Hundreds of nuclear and plastid loci yield novel insights into orchid relationships

Oscar Alejandro Pérez-Escobar1,12, Steven Dodsworth2,12, Diego Bogarín1, Sidonie Bellot1, Juan A. Balbuena4, Rowan J. Schley1, Izai A. Kikuchi5, Sarah K. Morris1, Niroshini Epitawalage1, Robyn Cowan1, Olivier Maurin1, Alexandre Zuntini1, Tatiana Arias4, Alejandra Serna-Sánchez6, Barbara Gravendeel1, María Fernanda Torres Jimenez2, Katharina Nargar6,9, Guillaume Chomicki10, Mark W. Chase1,11,*, Ilia J. Leitch1,*, Félix Forest1,* and William J. Baker1,*

PREMISE: The inference of evolutionary relationships in the species-rich family Orchidaceae has hitherto relied heavily on plastid DNA sequences and limited taxon sampling. Previous studies have provided a robust plastid phylogenetic framework, which was used to classify orchids and investigate the drivers of orchid diversification. However, the extent to which phylogenetic inference based on the plastid genome is congruent with the nuclear genome has been only poorly assessed.

METHODS: We inferred higher-level phylogenetic relationships of orchids based on likelihood and ASTRAL analyses of 294 low-copy nuclear genes sequenced using the Angiosperms353 universal probe set for 75 species (representing 69 genera, 16 tribes, 24 subtribes) and a concatenated analysis of 78 plastid genes for 264 species (117 genera, 18 tribes, 28 subtribes). We compared phylogenetic informativeness and support for the nuclear and plastid phylogenetic hypotheses.

RESULTS: Phylogenetic inference using nuclear data sets provides well-supported orchid relationships that are highly congruent between analyses. Comparisons of nuclear gene trees and a plastid supermatrix tree showed that the trees are mostly congruent, but revealed instances of strongly supported phylogenetic incongruence in both shallow and deep time. The phylogenetic informativeness of individual Angiosperms353 genes is in general better than that of most plastid genes.

CONCLUSIONS: Our study provides the first robust nuclear phylogenomic framework for Orchidaceae and an assessment of intragenomic nuclear discordance, plastid-nuclear tree incongruence, and phylogenetic informativeness across the family. Our results also demonstrate what has long been known but rarely thoroughly documented: nuclear and plastid phylogenetic trees can contain strongly supported discordances, and this incongruence must be reconciled prior to interpretation in evolutionary studies, such as taxonomy, biogeography, and character evolution.

KEY WORDS: Angiosperms353; incongruence; multilocus phylogenetic trees; nuclear-plastid discordance; Orchidaceae; recombination.
Phylogenetic inference based on nuclear and organellar DNA sequences has revolutionized plant systematics and evolution (e.g., Cameron et al., 1999; Eiserhardt et al., 2018; Solitis et al., 2000). From species complexes (Bogarin et al., 2018; Fernández-Mazuecos et al., 2018; Pérez-Escobar et al., 2020) to families and beyond (Bateman et al., 2018; Nauheimer et al., 2018; Wan et al., 2018; Wong et al., 2020), molecular phylogenetics has radically shaped our understanding of plant evolution at widely varying scales and has driven substantial changes to classifications to better reflect monophyly (e.g., APG IV, 2016). Historically, Sanger-sequenced DNA markers have been widely used to infer phylogenies at various taxonomic levels (e.g., Baldwin, 1992; Baldwin and Markos, 1998; Cameron et al., 1999; Solitis et al., 2000). In particular, plastid DNA loci such as *rbcL*, *matK*, and the *trnL-trnF* region, the nuclear ribosomal internal transcribed spacers (nrITS), and the low-copy nuclear gene *Xdh* have been used frequently (e.g., Chase et al., 1993; Baldwin et al., 1995; Solitis et al., 1999; Hilu et al., 2003; Görniak et al., 2010; Schley et al., 2018), reflecting the relative ease of sequencing these DNA regions, particularly those present in high copy numbers in the cell and genome (plastid regions and nrITS, but see Görniak et al. 2010).

Recently, implementation of high-throughput sequencing methods has expanded the number of DNA regions available for phylogenetic inference, but there has been a continuing focus on plastid genes (Edger et al., 2018; Li et al., 2019b) and plastomes (Ross et al., 2015; Vargas et al., 2017; Schley et al., 2020; Renner et al., 2021). The extent to which phylogenetic incongruence occurs in orchids relies on the universal Angiosperms353 target capture probe set (Johnson et al., 2019), which is increasingly being used in phylogenomic studies of flowering plants (Baker et al., 2021). We sampled 75 species representing all five recognized subfamilies and the majority of the tribes (Chase et al., 2015). To assess nuclear-plastid topological conflict, we also produced a plastid phylogenomic tree inferred from the published sequences of 78 genes for 264 species, also representing all the same sets of higher taxa. Lastly, we compared the phylogenetic informativeness of these nuclear and plastid loci. We addressed the following topics: (1) the extent to which the orchid plastid tree is congruent with that of the nuclear genome and (2) how well nuclear genes perform in recovering strongly supported relationships compared to plastid genes.

**MATERIALS AND METHODS**

**Taxon sampling**

We included 75 orchid species representing 69 genera, 24 sub-tribes (of 46 sensu Chase et al., 2015), 16 tribes (of 21), and all five subfamilies. In addition, 14 species from non-orchid monocot families were included as outgroups (Appendix S1). Newly generated nuclear DNA data were produced from 62 vouchered accessions stored in the DNA and tissue bank of the Royal Botanic Gardens, Kew (https://dna-bank.science.kew.org/homepage.html). These DNA samples had been previously extracted from silica-dried leaves using a modified CTAB method (Doyle and Doyle, 1987). Each sample has a voucher herbarium specimen in the Kew Herbarium (K; Appendix S1). Nuclear short illumina sequencing reads were data-mined from the Sequence Read Archive (SRA) using the fastq-dump software of the SRATool-kit package (available at https://ncbi.github.io/sra-tools/install_config.html) and the 1KP data repository (Wong et al., 2020; https://sites.google.com/a/ualbrrta.ca/onekp/) to expand our taxon sampling with 13 additional orchid species.
Library preparation, targeted enrichment, and sequencing

The quality of the DNA extractions, including concentration and distribution of fragment lengths, were checked using a Qubit 3.0 Fluorometer (Life Technologies, Carlsbad, CA, USA) and a TapeStation 42000 system (Agilent Technologies, Santa Clara, CA, USA). Genomic library preparation and enrichment were conducted following the protocols of Johnson et al. (2019). Dual-indexed Illumina genomic libraries were prepared for each DNA sample using an insert size of ~350 bp and the Ultra II Library Prep Kit (New England BioLabs, Ipswich, MA, USA) following the manufacturer’s protocol. We used the Angiosperms353 probe set to enrich each genomic library in 353 low-copy nuclear genes (Johnson et al., 2019; https://arborbiosci.com/genomics/targeted-sequencing/mybaits/mybaits-expert/mybaits-expert-angiosperm-s-353/). Sixty-two genomic libraries were pooled in equimolar quantities to make a 1 μg (total DNA) hybridization reaction, which was subsequently cleaned and sequenced on an Illumina MiSeq v3 (600 cycles, 300-bp paired-end reads) at the Royal Botanic Gardens, Kew to produce ~14.8 Gbp (~30 million paired-end reads). Raw sequence data generated in this study are accessible via the European Nucleotide Archive (https://www.ebi.ac.uk/ena/browser/home) under project PRJEB35285 and via the Kew Tree of Life Explorer (https://treeoflife.kew.org/; Baker et al., 2021).

Plastome phylogenomics

To assess the performance of the Angiosperms353 nuclear loci for resolving orchid relationships and investigating nuclear/plastid gene tree discordance at different taxonomic levels, we utilized the 78-plastid-gene data set of Serna-Sánchez et al. (2021; available at https://doi.org/10.6084/m9.figshare.14068892) for 264 species representing 117 genera, 28 subtribes, and 18 tribes. The taxon sampling of nuclear and plastid datasets overlaps for 34 genera, 14 tribes, and 22 subtribes (Appendix S1). Detailed information on the completeness of this plastid DNA data set is provided in Serna-Sánchez et al. (2021).

DNA sequence data analysis and phylogenomic inference

The quality of the newly generated sequencing data was assessed using the FastQC software (available at https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Paired-end DNA sequencing reads were adapter-trimmed and quality-filtered with the pipeline TrimGalore! v.0.6.5 (available at https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) using a Phred score quality threshold of 30 (flag = q), a minimum read length value of 20 (flag = -length), and only read pairs that passed all quality-filtering thresholds. Data obtained from the SRA were already adapter- and quality-filtered. For each sample, the Angiosperms353 loci were retrieved using the HybPiper v.1.3.1 pipeline (Johnson et al., 2016) by mapping the clean reads against template sequences of the 353 low copy nuclear genes (available at https://github.com/mossmates/Angiosperms353) using the Burrows-Wheeler Alignment (BWA) program v.0.7 (Li and Durbin, 2009) and then de novo assembling mapped reads for each gene separately using the software SPAdes v.3.13 (Bankevich et al., 2012), with a minimum coverage threshold of 8x. Given that our data set included samples produced by the 1KP initiative (i.e., read sequencing data produced for CDS DNA only) and the risk of introducing noise in our phylogenetic analyses due to potential homoplasy in noncoding regions (Bellot et al., 2020), we decided to harvest and utilize exonic regions only for all our phylogenetic analyses to reduce the proportion of missing data in individual gene and supermatrix data sets. For each gene, homologous sequences from each species were combined and aligned with the software MAFFT v. 7.4 (Katoh and Standley, 2013) using the FFT-NS-1 strategy. Gene alignments were trimmed for spurious sequences (flags = -resoverlap 0.75 and -seqoverlap 0.90) and positions with 90% missing data (flag = -gt 0.90) in the software trimAL v.1.2 (Capella-Gutierrez et al., 2009).

For each nuclear or plastid gene alignment, a maximum likelihood (ML) gene tree was computed using RAxML v8.0 (Stamatakis, 2014) with the GTR+GAMMA nucleotide substitution model, and branch support was estimated using 500 rapid bootstrap replicates (flags = -x and - # 500). The same ML approach was used to estimate two species trees based on a supermatrix made of the 353 concatenated nuclear gene alignments and a supermatrix of the 78 concatenated plastid gene alignments. We treated both concatenated matrices as a single partition and assigned the GTR+GAMMA substitution model. To account for possible topological incongruence between nuclear gene trees, we also inferred a species tree by analyzing the nuclear gene trees together under the multispecies-coalescent (MSC) framework implemented in the software ASTRAL-III v5.6 (Zhang et al., 2018). Given their nonrecombining nature, plastid genomes are often regarded as a single frozen linkage group, yet trees inferred from different plastid genes might still produce incongruent topologies (e.g., Walker et al., 2019; Zhang et al., 2020). Such intragenomic conflict could be driven by variation in substitution rates across loci, changes in organellar inheritance, heteroplasmy, and even incomplete lineage sorting (ILS) or horizontal gene transfer events, although evidence for the last two phenomena is scant (Gonçalves et al., 2019, 2020; Straub et al., 2013). The use of coalescence-based approaches such as ASTRAL on genes with the same evolutionary trajectory is inappropriate to infer species trees, but we nevertheless analyzed individual plastid gene trees in ASTRAL as a way to visualize potential topological conflicts among these loci (see the next section, “Quantification of intragenomic and nuclear-plastid discordance”). Branches with likelihood bootstrap support (LBS) < 20 in the gene trees were first collapsed using the Newick Utilities toolkit (Junier and Zdonov, 2010), as recommended by Sayyari and Mirarab (2016). The resulting ASTRAL topologies were annotated with quartet support values (Sayyari and Mirarab, 2016) for the main topology (q1), the first alternative topology (q2), and the second alternative topology (q3; flag = -t 2). In addition, we used the software SplitsTree4 (Huson, 1998) to infer neighbor-net networks based on uncorrected p-distances calculated from the nuclear supermatrix. Neighbor-nets (also known as splits graphs; dress and Huson, 2004) are suitable diagrams to represent evolutionary relationships in groups that have experienced reticulation (Rutherford et al., 2018) and are useful to identify relationships that exhibit some ambiguity (Solís-Lemus et al., 2017).

Quantification of intragenomic and nuclear-plastid discordance

The proportion of intragenomic discordance in the nuclear and plastid data sets was evaluated by looking at the normalized quartet scores produced by ASTRAL when inferring plastid and nuclear MSC species trees. The quartet score indicates the proportion of gene tree quartets that is in agreement with the species tree, and the magnitude is inversely proportional to incongruence, where a value of 1 indicates potential absence of gene tree discordance.
To assess the degree to which the evolutionary history reflected by the nucleus tracked that of the plastid genome, we compared Euclidean distances among terminals between each bootstrap replicate of each nuclear ML gene tree and each bootstrap replicate of the plastid ML species tree. The plastid ML species tree and associated bootstrap replicates were chosen over individual plastid ML gene trees because of the high degree of intragenomic congruence observed between the topologies derived from individual plastid loci (see Results). The comparisons were conducted with the Procrustean Approach to Cophylogenetics (PACo) pipeline implemented in R (Balbuená et al., 2013). The pipeline, originally designed to investigate cophylogenetic patterns between host and parasites, assesses the similarities between any two given trees by comparing the Euclidean distances separating terminals in both via Procrustean superimposition (i.e., the rotation and scaling of ordinations derived from the organelle distance matrix to fit those produced from the nuclear distance matrix; Balbuená et al., 2013). The efficiency of PACo to assess tree incongruence between nuclear–plastid associations (i.e., any given pair of nuclear–plastid terminals) was previously evaluated by Pérez-Escobar et al. (2016), and the pipeline has been widely used in other plant groups, including Asteraceae (Vargas et al., 2017), Fabaceae (Yang et al., 2018), Orchidaceae (Pérez-Escobar, 2016; Pérez-Escobar et al., 2016), and Rosaceae (Morales-Briones, Romoleroux, et al., 2018). The pipeline provides the sum of squared residuals (i.e., the disparity between an observed and a fitted value derived from a model) for each association and each pair of topologies evaluated (Balbuená et al., 2013). This sum of squared residuals can be interpreted as a concordance score because it is directly proportional to the magnitude of the topological conflict for the pair of terminals considered.

Because extremely long branches can bias the comparison of phylogenetic distances between terminals (De Vienne et al., 2011), we conducted PACo analyses on cladograms by assigning a value of 1 to each branch length in each tree, using the function br.length in the R package ape (Paradis et al., 2004). Differences in the position of terminals between nuclear and plastid trees were summarized in barplots using the R package ggplot2 (Wickham, 2016), for which the sum of squared residuals for each pair of terminals across nuclear genes was classified in quartiles. Here, the magnitude of the discordance was assessed by the proportion of genes binned in quartiles 3 and 4 (50% and 75%) in each terminal; the more genes binned in quartiles 3 and 4, the more discordant the terminals.

To test for intergenomic conflicts occurring across different taxonomic levels, we conducted the same analysis but with trees including one representative of each genus sampled in our nuclear and plastid species trees. For this analysis, we conducted sequence alignments and ML inference on each nuclear gene alignment and on the matrix of 78 plastid genes as before (see earlier section “DNA sequence data analysis and phylogenetic inference of methods”). We then produced subtribe- and tribe-level trees by keeping one representative of each clade in our nuclear gene trees and plastid ML species tree using the approach of Matzke (2013b) as implemented in the R package BioGeoBEARS (Matzke, 2013a; script available at http://phylo.wikidot.com/example-biogeobears-scripts#pruning-a_tree). Specifically, all tips belonging to the same taxon were pruned except for the first species in the list of taxa representing each clade. Terminals identified as potentially conflicting were depicted in tanglegrams using nuclear and plastid ML species trees derived from matrices sampled to genus level as implemented in the function tanglegram of the R package dndextend (Gallí, 2015). For subtribe- and tribe-level analyses, we relied on pruned trees originally derived from the nuclear and plastid supermatrices sampled to genus level.

Assessment of phylogenetic informativeness and support for nuclear and plastid data sets

We compared the performance of the Angiosperms353 low-copy nuclear genes and the 78 plastid genes for resolving orchid relationships in terms of support and phylogenetic informativeness (PI). The latter term refers to the potential of a given locus to resolve a four-terminal polytomy at a given time depth and is estimated from the substitution rates observed at all nucleotide sites of the locus (Townsend, 2007; Bellot et al., 2020). We first calculated the proportion of branches across genus-level nuclear and plastid gene trees that fell into nine discrete LBS categories defined by an interval length of 10 (excluding the interval LBS [81–100]). Secondly, we estimated the PI of each single nuclear and plastid gene alignment with regards to all species relationships recovered in the nuclear and plastid ML species trees, respectively. The trees were made ultrametric by assigning an arbitrary age of 1 to their root and of 0 to their tips (Townsend, 2007), using the software PATHd8 (https://www2.math.su.se/PATHd8/; Britton et al., 2007). Phylogenetic informativeness of the nuclear and plastid genes was computed in PhyDesign (López-Giraldez and Townsend, 2011; http://phydesign.townsend.yale.edu/) using the HyPhy algorithm recommended for DNA sequences (Kosakovsky Pond et al., 2005) and the ultrametric ML species trees and nuclear and plastid supermatrices with gene partition information as input.

RESULTS

Nuclear phylogenomics of Orchidaceae

Enrichment success of the target genes ranged from 5% (Neotia nidus-avis) to 87% (Dendrobium ellipsocephylum) gene recovery for samples sequenced in this study (Appendices S2, S3). The proportion of nuclear genes recovered from SRA and 1KP data-mined accessions ranged from 16% (Phalaenopsis equestris) to 75% (Muxipedium xeroxphyticum and Paphiopedilum malipoense). After excluding samples with fewer than 15 genes retrieved and gene alignments including fewer than 20 sequences (i.e., ~80% missing data), the final nuclear data set consisted of 294 genes and 89 species (Appendix S3). Splits graphs derived from the nuclear supermatrix revealed clear clustering between members of each orchid subfamily, tribe, and subtribe (Fig. 1). However, uncertainty regarding the phylogenetic placement of Neottieae, Nervilieae, and Xerorchideae representatives was reflected in an increased number of alternative splits connecting these groups to representatives of other subfamilies. Maximum likelihood inference of the nuclear supermatrix and MSC analyses converged on similar, strongly supported topologies (Fig. 2; Appendix S4). However, we found important differences between these analyses regarding the placement of Gastrodiaeae (represented by the mycoheterotrophic Gastrodia elata), which was placed as (Neottieae (Gastrodiaeae/other epidendroids))) in the coalescent tree, but as (Neottieae (Xerorchideae (Gastrodiaeae/other epidendroids))) in the ML results.

An overview of the quartet support obtained across the nuclear MSC indicated that for most branches between 40–96% of the gene trees agreed with the species tree topology (Fig. 2). For a few...
branches, the proportion of gene trees supporting the quartet displayed in the species tree was below 40%. This included the most recent common ancestors (MRCAs) of *Maxillaria*, *Zygopetalum*, *Stanhopea*, and *Coeliopsis* (36.27), *Bletia/Epidendreae* (34.52), and *Cymbidieae/Epidendreae* (38.76; Fig. 2). In particular, support for the *Neottia/Palmorchis* pair was flagged by ASTRAL as unreliable due to low numbers of gene trees recovered for *Neottia* (i.e., 15).

Likelihood bootstrap support (LBS) percentages for the ML tree were high, with only 15 branches displaying LBS <100, of which nine had LBS <85 (Appendix S4). The majority of branches with low support were located near the base of Epidendroideae (the pink clade), in line with our findings of branches with quartet support values <40 as inferred by the MSC estimation.

Intragenomic conflict in nuclear and plastid data sets

Estimation of the proportion of gene tree quartets that agree with the species tree through normalized quartet scores indicated that intragenomic incongruence was low. Here, the proportion of gene tree quartets in agreement with the species tree was 89% for the nuclear genome and 95% for the plastid genome. The proportion of gene quartets supporting topologies other than the species tree (i.e., quartet support) revealed that the majority of branches obtained values between 40 and 95% for the main topology ($q_1$) in MSC analyses conducted on nuclear and plastid data sets (Appendix S5). Exceptions to this pattern in the nuclear data set were a few branches with quartet values supporting alternative bipartitions linked to the MRCAs of Epidendreae/Cymbidieae ($q_1 = 38.6; q_2 = 38.6; q_3 = 22.69$) and *Gastrodia* remainder of epidendroids ($q_1 = 40.3; q_2 = 45.2; q_3 = 14$). In the plastid data set, only four branches obtained quartet values robustly supporting multiple quartets. Three other exceptions were in Pleurothallidinae (represented by *Dilomilis*, *Octomeria*, *Myoxanthus*, and *Specklinia*), and one represented the MRCA of Podochilieae/Collabieae plus the remainder of Epidendroideae (Appendix S5).
A similar trend was found in the proportion of gene trees supporting the species tree or alternative topologies in the nuclear data set sampled to the genus level (Appendix S6), in which the proportion of gene tree topologies conflicting with the MRCA of Epidendreae/Cymbidieae in the plastid data set was higher compared with nonconflicting terminals (Fig. 3; Appendix S5, S8). After inspecting the position of these terminals in nuclear and plastid ML and MSC trees, Cattleya, Coeloa, Calypso, and Earina appeared to be conflicting with moderate-strong (LBS 85–100) to weak-strong support (LBS 40–98) in the nuclear and plastid ML trees, respectively (Appendix S8, S9). These terminals were further linked to branches with quartet 1 (qt1) values >40 and <76 in the coalescent tree inferred from the nuclear dataset (Fig. 3, Appendix S5, S8). In the plastid dataset, however, the ASTRAL tree did not recover the same topology as ML, placing Earina and Coeloa as sister terminals to Epidendreae with qt1 > 50 (Appendix S5), Calypso as sister to Changnienia (qt1 = 60), and Cattleya as sister to Pleurothallidinae (qt1 = 73). Five other terminals identified as potentially conflicting (i.e., Angracrum, Catasetum, Eulophia, Phalaenopsis, and Vanda) were found in discordant positions with strong support in nuclear and plastid ML and coalescent trees (LBS > 88; qt1 > 48 < 76; Fig. 3; Appendices S5, S8, S9).

Comparisons of nuclear and plastid trees (subtribe level) revealed 13 terminals were potentially conflicting (Appendix S10). As in the genus level analysis, the squared residual values of terminals computed for each nuclear tree were overall higher for these taxa than the remainder of the tips (Appendix S10). However, only the placements of Catasetinae, Eulophiniae, Angracriniae, Aeridinae, Tropidieae, and Nervilieae were found to be conflicting with strong support in both nuclear and plastid ML and coalescent trees (Fig. 3; Appendix S11). Lastly, incongruence assessments conducted between trees sampled to tribe level revealed that Cymbidieae, Epidendreae, Nervilieae, Tropidieae, and Vandaeae were potentially conflicting (Appendix S12). The phylogenetic positions of Cymbidieae, Epidendreae, Nervilieae, Tropidieae, and Vandaeae were found to be conflicting with moderate to strong support (Fig. 3, Appendix S12, S13) in nuclear and plastid ML trees. Quartet support for the MRCA of Epidendreae/Cymbidieae in the nuclear dataset revealed almost equal support for three alternative bipartitions (qt1 = 38.6; qt2 = 38.6; qt3 = 22.69; Appendix S5).

**Phylogenetic informativeness and support of plastid and nuclear relationships**

Profiles of plastid phylogenetic informativeness (PI) for the nuclear and plastid data sets are shown in Appendix S14. Net and per-site PIs were in general higher in nuclear than plastid data sets, with average values of 82 and 0.192 versus 56 and 0.058 for nuclear and plastid data sets, respectively. The highest net PI values in nuclear alignments were attained between 0.2 and 0.6 in an arbitrary time scale of t of 0 (tips) to 1 (root; see Materials and Methods), broadly corresponding to the initial diversification of Vanilloideae, Orchidoideae, and Epidendroideae. In contrast, the highest per-site PI occurred at t(0.4–0.1), coinciding with the relative divergence times of most generic clades. Net PI values of plastid data sets overall attained uniform values from root to tips, with a notable decrease between t(0.4–0] and t(0.2–0] with plastid ycf1 as an exception to this pattern (Appendix S14). Per-site PI of plastid data sets presented a broadly similar distribution pattern to the nuclear PI per-site values, attaining their highest values at relative time intervals of t(0.4–0.01).

The distribution of LBS across gene tree branches strongly contrasted between nuclear and plastid data sets (Appendix S4). Here, a similar distribution pattern of LBS across the interval t(0 > 10 ≤ 20) and t(0 > 70 ≤ 80) was observed in nuclear and plastid data sets, with t(10 (LBS ≤ 10) and t(100 (LBS > 80) scoring the largest values. However, the number of gene tree branches receiving LBS > 80 (t100) in the nuclear data set was 1600 (vs. 800 in the plastid data set), whereas the interval t10 scored 1400 branches (vs. 3100 in the plastid data set).

**DISCUSSION**

**Limitations of plastid-only analyses**

With few exceptions (Bateman et al., 2018; Bogarin et al., 2018; Unruh et al., 2018; Brandrud et al., 2020; Pérez-Escobar et al., 2020), previous orchid studies (Cameron et al., 1999; Salazar et al., 2003; Neubig et al., 2012; Givnish et al., 2015; Y. Li et al., 2019; Serna-Sánchez et al., 2021) have relied on plastid data sets, although some have employed the low-copy nuclear gene Xdh (Górniak et al., 2010) and the nuclear ribosomal ITS (nrITS) region (Freudenstein and
Chase, 2015). Historically, nuclear genes have been difficult to sequence due to a combination of factors including insufficient amplification (due to degradation of DNA samples and/or large intronic regions), the need to clone amplified products due to paralogy or allelic diversity, the difficulty of sorting out paralogous sequences from orthologous ones, and the use of universal polymerase chain reaction (PCR) primers (Bailey et al., 2003; Feliner and Rosello, 2007; Sramkó et al., 2014).

High-throughput sequencing has revolutionized our ability to sequence DNA regions at a genomic scale and thus promises to enhance our understanding of phylogenetic relationships by utilizing various approaches to gain more sequence data per taxon sampled. One of the most commonly used methods is genome skimming (Straub et al., 2012; Dodsworth, 2015), which focuses on sequencing high-copy genomic partitions that are present even in low-coverage genomic sequencing. Most notably, such regions include the plastid genome, which has been used extensively for phylogenetics. The ease of this method has meant that it has surged in popularity over recent years in many organisms, including monocots (Givnish et al., 2018) and orchids specifically (Parks et al., 2009; Edger et al., 2018; H. Li et al., 2019a; Kim et al., 2020; Zavala-Páez et al., 2020). Genome skimming thus has perpetuated the bias toward the use of plastid markers in phylogenetic studies of orchids, but at the same time has improved our understanding of their relationships. Nevertheless, it has long been acknowledged that without an assessment of nuclear genes, it would be challenging to evaluate a number of important biological phenomena that have shaped angiosperm evolution and diversification, including hybridization and polyploidization, population structure, gene flow, and introgression (e.g., Rieseberg et al., 1996; Vargas et al., 2017; Schley et al., 2020; Pérez-Escobar et al., 2021). Given the current reliance on plastid analyses for classification/taxonomy and for studying diversification rates, biogeography, and trait evolution, it is important to ask how well these reflect organismal evolution. Considering a nuclear framework to investigate characters across time will refine and may also radically affect our understanding of the mode and tempo of their evolution. An epiphenomenon example of this phenomenon involves the gains and losses of deceit pollination, a trait thought to be linked with increased speciation and extinction rates in orchids (Givnish et al., 2015). Using a plastid tree of the orchid subtribes as a framework, deceit pollination appeared to have evolved in Laeliineae and Pleurothallidinae independently (Givnish et al., 2015). However, if the same character is optimized on our nuclear tree (Fig. 2), deceit pollination would be inferred to have evolved in the MRCA of these two subtribes; this result could admittedly be a reflection of the sparse taxonomic sampling of our study, but it nevertheless illustrates the problem.

Are nuclear and plastid evolutionary histories broadly congruent in orchids?

Our comparative phylogenomic analyses provide a solid evolutionary framework for the orchid family inferred from hundreds of low-copy nuclear genes and a detailed assessment of how much these relationships depart from those previously estimated with plastid DNA. An overview of our current understanding of the phylogenetic relationships of orchids was produced by Chase et al. (2015). The topology provided in this review is highly congruent with studies on entire coding plastid and mitochondrial genomes (Givnish et al., 2015; Li et al., 2019b; Serna-Sánchez et al., 2021) and with that of the low-copy nuclear gene Xdh (Görniak et al., 2010).

Nevertheless, the topological test for quantification of incongruence conducted here reveals that the positions of several Epidendroideae groups are potentially in conflict. One particular case is the incongruence of Coelia and Earina, which fall into moderately well-supported conflicting positions in nuclear ML and plastid analyses (Appendix S5). The plastid tree of Givnish et al. (2015) placed these genera as sister (LBS 90), a pattern also recovered in our ML plastid tree, albeit with lower support (LBS 40). The nuclear MSC analysis, in contrast, robustly places both genera as successively sister to the rest of Epidendreae, but with Earina recovering support for an alternative position supported by 25% of the gene quartets (vs. 65% of gene quartets supporting the species tree depicted in Appendix S5). Earina is often used for calibration in molecular clock analyses, as it is one of the three orchid macrofossils that has been unambiguously identified (Conran et al., 2009), and thus, its correct placement is important for estimating ages within the orchid tree of life.

Another notable exception to the general plastid–nuclear congruence includes the inter-relationships of Epidendreae, Cymbidieae, and Vandeae. These three clades account for nearly a third of the known orchid species diversity worldwide and are thought to be derived from multiple rapid diversifications (Givnish et al., 2015; Pérez-Escobar et al., 2017a). Previous plastid phylogenomic studies and our own analyses place Epidendreae as sister to Cymbidieae/Vandeae with maximum (Givnish et al., 2015; Y. Li et al., 2019b) to moderate support by (Fig. 3, Appendices S5, S12, S13), respectively. In contrast, our nuclear tree places Vandeae as sister to Cymbidieae/Epidendreae with strong LBS (Fig. 3; Appendix S13). However, we must be cautious about the extent to which the entire nuclear genome agrees with this topology. The quartet support generated for the MRCA of Cymbidieae and Epidendreae indicates that an equal proportion of gene quartets support this and an alternative topology (Fig. 3, Appendix S5), which could suggest that nuclear gene tree discordance might be responsible for an equally plausible arrangement, i.e., (Epidendreae(Cymbidieae/ Vandeae)). Biological phenomena responsible for gene tree incongruence include ILS and hybridization/reticulation (e.g., Rieseberg et al., 1996; van der Niet and Peter Linder, 2008; Joly, 2011; Schley et al., 2020). Teasing apart the signal that these phenomena leave on
a topology is methodologically challenging (e.g., Morales-Briones et al., 2018a, 2018b). Given the overall low proportion of genes supporting the main and alternative topologies (Appendices S5, S6), we suggest that increasing the number of genes representing this particular branch could help determine whether there is a dominant branching pattern overall, as well as the relative influence of ILS and reticulation (Nute et al., 2018).

**Potential of conserved low-copy nuclear genes for orchid phylogenetics**

Previously, it has been difficult to sequence more than a handful of low-copy nuclear genes in any plant group (Dodsworth et al., 2018), but genome complexity-reduction methods such as target enrichment have proven to be efficient for sequencing many nuclear loci (Bogarin et al., 2018; Brewer et al., 2019; Dodsworth et al., 2019; Johnson et al., 2019). This approach enables the capture of hundreds of nuclear genes simultaneously, and these loci are highly variable in their levels of conservation, making them useful as a set across all phylogenetic levels.

Our comparative analyses conducted on the per-gene informativeness and support offered by plastid coding data sets and the Angiosperms353 nuclear bait kit (Johnson et al., 2019) strongly point toward the higher performance of low-copy nuclear genes for resolving phylogenetic relationships in both shallow and deep time. These include the recent divergence of epidendroids and clades that experienced rapid radiations, such as Epidendreae/Laeliinae and Arethuseae/remainder of the epidendroids. Groups that remain problematic are those located toward the epidendroid MRCA and those that have been historically difficult to place, including Gastrodieae, Neottieae, and Nervilieae (Chase et al., 2015), which also proved difficult to sequence in our study (gene recovery was poor; Appendix S2). Finding a robust result for such clades will require an increase in their taxonomic representation as well as the number of genes included.

**CONCLUSIONS**

We present the first large-scale assessment of the congruence of nuclear and plastid evolutionary histories of the orchid family and provide a generally robust nuclear phylogenomic framework for the family. Comparative analyses of the performance of hundreds of low-copy nuclear genes versus plastid genes reliably demonstrated that the Angiosperms353 genes, in general, are more informative and resolve more relationships with higher support as expected given the usually slower substitution rates of individual plastid genes. We also discovered that, although the plastid genome largely tracks the evolution of the orchid family reflected by the similarity of its phylogenetic reconstruction to that produced by our analysis of the nuclear genome, there are several instances of incongruence at varying taxonomic levels that require further study. These incongruences are a clear indication that in spite of the overall congruence between nuclear and plastid data, they are not interchangeable, particularly when it comes to the study of character and trait evolution through time. Both approaches provide insights into orchid phylogeny, and the task before us is not to eliminate one as “flawed”, but rather to seek to integrate them to provide the best possible inferences about the evolution of this vast family. Our study also highlights the potential benefits of nuclear data sets for assessing the influence of hybridization and incomplete lineage sorting on patterns of diversification. Here, we found that nuclear gene tree discordance is limited, but nonetheless exists, and will be important to resolve in key groups for understanding the diversification of the orchids and the timing of this. We predict that our study will lead to more in-depth studies of the extent of orchid topological discordance both within the nuclear genome and between genomic compartments—and the phenomena driving this incongruence. By providing a well-supported nuclear phylogenomic backbone for most orchid tribes and subtribes, our study paves the way for future research to produce a densely sampled and strongly supported orchid family tree, building on and merging with decades of work that produced phylogenetic trees derived from Sanger-sequencing data with our phylogenomic framework.

**ACKNOWLEDGMENTS**

This work was supported by grants from the Calleva Foundation and the Sackler Trust to the Plant and Fungal Trees of Life project at the Royal Botanic Gardens, Kew. The Associate Editor (Sean Graham) and two anonymous reviewers provided excellent feedback that helped increasing the quality of our study. Sebastian Vieira kindly provided a photograph of *Maxillaria pereziana* (Rodr.-Mart. & M. A. Blanco) Schuit. & M. W. Chase, displayed on Figure 1. O.A.P.E. acknowledges support from the Sainsbury Orchid Fellowship at the Royal Botanic Gardens, Kew and the Swiss Orchid Foundation. O.A.P.E. dedicates this study in loving memory of Professor Jochen Heinrichs (14 March 1969–22 April 2018), a bryologist by profession (yet orchid lover at heart) and a source of tremendous inspiration to O.A.P.E.

**AUTHOR CONTRIBUTIONS**


**DATA AVAILABILITY**

Raw sequence data generated in this study are accessible via the European Nucleotide Archive (https://www.ebi.ac.uk/ena/browser/home) under project PRJEB35285 and via the Kew Tree of Life Explorer (https://treeoflife.kew.org/). High-resolution figures presented in the supplementary material, together with the raw per-site and net PI values are freely available at https://doi.org/10.6084/m9.figshare.14287538.

**SUPPORTING INFORMATION**

Additional Supporting Information may be found online in the supporting information tab for this article.

**APPENDIX S1.** Species names, taxonomic and voucher information (including herbarium specimens whenever available) for material used in this study. The number of Angiosperms353 low-copy
nuclear genes sampled in phylogenetic estimations and missing data per sample are also provided.

APPENDIX S2. Target enrichment success of de novo sequenced accessions. Heatmap denoting the proportion of the gene sequence length recovered for each sequenced accession.

APPENDIX S3. The number of reads produced for de novo sequenced accessions. The number of reads mapped and proportion of target gene sequenced are also provided.

APPENDIX S4. Maximum likelihood tree derived from a supermatrix of 292 low-copy nuclear genes. Likelihood bootstrap support (LBS) values is 100 unless shown otherwise (LBS <85 in red). (A) Number of branches per LBS interval derived from 292 ML nuclear gene trees. (B) Number of branches per LBS interval derived from 78 ML plastid gene trees.

APPENDIX S5. Multispecies coalescent (MSC) trees inferred from 292 ML nuclear gene trees (A) and 78 ML plastid gene trees (B). Pie diagrams at nodes represent quartet support, with q1 (deep blue portion) representing the proportion of gene tree quartets that support the depicted branch, q2 (blue portion) representing the proportion of quartets supporting the first alternative, and q3 (grey portion) representing the proportion of quartets supporting the second alternative. Tribes follow the classification of Chase et al. (2015).

APPENDIX S6. MSC trees inferred from 292 ML nuclear gene trees. Pie diagrams at nodes represent the proportion of gene trees supporting the species tree. Here, the blue portion represents the proportion of concordant gene trees with the species tree, the green portion represents the proportion of gene trees supporting the most common alternative topology, the red portion represents the proportion of gene trees supporting any alternative topology, and the grey portion represents gene trees that are uninformative. Tribes follow the classification of Chase et al. (2015).

APPENDIX S7. Experimental design of nuclear-plastid incongruence analyses. The maximum number of terminals sampled in genus, subtribe, and tribe-level analyses of incongruence are provided together with the corresponding number of nuclear-plastid terminal associations analyzed and those deemed potentially conflicting.

APPENDIX S8. (A) Conflicting phylogenetic positions between nuclear and plastid trees, depicted on ML trees. Tip names and their corresponding connections are taxa flagged as conflicting. Terminals highlighted in red are placed with strong branch support; those in gray are weakly supported in either the nuclear or plastid tree. Pie diagrams at nodes are provided only for branches with conflicting terminals. They represent quartet support, with q1 (deep blue portion) representing the proportion of gene tree quartets that support the depicted branch, q2 (blue portion) representing the proportion of quartets supporting the first alternative and q3 (grey portion) representing the proportion of quartets supporting the second alternative. All quartet support values are provided in Appendix S5. LBS values are 100 unless shown otherwise (LBS <85 in red). (B) Proportion of genes with summary of squared residual values binned in quartiles 1–4 for each nuclear-plastid association. Terminals with elevated numbers of genes binned in quartiles 3 and 4 were deemed potentially conflicting (labeled with an asterisk). Terminal names highlighted in bold and red denote those conflicting with strong support in nuclear and plastid trees whereas terminals in bold and connected with grey lines were deemed potentially conflicting albeit with low statistical support in either phylogeny.

APPENDIX S9. (A) Maximum likelihood tree derived from a supermatrix of 292 low-copy nuclear genes and (B) 78-coding plastid genes. LBS are 100 unless shown otherwise (LBS <85 in red). The color of branches connected to internal nodes denote LBSs. Tree tips are color coded by subfamilies (pink: Epidendroideae; orange: Orchidoideae light green: Cypripedioideae; light blue: Vanilloideae; deep blue: Apostasioideae).

APPENDIX S10. (A) Conflicting phylogenetic positions between nuclear and plastid phylogenies sampled to subtribe level, depicted on trimmed ML trees derived from concatenated supermatrices sampled to genus level. Tip names and their corresponding connections are taxa flagged as conflicting. Terminals highlighted in red are placed with strong branch support; those in gray are weakly supported in either the nuclear or plastid tree. Pie diagrams at nodes are provided only for branches interacting with conflicting terminals. They represent quartet support values, with q1 (deep blue portion) representing the proportion of gene tree quartets that support the main (depicted) branch, q2 (blue portion) representing the proportion of quartets supporting the first alternative branch, and q3 (grey portion) representing the proportion of quartets supporting the second alternative branch. All quartet support values are provided in Appendix S7. Experimental design of nuclear-plastid incongruence analyses. The maximum number of terminals sampled in genus, subtribe and tribe-level analyses of incongruence are provided together with the corresponding number of nuclear-plastid terminal associations analyzed and those deemed potentially conflicting. LBS values at nodes are 100 unless shown otherwise (LBS <85 are highlighted in red). (B) Proportion of genes with summary of squared residual values binned in quartiles 1–4 for each nuclear-plastid association. Terminals with elevated numbers of genes binned in quartiles 3 and 4 were deemed potentially conflicting (labeled with a black asterisk). Terminal names highlighted in bold and red denote tips found to be conflicting with strong support in nuclear and plastid phylogenies; terminals in bold and connected with gray lines were deemed potentially conflicting albeit with low statistical support in either phylogeny.

APPENDIX S11. (A) A subtribe-level trimmed maximum likelihood tree derived from a supermatrix of 292 low-copy nuclear genes and (B) 78-coding plastid genes sampled to genus level. LBS values are 100 unless shown otherwise (LBS <85 is highlighted in red). The color of branches connected to internal nodes denote LBS values. Tree tips are color coded by subfamilies (pink: Epidendroideae, orange: Orchidoideae, light green: Cypripedioideae, light blue: Vanilloideae, deep blue: Apostasioideae).

APPENDIX S12. (A) Conflicting phylogenetic positions between nuclear and plastid trees sampled to tribe level, depicted on trimmed maximum likelihood trees derived from supermatrices sampled to genus level. Tip names and their corresponding connections are taxa flagged as conflicting. Terminals highlighted in red are placed with strong branch support; those in gray are weakly supported in either the nuclear or plastid tree. Pie diagrams at nodes are provided only for branches interacting with conflicting terminals. They represent quartet support values, with q1 (deep
Chomicki, G., L. P. R. Bidel, F. Ming, M. Coiro, X. Zhang, Y. Wang, Y. Baissac, et al. 2015. The velamen protects photosynthetic orchid roots against UV-B...


Wickham, H. 2016. ggplot2: Elegant graphics for data analysis. Springer-Verlag, NY, NY, USA.


