Colonization and diversification of aquatic insects on three Macaronesian archipelagos using 59 nuclear loci derived from a draft genome

Sereina Rutschmann[a,b,c,*], Harald Detering[a,b,c], Sabrina Simon[d,e], David H. Funk[f], Jean-Luc Gattolliat[g], Samantha J. Hughes[h], Pedro M. Raposeiro[i], Rob DeSalle[4d], Michel Sartori[g], and Michael T. Monaghan[a,b]

[a]*Leibniz-Institute of Freshwater Ecology and Inland Fisheries (IGB), Müggelseedamm 301, 12587 Berlin, Germany*

[b]*Berlin Center for Genomics in Biodiversity Research, Königin-Luise-Straße 6-8, 14195 Berlin, Germany*

[c]*Department of Biochemistry, Genetics and Immunology, University of Vigo, 36310 Vigo, Spain*

[d]*Sackler Institute for Comparative Genomics, American Museum of Natural History, Central Park West and 79th St., New York, NY 10024, USA*

[e]*Biosystematics Group, Wageningen University, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands*

[f]*Stroud Water Research Center, Avondale, Pennsylvania 19311, USA*

[g]*Musée cantonal de zoologie, Palais de Rumine, Place de la Riponne 6, 1014 Lausanne, Switzerland*

[h]*Centro de Investigação e de Tecnologias Agro-Ambientais e Biológicas (CITAB), Universidade de Trás-os-Montes e Alto Douro, Quinta de Prados, Apartado 1013, 5001-801 Vila Real, Portugal*

[i]*Research Centre in Biodiversity and Genetic Resources (CIBIO)-Açores and the Biology Department, University of Azores, Rua Mãe de Deus 13A, 9501-855 Ponta Delgada, Portugal*

***Correspondence:** Sereina Rutschmann, Phylogenomics Lab, Department of Biochemistry, Genetics and Immunology, University of Vigo, 36310 Vigo, Spain, E-mail: sereina.rutschmann@gmail.com

Abstract

The study of processes driving diversification requires a fully sampled and well resolved phylogeny. Multilocus approaches to the study of recent diversification provide a powerful means to study the evolutionary process, but their application remains restricted because multiple unlinked loci with suitable variation for phylogenetic or coalescent analysis are not available for most non-model taxa. Here we identify novel, putative single-copy nuclear DNA (nDNA) phylogenetic markers to study the colonization and diversification of an aquatic insect species complex, *Cloeon dipterum* L. 1761 (Ephemeroptera: Baetidae), in Macaronesia. Whole-genome sequencing data from one member of the species complex were used to identify 59 nDNA loci (32,213 base pairs), followed by Sanger sequencing of 29 individuals sampled from 13 islands of three Macaronesian archipelagos. Multispecies coalescent analyses established six putative species. Three island species formed a monophyletic clade, with one species occurring on the Azores, Europe and North America. Ancestral state reconstruction indicated at least two colonization events from the mainland (Canaries, Azores) and one within the archipelago (between Madeira and the Canaries). Using random subsets of the 59 loci showed a positive linear relationship between number of loci and node support. In contrast, node support in the multispecies coalescent tree was negatively correlated with mean number of phylogenetically informative sites per locus, suggesting a complex relationship between tree resolution and marker variability. Our approach highlights the value of combining coalescent-based phylogeography, species delimitation, and phylogenetic reconstruction to resolve recent diversification events in an archipelago species complex.

Keywords: *Baetidae, island radiation, multispecies coalescent, phylogeny, phylogeography*

1. Introduction

Any inference about the ecological and evolutionary processes driving diversification requires a well sampled and fully resolved phylogeny upon which traits can be mapped. Molecular phylogenetic studies historically have been limited to a small number of loci. The majority of studies are based largely on mitochondrial DNA (mtDNA) loci (Avise et al., 2000; Garrick et al., 2015) which have the benefit of small population size and high levels of polymorphism but suffer from several characteristics that can limit their suitability to reconstruct the evolutionary process. These include an inability to detect processes that confound gene trees and species trees such as hybridization and introgression, the inference of oversimplified or unresolved evolutionary relationships based on their matrilineal history, underestimated genetic diversity (Zhang and Hewitt 2003), and overestimation of divergence times (Zheng et al., 2011). Another major drawback is the presence of mtDNA genes that have been transposed to the nuclear genome, forming nuclear mitochondrial DNA (Numt; Lopez et al., 1994) which may appear homologous but give very different evolutionary signals from those of the real mtDNA. Phylogenetics has begin to benefit from more widespread use of single-copy nuclear DNA (nDNA) loci, and several recent studies have applied greater numbers of nDNA loci with success at the species (e.g. *Ambystoma tigrinum* (O'Neill et al., 2013); *Triturus cristatus* (Wielstra et al., 2014)), genus (e.g. *Takydromus* (Tseng et al., 2014); *Heliconius* (Kozak et al., 2015)), and higher taxonomic levels (e.g. Plethodontidae (Shen et al., 2016)).

The phylogenetic resolution of closely related taxa enables crucial insights in studies of evolution. In particular, the investigation of recent or ongoing species radiations helps to explain how components such as adaptation and hybridization are involved in the diversification process (e.g. Monaghan et al., 2006; Morvan et al., 2013; Giarla and Esselstyn 2015; Toussaint et al., 2015). A number of model systems in evolutionary biology come from

closely related species groups that have diversified in island archipelagos (Schluter 2000; Gillespie and Roderick 2002, and references therein). Examples include Darwin's finches (Grant and Grant 2008), Anolis lizards (Losos and Ricklefs 2009), or Hawaiian spiders (Gillespie et al., 1994). While a robust phylogeny is needed to study diversification and adaptation in such groups, phylogenetic analysis of close relatives can be problematic. Discordance between gene trees and species trees is more likely when speciation is recent and the effective population size of the ancestral population is large relative to the age of the species (Kubatko and Degnan 2007; Degnan et al., 2012). This discordance can arise through hybridization, gene duplication and loss, and incomplete lineage sorting (Maddison 1997; Degnan and Rosenberg 2009; Knowles and Kubatko 2010; Nakhleh 2013). Increasing arrays of methods exist for examining multilocus data that account for these processes (Rannala and Yang 2003; Edwards 2009; Heled and Drummond 2010; Knowles and Kubatko 2010). Unfortunately, the appropriate data for these analyses can be lacking because it is difficult to generate sequence data for a sufficient number of suitable nDNA loci from non-model systems. Most nDNA loci exhibit low levels of polymorphism and therefore many loci are needed, whereas identification of novel nDNA loci that are suitable as phylogenetic markers is generally not straightforward. Here we use a whole-genome draft of a non-model species to develop nDNA markers suitable for phylogenetic reconstruction.

Macaronesia consists of four archipelagos (Azores, Madeira, Canary Islands, and Cape Verde) whose flora and fauna have been used in several studies as model systems for evolutionary research. Their distances to the adjacent continental mainland vary from 110 km (Fuerteventura in Canary Islands to Morocco) to more than 2000 km (Flores in the Azores to Portugal). Several colonization pathways have been identified (Juan et al., 2000; Emerson 2002; Emerson and Kolm 2005), including a single colonization event followed by stepping-stone dispersal (Juan et al., 1997; Emerson and Oromi 2005; Illera et al., 2007; Arnedo et al.,

100 2008; Dimitrov et al., 2008), or multiple independent colonization events within the Canary
101 Islands (Nogales et al., 1998; Ribera et al., 2003a; Díaz-Pérez et al., 2012; Rutschmann et al.,
102 2014; Gohli et al., 2015; Stervander et al., 2015; Faria et al., 2016). While much research has
103 been carried out on island evolution and endemism of terrestrial organisms, comparatively
104 limited information exists for aquatic invertebrates (e.g. Stauder 1995; Drotz 2003; Ribera et
105 al., 2003b, 2003c; Jordal and Hewitt 2004; Hughes and Malmqvist 2005). This is a large
106 discrepancy considering that aquatic insects contribute a disproportionally large amount of
107 global biodiversity despite the relatively small extent of their habitat (Dijkstra et al., 2014).

108 Mayflies are well suited for phylogeographic studies considering their ancient origins (300
109 million years (Ma)), global distribution, and limited dispersal ability due to the strict water
110 habitat fidelity of larvae and very short life of the winged adults (Monaghan et al., 2005;
111 Barber-James et al., 2008). Several studies have pointed out their unusual potential for
112 dispersion, reporting mayfly species on remote islands such as the Azores (Brinck and
113 Scherer 1961; Raposeiro et al., 2012), trans-oceanic dispersal between Madagascar and
114 continental Africa (Monaghan et al., 2005; Vuataz et al., 2013), and recent colonization
115 processes of several lineages on the Canary Islands and Madeira ≈ 14 Ma, including a close
116 link to the African mainland (Rutschmann et al., 2014).

117 The species complex of *Cloeon dipterum* L. 1761 is one of the most common and abundant
118 species of freshwater insects in European standing water. The taxonomic classification and
119 phylogenetic relationships within the *C. dipterum* s.l. species complex, including its
120 complicated synonymy, remain largely unknown. The species complex belongs to the
121 subgenus *Cloeon* Leach, 1815. In Europe, *Cloeon* consists of *C. dipterum*, two other currently
122 recognized species (*C. peregrinator* Gattolliat and Sartori, 2008, and *C. saharense* Soldán and
123 Thomas, 1983), and three species with unclear status (*species inquirenda; C. cognatum*
124 Stephens, 1836, *C. inscriptum* Bengtsson, 1914, and *C. rabaudi* Verrier, 1949) that are often

125 considered to be synonyms of *C. dipterum*. Its distribution ranges from North America, across
126 Europe to Northern Asia (excluding China), making it one of the largest known distributions
127 among mayflies (Bauernfeind and Soldán 2012, and references therein). Larvae are found in a
128 variety of aquatic habitats, including natural standing or slow-flowing waters, brackish water,
129 intermittent watercourses, and artificial biotopes across a wide range of climatic zones
130 (Bauernfeind and Soldán 2012, and references therein).

131 For this study we used a draft genome sequence of *Cloeon* to develop 59 nDNA loci
132 suitable for phylogenetic reconstruction of closely related members of the *C. dipterum* s.l.
133 species complex of mayflies. We identified target genes and designed primer pairs for them.
134 Standard PCR and Sanger sequencing were used to generate sequences. We then applied
135 Bayesian phylogenetic inference using concatenated sequence alignments and multispecies-
136 coalescent approaches to delineate species, examine their colonization from the mainland, and
137 understand their diversification throughout Atlantic oceanic islands (Fig. 1, Azores, Madeira,
138 and Canary Islands). Additionally, we quantitatively examined the effect of increasing
139 numbers of nDNA loci on tree resolution. Our analyses show how marker development can
140 proceed efficiently from draft whole genomes and that large numbers of nDNA loci can
141 produce fully resolved trees in closely related taxa, revealing the evolution and diversification
142 of the geographically widespread *C. dipterum* s.l. species complex. The disentangled
143 colonization routes of the three species occurring on the Macaronesian Islands highlight trans-
144 oceanic dispersal abilities of aquatic insects as an important driver of allopatric speciation,
145 including sympatric occurring sister-species on the islands and the mainland.

146 2. Material and methods

147 *2.1 Development of nuclear DNA loci*

148 To develop a set of nuclear loci we sequenced a newly created whole-genome library of *C.*
149 *dipterum* (see also Rutschmann et al., accepted). Libraries were generated from laboratory-
150 reared subimagos of *C. dipterum* specimens (full siblings). DNA was extracted from pooled
151 specimens (5-20) after removing eyes and wings using the Invisorb® Spin Tissue Mini kit
152 (STRATEC, Berlin, Germany). Extracted DNA was precipitated using Isopropanol and
153 pooled in order to obtain higher DNA yield. We prepared one 454 shotgun and one 454
154 paired-end library according to the manufacturer´s guidelines (Rapid Library Preparation
155 Method Manual, GS FLX+ Series - XL+, May 2011; Paired End Library Preparation Method
156 Manual – 20 kb and 8 kb Span, GS FLX Titanium Series, October 2009). The fragments were
157 amplified with an emulsion PCR (emPCR Method Manual - Lib-L SV, GS FLX Titanum
158 Series, October 2009; Rev. Jan 2010). Four lanes per library were sequenced on a Roche
159 (454) GS FLX machine). The sequence reads were trimmed and *de novo* assembled using
160 NEWBLER v. 2.5.3 (454 Life Sciences Corporation) under the default settings for large
161 datasets. We made two different assemblies, one with the reads from the shotgun library and
162 one with the reads from both shotgun and paired-end libraries. The newly sequenced draft
163 whole genome was combined with 4,197 expressed sequence tag (EST) sequences from
164 *Baetis* sp. (GenBank Acc. no. FN198828–FN203024). *Cloeon* and *Baetis* belong to the
165 Baetidae subfamilies Cloeoninae and Baetinae. Primer pairs were designed in the conserved
166 regions of orthologous sequences from included taxa. The above analysis procedures have
167 since been incorporated into the DISCOMARK pipeline for marker discovery and primer design
168 (Rutschmann at al., accepted; see Supplementary File 2).

169 ***2.2 Taxon sampling and DNA extraction***

170 We sampled individuals of the *C. dipterum* s.l. species complex from larval aquatic
171 habitats at 38 sampling sites on 13 islands including the Azorean archipelago, the Canary
172 Islands, Madeira (Fig. 1), and 32 sampling sites on the European and North American

173 mainland (Supplementary Tables 1 and 2). All samples were preserved in 99% ethanol in the
174 field and stored at 4°C until analysis. DNA was extracted from 107 individuals using
175 NucleoSpin® 96 tissue kits (Macherey-Nagel, Düren, Germany). Our analysis included
176 multiple populations of all currently recognized taxa (based on both morphological and
177 molecular data) on the islands (Brinck and Scherer 1961; Gattolliat et al., 2008; Rutschmann
178 et al., 2014).

179 ***2.3 PCR amplification, sequence alignment, and sequence heterogeneity***

180 We sequenced 60 loci for the study: the mtDNA barcoding gene (*cox1*) and 59 newly
181 developed nDNA loci. The *cox1* locus was amplified and sequenced using the procedure
182 described by Rutschmann et al. (2014). Based on a general mixed Yule-coalescent (gmyc)
183 model analysis (Fujisawa and Barraclough 2013) of *cox1*, we selected a representative set of
184 29 individuals for which we obtained nDNA sequences, using the 59 newly designed primer
185 pairs (Supplementary Table 3). Nuclear loci were amplified using standard polymerase chain
186 reaction (PCR) protocols with an annealing temperature of 55°C. The PCR products were
187 custom purified and sequenced at Beckman Coulter Genomics (Essex, UK) or Macrogen
188 (Amsterdam, The Netherlands). Forward and reverse sequences were assembled and edited
189 using GENEIOUS R7 v.7.1.3 (Biomatters Ltd.). Length variation (i.e. heterozygous indels) was
190 decoded using CODONCODE ALIGNER v.3.5.6 (CodonCode Corporation, Centerville MA,
191 USA). Additionally, we included previously published sequences from four individuals of 6
192 nDNA loci (KU971838-KU971840, KU971851, KU971919-KU971921, KU971933,
193 KU972490-KU972492, KU972503, KU972568-KU972570, KU972583, KU972653-
194 KU972654, KU972666, KU973060-KU973062, KU973074).

195 Multiple sequence alignments were made for each locus using MAFFT v.7.050b (L-INS-I
196 algorithm with default settings; Katoh and Standley 2013). The predicted orthologous
197 sequences of *Baetis* sp. were used to infer the correct exon-intron splicing boundaries

198 (canonical and non-canonical splice site pairs) of each alignment. Exon-intron boundaries of
199 locus 411912 could not be fully reconstructed and thus we used the exon sequence predicted
200 from tblastx searches for subsequent analyses. Locus alignments were split into coding and
201 non-coding parts using a custom script
202 (https://github.com/srutschmann/python_scripts/blob/master/extract_introns.py). All coding
203 alignments were checked for indels and stop codons using MESQUITE v.2.75 (Maddison and
204 Maddison 2011). Genotypes of the coding alignments were phased using the probabilistic
205 Bayesian algorithm implemented in PHASE v.2.1.1 (Stephens et al., 2001; Stephens and
206 Donnelly 2003) with a cutoff value of 0.6 (Harrigan et al., 2008; Garrick et al., 2010).
207 Multiple runs were performed for each alignment and phase calls checked for consistency.
208 Input and output files were formatted using the scripts from SEQPHASE (Flot 2010).
209 Heterozygous sites that could not be resolved were coded using ambiguity codes for
210 subsequent sequence analyses. All alignments were re-aligned after phasing with MAFFT.
211 We excluded introns for the haplotype phasing because the noncoding alignments contained
212 many gaps and missing data and thus the results of the sequence phasing were not
213 satisfactory.

214 For the subsequent analyses, we prepared three alignment sets (Table 1), whereby we used
215 all nDNA sequences (all_data), all coding genotypes (exon_all_data) and all coding
216 haplotypes (exonhap_all_data). Because data matrices were not 100% complete (see 3.1.
217 Development of nuclear DNA loci), we made a second set of matrices that were 100%
218 complete using only the 17 loci that were sequenced for all 29 individuals (complete_matrix,
219 exon_complete_matrix, exonhap_complete_matrix). The most appropriate substitution model
220 for each locus was determined according to Bayesian Information Criterion in the program
221 JMODELTEST v.2.1 (Guindon and Gascuel 2003; Darriba et al., 2012) (Supplementary Table
222 3).

223 To investigate the heterogeneity among the newly developed loci, we reconstructed
224 reticulation-free haplotype genealogies based on Fitch distances (Fitch 1970), using the
225 program FITCHI (Matschiner 2015). We used the exonhap_all_data matrix and calculated a
226 gene tree for each locus using RAXML v.8 (Stamatakis 2014) under the GTRCAT model with
227 1,000 bootstrap replicates using the rapid bootstrap algorithm. The number of variable sites,
228 informative sites, and Tajima´s D for each locus was assessed using the package DENDROPY
229 (Sukumaran and Holder 2010;
230 https://github.com/srutschmann/python_scripts/blob/master/alignment_stats.py) and a custom
231 script.

232 ***2.4 Species assignment and population structure analysis***

233 Most analyses that use phylogenetic or multilocus species tree approaches require *a priori*
234 species assignment. Because of the partly unknown and largely incomplete taxonomy of the
235 group, we used two approaches to first assign the 29 *C. dipterum* individuals to putative
236 species: the gmyc approach (Fujisawa and Barraclough 2013) and a Bayesian clustering
237 algorithm to assign individuals to 'populations' (STRUCTURE, Pritchard et al., 2000; Falush
238 et al., 2003). The gmyc approach was carried out using *cox1* from 147 specimens that
239 included all newly sequenced *Cloeon* individuals, published sequences that were available as
240 of February 2016 (Supplementary Table 2), six newly sequenced individuals of *C. simile*
241 Easton, 1870, and *Baetis rhodani* (GenBank Acc. no. KF438126) as an outgroup. The
242 analysis followed that of Rutschmann et al., (2014) except that we used BEAST v.2.3.2
243 (Bouckaert et al., 2014) and a 2-partition scheme in which the first two codon positions were
244 modeled with HKY + I and the third codon position with HKY + Γ. For the Bayesian
245 clustering approach we used the exon_all_data matrix. We assumed 1-10 genotypic clusters
246 (K) and ran nine replicate analyses for each K, using 1×10^6 MCMC generations with a burn-

247 in of 10%. All individuals were assigned probabilistically without *a priori* knowledge to
248 genetic clusters. We applied an admixture model with default settings (Supplementary File 3).

249 ***2.5 Phylogenetic reconstruction***

250 We performed Bayesian phylogenetic reconstructions using all data (exon_all_data) and
251 the complete matrix (exon_complete_matrix) using MRBAYES v.3.2.2 (Ronquist et al., 2012).
252 As outgroup we used *Baetis* sp.. All individual locus alignments were concatenated using a
253 custom Python script
254 (https://github.com/srutschmann/python_scripts/blob/master/fasta_concat.py). For the tree
255 reconstruction, we implemented the best-fit models for each locus, and unlinked the
256 nucleotide frequencies, gamma distributions, substitution rates and the proportion of invariant
257 sites across partitions. Two independent analyses of four MCMC chains, each with 1×10^7
258 generations and 25% burn-in were run.

259 To investigate how the number of loci analyzed affected node support values, we
260 performed phylogenetic reconstructions based on concatenated sets of varying numbers of
261 randomly selected loci (Supplementary Table 4). The analyses were performed and
262 summarized as above. Linear regressions were used to predict the number of supported nodes
263 for Bayesian posterior probability (PP) ≥ 0.95 and PP = 1 as a function of the number of loci
264 used in the analysis. The Pearson correlation between the number of loci and number of
265 supported nodes was calculated separately for both PP, using the stats package in R (R
266 Development Core Team, 2016).

267 Species tree reconstructions were carried out under a multispecies coalescent framework
268 (Drummond and Rambaut 2007; Heled and Drummond 2010) as implemented in the program
269 *BEAST v.2.1.3 (Bouckaert et al., 2014). All analyses were performed using exons, one
270 analysis using all data and one using only the complete matrix, as above (Table 1;

271 exonhap_all_data, exonhap_complete_matrix). All individuals were *a priori* assigned to
272 species based on the gmyc and Bayesian clustering analyses described above. In the Bayesian
273 clustering analysis, one Russian individual was considered to be admixed based on PP
274 assignment values > 0.05 for more than one cluster (Supplementary File 3, Supplementary
275 Fig. 1, Supplementary Table 5). This individual was therefore excluded from further analysis.
276 We used a relaxed uncorrelated lognormal clock for gene tree estimation at each locus and a
277 Yule speciation-process prior. We conducted six independent runs of 8×10^8 million
278 generations each. Runs were combined in LOGCOMBINER v.2.1.3 (Bouckaert et al., 2014),
279 whereby all parameters reached effective sample sizes (ESS) > 600. Maximum clade
280 credibility trees for each species trees were obtained using TREEANNOTATOR v.2.1.3
281 (Bouckaert et al., 2014). As for the concatenated phylogenetics (above), we examined how
282 the number of loci included in the multilocus species tree analysis affected node support by
283 re-running the analysis using subsets of differing numbers of randomly selected loci
284 (Supplementary Table 4).

285 ***2.5 Ancestral state reconstruction***

286 An ancestral state reconstruction approach was used to test the direction of the radiation
287 (i.e. Continental to Island or Island to Continental). Ancestral range patterns of each
288 individual were defined into four geographic areas: (1) a broadly defined Continental
289 referring to the European and North American mainland, (2) Canary Islands, (3) Madeira, and
290 (4) Azores. As input tree, we used the concatenated tree based on the exon_all_data inferred
291 with MRBAYES. A chronogram was fit to the tree using the chronos function in the ape v.3.4
292 (Paradis et al., 2004) package in R. Ancestral states were estimated under an equal-rates (ER)
293 model using the function ace, and the scaled likelihoods of each ancestral state were
294 calculated using the function lik.anc in ape v.3.4. A MCMC approach was used to sample
295 character histories from their PP distribution generating 1,000 stochastic character maps with

296 the function make.simmap of the phytools v.0.4.98 (Revell 2012) package in R
297 (Supplementary File 4).

298 # 3. Results

299 ### *3.1 Development of nuclear DNA loci*

300 Whole-genome sequencing resulted in 1,109,684 raw reads, including 651,306 reads for
301 the shotgun library and 458,378 reads for the paired-end library, with an average large contig
302 length of 1,187 and 736 bp, respectively (BioSample SAMN03202660, BioProject
303 PRJNA268073, Sequence Read Archive SRP050093). All reads were assembled into 68,473
304 contigs with an N50 of 1,116 bp. The reads of the shotgun library were assembled into 31,827
305 contigs with an N50 of 1,260 bp. We detected 918 putative orthologous gene sequences for *C.*
306 *dipterum* from the contigs derived from the shotgun library, 1,298 putative orthologous gene
307 sequences from the contigs of the combined assembly, and 416 for *Baetis* sp. (Supplementary
308 Table 6). We successfully designed primer pairs for 59 sequence alignments (Supplementary
309 Table 3), mostly based on orthologous sequences from both taxa.

310 Total fragment length per sequenced locus ranged from 210 - 1,007 bp with a mean of 545
311 bp. Exon sequence length ranged from 210 - 710 bp with a mean of 410 bp (Supplementary
312 Table 3) (KF438124-KF438125, KU757080-KU757184, and KU971616-KU973191). The
313 full data matrix of all 29 individuals and all 59 loci including exons and introns (all_data) was
314 32,213 bp in length when concatenated and when introns were removed (exon_all_data) it
315 was 24,168 bp (Table 1). All individuals were successfully sequenced for at least 44 loci, and
316 the above matrices were >75% complete. The 100% complete matrix included 17 loci that
317 were sequenced successfully for all 29 individuals. All heterozygous indels were located in
318 the intron sequences. However, 100 heterozygous sites could not been resolved and remained
319 in the exonhap alignments.

320 All haplotype genealogies showed clear structuring (Supplementary Fig. 2). For 33 loci,
321 we found haplotypes shared between putative species. The number of variable sites per locus
322 ranged from six to 65 (mean: 18.95). These values were lower than those reported above
323 because ambiguous sites were not considered variable in the haplotype analysis. In the
324 exon_all_data matrix, there was one SNP per every 21.62 nucleotides sequenced (i.e. total
325 length per total number of variable sites). The loci included between six and 54 informative
326 sites (mean: 16) and one to 26 ambiguous sites (mean: 8.4). Nucleotide diversity ranged from
327 0.007 to 0.04 (mean: 0.017), and Tajima´s D varied between -0.85 and 1.97 (mean: 0.29)
328 (Supplementary Table 7).

329 ***3.2 Species assignment and population structure***

330 There were 62 unique *cox1* haplotypes of *Cloeon* and the gmyc model was a significantly
331 better fit to the data than the null ($\chi^2 = 31.00$, $p < 0.001$). There were seven putative species
332 delineated within *C. dipterum* s.l. (Fig. 2a): One occurred only in Asia (South Korea) while
333 the remaining six included three species with distributions that included the Macaronesian
334 Islands (IS1- IS3) and three species only occurring on the European and North American
335 continents (CT1- CT3). The population assignments from the Bayesian clustering analyses of
336 nDNA (Supplementary File 3, Supplementary Fig. 1, and Supplementary Table 5) agreed
337 completely with the results from the gmyc analysis. Among these six, one widespread species
338 (IS1) was found on all Azorean islands, in Greece and Italy, and in North America, one on the
339 Canary Islands and Madeira (IS2), and one only on four of the Canary Islands (IS3). The
340 model recognized all seven *C. dipterum* gmyc species even when using the most conservative
341 estimate (95% CI based on two log likelihood units: 16-19 gmyc species). The two *C.*
342 *cognatum* specimens from the North American DNA barcoding project (Webb et al., 2012)
343 had *cox1* haplotypes identical to our gmyc species IS1.

344 ***3.3 Phylogenetic reconstruction***

345 Analyses based on both exon matrices (exon_all_data; exon_complete_matrix) recovered
346 the same tree topology with strong node support, resolving each of the three species occurring
347 on Macaronesia (IS1-IS3) as monophyletic and members of a monophyletic 'Island clade'
348 (Fig. 3a). The geographically widespread species IS1 was sister taxon to the two others.
349 Species CT2 and CT3 were both monophyletic and sister group to the Island clade (Fig. 3a).
350 All individuals in CT1 were monophyletic except for a single individual that was sister taxon
351 to the entire *C. dipterum* s.l. lineage. There were 27 resolved (PP ≥ 0.95) nodes in the full
352 matrix (all_data) tree; the only unresolved node was between the two Azorean individuals
353 (Fig. 3a). In contrast, the complete matrix tree contained only 19 resolved nodes, with lack of
354 resolution most pronounced in IS2 (Supplementary Fig. 3).

355 All species tree phylogenies had identical topologies and these matched the Bayesian
356 phylogenies (Figs. 2b, 3a) in that Island and Continental clade both were monophyletic, with
357 IS1 sister taxon to IS2 + IS3, and with CT1 sister taxon to CT2 + CT3. Using the
358 exon_complete_matrix, all nodes were highly supported (PP ≥ 0.99; Table 2). All individuals
359 clustered into six species in the same way in both the multilocus nDNA tree and the single-
360 locus (*cox1*) mtDNA tree (PP = 1), but the relationships among the species were different.
361 The mtDNA tree did not support the sister relationship of IS2 + IS3 or the monophyly of the
362 Continental clade (Fig. 2a; Table 2).

363 There was a strong positive relationship ($R^2 = 0.83$, $p < 0.001$ for PP ≥ 0.95, and $R^2 = 0.75$,
364 $p < 0.001$ for PP = 1) between the number of loci employed and the number of nodes resolved
365 for the concatenation approach (Fig. 3b). The relationship was less clear in the multispecies
366 coalescent analysis, with node support in the species tree varying more widely with the
367 number of loci employed. The highest overall support came from analysis of 17 and 40 loci,

368 although only the analysis using 20 loci failed to recover either node in the Macaronesian
369 clade and resulted in no resolution other than continental monophyly (Fig. 2b, Table 2).

370 ***3.4 Ancestral state reconstruction***

371 The ancestral state reconstruction identified four nodes showing marginal states with less
372 than 0.9 Bayesian PP for one character, including sister relationship between individual
373 CH010_SR21B07 and the remaining species, the ancestral node of IS2 + IS3 (Canary Islands
374 and Madeira), and the nodes separating Madeiran from Canarian individuals within IS2. The
375 Island clade had a continental origin, further a Canarian origin was estimated for IS2 + IS3.
376 The clade IS2 was estimated to have an ancestral state of 0.59 for Madeira and 0.4 for the
377 Canary Islands (Supplementary File 4).

378 4. Discussion

379 ***4.1 Number of loci for phylogenetics***

380 A recent study by O'Neill et al., (2013) examined how multilocus species tree inferences
381 varied with differing number of loci. In their study, analysis based on the 20 and 30 most
382 informative loci (using a parsimony criterion) resulted in high PPs, whereas node support
383 values were lower and likelihoods failed to converge when less informative loci were added
384 to the analysis. They concluded this was the result of the increasing number of parameters
385 while adding loci with decreasing levels of information. Our results are not directly
386 comparable to those of O'Neill et al. (2013) for the species tree reconstruction, because we
387 did not explicitly order loci by parsimony-informative sites in our tests. Nonetheless, we
388 found a strong negative correlation (Pearson R = -0.95) between the mean number of
389 informative sites per locus and mean node support in the coalescent species tree (data not
390 shown). This suggests that the number of informative sites was not able to explain variation in
391 support alone, and that multiple characteristics of individual loci play an important role in

392 whether or not analyses achieve convergence and tree resolution. For the concatenation
393 approach, we found a positive linear correlation between number of loci and node support.
394 This was despite the larger number of parameters. Simulations based on 200 to 300 loci
395 showed that the divergence time estimation using 50 loci are robust (Shen et al., 2016). In our
396 study we observed that the reduction in node support when using a reduced set of loci
397 (exon_complete_matrix *vs.* exon_all_data, see section 2.3) primarily affected the most
398 derived clade (IS2), which highlights the importance of large nDNA marker sets for the
399 reconstruction of shallow phylogenies.

400 ***4.2 Species delineation***

401 The perfect agreement of the Bayesian clustering and mitochondrial gmyc approaches for
402 *a priori* species delineation support the use of *cox1* as barcoding gene for the taxa studied
403 (e.g. Lucentini et al., 2011; Pereira-da-Conceicoa et al., 2012; Webb et al., 2012; Rutschmann
404 et al., 2014). The distant clustering of one individual (CH010_SR21B07) in the concatenated
405 tree analyses might be explained by incomplete lineage sorting since the species tree
406 inferences using *BEAST did result in a clear clustering of CT1 with low frequency of
407 different topology (Fig. 3a). Moreover, when incomplete lineage sorting is present, standard
408 methods for estimating species trees, such as concatenation and consensus methods, can be
409 statistically inconsistent (Degnan et al., 2009; Roch and Steel 2014), and produce highly
410 supported but incorrect trees (Kubatko and Degnan 2007). The majority of gene trees could
411 support an incorrect species tree if the phylogeny is in the anomaly zone (Degnan and
412 Rosenberg 2006). However, here this does not seem to be the case, otherwise one would
413 expect the concatenation and coalescent approach to support different topologies (Kubatko
414 and Degnan 2007; Liu and Edwards 2009). The inferred haplotype networks illustrate the
415 necessity of using several individuals per species. For example, the individuals of IS2 shared
416 several haplotypes with other species, indicating incomplete lineage sorting between the

417 different species.. Originally, it was thought that *BEAST analyses would be quite robust in
418 the presence of gene flow while migration is problematic (Heled et al., 2013). However,
419 Leaché et al., (2014) have shown that gene flow can alter species trees, ranging from
420 decreasing PPs for low gene flow up to altering the species tree topology when high levels of
421 gene flow occur.

422 ***4.3 Species diversity***

423 The use of nDNA and geographically extensive sampling uncovered a largely
424 underestimated species diversity for *C. dipterum* s.l. species complex, supporting the
425 existence of six geographically relevant species from our study (with a seventh in Asia).
426 Recent evidence from the study of another mayfly species found fine-scale ecological
427 differences among cryptic species detected with molecular methods (Macher et al., 2016),
428 lending support to the ecological and evolutionary significance of these and other DNA-based
429 findings. Another widespread species, *Baetis harrisoni*, was also found to consist of several
430 cryptic species (Pereira-da-Conceicoa et al., 2012). In light of the unusually broad ecological
431 tolerance (among mayflies) observed for *C. dipterum*, we also conclude that the lineage
432 clearly consists of multiple independent species, as has been recognized by morphological
433 taxonomy for some of the members (e.g. *C. peregrinator*). All of our analyses grouped two
434 specimens of *C. peregrinator* from Madeira with individuals from several Canary Islands into
435 species IS2. Gattolliat et al., (2008) described *C. peregrinator* as an endemic Madeiran
436 species based on morphological characters and support from mtDNA cytochrome-oxidase *b*
437 sequences. At the time, there were no nDNA sequences of Canarian *C. dipterum* s.l.
438 specimens available. Rutschmann et al., (2014) assigned all Madeiran *Cloeon* individuals to
439 *C. peregrinator* for their mtDNA phylogeny, but the specimens were not included in their
440 gmyc analysis because there were no *cox1* sequences. Based on our findings here, there is no
441 endemic *Cloeon* species on Madeira.

442 The focus of our study was Macaronesia and therefore nDNA results are only applicable to
443 these taxa, but the mtDNA gene tree provides evidence for broad cryptic diversity within the
444 subfamily Cloeoninae. Although these are single-locus data and must therefore be considered
445 preliminary, we note that the mtDNA and nDNA species delineation results were fully
446 congruent. *Cloeon simile* included two geographically widespread European gmyc species,
447 and *C. smaeleni* Lestage 1924 was two gmyc species, one with Saudi Arabian and one with
448 Afrotropical distribution. The species *C. praetexum* was clearly distinct from all other
449 examined European specimens, which was surprising because it is thought to belong to *C.*
450 *simile* s.l.. The two specimens of *C. cognatum*, which is thought to be a junior synonym of *C.*
451 *dipterum* by some authors, were nested within the IS1 clade. All of the above findings must
452 be considered preliminary because they are based on mtDNA, although we note that mtDNA
453 and nDNA markers agreed in all of the *Cloeon* species that were directly compared. Further
454 studies on these taxa with additional molecular markers, using morphological characteristics,
455 and including comparisons with previously described species that are now considered junior
456 synonyms or *species inquirenda* would be a valuable complement to the work presented here.

457 ***4.4 Evolution, colonization, and diversification***

458 For the species occurring in the Macaronesian region, one species appeared widely
459 distributed on all Canary Islands and Madeira (IS2), one species was found only on the
460 western group of the Canarian islands (IS3), and one species was found on five islands of the
461 Azores, in Italy, in Greece and in North America (IS1). The short branches and occurrence of
462 shared haplotypes of individuals from IS1 support very recent or perhaps ongoing gene flow.
463 Other studies have found evidence for recent or ongoing dispersal in *Cloeon* (e.g., Monaghan
464 et al., 2005) including a recent introduction of African *Cloeon* to South America (*C. smaeleni*,
465 Salles et al., 2014). This long-distance dispersal ability is probably at least partly related to
466 their reproductive flexibility including ovivipary and their ability to survive in anthropogenic

habitats. Our ancestral state reconstruction indicated that IS2 may have first colonized Madeira and then the Canaries from west to east. Colonization routes between these two archipelagos have been suggested for several taxa (Emerson et al., 2000a; Emerson et al., 2000b; Trusty et al., 2005; Illera et al., 2007; Dimitrov et al., 2008; Amorim et al., 2012). IS3 seems not to have reached La Palma and the two most eastern Canarian islands of Fuerteventura and Lanzarote. The dispersal of IS3 appears to have followed the progression rule, in which older islands are inhabited by older clades, which is further supported by stepping-stone dispersal along an east-western gradient.

Our data confirm at least three and possibly four independent colonization events of the islands studied, with a European origin for the Macaronesian *C. dipterum* s.l.; however, long branches between Continental clades and the Island clade suggest there may be missing intermediates. These may occur in the Iberian Peninsula or North Africa. Several studies have proposed a North African origin for both the Canarian and Madeiran fauna (Brunton and Hurst 1998; Kvist et al., 2005; Weingartner et al., 2006; Gohli et al., 2015; Stervander et al., 2015). The Continental clades are also distantly related to one another and the long branches within both clades, compared to the Island clades, suggest there may be additional European species that are not included here.

A strong effect of different habitat preferences between the two Canarian species, which might impact their colonization success, is evidenced. Although we recognize that our dataset was not quantitative, we observed that species IS3 generally occurred on islands with more potential habitats in comparison to IS2, which seems to have better dispersal abilities and might therefore be able to more successfully colonize islands with very little water occurrence. This pattern may be linked with the occurrence of suitable water habitats on the Canarian Islands, including the four islands of Gran Canaria, Tenerife, La Gomera, and La Palma, which all have permanent natural water sources, and the island of El Hierro where

492 several artificial standing water habitats exist due to mostly temperate climatic conditions,
493 whereas there are only very few habitats on Fuerteventura and Lanzarote due to the arid
494 climatic conditions. The effect of habitat use on species richness has been shown for aquatic
495 beetles (Ribera et al., 2003c), whereby running water bodies contain more species than
496 standing ones. This pattern also applies to the Macaronesian mayflies. The genus *Baetis*
497 occurs in running waters and is species-rich, including eight island endemic species on five
498 islands of Madeira and the Canary Islands (Rutschmann et al., 2014). In contrast, the genus
499 *Cloeon* comprises three species none of which are restricted to a single island. The impact of
500 agriculture and tourism on natural habitats (Malmqvist et al., 1995; Nilsson et al., 1998) has
501 clearly threatened the occurrence of species living in lotic habitats (*Baetis canariensis* and *B.*
502 *pseudorhodani*, Rutschmann et al., 2014), but it may have had less of an effect on *C.*
503 *dipterum*. The recent records of mayflies from El Hierro indicate a recent anthropogenic
504 import of the species, moreover because it is the youngest island of the Canarian archipelago
505 and its remote geographical position.

506 Interestingly, there were eight sampling sites (out of 32 examined sites, i.e. 25%) in which
507 both species IS2 and IS3 occurred sympatrically. Four of these localities were natural
508 habitats. However, more work needs to be done to make quantitative assessments on species
509 occurrence and local abundance of the two distinct species occurring on the same habitats. A
510 wider geographic sampling, focusing on the specimens from the European mainland and
511 North Africa will be needed to clarify the origin and distribution of the *C. dipterum* s.l.
512 species complex. We expect to find more individuals from distinct geographic localities
513 belonging to the species IS1, since this species seems to exhibit long distance trans-oceanic
514 dispersal abilities.

515

516 ## 5. Conclusion

517 Our aims were to delineate the species boundaries within the *C. dipterum* species complex
518 and place these lineages within a phylogenetic framework, in order to better understand their
519 evolution on the Macaronesian Islands. Robust phylogenetic reconstruction of such closely
520 related species can be challenging, but is a necessary step in the understanding of
521 evolutionary processes of diversification and adaptation. Most readily available nDNA loci
522 (e.g. rRNA) do not exhibit suitable polymorphism, resulting in a large dependence on mtDNA
523 for phylogenetics (Garrick et al., 2015). A distinct advantage of using multiple nDNA loci
524 comes from the advent of multilocus species tree reconstruction methods. These are important
525 tools in the reconstruction of relationships between close relatives, which is often intractable
526 based on single-locus (i.e. mtDNA) data. The difficulty in developing large numbers of
527 nDNA loci remains one of the primary reasons that there are few model systems available for
528 detailed studies of speciation and diversification processes. This is particularly true for
529 freshwater insects, despite their overwhelming contribution to global biodiversity (Dijkstra et
530 al., 2014). Here we developed a large set of nDNA loci using draft whole-genome sequencing
531 combined with sequences from a published EST library that was constructed using a different
532 subfamily. All of the procedures we used have since been incorporated into a single analysis
533 pipeline (Rutschmann et al., accepted). Our results show that even for taxa with very limited
534 available genomic resources, it is possible to develop sets of nuclear loci that produce fully
535 resolved and supported coalescent-based species trees and single-matrix phylogenetic trees.
536 Using these results, we were able to infer species boundaries within the largely cryptic *C.*
537 *dipterum* s.l. species complex and reconstruct the diversification and island colonization
538 history of these species with confidence.

539

540

Author contributions

S.R., M.S. and M.T.M. conceived the study. S.R., H.D., D.H.F., J.-L.G., S.J.H., and P.M.R., provided samples. S.R., H.D., S.S., and R.D. contributed analytical tools. S.R., and M.T.M performed and interpreted the analyses. S.R. and M.T.M wrote the manuscript. All authors provided comments and approved the final manuscript.

Acknowledgements

We are grateful to B. Ortiz Crespo, S. Mbedi, K. Preuß, and L.Wächter for laboratory work; D. Murányi, A. Wagner, A. Przhiboro, P. Manko, M. F. Geiger, K. Kurzrock, L.F. Pires Braz, K. C. Gritzalis, and M. Alp for field work; P. Rutschmann and the HPC Service of ZEDAT, Freie Universität Berlin for access to high-performance computing resources. We are greatly indebted to M. Báez for obtaining sampling permits for the Canary Islands and to the authorities who provided us with the collection permissions. We are very grateful to our research groups, especially to I. Lucas Lledó and M. Gamboa for their constructive comments on this work. This is publication number ### of the Berlin Center for Genomics in Biodiversity Research. This work was supported by the Leibniz Association (PAKT für Forschung und Innovation) project FREDIE (SAW-2011-ZFMK-3 to M.T.M.) and by a travel award from the Leibniz-Institute of Freshwater Ecology and Inland Fisheries to S.R.. Individual support was provided by the Japan Society for the Promotion of Science (Long-Term Research Fellowship L15543 to M.T.M.), the Swiss National Science Foundation (Early PostDoc.Mobility fellowship P2SKP3_158698 to S.R.), the European Investment Funds by FEDER/COMPETE/POCI – Operacional Competitiveness and Internationalisation Programme (POCI-01-0145-FEDER-006958 to S.J.H.), and the National Funds by FCT - Portuguese Foundation for Science and Technology (UID/AGR/04033/2013 to S.J.H. and SFRH/BPD/99461/2014 to P.M.R.).

References

Amorim, I.R., Emerson, B.C., Borges, P.A.V., Wayne, R.K., 2012. Phylogeography and molecular phylogeny of Macaronesian island *Tarphius* (Coleoptera: Zopheridae): why are there so few species in the Azores? J. Biogeogr. 39, 1583-1595.

Arnedo, M.A., Oromi, P., De Abreu, S.M., Ribera, C., 2008. Biogeographical and evolutionary patterns in the Macaronesian shield-backed katydid genus *Calliphona* Krauss, 1892 (Orthoptera : Tettigoniidae) and allies as inferred from phylogenetic analyses of multiple mitochondrial genes. Syst. Entom. 33, 145-158.

Avise, J.C., Nelson, W.S., Bowen, B.W., Walker, D., 2000. Phylogeography of colonially nesting seabirds, with special reference to global matrilineal patterns in the sooty tern (*Sterna fuscata*). Mol. Ecol. 9, 1783-1792.

Barber-James, H.M., Gattolliat, J.-L., Sartori, M., Hubbard, M.D., 2008. Global diversity of mayflies (Ephemeroptera, Insecta) in freshwater. Hydrobiologia 595, 339-350.

Bauernfeind, E., Soldán, T. 2012. The Mayflies of Europe. Ollerup, Apollo Books.

Bouckaert, R., Heled, J., Kuhnert, D., Vaughan, T., Wu, C.H., Xie, D., Suchard, M.A., Rambaut, A., Drummond, A.J., 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comp. Biol. 10, e1003537.

Brinck, P., Scherer, E. 1961. On the Ephemeroptera of the Azoreas and Madeira. Boletim do Museu Municipal do Funchal 47, 55-66.

Brunton, C.F.A., Hurst, G.D.D., 1998. Mitochondrial DNA phylogeny of Brimstone butterflies (genus *Gonepteryx*) from the Canary Islands and Madeira. Biol. J. Linn. Soc. Lond. 63, 69-79.

587 Darriba, D., Taboada, G.L., Doallo, R., Posada, D., 2012. jModelTest 2: more models, new
588 heuristics and parallel computing. Nat. Methods 9, 772.
589 Degnan, J.H., DeGiorgio, M., Bryant, D., Rosenberg, N.A., 2009. Properties of consensus
590 methods for inferring species trees from gene trees. Syst. Biol. 58, 35-54.
591 Degnan, J.H., Rosenberg, N.A., 2006. Discordance of species trees with their most likely
592 gene trees. PLoS Genet. 2, e68.
593 Degnan, J.H., Rosenberg, N.A., 2009. Gene tree discordance, phylogenetic inference and the
594 multispecies coalescent. Trends Ecol. Evol. 24, 332-340.
595 Degnan, J.H., Rosenberg, N.A., Stadler, T., 2012. A characterization of the set of species
596 trees that produce anomalous ranked gene trees. IEEE/ACM Trans. Comput. Biol.
597 Bioinform. 9, 1558-1568.
598 Díaz-Pérez, A.J., Sequeira, M., Santos-Guerra, A., Catalán, P., 2012. Divergence and
599 biogeography of the recently evolved Macaronesian red *Festuca* (Gramineae) species
600 inferred from coalescence-based analyses. Mol. Ecol. 21, 1702-1726.
601 Dijkstra, K.D., Monaghan, M.T., Pauls, S.U., 2014. Freshwater biodiversity and aquatic
602 insect diversification. Annu. Rev. Entomol. 59, 143-163.
603 Dimitrov, D., Arnedo, M.A., Ribera, C., 2008. Colonization and diversification of the spider
604 genus *Pholcus* Walckenaer, 1805 (Araneae, Pholcidae) in the Macaronesian archipelagos:
605 evidence for long-term occupancy yet rapid recent speciation. Mol. Phylogenet. Evol. 48,
606 596-614.
607 Drotz, M.K., 2003. Speciation and mitochondrial DNA diversification of the diving beetles
608 *Agabus bipustulatus* and *A. wollastoni* (Coleoptera, Dytiscidae) within Macaronesia. Biol.
609 J. Linn. Soc. Lond. 79, 653-666.
610 Drummond, A.J., Rambaut, A., 2007. BEAST: Bayesian evolutionary analysis by sampling
611 trees. BMC Evol. Biol. 7, 214.
612 Edwards, S.V., 2009. Is a new and general theory of molecular systematics emerging?
613 Evolution 63, 1-19.
614 Emerson, B.C., 2002. Evolution on oceanic islands: molecular phylogenetic approaches to
615 understanding pattern and process. Mol. Ecol. 11, 951-966.
616 Emerson, B.C., Kolm, N., 2005. Species diversity can drive speciation. Nature 434, 1015-
617 1017.
618 Emerson, B.C., Oromí, P., 2005. Diversification of the forest beetle genus *Tarphius* on the
619 Canary Islands, and the evolutionary origins of island endemics. Evolution 59, 586-598.
620 Emerson, B.C., Oromí, P., Godfrey, M.H., 2000a. Interpreting colonization of the *Calathus*
621 (Coleoptera: Carabidae) on the Canary Islands and Madeira through the application of the
622 parametric bootstrap. Evolution 54, 2081-2090.
623 Emerson, B.C., Oromí, P., Hewitt, G.M., 2000b. Tracking colonization and diversification of
624 insect lineages on islands: mitochondrial DNA phylogeography of *Tarphius canariensis*
625 (Coleoptera: Colydiidae) on the Canary Islands. Proc. Biol. Sci. 267, 2199-2205.
626 Falush, D., Stephens, M., Pritchard, J.K., 2003. Inference of population structure using
627 multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164,
628 1567-1587.
629 Faria, C.M.A., Machado, A., Amorim, I.R. Gage, M.J.G., Borges, P.A.V., Emerson, B.C.,
630 2016. Evidence for multiple founding lineages and genetic admixture in the evolution of
631 species within an oceanic island weevil (Coleoptera, Curculionidae) super-radiation. J.
632 Biogeogr.. 43, 178-191.
633 Fitch, W.M., 1970. Distinguishing homologous from analogous proteins. Syst. Zool. 19, 99-
634 113.
635 Flot, J.F., 2010. seqphase: a web tool for interconverting phase input/output files and fasta
636 sequence alignments. Mol. Ecol. Resour. 10, 162-166.

637 Fujisawa, T., Barraclough, T.G., 2013. Delimiting species using single-locus data and the
638 Generalized Mixed Yule Coalescent approach: a revised method and evaluation on
639 simulated data sets. Syst. Biol. 62, 707-724.
640 Garrick, R.C., Bonatelli, I.A., Hyseni, C., Morales, A., Pelletier, T.A., Perez, M.F., Rice, E.,
641 Satler, J.D., Symula, R.E., Thomé, M.T., Carstens, B.C., 2015. The evolution of
642 phylogeographic data sets. Mol. Ecol. 24, 1164-1171.
643 Garrick, R.C., Sunnucks, P., Dyer, R.J., 2010. Nuclear gene phylogeography using PHASE:
644 dealing with unresolved genotypes, lost alleles, and systematic bias in parameter
645 estimation. BMC Evol. Biol. 10, 118.
646 Gattolliat, J.-L., Hughes, S.J., Monaghan, M.T., Sartori, M., 2008. Revision of Madeiran
647 mayflies (Insecta, Ephemeroptera). Zootaxa 1957, 52-68.
648 Giarla, T.C., Esselstyn, J.A., 2015. The challenges of resolving a rapid, recent radiation:
649 empirical and simulated phylogenomics of philippine shrews. Syst. Biol. 64, 727-740.
650 Gillespie, R.G., Croom, H.B., Palumbi, S.R., 1994. Multiple origins of a spider radiation in
651 Hawaii. Proc. Natl. Acad. Sci. U.S.A. 91, 2290-2294.
652 Gillespie, R.G., Roderick, G.K., 2002. Arthropods on islands: colonization, speciation, and
653 conservation. Annu. Rev. Entomol. 47, 595-632.
654 Gohli, J., Leder, E.H., Garcia-Del-Rey, E., Johannessen, L.E., Johnsen, A., Laskemoen, T.,
655 Popp, M., Lifjeld, J.T., 2015. The evolutionary history of Afrocanarian blue tits inferred
656 from genome wide SNPs. Mol. Ecol. 24, 180-191.
657 Grant, P., Grant, R., 2008. How and why species multiply: the radiation of Darwin's finches.
658 New Jersey, Princeton University Press.
659 Guindon, S., Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large
660 phylogenies by maximum likelihood. Syst. Biol. 52, 696-704.
661 Harrigan, R.J., Mazza, M.E., Sorenson, M.D., 2008. Computation vs. cloning: evaluation of
662 two methods for haplotype determination. Mol. Ecol. Resour. 8, 1239-1248.
663 Heled, J., Bryant, D., Drummond, A.J., 2013. Simulating gene trees under the multispecies
664 coalescent and time-dependent migration. BMC Evol. Biol. 13, 44.
665 Heled, J., Drummond, A.J., 2010. Bayesian inference of species trees from multilocus data.
666 Mol. Biol. Evol. 27, 570-580.
667 Hughes, S.J., Malmqvist, B., 2005. Atlantic Island freshwater ecosystems: challenges and
668 considerations following the EU Water Framework Directive. Hydrobiologia 8, 289-297.
669 Illera, J.C., Emerson, B.C., Richardson, D.S., 2007. Population history of Berthelot's pipit:
670 colonization, gene flow and morphological divergence in Macaronesia. Mol. Ecol. 16,
671 4599-4612.
672 Jordal, B.H., Hewitt, G.M., 2004. The origin and radiation of Macaronesian beetles breeding
673 in Euphorbia: the relative importance of multiple data partitions and population sampling.
674 Syst. Biol. 53, 711-734.
675 Juan, C., Oromí, P., Hewitt, G.M., 1997. Molecular phylogeny of darkling beetles from the
676 Canary Islands: comparison of inter island colonization patterns in two genera. Biochem.
677 Syst. Ecol. 25, 121-130.
678 Juan, C., Emerson, B.C., Oromí, P., Hewitt, G.M., 2000. Colonization and diversification:
679 towards a phylogeographic synthesis for the Canary Islands. Trends Ecol. Evol. 15, 104-
680 109.
681 Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7:
682 improvements in performance and usability. Mol. Biol. Evol. 30, 772-780.
683 Knowles, L.L., Kubatko, L.S., 2010. Estimating Species Trees: Practical and Theoretical
684 Aspects. Wiley-Blackwell.

Kozak, K.M., Wahlberg, N., Neild, A.F., Dasmahapatra, K.K., Mallet, J., Jiggins, C.D., 2015. Multilocus species trees show the recent adaptive radiation of the mimetic *Heliconius* butterflies. Syst. Biol. 64, 505-524.

Kubatko, L.S., Degnan, J.H., 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst. Biol. 56, 17-24.

Kvist, L., Broggi, J., Illera, J.C., Koivula, K., 2005. Colonisation and diversification of the blue tits (*Parus caeruleus* teneriffae-group) in the Canary Islands. Mol. Phylogenet. Evol. 34, 501-511.

Leaché, A.D., Fujita, M.K., Minin, V.N., Bouckaert, R.R., 2014. Species delimitation using genome-wide SNP data. Syst. Biol. 63, 534-542.

Liu, L., Edwards, S.V., 2009. Phylogenetic analysis in the anomaly zone. Syst. Biol. 58, 452-460.

Lopez, J.V., Yuhki, N., Masuda, R., Modi, W., O'Brien, S.J., 1994. Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. J. Mol. Evol. 39, 174-190.

Losos, J.B., Ricklefs, R.E., 2009. Adaptation and diversification on islands. Nature 457, 830-836.

Lucentini, L., Rebora, M., Puletti, M.E., Gigliarelli, L., Fontaneto, D., Gaino, E., Panara, F., 2011. Geographical and seasonal evidence of cryptic diversity in the *Baetis rhodani* complex (Ephemeroptera, Baetidae) revealed by means of DNA taxonomy. Hydrobiologia 673, 215-228.

Macher, J.N., Salis, R.K., Blakemore, K.S., Tollrian, R., Matthaei, C.D., Leese, F., 2016. Multiple-stressor effects on stream invertebrates: DNA barcoding reveals contrasting responses of cryptic mayfly species. Ecol. Indic. 61, 159-169.

Maddison, W.P., 1997. Gene trees in species trees. Syst. Biol. 46:523.

Maddison, W.P., Maddison, D.R., 2011. Mesquite: a modular system for evolutionary analysis. http://mesquiteproject.org (accessed 25.03. 2016).

Malmqvist, B., Nilsson, A.N., Báez, M., 1995. Tenerife's freshwater macroinvertebrates: status and threats (Canary Islands, Spain). Aquat. Conserv. 5, 1-24.

Matschiner, M., 2015. Fitchi: Haplotype genealogy graphs based on the Fitch algorithm. Bioinformatics 32, 1250-1252.

Monaghan, M.T., Balke, M.M., Pons, J.J., Vogler, A.P., 2006. Beyond barcodes: complex DNA taxonomy of a South Pacific Island radiation. Proc. Biol. Sci. 273, 887-893.

Monaghan, M.T., Gattolliat, J.L., Sartori, M., Elouard, J.M., James, H., Derleth, P., Glaizot, O., de Moor, F., Vogler, A.P., 2005. Trans-oceanic and endemic origins of the small minnow mayflies (Ephemeroptera, Baetidae) of Madagascar. Proc. Biol. Sci. 272, 1829-1836.

Morvan, C., Malard, F., Paradis, E., Lefebure, T., Konecny-Dupre, L., Douady, C.J., 2013. Timetree of Aselloidea reveals species diversification dynamics in groundwater. Syst. Biol. 62, 512-522.

Nilsson, A.N., Malmqvist, B., Báez, M., Blackburn, J.H., Armitage, P.D., 1998. Stream insects and gastropods in the island of Gran Canaria (Spain). Ann. Limnol.-Int. J. Lim.. 34, 413-435.

Nogales, M., Delgado, J.D., Medina, F.M., 1998. Shrikes, lizards and *Lycium intricatum* (Solanaceae) fruits: a case of indirec seed dispersal on an oceanic island (Alegranza, Canary Islands). J. Ecol. 86, 866-871.

O'Neill, E.M., Schwartz, R., Bullock, C.T., Williams, J.S., Shaffer, H.B., Aguilar-Miguel, X., Parra-Olea, G., Weisrock, D.W., 2013. Parallel tagged amplicon sequencing reveals major lineages and phylogenetic structure in the North American tiger salamander (*Ambystoma tigrinum*) species complex. Mol. Ecol. 22, 111-129.

Paradis, E., Claude, J., Strimmer, K., 2004. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20, 289-290.

Pereira-da-Conceicoa, L.L., Price, B.W., Barber-James, H.M., Barker, N.P., de Moor, F.C., Villet, M.H., 2012. Cryptic variation in an ecological indicator organism: mitochondrial and nuclear DNA sequence data confirm distinct lineages of *Baetis harrisoni* Barnard (Ephemeroptera: Baetidae) in southern Africa. BMC Evol. Biol. 12, 26.

Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of population structure using multilocus genotype data. Genetics 155, 945-959.

R Core Team. 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org (acessed 26.03.2016).

Rannala, B., Yang, Z., 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164, 1645-1656.

Raposeiro, P.M., Cruz, A.M., Hughes, S.J., Costa, A.C., 2012. Azorean freshwater invertebrates: Status, threats and biogeographic notes. Limnetica 31, 13-22.

Revell, L., 2012. phytools: An R package for phylogenetic comparative biology (and other things). Methods Ecol. Evol. 3, 217-223.

Ribera, I., Bilton, D.T., Balke, M., Hendrich, L., 2003a. Evolution, mitochondrial DNA phylogeny and systematic position of the Macaronesian endemic *Hydrotarsus* Falkenström (Coleoptera: Dytiscidae). Syst. Entomol. 28, 493-508.

Ribera, I., Bilton, D.T., Vogler, A.P., 2003b Mitochondrial DNA phylogeography and population history of Meladema diving beetles on the Atlantic Islands and in the Mediterranean basin (Coleoptera, Dytiscidae). Mol. Ecol. 12, 153-167.

Ribera, I., Foster, G.N., Vogler, A.P., 2003c. Does habitat use explain large scale species richness patterns of aquatic beetles in Europe? Ecography 26, 145-152.

Roch, S., Steel, M., 2014. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. Theor. Popul. Biol. 100C, 56-62.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Hohna, S., Larget, B., Liu, L., Suchard, M.A., Huelsenbeck, J.P., 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61, 539-542.

Rutschmann, S., Detering, H., Simon, S., Fredslund, J., Monaghan, M.T. (accepted) DiscoMark: Nuclear marker discovery from orthologous sequences using draft genome data. Mol. Ecol. Resour.

Rutschmann, S., Gattolliat, J.-L., Hughes, S.J., Sartori, M., Monaghan, M.T. 2014. Evolution and island endemism of morphologically cryptic *Baetis* and *Cloeon* species (Ephemeroptera, Baetidae) on the Canary Islands and Madeira. Freshwater Biol. 59, 2516-2527.

Salles, F.F., Gattolliat, J.-L., Angeli, K.B., De-Souza, M.R., Goncalves, I.C., Nessimian, J.L., Sartori, M., 2014. Discovery of an alien species of mayfly in South America (Ephemeroptera). ZooKeys 1-16.

Schluter, D., 2000. The Ecology of Adaptive Radiation. New York, Oxford University Press.

Shen, X.X., Liang, D., Chen, M.Y., Mao, R.L., Wake, D.B., Zhang, P., 2016. Enlarged multilocus data set provides surprisingly younger time of origin for the Plethodontidae, the largest family of salamanders. Syst. Biol. 65, 66-81.

Soldán, T., Thomas, A., 1983. New a little-known species of mayflies (Ephemeroptera) from Algeria. Acta Entomol. Bohemos. 80, 356-376.

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312-1313.

Stauder, A., 1995. Survey of the Madeiran limnological fauna and their zoogeographical distribution. Boletim do Museu Municipal do Funchal 4, 715-723.

Stephens, M., Donnelly, P., 2003. A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am. J. Hum. Genet. 73. 1162-1169.

Stephens, M., Smith, N.J., Donnelly, P., 2001. A new statistical method for haplotype reconstruction from population data. Am. J. Hum. Genet. 68, 978-989.

Stervander, M., Illera, J.C., Kvist, L., Barbosa, P., Keehnen, N.P., Pruisscher, P., Bensch, S., Hansson, B., 2015. Disentangling the complex evolutionary history of the Western Palearctic blue tits (*Cyanistes* spp.) - phylogenomic analyses suggest radiation by multiple colonization events and subsequent isolation. Mol. Ecol. 24, 2477-2494.

Sukumaran, J., Holder, M.T., 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics 26, 1569-1571.

Toussaint, E.F., Condamine, F.L., Hawlitschek, O., Watts, C.H., Porch, N., Hendrich, L., Balke, M., 2015. Unveiling the diversification dynamics of australasian predaceous diving beetles in the cenozoic. Syst. Biol. 64, 3-24.

Trusty, J.L., Olmstead, R.G., Santos-Guerra, A., Sa-Fontinha, S., Francisco-Ortega, J., 2005. Molecular phylogenetics of the Macaronesian-endemic genus *Bystropogon* (Lamiaceae): palaeo-islands, ecological shifts and interisland colonizations. Mol. Ecol. 14, 1177-1189.

Tseng, S.-P., Li, S.-H., Hsieh, C.-H., Wang, H.-Y., Lin, S.-M., 2014. Influence of gene flow on divergence dating - implications for the speciation history of *Takydromus* grass lizards. Mol. Ecol. 23, 4770-4784.

Vuataz, L., Sartori, M., Gattolliat, J.-L., Monaghan, M.T., 2013. Endemism and diversification in freshwater insects of Madagascar revealed by coalescent and phylogenetic analysis of museum and field collections. Mol. Phylogenet. Evol. 66, 979-991.

Webb, J.M., Jacobus, L.M., Funk, D.H., Zhou, X,. Kondratieff, B., Geraci, C.J., DeWalt, R.E., Baird, D.J., Richard, B., Phillips, I., Herbert, P.D., 2012. A DNA barcode library for North American Ephemeroptera: progress and prospects. PLoS ONE 7, e38063.

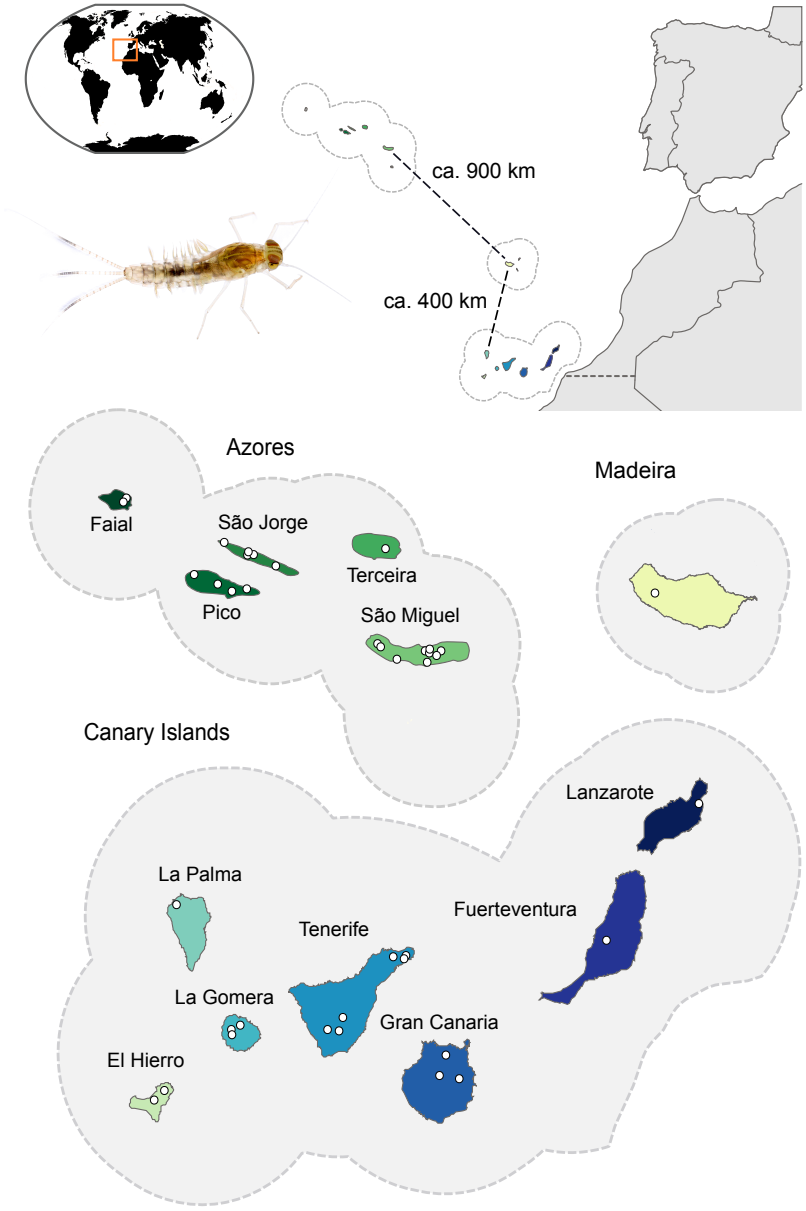
Weingartner, E., Wahlberg, N., Nylin, S., 2006. Speciation in *Pararge* (Satyrinae: Nymphalidae) butterflies - North Africa is the source of ancestral populations of all Pararge species. Syst. Entom. 31, 621-632.

Wielstra, B., Duijm, E., Lagler, P., Lammers, Y., Meilink, W.R., Ziermann, J.M., Arntzen, J.W., 2014. Parallel tagged amplicon sequencing of transcriptome-based genetic markers for *Triturus* newts with the Ion Torrent next-generation sequencing platform. Mol. Ecol. Resour. 14, 1080-1089.

Zhang, D.-X., Hewitt, G.M., 2003. Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. Mol. Ecol. 12, 563-584.

Zheng, Y., Peng, R., Kuro-o, M., Zeng, X,. 2011. Exploring patterns and extent of bias in estimating divergence time from mitochondrial DNA sequence data in a particular lineage: a case study of salamanders (order Caudata). Mol. Biol. Evol. 28, 2521-2535.

824 **Figures**



825

826 **Fig. 1.** Overview of the sampling localities on the Macaronesian region. The three
827 archipelagos of the Azores, Madeira, and the Canary Islands are shown in detail, whereby the
828 38 sampling sites are indicated by white dots. For the Azores and Madeira only islands with
829 sampling sites are shown in the detailed view. Photo of *Cloeon dipterum* s.l. larvae by
830 Amanda44 / CC BY 3.0.

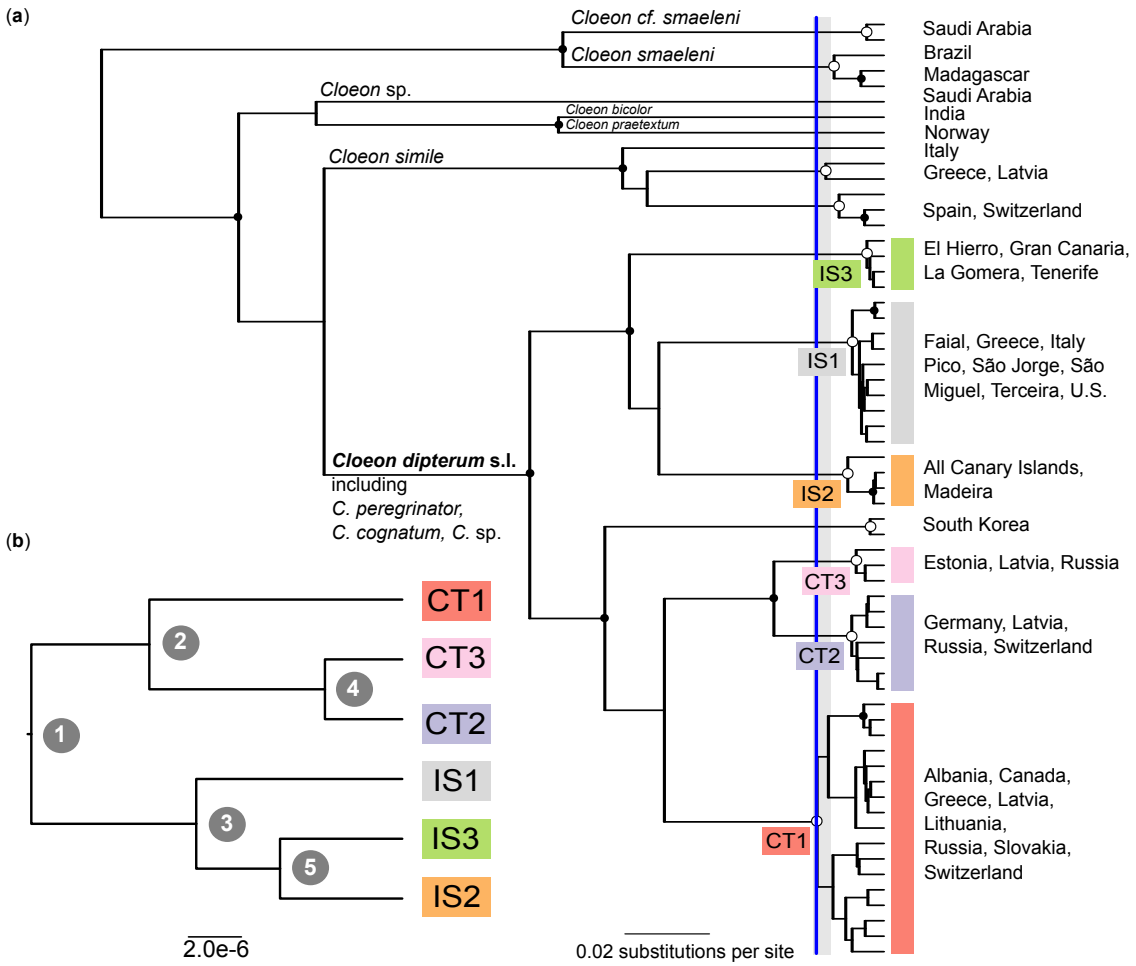


Fig. 2. Mitochondrial gene tree and nuclear species tree topology. Species delimitation from (**a**) the general mixed Yule-coalescent (gmyc) approach based on mitochondrial data (single-locus), and (**b**) the multispecies coalescent approach based on nuclear data (59 loci). (**a**), the ultrametric mitochondrial *cox1* gene tree was used as input for the gmyc analysis of *Cloeon* sp. All circles indicate well supported nodes (Bayesian posterior probability (PP) ≥ 0.95). Open circles at subtending nodes indicate sequence clusters corresponding to single gmyc species. Colors and alphanumeric codes indicate the six putative gmyc species. The outgroup *Baetis rhodani* is not shown. The blue line indicates the point of maximum-likelihood fit of the single-threshold gmyc model with 95% confidence interval in grey shading. Terminal labels indicate sampling regions (Supplementary Table 1). (**b**), the species trees of *C. dipterum* s.l. inferred by multispecies coalescent approach based on the exonhap_all_data matrix. Posterior probabilities of the five nodes are indicated in Table 2.

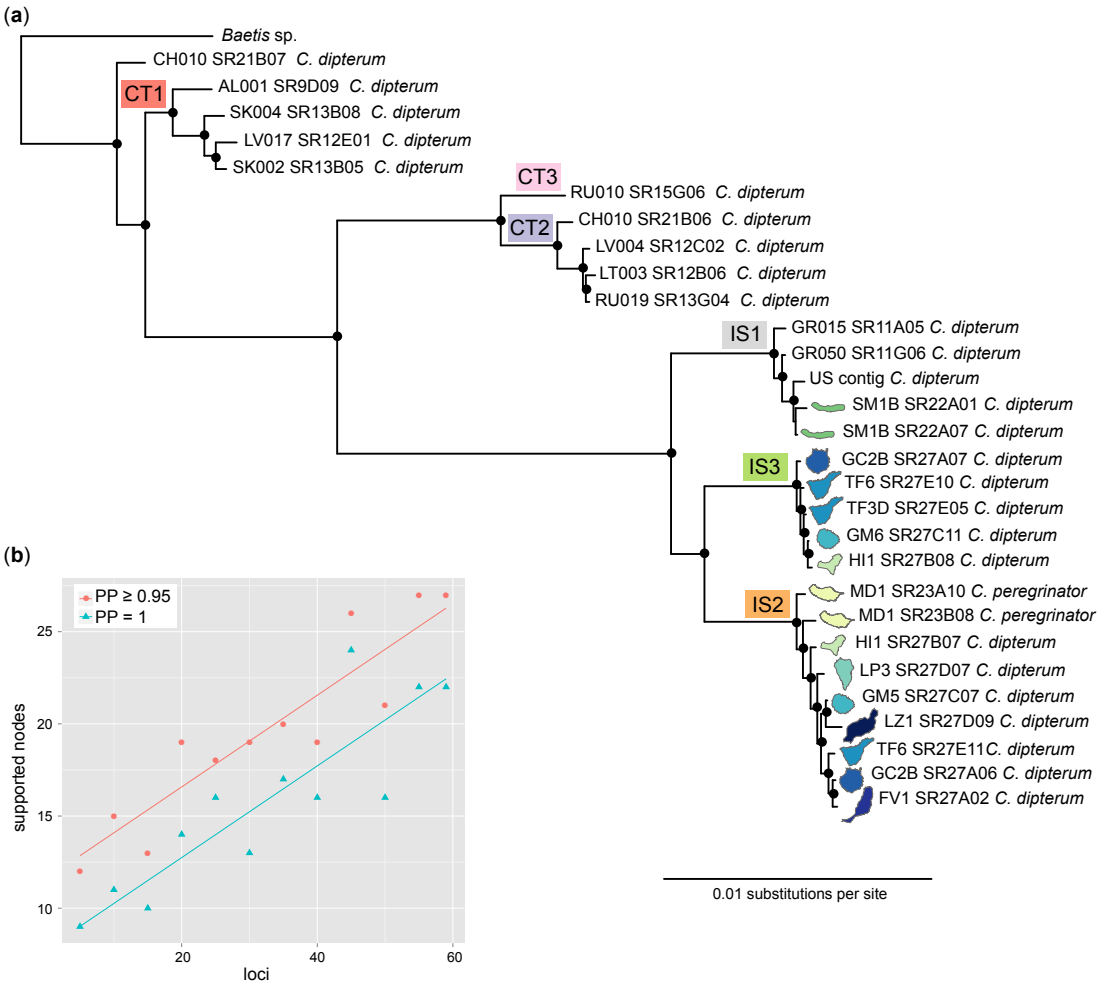


Fig. 3. Bayesian inference reconstruction of concatenation approach. (**a**), the phylogenetic relationships among *Cloeon dipterum* s.l. (including *C. peregrinator*) based on concatenated exon_all_data matrix. Filled circles indicate well supported nodes (Bayesian posterior probability (PP) ≥ 0.95). (**b**), the positive linear relationship between number of loci analyzed and number of supported nodes. Red circles represent nodes with PP ≥ 0.95 and blue triangles represent nodes with PP = 1. Sets of loci are reported in Supplementary Table 4. The R^2 for PP ≥ 0.95 were 0.83 ($p < 0.001$), and 0.75 ($p < 0.001$) for PP = 1.

853 **Tables**

854 **Table 1.** Overview of seven sequence alignments, including one based on mitochondrial sequences and six based on nuclear sequences. The
855 mitochondrial sequence alignment including only the species group of *Cloeon dipterum* s.l. comprised 130 taxa with 148 variable sites. Matrices
856 containing all taxa and all loci were >75% complete; the matrices containing only loci sequenced for all taxa were 100% complete. Exon
857 matrices refer to exon sequence alignments and the exonhap matrices refer to exon haplotype sequences.

858

Data Matrix	Concatenated Length [bp]	Number of Taxa	Number of Loci	Number of Variable Sites	Description
mitochondrial	658	148	1	240	All *cox1* sequences used for putative species assignment
all_data	32,213	29	59	2,481	All taxa and loci; introns and exons
exon_all_data	24,168	29	59	1,118	All taxa and loci; exons
complete_matrix	8,565	29	17	648	Only loci sequenced for all taxa; introns and exons
exon_complete_matrix	6,485	29	17	293	Only loci sequenced for all taxa; exons
exonhap_all_data		29	59	1,390	All taxa and loci; haplotypes of exons
exonhap_complete_matrix		29	17	361	Only loci sequenced for all taxa; haplotypes of exons

859

Table 2. Node support values of the species tree analysis (Fig. 2b) using six different sets of loci (Supplementary Table 4). Support values are given as Bayesian posterior probability (PP).

	Node Support					
Number of Loci	All (1)	Continental (2)	Island (3)	CT2+CT3 (4)	IS2+IS3 (5)	Mean
17	1.00	0.99	1.00	1.00	1.00	1.00
20	1.00	1.00	-	1.00	-	0.60
30	1.00	0.85	0.87	0.87	0.99	0.92
40	1.00	1.00	1.00	0.67	0.83	0.70
50	1.00	0.83	0.83	0.83	1.00	0.87
59	1.00	0.83	0.83	0.83	1.00	0.87