# A hybrid capture bait set for begonia

EDINBURGH JOURNAL OF BOTANY 79, *Begonia* special issue,
Article 409: 1–33 (2022). https://doi.org/10.24823/EJB.2022.409

Published by the Royal Botanic Garden Edinburgh
ISSN (online): 1474-0036, ISSN (print): 0960-4286

OPEN ACCESS
PEER REVIEWED

# A HYBRID CAPTURE BAIT SET FOR *BEGONIA*

T. Michel[1,2], Y.-H. Tseng[3], H. P. Wilson[2,4], K.-F. Chung[3] & C. A. Kidner[1,2*]

Hybrid capture with baits has proven to be a rich source of genetic data for many genera. The depth of information provided allows resolution of rapid radiations and of deep phylogenetic patterns. Retrieved data can also be used for population genetic studies and analysis of functional genetic diversity. To gain a better understanding of the evolutionary patterns across this large, diverse and fascinating genus through phylogenetics, population genetics and sequence analysis, we have designed and tested a set of 1239 baits covering low copy number and functionally annotated genes involved in shade adaptation and development and genetically linked to key traits. We demonstrate successful recovery of sequence data from species across *Begonia* and from fresh, silica-dried and older herbarium material.

*Keywords*. Begoniaceae, enrichment panel, probe set, target capture sequencing.
Received 10 August 2021 Accepted 19 April 2022 Published 18 August 2022

## Introduction

Hybrid capture is now a common method for retrieving sequence data for phylogenetics and population genetics in plants (Cronn *et al.*, 2012; Dodsworth *et al.*, 2019; Hale *et al.*, 2020; Larridon *et al.*, 2020; Slimp *et al.*, 2021). The method gives high coverage for hundreds of chosen loci across the genome, and the large numbers of loci recovered give a fuller picture of evolutionary history and allow exploration of reticulate patterns produced by hybridisation events. This high coverage gives confidence in variant calling and analysis of variation at population levels. The ability to choose loci allows testing of evolutionary hypotheses about the role of specific types of genes (McKain *et al.*, 2018).

The phylogenetic reach of bait sets varies. Cross-angiosperm baits have been produced and are becoming widely used for both deep and shallow phylogenomic studies (Johnson *et al.*, 2019; Larridon *et al.*, 2020; Slimp *et al.*, 2021). Family-wide sets have been useful in untangling relationships, particularly in large groups and recent radiations (e.g. Compositae, Mandel *et al.*, 2015; Euphorbiaceae, Villaverde *et al.*, 2018; mimosoid legumes, Koenen *et al.*, 2020; and Annonaceae, Couvreur *et al.*, 2018). Sets focused on specific genera have

[1] Institute of Molecular Plant Sciences, University of Edinburgh, The King's Buildings, Edinburgh EH9 3BF, Scotland, UK.
[2] Royal Botanic Garden Edinburgh, 20A Inverleith Row, Edinburgh EH3 5LR, Scotland, UK.
[3] Research Museum and Herbarium (HAST), Biodiversity Research Center, Academia Sinica, 128 Academia Road, Section 2, Taipei 115201, Taiwan.
[4] Institute of Biodiversity Animal Health & Comparative Medicine, Glasgow University, Graham Kerr Building, Glasgow G12 8QQ, Scotland, UK.
* Author for correspondence. E-mail: ckidner@rbge.ac.uk.

been useful in resolving relationships in difficult groups (Folk *et al.*, 2015; Pezzini, 2019; Soto Gomez *et al.*, 2019), and even species-specific sets have been designed for population genetics and breeding (e.g. barley, *Hordeum vulgare* L. [Hill *et al.*, 2019]; and wheat, *Triticum aestivum* L. [Gardiner *et al.*, 2019]).

*Begonia* L. is one of the largest plant genera (Frodin, 2004; Moonlight *et al.*, 2018) and presents a number of phylogenetic problems requiring increased resolution at both deep (sectional divisions) and shallow (recent radiations) levels. Additionally, there is a need to better understand the influence of past hybridisation events on present *Begonia* species diversity, given the evidence of multiple ancient and recent hybridisations (Goodall-Copestake *et al.*, 2010; Thomas *et al.*, 2012; Moonlight *et al.*, 2015; Tseng *et al.*, 2017; Liu *et al.*, 2019a). Hybrid capture provides a wealth of genetic data that can help resolve these issues and provide data for further studies of diversity and evolution across the genus.

In the present study, we aimed to provide background and advice on a genus-specific bait set for *Begonia* to encourage the use of this technology in addressing some of the questions about this exceptional genus. We describe how we designed a specific bait set for *Begonia* that incorporates developmental genes, differentially expressed genes, and genes linked to traits. Four projects have already used the bait set (Forrest *et al.*, 2019; Wilson, 2021; and two projects whose results are currently unpublished). We use data generated by these projects to characterise the performance of the bait set. We compare samples, captures and baits to determine the range of sample quality and species on which the bait set works, the range of hybridisation conditions, and the performance of individual baits in different captures. To compare sequence analysis pipelines, we analyse data from one capture using five pipelines to compare assembly metrics and phylogenies generated as concatenated data and gene tree analysis. The comparison of captures and analyses in this methods-focused paper will provide guidance for future projects using this bait set to answer the many questions about *Begonia* biology.

## Materials and methods

### *Bait design*

Our starting point was a transcriptome produced from mature leaves and male flower buds of the Asian species *Begonia luzhaiensis* T.C.Ku (Tseng *et al.*, 2017) (**Figure 1**). BLASTN of the transcriptome to itself was used to identify 15,349 sequences that share > 98% of their identity over 100 bp with another sequence. These were considered likely to be single-copy genes.

We then used BLASTN to compare the *Begonia luzhaiensis* transcriptome sequences with the set of annotated genes from the genome of the Central American *Begonia conchifolia* A.Dietr. (Campos-Dominguez, 2020). This identified 15,463 sequences with ‘good’ matches, that is, matches of > 90% identity over > 100 bp. A total of 11,261 sequences in this set were

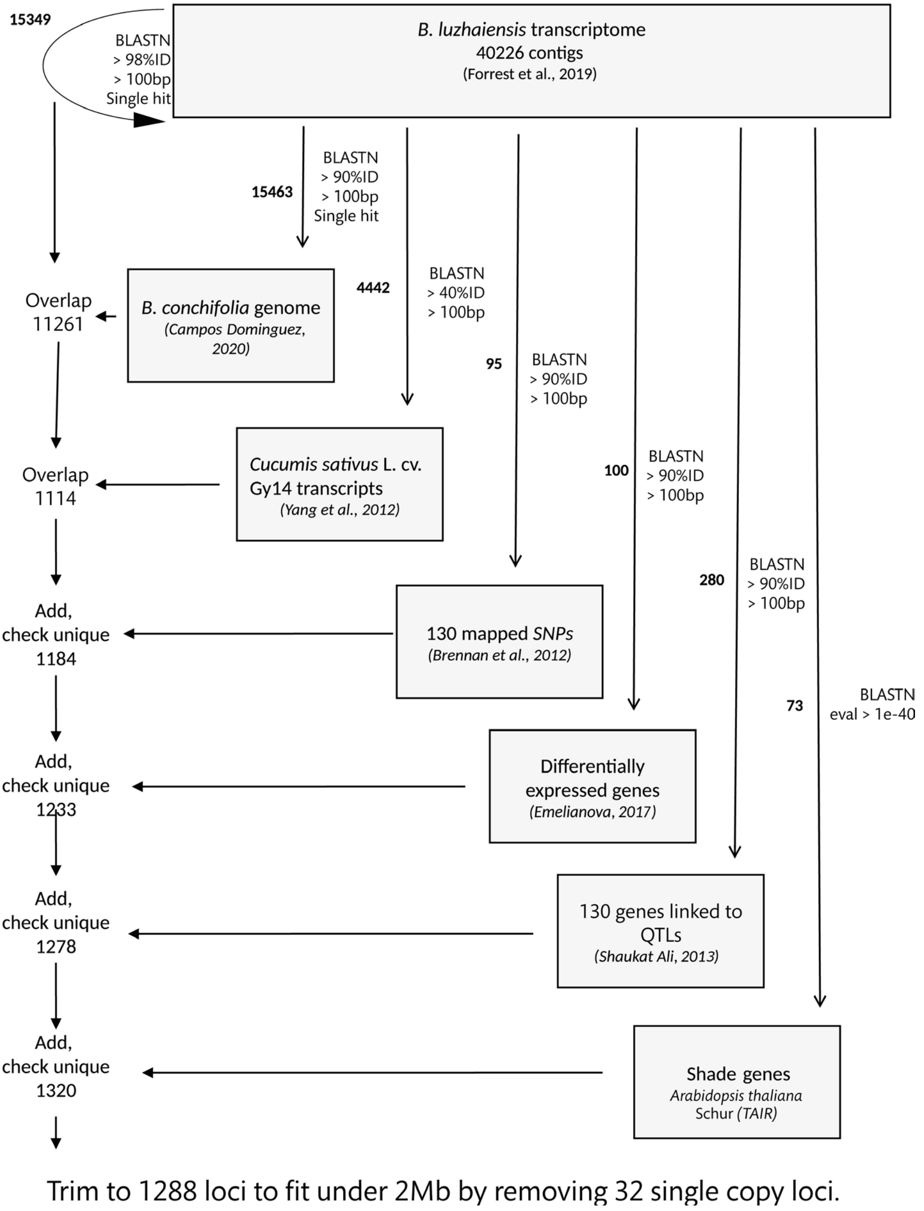

**Figure 1.** Design of a bait set for *Begonia*. Sequences from the *Begonia luzhaiensis* transcriptome were self-blasted to obtain possible single-copy genes; the overlap with the set of genes annotated in both the *B. conchifolia* genome and the *Cucumis sativus* genome was obtained. To this set were added sequences of the markers from a genetic map of *Begonia*, differentially expressed transcriptional factors, genes linked to quantitative trait loci (QTL), and a set of candidate genes for shade growth. Numbers in bold are the numbers of sequences retained at each step. ID, identity; SNP, single-nucleotide polymorphism.

determined to overlap with the ‘single-copy’ sequences identified in the previous step and therefore selected for further analysis. These 11,261 sequences were subsequently filtered by using the genome assembly method described by Yang *et al.* (2012) to identify only those with matches of > 90% over > 100 bp to an annotated cucumber gene. This step yielded 1114 sequences.

We added three sets of sequences that may be useful for functional studies in *Begonia*. The first consisted of matches to the 130 single-nucleotide polymorphism (SNP) markers used to generate the first *Begonia* genetic map (Brennan *et al.*, 2012). The second set was made up of matches to 100 transcription factors differentially expressed between *Begonia conchifolia* and *B. plebeja* Liebm. (Emelianova *et al.*, 2021). The third set comprised 280 sequences with good matches to genes falling within one logarithm of odds drop of significant quantitative trait loci for a range of traits in *Begonia conchifolia* × *B. plebeja* (Twyford *et al.*, 2014).

The final set of chosen sequences was a group of shade-associated genes. Having surveyed the literature on genes associated with light responses, particularly shade tolerance, we selected *PhyA*, *PhyB*, *PhyC*, *PhyD*, *PhyE*, *CRY1*, *CRY2*, *PIF3*, *PIF4*, *PIF5*, *GLK*, *PLASTID MOVEMENT* and *HAT4* (see the **Supplementary table** for details of these genes in *Arabidopsis* Schur). Using tBLASTN with a cut-off of 1e-40, we identified orthologues between the *Arabidopsis* proteins and the *Begonia luzhaiensis* transcriptome. These sequences were compared against those already chosen, and the newly identified sequences were then included to give a final set of 1320 target loci.

The final set of sequences exceeded the 2 MB of a standard kit size. Therefore, the number of sequences was trimmed to 1288 loci by removing some that had been identified only as single copy with a cucumber-annotated match. Daicel Arbor Biosciences (Ann Arbor, MI, USA) designed and synthesised 100-mer nucleotide baits, with 2.1 × tiling across our target sequences.

The first capture with this bait set was performed on samples from *Begonia* sect. *Coelocentrum* Irmsch. Genomic DNAs were extracted from fresh or dried material using the DNeasy Plant Mini Kit (Qiagen, Venlo, the Netherlands). Approximately 1 µg of DNA was sonicated in a Bioruptor Pico machine (Cosmo Bio, Tokyo, Japan) using a program to generate fragment sizes of 400–500 bp. Using 50 µL of sonicated DNA, a dual-indexed library for each sample was prepared using the NEBNext Ultra II DNA Library Prep kit (New England BioLabs, Ipswich, MA, USA), following the manufacturer’s protocol and selecting DNA in the range of 400–500 bp.

Each of the eight libraries was pooled into a 2 µg pool to perform hybridisation capture of target DNA using biotinylated RNA baits from the first custom-designed myBaits kit for *Begonia* (Arbor Biosciences, Ann Arbor, MI, USA). The custom bait set included 100-mer baits with 2.1 × tiling density across 1288 loci with 1,990,537 bp. The hybridisation procedure was performed at 65°C for 19 h following the myBaits v2.3.2 protocol. For post-capture PCR amplification, pools were amplified using 11 or 12 cycles. Target-enriched libraries were quantified using a Qubit 3.0 Fluorometer (Thermo Fisher Scientific, Waltham, MA, USA) and quality-checked using an Agilent Bioanalyzer (Agilent, Santa Clara, CA, USA). All samples were sequenced on the Illumina HiSeq platform (250 bp paired-end) (Illumina, San Diego, CA, USA) at the High Throughput Genomics Core at Biodiversity Research Centre, Academia Sinica, Taiwan. This data set is referred to as COEL.

Analysis of the targeted-capture sequencing output from this capture identified 13 loci with high paralogy and 83 loci with > 90% missing data. These loci were removed from the set and replaced with target sequences from the following two sets. For the first set, we identified interesting genes involved in regulating anthocyanin synthesis, flowering, leaf form, and epigenetic regulation, among other functions. We then used the *Arabidopsis* proteins to search for orthologues in the *Begonia conchifolia* genome using tBLASTN and a cut-off of 1e−40. The second set comprised sequences matching the Angiosperms353 panel (Johnson *et al.*, 2019). The original set of targets had an overlap of 23 loci with the Angiosperms353 enrichment panel. We added five further overlapping loci to take this number to 28. The final target sequences are presented in the **Supplementary data**, with annotation in the **Supplementary table**.

### *Capture experiments*

The bait set has been used in four capture projects so far (**Table 1**). The data for two, namely HAIR and PNG, have been published (Forrest *et al.*, 2019, and Wilson, 2021, respectively) and those for the other two, COEL and POP, are unpublished. We used the sequence data from these experiments to characterise the performance of the bait set.

### *HAIR*

These data derive from a study testing the degree to which various preservation techniques, including using a hairdryer, affected our ability to derive useful data from samples of three exemplar species (Forrest *et al.*, 2019). Sample extraction library preparation and capture methods have been described in detail by Forrest *et al.* (2019) and are summarised in **Table 1**.

### *PNG*

These data are taken from research in which samples from New Guinea begonias were used for a Ph.D. study of *Begonia*'s colonisation of the island (Wilson, 2021). Taxonomic sampling was wide and included many closely related species, multiple individuals for some species, and technical duplicates. Extractions were carried out using the standard Qiagen DNeasy Plant Mini Kit columns, following the manufacturer's protocol. Extracted DNA was quantified using Qubit dsDNA HS chemistry, with duplicate reads taken for all samples. Quality was

**Table 1.** Capture experiments carried out using the *Begonia* bait set

| Variable | HAIR | PNG | COEL | POP |
|---|---|---|---|---|
| Reference | Forrest *et al.* (2019) | Wilson (2021) | Tseng (in preparation) | Michel (in preparation) |
| No. of sections | 2 | 13 | 3 | 2 |
| Material | Fresh, silica-dried, herbarium | Fresh, silica-dried, herbarium | Fresh, silica-dried | Fresh, herbarium |
| DNA preparation | Qiagen | Qiagen | Qiagen | Qiagen |
| Library | NEBNext or TruSeq | NebNext Ultra II | NEBNext Ultra II | NEBNext |
| myBaits kit version | 4 | 4 | 3 | 4 |
| No. of samples pooled per capture | 11–17 | 4–10 | 8 | 10–19 |
| Hybridisation time, temperature | 14 h, 60°C | 16 h, 62°C | 19 h, 65°C | 24 h, 60°C |
| No. of post-capture PCR cycles | 13–20 | 9–22 | 11–12 | 16 |
| Sequencing | MiSeq 250 bp | HiSeq PE 150 bp | HiSeq PE 250 bp | NanoSeq6000 150 bp |
| Maximum capture efficiency (Bowtie2 to baits) (%) | 86 | 72 | 85 | 52 |
| Capture efficiency (mean coverage) | 634.9 | 105.0 | 206.9 | 437.3 |
| Sample no. (species, individuals) | 45 (3, 3) | 191 (152, 170) | 67 (67, 67) | 43 (4, 35) |

assessed on a DeNovix DS-11 (DeNovix, Wilmington, DE, USA) and an Agilent TapeStation. All DNA extractions were normalised to 1.9 ng/μL.

Samples containing high–molecular weight DNA were sheared to 300 bp using a Covaris M220 Focused-ultrasonicator (Covaris, Woburn, MA, USA). Library preparation was carried out as half-reactions using one NEBNext Ultra II DNA Library Prep Kit for Illumina, following the manufacturer's protocol. Samples that contained > 50 ng of DNA and had broad fragment size peaks on their ScreenTape were size-selected using Sera-Mag sample purification beads (Cytiva, Marlborough, MA, USA). The samples were then quality-checked using Agilent TapeStation High Sensitivity ScreenTapes, and then requantified on an Invitrogen Qubit 2.0 Fluorometer (Thermo Fisher Scientific) using Qubit dsDNA HS chemistry, such that each library could be normalised to 10 nM.

Libraries were grouped by quality into 32 pools (each containing 4–10 libraries). Hybridisation of the library pools followed the method described by Forrest *et al.* (2019), with some modifications, using a 16 h hybridisation at 62°C. The number of post-capture

PCR cycles depended on pool input DNA, varying from 9 to 22 cycles. Sequencing was carried out by NovogeneAIT Genomics, Singapore, on a HiSeq X, with 150 bp paired-end reads.

### *COEL*

These data come from a study of the radiation of *Begonia* sect. *Coelocentrum*, which included a wide range of samples, including samples from closely related species, but only a single sample per species. Sample preparation, library production and hybridisation followed the same workflow as described above. Details of these samples are being prepared for publication by Yu-Hsing Tseng and Kou-Fung Chung.

### *POP*

These data derive from a study of variant calling for population genetics using a mapping population (Brennan *et al.*, 2012) and closely related samples from *Begonia socotrana* Hook.f. and *B. samhaensis* M.Hughes & A.G.Mill. DNA extraction followed the standard Qiagen DNeasy Plant Mini Kit protocol. DNA was quantified using a Qubit 4 Fluorometer (Thermo Fisher Scientific) with the dsDNA HS chemistry kit, and a quality check performed on an Agilent TapeStation. All samples were normalised to 2 ng/µL before the fragmentation step.

Fresh and recent historical specimens were fragmented to 350 bp using a Covaris M220 Focused-ultrasonicator. Library preparation followed the protocol of the NEBNext Ultra II DNA Library Prep Kit for Illumina. Sera-Mag sample purification beads were used for size selection of samples above 50 ng, and clean-up was used for less concentrated samples. An Agilent TapeStation with High Sensitivity kit was used for a libraries quality check. Subsequently, libraries were normalised to 10 nM then pooled according to fragment size and quality.

Three pools of 10, 14 and 19 libraries were made. The hybridisation step followed the myBaits Hybridization Capture for Targeted NGS Manual version 4.01. According to the guidelines of the manual relating to degraded or contaminated DNA libraries, the hybridisation time was extended to 24 h with a temperature of 62°C. Sixteen post-amplification cycles were performed on all the samples. Pools were sequenced by Edinburgh Genomics (Edinburgh, UK) on a single lane of NovaSeq600 SP with 250 paired-end reads. Details of these samples are being prepared for publication by T. Michel and C. A. Kidner.

An overview of the hybridisations used is presented in **Table 1**. Comparisons of the capture by sample and by bait are shown in **Figures 2** and **3**.

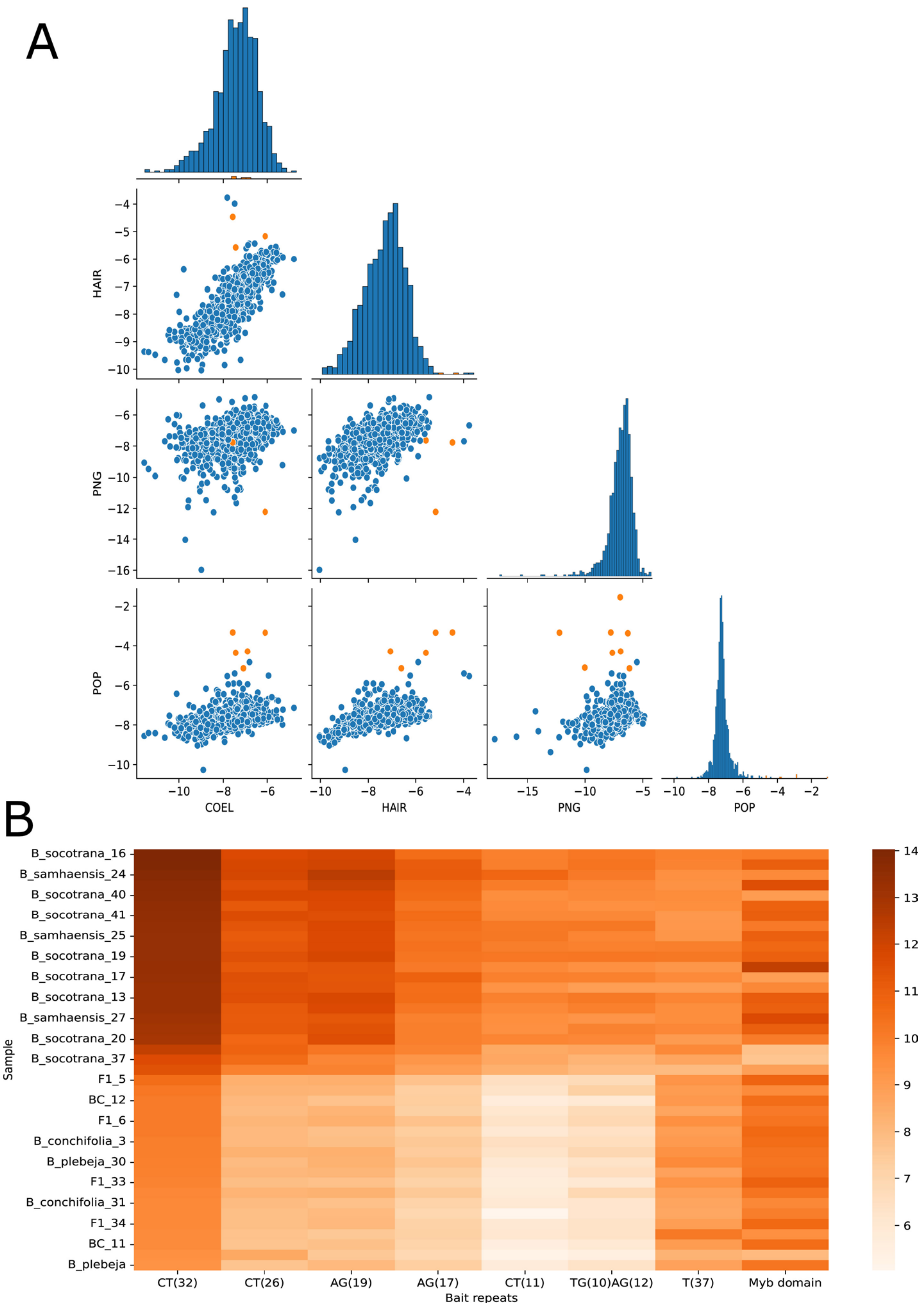

**Figure 2.** Variation in capture by target across experiments. A, Log percentage read capture per target per experiment. Orange data points are the eight targets with exceptionally high mean capture rates in experiment POP. B, Heat map of log read capture for experiment POP for the eight targets with exceptionally high mean capture rates.

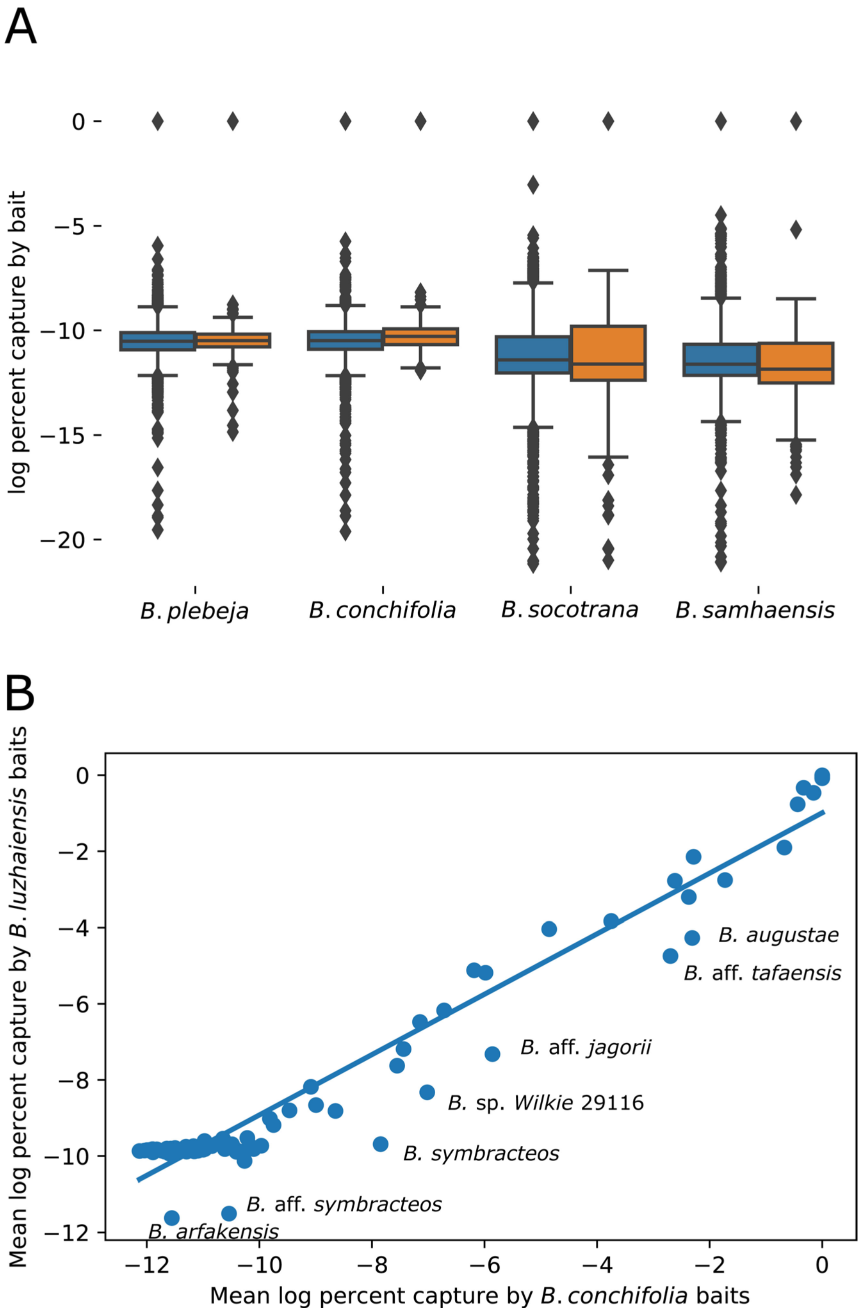

**Figure 3.** Capture by phylogenetic distance between target and sample. A, Log percentage read capture per target per species (data set POP). Blue: baits designed on *Begonia luzhaiensis*. Orange: baits designed on *Begonia conchifolia*. B, Mean log percentage read capture per target per species for data set PNG. Samples that show less capture by baits from *Begonia conchifolia* are labelled below the line.

### *Pipeline comparisons*

We compared five pipelines for generation of consensus sequence from captured reads on a subset of data from Wilson (2021): BASIC (Nicholls *et al.*, 2015), PALEOMIX (PAL) (Schubert *et al.*, 2014), HybPiper (Johnson *et al.*, 2016), SECAPR (Andermann *et al.*, 2018) and HybPhyloMaker (Fér & Schmickl, 2018). Pipelines were run on the Crop Diversity server (funded by BBSRC BB/S019669/1), except for HybPhyloMaker, which was run on the server of the Biodiversity Research Center, Academia Sinica.

The data we chose to compare the pipelines are from the PNG data set of Wilson (2021). This was a group of 47 samples of *Begonia* sect. *Symbegonia* (Warb.) L.L.Forrest & Hollingsw., a small section endemic to New Guinea. This data set was chosen to allow comparison of technical replicates with samples from the same populations and from closely related species. We had data for 15 species, with more than one sample for eight of these and 11 technical replicates. The choice of technical replicates was not optimal but represented cases in which low yield required the generation of a second set of data, or when a silica sample was used to confirm results from a herbarium sample.

The BASIC pipeline has been used for *Inga* Mill, *Begonia* and *Ceiba* Mill. (Nicholls *et al.*, 2015; Hart *et al.*, 2016; Forrest *et al.*, 2019; Pezzini, 2019). It uses a very conservative approach to align reads to the bait sequences using Bowtie2, using SAMtools and BCFtools (Li & Durbin, 2009; Li *et al.*, 2009). We followed the method described in Nicholls *et al.*, 2015). Reads were cleaned using Trimmomatic v0.30 (Bolger *et al.*, 2014). Bowtie2 v2.0.2 (Langmead & Salzberg, 2012), was used to align the reads back to bait sequences with an alignment score parameter of 140 to minimise mapping of paralogues. A VCF file was generated using BCFtools and filtered for a quality score of > 36 and to remove indels. A consensus sequence was generated with a custom Perl script (https://github.com/ckidner/Targeted_enrichment.git) and ambiguity codes converted to Ns.

HybPiper (Johnson *et al.*, 2016) maps reads to reference target sequences, using BWA (Li & Durbin, 2009). BLASTx (Altschul *et al.*, 1990) then extracts the reads that map to each locus, using SAMtools (Li *et al.*, 2009), and performs a *de novo* assembly for each locus, using SPAdes (Bankevich *et al.*, 2012). It then removes target flanking regions, using Exonerate (Slater & Birney, 2005), and selects the best contig to represent each locus. Alternatively, the intronerate.py script can be used to keep said flanking regions and create 'supercontigs', flagging putative paralogues on the process, which can later be investigated. Our run of HybPiper (Johnson *et al.*, 2016) used the default settings and the script 'reads_first.py' (as described in https://github.com/mossmatters/HybPiper/wiki), with BWA (Li & Durbin, 2009) as aligner and SPAdes for assembly (Bankevich *et al.*, 2012). The supercontigs were extracted for analysis with the corresponding Python script.

HybPhyloMaker (Fér & Schmickl, 2018) is a complete sequence to phylogeny pipeline that has been used for phylogenomic research at different taxonomic levels, such as in studies on *Ranunculus* L. (Tomasello *et al.*, 2020), Asteraceae (Jones *et al.*, 2019) and

Zingiberales (Carlsen *et al.*, 2018). Our run of the HybPhyloMaker pipeline (Fér & Schmickl, 2018) used the following settings: reads were cleaned with Trimmomatic v0.30 (Bolger *et al.*, 2014); mapped to the 'pseudoreference' based on the bait sequences using BWA (Li & Durbin, 2009); consensus sequence per locus generated with minimum relative abundance of the alternative base ('plurality' in the setting file of HybPhyloMaker) of 0.3; maximum number of heterozygous sites per exon of four; a minimum read coverage for ambiguity calling ('mincov') of 10; and 51% majority consensus for base calling using Kindel v.0.1.4 (Constantinides & Robertson, 2017).

SECAPR has been used to analyse data sets including those derived from palms (Helmstetter *et al.*, 2020), *Alchemilla* L. (Morales-Briones *et al.*, 2018) and Ochnaceae DC. (Schneider *et al.*, 2021). In this study it was run with default settings. Trimmed and cleaned reads were assembled *de novo* and contigs matching target regions were extracted using the original bait sequence as a reference. Sequence alignments were built using MAFFT (Katoh & Toh, 2008) for all loci present in at least three samples. Cleaned reads from the start of the analysis were then aligned to a file containing all alignments and a consensus produced for each locus, for each sample.

To investigate an approach designed for damaged DNA (potentially useful for herbarium samples), we used PALEOMIX (Schubert *et al.*, 2014) following the methods described online (https://paleomix.readthedocs.io/en/latest/) and generating consensus sequences per locus per sample, using SAMtools (Li *et al.*, 2009). PALEOMIX has been widely used on data from animals (e.g. Frantz *et al.*, 2016; Schubert *et al.*, 2017) but less frequently on data from plants (Vallebueno-Estrada *et al.*, 2016). The key step is the use of mapDamage 2.0 (Jónsson *et al.*, 2013) to identify the signatures of typical ancient DNA and rescale the quality scores based on this analysis.

We ran PALEOMIX using default parameters for cleaning and trimming reads and with BWA (Li & Durbin, 2009) for aligning reads to the reference bait sequences. PCR duplicates were marked and removed. The program mapDamage2 (Jónsson *et al.*, 2013) was used to recalibrate the BAM files to reduce the errors related to aDNA damage patterns. BCFtools was used to call variants, normalise indels, filter adjacent indels within 5 bp, and call consensus sequences for each locus for each sample (Li & Durbin, 2009).

A consensus per locus per sample was derived from each pipeline. The multi-FASTAs were rearranged by locus rather than by sample, using a custom python script (https://github.com/ckidner/Targeted_enrichment.git); aligned using MAFFT v7.475 (Katoh & Toh, 2008); and trimmed using trimAl v1.4.rev15 with strict settings (Capella-Gutiérrez *et al.*, 2009).

The alignments of each target were concatenated using AMAS (Borowiec, 2016). Phylogenies were produced for the concatenated matrix, using IQ-TREE multicore version 2.1.2 (parameters: -B 1000 -m MFP+MERGE -alrt 1000) (Nguyen *et al.*, 2015). We portioned the analysis by target locus except for SECAPR and BASIC, for which there were too many missing data to allow this. Using our standard pipeline, individual trees for each target were generated using FastTree version 2.1.10 (Price *et al.*, 2010) to produce a species tree with ASTRAL.5.7.7 (Zhang *et al.*, 2018). Metrics were collected using AMAS (Borowiec, 2016) and IQ-TREE (Nguyen *et al.*, 2015) (**Table 2**). We used PhyParts to analyse gene tree bipartitions (Smith *et al.*, 2015) and ETE3 to analyse Robinson–Foulds (RF) distances (Huerta-Cepas *et al.*, 2016). Phylogenies were visualised using FigTree v.1.4.4 (Rambaut, 2012: https://github.com/rambaut/figtree/releases/tag/v1.4.4) and ETE3 following directions available at https://github.com/mossmatters/MJPythonNotebooks/blob/master/PhyParts_PieCharts.ipynb to visualise bipartition analysis results. Tree comparisons used hierarchical clustering analysis carried out in SciPy Version 1.7.1 with scipy.cluster.hierarchy on the RF distances and tree space (Jombart *et al.*, 2017) using Kendall–Colijn distances (Kendall & Colijn, 2016).

## Results

### *Design of the bait set*

We have produced a bait set that works across *Begonia* and includes sequences for genes likely to be of interest in the genus. We started with likely single-copy genes from a transcriptome produced from the Asian species *Begonia luzhaiensis* (Tseng *et al.*, 2017).

We were concerned that the baits ought not to be anonymous, because for many downstream analyses it is important to understand the function of the genes used. At the time we were designing the baits, the closest related species with a well-annotated genome available was cucumber (*Cucumis sativus* L.), so we limited target sequences to those annotated in the cucumber genome assembly (Yang *et al.*, 2012) (see **Figure 1**).

We wished to link the bait set to work already done and to maximise our ability for use in functional studies. We added sequences that matched markers used to generate the first *Begonia* genetic map (Brennan *et al.*, 2012), along with genes linked to quantitative trait

**Table 2.** Alignment metrics per pipeline (post trimming)

| Metric | BASIC | HybPiper | HybPhyloMaker | SECAPR | PALEOMIX |
|---|---|---|---|---|---|
| Length of alignment (bp) | 1,806,453 | 1,552,006 | 1,814,262 | 1,002,748 | 1,940,728 |
| No. of patterns | 261,684 | 381,325 | 194,954 | 19,133 | 78,046 |
| No. of all-gap or ambiguous sites | 188,706 | 0 | 5518 | 79,510 | 965 |
| No. of phylogenetically informative sites | 28,110 | 78,788 | 194,954 | 42,555 | 143,386 |
| No. of singletons | 47,654 | 242,602 | 70,742 | 53,583 | 122,762 |
| No. of constant sites | 1,730,689 | 1,230,616 | 1,692,575 | 906,610 | 1,674,580 |
| No. of samples with > 50% gaps | 13 | 16 | 9 | 0 | 7 |

loci from an analysis of species-level variation (Twyford *et al*., 2014) and a DESeq analysis (Emelianova *et al.*, 2021) and sequences from transcription factors differentially expressed between *Begonia conchifolia* and *B. plebeja* (Emelianova, 2017).

Most *Begonia* are shade-adapted, and we wished to allow sequence analysis of key genes in the pathways of light perception and response. We added the *Begonia luzhaiensis* orthologues of genes associated with light responses, particularly shade tolerance. Based on performance in the *Begonia* sect. *Coelocentrum* capture (see **Table 1**), we refined the bait set, removing baits that captured many paralogues or that captured poorly, and replacing them with sequences of developmental genes and matches to the Angiosperms353 baits. We identified 13 target sequences with high paralogy and 83 that failed to capture. These sequences were removed from the set and replaced with sequences of several more developmental gene sequences matching the Angiosperms353 bait set (Johnson *et al.*, 2019). This set was used for the three further captures that generated the data analysed here.

### *Comparison of hybridisation protocols*

To determine how well the bait set worked, we compared four captures using the initial set and the revised set of baits (see **Table 1**). All the captures worked, although there was considerable variation depending on the quality of the sample, as has also been reported by Forrest *et al.* (2019) (see **Table 1**). This particularly affected very poor herbarium samples in the POP set (four with < 10 ng of DNA input) and the PNG set (eight with < 10 ng of DNA input).

### *Variation between baits*

**Figure 2A** shows the log percentage of reads captured for each target compared across experiments. Eight of the targets (labelled 'odd') showed very high capture rates in the POP data set. These eight targets did not have correspondingly high capture rates in the other experiments. Examination of these target loci revealed simple sequence repeats (SSRs) in six of them [Becon104Scf00540g0006.1;CT(32), Becon104Scf00540g0002.1;CT(26), ACmerged_contig_9951;AG(19), ACmerged_contig_1166;AG(17), ACmerged_contig_2307;CT(11), Becon104Scf01167g0029.1;TG(10)AG(12)], a poly-T motif in a seventh one [ACmerged_contig_5451;T(37)], and a myb domain in the eighth (ACmerged_contig_20957). The SSRs showed high capture rates in samples from *Begonia socotrana* and *B. samhaensis*. The poly-T motive and the myb domain showed high capture rates in three degraded samples from a specific species (see **Figure 2B**).

We wish these baits to work across *Begonia* without bias from how related the samples are to the species used for bait design. We tried to determine if targets from a closely related species were captured better than targets from a more distant species. **Figure 3** shows the comparison between log percentage capture by target for species in the POP data set. There was a greater range of capture efficiency in the targets based on sequences from *Begonia luzhaiensis* because there were many more targets (1192 compared with 47 *B. conchifolia* targets), but the baits designed from *B. conchifolia* targets did not capture better than the ones derived from *B. luzhaiensis*, even in *B. conchifolia* samples (**Figure 3A**). Some species in the PNG data set had fewer reads captured by the baits designed from *Begonia conchifolia* targets, but these species are not more phylogenetically distant to *B. conchifolia* than those that did not show this difference (**Figure 3B**).

### *Comparison of assembly pipelines*

We used sequence data from *Begonia* sect. *Symbegonia* generated as part of the PNG data set to compare the performance of five approaches to assembling consensus sequences from the captured reads. We chose to compare HybPiper (Johnson *et al.*, 2019), HybPhyloMaker (Fér & Schmickl, 2018), SECAPR (Andermann *et al.*, 2018), PALEOMIX (Schubert *et al.*, 2014) and the basic pipeline we had previously used on *Inga* (referred to here as BASIC) (Nicholls *et al.*, 2015). We compare ease of installation and use, and the amount and consistency of the results produced.

The BASIC pipeline was simple to set up and very fast. It generated a highly conservative consensus with relatively high proportions of missing data due to ambiguous (heterozygous) sites and indels being removed.

HybPiper (Johnson *et al.*, 2019) had the advantage of excellent instructions for installing and running (https://github.com/mossmatters/HybPiper/) but required very specific formatting of input files and extensive memory space due to requiring unzipped read files and generation of many interim files. Several samples needed to be re-run because initial memory settings on the cluster (32 G) were insufficient. Two paralogues (or allelic variants) were identified for most targets.

HybPhyloMaker (Fér & Schmickl, 2018) is also supported by an informative github (https://github.com/tomas-fer/HybPhyloMaker) and is relatively simple to set up and run. The pipeline performs all steps of Hyb-Seq data analysis from raw reads to species tree reconstruction, calculates and summarises the alignment and gene tree, and implements several species tree reconstruction methods. The preparation steps for renaming the raw reads and folder structure are more time-consuming in HybPhyloMaker than in other pipelines, but because it covers the whole process from raw reads to phylogeny, the time spent is worthwhile.

SECAPR is semi-automated and designed to be as easy as possible for new users; as such, it can be installed using Conda (Andermann *et al.*, 2018). Individual consensus sequences generated using SECAPR were shorter than those generated using the other pipelines, possibly due to the two-step process of *de novo* assembly followed by mapping to the assembled contigs, so the final length of consensus sequences depends on the success of the *de novo* assembly as well as the mapping step.

To investigate an approach designed for damaged DNA, which may be useful for herbarium samples, we also used the bam pipelines of PALEOMIX (Schubert *et al.*, 2014) to process bams from BWA alignment (Li & Durbin, 2009), followed by a basic consensus calling using SAMtools and BCFtools (Li & Durbin, 2009; Li *et al.*, 2009). This approach was fast and made no excessive memory demands.

We recovered extensive sequences using all approaches (see **Table 2**; **Figure 4**). The PALEOMIX pipeline generated the most data, and the BASIC pipeline the least (see **Table 2**). The BASIC pipeline removes all ambiguous sites and all indels, whereas the PALEOMIX pipeline assembles as long a sequence as possible using the reads that map to the reference. The BASIC pipeline also produced a concatenated matrix with the fewest number of phylogenetically informative (PI) sites and a high proportion of constant sites (see **Table 2**). PALEOMIX produces the targets with the highest distribution of PI sites (see **Figure 4C**). Both HybPiper and HybPhyloMaker output sequences only when a certain amount of data are present, therefore some targets are missing sequences for samples with poor data (see **Figure 4A**). Many of the SECAPR sequences are short (see **Figure 4B**) and, as expected, recovered targets have a low number of PI sites.

The maximum likelihood (ML) trees produced from the concatenated alignments vary between pipelines (**Figures 5**, **6**). HybPhyloMaker and PALEOMIX are least similar (RF distance, 0.69) and HybPiper and SECAPR are the most similar, despite the differences in tree length (RF distance, 0.27) (**Figure 6B,C**).

The BASIC pipeline has the least distance between technical replicates and the HybPiper pipeline the most (**Figure 6A**). This is probably due to the very conservative approach of the BASIC pipeline and our use of all consensus contigs from the HybPiper output (to facilitate comparison with other pipelines for the consensus calling) rather than a full paralogue-sensitive analysis. The placement of technical replicates as anything other than close sisters is concerning, but in this case, it may be due to the poor quality and quantity of the samples for the technical replicates.

HybPiper produced the longest tree and the BASIC pipeline the shortest. HybPiper produced long branches for many samples, contributing to a very long tree overall (see **Figure 5F**). These might be reduced by careful selection of which loci to include, removing all those identified as having paralogues. It is notable that different samples are placed on long branches by different pipelines, suggesting that the issue is not simply a high rate of gene duplications in a subset of species. This could represent random noise in the mapping leading to calling different paralogues, or sensitivity to different types of errors in HybPiper and PALEOMIX.

### *Variation among gene trees*

The ASTRAL quartet score analysis and the Phyparts bipartition analysis show the very high ratio of noise to signal in our data (**Figure 7**). Some nodes in our species trees are supported

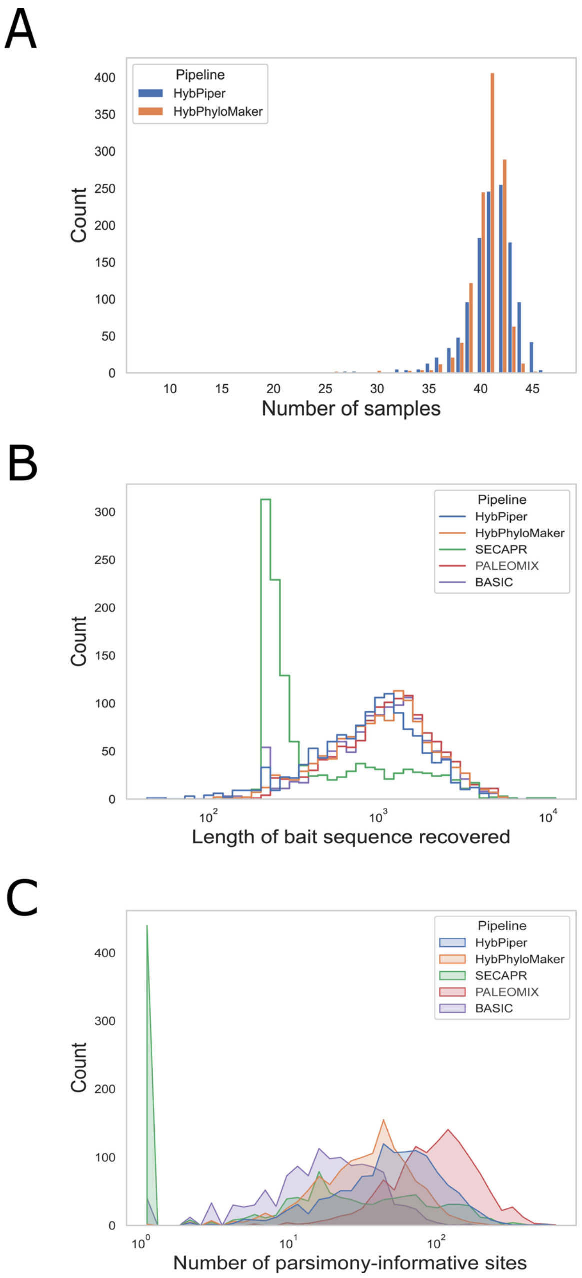

**Figure 4.** Alignment metrics for 47 *Begonia* sect. *Symbegonia* samples by pipeline. A, Number of consensus sequences recovered per sample for HybPiper and HybPhyloMaker pipelines (other approaches produced data for every locus). B, Distribution of mean length of sequence recovered per target across samples using each approach. C, Distribution of number of parsimony-informative sites per target using each approach.

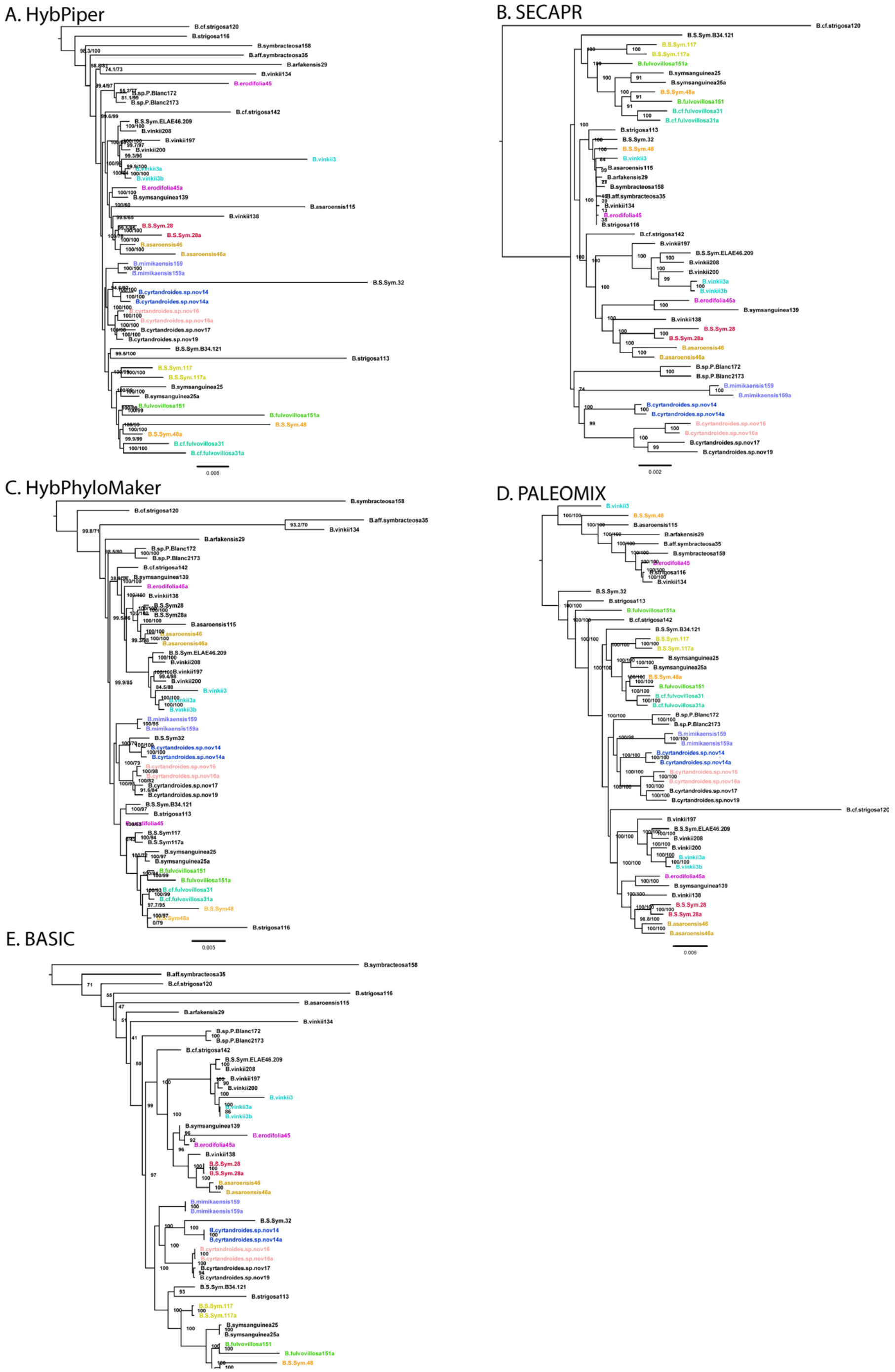

**Figure 5.** Phylogenies for *Begonia* sect. *Symbegonia* samples by pipeline. IQ-TREE concatenated maximum likelihood trees for each pipeline: A, HybPiper; B, SECAPR; C, HybPhyloMaker; D, PALEOMIX; E, BASIC.

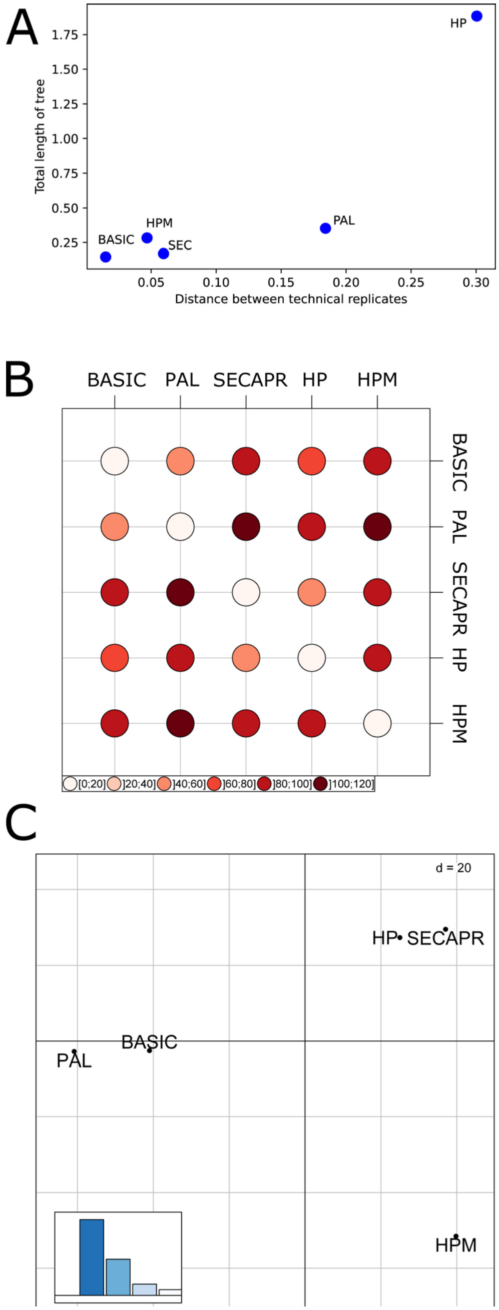

**Figure 6.** Comparisons between phylogenies produced from each pipeline. A, Distance between technical replicates by total tree length for each pipeline; B, pairwise comparison of trees by Kendall–Colijn metrics; C, principal coordinates analysis of tree distances. HP, HybPiper; HPM, HybPhyloMaker; PAL, PALEOMIX; SEC, SECAPR.

by very few gene trees; in all analyses, we see a minimum of 2/1239 gene trees supporting a given node, whereas only three nodes in each analysis are supported by a majority of genes. The well-supported nodes vary between analyses, and in many cases even the technical replicates are not supported by a majority of gene trees, suggesting that variation between gene trees of target loci reflects not only biological processes such as hybridisation and incomplete lineage sorting, but also noise and error in sequence assembly.

We investigated this further by examining the RF distance between each gene tree and the ASTRAL tree in each pipeline (Figure 8A). SECAPR had a large number of gene trees very different to the species tree, and HybPiper had a single gene tree (for target Becon104Scf03147g0009.1, a Dicer-like3 orthologue) that was a very close match to the ASTRAL tree (RF distance, 0.33). Overall, there is a positive correlation between the gene trees from each pipeline, supporting the conclusion that an overlapping set of genes are contributing to the ASTRAL species tree in each case. Loci that have a higher-than-average similarity to the species tree across all pipelines include Phytochrome A (ACcontig_8273), Phytochrome C (ACcontig_4025) and Plastid Movement Impaired (ACcontig_6745, AT5G26160.2).

Some of the similarities between pipelines in which gene trees are best related to the species tree derives from the influence of locus length. Longer loci produce genes trees that are closer to the species tree (Figure 8B). Consensus sequences produced by SECAPR were 30% shorter than the reference bait sequences, resulting in many gene trees that individually had little phylogenetic information. The variation seen between normalised RF distances for each bait points to the influence of noise, error and variation between consensus calling in each pipeline on the phylogeny produced.

We used the data from the pipeline with highest PI per target to look for structure within the gene tree space. We calculated the reciprocal RF distance between all the gene trees from the PALEOMIX pipeline and ran a hierarchical clustering analysis (in SciPy) to determine if there were distinct groups of trees. At least two and possibly four or more groups were detected (Figure 8C). Group 1 (orange in Figure 8C) comprises 465 trees generally distant to the species tree, group 2 (blue in Figure 8C) comprises 778 trees more similar to the species tree. To determine whether a similar structure existed in the gene trees produced by the other pipelines, we used the grouping from the PALEOMIX analysis to codify the targets across all pipelines and plotted the normalised RF distances between gene and species trees as box plots (Figure 8D). The group 1 set (the set of targets in orange in Figure 8C) were generally more distant from the species tree for all pipelines except SECAPR. This suggests that the groups detected in the clustering analysis reflect a true pattern and are not an artefact of a particular pipeline.

To examine the two groups of gene trees in more detail, we ran an ASTRAL analysis on each group, followed by PhyParts to illustrate the gene tree support at each node (Supplementary figure). The two trees differ with an RF distance of 0.49 (44/90). Nine of the

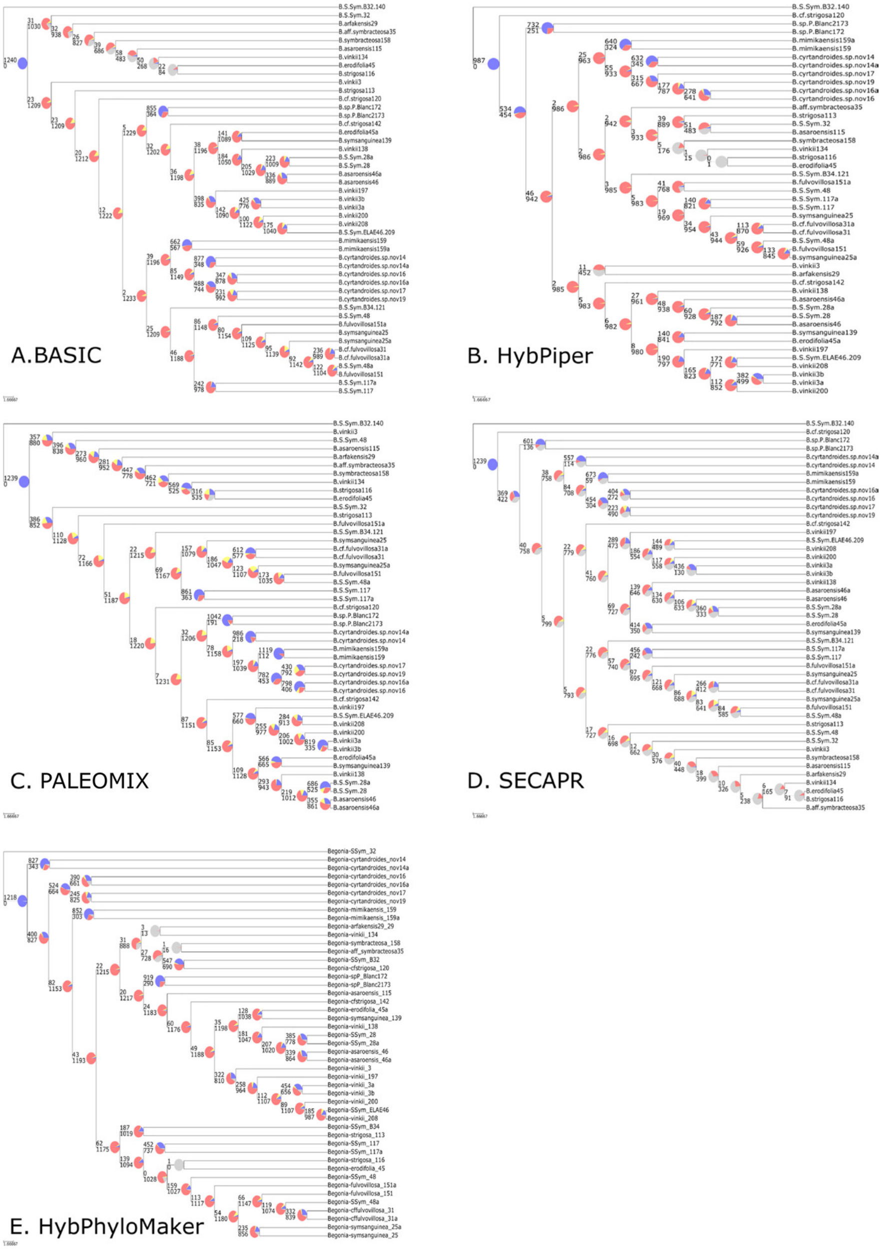

**Figure 7.** ASTRAL trees with PhyParts analysis for each pipeline: A, BASIC; B, HybPiper; C, PALEOMIX; D, SECAPR; and HybPhyloMaker. Numbers at each node show the numbers of supporting gene trees over the number of conflicting gene trees. At each node is a supporting gene trees/conflicting gene trees pie chart in which blue indicates the proportion supporting the topology; yellow, the proportion supporting the next most common bipartition; red, all other conflicting gene trees; and grey, the proportion with no support for a conflicting bipartition.

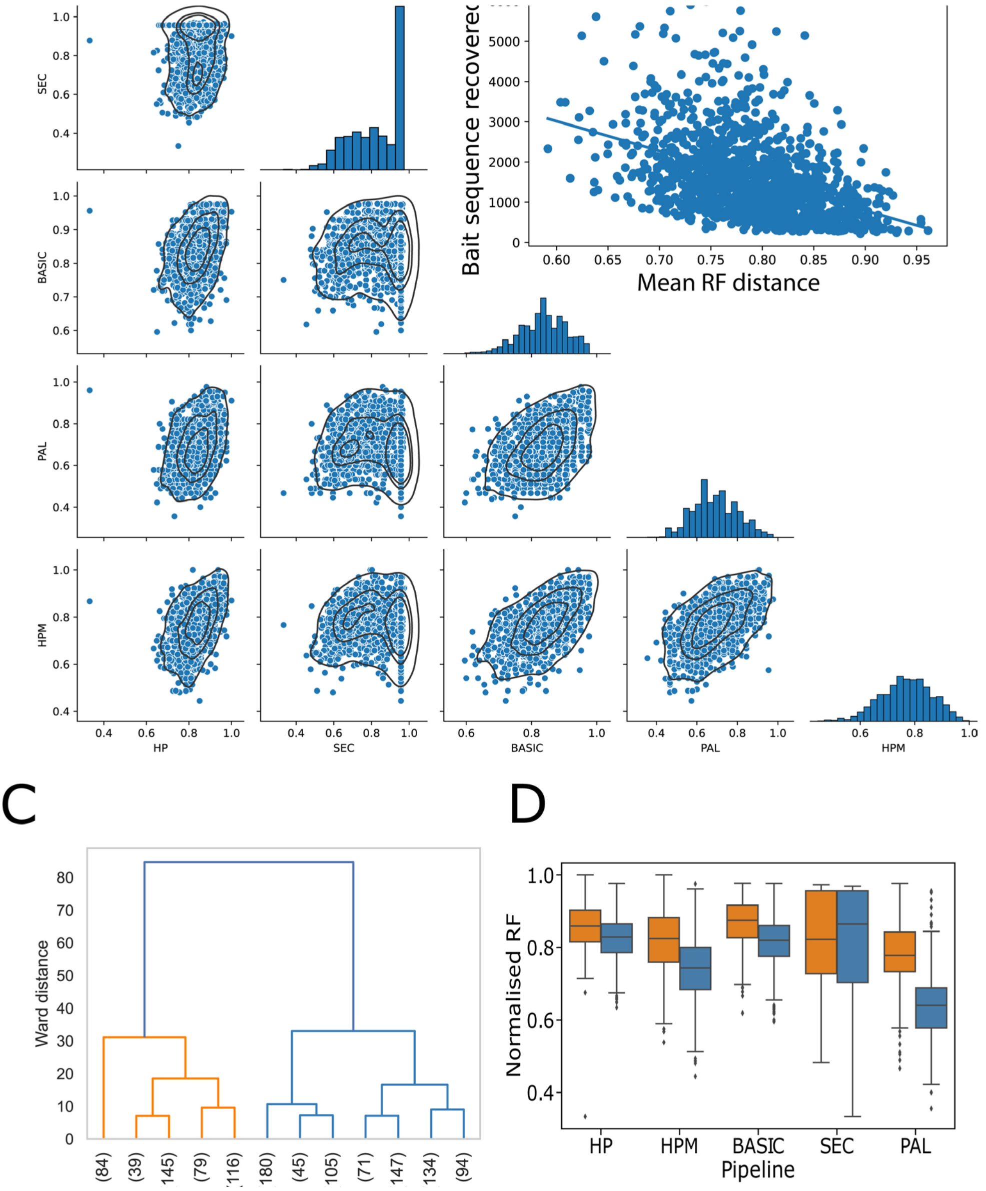

**Figure 8.** Patterns across gene trees. A, Robinson–Foulds (RF) tree distances between each gene tree and the species tree for each pipeline. As the points overlap, contour lines show the point density; B, mean RF distance across pipelines and the length of the bait sequence; C, hierarchical clustering of RF distance between individual gene trees and the species tree in the PALEOMIX pipeline; D, RF distance gene tree to species tree by pipeline for baits in the two groups identified in the PALEOMIX hierarchical clustering. Group 1, orange; group 2, blue. HPM, HybPhyloMaker; PAL, PALEOMIX; SEC, SECAPR.

11 sets of technical replicates map as sisters on the group 2 tree, seven with support from the majority of gene trees, but only six are sisters on the group 1 tree, and of these, three have support from the majority of gene trees. This suggests that group 1 contains trees with more noise and error than group 2.

## Discussion

### *Why use a target capture approach in* Begonia*?*

The results presented here show that the *Begonia* bait set resolves species-level phylogenies well, and the large number of variants identified supports its use on a population level. However, sequencing costs are decreasing to the point at which it is feasible to use genome skims for phylogenetics, which begs the question: why deal with the extra laboratory time and expense of hybrid capture? For some questions genome skims are a useful approach, but the complexity and large size of plant genomes mean that skims are an inefficient way of gathering functional genetic data. The prevalence of gene family expansions and partial and whole genome duplications in plants is also easier to deal with for a limited set of loci than with skim data.

Because there are more than 2000 *Begonia* species, few living collections hold anywhere near a representative collection of the genus needed for genomic studies. Additionally, field collection of samples is hampered by the remote and often difficult terrain in which many *Begonia* species are found. This has meant that *Begonia* phylogenetic studies must often rely on data from herbarium collections. For such old, degraded DNA, hybrid capture represents the only way to extract reliable sequence data across the nuclear genome (Hart *et al.*, 2016), and the *Begonia* baits set has been shown to retrieve useful data from even very poor herbarium samples (Forrest *et al.*, 2019).

Current phylogenetic studies on *Begonia* indicate a large number of rapid radiations and likely hybridisation events (Goodall-Copestake *et al.*, 2010; Thomas *et al.*, 2012; Moonlight *et al.*, 2015; Tseng *et al.*, 2017; Liu *et al.*, 2019a). The large number of unlinked SNPs recovered from hybrid capture (see **Figure 4B,C**) is ideal for analysis of rapid radiations and reticulate lineages (Shee *et al.*, 2020; Thomas *et al.*, 2021) and offers our best hope of understanding the complexities of *Begonia* evolution.

Studies have shown that broad bait sets such as the Angiosperms353 panel can be used to resolve recent radiations and have the potential for use in population genetic analysis even in the case of polyploids (Kates *et al.*, 2018; van Andel *et al.*, 2019; Larridon *et al.*, 2020; Melichárková *et al.*, 2020; Šlenker *et al.*, 2021; Slimp *et al.*, 2021). However, although general bait sets give excellent overlap between studies, providing useful data matrices, they can capture less efficiently than specific baits (Kadlec *et al.*, 2017; Liu *et al.*, 2019b; Larridon *et al.*, 2020).

Given the numbers of *Begonia* species and the poor preservation of DNA likely in many herbarium samples, it is important to focus on obtaining the highest proportion of usable data possible in each sequencing run. This is best achieved by using a specific bait set. Such a set also has the advantage of allowing focus on particular aspects of *Begonia* biology and inclusion of genomic regions known to vary between species (as recommended in Lee *et al.*, 2021). The high degree of resolution achieved in the present study for recent *Coelocentrum* radiation (see **Figure 5**) shows the utility of the *Begonia* bait set in this respect, but it also emphasises the need for careful consideration of analysis pipelines.

Integration with previous *Begonia* phylogeny studies can be achieved through use of the off-target reads to assemble sequences for plastid and mitochondrial genomes, and for nuclear repeats. This may require 'spiking' of the sequencing reactions with uncaptured libraries as a high capture efficiency, as observed in some cases in the present study (see **Figure 2A**), reduces the off-target reads to levels to a level too low for plastid genome assembly (Weitemier *et al.*, 2014).

Only 23 of the loci in the *Begonia* bait set are also present in the Angiosperms353 enrichment panel. Better integration with other studies could be achieved through combining our bait set with the Angiosperms353 set in the hybridisation step. This approach has worked well in Brassicaceae (Hendriks *et al.*, 2021).

### *Library preparation, hybridisation and sequencing*

The results of the four capture experiments carried out using our bait set show effective capture under a range of conditions. Forrest *et al.* (2019) showed no clear differences between the NEBNext Seq and TruSeq library preparation kits. Pooling is recommended to reduce costs. Here we report pooling up to 19, but up to 48 should be possible (Hale *et al.*, 2020). We suggest that the bait set is fairly robust to small changes in the hybridisation protocol, and that further work to optimise the protocol may be required for the most difficult samples. Unfortunately, the most difficult samples are usually those for which the least material is available, thus limiting the possibilities of multiple hybridisation attempts. Arbor Biosciences suggests 65°C and 16–24 h as the standard hybridisation conditions, with lower temperatures (55°C) for very fragmented samples and up to 40 h of incubation for samples with very low ratios of target to off-target DNA (Arbor Biosciences, 2018). Based on results presented here, we recommend 19 h at 60°C with 10–12 rounds of post-capture PCR as a good starting point.

Both 250 bp and 150 bp reads have been generated in the studies examined here. Because the reads are mapped to a known reference, there is little advantage to the longer length of read unless off-target sequences (either introns or organellar/ITS data) are also required. Given the depth of sequencing observed for the on-target reads, we suggest that the cheapest approach be used regardless of the read length generated. The *Begonia* bait set is 1.9 MB, longer than some others, such as the Angiosperms353 set, so more sequencing is required to obtain a similar sequence depth to that obtained with shorter bait sets. With 60% capture efficiency, 180 samples could be sequenced in one lane of MiSeq to 20 × mean coverage; this is good for species-level phylogenetics, although given the variation seen across loci and samples, half this number (90 samples per lane) might be recommended.

### *Analysis pipelines*

Of the five pipelines we trailed for calling consensus sequences from captured reads, the BASIC pipeline provided the least data but also the least distance between technical replicates, with all the technical replicates resolving as sisters in the concatenated tree (see **Table 2**, **Figure 5**, **6A**). However, the bipartition analysis shows that even these nodes at technical replicates were not supported by all the gene trees (see **Figure 7**). This confirms that all the pipelines tested produce errors that contribute to long branches and poor node support, despite the high level of support for each node in the ML analysis. It is possible that some of the discordance in our test set derives from hybridisation and incomplete lineage sorting in *Symbegonia*. The PALEOMIX results, particularly, show quite strong support for an alternative branching pattern in several of the deeper nodes in the tree, which could be related to hybridisation early in the colonisation of New Guinea, as suggested by morphology (Wilson, 2021).

We would recommend using at least two methods for deriving consensus bait sequences to allow comparison of results. We suggest one using reference-based consensus calling (such as BASIC or PALEOMIX) and one using a *de novo* assembly step (such as HybPiper, HybPhyloMaker or SECAPR). Analysis of the gene tree support in each approach could then be used to exclude ‘noisy’ loci.

### *Phylogenetic approaches using hybrid capture data*

One standard phylogenetic approach with hybrid capture data is to concatenate baits and generate a species tree using ML analysis. A second approach infers individual gene trees from each target (also under ML) and produces a species tree from the gene trees under the multispecies coalescent (MSC) using a program such as ASTRAL. Our bipartition analysis on the MSC phylogeny revealed very high variation between gene trees (see **Figure 7**). The choice of which loci to use for analysis and which to reject will clearly make large differences to support for particular nodes and the shape of the tree, even with the concatenated approach, because a few loci can have disproportionate effects on topology (Shen *et al.*, 2017) (see **Supplementary figure**).

There is clearly variation between captures in which baits capture well and those that capture poorly (see **Figure 2A**). Eight baits with some repetitive sequences gave very high capture in one experiment (POP, see **Figure 2B**), but behaved normally in the other three experiments (see **Figure 2A**). There is also extensive variation between baits in the number of parsimony-informative sites obtained (see **Figure 2C**). Patterns in bait capture need to be considered for each experiment, and baits with exceptionally high, low or variable coverage excluded from further analysis. Analysis of patterns of gene trees is key to reducing noise. PhyParts can be used to reveal the variation between gene trees and species trees, and this can be further explored using other software such as treespace (https://thibautjombart.github.io/treespace/) (see **Figure 7**, **Supplementary figure**).

The results of the present study show that some targets have a consistently good match to the species tree across pipelines (see **Figure 8A**). We have used clustering analysis to show that the targets can be divided into two groups, one of which is close to the species tree, and the other (which appears noisier-technical replicates are more distant) which is more distant to the species tree. It is possible that one set of targets is reflecting the 'true tree' and the other is capturing more paralogues and thus generating distorting noise. However, it is also possible that a single 'true tree' does not exist.

Striving to derive a single species tree from the mass of data in hundreds of loci is possibly not the best use of these data. The power of a hybrid capture approach lies in the ability to resolve complex evolutionary histories, and we suggest the use of approaches that include incongruence among gene trees in the analysis. In particular, given the prevalence of hybridisation and gene duplication in the evolution of *Begonia* (Brennan *et al.*, 2012; Hughes *et al.*, 2018; Liu *et al.*, 2019a; Tseng *et al.*, 2019), it is only by acknowledging the complex and reticulate nature of the phylogenies that we will begin to understand the evolutionary patterns in this genus. Such approaches are becoming more widespread (Morales-Briones *et al.*, 2018; Harris, 2019; Gagnon *et al.*, 2021; Lee *et al.*, 2021). The published pipelines accompanying these papers make this type of analysis more accessible, although the computing resources required can still be considerable for large numbers of trees and larger bait sets.

### *Further work*

Our choice of loci includes a set of key developmental, physiological and stress-related genes. The depth of coverage obtained in the captures from good-quality material (including many herbarium samples) is sufficient to allow sequence analysis enabling comparison of evolutionary patterns in functional genes. We hope that the wealth of data produced from capture experiments will not be limited to phylogenetic studies but will provide a greater knowledge of functional evolutionary patterns across the group.

## Conclusions

We recommend this bait set for *Begonia* phylogenetics and population-level studies. Arbor Biosciences can produce a copy of this bait set within 2 weeks. It is nearly as quick to obtain as an off-the-shelf general kit such as the Angiosperms353 set, but it has the advantages of better capture, more loci and loci chosen to study issues of *Begonia* biology.

General advice on a cost-effective approach to hybrid capture protocol has been published by Hale *et al.* (2020), making hybrid capture possible for budget-constrained laboratories and studies using hundreds of samples. We hope that by using a coordinated approach, the usefulness of the data will be increased and novel comparisons and overviews will be possible.

## Acknowledgements

The authors thank Michelle Hart and Laura Forrest (Royal Botanic Garden Edinburgh, RBGE) for laboratory help and general support, the staff of the Experimental Greenhouse and NGS High Throughput Genomic Core of BRCAS, Mark Hughes (RBGE) for sharing his *Begonia* expertise, the M. L. MacIntyre Begonia Trust for funding for H.W., PlantID (EU Horizon 2020 DTP 765000) for funding for T.M. and general support and advice, and CropDiversity servers and administrators for computing support (BBSRC grant BBSRC BB/S019669/1). Yu-Hsing Tseng acknowledges a postdoctoral grant (grant no. 107-2917-I-564-026) from the Minister of Science and Technology, Taiwan, for a visit to RBGE. Yu-Hsing Tseng and Kuo-Fang Chung were supported by research grants from Academia Sinica, Taiwan.

## ORCID iDs

T. Michel https://orcid.org/0000-0001-5717-5643
Y.-H. Tseng https://orcid.org/0000-0002-8166-5690
H. P. Wilson https://orcid.org/0000-0002-5231-3987
K.-F. Chung https://orcid.org/0000-0003-3628-2567
C. A. Kidner https://orcid.org/0000-0001-6426-3000

## Supplementary material

Supplementary material for this article is available from the *Edinburgh Journal of Botany* online portal.

**Supplementary data.** FASTA file of target sequences.

**Supplementary table.** Details of the bait set. Name, length, reason for choice, orthologues, annotation and sequence for the bait set of 1239 targets. Annotations produced using Blast2GO (Conesa and Götz, 2008).

**Supplementary figure.** ASTRAL trees with PhyParts analysis for each group of gene trees from the PALEOMIX pipeline. A, Group 1 (465 genes); B, group 2 (778 genes). Numbers at each node show the supporting gene tree over the conflicting gene trees. Supporting gene trees/conflicting gene trees shown as pie charts in which blue shows the proportion supporting the topology; yellow, the proportion supporting the next most common bipartition; red, all other conflicting gene trees; and grey, the proportion with no support for a conflicting bipartition.

## References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. Journal of Molecular Biology. 215(3):403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Andermann T, Cano Á, Zizka A, Bacon C, Antonelli A. 2018. SECAPR – a bioinformatics pipeline for the rapid and user-friendly processing of targeted enriched Illumina sequences, from raw reads to alignments. PeerJ. 6:e5175. https://doi.org/10.7717/peerj.5175

Arbor Biosciences. 2018. myBaits: Hybridization Capture for Targeted NGS. Manual, version 4.01. Ann Arbor, Michigan: Arbor Biosciences. https://arborbiosci.com/wp-content/uploads/2019/08/myBaits-Manual-v4.pdf

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology. 19(5):455–477. https://doi.org/10.1089/cmb.2012.0021

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 30(15):2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Borowiec ML. 2016. AMAS: a fast tool for alignment manipulation and computing of summary statistics. PeerJ. 4:e1660. https://doi.org/10.7717/peerj.1660

Brennan AC, Bridgett S, Shaukat Ali M, Harrison N, Matthews A, Pellicer J, Twyford AD, Kidner CA. 2012. Genomic resources for evolutionary studies in the large, diverse, tropical genus, *Begonia*. Tropical Plant Biology. 5:261–276. https://doi.org/10.1007/s12042-012-9109-6

Campos-Dominguez L. 2020. Does a dynamic genome drive speciation in a mega-diverse genus? Ph.D. thesis, University of Edinburgh.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 25(15):1972–1973. https://doi.org/10.1093/bioinformatics/btp348

Carlsen MM, Fér T, Schmickl R, Leong-Škorničková J, Newman M, Kress WJ. 2018. Resolving the rapid plant radiation of early diverging lineages in the tropical Zingiberales: pushing the limits of genomic data. Molecular and Phylogenetic Evolution. 128:55–68. https://doi.org/10.1016/j.ympev.2018.07.020

Conesa A, Götz S. 2008. Blast2GO: a comprehensive suite for functional analysis in plant genomics. International Journal of Plant Genomics. 2008:619832. https://doi.org/10.1155/2008/619832

Constantinides B, Robertson DL. 2017. Kindel: indel-aware consensus for nucleotide sequence alignments. Journal of Open Source Software. 2(15):282. https://doi.org/10.21105/joss.00282

Couvreur TLP, Helmstetter AJ, Koenen EJM, Bethune K, Brandão RD, Little SA, Sauquet H, Erkens RHJ. 2018. Phylogenomics of the major tropical plant family Annonaceae using targeted enrichment of nuclear genes. Frontiers in Plant Science. 9:1941. https://doi.org/10.3389/fpls.2018.01941

Cronn R, Knaus BJ, Liston A, Maughan PJ, Parks M, Syring JV, Udall J. 2012. Targeted enrichment strategies for next-generation plant biology. American Journal of Botany. 99(2):291–311. https://doi.org/10.3732/ajb.1100356

Dodsworth S, Pokorny L, Johnson MG, Kim JT, Maurin O, Wickett NJ, Forest F, Baker WJ. 2019.

Hyb-Seq for flowering plant systematics. Trends in Plant Science. 24(10):887–891. https://doi.org/10.1016/j.tplants.2019.07.011

Emelianova K. 2017. Using next generation sequencing to investigate the generation of diversity in the genus Begonia. Ph.D. thesis, University of Edinburgh.

Emelianova K, Martínez Martínez A, Campos-Dominguez L, Kidner C. 2021. Multi-tissue transcriptome analysis of two *Begonia* species reveals dynamic patterns of evolution in the chalcone synthase gene family. Scientific Reports. 11(1):17773. https://doi.org/10.1038/s41598-021-96854-y

Fér T, Schmickl RE. 2018. HybPhyloMaker: target enrichment data analysis from raw reads to species trees. Evolutionary Bioinformatics Online. 14:1176934317742613. https://doi.org/10.1177/1176934317742613

Folk RA, Mandel JR, Freudenstein JV. 2015. A protocol for targeted enrichment of intron-containing sequence markers for recent radiations: a phylogenomic example from *Heuchera* (Saxifragaceae). Applications in Plant Sciences. 3(8):1500039. https://doi.org/10.3732/apps.1500039

Forrest LL, Hart ML, Hughes M, Wilson HP, Chung KF, Tseng YH, Kidner CA. 2019. The limits of Hyb-Seq for herbarium specimens: impact of preservation techniques. Frontiers in Ecology and Evolution. https://doi.org/10.3389/fevo.2019.00439

Frantz LAF, Mullin VE, Pionnier-Capitan M, Lebrasseur O, Ollivier M, Perri A, Linderholm A, Mattiangeli V, Teasdale MD, Dimopoulos EA, *et al.* 2016. Genomic and archaeological evidence suggest a dual origin of domestic dogs. Science. 352(6290):1228–1231. https://doi.org/10.1126/science.aaf3161

Frodin DG. 2004. History and concepts of big plant genera. Taxon. 53(3):753–776. https://doi.org/10.2307/4135449

Gagnon E, Hilgenhof R, Orejuela A, McDonnell A, Sablok G, Aubriot X, Giacomin L, Gouvêa Y, Bragionis T, Stehmann JR, Bohs L, Dodsworth S, Martine C, Poczai P, Knapp S, Särkinen T. 2022. Phylogenomic discordance suggests polytomies along the backbone of the large genus *Solanum*. American Journal of Botany. 109(4):580–601. https://doi.org/10.1002/ajb2.1827

Gardiner LJ, Brabbs T, Akhunov A, Jordan K, Budak H, Richmond T, Singh S, Catchpole L, Akhunov E, Hall A. 2019. Integrating genomic resources to present full gene and putative promoter capture probe sets for bread wheat. Gigascience. 4(8):giz018. https://doi.org/10.1093/gigascience/giz018

Goodall-Copestake WP, Pérez-Espona S, Harris DJ, Hollingsworth PM. 2010. The early evolution of the mega-diverse genus *Begonia* (Begoniaceae) inferred from organelle DNA phylogenies. Biologial Journal of the Linnean Society, London. 101(2):243–250. https://doi.org/10.1111/j.1095-8312.2010.01489.x

Hale H, Gardner EM, Viruel J, Pokorny L, Johnson MG. 2020. Strategies for reducing per-sample costs in target capture sequencing for phylogenomics and population genomics in plants. Applications in Plant Science. 8(4):e11337. https://doi.org/10.1002/aps3.11337

Harris K. 2019. From a database of genomes to a forest of evolutionary trees. Nature Genetics. 51:1306–1307. https://doi.org/10.1038/s41588-019-0492-x

Hart ML, Forrest LL, Nicholls JA, Kidner CA. 2016. Retrieval of hundreds of nuclear loci from herbarium specimens. Taxon. 65:1081–1092. https://doi.org/10.12705/655.9

Helmstetter AJ, Kamga SM, Bethune K, Lautenschläger T, Zizka A, Bacon CD, Wieringa JJ, Stauffer

F, Antonelli A, Sonké B, Couvreur TLP. 2020. Unraveling the phylogenomic relationships of the most diverse African palm genus *Raphia* (Calamoideae, Arecaceae). Plants. 9(4):549. https://doi.org/10.3390/plants9040549

Hendriks KP, Mandáková T, Hay NM, Ly E, Hooft van Huysduynen A, Tamrakar R, Thomas SK, Toro-Núñez O, Pires JC, Nikolov LA, Koch MA, Windham MD, Lysak MA, Forest F, Mummenhoff K, Baker WJ, Lens F, Bailey CD. 2021. The best of both worlds: combining lineage-specific and universal bait sets in target-enrichment hybridization reactions. Applications in Plant Science. 9(7). https://doi.org/10.1002/aps3.11438

Hill CB, Wong D, Tibbits J, Forrest K, Hayden M, Zhang XQ, Westcott S, Angessa TT, Li C. 2019. Targeted enrichment by solution-based hybrid capture to identify genetic sequence variants in barley. Scientific Data. 6:12. https://doi.org/10.1038/s41597-019-0011-z

Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. Molecular Biology and Evolution. 33(6):1635–1638. https://doi.org/10.1093/molbev/msw046

Hughes M, Peng CI, Lin CW, Rubite RR, Blanc P, Chung KF. 2018. Chloroplast and nuclear DNA exchanges among *Begonia* sect. *Baryandra* species (Begoniaceae) from Palawan Island, Philippines, and descriptions of five new species. PLoS One. 13:e0194877. https://doi.org/10.1371/journal.pone.0194877

Johnson MG, Malley C, Goffinet B, Shaw AJ, Wickett NJ. 2016. A phylotranscriptomic analysis of gene family expansion and evolution in the largest order of pleurocarpous mosses (Hypnales, Bryophyta). Molecular Phylogenetics and Evolution. 98:29–40. https://doi.org/10.1016/j.ympev.2016.01.008

Johnson MG, Pokorny L, Dodsworth S, Botigué LR, Cowan RS, Devault A, Eiserhardt WL, Epitawalage N, Forest F, Kim JT, Leebens-Mack JH, Leitch IJ, Maurin O, Soltis DE, Soltis PS, Wong GKS, Baker WJ, Wickett NJ. 2019. A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-Medoids clustering. Systematic Biology. 68(4):594–606. https://doi.org/10.1093/sysbio/syy086

Jombart T, Kendall M, Almagro-Garcia J, Colijn C. 2017. treespace: statistical exploration of landscapes of phylogenetic trees. Molecular Ecology Resources. 17(6):1385–1392. https://doi.org/10.1111/1755-0998.12676

Jones KE, Fér T, Schmickl RE, Dikow RB, Funk VA, Herrando-Moraira S, Johnston PR, Kilian N, Siniscalchi CM, Susanna A, Slovák M, Thapa R, Watson LE, Mandel JR. 2019. An empirical assessment of a single family-wide hybrid capture locus set at multiple evolutionary timescales in Asteraceae. Applications in Plant Science. 7:e11295. https://doi.org/10.1002/aps3.11295

Jónsson H, Ginolhac A, Schubert M, Johnson PLF, Orlando L. 2013. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics. 29(13):1682–1684. https://doi.org/10.1093/bioinformatics/btt193

Kadlec M, Bellstedt DU, Le Maitre NC, Pirie MD. 2017. Targeted NGS for species level phylogenomics: "made to measure" or "one size fits all"? PeerJ. 5:e3569. https://doi.org/10.7717/peerj.3569

Kates HR, Johnson MG, Gardner EM, Zerega NJC, Wickett NJ. 2018. Allele phasing has minimal impact on phylogenetic reconstruction from targeted nuclear gene sequences in a case study of *Artocarpus*. American Journal of Botany. 105(3):404–416. https://doi.org/10.1002/ajb2.1068

Katoh K, Toh H. 2008. Recent developments in the MAFFT multiple sequence alignment program. Briefings in Bioinformatics. 9(4):286–298. https://doi.org/10.1093/bib/bbn013

Kendall M, Colijn C. 2016. Mapping phylogenetic trees to reveal distinct patterns of evolution. Molecular Biology and Evolution. 33(10):2735–2743. https://doi.org/10.1093/molbev/msw124

Koenen EJM, Ojeda DI, Steeves R, Migliore J, Bakker FT, Wieringa JJ, Kidner C, Hardy OJ, Pennington RT, Bruneau A, Hughes CE. 2020. Large-scale genomic sequence data resolve the deepest divergences in the legume phylogeny and support a near-simultaneous evolutionary origin of all six subfamilies. New Phytologist. 225(3):1355–1369. https://doi.org/10.1111/nph.16290

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods. 9:357–359. https://doi.org/10.1038/nmeth.1923

Larridon I, Villaverde T, Zuntini AR, Pokorny L, Brewer GE, Epitawalage N, Fairlie I, Hahn M, Kim J, Maguilla E, Maurin O, Xanthos M, Hipp AL, Forest F, Baker WJ. 2020. Tackling rapid radiations with targeted sequencing. Frontiers in Plant Science. 10:1655. https://doi.org/10.3389/fpls.2019.01655

Lee AK, Gilman IS, Srivastav M, Lerner AD, Donoghue MJ, Clement WL. 2021. Reconstructing Dipsacales phylogeny using Angiosperms353: issues and insights. American Journal of Botany. 108(7):1122–1142. https://doi.org/10.1002/ajb2.1695

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 25:1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25(16):2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Liu SH, Tseng YH, Zure D, Rubite RR, Balangcod TD, Peng CI, Chung KF. 2019a. *Begonia balangcodiae* sp. nov. from northern Luzon, the Philippines and its natural hybrid with *B. crispipila*, *B. × kapangan* nothosp. nov. Phytotaxa. 407(1):5–21. https://doi.org/10.11646/phytotaxa.407.1.3

Liu Y, Johnson MG, Cox CJ, Medina R, Devos N, Vanderpoorten A, Hedenäs L, Bell NE, Shevock JR, Aguero B, Quandt D, Wickett NJ, Shaw AJ, Goffinet B. 2019b. Resolution of the ordinal phylogeny of mosses using targeted exons from organellar and nuclear genomes. Nature Communications. 10:1485.

Mandel JR, Dikow RB, Funk VA. 2015. Using phylogenomics to resolve mega-families: an example from Compositae. Journal of Systematics and Evolution. 53(5):391–402. https://doi.org/10.1111/jse.12167

McKain MR, Johnson MG, Uribe-Convers S, Eaton D, Yang Y. 2018. Practical considerations for plant phylogenomics. Applications in Plant Sciences. 6(3):e1038. https://doi.org/10.1002/aps3.1038

Melichárková A, Šlenker M, Zozomová-Lihová J, Skokanová K, Šingliarová B, Kačmárová T, Caboňová M, Kempa M, Šrámková G, Mandáková T, Lysák MA, Svitok M, Mártonfiová L, Marhold K. 2020. So closely related and yet so different: strong contrasts between the evolutionary histories of species of the *Cardamine pratensis* polyploid complex in Central Europe. Frontiers in Plant Sciences. 11:588856. https://doi.org/10.3389/fpls.2020.588856

Moonlight PW, Richardson JE, Tebbitt MC, Thomas DC, Hollands R, Peng CI, Hughes M. 2015. Continental-scale diversification patterns in a megadiverse genus: the biogeography of Neotropical *Begonia*. Journal of Biogeography. 42(6):1137–1149. https://doi.org/10.1111/jbi.12496

Moonlight PW, Ardi WH, Padilla LA, Chung K-F, Fuller D, Girmansyah D, Hollands R, Jara-Muñoz A, Kiew R, Leong W-C, Liu Y, Mahardika A, Marasinghe LDK, O'Connor M, Peng C-I, Pérez ÁJ, Phutthai T, Pullan M, Rajbhandary S, Reynel C, Rubite RR, Sang J, Scherberich D, Shui Y-M, Tebbitt MC, Thomas DC, Wilson HP, Zaini NH, Hughes M. 2018. Dividing and conquering the fastest-growing genus: towards a natural sectional classification of the mega-diverse genus *Begonia* (Begoniaceae). Taxon 67(2):267–323. https://doi.org/10.12705/672.3

Morales-Briones DF, Liston A, Tank DC. 2018. Phylogenomic analyses reveal a deep history of hybridization and polyploidy in the Neotropical genus *Lachemilla* (Rosaceae). New Phytologist. 218(4):1668–1684. https://doi.org/10.1111/nph.15099

Nicholls JA, Pennington RT, Koenen EJM, Hughes CE, Hearn J, Bunnefeld L, Dexter KG, Stone GN, Kidner CA. 2015. Using targeted enrichment of nuclear genes to increase phylogenetic resolution in the neotropical rain forest genus *Inga* (Leguminosae: Mimosoideae). Frontiers in Plant Science. 6:710. https://doi.org/10.3389/fpls.2015.00710

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular Biology and Evolution. 32(1):268–274. https://doi.org/10.1093/molbev/msu300

Pezzini FF. 2019. Phylogeny, taxonomy and biogeography of *Ceiba* Mill. (Malvaceae: Bombacoideae). Ph.D. thesis, University of Edinburgh.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – approximately maximum-likelihood trees for large alignments. PLoS ONE. 5(3):e9490. https://doi.org/10.1371/journal.pone.0009490

Schneider JV, Jungcurt T, Cardoso D, Amorim AM, Töpel M, Andermann T, Poncy O, Berberich T, Zizka G. 2021. Phylogenomics of the tropical plant family Ochnaceae using targeted enrichment of nuclear genes and 250+ taxa. Taxon. 70(1):48–71. https://doi.org/10.1002/tax.12421

Schubert M, Ermini L, Der Sarkissian C, Jónsson H, Ginolhac A, Schaefer R, Martin MD, Fernández R, Kircher M, McCue M, Willerslev E, Orlando L. 2014. Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. Nature Protocols. 9:1056–1082. https://doi.org/10.1038/nprot.2014.063

Schubert M, Mashkour M, Gaunitz C, Fages A, Seguin-Orlando A, Sheikhi S, Alfarhan AH, Alquraishi SA, Al-Rasheid KAS, Chuang R, Ermini L, Gamba C, Weinstock J, Vedat O, Orlando L. 2017. Zonkey: a simple, accurate and sensitive pipeline to genetically identify equine F1-hybrids in archaeological assemblages. Journal of Archaeological Science. 78:147–157. https://doi.org/10.1016/j.jas.2016.12.005

Shaukat Ali M. 2013. Genetic architecture of species level differences in *Begonia* section *Gireoudia*. Ph.D. thesis, University of Edinburgh.

Shee ZQ, Frodin DG, Cámara-Leret R, Pokorny L. 2020. Reconstructing the complex evolutionary history of the papuasian *Schefflera* radiation through herbariomics. Frontiers in Plant Science. 11:258. https://doi.org/10.3389/fpls.2020.00258

Shen XX, Hittinger CT, Rokas A. 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. Nature Ecology and Evolution. 1:126. https://doi.org/10.1038/s41559-017-0126

Slater GSC, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 6:31. https://doi.org/10.1186/1471-2105-6-31

Šlenker M, Kantor A, Marhold K, Schmickl R, Mandáková T, Lysak MA, Perný M, Caboňová M, Slovák M, Zozomová-Lihová J. 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*. Fronteris in Plant Science. 12:659275. https://doi.org/10.3389/fpls.2021.659275

Slimp M, Williams LD, Hale H, Johnson MG. 2021. On the potential of Angiosperms353 for population genomic studies. Applications in Plant Science. https://doi.org/10.1002/aps3.11419

Smith SA, Moore MJ, Brown JW, Yang Y. 2015. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. BMC Evolutionary Biology. 15:150. https://doi.org/10.1186/s12862-015-0423-0

Soto Gomez M, Pokorny L, Kantar MB, Forest F, Leitch IJ, Gravendeel B, Wilkin P, Graham SW, Viruel J. 2019. A customized nuclear target enrichment approach for developing a phylogenomic baseline for *Dioscorea* yams (Dioscoreaceae). Applications in Plant Science. 7(6):e11254. https://doi.org/10.1002/aps3.11254

Thomas DC, Hughes M, Phutthai T, Ardi WH, Rajbhandary S, Rubite R, Twyford AD, Richardson JE. 2012. West to east dispersal and subsequent rapid diversification of the mega-diverse genus *Begonia* (Begoniaceae) in the Malesian archipelago. Journal of Biogeography. 39(1):98–113. https://doi.org/10.1111/j.1365-2699.2011.02596.x

Thomas AE, Igea J, Meudt HM, Albach DC, Lee WG, Tanentzap AJ. 2021. Using target sequence capture to improve the phylogenetic resolution of a rapid radiation in New Zealand *Veronica*. American Journal of Botany. 108(7):1289–1306. https://doi.org/10.1002/ajb2.1678

Tomasello S, Karbstein K, Hodač L, Paetzold C, Hörandl E. 2020. Phylogenomics unravels Quaternary vicariance and allopatric speciation patterns in temperate-montane plant species: a case study on the *Ranunculus auricomus* species complex. Molecular Ecology. 29:2031–2049. https://doi.org/10.1111/mec.15458

Tseng YH, Huang HY, Xu WB, Yang HA, Liu Y, Peng CI, Chung KF. 2017. Development and characterization of EST-SSR markers for *Begonia luzhaiensis* (Begoniaceae). Applications in Plant Science. 5(5):1700024. https://doi.org/10.3732/apps.1700024

Tseng YH, Huang HY, Xu WB, Yang HA, Peng CI, Liu Y, Chung KF. 2019. Phylogeography of *Begonia luzhaiensis* suggests both natural and anthropogenic causes for the marked population genetic structure. Botanical Studies. 60(1):20. https://doi.org/10.1186/s40529-019-0267-9

Twyford AD, Ennos RA, White CD, Ali MS, Kidner CA. 2014. The evolution of sex ratio differences and inflorescence architectures in *Begonia* (Begoniaceae). American Journal of Botany. 101(2):308–317. https://doi.org/10.3732/ajb.1300090

Vallebueno-Estrada M, Rodríguez-Arévalo I, Rougon-Cardoso A, Martínez González J, García Cook A, Montiel R, Vielle-Calzada JP. 2016. The earliest maize from San Marcos Tehuacán is a partial domesticate with genomic evidence of inbreeding. Proceedings of the National Academy of Sciences of the United States of America. 113(49):14151–14156. https://doi.org/10.1073/pnas.1609701113

Van Andel T, Veltman MA, Bertin A, Maat H, Polime T, Hille Ris Lambers D, Tjoe Awie J, De Boer H, Manzanilla V. Hidden rice diversity in the Guianas. Frontiers in Plant Science. 2019. 10:1161. https://doi.org/10.3389/fpls.2019.01161

Villaverde T, Pokorny L, Olsson S, Rincón-Barrado M, Johnson MG, Gardner EM, Wickett NJ, Molero J, Riina R, Sanmartín I. 2018. Bridging the micro- and macroevolutionary levels in phylogenomics: Hyb-Seq solves relationships from populations to species and above. New Phytol. 220(2):636–650. https://doi.org/10.1111/nph.15312

Weitemier K, Straub SCK, Cronn RC, Fishbein M, Schmickl R, McDonnell A, Liston A. 2014. Hyb-Seq: combining target enrichment and genome skimming for plant phylogenomics. Applications in Plant Science. 2(9):1400042. https://doi.org/10.3732/apps.1400042

Wilson H. 2021. Megadiversity and the New Guinea orogeny. Ph.D. thesis, University of Glasgow.

Yang L, Koo DH, Li Y, Zhang X, Luan F, Havey MJ, Jiang J, Weng Y. 2012. Chromosome rearrangements during domestication of cucumber as revealed by high-density genetic mapping and draft genome assembly. The Plant Journal. 71(6):895–906. https://doi.org/10.1111/j.1365-313X.2012.05017.x

Zhang C, Rabiee M, Sayyari E, Mirarab S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics. 19:153. https://doi.org/10.1186/s12859-018-2129-y