INTRODUCTION

With current estimates of more than 25,000 species, the Orchidaceae account for about 8% of all plant species (Givnish et al., 2015). The family is well known for its large number of epiphytes (>18,000 species) (Dressler, 1981; Gravendeel et al., 2004), extraordinary floral forms, unusual pollination strategies (Bohman et al., 2016; Cozzolino & Widmer, 2005; Schiestl & Schlüter, 2009; Tremblay et al., 2005), and a dependence on mycorrhizal fungi for germination (Phillips et al., 2020). As a consequence of this extreme diversity, the drivers of diversity in the Orchidaceae remain to be fully understood. Here, we outline a multitiered sequence capture strategy aimed at capturing hundreds of loci to enable phylogenetic resolution from subtribe to subspecific levels in orchids of the tribe Diurideae. For the probe design, we mined subsets of 18 transcriptomes, to give five target sequence sets aimed at the tribe (Sets 1 & 2), subtribe (Set 3), and within subtribe levels (Sets 4 & 5). Analysis included alternative de novo and reference-guided assembly, before target sequence extraction, annotation and alignment, and application of a homology-aware k-mer block phylogenomic approach, prior to maximum likelihood and coalescence-based phylogenetic inference. Our evaluation considered 87 taxa in two test data sets: 67 samples spanning the tribe, and 72 samples involving 24 closely related Caladenia species. The tiered design achieved high target loci recovery (>89%), with the median number of recovered loci in Sets 1–5 as follows: 212, 219, 816, 1024, and 1009, respectively. Interestingly, as a first test of the homologous k-mer approach for targeted sequence capture data, our study revealed its potential for enabling robust phylogenetic species tree inferences. Specifically, we found matching, and in one case improved phylogenetic resolution within species complexes, compared to conventional phylogenetic analysis involving target gene extraction. Our findings indicate that a customized multitiered sequence capture strategy, in combination with promising yet underutilized phylogenomic approaches, will be effective for groups where interspecific divergence is recent, but information on deeper phylogenetic relationships is also required.

KEYWORDS
bioinformatics/phyloinformatics, Orchidaceae, phylogenetic theory and methods, phylogeography, transcriptomics
Recent orchid wide plastid-based phylogenomic studies indicate that the evolution of pollinia, the epiphytic habitat, CAM photosynthesis in epiphytes, and tropical distributions are major contributors to the diversification of the family (Givnish et al., 2015, 2016). However, these factors are unlikely to fully explain local patterns of diversification. For example, neither epiphytism, nor a tropical distribution, can explain the rich terrestrial orchid flora of South Africa (>500 spp.) or temperate southern Australia (>1500 spp.) (Backhouse et al., 2019; Johnson et al., 2015). Instead, localized pollinator-driven speciation may be an important contributor to diversification in these orchid floras (Givnish et al., 2015; Van der Niet et al., 2014; Peakall et al., 2010). It follows that to more fully address the question of the drivers of species diversity across the Orchidaceae, investigations spanning different taxonomic and geographic scales will be required.

The first nuclear orchid genome, that of *Phalaenopsis equestris* (Epidendroideae), was published ahead of those in many plant families (Cai et al., 2015). Subsequently, genomes have been published for other members of the Epidendroideae: *Phalaenopsis aphrodite* (Chao et al., 2018), *Dendrobium catenatum* (Zhang, Xu, et al., 2016), and *Gastrodia elata* (Yuan et al., 2018). Draft genomes of *Apostasia shenzhenica* (Zhang et al., 2017), representing the Apostasioideae, and of the commercially important *Vanilla planifolia* in the sub-family Vanillaeae (Hu et al., 2019), have also been published (Figure 1). Genomes are presently unavailable for any representative of the remaining two orchid subfamilies, the Cypripedioideae and Orchidoideae. Beyond these orchid genomes, there is a growing number of published orchid transcriptomes (Figure 1). These span the family more widely, and include representatives of multiple genera from the Epidendroideae (e.g., Chao et al., 2017; Li et al., 2013; Xu et al., 2018; Zhang, He, et al., 2016); Orchidoideae (e.g., Piñeiro Fernández et al., 2019; Sedeek et al., 2013; Wong, Amarasinhe, et al., 2017; Xu et al., 2017); Vanillaeae (e.g., Rao et al., 2014; Zhang et al., 2017); Cypripedioideae (e.g., Guo et al., 2018; Unruh et al., 2018) and Apostasioideae (Chao et al., 2017; Deng et al., 2015) (Figure 1).

Recent initiatives such as the 1000 Plant Genomes Project (1KP) have generated a wealth of transcriptome data spanning >1000 plants, aiding both our understanding of the evolution of green plants (Leebens-Mack et al., 2019), and providing relevant sequences for the construction of a universal set of target enrichment probes for angiosperms (Johnson et al., 2019). Existing orchid genomic and transcriptomic resources also have the potential to be mined for phylogenetic and evolutionary investigations. For example, Zhang et al. (2017) utilized DNA sequence variation across 132 single-copy gene families identified in the draft genome of *Apostasia shenzhenica* (and confirmed across 14 other plant species), to build a high-confidence phylogenetic tree of 12 representative species spanning the five subfamilies of the Orchidaceae. Drawing on whole transcriptomes of eight *Cypripedium* species, Guo et al. (2018) were able to clarify phylogenetic relationships among six closely related species. Unruh et al. (2018) further confirmed the feasibility of using transcriptomes to investigate the phylogenetic relationships among five Cypripedioideae genera and to anchor the phylogeny within the Orchidaceae. In addition, previously uncertain phylogenetic relationships among some species of *Ophrys* and *Gymnadenia* orchid have also been clarified by the phylogenomic analysis of floral transcriptomes (Piñeiro Fernández et al., 2019).

Notwithstanding these few examples of orchid phylogenetic studies that have utilized genomic and transcriptomic resources, studies involving hundreds of samples appear to be lacking. NGS-based targeted sequence capture, also called hybrid enrichment, may offer one potential solution to achieve such a goal. The potential benefits of targeted sequence capture have been advocated in several reviews (e.g., de La Harpe et al., 2019; Grover et al., 2012; Jones & Good, 2016; Lemmon et al., 2012; Lemmon & Lemmon, 2013; Léveillé-Bourret et al., 2017; McCormack et al., 2013). Some key benefits include the flexibility of target selection (e.g., type, size, and numbers of regions targeted), the potential to recover phylogenetically informative content for enriched genomics, and the ability to use the same set of probe sequences for any angiosperm species.

![Figure 1](image-url)

**FIGURE 1** The availability of genomic and transcriptomic resources mapped onto a schematic phylogeny of the five Orchidaceae subfamilies. Species diversity is based on the estimated number of species provided in Chase et al. (2015). See the main text for the supporting literature on the various orchid genomes. The transcriptome data is based on publicly available data sets deposited in NCBI Sequence Reads Archive (last accessed 01 February, 2020).
The transcriptome species with names given in full, and their subtribes, providing coding DNA sequences for the selection of the target sequence Sets 1 to 5. The species/subtribes are shown overlaid onto a phylogeny from Weston et al. (2014), used as the working hypothesis of the subtribe and generic relationships within the orchid tribe Diurideae. Transcriptomes representing a total of 18 species were used to obtain the target gene set sequences. The coloured squares indicate the gene sets relevant to that given species. Transcriptomes representing 11 core species spanning the phylogeny, and including Pterostylis curta as the outgroup, were used in the construction of sequence Sets 1 and 2, aimed at the tribe Diurideae. A subset of six of the core species was also used to construct sequence Set 3, aimed at "other subtribes with uncertain placement within the Diurideae". The construction of sequence Sets 4 and 5, aimed at the tribes Drakaeinae and Caladeniinae, respectively, took advantage of additional transcriptomes from more closely related species. See Table 1 for additional details on the taxon targets and outgroups. The above phylogram is based on the topology and branch lengths of figure 3 in Weston et al. (2014), with the omission of some minor genera to simplify the presentation. The original phylogeny was based on the nuclear locus ITS-5.8S-ITS2 and chloroplast loci trnL-trnF, matK and rbcL, in combination with a morphological, anatomical and embryological character set. This phylogeny represented the most up-to-date study of the Diurideae at the onset of the present project, and remains the study with the widest sampling across the group. Note that many relationships are uncertain, as indicated by the lack of high bootstrap support on many nodes. Note that for sequence Set 3 we use the term "other subtribes" to encompass the subtribes Acianthinae, Diuridinae, Cryptostylidinae, Thelymitrinae, Megastylidinae, Prasophyllinae, which fall outside of our core research interests on the evolution of sexual deception within the subtribes Drakaeinae (Set 4) and Caladeniinae (Set 5). Nonetheless, this set was included due to uncertainties in the placement of these subtribes in existing studies. [Colour figure can be viewed at wileyonlinelibrary.com]
loci across wide phylogenetic depth and breadth, and missing data are often minimized relative to other methods. However, to fast-track a targeted sequence capture design, one potential constraint for many studies is the need for robust sequence knowledge, typically drawn from existing genomic/transcriptome resources.

We have generated transcriptomic resources as part of our ongoing research on the biochemistry and molecular biology of Australian terrestrial orchids (Wong et al., 2018, 2019; Wong, Amarasinghe, et al., 2017; Xu et al., 2017). Using these and additional targeted transcriptomes, our overarching objective was to develop customized NGS-based sequence capture for phylogenetic studies of Australian terrestrial orchids. Here, we outline the design and implementation of a multitiered sequence capture strategy for Australian terrestrial orchids belonging to the Diurideae (with >900 species). Our goal was to simultaneously capture in a single hybridization step hundreds of informative loci that collectively enable the phylogenetic resolution of relationships spanning among subtribes, through genera and species, to phylogeographic structure within species. We also evaluate the sequencing capture success and the preliminary phylogenetic outcomes. Our evaluation considers 134 samples, comprising 87 taxa across two test data sets: (i) A set of 67 samples spanning the Diurideae, and an outgroup, to assess informativeness across the entire tribe (D67). (ii) From within the species-rich subtribe Caladeniinae, a set of 72 Caladenia samples comprising 24 species, with duplicates for each species, and for some species additional samples from geographically distinct sites (C72) (Table S1).

We further consider the sequence capture success and the congruence of the phylogenetic results obtained via two widely-adopted approaches: de novo (Jones & Good, 2016; McCormack et al., 2013) and reference-guided (Johnson et al., 2016) assembly of target/exon sequence reads followed by sequence annotation and alignment. We also tested the application of a promising, yet underutilized, phylogenomic approach based on homologous sub-strings of sequence (k-mer) (Sanderson et al., 2017). Gene and species trees were generated by maximum likelihood in IQ-TREE 2 (Minh, Schmidt, et al., 2020) with branch support estimated by ultrafast bootstrap, and topological variation measured by gene concordance and site concordance factors (Minh, Hahn, et al., 2020). Complementary multispecies coalescence-based species tree inference was performed in ASTRAL III (Zhang et al., 2018). Finally, we conclude with a brief discussion of the phylogenetic insights uncovered for the study species, and conclude with consideration of the lessons learnt for developing customized sequence capture designs.

### Table 1

| Species              | Tribe                | Subtribe<sup>a</sup> | Sets      | Tissue | Core<sup>b</sup> | Alternative nomenclature<sup>c</sup> |
|----------------------|----------------------|-----------------------|-----------|--------|-----------------|--------------------------------------|
| *Pterostylis curta*  | Cranichideae         | Pterostylidinae       | 1 & 2     | L      | Yes             |                                      |
| *Drakaea thynniphila*| Diurideae            | Drakaeinae            | 4         | F      |                 |                                      |
| *Chiloglottis seminuda* | Diurideae      | Drakaeinae            | 4         | F      |                 |                                      |
| *Chiloglottis valida* | Diurideae            | Drakaeinae            | 4         | F      |                 | Simpliglottis valida                 |
| *Chiloglottis trapeziformis* | Diurideae     | Drakaeinae            | 1, 2 & 4  | L      | Yes             | Myrmechila trapeziformis             |
| *Caleana major*      | Diurideae            | Drakaeinae            | 1, 2 & 4  | L      | Yes             |                                      |
| *Rimacola elliptica* | Diurideae            | Megastyliniinae       | 1, 2, 3 & 4 | L      | Yes             |                                      |
| *Thelymitra angustifolia* | Diurideae    | Thelymitrininae       | 1, 2 & 3  | L      | Yes             |                                      |
| *Cryptostylis leptochila* | Diurideae      | Cryptostylidinae      | 1, 2 & 3  | L      | Yes             |                                      |
| *Microtis parviflora* | Diurideae            | Prasophyllinae        | 1, 2 & 3  | L      | Yes             |                                      |
| *Corybas aconitiflorus* | Diurideae      | Acianthininae         | 1, 2 & 3  | L      | Yes             |                                      |
| *Diuris orientis*    | Diurideae            | Diuridinae            | 1, 2 & 3  | L      | Yes             |                                      |
| *Caladenia denticulata* | Diurideae       | Caladeniinae          | 1, 2 & 5  | L      | Yes             | Jonesiopsis denticulata              |
| *Caladenia crebra*   | Diurideae            | Caladeniinae          | 5         | F      | Arachnorchis crebra                                  |
| *Caladenia attingens* | Diurideae            | Caladeniinae          | 5         | F      | Arachnorchis attingens                               |
| *Caladenia actensis* | Diurideae            | Caladeniinae          | 5         | F      | Arachnorchis actensis                               |
| *Caladenia plicata*  | Diurideae            | Caladeniinae          | 5         | L      | Arachnorchis plicata                                |
| *Caladenia longicauda* | Diurideae       | Caladeniinae          | 1, 2 & 5  | L      | Yes             | Arachnorchis longicauda              |

Note: The target gene sets in which the species were used, transcriptome tissue type (L, leaf; F, floral), and where applicable, alternative nomenclature, is also shown.

<sup>a</sup>Subtribes follow Chase et al. (2015).

<sup>b</sup>Transcriptomes of the core species were used as a source of target gene sequences in two or more target gene sets.

<sup>c</sup>Alternative generic nomenclature as used in Weston et al. (2014).
2 | MATERIALS AND METHODS

2.1 | Background to the study system

Australia exhibits a rich diversity of orchids, with a current estimate of more than 1860 species (Backhouse et al., 2019). The majority of Australian terrestrial orchids belong to the Diuridieae, with conservative estimates of between 900 and 1000 species in this tribe (Chase et al., 2015; Weston et al., 2014). Within the Diuridieae, the Caladeniinae represents the largest subtribe, with estimates of the number of species varying from 297 worldwide (Chase et al., 2015), to 425 species and subspecies in Australia and Island Territories (Backhouse et al., 2019). The genus Caladenia sensu lato accounts for the largest share of the species in the subtribe (e.g., Backhouse et al. (2019) list 382 taxa). Please see Methods S1 for more details.

2.2 | Transcriptome resources and sampling

With a focus on the Australian terrestrial orchids of the Diuridieae, our goal was to obtain phylogenetic resolution spanning subtribes to within species levels. To ensure wide applicability, additional transcriptomes, beyond those already in hand (Wong, Amarasinghe, et al., 2017; Wong, Pichersky, et al., 2017; Wong et al., 2018, 2019; Xu et al., 2017), were generated. All raw sequence reads have been deposited in NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) under the BioProject and SRA study accession of PRJNA661963 and SRP281918, respectively. The choice of species to target for these extra transcriptomes was guided by the inferred, albeit often uncertain, phylogenetic relationships indicated by Weston et al. (2014). (See Figure 2 and Methods S1 for more details).

2.3 | Target sequence sets

The final set of species for which transcriptomes were available or have been generated, is shown in Table 1 and Figure 2. Table S1 provides further details of the associated sample source, location, and voucher specimen details. Based on our objectives, and transcriptome availability, we identified five different target sequence sets for use in the final probe design (Tables 1 and 2). It is important to stress that while we will often refer to each of these target sequence sets individually, in practice a single DNA probe set was used in the hybridization step to simultaneously capture the target loci across all target sequence sets. Below we briefly outline the purpose of each sequence set:

2.3.1 | Target sequence set 1 – Diurideae

Genes matching loci from Deng et al. (2015), who identified 315 putative orchid wide single locus orthologues, were the target of this set, drawing on sequences from our 11 “core” transcriptomes (Figure 2 and Table 1). Target genes were located using the Deng et al. (2015) multispecies orthologue sequences data as the BLAST query. Set 1 ultimately targeted 230 loci (Table 2). The remaining 85 orthologues failed to pass the manual alignment checking phase, with many deemed to pose a high paralogue risk, as indicated by alignments containing evidence for gene duplicates. While in this study the target group was limited to the tribe Diuridieae and outgroup, the strategic inclusion of these target genes was intended to allow data sharing among studies and allow use beyond Australian orchids.

2.3.2 | Target Sequence Set 2: Diurideae

To supplement sequence Set 1, additional target genes common across our 11 core transcriptomes spanning the Diuridieae and outgroup (Figure 2, Table 1) were identified by reciprocal BLAST, yielding a final set of 245 additional gene loci (Table 2).

2.3.3 | Target Sequence Set 3: Other Diurideae

The goal of this set was to expand the number of genes targeting the Acianthinae, Diuridinae, Cryptostylidinae, Thelymitrinae, Megastylidinae, and Prasophyllinae, for which phylogenetic relationships at the tribe level were not fully resolved by Weston et al. (2014) (Figure 2, Table 1). Reciprocal BLAST across six of the 11 core transcriptomes, each representing one of these subtribes, yielded a final set of 916 gene loci (Table 2).

2.3.4 | Target Sequence Set 4: Drakaeninae

To obtain a sequence set aimed at the Drakaeninae, reciprocal BLAST was performed across six transcriptomes representing three of the 11 core transcriptomes, plus three additional transcriptomes. Collectively, the transcriptomes covered the genera Calanea, Chiloglottis, and Drakaea as representatives of the subtribe, and Rimacola as the outgroup (Figure 2). This yielded a final set of 1,058 gene loci (Table 2).

2.3.5 | Target Sequence Set 5: Caladeniinae

To obtain a sequence set aimed at the Caladeniinae, reciprocal BLAST was performed across seven transcriptomes, representing three of the 11 core transcriptomes, plus four additional transcriptomes. Collectively, the transcriptomes covered six Caladenia species, representing one species in the subgenus Phlebochilus, five species within subgenus Calonema, and Corybas as the outgroup (Figure 2 and Table 1), yielding 1052 gene loci. Given the uncertainty about the outgroup to Caladeniinae in the phylogenetic analysis of Weston et al. (2014), the use of alternative outgroups, including...
| Set | Target tribe/subtribe | Outgroup | No T | TNo Loci | TMed Len (bp) | TTot Seq (bp) | CNo. Loci | CNo. exons | CTot Seq (bp) | CAv exons/locus | CMed exon len (bp) | %Cv |
|-----|----------------------|----------|------|----------|---------------|---------------|-----------|------------|---------------|----------------|-----------------|-----------------|
| 1   | Diurideae            | Pterostylis curta | 11   | 230      | 1320          | 287,941       | 223       | 1238       | 247,290       | 5.6            | 132             | 84%             |
| 2   | Diurideae            | Pterostylis curta | 11   | 245      | 1501          | 321,597       | 236       | 1448       | 283,064       | 6.1            | 135             | 86%             |
|     | 1–2                  |           | 475  | 1428     | 609,538       |               | 459       | 2686       | 530,354       | 5.8            | 267             | 85%             |
| 3   | Other Diurideae§     | None      | 6    | 916      | 1434          | 1,236,746     | 843       | 3505       | 927,367       | 4.2            | 161             | 67%             |
| 4   | Drakaeiniae          | Rimacola elliptica | 6    | 1058     | 1401          | 1,411,148     | 1035      | 4208       | 1,090,900     | 4.1            | 160             | 71%             |
| 5   | Caladeniinae         | Corybas aconitiflorus | 7    | 1052     | 1455          | 1,414,274     | 1011      | 5172       | 1,179,041     | 5.1            | 149             | 80%             |
|     | 3–5                  |           | 3026 | 1428     | 4,062,168     | 2889          | 12,885    | 3,197,308  | 156           | 73%            |                 |                 |

**Note:** Target statistics based on the transcriptomes include the number of target loci (TNo Loci), median length of target loci (TMed Len), and total target sequence length (TTot Seq) for each gene set. Statistics based on the sequence capture outcomes, achieved via analysis Option 1, include the number of captured loci (CNo. Loci), the number of captured exons (CNo. Exons), total captured sequence length (CTot Seq), average number of captured exons per locus (CAv Exons/Locus), median capture exon length (CMed Exon Len) and the percentage of coverage (%Cv).  

*a* Target statistics are based on the analysis of the final coding DNA sequences submitted to Roche NimbleGen for the SeqCap EZ Developer Library design of the probes.  

*b* Capture statistics are based on the Illumina sequencing outcomes following the targeted enrichment of the sequencing libraries using the custom SeqCap EZ Developer Library probes for one representative sample of each species used in the original gene set. For example, the Set 1 analysis is based on targeted enrichment of the sequencing libraries representing the same 11 core transcriptome species (see Table 1). Data are shown for the outcomes obtained via the analysis Option 1 (see Figure 3), under high stringency parameters.  

§ The "Other Diurideae" encompass the subtribes Acianthinae, Diuridinae, Cryptostylidinae, Theylmitrinae, Megastylidinae, and Prasophyllinae. The relationships among these subtribes remain somewhat uncertain within the Weston et al. (2014) phylogeny used here as the working hypothesis of possible subtribe and generic relationships within the tribe Diurideae (see also Figure 2 and Table 1).  

d See Figure 2 for the provisional phylogenetic placement of the outgroup taxa, as indicated by the phylogenetic analysis of Weston et al. (2014).  

e Percentage coverage estimated as (CTot Seq/TTot Seq) × 100. The Roche NimbleGen provided estimate of Predicted Total Probe Design Coverage was specified as Probe coverage = 68%, with Estimated probe coverage = 86%. The realised coverage achieved during the sequence capture was thus very close to the Roche NimbleGen predictions.
Diuris and Microtis was also attempted, but yielded fewer target genes (full data not shown).

2.4 | Sequence selection for target sequence sets

Our methods for obtaining the transcriptomes follow methods we have already published (Wong et al., 2018; Wong, Amarasinghe, et al., 2017). Following the assembly of paired-end reads, trimming and conversion to CDS sequences, customized scripts taking advantage of the rich resource of Bioperl tools (https://bioperl.org/, last accessed July 2020), in combination with Standalone BLAST+ command line applications (2008) (https://www.ncbi.nlm.nih.gov/books/NBK279690/, last accessed July 2020), were used to mine the assembled CDS contigs for target gene sequences by iterative high stringency BLAST (Set 1) and reciprocal BLAST (Sets 2–5) (See Methods S1 for more details and BLAST parameters). Next, a combination of semi-automated and manual checks of the CDS sequence alignments viewed within Geneious V10 (Kearse et al., 2012) was used to refine the list of gene targets by removing loci with highly complex sequences that were difficult to align, low polymorphism or potentially containing paralogues. Finally, checks for shared DNA sequences across the target genes sets were run to remove any redundancy. Please see the Dryad repository (https://doi.org/10.5061/dryad.z08kprbj) for Fasta files for each of the target gene sequence Sets 1 to 5. This repository also provides the outcomes of functional annotation and assignment of plant-specific MapMan BIN ontology of the target gene sequences, which was performed using Mercator v2.0 (Schwacke et al., 2019).

2.5 | Probe design

This phase of the project started with the submission of the target gene sequences for each of the five sequence sets to Roche NimbleGen. Subsequently, SeqCap EZ Developer Library probe design was performed by Roche NimbleGen using proprietary algorithms. These algorithms also took full advantage of the orchid genomes of Phalaenopsis equestris (Cai et al., 2015) and Dendrobium catenatum (Zhang, Xu, et al., 2016). Following our review and approval of the probe design, as the customer, probe manufacture and delivery followed.

2.6 | DNA extraction, library preparation, and Illumina sequencing

Samples were sourced from both wild and cultivated plants (Table S1), with DNA extracted using the Qiagen DNeasy Plant mini kit (Cat. No. 69106) following the manufacturer's protocol. Comprehensive details on the library preparation to Illumina sequencing steps of: (i) library preparation, (ii) hybridization, (iii) capture and washing, (iv) enrichment, indexing and quality control, (v) preparation for multiplexed Illumina DNA sequencing, and (vi) initial post sequence processing, are provided in the Methods S1.

2.7 | Option 1—Target sequence assembly and extraction

Following the initial post-processing sequencing and quality control, the surviving sequence capture reads of each species were assembled using Trinity v2.6.5 with fastq format as input (\( \text{--seqType fq} \)), k-mer size of 25 (\( \text{--KMER_SIZE 25} \)), and the read normalization option disabled (\( \text{-- no_normalize_reads} \)) (Grabherr et al., 2011). Post assembly, to extract DNA sequences for the downstream phylogenetic analysis, we employed customized Bioperl scripts and Standalone BLAST+ in a similar way to that used in sequence selection for the target sequence sets (see Methods S1). Before phylogenetic analysis, the automated alignments were carefully checked by eye in Geneious V10 (Kearse et al., 2012). For a small fraction of the alignments (<2%), manual adjustments and/or locus exclusion, as deemed necessary, was performed at this stage. We also considered the outcomes of an independent assessment of paralogue risk loci via Option 3 (see below), leading us to exclude several additional loci not already excluded by our manual checks.

2.8 | Option 2—Homologous k-mer block discovery

Following trimming and de novo Trinity-assembling as described above for Option 1 (Figure 3), sets of homologous sequences (or k-mer blocks) in the assembled contigs were identified with Hakmer (Sanderson et al., 2017) and concatenated into a supermatrix for downstream phylogenetic analysis. Parameter settings \( k \) (k-mer length), \( q \) (the maximum number of mismatches allowed), \( N_{\text{min}} \) (minimum taxonomic coverage), \( w \) (width of flanking sequence) used for inferring respective phylogenies are shown in the figure captions (see also Methods S1 for further details).

2.9 | Option 3—HybPiper target sequence assembly and extraction

For comparison with analysis Options 1 and 2 (Figure 3), the standard HybPiper (Johnson et al., 2016), coupled with an automated phylogenomic data preprocessing pipeline was performed. We used the recommended (default) settings and the multiple relevant reference sequences (e.g., Set 1–2, 11 species; Set 3–4, six species, Set 5, seven species) for each target sequence set. As already noted, we also used
the HybPiper pipeline to identify loci with high paralogue risk, with this analysis performed across the 67 Diurideae wide samples (D67). By our criterion of requiring more than six of the 67 samples (>10%) at a locus to be flagged as potentially containing paralogues, very few paralogue risk loci were detected. See the Methods S1 for further details on the HybPiper workflow and results, and the Dryad repository (https://doi.org/10.5061/dryad.z08kprrbj) for additional paralogue analysis.

2.9.1 Phylogenetic analysis

To evaluate the phylogenetic utility of the sequence capture strategy, we conducted phylogenetic analysis of two test data sets with partially overlapping samples, as previously outlined above. In total, these phylogenetic analyses involved 134 samples, comprising 87 taxa. The D67 sample set, comprising 67 samples spanning the Diurideae, and an outgroup, was designed to assess phylogenetic informativeness across the entire tribe. Our C72 sample set, comprising 72 samples of Caladenia from subgenus Calonema, and C. denticulata from the subgenus Phlebochilus as the outgroup, was intended for exploratory analysis of sequence Set 5, targeting the Caladeniinae. One basis for sample inclusion in C72 was the requirement that two or more samples per species were successfully sequenced. Given a lack of species-level resolution in previous molecular studies (Clements et al., 2015; Swarts et al., 2014), we reasoned that a first test of the effectiveness of Set 5 would be if the phylogenetic analysis could group individuals of the same species together. We also included examples of species that, based on morphology, appear to be closely allied as members of “species complexes” but can still be confidently distinguished in the field (Brown et al., 2013). In the absence of morphological convergence, we predicted that taxa within species complexes should group together. To assess whether we could obtain a phylogeographic signal, for some species we also included replicate samples drawn from geographically distinct sites (Table S1).

For the purpose of this study, most phylogenetic inferences were based on a concatenated data set for each sequence set and data source (i.e., Option 1 exon and intron, as well as Options 2 and 3) to build a supermatrix of sequences. For analysis Option 1 data sets (Figure 3), this followed thorough manual checks of alignments at the locus level. For Options 2 and 3, only cursory checks were made so as to provide a test of the fully automated processes provided by these analysis pathways.

Phylogenetic gene trees and species trees based on the concatenated supermatrix were inferred by Maximum Likelihood analysis in IQ-TREE 2 (hereafter simply IQ-TREE) (Minh, Schmidt, et al., 2020; Nguyen et al., 2015). For in-frame exon data, the best substitution model for each locus was automatically selected by ModelFinder within IQ-TREE according to the Bayesian Information Criterium (Kalyaanamoorthy et al., 2017), before gene and species tree inference. Branch supports were obtained with the built-in ultrafast bootstrap algorithm (Hoang et al., 2018) from 1000 iterations. Subsequently, gene concordance (gCF) and site concordance (sCF) factors were estimated within IQ-TREE using the single locus (gene) and concatenated (species) tree as input (Minh, Hahn, et al., 2020). IQ-TREE is currently the only package available to estimate gCF while taking into account unequal taxon sampling, and provide the new concordance measure sCF (Minh, Schmidt, et al., 2020). The gene concordance factor (gCF) represents the percentage of gene trees containing that branch, while the site concordance factor (sCF) represents the percentage of parsimony informative sites supporting a branch. These concordance factors complement bootstrap estimates by offering additional information about the underlying variability in the data (Minh, Hahn, et al., 2020). For the in-frame exon data, the ML gene trees generated via IQ-TREE were also used as inputs into ASTRAL III (hereafter simply ASTRAL), to generate an alternative multispecies coalescent-based estimate of the species tree, with support estimated using local posterior probabilities with default parameters (Zhang et al., 2018). For the intron, k-mer (Option 2), HybPiper (Option 3) and combined data sets, a supermatrix was created before selection of the substitution model and IQ-TREE inference with ultrafast bootstrap.

All generated trees were visualized and rooted to the relevant outgroups for the data set in figtree v1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/, last accessed July 2020). To compare the network architecture of phylogenetic trees, tangle plots were visualized using PACO (Balbuena et al., 2013) in R version 4.0.2 (R Core Team, 2019).

3 RESULTS

3.1 Comparison of target and capture statistics

A summary of key target sequence assembly statistics for the 67 samples spanning the Diurideae (D67) and the 72 Caladenia samples (C72) is shown in Figure S1. For example, the median number of paired-end reads (trimmed and quality filtered) across our D67 target sequence assembly was approximately 636,000. Additionally, the median total number of assembled contigs across the 67 species was approximately 46,000 with respective median, average, and N50 contig lengths of 287, 416, and 471, respectively. Table 2 summarizes the target and capture statistics across the five sequence sets. The reported target statistics were based on the final coding DNA sequences submitted to Roche NimbleGen for the SeqCap EZ Developer Library design of the probes. For example, collectively the Set 1 and 2 target sequences spanned 475 loci, and a total sequence length of 609,538 bp. For the more specialized Sets 3 to 5, target sequences spanned a total of 3026 loci with a total sequence length of 4,062,168 bp. The target median length per locus was coincidentally 1428 bp each, across the two groups (Table 2).

The reported capture statistics are based on the Illumina sequencing outcomes, following the targeted enrichment of the sequencing libraries using the custom SeqCap EZ Developer Library probes. The analysis statistics reported per set in Table 2 were based on the inclusion of one representative of each of the species used
in the original design. For example, the Set 1 analysis was based on the sequencing outcomes for 11 samples each representing one of the core transcriptome species (see Table 1 for species involved). The results show very high capture rates averaging 85% of the target sequence for Sets 1 and 2, and an average of 73% coverage for Sets 3 to 5 (Table 2). The results fall within the Roche NimbleGen provided estimates of Predicted total probe design coverage of 68%, and Estimated probe coverage of 86%. Thus, the realized coverage achieved by the target sequence capture for the test samples was very close to the Roche NimbleGen predictions (Table 2). Note that these results are based on the outcomes from analysis pathway Option 1 with parameters that imposed high stringency for counting.

**Figure 4** The number of loci recovery by sequence Sets 1 to 5 for 18 genomic samples representing the transcriptomes used in the sequence capture design. Outcomes are shown for each of the three parallel analysis Options 1 to 3 (see Figure 3), with the Set relevance indicated for each species (see Figure 2). Please refer to Table S1 for the full species names [Colour figure can be viewed at wileyonlinelibrary.com]
the number of loci, number of exons, and sequence length recovered. Nonetheless, the recovered statistics such as the average number of captured exons per locus (CAv Exons/Locus) and median capture exon length (CMed Exon Len) are in line with those observed across other annotated orchid genomes (e.g., Apostasia shenzhenica, Phalaenopsis aphrodite, Phalaenopsis equestris) where they range between 228–283 and 3.9–4.8, respectively (Cai et al., 2015; Chao et al., 2018; Zhang et al., 2017; Zhang, Xu, et al., 2016).

The capture statistics shown in Table 2, which were achieved via analysis Option 1, cannot be directly compared with outcomes achieved via Options 2 and 3, because different methods and parameters are used. Nonetheless, it is of interest to compare the number of loci obtained by species for the three analysis options, across each sequence set. Figure 4 shows the number of loci recovered for each of the target sequence Sets 1 to 5 across 18 genomic samples representing the transcriptomes used in the sequence capture design (with replicates shown for two taxa). As expected, the number of loci recovered was more uniform across the test samples for Sets 1 and 2, aiming across the entire Diurideae, than for Sets 3–5, which aimed for phylogenetically narrower subsets of the tribe.

In a more extensive analysis, based on 67 samples spanning the tribe Diurideae (D67), box plots of the number of loci successfully captured are shown in Figure 5. This expanded analysis is further split into target (i.e., species belonging to the intended phylogenetic range as outlined above) and non-target species subsets. Target species represent those species falling within the intended phylogenetic target of the sequence set: Set 3, Other Diurideae; Set 4, Drakaeinae; Set 5, Caladeniinae (for more details, see the main text, Figures 2 and 4). Panel 1 via Option 1, after exclusion of exons that failed to be fully sequenced for 50% or more of the samples. Panel 2 via k-mer analysis Option 2. Panel 3 via HybPiper analysis Option 3 (see Figure 3) [Colour figure can be viewed at wileyonlinelibrary.com]

3.2 | Outcomes of phylogenetic analysis across the Diurideae

An IQ-TREE phylogeny obtained via analysis Option 1 for the 67 samples spanning the Diurideae, and outgroup (D67) is shown in Figure 6a with ultrafast bootstrap values, gene concordance factors (gCF) and site concordance factors (sCF) shown on the tree branches. This ML phylogeny is based on concatenated coding sequences with an alignment length of 250,431 bp from 221 loci targeted by Set 2. Three major clades are indicated:

Clade 1 comprising subtribe Prasophyllinae and represented by three samples drawn from two genera, Microtis and Prasophyllum. Clade 2 comprising the subtribes (Diuridinae(Cryptostylidinae(Theley mitrinae)(Megastylidinae(Drakaeinae))))). Clade 3 comprising subtribes (Acianthinae(Caladeniinae)).

Virtually all nodes had 100% bootstrap support, with a notable exception of lower support for the split of clades 2 and 3 (Figure 6a). Note that with the ultrafast bootstrap method of IQ-TREE (Hoang et al., 2018; Minh, Schmidt, et al., 2020), we consider strong support when bs ≥95. However, we show bootstrap values and concordance factors (where applicable) for bs ≥70, on our figures in order to provide more information on the uncertainty. As expected theoretically (Minh, Hahn, et al., 2020), across the entire D67 phylogeny the gene concordance factor (gCF) and site concordance factor (sCF) values were significantly correlated (R = 0.86, p < 10−16), with an average gCF of 66.6 (range 19.1 to 100) and average sCF of 59.3 (range sCF 31.8 to 98.8) (see Figures S3–S4 for more detailed analysis). The observed variation in concordance factors (both gCF and sCF) falls well within the range of similar observations by Minh, Hahn, et al.
(2020) for phylogenies from large data sets. The equivalent IQ-TREE phylogenies obtained via analysis Options 2 and 3, revealed highly comparable topologies as evident in the tangle plots shown in the inserts, respectively. These tangle plots reveal only one minor difference in tree topology between analysis Option 1 and 2, involving two closely related Drakaea species (Figure 6b,c, see also Figure S5 for full-size trees, and Figure S6 for full-size tangle plots).

Across the D67 samples almost identical groupings of species and genera into the same three major clades, were obtained via IQ-TREE for the Option 1 sequence Sets 1 to 3 and via ASTRAL (c.f. Figure 6a with Figure S7 and S8). However, in all cases there is considerable uncertainty about the precise relationships among the three major clades, with these adjoining nodes being characterized by very short branches, and low bootstrap and local posterior probability support values. Concordance factors were also low at these nodes. For example, a gCF value of 19.1, and sCF value of 33.5, was found at the node joining Clade 1, with Clades 2 and 3 in Figure 6a. This sCF value is very close to the expected minimum of 33, indicating there is no
consistent information in the alignment concerning the relationships among clades 1, 2 and 3 (Minh, Schmidt, et al., 2020). The ASTRAL results reinforce these findings with a low local posterior probability of 0.65 on the branch joining clades 2 and 3 (Figure S8).

### 3.3 Outcomes of phylogenetic analysis in Caladenia

A summary of key target sequence assembly statistics for the 72 Caladenia orchid samples (C72) is shown in Figure S1. For example, the median number of paired-end reads (trimmed and quality filtered) used for the target sequence assembly of the C72 sample set was approximately 1,236,000. Additionally, the median total number of assembled contigs across the 72 species was approximately 53,000 with respective median, average, and N50 contig lengths of 287, 430, 502.

Figure 7a shows the IQ-TREE phylogeny for the C72 samples based on coding exon sequences at 719 loci yielding an alignment length of 755,184 bp. The illustrated phylogeny, focused on the Caladenia subgenus Calonema, was based on sequence data where each locus was represented in each sample (although with some missing data within loci), demonstrating the high sequencing coverage achieved. With one minor exception (C. pectinata), within taxon (species or subspecies) replicates from a given collection site grouped together. The C. longiclavata complex formed a clade on a well-supported branch. A clade containing members of the C. hugelii complex and the C. longicadata complex was also supported, but within this clade support for the complexes and relationships among the species was weak. Nonetheless, of interest is the relationships indicated among four of the named subspecies of C. longicata, where samples of C. longicata subsp. redacta appeared to be more closely related to other species than the remaining three C. longicata subspecies included in our analysis (crassa, ridigula and insularis) (Figure 7).

The IQ-TREE bootstrap values, gene concordance factors (gCF) and site concordance factors (sCF) are also shown on the tree branches of the C72 phylogeny in Figure 7a (see also Figures S3 and S4). Given the genus level focus and our inclusion of replicate species and subspecies samples, it is not surprising that overall the concordance factors tended to be much lower (mean gCF = 9.8, range 0 to 95.3; mean sCF = 43.1, range 32.2–99.0) than observed across the tribe wide D67 phylogeny (Figure 6a), but with gCF and sCF remaining significantly correlated (R = 0.88, p < 10⁻¹⁵). Nonetheless, on close inspection there were interesting concordance factor patterns. For example, within each of the major clades, the concordance factors were higher on the nodes within taxa (species or subspecies) (mean gCF = 11.1, range 0–34.5; mean sCF = 43.7, range 33.0–74.7), than for the among species nodes within the clades (mean gCF = 3.2, range 0–11.5; mean sCF = 35.9, range 32.2–45.3), mirroring the very short branches and low bootstrap values. While, overall the highest concordance tended to be found on the nodes connecting the major clades (mean gCF = 18.6, range 0–95.3; mean sCF = 53.6, range 32.7–99.0). Note that these “major clades”, included the combined C. hugelii and C. longicata complex, the C. longicata complex, and each of the clades containing C. discoidea, C. reticulata and C. denticulata, among others (See Figure 7a which highlights in colour some, but not all of the major clades). The outcomes of ASTRAL species tree estimation revealed a similar pattern of higher local posterior probability support for within taxa replicate samples, but lower support for the relationships among taxa within the major clades, while revealing a very similar major clade structure to the IQ-TREE phylogenies (c.f. Figure 7a, Figure S9).

For those species with multiple samples from two or more sites, the pairing of samples by collection site was the norm in both the IQ-TREE and ASTRAL analyses. For example, the pair of C. reticulata samples taken from South Australia (SA), and the pair sampled from Victoria (Vic), some 400 km away, grouped together by site (Figures 7a and 8a, Figure S10). Other notable pairings by site were observed in the species C. denticulata and C. discoidea where the replicate sites were 260 and 200 km apart, respectively (Figure 7a). A pattern of clustering by geographic proximity was also evident over finer geographic scales (<100 km) in other species such as C. ensata, C. paludos, C. richardsoniorum and C. uliginosa (Figure 7a, Figure S10, see also Table S1 for site locations and species codes). Notably, these phylodogeographic groupings within species tended to show higher gCF and sCF values than the average within taxon values (see Figure 7a). Figure 8a shows an IQ-TREE phylogeny based on the homologous k-mer analysis Option 2, with a total alignment of 2,482,350 bp. Inserts show the tangle plot contrasts of exon (Figure 8b) and intron (Figure 8c) trees based on Set 5 loci only (see also Figure S10 for intron based tree, Figures S11–S13 for full-size tangle plots, and Figure...
From these plots it is evident that the k-mer based tree showed broad congruence in topology with both the exon and intron trees. The few differences evident in the tangle plots either comprised: (i) subtle differences in relationships among samples within a species (e.g., Caladenia longicauda), or (ii) variation in the precise relationships among species, but with paired replicates always clustering together (e.g., Caladenia longicauda subsp. insularis). From Table S1 the full species names. Photographs by Rod Peakall. (Colour figure can be viewed at wileyonlinelibrary.com)
C. thinnicola in the C. huegelli complex). Of interest, is that in the k-mer based tree (Figure 8a), bootstrap support for the relationships among the species belonging to the C. huegelli and C. longicauda complexes was stronger than found for the Set 5 exon tree (Figure 7a). Furthermore, all four species representing the C. huegelli complex grouped together with stronger support than obtained in the exon-based tree. The C. longicauda complex comprised two well supported clades, with C. longicauda redacta remaining distinct from the other subspecies, as observed in the exon-based tree. The pairing of samples by collection site across both the larger geographic scales of 200–400 km, and finer geographic scales (<100 km), was also observed (Figure 8a).
4 | DISCUSSION

4.1 | Overview

Despite early reviews outlining the potential benefits of sequence capture (e.g., Grover et al., 2012; Lemmon et al., 2012; Lemmon & Lemmon, 2013), it seems that its uptake for phylogenetic and phylogenomic studies of plants has not matched the initial enthusiasm (but a growing number of exceptions include: Barrett et al., 2018; Bogarin et al., 2018; de La Harpe et al., 2019; Folk et al., 2015; Heyduk et al., 2016; Johnson et al., 2019; Mendoza et al., 2020, among others). Furthermore, this apparent delay in wide uptake seems to be despite: (i) the availability of informative Angiosperm wide probe sets as a promising starting point (Bogarin et al., 2018; Buddenhagen et al., 2016; Johnson et al., 2019; Léveillé-Bourret et al., 2017; Mendoza et al., 2020) and (ii) increasing availability of “off the shelf” bioinformatic solutions and pipelines, such as HybPiper designed initially for plant sequence capture studies, but more broadly applicable to any group of organism (Johnson et al., 2016). This may suggest that for many plant families the lack of relevant genomic and transcriptomic resources has been perceived as a key impediment to building a custom sequence capture strategy for specific plant groups of interest.

On the other hand, within the diverse Orchidaceae, despite the availability of a growing number of genomic and transcriptomic resources (Figure 1), the uptake of sequence capture also appears to be rare. Mendoza et al. (2020) demonstrated that the application of the Angiosperm 1 “anchored hybrid enrichment (AHE)” probe set (Buddenhagen et al., 2016; Léveillé-Bourret et al., 2017), without modification, provided both phylogenetically informative target nuclear and off-target plastid loci across 18 Epidendrum species, and additional outgroups. Similarly, Bogarin et al. (2018) achieved a fully resolved phylogeny among 16 closely related Lepanthes orchid species using an orchid-specific modification of the Angiosperm 1 probe set. Chloroplast targeted sequence capture has also been used effectively to investigate the degree and rate of plastome degeneration in a complex of Corallorhiza heterotrophic orchids (Barrett et al., 2018).

The goal of this present study was to develop, implement, and evaluate a multitiered target sequence capture strategy for orchid phylogenetic studies spanning from subtribe through to phylogeographic analysis within species. Our target group was the rich terrestrial orchid flora belonging to the Diurideae within the Orchidoideae, with >900 species. At the onset of our project in 2015, no orchid genomes were available, and currently, there is still no genome for any species within the Orchidoideae (see Figure 1). Nonetheless, we were able to draw on our existing transcriptomes generated for other research purposes (Wong et al., 2018; Wong, Amarasinhe et al., 2017; Xu et al., 2017). To extend the applicability of our sequence capture design, we also completed additional transcriptome sequencing of a Pterostylis (Cranichideae) as the outgroup, and additional representatives of other Diurid subtribes (Figure 2, Table 1). The inclusion of Pterostylis is expected to extend applicability to the more than 350 species in that genus (Backhouse et al., 2019). Our Set 1, targeting putative Orchidaceae wide orthologues (Deng et al., 2015), should also allow the construction of phylogenies for a common set of orthologous loci beyond the target group, with the caveat that capture success is expected to decline with increasing phylogenetic distance. Thus, despite our prime focus on the Diurideae, we expect our sequence capture strategy will offer wider applicability.

Our evaluation of sequence capture success confirmed that the goal of a wide phylogenetic span was broadly achieved. For example, the phylogeny across the D67 test samples shown in Figure 6a is based on 221 of the 245 loci targeted by Set 2 (90%) with an alignment length of 250,431 bp. Similarly, the equivalent phylogeny shown in Figure S7A is based on 193 of the original 230 loci targeted by Set 1 (84%), with an alignment length of 200,935.

Our evaluation for Set 5, based on the C72 Caladenia samples, also confirmed the capacity for high sequence capture success. For example, for illustrative purposes, the exon-based phylogeny we show in Figure 7a is based on 719 of the 1052 maximum possible loci (68%), yielding a total alignment length of 755,184 bp. Allowing for even a small fraction of missing data (e.g., 10%) considerably expands the number of loci. It is thus evident that our multipurpose sequence capture strategy can achieve the uniform capture of a high proportion of the target sequences over samples spanning from the outgroup of the tribe Diurideae, across subtribes, to within species.

4.2 | An evaluation of homologous k-mer block phylogenies from target sequence capture derived data sets

With the increasingly widespread use of phylogenomic data sets (e.g., Johnson et al., 2019; Leebens-Mack et al., 2019), alternative solutions to the use of conventional annotation/alignment pipelines are now being developed and carefully evaluated (Bernard et al., 2019). This is in part due to the inherent difficulties of scaling up standard phylogenetic procedures to gigabase scales, but is also due to complications in sequence annotation, orthologue detection, and sequence alignment, among others (Sanderson et al., 2017). In this context, the use of k-mers (i.e., substrings of length k in a given biological sequence), is gaining popularity in phylogenomic analysis workflows as an independent way to infer phylogenies based on concatenated “supermatrices” and/or sets of “gene trees” (Bernard et al., 2019; Bertels et al., 2014; Fan et al., 2015; Leimeister & Morgenstern, 2014; Sanderson et al., 2017).

Sanderson et al. (2017) recently assessed the utility of homologous k-mers blocks across seven gigabase scale data sets (e.g., nuclear genomes and transcriptomes), confirming the potential to infer robust phylogenies across tens to hundreds of species spanning wide taxonomic breadths. By way of example using just transcriptomes, a fully resolved tree of 67 Caryophyllales species was obtained using Hakmer (with parameters $k = 28$; $N_{\text{min}} = 15$ of 67 taxa; and $q = 2$). Furthermore, the k-mer-based tree showed agreement at every node, except one, with the published tree obtained via the conventional approach (based on 1,122 putative orthologous loci shared among the Caryophyllales transcriptomes [Yang et al., 2015]).
To our knowledge, to date, no study appears to have applied homologous k-mer blocks to the analysis of sequence capture data for phylogenetic inference. Indeed, the need to apply a k-mer approach to such data may at first seem counterintuitive, since by design the target loci are known. Here, we considered the utility of this approach in three ways: (i) As part of an evaluation and validation of our sequence capture assembly and extraction strategies (analysis Options 1 and 3). (ii) As a way to more fully utilise both the target and off-target sequence data obtained by sequence capture. (iii) To test an alternative, complementary, and potentially faster way to enable species tree inferences.

To this end, we used Hakmer (Sanderson et al., 2017) to first identify homologous k-mer blocks within de novo assembled target sequence capture reads (analysis Option 2), before using the concatenated “supermatrix” of k-mer blocks in the same type of IQ-TREE analysis as applied in Options 1 and 3. While analysis Option 1 and 3 took weeks, this approach took just days. It also sidestepped the arduous annotation, alignment, and quality checks that in our view 3 took weeks, this approach took just days. It also sidestepped the arduous annotation, alignment, and quality checks that in our view are critical components of analysis Options 1.

Three key outcomes emerged from our evaluation of the homologous k-mer blocks approach (Option 2): (i) Congruent phylogenetic trees to those obtained via conventional analyses were recovered across the Diurideae (Figure 6, Figures S5 and S6), and within Caladenia (Figures 7–8, Figures S11 and S12). (ii) Considerably more sequence data was utilised in the Option 2 homologous k-mer analysis, than compared to Option 1 (D67: Option 1, Sets 1 to 5 = 1.34 million bp vs. Option 2 = 3.39 million bp. C72: Option 1, Set 5 = 0.75 million bp vs. Option 2 = 2.48 million bp). These results reflecting the inclusion of additional, potentially informative, homologous untranslated regions, introns, other off-target homologous nuclear, chloroplast and mitochondrial DNA sequences. (iii) Evidence for finer-scale phylogenetic resolution among closely related Caladenia species (c.f. Figures 7a and 8a). This may be attributable to the addition of informative off-target homologous sequences, as observed in a recent sequence capture study of Epidendrum orchids (Mendoza et al., 2020).

Thus, our preliminary evaluation of the application of homologous k-mer blocks approach to targeted sequence capture data revealed its potential to recover phylogenetic species tree inferences matching conventional analysis involving target gene extraction, alignment, and supermatrix phylogenetic analysis. This was also achieved with dramatically increased speed and simplicity of data analysis, over analysis Options 1 and 3. Furthermore, the inferred k-mer-based phylogeny appeared to better group morphologically similar members of species complexes. At the same time, consistent with the collective evidence, the polyphyly of some taxonomic entities was indicated.

4.3 | New phylogenetic insights

Our prime objective in this method focused paper was to evaluate the performance of our multipurpose targeted sequence capture strategy, rather than undertaking a comprehensive phylogenetic analysis. With an expanded data set, such analyses will be the subject of subsequent studies. Nonetheless, our exploratory phylogenetic analysis based on concatenated captured gene sequences from a set of 67 samples spanning the Diurideae and outgroup is of interest (Figure 6). While broadly recovering similar groupings of species and genera at the subtribe levels to Weston et al. (2014) (c.f. Figures 2 and 6), our results indicate better resolution, particularly among closely related genera (e.g., among genera within the Drakaeinae, Megastylidinae, and Caladeniinae, see Clements et al., 2013; Miller & Clements, 2014; Weston et al., 2014). However, our inability to resolve the relationships between the three major clades comprising the Prasophyllinae as Clade 1, Clade 2 as (Diuridinae(Cryptostylidinae(Theylmitrinae(Megastylidinae(Drakaeinae))))) and Clade 3 as (Acianthinae(Caladeniinae)), mirrors earlier findings (Clements et al., 2002; Kores et al., 2001; Weston et al., 2014, Figure 2), despite the very substantial increase in the number of loci. Furthermore, this phylogenetic pattern was found across the Option 1 to 3 IQ-TREE phylogenies, across sequence Sets 1 to 3 (e.g., Figure 6, Figures S5 and S7), and in the ASTRAL analysis (Figure S8). As already noted, very short branches, and low gCF and sCF scores characterize these nodes (Figure 6a). Unfortunately, low gCF scores can be due to either strong gene tree discordance or weak phylogenetic signal (Minh, Hahn, et al., 2020), which can be difficult to tease apart.

Similar findings of the inability to resolve relationships in the backbone of phylogenetic trees, particularly in plant groups with early radiations, are well known (Léveillé-Bourret et al., 2017). Furthermore, despite the substantial increases in the number of loci via sequence capture, deep backbone phylogenetic relationships with short branches continue to remain unresolved in plant groups such as the Cyperaceae (Léveillé-Bourret et al., 2017) and Campanulaceae (Bagley et al., 2020), but see both these studies for discussion on possible methodological and phylogenetic analysis solutions towards solving this problem. In common with our observations in orchids, these problematic backbone nodes are characterized by short branches, low bootstrap support, and high levels of gene tree discordance. This discordance may be due to several factors including incomplete lineage sorting (ILS), short times between divergences, and use of loci with low phylogenetic information. Notwithstanding these phylogenetic analysis challenges, collectively, our findings indicate that with expansion to additional samples, and expanded locus filtering (e.g., to remove loci with low phylogenetic information [Burbrink et al., 2019]) prior to further detailed phylogenetic analysis, this data set provides a valuable resource to gain insights into the phylogeny of the Diurideae.

While a more comprehensive phylogenetic analysis of Caladeniinae will follow, some new insights are already apparent from our analysis of the 72 Caladenia samples comprising 24 species. As already noted, we included both replicate samples per taxon (species or subspecies), and multiple examples of species that, based on morphology appear to be closely allied as members of “species complexes”, but can still be confidently distinguished in the field. While in most phylogenetic analyses (Set 5 exons—Figure 7a, k-mer—Figure 8a, ASTRAL—Figure S9, introns—Figure S10), replicate
samples grouped together, only in the k-mer based phylogeny did all members of the C. huegelii and C. longicauda species complexes, each fall into their own well supported clades. Across all analyses C. longicauda redacta did not group with the other subspecies. However, it did group consistently with Caladenia uliginosa, another species within the C. longicauda complex (Figures 7–8, Figures S9 and S10). Thus, as presently defined, based on morphology, the subspecies of Caladenia longicauda appear to be polyphyletic. Furthermore, there is also evidence for deeper phylogenetic structure within the more broadly defined C. longicauda species complex. With a total of 14 named Caladenia longicauda subspecies (Brown & Brockman, 2015), and more than 11 other morphologically similar species, it is evident that a more in depth phylogenetic analysis is not only warranted, but now feasible.

Nonetheless, resolving relationships among other closely related Caladenia species will probably remain an ongoing biological and technical challenge. This is evident from the trend of lower-than-average support values on the nodes joining taxa (species and subspecies) within the major clades (Figures 7 and 8, Figure S9–S10, S14) for all of the following measures: bootstrap, local posterior probabilities, gCF and sCF scores. It is likely that many of the clades within Caladenia have diversified rapidly, with much incomplete lineage sorting, and potentially with ongoing hybridization. However, as we have proposed for the Diurideae, smart locus filtering informed by more detailed analysis of phylogenetic informativeness and additional insights from gene and site level concordance analysis, may hold promise for improving resolution.

4.4 | Future phylogenetic directions

We have yet to evaluate how the IQ-TREE and ASTRAL species trees reported in this study will compare with the trees obtained after compositional heterogeneity testing (exon data sets), other forms of locus filtering, and with other coalescence-based phylogenetic analysis. We do note that target sequence capture is particularly well suited to a wide range of phylogenetic methods, given that in-frame exon regions are readily available. Not only do in-frame loci offer superior alignment opportunities based on their amino acids (Abascal et al., 2010; Bininda-Emonds, 2005; Hall, 2005), model inference performs better when codon positions are known (Shapiro et al., 2005). Different rates of evolution can also be allowed for each exonic region, and when DNA saturation is high (e.g., in phylogenetic inference of older lineages), amino-acid models may be used instead (Simmons, 2017). Finally, to maximize phylogenetic information, the evolutionary rate of loci can be matched to the phylogenetic scale being explored, with slow evolving sequences targeted for deep divergences, and fast evolving sequences for recently diverged lineages (Philippe et al., 2011). The sequence Sets 1 and 2 developed here, represent the two most conserved sets of loci, and should be highly suitable to resolve the deeper divergences in the Diurideae, whereas sequence Sets 4 and 5 are more appropriate for species level divergences in the subfamilies they target. The latter type of locus is ideally suited to coalescent-based phylogenetic methods, which can account for differential lineage sorting among loci in rapidly diverging lineages (Carstens and Knowles 2007). However, it remains to be seen whether these analyses will result in higher phylogenetic resolution compared to that already obtained with the k-mer data set.

4.5 | Lessons and tips for future studies

Achieving a high level of sequence capture success across moderately wide phylogenetic scales has been a common objective of many sequence capture projects. While noting there is no standardized approach, examples of customized sequence capture designs aimed at species to higher levels of phylogenetic scale have been reported for mammals (Bi et al., 2012), reptiles (Bragg et al., 2015), amphibians (Banker et al., 2019), insects (Zhang et al., 2019) and plants (de La Harpe et al., 2019; Jantzen et al., 2020). Alternatively, some sequence capture designs are focused on a narrower phylogenetic breadth, such as a single plant family (e.g., Jones et al., 2019; Larridon et al., 2020) or plant genus (e.g., Folk et al., 2015; Heyduk et al., 2016; Tomaszello et al., 2020).

With the recent emergence of Angiosperm wide probes sets (e.g., Buddenhagen et al., 2016; Dodsworth et al., 2019; Johnson et al., 2019; Léveillé-Bourret et al., 2017), other studies are now combining the use of both universal and specifically designed probe sets to leverage the benefits of both approaches (e.g., Chau et al., 2018; Larridon et al., 2020). While studies have demonstrated that universal probes can resolve the phylogenetic relationships among closely related species (e.g., among Carex species (Larridon et al., 2020; Léveillé-Bourret et al., 2017); and Lepanthes (Bogarin et al., 2018) and Epidendrum orchids (Mendoza et al., 2020), we predict that custom sequence capture strategies targeting additional loci may still be essential in many groups of plants. Fortunately, given the genomic and transcriptomic resources emerging for the Orchidaceae (Figure 1), and for a growing number of other families (Leebens-Mack et al., 2019), the opportunities for a rapid uptake of customized targeted sequence capture already seems feasible for many plant groups.

In this study, our goal was to simultaneously capture in a single hybridization step hundreds of informative loci that collectively enable the phylogenetic resolution of relationships spanning among subtribes, through genera and species to subspecific phylogeographic structure within species. To achieve this goal, we employed a tiered approach drawing on a total of 18 transcriptomes that we then subdivided into five different sets, spanning different phylogenetic scales, in an effort to achieve the sequence capture of loci offering phylogenetically informativeness at different levels.

Our evaluation of the performance of our five-target sequence sets in capturing sequences from target and non-target taxa, demonstrates the importance of ensuring probe design incorporates sequence variation spanning the phylogenetic breadth of interest.
For example, for Sets 1 and 2, targeting the entire tribe and outgroup, and for the target species relevant to Sets 3 to 5, sequence capture success was uniformly high (>89%). However, across Sets 3 to 5, sequence capture success fell (66%-70%) for the non-target species (Figures 4 and 5).

One alternative way to evaluate probe set performance across target and non-target taxa is to consider IQ-TREE estimates of concordance factors, and ASTRAL estimates of normalized quartet score across the D67 samples. As expected (Minh, Hahn, et al., 2020), gCF and sCF factors were correlated across the species trees for Set 1 to 5 (Figure S3), and showed a strong positive relationship with branch length (Figures S3 and S4). Nonetheless, average gCF values were highest for Set 2, which was designed to span the Diurideae (median of 77), and lowest for Set 5 (median of 58), which was designed for the much narrower taxonomic target of the Caladeniinae. Median sCF scores were similar across the sets (Figure S3). The ASTRAL normalized quartet scores were similar for Set 1 and 2 (0.903 vs. 0.906, respectively), marginally lower for Set 3 (0.898) and Set 4 (0.875), and lowest for Set 5 (0.858). Collectively, these findings demonstrate an increase in gene tree discordance as the number of non-target taxa increased. It follows, that in practice sequence Sets 4 and 5 should not be chosen as a whole for phylogenetic analysis beyond their target groups of the Drakaeinae and Caladeniinae, respectively. However, given that there was only a modest decrease in the median gCF scores and in the ASTRAL normalized quartet scores across the D67 samples for Sets 4 and 5, these sequence sets may still contain many loci that will be informative beyond the initial taxonomic targets. Therefore, in order to maximize high levels of uniform coverage we recommend that probes are designed from sequence data spanning the phylogenetic breadth of interest. Nonetheless, some probe loci are likely to be successfully captured and be potentially phylogenetically informative far beyond initial narrower phylogenetic targets.

5 | CONCLUSIONS

Three complementary sequence sets targeting the Diurideae (Sets 1 to 3) have been confirmed to successfully achieve the sequence capture of more than 1,000 nuclear loci. Our exploratory phylogenetic analyses, based on concatenated coding gene sequences, indicate congruence at deeper phylogenetic levels with earlier molecular studies based on just one or a few genes (Kores et al., 2000, 2001; Weston et al., 2014). At the same time, access to hundreds more nuclear loci promise to help fill remaining gaps in our understanding of the phylogenetic relationships across the tribe. Set 5, targeting closely related species of Caladenia, revealed unprecedented phylogenetic resolution when compared with earlier studies (e.g., Clements et al., 2015; Swarts et al., 2014) based on just a few loci, and also indicates potential to detect phylogeographic structure within some species. The uniform coverage of hundreds of loci across every sample was also achieved (Figure 7). Our exploration of homologous k-mer analysis reveals overlooked advantages for phylogenetic inference from supermatrices of sequence capture data, including: (i) the rapid assembly of data while maximizing the use of homologous sequences (e.g., target exons and introns, untranslated regions, and other informative off-target sequences), and (ii) the recovery of robust phylogenetic structure that is biologically meaningful.

Given that many phylogenetic, phylogenomic and taxonomic uncertainties remain across this hyperdiverse family (e.g., Chen et al., 2019; Hu et al., 2020; Li et al., 2019; Mendoza et al., 2020; Perez-Escobar et al., 2020; Salazar et al., 2018; Simo-Droissart et al., 2018), the Orchidaceae is a prime candidate for wider testing of the phylogenetic and phylogenomic utility of targeted sequence capture methods. Beyond orchids, our findings indicate that by drawing on the increasing availability of transcriptomes, a customized multi-tiered sequence capture strategy, in combination with promising, yet underutilized phylogenomic approaches, will be effective for many plant groups.

ACKNOWLEDGEMENTS

We thank Arild Ariffin, Thelma Bridle, Mark Clements, Katharina Nagar, Bev Overton, Nouchka Reiter and Alyssa Weinstein for providing some of the DNA samples, and Tobias Hayashi for providing photographs. Mark Clements generously shared his insights into both the known and uncertain relationships among the Australian orchids, which guided our selection of target species for building transcriptomes. Robert Lanfear, Bui Quang Minh and Renee Catullo provided advice on IQ-TREE and the interpretation of concordance factors. We also thank three anonymous referees for their helpful and insightful comments that improved the manuscript. This work was supported by Australian Research Council projects: DP1094453 to RP; LP130100162 to RP and CCL, DP150102762 to RP and CCL; DE150101720 to RDP; and DE190100249 to DCJW. We are also grateful to the ANU Centre for Biodiversity Analysis for an Ignition grant to RP and Katharina Nagar (CSIRO). Our ongoing collaborator, Eran Pichersky (University of Michigan), generously funded the cost of some of the early ancillary transcriptomes that were utilized in this study.

AUTHOR CONTRIBUTIONS

Rod Peakall, Darren Wong, Ryan Phillips, and Celeste Linde designed the research; all authors performed the research; Rod Peakall, Ryan Phillips, Monica Ruibal and Claudia Rodriguez-Delgado collected the specimens; Rod Peakall, Darren Wong, Claudia Rodriguez-Delgado and Celeste Linde analysed the data; Rod Peakall, Darren Wong, Ryan Phillips, Monica Ruibal and Celeste Linde wrote the manuscript.

DATA AVAILABILITY STATEMENT

The Dryad repository (https://doi.org/10.5061/dryad.z08kprbj) contains: (i) DNA sequences in Fasta format for each of sequence Sets 1 to 5, as provided to Roche NimbleGen for the probe design. (ii) Final sequence alignment and homologous k-mer block files for all phylogenetic trees provided in the main text and Supporting Information. (iii) Treefiles for all phylogenetic trees provided in the...
main text and Supporting Information. (iv) Functional annotation and MapMan BIN ontology of the target gene sequences. (v) Outcomes of HybPiper paralogue risk analysis.

ORCID
Rod Peakall https://orcid.org/0000-0001-9407-8404

REFERENCES
Abascal, F., Zardoya, R., & Telford, M. J. (2010). TranslatorX: Multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Research, 38, W7–W13.

Backhouse, G. N., Bates, R. J., Brown, A. P., & Copeland, L. M. (2019). A checklist of the orchids of Australia including its Island Territories. Melbourne, Australia: Gary Backhouse. http://www.anosovic.org.au/Checklist_of_Australian_Orchids_2nd_Ed_2019.pdf

Bagley, J. C., Uribe-Convers, S., Carlsen, M. M., & Muchhala, N. (2020). Utility of targeted sequence capture for phylogenomics in rapid, recent angiosperm radiations: Neotropical Burmeistera bellflowers as a case study. Molecular Phylogenetics and Evolution, 152, 106769.

Balbuena, J. A., Miguez-Lozano, R., & Blasco-Costa, I. (2013). PACAo: A novel procurites application to cophylogenetic analysis. PLoS One, 8, e61048.

Banker, S. E., Lemmon, A. R., Hassinger, A. B., Dye, M., Holland, S. D., Kortyna, M. L., Opsina, O. E., Ralicki, H., & Lemmon, E. M. (2019). Hierarchical hybrid enrichment: multi-tiltered genomic data collection across evolutionary scales, with application to choru frogs (Pseudacris). Systematic Biology, 69, 756–773.

Barrett, C. F., Wicke, S., & Sass, C. (2018). Dense intraspecific sampling reveals rapid and independent trajectories of plastome degradation in a heterotrophic orchid complex. New Phytophilolog, 218, 1192–1204.

Bernard, G., Chan, C. X., Chan, Y.-B., Chua, X.-Y., Hogan, J. M., Maetschke, S. R., & Ragan, M. A. (2019). Alignment-free inference of hierarchical and reticulate phylogenomic relationships. Briefings in Bioinformatics, 20, 426–435.

Bertels, F., Silander, O. K., Pachkov, M., Rainey, P. B., & Van Nimwegen, E. (2014). Automated reconstruction of whole-genome phylogenies from short-sequence reads. Molecular Biology and Evolution, 31, 1077–1088.

Bi, K. E., Vanderpool, D., Singhal, S., Linderoth, T., Moritz, C., & Good, J. (2012). Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. BMC Genomics, 13, 403.

Bininda-Emonds, O. R. P. (2005). transAlign: using amino acids to facilitate the multiple alignment of protein-coding DNA sequences. BMC Bioinformatics, 6, 156.

BLAST (2008). Command line applications user manual [Internet]. National Center for Biotechnology Information [US]. https://www.ncbi.nlm.nih.gov/books/NBK279690/

Bogarin, D., Pérez-Escobar, O. A., Groenenberg, D., Holland, S. D., Karremans, A. P., Lemmon, E. M., Lemmon, A. R., Pupulin, F., Smeds, E., & Gravenende, B. (2018). Anchored hybrid enrichment generated nuclear, plastid and mitochondrial markers resolve the Lepanthes hornia (Orchidaceae: Pleurothallidinae) species complex. Molecular Phylogenetics and Evolution, 129, 27–47.

Bohman, B., Fiematti, G. R., Barrow, R. A., Pichersky, E., & Peakall, R. (2016). Pollination by sexual deception—it takes chemistry to work. Current Opinion in Plant Biology, 32, 37–46.

Bragg, J. G., Potter, S., Bi, K., & Moritz, C. (2015). Exon capture phylogenomics: Efficacy across scales of divergence. Molecular Ecology Resources, 16, 1059–1068.

Brown, A. P., & Brockman, G. (2015). New taxa of Caladenia (Orchidaceae) from south-west Western Australia. Nuytsia, 25, 45–123.

Brown, A., Dixon, K., French, C., & Brockman, G. (2013). Field guide to the orchids of Western Australia: The definitive guide to the native orchids of Western Australia. Simon Neill Publications.

Buddenhagen, C., Lemmon, A. R., Lemmon, E. M., Bruhl, J., Cappa, J., Clement, W. L., Donoghue, M. J., Edwards, E. J., Hupp, A. L., Kortyna, M., Mitchell, N., Moore, A., Prychid, C. J., Segovia-Salcedo, M. C., Simmons, M. P., Sortis, P. S., Wanke, S., & Mast, A. (2016). Anchored Phylogenomics of Angiosperms I: Assessing the robustness of phylogenetic estimates. bioRxiv, 86298.

Burbrink, F. T., Ruane, S., Kuhn, A., Rabibisno, N., Randramanantahanto, B., Raselimanana, A. P., Andiranaramila, M. S. M., Cadle, J. E., Lemmon, A. R., Lemmon, E. M., Nussbaum, R. A., Jones, L. N., Pearson, R., & Raxworthy, C. J. (2019). The origins and diversification of the exceptionally rich Gembawks (Colubroidea: Lamprophiidae): Pseudoxyrophinae) in Madagascar. Systematic Biology, 68, 918–936.

Cai, J., Liu, X., Vanneste, K., Proost, S., Tsai, W.-C., Liu, K.-W., Chen, L.-J., He, Y., Xu, Q., Bian, C., Zheng, Z., Sun, F., Liu, W., Hsiao, Y.-Y., Pan, Z.-J., Hsu, C.-Y., Yang, Y.-P., Hsu, Y.-C., Chuang, Y.-C., ... Liu, Z.-J. (2015). The genome sequence of the orchid Phalaenopsis equestris. Nature Genetics, 47, 65–72.

Carstens, B. C., & Knowles, L. L. (2007). Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: An example from Melanopus grasshoppers. Systematic Biology, 56, 400–411.

Chao, Y.-T., Chen, W.-C., Chen, C.-Y., Ho, H.-Y., Yeh, C.-H., Kuo, Y.-T., Su, C.-L., Yen, S.-H., Hsueh, H.-Y., Yeh, J.-H., Hsu, H.-L., Tsai, Y.-H., Kuo, T.-Y., Chang, S. B., Chen, K.-Y., & Shih, M.-C. (2018). Chromosome-level assembly, genetic and physical mapping of Phalaenopsis aphrodite genome provides new insights into species adaptation and resources for orchid breeding. Plant Biotechnology Journal, 16, 2027–2041.

Chao, Y.-T., Yen, S.-H., Yeh, J.-H., Chen, W.-C., & Shih, M.-C. (2017). Orchidstrada 2.0—a transcriptomics resource for the orchid family. Plant and Cell Physiology, 58, e9.

Chase, M. W., Cameron, K. M., Freudenstein, J. V., Pridgeon, A. M., Salazar, G., van den Berg, C., & Schuittema, A. (2015). An updated classification of Orchidaceae. Botanical Journal of the Linnean Society, 177, 151–174.

Chau, J. H., Rahfeldt, W. A., & Omland, R. E. (2018). Comparison of taxon-specific versus general locus sets for targeted sequence capture in plant phylogenomics. Applications in Plant Sciences, 6, 1032.

Chen, S.-P., Tian, H.-Z., Guan, Q.-X., Zhai, J.-W., Zhang, G.-Q., Chen, L.-J., Liu, Z.-J., Lan, S.-R., & Li, M.-H. (2019). Molecular systematics of Goodyerinae (Cranichideae, Orchidoideae, Orchidaceae) based on multiple nuclear and plastid regions. Molecular Phylogenetics and Evolution, 139, 106542.

Clements, M. A., Howard, C. G., & Miller, J. T. (2015). Caladenia revisited: results of molecular phylogenetic analyses of Caladenia plastid and nuclear loci. American Journal of Botany, 102, 581–597.

Clements, M. A., Jones, D. L., Sharma, I. K., Garratt, M. J., Fitzgerald, K. J., Mackenzie, A. M., Mollov, B. P. J. (2002). Phylogenetics of Diurideae (Orchidaceae) based on the internal transcribed spacer (ITS) regions of nuclear ribosomal DNA. Lindleyana, 17, 135–171.

Cozzolino, S., & Widmer, A. (2005). Orchid diversity: an evolutionary concept in plant phylogenomics. Trends in Plant Science, 10, 65–72.

Deng, H., Zhang, G. Q., Lin, M., Wang, Y., & Liu, Z. J. (2015). Mining from transcriptomes: 315 single-copy orthologous genes concatenated for the phylogenetic analyses of Orchidaceae. Ecology and Evolution, 5, 3800–3807.

Dodsworth, S., Pokorny, L., Johnson, M. G., Kim, J. T., Maurin, O., Wickett, N. J., Forest, F., & Baker, W. J. (2019). Hyb-Seq for flowering plant systematics. Trends in Plant Science, 24, 887–891.
phylogeny and phylogenetics. *Molecular Phylogenetics and Evolution*, 66, 526–538.

Mendoza, C. G., Jost, M., Hagsater, E., Magallón, S., van den Berg, C., Lemmon, E. M., Lemmon, A. R., Salazar, G. A., & Wanke, S. (2020). Target nuclear and off-target plastid hybrid enrichment data inform a range of evolutionary depths in the orchid genus *Epidendrum*. *Frontiers in Plant Science*, 10, 1761.

Miller, J. T., & Clements, M. A. (2014). Molecular phylogenetic analyses of Drakaeinae: Diurideae (Orchidaceae) based on DNA sequences of the internal transcribed spacer region. *Australian Systematic Botany*, 27, 3–22.

Minh, B. Q., Hahn, M. W., & Lanfear, R. (2020). New methods to calculate concordance factors for phylogenomic datasets. *Molecular Biology and Evolution*, 37, 2727–2733.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*, 37, 1530–1534.

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32, 266–274.

Peekall, R. (2007). Speciation in the Orchidaceae: confronting the challenges. *Molecular Ecology*, 16, 2834–2837.

Peekall, R., Ebert, D., Poldy, J., Barrow, R. A., Francke, W., Bower, C. C., & Schiestl, F. P. (2010). Pollinator specificity, floral odour chemistry and the phylogeny of Australian sexually deceptive Chiloglottis orchids: Implications for pollinator-driven speciation. *New Phytologist*, 188, 437–450.

Perez-Escobar, O. A., Bogarin, D., Schley, R., Bateman, R. M., Gerlach, G., Harpke, D., Brassac, J., Fernández-Mazuecos, M., Dodsworth, S., Hagsater, E., Blanco, M. A., Gottschling, M., & Blattner, F. R. (2020). Resolving relationships in an exceedingly young Neotropical orchid lineage using Genotyping-by- Sequencing data. *Molecular Phylogenetics and Evolution*, 144, 106672.

Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T. J., Manuel, M., Wörheide, G., & Baurain, D. (2011). Resolving difficult phylogenetic questions: Why more sequences are not enough. *PLOS Biology*, 9, 1000602.

Phillips, R. D., Reiter, N., & Peakall, R. (2020). Orchard Conservation – from theory to practice. *Annals of Botany*, 126, 345–362.

Piñeiro Fernández, L., Byers, K. J., R. Francke, W., Bower, C. C., Kellenberger, R. T., Russo, A., Qi, W., Aquino Fournier, C., & Schlüter, P. M. (2019). A phylogenomic analysis of the floral transcriptomes of sexually deceptive and rewarding European orchids, *Ophrys* and *Gymnadenia*. *Frontiers in Plant Science*, 10, 1553.

R Core Team (2019). *R: A language and environment for statistical computing*, Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Rao, X., Krom, N., Tang, Y., Widiez, T., Havkin-Frenkel, D., Belanger, F. C., Dixon, R. A., & Chen, F. (2014). A deep transcriptomic analysis of pod development in the vanilla orchid (*Vanilla planifolia*). *BMC Genomics*, 15, 1–12.

Salazar, G. A., Batista, J. A. N., Cabrera, L. I., van Den Berg, C., Whitten, W. M., Smidt, E. C., Buzatto, C. R., Singer, R. B., Gerlach, G., Jiménez-Machorro, R., Radins, J. A., Insaurralde, I. S., Guimaraes, L. R. S., de Barros, F., Tobar, F., Linares, J. L., Mújica, E., Dressler, J. L., Blanco, M. A., … Chase, M. W. (2018). Phylogenetic systematics of subtribe Spiranthinae (Orchidaceae: Orchidoideae: Cranichideae) based on nuclear and plastid DNA sequences of a nearly complete generic sample. *Botanical Journal of the Linnean Society*, 186, 273–303.

Sanderson, M. J., Nicolae, M., & McMahon, M. M. (2017). Homology-aware phylogenomics at gigabase scales. *Systematic Biology*, 66, 590–603.

Schiestl, F. P., & Schlüter, P. M. (2009). Floral isolation, specialized pollination, and pollinator behavior in orchids. *Annual Review of Entomology*, 54, 425–446.

Schwacke, R., Ponce-Soto, G. Y., Krause, K., Bolger, A. M., Arsova, B., Hallab, A., Gruden, K., Stitt, M., Bolger, M. E., & Usadel, B. (2019). MapMan4: A refined protein classification and annotation framework applicable to multi-omics data analysis. *Molecular Plant*, 12, 879–892.

Sedex, K. E. M., Qi, W., Schauer, M. A., Gupta, A. K., Poveda, L., Xu, S., Liu, Z.-J., Grossniklaus, U., Schiestl, F. P., & Schlüter, P. M. (2013). Transcriptome and proteome data reveal candidate genes for pollinator attraction in sexually deceptive orchids. *PLoS One*, 8, e46421.

Shapiro, B., Rambaut, A., & Drummond, A. J. (2005). Choosing appropriate substitution models for the phylogenetic analysis of protein- coding sequences. *Molecular Biology and Evolution*, 23, 7–9.

Simmons, M. P. (2017). Relative benefits of amino-acid, codon, degeneracy, DNA, and purine-pyrimidine character coding for phylogenetic analyses of exons. *Journal of Systematics and Evolution*, 55, 85–109.

Simo-Droissart, M., Plunkett, V. M., Edwards, M. B., Farinhö, J. M., Jecken, V., D’Haijere, T., Lowry, P. P., Sonké, B., Michelena, C., Carlsson, B. S., Azandi, L., Verlynde, S., Hardy, O. J., Martos, F., Byeberle, B., Fischer, E., & Stévar, T. (2018). New phylogenetic insights toward developing a natural generic classification of African angrecoid orchids (Vandeae, Orchidaceae). *Molecular Phylogenetics and Evolution*, 126, 241–249.

Swarts, N. D., Clements, M. A., Bower, C. C., & Miller, J. T. (2014). Defining conservation units in a complex of morphologically similar, sexually deceptive, highly endangered orchids. *Biological Conservation*, 174, 55–64.

Tomasello, S., Karbstein, K., Hoda, C., 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.

Tremblay, R. L., Ackerman, J. D., Zimmerman, J. K., & Calvo, R. N. (2005). Variation in sexual reproduction in orchids and its evolutionary consequences: a spasmodic journey to diversification. *Biological Journal of the Linnean Society*, 84, 1–54.

Unruh, S. A., McKain, M. R., Lee, Y.-I., Yukawa, T., McCormick, M. K., Shefferson, R. P., Smithsonian, A., Leebens-Mack, J. H., & Pires, J. C. (2018). Phylotranscriptomic analysis and genome evolution of the Cypripedioideae (Orchidaceae). *American Journal of Botany*, 105, 631–640.

Van der Niet, T., Peakall, R., & Johnson, S. D. (2014). Pollinator-driven ecological speciation in plants: new evidence and future perspectives. *Annals of Botany*, 113, 199–212.

Weston, P. H., Perkins, A. J., Indso, J. O., & Clements, M. A. (2014). Phylogeny of Orchidaceae Tribe Diurideae and its implications for the evolution of pollination systems. In R. Edens-Meier, & P. Bernhardt (Eds.), Darwin’s *Orchids then and now*. The University of Chicago Press.

Wong, D. C. J., Amarasinghe, R., Falara, V., Pichersky, E., & Peakall, R. (2019). Duplication and selection in β-ketoacyl-ACP synthase gene lineages in the sexually deceptive *Chiloglottis* (Orchidaceae). *Annals of Botany*, 123, 1053–1066.

Wong, D. C. J., Amarasinghe, R., Pichersky, E., & Peakall, R. (2018). Evidence for the involvement of fatty acid biosynthesis and degradation in the formation of insect sex pheromone-mimicking chiloglottines in sexually deceptive *Chiloglottis* orchids. *Frontiers in Plant Science*, 9, 839.
unique UV-B dependent floral volatile, chiloglottone 1. *Frontiers in Plant Science*, 8, 1260.

Wong, D. C. J., Pichersky, E., & Peakall, R. (2017). The biosynthesis of unusual floral volatiles and blends involved in orchid pollination by deception: current progress and future prospects. *Frontiers in Plant Science*, 8, 1955.

Xu, D., Chen, H., Acı, M., Pan, Y., Shangguan, Y., Ma, J., Li, L., Qian, G., & Wang, Q. (2018). De Novo assembly, characterization and development of EST-SSRs from *Bletilla striata* transcriptomes profiled throughout the whole growing period. *PLoS One*, 13, 205954.

Xu, H., Bohman, B., Wong, D. C. J., Rodriguez-Delgado, C., Scaffidi, A., Flematti, G. R., Phillips, R. D., Pichersky, E., & Peakall, R. (2017). Complex sexual deception in an orchid is achieved by co-opting two independent biosynthetic pathways for pollinator attraction. *Current Biology*, 27, 1867–1877.

Yang, Y. A., Moore, M. J., Brockington, S. F., Soltis, D. E., Wong, G.-S., Carpenter, E. J., Zhang, Y., Chen, L. I., Yan, Z., Xie, Y., Sage, R. F., Covshoff, S., Hibberd, J. M., Nelson, M. N., & Smith, S. A. (2015). Dissecting molecular evolution in the highly diverse plant clade Caryophyllales using transcriptome sequencing. *Molecular Biology and Evolution*, 32, 2001–2014.

Yuan, Y., Jin, X., Liu, J., Zhao, X., Zhou, J., Wang, X., Wang, D., Lai, C., Xu, W., Huang, J., Zha, L., Liu, D., Ma, X., Wang, L. I., Zhou, M., Jiang, Z., Meng, H., Peng, H., Liang, Y., ... Huang, L. (2018). The *Gastrodia elata* genome provides insights into plant adaptation to heterothermy. *Nature Communications*, 9, 1615.

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.

Zhang, G.-Q., Liu, K.-W., Li, Z., Lohaus, R., Hsiao, Y.-Y., Niu, S.-C., Wang, J.-Y., Lin, Y.-C., Xu, Q., Chen, L.-J., Yoshida, K., Fujiwara, S., Wang, Z.-W., Zhang, Y.-Q., Mitsuda, N., Wang, M., Liu, G.-H., Pecoraro, L., Huang, H.-X., ... Liu, Z.-J. (2017). The Apostasia genome and the evolution of orchids. *Nature*, 549, 379–383.

How to cite this article: Peakall R, Wong DC, Phillips RD, et al. A multitiered sequence capture strategy spanning broad evolutionary scales: Application for phylogenetic and phylogeographic studies of orchids. *Mol Ecol Resour*. 2021;21:1118–1140. https://doi.org/10.1111/1755-0998.13327