The use of chloroplast genome sequences to solve phylogenetic incongruences in *Polystachya* Hook (Orchidaceae Juss)

Narjara Lopes de Abreu [Corresp., 1, 2] , Ruy José Válka Alves [2] , Sérgio Ricardo Sodré Cardoso [3] , Yann JK Bertrand [1] , Filipe Sousa [1, 4] , Climbiê Ferreira Hall [5] , Bernard E Pfeil [1] , Alexandre Antonelli [1, 6, 7, 8]

[1] Department of Biological and Environmental Sciences, University of Gothenburg, Göteborg, Sweden

[2] Museu Nacional, Universidade Federal do Rio de Janeiro, São Cristóvão, Rio de Janeiro, Brasil

[3] Instituto de Pesquisas, Jardim Botânico do Rio de Janeiro, Diretoria de Pesquisa Científica, Rio de Janeiro, Brasil

[4] Centro de Ciências do Mar, Universidade do Algarve, Faro, Portugal

[5] Campus Três Lagoas, Universidade Federal de Mato Grosso do Sul, Três Lagoas, Mato Grosso do Sul, Brasil

[6] Gothenburg Botanical Garden, Göteborg, Sweden

[7] Gothenburg Global Biodiversity Centre, Göteborg, Sweden

[8] Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts, United States

Corresponding Author: Narjara Lopes de Abreu
Email address: narjara.lopes@gmail.com

Background. Current evidence suggests that for more robust estimates of species tree and divergence times, several unlinked genes are required. However, most phylogenetic trees for non-model organisms are based on single sequences or just a few regions, using traditional sequencing methods. Techniques for massive parallel sequencing or Next Generation Sequencing are an alternative to traditional methods that allow access to hundreds of DNA regions. Here we use this approach to resolve the phylogenetic incongruence found in *Polystachya* Hook. (Orchidaceae), a genus that stands out due to several interesting aspects, including cytological (polyploid and diploid species), evolutionary (reticulate evolution) and biogeographical (species widely distributed in the tropics and high endemism in Brazil). The genus has a notoriously complicated taxonomy, with several sections that are widely used but probably not monophyletic.

Methods. We generated the complete plastid genome of 40 individuals from one clade within the genus. The method consisted in construction of genomic libraries, hybridisation to RNA probes designed from available sequences of a related species, and subsequent sequencing of the product. We also tested how well a smaller sample of the plastid genome would perform in phylogenetic inference in two ways: by duplicating a fast region and analysing multiple copies of this dataset, and by sampling without replacement from all non-coding regions in our alignment. We further examined the phylogenetic implications of non-coding sequences that appear to have undergone hairpin inversions (reverse complemented sequences associated with small loops).

Results. We retrieved 131,214 bp, including coding and non-coding regions of the plastid genome. The phylogeny was able to fully resolve the relationships among all species in the targeted clade with high support values. The first divergent species are represented by African accessions and the most recent ones are among Neotropical species.

Discussion. Our results indicate that using the entire plastid genome is a better option than screening highly variable markers, especially when the expected tree is likely to contain many short branches. The phylogeny inferred is consistent with the proposed origin of the genus, showing a probable origin in Africa, with later dispersal into the Neotropics, as evidenced by a clade containing all Neotropical individuals. The multiple positions of *Polystachya concreta* (Jacq.) Garay & Sweet in the phylogeny are explained by allotetraploidy. *Polystachya estrellensis* Rchb.f. can be considered a genetically distinct species from *P. concreta* and *P. foliosa* (Lindl.) Rchb.f., but the delimitation of *P. concreta* remains uncertain. Our study shows that next generation sequencing provides a powerful tool for inferring relationships at low taxonomic levels, even in taxonomically challenging groups with short branches and intricate morphology.

1   Manuscript submission to PeerJ.

3   **This paper is part of the Collection: "Endless forms: Advances in evolutionary analyses of**

4   **biodiversity"**

THE USE OF CHLOROPLAST GENOME SEQUENCES TO SOLVE
PHYLOGENETIC INCONGRUENCES IN *POLYSTACHYA* HOOK.
(ORCHIDACEAE JUSS.).

9   Narjara Lopes de Abreu[1,2*], Ruy José Válka Alves[1], Sérgio Ricardo Sodré Cardoso[3], Yann J.K.

10   Bertrand[2], Filipe Sousa[2,4], Climbiê Ferreira Hall[5], Bernard E. Pfeil[2§], & Alexandre

11   Antonelli[2,6,7,8§]

13   [1]Museu Nacional, Universidade Federal do Rio de Janeiro, São Cristóvão, Rio de Janeiro, Brasil

14   [2]Department of Biological and Environmental Sciences, University of Gothenburg, Box 463, 405

15   30 Göteborg, Sweden.

16   [3]Instituto de Pesquisas, Jardim Botânico do Rio de Janeiro, Diretoria de Pesquisa Científica, Rio

17   de Janeiro, Brasil.

18   [4]CCMAR (Centro de Ciências do Mar da Universidade do Algarve), Edifício 7, Gambelas 8005-

19   139, Faro, Portugal.

20   [5]Campus Três Lagoas, Universidade Federal de Mato Grosso do Sul, Três Lagoas, Mato Grosso

21   do Sul, Brasil.

22   [6]Gothenburg Botanical Garden, Carl Skottsbergs gata 22A, SE-41319 Göteborg, Sweden.

23   [7]Gothenburg Global Biodiversity Centre, Box 461, SE-405 30 Göteborg, Sweden.

24   [8]Department of Organismic and Evolutionary Biology, Harvard University, 26 Oxford St.,

25   Cambridge, MA 02138 USA

27   * Corresponding Author:

28   Narjara Abreu[1]. Email address: narjara.lopes@gmail.com

29   Bernard Pfeil[2]. Email address: bernard.pfeil@gu.se

30   § Both are senior authors in this paper.

31   **ABSTRACT**

32   **Background.** Current evidence suggests that for more robust estimates of species tree and

33   divergence times, several unlinked genes are required. However, most phylogenetic trees for

34   non-model organisms are based on single sequences or just a few regions, using traditional

35   sequencing methods. Techniques for massive parallel sequencing or Next Generation

36   Sequencing are an alternative to traditional methods that allow access to hundreds of DNA

37   regions. Here we use this approach to resolve the phylogenetic incongruence found in

38   *Polystachya* Hook. (Orchidaceae), a genus that stands out due to several interesting aspects,

39   including cytological (polyploid and diploid species), evolutionary (reticulate evolution) and

40   biogeographical (species widely distributed in the tropics and high endemism in Brazil). The

41   genus has a notoriously complicated taxonomy, with several sections that are widely used but

42   probably not monophyletic.

43

44   **Methods.** We generated the complete plastid genome of 40 individuals from one clade within

45   the genus. The method consisted in construction of genomic libraries, hybridisation to RNA

46   probes designed from available sequences of a related species, and subsequent sequencing of the

47   product. We also tested how well a smaller sample of the plastid genome would perform in

48   phylogenetic inference in two ways: by duplicating a fast region and analysing multiple copies of

49   this dataset, and by sampling without replacement from all non-coding regions in our alignment.

50   We further examined the phylogenetic implications of non-coding sequences that appear to have

51   undergone hairpin inversions (reverse complemented sequences associated with small loops).

52

53   **Results.** We retrieved 131,214 bp, including coding and non-coding regions of the plastid

54   genome. The phylogeny was able to fully resolve the relationships among all species in the

55   targeted clade with high support values. The first divergent species are represented by African

56   accessions and the most recent ones are among Neotropical species.

57

58   **Discussion.** Our results indicate that using the entire plastid genome is a better option than

59   screening highly variable markers, especially when the expected tree is likely to contain many

60   short branches. The phylogeny inferred is consistent with the proposed origin of the genus,

61   showing a probable origin in Africa, with later dispersal into the Neotropics, as evidenced by a

62   clade containing all Neotropical individuals. The multiple positions of *Polystachya concreta*

63   (Jacq.) Garay & Sweet in the phylogeny are explained by allotetraploidy. *Polystachya*

64   *estrellensis* Rchb.f. can be considered a genetically distinct species from *P. concreta* and *P.*

65   *foliosa* (Lindl.) Rchb.f., but the delimitation of *P. concreta* remains uncertain. Our study shows

66   that next generation sequencing provides a powerful tool for inferring relationships at low

67   taxonomic levels, even in taxonomically challenging groups with short branches and intricate

68   morphology.

69

70

71   **INTRODUCTION**

72   Orchidaceae is considered the largest family of Flowering Plants, with over 25,000

73   species (Dressler, 1990; Christenhusz & Byng, 2016). The family probably dates back to the

74   Late Cretaceous, as indicated by fossil-calibrated molecular phylogenies (Gustafsson, Verola &

75   Antonelli, 2010; Ramirez et al., 2007, 2011). *Polystachya* Hook. is an orchid genus containing

76   240 species, with most species found in Africa (Dressler, 1993). Thirteen species are reported

77   from the Neotropical region (Mytnik-Ejsmont, 2011), but this number may increase when

78   considering the endemic species from Brazil that were not accounted for in Mytnik-Ejsmont

79   (2011) or were considered synonymous (Barros et al., 2010).

80   Recent studies have shown a number of peculiar cytological, evolutionary and

81   biogeographic aspects of *Polystachya*. The genus has diploid and polyploid species, the latter

82   recently formed in the Neotropics and Madagascar (Rupp et al., 2010; Russell et al., 2010b).

83   Unlike most genera of Orchidaceae, *Polystachya* has a wide geographical distribution range

84   (Pridgeon et al., 2005; Fig. 1), having species that are Pantropical or have a transatlantic

85   distribution. On the other hand, the Neotropics presents a high level of endemism. Brazil, as an

86   example, has 12 species of which 10 are endemic (Barros et al., 2010). In addition, there is

87   evidence of reticulate evolution in the genus and hybridisation with independent origins (Russell

88   et al., 2010a).

89   The monophyly of the genus has been reported in the latest studies (Russell et al., 2010b;
90 Mytnik-Ejsmont, 2011), which contrasts starkly with the low level of monophyly observed in the
91 taxonomic sections described within the genus. Those 15 sections (Kraenzlin, 1926;
92 Summerhayes, 1942, 1947 *apud* Russell et al., 2011b; Brenan, 1954; Cribb, 1978) are based on
93 morphological characters and have been useful for field identification and inventories, but do not
94 find support as natural groupings in the molecular studies currently available (Russell et al.,
95 2010b; Mytnik-Ejsmont, 2011). According to those molecular studies, all sections are
96 polyphyletic or paraphyletic, except sect. *Isochiloides* (Russell et al., 2010b).
97   Section *Polystachya* has been described as comprising 32 species worldwide and is the
98 only section with species of pantropical distribution (Mytnik-Ejsmont, 2011). However,
99 according to molecular analyses, some species of this section appear to be more related to
100 species of other sections (Russell et al., 2010b; Mytnik-Ejsmont, 2011). These studies highlight
101 the need for new infrageneric divisions based on robust molecular evidence. Russell et al.
102 (2010b), using chloroplast markers, defined five different clades that could be used as the basis
103 for a revised classification of new sections within the genus. Clade III (sensu Russel et al.,
104 2010b) includes species from five different sections (sects. *Polystachya*, *Eurychilae*,
105 *Caulescentes*, *Superpositae*, *Polychaete*) and is divided into distinct subclades of
106 morphologically diverse plants. These species are Pantropical (such as *P. concreta*), Neotropical
107 (such as *P. foliosa*), Malagasy endemics (such as *P. henrici*) or African (such as *P. odorata*). The
108 relationships among Clade III species remain unresolved because specimens of *P. concreta*, *P.
109 foliosa*, *P. henrici*, and *P. modesta* form a large polytomy, due to low levels of divergence
110 between sequences. In addition, several distinct subclades are unresolved at the base. The
111 molecular phylogenetic studies produced so far have included about 35% of the recognized
112 species within the genus, and used only a small number of nuclear (*PgiC* between exons 11 and
113 15, *PhyC* exon 1, *Rpb2* intron 23 and ITS in Russell et al., 2010a; only ITS in Mytnik-Ejsmont,
114 2011) and chloroplast markers (*rps16* intron, *rps16* exon 2, *rps16-trnK* spacer, *trnK* intron
115 excluding *matK*, *matK* and *psbD-trnT* spacer in Russell et al., 2010b; *rps16-trnK*, *rps16* and
116 *rpl32-trnL* in Mytnik-Ejsmont, 2011).
117   New methods of DNA sequencing as well as the development of more powerful
118 algorithms are propelling the replacement of trees generated from one or a few genes to those
119 constructed from hundreds of them (Edwards, 2009). The improvement of massively parallel

120 sequencing techniques – or Next Generation Sequencing (NGS) – has increased the amount of

121 data available for biological research, whether the fully annotated reference genomes of species

122 under study have been sequenced or not (Bräutigam & Gowik, 2010). However, despite its

123 obvious potential, NGS technology is underused in most studies of plant systematics (Cronn et

124 al., 2012; Carstens et al., 2013; Eaton & Ree, 2013), probably as a result of a prevailing focus on

125 non-model organisms (which require *de novo* genomic sequencing and its inherent challenges),

126 the need to sample many individuals per species and the absence of well-established protocols

127 (McCormack, 2013). One method that increases the efficiency of NGS for non-model species

128 compared to other genomic partitioning strategies is sequence capture (or hybridisation-based

129 enrichment), which is based on the prior selection of loci of interest (Lemmon & Lemmon,

130 2013). The main benefit of this technique is that the number of specific sequences obtained can

131 be very high, which makes it an advantageous method compared to PCR-based approaches if the

132 objective is to sequence several individuals and multiple loci. Furthermore, sequence capture

133 when combined with NGS platforms, such as Illumina, also reduces the costs of the process

134 (Lemmon, Emme & Lemmon, 2012). Sequence capture methods have successfully been used in

135 other plant genera to generate large amounts of useful data for phylogenetic inference (Kamneva

136 et al., 2017; Sousa et al., 2014; Stephens et al., 2015; Weitemier et al., 2014)

137   The necessity of a molecular phylogenetic framework for (and a morphological

138 taxonomic revision of) *Polystachya* is clear. It requires a well-resolved phylogenetic hypothesis

139 in order to clarify the relationships between species and also to redefine new infrageneric

140 sections. In this paper, we explore the use of nearly complete plastid genomes (note that we use

141 chloroplast and plastid interchangeably when referring to these genomes), obtained by sequence

142 capture and massively parallel sequencing, to solve the phylogenetic inconsistencies found

143 within Clade III of *Polystachya* (sensu Russell et al., 2010b). We also explore whether

144 sequencing the entire chloroplast genome using NGS was worthwhile, compared to PCR and

145 Sanger sequencing of a few fast-evolving loci. We hope that the results generated here can be

146 extended to the rest of the genus and thus result in new interpretations of the evolutionary and

147 biogeographic history of the group.

148

149   **MATERIALS & METHODS**

150   **Sampling and DNA Extraction**

151   We sampled 19 species and 48 individuals (Table 1, Fig. 1), of which 15 were collected

152   in different locations in continental Brazil and three were collected on Trindade Island in the

153   South Atlantic. The DNA of the 15 Brazilian samples was extracted from 10 mg of tissue dried

154   with silica gel and using the DNeasy Plant Mini Kit (Qiagen). DNA samples for the remaining

155   33 individuals were provided by the University of Vienna and the DNA bank of the Royal

156   Botanic Gardens, Kew. To this sample of individuals, selected because they cluster in Clade III

157   (as defined in Russell et al., 2010b), we added *Polystachya tessellata* Lindl., supposedly

158   synonymous of *P. concreta*. We also included multiple samples of *P. concreta*, because previous

159   studies have reported a lack of monophyly for this species and several synonymous species.

160   Permits to collect were provided by the Ministério do Meio Ambiente (MMA), Instituto Chico

161   Mendes de Conservação da Biodiversidade (ICMBio) and Sistema de Autorização e Informação

162   em Biodiversidade (SISBIO), with registration number 29478-1.

163   *Polystachya bicolor* Rolfe and *P. melanantha* Schltr. were chosen as possible more

164   distant relatives of the species in focus, in order to provide additional context for the

165   phylogenetic inference. All studies conducted to date resolve *P. melanantha* as an outgroup with

166   respect to the Clade III species. *Polystachya bicolor* has already been treated as a synonym of *P.*

167   *rosea* (Mytnik-Ejsmont, 2011), with an uncertain position in the phylogenetic trees generated

168   thus far, being sometimes closely related to *P. concreta* and other associated species (Mytnik-

169   Ejsmont, 2011) and sometimes closely related to species from other clades (Russell et al., 2010b;

170   Mytnik-Ejsmont, 2011).

171

172   **Probe Design for DNA Capture**

173   We used complete chloroplast genome of *Phalaenopsis aphrodite* subsp. *formosana*

174   (NC_007499.1) (Chang et al., 2006) as the reference for the design of capture probes, because

175   there is no completely sequenced chloroplast genome of a *Polystachya*. According to molecular

176   analyses, *Polystachya* and *Phalaenopsis* belong to different sub-tribes but are closely related

177   within the Vandeae tribe (Van Den Berg et al., 2005; Górniak, Paun & Chase, 2010;

178   Freudenstein & Chase, 2015). The use of a quite distantly related species is made possible by the

179   DNA capture kit (MyBaits), which is able to support differences larger than 5% between probe

180   sequences and target sequences (e.g., Li et al., 2013). The complete sequence of the chloroplast

181   genome of *Phalaenopsis aphrodite* subsp. *formosana* was divided into blocks of 360 bp. Every

182   second block was used as the template for probe design; the probes consisting of 120 bp

183   sequences, three to each block without overlap. Given the genomic DNA fragment sizes of

184   between 300-400 bp (below) and that fragments can contain complementary sequence anywhere

185   on their length to attach to a probe, fragments can contain up to 200-300 bp of genomic sequence

186   into the flanking regions beyond the probes, or cover the probes with little extent into the

187   flanking sequence, or somewhere in between. In this way, captured sequences produce a series of

188   tiled overlapping sequence for high quality genomic assembly. Additionally, fragments with a

189   base repeated more than seven times in a row were avoided to reduce the capture of repetitive

190   sequences present in many places in the genome. Finally, the reference sequences in blocks used

191   for probe design totalled 63,720 bp and were brought together into a single FASTA file and sent

192   to MYcroarray (Ann Arbor, Michigan, USA) to produce the probes.

193   **Data Generation**

194   **Sonication and genomic library preparation**

195   Extracted DNA was randomly fragmented by sonication using a Covaris S220 instrument

196   (Covaris, Woburn, Massachusetts, USA), in order to evenly cover the full genome. Adapters

197   were incorporated into the fragmented DNA using NEXTflexTM DNA Sequencing Kit and

198   NEXflexTM Barcodes kit (BIOO Scientific, Austin, Texas, USA). Uniquely indexed adapters

199   were used for each sample. We selected fragments between 300 and 400 bp using

200   AgencourtAMPure XP magnetic beads kit (Beckman Coulter). The genomic library was

201   amplified following the program: 98ºC for 2 minutes; 14 cycles (98ºC for 30 seconds; 65ºC for

202   30 seconds; 72ºC for 60 seconds); 72ºC for 4 minutes. The products were purified using a

203   QIAquick PCR Purification Kit (Qiagen). The genomic DNA concentrations before and after

204   sonication and the amplification of the library were measured in a NanoDrop 2000c instrument

205   (Thermo Fisher Scientific, Waltham, MA, USA) (Table 1) to ensure that the final concentration

206   exceeded 400ng/ μL.

207 **Enrichment and sequencing**

208 Before the enrichment, equimolar amounts (400ng/μL) of each amplified library were

209 pooled into six reactions, each one containing eight indexed samples. The enrichment method

210 involves the selection of genomic regions and capture of DNA samples before sequencing

211 (Mamanova et al., 2010). The enrichment was performed with MYBaits target enrichment

212 system (MYcroarray, Ann Arbor, Michigan), following the manufacturer's instructions. The

213 probes were recovered using Dynabeads® MyOneTM Streptavidin C1 (Invitrogen Dynal AS,

214 Oslo, Norway).

215 To increase DNA concentration, 14 cycles of PCR were performed for each hybridisation

216 reaction using Herculase II Fusion DNA Polymerase (Agilent, Waldbronn, Germany) and the

217 following program: 98ºC for 30 seconds; 14 cycles (98ºC for 20 seconds; 60ºC for 30 seconds;

218 72ºC for 60 seconds), 72ºC for 5 minutes. Sequencing was performed on the Illumina MiSeq

219 platform (San Diego, California, USA) by the Genomics Core Facility (University of

220 Gothenburg, Sweden).

221 **Sequence editing**

222 Illumina reads were processed using the program CLC assembly cell (CLC Bio, Aarhus,

223 Denmark). Firstly, the Illumina adapter sequences were removed and low-quality reads were

224 excluded. Reads were then mapped against the reference sequence used for probe design

225 (*Phalaenopsis aphrodite*). Consensus sequences generated for each sample were converted into

226 FASTA format using the SAMTools software (Li et al. 2009) using the mpileup tool with

227 reference sequence option, allowing for the inclusion of indels in the consensus sequences. These

228 sequences were used as a new individual reference sequences for each sample in a second round

229 of mapping. Final consensus sequences were generated using mpileup, without the reference

230 sequence option, to avoid erroneous base calling in low read-depth portions of the read

231 alignment. Sequence alignment was performed using the auto strategy in MAFFT - Multiple

232 Sequence Alignment Software Version 7 (Katoh & Standley, 2013) and later manually refined

233 using Geneious Pro (Biomatters). In the last step we aligned the sequenced samples with the

234 *Phalaenopsis aphrodite* subsp. *formosana* (NC_007499.1) chloroplast genome to obtain the

235 sequenced region annotation."

236

237

Alignment and Phylogenetic Analysis

Hairpin inversions

Micro-structural features of chloroplast non-coding sequences can have a profound influence on the multiple sequence alignment, and hence also the phylogeny. Hairpins (short stem-loop structures in single stranded DNA or RNA), for example, can create sites that allow small inversions to occur at a high enough frequency that homoplasious inversions can be observed among sequences from closely related species (Kelchner & Wendel, 1996). Sometimes the inverted sequence is not so short and can disrupt phylogenetic analysis, leading to strongly supported but spurious groupings (Joly et al., 2010). Non-coding sequences, such as group II chloroplast introns, contain many such stem-loop structures (Kelchner, 2002).

We examined the non-coding sequences in our alignment for inverted (reverse complemented) sequences and tested for their effect on the phylogenetic inference. This was done by excluding all but one character of the inversion (to down-weight the inversion to a single event) and rerunning the analysis. The selected character to represent the inversion was arbitrarily chosen. This was done to avoid recoding the inversions as indel characters and creating a new, small partition (with only eight characters) that would have required many additional parameters, in comparison to our approach.

Site exclusions

The alignment process can sometimes be confronted with small regions that are difficult to align, probably most often due to overlapping indel events. We identified several such regions and excluded them using the nexus block commands from the Bayesian analysis. Poor alignment excluded sites: 54459-54465, 55220-55224, 55484-55489, 67162-67167, 67426-67431, 68186-68192, 118838-118843, 119266-119364, 119493-119726, 120012-120019, 120867-120872, 121295-121393, 121522-121755, 122041-122048. Inverted loop-associated excluded sites: 74668-74669, 88120-88128, 90474-90481, 94193-94205, 94696-94699, 97241-97245, 104712-104716, 106631-106635.

The aligned sequences were partitioned based on the chloroplast annotation of 116 functional genes, seven pseudogenes and two partitions that concatenated untranslated regions –

266 one partition containing introns with secondary structure and the other concatenating all

267 intergenic sequences. The files that contained the introns were submitted to trimAl v1.2

268 (Capella-Gutiérrez, Silla-Martínez & Gabaldón, 2009) to test different gap deletion settings.

Faster region assessment

270 We used the sequences from one sample on GenBank (FS1045) of *Polystachya*

271 *cultriformis* and examined two published markers, *psbD-trnT* (870 length aligned to our

272 samples) and *matK* (1521 length aligned to our samples). We compared these sequences pairwise

273 to one of our samples, *P. estrelensis*8, to check which was the faster-evolving region. We then

274 ran a MrBayes v.3.2.4 (Ronqvist et al., 2012) analysis on the faster region, as a representative of

275 a fast part of the chloroplast genome (fast cpDNA hereafter), with one copy of our dataset

276 trimmed to this region alone plus the GenBank sequence.

277 We then ran successive analyses, using additional copies of the dataset interleaved in the

278 same file, to explore the increase in support with an increase of characters evolving under the

279 same model. This was to discover how much of the fast cpDNA data would be needed to achieve

280 high support on most nodes in the phylogeny (i.e., among species but not necessarily within

281 species). An important assumption we made at this stage was that the single fast region would

282 contain mutations spread across most/all branches of the phylogeny. If this assumption was true,

283 then a single region could carry changes representative of the entire phylogenetic history that we

284 were exploring. The assumption is essentially one of i.i.d. (independently and identically

285 distributed sites) – in that the sites would be representative of many unsampled sites and that

286 double mutations would be rare – coupled with a sufficient dataset size to contain enough

287 changes overall to reflect the history. Although the i.i.d. assumption is rarely true across sites,

288 model-based analysis methods can cope with this, because the frequency of site patterns can be

289 modelled by an i.i.d. process (Steel and Penny, 2000). So we are in effect mainly testing whether

290 the original data set size was sufficient to carry changes reflective of the entire phylogeny under

291 investigation.

Random sample from all non-coding regions

293 A single region copied many times proved ineffective in recovering most nodes with

294 support (see Results). We therefore explored using random samples of characters without

295 replacement from among all of the non-coding regions in our dataset to test how much data from

296   faster regions would yield supported trees across most nodes. We expected this approach to be

297   less subject to the limitations caused by the stochastic nature of mutations coupled with the

298   limited size of any one region. By sampling across many regions (over 58 kb in this case), even

299   those few characters that have changed on short branches might be sampled occasionally. In

300   contrast, a single region, by chance, may simply not contain any characters changing on a

301   specific short branch.

302   We sampled without replacement 4%, 8% and 16% of the non-coding data using delete-

303   fraction jackknifing in the seqboot program v3.69 (from

304   http://evolution.genetics.washington.edu/phylip.html), excluding the poorly aligned parts and

305   with down-weighting of the inverted loops (by excluding all but one character of each loop), in

306   20 replicates each. The approximate average (and range) of PP per clade was taken across the 20

307   replicates to get an indication of the likely support for selected clades that a non-coding dataset

308   of these sizes would generate. These values were plotted on the whole alignment analysis to

309   compare to the support received when using the whole dataset. Given that the largest dataset we

310   used here (hereafter the "16% dataset", or ~9.2kb) failed to recover support for all nodes found

311   in the whole genome analysis (see Results), we did not end up analysing the smaller replicates.

312   **Phylogenetic analysis**

313   MrBayes v.3.2.4 analyses were used for phylogenetic inference. These analyses were run

314   for five million generations (two million for the random sample replicates), using a mixed

315   substitution model (plus gamma and invariant sites) to account for among-site rate variation.

316   Priors on branch lengths were set to unconstrained: exponential (100) to minimise the chance of

317   inferring incorrectly long branches (Marshall, 2010), otherwise with default settings. The paired

318   runs were checked for convergence and high effective samples sizes in the MrBayes output and

319   Tracer v.1.6 (Rambaut et al., 2014), respectively. Burn-in generations were removed by

320   discarding 10% of the samples of parameters and trees, while summarising in TreeAnnotator

321   v.1.8 (Rambaut and Drummond, 2010) to ascertain clade posterior probabilities. Trees were

322   rooted using the *Phalaenopsis* sequence. Analyses using the character partitions were also done,

323   returning nearly identical results to the analysis described above, so they are not reported further.

324

RESULTS

326   Our NGS approach allowed the capture of coding and non-coding regions throughout the
327 chloroplast genome. We recovered approximately 132kb, after the exclusion of gaps,
328 representing 116 genes, seven pseudogenes, as well as regions with intergenic sequences and
329 introns with secondary structure. Compared to the reference annotation, seven genes contained
330 frameshifts that are usually associated with pseudogenisation and corresponded to previously
331 reported pseudogenised genes in orchids (Luo et al., 2014).

332   We excluded eight of the 48 samples due to the low quality of the sequencing results
333 (Table1). These eight samples showed lower DNA concentrations after the genomic library
334 construction assembly, which may be the cause of low quality sequencing. The remaining 40
335 samples were submitted to the EMBL/ENA database under accession numbers ERS2203551-
336 ERS2203590. The coding regions have 48,308 polymorphic sites (38,4%). Introns with
337 secondary structure and regions with intergenic sequences have 21,264 (16,9%) and 56,226
338 (44,7%) polymorphic sites, respectively. The alignment of the concatenated data showed an
339 unbalanced (but fairly typical) mean nucleotide composition of A=29,9%, C=19,9%, G=19,4% e
340 T=30,8%.

341

342 **Analysis Results**

343   **Hairpin inversions**

344   In the non-coding part of alignment, we found evidence for eight putative small
345 inversions (Table 2), based on the presence of inverted repeated motifs that could form stems at
346 least four bp long. Stems of this length or longer are part of models of group II structures
347 (Michel, Umesono & Ozeki, 1989; Toor, Hausner & Zimmerly, 2001; Kelchner, 2002) and are
348 consistent with sequence patterns observed by one of us in the *rpL16* intron sequences in other
349 taxa (Pfeil et al., 2002).

350   **Faster region assessment**

351   The pairwise identity between *Polystachya concreta*5 and *P. concreta*8 (whose common
352 ancestor is relatively old and near the crown of *Polystachya*) for *psbD-trnT* was 98.7%. The
353 pairwise identity for these samples for *matK* was 99.2%. *PsbD-trnT* was therefore used as the

354   representative fast cpDNA region. The analysis with a single copy of this dataset (870 bp)

355   yielded a MCC tree (Fig. 2A) with only five nodes with high support (>0.95 PP). Increasing the

356   number of copies did not result in much improvement. The analysis using 16 copies of *psbD-*

357   *trnT* (~14 kb) produced a MCC tree (Fig. 2B) containing only eight highly supported nodes.

358   **Random sample from all non-coding regions**

359   The pairwise identity between *Polystachya concreta*5 and *P. concreta*8 was 98.7% across

360   58 236 bp of non-coding region contained in our alignment. This compares to 98.8% identity

361   between the same samples across the coding regions in our alignment.

362   The replicate datasets that sampled 16% of the original non-coding alignment (excluding

363   poorly aligned parts and down-weighting the inverted loops) failed to return all nodes/clades

364   found in the whole genome analysis. Of 13 selected nodes found in the tree from the whole

365   genome (four subtended by relatively long branches, four by medium length branches, and five

366   by short branches), only five were found with high support across most or all replicates (i.e., at

367   least 16 of 20 replicates had ≥0.95 PP). Three of the selected nodes instead had five or fewer

368   replicates with high support (≥0.95 PP), but only one or no replicates that contained highly

369   supported contradictory nodes (thus the support for the expected node was ≤0.05). Finally, five

370   of the nodes had generally poor support among replicates (i.e., five or fewer replicates had ≥0.95

371   PP along with six or more replicates with ≤0.05 PP for these nodes). The 16% datasets (~9.2 kb)

372   recovered from seven to 21 highly supported nodes among replicates (mean=14.8), with more

373   nodes recovered in 19 of 20 replicates than was the case with the larger repeated *psbD-trnT*

374   dataset (~14 kb and eight supported nodes). This character sampling strategy was probably more

375   reflective of the underlying support for various nodes than using repeated copies of a single

376   small dataset.

377   The mutually exclusive foliosa1/concreta2 versus foliosa1/foliosa2 clades (see below)

378   were also examined in the 16% datasets. In the first case (foliosa1/concreta2), just four replicates

379   contained this clade with high or moderate support (≥0.90 PP). The contradictory second

380   grouping (foliosa1/foliosa2) was found with a similar level of support (≥0.90 PP) in only two

381   replicates. The fact that both groupings could be recovered, with high support, in at least some

382   replicates suggests that the original dataset contains the signal of both clades. A NeighborNet

383   analysis (Fig. 3B) confirmed that a mixture of patterns exists in the original dataset involving

384   foliosa1, foliosa2, concreta1 and concreta2.

385   **Phylogenetic analyses**

386   Analyses with and without the inverted loops (the latter by down-weighting to a single

387   character) returned almost identical trees. The results of only the latter analysis is presented in

388   this section. The tree we recovered was able to resolve the phylogenetic relationships among the

389   groups of the large clade selected for this study, with high support values on almost every node

390   (Fig. 3). The tree was characterized by a large clade with relatively short branches containing

391   only sequences from the Neotropics, with a grade of a few small clades and single sequences

392   containing the remaining sequences (Fig. 3). The large clade contained 21 sequences from

393   Brazil, Dominica and Venezuela, whereas the grade included 19 sequences from tropical central

394   and eastern Africa, as well as Madagascar and the nearby islands (Fig. 3).

395   The grade recovered include a few geographically identifiable clades (Fig. 3). One of

396   these, attaching fairly deeply within the crown, consists of four Malagasy sequences

397   (*Polystachya humbertii*1, *P. humbertii*2, *P. oreocharis* and *P. tsinjoarivensis*2) that are sister to a

398   Kenyan sequence (*P. eurychila*). Another clade comprises a Kenyan sequence (*P. golungensis*)

399   and one from Reunion (*P. concreta*8). A third clade contains a pair of central African sequences,

400   one from Cameroon (*P. odorata*2) and one from Nigeria (*P. odorata*1). A fourth clade contains

401   sequences from central Africa (*P. concreta*5 from Cameroon), Madagascar (*P. tesselata*1), the

402   Comoros (*P. concreta*9) and two sequences without certain provenance. Finally, a fifth pair of

403   sequences were from samples collected from Mauritius (*P. concreta*7) and Madagascar (*P.*

404   *tesselata*2). Lineages containing only a single sequence in this grade included samples from

405   Kenya (*P. melanantha* and *P. steudneri*) and Cameroon (*P. dolichophylla*).

406   Sequences from the widely sampled and widely distributed *P. concreta* did not form a

407   monophyletic group and occurred on different branches of the tree, separated by several well

408   supported nodes (Fig. 3). Similarly, the two *P. tessellata* sequences from Madagascar did not

409   form a clade. *Polystachya estrellensis* sequences form a clade with *P. concreta* sequences

410   collected in Brazil. Although the sequences of *P. estrellensis* are thus paraphyletic, whether the

411   taxon itself is paraphyletic cannot be established for certain here. The identification of this *P.*

412   *concreta* sample could be wrong, given that the identification of these species is confused in

413   Brazil and sometimes they are considered synonymous (see also Discussion).

414   **With versus without loops**

415   The down weighting of the inversions we identified (by excluding all but one character

416   per inversion) resulted in a similar, but not identical, phylogenetic inference. The differences

417   among the maximum clade credibility (MCC) trees involved *P. foliosa1*, *P. foliosa2*, *P.*

418   *concreta1*, *P. concreta2*, *P. concreta3* and *P. concreta4*. The analysis with the inversions

419   included returned this tree (Fig. 3A and Fig. 4A):

420   ((concreta3,concreta4):1,((foliosa1,foliosa2):0.98,(concreta1,concreta2):0.98):0.97)

421   with clade posterior probabilities (PP) listed after each node. In contrast, the inference

422   resulting from down-weighted inversions returned this tree (Fig. 3 main panel and Fig. 4 main

423   panel):

424   ((foliosa1,concreta2):1,((concreta3,concreta4):1,(foliosa2,concreta1):0.76):1)

425   There are several supported differences between these trees, with at least one

426   corresponding to the way the inverted loops are weighted. When the entire loops are analysed, *P.*

427   *foliosa1* and *P. foliosa2* are supported as sisters, with these two sequences appearing to share two

428   loop inversions (if this topology is correct; Fig. 4A). However, down-weighting the inversions

429   produces a tree consistent instead with two independent inversions (Fig. 4 main panel).

430

431   **DISCUSSION**

432   **Chloroplast genome sequence provides a robust phylogeny**

433   In this work we used the nearly complete chloroplast sequences of 40 *Polystachya*

434   samples to infer a robust plastid phylogeny. The dataset significantly increased the phylogenetic

435   resolution within the genus. Thus, our results suggest that increasing the number of molecular

436   markers has the potential to solve not only the relationships among species, but also to identify

437   new *Polystachya* clades and define new sections. The delimitation of new sections will, however,

438   depend upon the inclusion of more taxa than was done by this study – in other words a higher

439   coverage of the genus. Below we highlight some of the clades recovered, their morphological

440   and/or geographical characterisation, and a comparison with previous studies.

441 ### *Polystachya bicolor/rosea* position contradicts Russell's Clade III

442   Of the two species selected as possible distant relatives to provide more context in the

443 phylogenetic inference, one of them, *P. bicolor* (=*P. rosea*), appears in a clade together with

444 samples of *P. concreta* (from Cameroon and from the Comoros), *P. tesselata* (= *P. concreta*) and

445 *P. modesta*. Not surprisingly, the clade that includes *P. bicolor/rosea* is deeply nested within the

446 ingroup, thus contradicting the monophyletic Clade III presented by Russell et al. (2010 b).

447   In prior studies, *Polystachya bicolor/rosea* has an uncertain position in the phylogenetic

448 trees. In an analysis using plastid markers and Bayesian inference, this species appears in a large

449 polytomy with *P. concreta* and other related species (Mytnik-Ejsmont, 2011), or related to

450 species of other clades (Russell et al., 2010b) depending on the marker used. A phylogeny using

451 nuclear data (ITS sequences) highlighted the lack of monophyly of this species (Mytnik-

452 Ejsmont, 2011), which may be connected to the difficulty in identifying it. *Polystachya*

453 *bicolor/rosea* is often mistaken for *P. concreta*, since differentiation between these is made by

454 subtle differences in the shapes of leaves, and the size and colour of the flowers. Unlike *P.*

455 *concreta*, which has a pantropical distribution, *P. bicolor/rosea* is restricted to Madagascar,

456 Comoros and the Seychelles (Mytnik-Ejsmont, 2011).

457 ### Brazilian sequences form a clade

458   The monophyletic nature of the group formed by the Brazilian accessions, contrasting

459 with the paraphyletic group made up of African accessions, is consistent with the hypothesis that

460 Africa is the centre of origin with a subsequent (i.e., more recent) dispersal into the Neotropics

461 (Russell et al., 2010 a; Russell et al., 2010b).

462 ### Hybrid origins of some taxa suggested

463   The hybrid origin of *P. concreta* is a possible explanation for this species being found in

464 different positions in the tree (Russell et al., 2010b). *Polystachya concreta* individuals that have

465 dispersed out of Africa are tetraploid, whereas plants found in continental Africa can be diploid

466 or tetraploid. The sister taxa of African *P. concreta* are diploid (Russell et al., 2010b), indicating

467 that tetraploidy is a derived state in *P. concreta*. Allotetraploidy in *P. concreta* has been

468 confirmed by analysis of low copy nuclear genes (Russell et al., 2010a).

469   Interspecific hybridisation events, as in *P. concreta*, are considered a source of

470 chloroplast genome exchange via introgression. Chloroplast genome exchange among species is

471   sometimes suggested as an explanation for the inconsistencies between phylogenetic trees based

472   on nuclear and plastid markers in, e.g., *Populus* (Salicaceae) (Smith & Sytsma, 1990; Tsitrone,

473   Kirkpatrick & Levin, 2003), *Nothofagus* (Nothofagaceae) and Crassulaceae (Mort et al., 2002;

474   Acosta & Premoli, 2010). In *Nothofagus*, chloroplast capture results in the association of

475   chloroplast genomes with geographic locations, rather than taxonomic relationships (Acosta &

476   Premoli 2010). Relationships based on geographic location could be explored as a possible

477   explanation for the proximity of *P. concreta* (accesses from Brazil) with *P. estrellensis* (also

478   from Brazil) and not with non-Brazilian accessions of *P. concreta*. In this case a study of nuclear

479   markers of these taxa would be needed.

480   **Neotropical species**

481   Relationships in the group that includes *P. concreta*, *P. foliosa*, *P. estrellensis* and other

482   species are not well resolved due to the low sequence divergence levels between species found in

483   both plastid and nuclear genes (Russell et al., 2010a; Russell et al., 2010b; Russell et al., 2011;

484   Mytnik-Ejsmont, 2011). Generally, the morphological variation observed in this group is

485   identified as *P. concreta*. Although *P. estrellensis* is considered a valid species on the official

486   plant list of Brazil (Barros et al., 2010), there is no consensus on synonymy with *P. concreta*.

487   This can be seen in the herbarium identifications that sometimes consider them as two distinct

488   species, but sometimes as the same species. The same occurs with *P. foliosa*, a name which

489   would only be correctly applied to plants from the Amazon basin, the Guyana Shield and the

490   West Indies (Peraza-Flores, Fernández-Concha & Romero-González, 2011). This

491   circumscription is not accepted by Mytnik-Ejsmont (2011), who considers *P. estrellensis* and *P.*

492   *foliosa* to be synonymous.

493   Genetic dissimilarity between African and Neotropical tetraploids was reported by

494   Russell et al. (2010) and Russell et al. (2011), but the delimitation *P. estrellensis*, *P. concreta*

495   and *P. foliosa* remained uncertain. According to our results, under a molecular perspective, *P.*

496   *estrellensis* should be considered distinct from *P. concreta*. Moreover, our results do not

497   corroborate the placement in synonymy of *P. estrellensis* and *P. foliosa* as proposed by Mytnik-

498   Ejsmont (2011). In our tree *P. foliosa* forms a highly supported group with some *P. concreta*

499   sequences (from samples collected in Brazil). Finally, although our results indicate a possible

500   separation of Brazilian and African *P. concreta*, the delimitation of this species remains

501   uncertain, considering that there is no generic taxonomic revision that has rigorously analysed

502   the morphological variation in this species. Moreover, considering the reticulated evolution by

503   Russell et al. (2010a), further investigation with nuclear markers would be necessary.

504   Taken together, our analysis suggests that *P. estrellensis* can be considered a distinct

505   species from *P. concreta* and *P. foliosa*, and that Brazilian and African *P. concreta* should

506   probably be treated as different species. Evidence of hybridisation influencing the evolution of *P.*

507   *concreta* (Russell et al., 2010a; Russell et al., 2010b) highlights how importance it will be to also

508   consider bi-parentally inherited nuclear DNA when inferring of phylogenetic relationships

509   between this species and other species of the genus. The placement in synonymy of *P.*

510   *estrellensis* and *P. foliosa* proposed by Mytnik-Ejsmont (2011) was not confirmed by this study.

511   In our results, *P. foliosa* forms a highly supported clade including Brazilian samples of *P.*

512   *concreta*.

513

514   **Implications for Data Requirements**

515   **The entire chloroplast is more useful than a fast subset**

516   By using a relatively large number of chloroplast sequences we were able to resolve the

517   polytomy involving the Neotropical species. But, if on one hand this dataset is promising in the

518   formulation of more robust phylogenetic hypotheses, on the other hand, the complete chloroplast

519   genome sequencing may be costly for systematic projects that consider genera with many species

520   (Särkinen & George, 2013), such as *Polystachya*, which has about 250 species. This was the

521   main motivation for testing how well a smaller sample of the chloroplast genome would perform

522   in phylogenetic inference. This was done in two ways: by duplicating a fast region and analysing

523   multiple copies of this dataset, and by sampling without replacement from all non-coding regions

524   in our alignment.

525   We found that sampling without replacement up to ~9 kb of non-coding sequence (16%

526   of our alignment) was not sufficient to return a robust inference across all nodes. This was in

527   contrast to the analysis of the entire chloroplast and showed that in the case of these samples of

528   *Polystachya*, more data were needed to resolve their relationships. The cost of primers,

529   amplification and Sanger sequencing of only three or four regions begins to exceed that of gene

530   capture of the entire chloroplast. It is therefore more cost effective and produces a more robust

531   result to undertake the collection of the entire chloroplast genome. That said, our 16% sample

532   did resolve some nodes with high support, and other nodes obtained moderate to high support

533   from a few of the replicates. This suggests that these data are on the way to resolving most

534   nodes, but a gradual increase in resolving power occurs as characters are added.

535   Duplicating a single fast region even 16 times, in this case *psbD-trnT* copies totalling ~14

536   kb, failed to achieve a robustly resolved phylogeny. The results for the *psbD-trnT* duplicated

537   analysis was poorer even than that of sampling fewer but more representative characters across

538   the non-coding region (above). It appears that a small sample size (only 870 bp of independent

539   sequence sites) is a serious source of stochastic error in this case. Sampling one versus 16 copies

540   of the same dataset only slightly increased the number of resolved nodes (but still falling short of

541   the number of nodes usually resolved with support by the smaller 16% sample), confirming the

542   limitations of the original dataset. It is likely that the original dataset simply did not contain sites

543   that changed on most branches of the phylogeny during the span of history that we investigated.

544   Numerous previous studies have also examined which regions of the plastid genome are

545   typically evolving faster than others (e.g., Small et al., 1998; Shaw et al., 2005; Shaw et al.,

546   2007; Shaw et al., 2014). Prior to NGS methods, the aim was to identify the "best" regions, when

547   sequencing only a limited number could be afforded in most projects. However, given current

548   technology, we should shift our focus to whether a few of the "best" regions are cost effective

549   compared to using the entire genome, as the latter has become affordable for even small

550   phylogenetic projects.

551   **Homoplastic hairpin inversions affect phylogenetic analysis**

552   One issue raised here that is rarely taken into account in analyses of whole chloroplasts is

553   that sequence patterns at the small scale, namely hairpin inversions of loops, can still have an

554   effect on phylogenetic inference, despite using very large data sets. Our results indicate that at

555   least some of the differences between the trees inferred using entire loops versus down-weighted

556   loops were driven by these hairpin loop inversions. This kind of phylogenetic effect has been

557   observed in other cases, although with smaller data sets (e.g., Kim and Lee, 2005; Joly et al.,

558   2010). If loops invert in a single molecular event (as is currently believed: Kelchner and Wendel,

559   1996; Kim and Lee 2005), such as an intra-molecular recombination, then there is no good

560   reason to use each character state difference found between sequences in the entire loop in an

561   analysis. This simply inflates the phylogenetic impact of a single event, treating it instead as

562   many independent events (corresponding to the number of character state differences in the

563   inversion), as also noted by Kim and Lee (2005). As shown here, a larger data set simply does

564   not give licence to ignore known analytical pitfalls.

565   Together, these findings show that sampling the entire chloroplast, analysed carefully, is

566   a better option than sampling a few (even a dozen or more) fast regions. This is true, at least in

567   *Polystachya*, but a similar result has also found by other studies, such as Parks, Cronn. & Liston

568   (2009) for *Pinus*. Based on cost alone, it seems there is no benefit to be gained by screening the

569   chloroplast for faster markers when there are many short branches in the particular tree, as there

570   are here. Whole chloroplast analyses are likely to be a better way forward than sampling

571   individual chloroplast markers in addressing many phylogenetic questions. If gene capture is

572   used, as it was here, it is also very easy to add probes to unlinked nuclear regions, further

573   increasing the power of this approach as a general solution to the issue of data sampling.

574

575   **CONCLUSIONS**

576   Our results show that significantly increasing the number of nucleotides can be an

577   effective option in the phylogenetic inference of taxonomic challenging taxa, such as the orchid

578   genus *Polystachya*. We generated complete chloroplast sequences of 40 *Polystachya* specimens

579   using a combination of Illumina NGS sequencing and a sequence capture, which solved a

580   notorious polytomy for Neotropical species. Our tests on how well a smaller sample of the

581   chloroplast genome would perform in phylogenetic inference shows that the whole chloroplast is

582   a better option than selecting just a few highly variable markers. Full plastid genomes appear

583   particularly powerful when the expected tree is likely to contain many short branches, but

584   nonetheless need to be analysed with care.

585

586   **ACKNOWLEDGEMENTS**

587   We thank our colleagues in our research groups for discussions, advice and support

588   throughout this project. We also thank the anonymous reviewers whose comments helped

589   improve the paper.

590

591   **REFERENCES**

592   Acosta MC, Premoli AC. 2010. Evidence of chloroplast capture in South American

593   *Nothofagus* (subgenus *Nothofagus*, Nothofagaceae). *Molecular Phylogenetics and Evolution* 54:

594   235-242

595   Barros F, Vinhos F, Rodrigues VT, Barberena FFVA, Fraga CN. 2010. Orchidaceae in:

596   Lista de Espécies da Flora do Brasil. Jardim Botânico do Rio de Janeiro.

597   (http://floradobrasil.jbrj.gov.br/2010/FB012071).

598   Bräutigam A, Gowik U. 2010. What can next generation sequencing do for you? Next

599   generation sequencing as a valuable tool in plant research. *Plant Biology* 12: 831–841

600   Brenan JPM (ed.). 1954. Plants collected by the Vernay Nyasaland expedition of 1946.

601   *Memoirs of the New York Botanical Garden* 9: 1–132

602   Capella-Gutiérrez S, Silla-Martínez J, Gabaldón T. 2009. trimAl: a tool for automated

603   alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25: 1972-1973

604   Carstens BC, Pelletier TA, Reid NM, Satler JD. 2013. How to fail at species delimitation.

605   *Molecular Ecology* 22: 4369–4383

606   Chang C-C, Lin H-C, Lin I-P, Chow T-Y, Chen H-H, Chen W-H, Cheng C-H-, Lin C-Y,

607   Liu S-M, Chang C-C, Chaw S-M. 2006. The chloroplast genome of *Phalaenopsis aphrodite*

608   (Orchidaceae): comparative analysis of evolutionary rate with that of grasses and Its

609   phylogenetic implications. *Molecular Biology and Evolution* 23, 279–291

610   https://doi.org/10.1093/molbev/msj029

611   Christenhusz MJM, Byng JW. 2016. The number of known plants species in the world

612   and its annual increase. *Phytotaxa* 261: 201-217. http://dx.doi.org/10.11646/phytotaxa.261.3.1

613   Cribb PJ. 1978. Studies in the genus *Polystachya* (Orchidaceae) in Africa. *Kew Bulletin*

614   32: 743–766

615   Cronn R, Knaus BJ, Liston AP, Maughan J, Parks M, Syring JV, Udall J. 2012. Targeted

616   enrichment strategies for next-generation plant biology. *American Journal of Botany* 99: 291-311

617   Daniell D, Lee S-B, Saski JGC, Guda TQ-VC, Tomkins J, Jansen RK. 2006. Complete

618   chloroplast genome sequences of *Solanum bulbocastanum*, *Solanum lycopersicum* and

619   comparative analyses with other Solanaceae genomes. *Theoretical and Applied Genetics* 112:

620   1503–1518

621   Dressler RL. 1990. The Orchids: natural historyand classification. Cambridge: Harvard

622   University Press.

623   Dressler RL. 1993. Phylogeny and classification of the orchid family. Cambridge:

624   Cambridge University Press.

625   Eaton DAR, Ree RH. 2013. Inferring phylogeny and introgression using RADseq data:

626   an example from 347 flowering plants (*Pedicularis*: Orobanchaceae). *Systematic Biology* 62:

627   689-706

628   Edwards SV. 2009. Is a new and general theory of molecular systematics emerging?

629   *Evolution* 63: 1-19

630   Freudenstein JV, Chase MW. 2015. Phylogenetic relationships in Epidendroideae

631   (Orchidaceae), one of the great flowering plant radiations: progressive specialization and

632   diversification. *Annals of Botany* 115: 665-681

633   Górniak M, Paun O, Chase MW. 2010. Phylogenetic relationships within Orchidaceae

634   based on a low-copy nuclear coding gene, Xdh: Congruence with organellar and nuclear

635   ribosomal DNA results. *Molecular Phylogenetics and Evolution* 56: 784–795

636   Gustafsson AL, Verola CF., Antonelli A. 2010. Reassessing the temporal evolution of

637   orchids with new fossil and a Bayesian relaxed clock, with implications for the diversification of

638   the rare South American genus *Hoffmanseggella* (Orchidaceae: Epidendroideae). *Evolutionary*

639   *Biology* 10:177

640   Joly S, Pfeil BE, Oxelman B, McLenachan PA, Lockhart PJ. 2010. A statistical approach

641   for distinguishing hybridisation and incomplete lineage sorting: correction. *American Naturalist*

642   175: 621-622

643   Kamneva OK, Syring J, Liston A, Rosenberg NA. 2017. Evaluating allopolyploid origins

644   in strawberries (*Fragaria*) using haplotypes generated from target capture sequencing. *BMC*

645   *Evolutionary Biology* 17(1):180. doi: 10.1186/s12862-017-1019-7

646   Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7:

647   improvements in performance and usability. *Molecular Biology and Evolution* 30: 772-780

648   Kelchner SA. 2002. Group II introns as phylogenetic tools: structure, function, and

649   evolutionary constraints. *American Journal of Botany* 89: 1651-1669

650   Kelchner SA, Wendel JF. 1996. Hairpins create minute inversions in non-coding regions
651   of chloroplast DNA. *Current Genetics* 30: 259-262

652   Kim K-J, Lee H-L. 2005. Widespread occurrence of small inversions in the chloroplast
653   genomes of land plants. *Molecules & Cells* 19: 104-113

654   Kraenzlin F. 1926. Monographie der Gattung Polystachya Hook. Repertorium Specierum
655   Novarum Regni Vegetabilis, Beihefte 39: 1-136

656   Lemmon A, Emme S, Lemmon E. 2012. Anchored hybrid enrichment for massively
657   high-throughput phylogenomics. *Systematic Biology* 61:727-744

658   Lemmon EM, Lemmon AR. 2013. High-throughput genomic data in systematics and
659   phylogenetics. *Annual Review of Ecology, Evolution and Systematics* 44: 99-121

660   Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G
661   Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078–
662   2079

663   Li C, Hofreiter M, Straube N, Corrigan S, Naylor GJP. 2013. Capturing protein-coding
664   genes across highly divergente species. *BioTechniques*, 54: 321-326

665   Luo J, Hou BW, Niu ZT, Liu W, Xue QY, Ding XY. 2014. Comparative chloroplast
666   genomes of photosynthetic orchids: insights into evolution of the Orchidaceae and development
667   of molecular markers for phylogenetic applications. *PLoS One* 9: e99016

668   Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E,
669   Shendure J, Turner DJ. 2010. Target-enrichment strategies for next generation sequencing.
670   *Nature Methods* 7: 111-118

671   Marshall DC. 2010. Cryptic failure of partitioned Bayesian phylogenetic analyses: lost in
672   the land of long trees. *Systematic Biology* 59: 108-117

673   McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT. 2013. Applications
674   of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics*
675   *and Evolution* 66: 526–538

676   Michel F, Umesono K, Ozeki H. 1989. Comparative and functional anatomy of group II
677   catalytic introns - a review. *Gene* 82: 5-30

678   Mort ME, Soltis DE, Soltis PS, Francisco-Ortega J, Santos-Guerra A. 2002.
679   Phylogenetics and evolution of the Macaronesian clade of Crassulaceae inferred from nuclear
680   and chloroplast sequence data. *Systematic Botany* 27: 271-288

681 Mytnik-Ejsmont J. 2011. A Monograph of the subtribe Polystachyinae Schltr.
682 (Orchidaceae). Koenigstein: Koeltz Scientific Books.

683 Parks M, Cronn R, Liston A. 2009. Increasing phylogenetic resolution at low taxonomic
684 levels using massively parallel sequencing of chloroplast genomes. *BMC Biology* 7:84

685 Peraza-Flores LN, Fernández-Concha GC, Romero-González GA. 2011. Taxonomic
686 notes in American *Polystachya* (Orchidaceae): the identity of *P. foliosa* (Hook.) Rchb.f. and the
687 reestablishment of *P. caracasana* Rchb.f. *Journal of the Torrey Botanical Society* 138: 366–380

688 Pfeil BE, Brubaker CL, Craven LA, Crisp MD. 2002. Phylogeny of Hibiscus and the tribe
689 Hibisceae (Malvaceae) using chloroplast DNA sequences of ndhF and the rpl16 intron.
690 *Systematic Botany* 27: 333-350

691 Pridgeon AM, Cribb PJ, Chase MW, Rasmussen FN (eds.). 2005. Genera
692 Orchidacearum, vol. 4, Epidendroideae, part 1. Oxford: Oxford University Press.

693 Rambaut A, Drummond AJ. 2010. TreeAnnotator version 1.6.1 http://beast.bio.ed.ac.uk

694 Rambaut A, Suchard MA, Xie D & Drummond AJ (2014) Tracer v1.6, Available from
695 http://tree.bio.ed.ac.uk/software/tracer/

696 Ramírez SR, Gravendeel B, Singer RB, Marshall CR, Pierce NE. 2007. Dating the origin
697 of the Orchidaceae from a fossil orchid with its pollinator. *Nature* 448: 1042-1045.
698 doi:10.1038/nature06039

699 Ramírez SR, Eltz T, Fujiwara MK, Gerlach G, Goldman-Huertas B, Tsutsui ND,
700 PierceNE. 2011. Asynchronous Diversification in a Specialized Plant-Pollinator Mutualism.
701 *Science* 333: 1742-1746. DOI: 10.1126/science.1209175

702 Ronquist F, Teslenko M, van der Mark P, Ayres D, Darling A, Höhna S, Larget B, Liu L,
703 Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: Efficient Bayesian phylogenetic inference
704 and model choice across a large model space. *Systematic Biology* 61:539-542

705 Rupp B, Samuel R, Russell A, Temsch EM, Chase MW, Leitch IJ. 2010. Genome size in
706 *Polystachya* (Orchidaceae) and its relationships to epidermal characters. *Botanical Journal of the*
707 *Linnean Society* 163: 223–233

708 Russell A, Samuel R, Klejna V, Barfuss MHJ, Rupp B, Chase MW. 2010 (a). Reticulate
709 evolution in diploid and tetraploid species of *Polystachya* (Orchidaceae) as shown by plastid
710 DNA sequences and low-copy nuclear genes. *Annals of Botany* 106: 37–56

711   Russell A, Samuel MR, Rupp B, Barfuss MHJ, Šafran M, Besendorfer V, Chase MW.

712   2010 (b). Phylogenetics and cytology of a pantropical orchid genus *Polystachya* (Polystachyinae;

713   Vandeae; Orchidaceae): evidence from plastid DNA sequence data. *Taxon* 59: 389-404

714   Russell A, Samuel R, Bogarín D, Fernando S, Wijesundera S, Klejna V, Chase MW.

715   2011. Genetic variation and phylogenetic relationships of a pantropical species group in

716   *Polystachya* (Orchidaceae). *Botanical Journal of the Linnean Society* 165: 235–250

717   Särkinen T, George M. 2013. Predicting plastid marker variation: can complete plastid

718   genomes from closely related species help? *PLoS One* 8: e82266

719   https://doi.org/10.1371/journal.pone.0082266

720   Shaw J, Lickey EB, Beck JT, Farmer SB, Liu W, Miller J, Siripun KC, Winder CT,

721   Schilling EE, Small RL. 2005. The tortoise and the hare II: relative utility of 21 noncoding

722   chloroplast DNA sequences for phylogenetic analysis. *American Journal of Botany* 92:142-166

723   Shaw J, Lickey EB, Schilling EE, Small RL. 2007. Comparison of whole chloroplast

724   genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the

725   tortoise and the hare III. *American Journal of Botany* 94:275-288

726   Shaw J, Shafer HL, Leonard OR, Kovach MJ, Schorr M, Morris AB. 2014. Chloroplast

727   DNA sequence utility for the lowest phylogenetic and phylogeographic inferences in

728   angiosperms: the tortoise and the hare IV. *American Journal of Botany* 101:1987-2004

729   Small RL, Ryburn JA, Cronn RC, Seelanan T, Wendel JF. 1998. The tortoise and the

730   hare: choosing between noncoding plastome and nuclear Adh sequences for phylogeny

731   reconstruction in a recently diverged plant group. *American Journal of Botany* 85:1301-1315

732   Smith RL, Sytsma KJ. 1990. Evolution of *Populusnigra* (sect.Aigeiros): introgressive

733   hybridisation and the chloroplast contribution of *Populusalba* (sect. Populus). *American Journal*

734   *of Botany* 77(9): 1176-1187

735   Song S, Liu L, Edwards SV, Wu S. 2012. Resolving conflict in eutherian mammal

736   phylogeny using phylogenomics and the multispecies coalescent model. *Proceedings of the*

737   *National Academy of Sciences, USA* 109: 14942-14947. doi: 10.1073/pnas.1211733109

738   Sousa F, Bertrand YKJ, Nylinder S, Oxelman B, Eriksson JS, Pfeil BE. 2014.

739   Phylogenetic properties of 50 nuclear loci in *Medicago* (Leguminosae) generated using

740   multiplexed sequence capture and next-generation sequencing. *PLoS One* 9(10):e109704. doi:

741   10.1371/journal.pone.0109704

742   Steel, M, Penny, D. 2000. Parsimony, likelihood, and the role of models in molecular

743   phylogenetics. *Molecular Biology and Evolution* 17:839-850.

744   doi:10.1093/oxfordjournals.molbev.a026364

745   Stephens JD, Rogers WL, Heyduk K, Cruse-Sanders JM, Determann RO, Glenn TC,

746   Malmberg RL. 2015. Resolving phylogenetic relationships of the recently radiated carnivorous

747   plant genus *Sarracenia* using target enrichment. *Molecular Phylogenetics & Evolution* 85:76-87.

748   doi: 10.1016/j.ympev.2015.01.015

749   Summerhayes VS. 1942. African orchids: XII. *Botanical Museum Leaflets* 10: 257-299

750   Summerhayes VS. 1947. African orchids: XVII. *Kew Bulletin* 2: 123-133

751   Timme RE, Kuehl JV, Boore JL, Jansen RK. 2007. A comparative analysis of the

752   *Lactuca* and *Helianthus* (Asteraceae) plastid genomes: Identification of divergent regions and

753   categorization of shared repeats. *American Journal of Botany* 94: 302-312

754   Toor N, Hausner G, Zimmerly S. 2001. Coevolution of group II intron RNA structures

755   with their intron-encoded reverse transcriptases. *RNA* 7: 1142-1152

756   Tsitrone A, Kirkpatrick M, Levin DA. 2003. A model for chloroplast capture. *Evolution*

757   57: 1776–1782

758   Van den Berg C, Goldman DH, Freudenstein JV, Pridgeon AM, Cameron KM, Chase

759   MW. 2005. An overview of the phylogenetic relationships within Epidendroideae inferred from

760   multiple DNA regions and recircumscription of Epidendreae and Arethuseae (Orchidaceae).

761   *American Journal of Botany* 92: 613-624Weitemier K, Straub SCK, Cronn RC, Fishbein M,

762   Schmickl R, McDonnell A, Liston A. 2014. Hyb-Seq: Combining target enrichment and genome

763   skimming for plant phylogenomics. *Applications in Plant Sciences* 2(9), 1400042

Figure 1

Distribution of *Polystachya* and the location of samples used in this study.

Grey shading shows the distribution of the genus. Coloured symbols show the location of samples used here, with the species determination of each sample as per Table 1.

🔴 *Polystachya concreta* 🟣 *Polystachya golungensis* 🟧 *Polystachya paniculata*
🟣 *Polystachya cornigera* ⚫ *Polystachya henrici* 🔺 *Polystachya pinicola*
🟡 *Polystachya dolicophylla* 🟨 *Polystachya humbertii* 🔺 *Polystachya steudneri*
🟢 *Polystachya estrellensis* 🟩 *Polystachya melanantha* 🔺 *Polystachya tessellata*
🟠 *Polystachya eurychila* 🟦 *Polystachya odorata* 🔺 *Polystachya tsinjoarivensis*
🔵 *Polystachya foliosa* 🟪 *Polystachya oreocharis*

Figure 2

Phylogeny of the psbD-trnT region estimated using Bayesian analysis and rooted using *Phalaenopsis aphrodite*.

Posterior probabilities are show above branches. Scale bar is in substitutions per site. The two branches leading to the root have been foreshortened to reduce space and are thus not to scale. *(A)* Phylogeny based on a single copy of *psbD-trnT*. *(B)* Phylogeny based on 16 identical copies of the *psbD-trnT* data set.

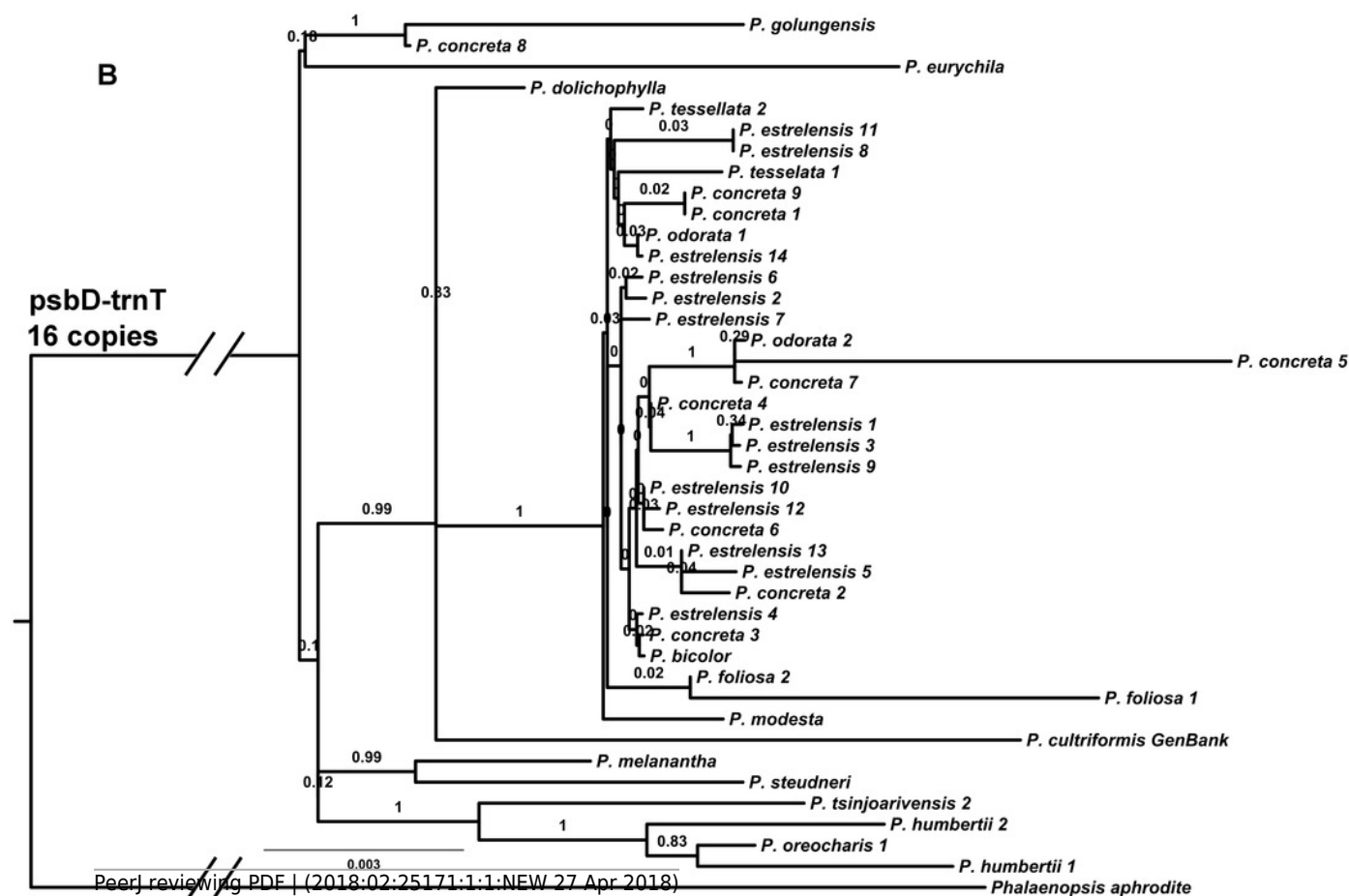
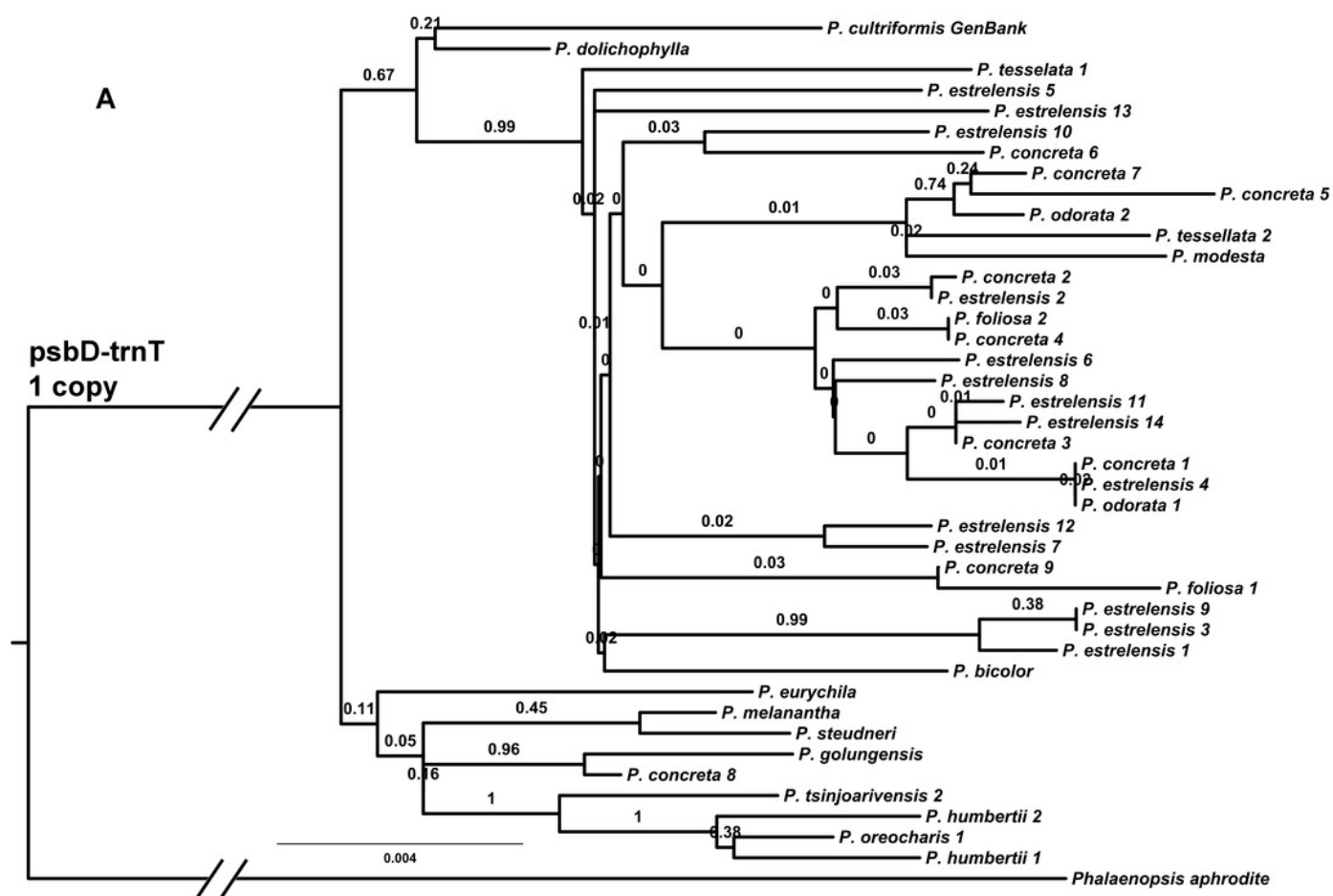



Figure 3

Plastid phylogeny from Bayesian analysis rooted using *Phalaenopsis aphrodite*.

Posterior probabilities are show above branches. The *Polystachya estrelensis* group has been collapsed to reduce detail. Scale bar is in substitutions per site. The two branches leading to the root have been foreshortened to reduce space and are thus not to scale. The two insets are not at the same scale as the main figure. *Main Figure*: Phylogeny based on the data set with poorly aligned regions excluded and loops down-weighted. *(A)* (in grey) Phylogeny of the foliosa/concreta group based on the full inclusion of loops. *(B)* (in black) NeighborNet network of the foliosa/concreta group based on the down-weighted loops. *Polystachya estrelensis* photo credit: N. Lopes de Abreu.

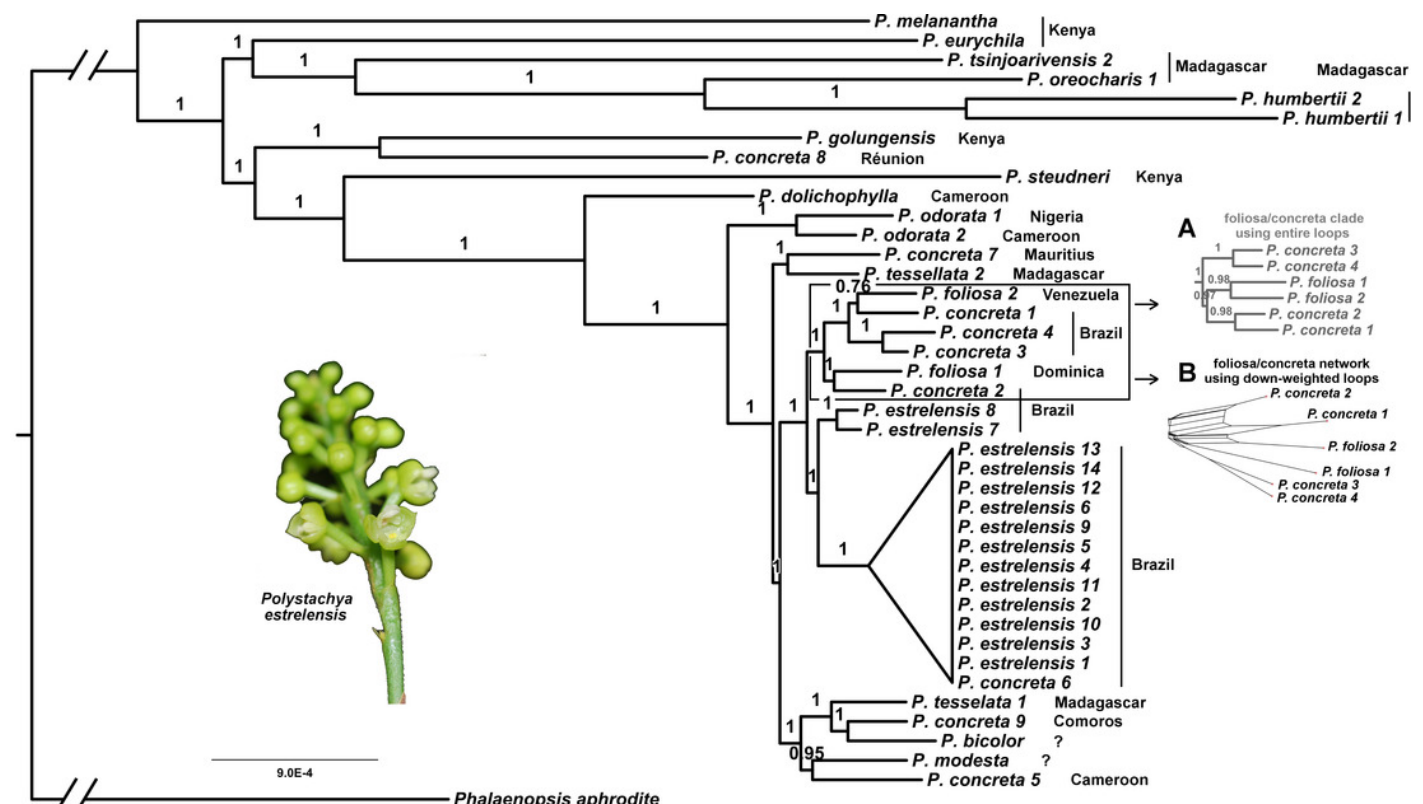
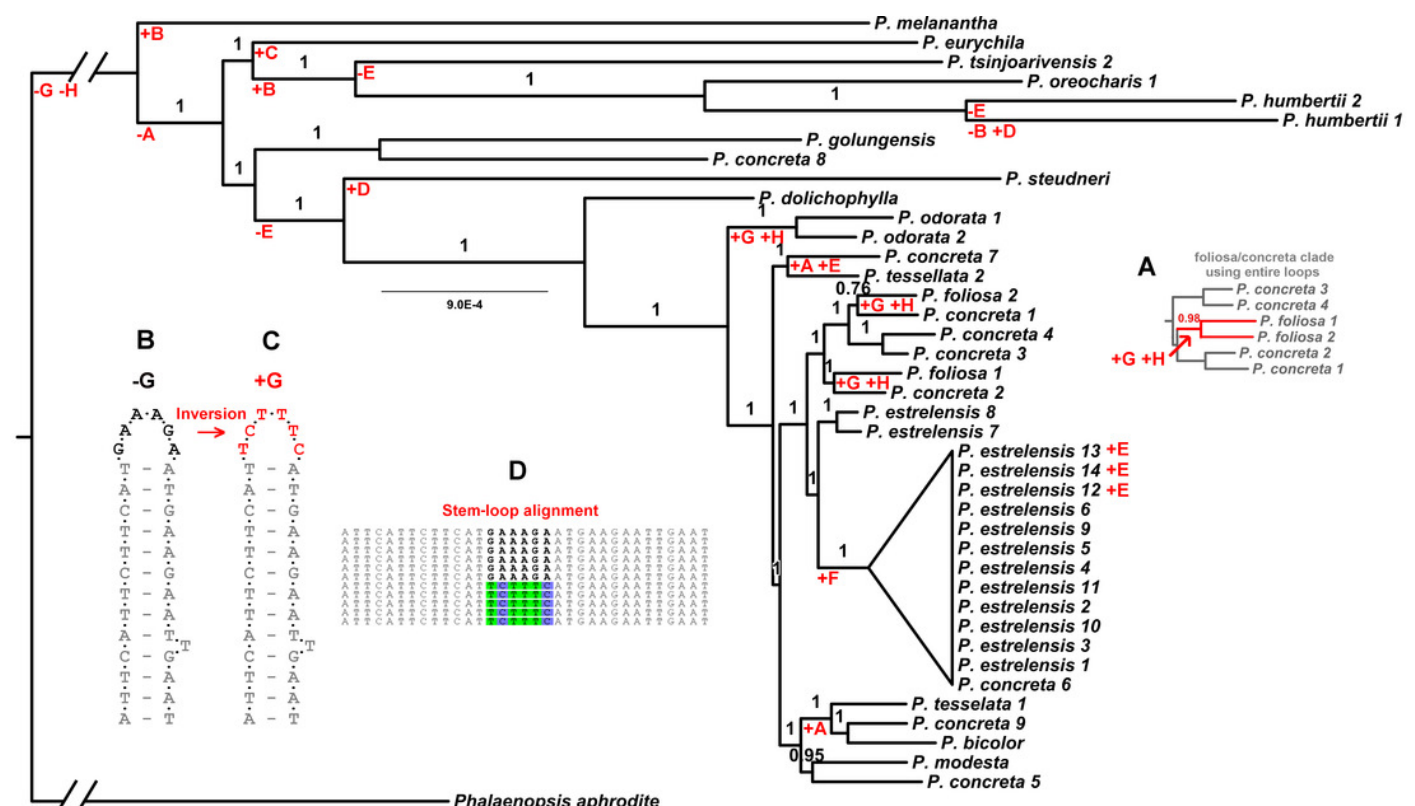



Figure 4

Parsimonious gains and losses of non-coding loop inversions in *Polystachya* relative to the outgroup sequence, *Phalaenopsis aphrodite*, mapped on to the plastid phylogeny.

The letter codes designate loops as per Table 2. *Main Figure*: Phylogeny estimated using down-weighted loops (from the main panel, Fig. 1). Where equally parsimonious interpretations were possible, accelerated transformation has been used. *(A)* Part of the phylogeny estimated using entire loops for the foliosa/concreta group. The mapping of gains of two loop inversions shared by foliosa1 and foliosa2 on this topology is in constrast to the mapping on the topology using down-weighted loops (main figure). *(B)* A diagramatic representation of the stem-loop structure with the majority form of the loop sequence (in black). *(C)* The stem-loop structure with the proposed inversion of the loop sequence (in red). *(D)* The consequence on the alignment before down-weighting of the loop sequence (loop seqeunce in bold – majority form; loop sequence with back colours – inverted form).

Species and sample information of the accessions used in the study.

Species analysed; location of the collection and voucher; DNA concentration and purity before and after genomic library assembly.

1 **Table 1: Species and sample information of the accessions used in the study.** Species analysed; location of the collection and
2 voucher; DNA concentration and purity before and after genomic library assembly.

SPECIES	LOCATION	VOUCHER	ENA REFERENCE	BEFORE GENOMIC LIBRARY ASSEMBLY			AFTER GENOMIC LIBRARY ASSEMBLY	
				Concentration (ng/uL)	Purity 260/280	Volume	Concentration (ng/uL)	Purity 260/280
P. bicolor (=P.rosea)	-	Kew 25884		20.1	1.51	50	31.7	1.93
P. concreta (1)	*Brazil, Distrito Federal*	N.L.Abreu 254		39.2	1.92	75	17.2	1.92
P. concreta (2)	*Brazil, Distrito Federal*	N.L.Abreu 254		33.6	1.96	75	31.9	1.9
P. concreta (3)	*Brazil, Bahia State*	N.L.Abreu 251		18	1.78	70	29.9	1.97
P. concreta (4)	*Brazil, Bahia State*	N.L.Abreu 251		8.5	1.52	130	38.3	1.9
P. concreta (5)	*Cameroon*	A. Russell 40 (YA)		35	1.82	8	15	1.75
P. concreta (6)	*Brazil*	HBV ORCH 066004		21.9	1.53	6	9.8	1.9
P. concreta (7)	*Mauritius*	HBV ORCH 07278		26	1.78	13	32.4	1.85
P. concreta (8)	*Réunion*	HBV "Chase & Samuel 1"		34	1.82	13	30	1.85
P. concreta (9)	*Comoros*	HBV ORCH 07417		23.7	1.83	13	30.4	1.86
**P. concreta (10)*	*Madagascar*	Fischer&Sieder FS3210 (WU)		21	1.81	13	9.4	1.63
**P. concreta (11)*	*Madagascar*	Kew 17854		70	1.65	67	9.6	1.6
**P. cornigera*	*Madagascar*	Fischer&Sieder FS3208 (WU)		22.2	1.76	11	7.7	1.58
P. dolichophylla	*Cameroon*	Kew 25886		108.9	1.05	55	30.3	1.89
P. estrellensis (1)	*Brazil, Minas Gerais State*	N.L.Abreu 255		29.8	1.81	65	33.7	1.9
P. estrellensis (2)	*Brazil, Minas Gerais State*	N.L.Abreu 255		17.8	2.06	90	26	1.54
P. estrellensis (3)	*Brazil, Minas Gerais State*	N.L.Abreu 255		11.9	1.82	75	35	1.86
P. estrellensis (4)	*Brazil, Espírito Santo State*	N.L.Abreu 253		12.8	1.77	65	18.7	1.84
P. estrellensis (5)	*Brazil, Espírito Santo State*	N.L.Abreu 253		22.4	1.87	70	21	1.96
P. estrellensis (6)	*Brazil, Espírito Santo State*	N.L.Abreu 253		28.2	1.79	70	26.3	1.96
P. estrellensis (7)	*Brazil, Bahia State*	N.L.Abreu 252		25.4	1.91	70	29.3	1.98

P. estrellensis (8)	Brazil, Bahia State	N.L.Abreu 252		14.6	1.79	70	22.4	2.02
P. estrellensis (9)	Brazil, Bahia State	N.L.Abreu 256		27.8	1.61	90	20.4	2.05
P. estrellensis (10)	Brazil, Bahia State	N.L.Abreu 256		23.8	1.81	75	23.1	1.93
P. estrellensis (11)	Brazil, Bahia State	N.L.Abreu 256		22.3	1.68	90	28	1.97
P. estrellensis (12)	Brazil, Trindade's Island	Nilber		10.1	1.64	130	30	1.88
P. estrellensis (13)	Brazil, Trindade's Island	Nilber		12.5	1.64	130	30	1.83
P. estrellensis (14)	Brazil, Trindade's Island	Nilber		10.3	1.48	130	34.7	1.84
P. eurychila	Kenya	Kew 17963		207.6	0.94	85	13.7	1.98
P. foliosa (1)	Dominica	Kew 25887		14.7	1.58	56	31.8	1.96
P. foliosa (2)	Venezuela	HBV ORCH 07082		30.2	1.73	8	18.6	1.79
P. golungensis	Kenya	Kew 17966		104.7	1.14	115	26.1	1.88
*P. henrici	Madagascar	Kew 17856		22	1.52	56	7.7	2.04
P. humbertii (1)	Madagascar	Fischer&Sieder FS2079 (WU)		116.3	1.82	13	29.4	1.77
P. humbertii (2)	Madagascar	Fischer&Sieder FS3017 (WU)		35.1	1.93	11	34.2	1.89
P. melanantha	Kenya	Kew 17954		207.2	0.94	50	12.2	1.8
P. modesta	-	HBV ORCH 05165		56.8	1.35	13	31.7	1.93
P. odorata (1)	Nigeria	Kew 17857		33.7	1.56	58	46.8	1.8
P. odorata (2)	-Cameroon	A. Russell 42 (YA)		36.9	1.91	13	26.2	1.91
P. oreocharis (1)	Madagascar	Fischer&Sieder FS2082 (WU)		58	1.74	13	30.2	1.87
*P. oreocharis (2)	Madagascar	Fischer&Sieder FS3152 (WU)		20.2	1.82	13	10.5	1.8
*P. paniculata (2)	Cameroon	L. Pearce 27 (YA)		56.1	1.4	8	8.1	1.98
*P. pinicola	Barzil	HBV ORCH 06606		31	1.11	7	9.2	1.98
P. steudneri	Kenya	Kew 17956		10.6	1.45	58	33.5	1.87
P. tessellata (1) (=P.concreta)	Madagascar	Kew 17859		176.2	1.1	58	32.7	1.89
P. tessellata (2) (=P.concreta)	Madagascar	Kew 17860		216.1	1.1	49	33.3	1.93
*P. tsinjoarivensis (1)	Madagascar	Fischer&Sieder FS3209 (WU)		19.3	1.83	13	9.3	1.79
P. tsinjoarivensis (2)	Madagascar	HBV FS4182		22.2	1.75	13	17.2	1.64

4   *Species excluded due to low quality sequencing.

Alignment co-ordinates and sequences of putative stem-loop structures that appear to have undergone inversion in *Polystachya*.

The loops that appear to have undergone inversions are underlined. Bases in red would be treated as phylogenetically informative if the alignment doesn't take into account the inversion. Bases not involved in stem formation are in subscript.

1 **Table 2: Alignment co-ordinates and sequences of putative stem-loop structures that appear to have undergone inversion in**
2 ***Polystachya***. The loops that appear to have undergone inversions are underlined. Bases in red would be treated as phylogenetically
3 informative if the alignment doesn't take into account the inversion. Bases not involved in stem formation are in subscript.

Loop	Alignment co-ordinates	Sequence (majority)	Sequence (minority)	Substitutions before/after inversion reverted	Sequences (minority)
A	74,660-74,678	TCTATCTA-GAA-TAGATAGA	TCTATCTA-TTC-TAGATAGA	3/0	bicolor, concreta7, concreta9, tesselata1, melanantha, Phalaenopsis
B	88,106-88,142	GGCCCAATCTTT$_C$-TTTTTTTGAGGA-AAAGATTGGGCC	GGCCCAATCTTT$_C$-TCCTCAAAAAAA-AAAGATTGGGCC	8/0	humbertii2, oreocharis1, tsinjoarivensis2, melanantha
C	90,456-90,496	AGTAAGAACTCAGCG-GGGTAAGGCCT-CGCTGAGTTCTTACT	AGTAAGAACTCAGCG-AGGCCTTACCC-CGCTGAGTTCTTACT*	7/0	eurychila
D	94,175-94,222	ATTGAAGTAATGAGCCC-CAAGATGAATATGA-GGGCTCATTACTTCAAT	ATTGAAGTAATGAGCCC-TCATATTCATCTTG-GGGCTCATTACTTCAAT	8/0	humbertii1, steudneri
E	94,679-94,717	GTATCTAAGGAAGATCC-AAAGG-GGATCTTCCTTAGATAC	GTATCTAAGGAAGATCC-CTTCT-GGATCTTCCTTAGATAC	5/2	estrelensis12, estrelensis13, estrelensis14, eurychila, concreta7, concreta8, golungensis, oreocharis1, melanantha, humbertii1, Phalaenopsis
F	97,230-97,258	AACGTCCAGTG-CCAAAGT-CACTGAATGGG	^CCCATTCAGTG-ACTTTGG-CACTGAATGGG	6/0	All 13 sequences of the estrelensis clade (cartooned in Figs. 2 and 4)
G	104,698-104,730	ATTC$_A$ATTCTTC-ATTCTTTCAA-GAAGAATGAAT	ATTC$_A$ATTCTTC-ATGAAAGAAA-GAAGAATGAAT	6/0	foliosa1, foliosa2, odorata1, odorata2, Phalaenopsis
H	106,618-106,650	ATTCATTCTTCAT-GAAAGA-ATGAAGAAT$_T$GAAT	ATTCATTCTTCAT-TCTTTC-ATGAAGAAT$_T$GAAT	6/0	foliosa1, foliosa2, odorata1, odorata2, Phalaenopsis

5 *Occurs in only one sequence. ^Only the sequences with the inversion appear to form the stem in this case.