Received: 9 January 2023 | Accepted: 21 June 2023

DOI: 10.1002/aps3.11557

GENOMIC RESOURCES ARTICLE

Applications in Plant Sciences

Calceolariaceae809: A bait set for targeted sequencing of nuclear loci

Nicolas Medina[1] | **David C. Tank**[2] | **Anahí Espíndola**[1]

[1]Department of Entomology, University of Maryland, College Park, Maryland 20742-4454, USA

[2]Department of Botany and Rocky Mountain Herbarium, University of Wyoming, Laramie, Wyoming 82071, USA

Correspondence

Nicolas Medina, Department of Entomology, University of Maryland, College Park, Maryland 20742-4454, USA.
Email: nicomed@umd.edu

Abstract

Premise: The genus *Calceolaria* (Calceolariaceae) is emblematic of the Andes, is hypothesized to have originated as a recent, rapid radiation, and has important taxonomic needs. Additionally, the genus is a model for the study of specialized pollination systems, as its flowers are nectarless and many offer floral oils as a pollination reward collected by specialist bees. Despite their evolutionary and ecological significance, obtaining a resolved phylogeny for the group has proved difficult. To address this challenge, we present a new bait set for targeted sequencing of nuclear loci in Calceolariaceae and close relatives.

Methods: We developed a bioinformatic workflow to use incomplete, low-coverage genomes of 10 *Calceolaria* species to identify single-copy loci suitable for phylogenetic studies and design baits for targeted sequencing.

Results: Our approach resulted in the identification of 809 single-copy loci (733 noncoding and 76 coding regions) and the development of 39,937 baits, which we validated in silico (10 specimens) and in vitro (29 Calceolariaceae and six outgroups). In both cases, the data allowed us to recover robust phylogenetic estimates.

Discussion: Our results demonstrate the appropriateness of the bait set for sequencing recent and historic specimens of Calceolariaceae and close relatives, and open new doors for further investigation of the evolutionary history of this hyperdiverse genus.

KEYWORDS

Andes, bioinformatic pipeline, *Calceolaria*, Lamiales, phylogenomics, target capture

The family Calceolariaceae is formed of two genera with contrasting diversity patterns. While *Jovellana* Ruiz & Pav. is a small genus containing only a handful of species (Nylinder, 2006) from Chile and New Zealand, *Calceolaria* L. is speciose and restricted to the Americas (Molau, 1988). Of the two, *Calceolaria* is among the most diverse genera in the order Lamiales (Molau, 1988) with over 250 accepted species and estimated to be recently diverged (6–15 mya; Renner and Schaefer, 2010; Frankel et al., 2022), suggesting a rapid mode of diversification. Native to the Andes and the Central American and Mexican mountains, *Calceolaria* is associated with diverse ecologies, such as alpine ecosystems, cloud forests, xeric environments, and coastal regions (Molau, 1988). Morphologically, *Jovellana* displays purple- or pink-spotted flowers with two lobes and a conserved open corolla across all of its species (Sérsic, 2004). Contrastingly, *Calceolaria*'s flowers are bilabiate, with each lobe varying strongly in shape, size, and coloration (e.g., yellow, orange, red, purple, white), making them one of the most recognizable plants in the Andes. *Calceolaria* is also well known for its pollination biology, which relies on oil-collecting bees for most species, leading to mutualistic interactions with bees of the genera *Chalepogenus* Holmberg and *Centris* Fabricius (Vogel, 1974; Sérsic, 2004); *Jovellana* instead produces no floral oils, and its pollinators are likely generalists (Sérsic, 2004).

While the relationships within the few species of *Jovellana* are well defined (Nylinder et al., 2012), this is not the case of *Calceolaria*. Even though the latter is an integral part of the Andean flora, baseline knowledge of attributes such as species delimitation, their evolutionary relationships, and their evolutionary trends and patterns is lacking. Previous attempts to infer a phylogeny for *Calceolaria* have resulted in limited estimations (Nylinder, 2006; Cosacov et al., 2009), likely due to the low variation of the genetic markers used and the rapid diversification suspected in the genus (Cosacov et al., 2009). More recently, Frankel et al. (2022) estimated a resolved backbone phylogeny based on the complete plastomes of 14 *Calceolaria* species. Their phylogeny revealed infrageneric relationships different from those obtained in previous results, but notably lacked information from the nuclear genome and was restricted in the number of species used. Despite this, Frankel's study provided a novel insight into the evolutionary history of the group and suggested that a larger number of genetic markers would likely resolve the ancestry relationships within the genus.

Understanding the evolutionary relationships of the family is critically important for future research efforts. In addition to advancing the systematics of the group, resolving its species history would improve our understanding of rapid diversification processes in one of the most diverse mountain regions of the world. For example, *Calceolaria*'s specialized pollination is suggested to be a main driver of its evolution (Molau, 1988; Sérsic, 2004; Cosacov et al., 2009; Murúa and Espíndola, 2015; Frankel et al., 2022), and its diversity also appears correlated with the history of the Andes in having two distinct centers of diversity (i.e., central Chile and the central Andes; Molau, 1988). Genome duplication also seems to play a role in *Calceolaria*'s speciation; while most southern species are diploids (Molau, 1988; Ehrhart, 2000), tropical *Calceolaria* are abundant in polyploid species. From another perspective, the genus harbors taxa of conservation concern with vulnerable and threatened statuses (León Yánez et al., 2011), and likely many undescribed species, as reflected by the frequent description of new species and species complexes in the group (i.e., Puppo and Novoa, 2012; Puppo, 2014, 2023; Romero-Hernández et al., 2017; Lavandero et al., 2021). Thus, our understanding of the evolution of the group will have direct impacts on the conservation of its diversity. Finally, the study of the evolution of *Calceolaria* may also help advance applied fields. For example, the genus includes species with many traditional uses, prompting the study of potential applications in pharmaceutics and agriculture (Khambay et al., 1999; Céspedes et al., 2014; Paniagua-Zambrana et al., 2020).

Given the potential impacts of resolving the phylogeny of the group and the leap made by Frankel et al. (2022) in determining the ancestry backbone of *Calceolaria*, albeit from a single locus, the time appears ripe to explore other genomic-level approaches for the family. Phylogenomics has revolutionized our understanding of plant evolution (Andermann et al., 2020); of the various data collection approaches available, one of the most promising is targeted sequencing, such as that made possible by the Angiosperms353 bait set (Johnson et al., 2019). In fact, this bait set was expected to boost plant systematics studies in the same way that ultraconserved element (UCE) markers catalyzed the understanding of animal evolution (Van Dam et al., 2021). Although the Angiosperms353 bait set has successfully worked in several groups, it has shown limitations in resolving infrageneric relationships of lineages with recent and rapid diversification such as that expected in Calceolariaceae (Siniscalchi et al., 2021). For this reason, molecular systematists have turned to taxon-specific bait sets that can recover higher genetic variation at lower taxonomic levels (Chau et al., 2018; Soto Gomez et al., 2019; Andermann et al., 2020; Koenen et al., 2020; Villaverde et al., 2020).

In this study, building on this body of knowledge and utilizing the low-coverage nuclear data not used by Frankel et al. (2022), our goal is to develop a bait set to target nuclear genomic markers and ultimately reconstruct a resolved, species-level phylogeny of Calceolariaceae. To do so, we first developed a bioinformatic pipeline to identify, select, and validate an appropriate capture bait set from low-coverage genomes. Finally, we tested (in vitro) the new bait set on a small set of DNA samples from Calceolariaceae and its close relatives, and compared our results to previous studies.

METHODS

General method overview

Our approach began with the assembly of low-coverage genomes of 10 *Calceolaria* species (Table 1). Based on these assemblies, our bioinformatic workflow (Figure 1) presented two parallel series of analyses. In the first series, we applied the phyluce pipeline (Faircloth, 2016) on the genome assemblies to identify shared single-copy loci. In the second series, we searched genes associated with floral development and pollinator cues by scanning ab initio predictions of proteins on the same genome assemblies. For this, we used MetaEuk (Levy Karin et al., 2020) to predict genes based on a custom database of proteins and inferred the ontologies of their exons applying ExOrthist (Márquez et al., 2021) to retain single-copy exons. Finally, we used MrBait (Chafin et al., 2018) to generate target-capture baits for the loci identified in the two series and validated them in silico and in vitro. All bioinformatic analyses were performed using the computational resources of the Research Computing and Data Services at the University of Idaho.

Sampling and sequencing

We built upon the sequencing work of Frankel et al. (2022) and used all their Illumina shotgun reads to assemble nuclear

TABLE 1 *Calceolaria* samples used for shotgun sequencing, accession information, and resulting genome assembly metrics.

Species	Section[a]	Sample type (Herbarium)[b]	Herbarium accession	Voucher no.	Country of origin	Assembly length (Mbp)	Contigs	Largest contig (bp)	N50 (bp)	L50 (bp)
C. bicolor Ruiz & Pav.	Urticopsis	Herbarium (MO)	MO 6071852	Puppo et al. 143	Peru	335.55	314,765	37,398	1726	53,579
C. boliviana (Britton ex Rusby) Pennell	Engleriana	Herbarium (MO)	MO 6353817	Araújo 4205	Bolivia	479.87	446,930	75,644	2472	48,389
C. corymbosa Ruiz & Pav.	Cheiloncos	Field silica-dried	—	Espíndola et al. NIC-8	Chile	367.89	370,182	99,508	3167	24,122
C. mexicana Benth.	Calceolaria	Herbarium (MO)	MO 6740412	Wortley et al. 234	Bolivia	361.77	170,533	65,944	7248	14,127
C. meyeniana Phil.	Cheiloncos	Field silica-dried	—	Espíndola et al. YER-7	Chile	338.79	278,109	57,032	4693	15,917
C. pedunculata Molau	Lehmannina	Herbarium (GB)	GB 014 9599	Croat & Hannon 88192	Ecuador	10.72	19,408	77,259	758	2374
C. phaseliifolia Edwin	Lobatae	Herbarium (GB)	GB 014 9542	Schmidt-Lebuhn 515	Peru	154.52	218,569	101,459	902	45,899
C. plectranthifolia Walp.	Rugosae	Herbarium (MO)	MO 6383566	Zárate 2302	Bolivia	264.71	188,078	42,090	3484	20,070
C. polyrhiza Cav.	Cheiloncos	Herbarium (MO)	MO 6581727	Zavala-Gallo et al. 89	Argentina	26.63	51,289	23,335	695	7121
C. tenella Poepp. & Endl.	Cheiloncos	Field silica-dried	—	Espíndola et al. 2012-12	Chile	248.80	167,177	65,660	5211	12,269

[a]Sensu Molau, 1988.
[b]Herbarium acronyms are according to Thiers (2023).

genome drafts of *Calceolaria*. The 10 species used span major taxonomic and ecological diversity of the genus (Table 1). As noted in Frankel et al. (2022), sequencing of these samples was performed by the Genomics Core Center of the University of Maryland School of Medicine on an Illumina HiSeq platform (Illumina, San Diego, California, USA).

Genome assemblies

Beginning from raw shotgun sequences, we filtered and trimmed low-quality reads using TrimGalore v0.6.7 (Krueger et al., 2021), applying default settings. We then applied the metagenomic pipeline SqueezeMeta (Tamames and Puente-Sánchez, 2019) to assemble the genomes and to exclude sequences that overlapped those of other organisms (e.g., bacteria, fungi). We used the *sequential-mode*, with the arguments *euk* and *doublepass* turned on, and selected MEGAHIT (Li et al., 2015) as the assembler. The pipeline returned contigs assigned to taxonomic groups of which we only retained Streptophyta sequences for the *Calceolaria* genome assemblies. Next, we used RepeatMasker (Smit et al., 2004) on each assembly with the *Arabidopsis* repeat reference to detect and mask repetitive and low-complexity sequences that are unsuitable for targeted sequencing. Finally, we assessed quality using the evaluation tools QUAST (Gurevich et al., 2013) and BUSCO v5 using the *eudicots10* database (Manni et al., 2021).

Series I: phyluce

We employed the phyluce pipeline (Faircloth, 2016) to search for anonymous single-copy loci across our 10 genome assemblies. First, we mapped the *Calceolaria* shotgun reads to our genome assembly of *C. boliviana* (Britton ex Rusby) Pennell using Stampy (Lunter and Goodson, 2011), and then merged overlapping mappings within each taxon with SAMtools (Danecek et al., 2021). We followed the phyluce pipeline to find regions shared by at least eight samples, extracting the central 160 bp of the shared regions from the reference. Next, we mapped the reference loci to the *Calceolaria* assemblies and extracted their sequences, extended by 200 bp on each flank. After extracting all loci from each sample, we mapped these sequences back on all the assemblies to remove cross-species duplicate loci. For these steps, we followed the phyluce strategy of breaking up target loci into shorter bait sequences and using LASTZ (Harris, 2007) for in silico hybridization. At the end of this process, all loci retained were single-copy and present in at least six of the genomes. To further filter these shared loci, we aligned each of them using MUSCLE (Edgar, 2004) and trimmed their less informative edges with Gblocks (Castresana, 2000). Then, for each locus we evaluated the number of taxa per alignment, locus length, and the number of phylogenetically informative sites using phyluce. Based on these data, we selected loci that were

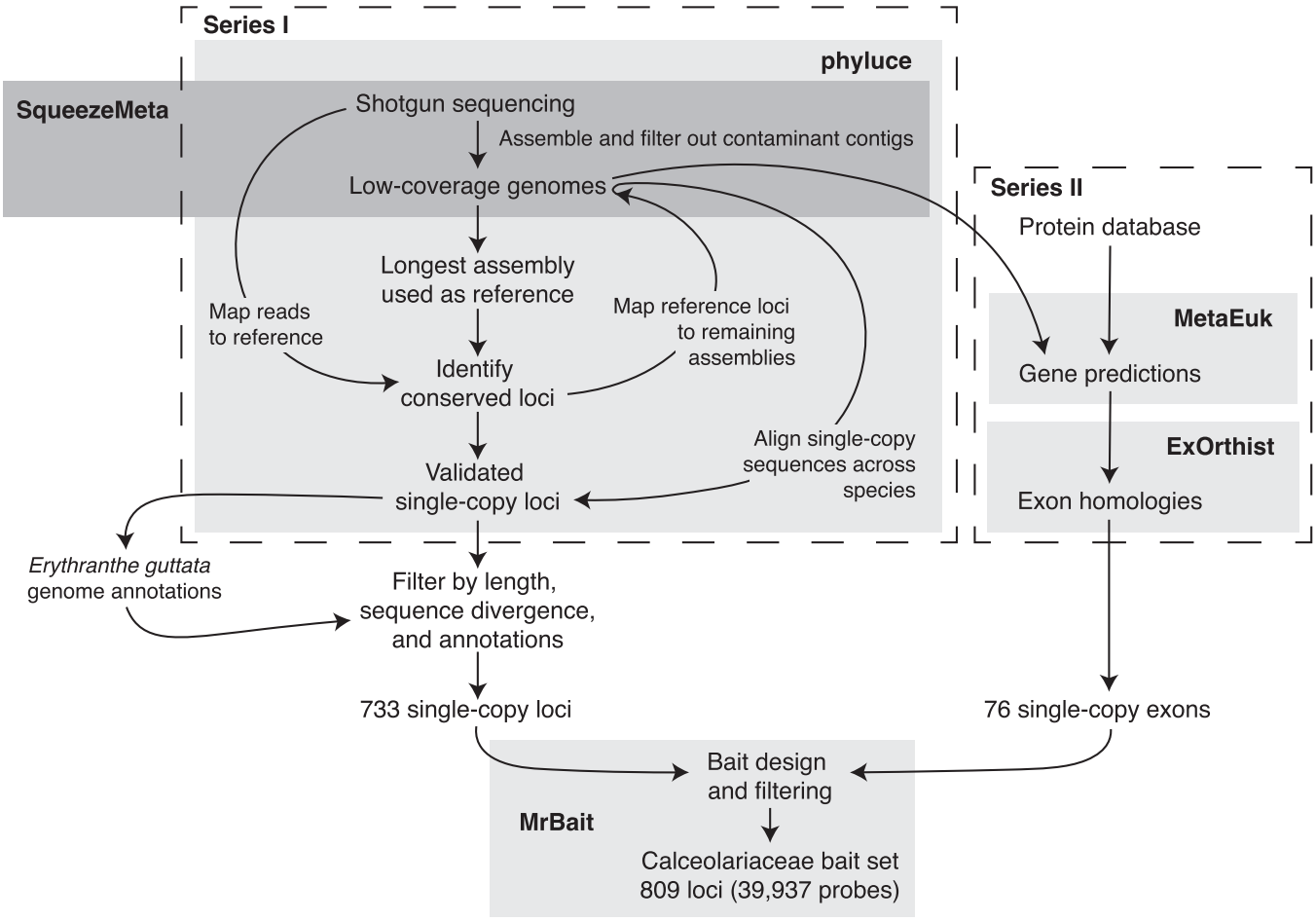

FIGURE 1 Bioinformatic pipeline used to identify informative targets from shotgun sequence data on which to develop capture baits. Sections enclosed in boxes with dashed lines correspond to steps associated with each series.

between 400–600 bp in length and that included 4–38 phylogenetically informative sites.

To understand the identity of these loci, we aligned them to the genome of the closest relative of *Calceolaria* for which a reference-level assembly exists (i.e., *Erythranthe guttata* (DC.) G. L. Nesom; Mower et al., 2012) and extracted overlapping coding gene annotations. Based on these, we only retained loci that overlapped with a coding sequence by less than 80%. To align and process the sequences and annotations, we used LASTZ, BEDtools (Quinlan and Hall, 2010), and the tidyverse R packages (Wickham et al., 2019; R Core Team, 2021). We also explored the potential molecular functions of the overlapping coding genes; based on the *Erythranthe* genome we queried their protein sequences to the InterPro database (Blum et al., 2021), extracted any associated Gene Ontology (GO) terms (Gene Ontology Consortium, 2021), and summarized them with a semantic analysis using GO-Figure! v1.0 (Reijnders and Waterhouse, 2021). Finally, we used MrBait (Chafin et al., 2018) to generate bait sequences based on all the sequences from the selected loci. We designed baits 80 bp in length and with 60 bp of overlap, filtering out redundant sequences at 95% similarity and those that mapped to *Calceolaria* organelles (Frankel et al., 2022). We selected this bait length and tiling arrangement to maximize the applicability of the set across a range of genomic DNA qualities obtained from fresh, silica gel–dried, and degraded (e.g., herbarium) samples.

Series II: Exon search

The evolution of *Calceolaria* is presumably at least partially affected by its specialized pollination system based on floral oil rewards (Molau, 1988; Cosacov et al., 2009; Murúa et al., 2014). For this reason, we sought to target genes associated with pollinator-relevant traits, such as those related to floral morphology and volatile production. To do this, we applied MetaEuk (Levy Karin et al., 2020) to predict genes, targeting protein domains of transcription factors controlling flower morphology (MADS-box and TCP; Danisman, 2016; Thomson and Wellmer, 2019), catalyzers of color (MYB and P450; Zhang et al., 2017), oils (P450; Yang et al., 2022), and volatiles associated with flowers (terpene synthase; Karunanithi and Zerbe, 2019), among others (Table 2). First, we built a query database of eight

TABLE 2 Selected protein domains for ab initio predictions.

Pfam accession	Pfam family/domain	Proteins in query	Function of interest
PF00319	SRF-type transcription factor (DNA-binding and dimerization domain; MADS-box)	164	Flower development
PF01161	Phosphatidylethanolamine-binding protein	17	Flower development
PF00067	Cytochrome P450	475	Flavonol/pigment and lipid synthesis
PF00195	Chalcone and stilbene synthases N-terminal domain (amino oxidase)	173	Flavonol/pigment synthesis
PF01593	Flavin containing amine oxidoreductase (CHS/STS)	48	Flavonol/pigment synthesis
PF00249	Myb-like DNA-binding domain	243	Flower development
PF03634	TCP family transcription factor	42	Flower development
PF01397	Terpene synthase, N-terminal domain	360	Volatile production

protein domains with sequences from the UniProt database (UniProt Consortium, 2021), and then queried each genome assembly with the *easy-predict* module of MetaEuk using default settings (Levy Karin et al., 2020). This step resulted in exon coordinates that were carried into the ExOrthist pipeline (Márquez et al., 2021). We applied this pipeline to cluster homologous exons at 70% of sequence similarity including adjacent introns. As we did in Series I, to detect and remove duplicate exons, we mapped these sequences on all the assemblies and removed exons with more than one hit per genome. Finally, we used MrBait to generate tiled 80-bp baits with 60-bp overlap based on exonic alignments, and filtered out organellar loci and redundant baits at 95% similarity.

In silico validation and comparison with the Angiosperms353 bait set

To evaluate the effect of the individual types of loci in our work, we conducted an in silico bait capture on four data schemes: (1) Series I loci: phyluce, (2) Series II loci: exons, (3) Series I + II combined, and (4) Angiosperms353. We used the phyluce tools to hybridize each group of baits in silico to the 10 low-coverage genomes used in bait design, and to generate and trim alignments with the aim of testing their ability to resolve phylogenetic relationships within *Calceolaria*. For each group of loci, we estimated maximum likelihood (ML) and coalescent-based phylogenies. For the ML approach, we concatenated the alignments of each group, applied them as a single partition on RAxML-NG (Kozlov et al., 2019) with a GTR+GAMMA model of sequence evolution, and evaluated support with 100 bootstrap replicates. For the coalescent approach, we first estimated individual trees for each locus using RAxML-NG with a GTR+GAMMA model of sequence evolution, and then used the resulting gene trees for species-tree estimation with ASTRAL (Zhang et al., 2018). For species-tree analyses, support was assessed with local posterior probability (Sayyari and Mirarab, 2016). We rooted the trees with the genome sequence of *Erythranthe guttata* (Mower et al., 2012). Finally, we compared the resulting nuclear phylogeny with the chloroplast topology from Frankel et al. (2022).

In vitro testing

To experimentally test our bait set, we extracted genomic DNA from 35 plant samples, including the ones used for probe development (Table 3). These samples were of different ages and extended the initial sampling to other Calceolariaceae and outgroups from the Lamiales clade. To evaluate the appropriateness of our bait set to sequence high-quality and degraded DNAs, we included genomic DNA extracted from freshly collected, historical, and rare collections (e.g., herbarium samples from 1905 to 2012). Genomic DNA was isolated using the Synergy 2.0 Plant DNA Extraction Kit (OPS Diagnostics, Lebanon, New Jersey, USA). Synthesis of the custom bait set, target capture, library preparation, and sequencing was conducted by Daicel Arbor Biosciences (Ann Arbor, Michigan, USA); sequencing was performed on an Illumina NovaSeq platform (2 × 150-bp paired-end reads). Upon receiving the short-read data, we used phyluce to assemble the data and concatenated the resulting alignments for tree estimation with RAxML-NG applying a GTR+GAMMA model and 100 bootstrap replicates to assess node support. We also evaluated the number of loci recovered across all samples and compared the tree estimate with that obtained by Frankel et al. (2022).

RESULTS

Genome assemblies

The SqueezeMeta pipeline recovered contigs from plant organisms, assigning a per-sample average of 27.5% (±15.5 standard deviation [SD]) of the reads to the phylum Streptophyta and 10.8% (±15.4 SD) to non-plant taxa

TABLE 3 Specimens used for in vitro validation and sequencing results.

Species[a]	Voucher (Herbarium)[b]	Year collected	Total sequenced loci	Sequenced single-copy	Single-copy exons	Single-copy phyluce
Calceolaria andina Benth.	Buchtien s.n. (US)	1903	795	603	57	546
C. asperula Phil.	Pennell 12272 (US)	1925	524	483	19	464
*C. bicolor**	Puppo et al. 143 (MO)	2007	808	117	6	111
C. bicolor	Molau et al. 547 (US)	1983	757	328	27	301
*C. boliviana**	Araújo et al. 4205 (MO)	2008	809	91	6	85
C. boliviana	Fuentes et al. 10050 (MO)	2006	784	288	36	252
*C. corymbosa**	Espíndola et al. NIC-8	2012	804	394	61	333
C. corymbosa	Ehrhart 1437 (MSB)	2001	808	409	54	355
C. georgiana Phil.	Werdermann 904 (US)	1925	809	624	67	557
*C. mexicana**	Wortley et al. 234 (MO)	2007	797	536	36	500
C. mexicana	García-Mendoza et al. 10299 (MO)	2013	780	571	46	525
*C. meyeniana**	Espíndola et al. YER-7	2012	798	632	62	570
C. meyeniana	Ehrhart 361 (MSB)	2001	809	559	58	501
C. nevadensis (Pennell) Standl.	de Bellard 219 (US)	1923	692	372	28	344
*C. pedunculata**	Croat & Hannon 88192 (MO)	2003	770	311	27	284
C. pedunculata	Croat et al. 93245 (MO)	2004	743	367	29	338
*C. phaceliifolia**	Schmidt-Lebuhn 515 (GB)	2002	26	NA	NA	26
C. phaceliifolia	Edwin & Schunke 3834 (US)	1966	696	350	19	331
*C. plectranthifolia**	Zárate 2302 (MO)	2006	802	771	69	702
C. plectranthifolia	van der Werff et al. 20809 (MO)	2006	782	740	47	693
*C. polyrhiza**	Zavala-Gallo et al. 89 (MO)	2011	393	336	11	325
C. polyrhiza	Ager 424 (US)	1975	786	651	57	594
*C. tenella**	Espíndola et al. 2012-12	2012	789	732	59	673
C. tenella	Zavala-Gallo et al. 85 (MO)	2011	804	741	66	675
C. triandra (Cav.) Vahl	Smith 8873 (MO)	1984	533	381	10	371
C. triandra	Ferreyra 18389 (MO)	1974	574	389	14	375
C. uniflora Lam.	Zavala-Gallo et al. 197 (MO)	2011	806	587	64	523
C. uniflora	Goodall 500 (US)	1967	797	659	62	597
Castilleja foliolosa Hook. & Arn. (Orobanchaceae)	Tank 2002-05	2016	236	171	22	149
Comoranthus minor H. Perrier (Oleaceae)	Dorr et al. 4135 (US)	1985	285	191	15	176
Episcia cupreata (Hook.) Hanst. (Gesneriaceae)	US Botanic Garden	2022	377	287	39	248
Jovellana punctata Ruiz & Pav. (Calceolariaceae)	Hutchison 262 (US)	1961	741	226	22	204
Monopyle ecuadorensis C. V. Morton (Gesneriaceae)	Cornejo & Mata 9187 (US)	2018	195	174	21	153
Neoeplingia leucophylloides Ramamoorthy, Hiriart & Medrano (Lamiaceae)	Medrano et al. 9596 (US)	1976	244	214	27	187
Schrebera alata (Hochst.) Welw. (Oleaceae)	Van Jaarsveld et al. 18883 (US)	2004	282	189	17	172

[a]Asterisks indicate samples initially used for shotgun sequencing and probe development.

[b]Herbarium acronyms are according to Thiers (2023).

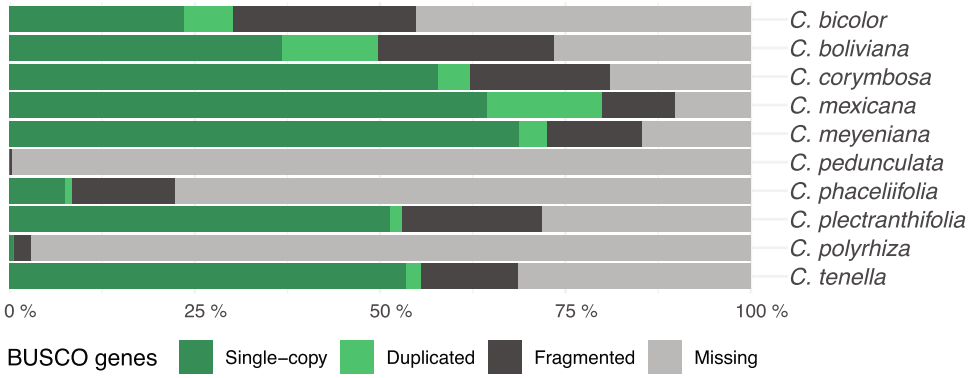

FIGURE 2 Fraction of types of BUSCO genes recovered by the assembly evaluation, per species.

(Appendix S1; see Supporting Information with this article). In contrast, a per-sample average of 48.4% (±13.1 SD) of reads were unclassified and 13.1% (±14.6 SD) did not map to the overall assemblies. This process resulted in 10 genome assemblies of variable size and fragmentation (Table 1), as well as varying BUSCO metrics of completeness (Figure 2). The assembly from *C. boliviana* was the longest (479 Mbp), but it was *C. mexicana* Benth. (361 Mbp) that resulted in the fewest missing BUSCO genes. This was in contrast with *C. pedunculata* Molau and *C. polyrhiza* Cav., for which the total assembled length was no more than 20 Mbp. It is important to note, however, that even with such poor coverage, these specimens still contained informative sequences (see below). This wide variation of genome metrics is not surprising given the multiple ages of the samples and limitations of short-read assemblies.

Loci selection and probe development

Series I

We identified 153,399 shared regions across a minimum of eight species. Of these, 1585 were cross-validated as single-copy and with potential for targeted capture (Table 4). After filtering by length, phylogenetic informativeness, and proportion within coding regions, our strategy yielded 733 loci 527.6 bp (±53.8 SD) in length and an average of 19.6 (±9.1 SD) informative sites per locus. A significant fraction of these loci (464; 63.4%) were 100% anonymous, and the remaining loci (269; 36.6%) were partially within coding regions with an average overlap of 52.6% (±16.4 SD). Finally, the semantic study of the GO terms associated with these loci resulted in a poor clustering of their functions, thus rejecting a common molecular role (Figure 3).

Series II

Our protein database included 1522 protein sequences from eight protein domains. Based on our genome assemblies, MetaEuk predicted an average of 616.6 (±345.8 SD) potential genes with coding regions of varying size but consistent protein references (785.6 bp ± 434.5 in average per gene). ExOrthist detected homology for a large proportion of exons across all samples, but many of them were discarded for presenting multiple copies. At this step, we only retained 46 genes and 76 single-copy exons of 282.1 bp (±52.9 SD) in average length (Table 4, Appendix S2).

Combined outputs

Once analyses of the two series were completed, the total length of the phyluce loci (Series I) was 386.8 Mbp, and that of the exons (Series II) was 20,632 bp, for a combined total of 809 targeted loci and 407.4 Mbp. MrBait generated an initial 72,763 baits for all the targeted regions; however, after redundancy filters were applied the set resulted in 39,937 baits, which we retained for our bait set.

In silico testing

In general, independent of the type of loci used (i.e., Series I: phyluce, Series II: exons), the inferred phylogenies were well supported (i.e., nodes with at least 90% bootstrap support and 0.7 posterior probability) and of similar topology, with only some differences in the branching order near the tips (Figure 4). We recovered two consistently supported clades: Clade 1, corresponding to the southern species (*C. corymbosa* Ruiz & Pav., *C. meyeniana* Phil., *C. polyrhiza*, and *C. tenella* Poepp. & Endl.), plus *C. plectranthifolia* Walp. from the central Andes and *C. pedunculata* from the northern region (and the only tetraploid in this mostly diploid clade), and Clade 2, including species from the central and northern Andes (*C. bicolor* Ruiz & Pav., *C. mexicana*, *C. phaceliifolia* Edwin, and *C. boliviana*), which are all tetraploids. Node support was lower (i.e., bootstrap support of less than 60% or 0.6 posterior probability) for the exonic trees than for the "neutral" phyluce trees, likely due to different numbers of loci (76 vs. 733, respectively), but also suggesting that coding regions could also

TABLE 4 Results of marker search and selection after in silico validation, per species and development strategy.

	Series I: phyluce			Series II: exons				Angiosperms353	
Species	Single-copy loci	Loci length (bp) (±SD)	Total length (bp)	Predicted genes	Single-copy exons	Exon length (bp) (±SD)	Total length (bp)	Loci in assembly	Single-copy loci
C. bicolor	449	449.98 (±66.57)	202,043	875	28	219.0 (±58.7)	6133	60	9
C. boliviana	297	465.97 (±45.84)	138,394	1076	0	0	0	73	7
C. corymbosa	703	474.44 (±44.65)	333,534	869	69	270.9 (±118.1)	18,696	43	7
C. mexicana	556	456.06 (±48.92)	253,570	820	38	275.6 (±105.0)	10,475	44	9
C. meyeniana	711	475.75 (±43.50)	338,260	795	72	274.3 (±115.3)	19,753	45	9
C. pedunculata	24	363.54 (±64.29)	8725	16	6	414.6 (±130.1)	2488	1	0
C. phaceliifolia	485	422.85 (±73.98)	205,084	488	55	259.3 (±87.5)	14,264	29	11
C. plectranthifolia	707	471.21 (±46.32)	333,148	597	73	273.1 (±114.9)	19,938	37	13
C. polyrhiza	130	361.18 (±66.65)	46,954	89	11	274.6 (±70.6)	3021	5	0
C. tenella	663	472.25 (±44.06)	313,104	541	68	277.8 (±116.8)	18,895	39	7

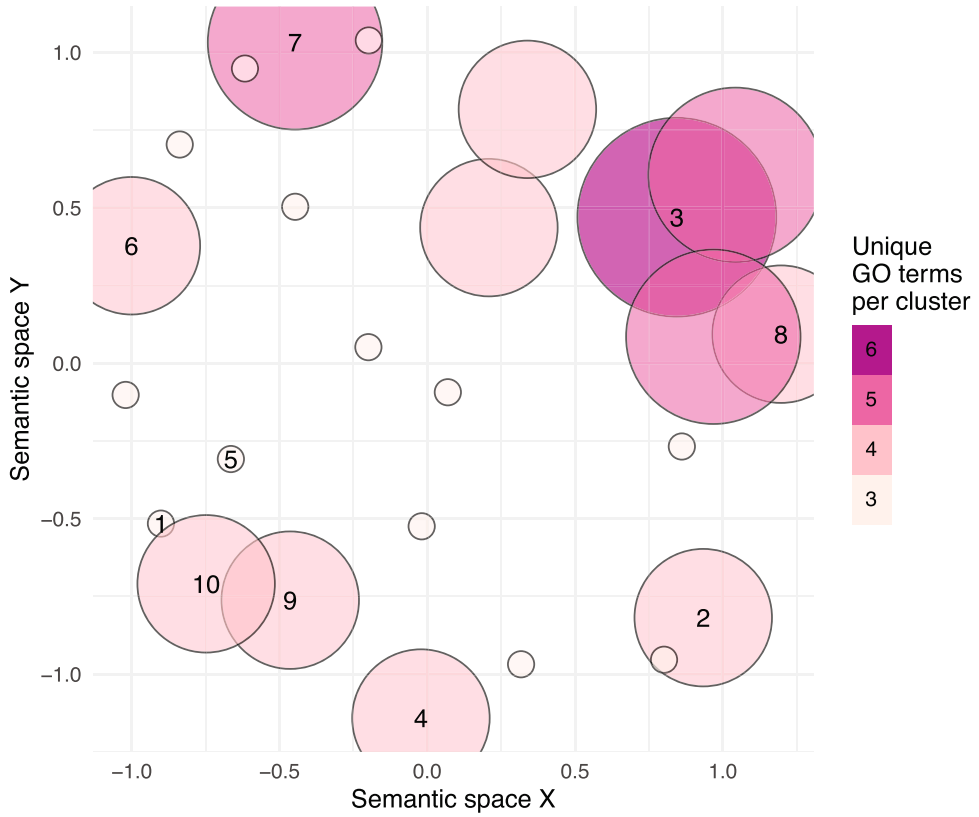

FIGURE 3 Semantic analysis of gene ontology (GO) terms associated with the phyluce loci (Series I). Circles represent clusters of molecular functions, with sizes proportional to the number of genes present in each group and colors representing the number of unique GO terms. Numbers correspond to the top 10 clusters. (1) Phosphatidate cytidylyltransferase, (2) iron ion transmembrane transporter, (3) calmodulin binding, (4) guanyl-nucleotide exchange, (5) UDP-glycosyltransferase activity, (6) ribonuclease III activity, (7) ATP-dependent peptidase activity, (8) GTP binding, (9) diacylglycerol *O*-acyltransferase, (10) 6-phosphofructokinase activity.

affect support of relationships. Finally, the Angiosperms353 bait set resulted in a poor recovery of single-copy loci, assembling only 25 loci with at least five samples, and missing two samples in the final assembly (Table 4).

In vitro testing

We successfully obtained data from the 35 samples included, resulting in approximately one million reads per

sample. The assembly recovered all 809 targeted loci, while 24 out of the 35 samples yielded over 600 loci (Table 3). We could not identify any evident correlation between the age of the samples and the number of loci assembled. As expected, the outgroup sequencing coverage was lower than that seen in Calceolariaceae, but with a remarkable average of 330 (±185 SD) loci per sample. After removing putative paralogues (i.e., loci with more than one matched contig), the final data set included 806 loci, and a per-sample average of 444 loci (±192 SD) for Calceolariaceae and 200 loci (±34 SD) for the outgroups. The phylogenetic estimation from these data resulted in a robust result (Figure 5). The topology we obtained is consistent with the in silico results, where a central split divides the genus into two main clades with a strong geographic signal. A notable difference between the two sets of results is the position of *C. pedunculata*, which in the in vitro analysis is included in the clade with all the other tropical *Calceolaria* (Clade 2). We also encountered stochastic variation; for example, the same specimen of *C. phaceliifolia* that was used in bait development displayed very low coverage in this sequencing run, while the opposite was true for *C. pedunculata*.

DISCUSSION

The phylogeny of Calceolariaceae, and in particular of *Calceolaria*, has been historically challenging to infer (Nylinder, 2006; Cosacov et al., 2009). In this context, our new bait set provides an optimistic outlook to accomplish this goal using a phylogenomic approach. Our tests indicate that our taxon-specific bait set is effective in resolving phylogenetic relationships within the group. In particular, the set is highly informative and able to resolve intrageneric relationships, even within groups that have been hard to resolve in previous studies (e.g., the southern Andean clade; Cosacov et al., 2009). From a technical perspective, the bait set targets 809 nuclear loci with significant phylogenetic information, integrates both anonymous loci and regions potentially under selection by pollinators, and will be useful for evolutionary studies of *Calceolaria* and relatives.

Of broader interest, our results are encouraging for the integration of biological collection material into phylogenomics, as we were able to produce usable sequences for DNA older than 100 years. For example, despite the low shotgun sequencing quality of *C. pedunculata* (Appendix S1),

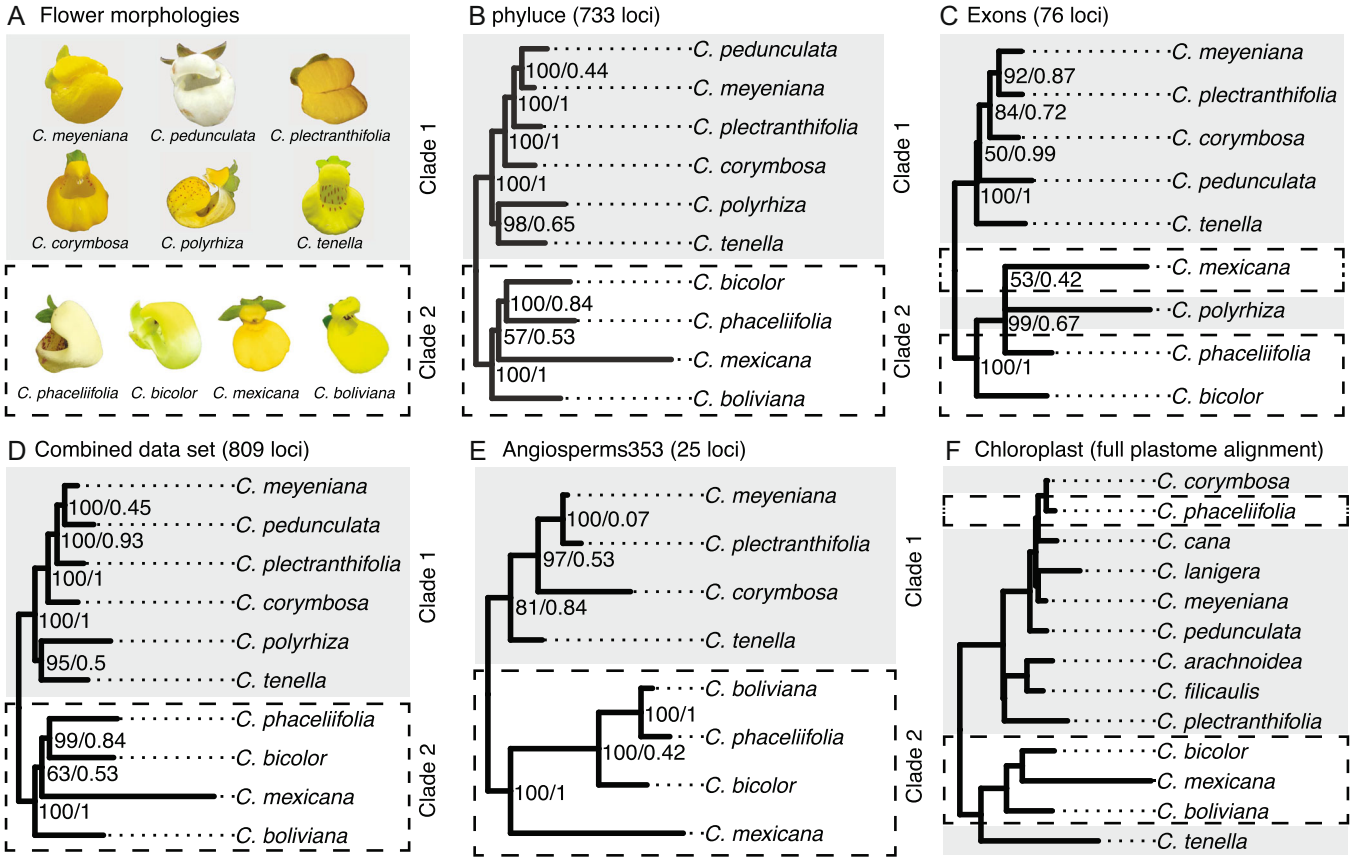

FIGURE 4 In silico phylogenetic estimates for a selection of species (A) using different bait set setups (B–E) and compared to the plastome inferences from Frankel et al. (2022) (F). The topologies were obtained with a maximum likelihood approach, and outgroups were pruned for visualization. Node supports are shown as RAxML bootstraps and ASTRAL posterior probabilities. Background colors highlight the position of species recovered in each clade across all inferences.

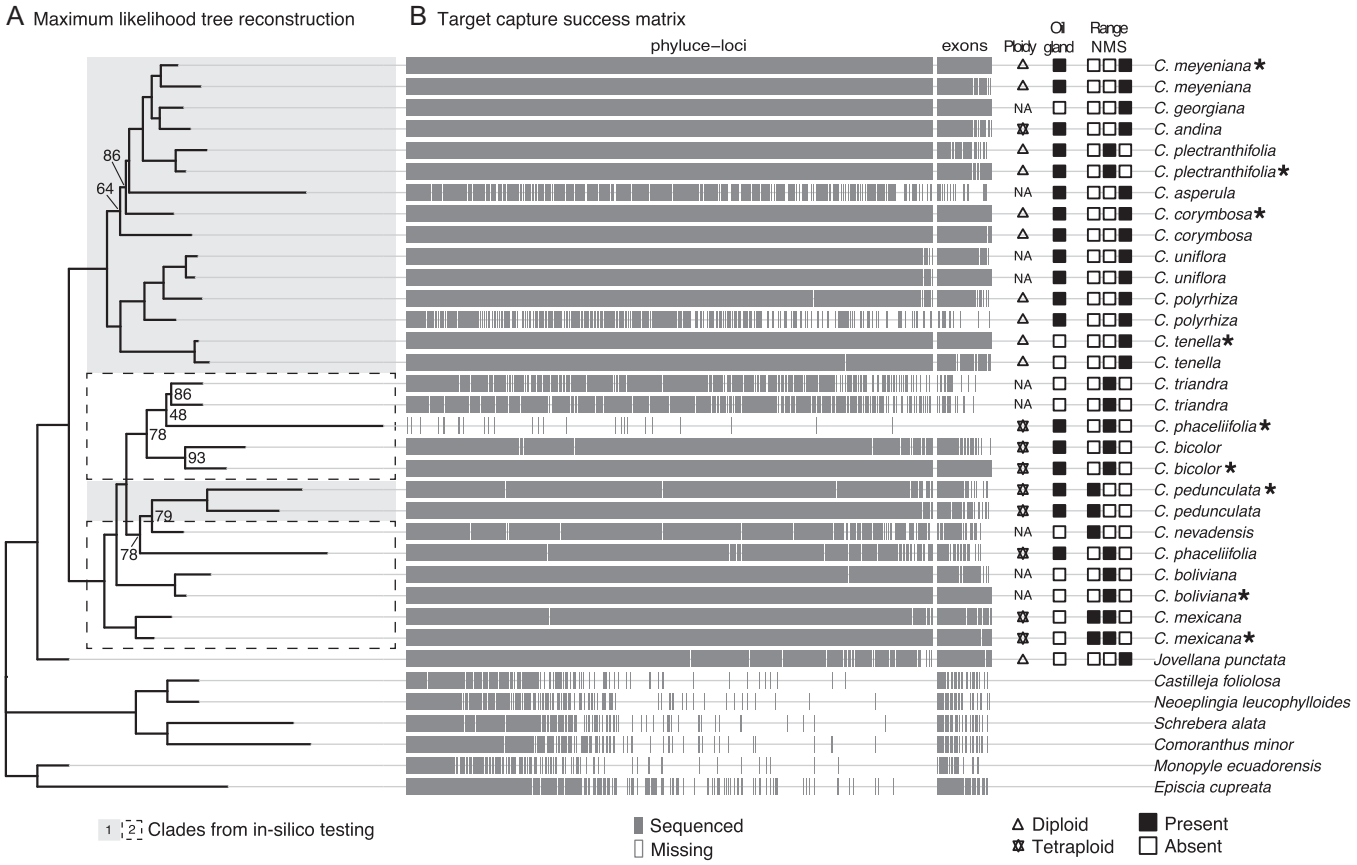

FIGURE 5 In vitro validation of the Calceolariaceae809 bait set on 35 *Calceolaria* and *Jovellana* species, and outgroups. (A) Maximum likelihood phylogenetic reconstruction based on single-copy loci and (B) occupancy matrix of the assembly for loci produced in Series I (phyluce) and Series II (exons). In (A), only nodes with bootstrap support lower than 95 are labeled. Background colors correspond to clades recovered in the in silico testing step. In (B), columns correspond to each expected locus; rows correspond to taxa included in the phylogeny. A selection of traits suspected to be important in the evolution of the group are mapped by the occupancy matrix. Asterisks indicate samples used for bait development.

our bait set successfully enriches our targets, leading to a very complete set of sequences (Table 3).

Workflow and bait set

Although bait-capture approaches are becoming more common in phylogenetics, the use of group-specific, custom bait sets is still limited in plants. Indeed, the Angiosperms353 bait set has improved phylogenetic inference in many groups (e.g., Crowl et al., 2022; Le et al., 2022; Schmidt-Lebuhn, 2022; Simões et al., 2022; Thureborn et al., 2022), although its ability to recover deep relationships in the angiosperms is counter-productive when the aim is to understand relationships at lower taxonomic levels (e.g., interspecific). In our case, the Angiosperms353 ML phylogenetic reconstruction recovered a similar topology to our bait set (Figure 4), but displayed many very short branches, especially in the already hard-to-resolve southern clade (clade I in Cosacov et al., 2009).

As a result, some studies have started to develop alternative bait sets that would recover infra-group variation (e.g., Ufimov et al., 2022). This study both develops a Calceolariaceae-specific bait set, and importantly, makes a significant contribution to this literature by developing a pipeline that can be used in other groups of closely related plants. By taking advantage of recent genomic tools (e.g., phyluce, MetaEuk, ExOrthist) and using shotgun sequencing data as input, this bioinformatic pipeline first assembles draft genomes and then develops baits targeting anonymous and functionally significant nuclear loci.

Our approach demonstrates that bait sets can be successfully developed even when the genomic quality is not exceptionally high. Indeed, our pipeline can filter sequences of non-plant DNA, a common output for most shotgun sequencing projects. Furthermore, our bait-tiling arrangement allows for the application of this set on a variety of DNA qualities and taxonomic placement, improving the inclusion of historical herbarium specimens, as well as of non-focal plant groups.

Along with these technical benefits, our use of two parallel approaches for identifying different target regions demonstrates our workflow's flexibility and ability to accommodate different research needs. For instance, while the methods used in Series I can be applied to other groups to identify

anonymous ("neutrally evolving") loci, those used in Series II permit targeting specific regions of the genome suspected to be associated with traits of interest (Figure 1). Although we focused Series II on floral characters, our pipeline can be used to develop bait sets seeking to address research questions related to other traits, in plants and other organisms, making this pipeline applicable to groups beyond our focal taxon.

Toward a fully resolved phylogeny of *Calceolaria*

Although studied for dozens of years, a definitive understanding of the relationships within Calceolariaceae and *Calceolaria* has been hindered by several challenges, including access to high-quality DNA and a lack of informative molecular data (Nylinder, 2006; Cosacov et al., 2009). With more than 250 described species, this genus presents particular opportunities and challenges given its tremendous diversity and likely rapid diversification in at least some of its clades (Molau, 1988; Frankel et al., 2022). Seeking to address these challenges, we present here taxon-specific genetic markers that provide a promising outlook for reconstructing robust and resolved phylogenies both in silico (Figure 4) and in vitro (Figure 5).

Our results indicated that, in general, markers obtained from Series I and II yielded similar phylogenetic results and recovered fully resolved and congruent trees (Figure 4B,C). When compared to one another, in silico and in vitro results provided very similar inferences, and including more species in the in vitro approach allowed us to gain a better insight of the relationships among species (see below). Importantly, through this process, we could confirm that the bait sets are likely to perform well on species not used in the bait set development, on outgroups and on low-quality DNA from historical/herbarium specimens. Finally, our integration and recovery of markers associated with genes involved in floral development in Series II will open doors to explicitly investigate potential pollinator-related evolutionary trends in the group, as well as provide a potential "baseline" for quantifying the neutrality of the anonymous markers selected here.

Compared to the plastome analysis by Frankel et al. (2022), the topology of the nuclear phylogeny recovered the same general split between species from the northern and southern Andes (Figure 5). However, our data set does not recover clades with representatives from all of the three main geographic regions known to be important to the family (i.e., middle, northern, and southern Andes; Molau, 1988), but rather yields a result similar to Cosacov et al. (2009): one clade including specimens from the northern and middle Andes, and another with specimens from the southern and middle regions (Figure 5). Furthermore, even though we only used 10 of the 14 samples used by Frankel et al. (2022), we found differences between Frankel's plastome tree (Figure 4F) and our nuclear inferences. First, some taxa included in both studies were assigned to different clades in each inference, such as we observed for the tetraploid *C. pedunculata*. This Ecuadorian species was embedded in a clade of mostly diploids from the southern Andes in Frankel et al. (2022), while our in vitro study recovers it with a clade of mainly tropical species. This difference could be due to different histories of the two genomes, but is likely an effect of the different sequence quality for this species in our data set compared to that of Frankel et al. (2022) (see above). Another remarkable difference is the position of the southern species *C. tenella* and *C. polyrhiza*. Our results recover them as sister to all southern taxa, while the plastid data set identifies them as sister to the northern clade. It is possible that our larger sequence and taxon sampling led to these differences, and/or that reticulated evolution is at play in the group.

An important aspect to consider in our phylogenetic reconstructions is the potential effect that polyploidization could have. Indeed, although we observed a higher proportion of single-copy loci among diploid species, we do not expect specimen ploidy to affect our phylogenetic reconstructions because we only used single-copy loci for tree estimation (see Methods section). This particular topic and its effect on the evolution of the group will be, however, further investigated in future studies, using allele phasing pipelines (i.e., Šlenker et al., 2021) and a larger taxon sampling.

When we compared the inferences using our in silico bait set and the "universal" Angiosperms353, we observed some important differences. Although the Angiosperms353 bait set yielded a similar tree topology (Figure 4E), we could only recover a very small number of single-copy markers, which likely led to an inability to properly resolve and support nodes associated with recent divergences (e.g., the southern Andes clade; Cosacov et al., 2009). Although it is possible that this low in silico locus recovery is due to the incomplete nature of the assemblies used to test the set, the fact that our bait set is able to target more than twice the number of loci likely makes it a much more informative set for understanding evolutionary processes occurring at the intergeneric and interspecific levels in Calceolariaceae. In this respect, our bait set represents a clear advance in our ability to understand the evolution of this group of plants and is opening new and exciting avenues to answer long-standing questions in the system and beyond.

Conclusions

In this work, we set out to develop a bait set for targeted sequencing in the historically challenging Calceolariaceae clade. Our 809-locus bait set allows us to do so, successfully recovering data for a wide range of DNA qualities and at multiple phylogenetic scales. The pipeline we present can be applied to other systems and is easily modified to target regions of interest in other groups, making it malleable and generalizable for other questions and study systems. Looking ahead, our next steps include implementing the use of this bait set to expand the sequencing to more Calceolariaceae taxa and ultimately reconstruct a completely sampled (or nearly

so) phylogeny of the clade that can be used to investigate the drivers of diversification (e.g., biogeographic history, effect of pollinators, climate, polyploidy) in a comparative framework, and to provide the tools for a taxonomic revision of *Calceolaria*. Furthermore, because of the presence of loci explicitly associated with floral traits, this bait set will also allow us to study the history of single-copy genes that are presumed to be under pollinator selection (Kellogg, 2003), providing a proxy for testing selective factors governing the evolution of this group.

AUTHOR CONTRIBUTIONS

N.M. and A.E. designed the bait set and sampling strategy for testing the bait set. N.M. and A.E. analyzed the data. N.M., A.E., and D.C.T. wrote the manuscript. All authors approved the final version.

ACKNOWLEDGMENTS

This work was funded by a National Science Foundation award (DEB 2050745) and University of Maryland, College Park, start-up funds allocated to A.E. We thank Dr. B. Oswald (Research Computing and Data Services, University of Idaho) for technical support and the herbaria GB, MO, MSB, and US (herbarium acronyms according to Thiers, 2023) for allowing sampling of leaf tissue from their collections.

DATA AVAILABILITY STATEMENT

DNA sequences of the targets and baits are available in Appendices S3 and S4.

ORCID

Nicolas Medina http://orcid.org/0000-0001-8983-5108
Anahí Espíndola http://orcid.org/0000-0001-9128-8836

REFERENCES

Andermann, T., M. F. Torres Jiménez, P. Matos-Maraví, R. Batista, J. L. Blanco-Pastor, A. L. S. Gustafsson, L. Kistler, et al. 2020. A guide to carrying out a phylogenomic target sequence capture project. *Frontiers in Genetics* 10: 1407.

Blum, M., H.-Y. Chang, S. Chuguransky, T. Grego, S. Kandasaamy, A. Mitchell, G. Nuka, et al. 2021. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Research* 49: D344–D354.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution* 17: 540–552.

Céspedes, C. L., J. R. Salazar, A. Ariza-Castolo, L. Yamaguchi, J. G. Ávila, P. Aqueveque, I. Kubo, and J. Alarcón. 2014. Biopesticides from plants: *Calceolaria integrifolia* s.l. *Environmental Research* 132: 391–406.

Chafin, T. K., M. R. Douglas, and M. E. Douglas. 2018. MrBait: Universal identification and design of targeted-enrichment capture probes. *Bioinformatics* 34: 4293–4296.

Chau, J. H., W. A. Rahfeldt, and R. G. Olmstead. 2018. Comparison of taxon-specific versus general locus sets for targeted sequence capture in plant phylogenomics. *Applications in Plant Sciences* 6: e1032.

Cosacov, A., A. N. Sérsic, V. Sosa, J. A. De-Nova, S. Nylinder, and A. A. Cocucci. 2009. New insights into the phylogenetic relationships, character evolution, and phytogeographic patterns of *Calceolaria* (Calceolariaceae). *American Journal of Botany* 96: 2240–2255.

Crowl, A. A., P. W. Fritsch, G. P. Tiley, N. P. Lynch, T. G. Ranney, H. Ashrafi, and P. S. Manos. 2022. A first complete phylogenomic hypothesis for diploid blueberries (*Vaccinium* section *Cyanococcus*). *American Journal of Botany* 109: 1596–1606.

Danecek, P., J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, et al. 2021. Twelve years of SAMtools and BCFtools. *GigaScience* 10: giab008.

Danisman, S. 2016. TCP transcription factors at the interface between environmental challenges and the plant's growth responses. *Frontiers in Plant Science* 7: 1930.

Edgar, R. C. 2004. MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5: 113.

Ehrhart, C. 2000. Die Gattung *Calceolaria* (Scrophulariaceae) in Chile. Schweizerbart'sche Verlagsbuchhandlung, Stuttgart, Germany.

Faircloth, B. C. 2016. PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics* 32: 786–788.

Frankel, L., M. Murúa, and A. Espíndola. 2022. Biogeography and ecological drivers of evolution in the Andes: Resolving the phylogenetic backbone for *Calceolaria* (Calceolariaceae). *Botanical Journal of the Linnean Society* 199: 76–92.

Gene Ontology Consortium. 2021. The Gene Ontology resource: Enriching a GOld mine. *Nucleic Acids Research* 49: D325–D334.

Gurevich, A., V. Saveliev, N. Vyahhi, and G. Tesler. 2013. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* 29: 1072–1075.

Harris, R. S. 2007. Improved pairwise alignment of genomic DNA. Ph.D. dissertation, The Pennsylvania State University, University Park, Pennsylvania, USA.

Johnson, M. G., L. Pokorny, S. Dodsworth, L. R. Botigué, R. S. Cowan, A. Devault, W. L. Eiserhardt, et al. 2019. A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. *Systematic Biology* 68: 594–606.

Karunanithi, P. S., and P. Zerbe. 2019. Terpene synthases as metabolic gatekeepers in the evolution of plant terpenoid chemical diversity. *Frontiers in Plant Science* 10: 1166.

Kellogg, E. A. 2003. What happens to genes in duplicated genomes. *Proceedings of the National Academy of Sciences, USA* 100: 4369–4371.

Khambay, B. P. S., D. Batty, M. Cahill, I. Denholm, M. Mead-Briggs, S. Vinall, H. M. Niemeyer, and M. S. J. Simmonds. 1999. Isolation, characterization, and biological activity of naphthoquinones from *Calceolaria andina* L. *Journal of Agricultural and Food Chemistry* 47: 770–775.

Koenen, E. J. M., C. Kidner, É. R. de Souza, M. F. Simon, J. R. Iganci, J. A. Nicholls, G. K. Brown, et al. 2020. Hybrid capture of 964 nuclear genes resolves evolutionary relationships in the mimosoid legumes and reveals the polytomous origins of a large pantropical radiation. *American Journal of Botany* 107: 1710–1735.

Kozlov, A. M., D. Darriba, T. Flouri, B. Morel, and A. Stamatakis. 2019. RAxML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* 35: 4453–4455.

Krueger, F., F. James, P. Ewels, E. Afyounian, and B. Schuster-Boeckler. 2021. FelixKrueger/TrimGalore: v0.6.7. Website: https://zenodo.org/record/5127899 [accessed 8 May 2021].

Lavandero, N., L. Santilli, and F. Pérez. 2021. *Calceolaria flavida* (Calceolariaceae) a new endemic species to central Chile. *PhytoKeys* 185: 99–116.

Le, H. T. T., L. N. Nguyen, H. L. B. Pham, H. T. M. Le, T. D. Luong, H. T. T. Huynh, V. T. Nguyen, et al. 2022. Target capture reveals the complex origin of Vietnamese ginseng. *Frontiers in Plant Science* 13: 814178.

León Yánez, S., R. Valencia, N. Pitman, L. Endara, C. Ulloa-Ulloa, and H. Navarrete. 2011. Libro rojo de las plantas endémicas del Ecuador. Publicaciones del Herbario QCA, Pontificia Universidad Católica del Ecuador, Quito, Ecuador.

Levy Karin, E., M. Mirdita, and J. Söding. 2020. MetaEuk—sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. *Microbiome* 8: 48.

Li, D., C.-M. Liu, R. Luo, K. Sadakane, and T.-W. Lam. 2015. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31: 1674–1676.

Lunter, G., and M. Goodson. 2011. Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research* 21: 936–939.

Manni, M., M. R. Berkeley, M. Seppey, F. A. Simão, and E. M. Zdobnov. 2021. BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Molecular Biology and Evolution* 38: 4647–4654.

Márquez, Y., F. Mantica, L. Cozzuto, D. Burguera, A. Hermoso-Pulido, J. Ponomarenko, S. W. Roy, and M. Irimia. 2021. ExOrthist: A tool to infer exon orthologies at any evolutionary distance. *Genome Biology* 22: 239.

Molau, U. 1988. Flora Neotropica, Vol. 47. Scrophulariaceae Part I. Calceolarieae. New York Botanical Garden, Bronx, New York, USA.

Mower, J. P., A. L. Case, E. R. Floro, and J. H. Willis. 2012. Evidence against equimolarity of large repeat arrangements and a predominant master circle structure of the mitochondrial genome from a monkeyflower (*Mimulus guttatus*) lineage with cryptic CMS. *Genome Biology and Evolution* 4: 670–686.

Murúa, M., and A. Espíndola. 2015. Pollination syndromes in a specialised plant-pollinator interaction: Does floral morphology predict pollinators in *Calceolaria*? *Plant Biology* 17: 551–557.

Murúa, M., J. Cisterna, and B. Rosende. 2014. Pollination ecology and breeding system of two *Calceolaria* species in Chile. *Revista Chilena de Historia Natural* 87: 1–3.

Nylinder, S. 2006. On the phylogeny of the genus *Calceolaria* (Calceolariaceae) as inferred from ITS and plastid matK sequences. *Taxon* 55: 125–137.

Nylinder, S., U. Swenson, C. Persson, S. B. Janssens, and B. Oxelman. 2012. A dated species–tree approach to the trans-Pacific disjunction of the genus *Jovellana* (Calceolariaceae, Lamiales). *Taxon* 61: 381–391.

Paniagua-Zambrana, N. Y., R. W. Bussmann, and J. Echeverría. 2020. *Calceolaria bartsiifolia* Wedd. *Calceolaria buchtieniana* Kraenzl. *Calceolaria engleriana* Kraenzl. *Calceolaria inamoena* Kraenzl. *Calceolaria stellariifolia* Phil. *Calceolaria rugulosa* Edwin Calceolariaceae. *In* N. Y. Paniagua-Zambrana and R. W. Bussmann [eds.], Ethnobotany of the Andes, 409–416. Springer International Publishing, Cham, Switzerland.

Puppo, P. 2014. Revision of the *Calceolaria tripartita* s. l. species complex (Calceolariaceae) using multivariate analyses of morphological characters. *Phytotaxa* 167: 61–78.

Puppo, P. 2023. *Calceolaria nivalis* subsp. *lanatifolia*, a new subspecies of Calceolariaceae from Northern Peru. *Phytotaxa* 591: 101–105.

Puppo, P., and P. Novoa. 2012. Revisión de la sección *Calceolaria* (Calceolariaceae) en Chile. *Gayana Botánica* 69: 275–285.

Quinlan, A. R., and I. M. Hall. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* 26: 841–842.

R Core Team. 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Website: http://www.R-project.org/ [accessed 26 October 2023].

Reijnders, M. J. M. F., and R. M. Waterhouse. 2021. Summary visualizations of gene ontology terms with GO-Figure! *Frontiers in Bioinformatics* 1: 638255.

Renner, S. S., and H. Schaefer. 2010. The evolution and loss of oil-offering flowers: New insights from dated phylogenies for angiosperms and bees. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365: 423–435.

Romero-Hernández, C., R. W. Bussmann, and P. Puppo. 2017. New species of *Calceolaria* (Calceolariaceae) from northern Peru. *Novon* 25: 316–321.

Sayyari, E., and S. Mirarab. 2016. Fast coalescent-based computation of local branch support from quartet frequencies. *Molecular Biology and Evolution* 33: 1654–1668.

Schmidt-Lebuhn, A. N. 2022. Sequence capture data support the taxonomy of *Pogonolepis* (Asteraceae: Gnaphalieae) and show unexpected genetic structure. *Australian Systematic Botany* 35: 317–325.

Sérsic, A. 2004. Pollination biology in the genus *Calceolaria* L. (Calceolariaceae). *Stapfia* 82: 1–121.

Simões, A. R. G., L. A. Eserman, A. R. Zuntini, L. W. Chatrou, T. M. A. Utteridge, O. Maurin, S. Rokni, et al. 2022. A bird's eye view of the systematics of Convolvulaceae: Novel insights from nuclear genomic data. *Frontiers in Plant Science* 13: 889988.

Siniscalchi, C. M., O. Hidalgo, L. Palazzesi, J. Pellicer, L. Pokorny, O. Maurin, I. J. Leitch, et al. 2021. Lineage-specific vs. universal: A comparison of the Compositae1061 and Angiosperms353 enrichment panels in the sunflower family. *Applications in Plant Sciences* 9(7): e11422.

Šlenker, M., A. Kantor, K. Marhold, R. Schmickl, T. Mandáková, M. A. Lysak, M. Perný, et al. 2021. Allele sorting as a novel approach to resolving the origin of allotetraploids using Hyb-Seq Data: A case study of the Balkan Mountain endemic *Cardamine barbaraeoides*. *Frontiers in Plant Science* 12: 659275.

Smit, A., R. Hubley, and P. Green. 2004. RepeatMasker. Website: https://www.repeatmasker.org/ [accessed 28 July 2021].

Soto Gomez, M., L. Pokorny, M. B. Kantar, F. Forest, I. J. Leitch, B. Gravendeel, P. Wilkin, et al. 2019. A customized nuclear target enrichment approach for developing a phylogenomic baseline for *Dioscorea* yams (Dioscoreaceae). *Applications in Plant Sciences* 7: e11254.

Tamames, J., and F. Puente-Sánchez. 2019. SqueezeMeta, A highly portable, fully automatic metagenomic analysis pipeline. *Frontiers in Microbiology* 9: 3349.

Thiers, B. 2023 (continuously updated). Index Herbariorum. Website: http://sweetgum.nybg.org/science/ih/ [accessed 27 October 2023].

Thomson, B., and F. Wellmer. 2019. Molecular regulation of flower development. *In* U. Grossniklaus [ed.], Plant development and evolution, 185–210. Academic Press, Cambridge, Massachusetts, USA.

Thureborn, O., S. G. Razafimandimbison, N. Wikström, and C. Rydin. 2022. Target capture data resolve recalcitrant relationships in the coffee family (Rubioideae, Rubiaceae). *Frontiers in Plant Science* 13: 967456.

Ufimov, R., J. M. Gorospe, T. Fér, M. Kandziora, L. Salomon, M. van Loo, and R. Schmickl. 2022. Utilizing paralogues for phylogenetic reconstruction has the potential to increase species tree support and reduce gene tree discordance in target enrichment data. *Molecular Ecology Resources* 22: 3018–3034.

UniProt Consortium. 2021. UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Research* 49: D480–D489.

Van Dam, M. H., J. B. Henderson, L. Esposito, and M. Trautwein. 2021. Genomic characterization and curation of UCEs improves species tree reconstruction. *Systematic Biology* 70: 307–321.

Villaverde, T., P. Jiménez-Mejías, M. Luceño, M. J. Waterway, S. Kim, B. Lee, M. Rincón-Barrado, et al. 2020. A new classification of *Carex* (Cyperaceae) subgenera supported by a HybSeq backbone phylogenetic tree. *Botanical Journal of the Linnean Society* 194: 141–163.

Vogel, S. 1974. Ölblumen und ölsammelnde Bienen [Oil flowers and oil-collecting bees]. *Tropische und subtropische Pflanzenwelt* 7: 283–547.

Wickham, H., M. Averick, J. Bryan, W. Chang, L. D'Agostino McGowan, R. Francois, G. Grolemund, et al. 2019. Welcome to the tidyverse. *Journal of Open Source Software* 4(43): 1686. https://doi.org/10.21105/joss.01686

Yang, Y., Q. Kong, A. R. Q. Lim, S. Lu, H. Zhao, L. Guo, L. Yuan, and W. Ma. 2022. Transcriptional regulation of oil biosynthesis in seed plants: Current understanding, applications, and perspectives. *Plant Communications* 3: 100328.

Zhang, J., Y. Liu, Y. Bu, X. Zhang, and Y. Yao. 2017. Factor analysis of MYB gene expression and flavonoid affecting petal color in three crabapple cultivars. *Frontiers in Plant Science* 8: 137.

Zhang, C., M. Rabiee, E. Sayyari, and S. Mirarab. 2018. ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* 19: 153.

SUPPORTING INFORMATION

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

Appendix S1. The taxonomic assignment and abundance of sequencing reads assigned to each sample by SqueezeMeta.

Appendix S2. List of genes and exons targeted from Series II.

Appendix S3. DNA sequences of targets obtained from this workflow.

Appendix S4. DNA sequences of baits obtained from this workflow.

How to cite this article: Medina, N., D. C. Tank, and A. Espíndola. 2023. Calceolariaceae809: A bait set for targeted sequencing of nuclear loci. *Applications in Plant Sciences* 11(6): e11557. https://doi.org/10.1002/aps3.11557