

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

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

<sup>a</sup>*Leibniz-Institute of Freshwater Ecology and Inland Fisheries (IGB), Müggelseedamm 301, 12587 Berlin, Germany*

<sup>b</sup>*Berlin Center for Genomics in Biodiversity Research, Königin-Luise-Straße 6-8, 14195 Berlin, Germany*

<sup>c</sup>*Department of Biochemistry, Genetics and Immunology, University of Vigo, 36310 Vigo, Spain*

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

<sup>e</sup>*Biosystematics Group, Wageningen University, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands*

<sup>f</sup>*Stroud Water Research Center, Avondale, Pennsylvania 19311, USA*

<sup>g</sup>*Musée cantonal de zoologie, Palais de Rumine, Place de la Riponne 6, 1014 Lausanne, Switzerland*

<sup>h</sup>*Department of Ecology and Evolution, Biophore, University of Lausanne, 1015 Lausanne, Switzerland*

<sup>i</sup>*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*

<sup>j</sup>*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, Department of Biochemistry, Genetics and Immunology, University of Vigo, 36310 Vigo, Spain, E-mail: [sereina.rutschmann@gmail.com](mailto: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). 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 begun 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

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

94 Macaronesia consists of four archipelagos (Azores, Madeira, Canary Islands, and Cape  
95 Verde) whose flora and fauna have been used in several studies as model systems for  
96 evolutionary research. Their distances to the adjacent continental mainland vary from 110 km  
97 (Fuerteventura in Canary Islands to Morocco) to more than 2000 km (Flores in the Azores to  
98 Portugal). Several colonization pathways have been identified (Juan et al., 2000; Emerson  
99 2002; Emerson and Kolm 2005), including a single colonization event followed by stepping-  
100 stone dispersal (Juan et al., 1997; Emerson and Oromi 2005; Illera et al., 2007; Arnedo et al.,

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

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

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

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

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

## 2. Material and methods

### 2.1 Development of nuclear DNA loci

To develop a set of nuclear loci we sequenced a newly created whole-genome library of *C. dipterum* (described in Rutschmann et al., 2016). Libraries were generated from laboratory-reared subimagos of *C. dipterum* specimens (full siblings). DNA was extracted from pooled specimens (five to 20) after removing eyes and wings using the Invisorb® Spin Tissue Mini kit (STRATEC, Berlin, Germany). Extracted DNA was precipitated using Isopropanol and pooled to obtain higher DNA yield. We prepared one 454 shotgun and one 454 paired-end library according to the manufacturer's guidelines (Rapid Library Preparation Method Manual, GS FLX+ Series - XL+, May 2011; Paired End Library Preparation Method Manual – 20 kb and 8 kb Span, GS FLX Titanium Series, October 2009). Fragments were amplified with an emulsion PCR (emPCR Method Manual - Lib-L SV, GS FLX Titanium Series, October 2009; Rev. Jan 2010). Four lanes per library were sequenced on a Roche (454) GS FLX machine. The sequence reads were trimmed and *de novo* assembled using NEWBLER v. 2.5.3 (454 Life Sciences Corporation) with default settings for large datasets. We generated two different assemblies, one using the reads from the shotgun library and one using the reads from both shotgun and paired-end libraries. For ortholog prediction, the newly assembled draft whole genomes were combined with 4,197 expressed sequence tag (EST) sequences from *Baetis* sp. (FN198828–FN203024; Simon et al., 2009). *Cloeon* and *Baetis* belong to the Baetidae subfamilies Cloeoninae and Baetinae. Primer pairs were designed in the conserved regions of orthologous sequences from included taxa. The above analysis procedures have been incorporated into the DISCOMARK pipeline for marker discovery and primer design (Rutschmann et al., 2016; see Supplementary File 2).

## 2.2 Taxon sampling and DNA extraction

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

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

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

We sequenced 60 loci for the study: the mtDNA barcoding gene (*cox1*) and 59 newly developed nDNA markers from protein-coding gene regions. The *cox1* locus was amplified and sequenced using the procedure described by Rutschmann et al., 2014. Based on a general mixed Yule-coalescent (gmyc) model analysis (Fujisawa and Barraclough 2013) of *cox1*, we selected a representative set of 29 individuals, including a set of one to nine individuals for each gmyc species, depending on the number of locations at which a given gmyc species was found (see 2.4 Species assignment and population structure analysis). For these individuals, we obtained nDNA sequences using the 59 newly designed primer pairs (Supplementary Table 3). Nuclear loci were amplified using standard polymerase chain reaction (PCR) protocols with an annealing temperature of 55°C. The PCR products were custom purified and sequenced at Beckman Coulter Genomics (Essex, UK) or Macrogen (Amsterdam, The Netherlands). Forward and reverse sequences were assembled and edited using GENEIOUS R7 v.7.1.3 (Biomatters Ltd.). Length variation (i.e. heterozygous indels) was decoded using CODONCODE ALIGNER v.3.5.6 (CodonCode Corporation, Centerville MA, USA). Additionally, we included published sequences of six nDNA loci from four individuals (KU971838-KU971840, KU971851, KU971919-KU971921, KU971933, KU972490-KU972492, KU972503, KU972568-KU972570, KU972583, KU972653-KU972654, KU972666, KU973060-KU973062, KU973074; Rutschmann et al., 2016).

Multiple sequence alignments were made for each locus using MAFFT v.7.050b (L-INS-I algorithm with default settings; Katoh and Standley 2013). The predicted orthologous sequences of *Baetis* sp. were used to check the correct exon-intron splicing boundaries (canonical and non-canonical splice site pairs) of each alignment. Exon-intron boundaries of locus 411912 could not be fully reconstructed and thus we used the exon sequence predicted from tblastx searches for subsequent analyses. Locus alignments were split into coding and non-coding parts using a custom script ([https://github.com/srutschmann/python\\_scripts/blob/master/extract\\_introns.py](https://github.com/srutschmann/python_scripts/blob/master/extract_introns.py)). All coding alignments were checked for indels and stop codons using MESQUITE v.2.75 (Maddison and Maddison 2011). Genotypes of the coding alignments were phased using the probabilistic Bayesian algorithm implemented in PHASE v.2.1.1 (Stephens et al., 2001; Stephens and Donnelly 2003) with a cutoff value of 0.6 (Harrigan et al., 2008; Garrick et al., 2010). Multiple runs were performed for each alignment and phase calls checked for consistency. Input and output files were formatted using the scripts from SEQPHASE (Flot 2010). Heterozygous sites that could not be resolved were coded using ambiguity codes and remained in the data set for subsequent sequence analyses. All alignments were re-aligned after phasing with MAFFT. As an outgroup we used *Baetis* sp.. The number of variable sites, informative sites, and Tajima's D for each locus were determined using a custom script and the package DENDROPY (Sukumaran and Holder 2010; [https://github.com/srutschmann/python\\_scripts/blob/master/alignment\\_stats.py](https://github.com/srutschmann/python_scripts/blob/master/alignment_stats.py)) and a custom script.

## 2.4 Species assignment and population structure analysis

Most analyses that use phylogenetic or multilocus species tree approaches require *a priori* species assignment. Because of the partly unknown and largely incomplete taxonomy of the group, we used two approaches to first assign the 29 *C. dipterum* individuals to putative

species. The first was a gmyc analysis (Fujisawa and Barraclough 2013) of mitochondrial *coxI* and the second was a Bayesian clustering algorithm using all nuclear loci to assign individuals to ‘populations’ (STRUCTURE, Pritchard et al., 2000; Falush et al., 2003). The gmyc approach was carried out using *coxI* from 147 specimens that included all newly sequenced *Cloeon* individuals, published sequences that were available as of February 2016 (Supplementary Table 2), six newly sequenced individuals of *C. simile* Eaton, 1870, and *Baetis rhodani* (KF438126) as an outgroup. The analysis followed that of Rutschmann et al., (2014) except that we used BEAST v.2.3.2 (Bouckaert et al., 2014) and the first two codon positions were modeled with HKY + I while the third codon position was modeled with HKY +  $\Gamma$ . For the Bayesian clustering we used the exon\_all\_data matrix (see 2.5 Phylogenetic and species tree reconstructions). We assumed 1-10 genotypic clusters (K) and ran nine replicate analyses for each K, using  $10^6$  MCMC generations with a burn-in of 10%. All individuals were assigned probabilistically without *a priori* knowledge to genetic clusters. We applied an admixture model with default settings (Supplementary File 3).

## 2.5 Phylogenetic and species tree reconstructions

We prepared three data matrices (Table 1), containing all nDNA sequences (all\_data), all coding genotypes (exon\_all\_data), and all coding haplotypes (exonhap\_all\_data). Because data matrices were not 100% complete (see 3.1 Development of nuclear DNA loci) we compiled a second set of matrices that were 100% complete using only the 17 loci that were successfully sequenced in all 29 individuals (complete\_matrix, exon\_complete\_matrix, exonhap\_complete\_matrix). All individual locus alignments were concatenated using a Python script ([https://github.com/srutschmann/python\\_scripts/blob/master/fasta\\_concat.py](https://github.com/srutschmann/python_scripts/blob/master/fasta_concat.py)). We used two partitioning schemes for the phylogenetic analysis. One used the most appropriate substitution model for each locus (“partition\_locus”) according to a Bayesian Information Criterion using the program JMODELTEST v.2.1.5 (Guindon and Gascuel 2003;

Darriba et al., 2012) (Supplementary Table 3). The other used the best-fit partitioning scheme identified by PARTITIONFINDER v.2 (<https://github.com/brettc/partitionfinder>) (“partition\_whole”). The latter was determined using the greedy algorithm (Lanfear et al., 2012), whereby PHYML (Guindon et al., 2010) was used for model evaluation prior to the Bayesian analysis, and RAXML v.8.2.8 (Stamatakis 2014) was used for model evaluation prior to maximum likelihood analysis (see 2.7 Relationship between node support and number of loci).

Bayesian phylogenetic analysis was carried out on each partitioning scheme using MRBAYES v.3.2.3 (Ronquist et al., 2012) using exons, one analysis using all data and one using only the complete matrix (Table 1; exon\_all\_data, exon\_complete\_matrix). We unlinked the nucleotide frequencies, gamma distributions, substitution rates and the proportion of invariant sites across partitions. Each run consisted of two independent analyses of four MCMC chains, each with  $10^7$  generations and 25% burn-in.

Species tree reconstructions were carried out under a multispecies coalescent framework (Drummond and Rambaut 2007; Heled and Drummond 2010) as implemented in the program \*BEAST v.2.1.3 (Bouckaert et al., 2014). All analyses were performed using exons, one analysis using all data and one using only the complete matrix (Table 1; exonhap\_all\_data, exonhap\_complete\_matrix). All individuals were *a priori* assigned to species based on the gmyc and Bayesian clustering analyses (see 2.4 Species assignment and population structure analysis). In the Bayesian clustering analysis, one individual sampled from Russia was considered to be admixed based on Bayesian posterior probability (PP) assignment values > 0.05 for more than one cluster (Supplementary File 3, Supplementary Fig. 1, Supplementary Table 5). This individual was therefore excluded from further analysis. We used a relaxed uncorrelated lognormal clock for gene tree estimation at each locus and a Yule speciation-process prior. Six independent runs of  $8 \times 10^8$  million generations each were conducted. Runs

were combined in LOGCOMBINER v.2.1.3 (Bouckaert et al., 2014), whereby all parameters reached effective sample sizes (ESS) > 600. Maximum clade credibility trees for each species tree were obtained using TREEANNOTATOR v.2.1.3 (Bouckaert et al., 2014).

## **2.6 Ancestral state reconstruction**

Ancestral state reconstruction was used to test the geographical direction of the radiation (i.e. Continental to Island or Island to Continental). Ancestral range patterns of each individual were defined into four geographic areas: (1) a broadly defined Continental referring to the European and North American mainland, (2) Canary Islands, (3) Madeira, and (4) Azores. As input tree, we used the concatenated tree based on the exon\_all\_data inferred with MRBAYES. A chronogram was fit to the tree using the chronos function in the ape v.3.4 (Paradis et al., 2004) package for R. Ancestral states were estimated under an equal-rates (ER) model using the function ace, and the scaled likelihoods of each ancestral state were calculated using the function lik.anc. An MCMC approach was used to sample character histories from their PP distribution generating 1,000 stochastic character maps with the function make.simmap of the phytools v.0.4.98 (Revell 2012) package in R (R Core Team, 2016; Supplementary File 4).

## **2.7 Relationship between node support and number of loci**

To investigate how the number of analyzed loci affected node support values, we performed phylogenetic reconstructions based on multiple subsets of randomly selected loci (Supplementary Table 4). This included twelve subsets for the phylogenetic analysis and four subsets for the multispecies coalescent analysis. For the Bayesian phylogenetic analysis of subsets of loci we used MRBAYES (see 2.5 Phylogenetic and species tree reconstructions), applying both partition schemes. As an alternative to Bayesian PP, we calculated Shimodaira-Hasegawa approximate likelihood ratio test (SH-aLRT; Guindon et al., 2010) supports of

each node using maximum likelihood phylogenies. For this, we first estimated the best tree with RAXML v.8.2.9 applying the “partition\_whole” partitioning scheme (Supplementary Table 4) and rapid bootstrap analysis with 1,000 replicates. Linear regressions were used to model the number of supported nodes for  $PP \geq 0.95$  and  $PP = 1$ , and the SH-aLRT supports  $\geq 0.95$  and SH-aLRT supports = 1 as a function of the number of loci used in the analysis. The Pearson correlation between the number of loci and number of supported nodes was calculated using the stats package for R.

The multispecies coalescent tree reconstructions were performed and summarized as described in section 2.5. We calculated the correlation between the mean number of parsimony-informative sites per locus and the mean node support values of the randomly selected loci and the resulting tree reconstruction using the cor.test function of the stats package for R.

### 3. Results

#### 3.1 Development of nuclear DNA loci

Whole-genome sequencing resulted in 1,109,684 raw reads, including 651,306 reads for the shotgun library and 458,378 reads for the paired-end library, with an average large contig length of 1,187 and 736 base pairs (bp), respectively (BioSample SAMN03202660, BioProject PRJNA268073, Sequence Read Archive SRP050093). All reads were assembled into 68,473 contigs with an N50 of 1,116 bp. The reads of the shotgun library were assembled into 31,827 contigs with an N50 of 1,260 bp. We detected 918 putative orthologous gene sequences for *C. dipterum* from the contigs derived from the shotgun library, 1,298 putative orthologous gene sequences from the contigs of the combined assembly, and 416 for *Baetis* sp. (Supplementary Table 6). From these, we haphazardly selected 65 markers for primer design, approximately 80% of which included orthologous sequences from both taxa. These

were chosen based on the presence of conserved regions and short introns suitable for primer design. Based on preliminary laboratory testing, 59 markers were selected that amplified consistently and had similar annealing temperatures in order to simplify the large number of PCR reactions (Supplementary Table 3).

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

The number of variable sites per locus ranged from six to 65 (mean: 18.95). In the exon\_all\_data matrix, there was one SNP per every 21.62 nucleotides sequenced (i.e. total length per total number of variable sites). The loci included between six and 54 informative sites (mean: 16) and one to 26 ambiguous sites (mean: 8.4). Nucleotide diversity ranged from 0.007 to 0.04 (mean: 0.017), and Tajima's D varied between -0.85 and 1.97 (mean: 0.29) (Supplementary Table 7).

### 3.2 Species assignment and population structure

There were 62 unique *cox1* haplotypes of *Cloeon* and the gmyc model was a significantly better fit to the data than the null model ( $\chi^2 = 31.00$ ,  $p < 0.001$ ). Seven putative species were delineated within *C. dipterum* s.l. (Fig. 2a): One occurred only in Asia (South Korea) while

the remaining six comprised three species with distributions that included the Macaronesian Islands (IS1, IS2, IS3) and three species only occurring on the European and North American continents (CT1, CT2, CT3). The population assignments from the Bayesian clustering analyses of nDNA agreed with the results from the gmyc analysis for the six species of interest (Supplementary File 3, Supplementary Fig. 1, and Supplementary Table 5). No nuclear data were available for the seventh gmyc species from Asia so it was not included in the clustering analysis. Among the six, one widespread species (IS1) was found on all Azorean islands, in Greece and Italy, and in North America; one was found on the Canary Islands and Madeira (IS2), and one was found only on four Canarian islands (IS3). The gmyc model delineated all seven *C. dipterum* species even when using the most conservative estimate (95% confidence interval based on two log likelihood units: 16-19 gmyc species). The two *C. cognatum* specimens from the North American DNA barcoding project (Webb et al., 2012) had *cox1* haplotypes identical to our gmyc species IS1.

### 3.3 Phylogenetic reconstruction

Analyses based on both exon matrices (exon\_all\_data; exon\_complete\_matrix) and both partition schemes recovered the same tree topology with strong node support, resolving each of the three species occurring on Macaronesia (IS1-IS3) as monophyletic and members of a monophyletic 'Island clade' (Fig. 3). The geographically widespread species IS1 was sister taxon to the two others. Species CT2 and CT3 were both monophyletic and sister group to the Island clade (Fig. 3). All individuals in CT1 were monophyletic except for a single individual that was sister taxon to the entire *C. dipterum* s.l. lineage. There were 27 resolved (PP  $\geq$  0.95) nodes in the tree resulting from the large, >75% complete matrix with all available data (exon\_all\_data, 59 loci). The only unresolved node was between the two Azorean individuals (Fig. 3). In contrast, the tree resulting from the smaller matrix with no missing data

(*exon\_complete\_matrix*, 17 loci) contained only 19 resolved nodes, with lack of resolution most pronounced in IS2 (Supplementary Fig. 2).

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

The analysis of subsets of loci showed a positive relationship between the number of loci employed and the number of nodes resolved for phylogenetic analyses (both Bayesian and maximum likelihood) (Fig. 4). The results for the multispecies coalescent were not as clear. Support of key nodes varied widely with the number of loci employed (Fig. 2b, Table 2). The highest overall support came from analysis of 17 and 40 loci, although only the analysis using 20 loci failed to recover either node in the Macaronesian clade and resulted in no resolution other than continental monophyly (Fig. 2b, Table 2). There was a strong negative correlation (Pearson  $R = -0.95$ ,  $p < 0.05$ ) between the mean number of informative sites per locus and mean node support (Supplementary Tables 4, 7).

### 3.4 Ancestral state reconstruction

The ancestral state reconstruction identified four nodes having marginal states with less than 0.9 PP for one character, including the sister relationship between the individual CH010\_SR21B07 and the remaining species, the ancestral node of IS2 + IS3 (Canary Islands and Madeira), and the nodes separating Madeiran from Canarian individuals within IS2 (Fig. 5). The Island clade had a continental origin, further a Canarian origin was estimated for IS2

+ IS3. The clade IS2 was estimated to have an ancestral state of 0.59 for Madeira and 0.4 for the Canary Islands (Supplementary File 4).

## 4. Discussion

### 4.1 Species delineation

The agreement of the multispecies coalescent and mitochondrial *gmyc* approaches for *a priori* species delineation support the use of *cox1* as barcoding gene for the taxa studied (e.g. Lucentini et al., 2011; Pereira-da-Conceicao et al., 2012; Webb et al., 2012; Rutschmann et al., 2014; Gattolliat et al., 2015). Nonetheless, the gene tree based on *cox1* did not resolve the relationships among the six closely related *C. dipterum* s.l., with no support for the sister relationship of IS2 + IS3 or the monophyly of the Continental clade. The distant clustering of one individual (CH010\_SR21B07) in the concatenated tree analyses might be explained by incomplete lineage sorting since the species tree inferences using \*BEAST did result in a clear clustering of CT1 with low frequency of different topology (Fig. 3). Moreover, when incomplete lineage sorting is present, standard methods for estimating species trees, such as concatenation and consensus methods, can be statistically inconsistent (Degnan et al., 2009; Roch and Steel 2014), and produce highly supported but incorrect trees (Kubatko and Degnan 2007). The majority of gene trees could support an incorrect species tree if the phylogeny is in the anomaly zone (Degnan and Rosenberg 2006). Here this does not appear to be the case, because one would expect the concatenation and coalescent approach to support different topologies (Kubatko and Degnan 2007; Liu and Edwards 2009).

### 4.2 Species diversity

The use of nDNA and geographically extensive sampling uncovered a largely underestimated species diversity for *C. dipterum* s.l. species complex, supporting the

existence of six geographically relevant species from our study (with a seventh in Asia). Recent evidence from the study of other mayfly species found fine-scale ecological differences among cryptic species detected with molecular methods (Leys et al., 2016; Macher et al., 2016), lending support to the ecological and evolutionary significance of these and other DNA-based findings. Another widespread species, *Baetis harrisoni*, was also found to consist of several cryptic species (Pereira-da-Conceicao et al., 2012). In light of the unusually broad ecological tolerance (among mayflies) observed for *C. dipterum*, we also conclude that the lineage clearly consists of multiple independent species, as has been recognized by morphological taxonomy for some of the members (e.g. *C. peregrinator*). All of our analyses grouped the two specimens of *C. peregrinator* from Madeira with individuals from several Canary Islands into species IS2. Gattolliat et al., (2008) described *C. peregrinator* as an endemic Madeiran species based on morphological characters and support from mtDNA cytochrome-oxidase *b* sequences. At the time, there were no nDNA sequences of Canarian *C. dipterum* s.l. specimens available. Rutschmann et al., (2014) assigned all Madeiran *Cloeon* individuals to *C. peregrinator* for their mtDNA phylogeny, but the specimens were not included in their gmyc analysis because there were no *coxI* sequences available. Based on our findings here, there is no endemic *Cloeon* species on Madeira.

The focus of our study was Macaronesia and therefore nDNA results are only applicable to these taxa, but the mtDNA gene tree provides evidence for broad cryptic diversity within the subfamily Cloeoninae. *Cloeon simile* included two geographically widespread European gmyc species, and *C. smaeleni* Lestage 1924 was two gmyc species, one with Saudi Arabian and one with Afrotropical distribution. The species *C. praetexum* was clearly distinct from all other examined European specimens, which was surprising because it is thought to belong to *C. simile* s.l.. The two specimens of *C. cognatum*, which is thought to be a junior synonym of *C. dipterum* by some authors, were nested within the IS1 clade. All of the above findings

must be considered preliminary because they are based on mtDNA, although we note that mtDNA and nDNA markers agreed in all of the *Cloeon* species that were directly compared. Further studies on these other *Cloeon* taxa with nuclear markers, using morphological characteristics, and including comparisons with previously described species that are now considered junior synonyms or *species inquirenda* would be a valuable complement to the work presented here.

### 4.3 Evolution, colonization, and diversification

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

Our results suggest a strong effect of different habitat preferences between the two Canarian species, which might affect their colonization success. Although 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. The four islands of Gran Canaria, Tenerife, La Gomera, and La Palma all have permanent natural water sources, and the island of El Hierro has several artificial water habitats due to the mostly temperate climatic conditions. In contrast, there are only a few habitats on Fuerteventura and Lanzarote due to the arid climatic conditions. The effect of habitat use on species richness has been shown for aquatic beetles (Ribera et al., 2003c), with running water bodies generally containing more species than standing ones. This pattern also applies to the Macaronesian mayflies. The genus *Baetis* occurs in running waters and is species-rich, including eight island endemic species on five islands of Madeira and the Canary Islands (Rutschmann et al., 2014). In contrast, the genus *Cloeon* comprises three species, none of which are restricted to a single island. The impact of agriculture and tourism on natural

habitats (Malmqvist et al., 1995; Nilsson et al., 1998) has clearly threatened the occurrence of species living in running water habitats (*Baetis canariensis* and *B. pseudorhodani*, Rutschmann et al., 2014), but it may have had less of an effect on *C. dipterum*. The records of mayflies from El Hierro indicate a recent anthropogenic import of the species, moreover because it is the youngest island of the Canarian archipelago and its remote geographical position.

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

#### **4.4 Number of loci for phylogenetics**

A recent study by O'Neill et al., (2013) examined how multilocus species tree inferences varied with differing numbers of loci. Their analysis based on the 20 and 30 most informative loci (using a parsimony criterion) in their data set resulted in high PPs, whereas node support values were lower and likelihoods failed to converge when loci that were less informative were added to the analysis. They concluded this was the result of the increasing number of parameters while adding loci with decreasing levels of information. Our results are not directly comparable to those of O'Neill et al. (2013) for the species tree reconstruction, because we did not explicitly order loci by parsimony-informative sites in our tests. Nonetheless, we found a strong negative correlation between the mean number of informative sites per locus and mean node support in the coalescent species tree. This suggests that the

number of informative sites was not able to explain variation in support alone, and that multiple characteristics of individual loci play an important role in whether or not analyses achieve convergence and tree resolution. For the species phylogeny, we found a positive linear correlation between number of loci and node support. This was despite the larger number of parameters. In our study we observed that the reduction in node support when using a reduced set of loci (exon\_complete\_matrix vs. exon\_all\_data, see 2.3 PCR amplification, sequence alignment, and sequence heterogeneity) primarily affected the most derived clade (IS2), which highlights the importance of large nDNA marker sets for the reconstruction of shallow phylogenies.

## 5. Conclusion

Our aims were to delineate the species boundaries within the *C. dipterum* species complex and place these lineages within a phylogenetic framework, in order to better understand their evolution on the Macaronesian Islands. Robust phylogenetic reconstruction of such closely related species can be challenging, but is a necessary step in the understanding of evolutionary processes of diversification and adaptation. Most of the widely used nDNA loci (e.g. rRNA) do not exhibit suitable polymorphism, resulting in a large dependence on mtDNA for phylogenetics (Garriek et al., 2015). A distinct advantage of using multiple nDNA loci comes from the advent of multilocus species tree reconstruction methods. These are important tools in the reconstruction of relationships between close relatives, which is often intractable based on single-locus (i.e. mtDNA) data. The difficulty in developing large numbers of nDNA loci remains one of the primary reasons that there are few model systems available for detailed studies of speciation and diversification processes. This is particularly true for freshwater insects, despite their overwhelming contribution to global biodiversity (Dijkstra et al., 2014). Here we developed a large set of nDNA loci using draft whole-genome sequencing. All of the procedures we used have since been incorporated into a single analysis

542 pipeline (Rutschmann et al., 2016). Our results show that even for taxa with very limited  
 543 available genomic resources, it is possible to develop sets of nuclear loci that produce fully  
 544 resolved and supported coalescent-based species trees and species-level phylogenetic trees.  
 545 Using these results, we were able to infer species boundaries within the largely cryptic *C.*  
 546 *dipterum* s.l. species complex and reconstruct the diversification and island colonization  
 547 history of these species with confidence.

## 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., P.M.R., and M.S. collected and identified specimens. H.D., S.S., and R.D. contributed analytical tools. S.R. and M.T.M. performed and interpreted the analyses, and wrote the manuscript. All authors commented and approved the final manuscript.

## Acknowledgements

We are grateful to Katrin Preuß, Susan Mbedi, Berta Ortiz Crespo, and Lydia Wächter for laboratory work; Matthias F. Geiger, Katharina Kurzrock, Konstantinos C. Gritzalis, Andrey Przhiboro, Maria Alp, Vicenc Acuna, Peter Manko, Dávid Murányi, André Wagner, Tomas Ruginis, Luis F. Pires Braz, and Verena Lubini for field work and providing samples; Peter Rutschmann and the HPC Service of ZEDAT, Freie Universität Berlin for access to high-performance computing resources. We are greatly indebted to Marcos Báez for the precious help for the fieldwork on the Canary Islands, the Canarian authorities who provided us with the collection permissions: Servicio Administrativo de Medio Ambiente, Cabildo de Tenerife (reg. number 2014-00200/2014), Ministerio de Medio Ambiente, Parque Nacional de Guajayón, La Gomera (reg. number 106051-17099/2014), Consejería de Medio Ambiente, Cabildo de Gran Canaria (reg. number 5480/2014), Servicio de Medio Ambiente, Cabildo de La Palma (reg. number 2014001631/2014), Ministerio de Medio Ambiente y Medio Rural y Marino, Parque Nacional de la Caldera de Taburiente (reg. number 269046-REUS 52472/2014), Fuerteventura Reserva de la Biosfera, Cabildo de Fuerteventura (reg. number 2348/2014), and the directors of the Parque Natural da Madeira and the Parque Ecológico do Funchal on Madeira for collecting permits. We are very grateful to our research groups, especially to Ignacio Lucas Lledó, Christian Wurzbacher, and Maribet Gamboa, and two anonymous reviewers for their constructive comments on this work. This is publication number 41 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 Swiss National Science Foundation (Early PostDoc.Mobility fellowship P2SKP3\_158698 to S.R.), the Japan Society for the Promotion of Science (Long-Term Research Fellowship L-15543 to M.T.M.), 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.
- Darriba, D., Taboada, G.L., Doallo, R., Posada, D., 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat. Methods 9, 772.
- Degnan, J.H., DeGiorgio, M., Bryant, D., Rosenberg, N.A., 2009. Properties of consensus methods for inferring species trees from gene trees. Syst. Biol. 58, 35-54.
- Degnan, J.H., Rosenberg, N.A., 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2, e68.
- Degnan, J.H., Rosenberg, N.A., 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24, 332-340.
- Degnan, J.H., Rosenberg, N.A., Stadler, T., 2012. A characterization of the set of species trees that produce anomalous ranked gene trees. IEEE/ACM Trans. Comput. Biol. Bioinform. 9, 1558-1568.
- Díaz-Pérez, A.J., Sequeira, M., Santos-Guerra, A., Catalán, P., 2012. Divergence and biogeography of the recently evolved Macaronesian red *Festuca* (Gramineae) species inferred from coalescence-based analyses. Mol. Ecol. 21, 1702-1726.
- Dijkstra, K.D., Monaghan, M.T., Pauls, S.U., 2014. Freshwater biodiversity and aquatic insect diversification. Annu. Rev. Entomol. 59, 143-163.
- Dimitrov, D., Arnedo, M.A., Ribera, C., 2008. Colonization and diversification of the spider genus *Pholcus* Walckenaer, 1805 (Araneae, Pholcidae) in the Macaronesian archipelagos: evidence for long-term occupancy yet rapid recent speciation. Mol. Phylogenet. Evol. 48, 596-614.
- Drotz, M.K., 2003. Speciation and mitochondrial DNA diversification of the diving beetles *Agabus bipustulatus* and *A. wollastoni* (Coleoptera, Dytiscidae) within Macaronesia. Biol. J. Linn. Soc. Lond. 79, 653-666.
- Drummond, A.J., Rambaut, A., 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7, 214.
- Edwards, S.V., 2009. Is a new and general theory of molecular systematics emerging? Evolution 63, 1-19.
- Emerson, B.C., 2002. Evolution on oceanic islands: molecular phylogenetic approaches to understanding pattern and process. Mol. Ecol. 11, 951-966.
- Emerson, B.C., Kolm, N., 2005. Species diversity can drive speciation. Nature 434, 1015-1017.
- Emerson, B.C., Oromí, P., 2005. Diversification of the forest beetle genus *Tarphius* on the Canary Islands, and the evolutionary origins of island endemics. Evolution 59, 586-598.
- Emerson, B.C., Oromí, P., Godfrey, M.H., 2000a. Interpreting colonization of the *Calathus* (Coleoptera: Carabidae) on the Canary Islands and Madeira through the application of the parametric bootstrap. Evolution 54, 2081-2090.
- Emerson, B.C., Oromí, P., Hewitt, G.M., 2000b. Tracking colonization and diversification of insect lineages on islands: mitochondrial DNA phylogeography of *Tarphius canariensis* (Coleoptera: Colydiidae) on the Canary Islands. Proc. Biol. Sci. 267, 2199-2205.

- Falush, D., Stephens, M., Pritchard, J.K., 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164, 1567-1587.
- Faria, C.M.A., Machado, A., Amorim, I.R., Gage, M.J.G., Borges, P.A.V., Emerson, B.C., 2016. Evidence for multiple founding lineages and genetic admixture in the evolution of species within an oceanic island weevil (Coleoptera, Curculionidae) super-radiation. *J. Biogeogr.* 43, 178-191.
- Flot, J.F., 2010. seqphase: a web tool for interconverting phase input/output files and fasta sequence alignments. *Mol. Ecol. Resour.* 10, 162-166.
- Fujisawa, T., Barraclough, T.G., 2013. Delimiting species using single-locus data and the Generalized Mixed Yule Coalescent approach: a revised method and evaluation on simulated data sets. *Syst. Biol.* 62, 707-724.
- Garrick, R.C., Bonatelli, I.A., Hyseni, C., Morales, A., Pelletier, T.A., Perez, M.F., Rice, E., Satler, J.D., Symula, R.E., Thomé, M.T., Carstens, B.C., 2015. The evolution of phylogeographic data sets. *Mol. Ecol.* 24, 1164-1171.
- Garrick, R.C., Sunnucks, P., Dyer, R.J., 2010. Nuclear gene phylogeography using PHASE: dealing with unresolved genotypes, lost alleles, and systematic bias in parameter estimation. *BMC Evol. Biol.* 10, 118.
- Gattolliat, J.-L., Cavallo, E., Vuataz, V., Sartori, M., 2015. DNA barcoding of Corsican mayflies (Ephemeroptera) with implications on biogeography, systematics and biodiversity. *Arthropod Systematics and Phylogeny* 73, 3-18.
- Gattolliat, J.-L., Hughes, S.J., Monaghan, M.T., Sartori, M., 2008. Revision of Madeiran mayflies (Insecta, Ephemeroptera). *Zootaxa* 1957, 52-68.
- Giarla, T.C., Esselstyn, J.A., 2015. The challenges of resolving a rapid, recent radiation: empirical and simulated phylogenomics of philippine shrews. *Syst. Biol.* 64, 727-740.
- Gillespie, R.G., Croom, H.B., Palumbi, S.R., 1994. Multiple origins of a spider radiation in Hawaii. *Proc. Natl. Acad. Sci. U.S.A.* 91, 2290-2294.
- Gillespie, R.G., Roderick, G.K., 2002. Arthropods on islands: colonization, speciation, and conservation. *Annu. Rev. Entomol.* 47, 595-632.
- Gohli, J., Leder, E.H., Garcia-Del-Rey, E., Johannessen, L.E., Johnsen, A., Laskemoen, T., Popp, M., Lifjeld, J.T., 2015. The evolutionary history of Afrocanarian blue tits inferred from genome wide SNPs. *Mol. Ecol.* 24, 180-191.
- Grant, P., Grant, R., 2008. How and why species multiply: the radiation of Darwin's finches. New Jersey, Princeton University Press.
- Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307-321.
- Guindon, S., Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* 52, 696-704.
- Harrigan, R.J., Mazza, M.E., Sorenson, M.D., 2008. Computation vs. cloning: evaluation of two methods for haplotype determination. *Mol. Ecol. Resour.* 8, 1239-1248.
- Heled, J., Drummond, A.J., 2010. Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol.* 27, 570-580.
- Hughes, S.J., Malmqvist, B., 2005. Atlantic Island freshwater ecosystems: challenges and considerations following the EU Water Framework Directive. *Hydrobiologia* 8, 289-297.
- Illera, J.C., Emerson, B.C., Richardson, D.S., 2007. Population history of Berthelot's pipit: colonization, gene flow and morphological divergence in Macaronesia. *Mol. Ecol.* 16, 4599-4612.

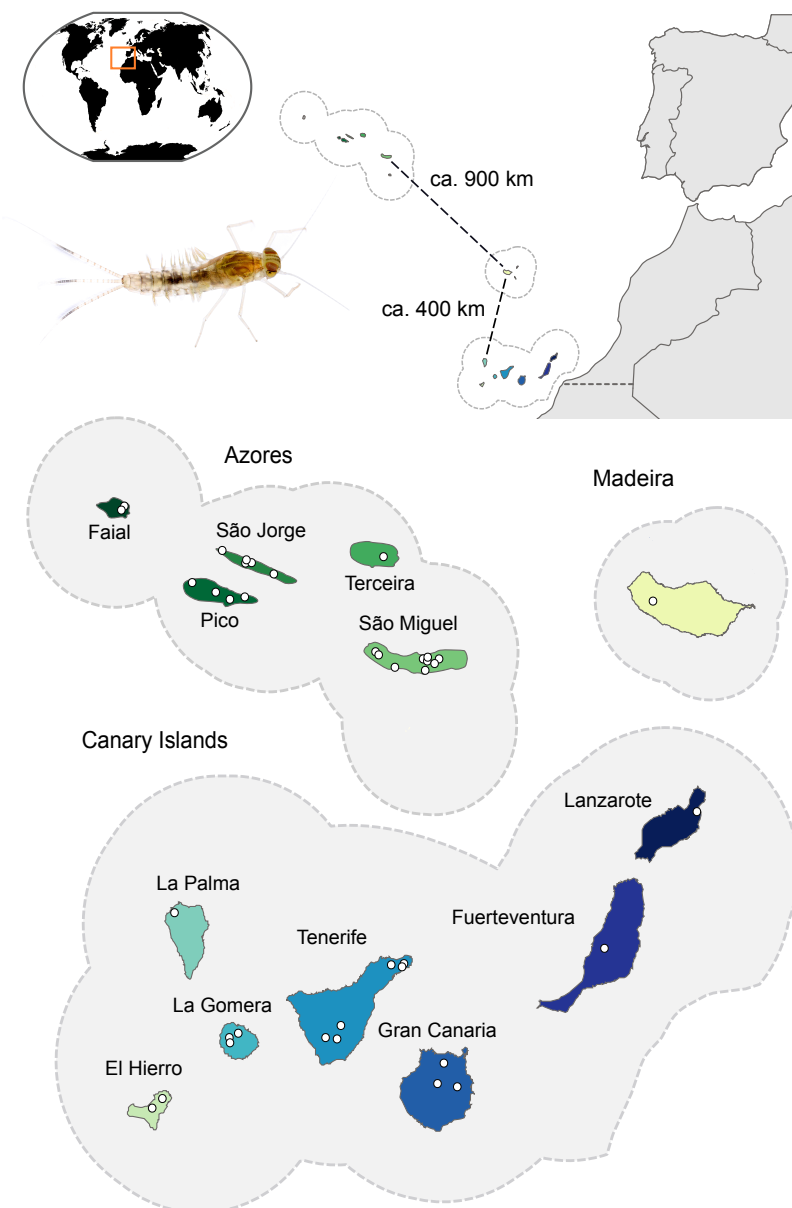
- Jordal, B.H., Hewitt, G.M., 2004. The origin and radiation of Macaronesian beetles breeding in Euphorbia: the relative importance of multiple data partitions and population sampling. Syst. Biol. 53, 711-734.
- Juan, C., Oromí, P., Hewitt, G.M., 1997. Molecular phylogeny of darkling beetles from the Canary Islands: comparison of inter island colonization patterns in two genera. Biochem. Syst. Ecol. 25, 121-130.
- Juan, C., Emerson, B.C., Oromí, P., Hewitt, G.M., 2000. Colonization and diversification: towards a phylogeographic synthesis for the Canary Islands. Trends Ecol. Evol. 15, 104-109.
- Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772-780.
- Knowles, L.L., Kubatko, L.S., 2010. Estimating Species Trees: Practical and Theoretical 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.
- Lanfear, R., Calcott, B., Ho, S.Y., Guindon, S., 2012. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol. Biol. Evol. 29, 1695-1701.
- Leys, M., Keller, I., Räsänen, K., Gattolliat, J.-L., Robinson, C.T., 2016. Distribution and population genetic variation of cryptic species of the Alpine mayfly *Baetis alpinus* (Ephemeroptera: Baetidae) in the Central Alps. BMC Evol. Biol. 16, 77.
- 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.
- 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. Limn. 34, 413-435.
- Nogales, M., Delgado, J.D., Medina, F.M., 1998. Shrikes, lizards and *Lycium intricatum* (Solanaceae) fruits: a case of indirect 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-Conceicao, 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> (accessed 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. 2016. DiscoMark: Nuclear marker discovery from orthologous sequences using draft genome data. Mol. Ecol. Resour. DOI: 10.1111/1755-0998.12576.

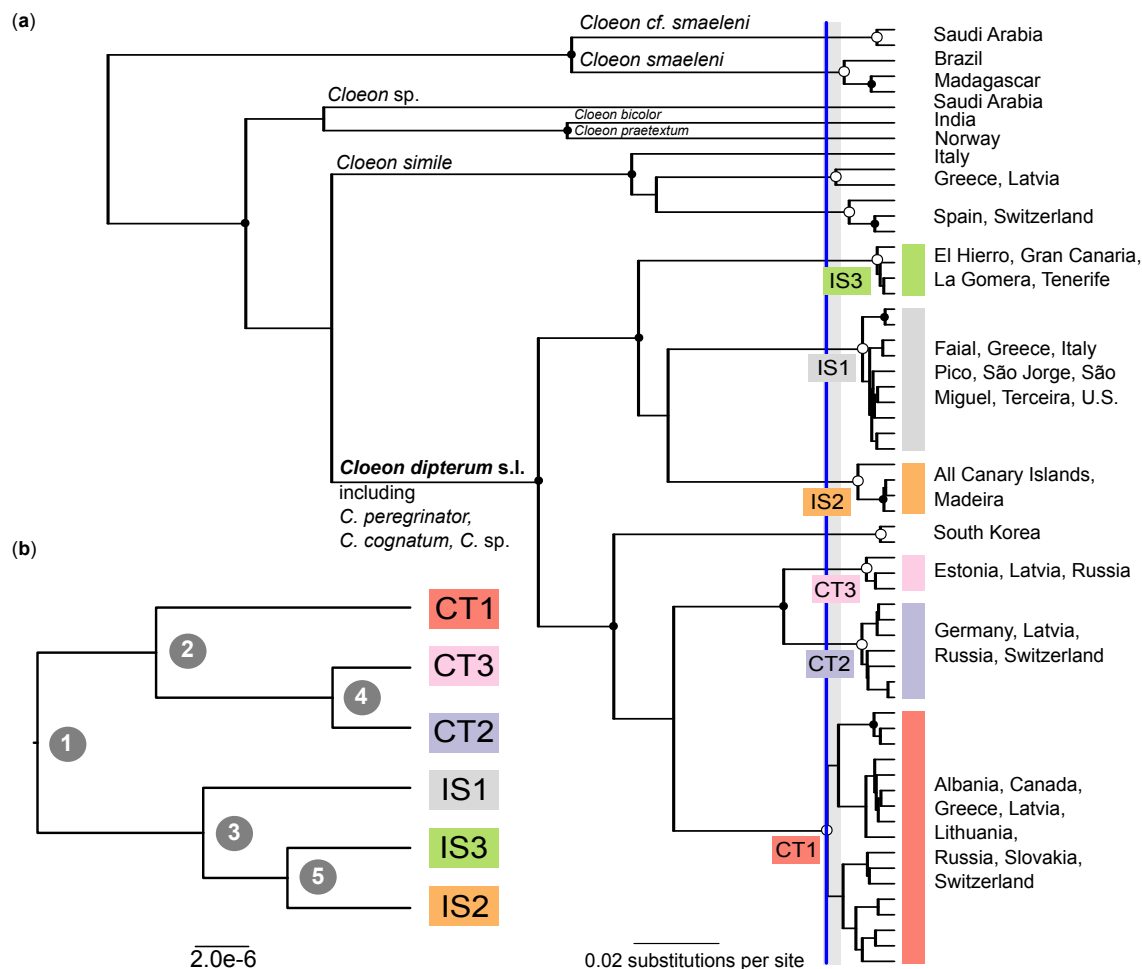
- Rutschmann, S., Gattolliat, J.-L., Hughes, S.J., Sartori, M., Monaghan, M.T. 2014. Evolution and island endemism of morphologically cryptic *Baetis* and *Cloeon* species (Ephemeroptera, Baetidae) on the Canary Islands and Madeira. *Freshwater Biol.* 59, 2516-2527.
- Salles, F.F., Gattolliat, J.-L., 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.
- Simon, S., Strauss, S., von Haeseler, A., Hadrys, H. 2009. A phylogenomic approach to resolve the basal pterygote divergence. *Mol. Biol. Evol.* 26, 2719-2730.
- 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., Pruijscher, 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

840 for *Triturus* newts with the Ion Torrent next-generation sequencing platform. Mol. Ecol.  
 841 Resour. 14, 1080-1089.  
 842 Zhang, D.-X., Hewitt, G.M., 2003. Nuclear DNA analyses in genetic studies of populations:  
 843 practice, problems and prospects. Mol. Ecol. 12, 563-584.  
 844 Zheng, Y., Peng, R., Kuro-o, M., Zeng, X., 2011. Exploring patterns and extent of bias in  
 845 estimating divergence time from mitochondrial DNA sequence data in a particular lineage:  
 846 a case study of salamanders (order Caudata). Mol. Biol. Evol. 28, 2521-2535.

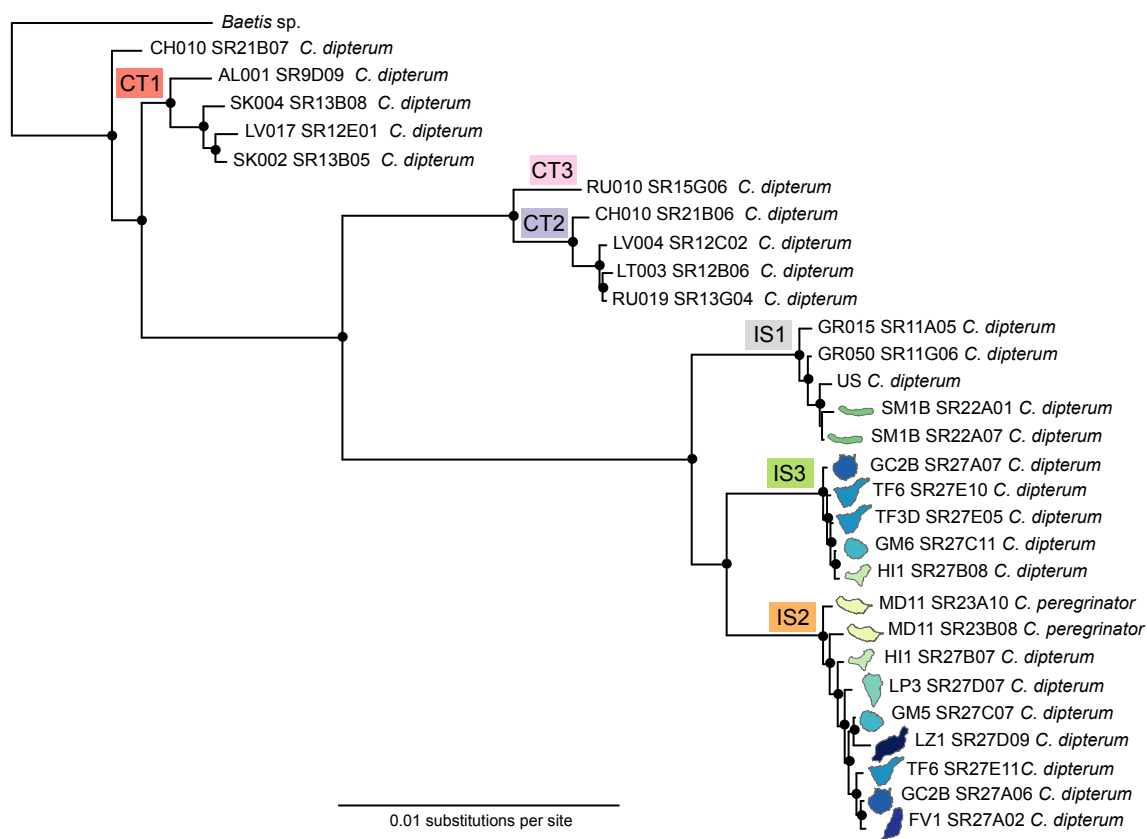
# Figure legends



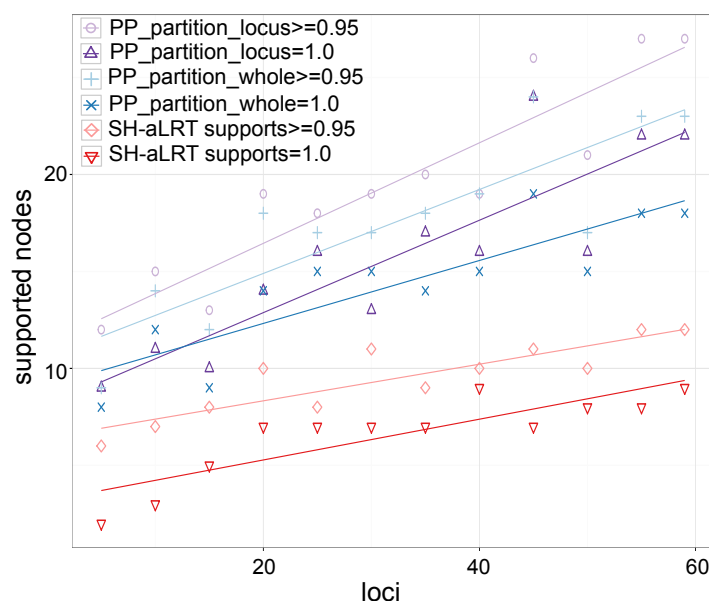
**Fig. 1.** Overview of the sampling localities in the Macaronesian region. The three archipelagos of the Azores, Madeira, and the Canary Islands are shown in detail, whereby the 38 sampling sites are indicated by white dots. For the Azores and Madeira only islands with sampling sites are shown in the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



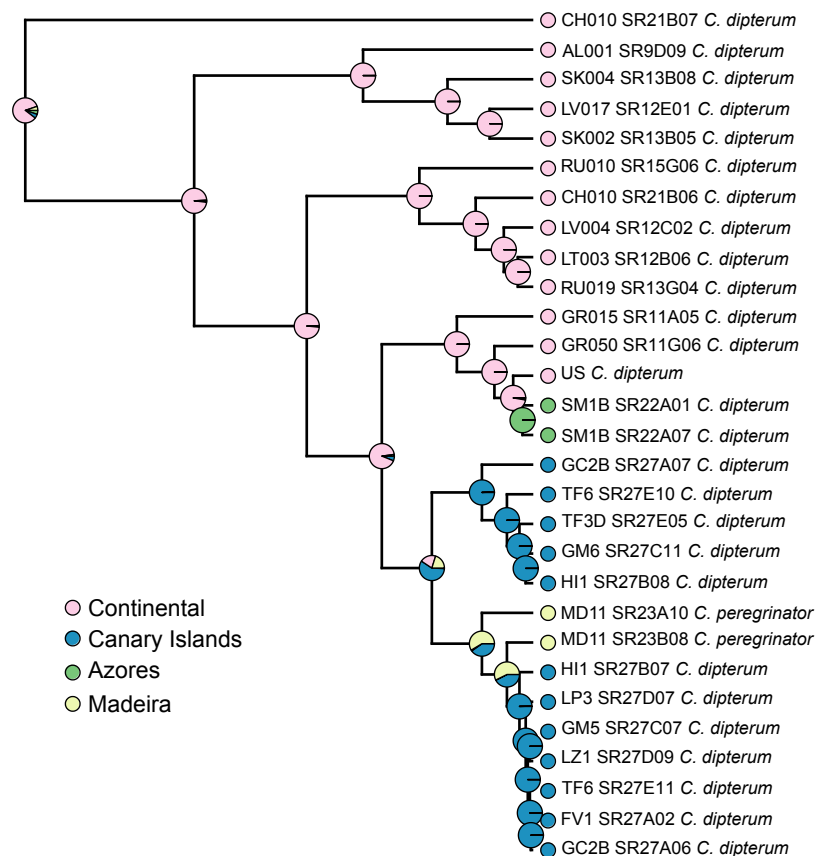
**Fig. 2.** Mitochondrial gene tree and nuclear species tree topology. Species delimitation based on (a) the general mixed Yule-coalescent (gmyc) approach using mitochondrial data (single-locus); and (b) the multispecies coalescent approach using nuclear data (59 loci). An ultrametric mitochondrial *cox1*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



**Fig. 3.** Phylogenetic relationships among *Cloeon dipterum* s.l. (including *C. peregrinator*) based on the Bayesian reconstruction using the concatenated exon\_all\_data matrix of 59 nuclear and one mitochondrial marker. Filled circles indicate nodes with Bayesian posterior probability  $\geq 0.95$ . Species inferred from multilocus coalescent and gmyc analyses are indicated with colored rectangles and alphanumeric codes as in Fig. 2.



**Fig. 4.** Linear relationships between number of supported nodes and number of loci analyzed, including as node support values the Posterior probability (PP) and Shimodaira-Hasegawa approximate likelihood ratio (SH-aLRT) supports for the two partition schemes, using each locus as partition (partition\_locus) and using the whole alignment as one partition (partition\_whole). Purple circles for PP\_partition\_locus  $\geq 0.95$  ( $R^2 = 0.83$ ,  $p < 0.001$ ), purple triangles for PP\_partition\_locus = 1 ( $R^2 = 0.75$ ,  $p < 0.001$ ), blue pluses for PP\_partition\_whole  $\geq 0.95$  ( $R^2 = 0.72$ ,  $p < 0.001$ ), blue crosses for PP\_partition\_whole = 1 ( $R^2 = 0.72$ ,  $p < 0.001$ ), red squares for SH-aLRT supports  $\geq 0.95$  ( $R^2 = 0.74$ ,  $p < 0.001$ ), and red triangles for SH-aLRT supports = 1 ( $R^2 = 0.71$ ,  $p < 0.001$ ). Sets of loci are reported in Supplementary Table 4.



**Fig. 5.** Ancestral state reconstruction of the origin estimated under an equal-rates model based on the concatenated tree. Colors highlight the four defined origins : Continental, including European and North American mainland (pink), Azores (green), Madeira (yellow), and Canary Islands (blue).

## 896 Tables

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

901

Data Matrix	Concatenated Length [bp]	Number of Taxa	Number of Loci	Number of Variable Sites	Description
mitochondrial	658	148	1	240	All <i>cox1</i> 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

**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).

Number of Loci	Node Support					Mean
	All (1)	Continental (2)	Island (3)	CT2+CT3 (4)	IS2+IS3 (5)	
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