference methods. Our results align with findings from other studies (Rubin et al., 2012; Wagner et al., 2013; Jiang et al., 2014; Huang and Lacey Knowles, 2016) that also concluded that including genes with a degree of missing data is more beneficial for estimating robust phylogenetic relationships than completely removing missing data. Our combined data set retrieved nodes throughout the phylogenetic tree as well resolved, which may be due to the Angiosperms353 data set providing resolution for more conserved sites with the ability to resolve relationships among deeper nodes, whilst the Ochnaceae-specific data set, allows inference of more recent radiations through the presence of more

rapidly evolving sites. Gene profiling can identify genes with high phylogenetic informativeness and minimal phylogenetic noise from each probe kit, which can be subsequently selected for future research for specific points in evolutionary history.

CONCLUSIONS

Data sets obtained with both probe kits resolved relationships with moderate to high confidence. The Angiosperms353 data set was more influenced by missing data and revealed more conflicts between the different approaches of phylogenetic inference than the Ochnaceae-specific data set. We found that data filtering is an important step in reducing phylogenetic noise. However, it is recommended that various thresholds of removal be tested to find the optimal threshold for a given data set. We found that for the Angiosperms353 data, using an overly stringent threshold for the removal of missing data (i.e., at 30% and 50%) may lead to a reduction of informative sites, causing a decrease in support and conflict in topology. Furthermore, we conclude that DNA sequence data obtained with the Angiosperms353 probe kit may not resolve relationships at shallow levels as successfully as the data obtained from the Ochnaceae-specific kit (e.g., in the subtribe Ochninae). We advise that when developing future taxon or family specific probe kits, the genes targeted by the Angiosperms353 probe kit be included in the specific probe kit to improve prospects of combining lineage-specific data sets across angiosperms. Alternatively, the Angiosperms353 probe set can be easily combined with a specific probe set at the hybridization step (Hendriks et al., 2021). Although the phylogenomic analyses based on the Angiosperms353 kit did not reveal novel relationships compared to those performed with the family-specific kit, they still provided important insights by corroborating relationships among the rapidly diverging Ochnineae as well as the three subfamilies that have long remained intractable.

ACKNOWLEDGMENTS

We thank Niroshini Epitawalage for her patience and clarity when providing laboratory guidance when generating the Angiosperms353 data set. The authors also thank Alexandre R. Zuntini, Sidonie Bellot, Rowan Schley, and Mathew Rees for advice and fruitful discussions on the bioinformatic analyses and use of scripts, as well as the two anonymous reviewers who provided constructive feedback that greatly helped improve the manuscript. This work was funded by grants from the Calleva Foundation and the Sackler Trust to the Plant and Fungal Tree of Life Project (PAFTOL) at the Royal Botanic Gardens, Kew, and from the Deutsche Forschungsgemeinschaft to G.Z. (ZI 557/14-1).

AUTHOR CONTRIBUTIONS

T.S. conducted the lab work for the Angiosperms353 data set, bioinformatics, phylogenetic inference, and drafted the manuscript. J.V.S. conducted the lab work for the Ochnaceae-specific data set. G.E.B. contributed to the lab work for the Angiosperms353 data set. J.V.S., G.Z., I.L., I.D., F.F., and W.B. contributed to the writing of the manuscript.

DATA AVAILABILITY

Raw sequence data for the Ochnaceae-specific probe kit are available from GenBank SRA under the Bioproject number PRJNA602196: <http://www.ncbi.nlm.nih.gov/bioproject/602196> (Schneider et al., 2020). Raw sequence data for the universal Angiosperms353 probe kit are available from European Nucleotide Archive under the umbrella project number PRJEB35285: <https://www.ebi.ac.uk/ena/browser/view/PRJEB35285>. Voucher information for all samples used are listed in Appendix 1. All the sequence alignments, gene trees and species trees for both probe kits and the combined dataset are available from the Dryad Digital Repository at <https://doi.org/10.5061/dryad.2547d7wsz>.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

APPENDIX S1. Table summarizing taxonomic representation across the family per probe kit.

APPENDIX S2. Table showing sample treatment and capture success for both probe kits.

APPENDIX S3. Summary statistics for the Angiosperms353 probe kit. Columns K–AL show the number of characters present for that gene.

APPENDIX S4. Summary statistics for the Ochnaceae-specific probe kit. Columns K–AL show the number of characters present for that gene.

APPENDIX S5. Proportion of total reference gene length recovered per gene and taxon for each probe kit.

LITERATURE CITED

- Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215: 403–410.
- Alvizu, A., M. H. Eilertsen, J. R. Xavier, and H. T. Rapp. 2018. Increased taxon sampling provides new insights into the phylogeny and evolution of the subclass Calcarea (Porifera, Calcarea). *Organisms Diversity and Evolution* 18: 279–290.
- Andrews, S. 2010. A quality control tool for high throughput sequence data. Website: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/> [Accessed 02 August 2020].
- Angiosperm Phylogeny Group. 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society* 181: 1–20.
- APG III. 2009. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Botanical Journal of the Linnean Society* 161: 105–121.
- Bankevich, A., S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* 19: 455–477.
- Barrett, C. F., C. D. Bacon, A. Antonelli, Á. Cano, and T. Hofmann. 2016. An introduction to plant phylogenomics with a focus on palms. *Botanical Journal of the Linnean Society* 182: 234–255.
- Bauckhage, C. 2015. NumPy / SciPy recipes for data science: *k*-medoids clustering machine learning. Technical report, University of Bonn, Bonn, Germany. <https://doi.org/10.13140/2.1.4453.2009>
- Bissengou, P. 2014. Systematics, evolution and historical biogeography of the family Ochnaceae with emphasis on the genus *Campylopermum*. Wageningen University, Wageningen, Netherlands.

- 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* 2016: e1660.
- Breinholt, J. W., S. B. Carey, G. P. Tiley, E. C. Davis, L. Endara, S. F. McDaniel, L. G. Neves, et al. 2020. A target enrichment probe set for resolving the flagellate plant tree of life. *bioRxiv*: 2020.05.29.124081 [Preprint].
- Brewer, G. E., J. J. Clarkson, O. Maurin, A. R. Zuntini, V. Barber, S. Bellot, N. Biggs, et al. 2019. Factors affecting targeted sequencing of 353 nuclear genes from herbarium specimens spanning the diversity of angiosperms. *Frontiers in Plant Science* 10: 1102.
- Buddenhagen, C., A. R. Lemmon, E. M. Lemmon, J. Bruhl, J. Cappa, W. L. Clement, M. Donoghue, et al. 2016. Anchored phylogenomics of angiosperms I: Assessing the robustness of phylogenetic estimates. *bioRxiv*: 086298 [Preprint].
- Chase, M. W., D. E. Soltis, R. G. Olmstead, D. Morgan, D. H. Les, B. D. Mishler, M. R. Duvall, et al. 1993. Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene *rbcL*. *Annals of the Missouri Botanical Garden* 80: 528.
- Chan, A. P., J. Crabtree, Q. Zhao, H. Lorenzi, J. Orvis, D. Puiui, A. Melake-Berhan, et al. 2010. Draft genome sequence of the oilseed species *Ricinus communis*. *Nature Biotech.* 28: 951–956.
- Chau, J. H., W. A. Rahfeldt, and R. G. Olmstead. 2018. Comparison of taxon-specific versus general locus sets for targeted sequence capture in plant phylogenomics. *Applications in Plant Sciences* 6: e1032.
- Christenhusz, M. J. M., M. F. Fay, and M. W. Chase. 2017. *Plants of the World: an illustrated encyclopedia of vascular plants*. University of Chicago Press, Chicago, IL, USA.
- Couvreur, T. L., A. J. Helmstetter, E. J. Koenen, K. Bethune, R. D. Brandão, S. A. Little, H. Sauquet, and R. H. Erkens. 2019. Phylogenomics of the major tropical plant family Annonaceae using targeted enrichment of nuclear genes. *Frontiers in Plant Science* 9: 1941.
- Crawley, S. S., and K. W. Hilu. 2012. Impact of missing data, gene choice, and taxon sampling on phylogenetic reconstruction: The Caryophyllales (angiosperms). *Plant Systematics and Evolution* 298: 297–312.
- Davis, C. C., C. O. Webb, K. J. Wurdack, C. A. Jaramillo, and M. J. Donoghue. 2005. Explosive radiation of Malpighiales supports a Mid-Cretaceous origin of modern tropical rain forests. *American Naturalist* 165: E36–E65.
- 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.
- Doyle, J. J., and J. L. Doyle. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* 19: 11–15.
- Farron, C. 1963. Contribution à la taxinomie des Ourateae Engl. (Ochnacées). *Berichte der Schweizerischen Botanischen Gesellschaft* 73: 196–217.
- Fragoso-Martínez, I., G. A. Salazar, M. Martínez-Gordillo, S. Magallón, L. Sánchez-Reyes, E. Moriarty Lemmon, A. R. Lemmon, et al. 2017. A pilot study applying the plant Anchored Hybrid Enrichment method to New World sages (*Salvia* subgenus *Calosphace*; Lamiaceae). *Molecular Phylogenetics and Evolution* 117: 124–134.
- Gnirke, A., A. Melnikov, J. Maguire, P. Rogov, E. M. LeProust, W. Brockman, T. Fennell, et al. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology* 27: 182–189.
- Goremykin, V. V., S. V. Nikiforova, and O. R. P. Bininda-Emonds. 2010. Automated removal of noisy data in phylogenomic analyses. *Journal of Molecular Evolution* 71: 319–331.
- Grover, C. E., A. Salmon, and J. F. Wendel. 2012. Targeted sequence capture as a powerful tool for evolutionary analysis. *American Journal of Botany* 99: 312–319.
- Hedtke, S. M., T. M. Townsend, and D. M. Hillis. 2006. Resolution of phylogenetic conflict in large data sets by increased taxon sampling. *Systematic Biology* 55: 522–529.
- Heled, J., and A. J. Drummond. 2010. Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution* 27: 570–580.
- Hendriks, K., T. Mandáková, N. M. Hay, E. Ly, A. Hooft van Huysduynen, R. Tamrakar, S. K. Thomas, et al. 2021. The best of both worlds: combining lineage specific and universal bait sets in target enrichment hybridization reactions. *Applications in Plant Sciences* 9 (in press). <https://doi.org/10.1002/aps3.11438>
- Hosner, P. A., B. C. Faircloth, T. C. Glenn, E. L. Braun, and R. T. Kimball. 2016. Avoiding missing data biases in phylogenomic inference: an empirical study in the landfowl (Aves: Galliformes). *Molecular Biology and Evolution* 33: 1110–1125.
- Huang, H., and L. Lacey Knowles. 2016. Unforeseen consequences of excluding missing data from next-generation sequences: simulation study of RAD sequences. *Systematic Biology* 65: 357–365.
- Jantzen, J. R., W. M. Whitten, K. M. Neubig, L. C. Majure, D. E. Soltis, and P. S. Soltis. 2019. Effects of taxon sampling and tree reconstruction methods on phylodiversity metrics. *Ecology and Evolution* 9: 9479–9499.
- Jiang, W., S. Y. Chen, H. Wang, D. Z. Li, and J. J. Wiens. 2014. Should genes with missing data be excluded from phylogenetic analyses? *Molecular Phylogenetics and Evolution* 80: 308–318.
- 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: e1600016.
- Johnson, M. G., L. Pokorny, S. Dodsworth, L. R. Botigué, R. S. Cowan, A. Devault, W. L. Eiserhardt, et al. 2019. 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.
- Kadlec, M., D. U. Bellstedt, N. C. Le Maitre, and M. D. Pirie. 2017. Targeted NGS for species level phylogenomics: 'made to measure' or 'one size fits all'? *PeerJ* 2017: e3569.
- Katoh, K., K. Misawa, K. Kuma, and T. Miyata. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* 30: 3059–3066.
- King, N., and A. Rokas. 2017. Embracing uncertainty in reconstructing early animal evolution. *Current Biology* 27: R1081–R1088.
- Korotkova, N., J. V. Schneider, D. Quandt, A. Worberg, G. Zizka, and T. Borsch. 2009. Phylogeny of the eudicot order Malpighiales: analysis of a recalcitrant clade with sequences of the *petD* group II intron. *Plant Systematics and Evolution* 282: 201–228.
- Larridon, I., T. Villaverde, A. R. Zuntini, L. Pokorny, G. E. Brewer, N. Epiawalage, I. Fairlie, et al. 2020. Tackling rapid radiations with targeted sequencing. *Frontiers in Plant Science* 10: 1655.
- Lee, E. K., A. Cibrian-Jaramillo, S.-O. Kolokotronis, M. S. Katari, A. Stamatakis, M. Ott, J. C. Chiu, et al. 2011. A functional phylogenomic view of the seed plants. *PLoS Genetics* 7: e1002411.
- Lemmon, A. R., S. A. Emme, and E. M. Lemmon. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology* 61: 727–744.
- Lemmon, E. M., and A. R. Lemmon. 2013. High-throughput genomic data in systematics and phylogenetics. *Annual Review of Ecology, Evolution, and Systematics* 44: 99–121.
- Léveillé-Bourret, É., J. R. Starr, B. A. Ford, E. M. Lemmon, and A. R. Lemmon. 2018. Resolving rapid radiations within angiosperm families using anchored phylogenomics. *Systematic Biology* 67: 94–112.
- Li, C., M. Hofreiter, N. Straube, S. Corrigán, and G. J. Naylor. 2013. Capturing protein-coding genes across highly divergent species. *Biotechniques* 54: 321–326.
- López-Giráldez, F., and J. P. Townsend. 2011. PhyDesign: an online application for profiling phylogenetic informativeness. *BMC Evolutionary Biology* 11: 152.
- Mamanova, L., A. J. Coffey, C. E. Scott, I. Kozarewa, E. H. Turner, A. Kumar, E. Howard, et al. 2010. Target-enrichment strategies for next-generation sequencing. *Nature Methods* 7: 111–118.
- 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.

- Matasci, N., L.-H. Hung, Z. Yan, E. J. Carpenter, N. J. Wickett, S. Mirarab, N. Nguyen, et al. 2014. Data access for the 1,000 Plants (1KP) project. *GigaScience* 3: 17.
- McCormack, J. E., J. M. Maley, S. M. Hird, E. P. Derryberry, G. R. Graves, and R. T. Brumfield. 2012. Next-generation sequencing reveals phylogeographic structure and a species tree for recent bird divergences. *Molecular Phylogenetics and Evolution* 62: 397–406.
- McKain, M. R., M. G. Johnson, S. Uribe-Convers, D. Eaton, and Y. Yang. 2018. Practical considerations for plant phylogenomics. *Applications in Plant Sciences* 6: e1038.
- Meyer, M., and M. Kircher. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols* 2010: pp.pdb-prot5448.
- Murphy, B., F. Forest, T. Barraclough, J. Rosindell, S. Bellot, R. Cowan, M. Golos, M. Jebb, and M. Cheek. 2020. A phylogenomic analysis of Nepenthes (Nepenthaceae). *Molecular Phylogenetics and Evolution* 144: 106668.
- Nguyen, L.-T., H. A. Schmidt, A. von Haeseler, and B. Q. Minh. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* 32: 268–274.
- Pollock, D. D., D. J. Zwickl, J. A. McGuire, and D. M. Hillis. 2002. Increased taxon sampling is advantageous for phylogenetic inference. *Systematic Biology* 51: 664–671.
- POWO. 2019. Plants of the World Online. Facilitated by the Royal Botanic Gardens, Kew. Published online: <http://www.plantsoftheworldonline.org/> [Retrieved 02 August 2020].
- Pupko, T., R. E. Bell, I. Mayrose, F. Glaser, and N. Ben-Tal. 2002. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics* 18(supplement 1): S71–S77.
- R Core Team. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Website: <https://www.R-project.org/> [accessed 26 July 2020].
- Revell, L. J. 2012. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* 3: 217–223.
- Roure, B., D. Baurain, and H. Philippe. 2013. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. *Molecular Biology and Evolution* 30: 197–214.
- Rubin, B. E. R., R. H. Ree, and C. S. Moreau. 2012. Inferring phylogenies from RAD sequence data. *PLoS One* 7: e33394.
- Sayyari, E., J. B. Whitfield, and S. Mirarab. 2017. Fragmentary gene sequences negatively impact gene tree and species tree reconstruction. *Molecular Biology and Evolution* 34: 3279–3291.
- Schneider, J. V., U. Swenson, and G. Zizka. 2002. Phylogenetic reconstruction of the neotropical family Quiinaceae (Malpighiales) based on morphology with remarks on the evolution of an androdioecious sex distribution. *Annals of the Missouri Botanical Garden* 89: 64–76.
- Schneider, J. V., U. Swenson, R. Samuel, T. Stuessy, and G. Zizka. 2006. Phylogenetics of Quiinaceae (Malpighiales): evidence from *trnL-trnF* sequence data and morphology. *Plant Systematics and Evolution* 257: 189–203.
- Schneider, J. V., P. Bissengou, M. do C. E. Amaral, A. Tahir, M. F. Fay, M. Thines, M. S. M. Sosef, et al. 2014. Phylogenetics, ancestral state reconstruction, and a new infrafamilial classification of the pantropical Ochnaceae (Medusagynaceae, Ochnaceae s.str., Quiinaceae) based on five DNA regions. *Molecular Phylogenetics and Evolution* 78: 199–214.
- Schneider, J. V., and G. Zizka. 2017. Phylogeny, taxonomy and biogeography of neotropical Quiinoideae (Ochnaceae s.l.). *Taxon* 66: 855–867.
- Schneider, J. V., T. Jungcurt, D. Cardoso, A. M. Amorim, M. Töpel, T. Andermann, O. Poncy, T. Berberich, and G. Zizka. 2020. Phylogenomics of the tropical plant family Ochnaceae using targeted enrichment of nuclear genes and 250+ taxa. *Taxon* 70: 48–71.
- Soltis, P. S., D. E. Soltis, and J. J. Doyle. 2012. Molecular systematics of plants II: DNA sequencing. Springer Science & Business Media, New York, NY, USA.
- Sosef, M. S. M. M. 2008. Révision du genre africain *Rhabdophyllum* Tiegh. (Ochnaceae), avec sa distribution au Cameroun et au Gabon. *Adansonia* 30: 119–135.
- Soto Gomez, M., L. Pokorny, M. B. Kantar, F. Forest, I. J. Leitch, B. Gravendeel, P. Wilkin, et al. 2019. A customized nuclear target enrichment approach for developing a phylogenomic baseline for *Dioscorea* yams (Dioscoreaceae). *Applications in Plant Sciences* 7: e11254.
- Straub, S. C. K., M. J. Moore, P. S. Soltis, D. E. Soltis, A. Liston, and T. Livshultz. 2014. Phylogenetic signal detection from an ancient rapid radiation: effects of noise reduction, long-branch attraction, and model selection in crown clade Apocynaceae. *Molecular Phylogenetics and Evolution* 80: 169–185.
- Streicher, J. W., J. A. Schulte, and J. J. Wiens. 2016. How should genes and taxa be sampled for phylogenomic analyses with missing data? An empirical study in iguanian lizards. *Systematic Biology* 65: 128–145.
- Summerer, D. 2009. Enabling technologies of genomic-scale sequence enrichment for targeted high-throughput sequencing. *Genomics* 94: 363–368.
- Thiers, B. 2014. Index Herbariorum: A global directory of public herbaria and associate staff. New York Botanical Garden's Virtual Herbarium, Bronx, NY, USA. [accessed August, 2020].
- Thomson, R. C., and H. B. Shaffer. 2010. Sparse supermatrices for phylogenetic inference: taxonomy, alignment, rogue taxa, and the phylogeny of living turtles. *Systematic Biology* 59: 42–58.
- Townsend, J. P. 2007. Profiling phylogenetic informativeness. *Systematic Biology* 56: 222–231.
- Townsend, J. P., and C. Leuenberger. 2011. Taxon sampling and the optimal rates of evolution for phylogenetic inference. *Systematic Biology* 60: 358–365.
- Townsend, J. P., Z. Su, and Y. I. Tekle. 2012. Phylogenetic signal and noise: predicting the power of a data set to resolve phylogeny. *Systematic Biology* 61: 835.
- Turner, E. H., S. B. Ng, D. A. Nickerson, and J. Shendure. 2009. Methods for genomic partitioning. *Annual Review of Genomics and Human Genetics* 10: 263–284.
- Vatanparast, M., A. Powell, J. J. Doyle, and A. N. Egan. 2018. Targeting legume loci: A comparison of three methods for target enrichment bait design in Leguminosae phylogenomics. *Applications in Plant Sciences* 6: e1036.
- Villaverde, T., P. Jiménez-Mejías, M. Luceño, M. J. Waterway, S. Kim, B. Lee, M. Rincón-Barrado, et al. 2020. A new classification of *Carex* (Cyperaceae) subgenera supported by a HybSeq backbone phylogenetic tree. *Botanical Journal of the Linnean Society* 194: 141–163.
- Villaverde, T., L. Pokorny, S. Olsson, M. Rincón-Barrado, M. G. Johnson, E. M. Gardner, N. J. Wickett, J. Molero, R. Riina, and I. Sanmartín. 2018. Bridging the micro- and macroevolutionary levels in phylogenomics: Hyb-Seq solves relationships from populations to species and above. *New Phytologist* 220(2): 636–650.
- Wagner, C. E., I. Keller, S. Wittwer, O. M. Selz, S. Mwaiko, L. Greuter, A. Sivasundar, and O. Seehausen. 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Molecular Ecology* 22: 787–798.
- 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: e1400042.
- 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.
- Wu, Y. 2012. Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. *Evolution* 66: 763–775.
- Xi, Z., L. Liu, and C. C. Davis. 2016. The impact of missing data on species tree estimation. *Molecular Biology and Evolution* 33: 838–860.
- Xi, Z., L. Liu, J. S. Rest, and C. C. Davis. 2014. Coalescent versus concatenation methods and the placement of *Amborella* as sister to water lilies. *Systematic Biology* 63: 919–932.
- Xi, Z., J. S. Rest, and C. C. Davis. 2013. Phylogenomics and coalescent analyses resolve extant seed plant relationships. *PLoS One* 8: e08070.
- Xi, Z., B. R. Ruhfel, H. Schaefer, A. M. Amorim, M. Sugumaran, K. J. Wurdack, P. K. Endress, et al. 2012. Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. *Proceedings of the National Academy of Sciences, USA* 109: 17519–17524.

- 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.
- Zwickl, D. J., and D. M. Hillis. 2002. Increased taxon sampling greatly reduces phylogenetic error. *Systematic Biology* 51: 588–598.

APPENDIX 1. Voucher information for material included in the study. Index Herbariorum (Thiers, 2014) abbreviations are in parentheses.

TAXON, Country of collection, *Voucher* (Herbarium).

Adenarake muriculata Maguire & Wurdack, Brazil, *Maguire 60447* (GH). *Blastemanthus gemmiflorus* (Mart.) Planch., Brazil, *Prance et al. 29826* (NY); Brazil, *Prance et al. 15495* (U). *Brackenridgea nitida* A.Gray, Australia, *Costion 1444* (CNS). *Brackenridgea zanguebarica* Oliv., Mozambique, *Burrows & Burrows 10975* (K). *Campylospermum elongatum* (Oliv.) Tiegh., Nigeria, *Onochie 9177* (K); Gabon, *Wieringa 6292* (WAG). *Campylospermum reticulatum* (P.Beauv.) Farron, Cameroon, *Cable 3433* (K); Gabon, *Bissiengou 827* (WAG). *Cespedesia spathulata* (Ruiz & Pav.) Planch., *Chase 1325* (K). *Elvasia calophyllea* DC., Brazil, *Kubitzki 84349* (NY); Brazil, *Kubitzki 84-349* (NY). *Fleurydora felicis* A.Chev., Guinea-Conakry, *Haba et al. 445* (K); Guinea, *Haba 18* (P). *Garcinia livingstonei* T.Anderson, *Chase 34490* (K). *Garcinia oblongifolia* T.Anderson, *Chen et al. 2010090804*. *Godoya obovata* Ruiz & Pav., Peru, *Weigend et al. 5695* (MO); Ecuador, *Neill and Quizphe 14978* (MO). *Hypericum perforatum* L., s.n. 132735 (ALTA). *Idertia axillaris* (Oliv.)

Farron, Sierra Leone, *Saradugu et al. 47* (K); Guinea, *Jongkind 11167* (WAG). *Lophira alata* Banks ex C.F.Gaertn., Cameroon, *Etuge & Mariana 5231* (K); n.a. 20110701A (RBGE). *Luxemburgia schwackeana* Taub., Brazil, *Esteves et al. s.n.* (CFCR no. 15466) (LZ). *Luxemburgia octandra* A.St.-Hil., Brazil, *Forzza et al. 3712* (K). *Mammea americana* L., *Soltis & Miles 3003*, (K). *Medusagyne oppositifolia* Baker, s.n. (NCY); Seychelles 20030393 (RBGE). *Ochna afzelii* R.Br. ex Oliv., Gabon, *de Wilde et al. 11413* (K); Guinea, *Haba 104* (WAG). *Ochna pulchra* Hook., Angola, *Crawford et al. FC845* (K); Namibia, *Silver SIL29* (WAG). *Ochna schweinfurthiana* F.Hoffm., Ethiopia, *Haile 840* (K) *Ouratea castaneifolia* (DC.) Engl., Brazil, *Sasaki et al. 1809* (K); Brazil, *Morawetz 22-22983* (LZ). *Perissocarpa steyermarkii* (Maguire) Steyererm. & Maguire, Venezuela, *Liesner and Gonzalez 10249* (MO). *Perissocarpa umbellifera* Steyererm. & Maguire, Brazil, *Prance et al. 29080* (K). *Philacra auriculata* Dwyer, Brazil, *Pipoly & Samuels 6867* (K); Venezuela, *Liesner 16657* (MO). *Quiina guianensis* Aubl., Guyana, *McDowell et al. 4356* (K). *Rhabdophyllum calophyllum* (Hook.f.) Tiegh., Cameroon, *Burgt et al. 1931* (K); Gabon, *Sosef 2685* (WAG). *Rhizophora mangle* L., s.n. *Sauvagesia erecta* L., Brazil, *Pereira-Silva 16200* (K); Brazil, *Benko-Iseppon 1790* (UFP/FR). *Sauvagesia serrata* (Korth.) Sastre, *Duangjai 47* (K); Indonesia, *Khairuddin (F.R.I) 31754* (L). *Schuermansia henningsii* K.Schum., *Hoogland 8954* (NY); Papua New Guinea, *Morawetz and Waha 13-2287* (LZ). *Testulea gabonensis* Pellegr., Gabon, s.n. 9420 (K); Gabon, *Wieringa 6171* (WAG). *Touroulia guianensis* Aubl., Brazil *de Souza & da Silva 172* (K); French Guiana, *Prévost 4595* (CAY). *Tyleria bicarpellata* Gleason, Venezuela, *Steyermark et al. 128556* (U). *Wallacea insignis* Spruce ex Benth. & Hook.f., *Kawasaki 243* (NY). *Wallacea multiflora* Ducke, Venezuela, *Berry 5926* (MO).