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

Justin C. Bagley\textsuperscript{a,b,⁎}, Simon Uribe-Convers\textsuperscript{a}, Mónica M. Carlsen\textsuperscript{c}, Nathan Muchhal\textsuperscript{a}

\textsuperscript{a} Department of Biology, University of Missouri-St. Louis, St. Louis, MO 63121, USA
\textsuperscript{b} Department of Biology, Virginia Commonwealth University, Richmond, VA 23284, USA
\textsuperscript{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 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, 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, 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, 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).
numbers of variable loci sequenced using HTS also increases chances of
turbation; Townsend et al., 2012; Straub et al., 2014
genetic noise due to homoplasy (convergence due to nucleotide sa
siting phylogenetic signal (Townsend et al., 2014; Spalink et al., 2016
periods less than 15 million years (Myr; e.g. Lagomarsino et al., 2016;
versification rates of angiosperm radiations pose significant challenges
and facilitates additional comparative analyses (I.e. not if plastomes only mapped to reference).

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, insertions-deletions (indels), and variations in nucleotide composition (e.g. GC content; reviewed by Ressaye 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, protein-coding sites may introduce biases that can mislead phylogenetic

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

<table>
<thead>
<tr>
<th>Study</th>
<th>Organisms</th>
<th>n</th>
<th>Loci</th>
<th>Capture success</th>
<th>Exons</th>
<th>Flanking</th>
<th>Supercontigs</th>
<th>Plastomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weitemier et al. (2014)</td>
<td>Asclepias (Apocynaceae)</td>
<td>12</td>
<td>768</td>
<td>99-100%</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Crowd et al. (2017)</td>
<td>Mediterranean Campanula (Campanulaceae)</td>
<td>105</td>
<td>246</td>
<td>95.70%</td>
<td>x</td>
<td>–</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Chau et al. (2018)</td>
<td>Buddleia (Lamiaceae)</td>
<td>48</td>
<td>1049</td>
<td>91-99%</td>
<td>x</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Gernandt et al. (2018)</td>
<td>Pinus (Pinaceae)</td>
<td>74</td>
<td>710+</td>
<td>96.6%</td>
<td>x</td>
<td>–</td>
<td>–</td>
<td>x</td>
</tr>
<tr>
<td>Herrande-Moraires et al. (2018)</td>
<td>Cardueae (Compositae)</td>
<td>85</td>
<td>1061</td>
<td>64-99%</td>
<td>x</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Jones et al. (2019)</td>
<td>Asteraceae</td>
<td>112</td>
<td>1061</td>
<td>66-99%</td>
<td>x</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Kates et al. (2018)</td>
<td>Artocarpus (Moraceae)</td>
<td>24</td>
<td>151</td>
<td>73.50%</td>
<td>x</td>
<td>x</td>
<td>–</td>
<td>x</td>
</tr>
<tr>
<td>Stubbis et al. (2018)</td>
<td>Micranthes (Saxifragaceae)</td>
<td>49</td>
<td>518</td>
<td>99-100%?</td>
<td>x</td>
<td>–</td>
<td>–</td>
<td>x</td>
</tr>
<tr>
<td>Villaverde et al. (2018)</td>
<td>Euphorbia (Euphorbiaceae)</td>
<td>121</td>
<td>431</td>
<td>45-86%</td>
<td>x</td>
<td>x</td>
<td>–</td>
<td>x</td>
</tr>
<tr>
<td>Vainiamparat et al. (2018)</td>
<td>Leguminosae</td>
<td>25</td>
<td>423</td>
<td>93%</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

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.

* Number of loci targeted with bait sets.

** 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).
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 ‘hot-spot’ (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 lobeliod genera, Centropogon and Siphocamptus. 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 Sondač 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 transctiptome 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 SeqClean v1.10.09 (https://github.com/ibest/seqclean) 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 ‘read_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 ‘intronrate.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. aerticaca, B. brachyandra E. Wimm., B. ceratocarpa Zahhr., 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 subsampled the full supercontig alignments using a 100% taxonomic completeness threshold to generate (2) a ‘100p supercontig’ dataset of...
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 + I 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 + I 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 (Savarey and Mirarab, 2016). To estimate a chronogram for Burmeistera

![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).](image)

---

### Table 2

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

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 et al. (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 PhylnormR 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 homoplasic 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 heat-maps 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; Rodriguez-Ezeleta 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 ($\chi^2$) test implemented in PAUP* v4.0a (Swoford, 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 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. Phylogenetic 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. ayacachensis split, while divergence between B. refracta 1 and 2 was slightly greater than that between B. almeidae 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. ayacachensis Jeppesen, B. glabrata Benth. & Hook. F. ex B. D. Jacis, B. vulgaris E. Wimm., and B. draconis Pérez & Muchhalá) 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. sodirona 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.)

decreases 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
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
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 con-
catenated 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 super-
contig 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 ac-
cumulating 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 tree-
space 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. 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).
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 (\( > 33\% \) 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 MRCA 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 (\( \approx 28–60\% \)), including trees derived from shorter and longer loci. The *sodiroana* epoch nodes (subclade 3-a) were supported by...
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
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 performed by more closely matching null expectations assuming equal intralocus substitution rates. Third, PI ratios indicated sharper declines performed for our 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 $(\lambda_{\omega_i})$, 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 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; Bates 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.
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. Rodriguez-Ezepeleta 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 saturation 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. ulayi, which lacks pubescent anthers (BP = 50–80), or to a clade comprised of B. ulayi (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 sordoiana 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 HybSeq 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%
BioProject database (cession numbers PRJNA646974 and PRJNA623031 in the NCBI have been accessioned in the NCBI Transcriptome Shotgun Assembly NCBI Short Read Archive database and transcriptome and nuclear genome sequences, have been accessioned in the

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

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

J.C. Bagley, et al.


