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

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

\textsuperscript{a}Leibniz-Institute of Freshwater Ecology and Inland Fisheries (IGB), M"uggelseedamm 301, 12587 Berlin, Germany
\textsuperscript{b}Berlin Center for Genomics in Biodiversity Research, K"onigin-Luise-Straße 6-8, 14195 Berlin, Germany
\textsuperscript{c}Department of Biochemistry, Genetics and Immunology, University of Vigo, 36310 Vigo, Spain
\textsuperscript{d}Sackler Institute for Comparative Genomics, American Museum of Natural History, Central Park West and 79\textsuperscript{th} St., New York, NY 10024, USA
\textsuperscript{e}Biosystematics Group, Wageningen University, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands
\textsuperscript{f}Stroud Water Research Center, Avondale, Pennsylvania 19311, USA
\textsuperscript{g}Musée cantonal de zoologie, Palais de Rumine, Place de la Riponne 6, 1014 Lausanne, Switzerland
\textsuperscript{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
\textsuperscript{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.,
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 disproportionally 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 (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. incriptum* 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* (see also Rutschmann et al., accepted). Libraries were generated from laboratory-reared subimagos of *C. dipterum* specimens (full siblings). DNA was extracted from pooled specimens (5-20) after removing eyes and wings using the Invisorb® Spin Tissue Mini kit (STRATEC, Berlin, Germany). Extracted DNA was precipitated using Isopropanol and pooled in order 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). The fragments were amplified with an emulsion PCR (emPCR Method Manual - Lib-L SV, GS FLX Titanum 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) under the default settings for large datasets. We made two different assemblies, one with the reads from the shotgun library and one with the reads from both shotgun and paired-end libraries. The newly sequenced draft whole genome was combined with 4,197 expressed sequence tag (EST) sequences from *Baetis* sp. (GenBank Acc. no. FN198828–FN203024). *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 since been incorporated into the DISCOMARK pipeline for marker discovery and primer design (Rutschmann at al., accepted; 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 loci. 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 for which 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 previously published sequences from four individuals of 6 nDNA loci (KU971838-KU971840, KU971851, KU971919-KU971921, KU971933, KU972490-KU972492, KU972503, KU972568-KU972570, KU972583, KU972653-KU972654, KU972666, KU973060-KU973062, KU973074).

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 infer 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). 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 for
subsequent sequence analyses. All alignments were re-aligned after phasing with MAFFT.
We excluded introns for the haplotype phasing because the noncoding alignments contained
many gaps and missing data and thus the results of the sequence phasing were not
satisfactory.

For the subsequent analyses, we prepared three alignment sets (Table 1), whereby we used
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 made a second set of matrices that were 100%
complete using only the 17 loci that were sequenced for all 29 individuals (complete_matrix,
exon_complete_matrix, exonhap_complete_matrix). The most appropriate substitution model
for each locus was determined according to Bayesian Information Criterion in the program
JMODELTEST v.2.1 (Guindon and Gascuel 2003; Darriba et al., 2012) (Supplementary Table
3).
To investigate the heterogeneity among the newly developed loci, we reconstructed reticulation-free haplotype genealogies based on Fitch distances (Fitch 1970), using the program FITCHI (Matschiner 2015). We used the exonhap_all_data matrix and calculated a gene tree for each locus using RAxML v.8 (Stamatakis 2014) under the GTRCAT model with 1,000 bootstrap replicates using the rapid bootstrap algorithm. The number of variable sites, informative sites, and Tajima’s D for each locus was assessed using the package DENDROPY (Sukumaran and Holder 2010; 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 gmyc approach (Fujisawa and Barraclough 2013) and a Bayesian clustering algorithm to assign individuals to ‘populations’ (STRUCTURE, Pritchard et al., 2000; Falush et al., 2003). The gmyc approach was carried out using cox1 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 Easton, 1870, and Baetis rhodani (GenBank Acc. no. 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 a 2-partition scheme in which the first two codon positions were modeled with HKY + I and the third codon position with HKY + Γ. For the Bayesian clustering approach we used the exon_all_data matrix. We assumed 1-10 genotypic clusters (K) and ran nine replicate analyses for each K, using \(1 \times 10^6\) 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 reconstruction

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

To investigate how the number of loci analyzed affected node support values, we performed phylogenetic reconstructions based on concatenated sets of varying numbers of randomly selected loci (Supplementary Table 4). The analyses were performed and summarized as above. Linear regressions were used to predict the number of supported nodes for Bayesian posterior probability (PP) $\geq 0.95$ and PP = 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 separately for both PP, using the stats package in R (R Development Core Team, 2016).

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, as above (Table 1;
exonhap_all_data, exonhap_complete_matrix). All individuals were a priori assigned to species based on the gmyc and Bayesian clustering analyses described above. In the Bayesian clustering analysis, one Russian individual was considered to be admixed based on 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. We conducted six independent runs of $8 \times 10^8$ million generations each. 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 trees were obtained using TREEANNOTATOR v.2.1.3 (Bouckaert et al., 2014). As for the concatenated phylogenetics (above), we examined how the number of loci included in the multilocus species tree analysis affected node support by re-running the analysis using subsets of differing numbers of randomly selected loci (Supplementary Table 4).

### 2.5 Ancestral state reconstruction

An ancestral state reconstruction approach was used to test the 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 in 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 in ape v.3.4. A 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 (Supplementary File 4).

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 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). We successfully designed primer pairs for 59 sequence alignments (Supplementary Table 3), mostly based on orthologous sequences from both taxa.

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.
All haplotype genealogies showed clear structuring (Supplementary Fig. 2). For 33 loci, we found haplotypes shared between putative species. The number of variable sites per locus ranged from six to 65 (mean: 18.95). These values were lower than those reported above because ambiguous sites were not considered variable in the haplotype analysis. 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 ($\chi^2 = 31.00, p < 0.001$). There were seven putative species delineated within C. dipterum s.l. (Fig. 2a): One occurred only in Asia (South Korea) while the remaining six included three species with distributions that included the Macaronesian Islands (IS1- IS3) and three species only occurring on the European and North American continents (CT1- CT3). The population assignments from the Bayesian clustering analyses of nDNA (Supplementary File 3, Supplementary Fig. 1, and Supplementary Table 5) agreed completely with the results from the gmyc analysis. Among these six, one widespread species (IS1) was found on all Azorean islands, in Greece and Italy, and in North America, one on the Canary Islands and Madeira (IS2), and one only on four of the Canary Islands (IS3). The model recognized all seven C. dipterum gmyc species even when using the most conservative estimate (95% CI 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) 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. 3a). 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. 3a). 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 ≥ 0.95) nodes in the full matrix (all_data) tree; the only unresolved node was between the two Azorean individuals (Fig. 3a). In contrast, the complete matrix tree contained only 19 resolved nodes, with lack of resolution most pronounced in IS2 (Supplementary Fig. 3).

All species tree phylogenies had identical topologies and these matched the Bayesian phylogenies (Figs. 2b, 3a) in that Island and Continental clade both were 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 ≥ 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).

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

3.4 Ancestral state reconstruction

The ancestral state reconstruction identified four nodes showing marginal states with less than 0.9 Bayesian PP for one character, including sister relationship between 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. 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 Number of loci for phylogenetics

A recent study by O'Neill et al., (2013) examined how multilocus species tree inferences varied with differing number of loci. In their study, analysis based on the 20 and 30 most informative loci (using a parsimony criterion) resulted in high PPs, whereas node support values were lower and likelihoods failed to converge when less informative loci 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 (Pearson R = -0.95) between the mean number of informative sites per locus and mean node support in the coalescent species tree (data not shown). 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 concatenation approach, we found a positive linear correlation between number of loci and node support. This was despite the larger number of parameters. Simulations based on 200 to 300 loci showed that the divergence time estimation using 50 loci are robust (Shen et al., 2016). 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 section 2.3) primarily affected the most derived clade (IS2), which highlights the importance of large nDNA marker sets for the reconstruction of shallow phylogenies.

4.2 Species delineation

The perfect agreement of the Bayesian clustering and mitochondrial gmyc approaches for a priori species delineation support the use of coxl as barcoding gene for the taxa studied (e.g. Lucentini et al., 2011; Pereira-da-Conceicoa et al., 2012; Webb et al., 2012; Rutschmann et al., 2014). 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. 3a). 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). However, here this does not seem to be the case, otherwise one would expect the concatenation and coalescent approach to support different topologies (Kubatko and Degnan 2007; Liu and Edwards 2009). The inferred haplotype networks illustrate the necessity of using several individuals per species. For example, the individuals of IS2 shared several haplotypes with other species, indicating incomplete lineage sorting between the
different species. Originally, it was thought that *BEAST analyses would be quite robust in the presence of gene flow while migration is problematic (Heled et al., 2013). However, Leaché et al., (2014) have shown that gene flow can alter species trees, ranging from decreasing PPs for low gene flow up to altering the species tree topology when high levels of gene flow occur.

4.3 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 another mayfly species found fine-scale ecological differences among cryptic species detected with molecular methods (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-Conceicoa 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 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 *cox1* sequences. 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 Cloeoninace. Although these are single-locus data and must therefore be considered preliminary, we note that the mtDNA and nDNA species delineation results were fully congruent. *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 taxa with additional molecular 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.

### 4.4 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 Canarian islands (IS3), and one species was found on five islands of the Azores, in Italy, in Greece and in North America (IS1). The short branches and occurrence of shared haplotypes 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 ovivipary and their ability to survive in anthropogenic conditions.
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
several artificial standing water habitats exist due to mostly temperate climatic conditions, whereas there are only very 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), whereby running water bodies contain 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 lotic habitats (*Baetis canariensis* and *B. pseudorhodani*, Rutschmann et al., 2014), but it may have had less of an effect on *C. dipterum*. The recent 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 the 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 seems to exhibit long distance trans-oceanic dispersal abilities.
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 readily available nDNA loci (e.g. rRNA) do not exhibit suitable polymorphism, resulting in a large dependence on mtDNA for phylogenetics (Garrick 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 combined with sequences from a published EST library that was constructed using a different subfamily. All of the procedures we used have since been incorporated into a single analysis pipeline (Rutschmann et al., accepted). Our results show that even for taxa with very limited available genomic resources, it is possible to develop sets of nuclear loci that produce fully resolved and supported coalescent-based species trees and single-matrix phylogenetic trees. Using these results, we were able to infer species boundaries within the largely cryptic *C. dipterum* s.l. species complex and reconstruct the diversification and island colonization history of these species with confidence.
Author contributions

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


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.


Fig. 1. Overview of the sampling localities on 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 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 R² for PP ≥ 0.95 were 0.83 (p < 0.001), and 0.75 (p < 0.001) for PP = 1.
Table 1. Overview of seven sequence alignments, including one based on mitochondrial sequences and six based on nuclear sequences. The mitochondrial sequence alignment including only the species group of *Cloeon dipterum* s.l. comprised 130 taxa with 148 variable sites. Matrices containing all taxa and all loci were >75% complete; the matrices containing only loci sequenced for all taxa were 100% complete. Exon matrices refer to exon sequence alignments and the exonhap matrices refer to exon haplotype sequences.

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

<table>
<thead>
<tr>
<th>Number of Loci</th>
<th>All (1)</th>
<th>Continental (2)</th>
<th>Island (3)</th>
<th>CT2+CT3 (4)</th>
<th>IS2+IS3 (5)</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>1.00</td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>20</td>
<td>1.00</td>
<td>1.00</td>
<td>-</td>
<td>1.00</td>
<td>-</td>
<td>0.60</td>
</tr>
<tr>
<td>30</td>
<td>1.00</td>
<td>0.85</td>
<td>0.87</td>
<td>0.87</td>
<td>0.99</td>
<td>0.92</td>
</tr>
<tr>
<td>40</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.67</td>
<td>0.83</td>
<td>0.70</td>
</tr>
<tr>
<td>50</td>
<td>1.00</td>
<td>0.83</td>
<td>0.83</td>
<td>0.83</td>
<td>1.00</td>
<td>0.87</td>
</tr>
<tr>
<td>59</td>
<td>1.00</td>
<td>0.83</td>
<td>0.83</td>
<td>0.83</td>
<td>1.00</td>
<td>0.87</td>
</tr>
</tbody>
</table>