Assessing the potential of RAD-sequencing to resolve phylogenetic relationships within species radiations: the fly genus *Chiastocheta* (Diptera: Anthomyiidae) as a case study

Tomasz Suchan[1,2], Anahí Espíndola[3*], Sereina Rutschmann[1,4*], Brent C. Emerson[5,6], Kevin Gori[7], Christophe Dessimoz[1,8,9,10], Nils Arrigo[1], Michał Ronikier[2], Nadir Alvarez[1]

[1] Department of Ecology and Evolution, University of Lausanne, Switzerland
[2] W. Szafer Institute of Botany, Polish Academy of Sciences, Kraków, Poland
[3] Department of Biological Sciences, University of Idaho, Moscow, ID, USA
[4] Department of Biochemistry, Genetics and Immunology, University of Vigo, Spain
[5] Island Ecology and Evolution Research Group, Instituto de Productos Naturales y Agrobiología (IPNA-CSIC), La Laguna, Tenerife, Canary Islands, Spain
[6] School of Biological Sciences, University of East Anglia, Norwich, UK
[7] Department of Veterinary Medicine, University of Cambridge, UK
[8] Center for Integrative Genomics, University of Lausanne, Switzerland
[9] Department of Genetics, Evolution & Environment and Department of Computer Science, University College London, UK
[10] Swiss Institute of Bioinformatics, Lausanne, Switzerland

[*] these authors are considered as second co-authors

corresponding authors:
Tomasz Suchan (t.suchan@botany.pl), W. Szafer Institute of Botany, Polish Academy of Sciences, Lubicz 46, 31-512 Kraków, Poland
Nadir Alvarez (nadir.alvarez@unil.ch), Department of Ecology and Evolution, University of Lausanne, Biophore Building, 1015 Lausanne, Switzerland; fax: +41 21 692 41 65

Abstract

Determining phylogenetic relationships among recently diverged species has long been a challenge in evolutionary biology. Cytoplasmic markers, which have been widely used notably in the context of molecular barcoding, have not always proved successful in resolving such phylogenies, but phylogenies for closely related species have been resolved at a much higher detail in the last couple of years with the advent of next-generation-sequencing technologies and associated techniques of reduced genome representation. Here we examine the potential and limitations of one of such techniques — Restriction-site Associated DNA (RAD) sequencing, a method that produces thousands of (mostly) anonymous nuclear markers, in disentangling the phylogeny of the fly genus *Chiastocheta* (Diptera: Anthomyiidae). This genus encompasses seven described species of seed predators, which have been widely studied in the context of their ecological and evolutionary interactions with the plant *Trollius europaeus* (Ranunculaceae). So far, phylogenetic analyses using mitochondrial markers failed to resolve monophyly of most of the species from this recently diversified genus, suggesting that their taxonomy may need to be revised. However, relying on a single, non-recombining molecule and ignoring potential incongruences between mitochondrial and nuclear loci may provide incomplete account of a lineage history. In this study, we apply both classical Sanger sequencing of three mtDNA regions and RAD-sequencing, for reconstructing the phylogeny of the genus. Contrasting with results based on mitochondrial markers, RAD-sequencing analyses retrieved the monophyly of all seven species, in agreement with the morphological species assignment. We found robust nuclear-based species assignment of individual samples, and low levels of estimated contemporary gene flow among them. However, despite recovering species' monophyly, interspecific relationships varied depending on the set of RAD loci considered, producing contradictory topologies. Moreover, coalescence-based phylogenetic analyses revealed low supports for most of the interspecific relationships. Our results indicate that despite the higher performance of RAD-sequencing in terms of species trees resolution compared to cytoplasmic markers, reconstructing inter-specific relationships may lie beyond the possibilities offered by large sets of RAD-sequencing markers in cases of strong gene tree incongruence.

Keywords: coalescent analysis; DNA barcoding; maximum likelihood; mito-nuclear incongruence; single nucleotide polymorphisms; quartet inference

1. Introduction

Recently diverged lineages pose a problem for traditional phylogenetic approaches that typically rely on a small set of relatively slowly evolving loci (DeFilippis 2000), often lacking resolution at narrower evolutionary scales (Cariou et al. 2013). In addition, complex processes such as incomplete lineage sorting (Avise et al. 2008; Maddison & Knowles 2006; Pollard et al. 2006) and gene flow among species (Leaché et al. 2013) increase incongruences among gene trees and topological deviations from the species tree (Dengan & Rosenberg 2009; Maddison 1997). This is especially true for lineages that have undergone rapid radiations, in which ancestral polymorphisms sorted idiosyncratically into the descendant taxa through short evolutionary nodes (Avise et al. 2008), and in cases where subsequent evolutionary events may blur phylogenetic signal (Whitfield & Kjer 2008; Whitfield & Lockhart 2007). Sampling more loci has been shown to be a promising approach in such cases (Rokas & Carroll 2005; Townsend et al. 2011; Wielstra et al. 2014; Williams et al. 2013), but the spectrum of genetic markers developed for phylogeny estimation is still limited (Whitfield & Kjer 2008).

Next-generation sequencing approaches, particularly reduced representation genome sequencing (Davey et al. 2011), offer the possibility to sample thousands of genomic markers from non-model species. Among them, Restriction site-Associated DNA (RAD; Baird et al. 2008) techniques rely on the sequencing of short DNA fragments flanking restriction sites, generating random anonymous genomic markers, homologous across the analyzed samples (Andrews et al. 2016; Davey & Blaxter 2010). From a phylogenetic perspective, an important aspect of RAD markers is the rise in the proportion of 'null alleles' as genome divergence across samples increases. This phenomenon is caused by random mutations occurring in the restriction sites that decrease the numbers of shared RAD loci among taxa, resulting in data matrices containing large amounts of missing data (Cariou et al. 2013; Chattopadhyay et al. 2014; Gautier et al. 2013). However, using an in-silico approach Rubin et al. (2012) and Cariou et al. (2013) have shown that RAD-seq data can be used successfully to resolve species relationships that transcend timescales up to 60 Mya (million years ago). Experimentally sampled RAD datasets have been applied to reconstruct phylogenetic relationships, mostly among recently diverged taxa (e.g., Eaton & Ree 2013; Harvey et al. 2016; Jones et al. 2013; Leaché et al. 2015; Nadeau et al. 2013; Wagner et al. 2013), with fewer studies involving more distantly related ones, even up to even 80 Mya (e.g., Cruaud et al. 2014; Eaton et al. 2016; Herrera and Shank 2016; Hipp et al. 2014; Pante et al. 2015). Although these genomic datasets improved phylogenetic inferences for groups that were ambiguous using classical markers (e.g., Escudero et al. 2014; Hipp et al. 2014), the potential utility of RAD loci for resolving more complex phylogenetic histories, such as those where historical introgression has occurred or those associated with incomplete lineage sorting, remains still poorly explored (Combosch & Vollmer 2015; Eaton & Ree 2013). Moreover, the use of RAD datasets as markers for evolutionary genetics has recently been heavily discussed (Lowry et al. 2017; McKinney et al. 2017).

In this study, we test the utility of RAD-sequencing to recover phylogenetic relationships in a genus of seed parasitic pollinators of *Trollius europaeus* (Ranunculaceae) — flies from the genus *Chiastocheta* Pokorny, 1889 (Diptera: Anthomyiidae). Here, sequencing of mitochondrial markers failed to reveal the monophyly and phylogenetic relationships among previously morphologically described species (Després et al. 2002; Espíndola et al. 2012). This discordance between morphology and mitochondrial phylogeny has been interpreted as a call for a taxonomic revision, and a possible reconsideration of conclusions from previous ecological and evolutionary studies (Espíndola et al. 2012). However, several mechanisms may cause mitochondria not to track species evolution (Funk & Omland 2003) and, indeed, there are many cases where mitochondrial and nuclear gene trees have been shown to be incongruent (e.g., Govindarajulu et al. 2015; Phillips et al. 2013; Seehausen et al. 2003). As relying on a single, non-recombining molecule may provide a misleading account of a species history (Ballard & Whitlock 2004), utilizing a large set of independent nuclear loci (sampled through RAD-sequencing) should allow us to test the monophyly of the morphologically described species and resolve phylogenetic relationships among them. Whether or not molecular markers are able to reveal scenarios of rapid radiations is still an open question (Giarla & Esselstyn 2015). In these, identifying a single species tree might lie beyond analytical possibilities due to pervasive conflicts among the gene trees, particularly when population sizes are large and speciation events happen at a higher rate than the mutation-drift equilibrium, eventually producing conflicting topologies. In order to explore gene and species trees, we applied both a concatenation-based phylogenetic approach (i.e., RAxML; Stamatakis 2014) and a coalescence-based inference method (i.e., SVDquartets; Chifman & Kubatko 2014) to a RAD-seq dataset encompassing specimens from 51 European populations, representative of the seven recognized *Chiastocheta* morphospecies. In order to examine the extent to which different combinations of RAD loci may produce distinct species trees, we used a newly developed algorithm that performs loci binning, using dissimilarity levels among phylogenetic patterns retrieved at single loci (treeCl; Gori et al. 2016). We also applied population genetics clustering algorithms (i.e., STRUCTURE; Pritchard et al. 2000) as a control. Eventually, we compared our results to those obtained with classical phylogenetic inference based on concatenation of three mitochondrial regions.

2. Materials and methods

2.1 Study system

The center of origin and diversity of *Chiastocheta* has been inferred to be the Western Palearctic, where seven fly species are involved in nursery pollination interactions with *Trollius europaeus* L. (Espíndola et al. 2012; Pellmyr 1989, 1992; Suchan et al. 2015). These seven morphologically delimited European *Chiastocheta* species, namely *C. dentifera* Hennig 1953; *C. inermella* (Zetterstedt, 1838); *C. lophota* Karl, 1943; *C. macropyga* Hennig, 1953; *C. rotundiventris* Hennig, 1953; *C. setifera* Hennig, 1953 and *C. trollii* (Zetterstedt, 1838) are ecologically very similar and often sympatric (Collin 1954; Hennig 1976; Michelsen 1985; Zetterstedt 1845; V. Michelsen pers. comm.). In his monograph of this plant-pollinator interaction, Pellmyr (1992) discussed another species, *C. abruptiventris* as a northern vicariant of *C. rotundiventris*, a taxon not supported by previous molecular studies (Espíndola et al. 2012) and never formally described. Although all *Chiastocheta* reproduce within the flowers of *T. europaeus*, with potential cross-species mating possibilities, no putative hybrids have been observed based on genital morphology (T. Suchan and A. Espíndola, pers. obs.).

Although the species are well defined morphologically, mitochondrial phylogenies recovered only three monophyletic clades – *C. rotundiventris*, *C. dentifera*, and *C. lophota* (Després et al. 2002; Espíndola et al. 2012), and suggested a polyphyletic origin for *C. inermella* and *C. setifera* (Després et al. 2002; Espíndola et al. 2012), with *C. macropyga* and *C. trollii* being paraphyletic (Espíndola et al. 2012). Molecular dating placed the most recent common ancestor of all European species at the end of the Pliocene (2-3.4 Ma; Després et al. 2002; Espíndola et al. 2012), and indicated that most diversification events occurred within the last 1.6 Ma.

2.2 Sampling

Chiastocheta specimens were sampled from 51 European populations during spring and summer 2006, 2007, and 2008 (Table 1; maps on Fig. S1). The flies were killed and preserved in 70% ethanol and stored at room temperature until DNA extraction. Collected specimens were identified to morphospecies following Hennig (1976) and unpublished keys (V. Michelsen). All identifications were confirmed by an expert (V. Michelsen, Natural History Museum of Denmark, Copenhagen), as the taxonomical revision of the genus is not yet published.

2.3 Sequencing mitochondrial regions and RAD markers

DNA was extracted from insect legs using a DNeasy Blood and Tissue Kit (Qiagen, Hilden, Germany), following the manufacturer's instructions. We amplified three mitochondrial regions: COI, COII, and the ultra-variable D-loop (control) region. We followed Espíndola et al. (2012) for sequencing of the COI and COII regions. For D-loop we used primers TM-N-193 and SR-J-14612 as desxcribed in Simon et al. (1994) as described by Espíndola et al. (2012) with the following modification of the PCR program: 5 min at 95°C, followed by 35 cycles of 1 min at 95°C, 1 min of annealing at 55°C and 2 min of elongation at 60°C, and 5 min of final elongation at 60°C. PCR products were sequenced at Macrogen Inc. (South Korea) and Fasteris SA (Switzerland). Chromatograms were visually corrected on ChromasPro 1.41 (Technelysium Pty. Ltd.). Alignment was performed using MUSCLE algorithm (Edgar 2004) in Geneious 10.1.3

(Biomatters, Auckland, New Zeland) and gaps with more than 50% missing data in the D-loop region were removed. Additionally, a dataset with D-loop removed was analyzed. Double digest RAD (ddRAD) libraries were prepared according to Mastretta et al. (2014), a modified protocol of Peterson et al. (2012), without performing the size-selection of DNA fragments, and other minor modifications (see Supporting Information). The enzymes used for DNA digestion were SbfI and MseI. Libraries were sequenced at the Lausanne Genomic Technologies Facility (Switzerland) on three lanes of the HiSeq2500 instrument (Illumina, San Diego, USA) using a 2x100 bp paired-end reads protocol. For RAD-sequencing we introduced technical replicates for optimizing *de novo* assembly and controls for the effects of sequencing errors and allele dropout on the final results (Mastretta-Yanes et al. 2015; samples with "REPL" suffix in Table S1), and DNA extraction replicates from the fly thoracic muscle (in order to control for flies' body contamination with pollen; samples with "MUS" suffix in Table S1).

2.4 RAD-seq loci assembly

Two important considerations for *de novo* RAD loci assembly are the parameters for clustering orthologous loci, while filtering out paralogs (Eaton 2014; Mastretta et al. 2015). If the sequence similarity required to consider sequence as orthologs is set too high, real heterozygous alleles will be split into more than one cluster, therefore creating false homozygous loci (Harvey et al. 2015). On the other hand, if the similarity is set too low, this will result in paralogous sequences being clustered together. Several methods were proposed for filtering out such sequences from the final dataset, including ploidy filtering (removing clusters that have more than two sequences per individual) and filtering out highly variable loci (Eaton 2014; Ilut et al. 2014). As there are no general guidelines for fine-tuning the parameters mentioned above, we empirically tested how the different clustering parameters affected the final dataset. Finally, we chose the dataset with clustering parameter that maximized the loci overlap between pairs of technical replicates (see below). Loci overlap among samples and pairs of technical replicates were calculated using the RADami 1.0 library in R (Hipp et al. 2014).

Read demultiplexing and *de novo* assembly of RAD loci was performed using the pyRAD 3.0.1 package (Eaton 2014), based on an alignment-clustering algorithm. This approach allows for indel variation among more diverged specimens. Moreover, it also allows for lower similarity among the clustered reads, making it well-suited for phylogenetic-scale analyses. First, the reads were demultiplexed according to the in-line 6-nucleotides barcode present at the beginning of each sequenced fragment, while allowing for one mismatch. Only reads with the restriction site present were retained for further analyses. All nucleotides with Phred quality score lower than 20 were converted to unknown bases and reads with more than four unknown sites were removed from the dataset. Reads were then clustered within and between individuals, with a minimum number of six reads to form a cluster and sequence similarity of 75%, 80%, 85%, 90%, and 95%. Possible clustered paralogs or repetitive sequences were removed by filtering out the loci that had more than five variable positions per locus or more than 10 shared polymorphic sites in a locus among individuals, and the loci for which more than two alleles were present per individual. Finally, datasets were produced by retaining the loci present in a minimum of 10, 20, and 100 individuals and compared for the total number of loci, proportion of missing data, loci overlap among replicates, and the mean number of individuals per locus.

2.5 Phylogenetic analyses

We performed Maximum-likelihood (ML) analyses using RAxML (Stamatakis 2014) with rapid bootstrap analyses and extended majority-rule consensus tree automatic bootstrap stopping criterion, following search for the best-scoring ML tree. The mitochondrial regions were partitioned using the PartitionFinder 2.1.1 software (Lanefar et al. 2016). Analyses were performed in the RAxML 8.2.4, for the mitochondrial dataset. For the dataset consisting of COI, COII, and D-loop regions the GTR+G+I model with all three nucleotide positions on coding genes considered as separate partitions and D-loop as a fourth partition. For the dataset consisting of COI and COII the GTR+G model with the first two nucleotide positions considered as a first and third nucleotide position considered as a second partition. Analyses for the RAD dataset were performed using the GTRCAT model in the RAxML 8.2.10 version on the CIPRES cluster (San Diego CA, USA; Miller et al. 2010). For the RAD-based dataset, replicated samples were retained in the phylogenetic analyses and the concatenated matrix was considered as a single partition. Additionally, ML analyses of the RAD datasets with other clustering parameters were performed in order to evaluate how this parameter affects the tree topology. The trees were rooted with *C. rotundiventris* as an outgroup, as identified previously (Després et al. 2002; Espíndola et al. 2012).

To account for the effects of incongruence among nuclear loci on the inferred phylogenies — for instance resulting from incomplete lineage sorting, we applied the method of Gori et al. (2016) implemented in the treeCl package (http://git.io/treeCl). Because the majority of RAD loci had sparse coverage over the individuals, we kept only loci present in more than 100 individuals for this part of analysis. The ML phylogenenies were first calculated for every locus using the GTR+G model as implemented in RAxML 8.1.11 (Stamatakis 2014). Then, pairwise geodesic distances between all the resulting single-locus phylogenies were measured, and the trees were grouped based on the distance matrix using spectral clustering (a protocol hereafter referred to as binning). The number of bins was estimated using the nonparametric bootstrapping stopping criterion. Support for each branch in each topology was calculated using aBayes in PhyML (Anisimova et al. 2011). We also analyzed the log-likelihood improvement when analyzing the data with *n*+1 splits vs. *n* splits, compared to the null expectation (i.e. random loci clustering).

Additionally, we applied a coalescent-based inference method using SVDquartets (Chifman and Kubatko 2014) as implemented in PAUP* v.4a150 and v.4a151 (Swofford 2002). This method infers the topology among randomly sampled quartets using a coalescent model, and assembles the randomly sampled quartets using a quartet amalgamation method. Breaking the sequence into quartets makes the analysis of large numbers of loci feasible. We randomly sampled the maximum of all possible quartets (i.e. 48,603,900 quartets = 200 taxa) with the multispecies coalescent option and 1,000 bootstrap replicates. The quartets were summarized with the QFM (Reaz et al. 2014) quartet amalgamation program as implemented in PAUP*. Phylogenetic trees were visualized using the ape 3.2 R package (Paradis et al. 2004).

2.6 Population structure

We inferred population structure using the admixture model implemented in STRUCTURE 2.3.4 (Pritchard et al. 2000), without prior population assignment and with allele frequencies correlated among populations. The software uses a Bayesian framework to estimate the likelihood of the data given a number of *a priori* defined *K* population clusters, outputting the likelihood of each sample to belong to each possible cluster. This analysis was performed after removing technical replicates from the dataset, retaining only the loci present in a minimum of 20 individuals, and selecting one random single SNP from each locus. Analyses were run for *K* values ranging between 1 and 8, with a burn-in of 200K cycles, followed by 1M cycles of sampling, with 3 replicates for each *K* value. The optimal *K* value was identified following Evanno et al. (2005), as implemented in STRUCTURE HARVESTER (Earl & vonHoldt 2012). To account for the phylogenetic component in the missing alleles, we ran STRUCTURE with the recessive alleles model, with missing data coded as recessive.

3. Results

3.1 *Chiastocheta* sampling

We analyzed a total of 272 *Chiastocheta* specimens sampled from the entire European range of the genus (see Table 1 and Fig. S1 for maps of the sampled specimens). Most species displayed a broad spatial distribution. Up to six species could be found in one single locality during a single visit (Table 1, mean = 2.7 species per locality, SD = 1.5), confirming the sympatric nature of the species and the existing opportunities for hybridization.

3.2 Sequencing and RAD loci assembly results

After initial screening, 21 samples were removed from the final dataset because of insufficient coverage or technical errors. We successfully analyzed 260 specimens for the mitochondrial dataset (255 for COI, 204 for COII, 141 for the first, and 152 for the second fragment of the D-loop), and 263 for the RAD dataset, while 251 samples were shared between the datasets (Table 1). For the RAD dataset, we also sequenced 22 technical replicates (samples with "REPL" suffix in Table S1), and 11 DNA extraction replicates from the fly muscle vs. extractions from legs (samples with "MUS" suffix in Table S1).

Sequencing of the mitochondrial regions yielded 1132 nucleotide positions for the COI+COII dataset (of which 120 were variable) and 2003 for the COI+COII+D-loop dataset (of which 334 were variable), after alignment and gap filtering. Three runs of RAD sequencing output 552'425'482 of 2 x 100 bp reads, from which 340'598'636 (62%) passed the restriction site and barcode quality filters (Table S1).

After comparing the number of loci, coverage, and overlap of loci among replicates in the obtained datasets (Fig. S2), we chose the dataset with a minimum of 75% sequence similarity required for the sequences to cluster in a locus and a minimum of 20 individuals per locus for the main analyses. This dataset contained 1724 loci after filtering and paralog removal, with 82'782 variable sites. The proportion of missing data in the dataset was 0.84, with a strong phylogenetic component in the distribution of missing loci (Fig. S3a). After sampling one SNP per locus for the STRUCTURE analysis, we obtained 1669 SNPs, of which 159 were bi-allelic.

For the dataset used for assessing loci incongruence in the RAD-seq based phylogeny (see below), we focused on loci present in at least 100 individuals. This resulted in a matrix of 176 loci (among 1724 overall number of loci identified; i.e., 10.2%) with missing data showing much less phylogenetic structuring (Fig. S3b).

3.3 Mitochondrial and nuclear-data phylogenies

The mitochondrial phylogeny on the COI+COII_D-loop dataset (Fig. 1a and Fig. S4a) failed to resolve four of the clades identified based on the RAD-seq data (see below), but retrieved well-supported monophyletic group for *C. rotundiventris* and, to the lesser extent for *C. dentifera* and *C. inermella*, as both of the latter had two specimens placed outside their clades. *C. inermella*, *C. setifera*, and *C. trollii* formed one clade with the species extensively interdispersed and a clade containing mostly *C. macropyga* nested within. As the analysis based based on the reduced COI+COII dataset recovered a similar pattern, except placing *C. lophota* as sister to *C. macropyga*, we refer to the results of the larger dataset in the rest of the paper. Most of the *C. lophota* samples also formed one clade with lower support values. The relationships among samples from the remaining four morphospecies remained unresolved, without clear support for the morphologically described species.

In contrast to the ML mtDNA phylogeny, both ML and SVDquartets analysis of the RAD analysis (Fig. 1b, c and Fig. S4b, c) confirmed monophyly of the seven morphologically defined taxa. RAxML analysis revealed relatively high bootstrap supports (> 90%) for all of the interspecific relationships, except the split between *C. setifera* and the clade (*C. lophota*, (*C. macropyga*, (*C. dentifera*, *C. trollii*))) with bootstrap support > 80%. The split of *C. rotundiventris* into two putative vicariant clades, informally proposed by Pellmyr (1992) — northern *C. abruptriventris* and southern *C. rotundiventris*, was not recovered. SVDquartets analysis also confirmed monophyly of the species, but only the split between *C. dentifera* and *C. trollii* had a bootstrap support > 90%; the clade (*C. macropyga*, (*C. dentifera*, *C. trollii*)) had bootstrap support > 80%; these two clades were the only ones supported by both SVDquartets and the RAxML analyses (Fig. 1c).

Moreover, SVDquartets revealed two well-supported clades within *C. rotundiventris*. These however do not show any pattern of vicariance and often occur together in a single population, thus most likely do not correspond to the two vicariant species of *C. abruptiventris* and *C. rotundiventris* as discussed by Pellmyr (1992). The technical replicates were consistent in the placement of the sample within the proper clade, and most replicates were placed as sister clades with both methods (Fig. S4b,c).

3.4 Incongruence among the RAD-sequencing loci

TreeCl analysis identified, in the most conservative interpretation, at least four clusters of loci, as the largest likelihood improvement was obtained when increasing the number of bins from three to four (Fig. S6). The bin sizes were of 29, 42, 47, and 58 loci, therefore the identified groups were not simply consisting of a few outliers. The trees inferred for the four bins confirmed the monophyly of the analysed species to a large extent, although few individuals appeared outside their expected clades. The largest departure from monophyly was observed for *C. lophota* in the smallest tree consisting of 29 loci (Fig. 2). The trees inferred for each cluster had branch supports for interspecific nodes larger than 95%, and differed substantially in terms of topology and branch lengths. Only one tree, with the largest number of loci (i.e., 58) supported the only clade that was supported by both RAML and SVDquartets analysis (*C. macropyga*, (*C. dentifera*, *C. trollii*)). Except that, the interspecific relationships retrieved with each of the treeCl bins were different than with the concatenated RAxML analysis and SVDquartets analysis (Fig. 1b, c).

3.5 Structure analysis

We found low levels of contemporary introgression, as shown by STRUCTURE analysis. The most likely *K* number of STRUCTURE groups was consistent with the number of morphological species (7), and all samples were assigned to their 'correct' morphospecies (Fig. 1d). Also for lower numbers of clusters, we did not observe signatures of introgression (Fig. S5).

4. Discussion

4.1 Utility and limits of RAD-sequencing for resolving phylogeny of a „difficult" genus

RAD-sequencing successfully discriminated all formally described European *Chiastocheta* species. The robust species delineation is striking when compared to mtDNA-based trees that failed to support monophyly of *C. inermella, C. macropyga*, *C. setifera,* and *C. trollii* (Fig. 1a and Fig. S4a; see also: Després et al. 2002; Espíndola et al. 2012). The ability to recover previously defined morphological species in our dataset, whatever analysis method used (i.e., maximum-likelihood phylogenetic reconstruction using a concatenated matrix with RAxML, coalescence-based phylogenetic inference with SVDquartets, or population-genetics clustering with STRUCTURE), supports the results of a previous simulation study by Hovmöller et al. (2013), that high amounts of missing data, typical for RAD-based datasets, should not interfere with clade (or cluster) identification. Recently, similar conclusions were drawn by Eaton et al. (2016) concerning the SVDquartets method.

In contrast, no consensus could be reached in retrieving inter-specific relationships. Whereas RAxML identified relationships with high bootstrap support in four of the five possible interspecific relationships, only two of them were also supported by the SVDquartets analysis (Fig. 1b,c). Incongruence in the phylogenetic signals associated with different sets of loci could explain the difficulty in resolving these interspecific relationships. When performing loci binning using treeCL (Gori et al. 2016), we found out that different subsets of loci (in our case, the optimal number of bins was equal to four) produced different topologies, while still being largely congruent in the sample assignment into species (Fig. 2).

Short interspecific branches in the resolved phylogenies confirm the conclusions of Espíndola et al. (2012) that most of the species from the *Chiastocheta* genus underwent a recent (less than 1.6 Mya), rapid radiation. These results highlight the fact that in such cases it may be impossible to retrieve some of the phylogenetic relationships among the taxa as fully bifurcating tree, because gene trees may depict different evolutionary histories due to incomplete lineage sorting (Avise et al. 2008; Maddison 1997). This is a limitation shared with classical markers (Walsh et al. 1999) and other NGS approaches (see below), pointing to a possible constitutive limitation in resolving rapid radiations. In rapidly diverging taxa, even the large number of nuclear markers, while being more successful here in recovering species boundaries than mitochondrial markers may not be informative-enough to retrieve all interspecific evolutionary relationships.

The extent to which the above limitation is the result of technical constraints of RAD datasets or a true biological limitation remains to be investigated. RAD-seq targets random, mostly neutral parts of the genome. This results in high number of lineage-specific mutations that bear a strong signal to delineate species or populations – within these fast-evolving parts of the genome, even varying allele copy-numbers (i.e. recent paralogs) can appear as population-specific (Mastretta-Yanes et al. 2014). The downside is however, missing data increases rapidly with evolutionary distance as a result of the loss of restriction sites (Cariou et al. 2013; Chattopadhyay et al. 2014; DaCosta & Sorenson 2016; Gautier et al. 2013; Rubin et al. 2012; Wagner et al. 2013). For instance, Leaché et al. (2015) found differences between phylogenies obtained using RAD-seq vs. target enrichment techniques, whereas other studies have shown the agreement among data types (Manthey et al. 2016). The latter techniques rely on capture of a predefined (Faircloth et al. 2014; McCormack et al. 2012) or random (Suchan et al. 2016, Schmid et al. 2017) subset of loci. By not relying on the presence of restriction sites, and thus having less missing data, enrichment techniques may be better suited for broader phylogenetic scales.

Nevertheless, it has been shown that even with hundreds of conserved loci, known substitution models and several individuals per species, trees with short branches are difficult to resolve, and ML analyses based on concatenated sequences may provide high bootstrap values despite incorrectly resolved topologies (Giarla & Esselstyn 2015; Kubatko & Degnan 2007; but see Gatesy & Springer 2013; Springer & Gatesy 2016; Roch & Warnow 2015). This is exemplified by our study, in which using all RAD loci, we obtained a ML phylogeny with highly supported interspecific nodes, whereas coalescence-based phylogenetic inference did not show strong supports for most of the interspecific relationships. Our exploration of explanations for such a discrepancy using the loci binning approach showed support for at least four different underlying gene tree topologies. In these analyses, we reduced the dataset to a non-random set of loci when filtering for high loci coverage among samples. The retained loci, present in at least 100 analyzed individuals, and with less phylogenetically-structured missing data (see Fig. S3b), should be characterized by lower mutation rates or being under stabilizing selection (Huang & Knowles 2014). Using binning, the best fit to the data was not obtained with a single bin of loci but with four. We could therefore not identify one single evolutionary history of the *Chiastocheta* genus, but rather equally-supported gene trees topologies. Importantly, these different topologies cannot be attributed to a few outlier loci, as their distribution was relatively even across the clusters (29, 58, 42 and 47 loci; Fig. S6), incongruence among these sets possibly impacting maximum-likelihood phylogenetic reconstruction using a concatenated matrix and coalescence-based phylogenetic inference. We have also confirmed that in such cases, ML methods provide elevated bootstrap support values, and that lower bootstrap support values resulting from coalescence-based methods may better reflect the biological uncertainty of interspecific relationships.

4.2 Mitonuclear discordance in the phylogeny of *Chiastocheta*

While our RAD-sequence dataset delineated seven clades, with full agreement with the morphological assignments, mitochondrial data failed to support species monophyly, except for *C. rotundiventris* and, to a lesser extent, *C. lophota* and *C. dentifera*. The other remaining species: *C. inermella*, *C. setifera* and *C. trollii* formed a large clade with the species extensively interdispersed and with the clade consisting mostly of *C. macropyga* nested within (Fig. 1a). Despite, on average, mitochondrial markers should be more suited for capturing relationships among recently diverged lineages, due to an effective population size four times less than that of nuclear genes (assuming neutral processes, equal sex ratios, and unbiased mating systems), and thus shorter coalescence times

(Zink & Barrowclough 2008), analyzing a large dataset of nuclear markers provided more power to discriminate the species in our case.

Mitonuclear discordance patterns can be explained either by the different biological properties of mitochondrial DNA (vegetative segregation, uniparental inheritance, intracellular selection, and reduced recombination; Birky 2001) or differences in the evolutionary histories of nuclear and mitochondrial markers [e.g., direct selection on the mitochondrial genes (Ballard et al. 2007; Ballard & Pichaud 2014; Boratyński et al. 2014; Dowling et al. 2008), incomplete lineage sorting, historical or ongoing gene flow among species, or hybrid speciation]. Indeed, it has been shown before that relying on a single, non-recombining mtDNA molecule may provide a misleading account of a species history (e.g., Ballard & Whitlock 2004; Govindarajulu et al. 2015; Phillips et al. 2013; Seehausen et al. 2003; and reviews by Funk & Omland 2003; Rubinoff & Holland 2005). While investigating the reasons for the mito-nuclear discordance was not within the scope of this paper, we could reject the hypothesis of a contemporary gene flow or hybrid origin of the taxa as responsible for this pattern. We did not detect signature of a genetic mosaic in the—mostly—nuclear RAD data, which would be expected in the case of hybrid origin (Ballard 2000; Brelsford et al. 2011; Mallet 2007; Pollard et al. 2006). Using RAD-sequencing data, the assignment of samples into species was concordant with morphology (Fig. 1b,c and S4b,c) and we did not detect significant levels of contemporary gene flow using population genetics-based approaches (Fig. 1d), despite apparent opportunities for hybridization. Most of *Chiastocheta* occur in sympatry (Fig. S1), they also have very similar biologies, reproducing and spending most of their time on or inside flowers of *Trollius europaeus* (Pellmyr 1989; Suchan et al. 2015). Although a temporal sequence in oviposition has been observed (Després & Jaeger 1999;

Johannesen & Loeschcke 1996; Pellmyr 1989), most species co-occur temporally. Despite these ecological similarities and the relatively young age of the genus (most of the clades emerging less than 1.6 Ma; Espindola et al. 2012), a lack of nuclear evidence for hybridization indicates strong contemporary reproductive barriers among the species.

5. Conclusions

This study demonstrates how a combination of RAD-seq and mtDNA data can provide insights into phylogenies of genera that are poorly resolved using mitochondrial markers alone and reveal complex picture of mitonuclear discordance. It also underlines the limits of RAD-seq-based phylogenies in case of rapid radiations. Our results show that a scenario of rapid radiation can affect many loci across the genome, leading to discordant gene trees, even when using methods controlling for incomplete lineage sorting. This may point to an inherent limitation of using molecular markers to resolve rapid radiations, at least at some of the inter-specific relationships, and suggests that this limitation is not necessarily due to technical issues (e.g. low number of shared markers).

Adding to the body of examples of mito-nuclear discordance (reviewed in Toews & Brelsford 2012), our study warns against relying solely on mitochondrial markers (e.g., COI barcoding; Herbert et al. 2003) for species delimitation, especially when they show incongruence with classical taxonomy. In the case presented here, mitochondrial markers suggested poly- or paraphyly for most species, and proposed the need to review the taxonomy of the genus (Espíndola et al. 2012). When tackled from the genomic point of view, the genetic support of species status for these seven entities was confirmed. Finally, we provide an example of how ML phylogenies based on large concatenated datasets can provide erroneously high bootstrap supports for incorrect or uncertain topologies (Giarla & Esselstyn 2015; Kubatko & Degnan 2007).

Acknowledgments

The authors would like to thank Y. Triponez, N. Magrou, P. Lazarevic, D. Gyurova, R. Lavigne, L. Juillerat, N. Villard, and R. Arnoux for their help during the fieldwork and V. Michelsen for his great help with species determination. We also thank J. Ollerton and J. Pannell for insightful discussions and comments on the manuscript. This work was supported by the grant 'Fonds des Donations' of the University of Neuchâtel (Switzerland) attributed to AE. TS acknowledges funding from the Rectors' Conference of the Swiss Universities (CRUS) through the "Scientific Exchange Program between Switzerland and the new EU member states" (Sciex NMS grant no. 10.116) and from Société Académique Vaudoise (Switzerland). This work was funded by the Swiss National Science Foundation (grants 3100A0-116778, PZ00P3-126624, PP00P3_144870 to N. Alvarez, and grant PP00P3_150654 to CD) and by a grant from Switzerland through the Swiss Contribution to the enlarged European Union (Polish-Swiss Research Program, project no. PSPB-161/2010).

References

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA (2016) Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, 17, 81-92.

Anisimova M, Gil M, Dufayard J-F, Dessimoz C, Gascuel O (2011) Survey of Branch Support Methods Demonstrates Accuracy, Power, and Robustness of Fast Likelihood-based Approximation Schemes. *Systematic Biology*, 60, 685–699.

Avise JC, Robinson TJ, Kubatko L (2008) Hemiplasy: a new term in the lexicon of phylogenetics. *Systematic Biology*, 57(3), 503-507.

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.

Ballard JWO (2000) When one is not enough: introgression of mitochondrial DNA in *Drosophila*. *Molecular Biology and Evolution*, 17, 1126-1130.

Ballard JWO, Melvin RG, Katewa SD, Maas K (2007) Mitochondrial DNA variation is associated with measurable differences in life-history traits and mitochondrial metabolism in *Drosophila simulans*. *Evolution*, 61, 1735-1747.

Ballard JWO, Pichaud N (2014) Mitochondrial DNA: more than an evolutionary bystander. *Functional Ecology*, 28, 218-231.

Ballard JWO, Whitlock MC (2004) The incomplete natural history of mitochondria. *Molecular Ecology*, 13, 729-744.

Birky CW (2001) The inheritance of genes in mitochondria and chloroplasts: laws, mechanisms, and models. *Annual Review of Genetics*, 35, 125-48.

Boratyński Z, Melo-Ferreira J, Alves PC, Berto S, Koskela E, Pentikäinen OT, Mappes T (2014) Molecular and ecological signs of mitochondrial adaptation: consequences for introgression. *Heredity*, 113, 277-286.

Brelsford A, Milá B, Irwin DE (2011) Hybrid origin of Audubon's warbler. *Molecular Ecology*, 20, 2380-2389.

Cariou M, Duret L, Charlat S (2013) Is RAD-seq suitable for phylogenetic inference? An *in silico* assessment and optimization. *Ecology and Evolution*, 3, 846-852.

Chattopadhyay B, Garg KM, Ramakrishnan U (2014) Effect of diversity and missing data on genetic assignment with RAD-Seq markers. *BMC research notes*, 7, 841.

Chifman J, Kubatko L (2014) Quartet inference from SNP data under the coalescent model. *Bioinformatics*, 30:3317-3324.

Collin JE (1954) The genus *Chiastocheta* Pokorny (Diptera: Anthomyiidae). *Proceedings of the Royal Entomological Society London (B)*, 23, 95-102.

Combosch DJ, Vollmer SV (2015) Trans-Pacific RAD-Seq population genomics confirms introgressive hybridization in Eastern Pacific *Pocillopora* corals. *Molecular Phylogenetics and Evolution*, 88, 154-162.

Cruaud A, Gautier M, Galan M, Foucaud J, Sauné L, Genson G, Rasplus JY (2014) Empirical assessment of RAD sequencing for interspecific phylogeny. *Molecular Biology and Evolution*, 31, 1272-1274.

DaCosta JM, Sorenson MD (2016) ddRAD-seq phylogenetics based on nucleotide, indel, and presence–absence polymorphisms: Analyses of two avian genera with contrasting histories. *Molecular Phylogenetics and Evolution*, 94, 122-135.

Davey JL, Blaxter MW (2010) RADSeq: Next-generation population genetics. *Briefings in Functional Genomics*, 9, 416-423.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, 12, 499-510.

DeFilippis VR, Moore WS (2000) Resolution of phylogenetic relationships among recently evolved species as a function of amount of DNA sequence: An empirical study based on woodpeckers (Aves: Picidae). *Molecular Phylogenetics and Evolution*, 16, 143-160.

Degnan, J.H., Rosenberg, N.A., 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology and Evolution* 24, 332–340.

Després L, Jaeger N (1999) Evolution of oviposition strategies and speciation in the globeflower flies *Chiastocheta* spp. (Anthomyiidae). *Journal of Evolutionary Biology*, 12, 822-831.

Després L, Pettex E, Plaisance V, Pompanon F (2002) Speciation in the globeflower fly *Chiastocheta* spp. (Diptera: Anthomyiidae) in relation to host plant species, biogeography, and morphology. *Molecular Phylogenetics and Evolution*, 22, 258-268.

Dowling DK, Friberg U, Lindell J (2008) Evolutionary implications of non-neutral mitochondrial genetic variation. *Trends in Ecology and Evolution*, 23, 546-554.

Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, 4, 359-361.

Eaton DAR (2014) PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics*, 30, 1844-1849.

Eaton DAR, Ree RH (2013) Inferring phylogeny and introgression using genomic RADseq data: An example from flowering plants (*Pedicularis*: Orobanchaceae). *Systematic Biology*, 62, 689-706.

Eaton, D. A., Spriggs, E. L., Park, B., & Donoghue, M. J. (2016). Misconceptions on Missing Data in RAD-seq Phylogenetics with a Deep-scale Example from Flowering Plants. *Systematic Biology*, syw092.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32, 1792-1797.

Escudero M, Eaton DA, Hahn M, Hipp AL (2014) Genotyping-by-sequencing as a tool to infer phylogeny and ancestral hybridization: A case study in *Carex* (Cyperaceae). *Molecular Phylogenetics and Evolution*, 79, 359-367.

Espíndola A, Buerki S, Alvarez N (2012) Ecological and historical drivers of diversification in the fly genus *Chiastocheta* Pokorny. *Molecular Phylogenetics and Evolution*, 63, 466-474.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*, 14, 2611-2620.

Faircloth BC, Branstetter MG, White ND, Brady SG (2014) Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among Hymenoptera. *Molecular Ecology Resources*, 15, 489-501

Funk DJ, Omland KE (2003) Species-level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA. *Annual Review of Ecology, Evolution and Systematics*, 34, 397-423.

Gatesy, J, Springer MS (2013) Concatenation versus coalescence versus "concatalescence". *Proceedings of the National Academy of Sciences*, 110, E1179-E1179.

Gautier M, Gharbi K, Cezard T, Foucaud J, Kerdelhué C, Pudlo P, Cornuet J-M, Estoup A (2013) The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Molecular Ecology*, *22*, 3165-3178.

Giarla TC, Esselstyn JA (2015) The challenges of resolving a rapid, recent radiation: Empirical and simulated phylogenomics of philippine shrews. *Systematic Biology*, 64, 727-740.

Gori K, Suchan T, Alvarez N, Goldman N, Dessimoz C (2016) Clustering genes of common evolutionary history. *Molecular Biology and Evolution*, 33,1590–1605

Govindarajulu R, Parks M, Tennessen JA, Liston A, Ashman TL (2015) Comparison of nuclear, plastid, and mitochondrial phylogenies and the origin of wild octoploid strawberry species. *American Journal of Botany*, 102, 544-554.

Harvey MG, Judy CD, Seeholzer GF, Maley JM, Graves GR, Brumfield RT (2015) Similarity thresholds used in DNA sequence assembly from short reads can reduce the comparability of population histories across species. *PeerJ*, 3, e895.

Harvey MG, Smith BT, Glenn TC, Faircloth BC, Brumfield RT (2016) Sequence capture versus restriction site associated DNA sequencing for shallow systematics. *Systematic Biology*, 65, 910-924.

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences*, 270, 313-322.

Hennig W (Ed.) (1976) Anthomyiidae. Die Fliegen der Palaearktischen Region. Stuttgart, E. Schweizerbart.

Herrera S, Shank TM (2016) RAD sequencing enables unprecedented phylogenetic resolution and objective species delimitation in recalcitrant divergent taxa. *Molecular Phylogenetics and Evolution*, 100, 70-79.

Hipp AL, Eaton DAR, Cavender-Bares J, Fitzek E, Nipper R, Manos PS (2014) A framework phylogeny of the american oak clade based on sequenced RAD data. *PLoS ONE*, 9, e93975.

Hovmöller R, Knowles LL, Kubatko LS (2013) Effects of missing data on species tree estimation under the coalescent. *Molecular Phylogenetics and Evolution*, 69, 1057-1062.

Huang H, Knowles LL (2014) Unforeseen consequences of excluding missing data from next-generation sequences: simulation study of RAD sequences. *Systematic Biology*, doi:10.1093/sysbio/syu046.

Ilut DC, Nydam ML, Hare MP (2014) Defining loci in restriction-based reduced representation genomic data from nonmodel species: Sources of bias and diagnostics for optimal clustering. *BioMed Research International*, 2014, 675158.

Johannesen J, Loeschcke V (1996) Distribution, abundance and oviposition patterns of four coexisting *Chiastocheta* species (Diptera: Anthomyiidae). *Journal of Animal Ecology*, 65, 567-576.

Jones JC, Fan S, Franchini P, Schartl M, Meyer A (2013) The evolutionary history of *Xiphophorus* fish and their sexually selected sword: a genome-wide approach using restriction site-associated DNA sequencing. *Molecular Ecology*, 22, 2986-3001.

Kubatko LS, Degnan JH (2007) Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology*, 56, 17-24.

Lanfear R, Frandsen PB, Wright AM, Senfeld T, Calcott B (2017) PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Molecular Biology and Evolution*. 34, 772-773.

Leaché AD, Chavez AS, Jones LN, Grummer JA, Gottscho AD, Linkem CW (2015) Phylogenomics of phrynosomatid lizards: conflicting signals from sequence capture versus restriction site associated DNA sequencing. *Genome Biology and Evolution*, 7, 706-719.

Leaché AD, Harris RB, Rannala B, Yang Z (2013) The influence of gene flow on species tree estimation: a simulation study. *Systematic Biology*, 63, 17-30.

Lowry DB, Hoban S, Kelley JL, Lotterhos KE, Reed LK, Antolin MF, Storfer A (2017) Breaking RAD: an evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. *Molecular Ecology Resources*, 17, 142–152.

Maddison WP (1997) Gene trees in species trees. *Systematic Biology*, 46, 523-536.

Maddison WP, Knowles LL (2006) Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology*, 55, 21-30.

Mallet J (2007) Hybrid speciation. *Nature*, 446, 279-283.

Manthey JD, Campillo LC, Burns KJ, Moyle RG (2016) Comparison of target-capture and restriction-site associated DNA sequencing for phylogenomics: a test in cardinalid tanagers (Aves, Genus: *Piranga*). *Systematic biology*, syw005.

Mastretta-Yanes A, Zamudio S, Jorgensen TH, Arrigo N, Alvarez N, Piñero D, Emerson BC (2014) Gene duplication, population genomics, and species-level differentiation within a tropical mountain shrub. *Genome biology and evolution*, 6, 2611-2624.

Mastretta-Yanes A, Arrigo N, Alvarez N, Jorgensen TH, Piñero D, Emerson BC (2015) Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Molecular Ecology Resources*, 15, 28-41.

McCormack JE, Faircloth BC, Crawford NG, Gowaty PA, Brumfield RT, Glenn TC (2012) Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species tree analysis. *Genome Research*, 22, 746-754.

McKinney GJ, Larson WA, Seeb LW, Seeb JE (2017) RADseq provides unprecedented insights into molecular ecology and evolutionary genetics: comment on Breaking RAD by Lowry *et al*. (2016). *Molecular Ecology Resources*, doi:10.1111/1755-0998.12649.

Michelsen V (1985) A revision of the Anthomyiidae (Diptera) described by J.W. Zetterstedt. *Steenstrupia*, 11, 37-65

Miller MA, Pfeiffer W, Schwartz T (2010) Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In: Proceedings of the Gateway Computing Environments Workshop (GCE), 14 Nov. 2010, New Orleans, LA pp. 1-8.

Nadeau NJ, Martin SH, Kozak KM, Salazar C, Dasmahapatra K, Davey JW, Baxter SW, Blaxter ML, Mallet J, Jiggins CD (2013) Genome-wide patterns of divergence and gene flow across a butterfly radiation. *Molecular Ecology*, 22, 814–826.

Pante E, Abdelkrim J, Viricel A, Gey D, France SC, Boisselier MC, Samadi S (2015) Use of RAD sequencing for delimiting species. *Heredity*, 114, 450-459.

Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20, 289-290.

Pellmyr O (1989) The cost of mutualism – interactions between *Trollius europaeus* and its pollinating parasites. *Oecologia*, 78, 53-59.

Pellmyr O (1992) The phylogeny of a mutualism: evolution and coadaptation between *Trollius* and its seed-parasitic pollinators. *Biological Journal of the Linnean Society*, 47, 337-365.

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.

Phillips MJ, Haouchar D, Pratt RC, Gibb GC, Bunce M (2013) Inferring Kangaroo Phylogeny from Incongruent Nuclear and Mitochondrial Genes. *PLoS ONE*, 8, e57745.

Pollard DA, Iyer VN, Moses AM, Eisen MB (2006) Widespread discordance of gene trees with species tree in *Drosophila*: Evidence for incomplete lineage sorting. *PLoS Genetics*, 2, e173.

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

Reaz R, Bayzid MS, Rahman MS (2014) Accurate phylogenetic tree reconstruction from quartets: A heuristic approach. *PloS One*, 9:e104008.

Roch S, Warnow T (2015) On the robustness to gene tree estimation error (or lack thereof) of coalescent-based species tree methods. *Systematic Biology*, syv016.

Rokas A, Carroll SB (2005) More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. *Molecular Biology and Evolution*, 22, 1337-1344.

Rubin BER, Ree RH, Moreau CS (2012) Inferring phylogenies from RAD sequence data. *PLoS ONE*, 7, e33394.

Rubinoff D, Holland BS (2005) Between two extremes: mitochondrial DNA is neither the panacea nor the nemesis of phylogenetic and taxonomic inference. *Systematic Biology*, 54, 952-961.

Schmid S, Genevest R, Gobet E, Suchan T, Sperisen C, Tinner W, Alvarez N (2017) HyRAD-X, a versatile method combining exome capture and RAD sequencing to extract genomic information from ancient DNA. *Methods in Ecology and Evolution*, doi:10.1111/2041-210X.12785

Seehausen O, Koetsier E, Schneider MV, Chapman LJ, Chapman CA, Knight ME, Turner GF, van Alphen JJM, Bills R (2003) Nuclear markers reveal unexpected genetic variation and a Congolese-Nilotic origin of the Lake Victoria cichlid species flock. *Proceedings of the Royal Society of London Series B*, 270, 129-137.

Springer MS, Gatesy J (2016) The gene tree delusion. *Molecular Phylogenetics and Evolution*, 94, 1-33.

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

Suchan T, Beauverd M, Trim N, Alvarez N (2015) Asymmetrical nature of the *Trollius-Chiastocheta* interaction: insights into the evolution of nursery pollination systems. *Ecology and Evolution*, 5, 4766-4777.

Suchan T, Pitteloud C, Gerasimova NS, Kostikova A, Schmid S, Arrigo N, Pajkovic M, Ronikier M, Alvarez N (2016) Hybridization Capture Using RAD Probes (hyRAD), a New Tool for Performing Genomic Analyses on Collection Specimens. *PLoS ONE*, 11, e0151651.

Suchan T, Espíndola A, Rutschmann S, Emerson BC, Gori K, Dessimoz C, Arrigo N, Ronikier M, Alvarez N (2017) Data from: Assessing the potential of RAD-sequencing to resolve phylogenetic relationships within species radiations: the fly genus *Chiastocheta* (Diptera: Anthomyiidae) as a case study. *Mendeley Data*, https://doi.org/XXX

Swofford DL (2002) PAUP*: Phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer, Sunderland, Massachusetts, USA.

Toews DP, Brelsford A (2012) The biogeography of mitochondrial and nuclear discordance in animals. *Molecular Ecology*, 21, 3907-3930.

Townsend TM, Mulcahy DG, Noonan BP, Sites JW, Kuczynski CA, Wiens JJ, Reeder TW (2011) Phylogeny of iguanian lizards inferred from 29 nuclear loci, and a comparison of concatenated and species-tree approaches for an ancient, rapid radiation. *Molecular Phylogenetics and Evolution*, 61, 363-380.

Wagner CE, Keller I, Wittwer S, Selz OM, Mwaiko S, Greuter L, Sivasundar A, Seehausen O (2013) Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Molecular Ecology*, 22, 787–798.

Walsh HE, Kidd MG, Moum T, Friesen VL (1999) Polytomies and the power of phylogenetic inference. *Evolution*, 53, 932-937.

Whitfield JB, Kjer KM (2008) Ancient rapid radiations of insects: challenges for phylogenetic analysis. *Annual Review of Entomology*, 53, 449-472.

Whitfield JB, Lockhart PJ (2007) Deciphering ancient rapid radiations. Trends in Ecology and Evolution, 22, 258-265.

Wielstra B, Arntzen JW, van der Gaag KJ, Pabijan M, Babik W (2014) Data Concatenation, Bayesian Concordance and Coalescent-Based Analyses of the Species Tree for the Rapid Radiation of *Triturus* Newts. *PLoS ONE*, 9, e111011.

Williams JS, Niedzwiecki JH, Weisrock DW (2013) Species tree reconstruction of a poorly resolved clade of salamanders (Ambystomatidae) using multiple nuclear loci. *Molecular Phylogenetics and Evolution*, 68, 671-682.

Zetterstedt JW (1845) Diptera scandinaviae disposita et descripta. IV, 1281-1738. Lund, Sweden.

Zink RM, Barrowclough GF (2008) Mitochondrial DNA under siege in avian phylogeography. *Molecular Ecology*, 17, 2107-2121.

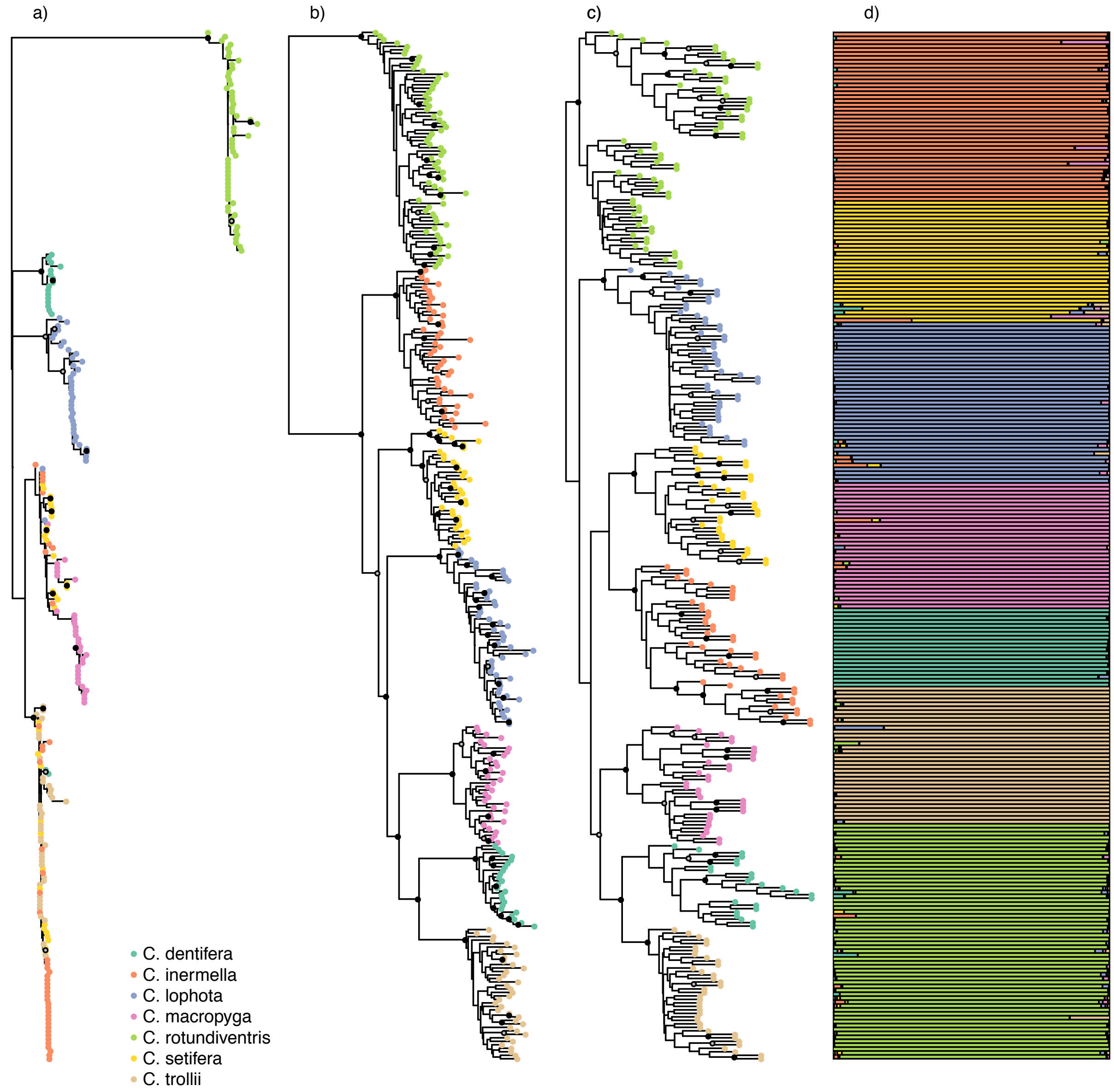
Data Accessibility

DNA sequences are available on Genbank under accessions no XXX-XXX. Nexus and STRUCTURE datasets used for the analyses, ML phylogeny inferred for the main RAD dataset, SVDquartet analyses, and ML phylogeny inferred for the mtDNA dataset are available on Mendeley Data https://doi .org/XXX (Suchan et al. 2017).

Author Contributions

TS, NA, AE, KG and CD designed research, TS, AE, KG, and SR performed research and analyzed data. All authors wrote the paper.

793 # Figures



794

795

796 Figure 1. Phylogenies obtained for a) ML analysis of the mtDNA dataset; b) ML analysis
797 of the RAD dataset; c) SVDquartets analysis of RAD dataset; bootstrap node supports >
798 80 are shown denoted by gray points, bootstrap node supports > 90 are shown denoted
799 by black points. d) Population clustering of the sampled *Chiastocheta* specimens,
800 estimated with STRUCTURE using $K = 7$ value.

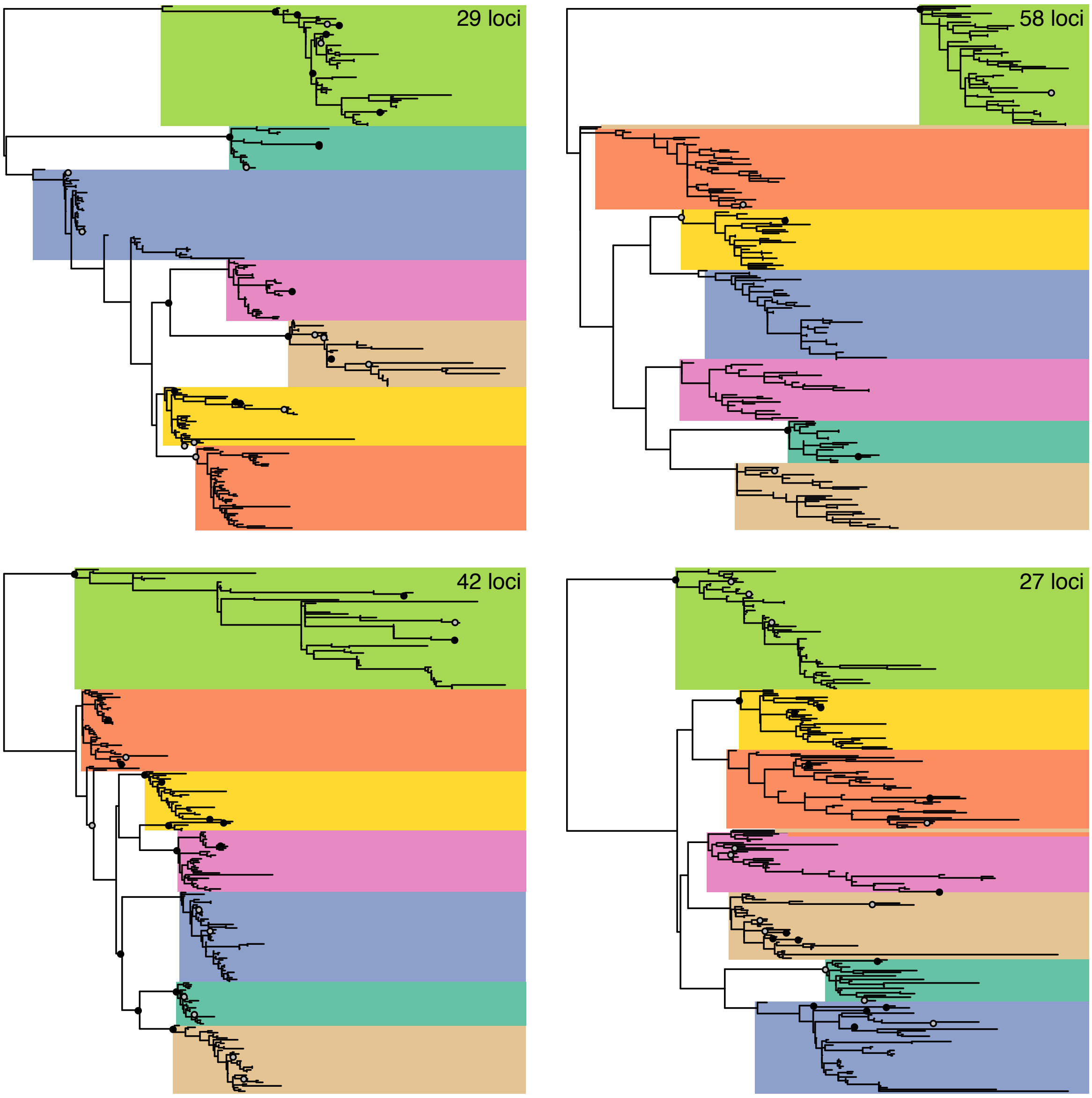


Figure 2. Phylogenetic trees on the four bins, as identified by the treeCl analysis, considering only the loci present in at least 100 specimens. Bootstrap node supports > 80 are shown denoted by gray points, bootstrap node supports > 90 are shown denoted by black points.

Tables

Table 1. Populations included in the study, with geographical coordinates and the number of specimens used in the final analyses. Letter codes denote *Chiastocheta* species: *C. dentifera* (D), *C. inermella* (I), *C. lophota* (L), *C. macropyga* (M), *C. rotundiventris* (R), *C. setifera* (S), and *C. trollii* (T).

code	site	latitude	longitude	year	D	I	L	M	R	S	T	sum
AMB	Ambri	46.50680	8.70292	2008			2	4	2		1	9
AMO	Amot	59.62199	8.42346	2007		4						4
BAY	Bayasse	44.30814	6.74067	2007		1	2	2	2		3	10
BEI	Beistohlen	61.20761	8.95473	2007		5						5
BID	Bidjovagge	69.29778	22.47808	2008		1			1			2
BON	Col de Bonnecombe	44.57557	3.11410	2007				3	1	1	2	7
BRA	Braas	57.09309	15.06817	2007		1						1
CCO	Col de la Co-lombière	45.98722	6.46972	2006			2	1	2		1	6
CDV	Creux du Van	46.93526	6.74119	2006		1	2		2	2		7
CHA	Chasseral	47.12569	7.02130	2006			5		3		1	9
CHE	Chemin	46.08993	7.08978	2006	3	1		3	2	1	2	12
CRA	Crans-Montana	46.34650	7.53890	2006		1	1		3	1	1	7
CRE	Cressbrook Dale	53.26724	-1.74041	2008						2		2
CTP	Colt Park	54.19365	-2.35247	2008		4			1		1	6
DON	Donovaly	48.88922	19.23068	2008			3	1	2			6
EID	Eidda Pastures	53.03720	-3.74190	2008						2		2
ELL	Ellingsrudelva	59.91771	10.91844	2007		1						1
EPO	Esposouille	42.62341	2.09450	2008	1		1		2	2	3	9
FRO	Froson	63.18205	14.60268	2007		2						2

GAL	Col du Galibier	45.08528	6.43861	2006		2	2		3		3	10
GLE	Glen Fender	56.78138	-3.79485	2008					2	2		4
HT1	Haute Tinee 1	44.29617	6.81871	2007			3					3
HT2	Haute Tinee 2	44.28426	6.85581	2007				3			1	4
KRA	Krasno Polje	44.80869	14.97271	2008						1		1
LAK	Laktatjakka	68.42931	18.40674	2007		1		4	2			7
LFE	Lough Fern	55.06569	-7.71130	2008						1		1
LOS	Loser	47.66052	13.78485	2007			4		3	1		8
MOE	Moerlimatt	47.90597	8.07760	2007			1		1	1	1	4
MTP	Monte Pizi	41.91524	14.16714	2008						2		2
NAV	Naverdal	62.70417	10.13002	2007		3					1	4
PAJ	Pajino Preslo	43.27799	20.81970	2008						2		2
PAN	Puerto de Panderrueda	43.12743	-4.97223	2008			1	1	3	2	1	8
PIL	Pila	48.90017	20.29449	2008	3		2	1				6
POD	Podlesok	48.94962	20.35190	2008				1			1	2
PPN	Petit Papa Noel	66.51647	25.79386	2007	1	3			2		3	9
PYD	Puy de Dome	45.77222	2.96333	2006			2	2	2			6
PYM	Puy Mary	45.11139	2.68083	2006			1					1
PYS	Puy de Sancy	45.53500	2.80972	2006			3	2	1			6
RAD	Radkow	50.46866	16.35321	2008	2	4			2			8
RIS	Risnjak - Snjeznik	45.43871	14.58494	2008					1	2		3
SAL	Salla	66.83020	28.65427	2007	1				2	1	4	8
SED	Sede de Pan	43.03949	-0.48651	2008			2					2
SET	Seterasen	65.53432	13.67744	2007	1	4		1	1		1	8
SOL	Solberga	57.95194	13.56116	2007	4	2			2		1	9
STE	Steingaden	47.59529	11.01296	2007				1	1	31		5

STR	Straumen	67.38440	15.64921	2007					2			2
SUS	Susch	46.74728	10.07473	2006	2		2		2	1	3	10
SVA	Svartla	65.99583	21.22062	2007	3	3						6
TAR	Tarasp	46.77730	10.25056	2006			1		2	1	3	7
VIT	Vitosha	42.59032	23.29342	2008						2		2
ZAL	Zali Log	46.20342	14.11080	2008				3	2	1		6
				total:	21	44	42	33	59	34	38	271

Appendix A. Supplementary material

Table S1. Summary statistics for the RAD-sequenced samples: number of RAD fragments clusters and mean coverage after retaining clusters with a coverage >5, estimated heterozygosities, number of consensus loci after paralog filtering, and the numbers of loci retained for each dataset after filtering for coverage among the samples.

Fig. S1. Map of the sampled *Chiastocheta* specimens used in the study.

Fig. S2. The effect of different clustering thresholds (*X*-axis) and minimum loci coverages (indicated by colors: red – 10, blue – 20, green – 100 individuals) on the total number of assembled loci, proportion of missing data, loci overlap among the technical replicates, and mean number of individuals per locus.

Fig. S3. Pattern of RAD-seq loci sharing among the sequenced individuals for datasets: a) the main dataset using clustering similarity of 75% and minimum loci coverage among individuals of 20; b) using clustering similarity of 75% and minimum loci coverage among individuals of 100.

Fig. S4. a) ML phylogeny inferred for the mtDNA dataset; b) ML phylogeny inferred for the RAD-seq dataset; c) SVDquartets phylogeny inferred for the RAD-seq dataset; bootstrap node supports > 80 are shown denoted by gray points, bootstrap node supports > 90 are shown denoted by black points.

Fig. S5. STRUCTURE runs for *K*=2 to 7, plotted against the RAD-seq based phylogeny.

Fig. S6. Phylogenetic trees on the loci partitioned into the sets of 2 to 6 clusters, considering only the loci present in at least 100 specimens. Bootstrap node supports > 80 are shown denoted by gray points, bootstrap node supports > 90 are shown denoted by black points. Numbers below the trees denote the number of clusters into which the dataset was divided. Plot of log-likelihood improvement versus the number of clusters is presented in the first box.

Appendix S1. RAD-sequencing protocol.