Evolutionary processes in mayflies (Ephemeroptera): genomics approaches to the study of ancient origins and recent diversification

Inaugural-Dissertation

to obtain the academic degree
Doctor rerum naturalium (Dr. rer. nat.)

submitted to the Department of Biology, Chemistry and Pharmacy
of Freie Universität Berlin

by
SEREINA RUTSCHMANN

from Winterthur, Switzerland

Berlin, 2015

The PhD research presented here was conducted from 01.07.2011 to 31.12.2014 as part of the FREDIE project (www.fredie.eu) at the Leibniz-Institute of Freshwater Ecology and Inland Fisheries Berlin (IGB) and the Berlin Center for Genomics in Biodiversity Research (BeGenDiv). The work was partly financed by the SAW/PAKT program of the Leibniz Association (WGL) and by the IGB. Dr. Michael T. Monaghan supervised the work.

1st Reviewer: Dr. Michael T. Monaghan
2nd Reviewer: Prof. Dr. Klement Tockner
Date of Defense: 16.02.2015

ACKNOWLEDGEMENTS

I am extremely thankful for the last three and a half years. I learned a lot doing exciting research, and I met great people within and besides research.

First of all, I am deeply thankful to Michael T. Monaghan for being a fantastic supervisor. I appreciated very much all the fruitful discussions that gave me input for my work, and for all the helpful advises he gave me in order to make my scientific life easier. I am very glad that you not only offered me this PhD position but also and moreover gave me a lot of support and freedom to develop my own research interests. Furthermore, I am very thankful for the generous funding to spend three months at the AMNH, to attend scientific conferences and workshops, and to perform all the sequencing needed fot this PhD research.

I thank Klement Tockner, director of the IGB and second examiner of my thesis, for his always very positive attitude, the support at the Freie Universität Berlin, and for reviewing this work.

I am very thankful to Rob DeSalle and Sabrina Simon for hosting me at the AMNH and their help with the marker design. I really enjoyed my time working together with Sabrina, for discovering together the museum and the city.

I very warmly thank Jean-Luc Gattolliat and Michel Sartori for sharing their knowledge with me and greatly supporting my work, sending me many mayfly specimens, providing me with detailed information, and their expertise.

I benefited much from the weekly Tuesday/Thursday meetings on Evolutionary Biology at the IGB and want to thank all present and past members. I got a lot of helpful comments on my own work and learned a lot. In particular I want to thank Maribet Gamboa, Christian Wurzbacher, Thomas Mehner, Maria Alp, Ignacio Lucas-Lledo, Elizabeth Bourne, and Paul Schmidt Yáñez for many enlightening and also critical discussions on phylogenetics, population genetics, aquatic insects, species tree reconstruction, specific software, and their help during the sampling trips to Spain and Norway.

The great laboratory support from Katrin Preuß, Lydia Wächter, Susan Mbedi, and Berta Ortiz Crespo was substantial for this work. I thank you all for your patience with me in

the laboratory, the moral support during the preparation of the high-throughput sequencing libraries, and your excellent work.

The BeGenDiv was a stimulating environment for me to work in the field of genomics/bioinformatics. I am particularly thankful to Camila Mazzoni, Harald Detering, and Felix Heeger for the bioinformatics support.

Many thanks go to all members of the FREDIE project. In particular, I am thankful to Matthias Geiger for discussion on BEAST analyses and collecting specimens, Katharina Kurzrock, and Jörg Freyhof for their help in the field, and to Fabian Herder as head of the project.

I want to thank Peter Manko for teaching me how to catch most efficiently mayflies and two wonderful field trips to the Carpathians. I am thankful to Marcos Báez for his help to retrieve the sample collecting permits on the Canary Islands and the wonderful day we had in the Barranco del Río.

Thanks to Kirsten Pohlman for coordinating the IGB doctoral program that helped me to structure my PhD and to Martin Allgaier and Christian Wurzbacher for co-supervising this PhD research. I especially appreciated the attendance of the scientific writing course held by Thomas Mehner.

I am grateful to 'my PhD fellows' and all members of the Kinderzimmer, in particular to Marlen Heinz, Francesca Pilotto, Magdalena Czarnecka, and Ann-Christin Honnen for sharing their experiences, for always having an open ear, and the fun moments in our spare time.

I am very thankful to my dear "Berlin buddies" and all my friends for all the great moments we had during the Thursday's mystery movies, the funny evenings in the Belgium beer bar, the beer festival, the carnivals of cultures, "short walks", delicious Pizza & Indian food…and so much more.

More overall, I am deeply indebted to my parents, my brothers, my grandparents, and the wonderful person who came into my life recently. Their incredible support gave me the strength to pursue my interest in science. I am very thankful that you always encouraged me and believed in me especially when I had doubts. Harry, Mum & Paps, Benji & Janine, Andri, Nani & Grospi, Grosmami & Grospapi thank you so much for always being there for me.

PREFACE

This thesis is a cumulative work and consists of three manuscripts within a general context. Thus this work contains four sections: Introduction, Objectives, Manuscripts, and Conclusions. The first three sections are structured into the two thematic subsections RECENT DIVERSIFICATION and ANCIENT ORIGINS.

At the beginning of my thesis, I start by providing an **INTRODUCTION** to the study of evolutionary histories and diversification of mayflies, and a section outlining the **OBJECTIVES** of the thesis.

The **MANUSCRIPTS** section is organized into ***CHAPTERS 1, 2 & 3***. Each of the chapters is an independent manuscript that is either already published (***CHAPTER 1***), or that will be submitted to peer-reviewed journals (***CHAPTERS 2 & 3***). Each chapter forms a stand-alone unit, with its own introduction, methods, results and discussion, and can therefore be read independently from the other chapters.

In the **CONCLUSIONS** section, I combine and summarize the main findings of my thesis and highlight future perspectives for research.

As result of this structure, the content of the sections overlap to some extent with the different chapters. The layout of the published manuscript has been reformatted to fit the consistent layout throughout this thesis. References are provided separately for each section.

The genomic data I generated during my PhD research have been deposited to the National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/), and can be downloaded with the accession numbers given in the chapters. Data included in manuscripts that have not been published yet are listed in this thesis as XXXXXXXX. The bioinformatics program DISCOMARK, which I created to facilitate the development of new genetic markers, can be downloaded from https://github.com/hdetering/discomark. All Python scripts that were written for the automation of bioinformatics processes are available from https://github.com/srutschmann/python_scripts.

Contents

SUMMARY

The field of molecular phylogenetics has benefited greatly from the recent advances of modern sequencing approaches that allow for the generation of large genomics data sets Nonetheless a lack of suitable genetic markers and incomplete taxon sampling remain common problems in studies of evolutionary relatedness. Most phylogenetic studies are based on mitochondrial DNA (mtDNA) because information about the nuclear genome and strategies to develop new genetic markers are often not available. The use of appropriate genetic markers and the inclusion of both a geographically and phylogenetically comprehensive taxon sampling are required for adequately reconstructing evolutionary histories among different taxa. This is particularly true for studies of recent diversification.

Mayflies (Ephemeroptera) are ancient freshwater insects, dating back more than 300 million years, but at the same time have been reported to successfully colonize and diversify on recently formed Atlantic oceanic islands. This combination of ancient origin and recent diversification makes them a fascinating study system for molecular phylogenetics. In the first part of my thesis, I investigated the recent diversification and colonization history of mayflies on 13 Atlantic oceanic islands of the Azores, Madeira, and the Canary Islands. The island fauna provides an ideal setting to understand how speciation and dispersal shape present-day freshwater biodiversity.

A first step in the research was an assessment of the species richness of the island fauna, because current taxonomic estimates are uncertain. Earlier research on mayflies in Europe, Africa, Madagascar, and North America has repeatedly uncovered otherwise cryptic diversity based on analysis of mtDNA. This suggests that past morphological estimates may underestimate species richness, and that a comprehensive understanding of island biodiversity and its evolution requires molecular-based taxonomy. In order to assess the biodiversity and date the origin of the island fauna, I used phylogenetic analyses based on universal mtDNA markers combined with a generalized mixed Yule-coalescent (gmyc) approach. In total, I found twelve island-endemic species within three species groups (*Baetis canariensis* s.l., *B. pseudorhodani* s.l., and *Cloeon dipterum* s.l.) that have diversified within the last 15 million years in parallel throughout the island archipelagos. While intriguing, the results also pointed out the limitations of mtDNA markers for the study of recent diversification events. The study clearly demonstrated a need for the development of new genetic markers that provide increased phylogenetic signal in order to resolve the relationships of closely related species groups.

To investigate relationships among newly diverged species, many polymorphisms are needed, and these should ideally be derived from multiple unlinked markers. Since mayflies are a non-model organism i.e. no reference genome is available, I generated a whole genome draft and used these data to design 59 nuclear DNA (nDNA) markers to establish a basis for inferring the evolutionary history of the *C. dipterum* s.l. species group. Prior to my work, there were only two suitably variable nuclear markers available, namely 28S ribosomal RNA (rRNA) and PEPCK. I applied species tree reconstruction methods using the multispecies coalescent approach, a phylogenetic framework developed within the last five years and suitable for large nDNA data sets. This model was used to overcome both the lack of phylogenetic signal and the potentially conflicting signal derived from gene tree incongruences. Using this approach, I delineated six different *Cloeon* species, three on the islands and three on the European mainland. The phylogeny resolved complex colonization routes on a large geographic scale (Macaronesian islands, the European mainland and North America). The three Macaronesian *Cloeon* species appear to have originated from European source populations and different species co-occur in the same freshwater habitats. The diversification within the *C. dipterum* s.l. species group was mainly promoted by allopatric speciation, whereby strong natural selection on ecological traits i.e. freshwater habitat adaptations and shifts in life history traits are presumed to play a key role. Future research identifying specific ecological, morphological, or behavioral traits, as well as genes that are under natural selection will be needed to understand the mechanistic basis of speciation.

The second part of my thesis focused on evolution over much longer temporal scales, namely ancient origins of the extant winged insects. It remains one of the open questions in the field of insect evolution and systematics, and is thought to act as foundation to understand the evolution of flight as one of the most fascinating evolutionary processes, leading to the development of the most diverse and successful animal group. All winged insects (Pterygota) are placed into one of two groups, based on wing function. The inability to fold back the wings, as seen in the Ephemeroptera and Odonata (dragonflies and damselflies), is considered to be an ancestral condition and these orders are therefore referred to as the Palaeoptera (old wings). In contrast, all other orders are able to fold their wings and as such referred to as Neoptera (new wings). The phylogenetic position of the Palaeoptera within the winged insects is one of the unresolved problems in insect systematics and is thus referred to as the ‘Palaeoptera problem’. Morphological and molecular data have provided support for three competing hypotheses: (1) the Palaeoptera hypothesis, stating the Ephemeroptera + Odonata as sister group to the Neoptera, (2) the basal Ephemeroptera hypothesis (Ephemeroptera + (Odonata + Neoptera)), and (3) the basal Odonata hypothesis (Odonata + (Ephemeroptera + Neoptera)). To date molecular phylogenetic reconstructions have been inferred with a limited number of genes, mostly mitochondrial and ribosomal genes, or a limited number of mayfly taxa (i.e. phylogenomic studies).

To resolve the ‘Palaeoptera problem’, I increased the taxon sampling to a total of 93 insect taxa, including 19 mayflies and I used as marker the protein-coding regions of the mitochondrial genomes (mitogenomes) in order to overcome the highly sensitive sequence alignment step. I applied two different phylogenetic tree reconstruction methods, namely Bayesian inference and maximum-likelihood. I identified taxa with unstable topological positions under the different statistical models, and tested the effects of excluding these taxa on the overall phylogenetic accuracy. First, I sequenced and annotated the mitogenomes of the three mayfly species *Baetis rutilocylindratus*, *Cloeon dipterum*, and *Habrophlebiodes zijinensis*. A comparison among mayfly mitogenomes showed that the gene content and gene orientation was conserved, including 37 protein-coding genes and low AT content. I found that the pruning of identified problematic taxa greatly improved the node support values of the tree reconstruction. Interestingly, also the chosen outgroup was identified as being a problematic taxon. The Bayesian inferences provided support for the basal Ephemeroptera hypothesis, whereas the maximum-likelihood phylogeny supported the basal Odondata hypothesis. The increased number of taxa, the exclusion of problematic taxa and the use of mitogenomes proved to be well suited to reconstruct ancient relationships. The contradicting results of the two phylogenetic methods support the growing evidences that phylogenetic methods based on Bayesian inference might be more appropriate for reconstructing ancient relationships. Thus, the relationships of the Palaeoptera remained unresolved but the results point out the need to investigate the suitability of currently used phylogenetic methods for resolving ancient splits.

Taken together, my thesis presents one of the first genetically comprehensive studies on aquatic insects, combining molecular phylogenetic approaches based on a large set of nDNA markers and mitogenomes. I found that the increase of nDNA markers and the development of bioinformatics approaches for recently evolved species groups and the use of mitogenomes for ancient taxa are extremely important for understanding evolution because of their capacity to reconstruct well supported phylogenetic trees.

ZUSAMMENFASSUNG

Das Feld der molekularen Phylogenetik hat bei der Generierung großer Mengen genomischer Daten stark von den aktuellen Fortschritten moderner Sequenzierungstechnologien profitiert. Dennoch mangelt es oft an geeigneten genetischen Markern und ausreichender Taxon-Abdeckung. Die meisten phylogenetischen Studien basieren auf mitochondrieller DNA (mtDNA), weil genomische Information und Strategien zur Entwicklung neuer genetischer Marker oft nicht verfügbar sind. Die Verwendung angemessener, genetischer Marker und die Einbeziehung sowohl umfassender geografischer und phylogenetischer Taxon-Proben sind Voraussetzungen zur adäquaten Rekonstruktion evolutionärer Entwicklungen unterschiedlicher Abstammungslinien.

Eintagsfliegen (Ephemeroptera) sind Süßwasserinsekten, deren Ursprung über 300 Millionen Jahre zurück liegt, welche sich erfolgreich spezialisieren und atlantische Inseln kolonisieren konnten. Diese Kombination aus ursprünglicher Herkunft und aktueller Diversifikation macht sie zu einem faszinierenden Studiensystem für die molekulare Phylogenetik. Im ersten Teil meiner Arbeit habe ich die aktuelle Diversifikation und Kolonisierungsgeschichte der Eintagsfliegen auf 13 atlantischen Inseln der Azoren, Madeira und der Kanarischen Inseln untersucht. Die Inselfauna bietet ideale Voraussetzungen, um zu verstehen, wie Speziation und Ausbreitung die heutige Süßwasser-Biodiversität geformt haben.

Ein erstes Zwischenziel der Forschungsarbeit war die Erfassung der Artenvielfalt der Inselfauna, denn aktuelle taxonomische Einschätzungen sind unsicher und werden hinterfragt. Frühere Untersuchungen über Eintagsfliegen in Europa, Afrika, Madagaskar und Nordamerika, die auf Analysen mittels mtDNA basieren, haben wiederholt eine andernfalls kryptische Diversität aufgedeckt. Dies suggeriert, dass vorherige morphologische Einschätzungen möglicherweise die Artenvielfalt unterschätzten, und dass ein umfassendes Verständnis von Biodiversität und Evolution auf endemischen Inseln molekularbasierte, taxonomische Untersuchungen erfordern. Um die Biodiversität zu bewerten und die Entstehung der Inselfauna zu datieren, habe ich phylogenetische Analysen durchgeführt, basierend auf universeller mtDNA Markern in Kombination mit einem "generalized mixed Yule-coalescent" (gmyc) Ansatz. Insgesamt fand ich zwölf insel-endemische Spezies in drei Spezies-Gruppen (*Baetis canariensis* s.l., *B. pseudorhodani* s.l. und *Cloeon dipterum* s.l.), die sich innerhalb der letzten 15 Millionen Jahre parallel auf den Insel-Archipelen diversifiziert haben. Obwohl aufschlussreich, unterstreichen die Ergebnisse dennoch die Notwendigkeit der Entwicklung neuer genetischer Marker, die ausreichende, phylogenetische Informationen enthalten, um die Verwandtschaftsverhältnisse der identifizierten, nahverwandten Speziesgruppen zu rekonstruieren.

Um die Verwandtschaft zwischen neu-divergierenden Spezies zu untersuchen, sind viele Polymorphismen nötig, und diese sollten idealerweise von einer Vielzahl unabhängigen Marker abstammen. Da Eintagsfliegen keinen Modellorganismus darstellen, weil kein Referenzgenom existiert, erstellte ich einen Ganzgenom-Draft. Dieses benutzte ich als Basis für 59 nukleäre DNA (nDNA) Marker zur Inferenz der Evolutionsgeschichte der *C. dipterum* s.l. Speziesgruppe). Vor meiner Arbeit gab es lediglich zwei geeignete nDNA Marker: 28S ribosomale RNA (rRNA) und PEPCK. Ich wendete Artbaum-Rekonstruktionsmethoden mit einem "multispecies coalescent"-Modell an, ein Ansatz, der in den letzten 5 Jahren entwickelt wurde und zur Analyse von großen nDNA Daten geeignet ist. Dieses Modell wurde gewählt, um sowohl den Mangel an phylogenetischem Signal zu bewältigen als auch um das widersprüchliche phylogenetische Signal aufgrund von Genbaum-Inkongruenzen zu überwinden. Ich grenzte sechs verschiedene *Cloeon*-Spezies ab, drei auf den Inseln und drei auf dem europäischen Festland. Die Phylogenetik konnte Kolonisationsrouten im großen geographischen Maßstab (Makaronesische Inseln, europäisches Festland und Nordamerika) rekonstruieren. Dabei scheint es, dass die drei makaronesischen *Cloeon* Spezies von europäischen Ursprungspopulationen abstammen, und Speziespaare in den selben Süßwasserhabitaten vorkommen. Die Diversifizierung innerhalb der *C. dipterum* s.l. Speziesgruppe wurde wesentlich durch allopatrische Speziation angetrieben, wobei starke natürliche Selektion ökologischer Merkmale (d.h. Süßwasserhabitat-Anpassung) und Verschiebungen von Lebenszyklus-Merkmalen vermutlich eine Schlüsselrolle spielten. Zukünftige Forschung zur Identifikation spezifischer Gene, die unter natürlicher Selektion stehen, und vergleichende Studien einschließlich morphometrischer und ökologischer Analysen werden nötig sein, um die grundlegende Basis von Speziationsmustern zu verstehen.

Der zweite Teil ist fokussiert auf den historischen Ursprung der heute lebenden geflügelten Insekten als eine der verbleibenden offenen Fragen in der Insektensystematik. Die Beantwortung bietet gleichsam das Fundament zum Verständnis der Evolution des Fluges als einem der faszinierendsten evolutionären Prozesse, der zur Entwicklung einer außerordentlich mannigfaltigen und erfolgreichen Tiergruppe geführt hat. Die geflügelten Insekten (Pterygota) werden in zwei Gruppen eingeteilt, basierend auf der Funktion ihrer Flügel. Die Unfähigkeit, ihre Flügel zurück zu falten, wie bei den Ephemeroptera und Odonata (Libellen) vorkommend, wird als ursprüngliche Eigenschaft betrachtet; sie werden daher als Palaeoptera (Altflügler) bezeichnet, im Gegensatz zu den Neoptera (Neuflügler), die jene Fähigkeit besitzen. Dabei ist die phylogenetische Stellung der Palaeoptera innerhalb der geflügelten Insekten eines der ungeklärten Probleme der Insektensystematik und wird daher als sogenanntes "Palaeoptera-Problem" bezeichnet. Morphologische und molekulare Daten haben Anhaltspunkte für drei konkurrierende Hypothesen geliefert: (1) die "Palaeoptera"-Hypothese, die Ephemeroptera + Odonata als Schwestergruppe zu den Neoptera beschreibt, (2) die "basale Ephemeroptera"-Hypothese (Ephemeroptera + (Odonata + Neoptera)) und (3) die "basale Odonata"-Hypothese (Odonata + (Ephemeroptera + Neoptera)). Bisher wurden molekular-phylogenetische Rekonstruktionen mit einer begrenzten Anzahl an Genen, hauptsächlich mitochondrielle und ribosomale Gene, oder mit einer geringen Anzahl an Taxa (d.h. phylogenomische Studien) durchgeführt. Bisher wurden molekular-phylogenetische Rekonstruktionen mit einer begrenzten Anzahl an Genen, hauptsächlich mitochondrielle und ribosomale Gene, oder mit einer geringen Anzahl an Taxa (d.h. phylogenomische Studien) durchgeführt.

Um das "Palaeoptera-Problem" aufzuklären, intensivierte ich das Taxon-Sampling zu einer Gesamtzahl von 93 Insekten-Taxa, inklusive 19 Eintagsfliegen. Als Marker wählte ich mitochondrielle Genome (Mitogenome), um den kritischen Schritt des Sequenz-Alignment zu bewältigen. Ich wendete zwei verschiedene phylogenetische Baum-Rekonstruktionsmethoden an, und zwar die Bayesian Inference und die Maximum Likelihood Methoden. Ich identifizierte Taxa mit unbeständiger, topologischer Positionierung unter den verschiedenen, statistischen Modellen und untersuchte die Effekte des Entfernens dieser Taxa auf die insgesamte phylogenetische Präzision. Zuerst sequenzierte und annotierte ich die Mitogenome der drei Eintagsfliegen-Spezies *Baetis rutilocylindratus*, *Cloeon dipterum* und *Habrophlebiodes zijinensis*. Ein Vergleich mit bekannten Eintagsfliegen-Mitogenomen zeigte die Konservierung von Genumfang und -

orientierung, mit 37 proteinkodierenden Genen und geringem AT-Gehalt. Ich fand heraus, dass das Entfernen der als problematisch identifizierten Taxa den Knoten-Support der Baum-Rekonstruktion stark verbessert. Interessanterweise wurde auch die gewählte Außengruppe (outgroup) als problematisches Taxon identifiziert. Die Bayesian Inference Methode unterstützte die "basale Ephemeroptera"-Hypothese, wohingegen die Maximum Likelihoood Phylogenie die "basale Odonata"-Hypothese unterstützte. Die höhere Anzahl an Taxa, das Ausschließen problematischer Taxa und die Verwendung von Mitogenomen erwies sich als gut geeignet, um ursprüngliche Verwandtschaftsverhältnisse zu rekonstruieren. Die widersprüchlichen Ergebnisse der beiden phylogenetischen Methoden verstärken weiter die zunehmenden Hinweise, dass auf Bayesian Inference basierende Methoden angemessener sind für die Rekonstruktion historischer Artverwandtschaften. Insgesamt können meine Ergebnisse das "Palaeoptera Problem" nicht lösen. Die Ergebnisse belegen aber das Potential und die Notwendigkeit, heute genutzte, phylogenetische Methoden zur Aufklärung ursprünglicher Aufspaltungen einzusetzen.

Zusammenfassend stellt meine Arbeit eine der ersten, genetisch umfassenden Studien über aquatische Insekten dar, die molekulare, phylogenetische Ansätze basierend auf einer großen Anzahl kombinierter nDNA Marker und Mitogenomen einsetzt. Die Ergebnisse zeigen eindrücklich, dass die Hinzunahme von nDNA Markern und die Entwicklung von bioinformatischen Methoden innerhalb nah verwandter Arten sowie die Verwendung von Mitogenomen für alte Gruppen, aufgrund ihrer Fähigkeit zur Rekonstruierung von phylogenetischen Stammbäumen, sehr wichtig sind, um die Evolution zu verstehen.

1 INTRODUCTION

More than 150 years after the publication of Charles Darwin's '*On the Origin of Species*', understanding the evolutionary processes that lead to the diversity of life remains a fundamental topic of research within the field of biology (Schluter 2000; Coyne and Orr 2004). Speciation is responsible for creating the diversity of life and can in general be driven by geographic isolation, reduction of gene flow, and local adaptation. Dispersal events can act to facilitate pronounced geographical expansion, after which populations may become isolated and allopatric speciation may take place. Dispersal can also inhibit speciation by maintaining gene flow among populations, counteracting any divergence by a constant input of new recruits. 'Intermediate dispersal' combines both the potential to colonize new geographic areas on one side and inhibit too much genetic exchange leading to reduced speciation on the other side (Mayr 1963). Lineages with low or high dispersal abilities have less potential for allopatric speciation. Successful colonization of new habitats and local adaptation are thought to promote evolutionary diversification, which are most prominent on oceanic islands due to the occurrence of relatively few species (Losos and Ricklefs 2009).

Our understanding of how species have diversified has been improved greatly through the application of phylogenetics i.e. the reconstruction of evolutionary histories among species. Several kinds of data are routinely used (e.g., morphological characters, nucleotide sequences (DNA), amino acid sequences) and analytical approaches based on distance methods, parsimony and maximum likelihood, including Bayesian inference have been developed. Many of the concepts developed to understand molecular evolution have also been adopted for other types of inheritable information such as language or cultures. For example, Gray et al. (2009) applied phylogenies based on 400 languages to investigate human prehistory, finding a Taiwan origin for the Austronesian settlers of the Pacific. However, most phylogenetic trees are generated by applying evolutionary models to molecular sequences i.e. deoxyribonucleic acid (DNA), ribonucleic acid (RNA) and amino acid sequences that derive from a common ancestor i.e. are homologous sequences that have arisen via speciation (orthologs). The components of DNA and RNA are referred as the four nucleotides G, C, A and T (U in RNA) while the proteins are composed of 20 different amino acids.

Phylogenetics is a fundamental tool for categorizing biological diversity over time. One of the advantages of phylogenetic trees is that they can be applied to the study of evolutionary processes at very long time scales (e.g., 100s million years) and relatively recent time scales (100s-1000s years). Phylogenetic trees also can be used for species discovery and biodiversity assessment of cryptic species groups i.e. morphologically indistinguishable species that are genetically and often ecologically distinct (Bickford et al. 2007). Species delineation uses phylogenetic trees but alternative statistical methods such as the general mixed Yule-coalescent model (gmyc, Pons et al. 2006; Fontaneto et al. 2007), which was developed for sequence-based delimitation of undescribed species and has been widely applied (e.g., Fujisawa and Barraclough 2013). The assessment of undiscovered genetic diversity plays a key role in the management and preservation of biodiversity (Bickford et al. 2007). Thus far, most studies on DNA-based species delimitation are based on mitochondrial genes (but see Moritz and Cicero 2004; Monaghan et al. 2009; Collins and Cruickshank 2013). Once we have a basic knowledge about the occurring biodiversity, the evolution of specific character states e.g. morphological or ecological traits can be investigated by interpreting its occurrence among a set of taxa along a phylogenetic tree. Phylogenetic reconstructions can calibrate the origin of extant and fossil taxa i.e. clock calibration, and reconstruct ancestral characters/sequences of extinct species based on extent descendants. For example, a very recent study based on molecular phylogenetics and historical records found that the HIV-1 virus in Africa, which is thought to have passed from chimpanzees to humans in Cameroon, remained a regional infection until it arrived in the 1920s in Kinshasa from where it became pandemic (Faria et al. 2014).

Until today, the most commonly used sequences for phylogenetic analyses are mitochondrial gene regions and ribosomal genes. This is due to their conserved flanking regions in which primers can be developed, high abundance in the cell, which makes them easy to amplify PCR, and variability within and among species. This is despite a long list of known limitations of mtDNA markers for inferring phylogenetic trees. Among these, the most important are that the mitogenomes are inherited uniparentally by the females. This means that using mtDNA for phylogenetic reconstructions only reflects the evolutionary history of maternal ancestors, causing significant bias when inferring demographic properties of populations or the evolutionary histories of species (Ballard and Whitlock 2004). As a single marker, mitochondrial genes are also limited in their ability to detect hybridization. Additionally, mitochondrial genes used in phylogenies can, in fact, be nuclear sequences that originally derived from the mitogenome (numt, Lopez et al. 1994). As a result, due to the high similarity of numts with the actual mitochondrial genes, it can happen that numts are unintentionally sequenced and compared with ‘real’ mitochondrial genes, misleading the interpretation of the resulting phylogenetic reconstructions. Despite these limitations, mitogenomes are one of the most appropriate markers for reconstructing evolutionary histories of distantly related lineages due to their conserved structure (e.g., Cameron 2014). In particular, the process of aligning a set of sequences is a hard computational problem for distantly related taxa such as whole orders (e.g. insect tree of life) and fast evolving markers such as ribosomal genes. The conserved nature of mitochondrial genes makes them ideally suited to this task.

In most taxa, a lack of suitable nDNA markers is the reason that mitochondrial genes continue to be so widely applied. The recent advances in novel sequencing approaches used in genomics, i.e. high-throughput sequencing technologies have led to an increased abundance of genomics data. The few available analyses of large nDNA data sets have found that individual genes can have different histories (Jeffroy et al. 2006; Galtier and Daubin 2008) and the histories of individual genes do not necessarily correspond to the history of the organismal lineage (Fitch 1970). Thus, when inferring phylogenetic relationships by the use of nDNA markers, many genes have to be used in order to obtain enough variable sites to resolve fine-scale phylogenetic relationships. Processes that can explain this discordance include hybridization and incomplete lineage sorting (see Maddison 1997; Knowles and Kubatko 2010). Incomplete lineage sorting, also known as deep coalescence, occurs when a gene persists in more than one form (allele) through a speciation event. Some alleles may subsequently be lost through selection while others are maintained and continue to evolve, resulting in a different history than the one of the species (Avise et al. 1983). This disagreement is more likely when the speciation events are close in time and when the effective population size of the ancestral population is large. A key advance has been developed with the multispecies coalescent model where histories of individual genes as reconstructed by phylogenetic gene trees are traced back in time employing a stochastic coalescent process constrained by the history of the species, i.e. the species tree (Rannala and Yang 2003; Heled and Drummond 2010).

THE STUDY SYSTEM - MAYFLIES (EPHEMEROPTERA)

Mayflies are an ecologically and morphologically well-studied insect order with intriguing evolutionary properties. This includes an ancient origin (more than 300 million years ago) and recent diversification on oceanic islands within the last million years. Nevertheless, they are genetically largely unstudied and thus a non-model organism in that no reference genome is available (Sartori 2001; Monaghan et al. 2005; Barber-James et al. 2008)). Previous phylogenetic studies have mostly been limited to mtDNA markers and rRNA with the exclusion of two recent studies where up to three nDNA markers were developed (e.g., Vuataz et al. 2011; Vuataz et al. 2013). Some recent phylogenomic studies i.e. phylogenetic reconstruction based on genome data, where lots of ortholog sequences derived from high-throughput sequencing data were used (e.g., Simon et al. 2009; Simon et al. 2012; Thomas et al. 2013; Misof et al. 2014), have included mayfly taxa, but the number of included species was limited in comparison to the total number of known mayfly species.

The order Ephemeroptera encompasses more than 3,000 described species within 42 families and more than 400 genera, including an almost worldwide distribution except Antarctica (Barber-James et al. 2008). The name Ephemeroptera, coming from Greek ‘ep’ = on and ‘hemera’ = day meaning ‘on one day’, and the German name ‘Eintagsfliege’ = ‘one-day-fly’ refer to their short life span as adults. Mayflies spend most of their lives (three to four weeks to more than two years) as aquatic nymphs, living in all types of freshwaters with most diversity in clear, running waters and few species in standing lakes or ponds (Brittain and Sartori 2003; Bauernfeind and Soldán 2012). They are a major component of the invertebrate drift in streams, an important link in the food chain between primary producers and vertebrate predators, and can be ecologically categorized into collectors and scrapers (Brittain and Sartori 2003). Their preferred occurrence in clear (oxygen-rich) waters makes them an important bioindicator of pollution and environmental change (Brittain and Sartori 2003; Bálint et al. 2011). The nymphs live through a different number of moults, depending on the species and external factors such as temperature, food availability and current velocity (Brittain and Sartori 2003). Unique among flying insects, mayflies undergo a hemimetabolous metamorphosis characterized by the development via an aquatic subimago, which is superficially similar to the adults, possessing microtrichia-covered wings and abdomen, but being sexually immature

(Edmunds and Mccafferty 1988; Barber-James et al. 2008). The adult's life duration depends on the species and varies between a few hours and few weeks (Barber-James et al. 2008). It is focused on reproduction whereby they do not feed, lacking mouthparts and relying on the nutritional buildup from their nymphal stages. Males typically form mating swarms at dawn or dusk. Ovoviviparity i.e. live offsprings (restricted to Baetidae, Gillies 1949; Barber-James et al. 2008) and parthenogenesis i.e. asexual reproduction without fertilization are known for some species (Harker 1997).

1.1 RECENT DIVERSIFICATION

The high dependence on aquatic habitats and their historically assumed reduced dispersal ability due to their short flying period as adults (mostly several days but in extreme cases up to one month) makes mayflies an ideal study system to investigate speciation processes. More specifically, populations on island can be used to test the role of dispersal for generating species diversity of different freshwater habitats on islands. Thereby as a first step it is important to infer the mode of speciation on islands i.e. weather the occurring fauna did arise by geographic isolation (allopatric) or as part of a widely distributed population (sympatric) from the neighboring mainland. In a second step the species occurring in lentic and lotic freshwater habitats can be compared in terms of species diversity and inferred colonization pathways as evidence for dispersal. Previous studies pointed a surprising potential for dispersal, reporting mayfly species on remote islands such as the Azores in the Northern Atlantic Ocean (Brinck and Scherer 1961), La Réunion in the Indian Ocean (Gattolliat 2004), or Vanuatu in the Pacific Ocean (Gattolliat and Staniczek 2011), and repeated trans-oceanic dispersal between Madagascar and continental Africa (Monaghan et al. 2005; Vuataz et al. 2013).

For my PhD research, I investigated the mayfly fauna on three Atlantic oceanic island archipelagos (Azores, Madeira, and Canary Islands). These three archipelagos along with Cape Verde, belong to the Macaronesian region. Throughout this thesis I refer to the Azores, Madeira, and the Canary Islands as Macaronesia or northeastern Macaronesia. The distances of the individual islands to the European and African mainland vary between 110 km (Fuerteventura, Canary Islands to Morocco) and more than 2000 km (Flores, Azores to Portugal). All three archipelagos are characterized by recent volcanic origin, dating back 0.8-21 million years (Carracedo et al. 1998), and the occurrence of relatively few freshwater macroinvertebrate species (Malmqvist et al. 1995; Hughes 2005; Gattolliat et al. 2008; Raposeiro et al. 2012). Permanent running water bodies, needed for the occurrence of most mayfly species (see TABLE I) are rare on the three archipelagos. Only the islands of Madeira, Gran Canaria, La Gomera, La Palma, and Tenerife are characterized by the presence of all-year streams (Stauder 1991; Malmqvist et al. 1993; Malmqvist et al. 1995; Stauder 1995; Nilsson et al. 1998).

While much research has been carried out on island evolution and endemism of terrestrial organisms, comparatively little information exists for aquatic invertebrates (e.g., Stauder 1991; Ribera et al. 2003a; Ribera et al. 2003b; Ribera et al. 2003c; Jordal and Hewitt 2004) and thus the colonization pathways are largely unknown. In contrast, several colonization pathways have been identified for terrestrial taxa on the oceanic islands such as the Macaronesian archipelagos (e.g., Juan et al. 2000; Emerson 2002; Emerson and Kolm 2005), including a single colonization event followed by stepping-stone diversification (e.g., 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 (e.g., Nogales et al. 1998; Ribera et al. 2003b; Díaz-Pérez et al. 2012).

Previous studies on the northeastern Macaronesian mayfly fauna provide evidence for the occurrence of an endemic species-pattern rather than a fauna consisting of widespread species (Müller-Liebenau 1971; Alba-Tercedor et al. 1987; Gattolliat and Sartori 2003; Raposeiro et al. 2012). Based on previous taxonomic work, seven recognized species of Baetidae have been described to occur on the three archipelagos (TABLE I). Historically it was thought that two of the species on the islands were identical to the continental species (Navás 1906; Brinck and Scherer 1961): *Baetis rhodani* (PICTET, 1843) and *Cloeon dipterum* L. 1761. These are two of the most common and abundant mayfly species in the western Palaearctic (Sowa 1975; Gattolliat and Sartori 2008). A number of studies have reported the presence of multiple clades based on mtDNA within *B. rhodani* on the European mainland that may indicate the presence of cryptic species (Williams et al. 2006; Lucentini et al. 2011; Sroka 2012). Müller-Liebenau (1971) demonstrated that the supposed presence of *B. rhodani* in the Canary Islands was erroneous and that two new species, namely *B. canariensis* MÜLLER-LIEBENAU, 1971 and *B. pseudorhodani* MÜLLER-LIEBENAU, 1971 were present. Further, Alba-Tercedor et al. (1987) found that the *Cloeon* specimens on the island of Tenerife showed morphological differences that slightly differentiate from their mainland counterparts. *Cloeon dipterum* was thought to be the only mayfly species occurring on all three archipelagos (Brinck and Scherer 1961; Müller-Liebenau 1971; Alba-Tercedor et al. 1987; Malmqvist et al. 1995; Nilsson et al. 1998; Borges et al. 2010; Raposeiro et al. 2012). However Gattolliat et al. (2008) recognized populations on Madeira to be an endemic species (C. peregrinator GATTOLLIAT & SARTORI, 2008), challenging the validity of the species name on any of the islands.

TABLE I Recognized Baetidae species groups on the Canary Islands, Madeira, and the Azores prior to this doctoral work, including geographical distribution, primary habitat type (lotic = perennial running water; lentic = permanent or temporary standing water), and historical local abundance

Species	Distribution	Habitat	Abundance
BAETIDAE			
Cloeon dipterum (LINNAEUS, 1761)[a]	Azores, Canaries, Palaearctic	Lentic	High
Cloeon peregrinator GATTOLLIAT & SARTORI, 2008[b]	Madeira	Lentic	High
Baetis atlanticus SOLDÁN & GODUNKO, 2006[c]	Madeira	Lotic	High
Baetis enigmaticus GATTOLLIAT & SARTORI, 2008[d]	Madeira	Lotic	Low
Baetis canariensis MÜLLER-LIEBENAU, 1971	Canaries	Lotic	High
Baetis pseudorhodani MÜLLER-LIEBENAU, 1971	Canaries	Lotic	Low
Baetis nigrescens NAVÁS, 1932[e]	Canaries, Iberia, North Africa	Lotic	High

[a]Including *Cloeon cognatum* Stephens, 1835
[b]*Cloeon dipterum* sensu Brinck and Scherer (1961)
[c]*Baetis rhodani* (Pictet, 1843) sensu Brinck and Scherer (1961)
[d]*Baetis pseudorhodani* sensu Stauder 1991, 1995); Hughes et al. (1998)
[e]*Baetis nigrescens* was last recorded by Malmqvist et al. (1995).
Bold highlighted species include cryptic species (i.e. species group, but see text below).

Both the mitochondrial gene analyses and morphological data indicate complex, unresolved species relationships underlying the Macaronesian mayfly diversity that are yet to be explained. Throughout this thesis, I refer by species group to a group of several morphologically very similar species, i.e. cryptic species. Thus, I use for example the term *C. dipterum* s.l. lineage (***CHAPTER 1***) or *C. dipterum* s.l. species group (***CHAPTER 2***) for referring to this group with unclear taxonomic status. The same applies for *B. canariensis*, *B. pseudorhodani* and *B. rhodani*. As species definition I use DNA-taxonomy sensu Vogler and Monaghan (2007). Thus when I use the term species, I refer to populations that I have assigned based on mitochondrial data (***CHAPTER 1***) and nuclear data (***CHAPTER 2***) to distinguish different groups that also apply to the morphological (***CHAPTER 1***) and geographic species criteria sensu (DeSalle et al. 2005) but have not been formally described yet (Gattolliat et al., unpublished).

1.2 ANCIENT ORIGINS

One of the unresolved questions in insect systematics is the relationship of the earliest-diverging winged insects, dating back to fossil records from the Early Carboniferous period (ca. 320 million years ago). A clear understanding of these relationships is crucial for the understanding of the evolution of winged insects (Kingsolver and Koehl 1985). Ephemeroptera are, together with the Odonata (dragonflies and damselflies), considered to be the most ancestral winged insects. In comparison to all other winged insects, these two taxa are unable to fold their wings flat over their abdomen and are thus classified as Palaeoptera (old wings), being sister taxa to all remaining insects (Neoptera, new wings). The monophyly of the winged insects (Pterygota), comprising 98% of all insects (Grimaldi and Engel 2005), is well established by molecular and morphological data (see review by Trautwein et al. 2012), but the basal relationships remain unresolved.

Despite the increasing amount of genomic data and phylogenetic approaches, the basal pterygote divergence is still not resolved and is thus referred to as the 'Palaeoptera problem'. Molecular and morphological data have provided conflicting signals, supporting three main hypotheses: (1) the Palaeoptera hypothesis, classifying the Ephemeroptera + Odonata as sister group to the Neoptera; (2) Metapterygota (Odonata + Neoptera); and Chiastomyaria (Ephemeroptera + Neoptera). All three hypotheses are still considered as valid and have received support from both molecular and morphological data (see review by Trautwein et al. 2012).

Until five years ago, all molecular phylogenetic reconstructions to resolve the basal insect relationships were entirely based on rRNA (12S and 16S (mitochondrial), 18S and 28S (nuclear)), mitochondrial protein-coding genes, and two nuclear protein-coding genes (H3, EF-1α, Hovmöller et al. 2002; Ogden and Whiting 2003; Mallatt and Giribet 2006). Ogden and Whiting (2003) already pointed out that the same genes, namely the ribosomal genes and H3, produce different tree topologies and, depending on the methods of sequence alignment and phylogenetic inference, provide support for all three hypotheses. More recent have used larger sets of nuclear protein-coding genes (i.e. phylogenomic studies) but only a very limited mayfly taxon sampling, including between one (Simon et al. 2009; Simon et al. 2012) to four mayfly species (Misof et al. 2014). Thomas et al. (2013) tested different methodological approaches based on seven 'universal genes' (see above) and up to 35 mayfly taxa. Their result highlighted taxon sampling, including the choice of outgroup, as another critical issue for resolving ancient phylogenetic reconstructions. Thereby, reconstructions based on limited numbers of taxa resulted in inconsistent and generally more unreliable results, being sensitive to taxon sampling and methodological changes (Thomas et al. 2013). On the other side, the monophyletic relationship of the mayflies is well established (see review by Monaghan and Sartori 2009). Apart from the SIphluriscidae, the Baetidae are the most basal member of the Ephemeroptera (Ogden et al. 2009; Lin et al. 2014). The Baetidae, i.e. small minnow mayflies, are divided into the two subfamilies Baetinae and Cloeoninae (Monaghan et al. 2005).

REFERENCES

Alba-Tercedor J, Báez M, Soldán T. 1987. New records of mayflies of the Canary Islands (Insecta, Ephemeroptera). Eos. 63:7-13.

Arnedo MA, Oromi P, De Abreu SM, 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. Systematic Entomology. 33:145-158.

Avise JC, Shapira JF, Daniel SW, Aquadro CF, Lansman RA. 1983. Mitochondrial DNA differentiation during the speciation process in *Peromyscus*. Mol Biol Evol. 1:38-56.

Bálint M, Domisch S, Engelhardt CHM, Haase P, Lehrian S, Sauer J, Theissinger K, Pauls SU, Nowak C. 2011. Cryptic biodiversity loss linked to global climate change. Nature Climate Change. 1:313-318.

Ballard JW, Whitlock MC. 2004. The incomplete natural history of mitochondria. Mol Ecol. 13:729-744.

Barber-James HM, Gattolliat JL, Sartori M, Hubbard MD. 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.

Bickford D, Lohman DJ, Sodhi NS, Ng PK, Meier R, Winker K, Ingram KK, Das I. 2007. Cryptic species as a window on diversity and conservation. Trends Ecol Evol. 22:148-155.

Borges PAV, Costa A, Cunha R, Gabriel R, Gonçalves V, Martins AF, Melo I, Parente M, Raposeiro P, Rodrigues P*, et al.* 2010. A list of the terrestrial and marine biota from the Azores.

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

Brittain JE, Sartori M. 2003. Ephemeroptera (Mayflies). In: Resh VH, Cardé RT editors. Encyclopedia of Insects. Amsterdam, Academic Press.

Cameron SL. 2014. Insect mitochondrial genomics: implications for evolution and phylogeny. Annu Rev Entomol. 59:95-117.

Carracedo JC, Day S, Guillou H, Badiola ER, Canas JA, Torrado FJP. 1998. Hotspot volcanism close to a passive continental margin: The Canary Islands. Geological Magazine. 135:591-604.

Collins RA, Cruickshank RH. 2013. The seven deadly sins of DNA barcoding. Mol Ecol Resour. 13:969-975.

Coyne JA, Orr HA. 2004. Speciation. Massachusetts, Sinauer Associates, Sunderland.

DeSalle R, Egan MG, Siddall M. 2005. The unholy trinity: taxonomy, species delimitation and DNA barcoding. Philos Trans R Soc Lond B Biol Sci. 360:1905-1916.

Díaz-Pérez AJ, 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.

Dimitrov D, Arnedo MA, 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.

Edmunds GF, Mccafferty WP. 1988. The Mayfly Subimago. Annual Review of Entomology. 33:509-529.

Emerson BC. 2002. Evolution on oceanic islands: molecular phylogenetic approaches to understanding pattern and process. Mol Ecol. 11:951-966.

Emerson BC, Kolm N. 2005. Species diversity can drive speciation. Nature. 434:1015-1017.

Emerson BC, Oromi P. 2005. Diversification of the forest beetle genus *Tarphius* on the Canary Islands, and the evolutionary origins of island endemics. Evolution. 59:586-598.

Faria NR, Rambaut A, Suchard MA, Baele G, Bedford T, Ward MJ, Tatem AJ, Sousa JD, Arinaminpathy N, Pepin J*, et al.* 2014. HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations. Science. 346:56-61.

Fitch WM. 1970. Distinguishing homologous from analogous proteins. Syst Zool. 19:99-113.

Fontaneto D, Herniou EA, Boschetti C, Caprioli M, Melone G, Ricci C, Barraclough TG. 2007. Independently evolving species in asexual bdelloid rotifers. PLoS Biol. 5:e87.

Fujisawa T, Barraclough TG. 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.

Galtier N, Daubin V. 2008. Dealing with incongruence in phylogenomic analyses. Philos Trans R Soc Lond B Biol Sci. 363:4023-4029.

Gattolliat J-L, Sartori M. 2003. An overview of the Baetidae of Madagascar. In: Gaino E editor. Research update on Ephemeroptera and Plecoptera, University of Perugia.

Gattolliat JL. 2004. First reports of the genus *Nigrobaetis* Novikova & Kluge (Ephemeroptera : Baetidae) from Madagascar and La Reunion with observations on afrotropical biogeography. Revue Suisse de Zoologie. 111:657-669.

Gattolliat JL, Hughes SJ, Monaghan MT, Sartori M. 2008. Revision of Madeiran mayflies (Insecta, Ephemeroptera). ZOOTAXA. 52-68.

Gattolliat JL, Sartori M. 2008. What is *Baetis rhodani* (Pictet, 1843) (Insecta, Ephemeroptera, Baetidae)? Designation of a neotype and redescription of the species from its original area. ZOOTAXA. 69-80.

Gattolliat JL, Staniczek A. 2011. New larvae of Baetidae (Insecta: Ephemeroptera) from Espiritu Santo, Vanuatu. Stuttgarter Beiträge zur Naturkunde A, Neue Serie. 4:75-82.

Gillies MT. 1949. Notes on some Ephemeroptera Baetidae from India and South-East Asia. Trans R Entomol Soc Lond. 161-177.

Gray RD, Drummond AJ, Greenhill SJ. 2009. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. Science. 323:479-483.

Grimaldi DA, Engel MS. 2005. Evolution of the insects. New York, Cambridge University Press.

Hovmöller R, Pape T, Källersjö M. 2002. The Palaeoptera Problem: Basal Pterygote Phylogeny Inferred from 18S and 28S rDNA Sequences. Cladistics. 18:313-323.

Hughes SJ. 2005. Application of the water framework directive to Macaronesian freshwater systems. Biology and Environment-Proceedings of the Royal Irish Academy. 105B:185-193.

Illera JC, Emerson BC, Richardson DS. 2007. Population history of Berthelot's pipit: colonization, gene flow and morphological divergence in Macaronesia. Mol Ecol. 16:4599-4612.

Jeffroy O, Brinkmann H, Delsuc F, Philippe H. 2006. Phylogenomics: the beginning of incongruence? Trends Genet. 22:225-231.

Jordal BH, Hewitt GM. 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, Emerson BC, Oromí P, Hewitt GM. 2000. Colonization and diversification: towards a phylogeographic synthesis for the Canary Islands. Trends Ecol Evol. 15:104-109.

Juan C, Oromí P, Hewitt GM. 1997. Molecular phylogeny of darkling beetles from the Canary Islands: comparison of inter island colonization patterns in two genera. Biochemical Systematics and Ecology. 25:121-130.

Kingsolver JG, Koehl MAR. 1985. Aerodynamics, Thermoregulation, and the Evolution of Insect Wings - Differential Scaling and Evolutionary Change. Evolution. 39:488-504.

Knowles LL, Kubatko LS. 2010. Estimating species trees: an introduction to concepts and models. In: Knowles LL, Kubatko LS editors. Estimating Species Trees: Practical and Theoretical Aspects. New York, Wiley-Blackwell.

Li D, Qin J-C, Zhou C-F. 2014. The phylogeny of Ephemeroptera in Pterygota revealed by the mitochondrial genome of *Siphluriscus chinensis* (Hexapoda: Insecta). Gene. 545:132-140.

Lopez JV, Yuhki N, Masuda R, Modi W, O'Brien SJ. 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 JB, Ricklefs RE. 2009. Adaptation and diversification on islands. Nature. 457:830-836.

Lucentini L, Rebora M, Puletti ME, 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.

Maddison WP. 1997. Gene trees in species trees. Syst Biol. 46:523-536.

Mallatt J, Giribet G. 2006. Further use of nearly complete 28S and 18S rRNA genes to classify Ecdysozoa: 37 more arthropods and a kinorhynch. Mol Phylogenet Evol. 40:772-794.

Malmqvist B, Nilsson AN, Báez M. 1995. Tenerife's freshwater macroinvertebrates: status and threats (Canary Islands, Spain). Aquatic Conservation-Marine and Freshwater Ecosystems. 5:1-24.

Malmqvist B, Nilsson AN, Báez M, Armitage PD, Blackburn J. 1993. Stream macroinvertebrate communities in the island of Tenerife. Archiv für Hydrobiologie. 128:209-235.

Mayr E. 1963. Animal Species and Evolution. Cambridge, Harvard University Press.

Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J, Flouri T, Beutel RG, *et al.* 2014. Phylogenomics resolves the timing and pattern of insect evolution. Science. 346:763-767.

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

Monaghan MT, Sartori M. 2009. Genetic contributions to the study of taxonomy, ecology, and evolution of mayflies (Ephemeroptera): review and future perspectives. Aquatic Insects. 31:19-39.

Monaghan MT, Wild R, Elliot M, Fujisawa T, Balke M, Inward DJ, Lees DC, Ranaivosolo R, Eggleton P, Barraclough TG*, et al.* 2009. Accelerated species inventory on Madagascar using coalescent-based models of species delineation. Syst Biol. 58:298-311.

Moritz C, Cicero C. 2004. DNA barcoding: promise and pitfalls. PLoS Biol. 2:e354.

Müller-Liebenau I. 1971. Ephemeroptera (Insecta) von den Kanarischen Inseln. Gewässer und Abwässer. 50/51:7-40.

Navás SJ. 1906. Catalogo descriptivo de los Insectos Neuropteros de las Islas Canarias. Revista de la Real Academia de Ciencias Exactas, Fısicas y Naturales. 4:1-24.

Nilsson AN, Malmqvist B, Báez M, Blackburn JH, Armitage PD. 1998. Stream insects and gastropods in the island of Gran Canaria (Spain). Annales De Limnologie-International Journal of Limnology. 34:413-435.

Nogales M, López M, Jiménez-Asensio J, Larruga JM, Hernández M, Gonzalez P. 1998. Evolution and biogeography of the genus *Tarentola* (Sauria : Gekkonidae) in the Canary Islands, inferred from mitochondrial DNA sequences. Journal of Evolutionary Biology. 11:481-494.

Ogden TH, Whiting MF. 2003. The problem with "the Paleoptera Problem:" sense and sensitivity. Cladistics. 19:432-442.

Ogden TH, Gattolliat JL, Sartori M, Staniczek AH, Soldán T, Whiting MF. 2009. Towards a new paradigm in mayfly phylogeny (Ephemeroptera): combined analysis of morphological and molecular data. Systematic Entomology. 34:616-634.

Pons J, Barraclough TG, Gomez-Zurita J, Cardoso A, Duran DP, Hazell S, Kamoun S, Sumlin WD, Vogler AP. 2006. Sequence-based species delimitation for the DNA taxonomy of undescribed insects. Syst Biol. 55:595-609.

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 PM, Cruz AM, Hughes SJ, Costa AC. 2012. Azorean freshwater invertebrates: Status, threats and biogeographic notes. Limnetica. 31:13-22.

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

Ribera I, Bilton DT, Vogler AP. 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 GN, Vogler AP. 2003c. Does habitat use explain large scale species richness patterns of aquatic beetles in Europe? Ecography. 26:145-152.

Sartori M. 2001. Current knowledge of mayfly research in Europe (Ephemeroptera). In: Dominguez E editor. Trends in Research in Ephemeroptera and Plecoptera Kluwer Academic/Plenum Publishers.

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

Simon S, Narechania A, Desalle R, Hadrys H. 2012. Insect phylogenomics: exploring the source of incongruence using new transcriptomic data. Genome Biol Evol. 4:1295-1309.

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.

Sowa R. 1975. What is *Cloeon dipterum* (Linneaeus, 1716)? The nomenclatural and morphological analysis of a group of the European species of *Cloeon* Leach (Ephemerida: Baetidae). Ent scand. 6:215-223.

Sroka P. 2012. Systematics and phylogeny of the West Palaearctic representatives of subfamily Baetinae (Insecta: Ephemeroptera): combined analysis of mitochondrial DNA sequences and morphology. Aquatic Insects. 34:23-53.

Stauder A. 1991. Water fauna of a Madeiran stream with notes on the zoogeography of the Macaronesian islands. Boletim do Museu Municipal do Funchal. 43:243-299.

Stauder A. 1995. Survey of the Madeiran limnological fauna and their zoogeogrpahical distribution. Boletim do Museu Municipal do Funchal. Sup. no. 4:715-723.

Thomas JA, Trueman JW, Rambaut A, Welch JJ. 2013. Relaxed phylogenetics and the palaeoptera problem: resolving deep ancestral splits in the insect phylogeny. Syst Biol. 62:285-297.

Trautwein MD, Wiegmann BM, Beutel R, Kjer KM, Yeates DK. 2012. Advances in insect phylogeny at the dawn of the postgenomic era. Annu Rev Entomol. 57:449-468.

Vogler AP, Monaghan MT. 2007. Recent advances in DNA taxonomy. Journal of Zoological Systematics and Evolutionary Research. 45:1-10.

Vuataz L, Sartori M, Gattolliat JL, Monaghan MT. 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.

Vuataz L, Sartori M, Wagner A, Monaghan MT. 2011. Toward a DNA taxonomy of Alpine *Rhithrogena* (Ephemeroptera: Heptageniidae) using a mixed Yule-coalescent analysis of mitochondrial and nuclear DNA. PLoS ONE. 6:e19728.

Williams HC, Ormerod SJ, Bruford MW. 2006. Molecular systematics and phylogeography of the cryptic species complex *Baetis rhodani* (Ephemeroptera, Baetidae). Mol Phylogenet Evol. 40:370-382.

2 Objectives

The aims of my thesis were to reconstruct the colonization of oceanic islands by mayflies (Recent Diversification, ***Chapters 1 & 2***) and to resolve the phylogenetic relationship of the oldest extant winged insects (Ancient Origins, ***Chapter 3***).

My specific research questions where:

- ***Chapter 1:*** *Does the species diversity of mayflies correlate with the habitat type i.e. lentic vs. lotic freshwater habitats?*
- ***Chapter 2:*** *What is the role of dispersal for generating species diversity?*
- ***Chapter 3:*** *Can the use of an increased mitochondrial genome data set and the exclusion of problematic taxa resolve the 'Palaeoptera problem'?*

My PhD research dealt with the entire evolutionary spectrum, including closely related species (***Chapters 1 & 2***) and divergent taxa i.e. insect orders (***Chapter 3***) and thus faced different challenges of molecular phylogenetics. From the methodological point of view, the main problems are the development of suitable genetic markers that containing enough phylogenetic signal to resolve the phylogenetic relationships on different evolutionary time-scales (e.g., a large set of nDNA markers, ***Chapter 2***). Furthermore, markers need to be present in a wide range of taxa, comprising for fine-scale phylogenetic reconstructions all species (i.e. geographic sampling, ***Chapter 1***) and for large-scale phylogenetic reconstructions several species per order (i.e. phylogenetic sampling, ***Chapter 3***). Thus as foundation for my PhD research, I generated a whole genome draft derived from high-throughput sequencing libraries of *C. dipterum* s.l. to (1) develop a large set of nDNA markers (59) for inferring species trees under the multispecies coalescent model (***Chapter 2***) and (2) extract the mitogenome for reconstructing the basal insect relationships (***Chapter 3***).

2.1 Recent Diversification

The first part of this subsection focused on quantifying the mayfly species diversity on the Canary Islands and Madeira (***Chapter 1***). In the second part, I extended the geographic sampling of the previously identified *C. dipterum* s.l. species from five up to 13 investigated islands and developed a set of 59 nDNA markers (for comparison: the number of previously used nDNA markers was two) in order to obtain a better understanding of how the group originated (e.g., number of colonization events) and diversified.

CHAPTER 1: *Which Baetidae species occur on the Canary Islands and Madeira and to what extent are they island-endemics, archipelago-endemics, or widespread continental species?*

As first step, I investigated the **biodiversity and origin of small minnow mayflies on the Canary Islands and Madeira**. Freshwater habitats on all islands are sparse and very valuable since they are highly threatened by agriculture and tourism, leading in the most extreme case to the absence of permanent freshwater habitats. Little is known about the freshwater diversity but cryptic species on the European mainland and previous taxonomic work based on morphological characters evidenced a diverse island fauna rather than a mayfly fauna consisting of geographically widespread species. To get a better knowledge on the species diversity, I applied universal mtDNA markers combined with a generalized mixed Yule-coalescent (gmyc) model analysis to delineate putative species within the morphologically cryptic species groups *Baetis* (*Rhodobaetis*) and *C. dipterum* s.l.. I further used a three-gene mitochondrial data set to infer the phylogenetic relationships and calibrate a molecular clock to date the colonization history. My predictions were that we would find several genetically distinct, morphologically cryptic species, being closely related with their counterparts on the European and North African mainland. Further, I assumed that mayflies recently colonized the Canary Islands and Madeira as result of several colonization processes. I assumed that there would be higher species diversity in running water than in standing waters. Moreover, allopatric speciation might be even more pronounced in species groups occurring in lotic waters and lead to higher genotypic and phenotypic diversification, resulting an island-endemic species pattern. While the results were intriguing for the species delimitation, they also pointed out the urgent need to develop a large set of nDNA markers and thus led immediately to ***CHAPTER 2.***

CHAPTER 2: *What is the evolutionary history of the species group C. dipterum s.l. on the Macaronesian archipelagos? Does the use of many nuclear markers provide better resolution or contradict previous mtDNA results?*

This chapter focuses on the full **reconstruction of the island colonization pathways of the *C. dipterum* s.l. species group on Macaronesia**. ***CHAPTER 1*** was a study of the whole Baetidae fauna of the islands, which included members of both subfamilies. An initial aim of the research in ***CHAPTER 2*** was to develop nDNA markers for both subfamilies using a high-throughput sequencing-derived draft genome. However, the *Baetis* library was of low coverage and the two genera were phylogenetically too distinct to find enough conserved gene regions using only *Cloeon*. Thus, I developed a large number of 59 new nDNA markers suitable for phylogenetic reconstruction of the *C. dipterum* s.l. species group, using a ***de novo* assembled draft genome**. The variety of steps (i.e. clustering of ortholog sequences, identification of conserved regions, primer design) used in the 'marker discovery approach' was implemented into a single bioinformatics program I developed for this thesis named DISCOMARK. I analyzed the gene sequences by combining species tree methods under the multispecies coalescent model, and concatenation approaches on the basis of Bayesian inference and maximum-likelihood methods. ***CHAPTER 1*** showed that the species group of *C. dipterum* s.l. has recently colonized the Macaronesian islands and thus it was highly likely that they still are in the process of speciation, showing incomplete lineage sorting. Species trees under the multispecies coalescent model can account for incomplete lineage sorting. With the concatenation approach, different genes might result in different tree topologies, but due to the use of 59 nDNA markers, the tree reconstructed based on the whole matrix should reflect an 'averaged' phylogenetic clustering. A distinct advantage of the concatenation approach is that it reveals the clustering of individual specimens rather than a priori species to which the analysis must be constrained using the multispecies coalescent approach. I aimed to design 50 nDNA markers based on the estimated numbers of polymorphisms and sequence alignment length that would be needed for the phylogenetic resolution on species-level. In 2011, Alföldi et al. published a species-level phylogeny of the *Anolis* lizards, using 46 markers with a total of 20,000 base pairs sequence alignment in the journal Nature. My personal motivation was to obtain a higher number of nDNA markers for *Cloeon*. The 'concatenated tree' provides the information on individual level and thus the dispersal within one species group whereas the species tree reconstruction confirms the relationships between the different species. The individual gene trees resulting from the multispecies coalescent approach can be inspected to identify markers that widely differ from the rest and thus are not suitable for phylogenetic reconstructions. In total, I developed 59 nDNA markers to gain enough phylogenetic signal for all phylogenetic reconstructions and 29 individuals, including for each previously identified species (***CHAPTER 1***) at least five representatives from different islands to reconstruct the evolutionary history via concatenation and to follow the general guidelines for ideal species tree reconstructions sensu (Heled and Drummond 2010). I hypothesized that (1) a large number of markers would resolve the fine-scale phylogenetic relationships of closely related species and that (2) their use would uncover different colonization routes to the oceanic islands of the Azores, Madeira and Canaries. I predicted that the use of a large set of nDNA markers would confirm the existence of several species and identify their source population on the European mainland.

2.2 ANCIENT ORIGINS

CHAPTER 3: *What is the phylogenetic position of the mayflies within the basal insects?*

This chapter investigated the **origin of the oldest extant winged insects** (Palaeoptera: mayflies, dragonflies and damselflies) and the **structure of the mayfly mitogenome.** For this purpose, I first extracted the mitogenome of *C. dipterum* from the whole genome draft, annotated the sequences, and compared it with 19 other mayfly mitogenomes. In the second step, I used the protein-coding genes of the mayfly species together with 74 other insect mitogenomes to infer the basal insect phylogenetic relationship. Since ancient phylogenetic relationships have shown to be highly sensitive to taxon sampling and phylogenetic methods (i.e. sequence alignment and phylogenetic method), I applied the following four criteria; I (1) increased the taxon sampling in mayflies to 19 taxa, (2) used the amino acid sequences of the mitochondrial protein-coding genes to minimize alignment problems, (3) used the two most widely used phylogenetic methods (i.e. Bayesian inference and maximum-likelihood), and (4) tested the increase of phylogenetic accuracy by removing 'problematic taxa' i.e. so-called rogue taxa. Rogue taxa refers to taxa with uncertain phylogenetic relationship and thus when reconstructing large sets of phylogenetic trees (e.g., Bayesian Markov Chain Monte Carlo sampling or Bootstrapping) these taxa vary in their inferred positions among the trees and thus may lead to lower overall performance (i.e. phylogenetic accuracy). The identification of rogue taxa is a relatively new approach (ROGUENAROK, Aberer et al. 2013) and not yet been applied analyses of the 'Palaeoptera problem'. Moreover, I assumed that specific taxa would lead to the decrease of the phylogenetic reconstruction and the identification of rogue taxa is a standardized approach to exclude taxa. In comparison, previous studies have often been based on subjective exclusion of taxa. I assumed that the mayfly mitogenomes would be conserved in terms of gene content and gene orientation. As evidenced by many previous studies, I expected that the divergences of the most basal winged insects would be challenging to resolve but that increased sampling on the Ephemeroptera and the removal of problematic 'rogue taxa' throughout the tree would result in a better-supported phylogeny, including the Baetidae as monophyletic clade and besides the Siphluriscidae as most ancestral mayflies.

REFERENCES

Aberer AJ, Krompass D, Stamatakis A. 2013. Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. Syst Biol. 62:162-166.

Alföldi J, Di Palma F, Grabherr M, Williams C, Kong L, Mauceli E, Russell P, Lowe CB, Glor RE, Jaffe JD*, et al.* 2011. The genome of the green anole lizard and a comparative analysis with birds and mammals. Nature. 477:587-591.

Heled J, Drummond AJ. 2010. Bayesian inference of species trees from multilocus data. Mol Biol Evol. 27:570-580.

3 Manuscripts

3.1 Recent Diversification

Chapter 1

Rutschmann S, Gattolliat J-L, Hughes SJ, Báez M, Sartori M, Monaghan MT (2014) Evolution and island endemism of morphologically cryptic *Baetis* and *Cloeon* species (Ephemeroptera, Baetidae) on the Canary Islands and Madeira. *Freshwater Biology*. **59**, 2516-2527. doi: 10.1111/fwb.12450

The original article is available at:
http://onlinelibrary.wiley.com/doi/10.1111/fwb.12450/full

Author contributions

S. Rutschmann collected the samples from the European mainland, performed all phylogenetic analyses, and drafted the manuscript. J-L. Gattolliat performed the morphological identification of the samples. S. J. Hughes, M. Báez, and M. Sartori designed and performed the fieldwork on the Canary Islands and Madeira. M. Sartori and M. T. Monaghan together with **S. Rutschmann** conceived the study, participated in its design and helped to draft the manuscript. All coauthors revised and contributed discussion to the final manuscript.

CHAPTER 2

Rutschmann S, Simon S, Detering H, DeSalle R, Funk DH, Gattolliat J-L, Raposeiro PM, Sartori M, Monaghan MT (*manuscript in preparation*) Colonization of Macaronesian Freshwaters: Phylogenetics using 59 Nuclear Markers derived from a Whole Genome Draft.

Author contributions

S. Rutschmann sampled the specimens from the Canary Islands, developed the laboratory protocols, performed some laboratory work, performed all phylogenetic analyses, and drafted the manuscript. S. Simon together with **S. Rutschmann** developed the genetic markers. H. Detering helped with bioinformatic analyses, implemented the bioinformatics program DISCOMARK in Python, and assisted during the fieldwork on the Canary Islands. R. DeSalle provided laboratory facilities and participated in the design of the marker development. D. H. Funk provided the laboratory specimens for the whole-genome sequencing. J-L. Gattolliat performed the morphological identification of the specimens. P. M. Raposeiro performed the fieldwork on the Azores. M. Sartori, M. T. Monaghan and S. Simon together with **S. Rutschmann** conceived the study, participated in its design and gave comments on the manuscript.

Colonization of Macaronesian Freshwaters: Phylogenetics using 59 Nuclear Markers derived from a Whole Genome Draft

SEREINA RUTSCHMANN, SABRINA SIMON, HARALD DETERING, ROB DESALLE, DAVE H. FUNK, JEAN-LUC GATTOLLIAT, PEDRO M. RAPOSEIRO, MICHEL SARTORI, AND MICHAEL T. MONAGHAN

Abstract. – The reconstruction of island colonization histories as evolutionary key processes rely on fully resolved phylogenetic tree reconstructions on a large geographic scale to cover all existing species. Standard phylogenetic markers are often unable to resolve evolutionary relationships among closely related species. This can hinder our ability to understand their evolutionary histories, in particular within groups that have undergone recent diversification. Mayflies (Ephemeroptera: Baetidae) are ancient freshwater insects that originated 300 million years ago (Ma) and have colonized the Macaronesian archipelagos (Azores, Madeira, Canary Islands) since their formation in the last 14 Ma. Thus, their phylogenetic relationships remained particularly difficult. Here we identified 59 nuclear protein-coding genes suitable for phylogenetic reconstruction across the range of the *Cloeon dipterum* L. 1716 species complex, using a whole genome draft. For this purpose we developed the bioinformatics program DISCOMARK to streamline the identification of nuclear DNA (nDNA) markers. We sequenced the 59 nDNA markers for 29 individuals and assigned them to the different species based on (1) the general mixed Yule-coalescent (gmyc) model and (2) the identification of genetic clusters based on the nuclear sequence alignment. Phylogenetic reconstructions were inferred using a coalescent-based species tree with the inferred species assignments, and Bayesian inference based on the concatenated exon sequence alignment of 24,168 base pairs (bp). Our results indicate that the Macaronesian islands have been colonized by three independent dispersal events that originated from the European mainland. We delineate six distinct species, of which three occurred on the Macaronesian islands. One species showed two trans-oceanic dispersal events from Greece to the U.S. and back to the Azorean islands. The second species seemed to have colonized Madeira and dispersed eastwards to all Canary Islands. In contrast the third species was restricted to the Canary Islands and its dispersal route was from Gran Canaria towards the western islands. Moreover, we found that several species co-occurred. The approach used was very successful for developing nDNA markers for a non-model organism. This study highlights the crucial combination of coalescent-based phylogeography, species delineation, and systematics for resolving recent diversification events. [*Keywords.* – Baetidae, Macaronesia, multispecies coalescent, marker development, nDNA, phylogeography, species tree, whole-genome sequencing]

The reconstruction of island colonization histories relies on fully resolved phylogenetic tree reconstructions on a large geographic scale. Most phylogeographic studies have used relatively small sets of molecular DNA markers and/or a geographically restricted taxon sampling. Key advances have been made with the increasing abundance of genomic data derived from high-throughput sequencing and with the development of the multispecies coalescent model (Edwards 2009; Heled and Drummond 2010; Knowles and Kubatko 2010) (Rannala and Yang 2003), accounting for incomplete lineage sorting and species tree *vs.* gene tree conflicts due to ancestral polymorphism. In general, multiple processes can explain discordances in phylogenetic reconstructions, namely hybridization, gene duplication and loss, and incomplete lineage sorting (Degnan and Rosenberg 2009; Knowles and Kubatko 2010) (Maddison 1997; Nakhleh 2013). A disagreement between gene and species tree is more likely when the speciation events has occurred recently and when 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). Thus the phylogenetic reconstruction of closely related species remains challenging. For few groups, large sets of nDNA markers have been successfully used within closely related species groups as for example by O'Neill et al. (2013) to reconstruct the phylogenetic relationship of tiger salamander, using 95 nuclear markers designed with parallel tagged amplicon sequencing. Ruane et al. (2014) used eleven nuclear markers to reconstruct the relationships on the genus level of milksnakes.

Here we present one of the first studies using a large set of nuclear DNA markers to reconstruct island colonization processes of aquatic insects (Ephemeroptera, Baetidae) on three Atlantic Ocean island archipelagos (Azores, Madeira, and Canary Islands). These three oceanic archipelagos, along with Cape Verde, belong to the Macaronesian region. Their distances to the adjacent continental mainland vary from 110 (Fuerteventura, Canary Islands to Morocco) to more than 2000 km (Flores, 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 Oromí 2005; Illera et al. 2007; Arnedo et al. 2008; Dimitrov et al. 2008), or multiple independent colonization events (Nogales et al. 1998; Ribera et al. 2003a; Diaz-Perez et al. 2012). While much research has been carried out on island evolution and endemism of terrestrial organisms, comparatively limited information exists for aquatic invertebrates (Stauder 1991; 1995; Drotz 2003; Ribera et al. 2003b; Ribera et al. 2003c; Jordal and Hewitt 2004). However, considering that aquatic invertebrates have a disproportional contribution to biodiversity given the relatively small extent of their habitat, this is a huge discrepancy (Dijkstra et al. 2014).

Mayflies are well suited for phylogeographic studies because of their ancient origins (300 Ma), global distribution, and limited dispersal ability due to the strict water habitat fidelity of larvae and very short life of the winged adults (Sartori 2001; Monaghan et al. 2005; Barber-James et al. 2007). Recent studies pointed out unusual potential for dispersion, reporting mayfly species on remote islands such as the Azores in the Northern Atlantic Ocean (Brinck and Scherer 1961), La Réunion in the Indian Ocean (Gattolliat et al. 2004), or Vanuatu in the Pacific Ocean (Gattolliat and Staniczek 2011), 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 ≈ 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 most abundant among freshwater insects. The species complex may be considered more or less equivalent to the subgenus *Cloeon* LEACH, 1815, which comprises besides *C. dipterum* currently two recognized species (*C. peregrinator* GATTOLLIAT & SARTORI, 2008, and *C. saharense* SOLDÁN & THOMAS, 1983) and three species with unclear taxonomy (*species inquirenda; C. cognatum* STEPHENS, 1836, *C. inscriptum* BENGTSSON, 1914, and *C. rabaudi* VERRIER, 1949). Its distribution ranges from North America, across Europe to Northern Asia (excluding China), making the inter-continental distribution one of the largest known among mayflies (Bauernfeind and Soldán 2012, and references therein). The specimens are found in a variety of aquatic habitats, including any natural habitat, brackish water, periodical watercourses, as well as artificial biotopes from a wide range of climatic zones (Bauernfeind and Soldán 2012 and references therein; Barber-James et al.

2013, http://fada.biodiversity.be/group/show/35). The larvae have remarkable resistances against low oxygen conditions, surviving three to four months of complete anoxia (Nagell 1977, 1980, 1981), water temperatures between ≈0-35°C (Nagell 1977; Cianciara 1979, 1980; Nagell 1981; Soldán and Thomas 1983), high salinity, short-term freezing in ice, and a pH range of ≈4.5-10.3 (Soldán and Zahrádková 2000). A high reproductive fitness is supported by their ovoviviparous lifestyle (Gillies 1949), whereby eggs do not require direct contact with water. Embryogenesis can be completely finished in the female body and parthenogenesis i.e. reproduction via unfertilized viable eggs from females without having copulated (Harker 1997), promoting dispersal abilities. The role of parthenogenesis might differ between populations, leading in the most extreme case to populations only consisting of females. The females have an extraordinary long life span between ten and 14 days (Degrange 1960; Oehme 1972) and sometimes surviving up to 48 days (Silina 1994). *Cloeon dipterum* s.l. was thought to be the only mayfly species occurring on all three Macaronesian archipelagos (Brinck and Scherer 1961; Müller-Liebenau 1971; Alba-Tercedor et al. 1987; Soldán et al. 1987; Stauder 1991, 1995; Malmqvist et al. 1995; Nilsson et al. 1998; Borges et al. 2005, 2010; Gattolliat et al. 2008; Raposeiro et al. 2012; Rutschmann et al. 2014). However, early taxonomical studies on the Canary Islands reported besides *C. dipterum* the presence of *C. cognatum* (Alba-Tercedor et al. 1987; Malmqvist et al. 1995), and the occurrence of one, respective two, additional *Cloeon* sp. (Malmqvist et al. 1995; Nilsson et al. 1998). More recent work based on mtDNA analyses identified populations on Madeira to be an endemic species (*C. peregrinator*, Gattolliat et al. 2008), and also found several Canarian species (*C.* sp1 and *C.* sp2, Rutschmann et al. 2014), challenging the validity of the species name of *C. dipterum* on any of the islands. Alba-Tercedor et al. (1987) reported phenotypic differences (arrangements of gills, mouthparts, and color patterns) within the specimens on Tenerife even being more pronounced than those found among their counterparts on the European mainland (*C. cognatum, C. dipterum, and C. inscriptum*; Sowa 1975). *Cloeon dipterum* has first been reported from the U.S. based on a single female from Illinois (Burks 1953) but its presence in North America remained controversial, being either regarded as non-native (Traver 1962; McCafferty et al. 2008) or possessing an old Holarctic distribution (Randolph et al. 2003). However, in a recent barcoding project by Webb et al. (2012), the *Cloeon* specimens have been assigned to *C. cognatum*. In conclusion, the exact taxonomic classification and phylogenetic relationships within the

C. dipterum species complex, including its complicated synonymy, remain largely unknown. In addition, the ecological and phenotypic differences provide strong evidence for the occurrence of multiple unrecognized species.

We developed a large multilocus dataset (59 nDNA markers) derived from whole-genome data in order to reconstruct the colonization of Atlantic oceanic islands by mayflies. The use of a multilocus data set enabled us to infer fine-scale evolutionary histories of closely related species that will explain the colonization pathways of the *C. dipterum* s.l. species group. More specifically, our objectives were: (i) to reconstruct the colonization of the Macaronesian archipelagos by the *C. dipterum* s.l. species group, (ii) to delineate the species boundaries within *Cloeon sp*., (iii) to investigate genetic differences between the different *Cloeon* species, (iv) to develop a large set of nDNA markers derived from a draft genome, and (v) to streamline the *ab initio* development of nDNA markers derived from high-throughput sequencing data by introducing the bioinformatics program DISCOMARK.

MATERIAL AND METHODS

Sampling and DNA Extraction

We sampled individuals from the Azorean archipelago (Faial, Pico, São Jorge, São Miguel, and Terceira islands), the Canary Islands (islands of El Hierro, Fuerteventura, Gran Canaria, La Gomera, Lanzarote, La Palma, and Tenerife), Madeira, and the European mainland. In total, we included for this study 107 newly sampled *C. dipterum* s.l. individuals from larval aquatic habitats. From the Macaronesian islands, we included 90 individuals from 38 sampling sites on the 13 islands (Fig. 1; APPENDIX 1). All samples were preserved in 99% ethanol in the field and stored at 4°C until analysis. Genomic DNA was extracted from whole specimens using the NucleoSpin® 96 (Macherey-Nagel, Düren, Germany) tissue kits. Our sampling included all the currently morphologically recognized taxa on the islands (Brinck and Scherer 1961; Gattolliat et al. 2008; Rutschmann et al. 2014).

Marker Development from Genomic Data

To develop a whole set of new nuclear markers, we used sequences from one newly created whole-genome library of *C. dipterum* and from 4,197 expressed sequence tag sequences (EST) of *Baetis* sp. (GenBank Acc. no. FN198828–FN203024; (Simon et al. 2009), being the sister taxa of *Cloeon* (Monaghan et al. 2005). In order to facilitate the design of new markers from predicted orthologous genes, we further developed the bioinformatics program DISCOMARK (discovery markers), which designs primer pairs in the conserved parts of the orthologous sequences from the included taxa.

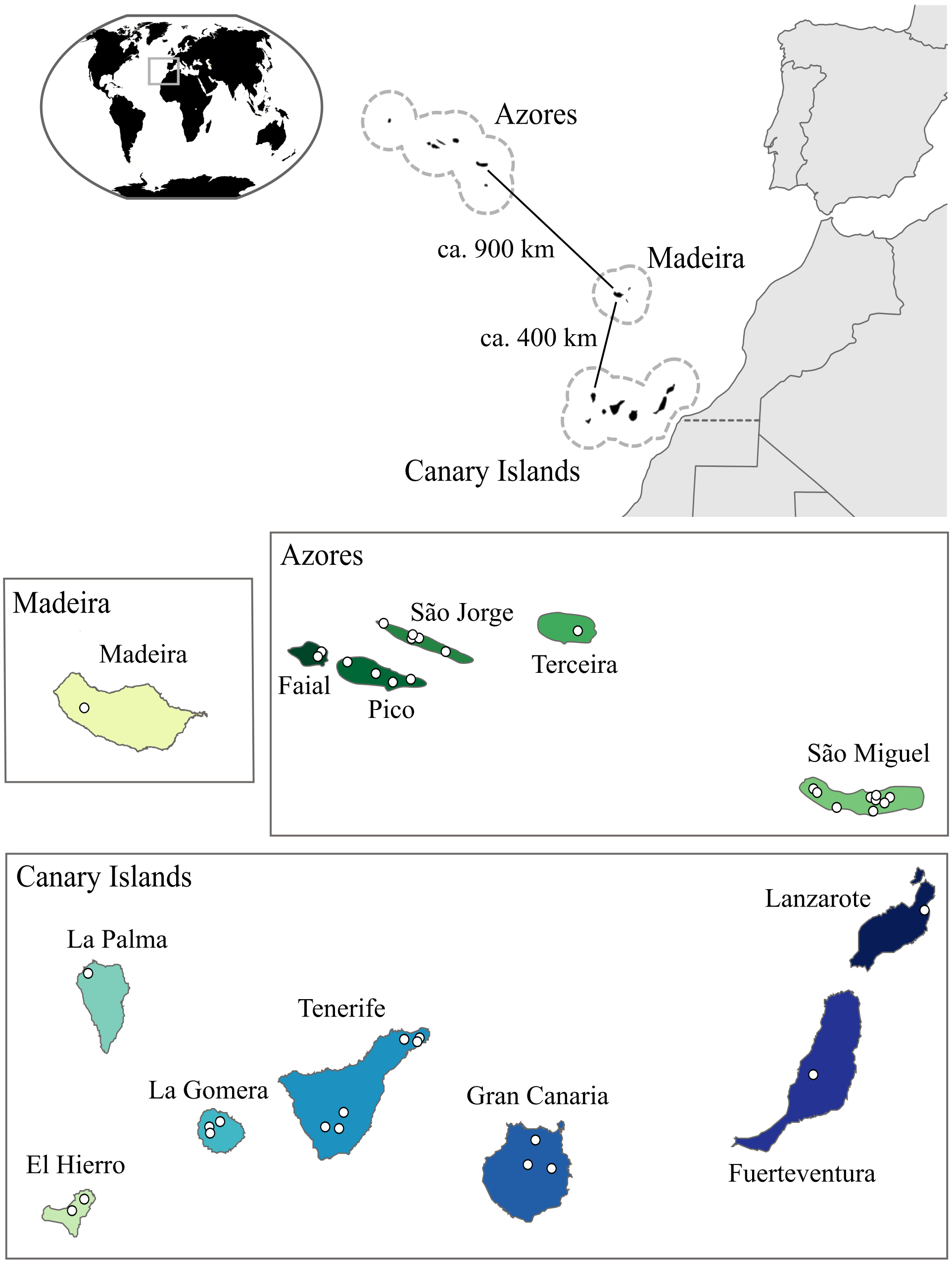


FIGURE 1. Map of the sampling localities in the Macaronesian region. Sites are indicated by filled white circles. Islands are colored differently and the same colors are used in FIGURES 4, 5 and APPENDIX 5. Notes: Only islands with sampling sites are shown; thus only the central Azorean islands.

For generating the whole-genome sequencing libraries, we used reared subimago *C. dipterum* specimen. The Invisorb® Spin Tissue Mini (STRATEC, Berlin, Germany) was used to extract the DNA from five to 20 pooled specimens after removing their eyes and wings. The extracted DNA was precipitated (Isopropanol precipitation of DNA, QIAGEN, Leipzig, Germany), and pooled according manufacturer's guidelines in order to obtain higher DNA yield. We prepared a shotgun and paired-end library according to the manufacturer 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 at the Berlin Center for Genomics in Biodiversity Research (BeGenDiv, Berlin, Germany) according to manufacturer's guideline (Sequencing Method Manual, GS FLX Titanum Series, October 2009 (Rev. Jan 2010)). The sequence reads were trimmed and *de novo* assembled using the 454-provided software NEWBLER v. 2.5.3 (454 Life Sciences Corporation) under the default settings for large datasets. Thereby, we made two different assemblies, one with the reads from the shotgun library and one with the reads from both shotgun and paired-end library.

The obtained assembled genomic data were used for orthologous sequence prediction. Identification of orthologous genes from assembled *C. dipterum* contigs and the *Baetis* sp. EST sequences was carried out using HAMSTR (Ebersberger et al. 2009), v.9 (http://www.deep-phylogeny.org/hamstr/download/archive/hamstrsearch_local_v9.tar.gz). We used the insecta_hmmer3-2 core reference taxa set (http://www.deep-phylogeny.org/hamstr/download/datasets/hmmer3/insecta_hmmer3-1.tar.gz), including 1,579 orthologous genes. HAMSTR is designed primarily for use with EST sequences data rather than whole-genome data, but produced the most reliable results among the available tools. Moreover, the results were carefully inspected and compared with those of the *Baetis* sp. EST dataset for quality assurance.

We first manually designed exon-primed intron-crossing (EPIC) markers based on the predicted orthologous genes and in a second step we developed the bioinformatics program DISCOMARK. We combined predicted orthologous genes from both species and aligned them with MAFFT v.7.050b under the default parameters (L-INS-I algorithm with default settings; Katoh and Standley 2013). In order to infer the exon-intron boundaries, which was important to estimate intron length and essential to design EPIC markers, we performed BLAST searches (Altschul et al. 1997) of each orthologous sequence alignment against the assembled *C. dipterum* contigs. The predicted orthologous genes and the matching contigs were then re-aligned using MAFFT. All alignments were inspected by eye and we filtered out alignments containing introns with more than 1000 base pairs (bp) lengths. Primer pairs for each alignment were designed using PRIFI (Fredslund et al. 2005), specifying an estimated PCR product length between 100 and 1000 bp, and melting temperature between 50°C and 60°C. All obtained primer pairs were tested for their specificity with PRIMER-BLAST (Ye et al. 2012), whereby we excluded primer sequences mapping to primates or bacteria, potentially leading to contamination. The function of the genes was assessed through BLAST against the eukaryotic orthologous groups (KOGs) database (http://biotec.icb.ufmg.br/K-EST/begin.html) and assigned to the four major KOG categories: cellular processes and signaling, information storage and processing, metabolism, and poorly characterized (Table 1).

TABLE 1. Nuclear markers designed for *Cloeon dipterum* s.l.

Marker	KOG	Function, (Category)	FW Primer Name: Primer Sequence (5'-3')	RV Primer Name: Primer Sequence (5'-3')
411892	KOG1570	60S ribosomal protein L10A, (IP)	411892-FW: CTCAAACACATTCCTCGTCCC	411892-RV: ACATTCTGCCARTGCTTCTTC
411912	KOG0307	Vesicle coat complex COPII, subunit SEC31, (CS)	411912-FW*: AAATGCCTCAGAATCAGATGAG	411912-RV: AAAAAGAATTTCCAATTTCCTGCC
411913	KOG0660	Mitogen-activated protein kinase, (CS)	411913-FW: CAGATTTGTGACTTTGGTCTCG	411913-RV: CTGGGTCATAATACTGCTCTAAG
411925	KOG4067	Uncharacterized conserved protein, (PC)	411925-FW: ATCACCGAAACACAATCAGTCTTC	411925-RV: AAAGTCCGGATTTTGTGCTAG
411939	KOG1101	Apoptosis inhibitor IAP1 and related BIR domain proteins, (CS, PC)	411939-FW: ATCGTCTCTATTCTYTGCTG	411939-RV: ACTTTTACCACGAATGAAGGTCCC
411945	KOG2574	mRNA splicing factor PRP31, (IP)	411945-FW: CCTCCAGTGAAATTCATCAAACCC	411945-RV: TCCTCTCCACCCACTTTCTC
411965	KOG0027	Calmodulin and related proteins (EF-Hand superfamily), (CS)	411965-FW: GAGAGTGCTTTTACCTGTTTGC	411965-RV: GTAGTAGTCAGGCACTGGTG
411989	KOG0332	ATP-dependent RNA helicase, (IP)	411989-FW: GTACCAGATCCAATCATCATCAGG	411989-RV: TCTTTGGAGGTCTATAGGAAGGTC
412045	KOG3167	Box H/ACA snoRNP component, involved in ribosomal RNA pseudouridinylation, (IP)	412045-FW: AAGGCGAAGTTTCGTACACC	412045-RV: TTGTAGTCTGGTTTTGGCTTGATC
412048	KOG3434	60S ribosomal protein L22, (IP)	412048-FW: GCAGAAAAAGAAGAAGGTCCAG	412048-RV: TCTTCCTCCTCATCTTCCTG
412085	KOG4036	Uncharacterized conserved protein, (PC)	412085-FW: GAGGAACAGAAGAAAAAGCGTC	412985-RV: TGTCATCCTGAAGACTATTCACAG
412111	KOG1147	Glutamyl-tRNA synthetase, (IP)	412111-FW: CTTACGCCTACCAAAAGGAAG	412111-RV: GCTCAACTCGAATGGGTACAC
412148	KOG3101	Esterase D, (PC)	412148-FW: CTACCAGATGTTCTCTTACGTCAC	412148-RV: GTGATGCTTAATGTGGTCATCCAC
412168	KOG0875	60S ribosomal protein L5, (IP)	412168-FW: AGGCATACTTCAAAAGATTCCAAG	412168-RV: TGTGAGGAATGTTCAGTCCACC
412192	KOG0714	Molecular chaperone (DnaJ superfamily), (CS)	412192-FW: TACACGAGTGATCAATTGGAGG	412192-RV: GCGCTTTCCATAAYATAGACTCC
412199	KOG3317	Translocon-associated complex TRAP, beta subunit, (CS)	412199-FW: TCGAAGCACATTTTGAACCG	412199-RV: GTCTTTGAAAGCAACAATAGCCC

412207	KOG3285	Spindle assembly checkpoint protein, (CS)	412207-FW: TCTTCGAAATCTGCATCAATCAAC	412207-RV: AATTTTAAAAGTGCATCCTTCAGG
412211	KOG3418	60S ribosomal protein L27, (IP)	412211-FW: ATGGGTAAAATTATGAAGTCGGGC	412211-RV: CTGGAAGAACCACTTGTTCTTTCC
412221	KOG0357	Chaperonin complex component, TCP-1 epsilon subunit (CCT5), (CS)	412221-FW: GTCATCAGAAACCTGGTGAAGG	412221-RV: CTTAAGGATCATCTTTACCAGCTG
412236	KOG0313	Microtubule binding protein YTM1 (contains WD40 repeats), (CS)	412236-FW: AACCAACTGATTGTCTTGAACACG	412236-RV: GTCGTAAAGTCTAACGTGTCTGTC
412242	KOG2684	Sirtuin 5 and related class III sirtuins (SIR2 family), (IP)	412242-FW: CCATTGATGGAGTTCATCCAGTC	412242-RV: AATATTTTCACAGTTCATGCACCG
412250	KOG2030	Predicted RNA-binding protein, (PC)	412250-FW: AACAGCAACAGATGAAGAGAG	412250-RV: CTTCAATTTCACCATTTGTGGAGC
412320	KOG2000	Gamma-tubulin complex, DGRIP91/SPC98 component, (CS)	412320-FW: CAAATCCGCATCATCACACATC	412320-RV: AACACYGGGTGAGAAAAGCC
412334	KOG2577	Transcription factor E2F/dimerization partner (TDP), (IP)	412334-FW: TGCAGAGCATGAAAAATGTCAC	412334-RV: ATAGTCTTGCGCTGTGGTAGG
412343	KOG3083	Prohibitin, (CS)	412343-FW: GGTCACGACCCAGAAACATACC	412343-RV: GGAAGGTAGATGACGTTCCTTG
412379	KOG2212	Alpha-amylase, (M)	412379-FW: AGGATCACCGAGTTCAGATTC	412379-RV: ACCACGTTTTTAAATTCAATCATG
412426	KOG0940	Ubiquitin protein ligase RSP5/NEDD4, (CS)	412426-FW: AAAGCGATCTACGACAACAAG	412426-RV: GTCATTTTCATCTGGCACACG
412438	KOG0183	20S proteasome, regulatory subunit alpha type PSMA7/PRE6, (CS)	412438-FW: CGGTACAGCTTTTCGTTGAC	412438-RV: AAATTCCCACCTCAATGTTGTCAG
412519	KOG2558	Negative regulator of histones, (IP)	412519-FW: TCGCCTTCTTATACCACAAGG	412519-RV: AGCATGAATACTATTCGATGGC
412665	NO related KOG	NO related KOG, (PC)	412665-FW: TTGATAGCCACAAACARAGCC	412665-RV: TGACCTTTTCTTTCAATTCGCTTC
412670	KOG3237	Uncharacterized conserved protein, (PC)	412670-FW: TCTTGGAAAAATGCATCGAAGG	412670-RV: TGAAGTCTCTTCAGCTTGTTTC
412679	KOG0270	WD40 repeat-containing protein, (PC)	412679-FW: ACAGTTTATTCAATGAGTGGGAGC	412679-RV: GTGTCCTTTGCATCCACCTTC
412698	KOG3038	Histone acetyltransferase SAGA associated factor SGF29, (PC)	412698-FW: GATGAAAATGCTGCAAATATCTGC	412698-RV: AGCCACTGTTAAAGGAGGAG
412704	KOG3024	Uncharacterized conserved protein, (PC)	412704-FW: TCGTGAATATCAGTTGACGGATTC	412704-RV: GCATGATGGTATCTTTGACAGC
412727	KOG1712	Adenine phosphoribosyl transferases, (M)	412727-FW: TTTTTAAAGGGCTGTCTGAAAAGC	412727-RV: AGTCAATTCAATTACGACCAGGC
412741	NO related KOG	NO related KOG, (PC)	412741-FW: GCGAATCCAGAAAATAGTAGCC	412741-RV: GATGAGAGTCCGTTCTTTTGGTC
412757A	KOG0822	Protein kinase inhibitor, (CS)	412757A-FW: GGAATGTGTTTCGCTCAGTAG	412757A-RV: TGGATTTTTCTCCACTGCATAGAC
412757B	KOG0822	Protein kinase inhibitor, (CS)	412757B-FW: GACTGTTATGGTCGTTGGGG	412757B-RV: CTCATGATCACATCTTTGTACAGG
412825	KOG3256	NADH:ubiquinone oxidoreductase, NDUFS8/23 kDa subunit, (M)	412825-FW: GGGGTTGCTACAAAATTGTGACTC	412825-RV: TGTTAGAAGCAATTTCAGACTCCC
412828	KOG0829	60S ribosomal protein L18A, (IP)	412828-FW: ATACAAAATGAGGATTTTCGCCCC	412828-RV: GATGGAAAATCTGGGAAGGGTG
412840	KOG4009	NADH-ubiquinone oxidoreductase, subunit NDUFB10/PDSW, (M)	412840-FW: TTTTGATGGTTTCATCAATGCTGC	412840-RV: GCTTCATGTAAGCATCCTTCAC
412852	KOG1666	V-SNARE, (CS)	412852-FW: GCTGACAGAAAAGATGCCAC	412852-RV: GATCTCTTCAGTCTCAAGAGC
412884	KOG0898	40S ribosomal protein S15, (IP)	412884-FW: AACTAAGAAGAAGAGGGCTTTCC	412884-RV: GAGGAATAAATCTAGAGCTGTGAG
412894	KOG2764	Putative transcriptional regulator DJ-1, (CS, PC)	412894-FW: CCTTTTGCTTCTGGCTTATG	412894-RV: TAAACAGGAAGAAGTTAGCCAGG
412937	KOG1751	60s ribosomal protein L23, (IP)	412937-FW: AAGCCGAGGAGAAGAAAAAG	412937-RV: CAAGGTGTTGACCTTAGCCAC
412964	KOG4018	Uncharacterized conserved protein, contains RWD domain, (PC)	412964-FW: AAGGCTCAAGAAAATTTGGGAATG	412964-RV: CATCCATGTCCTCAAACAGAGTC
412985	KOG1690	emp24/gp25L/p24 family of membrane trafficking proteins, (CS)	412985-FW*: GTTCAGCTCTATGATCCGAG	412985-RV: AGTTTTTTGGCCTCRAAGAAGC
412986	KOG0385	Chromatin remodeling complex WSTF-ISWI, small subunit, (IP)	412986-FW: CAAGTCTGTCGGATACAAAGTCC	412986-RV: GTTGTTAGAAGTTGCCTGTGGG
413065	KOG3095	Transcription initiation factor IIE, beta subunit, (IP)	413065-FW: AAAACTTCAAAAGAAGRGCAATCG	413065-RV: TTTGAACAGAAACCTTTTCTCGC
413094	KOG3424	40S ribosomal protein S24, (IP)	413094-FW: CGATAGCCAAGACTGTCATC	413094-RV: CTTGGTCTTCTTGGTTCCTC

413147	KOG0438	Mitochondrial/chloroplast ribosomal protein L2, (IP)	413147-FW: GGTTAGCCGAYTGGCCAT	413147-RV: ATTCCCAYTGCCAACGAG
413200	KOG1897	Damage-specific DNA binding complex, subunit DDB1, (IP)	413200-FW: GCCTATCAAGAAACATCCCAAACC	413200-RV: TGTTGTTGAAGTGACTGCACTC
413263	KOG1624	Mitochondrial/chloroplast ribosomal protein L4, (IP)	413263-FW: AAGTTGGAATTCATGCCACCAAAC	413263-RV: CCTACGTTTTCTGTTGCCTTG
413280	KOG1339	Aspartyl protease, (CS)	413280-FW: TCAATACTATGGCCCAATCAGC	413280-RV: AGGTAAAACGAGAAGACAGGAG
413294	KOG0052	Translation elongation factor EF-1 alpha/Tu, (IP)	413294-FW: ATCAACATCGTGGTCATCGG	413294-RV: TGGAGTCCATCTTGTTGACACC
413321	KOG3129	26S proteasome regulatory complex, subunit PSMD9, (CS)	413321-FW: ATCGACGTTTAYCAAGTTCGAC	413321-RV: TTGACGTGGATTGGCATGTTC
413388	KOG2020	Nuclear transport receptor CRM1/MSN5 (importin beta superfamily), (CS)	413388-FW: CAGACCTACTTCACCGACATTC	413388-RV: CTCGTGAGGGTTCAGAATACC
413390	NO related KOG	NO related KOG, (PC)	413390-FW: CGAAGTGTGTCAGCTTGTTC	413390-RV: TTCACAACATTGGTGACAAACCAG
413415	KOG0517	Beta-spectrin, (CS)	413415-FW: AGGAACAACTCAACGAGTTCC	413415-RV: ATAAAGGGCTGTAGAGAAGGAC

Notes: *, Primers do not work for sequencing. Categories are listed in parentheses (IP: information storage and processing, CS: cellular processes and signaling, M: metabolism, PC: poorly characterized)

The DISCOMARK program is written in Python and uses predicted orthologous genes optionally combined with genomic data such as whole-genome sequencing data as used for this study, to design primer pairs. Therefore, DISCOMARK performs seven steps combining Python scripts with widely used bioinformatics programs. In short, the steps include: (1) parsing of input files, (2) aligning the sequences of each orthologous gene using MAFFT, (3) trimming of orthologous sequence alignments with TRIMAL (Capella-Gutierrez et al. 2009), (4) mapping of orthologous sequence alignments against reference databank (e.g. whole-genome contigs or EST dataset ideally from the same or closely related species) with local BLAST searches and realignment using MAFFT, (5) design primer pairs on sequence alignments using PRIFI, (6) checking primer specificity with PRIMER-BLAST, (7) producing HTML summary output. DISCOMARK can be downloaded from GITHUB (https://github.com/hdetering/discomark).

PCR Amplification, Sequence Alignment and Haplotype Reconstruction

In total, we sequenced 59 newly developed markers for a representative set of *C. dipterum* s.l specimens. To identify these specimens, we sequenced the mitochondrial DNA barcoding gene (cytochrome *c* oxidase subunit 1 (*cox1*) gene) of all individuals using the procedure described by Rutschmann et al. (2014). Based on these preliminary barcoding results, we selected a representative set of 29 individuals for which we obtained sequences of these 59 manually designed primer pairs (Table 1; APPENDIX 2).

All markers were amplified using standard polymerase chain reactions (PCR) protocols with an annealing temperature of 55°C. All 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.). Heterozygous indels were detected and resolved using the find and split heterozygous function implemented in CODONCODE ALIGNER v.3.5.6 (CodonCode Corporation). Multiple sequence alignments were made for each locus using MAFFT. The predicted orthologous sequences of *Baetis* sp. (see above) were used to infer the correct exon-intron splicing boundaries (canonical and non-canonical splice site pairs) of each alignment. For alignments with different exon-intron boundaries between the predicted orthologs from the whole-genome shotgun library and the EST data, we used the predicted boundaries according to the EST data. Exon-intron boundaries of the marker 411912 could not be fully reconstructed and thus we used the exon sequence predicted from TBLASTX searches for subsequent analyses. We used a Python script (extract_introns.py; https://github.com/srutschmann/python_scripts) to split the gene alignments into coding and noncoding parts. All coding alignments were checked for indels and stop codons using MESQUITE v.2.75 (Maddison and Maddison 2011). Haplotypes from 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 Perl scripts seqphase1.pl and seqphase2.pl from SEQPHASE (Flot 2010). Heterozygous sites that could not be resolved were coded as N or ambiguity codes for subsequent sequence analyses. All alignments were re-aligned with MAFFT after phasing. 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 phylogenetic analyses, we prepared three alignment sets, whereby we used full genes (= full), coding genotypes (= exon) and coding haplotypes (= exonhap). All alignment sets were not 100% complete, so we made two matrices for the manually developed markers, one with all genes, consisting of the 59 sequenced genes (= all_gene), and one with all taxa, including 17 genes for which we had data from all 29 individuals (= all_taxa). Thus we ended up with the following matrices: full_all_gene, full_all_taxa, exon_all_gene, exon_all_taxa, exonhap_all_gene, and exonhap_all_taxa (Tables 2, 3).

TABLE 2. Overview of Data Sets

Data Set	Number of Taxa	Number of Markers	Concatenated Length [bp]
mitochondrial gene (*cox1*)	141	1	658
full_all_gene	29	59	32,213
full_all_taxa	29	17	8,565
exon_all_gene	29	59	24,168
exon_all_taxa	29	17	6,485
exonhap_all_gene	29	59	-
exonhap_all_taxa	29	17	-

To investigate the heterogenity among the newly developed markers, we reconstructed reticulation-free haplotype genealogies based on Fitch distances (Fitch 1970; Salzburger et al. 2011), and assesed variable sites of each marker across populations using the program FITCHI (Matschiner 2014; http://www.evoinformatics.eu/fitchi). Therefore, we used the exonhap_all_gene dataset and calculated for each marker a gene tree using the program RAXML v.8 (Stamatakis 2014) with the GTRCAT model for each gene tree with 1000 bootstrap replicates under the rapid bootstrap algorithm.

TABLE 3. Nuclear markers with corresponding lengths and best-fitting model of molecular evolution inferred with JMODELTEST. Models in parentheses refer to different inferred models for the species assignment based on the STRUCTURE analyses

Marker	Full Length [bp]	Exon Length [bp]	Exon Model	Exon Variable Sites *Cloeon*	Number of Sequences *Cloeon*
411892	575	397	SYM + Γ	47	27
411912*	210	210	JC	28	29
411913*	543	423	K80	10	29
411925	485	485	JC	46	27
411939	631	631	K80	76	20
411945	486	374	F81	43	25
411965*	589	364	K80 + I	25	29
411989	335	335	K80	14	29
412045	486	313	K80 + I	25	27
412048	536	299	JC	24	26
412085	433	266	HKY	13	28
412111	738	542	K80 + I	33	28
412148*	541	426	K80	26	29
412168	699	451	K80 + Γ	52	28
412192*	819	688	K80	38	29
412199*	568	298	K80 (K80 + I)	11	29
412207*	454	331	K80	20	29
412211	466	344	JC	20	22
412221	481	319	K80	16	20

412236	792	619	K80 + I	58	27
412242	345	345	K80 (K80 + I)	21	26
412250	658	547	F81	35	28
412320	502	351	JC + Γ	15	28
412334	525	468	K80	28	23
412343	646	529	K80 + I (K80)	33	26
412379*	373	373	K80	18	29
412426	259	259	K80 + I	22	28
412438	1007	580	K80	46	27
412519*	321	321	K80	25	29
412665	277	219	K80	41	28
412670*	320	320	K80	21	29
412679*	298	240	K80 + I	25	29
412698	548	436	K80 + I	33	27
412704	824	558	K80	25	27
412727	850	408	K80	15	24
412741*	458	341	JC + Γ	21	29
412757A	675	568	K80	21	28
412757B	756	576	K80	23	28
412825	726	461	JC + Γ	35	27
412828	803	380	K80	18	28
412840*	608	339	K80	24	25
412852*	432	315	HKY + I	15	25
412884	494	368	K80	20	29
412894	753	516	K80	40	29
412937*	728	482	HKY (F81)	22	21
412964	451	331	F81 + Γ + I (F81 + Γ)	10	27
412985	641	457	K80	37	29
412986*	884	711	HKY + I (HKY)	35	27
413065	587	347	K80 (HKY)	23	28
413094	231	231	K80 + I	15	29
413147	315	315	JC + Γ (JC)	21	24
413200	503	503	K80	23	17
413263	625	514	K80	34	27
413280	471	390	K80 + I (K80)	17	15
413294	493	408	K80	16	28
413321*	419	303	K80	30	27
413388	476	409	K80	11	27
413390	558	450	K80 + I	20	29
413415	506	384	K80	14	28

Note: *, Markers used for the all_taxa matrices. F81 implemented as JC.

Species Assignment and Population Structure Analysis

We used two approaches to assign the 29 *C. dipterum* specimen to putative species/genetic clusters: (i) the general mixed Yule-coalescent (gmyc) model analysis (Fujisawa and Barraclough 2013) , and (ii) a Bayesian clustering algorithm to assign individuals to populations (Pritchard et al. 2000; Falush et al. 2003). The gmyc approach was carried out based on the mitochondrial *cox1* alignment of 141 specimens (APPENDIX 2). Therefore we included all available published *cox1* sequences of the genus *Cloeon*, six own *Cloeon simile* EASTON, 1870 sequences, since the taxonomy of the whole genus remains largely unknown, and as outgroup the damselfly *Euphaea formosa* (GenBank Acc. no. NC_014493). The analysis followed that of Rutschmann et al. (2014) with the only difference, that we used BEAST v.2.1.3 (Bouckaert et al. 2014) and as partition scheme the codon positions (1, 2) + 3 instead of using each codon position as partition (i.e. 1 + 2 + 3). Population structure was inferred based on the nDNA data. Therefore, we applied a Python script to format the concatenated exon_all_gene alignment to use as an input file for STRUCTURE v.2.3 (msa2structure.py; https://github.com/srutschmann/python_scripts). We assumed 1-10 genotypic clusters (K) and ran nine replicate analyses for each K, using 1 × 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. The STRUCTURE HARVESTER v.6.94 Python script (Earl and vonHoldt 2011) was used to monitor convergence among runs, and asses mean InPr(X'K) across replictes for each K to calculate ΔK according the Evanno method (Evanno et al. 2005). After a first round of STRUCTURE analyses on the full dataset, we created separate data partitions to explore fine-scale population structure (sensu O'Neill et al. 2013). Therefore we used the following two partitions: (i) a cluster of ten individuals from the European mainland, which were detected as one cluster in the full data analysis, and (ii) a cluster of 19 individuals from the Macaronesian archipelagos, U.S., and Greece, which were also assigned to one cluster in the full data analysis. These two datasets were analyzed as described above with five replicate analyses.

Phylogenetics

Concatenated Phylogenetics. – We performed Bayesian phylogenetic reconstructions using the program MRBAYES v.3.2.2 (Ronquist et al. 2012). We used the exon_all_gene, the exon_all_taxa. As outgroup we used the predicted orthologous sequences from *Baetis* sp.. The most appropriate substitution models for each gene were determined using Bayesian Information Criterion in the program JMODELTEST v.2.1 (Guindon and Gascuel 2003; Darriba et al. 2012) (Table 2). All individual gene alignments were concatenated using a Python script (fasta_concat.py; https://github.com/srutschmann/python_scripts). For the tree reconstruction, we implemented the best-fit models for each gene, 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 × 10^7 generations and 25% burn-in were run.

Multilocus Species Tree Phylogenetics. – Phylogenetic reconstruction was 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 two different sets of genes (exonhap_all_gene and exonhap_all_taxa), whereby we calculated the best-fit model of molecular evolution with JMODELTEST (Table 3). As required by *BEAST, all individuals were *a priori* assigned to population. Therefore we used the results based on the gmyc approach and the STRUCTURE analyses. Thus we pre-defined six populations based on the gmyc analysis and five populations based on the STRUCTURE analysis (see Results). For the latter, individuals with posterior probability (PP) assignment values >0.05 for more than one cluster were considered to be admixed and were excluded from the analysis. For each 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 × 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 gene and the species trees were obtained using TREEANNOTATOR v.2.1.3 (Bouckaert et al. 2014).

RESULTS

Marker Development

The sequences were deposited with BioSample Acc. no. SAMN03202660, BioProject ID PRJNA268073, and Sequence Read Archive no. XXXXXXXX. The 454 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 1187 and 736 bp, repsectively. All reads were assembled into 68,473 contigs with an N50 of 1116 bp. The reads of the shotgun library were assembled into 31,827 contigs with an N50 of 1260 bp.

The HAMSTR approach detected for 918 orthologous gene sequences for *C. dipterum* from the contigs derived from the shotgun library, 1,298 orthologous gene sequences from the contigs of the combined assembly, and 416 for *Baetis* sp. We successfully designed primer pairs for 59 sequence alignments (Table 1), mostly based on orthologous gene sequences from both species. Most of these orthologous sequences were assigned to the categories cellular processes and signaling (22) and information storage and processing (21). For the alignment of marker 412757, we designed two primer pairs (412757A, 412757B) due to its length (>1000 bp). DISCOMARK found 326 orthologous sequences of both species and designed primer pairs for 73 sequence alignments.

The concatenated alignment lengths were 32,213 bp for the full_all_genes, 24,168 bp for the exon_all_genes, 8,565 bp for the full_all_taxa, and 6,485 bp for the exon_all_taxa matrix, ranging per marker from 210 to 711 bp exon length with an average exon length of 410 bp per marker (Tables 2, 3). Our final seqeunce alignment contained for each specimen >75% of all markers (>44.25 markers). The marker 411912 included a three bp indel in the coding sequence of the specimens belonging to EU1. All heterozygous indels were located in the intron sequences. However, 100 heterozygous sites could not been resolved and remained in the exonhap alignments. The exon_all_gene alignment included 1,384 single nucleotide polymorphisms (SNPs). We found on average per marker 26.66 variable sites (range: 10-76). All haplotype genealogies showed clear structering (Fig. 2; APPENDIX 3). For 26 markers, we found shared haplotpyes between different gmyc species.

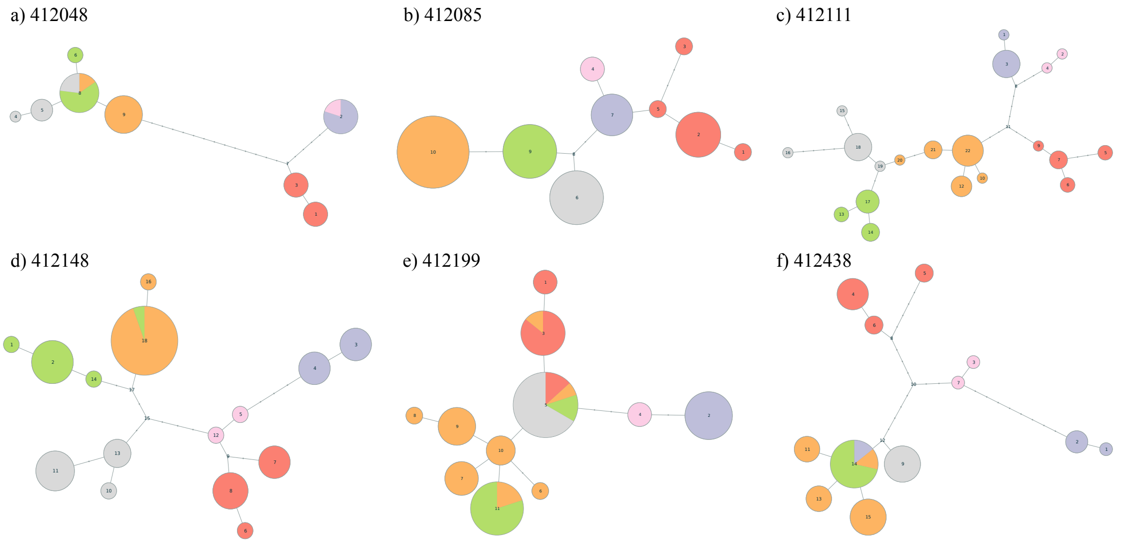


FIGURE 2. Haplotype networks of nDNA markers based on exonhap_all_gene data set (see TABLES 1,2); a) 412048, b) 412085, c) 412111, d) 412148, e) 412199, and f) 412438 (all haplotypes are shown in APPENDIX 3) using Fitch distances. Shown are the genealogical relationships between the haplotypes in the six putative gmyc species (green = CA2, grey = AZ1, orange = CA1, pink = EU3, purple = EU1, and red = EU2). Missing mutational steps connecting haplotypes are represented by non-colored dots. The size of the circles correlates with haplotype frequency within each network.

Species Assignment and Population Structure

The population assignments derived from the gmyc approach and the STRUCTURE analyses devided our specimens into six (gmyc) and five (STRUCTURE analyses) putative species (Fig. 3) on the Macaronesian islands. For the gmyc approach, we found 57 unique *cox1* haplotypes across all *Cloeon* sp. specimens. The gmyc model was significant (χ^2: 30.19, *P*<0.001) and delineated a total of 14 putative species, composed of ten distinct clusters and four singletons, whereby the 95% confidence intervals (CI as 2 log likelihood units) ranged from 14 to 17 species (all numbers including the damselfly outgroup; Fig. 3a). The gmyc approach delineated seven putative *C. dipterum* species, including one widespread gmyc species (North America, all Azorean islands and Greece; AZ1), three species occurring on the European mainland (EU1, EU2, EU3), one occurring on the Canary Islands and Madeira (CA1), one Canarian gmyc species (CA2), and one species in Asia (Korea). All seven different *C. dipterum* gmyc species were recognized by the gmyc model even when using the most conservative estimate (based on the lower 95% CI). The upper 95% CI detected ten *C. dipterum* species, splitting the European mainland species into five putative gmyc species. Additionally, two *C. smaeleni* species (including one from Brazil and Madagascar, and one from Saudi Arabia), one *C. praetextum* species (European mainland), two *C. simile* species (both European mainland), and one unknown (*Cloeon* sp.) from Saudi Arabia were recovered. The two *C. cognatum* specimens derived from the North American barcoding project (Webb et al. 2012) had identical *cox1* haplotypes as our Azorean and American specimens belonging to AZ1.

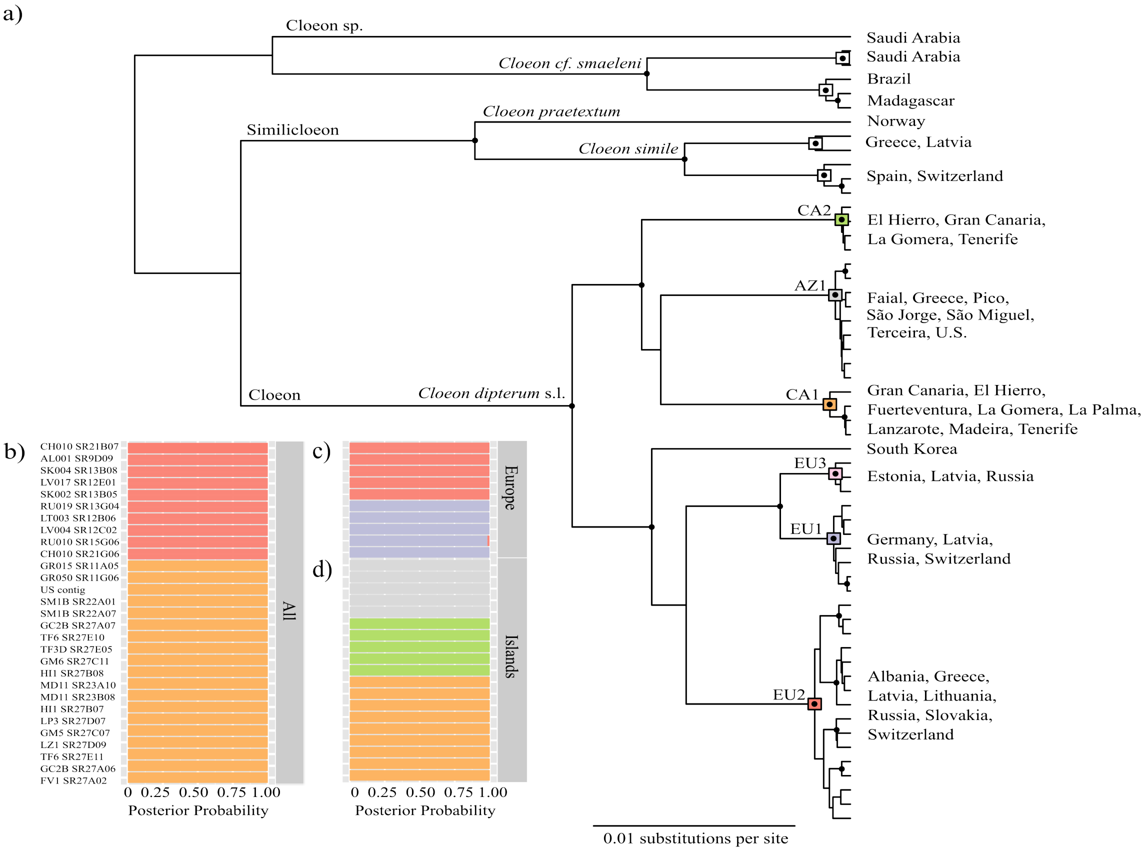


FIGURE 3. Species assignment from a) the gmyc approach and b), c), d) STRUCTURE analyses. For a), the mitochondrial *cox1* gene tree was used as input for the gmyc analysis of *Cloeon* sp.. Sequence clusters corresponding to single gmyc species are indicated by squares at subtending nodes. Colors indicate the six putative gmyc species (green = CA2, grey = AZ1, orange = CA1, pink = EU3, purple = EU1, and red = EU2). Terminal labels indicate sampling regions (see APPENDIX 2). Filled circles indicate well supported nodes (PP ≥ 0.95). For b), c), d) Results from the STRUCTURE analyses of the nuclear exon_all_gene data set. In all plots, horizontal bars represent an individual's assignment to a genotypic cluster with colors designating the different clusters (green = CA2, grey = AZ1, orange = CA1, purple = EU1, and red = EU2). For b) K = 2 STRUCTURE plot resulting from analysis of the full 29 specimens. For c) K = 2 STRUCTURE plot resulting from ten European specimens, whereby the individual RU010_SR15G06 (EU3) was detected as admixed. For d) K = 3 STRUCTURE plot for the 19 Island specimens (including two Greece and one U.S. specimens).

Replicate STRUCTURE analyses on the exon_all_gene data set produced a ΔK that favoured K = 2 (APPENDIX 4). All 29 individuals were assigned with PP ≈ 1.0 to one of these clusters, of which the first contained ten individuals from the European mainland, and the second 19 individuals from the Macaronesian archipelago, U.S. and Greece (Fig. 3b). However, all runs with K>2 evidenced the occurrence of more than two distinct genetic clusters (APPENDIX 4). Separate STRUCTURE analyses for these first two genetic clusters effectively resolved further levels of genetic structure. The analysis of the ten individuals from the European mainland produced a ΔK that favored K = 2 (Fig. 3c). Clustering was geographically specific. The two individuals RU010_SR15G06 and CH010_SR21B07 showed signs for admixture between two populations when K>2 (Fig. 3c; APPENDIX 4). Similarily, separate analyses of the second cluster, including Macaronesia, U.S. and Greece individuals supported K = 3, resulting in one genetic cluster including individuals geographically restricted to the Canaries, one with individuals from the Canary Islands and Madeira, and one comprising geographically widespread individuals from the Azores, the U.S. and Greece (Fig. 3d).

Phylogenetics

All phylogenetic trees based on the nDNA resulted in well-supported phylogenetic relationships within the *C. dipterum* species group. However, the European specimens, clustered differently, paraphyletic for the concatenated analyses conducted in MRBAYES (Fig. 4) and monophyletic based on the *BEAST analyses (Fig. 5). In contrast, the *cox1* gene tree supported the existence of six species but not the relationships among them (Fig. 3a). The analyses of the concatenated nDNA and the coalescent-based species tree reconstructions both resulted in very high PPs for the phylogenetic relationships among these species.

Concatenated Phylogenetics. – Both analyses based on the exon_all_gene and exon_all_taxa concatenated nDNA dataset showed the same tree topology, resolving the four *C. dipterum* species EU1, AZ1, CA2 and CA1 as well-supported monophyletic clades (Fig. 4; APPENDIX 5). The individuals belonging to EU2 were detected as paraphyletic group, whereby four out of five individuals clustered monophyletically. The species EU2 appeared as most ancestral. EU1 clustered as sister taxa to one clade, containing all Macaronesian specimens. The geographically widespread species AZ1 was detected as sister taxon to the two species CA1 and CA2. In general, the concatenated analysis based on the full_all_gene alignment well-resolved all nodes except between the two Azorean individuals whereas the full_all_taxa tree contained several polytomies and in general lower PPs especially within the CA1 species.

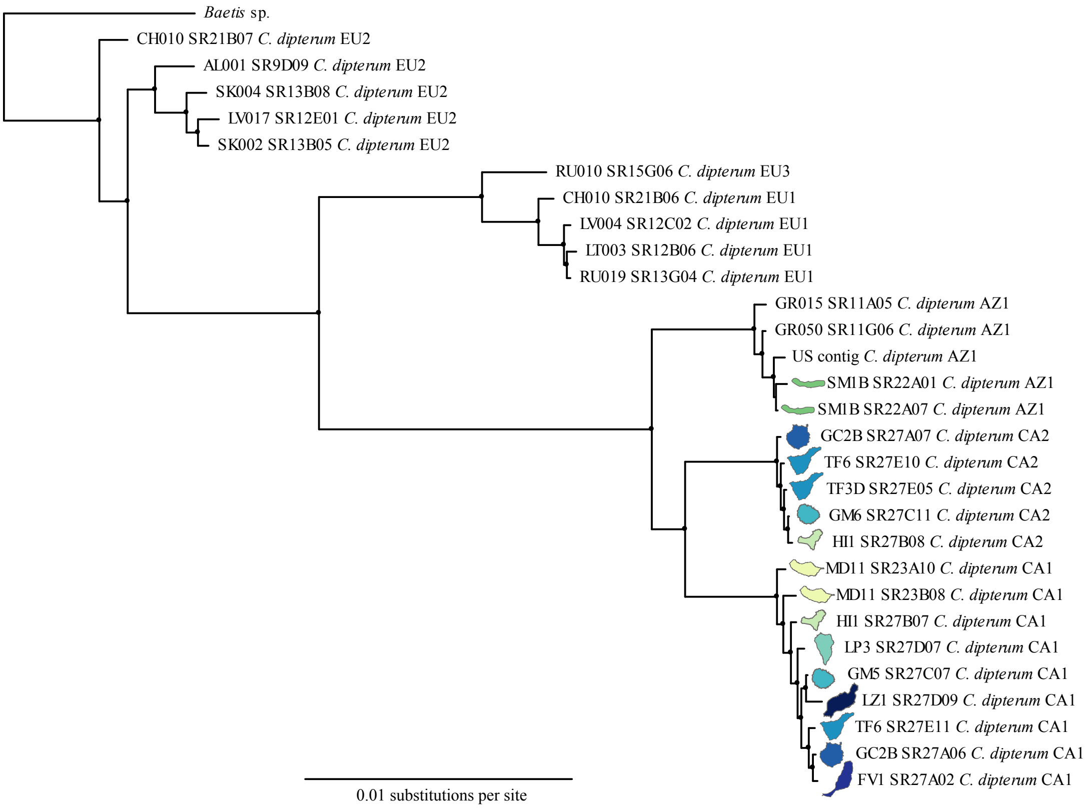

Baetis sp.
CH010 SR21B07 *C. dipterum* EU2
AL001 SR9D09 *C. dipterum* EU2
SK004 SR13B08 *C. dipterum* EU2
LV017 SR12E01 *C. dipterum* EU2
SK002 SR13B05 *C. dipterum* EU2
RU010 SR15G06 *C. dipterum* EU3
CH010 SR21B06 *C. dipterum* EU1
LV004 SR12C02 *C. dipterum* EU1
LT003 SR12B06 *C. dipterum* EU1
RU019 SR13G04 *C. dipterum* EU1
GR015 SR11A05 *C. dipterum* AZ1
GR050 SR11G06 *C. dipterum* AZ1
US contig *C. dipterum* AZ1
SM1B SR22A01 *C. dipterum* AZ1
SM1B SR22A07 *C. dipterum* AZ1
GC2B SR27A07 *C. dipterum* CA2
TF6 SR27E10 *C. dipterum* CA2
TF3D SR27E05 *C. dipterum* CA2
GM6 SR27C11 *C. dipterum* CA2
HI1 SR27B08 *C. dipterum* CA2
MD11 SR23A10 *C. dipterum* CA1
MD11 SR23B08 *C. dipterum* CA1
HI1 SR27B07 *C. dipterum* CA1
LP3 SR27D07 *C. dipterum* CA1
GM5 SR27C07 *C. dipterum* CA1
LZ1 SR27D09 *C. dipterum* CA1
TF6 SR27E11 *C. dipterum* CA1
GC2B SR27A06 *C. dipterum* CA1
FV1 SR27A02 *C. dipterum* CA1
0.01 substitutions per site

FIGURE 4. Bayesian inference reconstruction of the phylogenetic relationships among *Cloeon dipterum* s.l. based on the concatenated exon_all_gene data set using a separate substitution model for each nDNA marker (see TABLE 2). Filled circles indicate well supported nodes (PP ≥ 0.95). Shapes to the left of terminal labels indicate the origin of individuals (see FIGURE 1).

Multilocus Species Tree Phylogenetics. – The *BEAST runs resulted in identical tree topologies (Fig. 5). All nodes were highly supported (PP ≥ 95). The European species clustered well supported together (PP = 1 for EU1, EU2 and EU3 from the species assignment according gmyc species (Fig. 5a) and PP = 1 for EU1 and EU2 from the genotypic clusters of the STRUCTURE anlyses (Fig. 5b). For the three species on the Macaronesian region, AZ1 was detected as sister clade to the two Canarian species CA1 and CA2 (PP = 1).

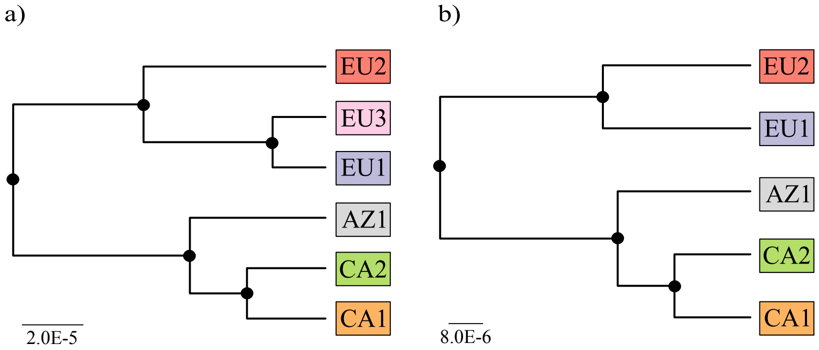


FIGURE 5. Species trees of *Cloeon dipterum* s.l. inferred with *BEAST based on the exonhap_all_gene data set (see TABLE 2) using a) the species assignment derived from the gmyc approach and b) the species assignment derived from the STRUCTURE analyses (see FIGURE 3). Filled circles indicate well supported nodes (PP ≥ 0.95).

DISCUSSION

A large multilocus dataset fully resolved the phylogenetic relationships of at least five previously poorly known *Cloeon* species, three of them occurring on the Macaronesian archipelago. We have shown that even for taxa with very limited available genomic resources, it is possible to develop sets of nuclear markers for both coalescent-based and concatenated-based phylogenetic approaches, producing well-supported phylogenies. Finally, we developed the bioinformatics program, DISCOMARK, that automatically and reliably designs target-specific primer pairs based on pervious whole genome and/or RNA-sequencing (RNA-seq) data sets. As a result, we can infer the colonization history and the species boundaries within the species complex *C. dipterum* including their respective genetic structure with higher confidence and resolution than before.

Accuracy of Phylogenetics

The pyhlogenetic relationships based on the nuclear dataset resulted in identical topologies for the specimens occuring on the Macaronesian region, applying different phylogenetic approches. The identical species-level relationships derived from the *cox1* (Fig. 3) and nuclear gene trees were unexpected since it has been shown for several taxa, that the use of mtDNA and nDNA result in contrasting phylogenetic tree topologies (e.g. Degnan and Rosenberg 2009; Hailer et al. 2012; Ruane et al. 2014). Although only very few nodes were supported for the mtDNA gene tree, the accordance of the gmyc approach and the nuclear data support the use of *cox1* as barcoding gene for the taxa studied for this work because individuals clustered together consistently fot mtDNA and nDNA markers. The distant clustering of the 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 EU2 with low frequency of different topology (Fig. 5). Moreover, the nesting within EU2 for the *cox1* tree and the outgroup to all other *Cloeon* for the nDNA trees evidence mitochondrial introgression. In contrast to our inferred phylogenies based on different amounts of genes, O'Neill et al. (2013) have shown that inferences based on 20 and 30 loci result in high PPs whereas the use of >50 loci lead to overall lower PPs and lower log-likelihood values, probably as result of the increasing number of parameters. Further studies investigating the quality of new markers, including comparisons between SNPs and gene based approaches, and testing the fit of markers for coalescent-based approaches (e.g. Reid et al. 2014) are required.

Haplotype networks further illustrate the necessity of using several individuals per species. For example, the individuals of CA1 share several haplotypes with other species (e.g. Fig. 2e), indicating incomplete lineage sorting between the different species, especially the co-ocurring or geographically close European and Macaronesian individuals. Originally, it was thought that *BEAST analyses might 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.

Marker Development

We used genomic sequence information obtained from one whole-genome library to infer species-level phylogenetic relationships of six *Cloeon* species. Generally, the screening of suitable markers and the availability of genomic information (e.g. reference taxon, sequence information) are crucial factors for the development of new markers. Advances in barcoding/indexing strategies for multiple libraries made parallel sequencing of both multiple individuals and markers within one single high-throughput sequencing run feasible.

The screening of suitable genes can be a quite time consuming process and highly depends on the available genomic resources (sequences and tools) and bioinformatics skills. The use of bioinformatics programs such as DISCOMARK will streamline this process by parsing orthologous sequence data, aligning orthologous sequence data from several data sets (different types of data, different taxa), and designing primer pairs at conservative positions of the data set. However, the program performs less efficient than the manual approach because the trimming step (step 3) is highly sensitive to sequence variability (e.g. occurrence of introns). Nevertheless, we have shown that the program produces accurate results for a smaller set of primer pairs. Moreover, all steps can be performed independently, producing output files that can be visually checked, and re-run with adjusted parameters without repeating the whole process.

Recent approaches targeted large numbers of conserved elements across a divergent range of taxa including individuals with no or limited genomic sequence information available (e.g. Faircloth et al. 2012; Lemmon et al. 2012). Target enrichment approaches can be used to specifically amplify known gene regions (Mamanova et al. 2010). Previously, a priori genomic knowledge was needed in order to develop new markers within closely related species (e.g. Nadeau et al. 2012; O'Neill et al. 2013). For organisms with no genomic knowledge, transcribed RNA-seq (e.g. Marioni et al. 2008; Hittinger et al. 2010; McCormack et al. 2013) or reduced sequencing technologies such as restriction-site associated DNA sequencing (RAD-seq; Baird et al. 2008; Peterson et al. 2012) have been widely applied. However, the challenges for these approaches are the large number of reads that cannot be assembled into contigs, which may be especially pronounced in large genomes, and that the number of loci in which the majority of individuals are represented can often be quite low (e.g. McCormack et al. 2013). Thus, the decision of using Sanger sequencing or high-throughput sequencing approaches depends on the research question (number of individuals, number of genes, length of genes, coverage of genes).

Species Delineation

The use of nDNA and a geographically extensive sampling pointed out largely underestimated species diversity for *C. dipterum* s.l., supporting the existence of five/six distinct species so far considered to be *C. dipterum* s.l.. The *cox1* gene tree showed a clear separation of the subgenera *Similicloeon* (*C. simile* and *C. praetextum*), *Cloeon* and

Cloeodes, supporting previous work by Monaghan et al. (2005). We propose that the five species AZ1, EU1, EU2, CA1, and CA2 should be formally recognized. The occurrence of a third species on the European mainland (EU3), reported from Eastern Europe (Estonia, Latvia, and Russia), has been evidenced by the gmyc approach, from admixed clustering for the STRUCTURE analyses, and increased phylogenetic distance based on branch length to all individuals of the EU1 species. However, our sampling was limited for Eastern Europe and thus the few sampled populations could cause an artificial splitting of clades, leading to an overestimate of species number (e.g. Lohse 2009; Papadopoulou et al. 2009). On the other side, Bergsten et al. (2012) have demonstrated that widespread allopatric speciation could lead to an underestimate of species when large-scale sampling uncovers more sister taxa than restricted sampling, which might be the case for the *C. dipterum* species from the European mainland. The gmyc analyses and genetic clustering did not identify the specimens of *C. peregrinator* as a separate species. Gattolliat et al. (2008) described the species based on mtDNA (cytochrome-oxidase *b* (*cob*)) sequences and morphological characters. However, they point out that the species is morphologically very similar to *C. dipterum* and at the time they described the species, there were no DNA sequences of Canarian *C. dipterum* specimens available. Rutschmann et al. (2014) also assigned all Madeiran individuals to *C. peregrinator* since two mtDNA gene trees (*cob* and *rrnL*) evidenced its monophyly but the specimens were not included in their gmyc analysis because the *cox1* sequences were missing. Based on this data, it seems that there is no endemic *Cloeon* species on Madeira; although the material examined here was collected at Ribeira do Alecrim and the species *C. peregrinator* is described from the Funchal, Jardim Botánico. The identical genetic clustering with the Canarian specimens, the similar morphology with European *Cloeon* (Gattolliat et al. 2008), and the same habitat preferences make its species status questionable. On the other hand it could be that the species *C. peregrinator* also occurs on the Canary Islands and then this name should be applied to CA1. This work clearly pointed out an unexpected genetic diversity within *C. dipterum*, which was treated in past studies as one species and thus this will impact future studies on freshwater biodiversity. The mitochondrial gene tree evidenced the occurrence of cryptic diversity within the subfamily Cloeoninae. Thereby, *C. simile* included two geographically widespread European gmyc species, and *C. smaeleni* LESTAGE 1924 two gmyc species with Saudi Arabian and Afrotropical (Madagascar and Brazil) origin, respectively. The species *C. praetexum* is clearly distinct from all other examined European specimens and therefore we propose its formal species recognition. The two specimens *C. cognatum* (North American barcoding project, Webb et al. 2012), which is thought to be a junior synonym of *C. dipterum* were nested within the AZ1 clade. However, all these evidences are preliminary because they are based entirely on mtDNA. Further studies on morphological characteristics, including comparisons with previously described species but nowadays listed junior synonyms or *species inquirenda* would be a valuable complement to the "molecular identification" presented here.

Species Origin and Colonization

Our data confirm three independent colonization events, all with European origin, for the Macaronesian *C. dipterum* specimens. The species from the European mainland (excluding two individuals from Greece) are clearly distinct from the ones on the Macaronesian archipelago, supported by the genetic clustering, gmyc analysis, and the phylogenetic approach. The close relationships between the European and Macaronesian clade demonstrate European origin for all Macaronesian *Cloeon*. Moreover also within the clade AZ1, the most ancestral specimens originate from Greece. However, the long phylogenetic branches between the European clades and Macaronesian clade indicate that the source populations, most likely from the Iberian Peninsula, are missing. Also, the European clades are relatively distantly related to each other and the longer phylogenetic branches within both clades, compared to the island clades, support the occurrence of additional European species as already shown by the gmyc approach and Structure analyses. Moreover, the two detected as admixed individuals appeared as sister taxa to all other individuals from the same species (RU010_SR15G06 and CH010_SR21B07). This contrasts several other studies that have proposed an African origin for both the Canarian as well as Madeiran fauna (e.g. Brunton and Hurst, 1998; Weingartner et al. 2006; Kvist et al. 2005).

For the species occurring in the Macaronesian region, it seems that one species is widely distributed on all Canary Islands and Madeira (CA1), one species occurs on the western Canarian islands (CA2), and one on all five islands of the Azores, Greece and U.S. (AZ1). The short phylogenetic branches and identical haplotypes of individuals from the Azorean islands, Greece and U.S. further support a very recent and therefore probably anthropogenic introduction of the species. Similarly, a recent introduction of a *Cloeon* representative has been reported from South America (*C. smaeleni*, Salles et al. 2014). Clearly, this long-distance dispersal ability has arisen from their abilities to survive in extreme habitats and their reproductive flexibility (parthenogenesis and ovivipary), which both among other factors made *Cloeon* a competitive early pioneer. However, this study demonstrates slight differences between the different species. CA2 seems not to have reached La Palma and the two most eastern Canarian islands of Fuerteventura and Lanzarote. The dispersal of CA2 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. In contrast, the species CA1 originated from Madeira and colonized the Canaries from west to east. Colonization routes between these two archipelagos have been suggested for several taxa (e.g. Emerson et al. 2000; Dimitrov et al. 2008; Illera et al. 2007; Trusty et al. 2005; Amorim et al. 2012). Although in most cases the colonization occurred from the Canary Islands towards Madeira and only few cases are known where it occurred the other way around.

Habitat Preference

Our data suggest a strong effect of different habitat preferences between the two Canarian species, which might impact their colonization success. Although we do acknowledge, that our dataset was not quantitative, it seems that the species CA2 occurred only on islands with more potential habitats in comparison to CA1, 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 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 comprise more species than standing ones. This pattern 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 not being restricted to one island. The low number of *Cloeon* species also has not promoted their diversification. The number of species is strongly related with the degree of diversification (Emerson and Kolm 2005), hence this might explain the low degree of diversification within *Cloeon* in comparison to *Baetis*. The impact of agriculture and tourism on natural habitats (Malmqvist et al. 1993; Malmqvist et al. 1995; Nilsson et al. 1998) has clearly threatened the occurrence of the lineages living in lotic habitats (*Baetis canariensis* and *B. pseudorhodani*, Rutschmann et al. 2014), but it seems to have less affected *C. dipterum*. Reports of mayflies from the other islands are sparse, but populations from natural habitats have been collected for this work from Fuerteventura and Lanazarote and previously also from La Palma (Müller-Liebenau 1971; Alba-Tercedor et al. 1987). So far, there are no records of mayflies from natural habitats from El Hierro, indicating a recent anthropogenic import of the species, moreover because of its remote geographical position.

Interestingly, there were eight sampling sites, where two species occurred sympatrically and half of these localities were natural aquatic 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 group. We expect to find more individuals from distinct geographic localities belonging to the species AZ1, since this species seems to exhibit long distance trans-oceanic dispersal abilities.

CONCLUSIONS

Our data demonstrate a European origin of the species, whereby the Macaronesian islands act as end points of three colonization processes (sensu Bellemain and Ricklefs 2008). However, the island populations might also act as source populations for further dispersal (e.g. the initial occurrence of CA1 on Madeira). The enormous dispersal abilities of *Cloeon* make its occurrence or if not present yet its dispersal to the North African mainland more than likely. The approach applied here is widely applicable to other taxa with no genome data available due to the development of the bioinformatics program DISCOMARK. Therefore it might be especially interesting for non-model organisms and is encouraging for future studies on non-model organisms.

SUPPLEMENTARY MATERIAL

Supplementary material, including data files and/or online-only appendices, can be found in the Dryad Digital Respository at http://dx.doi.org/10.5061/dryad.[NNNN] and TreeBASE data repository at http://purl.org/phylo/treebase/phylows/study/NNNN.

FUNDING

This work was supported by the Leibniz Association (PAKT für Forschung und Innovation), and an award from the Leibniz-Institute of Freshwater Ecology and Inland Fisheries to SR for partially funding her visit at the Sackler Institute for Comparative Genomics, American Museum of Natural History, New York, USA.

ACKNOWLEDGEMENTS

We are grateful to Berta Ortiz Crespo, Susan Mbedi, Katrin Preuß, and Lydia Wächter for excellent laboratory work, Samantha J. Hughes, David Murányi, André Wagner, Tomas Ruginis, and Andrey Przhiboro for contributing specimens, Peter Rutschmann and the HPC Service of ZEDAT, Freie Universität Berlin, Germany for access to high-performance computing resources, and Jasmine Smith for assistance with the sequence submission to NCBI. We are thankful to Jakob Fredslund for providing the Python source code of PRIFI. We are greatly indebted to Marcos Báez for his help to obtain all the sampling permits from the Canary Islands and the authorities who provided us with the appropriate permissions to collect the material: Servicio Administrativo de Medio Ambiente, Cabildo de Tenerife (reg. number 2014-00200/2014), Ministerio de Medio Ambiente, Parque Nacional de Guarajonay, La Gomera (reg. number 106051-17099/2014), Consejeria 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. Thanks to many anonymous people from the Canary Islands who gave helpful recommendations about the sampling sites. SS would like to thank the German Academic Exchange Service (DAAD) and the Biosystematics Group at Wageningen University for support. We are very grateful to current and past members of the Monaghan research group, especially to Ignacio Lucas Lledó and Maribet Gamboa, all participants of the weekly meeting on "Evolutionary Biology" at the IGB, and the BeGenDiv team for the constructive comments on this work and the familiar working atmosphere.

REFERENCES

Alba-Tercedor J, Báez M, Soldán T. 1987. New records of mayflies of the Canary Islands (Insecta, Ephemeroptera). Eos. 63:7-13.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389-3402.

Amorim IR, Emerson BC, Borges PAV, Wayne RK. 2012. Phylogeography and molecular phylogeny of Macaronesian island *Tarphius* (Coleoptera: Zopheridae): why are there so few species in the Azores? Journal of Biogeography. 39:1583-1595.

Arnedo MA, Oromi P, De Abreu SM, 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. Systematic Entomology. 33:145-158.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA. 2008. Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. PLoS One. 3:e3376.

Barber-James H, Sartori M, Gattolliat J-L, Webb J. 2013. World checklist of freshwater Ephemeroptera species. http://fada.biodiversity.be/group/show/35.

Barber-James HM, Gattolliat J-L, Sartori M, Hubbard MD. 2007. Global diversity of mayflies (Ephemeroptera, Insecta) in freshwater. Hydrobiologia. 595:339-350.

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

Bellemain E, Ricklefs RE. 2008. Are islands the end of the colonization road? Trends Ecol Evol. 23:461-468.

Bergsten J, Bilton DT, Fujisawa T, Elliott M, Monaghan MT, Balke M, Hendrich L, Geijer J, Herrmann J, Foster GN*, et al.* 2012. The effect of geographical scale of sampling on DNA barcoding. Syst Biol. 61:851-869.

Borges PAV, Costa A, Cunha R, Gabriel R, Gonçalves V, Martins AF, Melo I, Parente M, Raposeiro P, Rodrigues P*, et al.* 2010. A list of the terrestrial and marine biota from the Azores.

Bouckaert R, Heled J, Kuhnert D, Vaughan T, Wu CH, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput 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 CFA, Hurst GDD. 1998. Mitochondrial DNA phylogeny of Brimstone butterflies (genus *Gonepteryx*) from the Canary Islands and Madeira. Biol J Linn Soc Lond. 63:69-79.

Burks BD. 1953. The mayflies, or Ephemeroptera of Illinois. Bulletin of the Illinois Natural History Survey. 26:1-216.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 25:1972-1973.

Cianciara S. 1979. Life cycles of *Cloeon dipterum* (L.) in natural environment. Pol Arch Hydrobiol. 4:501-513.

Cianciara S. 1980. Stages and physiological periods in the development of Cloeon dipterum (L.) (Baetidae). Advances in Ephemeropterean Biology. 265-276.

Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods. 9:772.

Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol Evol. 24:332-340.

Degnan JH, Rosenberg NA, 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.

Degrange C. 1960. Recherches sur la reproduction des Ephéméroptères. Travaux Laboratoire de Pisciculture de l'Université de Grenoble. 50-51:7-193.

Díaz-Pérez AJ, 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 KD, Monaghan MT, Pauls SU. 2014. Freshwater biodiversity and aquatic insect diversification. Annu Rev Entomol. 59:143-163.

Dimitrov D, Arnedo MA, 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 MK. 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 AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol. 7:214.

Earl DA, vonHoldt BM. 2011. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genetics Resources. 4:359-361.

Ebersberger I, Strauss S, von Haeseler A. 2009. HaMStR: profile hidden markov model based search for orthologs in ESTs. BMC Evol Biol. 9:157.

Edwards SV. 2009. Is a new and general theory of molecular systematics emerging? Evolution. 63:1-19.

Emerson BC. 2002. Evolution on oceanic islands: molecular phylogenetic approaches to understanding pattern and process. Mol Ecol. 11:951-966.

Emerson BC, Kolm N. 2005. Species diversity can drive speciation. Nature. 434:1015-1017.

Emerson BC, Oromi 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 BC, Oromí P, Godfrey MH. 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 BC, Oromi P, Hewitt GM. 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.

Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol. 14:2611-2620.

Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC. 2012. Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary Timescales. Systematic Biology. 61:717-726.

Falush D, Stephens M, Pritchard JK. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics. 164:1567-1587.

Fitch WM. 1970. Distinguishing homologous from analogous proteins. Syst Zool. 19:99-113.

Flot JF. 2010. seqphase: a web tool for interconverting phase input/output files and fasta sequence alignments. Mol Ecol Resour. 10:162-166.

Fredslund J, Schauser L, Madsen LH, Sandal N, Stougaard J. 2005. PriFi: using a multiple alignment of related sequences to find primers for amplification of homologs. Nucleic Acids Res. 33:W516-520.

Fujisawa T, Barraclough TG. 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 RC, Sunnucks P, Dyer RJ. 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, Hughes SJ, Monaghan MT, Sartori M. 2008a. Revision of Madeiran mayflies (Insecta, Ephemeroptera). ZOOTAXA, p. 1-17.

Gattolliat JL. 2004. First reports of the genus *Nigrobaetis* Novikova & Kluge (Ephemeroptera : Baetidae) from Madagascar and La Reunion with observations on afrotropical biogeography. Revue Suisse de Zoologie. 111:657-669.

Gattolliat JL, Hughes SJ, Monaghan MT, Sartori M. 2008b. Revision of Madeiran mayflies (Insecta, Ephemeroptera). ZOOTAXA. 52-68.

Gattolliat JL, Staniczek A. 2011. New larvae of Baetidae (Insecta: Ephemeroptera) from Espiritu Santo, Vanuatu. Stuttgarter Beiträge zur Naturkunde A, Neue Serie. 4:75-82.

Gillies MT. 1949. Notes on some Ephemeroptera Baetidae from India and South-East Asia. Trans R Entomol Soc Lond. 161-177.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 52:696-704.

Hailer F, Kutschera VE, Hallstrom BM, Klassert D, Fain SR, Leonard JA, Arnason U, Janke A. 2012. Nuclear genomic sequences reveal that polar bears are an old and distinct bear lineage. Science. 336:344-347.

Harker JE. 1997. The role of parthenogenesis in the biology of two species of mayfly (Ephemeroptera). Freshwater Biology. 37:287-297.

Harrigan RJ, Mazza ME, Sorenson MD. 2008. Computation vs. cloning: evaluation of two methods for haplotype determination. Mol Ecol Resour. 8:1239-1248.

Heled J, Bryant D, Drummond AJ. 2013. Simulating gene trees under the multispecies coalescent and time-dependent migration. BMC Evol Biol. 13:44.

Heled J, Drummond AJ. 2010. Bayesian inference of species trees from multilocus data. Mol Biol Evol. 27:570-580.

Hittinger CT, Johnston M, Tossberg JT, Rokas A. 2010. Leveraging skewed transcript abundance by RNA-Seq to increase the genomic depth of the tree of life. Proc Natl Acad Sci U S A. 107:1476-1481.

Illera JC, Emerson BC, Richardson DS. 2007. Population history of Berthelot's pipit: colonization, gene flow and morphological divergence in Macaronesia. Mol Ecol. 16:4599-4612.

Jordal BH, Hewitt GM. 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 GM. 1997. Molecular phylogeny of darkling beetles from the Canary Islands: comparison of inter island colonization patterns in two genera. Biochemical Systematics and Ecology. 25:121-130.

Juan II, Emerson BC, Orom II, Hewitt GM. 2000. Colonization and diversification: towards a phylogeographic synthesis for the Canary Islands. Trends Ecol Evol. 15:104-109.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30:772-780.

Knowles LL, Kubatko LS. 2010. Estimating species trees: an introduction to concepts and models. 1-14.

Kubatko LS, Degnan JH. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst Biol. 56:17-24.

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

Leaché AD, Fujita MK, Minin VN, Bouckaert RR. 2014. Species delimitation using genome-wide SNP data. Syst Biol. 63:534-542.

Lemmon AR, Emme SA, Lemmon EM. 2012. Anchored Hybrid Enrichment for Massively High-Throughput Phylogenomics. Syst Biol. 61:727-744.

Lohse K. 2009. Can mtDNA barcodes be used to delimit species? A response to Pons et al.(2006). Syst Biol. 58:439-442.

Maddison WP. 1997. Gene trees in species trees. Syst Biol. 46:523-536.

Maddison WP, Maddison DR. 2011. Mesquite: a modular system for evolutionary analysis. Version 2.75. http://mesquiteproject.org.

Malmqvist B, Nilsson AN, B áez M, Armitage PD, Blackburn J. 1993. Stream macroinvertebrate communities in the island of Tenerife. Archiv für Hydrobiologie. 128:209-235.

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

Malmqvist B, Nilsson AN, Báez M. 1995b. Tenerife's freshwater macroinvertebrates: status and threats (Canary Islands, Spain). Aquatic Conservation-Marine and Freshwater Ecosystems. 5:1-24.

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

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 18:1509-1517.

Matschiner M. 2014. http://www.evoinformatics.eu/fitchi.

McCafferty WP, Jacobus LM, Webb JM, Meyer MD. 2008. Insecta, Ephemeroptera: range extensions and new records for Ontario and Canada. Check List, p. 445-448.

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT. 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. Mol Phylogenet Evol. 66:526-538.

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

Müller-Liebenau I. 1971. Ephemeroptera (Insecta) von den Kanarischen Inseln. Gewässer und Abwässer. 50/51:7-40.

Nadeau NJ, Whibley A, Jones RT, Davey JW, Dasmahapatra KK, Baxter SW, Quail MA, Joron M, ffrench-Constant RH, Blaxter ML*, et al.* 2012. Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. Philos Trans R Soc Lond B Biol Sci. 367:343-353.

Nagell B. 1977. Survival of Cloeon dipterum (Ephemeroptera) larvae under anoxic conditions in winter. Oikos. 29:161-165.

Nagell B. 1980. Overwintering strategy of Cloeon dipterum (L.) larvae. In: Flannigan JF, Marshall KE editors. Advances in Ephemeroptera Biology. New York and London, Plenum Press.

Nagell B. 1981. Overwintering strategy of two closely related forms of *Cloeon* (*dipterum*?) (Ephemeroptera) from Sweden and England. Freshwater Biology. 11:237-244.

Nakhleh L. 2013. Computational approaches to species phylogeny inference and gene tree reconciliation. Trends Ecol Evol. 28:719-728.

Nilsson AN, Malmqvist B, Báez M, Blackburn JH, Armitage PD. 1998. Stream insects and gastropods in the island of Gran Canaria (Spain). Annales De Limnologie-International Journal of Limnology. 34:413-435.

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

O'Neill EM, Schwartz R, Bullock CT, Williams JS, Shaffer HB, Aguilar-Miguel X, Parra-Olea G, Weisrock DW. 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.

Oehme G. 1972. Zur maximualen Lebensdauer von Cloeon dipterum L. (Eph. Baetidae). Entomologische nachrichten. 16:131-133.

Papadopoulou A, Monaghan MT, Barraclough TG, Vogler AP. 2009. Sampling Error Does Not Invalidate the Yule-Coalescent Model for Species Delimitation. A Response to Lohse (2009). Syst Biol. 58:442-444.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. 2012. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS ONE. 7:e37135.

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

Randolph RP, McCafferty WP, Zaranko D, Jacobus LM, Webb JM. 2003. New Canadian records of Baetidae (Ephemeroptera) and adjustments to North American *Cloeon*. Entomological News. 113:306-308.

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 PM, Cruz AM, Hughes SJ, Costa AC. 2012. Azorean freshwater invertebrates: Status, threats and biogeographic notes. Limnetica. 31:13-22.

Reid NM, Hird SM, Brown JM, Pelletier TA, McVay JD, Satler JD, Carstens BC. 2014. Poor fit to the multispecies coalescent is widely detectable in empirical data. Syst Biol. 63:322-333.

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

Ribera I, Bilton DT, Vogler AP. 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 GN, Vogler AP. 2003c. Does habitat use explain large scale species richness patterns of aquatic beetles in Europe? Ecography. 26:145-152.

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

Ruane S, Bryson RW, Jr., Pyron RA, Burbrink FT. 2014. Coalescent species delimitation in milksnakes (genus *Lampropeltis*) and impacts on phylogenetic comparative analyses. Syst Biol. 63:231-250.

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

Salles FF, Gattolliat JL, Angeli KB, De-Souza MR, Goncalves IC, Nessimian JL, Sartori M. 2014. Discovery of an alien species of mayfly in South America (Ephemeroptera). ZooKeys. 1-16.

Salzburger W, Ewing GB, Von Haeseler A. 2011. The performance of phylogenetic algorithms in estimating haplotype genealogies with migration. Mol Ecol. 20:1952-1963.

Sartori M. 2001. Current knowledge of mayfly research in Europe (Ephemeroptera). In: Dominguez E editor. Trends in Research in Ephemeroptera and Plecoptera Kluwer Academic/Plenum Publishers.

Silina AE. 1993. Ecological peculiarities of the sympatric species of mayflies, *Cloeon dipterum* L. and *C. inscriptum* Btss. (Ephemeroptera, Baetidae)*. Entomological Review. 72:776-781.

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. 1987. Adaptation of the subimaginal life span of Cloeon (Ephemeroptera, Baetidae) in the arid areas of North Africa and the Canary Islands. Acta Entomologica Bohemoslov. 84:62-65.

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

Soldán T, Zahrádková S. 2000. Ephemeroptera of the Czech Republic: Atlas of distribution. In: Helesic J, Zahrádková S editors. Fauna Aquatica Europea Centralis. Masaryk University Brno, Biodiversity Working Group.

Sowa R. 1975. What is *Cloeon dipterum* (Linneaeus, 1716)? The nomenclatural and morphological analysis of a group of the European species of *Cloeon* Leach (Ephemerida: Baetidae). Ent scand. 6:215-223.

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

Stauder A. 1991. Water fauna of a Madeiran stream with notes on the zoogeography of the Macaronesian islands. Boletim do Museu Municipal do Funchal. 43:243-299.

Stauder A. 1995. Survey of the Madeiran limnological fauna and their zoogeogrpahical distribution. Boletim do Museu Municipal do Funchal. Sup. no. 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 NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet. 68:978-989.

Traver JR. 1962. *Cloeon dipterum* (L.) in Ohio (Ephemeroptera: Baetidae). Bulletin of the Brooklyn Entomological Society. 57:47-50.

Trusty JL, Olmstead RG, 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.

Vuataz L, Sartori M, Gattolliat JL, Monaghan MT. 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 JM, Jacobus LM, Funk DH, Zhou X, Kondratieff B, Geraci CJ, DeWalt RE, Baird DJ, Richard B, Phillips I*, et al.* 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. Systematic Entomology. 31:621-632.

Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL. 2012. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics. 13:134.

APPENDICES

APPENDIX 1. Sampling localities of included specimens with corresponding region, sampling site description, latitude, longitude, collector and sampling date

Locality ID	Region	Sampling Site Description	Latitude	Longitude	Collector(s)	Date of Sampling
FA1	Azores, Faial	Near Salão village, small trough	38.6031	-28.6329	P. Raposeiro	19.03.2014
FA4	Azores, Faial	Flamengos, small trough	38.5590	-28.6595	P. Raposeiro	19.03.2014
PI1	Azores, Pico	Paul Lake	38.4281	-28.2331	P. Raposeiro	17.03.2014
PI2	Azores, Pico	Planalto Central, small tank	38.4353	-28.1306	P. Raposeiro	17.03.2014
PI3	Azores, Pico	Planalto Central, small trough	38.4728	-28.3466	P. Raposeiro	17.03.2014
PI4	Azores, Pico	Cabeço Chão, small trough	38.5374	-28.4797	P. Raposeiro	17.03.2014
SJ1	Azores, São Jorge	Sete Fontes, artificial pond	38.7319	-28.2673	P. Raposeiro	18.03.2014
SJ2	Azores, São Jorge	Planalto Central, small tank	38.6581	-28.1119	P. Raposeiro	18.03.2014
SJ3	Azores, São Jorge	Fajã dos Cubres, small tank	38.6397	-27.9644	P. Raposeiro	18.03.2014
SJ4	Azores, São Jorge	Pico do Carvão, small trough	38.6714	-28.0962	P. Raposeiro	18.03.2014
SJ5	Azores, São Jorge	Near Ponta dos Rosais, small trough	38.7437	-28.2849	P. Raposeiro	18.03.2014
SM10	Azores, São Miguel	Near Furnas, small tank	37.7839	-25.3543	P. Raposeiro	03.03.2014
SM1A	Azores, São Miguel	University of the Azores, Campus, pond	37.7460	-25.6643	P. Raposeiro	23.02.2006
SM1B	Azores, São Miguel	University of the Azores, Campus, pond	37.7460	-25.6643	P. Raposeiro	02.2013
SM2	Azores, São Miguel	Near Rasa Lake (Sete Cidades), small tank	37.8381	-25.7815	P. Raposeiro	02.2014
SM3	Azores, São Miguel	Sete Cidades, Verde Lake	37.8489	-25.7850	P. Raposeiro	03.03.2014
SM4	Azores, São Miguel	Near São Brás lake, small trough	37.7910	-25.4293	P. Raposeiro	03.03.2014
SM5	Azores, São Miguel	Near São Brás lake, small pond	37.7910	-25.4293	P. Raposeiro	03.03.2014
SM7	Azores, São Miguel	Near Lomba da Maia, small trough	37.8046	-25.3692	P. Raposeiro	03.03.2014
SM8	Azores, São Miguel	Near Ponta Garca, small tank	37.7402	-25.3798	P. Raposeiro	03.03.2014
SM9	Azores, São Miguel	Near Furnas Golf, small trough	37.7778	-25.3601	P. Raposeiro	03.03.2014
TE1	Azores, Terceira	Serra do Cume	38.7014	-27.1027	P. Raposeiro	02.2014
MD11	Madeira, Madeira	Rabaçal, Ribeira do Alecrim	32.7533	-17.1291	S. J. Hughes & L. F. Peres Braz	16.11.2013
FV1	Canary Islands, Fuerteventura	Betancuria, water reservoir	28.4215	-14.0587	S. Rutschmann & H. Detering	16.03.2014
GC1	Canary Islands, Gran Canaria	Telde, Barranco de Los Cernícalos	27.9650	-15.4961	M. Sartori & M. Báez	25.01.2009
GC2B	Canary Islands, Gran Canaria	Las Lagunetas, Barranco de La Mina	28.0007	-15.5850	S. Rutschmann & H. Detering	14.03.2014
GC7	Canary Islands, Gran Canaria	Moya, water reservoir	28.1050	-15.5841	S. Rutschmann & H. Detering	18.03.2014
HI1	Canary Islands, El Hierro	Valvedere, Embalse de Terifabe, water reservoir	27.8043	-17.9242	S. Rutschmann & H. Detering	20/21.03.2014
HI2	Canary Islands, El Hierro	La Frontera, water reservoir in former channel	27.7596	-17.9995	S. Rutschmann & H. Detering	20.03.2014

GM4	Canary Islands, La Gomera	Las Rosas, Barranco de las Rosas	28.1867	-17.2197	S. Rutschmann & H. Detering	07.03.2014
GM5	Canary Islands, La Gomera	Arure, Barranco de Arure	28.1329	-17.3202	S. Rutschmann & H. Detering	08.03.2014
GM6	Canary Islands, La Gomera	El Guro, Barranco de Arure	28.1068	-17.3260	S. Rutschmann & H. Detering	08.03.2014
LP3	Canary Islands, La Palma	Cueva del Agua, water reservoir	28.8117	-17.9580	S. Rutschmann & H. Detering	11.03.2014
LZ1	Canary Islands, Lanzarote	Mala, Barranco Valle del Palomo, Presa de la Mala	29.1083	-13.4760	S. Rutschmann & H. Detering	17.03.2014
TF1	Canary Islands, Tenerife	Afur, Barranco de Afur	28.5550	-16.2506	M. Báez	04.04.2007
TF1B	Canary Islands, Tenerife	Afur, Barranco de Afur	28.5544	-16.2498	S. Rutschmann & H. Detering	13.03.2014
TF2	Canary Islands, Tenerife	San Andrés, Barranco de Igueste	28.5397	-16.1581	M. Sartori & M. Báez	18.03.2007
TF3B	Canary Islands, Tenerife	Adeje, Barranco del Infierno	28.1330	-16.7108	S. Rutschmann, H. Detering & M. Báez	06/22.03.2014
TF3D	Canary Islands, Tenerife	Adeje, Barranco del Infierno	28.1334	-16.7052	S. Rutschmann, H. Detering & M. Báez	06/22.03.2014
TF4B	Canary Islands, Tenerife	San Andrés, Barranco de Igueste	28.5380	-16.1560	S. Rutschmann & H. Detering	13.03.2014
TF6	Canary Islands, Tenerife	Vilaflor, Barranco del Río	28.1930	-16.5722	S. Rutschmann & H. Detering	12.03.2014
TF7	Canary Islands, Tenerife	Vilaflor, water reservoir	28.1287	-16.6551	S. Rutschmann & H. Detering	12.03.2014
AL001	Albania	Bajzë, Syri i Sheganit Spring	42.2727	19.3960	Z. Fehér, T. Kovács & D. Murányi	17.06.2012
CH010	Switzerland	Kleinandelfingen, Räubrichsee	47.6129	8.6764	V. Lubini	15.05.2012
CH041*	Switzerland	L'Abbaye, Lac de Joux	46.6427	6.3075	A. Wagner	30.04.2013
ES012*	Spain	near la Farga, Ríu de Nuria	42.3640	2.1743	M. Alp & V. Acuna	12.06.2012
EE003	Spain	near Metsküla, Nõrga oja	58.4079	25.4153	S. Rutschmann, M.F. Geiger & K. Kurzrock	10.09.2011
ES012*	Spain	near la Farga, Ríu de Nuria	42.3640	2.1743	M. Alp & V. Acuna	12.06.2012
GR014*	Greece	Koma Village, Spercheios River	38.8549	22.4577	S. Rutschmann, M.F. Geiger & K.C. Gritzalis	24.09.2011
GR015	Greece	Kalipefki Village, main artificial channel of Kalipefki Village	39.9517	22.4563	S. Rutschmann, M.F. Geiger & K.C. Gritzalis	24.09.2011
GR016	Greece	Karia Village, Skamnias River	39.9617	22.3828	S. Rutschmann, M.F. Geiger & K.C. Gritzalis	24.09.2011
GR018	Greece	Mouries Village, Doirani Lake	41.2412	22.7667	S. Rutschmann, M.F. Geiger & K.C. Gritzalis	25.09.2011
GR020	Greece	Mandraki Village, Strymon River	41.2561	23.1401	S. Rutschmann, M.F. Geiger & K.C. Gritzalis	25.09.2011
GR032	Greece	Amfithea Village, Lake Pamvotis	39.6804	20.8898	S. Rutschmann, M.F. Geiger & K.C. Gritzalis	28.09.2011
GR033	Greece	Perama Village, Lake Pamvotis	39.6903	20.8495	S. Rutschmann, M.F. Geiger & K.C. Gritzalis	28.09.2011
GR050	Greece	Psari Village, Koprinitsa Springs	37.3082	21.8659	S. Rutschmann, M.F. Geiger &	02.10.2011

					K.C. Gritzalis	
LT003	Lithuania	near Kaimynai, pond	54.9367	24.9068	S. Rutschmann, M.F. Geiger & K. Kurzrock	11.09.2011
LT019	Lithuania	Klaipėda, Danė	55.7157	21.1613	T. Ruginis	27.08.2013
LV001	Latvia	between boarder and Nidasciems, small ditch	56.0807	21.1135	S. Rutschmann, M.F. Geiger & K. Kurzrock	05.09.2011
LV002	Latvia	between Pape and Rucava, Paupes Kanals	56.1547	21.0946	S. Rutschmann, M.F. Geiger & K. Kurzrock	05.09.2011
LV003*	Latvia	near Papesciems, Paurupes Kanals	56.1508	21.0314	S. Rutschmann, M.F. Geiger & K. Kurzrock	05.09.2011
LV004	Latvia	near Durbe, pond	56.5956	21.3609	S. Rutschmann, M.F. Geiger & K. Kurzrock	06.09.2011
LV010	Latvia	near Darzini, Daugava River	56.8587	24.2854	S. Rutschmann, M.F. Geiger & K. Kurzrock	07.09.2011
LV011	Latvia	near Darzini, Daugava Kanal	56.8575	24.2926	S. Rutschmann, M.F. Geiger & K. Kurzrock	07.09.2011
LV016	Latvia	between Salas and Lapmežciems, Akacis Lake	56.9422	23.5549	S. Rutschmann, M.F. Geiger & K. Kurzrock	08.09.2011
LV017	Latvia	between Salas and Lapmežciems, Ozero Sloka Lake	56.9575	23.5466	S. Rutschmann, M.F. Geiger & K. Kurzrock	08.09.2011
LV019	Latvia	between Lapmežciems and Antinciems, Kanieris Lake	56.9977	23.4762	S. Rutschmann, M.F. Geiger & K. Kurzrock	08.09.2011
RU010	Russia	Sankt-Peterburg, Lordanskii Pond	59.9936	30.3365	A. Przhiboro	27.05.2012
RU019	Russia	Osinovaya Roshcha, Sankt-Petersburg, Lake Glukhoe	60.1201	30.2590	A. Przhiboro	01.07.2012
RU020	Russia	Osinovaya Roshcha, Sankt-Petersburg, Lake Srednee	60.1138	30.2528	A. Przhiboro	01.07.2012
RU038	Russia	village Abdulgazino, Bol´shoy Kizil River	53.3487	58.3213	not known	25.08.2012
RU046	Russia	Polazna, Floodplain Lake of Polazna River	58.2852	56.4567	not known	29.10.2012
SK002	Slovakia	Kucany, Starý Laborec	48.5307	21.8638	S. Rutschmann, P. Manko & K. Kurzrock	15.08.2011
SK004	Slovakia	Somotor, Somotorský Kanál	48.3970	21.8079	S. Rutschmann, P. Manko & K. Kurzrock	16.08.2011

Notes: *, sampling localities of *Cloeon simile*

APPENDIX 2. Included specimens with GenBank. Acc. no. For the *cox1* sequence, gmyc species assignment, geographic region, locality id, and voucher number

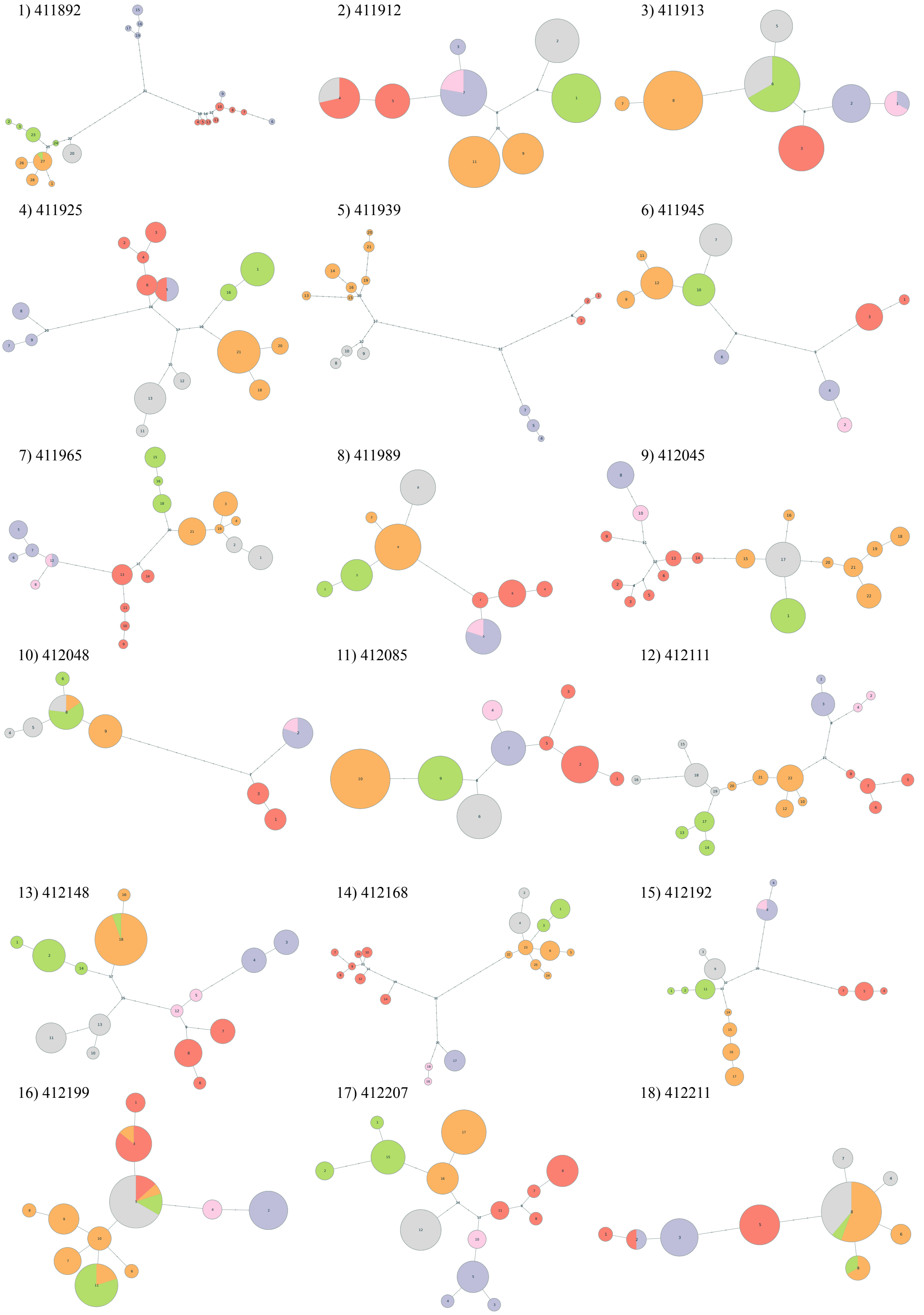
GI	Species	Gmyc Species	Region	Locality ID	Voucher	Reference
XXXXXXXX	***C. dipterum***	**EU2**	**Albania**	**AL001**	**SR9D09**	**this study**
KJ631625	***C. dipterum***	**EU1**	**Switzerland**	**CH010**	**SR21B06**	**Rutschmann et al. 2014**
KJ631626	***C. dipterum***	**EU2**	**Switzerland**	**CH010**	**SR21B07**	**Rutschmann et al. 2014**
XXXXXXXX	*C. dipterum*	EU3	Estonia	EE003	SR9H03	this study
XXXXXXXX	*C. dipterum*	AZ1	Azores, Faial	FA1	SR27F04	this study
XXXXXXXX	*C. dipterum*	AZ1	Azores, Faial	FA1	SR27F05	this study
XXXXXXXX	*C. dipterum*	AZ1	Azores, Faial	FA4	SR27F08	this study
XXXXXXXX	***C. dipterum***	**CA1**	**Canary Islands, Fuerteventura**	**FV1**	**SR27A02**	**this study**
KF438141	*C. dipterum**	CA1	Canary Islands, Gran Canaria	GC1	100046	Rutschmann et al. 2014
KF438144	*C. dipterum**	CA1	Canary Islands, Gran Canaria	GC1	250010	Rutschmann et al. 2014
XXXXXXXX	***C. dipterum***	**CA1**	**Canary Islands, Gran Canaria**	**GC2B**	**SR27A06**	**this study**
XXXXXXXX	***C. dipterum***	**CA2**	**Canary Islands, Gran Canaria**	**GC2B**	**SR27A07**	**this study**
XXXXXXXX	*C. dipterum*	CA1	Canary Islands, Gran Canaria	GC7	SR27B05	this study
XXXXXXXX	*C. dipterum*	CA1	Canary Islands, Gran Canaria	GC7	SR27B06	this study
XXXXXXXX	*C. dipterum*	CA1	Canary Islands, Gran Canaria	GC7	SR26B10	this study
XXXXXXXX	*C. dipterum*	CA1	Canary Islands, Gran Canaria	GC7	SR26B11	this study
KF438134	*C. dipterum*	EU1	Germany	DE	ZSM00215	Rutschmann et al. 2014
XXXXXXXX	*C. dipterum*	CA1	Canary Islands, La Gomera	GM4	SR27C05	this study
XXXXXXXX	*C. dipterum*	CA1	Canary Islands, La Gomera	GM5	SR27C06	this study
XXXXXXXX	***C. dipterum***	**CA1**	**Canary Islands, La Gomera**	**GM5**	**SR27C07**	**this study**
XXXXXXXX	*C. dipterum*	CA1	Canary Islands, La Gomera	GM5	SR26E02	this study
XXXXXXXX	*C. dipterum*	CA1	Canary Islands, La Gomera	GM5	SR26E03	this study
XXXXXXXX	***C. dipterum***	**CA2**	**Canary Islands, La Gomera**	**GM6**	**SR27C11**	**this study**
XXXXXXXX	***C. dipterum***	**AZ1**	**Greece**	**GR015**	**SR11A05**	**this study**
XXXXXXXX	*C. dipterum*	EU2	Greece	GR016	SR11A06	this study
XXXXXXXX	*C. dipterum*	EU2	Greece	GR018	SR11A08	this study
XXXXXXXX	*C. dipterum*	EU2	Greece	GR020	SR11A11	this study
XXXXXXXX	*C. dipterum*	EU2	Greece	GR020	SR11A12	this study
XXXXXXXX	*C. dipterum*	EU2	Greece	GR032	SR11D03	this study
XXXXXXXX	*C. dipterum*	EU2	Greece	GR033	SR11D05	this study
XXXXXXXX	***C. dipterum***	**AZ1**	**Greece**	**GR050**	**SR11G06**	**this study**
XXXXXXXX	*C. dipterum*	EU2	Greece	GR050	SR11G07	this study
XXXXXXXX	***C. dipterum***	**CA1**	**Canary Islands, El Hierro**	**HI1**	**SR27B07**	**this study**
XXXXXXXX	***C. dipterum***	**CA2**	**Canary Islands, El Hierro**	**HI1**	**SR27B08**	**this study**
XXXXXXXX	*C. dipterum*	CA1	Canary Islands, El Hierro	HI2	SR27B09	this study
XXXXXXXX	*C. dipterum*	CA2	Canary Islands, El Hierro	HI2	SR27B10	this study

KF257125	*C. dipterum*	KR	South Korea	KR	EPH02	Kim et al. 2014
XXXXXXXX	*C. dipterum*	CA1	Canary Islands, La Palma	LP3	SR27D06	this study
XXXXXXXX	***C. dipterum***	**CA1**	**Canary Islands, La Palma**	**LP3**	**SR27D07**	**this study**
KJ631627	***C. dipterum***	**EU1**	**Lithuania**	**LT003**	**SR12B06**	**Rutschmann et al. 2014**
KJ631628	*C. dipterum*	EU2	Lithuania	LT019	SR22G08	Rutschmann et al. 2014
KJ631629	*C. dipterum*	EU1	Latvia	LV001	SR12B09	Rutschmann et al. 2014
XXXXXXXX	*C. dipterum*	EU3	Latvia	LV001	SR12B10	this study
KJ631630	*C. dipterum*	EU3	Latvia	LV002	SR12B11	Rutschmann et al. 2014
KJ631631	*C. dipterum*	EU1	Latvia	LV004	SR12C01	Rutschmann et al. 2014
KJ631632	***C. dipterum***	**EU1**	**Latvia**	**LV004**	**SR12C02**	**Rutschmann et al. 2014**
KJ631633	*C. dipterum*	EU2	Latvia	LV010	SR12D01	Rutschmann et al. 2014
KJ631634	*C. dipterum*	EU2	Latvia	LV011	SR12D03	Rutschmann et al. 2014
KJ631635	*C. dipterum*	EU1	Latvia	LV016	SR12D12	Rutschmann et al. 2014
KJ631636	***C. dipterum***	**EU2**	**Latvia**	**LV017**	**SR12E01**	**Rutschmann et al. 2014**
KJ631637	*C. dipterum*	EU2	Latvia	LV019	SR12E06	Rutschmann et al. 2014
XXXXXXXX	*C. dipterum*	CA1	Canary Islands, Lanzarote	LZ1	SR27D08	this study
XXXXXXXX	***C. dipterum***	**CA1**	**Canary Islands, Lanzarote**	**LZ1**	**SR27D09**	**this study**
XXXXXXXX	***C. dipterum***	**CA1**	**Madeira, Madeira**	**MD11**	**SR23A10**	**this study**
XXXXXXXX	***C. dipterum***	**CA1**	**Madeira, Madeira**	**MD11**	**SR23B08**	**this study**
XXXXXXXX	*C. dipterum*	AZ1	Azores, Pico	PI1	SR27G04	this study
XXXXXXXX	*C. dipterum*	AZ1	Azores, Pico	PI1	SR27G05	this study
XXXXXXXX	*C. dipterum*	AZ1	Azores, Pico	PI2	SR27G06	this study
XXXXXXXX	*C. dipterum*	AZ1	Azores, Pico	PI3	SR27G03	this study
XXXXXXXX	*C. dipterum*	AZ1	Azores, Pico	PI4	SR27G07	this study
XXXXXXXX	***C. dipterum***	**EU3**	**Russia**	**RU010**	**SR15G06**	**this study**
XXXXXXXX	***C. dipterum***	**EU1**	**Russia**	**RU019**	**SR13G04**	**this study**
XXXXXXXX	*C. dipterum*	EU1	Russia	RU020	SR13G05	this study
XXXXXXXX	*C. dipterum*	EU1	Russia	RU038	SR22B10	this study
XXXXXXXX	*C. dipterum*	EU2	Russia	RU046	SR22D02	this study
XXXXXXXX	*C. dipterum*	AZ1	Azores, São Jorge	SJ1	SR27F10	this study
XXXXXXXX	*C. dipterum*	AZ1	Azores, São Jorge	SJ2	SR27F11	this study
XXXXXXXX	*C. dipterum*	AZ1	Azores, São Jorge	SJ3	SR27F12	this study
XXXXXXXX	*C. dipterum*	AZ1	Azores, São Jorge	SJ4	SR27G01	this study
XXXXXXXX	*C. dipterum*	AZ1	Azores, São Jorge	SJ5	SR27G02	this study
KJ631638	***C. dipterum***	**EU2**	**Azores, São Miguel**	**SK002**	**SR13B05**	**Rutschmann et al. 2014**
KJ631639	***C. dipterum***	**EU2**	**Azores, São Miguel**	**SK004**	**SR13B08**	**Rutschmann et al. 2014**
XXXXXXXX	*C. dipterum*	AZ1	Azores, São Miguel	SM10	SR27H04	this study
KF438124	*C. dipterum*	AZ1	Azores, São Miguel	SM1A	745833	this study
KF438125	*C. dipterum*	AZ1	Azores, São Miguel	SM1A	745834	this study
XXXXXXXX	***C. dipterum***	**AZ1**	**Azores, São Miguel**	**SM1B**	**SR22A01**	**this study**
XXXXXXXX	*C. dipterum*	AZ1	Azores, São Miguel	SM1B	SR22A02	this study
XXXXXXXX	*C. dipterum*	AZ1	Azores, São Miguel	SM1B	SR22A03	this study
XXXXXXXX	*C. dipterum*	AZ1	Azores, São Miguel	SM1B	SR22A04	this study
XXXXXXXX	*C. dipterum*	AZ1	Azores, São Miguel	SM1B	SR22A05	this study

XXXXXXXX	*C. dipterum*	AZ1	Azores, São Miguel	SM1B	SR22A06	this study
XXXXXXXX	***C. dipterum***	**AZ1**	**Azores, São Miguel**	**SM1B**	**SR22A07**	**this study**
XXXXXXXX	*C. dipterum*	AZ1	Azores, São Miguel	SM1B	SR22A08	this study
XXXXXXXX	*C. dipterum*	AZ1	Azores, São Miguel	SM1B	SR22A09	this study
XXXXXXXX	*C. dipterum*	AZ1	Azores, São Miguel	SM1B	SR22A10	this study
XXXXXXXX	*C. dipterum*	AZ1	Azores, São Miguel	SM1B	SR22A11	this study
XXXXXXXX	*C. dipterum*	AZ1	Azores, São Miguel	SM1B	SR22A12	this study
XXXXXXXX	*C. dipterum*	AZ1	Azores, São Miguel	SM1B	SR22B01	this study
XXXXXXXX	*C. dipterum*	AZ1	Azores, São Miguel	SM1B	SR22B02	this study
XXXXXXXX	*C. dipterum*	AZ1	Azores, São Miguel	SM1B	SR22B03	this study
XXXXXXXX	*C. dipterum*	AZ1	Azores, São Miguel	SM1B	SR22B04	this study
XXXXXXXX	*C. dipterum*	AZ1	Azores, São Miguel	SM1B	SR22B05	this study
XXXXXXXX	*C. dipterum*	AZ1	Azores, São Miguel	SM1B	SR22B06	this study
XXXXXXXX	*C. dipterum*	AZ1	Azores, São Miguel	SM1B	SR22B07	this study
XXXXXXXX	*C. dipterum*	AZ1	Azores, São Miguel	SM2	SR24A04	this study
XXXXXXXX	*C. dipterum*	AZ1	Azores, São Miguel	SM3	SR27H02	this study
XXXXXXXX	*C. dipterum*	AZ1	Azores, São Miguel	SM4	SR27G08	this study
XXXXXXXX	*C. dipterum*	AZ1	Azores, São Miguel	SM5	SR27G09	this study
XXXXXXXX	*C. dipterum*	AZ1	Azores, São Miguel	SM7	SR27G11	this study
XXXXXXXX	*C. dipterum*	AZ1	Azores, São Miguel	SM8	SR27G12	this study
XXXXXXXX	*C. dipterum*	AZ1	Azores, São Miguel	SM9	SR27H03	this study
KC135930	*C. dipterum*	KR	South Korea	KR	P005	Park et al. unpublished
XXXXXXXX	*C. dipterum*	AZ1	Azores, Terceira	TE1	SR24A05	this study
XXXXXXXX	*C. dipterum*	AZ1	Azores, Terceira	TE1	SR24A06	this study
XXXXXXXX	*C. dipterum*	AZ1	Azores, Terceira	TE1	SR24A07	this study
XXXXXXXX	*C. dipterum*	AZ1	Azores, Terceira	TE1	SR24A09	this study
XXXXXXXX	*C. dipterum*	AZ1	Azores, Terceira	TE1	SR24A10	this study
XXXXXXXX	*C. dipterum*	AZ1	Azores, Terceira	TE1	SR24A12	this study
XXXXXXXX	*C. dipterum*	AZ1	Azores, Terceira	TE1	SR26H09	this study
KF438163	*C. dipterum***	CA2	Canary Islands, Tenerife	TF1	745839	Rutschmann et al. 2014
XXXXXXXX	*C. dipterum*	CA1	Canary Islands, Tenerife	TF1B	SR27D10	this study
XXXXXXXX	*C. dipterum*	CA2	Canary Islands, Tenerife	TF1B	SR27D11	this study
KF438164	*C. dipterum***	CA2	Canary Islands, Tenerife	TF2	745827	Rutschmann et al. 2014
XXXXXXXX	*C. dipterum*	CA2	Canary Islands, Tenerife	TF3B	SR27H10	this study
XXXXXXXX	***C. dipterum***	**CA2**	**Canary Islands, Tenerife**	**TF3D**	**SR27E05**	**this study**
XXXXXXXX	*C. dipterum*	CA1	Canary Islands, Tenerife	TF4B	SR27E07	this study
XXXXXXXX	***C. dipterum***	**CA2**	**Canary Islands, Tenerife**	**TF6**	**SR27E10**	**this study**
XXXXXXXX	***C. dipterum***	**CA1**	**Canary Islands, Tenerife**	**TF6**	**SR27E11**	**this study**
XXXXXXXX	*C. dipterum*	CA1	Canary Islands, Tenerife	TF7	SR27F01	this study
XXXXXXXX	*C. dipterum*	AZ1	U.S.	US	250001	this study
XXXXXXXX	*C. dipterum*	AZ1	U.S.	US	250002	this study
XXXXXXXX	*C. dipterum*	AZ1	U.S.	US	250004	this study
XXXXXXXX	***C. dipterum***	**AZ1**	**U.S.**	**US**	**contig**	**this study**
HM900399	*C. cognatum*	AZ1	U.S.	US	BIOUG<CAN>:10BGMAY-054	Webb et al. 2012

HM900400	*C. cognatum*	AZ1	U.S.	US	BIOUG<CAN>:10BGMAY-055	Webb et al. 2012
HG935110	*Cloeon* sp.	-	Saudia Arabia	SA	GL25	Salles et al. 2014
HG935108	*C. cf.smaeleni*	-	Saudia Arabia	SA	GL29	Salles et al. 2014
HG935109	*C. cf.smaeleni*	-	Saudia Arabia	SA	GL30	Salles et al. 2014
HG935104	*C. smaeleni*	-	Brazil	BR	GL21	Salles et al. 2014
HG935105	*C. smaeleni*	-	Brazil	BR	GL22	Salles et al. 2014
HG935106	*C. smaeleni*	-	Madagascar	MG	GL23	Salles et al. 2014
HG935107	*C. smaeleni*	-	Madagascar	MG	GL24	Salles et al. 2014
JN299149	*C. praetextum*	-	Norway	NO	NO-EPH89	Kjærstad et al. 2012
JN299150	*C. praetextum*	-	Norway	NO	NO-EPH90	Kjærstad et al. 2012
XXXXXXXX	*C. simile*	-	Switzerland	CH041	SR21H01	this study
XXXXXXXX	*C. simile*	-	Switzerland	CH041	SR21H02	this study
XXXXXXXX	*C. simile*	-	Switzerland	CH041	SR25H10	this study
XXXXXXXX	*C. simile*	-	Spain	ES012	SR10B08	this study
XXXXXXXX	*C. simile*	-	Greece	GR014	SR11A03	this study
XXXXXXXX	*C. simile*	-	Latvia	LV003	SR12B12	this study

Notes: bold, 29 individuals used for phylogeny. *, *C.* sp1 (Rutschmann et al. 2014). **, *C.* sp2 (Rutschmann et al. 2014).



1) 411892

2) 411912

3) 411913

4) 411925

5) 411939

6) 411945

7) 411965

8) 411989

9) 412045

10) 412048

11) 412085

12) 412111

13) 412148

14) 412168

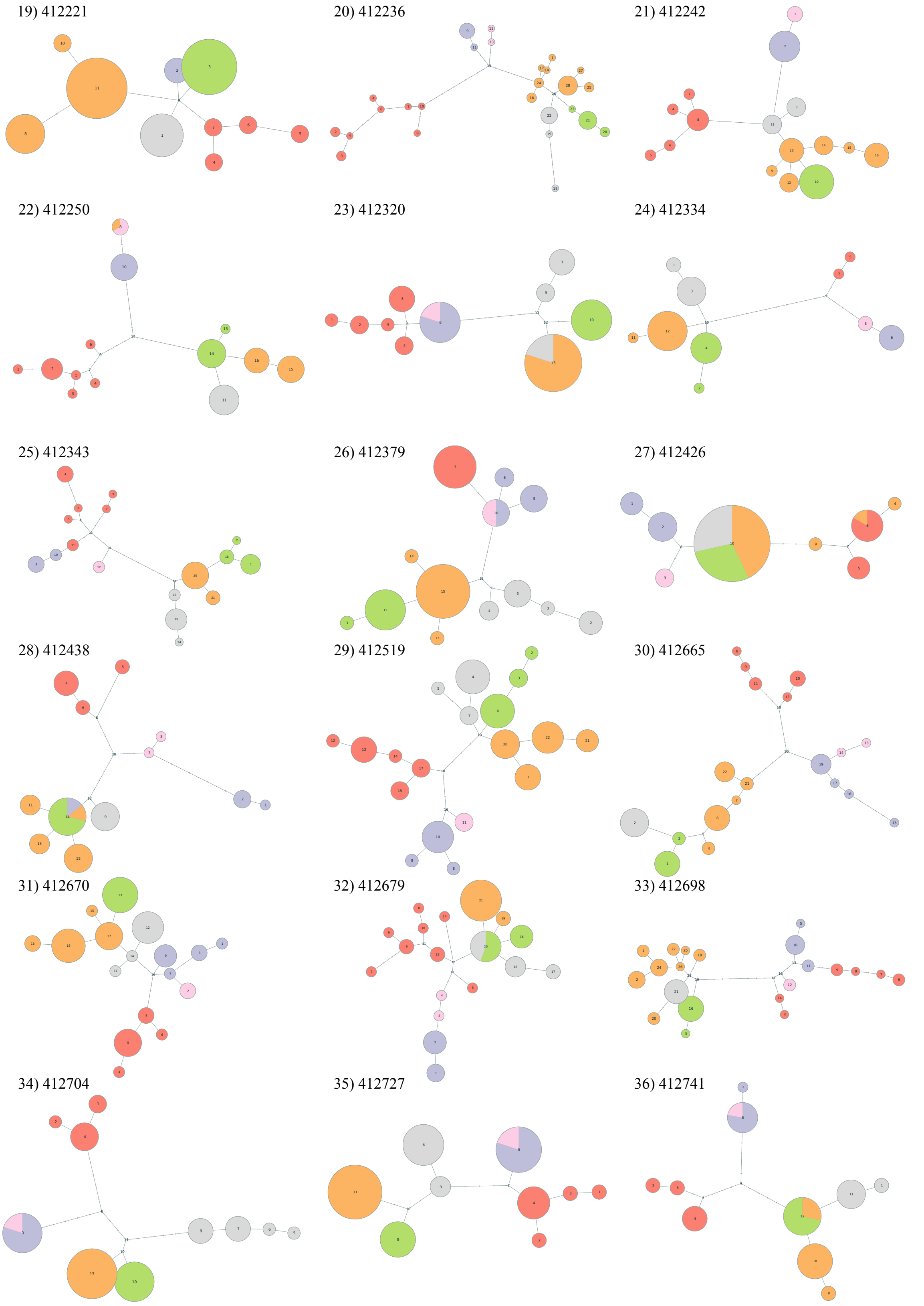
15) 412192

16) 412199

17) 412207

18) 412211

APPENDIX 3. Haplotype networks of nDNA markers (see TABLE 1) based on Fitch distances. Shown are the genealogical relationships between the haplotypes in the six putative gmyc species (green = CA2, grey = AZ1, orange = CA1, pink = EU3, purple = EU1, and red = EU2). Missing mutational steps connecting haplotypes are represented by non-colored dots. The size of the circles correlates with haplotype frequency within each network.



19) 412221

20) 412236

21) 412242

22) 412250

23) 412320

24) 412334

25) 412343

26) 412379

27) 412426

28) 412438

29) 412519

30) 412665

31) 412670

32) 412679

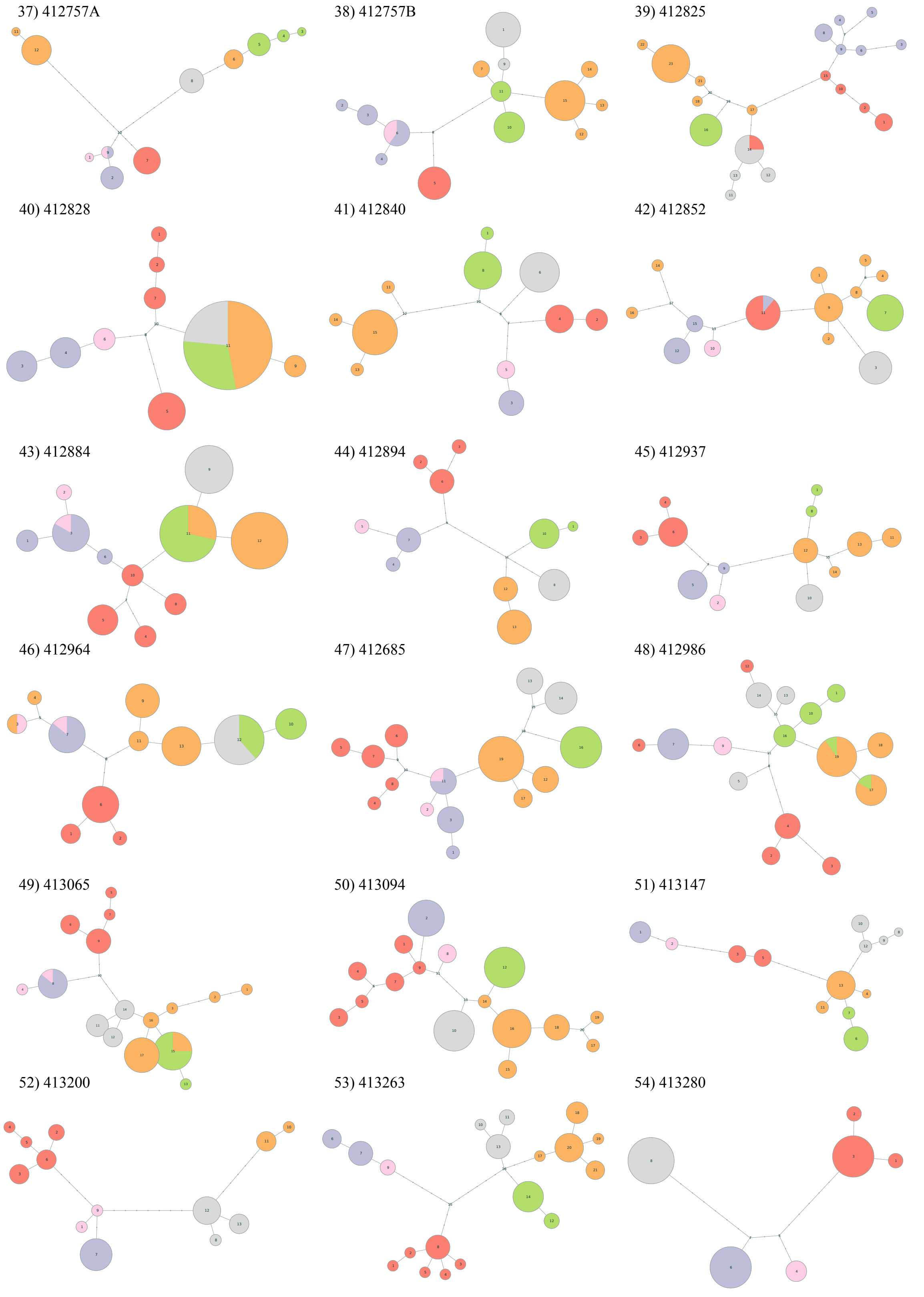
33) 412698

34) 412704

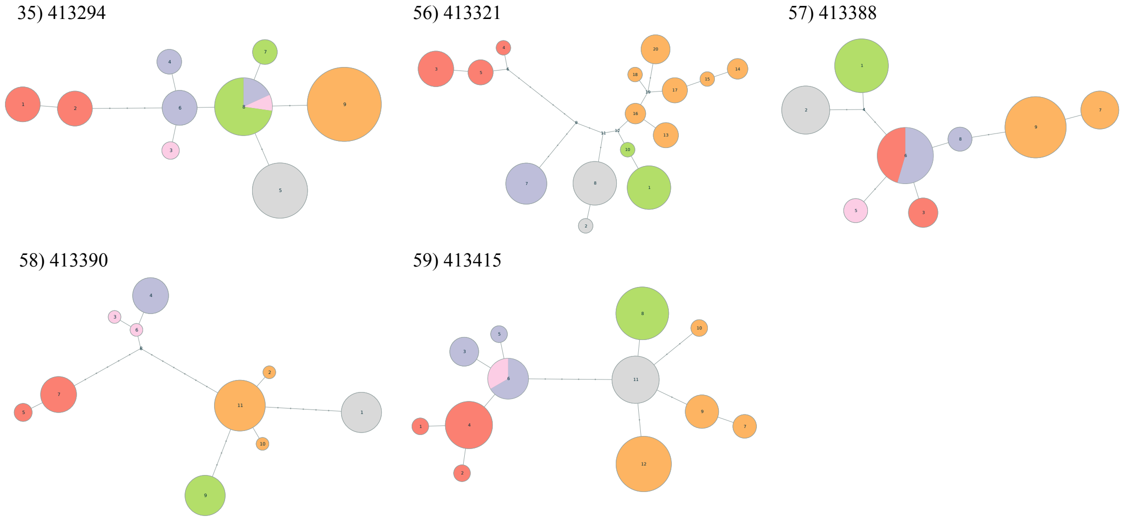
35) 412727

36) 412741

APPENDIX 3. Continued



APPENDIX 3. Continued



35) 413294

56) 413321

57) 413388

58) 413390

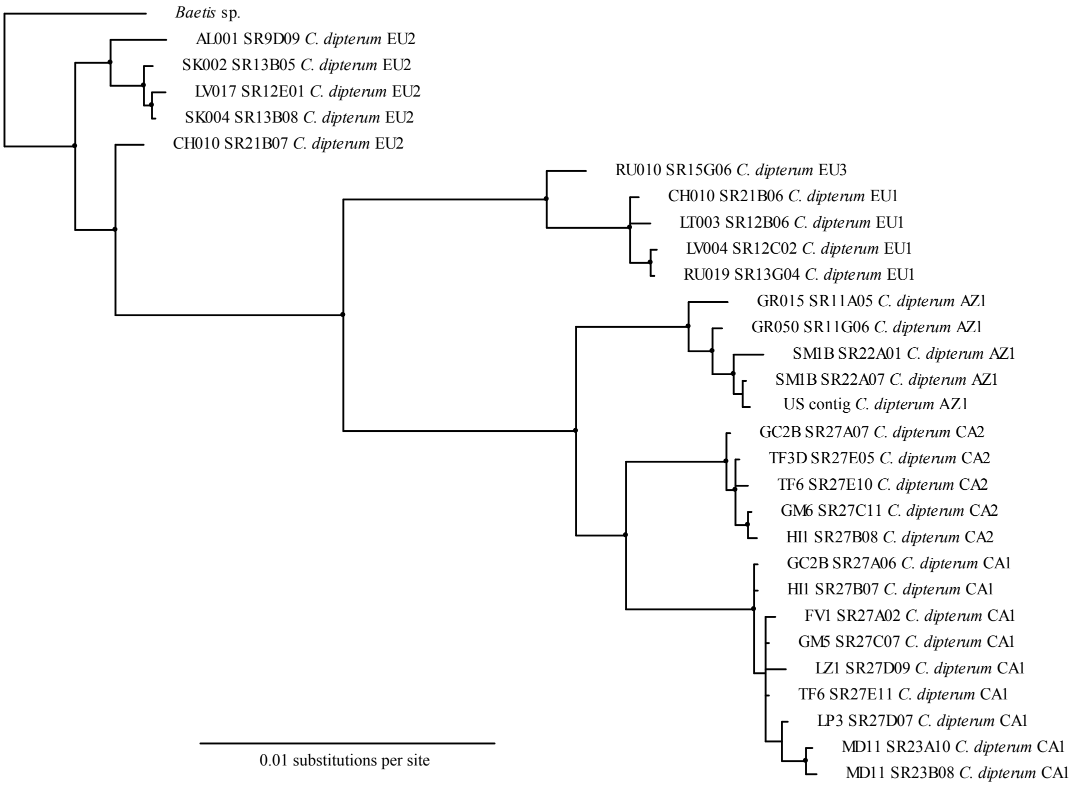
59) 413415

Appendix 3. Continued

Appendix 4. Details about Structure analyses

Dataset	K	Replicates	Mean LnP(K)	Stdev LnP(K)	Ln'(K)	|Ln''(K)|	Delta K
All	1	9	-60588.43	2.31	–	–	–
	2	**9**	**-51428.61**	**2.80**	**9159.82**	**75975.17**	**27101.84**
	3	9	-118243.96	134213.14	-66815.34	134898.81	1.01
	4	9	-50160.49	1059.98	68083.47	346339.89	326.74
	5	9	-328416.91	336860.74	-278256.42	292112.42	0.87
	6	9	-314560.91	307175.72	13856.00	74576.72	0.24
	7	9	-226128.19	359022.51	88432.72	304775.07	0.85
	8	9	-442470.53	454414.91	-216342.34	71244.83	0.16
	9	9	-730057.71	459675.00	-287587.18	626537.73	1.36
	10	9	-391107.16	418453.61	338950.56	–	–
Europe	1	5	-11505.28	4.23	–	–	–
	2	**5**	**-9252.28**	**1.58**	**2253.00**	**2219.24**	**1400.21**
	3	5	-9218.52	3.11	33.76	45.62	14.65
	4	5	-9230.38	5.13	-11.86	13.82	2.70
	5	5	-9228.42	3.51	1.96	6.74	1.92
	6	5	-9233.20	1.81	-4.78	5.16	2.86
	7	5	-9232.82	3.58	0.38	2.86	0.80
	8	5	-9235.30	1.66	-2.48	0.84	0.51
	9	5	-9236.94	3.45	-1.64	1.10	0.32
	10	5	-9239.68	3.06	-2.74	–	–
Islands	1	5	-15670.56	2.71	– –	–	–
	2	5	-13774.42	275.47	1896.14	183.38	0.67
	3	**5**	**-12061.66**	**0.42**	**1712.76**	**1716.40**	**4068.26**
	4	5	-12065.30	1.05	-3.64	0.92	0.88
	5	5	-12069.86	0.59	-4.56	0.38	0.64
	6	5	-12074.80	1.57	-4.94	2.04	1.30
	7	5	-12077.70	3.03	-2.90	5.88	1.94
	8	5	-12086.48	5.48	-8.78	9.70	1.77
	9	5	-12085.56	2.99	0.92	7.62	2.55
	10	5	-12092.26	3.78	-6.70	–	–

Notes: bold, best supported K


Baetis sp.
AL001 SR9D09 *C. dipterum* EU2
SK002 SR13B05 *C. dipterum* EU2
LV017 SR12E01 *C. dipterum* EU2
SK004 SR13B08 *C. dipterum* EU2
CH010 SR21B07 *C. dipterum* EU2
RU010 SR15G06 *C. dipterum* EU3
CH010 SR21B06 *C. dipterum* EU1
LT003 SR12B06 *C. dipterum* EU1
LV004 SR12C02 *C. dipterum* EU1
RU019 SR13G04 *C. dipterum* EU1
GR015 SR11A05 *C. dipterum* AZ1
GR050 SR11G06 *C. dipterum* AZ1
SM1B SR22A01 *C. dipterum* AZ1
SM1B SR22A07 *C. dipterum* AZ1
US contig *C. dipterum* AZ1
GC2B SR27A07 *C. dipterum* CA2
TF3D SR27E05 *C. dipterum* CA2
TF6 SR27E10 *C. dipterum* CA2
GM6 SR27C11 *C. dipterum* CA2
HI1 SR27B08 *C. dipterum* CA2
GC2B SR27A06 *C. dipterum* CA1
HI1 SR27B07 *C. dipterum* CA1
FV1 SR27A02 *C. dipterum* CA1
GM5 SR27C07 *C. dipterum* CA1
LZ1 SR27D09 *C. dipterum* CA1
TF6 SR27E11 *C. dipterum* CA1
LP3 SR27D07 *C. dipterum* CA1
MD11 SR23A10 *C. dipterum* CA1
MD11 SR23B08 *C. dipterum* CA1
0.01 substitutions per site

Appendix 5. Bayesian inference reconstruction of the phylogenetic relationships among *Cloeon dipterum* s.l. based on the concatenated exon_all_taxa data set using a separate substitution model for each gene (see Table 2). Filled circles indicate well supported nodes (PP ≥ 0.95). Shapes to the left of terminal labels indicate the origin of individuals (see Figure 1).

3.2 ANCIENT ORIGINS

CHAPTER 3

Rutschmann S, Chen P, Zhou C, Monaghan MT (*manuscript in preparation*) Using mitochondrial mayfly (Ephemeroptera) genomes to resolve phylogenetic relationships of the oldest extant winged insects.

Author contributions

S. Rutschmann gathered and analysed the *Cloeon dipterum* genome, performed all phylogenetic analyses, and drafted the manuscript. P. Chen and C. Zhou obtained and analysed the *Baetis rutilocylindratus* and *Habrophlebiodes zijinensis* genomes. M. T. Monaghan together with **S. Rutschmann** conceived the study, participated in its design and coordination, contributed discussion on the phylogenetic analyses and helped to draft the manuscript.

Using mitochondrial mayfly (Ephemeroptera) genomes to resolve phylogenetic relationships of the oldest extant winged insects

Sereina Rutschmann, Ping Chen, Changfa Zhou and Michael T. Monaghan

Abstract

Background: The relationships among the oldest winged insects (Palaeoptera), including the Ephemeroptera (mayflies) and Odonata (dragonflies and damselflies), remain unclear. The understanding of the phylogenetic relationships among major insect orders has greatly benefited from the development of high-throughput sequencing technologies. These two orders together with the Neoptera have arisen as result of a rapid divergence from a common ancestor in the distant past and are thus thought to be more susceptible to systematic inadequacies, including taxon sampling, choice of outgroup, marker selection, and phylogenetic methods.

Results: Our aim was (1) to reconstruct the origin of winged insects and (2) to investigate the impact of different systematic inadequacies, namely used phylogenetic framework (Bayesian inference (BI) vs. maximum-likelihood (ML)), and taxon sampling. We present the three newly sequenced mitochondrial genomes of *Baetis rutilocylindratus*, *Cloeon dipterum*, and *Habrophlebiodes zijinensis*, and phylogenetic reconstructions based on 93 taxa. The two different phylogenetic approaches resulted in distinct, highly supported trees, providing evidence for both the Ephemeroptera as well as the Odonata as most ancestral winged insect order. Regarding the structure of the newly sequenced genomes, we found that the gene orientation and gene content were conserved, including the complete set of 37 genes.

Conclusions: The choice of phylogenetic framework and outgroup selection were crucial to infer phylogenetic relationships within ancient insect taxa (sensu Thomas et al. 2013). Further, highly supported nodes have to be considered carefully and the phylogenetic relationships among the Palaeoptera remains a challenge for future studies. However the pruning of rogue taxa significantly improved the overall node support values.

Keywords: Baetidae, Mitochondrial genome, Pterygota, Palaeoptera problem, Rogue taxa

Background

One of the open questions within insect systematics is the onset and relationship of the Palaeoptera (Ephemeroptera and Odonata) to the modern Pterygota, and therefore also called the "Palaeoptera problem". The winged insects are divided into two groups based on their wing function: the Palaeoptera and the Neoptera. The inability of the Ephemeroptera (mayflies) and Odonata (dragon- and damselflies) to fold their wings flat over the abdomen has been considered to be an ancestral condition and therefore they are called the Palaeoptera (old wings) in contrast to the more modern Neoptera (new wings), which possess this ability (reviewed by [1]). The monophyly of the Neoptera, including the three lineages: Polyneoptera, Paraneoptera, and Holometabola, is widely accepted. Although the two earliest-branching lineages namely Polyneoptera and Paraneoptera, lack support and the phylogenetic relationships among many orders is largely unconfirmed [1].

Today, three competing hypotheses on the 'Palaeoptera problem' have been established: (i) the Palaeoptera hypothesis, classifying the Ephemeroptera + Odonata as sister group to the Neoptera, (ii) the Metapterygota hypothesis (Ephemeroptera + (Odonata + Neoptera)), and (iii) the Chiastomyaria hypothesis (Odonata + (Ephemeroptera + Neoptera)) (for a review see [1, 2]). All hypotheses are to varying degrees supported by morphological as well as molecular data. Interestingly, different authors, using the same set of genes but different phylogenetic approaches (ML and BI approaches), have supported these hypotheses (e.g. [3-8]). The first hypothesis, clustering the Ephemeroptera and Odonata as monophyletic clade has also received support from several molecular studies based on ribosomal RNAs (rRNA) [3, 9, 10], multi-gene data sets [11, 12], and phylogenomic data [6, 8, 13]. The basal Ephemeroptera hypothesis (ii) was supported by previous studies using mitogenome data [14], rRNA [10], and a combined analysis of rRNA and one nuclear gene [5]. Notably, all these studies included either a limited number of mayfly species (1, [14]) or a limited number of genes (3, [5]). In contrast, the basal Odonata (iii) hypothesis has also received a lot of support from previous studies based on mitogenome data [15], rRNA [4, 16-18], and phylogenomic studies including over 125 genes [7].

The conflicting phylogenetic signals may result from the ancient radiation of the lineages Ephemeroptera, Odonata, and Neoptera from a common ancestor in the distant past [19], leading to weak phylogenetic signal, being more susceptible to systematic errors [8, 19-22]. Thus, much of the conflicting signal among the relationships between these three orders may result from unbalanced taxon sampling, the sequence data, sequence alignments methods, and the phylogenetic methods and their models. [1, 8]. Thereby, mitogenomes as conservative marker are thought to overcome some of these systematic errors, namely they are easier to align and also appropriate models of molecular evolution are well established [23, 24]. On the other hand, mitochondrial genes include several drawbacks, most importantly the possible presence of pseudogenes [25-27]. However, mitogenomes are well studied and therefore the most widely employed genetic markers in insects and seemed to be a promising 'instrument' for insect systematics, as reviewed by Cameron 2014 [28].

Arthropod mitogenomes are highly conserved, ranging from 15 to 18 kb in length and containing 37 genes: 13 protein-coding genes (PCGs) of four complexes of the respiratory chain, two rRNAs (*rrnL* and *rrnS*) and 22 transfer RNAs (tRNAs, *trn**) [29]. A non-coding region (CR) of variable length, thought to be the origin of initiation of transcription and replication, is typically present [30, 31], and referred in insects as the AT-rich region. The typical ancestral insect mitogenome differs from the ancestral arthropod mitogenome only by the location of one gene (*trnL* [32]). Significant differences in terms of structure, gene content, and gene arrangement have been found to be the exception for highly derived taxa. Until today, 18 complete or nearly complete mayfly mitogenomes from eleven families are available on GenBank, of which ten are included in publications [14, 33-35]. Compared to the number of species/families (3046/42, [36]) this is still a relatively low number. With the development of high-throughput sequencing technologies, such as pooled multiplex sequencing, the number of mitogenomes is supposed to increase dramatically within the next years. Recent studies have sequenced pooled DNA samples in order to obtain complete or nearly complete mitogenomes from 92 weevil beetles (Coleoptera, [37]), or mitogenomes derived from a pooled DNA metabarcoding sample, including species from a wide range of taxa [35, 38].

Here, we investigate the relationships of the earliest winged insects with a special emphasis on mayflies using mitogenomes as robust markers that are less sensitive to phylogenetic systematic errors. Therefore, we investigated the impact of different systematic inadequacies, namely phylogenetic approaches (Bayesian inference (BI) vs. maximum-likelihood (ML)), and taxon sampling. We included a large set of palaeopteran sequences (29) together with 64 other insect mitogenomes, and pruned taxa being assumed to show varying and often contradictory topological positions (rogue taxa, [39]) in a set of trees using the program RogueNaRok [40]. This approach has never been tested before to resolve the 'Palaeoptera problem'. Further, we excluded taxa leading to the phenomenon of long-branch attractions (LBA, [41, 42]). We newly sequenced and characterized the mitogenomes of three mayfly species, including one representative of the family Leptophlebiidae for which no mitogenome data was available so far and two baetid specimens. The genomic data were produced using standard Sanger sequencing as well as 454 pyrosequencing.

Results and discussion

Mitogenome assembly

The pyrosequencing run using the 454 GS FLX system resulted in 651.306 reads, of which 1.14% mapped to the mitogenome. The depth of coverage for the *C. dipterum* mitogenome was 249.2x (± 80.6 SD, Figure 1). The coverage of the *C. dipterum* is in agreement with other studies, using the same sequencing platform (e.g. 59-281x, [43]). The missing of part of the AT-rich region is due to reduced sequencing and assembly efficiency of this low complexity region and common among insect mitogenomes [34, 35].

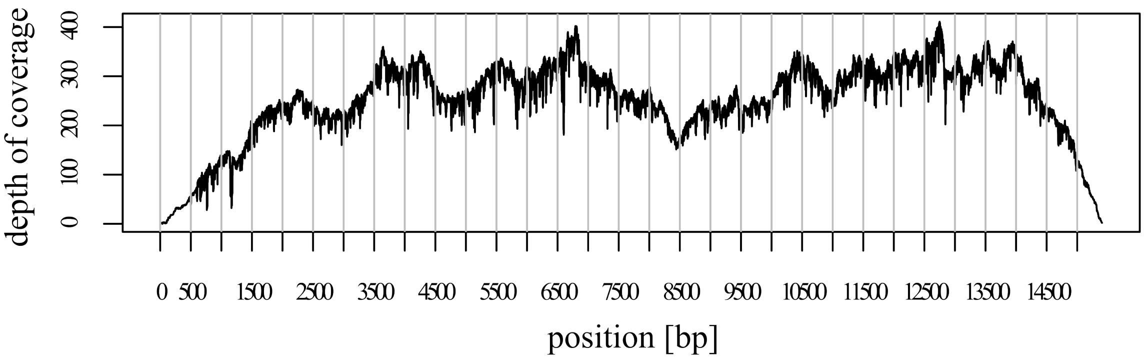


Figure 1 Coverage depth of *Cloeon dipterum* mitochondrial genome.

Mitogenome organization

The three mayfly mitogenomes were 14,355 base pairs (bp) (*H. zijinensis*), 14,883 bp (*B. rutilocylindratus)*, and 15,408 bp (*C. dipterum*) long, whereby the genomes of *H. zijinensis* and *C. dipterum* were incomplete due to incomplete AT-rich regions and three missing tRNAs for *H. zijinensis* (Figure 2). The mitogenome sequences have been deposited at GenBank accession numbers GU936204, GU936203, and XXXXXXXX.

All three sequenced mitogenomes contained the entire set of 13 PCGs, two rRNAs, and 19 tRNAs (*H. zijinensis*) respective 22 tRNAs (*B. rutilocylindratus*, *C. dipterum*), with 21 coded at the (+) strand (19 for *H. zijinensis*) and 16 at the (-) strand (15 *H. zijinensis*) (Figure 2). The gene order and orientation were identical to the ancestral insect mitogenome [29, 44]. Typically, all PCGs started with the ATN codons (ATT, ATG, ATA), and mostly ended with the complete termination codon (TAA or TAG). In *B. rutilocylindratus* we found an incomplete T termination codons for *nad4*, and in *C. dipterum* the gene *cox1* started with CTC. Other mayflies are also reported to miss complete T termination codons in the genes *cox2* and *nad5* [14, 33-35]. The two rRNAs were located between *trnL1* and *trnV* (*rrnL*), and between *trnV* and the CR (*rrnS*), respectively. The AT-rich region (CR) of all ephemeropteran mitogenomes is placed between the *rrnS* (- strand) and *trnI* (+ strand). Li et al. [34] reported two distinct parts within the AT-rich region in *Siphluriscus chinensis,* which also seems to be present in *C. dipterum*. Therein, they described the so called $CR_{1,}$ which is located close to the *rrnS* and has a high AT content (71.6%), including six identical 140 bp sequences, and the CR_2, which is close to the *trnI* and has a lower AT content (58.1%). *Ephemera orientalis* contains two identical 55 bp long sequences in the AT-rich region [33]. Few mayfly mitogenomes differ in their gene content from the ancestral insect mitogenome, possessing one additional tRNA. The two heptageniid species *Parafronurus youi* and *Epeorus* sp. encode an additional copy of the *trnM* (AUG, *trnM2*) gene located between *trnI* and *trnQ* [14, 35]. For *S. chinensis*, an additional *trnK2* (AAA) gene is described [34]. All tRNAs can be folded into the typical overleaf structure with amino-acyl stem (7 bp), anticodon arm (5 bp), anticodon loop (7 bp), a variable loop, DHU arm and a TyC arm.

Base composition

The overall AT content across mayflies ranges from 60.1 in *B. rutilocylindratus* to 72.7% in *Ephemera orientalis* (Table 1). The average value for all mayflies was 66% for the whole sequences. The values were as expected slightly lower for the common_sequence data set except for *Alainites yixiani* (Table 1). These high AT contents are typical for insect species, ranging from 64% in termites to 86.7% in bees [45]. The average whole mitogenome AT-skew was -0.03 (± 0.04 SD). The average GC-skew was -0.14 (-0.30 to 0.14) with most mitogenomes displaying negative skews.

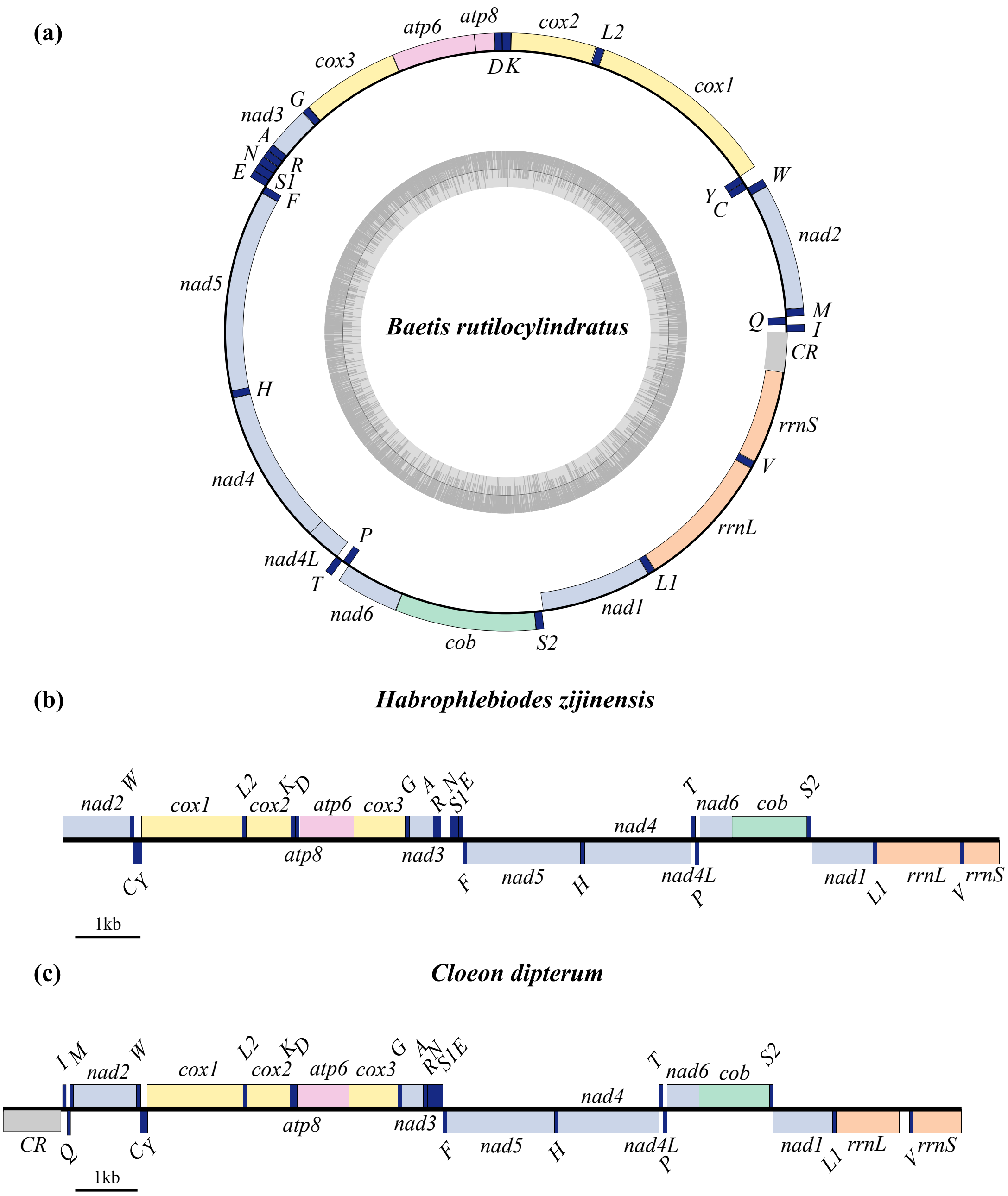


Figure 2 Mitochondrial genome maps. Complete mitochondrial genome of (a) *Baetis rutilocylindratus,* and nearly complete mitochondrial genomes of (b) *Habrophlebiodes zijinensis*, and (c) *Cloeon dipterum*. Transfer RNA genes are indicated by single-letter IUPAC-IUB abbreviations for their corresponding amino acid. Protein coding genes and ribosomal RNA genes are listed and colored in the following way: *atp6*, *atp8*, ATP synthase subunits 6 and 8 genes (pink); *cob*, cytochrome oxidase *b* gene (green); *cox1-cox3*, cytochrome oxidase *c* subunit 1-3 genes (yellow); *nad1-6, nad4L*, NADH dehydrogenase subunits 1-6 and 4L (blue); *rrnS, rrnL*, small and large ribosomal RNA subunits (red); CR, control region/AT-rich region (grey). Genes located at the (-) strand appear in the outer circle for (a) respective above the central line in (b) and (c). Genes located on the (+) strand appear in the inner circle for (a) respective below the central line in (b) and (c).

Table 1 Ephemeropteran mitochondrial genomes information including nucleotide compositions calculated based on the whole available sequences and the common_sequence data set (above, see Methods)

Family	Species	Length	A%	C%	G%	T%	GC%	AT%	AT-skew	GC-skew
Ameletidae	*Ameletus* sp1	15,141	33.2	13.7	20.6	32.5	34.2	65.7	0.011	0.201
		12,305	33.1	13.9	20.7	32.3	34.6	65.4	0.012	0.197
Baetidae	*Alainites yixiani*	**14,589**	29.1	15.5	20.5	34.9	36.0	64.0	-0.091	0.139
		12,130	28.8	14.7	21.0	35.5	35.7	64.3	-0.104	0.176
Baetidae	*Baetis rutilocylindratus*	**14,883**	27.3	19.8	20.1	32.8	39.9	60.1	-0.092	0.008
		12,234	26.7	19.8	20.4	33.2	40.2	59.9	-0.109	0.015
Baetidae	*Cloeon dipterum*	14,355	30.6	14.7	16.3	38.4	30.9	69.0	-0.113	0.052
		12,197	29.9	14.5	16.9	38.7	31.4	68.6	-0.128	0.076
Caenidae	*Caenis pycnacantha*	**15,351**	32.4	20.8	14.2	32.5	35.1	64.9	-0.002	-0.189
		12,301	31.8	21.4	15.3	31.6	36.7	63.4	0.003	-0.166
Ephemerellidae	*Ephemerella* sp.	14,896	30.4	21.9	16.3	31.4	38.2	61.8	-0.016	-0.147
		12,233	30.4	21.6	17.1	31.0	38.6	61.4	-0.010	-0.116
Ephemerellidae	*Vietnamella dabieshanensis*	**15,761**	32.1	17.6	11.8	38.6	29.4	70.7	-0.092	-0.197
		12,286	32.5	18.1	12.3	37.1	30.4	69.6	-0.066	-0.191
Ephemerellidae	*Vietnamella* sp.	15,043	30.8	20.5	13.5	35.2	34.0	66.0	-0.067	-0.206
		12,788	30.8	21.0	13.9	34.3	34.9	65.1	-0.054	-0.203
Ephemeridae	*Ephemera orientalis*	**16,463**	37.2	17.4	10.3	35.0	27.7	72.2	0.030	-0.256
		12,366	36.3	18.2	11.4	34.1	29.6	70.4	0.031	-0.230
Heptageniidae	*Epeorus* sp.	15,456	31.7	22.1	13.8	32.3	35.9	64.0	-0.009	-0.231
		12,371	31.0	22.8	14.4	31.8	37.2	62.8	-0.013	-0.226
Heptageniidae	*Paegniodes cupulatus*	**15,715**	32.5	20.7	13.7	33.1	34.4	65.6	-0.009	-0.203
		12,366	32.2	20.8	13.6	33.4	34.3	65.6	-0.018	-0.209
Heptageniidae	*Parafronurus youi*	**15,481**	32.7	20.5	13.1	33.7	33.6	66.4	-0.015	-0.220
		12,361	32.9	20.2	12.9	34.0	33.0	66.9	-0.016	-0.221
Isonychiidae	*Isonychia ignota*	**15,105**	30.5	23.4	14.9	31.2	38.3	61.7	-0.011	-0.222
		12,318	30.6	23.6	15.0	30.8	38.6	61.4	-0.003	-0.223
Leptophlebiidae	*Habrophlebiodes zijinensis*	15,407	33.8	20.2	11.0	35.1	31.2	68.9	-0.019	-0.295
		12,417	33.6	20.6	11.5	34.3	32.1	67.9	-0.010	-0.283
Potamanthidae	*Potamanthus* sp.	14,937	32.8	19.4	13.6	34.1	33.1	66.9	-0.019	-0.176
		12,289	32.5	19.5	14.2	33.9	33.6	66.4	-0.021	-0.157
Siphlonuridae	*Siphlonurus immanis*	**15,529**	35.2	17.6	11.7	35.4	29.3	70.6	-0.003	-0.201
		12,339	34.7	18.0	12.2	35.1	30.2	69.8	-0.006	-0.192
Siphlonuridae	*Siphlonurus* sp.	14,745	31.5	20.9	14.7	32.9	35.6	64.4	-0.022	-0.174
		12,333	31.0	21.0	15.1	32.9	36.1	63.9	-0.030	-0.163
Siphluriscidae	*Siphluriscus chinensis*	16,616	33.9	19.0	14.4	32.6	33.4	66.5	0.020	-0.138
		12,394	33.5	19.6	14.3	32.7	33.9	66.2	0.012	-0.156
Teloganodidae	sp.	15,252	30.1	21.9	14.4	33.6	36.3	63.7	-0.055	-0.207
		12,208	29.8	22.6	15.1	32.6	37.6	62.4	-0.045	-0.199

Numbers in bold indicate complete mitochondrial genomes; rest are incomplete mitochondrial genomes.

Palaeoptera relationships

The different phylogenetic analyses supported the basal Ephemeroptera hypothesis (BI) respective the basal Odonata hypothesis (ML) (Figure 3, Additional file 1: Figure S1) with highest support values (Bayesian posterior probability (BPP) = 1.00 and Bootstrap support (BS) = 100%). This is not surprising given the long list of phylogenetic studies that have failed to resolve their phylogenetic relationships in the past. The total length of the concatenated amino acid sequence alignment was 3,943 bp for the data set including all taxa, and 3,917 bp for the data set, excluding rogue taxa and taxa with LBA (see Table 2).

The monophyly of the Ephemeroptera was highly supported for all analyses BPP = 1.00 and BS = 100%. Also the Odonata clustered as monophyletic clade (BPP = 1.00 and BS = 100%). The ML and BI approaches recovered besides the 'Palaeoptera problem' a different tree topology for three taxa: *Tricholepidion gertischi* (Zygentoma), *Eupolyphaga sinensis* (Blattodea), and *Bacillus rossius* (Phasmatodea), whereby the positions of the first two were not supported. The two orders of the Zygentoma and Orthoptera appeared as paraphyletic clades. The close phylogenetic relationship of the Dermaptera and Ephemeroptera (Additional file 2: Figure S1) might be misleading due to LBA. However, Li et al. [34] found the order Dermaptera as being closely related to the mayflies. On the other hand studies based on phylogenomic data reported the Dermaptera as sister taxa to the orders Plecoptera [46], and Zoraptera [47]. However, more phylogenetic studies will be needed to clarify this issue; mostly also because the use of few anomalous taxa such as the Dermaptera or Embioptera tend to evoke LBA [41, 42]. The inclusion of several mitogenomes of the orders Ephemeroptera and Plecoptera has contributed evidence to consider these two orders as distantly related and not sister taxa as proposed by previous studies [14, 15, 34]. However, all these studies have used limited sets of taxa, including between one to four mayfly species, one stonefly species, and were based on mitochondrial DNA (mtDNA).

The order of the Ephemeroptera was found to be one of the most phylogenetically diverse orders as measured by comparisons of the branch lengths (Figure 1, Additional file 2: Figure S1). We recovered *S. chinensis* as sister taxa to all other mayfly taxa (also found by Li et al. [34]). The Baetidae are the most species-rich mayfly family, with worldwide 833 described species [36] and for a long time they were thought to be one of the most ancestral mayfly families [47, 48]. Within the Baetidae, *C. dipterum* has been detected as sister taxa to *B. rutilocylindratus* and *A. yixiani* (BPP = 1.00, BS = 100).

Pruning of problematic taxa

Pruning of taxa with LBA (*Aposthonia japonica* (Embioptera); and *Challia fletcheri* (Dermaptera)) and rogue taxa improved the accuracy of our phylogenetic reconstructions. The use of the all_taxa set using BI resulted in four unresolved nodes whereas the optimized_taxa set only contained two uncertainties. For several nodes, the overall node support values increased from BPP > 0.95 to BPP = 1.00 (see Figure 3, Additional file 2: Figure S1). The RogueNaRok analysis identified the outgroup species *Tetrodontophora bielanensis* (Collembola) together with the species *Gryllotalpa orientalis* (Orthoptera), and *Phraortes* sp. (Phasmatodea) as taxa with uncertain phylogenetic position leading to less accurate overall phylogenetic reconstructions. Interestingly, the choice of the outgroup species was reported as being crucial for resolving problematic splits in the tree of life such as insects origin [8]. Thus, these findings are not surprising but clearly add more evidences that the choice of outgroup is crucial especially for phylogenetic reconstructions within ancient lineages.

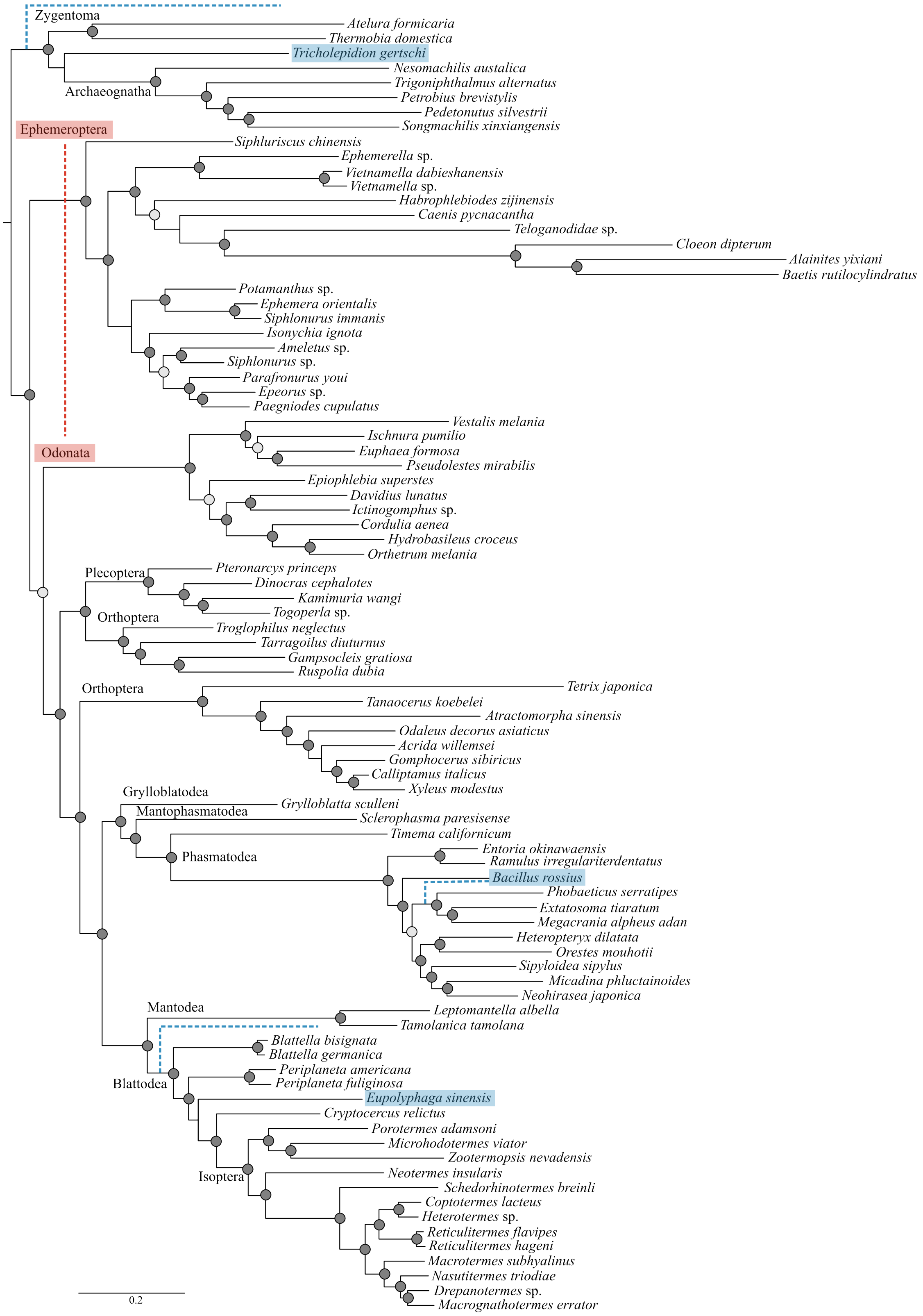


Figure 3 (See legend on next page.)

(See figure above and on previous page.)

Figure 3 Phylogenetic relationships of insect orders. Bayesian inference reconstruction based on mitochondrial genomes data using the concatenated amino acid sequences of the optimized_taxa data set (see Table 2). Filled circles indicate well-supported nodes; whereby dark grey circles represent Bayesian posterior probability (BPP) = 1, and light grey circles BPP > 0.95. Scale bar indicates substitutions per site. Coloured taxa represent differences to the maximum-likelihood (ML) phylogeny. Specimens with different topology are highlighted in blue and the different clustering is indicated by dotted lines. The red highlighted orders Ephemeroptera and Odonata appeared in reversed clustering for the ML phylogeny (for details see Results and discussion section).

Conclusions

We here present the first study including mitogenomes of representatives from eleven mayfly families. The increasing number of mayfly mitogenomes sheds more light on the structure of the ancestral winged insect mitogenomes. However, the evolution of the oldest extant insect orders could not be resolved since we found support for both the basal Ephemeroptera (BI) and basal Odonata hypotheses (ML). Most phylogenetic studies on the Palaeoptera only included few mayfly specimens and phylogenetic studies using mitogenome data with incomplete representation of major orders/taxa should be interpreted with caution even if the phylogenetic relationships are highly supported [49]. Although the Chiastomyaria and Metapterygota hypotheses have received support from recent studies based on large nuclear data sets [6, 7, 18, 50]. However the most recent phylogenetic reconstruction by Misof et al. [47] based on 1,478 PCGs failed to resolve the relationships among the oldest extant winged insects.

The Palaeoptera problem remains a challenge demanding further phylogenomic studies, whereby more knowledge about the limitation of individual markers are needed. For example Simon et al. [46] found that proteins involved in cellular processes and signaling harbor the most phylogenetic signal. Odgen and Whiting [51] already pointed out the ambiguous resolution of the mayflies based on molecular data. The ancient lineages Ephemeroptera, Odonata, and Neoptera appear to have diverged rapidly, leaving few characteristics to determine their phylogenetic relationships [19]. Evolutionary rate heterogeneity across clades and the representation of old clades by recent extant taxa make ancient relationships such as these of the pterygotes or the mammalian orders comprising the Paenungulata difficult to resolve [19, 52].

Methods

Taxon sampling

We chose to sequence the species *Habrophlebiodes zijinensis* Gui, Zhang and Wu, 1996 as the first representative of the family Leptophlebiidae, and one representative of each subfamily within the Baetidae distantly related species: *Baetis rutilocylindratus* Wang, Qin, Chen & Zhou, 2011 and *Cloeon dipterum* L. 1761. Taxon selection for integrated phylogenetic analyses was focused on available mitogenomes of related basal pterygote (Table 2). All nucleotide and amino acid sequences were obtained from NCBI (November 2014) using a Python script (mitogenome_ncbi.py, https://github.com/srutschmann/python_scripts).

Table 2 Set of mitochondrial genomes with according GenBank accession number

Order	Family	Species	Accession
Archaeognatha	Machilidae	*Pedetontus silvestrii*	NC_011717
Archaeognatha	Machilidae	*Petrobius brevistylis*	NC_007688
Archaeognatha	Machilidae	*Songmachilis xinxiangensis*	NC_021384
Archaeognatha	Machilidae	*Trigoniophthalmus alternatus*	NC_010532
Archaeognatha	Meinertellidae	*Nesomachilis australica*	NC_006895
Blattodea	Blattidae	*Periplaneta americana*	NC_016956
Blattodea	Blattidae	*Periplaneta fuliginosa*	NC_006076
Blattodea	Corydiidae	*Eupolyphaga sinensis*	NC_014274
Blattodea	Cryptocercidae	*Cryptocercus relictus*	NC_018132
Blattodea	Ectobiidae	*Blattella bisignata*	NC_018549
Blattodea	Ectobiidae	*Blattella germanica*	NC_012901
Collembola	Tetrodontophorinae	***Tetrodontophora bielanensis***	NC_002735
Dermaptera	Pygidicranidae	***Challia fletcheri***	NC_018538
Embioptera	Oligotomidae	***Aposthonia japonica***	AB639034
Ephemeroptera	Ameletidae	*Ameletus* sp1	KM244682
Ephemeroptera	Baetidae	*Alainites yixiani*	NC_020034
Ephemeroptera	Baetidae	*Baetis rutilocylindratus*	GU936204
Ephemeroptera	Baetidae	*Cloeon dipterum*	XXXXXXXX
Ephemeroptera	Caenidae	*Caenis pycnacantha*	GQ502451
Ephemeroptera	Ephemerellidae	*Ephemerella* sp.	KM244691
Ephemeroptera	Ephemerellidae	*Vietnamella dabieshanensis*	NC_020036
Ephemeroptera	Ephemerellidae	*Vietnamella* sp.	KM244655
Ephemeroptera	Ephemeridae	*Ephemera orientalis*	NC_012645
Ephemeroptera	Heptageniidae	*Epeorus* sp.	KM244708
Ephemeroptera	Heptageniidae	*Paegniodes cupulatus*	NC_020035
Ephemeroptera	Heptageniidae	*Parafronurus youi*	NC_011359
Ephemeroptera	Isonychiidae	*Isonychia ignota*	NC_020037
Ephemeroptera	Leptophlebiidae	*Habrophlebiodes zijinensis*	GU936203

Ephemeroptera	Potamanthidae	*Potamanthus* sp.	KM244674
Ephemeroptera	Siphlonuridae	*Siphlonurus immanis*	NC_013822
Ephemeroptera	Siphlonuridae	*Siphlonurus* sp.	KM244684
Ephemeroptera	Siphluriscidae	*Siphluriscus chinensis*	HQ875717
Ephemeroptera	Teloganodidae	sp.	KM244670, KM244703*
Grylloblattodea	Grylloblattidae	*Grylloblatta sculleni*	DQ241796
Isoptera	Hodotermitidae	*Microhodotermes viator*	NC_018122
Isoptera	Kalotermitidae	*Neotermes insularis*	NC_018124
Isoptera	Rhinotermitidae	*Coptotermes lacteus*	NC_018125
Isoptera	Rhinotermitidae	*Heterotermes* sp.	NC_018127
Isoptera	Rhinotermitidae	*Reticulitermes flavipes*	NC_009498
Isoptera	Rhinotermitidae	*Reticulitermes hageni*	NC_009501
Isoptera	Rhinotermitidae	*Schedorhinotermes breinli*	NC_018126
Isoptera	Termitidae	*Drepanotermes* sp.	NC_018129
Isoptera	Termitidae	*Macrognathotermes errator*	NC_018130
Isoptera	Termitidae	*Macrotermes subhyalinus*	NC_018128
Isoptera	Termitidae	*Nasutitermes triodiae*	NC_018131
Isoptera	Termopsidae	*Porotermes adamsoni*	NC_018121
Isoptera	Termopsidae	*Zootermopsis nevadensis*	NC_024658
Mantodea	Caliridinae	*Leptomantella albella*	NC_024028
Mantodea	Mantidae	*Tamolanica tamolana*	NC_007702
Mantophasmatodea	Mantophasmatidae	*Sclerophasma paresisensis*	NC_007701
Odonata	Calopterygidae	*Vestalis melania*	NC_023233
Odonata	Coenagrionidae	*Ischnura pumilio*	NC_021617
Odonata	Corduliidae	*Cordulia aenea*	JX963627
Odonata	Epiophlebiidae	*Epiophlebia superstes*	NC_023232
Odonata	Euphaeidae	*Euphaea formosa*	NC_014493
Odonata	Gomphidae	*Davidius lunatus*	NC_012644
Odonata	Gomphidae	*Ictinogomphus* sp.	KM244673
Odonata	Libellulidae	*Hydrobasileus croceus*	KM244659
Odonata	Libellulidae	*Orthetrum melania*	AB126005
Odonata	Pseudolestidae	*Pseudolestes mirabilis*	NC_020636
Orthoptera	Acrididae	*Acrida willemsei*	NC_011303
Orthoptera	Acrididae	*Calliptamus italicus*	NC_011305
Orthoptera	Acrididae	*Gomphocerus sibiricus*	NC_021103
Orthoptera	Acrididae	*Oedaleus decorus asiaticus*	NC_011115
Orthoptera	Gryllotalpidae	***Gryllotalpa orientalis***	NC_006678
Orthoptera	Pneumoridae	*Tanaocerus koebelei*	NC_020777
Orthoptera	Prophalangopsidae	*Tarragoilus diuturnus*	NC_021397
Orthoptera	Pyrgomorphidae	*Atractomorpha sinensis*	NC_011824
Orthoptera	Rhaphidophoridae	*Troglophilus neglectus*	NC_011306
Orthoptera	Romaleidae	*Xyleus modestus*	NC_014490
Orthoptera	Tetrigidae	*Tetrix japonica*	NC_018543
Orthoptera	Tettigoniidae	*Gampsocleis gratiosa*	NC_011200

Orthoptera	Tettigoniidae	*Ruspolia dubia*	NC_009876
Phasmatodea	Bacillidae	*Bacillus rossius*	GU001956
Phasmatodea	Diapheromeridae	*Micadina phluctainoides*	NC_014673
Phasmatodea	Diapheromeridae	*Sipyloidea sipylus*	AB477470
Phasmatodea	Heteropterygidae	*Heteropteryx dilatata*	NC_014680
Phasmatodea	Heteropterygidae	*Orestes mouhotii*	AB477462
Phasmatodea	Phasmatidae	*Entoria okinawaensis*	NC_014694
Phasmatodea	Phasmatidae	*Extatosoma tiaratum*	NC_017748
Phasmatodea	Phasmatidae	*Megacrania alpheus adan*	NC_014688
Phasmatodea	Phasmatidae	*Neohirasea japonica*	AB477469
Phasmatodea	Phasmatidae	*Phobaeticus serratipes*	NC_014678
Phasmatodea	Phasmatidae	***Phraortes* sp.**	NC_014705
Phasmatodea	Phasmatidae	*Ramulus irregulariterdentatus*	NC_014702
Phasmatodea	Timematidae	*Timema californicum*	DQ241799
Plecoptera	Perlidae	*Dinocras cephalotes*	NC_022843
Plecoptera	Perlidae	*Kamimuria wangi*	NC_024033
Plecoptera	Perlidae	*Togoperla* sp.	KM409708
Plecoptera	Pteronarcyidae	*Pteronarcys princeps*	NC_006133
Zygentoma	Lepidotrichidae	*Tricholepidion gertschi*	NC_005437
Zygentoma	Lepismatidae	*Thermobia domestica*	NC_006080
Zygentoma	Nicoletiidae	*Atelura formicaria*	NC_011197

Bold specimens were excluded in the optimized_taxa set (see Methods). Star represents sequence consisting of two genomic fragments (contigs) from the same sample.

DNA extraction and sequencing

The three mayfly species were sequenced with two different approaches: (i) 454/FLX pyrosequencing of *C. dipterum* (Baetidae), and (ii) Sanger sequencing for *B. rutilocylindratus* (Baetidae) and *H. zijinensis* (Leptophlebiidae).

(i) We extracted the DNA of *C. dipterum* from twelve to twenty pooled reared subimago specimens using the Invisorb® Spin Tissue Mini (Stractec, Berlin, Germany) kit. The extracted DNA was precipitated (Isopropanol precipitation of DNA, Qiagen, Leipzig, Germany), and pooled according manufacturer's guidelines in order to obtain a higher DNA yield. We prepared a shotgun library according to the manufacturer's guidelines (Rapid Library Preparation Method Manual, *GS FLX+ Series - XL+*, May 2011). 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 were sequenced on a Roche (454) *GS FLX* machine at the Berlin Center for Genomics in Biodiversity Research (BeGenDiv, Berlin, Germany) according to manufacturer's guideline (Sequencing Method Manual, *GS FLX Titanum Series*, October 2009 (Rev. Jan 2010)). The obtained sequence reads were trimmed and *de novo* assembled using the software Newbler v.2.5.3 (454 Life Science Cooperation) under default settings for large datasets. In order to extract the mitogenome of *C. dipterum*, we performed BLASTN searches [53] using as query all assembled contigs against the NCBI database. We mapped all reads back to the mitogenome with BWA [54], using the BWA-SW algorithm [55] with settings suggested for 454 data by CORAL (match score = 2, mismatch penalty = 2, and gap open penalty = 3, [56]).

(ii) Specimens of *B. rutilocylindratus* and *H. zijinensis* were collected in Zijin Hill, Nanjing, China. The DNA was extracted from between two and four larvae using the DNeasy® Blood & Tissue (Qiagen, Leipzig, Germany) kit. Four DNA fragments of each species were amplified with universal primers (*B. rutilocylindratus*: *cox1*, *cox3*, *cob*, *rrnL*; *H. zijinensis*: *cox1*, *cob*, *nad4*, *rrnL*; [44]). Subsequently, specific primers were designed based on the previously obtained sequence information (Additional file 1: Table S1). Standard and long polymerase chain reactions (PCRs) were performed on a DNA Engine Peltier Thermal Cycler (Bio-Rad, Shanghai, China). Therefore, we used the rTaq^{TM} DNA polymerase (TaKaRa Bio, Dalian, China) for fragments smaller than two kb and the LA Taq^{TM} polymerase (TaKaRa Bio, Dalian, China) for fragments larger than two kb. All PCR products were purified with the Axygen agarose-out kit. When the PCR amplification signal was too weak to sequence or sequencing resulted in overlap peaks, the products were ligated to pGEM®-T Easy Vector (Promega, Southampton, UK) by *Escherichia coli*, and each resulting clone was sequenced. Purified amplification products were sequenced successively in both directions step by step on an ABI3130xl capillary sequencer. Forward and reverse sequences were assembled and edited using CodonCode Aligner v.3.5.6 (CodonCode Corporation).

Mitogenome annotation and characterization

The mitogenomes were annotated using the MITOS webserver (http://mitos.bioinf.uni-leipzig.de/index.py, [57]), DOGMA [58], and MFannot (http://megasun.bch.umontreal.ca/cgi-bin/mfannot/mfannotInterface.pl). All predicted PCGs were checked for stop codons and manually elongated by comparison with predicted open reading frames (ORF) as implemented in Geneious R7 v.7.1.3 (Biomatters Ltd.), and by comparison with homologous insect sequence alignments. For the tRNA prediction, we additionally used ARWEN v.1.2.3 [59] and tRNAscan-SE v.1.21 (http://lowelab.ucsc.edu/tRNAscan-SE/, [60]). All annotated mitogenomes were visualized with OrganellarGenomeDRAW (http://ogdraw.mpimp-golm.mpg.de/cgi-bin/ogdraw.pl, [61]) and manually edited.

Nucleotide contents for were retrieved using Geneious R7 v.7.1.3. The AT and GC composition skewness of all mayfly specimens were calculated as follows: AT-skew = (A - T) / (A + T), and GC-skew = (G - C) / (G + C) [62]. To correct biases in the AT content due to incomplete mitogenomes mostly missing the AT-rich region, we removed all AT-rich regions and the two rRNAs including the five close by tRNAs (*trnL1-trnM*) i.e. common_sequence data set and recalculated the base pair compositions.

Phylogenetic reconstruction

Phylogenetic reconstructions were inferred with different approaches: Maximum Likelihood (ML) using the program RAxML v.8.1.5 [63] with subsequent RogueNaRok analysis, and Bayesian inference (BI) with MrBayes v.3.2.2 [64]. We used two different taxa sets: all_taxa set and optimized_taxa set (details see below). The amino acid sequences were aligned using the program MAFFT v.7.050b under the default settings (L-INS-I algorithm with default settings, [65]). The best-fitting models of molecular evolution for each PCG were estimated with ProtTest v.2.4 [24], using the Bayesian information criterion (BIC).

The ML reconstructions were calculated with the program RAxML conducting 1000 bootstrap searches under the GAMMA model of rate heterogeneity. As models of molecular evolution, we used for each PCG the best-fitting models as inferred by ProtTest (MtArt + Γ + I for *atp6*, *cox1*, *cox2*, *cox3*, *cob*, *nad1*, *nad2*, *nad4*, *nad5*; and MtREV + Γ for *atp8*, *nad3*, *nad4L*, *nad6*). For the first tree reconstruction, the collembolan species *Tetrodontophora bielanensis* (NC_002735) was set as outgroup (all_taxa set). All inferred bootstrap trees and the best-known ML tree were used for the RogueNaRok analysis. After pruning the rogue taxa, we calculated a second ML phylogeny, using the so called optimized_taxa set, excluding the inferred rogue taxa and specimens with LBA sensu Bergsten [66]. After the first phylogenetic analyses based on the all_taxa set, we manually selected taxa with long branches (*Aposthonia japonica* (Embioptera), and *Challia fletcheri* (Dermaptera) sensu Wan et al. [67]) and excluded them. For the tree reconstructions, we used identical settings as above except that we excluded the outgroup since it was identified as rogue taxa.

For the MrBayes analysis, all individual amino acid sequence alignments were concatenated using a Python script (fasta_concat.py, https://github.com/srutschmann/python_scripts). We implemented the MtREV model of evolution for each PCG separately, and unlinked the frequencies, gamma distributions, substitution rates and the proportion of invariant sites across partitions. Two independent analyses of four MCMC chains were run for each data set, including 7 × 10^6 generations for the all-taxa set and 8 × 10^6 generations for the optimized_taxa set, and 25% burn-in for both runs.

Availability of supporting data

Newly generated sequences are available on GenBank (GU936203, GU936204, and XXXXXXXX).

Additional files

Additional file 1: Table S1. - Universal primers for Sanger sequencing and PCR amplification conditions. (Note: These data will be provided by Changfa Zhou, zhouchangfa@njnu.edu.cn).

Additional file 2: Figure S1. - **(a)** Bayesian inference reconstruction based on mitochondrial genomes data using the concatenated amino acid sequences of the all_taxa data set (see Table 2). Filled circles indicate well-supported nodes; whereby dark grey circles represent Bayesian posterior probability (BPP) = 1, and light grey circles BPP > 0.95. Scale bar indicates substitutions per site. Coloured taxa represent differences to the maximum-likelihood (ML) phylogeny; whereby the red highlighted orders Ephemeroptera and Odonata are shown in (b), and alternative clustering of the blue taxa is indicated by dotted lines. **(b)** Maximum-likelihood (ML) phylogeny of selected orders that differ from (a) and different relationships of the mayfly specimens.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SR and MTM conceived the study; PC and CZ obtained and analyzed the *Baetis rutilocylindratus* and *Habrophlebiodes zijinensis* genomes; SR and MTM gathered and analyzed the *C. dipterum* genome, performed all combined analyses, and drafted the manuscript. All authors made contributions to subsequent revisions.

Acknowledgements

We are grateful to D. H. Funk from the Stround Water Research Center for providing the *Cloeon dipterum* specimens, S. Mbedi and K. Preuß for their great help with laboratory work, G. Glöckner, C. Mazzoni, and S. Simon for constructive comments about the analysis of the mitochondrial genomes and the study design, P. Rutschmann and the Freie Universität Berlin, Germany for providing access to high-performance computing clusters. Additionally, we want to thank all members of our research groups for constructive working environments.

Author details

[1]Leibniz-Institute of Freshwater Ecology and Inland Fisheries (IGB), Müggelseedamm 301, 12587 Berlin, Germany. [2]Berlin Center for Genomics in Biodiversity Research (BeGenDiv), Königin-Luise-Straße 6-8, 14195 Berlin, Germany. [3]The Key Laboratory of Jiangsu Biodiversity and Biotechnology, College of Life Sciences, Nanjing Normal University, Nanjing 210046, China.

References

1. Trautwein MD, Wiegmann BM, Beutel R, Kjer KM, Yeates DK: **Advances in insect phylogeny at the dawn of the postgenomic era**. *Annu Rev Entomol* 2012, **57**(1):449-468.
2. Yeates DK, Cameron SL, Trautwein M: **A view from the edge of the forest: recent progress in understanding the relationships of the insect orders**. *Australian Journal of Entomology* 2012, **51**(2):79-87.
3. Hovmöller R, Pape T, Källersjö M: **The Palaeoptera Problem: Basal Pterygote Phylogeny Inferred from 18S and 28S rDNA Sequences**. *Cladistics* 2002, **18**:313-323.
4. Mallatt J, Giribet G: **Further use of nearly complete 28S and 18S rRNA genes to classify Ecdysozoa: 37 more arthropods and a kinorhynch**. *Mol Phylogenet Evol* 2006, **40**:772-794.
5. Ogden T: **The problem with "the Paleoptera Problem:" sense and sensitivity**. *Cladistics* 2003, **19**(5):432-442.
6. Regier JC, Shultz JW, Zwick A, Hussey A, Ball B, Wetzer R, Martin JW, Cunningham CW: **Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences**. *Nature* 2010, **463**(7284):1079-1083.
7. Simon S, Strauss S, von Haeseler A, Hadrys H: **A phylogenomic approach to resolve the basal pterygote divergence**. *Mol Biol Evol* 2009, **26**(12):2719-2730.

8. Thomas JA, Trueman JW, Rambaut A, Welch JJ: **Relaxed phylogenetics and the palaeoptera problem: resolving deep ancestral splits in the insect phylogeny**. *Syst Biol* 2013, **62**(2):285-297.
9. Giribet G, Ribera C: **A Review of Arthropod Phylogeny: New Data Based on Ribosomal DNA Sequences and Direct Character Optimization**. *Cladistics* 2000, **16**(2):204-231.
10. Wheeler WC, Whiting M, Wheeler QD, Carpenter JM: **The phylogeny of the extant hexapod orders**. *Cladistics* 2001, **17**(2):113-169.
11. Ishiwata K, Sasaki G, Ogawa J, Miyata T, Su ZH: **Phylogenetic relationships among insect orders based on three nuclear protein-coding gene sequences**. *Mol Phylogenet Evol* 2011, **58**(2):169-180.
12. Kjer KM, Carle FL, Litman J, Ware J: **A molecular phylogeny of Insecta**. *Arthropod Syst Phylogeny* 2006, **64**:35-44.
13. von Reumont BM, Jenner RA, Wills MA, Dell'ampio E, Pass G, Ebersberger I, Meyer B, Koenemann S, Iliffe TM, Stamatakis A *et al*: **Pancrustacean phylogeny in the light of new phylogenomic data: support for Remipedia as the possible sister group of Hexapoda**. *Mol Biol Evol* 2012, **29**(3):1031-1045.
14. Zhang J, Zhou C, Gai Y, Song D, Zhou K: **The complete mitochondrial genome of *Parafronurus youi* (Insecta: Ephemeroptera) and phylogenetic position of the Ephemeroptera**. *Gene* 2008, **424**(1-2):18-24.
15. Lin CP, Chen MY, Huang JP: **The complete mitochondrial genome and phylogenomics of a damselfly, *Euphaea formosa* support a basal Odonata within the Pterygota**. *Gene* 2010, **468**(1-2):20-29.
16. Kjer KM: **Aligned 18S and insect phylogeny**. *Syst Biol* 2004, **53**(3):506-514.
17. Misof B, Niehuis O, Bischoff I, Rickert A, Erpenbeck D, Staniczek A: **Towards an 18S phylogeny of hexapods: accounting for group-specific character covariance in optimized mixed nucleotide/doublet models**. *Zoology* 2007, **110**(5):409-429.
18. von Reumont BM, Meusemann K, Szucsich NU, Dell'Ampio E, Gowri-Shankar V, Bartel D, Simon S, Letsch HO, Stocsits RR, Luan YX *et al*: **Can comprehensive background knowledge be incorporated into substitution models to improve phylogenetic analyses? A case study on major arthropod relationships**. *BMC Evol Biol* 2009, **9**:119.
19. Whitfield JB, Kjer KM: **Ancient rapid radiations of insects: challenges for phylogenetic analysis**. *Annu Rev Entomol* 2008, **53**:449-472.
20. Baurain D, Brinkmann H, Philippe H: **Lack of resolution in the animal phylogeny: closely spaced cladogeneses or undetected systematic errors?** *Mol Biol Evol* 2007, **24**(1):6-9.
21. Rokas A, Carroll SB: **Bushes in the tree of life**. *PLoS Biol* 2006, **4**(11):e352.
22. Whitfield JB, Lockhart PJ: **Deciphering ancient rapid radiations**. *Trends in ecology & evolution* 2007, **22**(5):258-265.
23. Abascal F, Posada D, Zardoya R: **MtArt: a new model of amino acid replacement for Arthropoda**. *Mol Biol Evol* 2007, **24**(1):1-5.
24. Abascal F, Zardoya R, Posada D: **ProtTest: selection of best-fit models of protein evolution**. *Bioinformatics* 2005, **21**(9):2104-2105.

25. Bensasson D, Zhang D, Hartl DL, Hewitt GM: **Mitochondrial pseudogenes: evolution's misplaced witnesses**. *Trends in ecology & evolution* 2001, **16**(6):314-321.
26. Rogers HH, Griffiths-Jones S: **Mitochondrial pseudogenes in the nuclear genomes of *Drosophila***. *PLoS ONE* 2012, **7**(3):e32593.
27. Song H, Buhay JE, Whiting MF, Crandall KA: **Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified**. *Proc Natl Acad Sci U S A* 2008, **105**(36):13486-13491.
28. Cameron SL: **Insect Mitochondrial Genomics: Implications for Evolution and Phylogeny**. *Annu Rev Entomol* 2014, **59**(1):95-117.
29. Boore JL: **Animal mitochondrial genomes**. *Nucleic Acids Res* 1999, **27**(8):1767-1780.
30. Saito S, Tamura K, Aotsuka T: **Replication origin of mitochondrial DNA in insects**. *Genetics* 2005, **171**(4):1695-1705.
31. Zhang D-X, Hewitt GM: **Insect mitochondrial control region: a review of its structure, evolution and usefulness in evolutionary studies**. *Biochem Sys Ecol* 1997, **25**:99-120.
32. Boore JL, Collins TM, Stanton D, Daehler LL, Brown WM: **Deducing the pattern of arthropod phylogeny from mitochondrial DNA rearrangements**. *Nature* 1995, **376**(6536):163-165.
33. Lee EM, Hong MY, Kim MI, Kim MJ, Park HC, Kim KY, Lee IH, Bae CH, Jin BR, Kim I: **The complete mitogenome sequences of the palaeopteran insects *Ephemera orientalis* (Ephemeroptera: Ephemeridae) and *Davidius lunatus* (Odonata: Gomphidae)**. *Genome* 2009, **52**(9):810-817.
34. Li D, Qin J-C, Zhou C-F: **The phylogeny of Ephemeroptera in Pterygota revealed by the mitochondrial genome of *Siphluriscus chinensis* (Hexapoda: Insecta)**. *Gene* 2014, **545**(1):132-140.
35. Tang M, Tan M, Meng G, Yang S, Su X, Liu S, Song W, Li Y, Wu Q, Zhang A *et al*: **Multiplex sequencing of pooled mitochondrial genomes--a crucial step toward biodiversity analysis using mito-metagenomics**. *Nucleic Acids Res* 2014.
36. Barber-James HM, Gattolliat JL, Sartori M, Hubbard MD: **Global diversity of mayflies (Ephemeroptera, Insecta) in freshwater**. *Hydrobiologia* 2008, **595**(1):339-350.
37. Gillett CP, Crampton-Platt A, Timmermans MJ, Jordal BH, Emerson BC, Vogler AP: **Bulk de novo mitogenome assembly from pooled total DNA elucidates the phylogeny of weevils (Coleoptera: Curculionoidea)**. *Mol Biol Evol* 2014, **31**(8):2223-2237.
38. Timmermans MJ, Roelofs D, Marien J, van Straalen NM: **Revealing pancrustacean relationships: phylogenetic analysis of ribosomal protein genes places Collembola (springtails) in a monophyletic Hexapoda and reinforces the discrepancy between mitochondrial and nuclear DNA markers**. *BMC Evol Biol* 2008, **8**(1):83.
39. Wilkinson M: **Majority-rule reduced consensus trees and their use in bootstrapping**. *Mol Biol Evol* 1996, **13**(3):437-444.
40. Aberer AJ, Krompass D, Stamatakis A: **Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice**. *Syst Biol* 2013, **62**(1):162-166.
41. Felsenstein J: **Cases in Which Parsimony or Compatibility Methods Will Be Positively Misleading**. *Systematic zoology* 1978, **27**(4):401-410.

42. Hedtke H, Townsend TM, M HD: **Resolution of phylogenetic conflict in large datasets by increased taxon sampling**. *Syst Biol* 2006, **55**:522-529.

43. Pons J, Bauzà-Ribot MM, Jaume D, Juan C: **Next-generation sequencing, phylogenetic signal and comparative mitogenomic analyses in Metacrangonyctidae (Amphipoda: Crustacea)**. *BMC Genomics* 2014, **15**:566.

44. Simon C, Frati F, Beckenbach A, Crespi B, Liu H, Flook P: **Evolution, Weighting, and Phylogenetic Utility of Mitochondrial Gene-Sequences and a Compilation of Conserved Polymerase Chain-Reaction Primers**. *Annals of the Entomological Society of America* 1994, **87**(6):651-701.

45. Silvestre D, Dowton M, Arias MC: **The mitochondrial genome of the stingless bee *Melipona bicolour* (Hymenoptera, Apidae, Meliponini): sequence, gene organisation and a unique tRNA translocation event conserved across the tribe Meliponini**. *Genet Mol Biol* 2008, **31**:451-460.

46. Simon S, Narechania A, DeSalle R, Hadrys H: **Insect phylogenomics: exploring the source of incongruence using new transcriptomic data**. *Genome Biol Evol* 2012, **4**(12):1295-1309.

47. Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J, Flouri T, Beutel RG *et al*: **Phylogenomics resolves the timing and pattern of insect evolution**. *Science* 2014, **346**(6210):763-767.

48. Ogden TH, Gattolliat JL, Sartori M, Staniczek AH, Soldan T, Whiting MF: **Towards a new paradigm in mayfly phylogeny (Ephemeroptera): combined analysis of morphological and molecular data**. *Syst Entomol* 2009, **34**(4):616-634.

49. Simon S, Hadrys H: **A comparative analysis of complete mitochondrial genomes among Hexapoda**. *Mol Phylogenet Evol* 2013, **69**(2):393-403.

50. Meusemann K, von Reumont BM, Simon S, Roeding F, Strauss S, Kuck P, Ebersberger I, Walzl M, Pass G, Breuers S *et al*: **A phylogenomic approach to resolve the arthropod tree of life**. *Mol Biol Evol* 2010, **27**(11):2451-2464.

51. Ogden TH, Whiting MF: **Phylogeny of Ephemeroptera (mayflies) based on molecular evidence**. *Mol Phylogenet Evol* 2005, **37**(3):625-643.

52. Nishihara H, Satta Y, Nikaido M, Thewissen JG, Stanhope MJ, Okada N: **A retroposon analysis of Afrotherian phylogeny**. *Mol Biol Evol* 2005, **22**(9):1823-1833.

53. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**(17):3389-3402.

54. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2009, **25**(14):1754-1760.

55. Li H, Durbin R: **Fast and accurate long-read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2010, **26**(5):589-595.

56. Salmela L, Schröder J: **Correcting errors in short reads by multiple alignments**. *Bioinformatics* 2011, **27**(11):1455-1461.

57. Bernt M, Donath A, Juhling F, Externbrink F, Florentz C, Fritzsch G, Putz J, Middendorf M, Stadler PF: **MITOS: improved de novo metazoan mitochondrial genome annotation**. *Mol Phylogenet Evol* 2013, **69**(2):313-319.

58. Wyman SK, Jansen RK, Boore JL: **Automatic annotation of organellar genomes with DOGMA**. *Bioinformatics* 2004, **20**(17):3252-3255.

59. Laslett D, Canbäck B: **ARWEN: a program to detect tRNA genes in metazoan mitochondrial nucleotide sequences**. *Bioinformatics* 2008, **24**(2):172-175.

60. Lowe TM, Eddy SR: **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence**. *Nucleic Acids Res* 1997, **25**(5):955-964.

61. Lohse M, Drechsel O, Kahlau S, Bock R: **OrganellarGenomeDRAW--a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets**. *Nucleic Acids Res* 2013, **41**(Web Server issue):W575-581.

62. Perna NT, Kocher TD: **Patterns of nucleotide composition at fourfold degenerate sites of animal mitochondrial genomes**. *J Mol Evol* 1995, **41**(3):353-358.

63. Stamatakis A: **RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies**. *Bioinformatics* 2014, **30**(9):1312-1313.

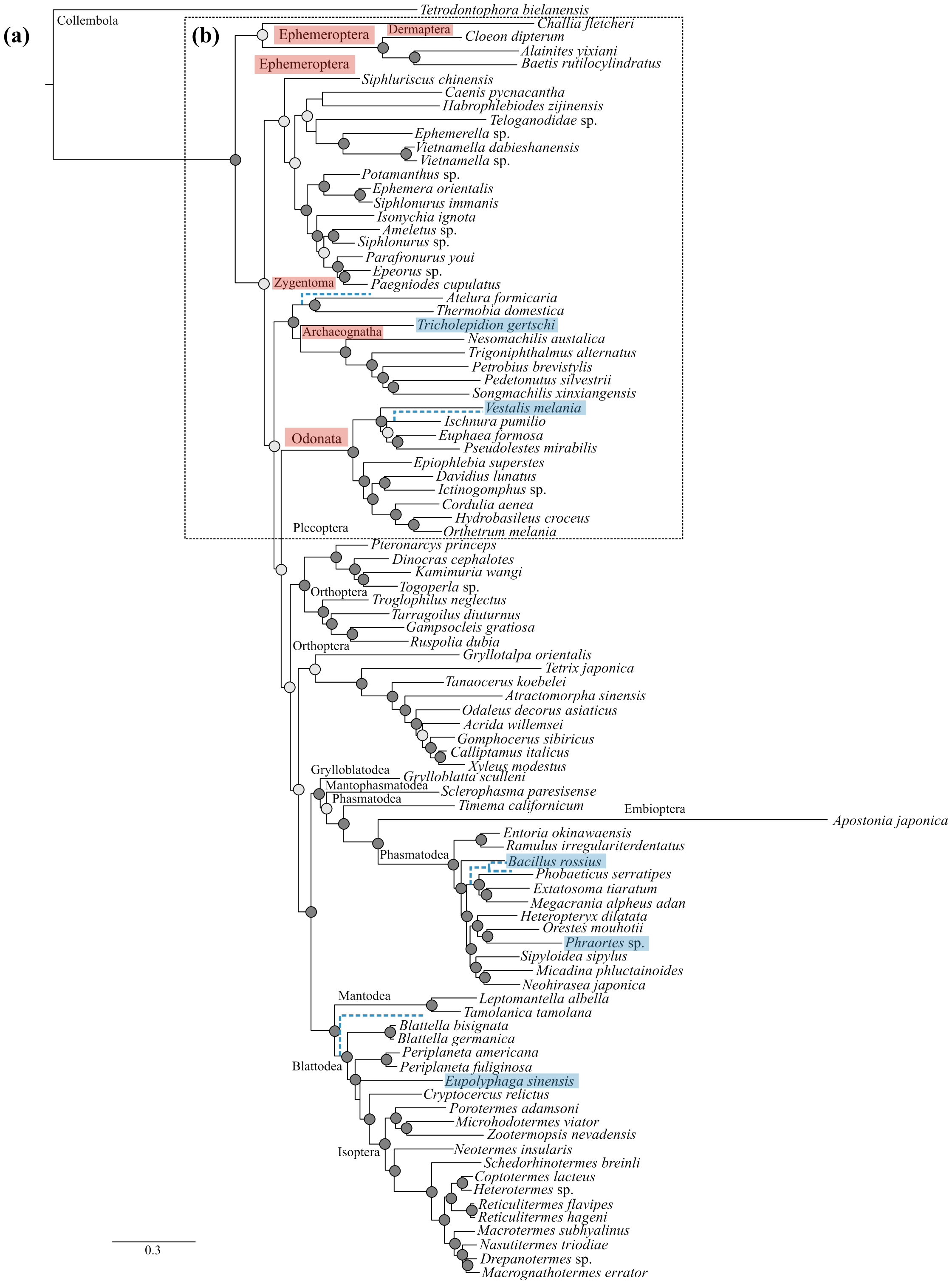
64. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP: **MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space**. *Syst Biol* 2012, **61**(3):539-542.

65. Katoh K, Standley DM: **MAFFT multiple sequence alignment software version 7: improvements in performance and usability**. *Mol Biol Evol* 2013, **30**(4):772-780.

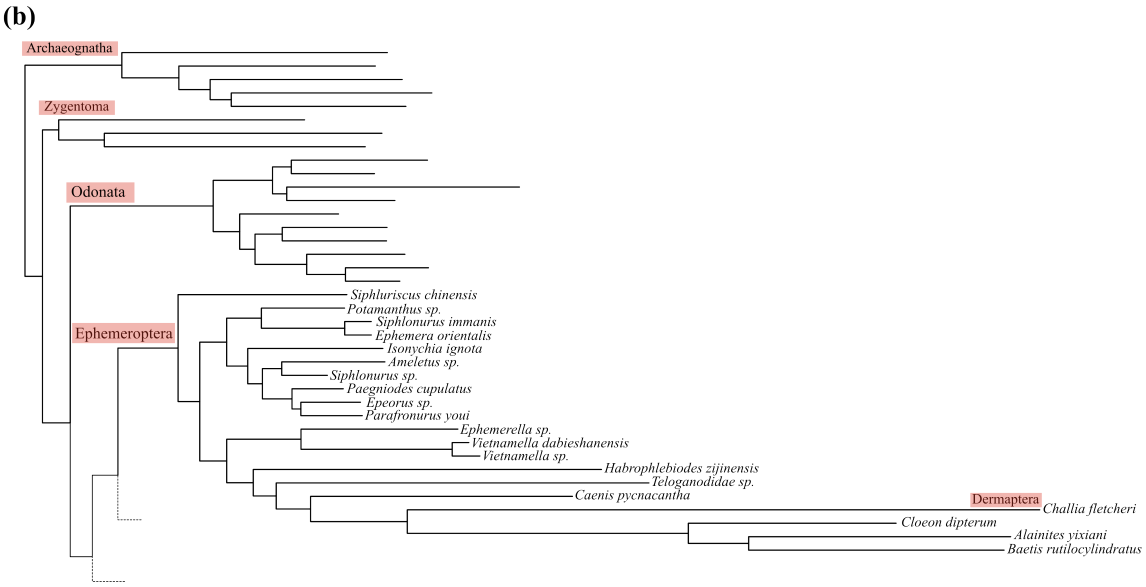
66. Bergsten J: **A review of long-branch attraction**. *Cladistics* 2005, **21**:163-193.

67. Wan X, Kim MI, Kim MJ, Kim I: **Complete Mitochondrial Genome of the Free-Living Earwig, *Challia fletcheri* (Dermaptera: Pygidicranidae) and Phylogeny of Polyneoptera**. *PLoS One* 2012, **7**(8):e42056.

Additional files



Additional file 2: Figure S1 (See legend on next page.)



(See figure above and on previous page.)

Additional file 2: Figure S1 (a) Bayesian inference reconstruction based on mitochondrial genomes data using the concatenated amino acid sequences of the all_taxa data set (see Table 2). Filled circles indicate well-supported nodes; whereby dark grey circles represent Bayesian posterior probability (BPP) = 1, and light grey circles BPP > 0.95. Scale bar indicates substitutions per site. Coloured taxa represent differences to the maximum-likelihood (ML) phylogeny; whereby the red highlighted orders Ephemeroptera and Odonata are shown in (b), and alternative clustering of the blue taxa is indicated by dotted lines. **(b)** Maximum-likelihood (ML) phylogeny of selected orders that differ from (a) and different relationships of the mayfly specimens.

4 Conclusions

My PhD research has led to a number of important insights into the evolution of mayflies, showing that global mayfly diversification is a result of complex processes, including allopatric and sympatric speciation processes with an important role of dispersal across large distances. The role of dispersal ability and gene flow among species appears to be different for species groups occurring in lentic freshwater habitats (*C. dipterum* s.l.) and lotic freshwater habitats (*B. canariensis* s.l. and *B. pseudorhodani* s.l.). Clearly, this work has shown that all species groups posses dispersal abilities. It seems that the species groups of *Baetis* possess more ‘intermediate’ dispersal abilities and presumably adapt better to the more ecologically stable lotic freshwater habitats as supported by the island-endemic-species pattern (Figure I) and their morphological diversification i.e. relative length of gills, shape of the labial palp, and the ornamentation of the tergites (***Chapter 1,*** Gattolliat et al. unpublished). On the other hand, the species of the *C. dipterum* s.l. species group disperse more frequently, resulting in wider geographic distributions with trans-oceanic colonization pathways and fewer local adaptations to their habitats (Figure II, ***Chapter 1***). However, the role of local habitat adaptations in small minnow mayflies remains to be addressed in future studies.

Drivers of Speciation

I found complex colonization pathways of small minnow mayflies rather than unidirectional dispersal from a continental source (sensu Monaghan et al. 2005). All small minnow mayflies showed dispersal abilities on different geographic scales, but dispersal was more pronounced in the lentic *C. dipterum* s.l. species group (compare Figures I, II). Recent studies have also uncovered surprising dispersal abilities for other *Cloeon* species on Madagascar (Monaghan et al. 2005) and in Brazil (Salles et al. 2014), and large-scale geographic dispersal events from other insects possessing aquatic larvae (e.g, caddisflies, Groeneveld et al. 2007; e.g., damselflies, Gíslason et al. 2015). However, all these studies were based primarily on mtDNA markers, leaving any conclusion derived from it ambiguous. Very small genetic differences (measured by short branch lengths) between trans-oceanic occurring *C. dipterum* s.l. specimens, i.e. AZ1 on Greece, the Azores and

U.S, suggest a large, continuous geographic occurrence, including *Cloeon* species that disperse frequently (FIGURE II).

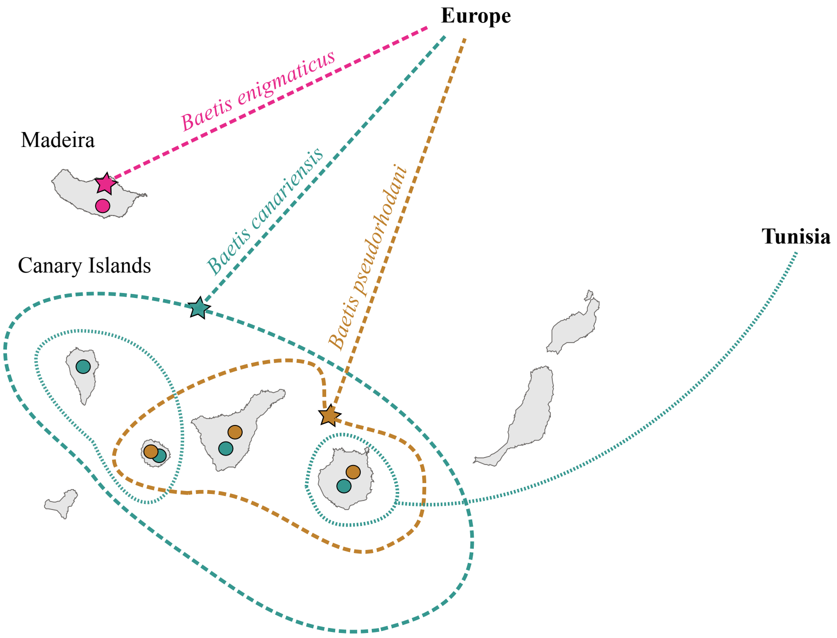


FIGURE I Schematic overview of distributions of the three *Rhodobaetis* species groups *Baetis enigmaticus* (pink), *Baetis canariensis* s.l. (turquoise), and *Baetis pseudorhodani* (brown) based on three mitochondrial DNA markers; whereby the species group *Baetis canariensis* s.l. clustered into two distinct clades. Dots represent the delineated species based on DNA-taxonomy and morphological differences. Stars indicate the colonization events on Madeira and the Canary Islands, which could not be assigned to a single island due to the lack of phylogenetic resolution by using mitochondrial DNA markers.

Dispersal abilities in *Cloeon* might clearly be enhanced by unique reproduction modes (ovoviviparity, Gillies 1949; parthenogenesis, Harker 1997) and the relatively long-lived adult life stage. The complex colonization pathways of *Cloeon* species seem to be driven by random dispersal opportunities such as wind- and anthropogenic-associated dispersal. However, a species-level phylogeny including more than five individuals per population (i.e. five representatives of AZ1 from the each of the Azorean islands, five from the U.S., and five from Greece; sensu Heled and Drummond 2010) will be needed to shed more light on the population demographic structure since the inferred colonization pathways rely on a concatenated sequence alignment, and thus may not reflect the fine-scale relationships of the individual populations. For example, based on the results from ***CHAPTER 2*** it is not clear whether the populations in Europe, the Azores or the U.S. are the most ancestral ones within the species AZ1. It could be that the species dispersed from Greece via the Azores to the U.S., the opposite way, or a mix. For a definite resolution, a phylogenetic reconstruction would be needed. The dispersal ability of the two *Baetis* species groups seems to be less strong as evidenced by the more narrow geographic distributions of the individual species, i.e. for *B. pseudorhodani*: *B.* sp3 on La Gomera, *B.* sp4 on Tenerife, and *B. pseudorhodani* on Gran Canaria (***CHAPTER 1***). However, the results should be interpreted carefully since they are based on mtDNA markers. Nevertheless, the fact that the results from the phylogenetic relationships of *C. dipterum* s.l. based on the large nDNA markers and the mtDNA tree were congruent, is strong evidence that the *Baetis* tree based on mtDNA might also reflect the true species relationships.

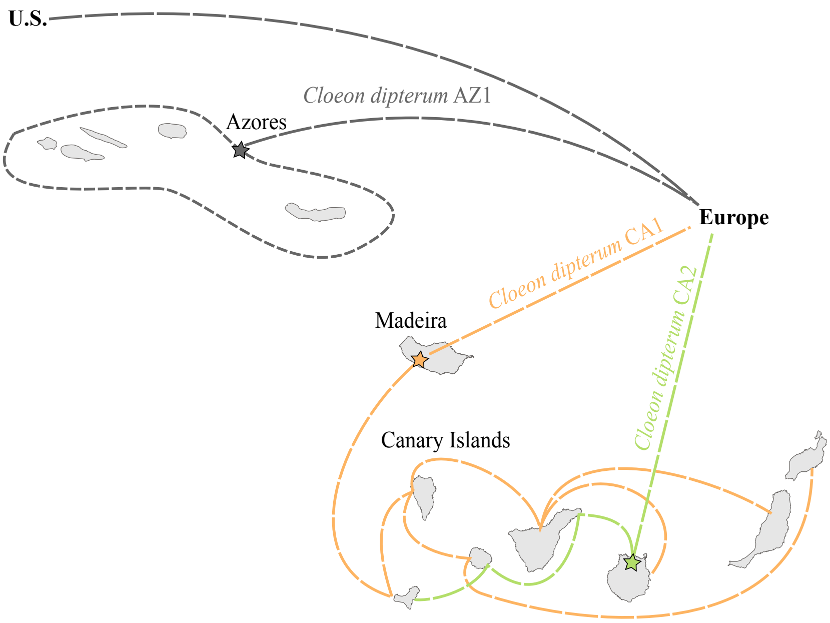


FIGURE II Schematic colonization routes of the three *Cloeon dipterum* s.l. species: AZ1 (grey), CA1 (orange), and CA2 (green) based on 59 nuclear DNA markers. Stars indicate the proposed colonization events on the Azorean archipelago, Madeira, and Gran Canaria.

Early allopatric speciation followed by colonization processes as main driver for the Macaronesian mayfly fauna seems likely because the divergence date estimates of all occurring species are far younger than those of the geological island ages (***CHAPTER 1***). Moreover, molecular clock calibrations based on species trees inferred from nDNA should result in even younger ages than inferences based on gene trees (e.g., Carstens and

Knowles 2007; Edwards 2009; Burbrink and Pyron 2011). The overall dispersal abilities contradict early biogeographic studies suggesting ancient continental separations as the main driver for global mayfly diversification (Edmunds 1972; Edmunds 1975) assuming very limited dispersal abilities (Brittain 1982; Hershey et al. 1993; Monaghan et al. 2002; Caudill 2003; Hughes et al. 2003). However, most studies on mayfly dispersal are based on small minnow mayflies and studies of other families are sparse (e.g., Heptageniidae, Vuataz et al. 2013).

My work underlines the existence of cryptic species groups, increasing the number of the previously seven recognized species to twelve species (***CHAPTERS 1 & 2***). In total, five clades contained sister taxa from geographically separated origins: *B. rhodani* s.l. (including *B. atlanticus*), occurring on the European mainland and Madeira, *B. canariensis* s.l. and *B. rhodani* s.l. on the Canaries and Tunisia, and three *C. dipterum* s.l. species, including CA1 on Madeira and the Canary Islands, CA2 on the Canaries, and AZ1 on the Azores, Greece, and the U.S. (***CHAPTERS 1 & 2***). The clustering of *B. atlanticus* within a geographically widely distributed *B. rhodani* s.l. species group evidences a synonymy of *B. atlanticus*, which was also found for several *Rhithrogena* species that were historically thought to be locally endemic and were later found to be more widespread (Vuataz et al. 2011). The mayfly species distribution patterns on the northeastern Macaronesian islands are in agreement with previous work on Madagascar, where all occurring mayfly species except *C. smaeleni* were found as endemic (Gattolliat and Rabeantoandro 2002; Elouard et al. 2003). Moreover, several recent studies reported cryptic species complexes (but see ***CHAPTERS 1***).

Incomplete sampling from the Iberian Peninsula and the neighboring North Africa may limit the reconstruction of the colonization pathways. I assume that the *C. dipterum* s.l. species CA1 successfully colonized the North African mainland via Lanzarote. A close link between the North African and Canarian mayfly fauna is further evidenced by the sister taxa relationship of *B. canariensis* s.l. and *B. rhodani* s.l. from Tunisia (***CHAPTER 1***, FIGURE 3; FIGURE I). Especially the acquirement of specimens from the North African mainland is of high interest to investigate the phylogenetic relationships of the known *Cloeon* species and their North African counterparts. However, a North African origin of the species groups occurring on the Macaronesian archipelagos seems to be unlikely from the phylogenetic reconstructions (***CHAPTERS 1 & 2***) due to the overall short branch lengths. However, I expect a close link due to the short geographic distance. Moreover, I collected the ‘North African species’ *B. nigrescens*, which was last reported in 1995 by Malmqvist and colleagues, during the large fieldwork in March 2014 on the two Canarian islands of Gran Canaria and La Gomera. The occurrence of this species on the Canaries substantiates the hypothesis of a close link between the Canarian and African mayfly fauna (Rutschmann et al. in preparation). The question remains how the *C. dipterum* s.l. species could have obtained their clear genetic structuring despite the lack of obvious gene flow barriers (e.g. geographical or physical isolation) and the presence of frequent dispersal. Generally, it would be interesting to investigate possible discordances among mtDNA and nDNA markers that could evidence contrasting patterns of female and male gene flow (Chan and Levin 2005), which may be enhanced in *Cloeon* due to the specialized reproduction modes (i.e. parthenogenesis and ovoviviparity).

The very recent and frequent dispersal within the *C. dipterum* s.l. species group sensu Monaghan et al. (2005) supports the hypothesis that evolutionary unstable habitats may select for ecological traits that promote dispersal in insects. The facts that four *C. dipterum* s.l. species pairs co-occur in identical permanent and temporary aquatic habitats (***CHAPTER 2***, FIGURE 4) indicate that very strong ecological selection drove speciation in this species group as for example the tolerance of the larvae to harsh conditions such as periods of anoxia (Nagell and Fagerström 1978; Nagell 1980; 1981), high water temperatures (Cianciara 1979; 1980; Soldán and Thomas 1983), and high salinity (Soldán and Zahrádková 2000) among others. It seems that the *Baetis* species groups have evolved more local adaptations, measured by the higher genetic distances, the clear allopatric species distribution pattern and the clear morphological differences within each species group. Moreover, based on the molecular clock calibration, using the a priori substitution rate, it seems that *Baetis* has colonized the Canary Islands before *Cloeon* and thus a possible higher local adaptation might result from the longer time period or a better ability for adapting to very sparse, specialized habitats.

SPECIATION PROCESS

I found strong evidences for allopatric speciation in both *Baetis* and *Cloeon* genera, whereby the speciation process in *Baetis* seems to be in a more derived stage whereas the speciation within the *C. dipterum* s.l. species group seems to be shaped by allopatric and possibly sympatric modes. The two species groups *B. canariensis* s.l. and *B.*

pseudorhodani s.l. have radiated into four, respective three, endemic species, showing genetic and morphological differences (taxonomic review for *sp3*-*sp7*, in preparation; (FIGURE I)), which all together support the biological species concept sensu DeSalle et al. (2005). I hypothesize that the co-occurring *Rhodobaetis* species groups show different feeding types that could be investigated by morphological analyses of the mouthparts, metagenomic sequencing of the gut microbiome, and stable isotope analyses. Taken all this criteria together, the mode of speciation could be investigated more detailed. Moreover, several previous studies on *B. rhodani* s.l. have highlighted the occurrence of multiple lineages on small geographic scales i.e. same stream/drainage (Rebora et al. 2005; Williams et al. 2006).

In contrast, none of the *Cloeon* species was restricted to occur on one island (FIGURE I) and it seems that the speciation mode of *Cloeon* is more complex, possible combining allopatric and sympatric speciation and there are two main explanations. First, the *Cloeon* species could have evolved before they have colonized the Canary Islands respective Europe (i.e. allopatric speciation on for example the European mainland followed by subsequent colonization). Likewise, they could have arisen from ancestral, extinct species on the mainland rather than trans-oceanic dispersal. However, this remains speculative and is not really testable without fossil records. The fact that representatives from different species from identical freshwater habitats do not cluster monophyletically supports the latter (***CHAPTER 2***). A time-calibrated species level phylogeny based on comprehensive taxon sampling, including unsampled regions from the European and African mainland where source populations are thought to occur, and an assessment of ancestral species distribution areas with subsequent fine-scale diversification analyses (e.g. disparity-through-time analyses i.e. modeling the evolution of specific characters over time in comparison to neutral evolution (sensu Toussaint et al. 2015) would shed more light on whether ancient allopatric speciation might be important. Likewise, the *Cloeon* species possibly have evolved in parallel due to strong natural selection via sympatric speciation. This hypothesis could be tested with large population genetic analyses and species tree reconstructions using several representatives of each species per habitat. The shared nuclear haplotypes between the different species indicate incomplete lineage sorting as well as a possible presence of gene flow (***CHAPTER 2***, FIGURE 2). Moreover, it would be very interesting to test for gene flow, being more likely to occur in the absence of geographic barriers unless the species occupy different ecological niches.

For example, it might be that they have different feeding and/or life history preferences that can not be investigated by only using phylogenetic approaches. Taken together, it might be that the gene flow of both genera is equal or that *Cloeon* species have more gene flow and/or disperse more and at the same time the *Baetis* species adapt faster to the local habitats.

PALAEOPTERA PROBLEM

I found that mitogenomes might be the 'key' genetic marker to investigate ancient radiations. All phylogenetic reconstructions resulted in highly supported phylogenetic trees (***CHAPTER 3***, FIGURE 3). The support for two alternative clustering hypotheses (***CHAPTER 3***) is not surprising, given the fact that several other studies have failed to resolve the early pterygote divergence. Based on the results of ***CHAPTER 2*** we gained evidences that within Baetidae the occurrence of numts might be less problematic. It seems that mitogenomes are at the moment the most appropriate marker for deep phylogenetic splits. Moreover, because also a very recent phylogenomic study by Misof et al. (2014), including 1,478 protein-coding genes among a wide range of insects, failed to resolve the 'Palaeoptera proble'm suggests that the use of more genetic markers might also not solve the this problem. Disadvantages of large phylogenomic studies are the often considerable amount of missing data, which can decrease accuracy and result in unsupported phylogenetic tree relationships (Kearney 2002; Hartmann and Vision 2008; Lemmon et al. 2009). I hypothesize that some nuclear protein-coding genes might be subject to recombination and thus at this state of knowledge, mtDNA genes are better suited to resolve ancient splits because they are more conserved among a wide range of taxa and also due to their smaller population size given their uniparental inheritance. However, in general nuclear protein-coding genes are the 'upcoming-preferred' markers in the field of phylogenetics but future studies might be needed to identify those nDNA markers most suitable for ancient splits so that they can be sequenced.

FUTURE PERSPECTIVES

The co-occurring *C. dipterum* s.l. species and the parallel-evolved *Baetis* lineages provide an ideal study system for future studies focusing on speciation within freshwater habitats. In fact, by comparatively studying these species groups it would be possible to investigate diversification within and among lotic and lentic freshwater habitats. Understanding speciation is difficult, partly because it is not observable in real-time and the process of speciation is often very complex and varies between different taxa. Moreover, we need several layers of data sets (e.g., genomics, ecology, and behavioral biology) to get a better understanding of how speciation acts.

It will be important for future studies to complement the genetic data with ecological and morphological data such as habitat parameters providing evidences for local adaptations and morphometrics to test for phenotypic divergence. Given the occurrence of multiple *Cloeon* species pairs in very different habitats i.e., species CA1 and CA2 co-occurring in artificial ponds on El Hierro as well as natural ponds on Tenerife, I assume that hydromorphological characteristics i.e., habitat structure, water depth or water temperature of the freshwater habitat do not play a role for the diversification of *Cloeon* species. This is supported by the ability of the species to colonize new habitats easily and the fact that they can survive under harsh conditions i.e, anoxic conditions. Moreover, the preferred microhabitats of *Cloeon* include vascular plants or mosses and riparian vegetation, and they do not like stony substrates (Bauernfeind and Soldán 2012). Therefore I assume that the *Cloeon* species colonize most of the artificial habitats on the Macaronesian islands by chance. All this taken together, it seems most unlikely that the different *Cloeon* species develop local adaptations to their habitats. On the other hand it might be possible that they develop adaptations e.g. to some specific chemical parameters and that we are not able to trace at the moment.

The question remains how species pairs co-occur in the same freshwater habitats on small scale and whether there is gene flow occurring or not. So far, the data suggest, that the species are genetically distinct and thus only very little gene flow might be present. Therefore, the species must be clearly separated. Most likely seems a shift in the life history traits i.e. flying periods, which can be seasonal uni- (one flying period) or bi- (two flying periods) or polyvioline (several flying periods) depending on the climatic zone (review by Bauernfeind and Soldán 2013). Since the flying period is very short (up to two weeks) it might be, that co-occurring species have small shifts in their flying periods. Another possibility would be a specialization on their food sources. The larvae are generally herbivorous, feeding on detritus, diatoms, and small algae. It might be possible that the species feed on different types of diatoms or detritus.

On the other hand, the two *Baetis* species groups seem to have adapted to hydromorphological characteristics, mostly water depth and velocity. Also morphological diversification seems to be more advanced. . For example, the *Baetis* the species group *B. pseudorhodani* seems to occur in higher depths with strong velocity and also is bigger in size (Gattolliat et al. unpublished). Thus it might be that the adaptation to hydromorphological characteristics promotes species diversity

Once we have a better knowledge of the life history, ecological and morphological traits, it would be interesting to look at specific genes/polymorphisms to link them with speciation. Thereby, the most widely used approach is restriction-digest based methods such as restriction-site associated (RAD) DNA sequencing (RAD-seq; Baird et al. 2008). For RAD-seq, restriction enzyme digestion is used to cut the genome at defined sites into smaller fragments, which will then be sequenced. RAD-seq produces shorter reads ideal to target polymorphisms in both coding and non-coding regions and could be used to detect regions that are under natural selection such as for example genes in *Cloeon* that allow this species group to live in temporary freshwater habitats.

A phylogenomic approach would be interesting in order to reconstruct the phylogenetic relationships among the different populations i.e., between AZ1 in Greece, AZ1 in the Azores, and AZ1 in U.S. With the possibilities to scan large proportions of the genome it now became feasible to move away from the use of few genes towards tackling genome-wide, i.e. hundreds/thousands genes, patterns to specifically determine genes or single nucleotide polymorphisms under selection and associate them with phenotypic traits. In order to generate the necessary genomics data, target enrichment approaches like anchored hybrid enrichment sequencing (AE-seq; Lemmon et al. 2012) could be applied. Anchored enrichment is based on the enrichment of genomic DNA for target regions prior to sequencing, whereby probes for the target regions are designed from known sequences and hybridized to genomic DNA. Using this approach, it would be feasible to sequence more nDNA markers for more species within shorter time. For example, Leaché et al. (2014) used a hybrid phylogenetic approach of Sanger sequencing and AE-seq for species tree estimation in African *Agama* lizards. Thereby, they obtained data of 215 nDNA markers from 23 species by using AE-seq. In contrast to the approach I applied in ***CHAPTER 2**,* this method can be applied simultaneously to more distantly related taxa and would presumably produce long sequence fragments of *Cloeon* and *Baetis*.

REFERENCES

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA. 2008. Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. PLoS One. 3:e3376.

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

Brittain JE. 1982. Biology of mayflies. Rev Entomol. 27:119-147.

Burbrink FT, Pyron RA. 2011. The impact of gene-tree/species-tree discordance on diversification-rate estimation. Evolution. 65:1851-1861.

Carstens BC, Knowles LL. 2007. Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from *Melanoplus* grasshoppers. Syst Biol. 56:400-411.

Caudill CC. 2003. Measuring dispersal in a metapopulation using stable isotope enrichment: high rates of sex-biased dispersal between patches in a mayfly metapopulation. Oikos. 101:624-630.

Chan KM, Levin SA. 2005. Leaky prezygotic isolation and porous genomes: rapid introgression of maternally inherited DNA. Evolution. 59:720-729.

Cianciara S. 1979. Life cycles of *Cloeon dipterum* (L.) in natural environment. Pol Arch Hydrobiol. 4:501-513.

Cianciara S. 1980. Food preference of *Cloeon dipterum* (L) larvae and dependence of their development and growth on the type of food. Pol Arch Hydrobiol. 27:143-160.

DeSalle R, Egan MG, Siddall M. 2005. The unholy trinity: taxonomy, species delimitation and DNA barcoding. Philos Trans R Soc Lond B Biol Sci. 360:1905-1916.

Edmunds GF. 1972. Biogeography and Evolution of Ephemeroptera. Annual Review of Entomology. 17:21-42.

Edmunds GF. 1975. Phylogenetic Biogeography of Mayflies. Annals of the Missouri Botanical Garden. 62:251-231.

Edwards SV. 2009. Is a new and general theory of molecular systematics emerging? Evolution. 63:1-19.

Elouard J-M, Gattolliat J-L, Sartori M. 2003. Ephemeroptera, mayflies. In: Goodman SM, Benstead JP editors. The natural history of Madagascar, University of Chicago Press, p. 639-645.

Gattolliat J-L, Rabeantoandro SZ. 2002. The genus *Cloeon* (Ephemeroptera, Baetidae) in Madagascar. Mitteilungen der Schweizerischen Entomologischen Gesselschaft, Bulletin de la Sociéte´ Entomologique Suisse. 74:195-209.

Gillies MT. 1949. Notes on some Ephemeroptera Baetidae from India and South-East Asia. Trans R Entomol Soc Lond. 161-177.

Gíslason GM, Hannesdóttir ER, Munoz SS, S P. 2015. Origin and dispersal of *Potamophylax cingulatus* (Trichoptera: Limnephilidae) in Iceland. Freshwater Biology. 60:387-394.

Groeneveld LF, Clausnitzer V, Hadrys H. 2007. Convergent evolution of gigantism in damselflies of Africa and South America? Evidence from nuclear and mitochondrial sequence data. Mol Phylogenet Evol. 42:339-346.

Harker JE. 1997. The role of parthenogenesis in the biology of two species of mayfly (Ephemeroptera). Freshwater Biology. 37:287-297.

Hartmann S, Vision TJ. 2008. Using ESTs for phylogenomics: can one accurately infer a phylogenetic tree from a gappy alignment? BMC Evol Biol. 8:95.

Heled J, Drummond AJ. 2010. Bayesian inference of species trees from multilocus data. Mol Biol Evol. 27:570-580.

Hershey AE, Pastor J, Peterson BJ, Kling GW. 1993. Stable isotopes resolve the drift paradox for Baetis mayflies in an Arctic River. Ecology. 2315-2325.

Hughes JM, Mather PB, Hillyer MJ, Cleary C, Peckarsky B. 2003. Genetic structure in a montane mayfly *Baetis bicaudatus* (Ephemeroptera: Baetidae), from the Rocky Mountains, Colorado. Freshwater Biology. 2149-2162.

Kearney M. 2002. Fragmentary taxa, missing data, and ambiguity: mistaken assumptions and conclusions. Syst Biol. 51:369-381.

Leaché AD, Wagner P, Linkem CW, Böhme W, Papenfuss TJ, Chong RA, Lavin BR, Bauer AM, Nielsen SV, Greenbaum E*, et al.* 2014. A hybrid phylogenetic-phylogenomic approach for species tree estimation in African *Agama* lizards with applications to biogeography, character evolution, and diversification. Mol Phylogenet Evol. 79:215-230.

Lemmon AR, Brown JM, Stanger-Hall K, Lemmon EM. 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. Syst Biol. 58:130-145.

Lemmon AR, Emme SA, Lemmon EM. 2012. Anchored Hybrid Enrichment for Massively High-Throughput Phylogenomics. Syst Biol. 61:727-744.

Malmqvist B, Nilsson AN, Báez M. 1995. Tenerife's freshwater macroinvertebrates: status and threats (Canary Islands, Spain). Aquatic Conservation-Marine and Freshwater Ecosystems. 5:1-24.

Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J, Flouri T, Beutel RG*, et al.* 2014. Phylogenomics resolves the timing and pattern of insect evolution. Science. 346:763-767.

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

Monaghan MT, Spaak P, Robinson CT. 2002. Population genetic structure of 3 Alpine stream insects: influences of gene flow, demographics, and habitat fragmentation. J North Am Benthol Soc. 21:114-131.

Nagell B. 1980. Overwintering strategy of Cloeon dipterum (L.) larvae. In: Flannigan JF, Marshall KE editors. Advances in Ephemeroptera Biology. New York and London, Plenum Press.

Nagell B. 1981. Overwintering strategy of two closely related forms of *Cloeon* (*dipterum*?) (Ephemeroptera) from Sweden and England. Freshwater Biology. 11:237-244.

Nagell B, Fagerström T. 1978. Adaptations and resistance to anoxia in *Cloeon dipterum* (Ephemeroptera) and *Nemoura cinerea* (Plecoptera). Oikos. 30:95-99.

Rebora M, Lucentini L, Palomba A, Panara F, Gaino E. 2005. Genetic differentiation among populations of *Baetis rhodani* (Ephemeroptera, Baetidae) in three Italian streams. Italian Journal of Zoology. 72:121-126.

Salles FF, Gattolliat JL, Angeli KB, De-Souza MR, Goncalves IC, Nessimian JL, Sartori M. 2014. Discovery of an alien species of mayfly in South America (Ephemeroptera). ZooKeys. 1-16.

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

Soldán T, Zahrádková S. 2000. Ephemeroptera of the Czech Republic: Atlas of distribution. In: Helesic J, Zahrádková S editors. Fauna Aquatica Europea Centralis. Masaryk University Brno, Biodiversity Working Group.

Toussaint EF, Condamine FL, Hawlitschek O, Watts CH, Porch N, Hendrich L, Balke M. 2015. Unveiling the diversification dynamics of australasian predaceous diving beetles in the cenozoic. Syst Biol. 64:3-24.

Vuataz L, Sartori M, Gattolliat JL, Monaghan MT. 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.

Vuataz L, Sartori M, Wagner A, Monaghan MT. 2011. Toward a DNA taxonomy of Alpine *Rhithrogena* (Ephemeroptera: Heptageniidae) using a mixed Yule-coalescent analysis of mitochondrial and nuclear DNA. PLoS ONE. 6:e19728.

Williams HC, Ormerod SJ, Bruford MW. 2006. Molecular systematics and phylogeography of the cryptic species complex *Baetis rhodani* (Ephemeroptera, Baetidae). Mol Phylogenet Evol. 40:370-382.

CURRICULUM VITAE