Molecular Phylogenetics and Evolution 152 (2020) 106769

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Utility of targeted sequence capture for phylogenomics in rapid, recent angiosperm radiations: Neotropical *Burmeistera* bellflowers as a case study

Justin C. Bagley[a,b,*], Simon Uribe-Convers[a], Mónica M. Carlsen[c], Nathan Muchhala[a]

[a] Department of Biology, University of Missouri–St. Louis, St. Louis, MO 63121, USA
[b] Department of Biology, Virginia Commonwealth University, Richmond, VA 23284, USA
[c] Research Department, Science and Conservation Division, Missouri Botanical Garden, St. Louis, MO 63110, USA

ARTICLE INFO

Keywords:
Angiosperms
Campanulaceae
Coalescent-based species trees
Gene tree conflict
Phylogenetic informativeness
Phylogenomics
Targeted sequence capture

ABSTRACT

Targeted sequence capture is a promising approach for large-scale phylogenomics. However, rapid evolutionary radiations pose significant challenges for phylogenetic inference (e.g. incomplete lineages sorting (ILS), phylogenetic noise), and the ability of targeted nuclear loci to resolve species trees despite such issues remains poorly studied. We test the utility of targeted sequence capture for inferring phylogenetic relationships in rapid, recent angiosperm radiations, focusing on *Burmeistera* bellflowers (Campanulaceae), which diversified into ~130 species over less than 3 million years. We compared phylogenies estimated from supercontig (exons plus flanking sequences), exon-only, and flanking-only datasets with 506–546 loci (~4.7 million bases) for 46 *Burmeistera* species/lineages and 10 outgroup taxa. Nuclear loci resolved backbone nodes and many congruent internal relationships with high support in concatenation and coalescent-based species tree analyses, and inferences were largely robust to effects of missing taxa and base composition biases. Nevertheless, species trees were incongruent between datasets, and gene trees exhibited remarkably high levels of conflict (~4–60% congruence, ~40–99% conflict) not simply driven by poor gene tree resolution. Higher gene tree heterogeneity at shorter branches suggests an important role of ILS, as expected for rapid radiations. Phylogenetic informativeness analyses also suggest this incongruence has resulted from low resolving power at short internal branches, consistent with ILS, and homoplasy at deeper nodes, with exons exhibiting much greater risk of incorrect topologies due to homoplasy than other datasets. Our findings suggest that targeted sequence capture is feasible for resolving rapid, recent angiosperm radiations, and that results based on supercontig alignments containing nuclear exons and flanking sequences have higher phylogenetic utility and accuracy than either alone. We use our results to make practical recommendations for future target capture-based studies of *Burmeistera* and other rapid angiosperm radiations, including that such studies should analyze supercontigs to maximize the phylogenetic information content of loci.

1. Introduction

Genome reduction approaches to high-throughput sequencing (HTS), including multiplex PCR (Turner et al., 2009; Uribe-Convers et al., 2016), RAD-seq (e.g. ddRAD-seq, Peterson et al., 2012), RNA-seq (e.g. Timme et al., 2012), and target capture-based approaches (Weitemier et al., 2014), are revolutionizing our ability to generate large-scale phylogenomic datasets (reviewed by Cronn et al., 2012; Lemmon and Lemmon, 2013; Andrews et al., 2016; McKain et al., 2018). These approaches are more cost-effective than Sanger sequencing, and the resulting datasets typically contain hundreds to thousands of low-copy nuclear loci. By greatly expanding the number of phylogenetically informative sites available for multilocus analyses, such advances provide a crucial basis for resolving species trees while overcoming phylogenetic noise and gene tree discordance (e.g. Rokas et al., 2003; Leaché and Rannala, 2011; Salichos and Rokas, 2013; Straub et al., 2012, 2014; Townsend et al., 2012).

Targeted sequence capture has emerged as a promising approach for phylogenomic studies of plant and animal taxa. One method focuses on sequencing ultraconserved elements (UCEs), genomic regions that are conserved across a broad taxonomic range of organisms but which have highly variable flanking regions (Bejerano et al., 2004). The UCE approach has emphasized probe sets and experimental procedures tailored to animal genomes (e.g. Faircloth et al., 2012; Lemmon and Lemmon,

[*] Corresponding author at: Department of Biology, University of Missouri-St. Louis, One University Boulevard, 223 Research Building, St. Louis, MO 63121-4499, USA.
E-mail address: bagleyj@umsl.edu (J.C. Bagley).

https://doi.org/10.1016/j.ympev.2020.106769
Received 9 December 2019; Received in revised form 10 February 2020; Accepted 12 February 2020
Available online 17 February 2020

Table 1
Review of recent targeted sequence capture studies of plants employing a Hyb-Seq (Weitemier et al., 2014) approach.

Study	Organisms	*n*	Loci [a]	Capture success	Exons	Flanking	Supercontigs	Plastomes [b]
Weitemier et al. (2014)	*Asclepias* (Apocynaceae)	12	768	99–100%	x	x	–	x
Crowl et al. (2017)	Mediterranean *Campanula* (Campanulaceae)	105	246	95.70%	x	–	x	x
Chau et al. (2018)	*Buddleja* (Lamiales)	48	1049	91–99%	x	–	–	–
Gernandt et al. (2018)	*Pinus* (Pinaceae)	74	710 +	96.6%	x	–	–	x
Herrando-Moraira et al. (2018)	Cardueae (Compositae)	85	1061	64–99%	x	–	–	–
Jones et al. (2019)	Asteraceae	112	1061	66–99%	x	–	x	–
Kates et al. (2018)	*Artocarpus* (Moraceae)	24	151	73.50%	x	x	x	–
Stubbs et al. (2018)	*Micranthes* (Saxifragaceae)	49	518	99–100%?	x	–	–	x
Villaverde et al. (2018)	*Euphorbia* (Euphorbiaceae)	121	431	45–86%	x	x	x	x
Vatanparast et al. (2018)	Leguminosae	25	423	93%	x	–	x	x

Single-digit values for "Capture success" are generally averages for exons. Terms and abbreviations: *n*, number of samples; No., number of; supercontig, contig of targeted exons and flanking nuclear regions.

[a] Number of loci targeted with bait sets.

[b] Partial or whole plastome sequences obtained via 'genome-skimming' (Straub et al., 2012) from Hyb-Seq reads. Only counted if plastome phylogenetic results were reported (i.e. not if plastomes only mapped to reference).

2013) but may not be suitable for some plants due to the non-syntenic and non-orthologous nature of plant UCEs (e.g. Reneker et al., 2012). Accordingly, plant phylogenomic studies have largely eschewed UCEs, turning to taxon- or lineage-specific targeted sequence capture datasets, or universal datasets, based on transcriptomic or genomic resources (e.g. Mandel et al., 2014; Johnson et al., 2019; but see Léveillé-Bourret et al., 2018, refs. therein). A related approach that is useful for non-model plant taxa, Hyb-Seq (Weitemier et al., 2014), combines targeted sequence capture and HTS with 'genome-skimming' of the data to obtain high-copy plastome, mitochondrial DNA, and ribosomal internal transcribed spacer sequences (Straub et al., 2012; Johnson et al., 2016), or even low-copy nuclear sequences flanking the targeted exons (Kates et al., 2018). Recent applications to angiosperm and gymnosperm lineages showcase the feasibility of Hyb-Seq, with or without genome-skimming, over varying taxonomic and temporal scales of evolution (summarized in Table 1). However, relatively little work has tested this targeted sequence capture approach for rapid, recent angiosperm radiations. Additionally, phylogenomic studies of rapid angiosperm evolutionary radiations have relied on relatively sparse taxon sampling, sequencing up to ~25–40% of genera or species across megadiverse clades (e.g. Mandel et al., 2014, 2015; Chau et al., 2018; Villaverde et al., 2018; Folk et al., 2019).

Evolutionary radiations occur when many species arise from a single common ancestor over short, explosive periods of diversification (Givnish, 1997, 2015) resulting in higher-than-background rates of speciation (e.g. Soltis et al., 2019). Rapid accumulations of species may result from many non-mutually exclusive ecological and evolutionary processes including geographic isolation, sexual selection, or classical adaptive radiation (e.g. Schluter, 2000; Givnish, 2015; Uribe-Convers & Tank, 2015; Simões et al., 2016). Given that many such radiations have occurred in different angiosperm clades (e.g. Hughes et al., 2015; Soltis et al., 2004), inferring phylogenetic relationships in species-rich radiations of flowering plants is an important goal in evolutionary biology and facilitates additional comparative analyses (Bell et al., 2010; Abrahamczyk et al., 2014; Givnish et al., 2014; Lagomarsino et al., 2016, 2017; Folk et al., 2019; Soltis et al., 2019). Still, the rapid diversification rates of angiosperm radiations pose significant challenges for phylogenomic studies, as they typically span geologically brief periods less than 15 million years (Myr; e.g. Lagomarsino et al., 2016; Spalink et al., 2016). Short internode distances during rapid radiations increase the probability of incomplete lineage sorting (ILS) while limiting phylogenetic signal (Townsend, 2007; Whitfield and Lockhart, 2007), and longer terminal branches can generate substantial phylogenetic noise due to homoplasy (convergence due to nucleotide saturation; Townsend et al., 2012; Straub et al., 2014). Increasing the numbers of variable loci sequenced using HTS also increases chances of gene tree conflicts due to ILS, introgression, gene duplication/loss, or horizontal gene transfer (Maddison, 1997; Degnan and Rosenberg, 2009). Finally, undetected systematic biases due to rate heterogeneity among loci, incorrect alignments, or compositional biases could cause issues such as long-branch attraction (Felsenstein, 1978) or topological incongruence (e.g. Dávalos and Perkins, 2008; Rodríguez-Ezpeleta et al., 2007). In one recent study by Léveillé-Bourret et al. (2018), analyses of anchored hybrid enrichment loci resolved the backbone phylogeny and relationships of early-radiating lineages that diverged over the initial 10 Myr period of the ancient Eocene Cyperaceae evolutionary radiation. Still, the ability of nuclear loci from targeted sequence capture to overcome the phylogenetic challenges above and resolve species trees for geologically recent angiosperm radiations remains poorly studied.

Different types of nuclear data obtained from targeted sequence capture may contribute differently to overcoming the above challenges in recent angiosperm radiations. As exons are more functionally constrained, they tend to evolve more slowly in plants and animals than non-coding introns or intergenic sequences flanking the targeted exons (e.g. reviewed by Graur and Li, 2000; Avise, 2004). Lower substitution rates in exons are thus assumed to yield lower levels of homoplasy, predicting that they should be more useful in resolving deeper nodes, while their flanking sequences should perform better in resolving shallower nodes. Additionally, non-coding sequences flanking plant exons may be more likely to exhibit length polymorphisms, insertion-deletions (indels), and variations in nucleotide composition (e.g. GC content; reviewed by Ressayre et al., 2015), making them significantly more difficult to align into homologous sequences than coding regions. These characteristics may decrease the utility of flanking sequences for phylogenetics, particularly when attempting to align and analyze data from more distant lineages. By contrast, plant systematists have become increasingly interested in non-coding sequences, because exons may contain insufficient phylogenetic signal, and selection at degenerate protein-coding sites may introduce biases that can mislead phylogenetic inference from exons (Castoe et al., 2009). Only rarely have studies of angiosperms explicitly compared the phylogenetic utility of exons, flanking regions, and 'supercontig' alignments containing both exons and flanking sequences (e.g. see Table 1). Where available, such comparisons have largely been qualitative (e.g. Kates et al., 2018), lacking quantitative rigor that could be provided by approaches such as phylogenetic signal-to-noise theory (e.g. Townsend, 2007; Townsend et al., 2012).

With ~130 species (Mashburn, 2019) that diverged only in the last ~2.6 Myr (Lagomarsino et al., 2016, 2017), Neotropical bellflowers in the genus *Burmeistera* Karsten & Triana (Campanulaceae) present an ideal opportunity for evaluating the utility of different targeted sequence capture datasets to resolve phylogenetic relationships in a rapid, geologically recent angiosperm radiation. *Burmeistera* are semi-woody terrestrial shrubs or hemi-epiphytic subshrubs that inhabit cloud forest ecosystems from Guatemala to Perú (Lammers, 2007) and contribute to floral and ecological diversity in the Tropical Andes biodiversity 'hotspot' (Myers et al., 2000). Due to pollination by nectarivorous bats and hummingbirds, *Burmeistera* have attracted attention in comparative and ecological studies of pollination syndrome evolution and plant–pollinator interactions (Muchhala, 2006, 2008; Muchhala and Potts, 2007; Lagomarsino et al., 2017). Previous phylogenetic work has often incorporated small numbers of *Burmeistera* samples into broader treatments of Campanulaceae subfamily Lobelioideae (> 550 species; Antonelli, 2008; Knox et al., 2008; Lagomarsino et al., 2014), although one such study increased taxon sampling up to 33 *Burmeistera* species (Lagomarsino et al., 2016). Uribe-Convers et al. (2017) sequenced 45 *Burmeistera* for multiple plastid genes but also increased numerical sampling up to full plastome sequences for a subset of 17 *Burmeistera* and two outgroups. Unfortunately, all of these previous studies relied largely on plastid sequences, and even the whole plastomes (> 163 kb) provided insufficient numbers of informative characters to resolve the backbone topology in maximum-likelihood (ML) analyses (Uribe-Convers et al., 2017). As the plastome comprises a single, uniparentally-inherited locus, these analyses also suffer from depending on a single realization of the evolutionary process (Hudson, 1990; Brito and Edwards, 2009). Clearly, data from many unlinked nuclear genes will be needed to infer the backbone phylogeny and species relationships in *Burmeistera* with high confidence.

Here, we provide the first multilocus phylogenomic perspective on *Burmeistera* evolution by inferring phylogenetic relationships among 46 species/lineages of *Burmeistera* plus 10 outgroup taxa, using HTS data obtained by targeting single- to low-copy nuclear loci from across the genome (1.35 Mbp). Our main goal was to test the utility of targeted sequence capture to resolve the phylogeny of rapid, recent angiosperm radiations, using *Burmeistera* as an exemplary case of high Andean speciation rates, given that its center of diversity lies in Colombia and Ecuador (Lammers, 2007; Mashburn, 2019). We also quantitatively evaluated gene tree heterogeneity, as well as potential effects of phylogenetic signal, missing taxa, and base frequency compositional biases on our inferences. Specifically, we mapped patterns of gene tree congruence and conflict over coalescent-based species trees (Mirarab et al., 2014) using recent methods (Smith et al., 2015; Kates et al., 2018). Additionally, we compared phylogenetic informativeness (PI) and related internode statistics (Townsend, 2007; Townsend et al., 2012) against null expectations for our supercontig, exon, and flanking sequence alignments, and we evaluated relationships between congruence/conflict and fluctuating PI. We discuss the feasibility and value of targeted sequence capture for resolving rapid, recent angiosperm radiations, and we use our results to make practical recommendations for phylogenetic experimental design.

2. Material and methods

2.1. Taxon sampling and HTS data generation and processing

We obtained silica-dried leaf material for 60 samples (Table S1) representing 50 species/lineages of *Burmeistera* and 10 outgroup species from Campanulaceae subfamily Lobelioideae based on previous fieldwork in Ecuador, Costa Rica, Colombia, and Panama (e.g. see sampling in Lagomarsino et al., 2016; Uribe-Convers et al., 2017), as well as herbarium specimens from the Missouri Botanical Garden (MO), Chicago Field Museum (F), and New York Botanical Garden (NY). Samples were selected to maximize coverage of the taxonomic, geographic, and genetic diversity within the genus, as well as the other two major lobelioid genera, *Centropogon* and *Siphocampylus*. For *B. aspera* E. Wimm. and *B. refracta* E. Wimm., we obtained leaf material for $n = 2$ intraspecific samples with confirmed morphological identifications. However, our results showed that each sample of these species represented a genetically distinct lineage (see Results); thus, our sampling encompassed described species and undescribed but genetically distinct forms.

To identify suitable nuclear loci, we sequenced transcriptomes from leaf tissue of three *Burmeistera* and three outgroup species and used the resulting data alongside 17 available *Burmeistera* shotgun libraries to design capture probes for hybrid enrichment target capture. Using the Sondovač pipeline (Schmickl et al., 2015) with custom scripts, we identified 800 putatively single- to low-copy nuclear genes for probe development. These included 500 genes with intron–exon boundaries (2198 total exons), and 300 genes lacking intron-exon boundaries. MarkerMiner v1.0 (Chamala et al., 2015) was used to compare transcriptome data from two *Burmeistera* species and two outgroups against reference databases of known single-copy nuclear genes previously identified in other angiosperm genomes (De Smet et al., 2013). Using MarkerMiner, we were able to add 158 genes (containing 517 exons) to our set of targeted low-copy nuclear genes for *Burmeistera*. After filtering the 958-nuclear gene set, we identified 745 putatively single- to low-copy nuclear loci that were over 600 bp long, with a maximum exon length of 3764 bp. We targeted this final set of loci with 120-bp probes at 2× tile density.

We extracted total genomic DNA from all 60 samples using the 2× CTAB method (Doyle and Doyle, 1987). Subsequently, DNA libraries were constructed, enriched, and sequenced by RAPiD Genomics (Gainesville, FL) on an Illumina HiSeq 3000 sequencer (Illumina Inc., San Diego, CA, USA) to generate 100-bp single-end reads as well as 150-bp paired-end reads with a minimum sequencing depth of coverage of 40× per sample. Raw reads were quality-filtered using Seqyclean v1.10.09 (https://github.com/ibest/seqyclean) to remove Illumina adapters, low quality bases (PHRED scores < Q20), and short reads (< 40 bp). Remaining reads were assembled, and target sequences and flanking sequences were extracted, in the HybPiper v1.3.1 pipeline (Johnson et al., 2016), as summarized in Fig. 1. We used the 'reads_first.py' Python script to (1) conduct quality filtering, (2) align reads to target gene reference sequences using BWA v0.7.17 (Li and Durbin, 2009), (3) assemble contigs *de novo* using SPAdes v3.6.1 (Bankevich et al., 2012), and (4) automate extraction of exon-only sequences and supercontigs (by combining overlapping contigs) for each gene using Exonerate v2.4.0 (Slater and Birney, 2005) (Fig. 1A). We ran Exonerate again on the supercontigs to identify and extract flanking-only sequences, as automated by the 'intronerate.py' script, and this yielded full or partial introns and intergenic sequences. Last, we retrieved multi-individual FASTA files of exon-only, flanking-only, and supercontig sequences for each gene using the 'retrieve_sequences.py' script, and we calculated summary statistics for target enrichment and gene recovery using the 'hybpiper_stats.py' script (Fig. 1B). We used a combination of orthology and data quality filters to reduce the datasets obtained from HybPiper. As HybPiper flagged 135 loci as potential paralogs based on multiple assembled contigs with lengths > 85% of the target sequence length, we removed these from the 681 successfully assembled loci (maximum across individuals; see Results Section 3.1), leaving 546 putatively orthologous loci for analysis. We also removed four problematic species (*B.* sp. *cf. aeribacca*, *B. brachyandra* E. Wimm., *B. ceratocarpa* Zahlbr., and *B. quercifolia* Gómez & Gómez) with substantial amounts of missing data (> 50%) and reduced the dataset to 56 species/lineages (Supplementary Table S1).

We generated five datasets for downstream genetic analyses (Table 2), each composed of single-locus alignments plus a concatenated 'supermatrix' (de Queiroz and Gatesy, 2007). First, we generated (1) a 'full supercontig' dataset by aligning the supercontig sequences obtained by Exonerate for 542 loci using MAFFT v7.294b (Katoh and Standley, 2013; (–auto option) and then cleaning alignments with Phyutility (Smith and Dunn, 2008) at 50% occupancy. Second, to evaluate potential effects of missing data and whether inferences could be improved by using complete taxon sampling, we subsetted the full supercontig alignments using a 100% taxonomic completeness threshold to generate (2) a '100p supercontig' dataset of

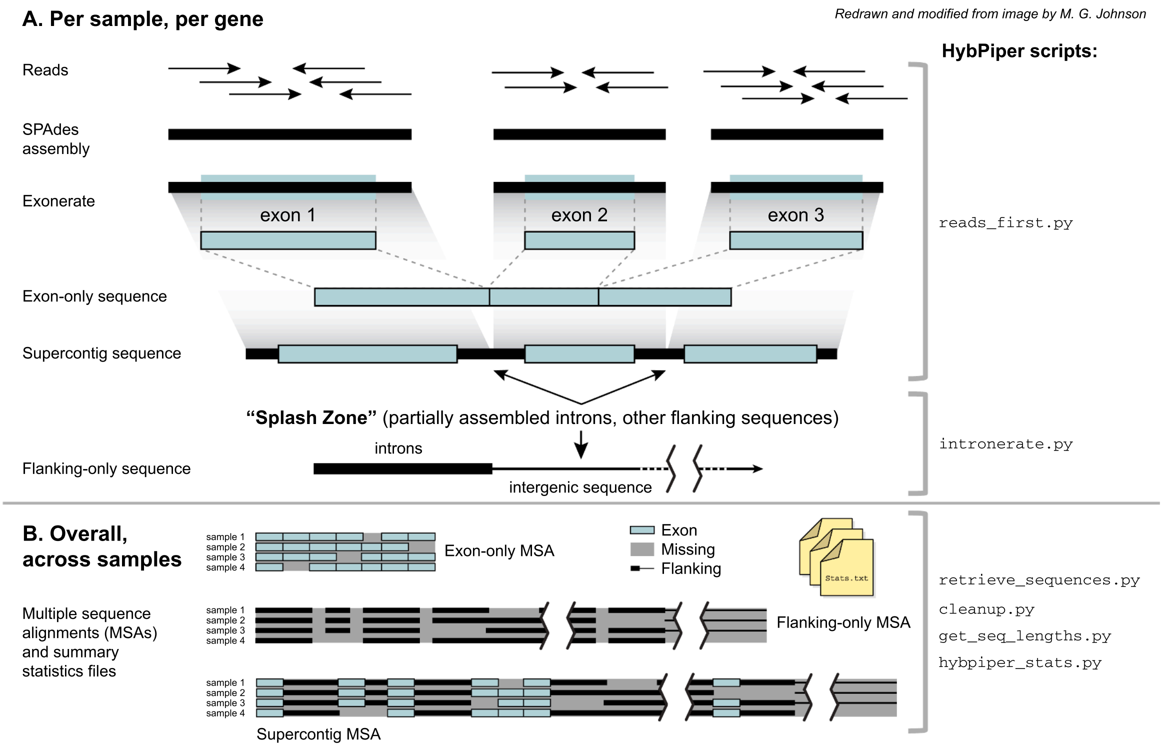

Fig. 1. Graphical representation of workflow for data assembly and extraction of target gene sequences into exon-only, supercontig (exons plus flanking sequences, including introns and intergenic regions), and flanking-only sequence sets using the HybPiper pipeline (Johnson et al., 2016). The main HybPiper scripts used for data processing are shown to the right of brackets representing analysis phases conducted per sample for each gene [A; redrawn and modified from a HybPiper wiki image (https://github.com/mossmatters/HybPiper/wiki) available under GNU General Public License v3.0], and across samples and genes (B). See text Section 2.1 and the accompanying Mendeley Data accession for additional details, including licensing and multiple sequence alignment (MSA) and data filtering procedures. (PDF).

515 supercontig alignments, with no missing loci within individuals. Following procedures similar to those above, we created (3) an 'exon-only' dataset containing cleaned and aligned exon-only sequences from Exonerate for all 546 loci, which we subsetted to generate (4) a '100p exon' dataset of 519 exon alignments. To assess the phylogenetic utility of "splash zone" sequences flanking the targeted exons (Weitemier et al., 2014; Fig. 1), we created (5) a 'flanking-only' dataset by cleaning and aligning flanking sequences from Exonerate, removing 10 individuals with substantial missing flanking sequence data, and applying a 100% taxonomic completeness threshold for the remaining 46 species/lineages, yielding alignments for 506 loci (93% of full supercontig dataset) with no missing loci within individuals (Table 2). For each dataset, we used the 'completeConcatSeqs' function in PIrANHA v0.3a2 (Bagley, 2019) to concatenate gene alignments into supermatrices and automatically generate partition blocks for downstream analyses.

2.2. Phylogenomic inference and divergence dating

Taking a model-based supermatrix approach, we first conducted concatenation + ML analysis (CAML) on each dataset's supermatrix in RAxML v8.2.8 (Stamatakis, 2014), while estimating parameters of separate GTR + Γ models for each locus partition and calculating nodal support from 250 rapid bootstrap pseudoreplicates (-f a -x option). We also estimated gene trees independently for each locus by dataset in RAxML while using the GTR + Γ model and 100 rapid bootstrap pseudoreplicates, as automated in the 'MAGNET' v1.1.0 function of PIrANHA. Given supermatrix approaches can be inconsistent (i.e. incongruent relative to true species tree; Roch and Steel, 2015), we estimated species trees using a summary method known to be consistent under the multispecies coalescent, ASTRAL-III v5.6.3 (Mirarab et al., 2014; Zhang et al., 2018; hereafter, 'ASTRAL'). We chose ASTRAL over related methods such as MP-EST (Liu et al., 2010) because ASTRAL is less sensitive to gene tree estimation error (e.g. Mirarab and Warnow, 2015). Before running ASTRAL, we collapsed nodes in the RAxML gene trees that had < 33% bootstrap using Newick utilities (Junier and Zdobnov, 2010) to improve inference and avoid spurious species trees (Smith et al., 2015; Kates et al., 2018; Zhang et al., 2018). This yielded species trees with local posterior probability (LPP) branch support, which is more accurate and precise than multi-locus bootstrapping (Sayyari and Mirarab, 2016). To estimate a chronogram for *Burmeistera*

Table 2
Summary statistics for the five targeted sequence capture datasets for species/lineages of *Burmeistera* and outgroups analyzed in this study.

Dataset	*n* (ingroup/outgroup)	No. loci	bp	% nucleotide	% missing	% gap	Base frequencies (a, c, g, t)
1. Full supercontig	56 (46/10)	542	4,211,052	84%	0.12%	16%	(0.30, 0.19, 0.20, 0.31)
2. 100p supercontig	56 (46/10)	515	4,144,703	84%	0.00%	16%	(0.30, 0.19, 0.20, 0.31)
3. Exon-only	56 (46/10)	546	1,057,755	96%	0.13%	4%	(0.29, 0.21, 0.21, 0.29)
4. 100p exon	56 (46/10)	519	1,038,025	96%	0.00%	4%	(0.29, 0.21, 0.21, 0.29)
5. Flanking-only	46 (40/6)	506	4,754,799	74%	0.30%	26%	(0.31, 0.19, 0.19, 0.31)

See Supplementary Table S1 for additional sampling information. Terms and abbreviations: bp, base pairs; *n*, number of samples; No., number of; prop., proportion; supercontig, contig of targeted exons and flanking nuclear regions.

and outgroups for downstream comparative analyses of phylogenetic information content, we used penalized likelihood (PL; Sanderson, 2002) as implemented in the 'chronos' function of APE v5.0 (Paradis and Schliep, 2018) in R v3.5.3 (R Core Team, 2018). We calibrated the best CAML tree for the full supercontig dataset using two secondary calibration points defined by 95% credible intervals of molecular divergence times from a previous Bayesian analysis of centropogonid diversification (Lagomarsino et al., 2016; see details in Supplementary Data S1).

2.3. Substitution rates and phylogenetic informativeness analyses

We estimated site-by-site relative substitution rates under ML with the LEISR method (Spielman and Pond, 2018) in HyPhy v2.5.0 (Pond and Muse, 2005; https://hyphy.org). For each locus in the full supercontig, exon-only, and flanking-only datasets, alignment-wide branch lengths were optimized over a version of the PL chronogram rescaled to a total height of 1, and then relative rates were obtained as site-specific uniform tree scalars (Spielman and Pond, 2018). Branch-length optimization was performed under the JC69 substitution model (Jukes and Cantor, 1969) to match assumptions of Townsend (2007) and Townsend et al.'s (2012) signal-to-noise theory calculations. To get a sense of rate variation, we qualitatively compared relative rate frequency histograms for the first 50,000 sites in each dataset in R.

To assess phylogenetic information content, we estimated PI profiles for each full-supercontig, exon, and flanking-only alignment, using the corresponding site rates and the rescaled PL chronogram, in the R package PhyInformR v1.0 (Dornburg et al., 2016). Each PI profile represents the probability density of a true parsimony-informative synapomorphy through time, calculated from empirical rates of character evolution in a locus or dataset (Townsend, 2007; Townsend and Leuenberger, 2011; Townsend et al., 2012). As PI profiles do not quantify the influence of homoplasy on phylogenetic resolution, we also estimated and plotted Quartet Internode Resolution Probability (QIRP; Townsend et al., 2012) of each alignment at each guide chronogram node (Hwang et al., 2015; Prum et al., 2015) under a three-character state model (Simmons et al., 2004). QIRP uses a Gaussian approximation to estimate the probability of correct resolution based on signal and noise theory defining a predictive relationship between known site-specific rates and the depth and internode distance of a given node, or whole topology (Townsend et al., 2012; Dornburg et al., 2016). To evaluate the resulting PI and QIRP profiles against a null expectation of equal site rates, we generated 'dummy' sets of relative rates, set to 1.0 (equal rates) for each site and mimicking the precise patterns of non-gap sites in all loci in the three analyzed datasets, using MEGA-CC v7.0.26 (Kumar et al., 2016). We then reanalyzed these dummy datasets using PI and QIRP analyses identical to those above.

Finally, we conducted several analyses on each dataset to distinguish whether low PI estimates at five key nodes in two time-epochs (see Results Section 3.3) were driven primarily by increases in homoplasious sites, or by low resolution probabilities (low statistical power). Analyses were conducted using custom code and R scripts modified from Prum et al. (2015). First, we quantified the ratio of PI at the younger ends versus PI at the older ends of internodes subtending the five focal nodes, i.e. across the two epochs. Ratios less than 1 indicate a rootward rise in PI, values near 1 indicate constant PI, and values greater than 1 indicate declines in PI towards the root likely due to homoplasy, and possibly a "rain shadow of noise" following peak PI (Townsend and Leuenberger, 2011). Second, we evaluated and counted loci with 'phantom spike' patterns reflecting artificially high PI values due to the presence of unusually fast-evolving sites (Townsend et al., 2008; Herrando-Moraira et al., 2018). Third, we created QIRP heatmaps for 20 alignments of increasing sequence length, across varying internode lengths (*cf.* Prum et al., 2015). Heatmap loci included (1) the concatenated supermatrices and (2) 19 additional alignments selected by reverse sorting individual loci in each dataset by number of nucleotides into 19 equal-sized bins, and taking the first locus from each bin. By probing resolution probabilities over different hypothetical internode lengths, this approach accounts for uncertainty in the guide tree, which may contain errors in topology or branch lengths but is assumed in PI analyses to represent the 'true' tree (Prum et al., 2015).

2.4. Treespace visualization of effects of missing taxa

We qualitatively compared the effects of missing data, in the form of missing taxa, on our phylogenetic gene tree and species tree inferences using 2-dimensional visualizations of treespace in the TreeSetViz module (Hillis et al., 2005; Amenta et al., 2012) of Mesquite v3.5.1 (Maddison and Maddison, 2018). In TreeSetViz, we plotted similarly rooted and collapsed versions of the ML gene trees and the ASTRAL species trees, and these analyses focused on regular versus '100p' supercontig and exon-only dataset pairs (datasets 1–4), which provided comparisons with and without missing taxa. This procedure yielded coordinate plots of each tree in treespace with spacing based on pairwise Robinson–Foulds distances (topological distances) from an arbitrarily selected species tree, the full supercontig ASTRAL tree.

2.5. Effects of compositional biases

When different lineages have different overall frequencies of base pairs, such compositional biases can negatively influence phylogenomic analyses (e.g. Dávalos and Perkins, 2008; Rodríguez-Ezpeleta et al., 2007; Ishikawa et al., 2012; Longo et al., 2017). We calculated base frequencies and tested for the presence of base compositional biases across taxa in our five datasets using the chi-squared (χ^2) test implemented in PAUP* v4.0a (Swofford, 2002). Subsequently, we evaluated the potential for significant deviations from homogeneous base frequencies in a given dataset to produce systematic biases leading to conflicting groupings of taxa with similar base frequencies. We performed ML phylogenetic analyses on versions of the concatenated supermatrices for all five datasets converted to binary 'RY' coding (Woese et al., 1991), with purines coded as 0′s and pyrimidines coded as 1′s. We analyzed the binary supermatrices in RAxML using the BINGAMMA model for binary state data, while estimating nodal support with 100 fast bootstrap pseudoreplicates.

2.6. Assessing and mapping tree congruence and conflict

We assessed patterns of congruence and conflict among our gene trees and species trees using PhyParts (Smith et al., 2015; available at: https://bitbucket.org/blackrim/phyparts), which summarizes the congruence and conflict of bipartitions (shared internal edges) across a set of trees by comparison to a reference tree. For each clade, PhyParts tabulates the proportion of gene trees that (1) support the reference tree, (2) support the main alternative topology, (3) support all remaining alternatives, and (4) support the proportion of gene trees that are informative for the clade but have less than a user-specified bootstrap support level (i.e. gene trees that are uncertain for each node; Smith et al., 2015). For the full supercontig, exon-only, and flanking-only datasets, we ran PhyParts (-a 1 -v option) on rooted ML gene trees for each locus with nodes with < 33% bootstrap support collapsed, while using a rooted version of the corresponding ASTRAL species tree as the reference (*cf.* Kates et al., 2018). *Siphocampylus jelskii* Zahlbr. was used as the outgroup, because it was present in the greatest number of alignments, and gene trees lacking the outgroup were excluded from the analysis (thus, in results figures, we give number of genes/total). We summarized and plotted PhyParts results over ASTRAL species trees using a modified Python notebook from M. G. Johnson (https://github.com/mossmatters/MJPythonNotebooks/). We tested for statistically significant relationships between mean PI and the numbers of congruent gene trees for each node in the ASTRAL species trees using generalized linear modeling in R. We also statistically tested whether patterns of congruence were significantly correlated to node depths, as judged by divergence times estimated under PL above, using linear modeling.

3. Results

3.1. Taxon sampling and HTS data generation and processing

Targeting nuclear loci using RNA probes and sequencing them on an Illumina HiSeq 3000 yielded ~359.5 million 100-bp and 2× 150-bp reads, with an average of ~6.3 million reads per sample (range: 1.8 million to 15.6 million reads). Probes had high accuracy, and the majority of raw reads mapped to a target gene (mean: 87%, range: 81–94%). We assembled 617–681 loci, with an average of 676 loci (91% of targets; range 83–91%) per sample, and after checking alignments 'by-eye' for quality issues (e.g. missing data or short sequences), we deemed 677 of these loci to be of high quality. Most targeted loci were present as single copies within each species, with only 135 (20%) of the 681 total loci triggering paralog warnings in HybPiper due to multiple long contigs, and subsequently removed (see Section 2.1).

Final concatenated alignments contained 506–546 loci and ranged from ~1 Mbp to ~4.7 Mbp in length, with supercontig and flanking-only supermatrices being ~4-fold larger than the exon-only supermatrix, but the full and 100p supermatrices from the same data type being similar in size (Table 2). Datasets were highly complete, with average amounts of missing data and gap characters being only 0.11% and 12.9%, respectively (Table 2). Consistent with our expectations, individual supercontig loci were much longer (range: 511 bp to 80,057 bp, mean: 7769 bp) than exon loci (range: 141 bp to 13,248 bp, mean: 1937 bp) (Supplementary Fig. S1A). Flanking-only loci were generally slightly longer than the supercontigs (Fig. S2A), reflecting shorter sequences in combination with the roughly 60% increase in the proportion of gap characters (Table 2). While the proportion of parsimony-informative sites (PIS) was similar between full and 100p datasets from the same data type, loci in the supercontig and flanking-only datasets contained thousands more PIS (full supercontig range: 44 to 55,448 PIS, mean: 2089 PIS; flanking-only range: 23 to 95,203 PIS, mean: 2532 PIS) than those in the full exon-only dataset (range: 5 to 4253 PIS, mean: 241 PIS) (Fig. S1B). Despite comprising realigned subsets of supercontig sequences, the flanking-only dataset had the most PIS overall (Fig. S2B) because it contained at least one locus with very long flanking sequences (Gene000000007817) that was absent from supercontig dataset 1 (full supercontig) and 2 (100p supercontig) due to filtering procedures.

3.2. Phylogenomic inference and divergence dating

Topologies from CAML and coalescent-based analyses were overall highly congruent and provided a well-resolved phylogeny of *Burmeistera* defining clades and relationships among closely related species. In results for our main datasets 1 (full supercontig), 3 (exon-only), and 5 (flanking-only) (Section 2.1), *Burmeistera* was unambiguously monophyletic with most ingroup nodes receiving definitive bootstrap proportion (BP) support (BP = 100%) in concatenated supermatrix results (Fig. 2A, 2B, 3A and 3B), and strong LPP support ($>$ 0.9) in ASTRAL species trees (Fig. 2C, 2D, 3C and 3D). Within *Burmeistera*, we consistently resolved four well-supported major clades (clades 1–4) plus a distinct lineage formed by *B. xerampelina* E. Wimm that was usually sister to all other *Burmeistera*, yielding an ingroup topology of the form, ((((clade 1, clade 2), clade 3), clade 4), *B. xerampelina*) (Figs. 2 and 3). Maximum-likelihood branch lengths generally increased from the root towards the tips of the tree; however, despite some very long terminal branches (e.g. *B. huacamayensis* Jeppesen, *B. xerampelina*; Fig. 2A), we found no clear patterns of long-branch attraction. Branch lengths also demonstrated that genetic divergences within species with intraspecific sampling were relatively deep, achieving that seen between species pairs: divergence between *B. aspera* samples 1 and 2 was similar to the *B. borjensis–B. oyacachensis* split, while divergence between *B. refracta* 1 and 2 was slightly greater than that between *B. almedae* and *B. obtusifolia*. Relative to supercontig results, exon-only results generally had lower nodal support values across analyses (Figs. S1C–S1E), and the flanking-only results yielded mixed nodal support (Figs. S2C–S2E). Results for '100p' datasets 2 and 4 were nearly identical to those of corresponding full datasets 1 and 3 in their relationships and nodal support, suggesting that the degree of completeness of taxon sampling across datasets and genes did not negatively influence results. We thus provide the 100p results as Supplementary material (see Figs. S3 and S4) and, hereafter, emphasize results from main datasets 1, 3, and 5.

Contrasting general trends of high support and congruence, several nodes with short subtending internodes exhibited conspicuously lower support values or incongruence. These included internal nodes near the bases of clades 2–4 and one backbone node, and the corresponding tips generated most cases of topological incongruence between CAML trees or species trees from different datasets (Figs. 2 and 3). Striking cases of backbone incongruence included (1) *B. xerampelina* placed with definitive support in the outgroup clade of the flanking-only CAML tree (Fig. 3B) and (2) *B. brighamoides* Lammers placed sister to a clade containing clades 1–3 plus all other members of clade 4 in the exon-only ASTRAL species tree, but with very low support (LPP = 0.58; Fig. 2D). Two notable sets of internal nodes created incongruence between datasets. The first was subclade '1-a' (*B. borjensis* Jeppesen, *B. oyacachensis* Jeppesen, *B. glabrata* Benth. & Hook. F. ex B. D. Jacls, *B. vulgaris* E. Wimm., and *B. draconis* Pérez & Muchhala) that received low support (LPP = 0.58–0.78) in the exon-only ASTRAL species tree, where positions of *B. glabrata* and *B. vulgaris* were switched (Fig. 2D); however, this subclade was supported in all other ASTRAL trees. Additionally, subclade '3-a' (*B. crispiloba* Zahlbr., *B. sodiroana* Zahlbr., and *B. succulenta* Triana) varied in its position within clade 3, and was resolved with varying internal relationships between supercontig and exon-only results (Fig. 2).

Comparing penalized log-likelihoods of different PL (Sanderson, 2002) models run over a range of λ values in APE showed that λ = 0.1 represented the optimal smoothing parameter, and the time-calibrated phylogeny of *Burmeistera* and outgroup taxa derived from the best-supported PL model (penalized log $L = -14.35293$) indicated a largely Pliocene–recent timescale of *Burmeistera* diversification (Fig. S5). All five major lineages of *Burmeistera* diverged from one another between ~3.15 Ma and ~2 Ma in the late Pliocene to early Pleistocene (Gelasian; Gibbard et al., 2010; additional details in Supplementary Data S1).

3.3. Substitution rates and phylogenetic informativeness analyses

Site-specific relative substitution rates across loci in the full supercontig, full exon-only, and flanking-only datasets mostly fell between 0 and 1 (equal rates), although thousands of sites had rates between 1 and 5 and small numbers of sites had rates between 5 and 24 (Fig. S6). Per-locus PI profiles from different datasets were broadly similar through time, peaking at upper- to mid-crown depths before declining towards the root of the tree (Fig. 4). PI profiles were also similar to 'null' expectations based on equal rates, and anomalous 'phantom spike' patterns due to very fast-evolving sites were not apparent. While exon-only loci most closely matched null expectations, consistent with lower noise potential, 25 full supercontig loci and 153 flanking-only loci had higher peak PI values (Fig. 4) than an equal-rates distribution, and these loci tended to have more drastic declines in PI following their peaks, consistent with higher potential for phylogenetic noise. The exon-only dataset contained 5-fold more PI profiles (55 loci) with phantom spikes indicating fast-evolving sites than other datasets (full supercontig dataset: 11 loci; flanking-only dataset: 12 loci). While PI profiles do not discount noise due to homoplasy (Townsend, 2007), quantifying PI

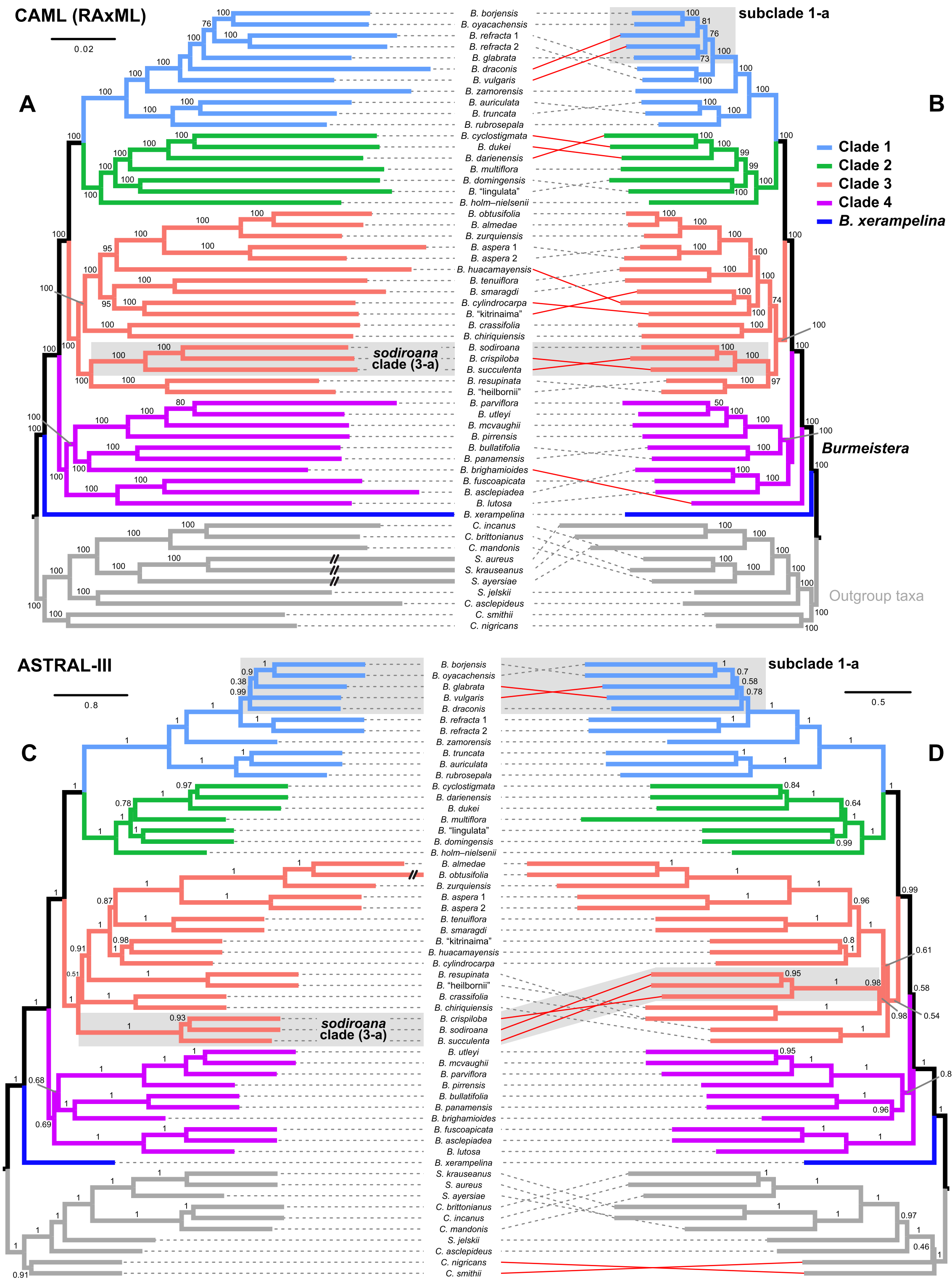

Fig. 2. Tanglegram comparisons of phylogenies from RAxML (Stamatakis, 2014) concatenation + ML (CAML) analyses and ASTRAL-III (Zhang et al., 2018) species tree analyses of the full supercontig dataset (A, C; 542 loci) and exon-only dataset (B, D; 546 loci) for 46 species/lineages of *Burmeistera* plus 10 outgroup species (Table S1). RAxML results include bootstrap proportion (BP; %) support values along nodes and scale bar in units of substitutions/site. ASTRAL species trees are labeled with local posterior probabilities (LPP) and scale bars in coalescent units. Branches are colored for five major lineages we identified, including four major clades (clades 1–4) plus *B. xerampelina*, and red tanglegram lines indicate incongruent tip taxon placements or groupings. Shaded boxes enclose subclades discussed in the text. (PDF). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

declines for 'backbone epoch' nodes comprising the earliest ingroup divergences (~3.15–2 Ma) gave PI ratios that were frequently far above 1 (Fig. S7), indicating much higher potential phylogenetic noise (Townsend and Leuenberger, 2011). By contrast, PI ratios of mainly 1 to 1.5 for the '*sodiroana* epoch' internode (~1.7–1.03 Ma) signified more limited noise potential for mid-crown nodes (Fig. S7).

The noise interpretations above were confirmed by our QIRP results, which revealed generally lower node-resolving power, with greater

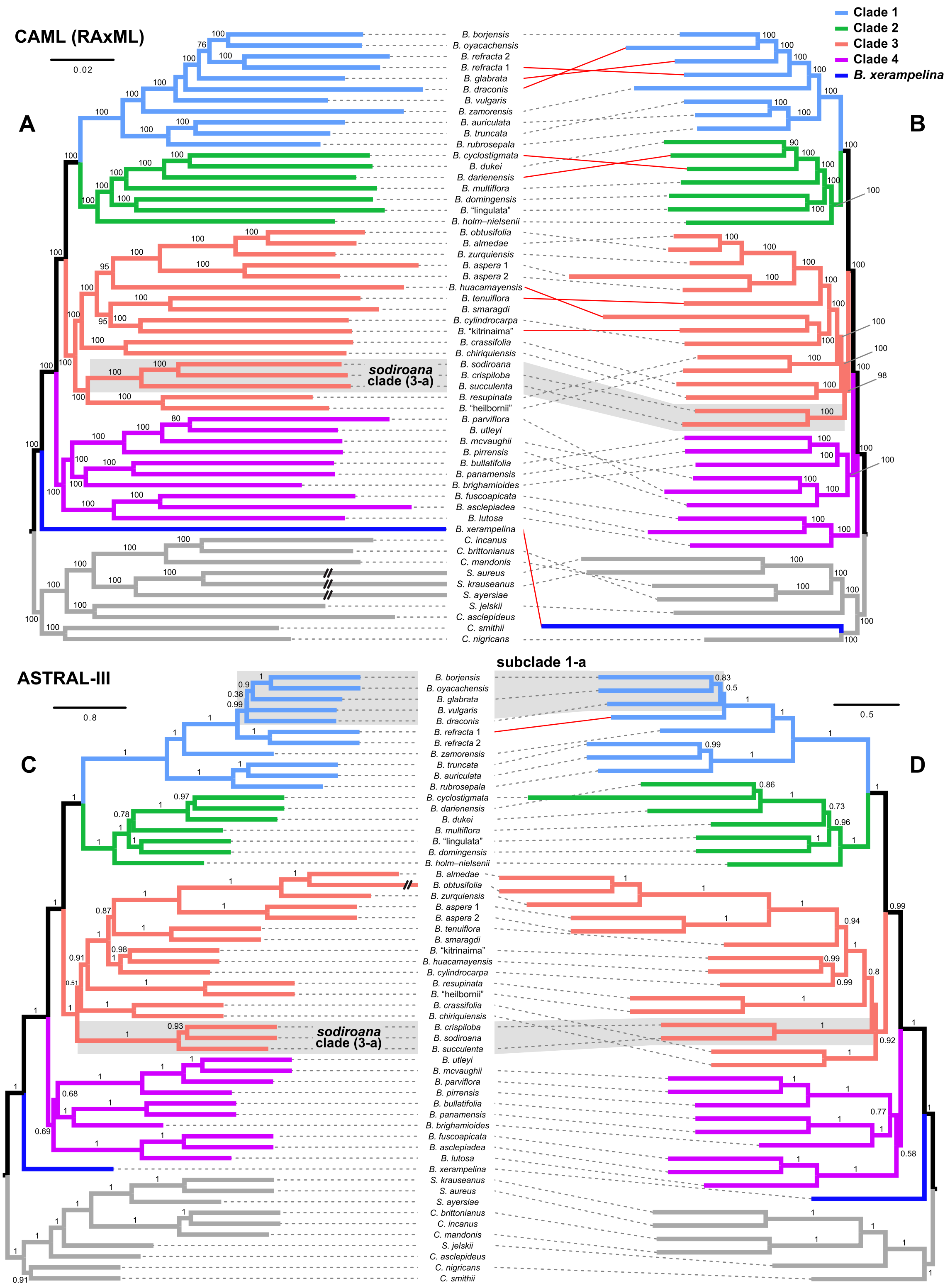

Fig. 3. Tanglegram comparisons of CAML gene tree and ASTRAL species tree results for the full supercontig dataset (A, C; 542 loci) and flanking-only dataset (B, D; 506 loci). Nodal support values and formatting follow details in the caption to Fig. 2. (PDF).

variance, at deeper nodes (Fig. 4). Still, while noise due to homoplasy affected all nuclear loci during early *Burmeistera* diversification, QIRP sensitivity analyses indicated that targeted sequence capture datasets varied in their phylogenetic utility, such that we should much more frequently obtain high support for correct backbone relationships from the supercontig and flanking-only loci than from the exon-only loci (Fig. 5A–C and S8). The same was also true to a lesser degree for the generally longer mid-crown internodes, including that of the *sodiroana* epoch (Fig. 5D–F and S8). But near-universally worse performance of exons at backbone nodes (Fig. 5B and S8) suggested a particularly high risk of obtaining spurious results; hence, incongruent backbone relationships in trees derived from this dataset (e.g. Fig. 2B, D) are more likely to be incorrect.

Null QIRP expectations ranged from 0 to 1 through time, with

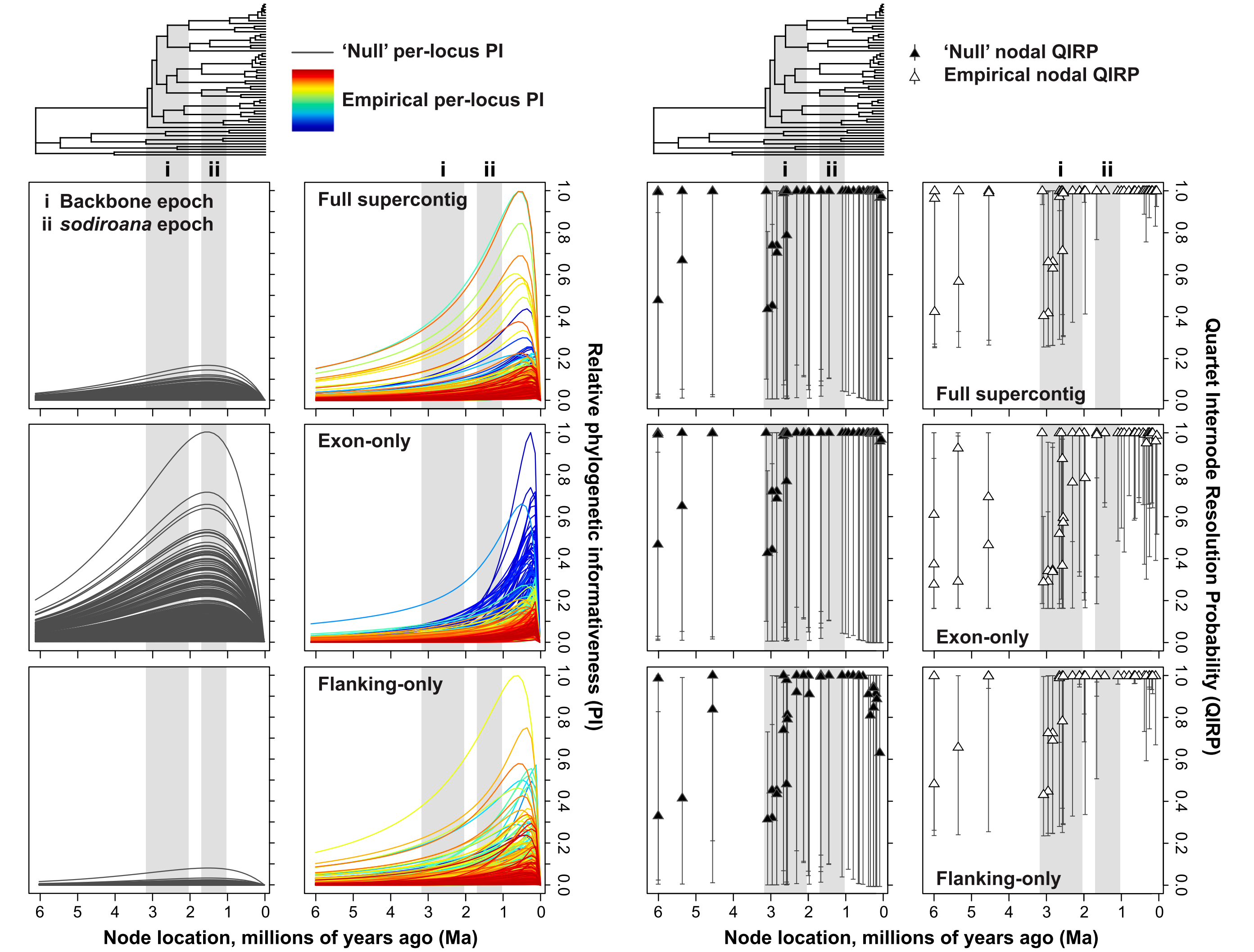

Fig. 4. Phylogenetic informativeness (PI; Townsend, 2007) profiles and node-resolving power estimates (QIRP; Townsend et al., 2012) for individual loci in the full supercontig, exon-only, and flanking-only datasets. Columns 1 and 3 show PI profiles (gray lines, per-locus values) and nodal QIRP values (dark triangles, medians), respectively, for 'dummy' datasets (rates = 1, mimicking empirical data patterns in our alignments) representing expectations under an assumption of equal rates. Columns 2 and 4 show PI profiles (each locus assigned a different color line) and nodal QIRP values (white triangles, medians), respectively, for the empirical datasets. Results are plotted over node locations from the chronogram (Fig. S5), with QIRP values aligned to parent nodes. Two key epochs, (i) the backbone epoch and (ii) the *sodiroana* epoch (discussed in Section 3.3), are demarcated with vertical shaded boxes. (PDF).

median values being more dispersed over the long branch or 'fuse' leading to *Burmeistera* and the ingroup backbone, but consistently near 1 from mid-crown to tips (Fig. 4). Despite ranging slightly higher (~0.2 to 1.0), median empirical QIRPs matched well to expectations, even along the fuse, consistently rising up through the backbone towards median values near 1 for mid- to high-crown nodes. Consistent with higher potential for phylogenetic signal leading to a correct topology in the supercontig and flanking-only loci, median QIRP values for these loci were higher for virtually all ingroup nodes, and especially for backbone epoch and *sodiroana* epoch nodes, than those of exon-only loci. Our analysis evaluating QIRP sensitivity to varying sequence lengths and varying hypothetical internode distances showed that homoplasy or low node-resolving power present substantial difficulties for shorter internodes, shorter loci, and for exon-only loci in particular (Figs. 5 and S8). Node-resolving power was universally worse in the exon-only dataset for all assessed nodes. As expected, the longest concatenated supermatrix alignments had the highest node-resolving power, with QIRPs of 1. Node-resolving power generally increased proportional to sequence length and internode length for full supercontig alignments (skewed towards 1) and peaked at concatenation (Fig. 5A and 5D). Conversely, flanking-only alignments showed the inverse patterns, with worse node-resolving power at greater sequence lengths probably driven by artificially increased homoplasy due to accumulating misalignments; still, the flanking-only data reached peak QIRP at concatenation across all internode lengths (Fig. 5C and 5F). Node-resolving power was universally worse for exon-only loci (Figs. 5 and S8) and not rescued by increasing sequence data except in the concatenation case (Fig. 5B and 5E). These results suggest that exon-only phylogenetic inference may have been misled by low statistical power.

3.4. Treespace visualization of effects of missing taxa

Treespace visualizations of rooted ML gene trees and ASTRAL species trees for the supercontig and exon-only datasets illustrated that missing taxa have had very limited effects on our phylogenetic results (Fig. S9). Gene trees from 100p dataset results mapped as completely nested within the treespace of their corresponding full datasets, indicating that treespace was more similar between datasets from the same rather than different HybPiper assemblies. Species trees from full supercontig versus exon-only analyses were also located close to one another in treespace, as expected given their high topological similarity (Fig. 2C and 2D).

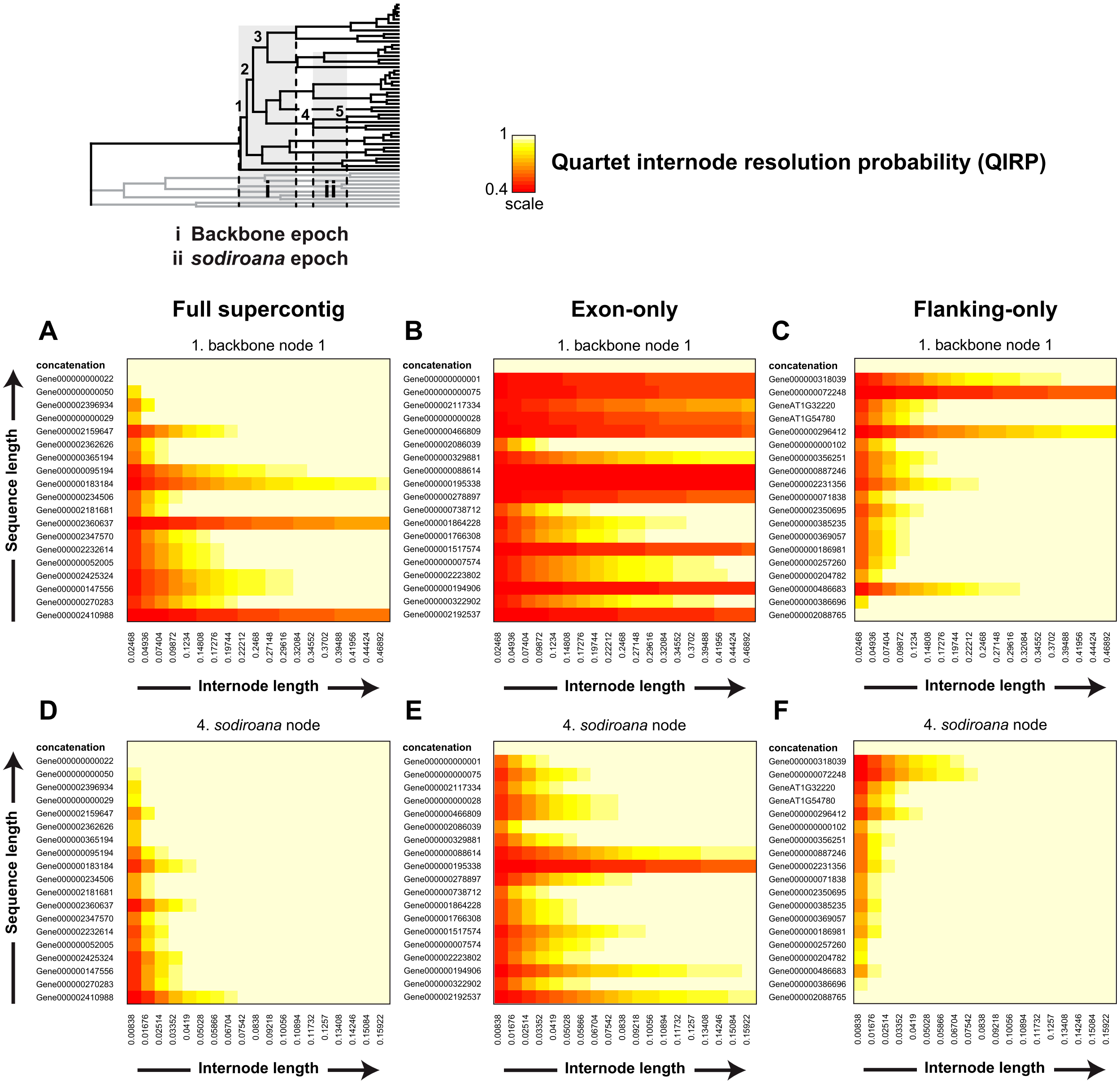

Fig. 5. Sensitivity of node-resolving power (QIRP; Townsend et al., 2012) in the full supercontig (A), exon-only (B), and flanking-only (C) datasets to varying locus lengths and internode distances. Changes in QIRP were evaluated at five focal internal nodes, including three backbone epoch nodes (nodes 1–3) and two *sodiroana* epoch nodes (nodes 4 and 5) labeled on the inset chronograms (see Fig. 4 and Section 3.3). Results for nodes 1 and 4 are plotted here as heatmaps, with rows containing names of 20 alignments with increasing lengths up to concatenation, followed by the corresponding QIRPs over varying hypothetical internode lengths at that node, colored according to increasing probability values (scale bar). For node 2, 3, and 5 results, see Fig. S7. (PDF).

3.5. *Effects of compositional biases*

We found evidence for significant heterogeneity of base frequencies in the supercontig supermatrices (full supercontig dataset: $\chi^2 = 7413.51$, df = 165, $p = 0.00$; 100p supercontig dataset: $\chi^2 = 7386.08$, df = 165, $p = 0.00$). However, this was driven by variable flanking sequence sites, as base frequencies were equal in the exon supermatrices (exon-only dataset: $\chi^2 = 89.70$, df = 165, $p = 0.99$; 100p exon dataset: $\chi^2 = 88.46$, df = 165, $p = 0.99$) but heterogeneous in the flanking-only supermatrix ($\chi^2 = 5832.05$, df = 135, $p = 0.00$; see Table 2 for base frequencies). Optimal CAML trees from analyses of RY-coded versions of the five supermatrices were highly similar to the original CAML trees in all cases (Supplementary Figs. S10–S14), indicating deviations from stationary base frequencies likely have not exerted an undue influence on resolution of topological relationships. Still, the RY-coded topologies had lower bootstrap support for some nodes at short internodes near the bases of clades 2 and 3, in RY-coding analyses of datasets 1, 3, and 5.

3.6. *Assessing and mapping tree congruence and conflict*

Visually and quantitatively summarizing gene tree congruence/conflict at each species tree node revealed *Burmeistera* monophyly was supported by only 32–43 gene trees (~6–8%) with > 33% BP support in main datasets 1, 3, and 5 (Figs. 6 and 7). Other backbone nodes exhibited even higher incongruence, with the most recent common ancestor (MRCA) of all *Burmeistera* excluding *B. xerampelina* supported by 12–23 gene trees, and MRCAs for clades 1–3 supported by 3 or 4 gene trees (Fig. 6). By contrast, mid- to high-crown nodes (< 2 Ma in Fig. S5) showed higher congruence, increasing support up to 153–324 gene trees (~28–60%), including trees derived from shorter and longer loci. The *sodiroana* epoch nodes (subclade 3-a) were supported by

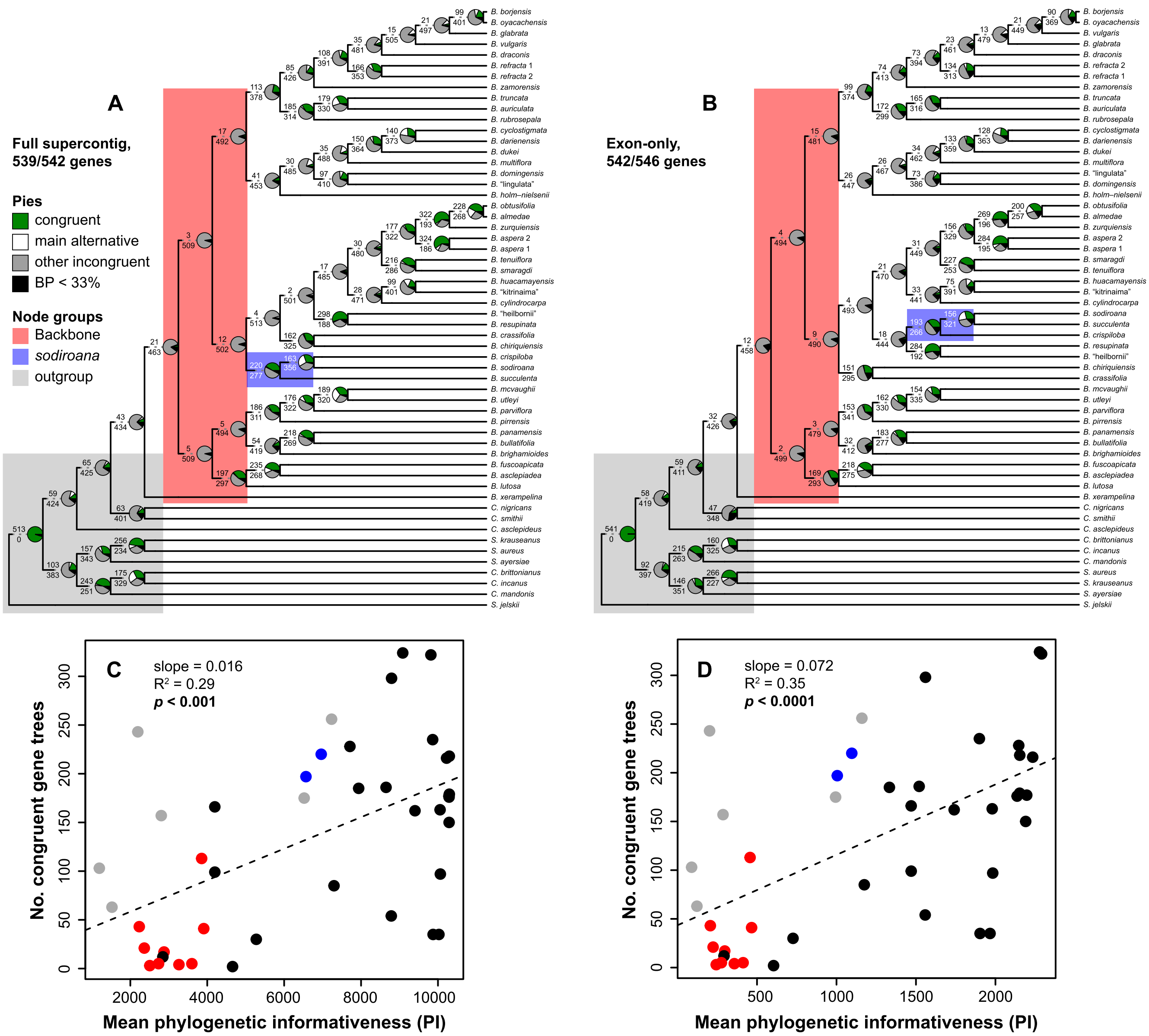

Fig. 6. Gene tree–species tree congruence and conflict, as summarized in PhyParts (Smith et al., 2015). Top panels show full supercontig (A) and exon-only (B) dataset ASTRAL species trees annotated with the number of congruent gene trees (upper branch values) and conflicting gene trees (lower branch values) at each node. Pie charts show the proportion of genes supporting each clade (green), supporting the main alternative topology (white), supporting all other topologies (gray), supporting the clade but having low support with BP < 33% (black). Bottom panels show linear modeling results indicating significant positive relationships between number of congruent gene trees and mean phylogenetic informativeness (PI) for the full supercontig (C) and exon-only (D) datasets. Observed values (dots) are colored according to three tree regions marked with shaded boxes in the top panels. (PDF). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

117–220 gene trees (~23–40%) (Fig. 6). Linear modeling in R revealed significant positive relationships between gene tree congruence and mean nodal PI for full supercontig and exon-only datasets (Fig. 6C and 6D), but the congruence–PI relationship was non-significant for the flanking-only dataset (Fig. 7B). Relationships between congruence and divergence times were greater in magnitude (slope) but inverse in sign, with negative correlations (Fig. S15 and Data S1).

4. Discussion

We assessed the utility of nuclear loci from targeted sequence capture to resolve the phylogeny of *Burmeistera* bellflowers (Lobelioideae, Campanulaceae), which present an ideal test case for rapid, recent angiosperm radiations and associated phylogenetic challenges. Previous studies were unable to resolve the backbone phylogeny of the genus using plastid markers and whole-plastome alignments (e.g. Knox et al., 2008; Lagomarsino et al., 2014, 2016; Uribe-Convers et al., 2017). Our targeted sequence capture dataset is 25-fold larger and much more variable, for example with ~1.1. million parsimony-informative sites out of ~4.2 million bases in the full supercontig dataset (Table 2). Targeted capture success was high, averaging 91%, and large amounts of data were maintained after quality control. While UCEs are typically conserved, with relatively low phylogenetic signal (e.g. Fragoso-Martínez et al., 2017; Molloy and Warnow, 2017; Herrando-Moraira et al., 2018), the > 500 targeted nuclear loci in our final datasets had high phylogenetic signal. The longer and more variable supercontigs (coding plus flanking sequences) and flanking-only loci (Figs. S1 and S2) had even higher phylogenetic signal, yielding higher node-resolving power than the exon-only loci (Figs. 4 and 5). Gene alignments were long and highly complete (usually > 90% taxa), thus

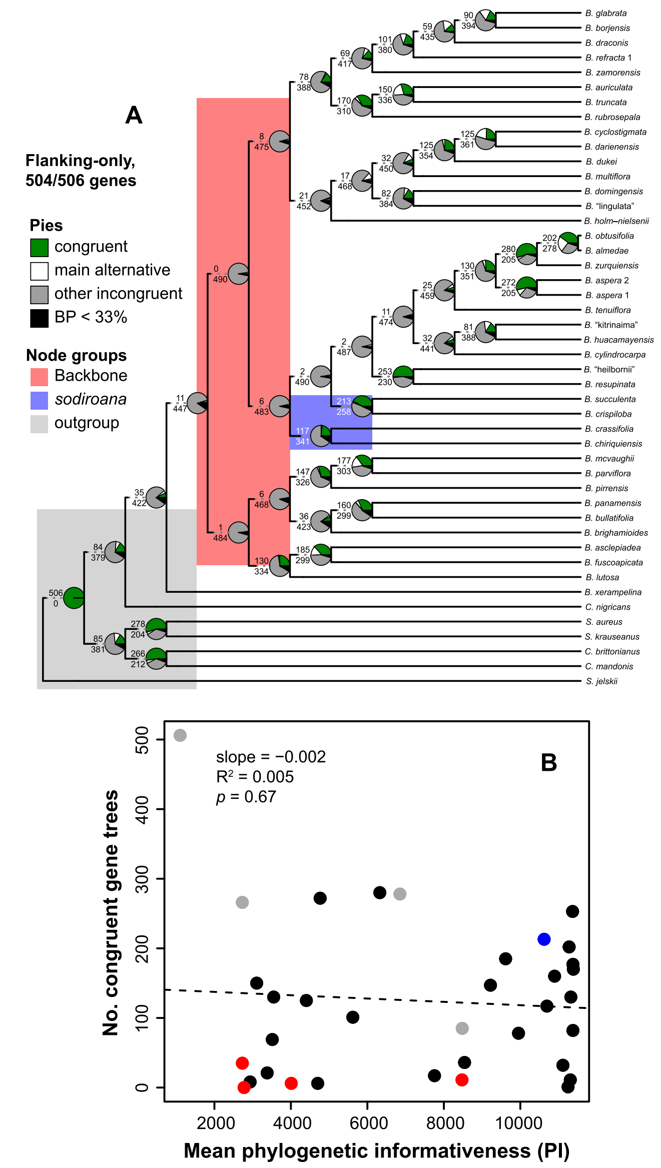

Fig. 7. Gene tree–species tree congruence and conflict results for the flanking-only dataset. The ASTRAL species tree for the flanking-only dataset (A) is shown annotated with information on the numbers of congruent versus conflicting gene trees at each node from PhyParts; formatting and pie chart descriptions follow Fig. 6. The bottom panel shows linear modeling results indicating a non-significant relationship between congruent gene trees and mean PI for the flanking-only dataset (B). Formatting and descriptions for both panels follow that in Fig. 6. (PDF).

amenable to 'total-evidence' concatenation analyses, and they contained sufficient informative sites to estimate gene trees for two-step species tree inference in ASTRAL (Mirarab et al., 2014; Zhang et al., 2018). These desirable properties are less common in datasets from GBS (Elshire et al., 2011) or RAD-seq (Peterson et al., 2012), which also tend to be more sensitive to bioinformatics processing steps (e.g. Eaton and Ree, 2013; Leaché et al., 2015; Harvey et al., 2016). Our results provided the first well-resolved multilocus phylogeny of *Burmeistera* species with substantial taxonomic coverage (36%), including highly supported relationships along internal nodes and backbone nodes. Species trees were largely congruent in the arrangements of major clades across supercontig, exon-only, and flanking-only datasets, and across concatenation approaches versus coalescent-based species tree methods, although internal patterns of evolutionary relationships varied. This incongruence was mainly localized to short branches associated with rapid divergence near the phylogenetic backbone and bases of our major clades, especially within clades 1 and 3 (Figs. 2, 3, S3, and S4). Overall, our results suggest that targeted sequence capture has great potential for resolving relationships in rapid angiosperm radiations, particularly when combining data from exon and non-coding flanking sequences into supercontigs, and provides several advantages over UCE or RAD-seq-based approaches.

4.1. Supercontig and flanking region loci outperform exons

The fact that different characters (e.g. genome regions) and taxon sampling strategies yield varying information content and performance has long fueled debates on phylogenetic experimental design (e.g. Dornburg et al., 2019; Graybeal, 1993; Heath et al., 2008; Townsend et al., 2012; Townsend and Leuenberger, 2011). Recent targeted sequence capture studies of angiosperms have shown considerable interest in the more variable sequences in the "splash zone" flanking targeted exons (e.g. Weitemier et al., 2014; Folk et al., 2015; Johnson et al., 2016; Gernandt et al., 2018; Kates et al., 2018). If these regions do in fact have higher rates of evolution, then they should lead to better phylogenetic resolution, particularly over shallower time scales (e.g. Sang, 2002; Folk et al., 2015). Non-coding flanking sequences should also be less susceptible to selective pressures, such as selection-driven convergence, which can mislead phylogenetic inference (e.g. Castoe et al., 2009). However, ours is the first angiosperm study we are aware of to use quantitative methods for assessing phylogenetic utility to directly test the performance of coding vs. non-coding regions of targeted loci. The few studies to date comparing these marker classes with conventional approaches have provided equivocal results, with similar performance and phylogenetic congruence between supercontig, exon, and flanking loci in some cases (Kates et al., 2018; Gernandt et al., 2018; Villaverde et al., 2018), but less optimal performance of exons in others (Folk et al., 2015).

In this context, our results provide multiple lines of evidence that the longer and more variable supercontig and flanking-only loci outperformed exon-only loci. First, support levels were almost always higher for supercontig and flanking-only trees relative to exon-only trees, in terms of BP support levels in CAML gene trees as well as LPPs from ASTRAL species trees (Figs. 2, 3, S1 and S2). Supercontig and flanking-only datasets also yielded the most congruent ASTRAL species trees, with only 1 case of supercontig–flanking tip incongruence, compared with 7 cases of supercontig–exon-only tip incongruence, albeit there were additional cases of incongruence at internal nodes. Second, median nodal QIRPs were higher for supercontig and flanking-only loci than exon-only loci, especially at backbone nodes and some recent nodes (Fig. 4), although we note that exon PI profiles outperformed by more closely matching null expectations assuming equal intralocus substitution rates. Third, PI ratios indicated sharper declines in PI, or more potential noise, in some exon loci (Fig. S7), and five times as many exon loci had phantom spikes in PI, consistent with elevated homoplasy. Fourth, heatmaps from QIRP sensitivity analyses showed that supercontig and flanking-only alignments universally outperformed over varying sequence lengths and hypothetical internode lengths (Fig. 5). QIRPs are analytically approximated, and represent the probability that a set of characters with known rates ($\lambda_{i\ldots j}$), and state space will estimate the correct topology (Townsend et al., 2012). The high QIRP values for our supercontig and flanking-only topologies correspond well with high congruence between these topologies, and low QIRP values for our exon-only loci agree with low congruence when comparing the exon results to the other topologies (Figs. 2, 3, S1–S4). Thus, our QIRP findings suggest that exon-only loci are far less robust to errors in guide tree topology and branch lengths, and more likely to mislead phylogenetic inference due to low resolution resulting from homoplasy, rather than low phylogenetic signal per se (Townsend et al., 2012; Prum et al., 2015). A final line of evidence for varying performance of different genome regions stems from examining the extent to which the various gene trees agree with each other, as shown in Figs. 6 and 7. In this regard, supercontig and exon-only results did not differ greatly, while flanking-only gene trees showed relatively higher incongruence. Given all of the above, we interpret phylogenetic relationships based on the supercontigs (Fig. 2A and 2C) as our preferred hypotheses, and we consider supercontigs superior to both other data types for phylogenetic inference.

4.2. Robustness to effects of missing taxa and base compositional biases

Missing data and deviations from the standard phylogenetic assumption of a stationary distribution of nucleotide base frequencies represent two important factors known to potentially mislead phylogenetic inference. The problem of missing taxa can contribute to accumulations of homoplasy and other systematic biases on certain parts of the tree, and is also known to contribute to longer branches, potentially causing long-branch attraction (Felsenstein, 1978; reviewed by Heath et al., 2008). In our case, there were only low levels of missing taxa (∼1–10% missingness), and these had limited impact on phylogenetic inference based on our comparisons of our full datasets to corresponding datasets with 100% taxonomic completeness, as demonstrated by the similarity of topological relationships in tanglegrams (Fig. 2, S1–S3) and treespace visualizations using multidimensional scaling (Fig. S4). We also found no evidence for long-branch attraction in our phylogenetic hypotheses, given that taxa with the longest branches (e.g. *B. zamorensis* Muchhala & Perez, *B. huacamayensis*, *B. xerampelina*) were not placed as sister to one another.

Compositional heterogeneity in base frequencies in phylogenomic datasets is correlated with saturation (e.g. Rodríguez-Ezpeleta et al., 2007), and when superimposed on saturated alignments, such compositional biases can lead to incorrect but strongly supported topologies (Dávalos and Perkins, 2008). Whereas the supercontigs and flanking-only datasets in our study exhibited significant base heterogeneity owing to variable sites and differing functional constraints of flanking sequences (Results Section 3.5), analyses of binary matrices with RY-coding effectively normalizing the base frequencies (Ishikawa et al., 2012) showed very minimal effects of base compositional biases on our inferences (Figs. S10–S14). Our RY-coded results should have reduced saturational effects as well; however, topological incongruence between datasets was not completely removed by RY-coding. Another potential issue with the RY-coded results is that they contained slightly lower bootstrap support, typically on short major clade internodes, as well as cases of incongruent groupings not found in the original trees (e.g. low-supported *B. draconis* + *B. vulgaris* relationship in RY-coded ML tree from the full supercontig dataset; Fig. S5B). Lower support in RY-coded matrix analyses is expected to some extent, given that binary coding reduces the number of characters from four to two, but may also reflect a bias such that similarities in base frequencies are inflating bootstrap support values (e.g. Longo et al., 2017).

4.3. Gene tree estimation error and ILS

Coalescent-based methods for species tree inference using gene tree summarization approaches, such as ASTRAL, rely critically on the assumption that gene trees have been correctly estimated (Mirarab et al., 2014; Roch and Warnow, 2015). Poor gene tree estimation may have contributed to poor performance of our exon-only dataset. To assess this possibility, we conducted *a posteriori* CAML analyses of the exon-only dataset in RAxML using only conflicting loci (*cf.* Léveillé-Bourret et al., 2018); that is, loci that disagreed with the main backbone node (the ingroup MRCA) of the original CAML tree. The resulting tree topology (Fig. S16) was nearly identical to the original CAML tree (Fig. 2B). If these conflicting loci had been incorrectly estimated, due to, for example, mutational errors, model mis-specification, or methodological artifacts, then we would expect a tree inferred from these loci to be highly incongruent with the original tree. Results instead strongly suggest that low support values in the exon-only trees reflect ‘hard incongruence’ driven by intrinsic factors such as ILS or introgression, rather than ‘soft incongruence’ due to gene tree estimation error (Léveillé-Bourret et al., 2018). Similar CAML analyses of supercontig loci conflicting at the same node also yielded a CAML tree (Fig. S17) that was identical to the original supercontig CAML tree (Fig. 2A), indicating that supercontig analyses were also not seriously compromised due to gene tree estimation error. While introgression may contribute to the ‘hard incongruence’ in our ASTRAL trees, the fact that we find greater gene tree heterogeneity at shorter backbone and mid-crown branches suggests a central role for ILS, as is expected for rapid, recent radiations with short internodes (e.g. Maddison, 1997; Whitfield and Lockhart, 2007; Brito and Edwards, 2009).

4.4. Monophyly and species relationships within Burmeistera

Monophyly of *Burmeistera* was definitively supported across our analyses, in agreement with previous molecular studies (Knox et al., 2008; Lagomarsino et al., 2014, 2016; Uribe-Convers et al., 2017). Our results agree with Lagomarsino et al. (2014) and Uribe-Convers et al. (2017) in showing polyphyly of the two taxonomic sections of *Burmeistera* previously recognized by Wimmer (1943) based on morphological differentiation in anther pubescence. In view of available presence/absence data on anther pubescence (tufted hairs on the ventral two anthers) in *Burmeistera* (Uribe-Convers et al., 2017; Mashburn, 2019), the potential monophyly of Wimmer’s (1943) section *Barbatae* E. Wimm. is invalidated in our study by the placement of *B. parviflora* with pubescent anthers as sister to either *B. utleyi*, which lacks pubescent anthers (BP = 50–80), or to a clade comprised of *B. utleyi* (pubescent anthers absent) + *B. mcvaughii* (pubescent anthers present) (LPP = 1; Figs. 2 and 3). Similar to Uribe-Convers et al. (2017), taxa forming our clade 1 are characterized as lacking pubescent anthers, and taxa in the *sodiroana* group consistently form a highly supported monophyletic group with recurved petals, indicating that these characters may be reliable synapomorphies for these clades. Our results also provide important clarification of recalcitrant relationships within *Burmeistera*. In particular, Uribe-Convers et al.’s (2017) clade “D” is shown herein to be non-monophyletic, placements of their clades “A” and “B” are clarified, and we were also able to confidently place all taxa that collapsed into a polytomy within their clade D into our clades 2 and 3, with high support, although with some topological variations (see additional details and discussion of phylogenetic results in Supplementary Data S1).

4.5. Practical recommendations for phylogenomic studies of rapid angiosperm radiations

Overall, our results justify several practical recommendations for future phylogenomic studies of angiosperm radiations based on Hyb-Seq and related targeted enrichment approaches. First, supercontigs should be assembled and analyzed, rather than solely the coding sequences, in order to maximize the phylogenetic utility and node-resolving power of loci. In the context of other recent targeted sequence capture studies reviewed above and in Table 1, it seems that the utility of flanking sequences, themselves, may prove to be lineage-specific (see additional discussion in Data S1). Nevertheless, our findings show that taking the extra time and effort to extract and align flanking sequences and combine them with exons using recently developed techniques that detect and account for intron-exon junctions [e.g. Exonerate analyses, as incorporated into pipelines by Weitemier et al. (2014) and Johnson et al. (2016)] can be highly valuable for improving phylogenetic inference in rapid, recent angiosperm radiations. Second, we recommend procedures to reduce missing-data effects on phylogenetic inference, which, although only partly tested in this study, may be important for improving accuracy in supercontig analyses. We recommend quality-filtering steps similar to those implemented here to remove taxa and alignment positions with large amounts of missing data (> 50%

missing), as supercontigs may not outperform exons under other conditions. Third, after obtaining supercontig and exon alignments, we encourage applications of similar PI analyses to those herein (e.g. combining profiling and QIRP estimation; Dornburg et al., 2016, 2019; Prum et al., 2015; Townsend et al., 2012) in order to validate whether adding flanking sequences does, in fact, improve node-resolving power in other specific use cases. Finally, our study leaves a great deal of room for additional work to determine the impact of data-filtering strategies on the accuracy of phylogenetic inferences under concatenation and coalescent-based approaches in phylogenomic studies of rapid, recent angiosperm radiations. We recommend that future studies build on the present work to determine which combination(s) of supercontigs and data filtration strategies (e.g. removing fast-evolving sites producing phantom spikes in PI profiles) best maximize phylogenetic signal while reducing phylogenetic noise (reviewed by Fragoso-Martínez et al., 2017; Herrando-Moraira et al., 2018).

5. Conclusion

We tested the phylogenetic utility of targeted sequence capture to resolve the phylogeny of rapid, recent angiosperm radiations, using *Burmeistera* bellflowers as a test case representative of high Andean speciation rates. We targeted 745 low-copy nuclear loci, successfully generating supercontigs of exon and flanking sequences for up to 681 these, with an average 91% capture success. This allowed us to infer the first well-supported multilocus phylogenetic hypothesis for a substantial coverage (~36%) of *Burmeistera* species, with results demonstrating overarching congruence in support of four major clades and emerging patterns for potential higher-level synapomorphies. Our study adds to a burgeoning literature illustrating the feasibility of targeted sequence capture for phylogenomics, extending the approach for the first time to flowering plants from the Tropical Andes biodiversity hotspot (Myers et al., 2000). Results were robust to effects of missing taxa and base compositional biases. However, we also found instances of topological incongruence between supercontig, exon-only, and flanking-only datasets, as well as widespread underlying gene tree heterogeneity in each of these datasets. Detailed data interrogation, including PI profiling and sensitivity analyses plus additional phylogenetic analyses, strongly indicated that this incongruence was due to homoplasy and low node-resolving power of shorter loci, particularly for deeper nodes and shorter branches, rather than low phylogenetic utility or gene tree estimation error. For all datasets, node-resolving power was rescued by concatenation. Exon-only datasets consistently performed worse than the ~4-fold longer and more variable supercontig loci. Our study suggests that targeted sequence capture can overcome the significant challenges for phylogenetic inference in rapid, recent angiosperm radiations, particularly when using supercontig loci that combine exon and flanking sequences, and provides several advantages over UCE or RAD-seq-based approaches. Additional angiosperm studies using similar assessments of phylogenetic utility will allow us to test the generality of many details here, such as the low bootstrap support in RY-coded analyses, the lower QIRP values for exon-only data, and the high support in CAML results obscuring underlying gene tree heterogeneity, and we predict that supercontig assembly will be comparably effective in other rapid angiosperm radiations.

6. Data accessibility

Raw sequence reads generated during this study, including transcriptome and nuclear genome sequences, have been accessioned in the NCBI Short Read Archive database and transcriptome assembly files have been accessioned in the NCBI Transcriptome Shotgun Assembly database. These data have been deposited with links to BioProject accession numbers PRJNA646974 and PRJNA623031 in the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/). Target gene sequences, analysis code, phylogenetic alignments, and additional information are made available through a Mendeley Data accession (doi: 10.17632/wsbjwr3p42.1).

CRediT authorship contribution statement

Justin C. Bagley: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Supervision, Validation, Visualization, Writing - original draft. **Simon Uribe-Convers:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Writing - review & editing. **Mónica M. Carlsen:** Investigation, Methodology, Software, Writing - review & editing. **Nathan Muchhala:** Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Writing - review & editing. All authors have read and approved this version of the manuscript for submission.

Acknowledgements

This research was supported by funding from U.S. National Science Foundation grant DEB- 1754802 to NM. We thank the numerous people who participated in field sampling or provided access to herbarium specimens, especially John L. Clark and Laura Lagomarsino. We are also grateful to the Virginia Commonwealth University Center for High Performance Computing (CHiPC) and the Brigham Young University Office of Research Computing for providing generous computational resources used during our analyses.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ympev.2020.106769.

References

Abrahamczyk, S., Souto-Vilaros, D., Renner, S.S., 2014. Escape from extreme specialization: passionflowers, bats and the sword-billed hummingbird. Proc. R. Soc. B. 281, 20140888.

Amenta, N., St. John, K., Klingner, J., Maddison, W., Maddison, D., Clarke, F., Edwards, D., Guzman, D., Mahindru, R., Ivanov, P., Prabhum, U., Postarnakevich, N., Heath, T., Hillis, D., 2012. Tree Set Visualization: a package for Mesquite. Version 3.0.

Andrews, K.R., Good, J.M., Miller, M.R., Luikart, G., Hohenlohe, P.A., 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. Nat. Rev. Genet. 17 (2), 81–92.

Antonelli, A., 2008. Higher level phylogeny and evolutionary trends in Campanulaceae subfam. Lobelioideae: molecular signal overshadows morphology. Mol. Phylogenet. Evol. 46, 1–18.

Avise, J.C., 2004. Molecular Markers, Natural History, and Evolution, second ed. Sinauer Associates, Sunderland, MA.

Bagley, J.C., 2019. PIrANHA v0.3a2. GitHub repository, Available at: < https://github.com/justincbagley/piranha > .

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19 (5), 455–477.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., Haussler, D., 2004. Ultraconserved elements in the human genome. Science 304, 1321–1325.

Bell, C.D., Soltis, D.E., Soltis, P.S., 2010. The age and diversification of the angiosperms re-revisited. Am. J. Bot. 97, 1296–1303.

Brito, P.H., Edwards, S.V., 2009. Multilocus phylogeography and phylogenetics using sequence-based markers. Genetica 135 (3), 439–455.

Castoe, T.A., de Koning, A.J., Kim, H.M., Gu, W., Noonan, B.P., Naylor, G., Jiang, Z.J., Parkinson, C.L., Pollock, D.D., 2009. Evidence for an ancient adaptive episode of convergent molecular evolution. Proc. Natl. Acad. Sci. USA 106 (22), 8986–8991.

Chamala, S., García, N., Godden, G.T., Krishnakumar, V., Jordon-Thaden, I.E., De Smet, R., Barbazuk, W.B., Soltis, D.E., Soltis, P.S., 2015. MarkerMiner 1.0: A new application for phylogenetic marker development using angiosperm transcriptomes. Appl. Plant Sci. 3 (4), 1400115.

Chau, J.H., Rahfeldt, W.A., Olmstead, R.G., 2018. Comparison of taxon-specific versus general locus sets for targeted sequence capture in plant phylogenomics. Appl. Plant Sci. 6 (3), e1032.

Cronn, R., Knaus, B.J., Liston, A., Maughan, P.J., Parks, M., et al., 2012. Targeted enrichment strategies for next-generation plant biology. Am. J. Bot. 99, 291–311.

Crowl, A.A., Myers, C., Cellinese, N., 2017. Embracing discordance: phylogenomic analyses provide evidence for allopolyploidy leading to cryptic diversity in a Mediterranean *Campanula* (Campanulaceae) clade. Evolution 71 (4), 913–922.

Dávalos, L.M., Perkins, S.L., 2008. Saturation and base composition bias explain phylogenomic conflict in *Plasmodium*. Genomics 91, 433–442.

de Queiroz, A., Gatesy, J., 2007. The supermatrix approach to systematics. Trends Ecol. Evol. 22, 34–41.

De Smet, R., Adams, K.L., Vandepoele, K., Van Montagu, M.C.E., Maere, S., Van de Peer, Y., 2013. Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. Proc. Natl. Acad. Sci. USA 110, 2898–2903.

Degnan, J.H., Rosenberg, N.A., 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24 (6), 332–340.

Dornburg, A., Fisk, J.N., Tamagnan, J., Townsend, J., 2016. PhyInformR: Rapid Calculation of Phylogenetic Information Content. R package version 1.0. Available at: < https://CRAN.R-project.org/package=PhyInformR > .

Dornburg, A., Su, Z., Townsend, J.P., 2019. Optimal rates for phylogenetic inference and experimental design in the era of genome-scale data sets. Syst. Biol. 68 (1), 145–156.

Doyle, J.J., Doyle, J.L., 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochemical Bull. 19, 11–15.

Eaton, D.A., Ree, R.H., 2013. Inferring phylogeny and introgression using RADseq data: an example from flowering plants (*Pedicularis*: Orobanchaceae). Syst. Biol. 62 (5), 689–706.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., Mitchell, S.E., 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PloS One 6 (5), e19379.

Faircloth, B.C., McCormack, J.E., Crawford, N.G., Harvey, M.G., Brumfield, R.T., Glenn, T.C., 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Syst. Biol. 61, 717–726.

Felsenstein, J., 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Biol. 27, 401–410.

Folk, R.A., Mandel, J.R., Freudenstein, J.V., 2015. A protocol for targeted enrichment of intron-containing sequence markers for recent radiations: a phylogenomic example from *Heuchera* (Saxifragaceae). Appl. Plant Sci. 3 (8), 1500039.

Folk, R.A., Stubbs, R.L., Mort, M.E., Cellinese, N., Allen, J.M., Soltis, P.S., Soltis, D.E., Guralnick, R.P., 2019. Rates of niche and phenotype evolution lag behind diversification in a temperate radiation. Proc. Natl. Acad. Sci. USA 116 (22), 10874–10882.

Gibbard, P.L., Head, M.J., Walker, M.J., Subcommission on Quaternary Stratigraphy., 2010. Formal ratification of the Quaternary system/period and the Pleistocene series/epoch with a base at 2.58 Ma. J. Quaternary Sci. 25(2), 96–102.

Fragoso-Martínez, I., Salazar, G.A., Martínez-Gordillo, M., Magallón, S., Sánchez-Reyes, L., Lemmon, E.M., Lemmon, A.R., Sazatornil, F., Mendoza, C.G., 2017. A pilot study applying the plant Anchored Hybrid Enrichment method to New World sages (Salvia subgenus Calosphace; Lamiaceae). Mol. Phylogenet. Evol. 117 (2017), 124–134.

Gernandt, D.S., Aguirre Dugua, X., Vázquez-Lobo, A., Willyard, A., Moreno Letelier, A., Pérez de la Rosa, J.A., Piñero, D., Liston, A., 2018. Multi-locus phylogenetics, lineage sorting, and reticulation in Pinus subsection Australes. Am. J. Bot. 105 (4), 711–725.

Givnish, T.J., 1997. Adaptive radiation and molecular systematics: issues and approaches. In: Givnish, T., Sytsma, K. (Eds.), Molecular Evolution and Adaptive Radiation. Cambridge University Press, Cambridge, UK, pp. 1–54.

Givnish, T.J., 2015. Adaptive radiation versus 'radiation' and 'explosive diversification': why conceptual distinctions are fundamental to understanding evolution. New Phytol. 207 (2), 297–303.

Givnish, T.J., Barfuss, M.H.J., Van Ee, B., Riina, R., Schulte, K., Horres, R., Gonsiska, P.A., Jabaily, R.S., Crayn, D.M., Smith, J.A.C., et al., 2014. Adaptive radiation, correlated and contingent evolution, and net species diversification in Bromeliaceae. Mol. Phylogenet. Evol. 71, 55–78.

Graur, D., Li, W.-H., 2000. Fundamentals of Molecular Evolution. Sinauer Associates, Sunderland, MA.

Graybeal, A., 1993. The phylogenetic utility of cytochrome *b*: lessons from bufonid frogs. Mol. Phylogenet. Evol. 2 (3), 256–269.

Harvey, M.G., Smith, B.T., Glenn, T.C., Faircloth, B.C., Brumfield, R.T., 2016. Sequence capture versus restriction site associated DNA sequencing for shallow systematics. Syst. Biol. 65 (5), 910–924.

Heath, T.A., Hedtke, S.M., Hillis, D.M., 2008. Taxon sampling and the accuracy of phylogenetic analyses. J. Syst. Evol. 46 (3), 239–257.

Herrando-Moraira, S., Calleja, J.A., Carnicero, P., Fujikawa, K., Galbany-Casals, M., Garcia-Jacas, N., Im, H.T., Kim, S.C., Liu, J.Q., Lopez-Alvarado, J., López-Pujol, J., 2018. Exploring data processing strategies in NGS target enrichment to disentangle radiations in the tribe Cardueae (Compositae). Mol. Phylogenet. Evol. 128, 69–87.

Hillis, D.M., Heath, T.A., John, K.S., 2005. Analysis and visualization of tree space. Syst. Biol. 54 (3), 471–482.

Hudson, R.R., 1990. Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol. 7, 1–44.

Hughes, C.E., Nyffeler, R., Linder, H.P., 2015. Evolutionary plant radiations: where, when, why and how? New Phytol. 207 (2), 249–253.

Hwang, J., Jonathan, H., Qi, Z., Yang, Z.L., Zheng, W., Townsend, J.P., 2015. Solving the ecological puzzle of mycorrhizal associations using data from annotated collections and environmental samples–an example of saddle fungi. Environ. Microbiol. Rep. 7, 658–667.

Ishikawa, S.A., Inagaki, Y., Hashimoto, T., 2012. RY-coding and non-homogeneous models can ameliorate the maximum-likelihood inferences from nucleotide sequence data with parallel compositional heterogeneity. Evol. Bioinforma. 8, 357–371.

Johnson, M.G., Gardner, E.M., Liu, Y., Medina, R., Goffinet, B., Shaw, A.J., Zerega, N.J.C., Wickett, N.J., 2016. HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. Appl. Plant Sci. 4, 1600016.

Johnson, M.G., Pokorny, L., Dodsworth, S., Botigue, L.R., Cowan, R.S., Devault, A., Eiserhardt, W.L., Epitawalage, N., Forest, F., Kim, J.T., Leebens-Mack, J.H., 2019. A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. Systemat. Biol. 68 (4), 594–606.

Jones, K.E., Fér, T., Schmickl, R.E., Dikow, R.B., Funk, V.A., Herrando-Moraira, S., Johnston, P.R., Kilian, N., Siniscalchi, C.M., Susanna, A., Slovák, M., 2019. An empirical assessment of a single family-wide hybrid capture locus set at multiple evolutionary timescales in Asteraceae. Appl. Plant Sci. 7 (10), e11295.

Jukes, T.H., Cantor, C.R., 1969. Evolution of protein molecules. In: Munro, H. (Ed.), Mammalian Protein Metabolism. Academic Press, New York, NY, pp. 21–132.

Junier, T., Zdobnov, E.M., 2010. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. Bioinformatics 26 (13), 1669–1670.

Kates, H.R., Johnson, M.G., Gardner, E.M., Zerega, N.J., Wickett, N.J., 2018. Allele phasing has minimal impact on phylogenetic reconstruction from targeted nuclear gene sequences in a case study of *Artocarpus*. Am. J. Bot. 105 (3), 404–416.

Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30 (4), 772–780.

Knox, E.B., Muasya, A.M., Muchhala, N., 2008. The predominantly South American clade of Lobeliaceae. Syst. Bot. 33, 462–468.

Kumar, S., Stecher, G., Tamura, K., 2016. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 33 (7), 1870–1874.

Lagomarsino, L.P., Antonelli, A., Muchhala, N., Timmermann, A., Mathews, S., Davis, C.C., 2014. Phylogeny, classification, and fruit evolution of the species-rich Neotropical bellflowers (Campanulaceae: Lobelioideae). Am. J. Bot. 101, 2097–2112.

Lagomarsino, L.P., Condamine, F.L., Antonelli, A., Mulch, A., Davis, C.C., 2016. The abiotic and biotic drivers of rapid diversification in Andean bellflowers (Campanulaceae). New Phytol. 210, 1430–1442.

Lagomarsino, L.P., Forrestel, E.J., Muchhala, N., Davis, C.C., 2017. Repeated evolution of vertebrate pollination syndromes in a recently diverged Andean plant clade. Evolution 71 (8), 1970–1985.

Lammers, T.G., 2007. World Checklist and Bibliography of Campanulaceae. Royal Botanical Gardens, Kew, UK.

Leaché, A.D., Chavez, A.S., Jones, L.N., Grummer, J.A., Gottscho, A.D., Linkem, C.W., 2015. Phylogenomics of phrynosomatid Lizards: conflicting signals from sequence capture versus restriction site associated DNA sequencing. Genome Biol. Evol. 7 (3), 706–719.

Leaché, A.D., Rannala, B., 2011. The accuracy of species tree estimation under simulation: a comparison of methods. Syst. Biol. 60 (2), 126–137.

Lemmon, E.M., Lemmon, A.R., 2013. High-throughput genomic data in systematics and phylogenetics. Annu. Rev. Ecol. Evol. Syst. 44, 99–121.

Léveillé-Bourret, É., Starr, J.R., Ford, B.A., Moriarty Lemmon, E., Lemmon, A.R., 2018. Resolving rapid radiations within angiosperm families using anchored phylogenomics. Syst. Biol. 67 (1), 94–112.

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25 (14), 1754–1760.

Liu, L., Yu, L., Edwards, S.V., 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol. Biol. 10, 302.

Longo, S.J., Faircloth, B.C., Meyer, A., Westneat, M.W., Alfaro, M.E., Wainwright, P.C., 2017. Phylogenomic analysis of a rapid radiation of misfit fishes (Syngnathiformes) using ultraconserved elements. Mol. Phylogenet. Evol. 113, 33–48.

McKain, M.R., Johnson, M.G., Uribe-Convers, S., Eaton, D., Yang, Y., 2018. Practical considerations for plant phylogenomics. Appl. Plant Sci. 6 (3), e1038.

Maddison, W., 1997. Gene trees in species trees. Syst. Biol. 46, 523–536.

Maddison, W.P., Maddison, D.R., 2018. Mesquite: a modular system for evolutionary analysis. Version 3.51. Available at: < http://www.mesquiteproject.org > .

Mandel, J.R., Dikow, R.B., Funk, V.A., 2015. Using phylogenomics to resolve mega-families: an example from Compositae. J. System. Evol. 53, 391–402.

Mandel, J.R., Dikow, R.B., Funk, V.A., Masalia, R.R., Staton, S.E., Kozik, A., Michelmore, R.W., Rieseberg, L.H., Burke, J.M., 2014. A target enrichment method for gathering phylogenetic information from hundreds of loci: an example from the Compositae. Appl. Plant Sci. 2, 1300085.

Mashburn, B., 2019. Taxonomic Revision of the *Burmeistera* (Campanulaceae) of Ecuador. Master's Thesis. University of Missouri-St. Louis, St, Louis, MO.

Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T., Swenson, M.S., Warnow, T., 2014. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30, i541–i548.

Mirarab, S., Warnow, T., 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics 31 (12), i44–i52.

Molloy, E.K., Warnow, T., 2017. To include or not to include: the impact of gene filtering on species tree estimation methods. Syst. Biol. 67 (2), 285–303.

Muchhala, N., 2006. The pollination biology of *Burmeistera* (Campanulaceae): specialization and syndromes. Am. J. Bot. 93, 1081–1089.

Muchhala, N., 2008. Functional significance of extensive interspecific variation in *Burmeistera* floral morphology: evidence from nectar bat captures in Ecuador. Biotropica 40, 332–337.

Muchhala, N., Potts, M.D., 2007. Character displacement among bat-pollinated flowers of the genus *Burmeistera*: analysis of mechanism, process and pattern. Proc. R. Soc. B. 274, 2731–2737.

Myers, N., Mittermeier, R.A., Mittermeier, C.G., Da Fonseca, G.A., Kent, J., 2000. Biodiversity hotspots for conservation priorities. Nature 403 (6772), 853–858.

Paradis, E., Schliep, K., 2018. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35 (3), 526–528.

Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S., Hoekstra, H.E., 2012. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS One 7, e37135.

Pond, S.L.K., Muse, S.V., 2005. HyPhy: hypothesis testing using phylogenies. In: Statistical Methods in Molecular Evolution. Springer, New York, NY, pp. 125–181.

Prum, R.O., Berv, J.S., Dornburg, A., Field, D.J., Townsend, J.P., Lemmon, E.M., et al., 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation

DNA sequencing. Nature 526, 569–573.

R Core Team, 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: < https://www.R-project.org/ > .

Reneker, J., Lyons, E., Conant, G.C., Pires, J.C., Freeling, M., Shyu, C.-R., Korkin, D., 2012. Long identical multispecies elements in plant and animal genomes. Proc. Natl. Acad. Sci. USA 109, E1183–E1191.

Ressayre, A., Glémin, S., Montalent, P., Serre-Giardi, L., Dillmann, C., Joets, J., 2015. Introns structure patterns of variation in nucleotide composition in *Arabidopsis thaliana* and rice protein-coding genes. Genome Biol. Evol. 7 (10), 2913–2928.

Roch, S., Steel, M., 2015. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. Theor. Popul. Biol. 100, 56–62.

Roch, S., Warnow, T., 2015. On the robustness to gene tree estimation error (or lack thereof) of coalescent-based species tree methods. Syst. Biol. 64 (4), 663–676.

Rodríguez-Ezpeleta, N., Brinkmann, H., Roure, B., Lartillot, N., Lang, B.F., Philippe, H., 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. Syst. Biol. 56, 389–399.

Rokas, A., Williams, B.L., King, N., Carroll, S.B., 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425 (6960), 798–804.

Salichos, L., Rokas, A., 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497 (7449), 327.

Sanderson, M.J., 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Mol. Biol. Evol. 19 (1), 101–109.

Sang, T., 2002. Utility of low-copy nuclear gene sequences in plant phylogenetics. Cr. Rev. Biochem. Mol. Biol. 37, 121–147.

Sayyari, E., Mirarab, S., 2016. Fast coalescent-based computation of local branch support from quartet frequencies. Mol. Biol. Evol. 33 (7), 1654–1668.

Schluter, D., 2000. The Ecology of Adaptive Radiation. Oxford University Press, Oxford, UK.

Schmickl, R., Liston, A., Zeisek, V., Oberlander, K., Weitemier, K., Straub, S.C., Cronn, R.C., Dreyer, L.L., Suda, J., 2015. Phylogenetic marker development for target enrichment from transcriptome and genome skim data: the pipeline and its application in southern African *Oxalis* (Oxalidaceae). Mol. Ecol. Resour. 16 (5), 1124–1135.

Simmons, M.P., Carr, T.G., O'Neill, K., 2004. Relative character-state space, amount of potential phylogenetic information, and heterogeneity of nucleotide and amino acid characters. Mol. Phylogenet. Evol. 32, 913–926.

Simões, M., Breitkreuz, L., Alvarado, M., Baca, S., Cooper, J.C., Heins, L., Herzog, K., Lieberman, B.S., 2016. The evolving theory of evolutionary radiations. Trends Ecol. Evol. 31 (1), 27–34.

Slater, G., Birney, E., 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6, 31.

Smith, S.A., Dunn, C.W., 2008. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. Bioinformatics 24, 715–716.

Smith, S.A., Moore, M.J., Brown, J.W., Yang, Y., 2015. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. BMC Evolut. Biol. 15 (1), 150.

Soltis, P.S., Folk, R.A., Soltis, D.E., 2019. Darwin review: angiosperm phylogeny and evolutionary radiations. Proc. R. Soc. B 286 (1899), 20190099.

Soltis, P.S., Soltis, D.E., Chase, M.W., Endress, P.K., Crane, P.R., 2004. The diversification of flowering plants. In: Cracraft, J., Donoghue, M. (Eds.), The Tree of Life. Oxford University Press, New York, NY, pp. 154–167.

Spalink, D., Drew, B.T., Pace, M.C., Zaborsky, J.G., Starr, J.R., Cameron, K.M., Givnish, T.J., Sytsma, K.J., 2016. Biogeography of the cosmopolitan sedges (Cyperaceae) and the area-richness correlation in plants. J. Biogeogr. 43, 1893–1904.

Spielman, S.J., Kosakovsky Pond, S.L., 2018. Relative evolutionary rate inference in HyPhy with LEISR. PeerJ 6, e4339.

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., Moore, M.J., Soltis, P.S., Soltis, D.E., Liston, A., Livshultz, T., 2014. Phylogenetic signal detection from an ancient rapid radiation: effects of noise reduction, long-branch attraction, and model selection in crown clade Apocynaceae. Mol. Phylogenet. Evol. 80, 169–185.

Straub, S.C., Parks, M., Weitemier, K., Fishbein, M., Cronn, R.C., Liston, A., 2012. Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. Am. J. Bot. 99, 349–364.

Stubbs, R.L., Folk, R.A., Xiang, C.-L., Soltis, D.E., Cellinese, N., 2018. Pseudo-parallel patterns of disjunctions in an Arctic-alpine plant lineage. Mol. Phylogenet. Evol. 123, 88–100.

Swofford, D.L., 2002. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, MA.

Timme, R.E., Bachvaroff, T.R., Delwiche, C.F., 2012. Broad phylogenetic sampling and the sister lineage of land plants. PLoS One 7, e29696.

Townsend, J.P., 2007. Profiling phylogenetic informativeness. Syst. Biol. 56, 222–231.

Townsend, J.P., Leuenberger, C., 2011. Taxon sampling and the optimal rates of evolution for phylogenetic inference. Syst. Biol. 60, 358–365.

Townsend, J.P., López-Giráldez, F., Friedman, R., 2008. The phylogenetic informativeness of nucleotide and amino acid sequences for reconstructing the vertebrate tree. J. Mol. Evol. 67 (5), 437–447.

Townsend, J.P., Su, Z., Tekle, Y.I., 2012. Phylogenetic signal and noise: predicting the power of a data set to resolve phylogeny. Syst. Biol. 61, 835–849.

Turner, E.H., Ng, S.B., Nickerson, D.A., Shendure, J., 2009. Methods for genomic partitioning. Annu. Rev. Genomics Hum. Genet. 10, 264–284.

Uribe-Convers, S., Tank, D.C., 2015. Shifts in diversification rates linked to biogeographic movement into new areas: an example of disparate continental distributions and a recent radiation in the Andes. Am. J. Bot. 102, 1854–1869.

Uribe-Convers, S., Settles, M.L., Tank, D.C., 2016. A phylogenomic approach based on PCR target enrichment and high throughput sequencing: resolving the diversity within the South American species of *Bartsia* L. (Orobanchaceae). PLoS One 11 (2), e0148203.

Uribe-Convers, S., Carlsen, M.M., Lagomarsino, L.P., Muchhala, N., 2017. Phylogenetic relationships of *Burmeistera* (Campanulaceae: Lobelioideae): combining whole plastome with targeted loci data in a recent radiation. Mol. Phylogenet. Evol. 107, 551–563.

Vatanparast, M., Powell, A., Doyle, J.J., Egan, A.N., 2018. Targeting legume loci: a comparison of three methods for target enrichment bait design in Leguminosae phylogenomics. Appl. Plant Sci. 6 (3), e1036.

Villaverde, T., Pokorny, L., Olsson, S., Rincón-Barrado, M., Johnson, M.G., Gardner, E.M., Wickett, N.J., Molero, J., Riina, R., Sanmartín, I., 2018. Bridging the micro-and macroevolutionary levels in phylogenomics: Hyb-Seq solves relationships from populations to species and above. New Phytol. 220 (2), 636–650.

Weitemier, K., Straub, S.C.K., Cronn, R.C., Fishbein, M., Schmickl, R., McDonnell, A., Liston, A., 2014. Hyb-Seq: combining target enrichment and genome skimming for plant phylogenomics. Appl. Plant Sci. 2, 1400042.

Whitfield, J.B., Lockhart, P.J., 2007. Deciphering ancient rapid radiations. Trends Ecol. Evol. 22, 258–265.

Wimmer, F.E., 1943. Campanulaceae-Lobelioideae. I. Das Pflanzenreich. IV. 276b(Heft 106), 1–260.

Woese, C.R., Achenbach, L., Rouviere, P., Mandelco, L., 1991. Archaeal phylogeny: reexamination of the phylogenetic position of *Archaeoglohus fulgidus* in light of certain composition-induced artifacts. Syst. Appl. Microbiol. 14, 364–371.

Zhang, C., Rabiee, M., Sayyari, E., Mirarab, S., 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinforma. 19 (Suppl. 6), 153.