

1 **ORIGINAL ARTICLE**

2 **DNA barcoding British *Euphrasia* reveals deeply divergent polyploids but lack of**
3 **species-level resolution**

4 Xumei Wang¹, Galina Gussarova^{2,3,4}, Markus Ruhsam⁵, Natasha de Vere⁶, Chris Metherell⁷,
5 Peter M. Hollingsworth⁵, Alex D. Twyford^{8*}

6 *1* School of Pharmacy, Xi'an Jiaotong University, 76 Yanta West Road, Xi'an, 710061,
7 P.R.China

8 *2* Tromsø University Museum, UiT The Arctic University of Norway, PO box 6050 Langnes,
9 NO-9037, Tromsø, Norway

10 *3* CEES-Centre for Ecological and Evolutionary Synthesis, Dept. of Biosciences, University
11 of Oslo, PO BOX 1066, Blindern, NO-0316, Norway

12 *4* Dept. of Botany, Faculty of Biology, St Petersburg State University, Universitetskaya nab.
13 7/9, 199034 St Petersburg, Russia

14 *5* Royal Botanic Garden Edinburgh, 20A Inverleith Row, Edinburgh, EH3 5LR, UK

15 *6* National Botanic Garden of Wales, Llanarthne, Carmarthenshire, SA32 8HG, UK and

16 Institute of Biological, Environmental & Rural Sciences (IBERS), Aberystwyth University,
17 Penglais, Aberystwyth, Ceredigion, SY23 3DA, UK

18 *7* 57 Walton Rd, Bristol BS11 9TA

19 *8* University of Edinburgh, Institute of Evolutionary Biology, West Mains Road, Edinburgh,
20 EH3 9JT, UK

21

22 **Running title:** Complex patterns of diversity in *Euphrasia*

23 For correspondence: Alex.Twyford@ed.ac.uk

24

25

26 **ABSTRACT**

27 **Background and aims** DNA barcoding is emerging as a useful tool not only for species
28 identification but for studying evolutionary and ecological processes. Although plant DNA
29 barcodes do not always provide species-level resolution, the generation of large DNA
30 barcode datasets can provide insights into the mechanisms underlying the generation of
31 species diversity. Here, we use DNA barcoding to study evolutionary processes in
32 taxonomically complex British *Euphrasia*, a group with multiple ploidy levels, frequent self-
33 fertilization, young species divergence and widespread hybridisation.

34 **Methods** We sequenced the core plant barcoding loci, supplemented with additional nuclear
35 and plastid loci, in representatives of all 19 British *Euphrasia* species. We analyse these data
36 in a population genetic and phylogenetic framework. We then date the divergence of
37 haplotypes in a global *Euphrasia* dataset using a time-calibrated Bayesian approach
38 implemented in BEAST.

39 **Key results** No *Euphrasia* species has a consistent diagnostic haplotype. Instead, haplotypes
40 are either widespread across species, or are population specific. Nuclear genetic variation is
41 strongly partitioned by ploidy levels, with diploid and tetraploid British *Euphrasia* possessing
42 deeply divergent ITS haplotypes ($D_{XY} = 5.1\%$), with haplotype divergence corresponding to
43 the late Miocene. In contrast, plastid data show no clear division by ploidy, and instead reveal
44 weakly supported geographic patterns.

45 **Conclusions** Using standard DNA barcoding loci for species identification in *Euphrasia* will
46 be unsuccessful. However, these loci provide key insights into the maintenance of genetic
47 variation, with divergence of diploids and tetraploids suggesting that ploidy differences act as
48 a barrier to gene exchange in British *Euphrasia*, with rampant hybridisation within ploidy
49 levels. The scarcity of shared diploid-tetraploid ITS haplotypes supports the polyploids being

50 allotetraploid in origin. Overall, these results show that even when lacking species-level
51 resolution, DNA barcoding can reveal insightful evolutionary patterns in taxonomically
52 complex genera.

53

54 **Keywords:** DNA barcoding, *Euphrasia*, Orobanchaceae, polyploidy, taxonomic complexity,

55 British flora, phylogeny

56

57

58 **INTRODUCTION**

59

60 DNA barcoding is a valuable tool for discriminating among species, and these data often give
61 insights into identity that are overlooked based on morphology alone (Hebert and Gregory,
62 2005). Plant DNA barcoding has been used for species discovery, reconstructing historical
63 vegetation types from frozen sediments, surveying environmental variation, and many other
64 research topics (reviewed in Hollingsworth et al., 2016). Despite the extensive uptake of
65 DNA barcoding, there are numerous reports of taxon groups where DNA barcodes do not
66 provide exact plant species identification, and where DNA barcode sequences are shared
67 among related species (Percy et al., 2014, Yan et al., 2015a, Yan et al., 2015b, Spooner,
68 2009). Identifying the causes of why DNA barcoding ‘fails’ is essential to help guide the
69 development of future DNA barcoding systems. If species discrimination is usually limited
70 by information content of the core barcoding loci, then improvements can be made by
71 extending the DNA barcode to harness variation across the whole plastid genome. In contrast,
72 if limits to species discrimination are caused by hybridisation, incomplete lineage sorting, and
73 polyploidy, then we may see no improvement with more plastid data, and future barcoding
74 systems will need to target the nuclear genome, as well as use new analytical methods to cope
75 with many independent loci (Coissac et al., 2016, Hollingsworth et al., 2016). As such, more
76 genetic studies of complex taxa groups are required in order to identify the underlying
77 evolutionary processes that can cause DNA barcoding to fail. In addition, the generation of
78 large data sets of DNA sequence data from multiple individuals of multiple species in such
79 groups also provide datasets that can shed light onto evolutionary relationships and
80 evolutionary divergence, without a need for the barcode markers to track species boundaries.

81

82 Postglacial species radiations of taxonomically complex groups in Northern Europe are a case
83 where we may not expect a clear cut-off between intraspecific variation and interspecific
84 divergence and thus DNA barcoding may provide limited discriminatory power. Despite this,
85 DNA barcoding may still be valuable if used to identify evolutionary and ecological
86 processes that result in shared sequence variation. For example, many postglacial taxa are
87 characterised by a combination of: (1) recent postglacial speciation, (2) extensive
88 hybridisation, (3) frequent self-fertilization, (4) divergence involving polyploidy. Our
89 expectation is that factors 1 + 2 will cause DNA barcode sequences to be shared among
90 geographically proximate taxa, while 3 will cause barcodes to be population rather than
91 species-specific (Hollingsworth et al., 2011). Factor 4, polyploidy, will manifest as shared
92 haplotypes between recent polyploids and their parental progenitors, or deep haplotype
93 divergence in older polyploid groups, where ploidy act as a reproductive isolating barrier and
94 allows congeneric taxa to accumulate genetic differences. Many of these interacting factors
95 are common across taxonomically complex postglacial groups (Ennos, 2005), which include
96 the *Arabidopsis arenosa* complex (Schmickl et al., 2012), *Cerastium* (Brysting et al., 2007),
97 *Epipactis* (Squirrell et al., 2002) and *Galium* (Kolář et al., 2015). As such, the application of
98 DNA barcoding to such groups will not only reveal the prevalence of haplotype sharing and
99 the potential for species discrimination, but improve our understanding of processes
100 underlying shared variation.

101

102 One example of a taxonomically challenging group showing postglacial divergence is British
103 *Euphrasia* species. This group of 19 taxa are renowned for their difficult species
104 identification, and at present only a handful of experts can identify these species in the field.
105 Morphological species identification is difficult due to their small stature, combined with
106 species being defined by a complex suite of overlapping characters (Yeo, 1978). They are

107 also generalist hemiparasites and thus phenotypes are plastic and depend upon host quality
108 (Svensson and Carlsson, 2004). DNA barcode-based identification could partly resolve these
109 identification issues, and lead to a greater understanding of species diversity and distributions
110 in this under-recorded group. This is particularly important as a number of *Euphrasia* species
111 are critically rare and of conservation concern, while others are ecological specialists that are
112 useful indicators of habitat type (French et al., 2008). More generally, DNA barcoding data
113 could reveal the processes structuring genetic diversity and those that are responsible for
114 recent speciation.

115

116 Previous broad-scale surveys of *Euphrasia* using AFLP and microsatellites have shown a
117 significant proportion of genetic variation is partitioned by ploidy groups and by species,
118 despite extensive hybridisation (French et al., 2008). Here, we follow-on from previous
119 population genetic studies by testing the utility of DNA barcoding across British *Euphrasia*.
120 Our first aim is to assess whether DNA barcoding is informative for species recognition in
121 this young postglacial group. Our second aim is to understand the evolutionary factors that
122 may explain patterns of shared DNA sequences. We first sequence British populations for the
123 standard DNA barcoding loci, and analyse the distribution of haplotypes across populations
124 and species. To better understand the historical context of haplotype sharing we sequence
125 additional loci and use dated phylogenetic analyses to infer divergence ages. Overall, these
126 results are used to understand the efficacy of genetic tools for studying species-level variation
127 in a taxonomically complex group.

128

129 **MATERIALS AND METHODS**

130

131 *Specimen sampling*

132

133 The 19 currently recognised British *Euphrasia* species are all annuals, selfers or mixed-
134 mating small herbaceous plants, which occur in a range of habitats including coastal turf,
135 chalk downland, mountain ridges and heather moorland. The species can be divided into two
136 groups, glabrous or short-glandular hairy tetraploids (15 species, Fig 1A), or long glandular
137 hairy diploids (4 species, Fig. 1B). Our sampling included representatives of all British
138 species, with samples collected from across widespread populations (Fig. 1c, Table S1,
139 **[Supplementary Information]**). Samples were collected in South West England and Wales
140 to allow us to include mixed populations of diploids and tetraploids, early generation diploid
141 x tetraploid hybrids, and two diploid hybrid species hypothesised to be derived from diploid x
142 tetraploid crosses (*E. vigursii*, parentage: *E. rostkoviana* x *E. micrantha*), *E. rivularis*
143 (parentage: *E. anglica* x *E. micrantha*; Yeo, 1956). Samples from Scotland allowed us to
144 sample complex tetraploid taxa and tetraploid hybrids, plus scarcer Scottish diploids. Our
145 sampling scheme investigated range-wide variation by targeting many taxa and populations,
146 but without intrapopulation sampling. This is because prior work has shown low
147 intrapopulation diversity, with populations frequently fixed for single haplotypes (French et
148 al., 2008). All samples were identified by *Euphrasia* experts Alan Silverside or Chris
149 Metherell.

150

151 For our population-level DNA barcoding study, we analysed a total of 133 individuals, with
152 106 samples representing 19 species, as well as 27 samples from 14 putative hybrids. We
153 sequenced samples for the core plant DNA barcoding loci, *matK* and *rbcL* (Hollingsworth et
154 al., 2009), as well as ITS2 (China Plant BOL Group, 2011), and supplemented these loci with
155 *rpl32-trnL^{UAG}*, which has been informative in prior population studies of *Euphrasia* (Stone,
156 2013).

157

158 For our broader molecular phylogenetic and molecular dating analysis, we expanded the
159 sampling in the molecular phylogeny of the genus by Gussarova et al. (2008), to include a
160 detailed sample of British taxa. The previous analysis included 40 taxa for the nuclear
161 ribosomal internal transcribed spacer (ITS), and 50 taxa for plastid DNA (Gussarova et al.,
162 2008). We sequenced samples to match the previous data matrix, which included: the *trnL*
163 intron (Taberlet et al., 1991), intergenic spacers *atpB-rbcL* (Hodges and Arnold, 1994) and
164 *trnL-trnF* (Taberlet et al., 1991), and *ITS* (White et al. 1990).

165

166 *DNA extraction, PCR amplification, and sequencing*

167

168 DNA was extracted from silica dried tissue using the DNeasy Plant Mini kit (Qiagen, Hilden,
169 Germany) following the manufacturer's protocol, but with an extended incubation of 1 hour
170 at 65 °C. These DNA samples were added to existing DNA extractions from 68 individuals
171 from French et al. (2008).

172

173 PCRs were done in two laboratories following separate protocols. We applied the following
174 conditions for most taxa. We performed PCRs in 10 µL reactions, with DNA amplification
175 and PCR conditions for each primer given in Table S2 [**Supplementary Information**]. We
176 visualised PCR products on a 1% agarose gel, with 5 µL of PCR product purified for
177 sequencing with ExoSAP-IT (USB Corporation, Cleveland, OH, USA) using standard
178 protocols. Sequencing was performed in 10 µL reactions containing 1.5 µL 5 × BigDye
179 buffer (Life Technologies, Carlsbad, CA, USA), 0.88 µL BigDye enhancing buffer BD × 64
180 (MCLAB, San Francisco, CA, USA), 0.125 µL BigDye v3.1 (Life Technologies), 0.32 µM
181 primer and 1 µM of purified PCR product. We sequenced PCR products on the ABI 3730

182 DNA Analyser (Applied Biosystems, Foster City, CA, USA) at Edinburgh Genomics. A
183 subset of sequences were generated as part of the effort to DNA barcode the UK Flora, and
184 followed a different set of protocols, detailed in de Vere et al. (2012). We assembled, edited
185 and aligned sequences using Geneious v. 8 (Biomatters, Auckland, New Zealand), with
186 manual editing. Indels were coded as unordered binary characters and appended to the
187 matrices. We used gap coding as implemented in GapCoder (Young and Healy, 2003), with
188 indels treated as point mutations and equally weighted with other mutations.

189

190 *Population genetic analyses of British samples*

191 We examined patterns of sequence variation using a range of population genetic methods. We
192 investigated the amount of sequence diversity in these recently diverging species using
193 descriptive statistics, and then tested the cohesiveness of taxa using analysis of molecular
194 variance (AMOVA) and related methods. Analyses were performed separately on ITS2 and a
195 concatenated matrix of plastid data. For plastid data, haplotypes were determined from
196 nucleotide substitutions and indels of the aligned sequences. Basic population genetic
197 statistics were performed in Arlequin Version 3.0 (Excoffier and Lischer, 2010), and this
198 included the number of haplotypes, as well as hierarchical AMOVA in groups according to:
199 (1) ploidy levels (diploid vs tetraploid); (2) geographic regions (Wales, England, Scotland);
200 (3) species. AMOVAs were performed on all taxa, and repeated for ploidy levels and
201 geographic regions on a dataset only including confirmed species (i.e. excluding hybrids).
202 Sequence diversity and divergence statistics were estimated with DnaSP (Librado and Rozas,
203 2009), which included: average nucleotide diversity across taxa (Π), Watterson's theta (per
204 site), Tajima's D and divergence between ploidy levels (D_{XY}).

205

206 Genetic divergence among sampling localities were explored with Spatial Analysis of
207 Molecular Variance (SAMOVA; Dupanloup et al., 2002), implemented in SPADS v.1.0
208 (Dellicour and Mardulyn, 2014). SAMOVA maximises the proportion of genetic variance
209 due to differences among populations (F_{CT}) for a given number of genetic clusters (K -value).
210 We considered the best grouping to have the highest F_{CT} value after 100 repetitions. This
211 analysis investigated interspecific differentiation, thus only used species samples, excluding
212 hybrids.

213

214 The relationships between haplotypes was inferred by constructing median-joining networks
215 (MJM; Bandelt et al., 1999) with the program NETWORK v.4.6.1.1 (available at
216 <http://www.fluxus-engineering.com/>), treating gaps as single evolutionary events.

217

218 *Phylogenetic analyses and molecular dating*

219 We used Bayesian phylogenetic analyses in MrBayes v. 3.1.2 (Huelesenbeck & Ronquist
220 2001) to infer species relationships and broad-scale patterns of colonisation. Our analyses
221 used a sequence matrix that included our newly sampled British taxa in addition to previous
222 global *Euphrasia* samples from Gussarova et al. (2008). We selected the best fitting model of
223 nucleotide substitution using the Akaike Information Criterion (AIC) with an empirical
224 correction for small sample sizes implemented in MrAIC (Nylander, 2004). Using GTR +G
225 as the best model for the plastid dataset and SYM + G for the ITS dataset we ran two sets of
226 four Markov Chain Monte Carlo (MCMC) runs for 5,000,000 generations. Indels were
227 included as a separate partition with a restriction site (binary) model. We sampled every
228 1000th generation and used a burnin of 2,500,000, and default priors. We confirmed chain
229 convergence by observing the average standard deviation of split frequencies and by plotting
230 parameter values in Tracer v. 1.6 (Rambaut & Drummond 2013).

231

232 We used BEAST v. 1.8.2 (Drummond et al. 2012) to estimate the divergence ages of major
233 *Euphrasia* lineages occurring in Britain. This analysis used our British samples in
234 conjunction with the full global dataset of *Euphrasia* (Gussarova et al., 2008). Our analysis
235 gives crude dates due to the lack of available fossils for calibration, but these estimates allow
236 us to compare between very recent (postglacial) divergence, and much older divergence
237 events. We only analysed ITS sequences, due to the lack of support obtained for the plastid
238 phylogeny (see results). The analysis included two partitions: one containing all ITS
239 sequences (638 bp) and the other containing 38 gap-coded indels. We applied the same
240 substitution models as those used in the MrBayes analyses. For the binary characters, we used
241 a stochastic Dollo model. Separate analyses were run testing for the strict vs the uncorrelated
242 lognormal (UCLN) molecular clock. The branching was modelled using the Yule tree prior,
243 which assumes a constant speciation rate per lineage. The root age was set with a normal
244 prior of mean = 2.65×10^7 and SD = 5×10^5 , according to the results obtained based on a
245 calibrated phylogeny of *Euphrasia* by Gussarova et al. (2008). This prior analysis used
246 geological dating of volcanic islands to set priors as maximum ages for colonization of
247 endemic *Euphrasia* species. Substitution rate priors were based on Key et al. (2006), ranging
248 from 0.38×10^{-9} to 8.34×10^{-9} substitutions/site/year. Data analyses were run for 50,000,000
249 generations, logging every 50,000th generation. We compared the fit of the clock models
250 using AICM criterion implemented in Tracer. The AICM values obtained were very similar
251 between the two models: 8723.681+/-0.685 vs. 8882.911+/-1.582. The strict clock was
252 chosen as it is a simpler model and with lower AICM. These phylogenetic results were
253 compared with those from an empty alignment using the same values for priors.

254

255 **RESULTS**

256

257 *ITS haplotype diversity in British populations*

258

259 The final ITS2 alignment contained 130 individuals representative of all British taxa, and was
260 380 bp in length. Only two samples (of *E. scottica*) presented double peaks, and were
261 excluded from analysis. Overall diversity across taxa was modest, with a nucleotide diversity
262 (Π) of 2.3%, and theta (per site) of 0.01781. There were 33 nucleotide substitutions and one
263 indel, from which we called 23 haplotypes. All sequences have been deposited in GenBank
264 (Accession numbers given on acceptance).

265

266 The ITS haplotypes revealed strong partitioning by ploidy. Of the 23 haplotypes, three (H1,
267 H20 and H21) were restricted to diploids, and 19 to tetraploids, with only one haplotype (H2)
268 shared across ploidy levels (Table 1). Haplotype H2 was not only shared across ploidy levels
269 but was also the most widespread haplotype, found in 67 samples across 34 populations. This
270 included geographically distinct species such as the Scottish endemic *E. marshallii* and the
271 predominantly English and Welsh *E. anglica*, and ecologically contrasting taxa such as the
272 dry heathland specialist *E. micrantha* and the (currently unpublished) obligate coastal “*E.*
273 *fharaidensis*”. Overall, 86% of taxa had one of six widespread haplotypes. There were also a
274 large number of rare haplotypes, with over two thirds restricted to a single population (17
275 haplotypes: H4, H7-H10, H12-H21 and H23; Table 1). The remaining haplotypes found in
276 multiple populations (H1, H2, H3, H5, H6, H11) showed no clear pattern of geography, with
277 three found in all geographic regions (England, Scotland, Wales) and the remaining three
278 shared between two geographic regions. Similarly, patterns of haplotype sharing do not
279 follow species boundaries. Of the eight species with multiple populations (excluding
280 hybrids), none of them had diagnostic ITS haplotypes. Despite haplotype sharing across taxa,

281 there was no evidence for this being due to non-neutral processes, as the value of Tajima's D
282 (-0.17) was not significantly different from zero.

283

284 The putative hybrid species, *E. vigursii* and *E. rivularis*, possessed ITS haplotype H1, which
285 is common to other diploid taxa, or population specific haplotypes (H20, H21), but no
286 tetraploid haplotypes. The two sampled diploid-tetraploid hybrids (*E. arctica* x *rostkoviana*,
287 *E. tetraquetra* x *vigursii*) possessed the full range of haplotypes: haplotype H2, which is
288 common across ploidy levels, tetraploid specific haplotype H3, and diploid specific haplotype
289 H1. Most (9/12) tetraploid hybrid populations had haplotypes shared with their putative
290 parents, while the other populations had unique haplotypes.

291

292 The highest F_{CT} values in the SAMOVA were when $K = 2$ (Table S3 [**Supplementary**
293 **Information**]), and this corresponded to the diploid-tetraploid divide described above. At K
294 = 3, SAMOVA distinguished clusters corresponding to the two ploidy groups, and a third
295 group of hybrid species derived from inter-ploidy level mating. AMOVA also supported the
296 strong division by ploidy, with 88.2% of variation attributed to ploidy differences (Table 2; P
297 < 0.001). A high proportion of variation was also partitioned by taxa (63.2%), and regions
298 (25%), though this may be inflated by limited sampling within species.

299

300 Summary statistics and haplotype networks revealed substantial divergence between diploid
301 and tetraploid haplotypes. An average of 18.3 site differences were found between diploids
302 and tetraploids, with divergence measured as $D_{XY} = 0.051$ (5.1%). The haplotype network
303 revealed clusters corresponding to diploid and tetraploid haplotypes, separated by many
304 mutations (Figure 2). The diploid cluster centres round haplotype H1, found in 15 individuals
305 from 5 diploid species and one diploid-tetraploid hybrid. The only haplotype from this part of

306 the network present in tetraploids is haplotype H18, found in a single sample of *E. ostenfeldii*.
307 Within the tetraploid cluster, widespread haplotype H2 is at the centre, surrounded by other
308 widespread haplotypes (H3, 8 populations, 12 samples; H6, 7 populations, 11 samples), and
309 singleton haplotypes.

310

311 *ITS phylogeny and molecular dating*

312

313 Our broad-scale global *Euphrasia* phylogenetic analyses performed using MrBayes gave
314 meaningful clusters of species, though the tree topology was generally poorly supported with
315 many polytomies (Fig. 3). Haplotypes occurring in Britain were predominantly found in two
316 main clusters: a tetraploid clade of Holarctic taxa from Sect. *Euphrasia* (posterior probability
317 support, pp = 1.00, Clade A, Fig. 3), and a well-supported geographically restricted Palearctic
318 diploid lineage (pp = 1.0, Clade B, Fig. 3). The tetraploid clade included a mix of British and
319 European taxa, and is sister to a mixed clade of alpine diploid species and tetraploid *E.*
320 *minima* (Clade IVc, Fig. 3). The diploid clade includes British diploids *E. anglica*, *E.*
321 *rivularis*, *E. rostkoviana*, *E. vigursii* and European relatives (diploids or taxa without
322 chromosome counts). The only non-diploid in the clade is one individual of tetraploid British
323 *E. ostenfeldii*, which appears to be correctly identified and thus may have captured the diploid
324 ITS haplotype through historical hybridisation. Overall, terminal branches of the tree are
325 short, indicative of limited variation between related haplotypes. The only exception was the
326 long branch of *E. disperma* from New Zealand, a result seen in previous Bayesian analyses
327 (cf. Gussarova et al., 2008, fig 2) but not in Parsimony analyses, where it clusters together
328 with the other southern hemisphere species on a shorter branch (Gussarova et al., 2008).

329

330 Molecular dating with BEAST yielded a similar tree topology to MrBayes (results not
331 shown), with many nodes having a low posterior support values. The clade of tetraploids and
332 alpine relatives had an estimated crown age of 2.2 Ma (95% HPD = 2.0 - 2.4 Ma, Fig. 3,
333 Node 1). The median crown age of the major group of diploids (also including Nearctic
334 endemics *E. oakesii* and *E. randii*) was estimated as 1.0 Ma (95% HPD = 0.7 – 1.4 Ma; pp =
335 1.00, Fig. 3, Node 2). Due to low support of internal branches, the age of the most recent
336 common ancestor of the diploid and tetraploid lineages could not be inferred directly. The
337 nearest dated node gaining support is the broader clade of *Euphrasia*, which includes the
338 diploid and tetraploid groups, in addition to two additional clades that include divergent
339 species of unknown ploidy such as *E. insignis* (Japan), which has a crown age of 8.0 Ma,
340 95% HPD = 6.5 – 9.8 Ma (Fig. 3, Node 3).

341

342 *Plastid haplotype diversity*

343

344 Initial sequencing of *rbcL* in 48 samples revealed no polymorphism, and so no further
345 sequencing was performed for this region and it was excluded from further analyses.
346 The final *matK* alignment was 844 bp with one indel, and the *rpl32-trnL* region was 630 bp
347 with nine indels. The final concatenated plastid alignment was 1474 bp for 130 samples, with
348 2.7% segregating sites (40 sites) across 38 haplotypes. Nucleotide diversity was exceptionally
349 low with $\Pi = 0.3\%$, and theta (per site) was similarly low at 0.00381. Tajima's D was not
350 significantly different from zero (-0.27).

351

352 Similar to ITS, most plastid haplotypes were individual or population specific (63%, 24/38
353 haplotypes found in one population only), with only 4 haplotypes being widespread (H4: 15
354 populations, 24 samples; H5: 13 populations, 18 samples; H2: 11 populations, 16 samples;

355 H1, 10 populations, 14 samples, Table S4 [**Supplementary Information**]). However, unlike
356 ITS, plastid haplotypes revealed complex patterns unrelated to ploidy. Most widespread
357 haplotypes were shared across ploidy levels. An AMOVA found a moderate degree of
358 genetic diversity was partitioned by ploidy (18.7%) and species (26.5%), with these values
359 being reduced when hybrids were included (Table 2 [**Supplementary Information**]).
360 Despite geography explaining little of the variation across the total dataset (4.9%), or being
361 evident in the SAMOVA (Table S5), localised haplotype sharing was apparent in Scottish
362 tetraploids. For example, haplotype H7 is shared across Scottish populations of *E. arctica*
363 (and its hybrids), *E. foulaensis* and *E. micrantha* (and its hybrids), while H10 is also shared
364 across three species in Scotland.

365

366 *Relationship among plastid haplotypes*

367

368 The final concatenated plastid alignment was 1692 bp in length, for a total of 82 *Euphrasia*
369 samples, including those from Gussarova et al. (2008). This alignment included the *trnL*
370 intron (517 bp, 73 variable sites), *trnL-trnF* (420 bp, 85 variable sites) and *atpB-rbcL* (754
371 bp, 89 variable sites). The plastid tree (Fig. 4) successfully recovered the geographic clades
372 reported in Gussarova et al. (2008). All diploid and tetraploid British samples possessed
373 plastid haplotypes from the broad Palearctic taxa clade, which also includes *E. borneensis*
374 (Borneo) and *E. fedtschenkoana* (Tian Shan). This clade received moderate support in our
375 analysis (pp = 0.85). While informative of broad-scale relationships, most terminal branches
376 were extremely short, and gave no information on interspecific relationships.

377

378 **DISCUSSION**

379

380 We have investigated the utility of DNA barcoding data for the study of taxonomically
381 complex *Euphrasia* in Britain. We find that *Euphrasia* species do not possess diagnostic ITS
382 and plastid sequence profiles, and instead haplotypes are either widespread across taxa, or are
383 individual or population specific. While DNA barcoding is of limited value as a molecular
384 tool for identifying *Euphrasia* species, we show these data to be informative of the
385 evolutionary processes underlying the generation and maintenance of diversity. Most notably,
386 deep divergence of ITS haplotypes between ploidy groups ($P_i = 5.1\%$, >2 Ma divergence),
387 support British *Euphrasia* being assembled from diverse pre-glacial continental taxa, and
388 with ploidy differences acting as an important reproductive barrier partitioning genetic
389 variation. Overall our results shed light on the maintenance of genetic variation in one of the
390 most renowned taxonomically challenging plant groups.

391

392 *DNA barcoding in taxonomically complex genera*

393

394 Taxonomically complex *Euphrasia* have many characteristics that would make a DNA-based
395 identification system desirable. In particular, molecular identification tools could be used to
396 confirm species identities and subsequently revise the distribution of these under-recorded
397 taxa. More generally, genetic data could be used to investigate which British species are
398 genetically cohesive ‘good’ taxa. Our data show, however, that barcode sequences do not
399 correlate with species boundaries as defined by morphology. Instead, haplotypes are often
400 individual or population specific (for both ITS and plastid DNA), and where widespread are
401 either shared across species within a ploidy level (for ITS) or across ploidy levels and species
402 (plastid DNA). Patterns of haplotype sharing are often surprising, including between
403 geographically disparate populations 100+ km apart, and between ecologically specialised
404 taxa that seldom co-occur in the wild.

405

406 One possibility is that the species are not discrete genetic entities, and that the current
407 taxonomy reflects a blend of discrete lineages, polytopic taxa, and morphotypes determined
408 by a small number of genes. Alternatively, the species may represent meaningful biological
409 entities, but with boundaries permeable to gene flow. In either case, the lack of species-
410 specific or morphotype-specific barcodes may be affected by numerous evolutionary
411 processes. Low sequence diversity associated with recent diversification is the first important
412 factor. While elevated plastid diversity is a hallmark of some parasitic taxa, this is not the
413 case for facultative hemiparasites like *Euphrasia*, which generally show similar patterns of
414 mutation to autotrophic taxa (Wicke et al., 2016). Despite observing low plastid diversity
415 (<0.5%) across taxa, we still detected 38 plastid haplotypes. This is sufficient variation for
416 population genetic analysis, but lack of nucleotide diversity could explain the poor
417 performance of our phylogenetic analyses. A second factor that reduces the ability of DNA
418 data to distinguish species is selective sweeps, which can act on any genomic region
419 including the plastid (Twyford, 2014, Muir and Filatov, 2007). While our data collection was
420 not designed to measure the strength of selection, values for Tajima's D were not
421 significantly different from zero thus do not point to a strong selective regime.

422

423 In contrast to the factors above, it seems that self-fertilization, hybridity and incomplete
424 lineage sorting are dominant factors shaping haplotype distributions across British *Euphrasia*
425 species and populations. We observed many haplotypes that are restricted to individual
426 samples or populations, consistent with French et al. (2008), who used extensive
427 intrapopulation sampling to show most *Euphrasia* populations are fixed for a single plastid
428 haplotype. This pattern of local fixation of haplotypes, and scarcity of widespread haplotypes,
429 is likely due to many *Euphrasia* species being (at least partly) self-fertilising ($F_{is} > 0.4$,

430 French et al., 2004), and localised gravity-mediated seed dispersal. Where widespread
431 haplotypes are present, these are shared amongst taxa and geographic areas. In these cases,
432 dispersal of haplotypes has clearly occurred and their sharing among species may be due to
433 hybridisation and incomplete lineage sorting. The large number of reported hybrids and
434 hybrid species based on morphology (Preston and Pearman, 2015, Stace et al., 2015), as well
435 as the prevalence of hybridisation in genetic data (Stone, 2013, Liebst, 2008), point to
436 hybridisation being a key factor shaping genetic diversity in *Euphrasia*. Future genomic
437 surveys will estimate the proportion of loci introgressing across species barrier in models that
438 explicitly account for incomplete lineage sorting (Twyford and Ennos, 2012).

439

440 The sharing of DNA barcodes among *Euphrasia* species parallels a number of other studies
441 where DNA barcoding has failed to provide species-level information. A notable example is
442 willows, where hybridisation and selective sweeps have caused a single haplotype to spread
443 across highly divergent taxa and between geographic regions (Percy et al., 2014). Poor
444 species discrimination from DNA barcoding is also seen in the rapid radiation of Chinese
445 *Primula* (Yan et al., 2015a) and *Rhododendron* (Yan et al., 2015b), which is likely a product
446 of hybridisation and recent species divergence. In each of these groups, future DNA
447 barcoding systems that target large quantities of nuclear sequence variation may provide
448 resolution (Coissac et al., 2016, Hollingsworth et al., 2016). These data have the joint benefit
449 of providing many nucleotide characters from unlinked loci, while also moving away from
450 genomic regions that have atypical inheritance and patterns of evolution (i.e. plastids).
451 Analyses of many nuclear genes (or entire genomes) would be particularly valuable for
452 *Euphrasia*, where it may be possible to identify adaptive variants maintained in the face of
453 hybridisation. These adaptive genes may underlie differences between species or ecotypes

454 (Twyford and Friedman, 2015), and these loci could then potentially be used for future
455 species identification.

456

457 *Polyploidy and the maintenance of genetic variation*

458

459 The deep divergence of ITS haplotypes between diploid and tetraploid *Euphrasia* suggest
460 strong reproductive barriers between ploidy levels, a result consistent with previous
461 population surveys with AFLPs (French et al., 2008). The lack of support of internal nodes in
462 our phylogeny make the divergence between ploidy groups difficult to date, but it must
463 predate the origin of the tetraploid clade (2.2 Ma) and the diploid clade (1 Ma), with a date
464 likely to be closer to 8 Ma (similar to a previous global *Euphrasia* phylogeny, Gussarova et
465 al., 2008). Our divergence age estimate suggests that genetic diversity in British *Euphrasia*
466 long pre-dates recent glacial divergence and the origin of young British endemic taxa. As
467 such, British *Euphrasia* diversity has been assembled from a diverse pool of genetic diversity
468 from European and Amphi-Atlantic taxa. The presence of distinct ITS haplotypes in each
469 ploidy group may also be a consequence of concerted evolution of this multi-copy region
470 (Sang et al., 1995), rather than absolute reproductive isolation. Further work will be required
471 to test the extent of gene flow across ploidy levels, which is now emerging as a common
472 feature of other plant groups (Pinheiro et al., 2010, Slotte et al., 2008).

473

474 The divergence between diploid and tetraploid ITS sequences adds further weight to the
475 British tetraploids not being young autopolyploids, and instead being allopolyploids. While it
476 is difficult to decipher the parentage of British tetraploids from our data, our phylogenetic
477 analysis place these British taxa in a clade composed exclusively of tetraploids (Clade IVd,
478 Gussarova et al., 2008), and thus these species may have originated elsewhere before dispersal

479 to the UK. Alternatively, plastid haplotypes shared between diploids and tetraploids could
480 point to a British diploid parent to the tetraploids, with this parent not contributing an ITS
481 haplotype. An allotetraploid origin is further supported by the high number of tetraploid-
482 specific AFLP bands (French et al., 2008), fixed microsatellite heterozygosity indicative of
483 disomic inheritance (Stone, 2013), and tetraploid genome assemblies of double the size of
484 diploids (Twyford and Ness, Unpublished data). Future genomic and cytogenetic studies will
485 clarify the origin of British tetraploids.

486

487 **Conclusions**

488

489 This study highlights how DNA barcoding data may fail to distinguish between species in
490 taxonomically complex groups such as *Euphrasia*. No species in our study possessed a
491 consistent diagnostic haplotype. Widespread haplotype sharing among species, in conjunction
492 with high levels of intraspecific variation, make *Euphrasia* a particular challenge for DNA
493 barcoding. However, our results are able to help us understand the maintenance of diversity,
494 and in particular allow us to comment on the origins of British tetraploid species.

495

496 **ACKNOWLEDGEMENTS**

497

498 Thanks to Richard Ennos for his assistance with this research. This study formed part of an
499 international exchange programme for XW, funded by a Chinese Scholarship Council (CSC).
500 The Royal Botanic Garden Edinburgh (RBGE) is supported by the Scottish Government's
501 Rural and Environmental Science and Analytical Services Division. Research by ADT is
502 supported by NERC Fellowship NE/L011336/1.

503

504 **LITERATURE CITED**

505

506 **Bandelt H-J, Forster P, Röhl A. 1999.** Median-joining networks for inferring intraspecific
507 phylogenies. *Molecular Biology and Evolution*, **16**: 37-48.

508 **Brysting AK, Oxelman B, Huber KT, Moulton V, Brochmann C, Renner S. 2007.**

509 Untangling complex histories of genome mergings in high polyploids. *Systematic*
510 *Biology*, **56**: 467-476.

511 **Coissac E, Hollingsworth PM, Lavergne S, Taberlet P. 2016.** From barcodes to genomes:
512 extending the concept of DNA barcoding. *Molecular Ecology*, **25**: 1423-1428.

513 **de Vere N, Rich TC, Ford CR, Trinder SA, Long C, Moore CW, Satterthwaite D,**

514 **Davies H, Allainguillaume J, Ronca S. 2012.** DNA barcoding the native flowering
515 plants and conifers of Wales. *PLoS One*, **7**: e37945.

516 **Dellicour S, Mardulyn P. 2014.** SPADS 1.0: a toolbox to perform spatial analyses on DNA
517 sequence data sets. *Molecular Ecology Resources*, **14**: 647-651.

518 **Dupanloup I, Schneider S, Excoffier L. 2002.** A simulated annealing approach to define the
519 genetic structure of populations. *Molecular Ecology*, **11**: 2571-2581.

520 **Drummond AJ, Suchard MA, Xie D & Rambaut A. 2012.** Bayesian phylogenetics with
521 BEAUti and the BEAST 1.7 *Molecular Biology And Evolution* **29**: 1969-1973.

522 **Ennos RA, French GC, Hollingsworth PM. 2005.** Conserving taxonomic complexity.
523 *Trends in Ecology & Evolution*, **20**: 164-168.

524 **Excoffier L, Lischer HE. 2010.** Arlequin suite ver 3.5: a new series of programs to perform
525 population genetics analyses under Linux and Windows. *Molecular Ecology*
526 *resources*, **10**: 564-567.

- 527 **French G, Hollingsworth P, Silverside A, Ennos R. 2008.** Genetics, taxonomy and the
528 conservation of British *Euphrasia*. *Conservation Genetics*, **9**: 1547-1562.
- 529 **French GC, Ennos RA, Silverside AJ, Hollingsworth PM. 2004.** The relationship between
530 flower size, inbreeding coefficient and inferred selfing rate in British *Euphrasia*
531 species. *Heredity*, **94**: 44-51.
- 532 **China Plant BOL Group. 2011.** Comparative analysis of a large dataset indicates that
533 internal transcribed spacer (ITS) should be incorporated into the core barcode for seed
534 plants. *Proceedings of the National Academy of Sciences*, **108**: 19641-19646.
- 535 **Gussarova G, Popp M, Vitek E, Brochmann C. 2008.** Molecular phylogeny and
536 biogeography of the bipolar *Euphrasia* (Orobanchaceae): Recent radiations in an old
537 genus. *Molecular Phylogenetics and Evolution*, **48**: 444-460.
- 538 **Hebert PD, Gregory TR. 2005.** The promise of DNA barcoding for taxonomy. *Systematic*
539 *Biology*, **54**: 852-859.
- 540 **Hodges SA, Arnold ML. 1994.** Columbines: a geographically widespread species flock.
541 *Proceedings of the National Academy of Sciences*, **91**: 5129-5132.
- 542 **Hollingsworth P, De-Zhu L, Van der Bank M, Twyford A. 2016.** Telling plant species
543 apart with DNA: from barcodes to genomes. *Philosophical Transactions of the Royal*
544 *Society B*.
- 545 **Hollingsworth PM, Forrest LL, Spouge JL, Hajibabaei M, Ratnasingham S, van der**
546 **Bank M, Chase MW, Cowan RS, Erickson DL. et al. 2009.** A DNA barcode for
547 land plants. *Proceedings of the National Academy of Sciences*, **106**: 12794-12797.
- 548 **Hollingsworth PM, Graham SW, Little DP. 2011.** Choosing and using a plant DNA
549 barcode. *PLoS ONE*, **6**: e19254.
- 550 **Kolář F, Píšová S, Závěská E, Fér T, Weiser M, Ehrendorfer F, Suda J. 2015.** The origin
551 of unique diversity in deglaciated areas: traces of Pleistocene processes in

- 552 north-European endemics from the *Galium pusillum* polyploid complex (Rubiaceae).
553 *Molecular Ecology*, **24**: 1311-1334.
- 554 **Librado P, Rozas J. 2009.** DnaSP v5: a software for comprehensive analysis of DNA
555 polymorphism data. *Bioinformatics*, **25**: 1451-1452.
- 556 **Liebst B. 2008.** Do they really hybridize? A field study in artificially established mixed
557 populations of *Euphrasia minima* and *E. salisburgensis* (Orobanchaceae) in the Swiss
558 Alps. *Plant Systematics and Evolution*, **273**: 179-189.
- 559 **McNeal JR, Bennett JR, Wolfe AD, Mathews S. 2013.** Phylogeny and origins of
560 holoparasitism in Orobanchaceae. *American Journal of Botany*, **100**: 971-983.
- 561 **Muir G, Filatov D. 2007.** A selective sweep in the chloroplast DNA of dioecious *Silene*
562 (Section *Elisanthe*). *Genetics*, **177**: 1239-1247.
- 563 **Percy DM, Argus George W, Cronk Quentin C, Fazekas Aron J, Kesanakurti Prasad R,**
564 **Burgess Kevin S, Husband Brian C, Newmaster Steven G, Barrett Spencer CH,**
565 **Graham Sean W. 2014.** Understanding the spectacular failure of DNA barcoding in
566 willows (*Salix*): Does this result from a trans-specific selective sweep? *Molecular*
567 *Ecology*, **23**: 4737-4756.
- 568 **Pinheiro F, De Barros F, Palma-Silva C, Meyer D, Fay MF, Suzuki RM, Lexer C,**
569 **Cozzolino S. 2010.** Hybridization and introgression across different ploidy levels in
570 the Neotropical orchids *Epidendrum fulgens* and *E. puniceoluteum* (Orchidaceae).
571 *Molecular Ecology*, **19**: 3981-3994.
- 572 **Preston CD, Pearman DA. 2015.** Plant hybrids in the wild: evidence from biological
573 recording. *Biological Journal of the Linnean Society*, **115**: 555-572.
- 574 **Sang T, Crawford DJ, Stuessy TF. 1995.** Documentation of reticulate evolution in peonies
575 (*Paeonia*) using internal transcribed spacer sequences of nuclear ribosomal DNA:

- 576 implications for biogeography and concerted evolution. *Proceedings of the National*
577 *Academy of Sciences*, **92**: 6813-6817.
- 578 **Schmickl R, Paule J, Klein J, Marhold K, Koch MA. 2012.** The evolutionary history of the
579 *Arabidopsis arenosa* complex: diverse tetraploids mask the western carpathian center
580 of species and genetic diversity. *PLOS ONE*, **7**: e42691.
- 581 **Slotte T, Huang H, Lascoux M, Cephitis A. 2008.** Polyploid speciation did not confer
582 instant reproductive isolation in *Capsella* (Brassicaceae). *Molecular Biology and*
583 *Evolution*, **25**: 1472-1481.
- 584 **Spooner DM. 2009.** DNA barcoding will frequently fail in complicated groups: An example
585 in wild potatoes. *American Journal of Botany*, **96**: 1177-1189.
- 586 **Squirrell J, Hollingsworth PM, Bateman RM, Tebbitt MC, Hollingsworth ML. 2002.**
587 Taxonomic complexity and breeding system transitions: conservation genetics of the
588 *Epipactis leptochila* complex (Orchidaceae). *Molecular Ecology*, **11**: 1957-1964.
- 589 **Stace CA, Preston CD, Pearman DA. 2015.** *Hybrid flora of the British Isles*: Botanical
590 Society of Britain and Ireland.
- 591 **Stone H. 2013.** *Evolution and conservation of tetraploid Euphrasia L. in Britain*.
592 Unpublished Thesis, The University of Edinburgh.
- 593 **Svensson BM, Carlsson BÅ. 2004.** Significance of time of attachment, host type, and
594 neighbouring hemiparasites in determining fitness in two endangered grassland
595 hemiparasites. *Annales Botanici Fennici*, **41**: 63-75.
- 596 **Taberlet P, Gielly L, Pautou G, Bouvet J. 1991.** Universal primers for amplification of
597 three non-coding regions of chloroplast DNA. *Plant Molecular Biology*, **17**: 1105-
598 1109.
- 599 **Twyford A, Ennos R. 2012.** Next-generation hybridization and introgression. *Heredity*, **108**:
600 179-189.

- 601 **Twyford AD. 2014.** Testing evolutionary hypotheses for DNA barcoding failure in willows.
602 *Molecular Ecology*, **23**: 4674-4676.
- 603 **Twyford AD, Friedman J. 2015.** Adaptive divergence in the monkey flower *Mimulus*
604 *guttatus* is maintained by a chromosomal inversion. *Evolution*, **69**: 1476-1486.
- 605 **Wicke S, Müller KF, Quandt D, Bellot S, Schneeweiss GM. 2016.** Mechanistic model of
606 evolutionary rate variation en route to a nonphotosynthetic lifestyle in plants.
607 *Proceedings of the National Academy of Sciences*: 201607576.
- 608 **Yan H-F, Liu Y-J, Xie X-F, Zhang C-Y, Hu C-M, Hao G, Ge X-J. 2015a.** DNA barcoding
609 evaluation and its taxonomic implications in the species-rich genus *Primula* L. in
610 China. *PLOS ONE*, **10**: e0122903.
- 611 **Yan L-J, Liu J, Möller M, Zhang L, Zhang X-M, Li D-Z, Gao L-M. 2015b.** DNA
612 barcoding of *Rhododendron* (Ericaceae), the largest Chinese plant genus in
613 biodiversity hotspots of the Himalaya–Hengduan Mountains. *Molecular Ecology*
614 *Resources*, **15**: 932-944.
- 615 **Yeo P. 1956.** Hybridization between diploid and tetraploid species of *Euphrasia*. *Watsonia*,
616 **3**: 253-269.
- 617 **Yeo PF. 1978.** A taxonomic revision of *Euphrasia* in Europe. *Botanical Journal of the*
618 *Linnean Society*, **77**: 223-334.
- 619 **Young ND, Healy J. 2003.** GapCoder automates the use of indel characters in phylogenetic
620 analysis. *BMC Bioinformatics*, **4**: 6.

621

622

623

624

625 **Figure legends**

626

627 Figure 1. *Euphrasia* samples used in this study. (A) Tetraploid British *Euphrasia* (here *E.*
628 *arctica*) have glabrous leaves sometimes with sparse short eglandular hairs or bristles. (B)
629 Diploid British *Euphrasia* have long glandular hairs. (C) Collection sites of *Euphrasia* DNA
630 samples. Diploids are shown in red, tetraploids in blue. Orange boxes correspond to the three
631 broad sampling areas.

632

633 Figure 2. Median-joining network of ITS haplotype relationships in British *Euphrasia*.
634 Numbers correspond to the ITS haplotypes in Table 1. Haplotypes are coloured by ploidy,
635 with diploids in red and tetraploids in blue. Hypothetical (unsampled) haplotypes are
636 represented by filled black circles.

637

638 Figure 3. Majority rule consensus phylogeny of *Euphrasia* inferred from the ITS region using
639 MrBayes. Posterior probabilities greater than 0.85 are indicated. Individuals are coloured by
640 ploidy and geography: British diploids (red), British tetraploid (blue), other geographic areas
641 (black). Green circled nodes have divergence dates estimated with BEAST, as given in the
642 text. Clade A and Clade B correspond to the main study groups, with additional clades
643 corresponding to Gussarova et al. (2008) also marked: II northern tetraploids; III Taiwan; IVa
644 South American/Tasmanian; IVb complex (S. American, N. Zealand, Japan); IVc Alpine
645 European.

646

647 Figure 4. Majority rule consensus phylogeny of *Euphrasia* inferred from a concatenation of
 648 plastid *trnL* intron, *trnL-trnF* and *atpB-rbcL* using MrBayes. Posterior probabilities greater
 649 than 0.85 are indicated, and British diploids (red) and British tetraploid (blue) are coloured.

650

651 Table 1. The distribution of ITS2 haplotypes between species and geographic regions in
 652 British *Euphrasia*. Haplotype numbers correspond to the haplotype network Figure 2.

653 Population specific haplotypes are aggregated under one column. n = number of samples.

654

Species	Region	n	Widespread haplotypes						Population-specific haplotypes
			H1	H2	H3	H5	H6	H11	
<i>E. anglica</i>	SW	4	4						0
<i>E. anglica</i>	W	3	2	1					0
<i>E. arctica</i>	S	3		1	2				0
<i>E. arctica</i>	SW	2		2					0
<i>E. arctica</i>	W	2		1	1				0
<i>E. arctica</i> x <i>confusa</i>	S	1							1
<i>E. arctica</i> x <i>foulaensis</i>	S	1		1					0
<i>E. arctica</i> x <i>micrantha</i>	S	3		1	1	1			0
<i>E. arctica</i> x <i>nemorosa</i>	S	2		1			1		0
<i>E. arctica</i> x <i>rostkoviana</i>	S	3		1	2				0
<i>E. cambrica</i>	W	3				1	1		1
<i>E. campbelliae</i>	S	3					3		0
<i>E. confusa</i>	S	4		4					0
<i>E. confusa</i>	SW	3			1	1			1
<i>E. confusa</i>	W	3		2					1
<i>E. confusa</i> x <i>micrantha</i>	S	2		1					1
<i>E. "fharaidensis"</i>	S	2		1				1	0
<i>E. foulaensis</i>	S	6		2	2		2		0

<i>E. foulaensis x marshllii</i>	S	2	1		1			0	
<i>E. foulaensis x nemorosa</i>	S	1	1					0	
<i>E. foulaensis x ostenfeldii</i>	S	1	1					0	
<i>E. frigida</i>	S	6	5		1			0	
<i>E. heslop-harrisonii</i>	S	6	5					1	
<i>E. marshllii</i>	S	3	3					0	
<i>E. marshallii x micrantha</i>	S	2	2					0	
<i>E. micrantha</i>	S	5	4					1	
<i>E. micrantha</i>	SW	3	2					1	
<i>E. micrantha</i>	W	3				2		1	
<i>E. micrantha x nemorosa</i>	SW	1		1				0	
<i>E. micrantha x scottica</i>	W	4	4					0	
<i>E. nemorosa</i>	S	3	3					0	
<i>E. nemorosa</i>	SW	2	1					1	
<i>E. nemorosa</i>	W	3	2					1	
<i>E. nemorosa x tetraquetra</i>	SW	1	1					0	
<i>E. ostenfeldii</i>	S	5	3					2	
<i>E. ostenfeldii</i>	W	1	1					0	
<i>E. pseudokernerii</i>	W	3		2		1		0	
<i>E. rivularis</i>	W	3	2					1	
<i>E. rostkoviana</i>	W	3	2					1	
<i>E. rotundifolia</i>	S	1	1					0	
<i>E. scottica</i>	S	3	3					0	
<i>E. scottica</i>	W	3						3	
<i>E. tetraquetra</i>	SW	3	3					0	
<i>E. tetraquetra</i>	W	3	1			2		0	
<i>E. tetraquetra x vigursii</i>	SW	2	1	1				0	
<i>E. vigursii</i>	SW	4	4					0	
Total		130	15	67	12	4	11	3	18

655

656

657

658

659 Table 2. Hierarchical analysis of molecular variance (AMOVA) of British *Euphrasia*
 660 populations. Analyses performed between (A) species, (B) 3 geographic locations (Wales,
 661 South-West England, Scotland, (C) diploids and tetraploids. Number in parentheses are the
 662 results only including species (excluding hybrids). d.f. = Degrees of freedom. ** $P < 0.001$; * P
 663 < 0.05 .

664

Source of variation	ITS		Plastid DNA	
	d.f.	% Total variance	d.f.	% Total variance
(A) Taxa				
Between taxa	33 (19)	63.17** (65.62**)	33 (19)	15.87** (26.48**)
Within taxa	96 (77)	36.83 (34.38)	96 (68)	84.13 (73.52)
(B) Location				
Between regions	2 (2)	25.52** (25.92**)	2 (2)	5.40** (4.91*)
Within regions	127 (94)	74.48 (74.08)	127 (85)	94.60 (95.09)
(C) Ploidy				
Between ploidy groups	1 (1)	88.24** (84.39**)	1 (1)	11.05* (18.76*)
Within Diploid and tetraploid	123 (95)	11.76 (15.61)	123 (86)	88.95 (81.24)

665

666

667 Supplementary Table S1. Voucher information and population details of *Euphrasia* samples

668 used in the study. Columns headed cpDNA and ITS 2 refer to the number of individuals

669 sequenced for these regions.

670 Supplementary Table S2. PCR conditions and primer sequences for regions sequenced in this

671 study.

672 Supplementary Table S3. Spatial analysis of molecular variation (SAMOVA) of ITS

673 sequence data across British *Euphrasia* populations.

674 Supplementary Table S4. Plastid haplotype frequencies across British *Euphrasia* species. The

675 column cpDNA indicates the number of individuals sampled.

676 Supplementary Table S5. Spatial analysis of molecular variation (SAMOVA) of plastid

677 sequence data across British *Euphrasia* populations.

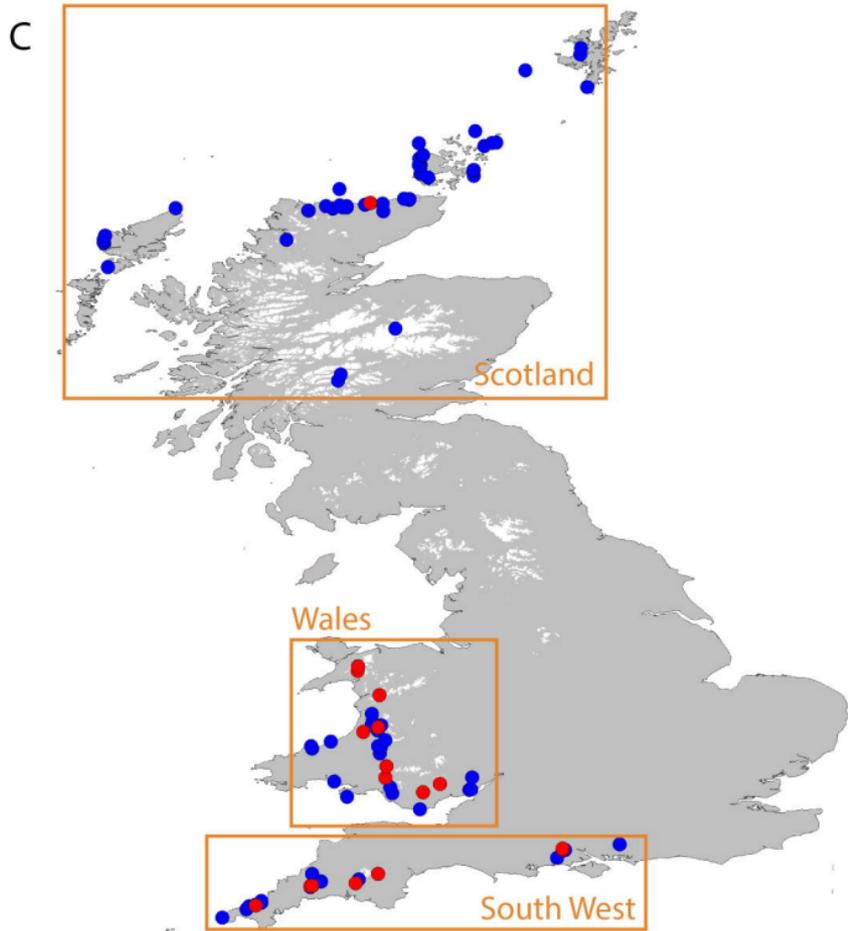
678

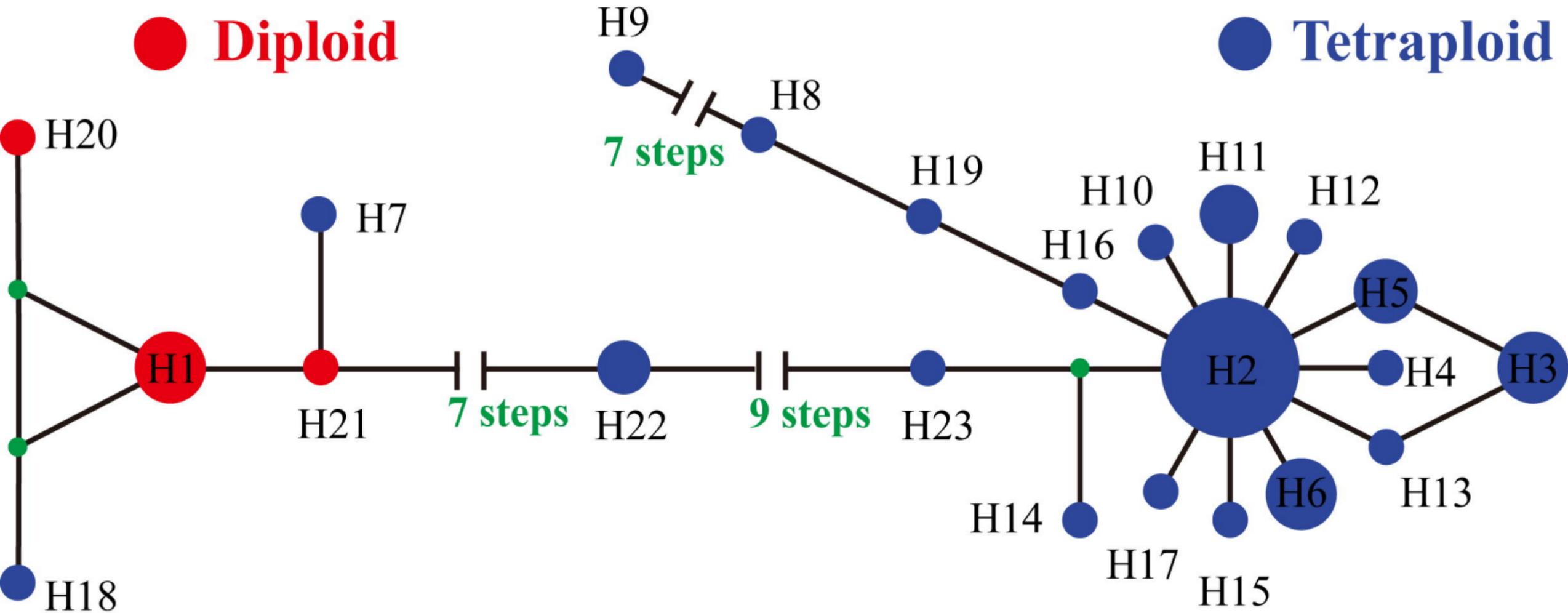
679

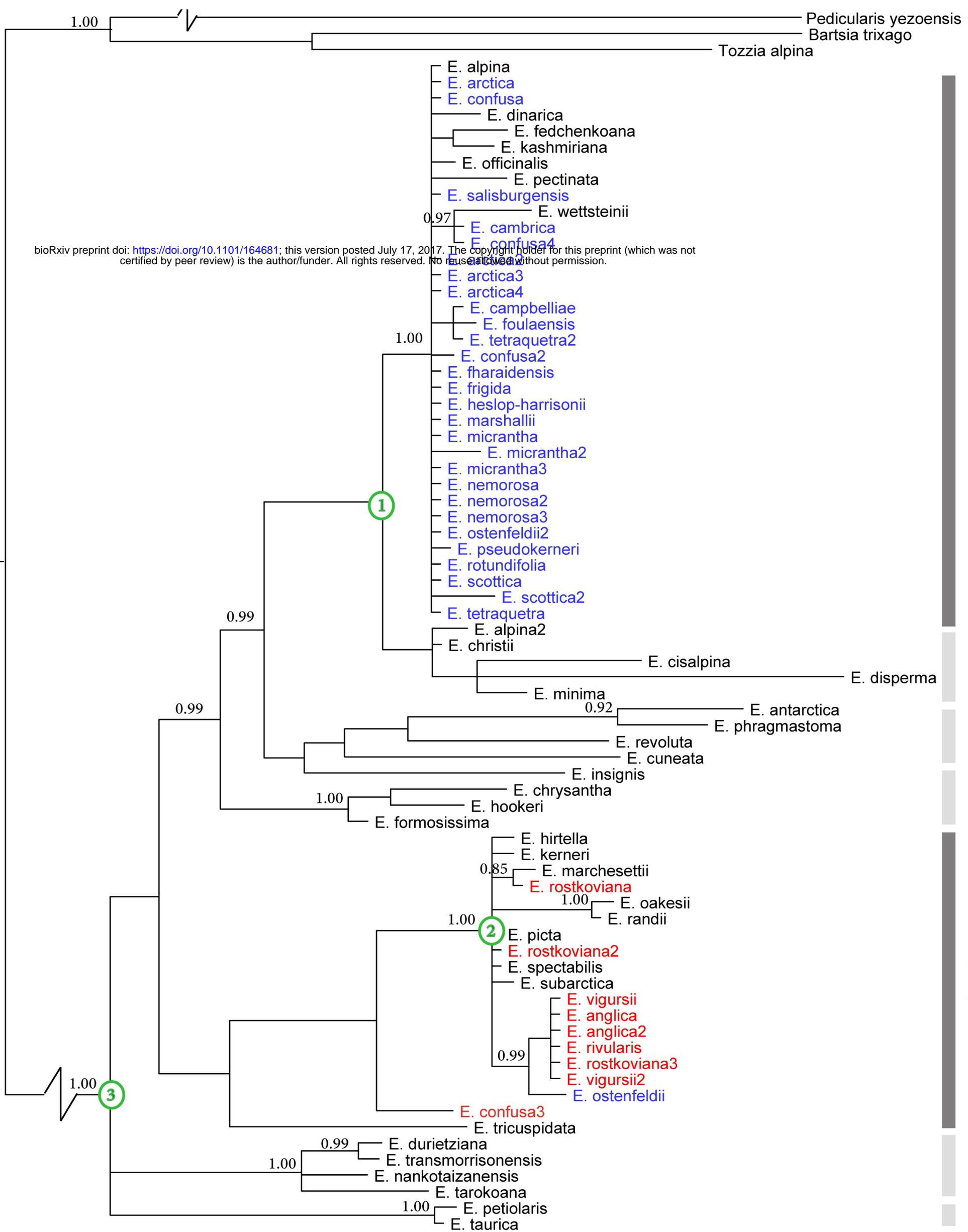
680

681

682







bioRxiv preprint doi: <https://doi.org/10.1101/164681>; this version posted July 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Palearctic
+ Alps
Clade A

IVc

IVb

IVa

Palearctic
+ Alps
Clade B

III

II

1.00

bioRxiv preprint doi: <https://doi.org/10.1101/163861>; this version posted July 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Palearctic
+ Alps

0.86

0.95

0.99

1.00

0.99

0.94

0.99

0.01

