PeerJ Preprints NOT PEER-REVIEWED

Resolving relationships in an exceedingly young orchid lineage using Genotyping-by-sequencing data

Oscar Alejandro Pérez-Escobar [a,b,c*]; Diego Bogarín [d]; Rowan Schley [a,k]; Richard M. Bateman [a]; Günter Gerlach [e]; Dörte Harpke [f]; Jonathan Brassac [f]; Mario Fernández-Mazuecos [g]; Steven Dodsworth [h*]; Eric Hagsater [i]; Marc Gottschling [j]; Frank R. Blattner [f*]

[a] Royal Botanic Gardens Kew, Richmond TW9 3AE, UK
[b] Department of Biological and Environmental Sciences, University of Gothenburg, 405 30 Gothenburg, Sweden
[c] Gothenburg Global Biodiversity Centre, Gothenburg, Sweden
[d] Jardín Botánico Lankester, Universidad de Costa Rica, P.O. Box 302-7050, Cartago, Costa Rica
[e] Botanischer Garten München-Nymphemburg, Menzinger Str. 61, D-86 638 Munich, Germany
[f] Institute of Plant Genetics and Crop Plant Research (IPK), D-06 466 Gatersleben, Germany
[g] Department of Plant Sciences, University of Cambridge, Cambridge CB2 3AE, UK
[h] School of Life Sciences, University of Bedfordshire, Luton LU1 3JU, UK
[i] Herbarium AMO, Montañas Calizas 490, Lomas Barrilaco, Mexico
[j] Department of Biology, Systematic Botany and Mycology, GeoBio-Center, Ludwig-Maximillians University, Menzinger Str. 67, D-80 638 Munich, Germany
[k] Department of Life Sciences, Imperial College London, Silwood Park, Ascot, Berkshire, SL5 7PY, UK

* Corresponding author.
E-mail address: o.perez-escobar@kew.org; steven.dodsworth@beds.ac.uk; blattner@ipk-gatersleben.de

Abstract [max 300 words]:

Poor morphological and molecular differentiation in recently diversified lineages is a widespread phenomenon in plants. Phylogenetic relationships within such species complexes are often difficult to resolve because of the low variability in traditional molecular loci, as well as various other biological phenomena responsible for topological incongruence such as ILS and hybridization. In this study, we employ a Genotyping-by-sequencing (GBS) approach to disentangle evolutionary relationships within a species complex belonging to the Neotropical orchid genus *Cycnoches*. The complex includes seven taxa distributed in Central America and the adjacent Chocó biogeographic region, nested within a clade estimated to have first diversified in the early Quaternary. Previous phylogenies inferred from a handful of loci failed to provide support for internal relationships within the complex. Our Neighbor-net and coalescent-based analyses inferred from *ca*. 13,000 GBS loci obtained from 31 individuals belonging to six of the seven traditionally accepted *Cycnoches* species provided a robustly supported network. The resulting three main clades are corroborated by morphological traits and geographical

PeerJ Preprints NOT PEER-REVIEWED

distributions. Similarly, Maximum Likelihood (ML) inferences of concatenated GBS-loci produced results comparable with those derived from coalescence and network-based methods, albeit always with poor statistical support. The low support evident in the ML phylogeny might be attributed to the abundance of uninformative GBS loci, which can account for up to 50% of the total number of loci recovered. The phylogenomic framework provided here, as well as morphological evidence and geographical patterns, suggest that the six entities previously thought to be different species might actually represent only three distinct segregates. Our study is the first to demonstrate the utility of GBS data in phylogenomic research of a very young Neotropical plant clade (~2 Ma), and it paves the way for the study of the many other species complexes that populate the species-rich orchid family.

Keywords: American Tropics, high-throughput-sequencing, Orchidaceae, phylogenetic incongruence, rapid diversification.

1. Introduction

Species complexes are aggregates of putative species that exhibit little morphological differentiation or genetic divergence. They present a challenge to systematists and molecular biologists because their phylogenetic relationships are particularly difficult to resolve (Després et al., 2003; Escudero et al., 2014). This is particularly true when attempts to disentangle genetically the evolutionary history within such lineages are based on only a small number of loci (Taberlet et al., 2007), because the number of informative positions they provide is very limited. Moreover, past and/or ongoing biological phenomena that are known to produce discordance among gene trees (e.g. hybridization and incomplete lineage sorting [ILS]: (Fehrer et al., 2007; Pérez-Escobar et al., 2016a)) further exacerbate the difficulty of species-tree inference for species complexes (Fernández-Mazuecos et al., 2018; Salichos et al., 2014).

The recent advent of several high-throughput sequencing methods has facilitated the generation of millions of base pairs of DNA sequence data (Dodsworth, 2015). These approaches have enabled researchers to sequence hundreds or thousands of loci in parallel for multiple individuals (Hipp et al., 2014), potentially leading to a great increase in the number of informative sites with which to infer evolutionary relationships inside lineages with little divergence (e.g. Escudero et al., 2014). Among the plethora of high-throughput sequencing techniques, Genotyping-by-sequencing approaches (GBS, also referred as to Restriction-site Associated DNA sequencing, RADseq; Andrews et al. 2016) reduce the complexity of a genome by means of restriction enzymes (Andrews et al., 2016; Elshire et al., 2011). Enzymes specific to restriction cut sites digest a genome at specific lengths, thus enabling the sequencing of fragments adjoining enzyme cut sites and the generation of huge yet tractable numbers of genetic markers per individual (Hipp et al., 2014). For the particular case of GBS, methylation-sensitive restriction enzymes are employed for genome digestion, thus enabling more efficient access to low-copy regions (Elshire et al., 2011).

PeerJ Preprints NOT PEER-REVIEWED

Several studies have demonstrated the utility of reduced-representation genomic data for resolving the phylogenetic relationships of plant and animal lineages derived from both ancient (e.g. oaks: Hipp et al., 2014) and very recent rapid diversifications (e.g. cichlid fishes: Wagner et al., 2013, toadflaxes (*Linaria*): Fernández-Mazuecos et al., 2018; *Diospyros*: Paun et al., 2016). RADseq has also been successfully employed to tease apart the inter- and intraspecific relationships of a recently diversified clade of *Heliconius* butterflies (Nadeau et al., 2013). However, the extent to which such data are useful for inferring phylogenetic relationships in plant species complexes has been little studied (Anderson et al., 2017).

Orchidaceae (the orchid family) is a hyper-diverse lineage of flowering plants distributed worldwide (Chase et al., 2015), which is known for its rich and often intricate morphological diversity (Mondragón-Palomino, 2013; Gramzow and Theißen, 2010; Mondragón-Palomino and Theißen, 2008). Species complexes are common across the orchid family (e.g. Bateman et al., 2017; Gale et al., 2018; Johnson and Steiner, 1997). As such, orchids provide a good opportunity to test the utility of GBS data for resolving phylogenetic relationships among taxa with little or overlapping morphological variation. Notable among orchids with intricate reproductive morphology (Fig. 1, inset) is the Neotropical genus *Cycnoches* (Darwin, 1877), which encompasses 34 epiphytic species distributed from Southern Mexico to Bolivia and Central Brazil (Pérez-Escobar, 2016a).

Previous studies using a combination of three nuclear and two plastid loci failed to resolve the phylogenetic relationships of very young clades within *Cycnoches* (Batista et al., 2014; Gerlach and Pérez-Escobar, 2014; Pérez-Escobar et al., 2017a; Pérez-Escobar et al., 2017b). This statement is particularly true for the *C. egertonianum* species alliance, a complex consisting of seven species and two subspecies restricted to Central America and the Chocó region of South America. Their members are characterized by male flowers with narrow, elongated tepals (which can be pale green to deep purple), and by a labellum that is divided into oblong to capitate dactylar processes (Fig. 1) (Carr, 2012; Romero and Gerlach, n.d.). Ample intraspecific morphological variability exists and is most evident in tepal coloration patterns and morphology of the dactylar processes that extend radially from the labellar margin (Fig. 1). Moreover, herbarium material of *Cycnoches* is very scarce; some putative species of the *egertonianum* complex are known only from drawings of the type specimen (*C. pachydactylon*, *C. amparoanum*, *C. pauciflorum*, *C. powellii*, *C. stenodactylon*), thus effectively precluding species delimitation by means of morphology.

The available molecular phylogenies of *Cycnoches* recovered the *C. egertonianum* complex as a monophyletic group composed of two clades. Further comparison between nuclear and plastid phylogenies revealed strongly supported incongruences within the complex, thus suggesting ILS and/or that primary hybridization or introgression events have taken place (Pérez-Escobar et al., 2016b). In this study, we test the utility of GBS data for resolving the evolutionary history of the *C. egertonianum* complex by investigating six members (four species and two

subspecies; Carr 2012) using phylogenomic analyses. We specifically ask: (*i*) do all species in the *C. egertonianum* complex conform to monophyletic groups? (*ii*) are ML phylogenies congruent with trees derived from coalescent-based methods? And (*iii*) how informative are the loci produced by GBS reduced-representation genomic data for resolving phylogenetic relationships in the ~2 Myr-old *C. egertonianum* complex? Answering these questions allows us to draw more general conclusions regarding the applicability of GBS methods to recent and rapid diversifications.

2. Material and Methods

2.1. Taxon sampling, DNA extraction and Genotyping-by-sequencing (GBS) library preparation

We analysed 29 accessions together representing four species and two subspecies of the *C. egertonianum* complex, collectively sampled from four localities. Additionally, we included as outgroup single accessions of *C. barthiorum* and *C. herrenhusanum*. The number of morphotypes sampled per species and voucher information are provided in Table S1. Genomic DNA was extracted from fresh leaf tissue (preserved in silica gel) with the NucleoSpin® plant kit (Macherey-Nagel; Düren, Germany), following the manufacturer's protocol. Total genomic DNA was analysed on a 1.5% Agarose gel, and concentration and fragment length distribution were assessed relative to a DNA standard. Genomic library preparation for GBS was performed at the Institute for Plant Genetics and Crop Plant Research (IPK), Germany, following a protocol of Elshire et al. (2011) as modified by Weltman (2016), using the rare-cutting enzyme *PstIHF*® (recognition site: CTGCA'G) and the methylation-sensitive enzyme *MspI* (recognition site: C'CGG). Genomic libraries were sequenced at the IPK on an Illumina HiSeq 2000, generating single-end reads of 100 bp.

2.2. DNA sequence data analysis, SNP detection and phylogenomic inference

Illumina reads were trimmed, filtered and *de novo* assembled with the pipeline ipyRAD v.3.0 (Eaton, 2014) using the default parameters recommended for single-end read data. We produced GBS loci assemblies using a clustering value (*c*) of 0.95 and the statistical base-calling parameters of (Li et al., 2008), specifically a minimum depth coverage of six and a maximum number of five Ns in the consensus sequence. Paralogous loci were filtered by setting two alleles and eight heterozygous positions per consensus sequence. To assess whether missing data had an impact on tree topologies (Huang and Knowles, 2016), we employed three contrasting thresholds of taxon coverage (i.e. minimum number of species per locus – *s*), specifically 15%, 50% and 100% of the total number of species, each in combination with $c = 0.95$. Thus, three DNA matrices were set for phylogenomic inference: c95s15, c95s50 and c95s100.

We inferred Maximum Likelihood trees from the three DNA matrices using RAxML v.8.0 (Stamatakis, 2014) with the GTR+G substitution model and 1000 bootstrap replicates, operating via the CIPRES Science Gateway computing facility (Miller et al., 2015). We also

analysed the same DNA matrices in SplitsTree4 (Huson, 1998) to produce Neighbor-net networks (i.e. split graphs) derived from uncorrected P-distances. Split graphs are considered more suitable than phylograms or ultrametric trees to represent evolutionary histories that are still subject to reticulation (Rutherford et al., 2018).

Maximum Likelihood analyses of concatenated loci might produce biased topologies (Rokas et al., 2003), particularly in the presence of biological phenomena responsible for gene tree incongruence such as hybridization and ILS (Fernández-Mazuecos et al., 2018). Thus, to obtain a species tree from the complete multi-locus data sets while considering ILS, we performed the coalescence-based analysis SVDquartets (Chifman and Kubatko, 2014). To test the monophyly of the putative species and subspecies in *C. egertonianum* complex, we carried out analyses assigning tips to delimitations indicated by concatenated and network analyses produced in this study (i.e. with the 'taxon partition' option) and also without the 'taxon partition' option. We executed SVDquartets in the software PAUP* v.4.0a (Swofford, 2001), and for each dataset, we evaluated 100,000 random quartets and performed 100 bootstrap replicates under the multispecies coalescent tree model.

2.3. Assessing per-locus phylogenetic informativeness

We assessed the performance of every GBS locus to resolve phylogenetic relationships in *C. egertonianum* complex by estimating the net and per-site phylogenetic informativeness (PI) across an arbitrary time scale (tips assigned to time 0 and root to 1) following the method of Townsend (2007) and using the web service PhyDesign (Townsend, 2007) (available at http://phydesign.townsend.yale.edu/). This package requires locus alignments with complete taxon sampling and an ultrametric tree. Thus, we sampled sequences of 29 individuals (one outgroup and 28 ingroup samples) from 2297 locus alignments derived from the c95s100 dataset. We concatenated the sampled alignments to form a new super matrix, and inferred an ML tree with the same settings as specified above; the tree was later converted into a chronogram with PATHd8, a program for phylogenetic dating without a molecular clock (https://www2.math.su.se/PATHd8/) (Britton et al., 2007; Schoch et al., 2009). The PI profiles were estimated with the HyPhy substitution rates algorithm for DNA sequences (Kosakovsky-Pond et al., 2005). We identified the sites with unusually high substitution rates that may cause phylogenetic noise with the R script and filtering method described by Fragoso-Martínez et al. (2016) using a cut-off value of five. The identified sites were removed manually from the alignments using the software Geneious v8.1 (Biomatters Ltd.; Kearse et al., 2012) and these corrected matrices were uploaded again to PhyDesign as described above.

The net and per-site PI of the GBS loci was compared with the performance of an Internal Transcribed Spacer (ITS) alignment for the same set of species. Here, ITS sequences for each individual were obtained by mapping trimmed reads (GBS sequence data) for each species against an nrITS reference sequence of *C. herrenhusanum* (GenBank acc. MF285490) and

producing a majority rule consensus sequence in Geneious v8.1. The ITS sequences were then aligned in MAFFT (Katoh et al., 2017) using the Q-INS-I strategy, which is optimized to consider secondary structures during the alignment process. A ML tree from the ITS alignment was inferred and then converted to an ultrametric tree, after which it was analysed in PhyDesign using the aforementioned settings.

To understand the extent to which particular GBS loci resolve phylogenetic relationships at any node with confidence, we computed (*i*) the number of nodes receiving null to maximum Likelihood Bootstrap Support (LBS) values, and (*ii*) the distribution of LBS across node depth. Here, we obtained ML trees for every locus using RAxML v.8.0 (Stamatakis, 2014) with the GTR+G substitution model and 100 bootstrap replicates. We classified LBS values at nodes into the following intervals: [0–10], [11–20], [21–30], [31–40], [41–50], [51–60], [61–70], [71–80], [81–100]. Node depth values were obtained for every locus tree by converting the ML phylogeny into an ultrametric tree with a root age of 1 using the penalized likelihood method implemented in the function *chronos* from the R package APE (Paradis et al., 2004) which is faster and easier to parallelize than PATHd8 for analyses of large numbers of trees. Here, we set a root age of 1, following the approach of Lee et al. (2018), and used a lambda value of 0 for rate smoothing. Node number vs LBS intervals and node depth vs LBS values were plotted using the R package GGPLOT2 (Wickham, 2009).

3. Results

3.1. Genotyping-by-sequencing data recovery

Illumina sequencing produced *ca*. 66 million raw reads, bases with a quality score of Q30 for 96.3% reads and a GC content of 45.7%. The number of loci recovered under different taxon coverages and the proportion of missing data are summarised in Table 1. As expected, the highest total number of loci recovered (13,960) was achieved under the smallest taxon coverage (i.e. s = 15) whereas the smallest number of sequenced loci (2,297) was obtained under the highest taxon coverage, s = 100. These loci have almost the same length (median=102 bp, mean=100.8, SD=11.5).

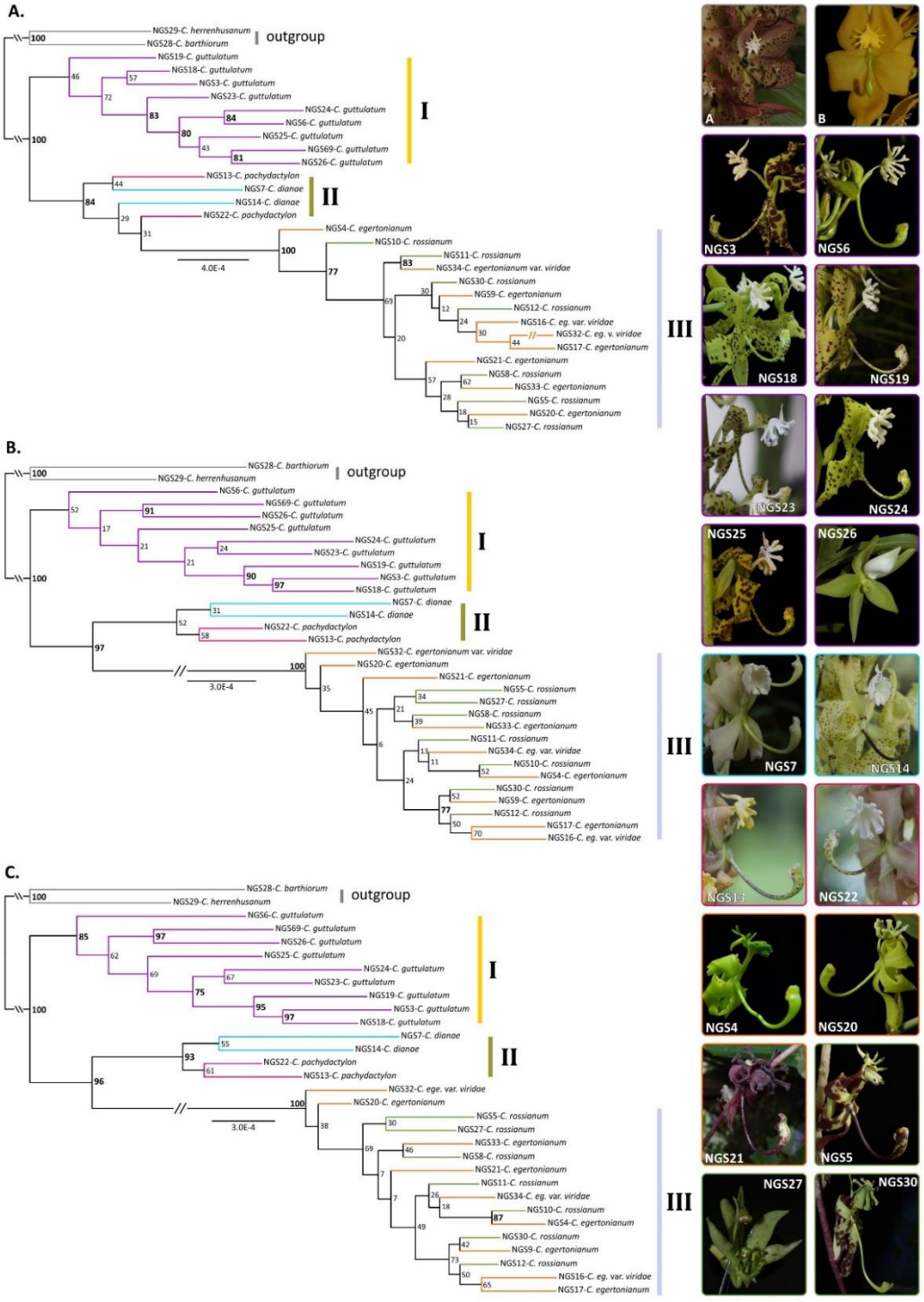
Data set	Minimum taxon coverage	Number of loci	Concatenated length	Missing data (%)
c95s15	15%	13,960	1,382,653	35.2%
s95s50	50%	9,626	949,726	17.15%
c95s100	100%	2,297	231,238	0%

Table 1
Characteristics of the three assembled GBS datasets under ipyRAD.

3.2. Phylogenomic inferences

Maximum Likelihood (ML) analyses of the c95s15, c95s50 and c95s100 concatenated DNA matrices produced similar topologies, and provided strong statistical support for the *C.*

egertonianum complex (LBS = 100; Fig. 1). They recovered three weakly to moderately supported main groups (I–III; Fig. 1). Group I included all samples identified as *C. guttulatum* (LBS = 46–85) and was placed as sister to groups II plus III. Group II clustered all specimens assigned to *C. dianae* and *C. pachydactylon* (LBS = 52–93), except in the ML tree of c95s100, which rendered the group paraphyletic (Fig. 1A). However, the topologies of c95s15 and c95s50 grouped *C. dianae* and *C. pachydactylon* samples as reciprocally monophyletic, albeit with weak statistical support (LBS = 31–55, and 58–61, respectively). Group III included samples of *C. rossianum* and the two subspecies of *C. egertonianum* (LBS = 100), all intermingled across the clade rather than clustering according to traditional taxonomy.



PeerJ Preprints NOT PEER-REVIEWED

Fig. 1. Best scoring Maximum Likelihood trees inferred under contrasting proportions of missing data: (A) c95s100 (i.e. no missing data); (B) c95s50 (minimum number of 50% of species per locus included); (C) c95s15 (minimum number of 15% of species per locus included). Likelihood Bootstrap Support (LBS) values > 75 are displayed in boldface. Branches are colour coded according to taxonomic identities. Clades I–III are indicated by vertical bars. (Inset) Pictures of individuals sampled, colour coded according to taxonomic identity. Left: *Cycnoches barthiorum*; Right: *C. herrenhusanum*.

Uncorrected P-distance split networks of all DNA matrices recovered the same groups (i.e. clades I–III; Fig. 2) produced by the ML phylogenies and revealed a clear distinction between each cluster. Here, groups I and II were distanced from each other by very short splits, whereas group III was placed more distantly from both groups I plus II and the outgroup, a topology consistent with the ML trees. The bootstrap support (B) of splits was proportional to the amount of missing data; the c95s100 network contained nine splits with support values exceeding 50, compared with 18 such splits obtained in the network derived from c95s15. The splits leading to group III from groups I plus II were in all instances strongly supported (B = 100), while splits joining groups I and II were only moderately supported (i.e. B = 85, 88) in the network derived from c95s15 (Fig. 2C). None of the split networks produced clear clustering patterns between samples of the same named taxon within either group II or group III.

3.3. Coalescence-based phylogenies

The SVDquartet analyses produced backbone topologies very similar to those derived from ML analyses (Figs. S1, S2). Inferences with and without the taxon partition option produced similar phylogenies in both topology and bootstrap support values across the different taxon coverages (Fig. 3), and strongly supported monophyly of the *C. egertonianum* complex. All analyses without the taxon partition recovered all accessions of *C. guttulatum* as a well-supported monophyletic group (I in Figs. S1, S2) and placed it as sister to the rest of the alliance. The samples of *C. dianae* were revealed as a paraphyletic group, and one of them (NGS14) formed a moderately supported clade (B = 81–88) together with the two accessions of *C. pachydactylon* (group II). The accessions of *C. rossianum* and the two subspecies of *C. egertonianum* were intermingled in a strongly supported clade (together constituting group III; B = 100). No clustering pattern between samples of the same species was evident within this clade.

The trees produced using taxon partitions converged on the same topology derived from analyses without taxon partitions, albeit with some differences in bootstrap support values (Fig. 3). Here, *C. guttulatum* was recovered as sister to rest of the species alliance (group I). *Cycnoches dianae* and *C. pachydactylon* were placed as sister lineages in a moderate-to-strongly supported clade (group II, B = 84–94). Group II was in turn recovered as sister to the *C. egertonianum* + *C. rossianum* clade (group III) with high statistical support (B = 99–100). Group III received similarly strong statistical support (B = 100).

PeerJ Preprints NOT PEER-REVIEWED

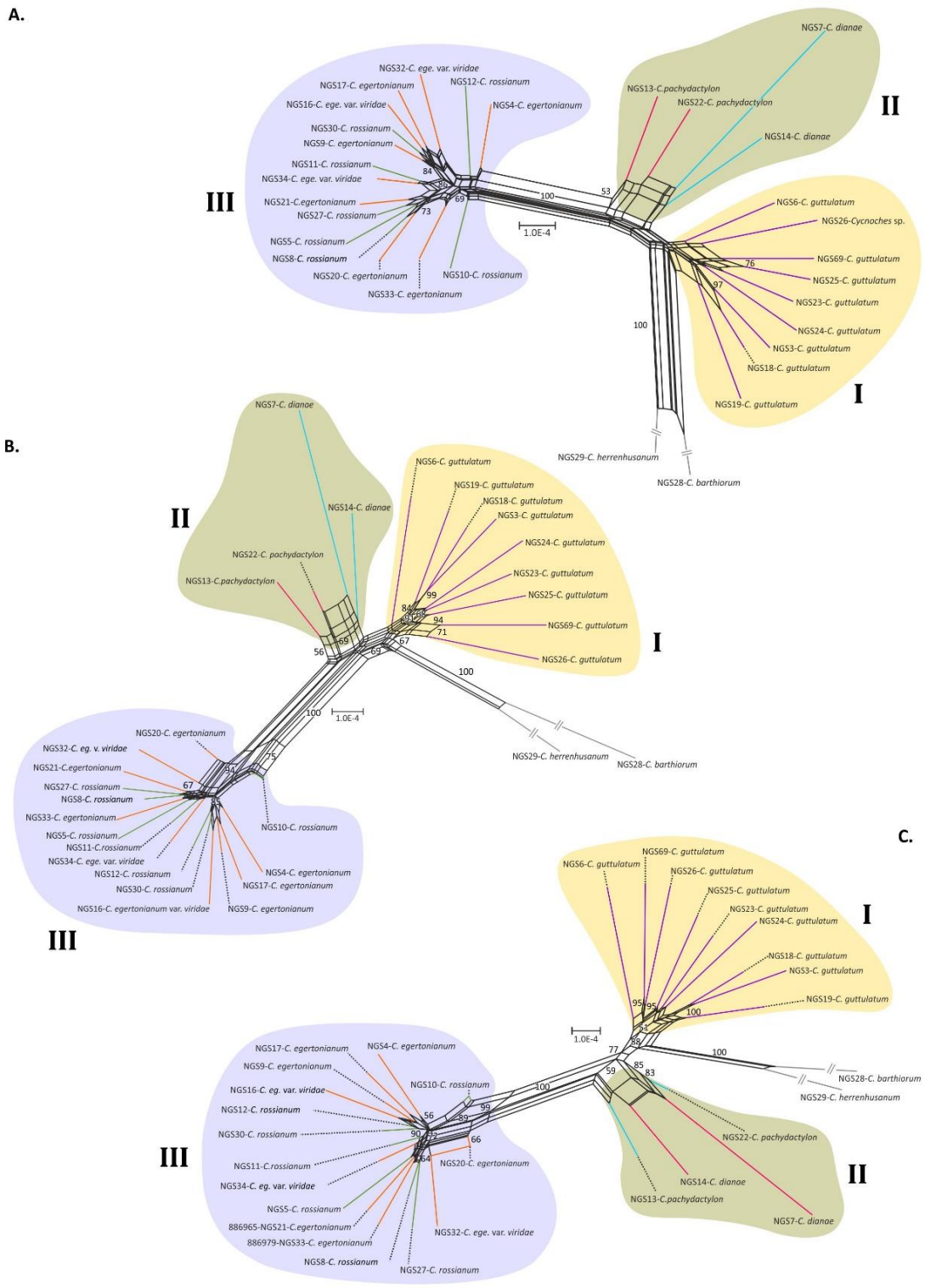


Fig. 2. Uncorrected P-distance split networks under different proportions of missing data: (A) c95s100 (i.e. no missing data); (B) c95s50; (C) c95s15. Bootstrap support values (B) > 50 at splits are shown. Main clusters are colour coded according to main clades (I–III) shown in Fig. 1.

PeerJ Preprints NOT PEER-REVIEWED

3.4. Phylogenetic informativeness and node support of GBS loci

The net phylogenetic informativeness (PI) analyses of the GBS loci generated a total of 231,997 observations. The per-site and net PI values for all loci varied dramatically, ranging from 0 to 0.13, and from 0 to 135.09, respectively. Most of the loci reached their highest PI values between a time interval of 0.06–0.20, which includes the most recent common ancestors (MRCAs) of *C. egertonianum* + *C. rossianum* accessions. 1,296 out of 2,297 loci (56.4%) were completely uninformative (net and per-site PI = 0). Only 346 loci had net and per-site PI values higher than the mean and peaked at an average time of 0.3, though the most informative peaked closer to the present. Of these, two loci (locus679 and locus787) showed disproportionate increases in net and per-site PI values towards the present (between times 0.08 and 0.15). Both net and per-site PI values of the GBS loci yielded almost identical results. The mean net PI value of nrITS was ~210, and it peaked at time 0.5, near the MRCAs of *C. pachydactylon* + *C. dianae* grade + *egertonianum* + *C. rossianum* and *C. guttulatum* (Fig. 4). This locus was the most informative compared with all GBS loci for the net PI. In contrast, the maximum per-site PI value of the nrITS (~0.028) was lower than 42 GBS loci but still much higher than the average of the remaining GBS.

The distribution pattern of LBS values at nodes was virtually identical across the c95s100, c95s50 and c95s15 datasets, most of the node support values being strongly skewed towards the 0–10 LBS interval (Figs. S3–S5). Very few nodes were moderately to strongly supported across gene trees (LBS interval 81–100), and they accounted for only ~1% of the total number of nodes across gene trees in each dataset. Likewise, the distribution of LBS values across node depth was strikingly similar across the three DNA datasets (Figs. S6–S8). Here, we could detect no clear distribution pattern of LBS support at a given node depth.

4. Discussion

4.1. GBS-loci and ITS phylogenetic informativeness in the C. egertonianum *complex*

Several studies have convincingly demonstrated the utility of SNP discovery methods coupled with high-throughput sequencing to resolve rapid plant diversifications as recent as ~5 Myr old (Fernández-Mazuecos et al., 2018; Hou et al., 2015). More recently, Lee et al. (2018) showed that ddRADseq is subject to a dramatic decrease in the number of parsimony-informative SNPs towards both shallow and deeper phylogenetic levels (i.e. suffers from locus dropout). Our comprehensive analysis of LBS values across nodes in gene trees and species trees, and the phylogenetic informativeness of GBS data, at least partially follows the locus dropout behaviour previously observed by Lee et al. (2018). However, our results differ in that the lack of support and informativeness is not confined to very shallow nodes but rather appears to occur at all phylogenetic depths. Perhaps more importantly, their comparable average PI values indicate that a widely used marker such as nrITS might be almost as informative as the majority of the GBS loci recovered here. However, some shorter GBS loci may be good

candidates to improve resolution as they have higher per-site PI values than nrITS and peaked closer to the present (locus679 and locus787). The higher net PI values of nrITS are associated with the length of this locus (612 bp), which is six times the average of the GBS loci length. Thus, the concatenation of all GBS loci with positive PI values might bring considerably more phylogenetic signal to a particular dataset than would nrITS alone.

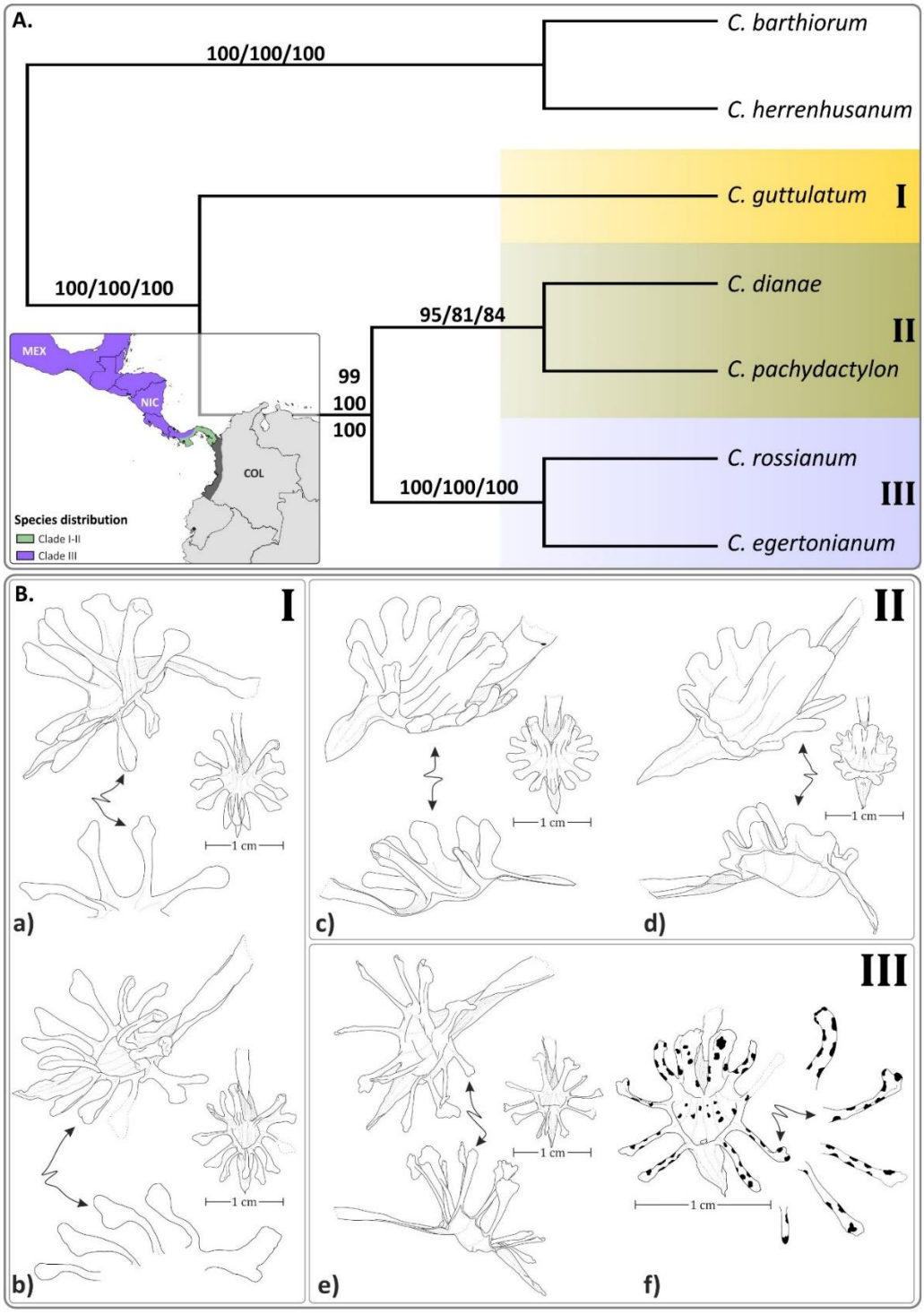


Fig. 3. (A) The 50% majority rule consensus tree derived from SVDQuartets analysis for all DNA matrices (c95s100, c95s50 and c95s15) using taxon partition. Bootstrap support values > 75 are displayed in boldface. Bootstrap values are listed in the following order: c95s100/c95s50/c95s15. Identical B values produced from both datasets are shown only once. Clades I–III are indicated with vertical bars. (Inset) Geographical distribution of terminals nested in clades I, II, and III. (B) Detailed illustrations of the labellum (lateral and ventral view, including details of the dactylar processes indicated by the black arrows) of preserved specimens of selected members of the *Cycnoches egertonianum* complex. (a) *C. guttulatum* (drawn from the Isotype Powell 14 [AMES]); (b) *C. guttulatum* (Powell 20 [AMES]); (c) *C. pachydactylon* (Gerlach 00/3415 [M]); (d) *C. dianae* (Powell 186 [AMES]); (e) *C. egertonianum* var. *viridae* (Hamer 87 [AMES]); (f) *C. rossianum* (Tuerckheim 7777 [AMES]). Illustrations by O. Pérez-Escobar extrapolated from rehydrated herbarium material or flowers preserved in spirits.

A striking result of our study is the large number of GBS loci with null PI values, which can be more than 50% of the total GBS loci produced by ipyRAD. A similar lack of resolution and phylogenetic structure have been reported within each of the nine very young clades that together constitute the temperate terrestrial orchid genus *Ophrys* (Bateman et al., 2018). Here, phylogenetic relationships within the young (< 1 Myr old) *O. sphegodes* and *O. fuciflora* complexes were rendered poorly supported in ML analysis of concatenated RAD loci. The same pattern was also recovered in a split network analysis, which revealed large phylogenetic uncertainty within these clades and some residual uncertainty between them (Bateman et al., 2018). Broadly similar results have been obtained from species complexes within the terrestrial orchid genera *Dactylorhiza*, *Gymnadenia* and *Epipactis*.

Low statistical support is often attributed to the presence of phylogenetic incongruence among gene trees (Aberer et al., 2013; Wilkinson, 1996). Alternatively, the exclusion of loci with missing data might also negatively affect the phylogenetic accuracy and the probability of recovering taxa as monophyletic (Huang and Lacey Knowles, 2016). The distribution of LBS values across nodes in trees derived from the three different matrices with different degrees of missing data (i.e. c95s100, c95s50, c95s15) is virtually identical; the majority of gene-tree nodes have very low LBS values, and low LBS values occur at all phylogenetic depths. Likewise, the phylogenies produced from the three datasets resulted in similar topologies, despite the fact that each incurred poor statistical support (Fig. 1). Here, the only notable difference between the datasets is the monophyly of group II, which is recovered as monophyletic (with weak to moderate support) in c95s50 and c95s15 (i.e. datasets with greater proportions of missing data). Thus, in the case of *Cycnoches*, the proportion of missing data did not have a notable impact on the node support of any of our phylogenetic estimations, though it did affect to some extent the topology. Additionally, given the overall low LBS support across loci trees, it is difficult to assess whether there are statistically supported phylogenetic conflicts among our loci trees.

Lack of phylogenetic signal in the loci data, or rapid diversification events, arefurther thought to produce low statistical support (Bogarín et al., 2018). Low PI profiles are derived from sites evolving at rates that are either extremely fast or extremely slow (Townsend 2007).

Even though we do not disregard the influence of the biological phenomena responsible for among-loci tree incongruence previously reported for *Cycnoches* by Pérez-Escobar et al. (2016b, 2016c), we speculate that poor statistical support in our ML phylogenies is indeed derived from the overall low PI of the GBS loci here sequenced. The low PI values and poor bootstrap support in our data indicate that perhaps more variable regions are needed to disentangle species relationships resulted from rapid radiations (López-Giráldez et al., 2013; López-Giráldez and Townsend, 2011; Townsend, 2007).

In this way, alternative reduced-representation approaches such as anchored hybrid enrichment (AHE) have recently been employed to resolve phylogenetic relationships at very shallow phylogenetic levels in both plant and animal lineages (Fragoso-Martínez et al., 2016; Granados Mendoza et al., 2013; Lemmon and Lemmon, 2012; Wanke et al., 2017). In the case of orchids, AHE did not yield loci alignments with PI values equal to 0. Therefore, they were more informative than the GBS loci sequenced here (Bogarín et al., 2018). Moreover, the AHE loci resolved intricate phylogenetic relationships with confidence in a taxonomically difficult complex of a recent linage of *Lepanthes* orchids. Thus, we suggest that alternative reduced-representation methods such as targeted enrichment might also be also suitable to disentangle phylogenetic relationships in very recent orchid diversifications. Here, lineage-specific sets of targeted sequencing probes can retrieve hundreds (~500) of orthologous known-exons from all plant genomic compartments plus their partial, contiguous intronic regions (Bogarín et al., 2018; Johnson et al., 2018). Such intronic regions certainly enhance the phylogenetic informativeness of gene alignments (~1000 bp) at shallow phylogenetic levels. The advantage of targeted enrichment methods and the well-defined regions that they target stands in stark contrast with the greatest major drawback of GBS approaches; GBS samples unknown genomic regions (Andrews et al., 2016), some of which might be under contrasting selective pressures.

4.2. Paraphyly in the C. egertonianum *complex*

Previous maximum Likelihood and Bayesian analyses of *Cycnoches* inferred from three nuclear and two plastid DNA loci have revealed two clades within the *C. egertonianum* complex, and strongly supported the previously demonstrated monophyly of the alliance. However, they failed to disentangle the internal relationships of these clades (Batista et al., 2014; Gerlach and Pérez-Escobar, 2014; Pérez-Escobar et al., 2017). The copious amount of GBS loci obtained from our multiple-accessions-per-species approach, together with comparative phylogenomic analyses, do not support the monophyly of the three species (i.e. *C. dianae*, *C. pachydactylon* and *C. rossianum*) as well as the two subspecies of *C. egertonianum*. Only the monophyly of *C. guttulatum* is (weakly to moderately) supported by all analyses. More importantly, they also consistently point to the monophyly of two species aggregates, namely *C. pachydactylon* + *C. dianae* (clade II) and *C. egertonianum* + *C. rossianum* (clade III), implying that clades II and III may in fact represent two *bona fide* species.

PeerJ Preprints NOT PEER-REVIEWED

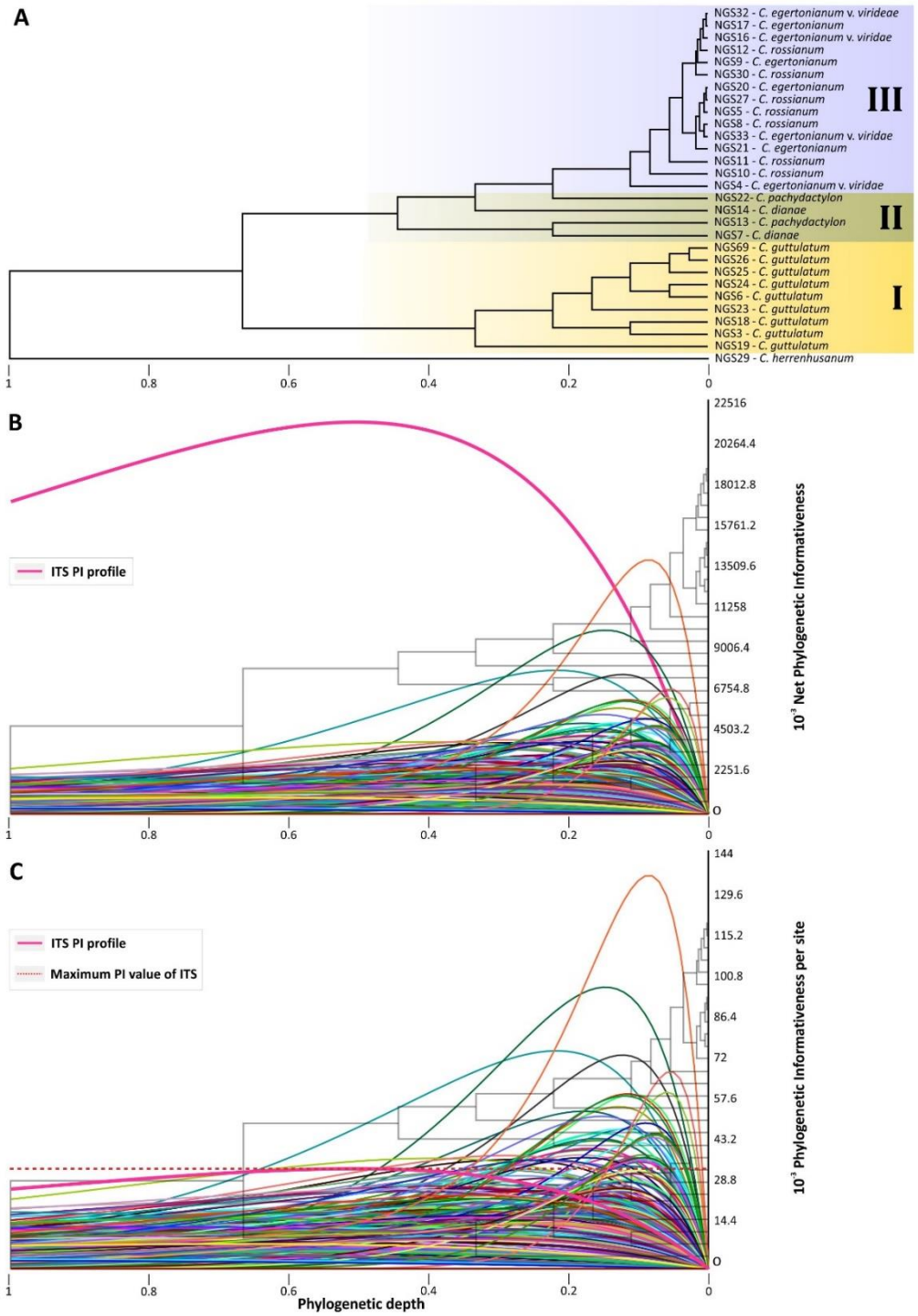


Fig. 4. (A) Ultrametric tree derived from a ML phylogeny computed from the concatenated c95s100 dataset. (B) Net phylogenetic informativeness (PI) profile per locus of the filtered c95s100 GBS loci (for details of filtering strategy see Section 2.2). (C) Per-site PI profiles of filtered c95s100 GBS loci. The dashed red line indicates the maximum net PI of the nrITS for the same phylogeny and phylogenetic depth. The maximum PI value obtained for the filtered loci alignments is indicated. Note the lower maximum PI value of the nrITS.

PeerJ Preprints NOT PEER-REVIEWED

The three clades confidently supported and recovered by all ML inferences and by coalescence-based analyses with taxon partitions also exhibit a strong morphological and geographical structure. Here, *C. guttulatum* and *C. pachydactylon* + *C. dianae* are restricted to the lowland South-eastern wet forest of Panama. *Cycnoches powellii*, another member of the *C. egertonianum* complex (only known from the type specimen, therefore not available for sampling in this study), greatly resembles *C. dianae* and occurs within the distribution range of this species. We therefore hypothesise that *C. powellii* belongs to the *C. dianae* + *C. pachydactylon* clade. *Cycnoches guttulatum* can easily be distinguished from *C. dianae* + *C. pachydactylon* by the elongated, capitate to oblong dactylar processes of the labellum (Figs. 1, 3A, B) and the green to pale yellow sepals and lateral petals bearing conspicuous macules (Carr, 2012). Members of clade II in turn have vestigial to short, rectangular dactylar processes, and the sepals and lateral petals are either pink or pale green, usually lacking macules (when present, these are inconspicuous; Figs. 1, 3C, D). The *C. egertonianum* + *C. rossianum* aggregate (clade III) is distributed from the Cordillera de Talamanca in Panama northwards to Southern Mexico. The morphology of the dactylar processes in this clade broadly resembles that of *C. guttulatum*, though in this aggregate the sepals and lateral petals are usually deep purple, plain green or green with conspicuous purple markings (Figs. 1, 3E, F).

Because phylogenetic incongruence is a pervasive phenomenon across the plant tree of life (Eiserhardt et al., 2018; Rokas et al., 2003), species delimitation based on multilocus phylogenies should always be accompanied by contrasting lines of evidence, preferably including detailed morphological analyses (Bogarín et al., 2018). Thus, whether the three clades here recovered represent three genuinely distinct entities merits further study. Even though our coalescence-based phylogenies with taxon partition reject the monophyly of almost all of the species previously recognised in the *C. egertonianum* complex, and both morphology and geographical distribution clearly delimit clades I–III, further evidence is required to circumscribe with sufficient confidence these three entities as distinct species. Our phylogenomic framework does pave the way for further integrative studies combining morphological and ecological information (e.g. plant-pollinator interactions) as well as presaging extended sampling to further investigate the genome divergence among populations of members of the *C. egertonianum* complex.

5. Conclusions

Our well-founded phylogenomic framework for the *C. egertonianum* complex suggests that six of the seven sampled taxa recognised by traditional taxonomy more likely constitute only three genuine species. Further testing through genomic, morphological and ecological studies is desirable to confirm their circumscription and most appropriate rank. Genotyping-by-sequencing reduced-representation genomic data coupled with coalescence-based inferences of relationships is a useful tool to shed light on the phylogenetic relationships in recently diversified orchid

species complexes. However, given the lack of PI in a large proportion of the GBS loci sequenced, alternative methods such as targeted enrichment – known to have effectively resolved intricate phylogenetic relationships in equally young orchid lineages – might prove also suitable for disentangling relationships within a species complex. In this study, exclusion of loci with missing data did not notably affect node support, but rather clade recovery in concatenated ML and coalescence-based inferences. Nevertheless, our results suggest that coalescence methods with taxon partitioning produce better supported phylogenies compared with ML analyses, likely due to the fact that coalescence methods can better account for topological incongruence among loci, such as that resulting from ILS.

Acknowledgements

We are grateful to R. Jimenez-Machorro and Norman Cash for assistance during fieldwork. We thank G. Romero (AMES), the Lankester Botanic Garden (Cartago) and Comisión Institucional de Biodiversidad of the University of Costa Rica for providing vegetal material and the permits for the access to the genetic information. Autoridad Nacional del Ambiente of Panama (ANAM) and the Smithsonian Tropical Research Institute (Panama City) and Costa Rican ministry of Environment (MINAET) kindly provided scientific research permits (SE/AP-20-13 and FOI-004-001, respectively). We thank the Deutsche Forschungsgemeinschaft for financial support (grants BL 462/14-1, GE 828/12-1, GO 1459/8). MFM is supported by the Juan de la Cierva fellowship of the Spanish Ministry of Economy (IJCI-2015-23459). OAPE is supported by the Lady Sainsbury Orchid Fellowship at the Royal Botanic Gardens Kew.

Appendix A

Table S1

Species names and voucher information, including herbarium of voucher deposition for material used in this study.

Fig. S1. The 50% majority rule consensus tree derived from SVDQuartets analysis for the DNA matrix c95s100 (i.e. no missing data) without taxon partition. Bootstrap support values (B) are provided at branches. Clades I–III are indicated with vertical bars.

Fig. S2. The 50% majority rule consensus tree derived from SVDQuartets analysis for the DNA matrix c95s50 and c95s15 without taxon partition. Bootstrap support values > 75 are displayed in boldface; values are shown in the order c95s50/c95s15. Clades I–III are indicated with vertical bars.

Fig. S3. Distribution of LBS values in six different intervals at nodes for the c95s100 dataset.

Fig. S4. Distribution of LBS values in six different intervals at nodes for the c95s50 dataset.

Fig. S5. Distribution of LBS values in six different intervals at nodes for the c95s15 dataset.

Fig. S6. Distribution of LBS values in relation to phylogenetic depth for the c95s100 dataset.

Fig. S7. Distribution of LBS values in relation to phylogenetic depth for the c95s50 dataset.

Fig. S8. Distribution of LBS values in relation to phylogenetic depth for the c95s15 dataset.

References

Aberer, A.J., Krompass, D., Stamatakis, A., 2013. Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. Syst. Biol. 62, 162–166. https://doi.org/10.1093/sysbio/sys078

Anderson, B.M., Thiele, K.R., Krauss, S.L., Barrett, M.D., 2017. Genotyping-by-sequencing in a species complex of Australian hummock grasses (*Triodia*): Methodological insights and phylogenetic resolution, PLoS ONE 12, e0171053. https://doi.org/10.1371/journal.pone.0171053

Andrews, K.R., Good, J.M., Miller, M.R., Luikart, G., Hohenlohe, P.A., 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. Nat. Rev. Genet. 17, 81–92. https://doi.org/10.1038/nrg.2015.28

Bateman, R.M., Molnár V., A., Sramkó, G., 2017. *In situ* morphometric survey elucidates the evolutionary systematics of the Eurasian *Himantoglossum* clade (Orchidaceae: Orchidinae). PeerJ 5, e2893. https://doi.org/10.7717/peerj.2893

Bateman, R.M., Sramkó, G., Paun, O., 2018. Integrating restriction site-associated DNA sequencing (RAD-seq) with morphological cladistic analysis clarifies evolutionary relationships among major species groups of bee orchids. Ann. Bot. 121, 85–105. https://doi.org/10.1093/aob/mcx129

Batista, J.A.N., Mota, A.C.M., Proite, K., Bianchetti, L. de B., Romero-González, G.A., Espinoza, H.M.H., Salazar, G.A., 2014. Molecular phylogenetics of Neotropical *Cyanaeorchis* (Cymbidieae, Epidendroideae, Orchidaceae): Geographical rather than morphological similarities plus a new species. Phytotaxa 156, 251–272. https://doi.org/10.11646/phytotaxa.156.5.1

Bogarín, D., Pérez-Escobar, O.A., Groenenberg, D., Holland, S.D., Karremans, A.P., Moriarty Lemmon, E., Lemmon, A.R., Pupulin, F., Smets, E., Gravendeel, B., 2018. Anchored Hybrid Enrichment generated nuclear, plastid and mitochondrial markers resolve the *Lepanthes horrida* (Orchidaceae: Pleurothallidinae) species complex. Mol. Phylogenet. Evol. 129, 27–47. https://doi.org/10.1016/j.ympev.2018.07.014

Britton, T., Lisa, C., Jacquet, D., Lundqvist, S., Bremer, K., 2007. Estimating divergence times

in large phylogenetic trees estimating divergence times in large phylogenetic trees. Syst. Biol. 56, 741–752. https://doi.org/10.1080/10635150701613783

Carr, G.F.J., 2012. The genus *Cycnoches*: species and hybrids. Orchid Rev. 1–31.

Chase, M.W., Cameron, K.M., Freudenstein, J. V., Pridgeon, A.M., Salazar, G., van den Berg, C., Schuiteman, A., 2015. An updated classification of Orchidaceae. Bot. J. Linn. Soc. 177, 151–174. https://doi.org/10.1111/boj.12234

Chifman, J., Kubatko, L., 2014. Quartet inference from SNP data under the coalescent model. Bioinformatics 30, 3317–3324. https://doi.org/10.1093/bioinformatics/btu530

Darwin, C., 1877. On the various contrivances by which british and foreign orchids are fertilised by insects. Appleton and Co., New York.

Després, L., Gielly, L., Redoutet, B., Taberlet, P., 2003. Using AFLP to resolve phylogenetic relationships in a morphologically diversified plant species complex when nuclear and chloroplast sequences fail to reveal variability. Mol. Phylogenet. Evol. 27, 185–196. https://doi.org/10.1016/S1055-7903(02)00445-1

Dodsworth, S., 2015. Genome skimming for next-generation biodiversity analysis. Trends Plant Sci. 20, 525–527. https://doi.org/10.1016/j.tplants.2015.06.012

Eaton, D. a R., 2014. PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. Bioinformatics 30, 1844–1849. https://doi.org/10.1093/bioinformatics/btu121

Eiserhardt, W.L., Antonelli, A., Bennett, D.J., Botigué, L.R., Burleigh, J.G., Dodsworth, S., Enquist, B.J., Forest, F., Kim, J.T., Kozlov, A.M., Leitch, I.J., Maitner, B.S., Mirarab, S., Piel, W.H., Pérez-Escobar, O.A., Pokorny, L., Rahbek, C., Sandel, B., Smith, S.A., Stamatakis, A., Vos, R.A., Warnow, T., Baker, W.J., 2018. A roadmap for global synthesis of the plant tree of life. Am. J. Bot. https://doi.org/10.1002/ajb2.1041

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., Mitchell, S.E., 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6, 1–10. https://doi.org/10.1371/journal.pone.0019379

Escudero, M., Eaton, D.A.R., Hahn, M., Hipp, A.L., 2014. Genotyping-by-sequencing as a tool to infer phylogeny and ancestral hybridization: A case study in *Carex* (Cyperaceae). Mol. Phylogenet. Evol. 79, 359–367. https://doi.org/10.1016/j.ympev.2014.06.026

Fehrer, J., Gemeinholzer, B., Chrtek, J., Bräutigam, S., 2007. Incongruent plastid and nuclear DNA phylogenies reveal ancient intergeneric hybridization in *Pilosella* hawkweeds (*Hieracium*, Cichorieae, Asteraceae). Mol. Phylogenet. Evol. 42, 347–361. https://doi.org/10.1016/j.ympev.2006.07.004

Fernández-Mazuecos, M., Mellers, G., Vigalondo, B., Sáez, L., Vargas, P., Glover, B.J., 2018. Resolving recent plant radiations: power and robustness of Genotyping-by-Sequencing. Syst. Biol. 67, 250–268. https://doi.org/10.1093/sysbio/syx062

633 Fragoso-Martínez, I., Salazar, G.A., Martínez-Gordillo, M., Magallón, S., Sánchez-Reyes, L.,
634 Moriarty Lemmon, E., Lemmon, A.R., Sazatornil, F., Granados Mendoza, C., 2016. A pilot
635 study applying the plant Anchored Hybrid Enrichment method to New World sages (*Salvia*
636 subgenus *Calosphace*; Lamiaceae). Mol. Phylogenet. Evol. 117, 124–134.
637 https://doi.org/10.1016/j.ympev.2017.02.006

638 Gale, S.W., Duangjai, S., Li, J., Ito, Y., Watthana, S., Termwutthipreecha, P., Cheuk, M.L.,
639 Suddee, S., 2018. Integrative analyses of *Nervilia* (Orchidaceae) section *Linervia* reveal
640 further undescribed cryptic diversity in Thailand. Syst. Biodivers. 2000, 1–20.
641 https://doi.org/10.1080/14772000.2017.1415233

642 Gerlach, G., Pérez-Escobar, O.A., 2014. Looking for missing swans: phylogenetics of
643 *Cycnoches*. Orchids 83, 434–437. https://doi.org/10.1126/science.337.6100.1274-a

644 Gramzow, L., Theißen, G., 2010. A hitchhiker's guide to the MADS world of plants. Genom
645 Biol. 11, 214. https://doi.org/10.1186/gb-2010-11-6-214

646 Granados-Mendoza, C., Wanke, S., Salomo, K., Goetghebeur, P., Samain, M.S., 2013.
647 Application of the phylogenetic informativeness method to chloroplast markers: A test case
648 of closely related species in tribe Hydrangeeae (Hydrangeaceae). Mol. Phylogenet. Evol.
649 66, 233–242. https://doi.org/10.1016/j.ympev.2012.09.029

650 Hipp, A.L., Eaton, D.A.R., Cavender-Bares, J., Fitzek, E., Nipper, R., Manos, P.S., 2014. A
651 framework phylogeny of the American oak clade based on sequenced RAD data. PLoS One
652 9. https://doi.org/10.1371/journal.pone.0093975

653 Hou, Y., Nowak, M.D., Mirré, V., Bjorå, C.S., Brochmann, C., Popp, M., 2015. Thousands of
654 RAD-seq loci fully resolve the phylogeny of the highly disjunct arctic-alpine genus
655 *Diapensia* (Diapensiaceae). PLoS One 10, 1–14.
656 https://doi.org/10.1371/journal.pone.0140175

657 Huang, H., Lacey Knowles, L., 2016. Unforeseen consequences of excluding missing data from
658 next-generation sequences: Simulation study of RAD sequences. Syst. Biol. 65, 357–365.
659 https://doi.org/10.1093/sysbio/syu046

660 Huson, D.H., 1998. SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics 14,
661 68–73. https://doi.org/10.1093/bioinformatics/14.1.68

662 Johnson, M., Pokorny, L., Dodsworth, S., Botigue, L., Cowan, R.S., Devault, A., Eiserhardt,
663 W.L., Epitawalage, N., Forest, F., Jan T, K., Leebens-Mack, J., Leitch, I., Maurin, O.,
664 Soltis, D., Soltis, P., Wong, G.., Baker, W.J., Wickett, N., 2018. A universal probe set for
665 targeted sequencing of 353 nuclear genes from any flowering plant designed using k-
666 medoids clustering. BioRxiv 361618, 1–36. https://doi.org/10.1101/361618

667 Johnson, S.D., Steiner, K.E., 1997. Long-Tongued fly pollination and evolution of floral spur
668 length in the *Disa draconis* complex (Orchidaceae). Evolution 51, 45–53.
669 https://doi.org/10.2307/2410959

670 Katoh, K., Rozewicki, J., Yamada, K.D. 2017. MAFFT online service: multiple sequence
671 alignment , interactive sequence choice and visualization. *Briefings in Bioinformatics*
672 bbx108. https://doi.org/10.1093/bib/bbx108

673 Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S.,
674 Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., Drummond, A.,
675 2012. Geneious Basic: An integrated and extendable desktop software platform for the
676 organization and analysis of sequence data. Bioinformatics 28, 1647–1649.
677 https://doi.org/10.1093/bioinformatics/bts199

678 Kosakovsky Pond, S.L., Frost, S.D.W., Muse, S. V., 2005. HyPhy: Hypothesis testing using
679 phylogenies. Bioinformatics 21, 676–679. https://doi.org/10.1093/bioinformatics/bti079

680 Lee, K.M., Kivelä, S.M., Ivanov, V., Hausmann, A., Kaila, L., Wahlberg, N., Mutanen, M.,
681 2018. Information dropout patterns in RAD phylogenomics and a comparison with
682 multilocus sanger data in a species-rich moth genus. Syst. Biol. 67, 925-939.
683 https://doi.org/10.1093/sysbio/syy029

684 Lemmon, A.R., Lemmon, E.M., 2012. High-throughput identification of informative nuclear loci
685 for shallow-scale phylogenetics and phylogeography. Syst. Biol. 61, 745–761.
686 https://doi.org/10.1093/sysbio/sys051

687 Li, H., Ruan, J., Durbin, R., 2008. Mapping short DNA sequencing reads and calling variants
688 using mapping. Genome Res. 18, 1851–1858. https://doi.org/10.1101/gr.078212.108.

689 López-Giráldez, F., Moeller, A.H., Townsend, J.P., 2013. Evaluating phylogenetic
690 informativeness as a predictor of phylogenetic signal for metazoan, fungal, and mammalian
691 phylogenomic data sets. Biomed Res. Int. 2013, 7–11. https://doi.org/10.1155/2013/621604

692 López-Giráldez, F., Townsend, J.P., 2011. PhyDesign: An online application for profiling
693 phylogenetic informativeness. BMC Evol. Biol. 11, 2–5. https://doi.org/10.1186/1471-2148-
694 11-152

695 Miller, M.A., Schwartz, T., Pickett, B.E., He, S., Klem, E.B., Scheuermann, R.H., Passarotti, M.,
696 Kaufman, S., Oleary, M.A., 2015. A RESTful API for access to phylogenetic tools via the
697 CIPRES science gateway. Evol. Bioinform. 11, 43–48. https://doi.org/10.4137/EBO.S21501

698 Mondragón-Palomino, M., 2013. Perspectives on MADS-box expression during orchid flower
699 evolution and development. Front. Plant Sci. 4, 377.
700 https://doi.org/10.3389/fpls.2013.00377

701 Mondragón-Palomino, M., Theißen, G., 2008. MADS about the evolution of orchid flowers.
702 Trends Plant Sci. 13, 51–59. https://doi.org/10.1016/j.tplants.2007.11.007

703 Nadeau, N.J., Martin, S.H., Kozak, K.M., Salazar, C., Dasmahapatra, K.K., Davey, J.W., Baxter,
704 S.W., Blaxter, M.L., Mallet, J., Jiggins, C.D., 2013. Genome-wide patterns of divergence
705 and gene flow across a butterfly radiation. Mol. Ecol. 22, 814–826.
706 https://doi.org/10.1111/j.1365-294X.2012.05730.x

Paradis, E., Claude, J., Strimmer, K., 2004. APE: Analyses of phylogenetics and evolution in R language. Bioinformatics 20, 289–290. https://doi.org/10.1093/bioinformatics/btg412

Paun, O., Turner, B., Trucchi, E., Munzinger, J., Chase, M.W., Samuel, R., 2016. Processes driving the adaptive radiation of a tropical tree (*Diospyros*, Ebenaceae) in New Caledonia, a biodiversity hotspot. Syst. Biol. 65, 2012-227. https://doi.org/10.1093/sysbio/syv076

Pérez-Escobar, O.A., 2016a. Molecular phylogenetics, evolution of sexual systems and historical biogeography of Darwin's favorite orchids (Catasetinae) and Swan orchids (*Cycnoches* Lindl.). Doctoral thesis, Ludwig-Maximilians Universität.

Pérez-Escobar, O.A., Balbuena, J.A., Gottschling, M., 2016b. Rumbling orchids: How to assess divergent evolution between chloroplast endosymbionts and the nuclear host. Syst. Biol. 65, 51-65. https://doi.org/10.1093/sysbio/syv070

Pérez-Escobar, O.A., Chomicki, G., Condamine, F.L., de Vos, J.M., Martins, A.C., Smidt, E.C., Klitgård, B., Gerlach, G., Heinrichs, J., 2017a. Multiple geographical origins of environmental sex determination enhanced the diversification of Darwin's favourite orchids. Sci. Rep. 7, 12878. https://doi.org/10.1038/s41598-017-12300-y

Pérez-Escobar, O.A., Gottschling, M., Chomicki, G., Condamine, F.L., Klitgård, B.B., Pansarin, E., Gerlach, G., 2017b. Andean mountain building did not preclude dispersal of lowland epiphytic orchids in the Neotropics. Sci. Rep. 7. https://doi.org/10.1038/s41598-017-04261-z

Pérez-Escobar, O.A., Gottschling, M., Whitten, W.M., Salazar, G., Gerlach, G., 2016c. Sex and the Catasetinae (Darwin's favourite orchids). Mol. Phylogenet. Evol. 97, 1–10. https://doi.org/10.1016/j.ympev.2015.11.019

Rokas, A., Williams, B.L., King, N., Carroll, S.B., 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425, 798–804. https://doi.org/10.1038/nature02053

Romero, G.A., Gerlach, G., in press. *Cycnoches*, in: Flora Mesoamericana. Missouri Botanical Garden Press, Saint Louis.

Rutherford, S., Rossetto, M., Bragg, J.G., McPherson, H., Benson, D., Bonser, S.P., Wilson, P.G., 2018. Speciation in the presence of gene flow: population genomics of closely related and diverging *Eucalyptus* species. Heredity 121, 126–141. https://doi.org/10.1038/s41437-018-0073-2

Salichos, L., Stamatakis, A., Rokas, A., 2014. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. Mol. Biol. Evol. 31, 1261–1271. https://doi.org/10.1093/molbev/msu061

Schoch, C.L., Sung, G.H., López-Giráldez, F., Townsend, J.P., Miadlikowska, J., Hofstetter, V., Robbertse, B., Matheny, P.B., Kauff, F., Wang, Z., Gueidan, C., Andrie, R.M., Trippe, K., Ciufetti, L.M., Wynns, A., Fraker, E., Hodkinson, B.P., Bonito, G., Groenewald, J.Z.,

744 Arzanlou, M., Sybren De Hoog, G., Crous, P.W., Hewitt, D., Pfister, D.H., Peterson, K.,
745 Gryzenhout, M., Wingfield, M.J., Aptroot, A., Suh, S.O., Blackwell, M., Hillis, D.M.,
746 Griffith, G.W., Castlebury, L.A., Rossman, A.Y., Lumbsch, H.T., Lücking, R., Büdel, B.,
747 Rauhut, A., Diederich, P., Ertz, D., Geiser, D.M., Hosaka, K., Inderbitzin, P., Kohlmeyer,
748 J., Volkmann-Kohlmeyer, B., Mostert, L., O'Donnell, K., Sipman, H., Rogers, J.D.,
749 Shoemaker, R.A., Sugiyama, J., Summerbell, R.C., Untereiner, W., Johnston, P.R.,
750 Stenroos, S., Zuccaro, A., Dyer, P.S., Crittenden, P.D., Cole, M.S., Hansen, K., Trappe,
751 J.M., Yahr, R., Lutzoni, F., Spatafora, J.W., 2009. The ascomycota tree of life: A phylum-
752 wide phylogeny clarifies the origin and evolution of fundamental reproductive and
753 ecological traits. Syst. Biol. 58, 224–239. https://doi.org/10.1093/sysbio/syp020

754 Stamatakis, A., 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of
755 large phylogenies. Bioinformatics 30, 1312–1313.
756 https://doi.org/10.1093/bioinformatics/btu033

757 Swofford, D.L., 2001. PAUP*: phylogenetic analysis using parsimony (*and other methods).
758 Version 4 Sinauer Associates, Sunderland, Massachusetts.
759 https://doi.org/10.1007/BF02198856

760 Taberlet, P., Coissac, E., Pompanon, F., Gielly, L., Miquel, C., Valentini, A., Vermat, T.,
761 Corthier, G., Brochmann, C., Willerslev, E., 2007. Power and limitations of the chloroplast
762 *trnL* (UAA) intron for plant DNA barcoding. Nucleic Acids Res. 35, e14.
763 https://doi.org/10.1093/nar/gkl938

764 Townsend, J.P., 2007. Profiling phylogenetic informativeness. Syst. Biol. 56, 222–231.
765 https://doi.org/10.1080/10635150701311362

766 Wagner, C.E., Keller, I., Wittwer, S., Selz, O.M., Mwaiko, S., Greuter, L., Sivasundar, A.,
767 Seehausen, O., 2013. Genome-wide RAD sequence data provide unprecedented resolution
768 of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. Mol.
769 Ecol. 22, 787–798. https://doi.org/10.1111/mec.12023

770 Wanke, S., Granados Mendoza, C., Müller, S., Paizanni Guillén, A., Neinhuis, C., Lemmon,
771 A.R., Lemmon, E.M., Samain, M.S., 2017. Recalcitrant deep and shallow nodes in
772 *Aristolochia* (Aristolochiaceae) illuminated using anchored hybrid enrichment. Mol.
773 Phylogenet. Evol. 117, 111–123. https://doi.org/10.1016/j.ympev.2017.05.014

774 Weltman, B. 2016. Systematics, taxonomy, and evolution of *Urtica* L. (Urticaceae). Doctoral
775 thesis, Rheinischen Freidrich-Wilhelms-Universität Bonn.

776 Wickham, H., 2009. GGPLOT2: elegant graphics for data analysis. Springer-Verlag, New York.
777 https://doi.org/10.1007/978-0-387-98141-3

778 Wilkinson, M., 1996. Majority-rule reduced consensus trees and their use in bootstrapping. Mol.
779 Biol. Evol. 13, 437–444. https://doi.org/10.1093/oxfordjournals.molbev.a025604

780