Version dated: December 21, 2017

RH: Phasing improves utility of UCEs

Allele Phasing Greatly Improves the Phylogenetic Utility of Ultraconserved Elements

Tobias Andermann[1,2], Alexandre M. Fernandes[3], Urban Olsson[1,2], Mats Töpel[2,4], Bernard Pfeil[1,2], Bengt Oxelman[1,2], Alexandre Aleixo[5], Brant C. Faircloth[6] and Alexandre Antonelli[1,2,7]

[1] *Department of Biological and Environmental Sciences, University of Gothenburg, SE-41319, Göteborg, Sweden;*

[2] *Gothenburg Global Biodiversity Centre, Box 461, SE-405 30, Göteborg, Sweden*

[3] *Universidade Federal Rural de Pernambuco, Serra Talhada, Brazil*

[4] *Department of Marine Sciences, University of Gothenburg, SE-41319, Göteborg, 41319, Sweden;*

[5] *Museu Paraense Emílio Goeldi, Collection of Birds, Belém, Brazil*

[6] *Department of Biological Sciences and Museum of Natural Science, Louisiana State University, Baton Rouge, LA, U.S.A.*

[7] *Gothenburg Botanical Garden, Göteborg, SE-41319, Sweden*

Corresponding author: Tobias Andermann, Department of Biological and Environmental Sciences, University of Gothenburg, Carl Skottsbergs Gata 22B, Göteborg, 41319, Sweden; E-mail: tobias.hofmann@bioenv.gu.se

Abstract.— Advances in high-throughput sequencing techniques now allow relatively easy and affordable sequencing of large portions of the genome, even for non-model organisms. Many phylogenetic studies prefer to reduce costs by focusing their sequencing efforts on a selected set of targeted loci, commonly enriched using sequence capture. The advantage of this approach is that it recovers a consistent set of loci, each with high sequencing depth, which leads to more confidence in the assembly of target sequences. High sequencing depth can also be used to identify phylogenetically informative allelic variation within sequenced individuals, but allele sequences are infrequently assembled in phylogenetic studies. Instead, many scientists perform their phylogenetic analyses using contig sequences which result from the *de novo* assembly of sequencing reads into contigs containing only canonical nucleobases, and this may reduce both statistical power and phylogenetic accuracy. Here, we develop an easy-to-use pipeline to recover allele sequences from sequence capture data, and we use simulated and empirical data to demonstrate the utility of integrating these allele sequences to analyses performed under the Multispecies Coalescent (MSC) model. Our empirical analyses of Ultraconserved Element (UCE) locus data collected from the South American hummingbird genus *Topaza* demonstrate that phased allele sequences carry sufficient phylogenetic information to infer the genetic structure, lineage divergence, and biogeographic history of a genus that diversified during the last three million years, support the recognition of two species, and suggest a high rate of gene flow across large distances of rainforest habitats but rare admixture across the Amazon River. Our simulations show that analyzing allele sequences leads to more accurate estimates of tree topology and divergence times than the more common approach of using contig sequences. We conclude that allele phasing may be the most appropriate processing scheme for phylogenetic analyses of UCE data in particular, and sequence capture data, more generally.

(Keywords: UCE, SNP, heterozygous sites, Multispecies Coalescent, gene tree, species tree, Mitochondrial Genome, Trochilidae, Birds, Amazon)

Massive Parallel Sequencing (MPS) techniques enable time- and cost-efficient generation of DNA sequence data. Instead of using MPS to sequence complete genomes, many researchers choose to focus their sequencing efforts on a set of target loci to lower costs while achieving higher coverage and more reliable sequencing of these target regions (Faircloth et al. 2012, 2013; Mirarab et al. 2014; Smith et al. 2014; Faircloth 2015; Harvey et al. 2016; Meiklejohn et al. 2016). These multilocus datasets typically contain hundreds or thousands of target loci, and most are generated through enrichment techniques such as sequence capture (synonym: target enrichment, Gnirke et al. (2009)). After collecting sequence data from these targeted loci, many researchers assemble their high coverage sequence reads into “contigs” using *de novo* genome assembly software, and the “contigs” output by these assemblers often ignore the variants at heterozygous positions that are expected in diploid organisms. Typically, variable positions are treated as sequencing errors and assembly algorithms output the contig containing the more probable (i.e., numerous) variant while discarding the alternative (Iqbal et al. 2012). As a result, the contigs that are produced contain only canonical nucleobases, losing the information for each alternative allele present at each variable position (Fig. 4). Hereafter, we use “contigs” and “contig sequences” to refer to the sequences that are output by *de novo* assemblers.

One alternative approach to generating contig sequences uses the depth of sequencing coverage to programatically identify variable positions within a targeted locus (also known as “calling” single nucleotide polymorphism (SNPs)) and subsequently sorting (or “phasing”) these SNPs into two allele sequences or “haplotypes” which represent alleles on the same chromosome present at that locus. These approaches have been used to estimate demographic parameters such as effective population size, rate of migration, and the amount of gene flow between and within populations. However, it is rarely acknowledged (*c.f.* Lischer et al. 2014; Potts et al. 2014; Schrempf et al. 2016; Eriksson et al. 2017) that allelic sequences are useful for phylogenetic studies to improve the estimation of gene trees, species trees, and divergence times (Garrick et al. 2010; Potts et al. 2014; Lischer et al. 2014). The common practice of neglecting allelic information in phylogenetic studies possibly results from historical inertia and a lack of computational pipelines to prepare allele sequences for phylogenetic analysis using MPS data.

In addition to the problems of determining allelic sequences, the proper analysis of allelic information in phylogenetic studies remains a challenging and intensively discussed topic (Garrick et al. 2010; Lischer et al. 2014; Potts et al. 2014; Schrempf et al. 2016; Leaché and Oaks 2017). Various approaches have been proposed to include this information into phylogenetic methods (Lischer et al. 2014; Potts et al. 2014; Schrempf et al. 2016). One is to code heterozygous sites using IUPAC ambiguity codes and to include these as additional characters in existing substitution models for gene tree and species tree inference (Potts et al. 2014; Schrempf et al. 2016). While these studies demonstrate that integrating additional allelic information in this manner increases accuracy in phylogenetic inference, Lischer et al. (2014) found that coding heterozygous sites as IUPAC ambiguity codes in phylogenetic models biases the results toward older divergence time estimates. Instead, Lischer et al. (2014) introduced a method of repeated random haplotype sampling (RRHS) in which allele sequences are repeatedly concatenated across many loci, using a random haplotype for any given locus in each replicate. In their approach they then analyzed thousands of concatenation replicates separately for phylogenetic tree estimation and summarized the results between replicates, thereby integrating the allelic information in form of uncertainty intervals. However there are two important shortcomings of this approach: 1. concatenating unlinked loci (and in particular allele sequences from unlinked loci) in a random manner is known to produce incorrect topologies (Degnan and Rosenberg 2009) often with false confidence (Edwards et al. 2007; Kolaczkowski and Thornton 2004; Kubatko and Degnan 2007; Mossel and Vigoda 2005), which is not accounted for when doing so repeatedly and summarizing the resulting trees, and 2. running thousands of tree estimation replicates based on extensive amounts of sequence data results in unfeasibly long computation times, particularly for Markov-Chain Monte Carlo (MCMC) based softwares such as MrBayes or BEAST. Hence there is need to find proper solutions to include heterozygous information in phylogenetic analyses, as concluded by Lischer et al. (2014).

Here, we introduce the bioinformatic assembly of allele sequences from UCE data and demonstrate a full integration of allele sequences to species tree estimation under the multispecies coalescent (MSC) model using empirical and simulated data. In our approach, we treat each allelic sequence of an individual at a given locus as an independent sample from the population, and we analyze these sequences using the species tree and delimitation software STACEY (Jones et al. 2014; Jones 2017), which does not require *a priori* clade- or species-assignments. We first demonstrate the empirical utility of our approach by resolving the shallow genetic structure (<1 Ma) within two recognized morphospecies of the South American hummingbird genus *Topaza* by analyzing a set of 2,386 ultraconserved elements (UCEs, see Faircloth et al. (2012)) collected using sequence capture of the 2.5k tetrapod baitset (see `http://ultraconserved.org`). We then validate this approach, using simulations, and show that allele sequences yield more accurate results in terms of species tree estimation and species delimitation than the contig sequence approach that ignores heterozygous information. Our simulation results further demonstrate that proper phasing of allele sequences outperforms alternative approaches of coding heterozygous information, such as analyzing sequences containing IUPAC ambiguity codes or SNPs. We conclude by demonstrating that phasing sequence capture data can be critical for correct species delimitation and phylogeny estimation, particularly in recently diverged groups, and that analyses using phased alleles should be considered as one “best practice” for analyzing sequence capture datasets in a phylogenetic context.

Materials and Methods

Study System

The genus *Topaza* and its sister genus *Florisuga* form the Topazes group, which together with the Hermits represent the most ancient branch within the hummingbird family (Trochilidae) (McGuire et al. 2014). Topazes are estimated to have diverged as a separate lineage from all other hummingbirds around 21.5 Ma, whereas the most recent common ancestor (MRCA) of *Topaza* and *Florisuga* lived approximately 19 Ma (McGuire et al. 2014). At present, there are two morphospecies recognized within *Topaza*, namely the Fiery Topaz, *T. pyra* (Gould, 1846), and the Crimson Topaz, *T. pella* (Linnaeus, 1758). However, the species status of *T. pyra* has been challenged by some authors (Schuchmann 1999; Ornés-Schmitz and Schuchmann 2011), who consider this genus to be monotypic. Topaz hummingbirds are endemic to the Amazonian rainforest and are some of the most spectacular and largest hummingbirds worldwide, measuring up to 23 cm (adult males, including tail feathers) and weighing up to 12 g (Schuchmann et al. 2016; del Hoyo et al. 2016a). These birds are usually found in the forest canopy along forest edges and clearings, and are often seen close to river banks (Ornés-Schmitz and Schuchmann 2011). There is morphological evidence for several subspecies within both currently recognized *Topaza* species (Peters 1945; Schuchmann 1999; Hu et al. 2000; Ornés-Schmitz and Schuchmann 2011) that we investigate using genetic data.

Sequence Data Generation

We extracted DNA from the muscle tissue of 10 vouchered hummingbirds (9 *Topaza*, one *Florisuga*, see Table 1) using the Qiagen DNeasy Blood and Tissue Kit according to the manufacturer's instructions (Qiagen GmbH, Hilden, Germany). These samples cover most of the genus' total geographic range (Fig. 1) and all morphologically recognized intraspecific taxa (Schuchmann et al. 2016; del Hoyo et al. 2016a). All samples

Table 1: Specimens sequenced. Subspecies identifications based on morphological characters. Abbreviation for sample providers: INPA = Instituto Nacional de Pesquisas da Amazônia, MPEG = Museum Paraense Emílio Goeldi, USNM = NMNH, Smithsonian Institution, Washington DC, USA.

ID	Taxon	Subspecies	Voucher number	Latitude	Longitude
1	*Topaza pyra*	*amaruni*	INPA A1106	-0.044167	-66.94944
2	*T. pyra*	*pyra*	MPEG 62475	-1.559444	-65.88006
3	*T. pyra*	*pyra*	MPEG 62474	-4.083889	-60.66050
4	*T. pyra*	*pyra*	MPEG 52721	-7.350000	-73.66667
5	*T. pella*	NA	USNM 586322	7.220000	-60.29000
6	*T. pella*	*pella*	INPA A3319	-1.927900	-59.41600
7	*T. pella*	*smaragdula*	MPEG 61688	-1.950000	-51.60000
8	*T. pella*	*microrhyncha*	MPEG 65603	-5.352417	-57.47500
9	*T. pella*	NA	INPA A6233	-9.028550	-64.24231
10	*Florisuga fusca*	NA	MPEG 70697	-15.15972	-39.04500

were sonicated with a Covaris S220 to a fragment length of 800 bp. Paired-end, size-selected (range 600-800bp) DNA libraries were prepared for sequencing on the Illumina MiSeq platform, using the magnetic-bead based NEXTflexTM Rapid DNA-Seq Kit (Bioo Scientific Corporation, Austin, TX, USA), following the user's manual (v14.02).

We used the "Tetrapods-UCE-2.5Kv1" bait set (`uce-2.5k-probes.fasta`), consisting of 2,560 baits (each 120 bp), targeting 2,386 UCEs, as described by Faircloth et al. (2012). The bait sequences were downloaded from `http://ultraconserved.org` and synthesized by MYcroarray (Biodiscovery LLC, Ann Arbor, MI, USA). Sequence enrichment was performed using a MYbaits kit according to the enclosed user manual (v1.3.8). The enriched libraries were then sequenced using 250 bp, paired-end sequencing on an Illumina MiSeq machine (Illumina Inc., San Diego, CA, USA). Library preparation, sequence enrichment and sequencing were performed by Sahlgrenska Genomics Core Facility in Gothenburg, Sweden.

Mitochondrial Genome

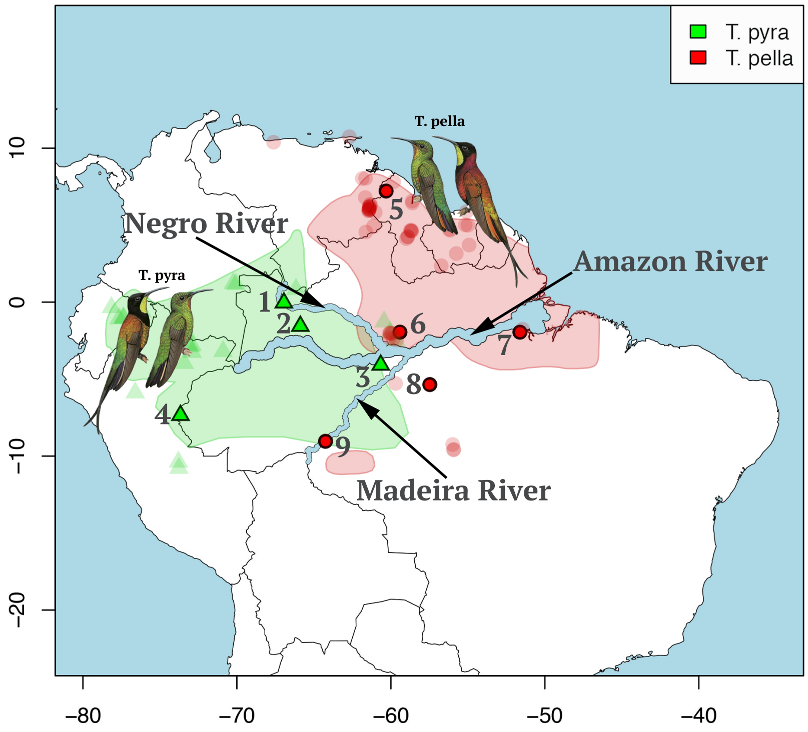


Figure 1: Sample locations of *Topaza* specimens (numbered symbols) in northern South America. Numbers represent sample IDs (Table 1). The colored polygons show the distribution range of the two morphospecies (*T. pyra* = green, *T. pella* = red) as estimated by BirdLife International (`http://www.birdlife.org`). Transparent symbols (triangles and circles) represent *Topaza* sightings, which were downloaded from the eBird database (Sullivan et al. 2009). The major river systems in the Amazon drainage basin are marked in blue (not in proportion). *Topaza* illustrations were provided by del Hoyo et al. (2016b).

To infer a dated mitochondrial phylogeny for the genus *Topaza* to compare with the nuclear phylogeny, we used off-target mitochondrial reads to assemble the complete mitochondrial genome for all samples. We found that as many as 4.5% of all sequence reads were of mitochondrial origin, even though no baits targeting mitochondrial loci were used during sequence capture. An alignment of the assembled mitochondrial genomes for all samples was analyzed in BEAST (Drummond et al. 2012). Dating priors included clock-rate priors for three mitochondrial genes, estimated for honeycreepers by Lerner et al. (2011) and node-age priors within the genus *Topaza* that were estimated by McGuire et al.

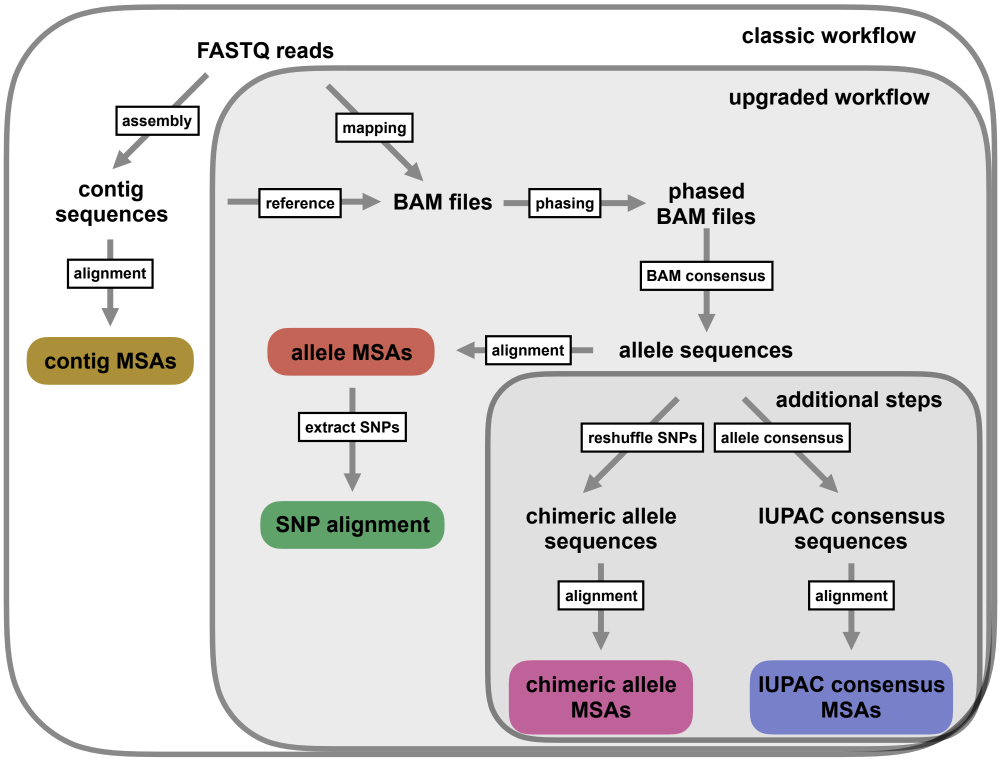


Figure 2: Depiction of the workflow developed here. Colored boxes represent different types of multiple sequence alignments (MSAs) used for phylogenetic inference in this study. In addition to the standard UCE workflow of generating contig MSAs (Faircloth et al. 2012; Smith et al. 2014; Faircloth 2015), we extended the bioinformatic processing in order to generate UCE allele MSAs, and to extract SNPs from these allele MSAs. We added these new functions to the PHYLUCE pipeline (Faircloth 2015). Additional data processing steps were executed in this study in order to test different codings of heterozygous positions.

(2014). A detailed description of the assembly and analysis of the mitochondrial genome data can be found in online Appendix 1 (Supplemental Material available on Dryad).

UCE Data Processing

For this study we generated five different types of datasets, which we analyzed under the MSC. These five datasets represent different coding schemes for heterozygous information and are listed and described in the following sections.

1. UCE contig alignments.— Because contig sequences are commonly used in phylogenetic analyses of MPS datasets (e.g. Faircloth et al. (2012); Smith et al. (2014); Faircloth (2015)), we generated multiple sequence alignments (MSAs) of contigs for all UCE loci in order to test the accuracy of the phylogenetic estimation of this approach.

To create MSAs from UCE contig data, we followed the suggested workflow from the PHYLUCE documentation (`http://phyluce.readthedocs.io/en/latest/tutorial-one.html`). We applied the PHYLUCE default settings unless otherwise stated. First we quality-filtered and cleaned raw Illumina reads of adapter contamination with Trimmomatic (Bolger et al. 2014), which is implemented in the PHYLUCE function `illumiprocessor`. The reads were then assembled into contigs using the software ABYSS (Simpson et al. 2009) as implemented in the PHYLUCE pipeline. In order to identify contigs representing UCE loci, all assembled contigs were mapped against the UCE reference sequences from the bait sequence file (`uce-2.5k-probes.fasta`), using the PHYLUCE function `match_contigs_to_probes.py`. We extracted only those sequences that matched UCE loci and that were present in all samples (n=820). These UCE sequences were then aligned for each locus (Fig. 2) using MAFFT (Katoh et al. 2009).

2. UCE allele alignments.— We altered the typical UCE workflow in order to retrieve the allelic information that is lost when collapsing multiple reads into a single contig sequence (Fig. 2). To create this new workflow, we extracted all UCE contigs for each sample separately and treated each resulting contig set as a sample-specific reference library for read mapping. We then mapped the cleaned reads against each reference library on a per sample basis, using CLC-mapper from the CLC Workbench software. The mapped reads were sorted and then phased with SAMtools v0.1.19 (Li et al. 2009), using the commands `samtools sort` and `samtools phase`, respectively. This phasing function is based on a dynamic programming algorithm that uses read connectivity across multiple variable sites to determine the two phases of any given diploid locus (He et al. 2010). Further, this algorithm uses paired-end read information to reach connectivity over longer distances and it minimizes the problem of accidentally phasing a sequencing error, by applying the minimum error correction function (He et al. 2010).

UCE data provide an excellent dataset for allele phasing based on read connectivity, because the read coverage across any given UCE locus typically is highest in the center and decreases toward the ends. This makes it possible to phase throughout the complete locus without any breaks in the sequence. Even in cases where the only variable sites are found on opposite ends of the locus, the insert size we targeted in this study (800 bp), in combination with paired-end sequencing, enabled the phasing process to bridge the complete locus. The two phased output files (BAM format) were inspected for proper variant separation for all loci using Tablet (Milne et al. 2013). We then collapsed each BAM file into a single sequence and exported the two resulting allele sequences for each sample in FASTA format. In the next, step we aligned the allele sequences between all samples, separately for each UCE locus, using MAFFT (Fig. 2). We integrated this complete workflow into the UCE processing software PHYLUCE (Faircloth 2015) with slight alterations, one of which is the use of the open-source mapping program bwa (Li and Durbin 2010) in place of CLC-mapper.

3. UCE IUPAC consensus sequence alignments.— We generated an additional set of alignments by merging the two allele sequences for each individual into one consensus sequence with heterozygous sites coded as IUPAC ambiguity codes (`merge_allele_sequences_ambiguity_codes.py`, available from: github.com/tobiashofmann88/UCE-data-management/). We used this dataset to test whether our allele phasing approach improved phylogenetic inference when compared to the IUPAC consensus approach applied in other studies (where heterozygous positions are simply coded as IUPAC ambiguity codes in a consensus sequence for each locus and individual (Potts et al. 2014; Schrempf et al. 2016)).

4. UCE chimeric allele alignments.— To investigate whether correct phasing of heterozygous sites is essential or if similar results are achieved by randomly placing variants in either allele sequence, we generated a dataset with chimeric allele sequence alignments. We created these alignments by applying a custom python script (`shuffle_snps_in_allele_alignments.py`, available from: github.com/tobiashofmann88/UCE-data-management/) to the phased allele sequence alignments and randomly shuffling the two variants at each polymorphic position between the two allele sequences for each individual. This process leads, in many cases, to an incorrect combination of variants on each allele sequence, thereby creating chimeric allele sequences. The resulting alignments contain the same number of sequences as the phased allele alignments (two sequences per individual), whereas the contig alignments and the IUPAC consensus alignments contain only half as many sequences (one sequence per individual).

5. UCE SNP alignment.— A common approach to analyze heterozygous information is to reduce the sequence information to only a single variant SNP per locus. This data-reduction approach is often chosen because multilocus datasets of the size generated in this study can be incompatible with Bayesian MSC methods applied to the full sequence data, due to extremely long computational times. Instead, alignments of unlinked SNPs can be used to infer species trees and species demographics under the MSC model with the BEAST2 package SNAPP (Bryant et al. 2012), a program specifically designed for such data. However, extracting and filtering SNPs from BAM files with existing software (such as the Genome Analysis Toolkit (GATK), McKenna et al. (2010)) and converting these into a SNAPP compatible format can be cumbersome, because SNAPP requires positions with exactly two different states, coded in the following manner: individual homozygous for the original state = "0", heterozygous = "1", and homozygous for the derived state = "2".

To alleviate this problem, we developed a python function that extracts biallelic SNPs directly from allele sequence MSAs (`snps_from_uce_alignments.py`, available from: github.com/tobiashofmann88/UCE-data-management/). Extracting SNPs from MSAs in this manner is a straightforward and simple way to generate a SNP dataset compatible with SNAPP, and does not require re-visiting the BAM files. Although a similar program already exists, which is implemented in the R-package `phrynomics` (Leaché et al. 2015), we integrated the SNP extraction from allele sequence MSAs into the PHYLUCE pipeline, and used this approach to extract one position per alignment (to ensure unlinked SNPs) that had exactly two different states among all *Topaza* samples, not allowing for positions with missing data or ambiguities. This produced a SNP dataset of 598 unlinked SNPs.

Generation of Simulated UCE Data

To assess the accuracy of the phylogenetic inferences resulting from different data processing approaches, we simulated UCE data similar to those discussed in the five processing schemes we applied to the empirical *Topaza* data. However, because this approach required us to simulate allele alignments before generating contig alignments, steps one and two, below, are reversed from their order, above. For each of the five processing schemes, we generated and analyzed ten independent simulation replicates.

1. Simulated allele alignments.— From the empirical UCE allele alignments, we estimated species divergence times and population sizes under the MSC model (Rannala and Yang 2003) using the Bayesian MCMC program BPP v3.1 (Yang 2015). To do this, we used the A00 model with the species tree topology from the analysis of the allele sequence data in

STACEY, assigning the *Topaza* samples to five separate taxa (corresponding to colored clades in Fig. 6b). An initial BPP analysis did not converge in reasonable computational time, a problem that has previously been reported for UCE datasets containing several hundred loci (Giarla and Esselstyn 2015). To avoid this issue, we split the 820 UCE alignments randomly into 10 subsets of equal size (n=82) and analyzed these separately with identical settings in BPP. The MCMC was set for 150,000 generations (burn-in 50,000), sampling every 10 generations. We summarized the estimates for population sizes and divergence times across all 10 individual runs. We then applied the mean values of these estimates to the species tree topology, by using the estimated divergence times as branch lengths and estimated population sizes as node values, resulting in the species tree in Fig. 6g. This tree was used to simulate sequence alignments with the MCcoal simulator, which is integrated into BPP. Equivalent to the empirical data, we simulated sequence data for five taxa (D, E, X, Y, and Z) and one outgroup taxon (F, not shown in Fig. 6g). In the simulations, these taxa were simulated as true species under the MSC model. In order to mimic the empirical allele data, we simulated four individuals for species 'D' (equivalent to two allele sequences for 2 samples), four for species 'E', four for species 'X', two for species 'Y' (two allele sequences for one sample), four for species 'Z', and two for the outgroup species 'F'. In this manner we simulated 820 UCE allele MSAs of 848 bp length (a value equal to the average alignment length of the empirical allele alignments).

2. Simulated contig alignments.— To simulate UCE contig MSAs similar to those used in previous studies (Faircloth et al. 2012; McCormack et al. 2012; Smith et al. 2014; Faircloth 2015) and output by assemblers like Velvet or Trinity which pick only one of the two variants at a heterozygous site, we merged the sequences within each coalescent species in pairs of two (equivalent to pairs of allele sequences). Each pair of allele sequences was joined into one contig sequence by randomly picking one of the two variants at each heterozygous site across all loci. As in the empirical contig assembly approach, our simulation approach may generate chimeric contig sequences.

3. Simulated IUPAC consensus alignments.— Next, we generated IUPAC consensus MSAs in the same manner as we generated the simulated contig MSAs in the previous step, with the exception that all heterozygous sites were coded with IUPAC ambiguity codes instead of randomly picking one of the two variants.

4. Simulated chimeric allele alignments.— We generated chimeric allele sequence MSAs from the simulated allele MSAs by randomly shuffling the heterozygous sites between each pair of sequences using the same pairs as in the previous two steps.

5. Simulated SNP alignment.— Finally, we extracted two different SNP datasets from the simulated phased allele MSAs. The first SNP dataset (SNPs complete) was extracted in the same manner as described for the empirical data (one SNP per locus for all loci) which resulted in a total alignment length of 820 SNPs for the simulated data. We extracted an additional SNP dataset (SNPs reduced) from only the subset of the 150 simulated allele alignments that were used for the sequence-based MSC analyses (see next section below). The resulting dataset of 150 SNPs was used to compare the phylogenetic inference based on SNP data versus that based on full sequence data, if the same number of loci is being analyzed. This enabled us to evaluate the direct effect of reducing the full sequence information in the MSAs to one single SNP for each of the selected 150 loci.

MSC Analyses of Empirical and Simulated UCE Data

Sequence-based tree estimation.— To jointly infer gene trees and species trees, we analyzed each of the generated sets of MSAs (processing schemes 1-4 for empirical and simulated) under the MSC model, using the DISSECT method (Jones et al. 2014) implemented in

STACEY (Jones 2017), which is available as a BEAST2 (Bouckaert et al. 2014) package. STACEY allows *BEAST analyses without prior taxonomic assignments, searching the tree space while simultaneously collapsing very shallow clades in the species tree (controlled by the parameter collapseHeight). This collapsing avoids a common violation of the MSC model that occurs when samples belonging to the same coalescent species are assigned to separate taxa in *BEAST. This feature makes STACEY suitable for analyzing allele sequences, because they do not have to be constrained to belong to the same taxon and can be treated as independent samples from a population. STACEY runs with the usual *BEAST operators, but integrates out the population size parameter and has new MCMC proposal distributions to more efficiently sample the species tree, which decreases the time until convergence. In order to reach even faster convergence, we reduced the number of loci for this analysis by selecting the 150 allele MSAs with the most parsimony informative sites. This selection was made for both the empirical and the simulated allele MSAs. The same 150 loci were selected for all other processing schemes.

Prior to analysis, we estimated the most appropriate substitution model for each of the 150 loci with jModeltest (Supplementary Table S1 available on Dryad) using BIC. We used BEAUTI v2.4.4 to create an input file for STACEY in which we unlinked substitution models, clock models and gene trees for all loci. We did not apply any taxon assignments, thereby treating every sequence as a separate taxon. We chose a strict clock for all loci and fixed the average clock rate for one random locus to 1.0, while estimating all other clock rates in relation to this locus. To ensure that all resulting species trees were scaled to an average clock rate of 1.0, we rescaled every species tree from the posterior distribution using the average clock rate of the respective MCMC step. We applied the STACEY-specific BirthDeathCollapse model as a species tree prior, choosing a value of 1e-5 for the collapseHeight parameter. Other settings were: bdcGrowthRate = log normal (M=4.6, S=1.5); collapseWeight = beta (alpha=2, beta=2); popPriorScale = log normal

(M=-7, S=2); relativeDeathRate = beta (alpha=1.0, beta=1.0). For the IUPAC consensus data, we enabled the processing of ambiguous sites by adding `useAmbiguities="true"` to the gene tree likelihood priors for all loci in the STACEY XML file. All analyses were run for 1,000,000,000 MCMC generations or until convergence (ESS values >200), logging every 20,000 generations. Convergence was assessed using Tracer v1.6 (Rambaut et al. 2013). We then summarized the posterior tree distribution into one maximum clade credibility tree with TreeAnnotator v2.4.4, discarding the first 10% of trees as burn-in.

For the simulated data, we analyzed the posterior species tree distributions of each analysis with the program SpeciesDelimitationAnalyser (part of the STACEY distribution). This program produces a similarity matrix that contains the posterior probabilities of belonging to the same cluster for each pair of sequences. This analysis was run with a collapseHeight value of 1e-5 (identical to the collapseHeight used in the STACEY analysis), while discarding the first 10% of trees as burn-in.

SNP-based tree estimation.— To estimate the species tree phylogeny from the extracted SNP data, we analyzed the empirical and simulated SNP data in SNAPP. We did not apply prior clade assignments to the samples in the SNP alignment (each sample was assigned as its own taxon), we set coalescent rate and mutation rates to be estimated based on the input data, and we chose a Yule species tree model with default settings ($\lambda =$ 0.00765). We ran the analysis for 10,000,000 generations, sampling trees and other parameters from the posterior every 1,000 generations. Unlike STACEY, SNAPP assumes correct assignments of all sequences to coalescent species. Using the simulated SNP data, we therefore tested how our approach of assigning every individual as its own coalescent species affects the resulting phylogenetic inference. We did so by running a separate analysis for both simulated SNP datasets (complete and reduced) with correct species assignments (assignments as in Fig. 6g).

Results

Mitochondrial Tree (BEAST)

The BEAST analysis of complete mitochondrial genomes (see online Appendix 1) produced a fully resolved topology (Fig. 3). All nodes were supported by 100% Bayesian posterior probability (PP). We inferred the divergence between the two lineages *T. pyra* and *T. pella* at 2.36 Ma, with 95% of the highest posterior density (HPD) ranging between 1.96 and 2.78 Ma. The tree also shows a separation of two distinct lineages within *T. pyra* at 0.68 Ma (95% HPD: 0.54 - 0.84 Ma), dividing the samples of this morphospecies into a northern and a southern clade, separated by the Amazon River (Fig. 1). A similar, yet slightly more recent split can be seen within *T. pella*. We inferred the age of this split to 0.39 Ma (95% HPD: 0.30 - 0.48 Ma), revealing the same pattern of one northern and one southern clade with the exception of sample 7; this sample from the southern bank of the Amazon River delta is placed together with the samples derived from localities north of the Amazon (samples 5 and 6). Below, we refer to those individuals sampled north of the Amazon River as “northern” and to those sampled south of the Amazon as “southern”.

UCE Summary Statistics

Alignment statistics.— We use the term “polymorphic sites” for those positions within a MSA alignment of a given locus where we find at least two different states at a particular position among the sequences for all samples. This does not require a particular individual being heterozygous for the given position, since we do not search for SNPs on a per sample basis but rather for SNPs within the genus *Topaza* (for the following statistics we are excluding the outgroup). In this manner, we found that the empirical UCE contig sequence

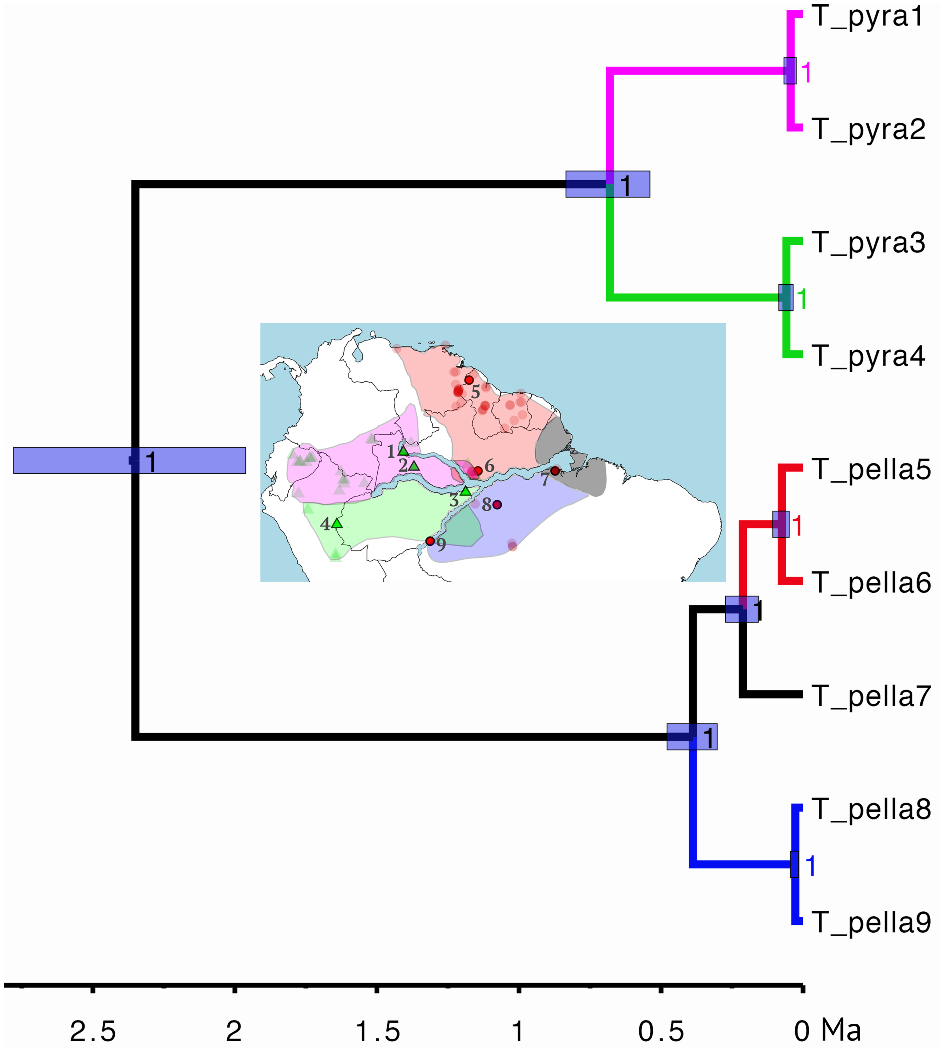


Figure 3: Phylogeny estimated from complete mitochondrial genomes in BEAST. Node support values represent PP. The blue bars at nodes represent the 95% HPD of divergence times. Scale axis shows time units in millions of years. The map in the center shows the potential ranges of the clades that are found in the mitochondrial tree (color-coded). The ranges are based on the BirdLife distribution ranges (Fig. 1) and have been expanded in order to accommodate all *Topaza* occurrence data.

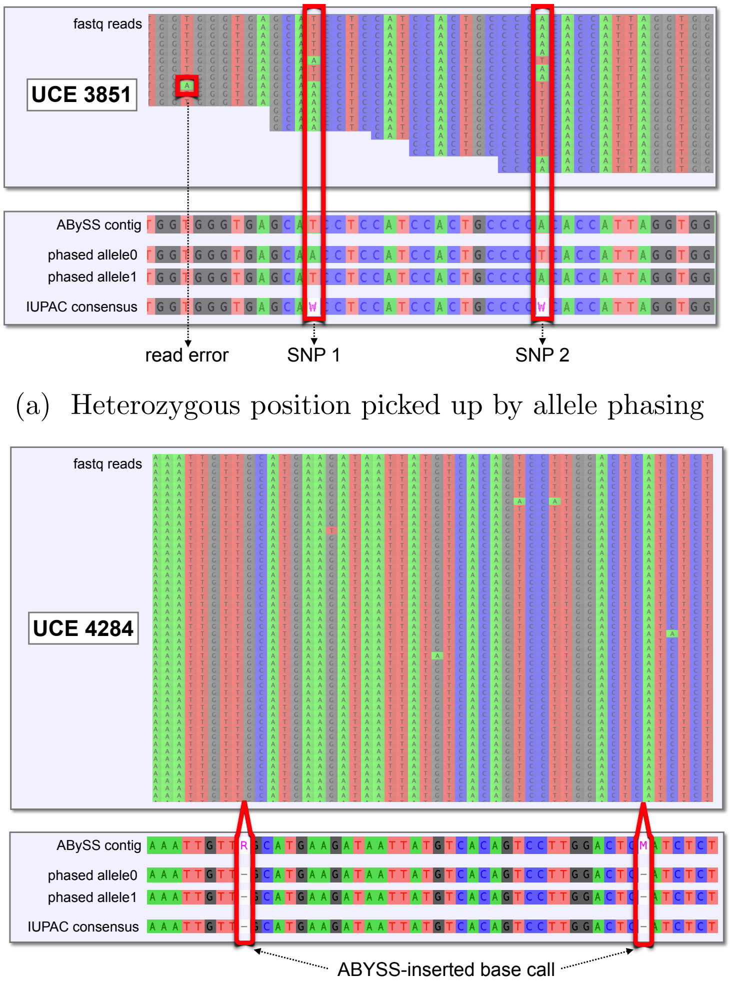
alignments had an average of 2.8 polymorphic sites per locus and an average alignment length of 870 bp. In contrast, phasing the empirical UCE data to create allele alignments led to 4.5 polymorphic sites per locus and an average alignment length of 848 bp, representing a 60% increase in polymorphic sites per locus. This increase of polymorphic sites was attributable to the fact that many variants get lost during contig assembly, because ABYSS and other tested contig assemblers, namely Trinity and Velvet, often eliminate one of the two variants at heterozygous positions (see below). The reduced length of the allele alignments in comparison to the contig alignments was due to conservative alignment clipping thresholds implemented in PHYLUCE, which clip alignment ends if less than 50% of sequences are present. Because the allele phasing algorithm divides the FASTQ reads into two allele bins and because a nucleotide is only called if it is supported by at least three high-quality FASTQ reads, we lost some of the nucleotide calls at areas of low read coverage (mostly at the ends of a locus) when comparing the allele sequences to the contig sequences. More information about the distribution of lengths and variable sites within the empirical UCE data can be found in the Supplementary Figs. S1 and S2 available on Dryad.

The simulated contig MSAs had an average of 3.2 polymorphic sites per locus, after excluding the outgroup. The simulated allele MSAs, on the other hand, contained an average of 5.4 polymorphic sites (69% increase). An overview of parsimony informative sites, variable sites and length of each alignment (simulated and empirical data) can be found in Supplementary Table S2 available on Dryad.

ABYSS does not detect heterozygous sites.— ABYSS occasionally produces contig sequences containing IUPAC ambiguity codes, which suggests that these sites may accurately represent heterozygosity in the read data and that assembly with ABYSS may be preferred to using other assembly algorithms because the resulting contigs contain more information. To validate this assumption, we checked one randomly selected sample (sample 5, *T. pella*) to see if degenerate sites in the contig sequences produced by ABYSS were heterozygous in the phased allele sequences. The results are striking, because there are zero heterozygous sites within the allele sequences for sample 5 that were correctly coded as IUPAC ambiguity codes in the ABYSS contigs (e.g. Fig 4a). Moreover, our phasing approach revealed 343 heterozygous UCE loci with a total of 728 SNPs in sample 5 while contig sequences from the ABYSS *de novo* assembly only contained IUPAC ambiguity codes (degenerate bases) at 26 UCE loci. For all other loci, ABYSS output homozygous contig sequences, indicated by the fact that all are free of ambiguity codes. Within the 26 loci containing IUPAC ambiguity codes, ABYSS introduced 473 degenerate bases, most of which constitute blocks of N's. Effectively, all of these ambiguous positions are in places of extremely low FASTQ read coverage (<2 reads per haplotype), with the exception of six positions that are covered by greater than two reads per haplotype. However, even those six positions do not represent true heterozygous sites within sample 5, which becomes apparent when comparing aligned FASTQ reads at those loci with the phased allele sequences with the contig sequences produced by ABYSS (e.g. Fig. 4b).

MSC Results of Empirical UCE Data

The MSC species tree results for all tested processing schemes of the empirical UCE data (contig sequences, allele sequences, IUPAC consensus sequences, chimeric allele sequences and SNPs) converge on similar topologies for relationships among *T. pella*, yet the relationships inferred among *T. pyra* are less clear (Fig. 5 and Supplementary Fig. S3 available on Dryad). All analyses strongly support the monophyly of both *T. pyra* and *T. pella* with 100% PP. In all MSC analyses, we also see strongly supported genetic structure within *T. pella* ($\geq$ 97% PP), separating the northern samples (5 and 6) from the southern ones (7, 8 and 9). Additionally, within the shallow southern *T. pella* clade, all datasets, with exception of the IUPAC consensus data (Fig. 5c), strongly support a genetic distinction ($\geq$ 99% PP) between sample 7 from the Amazon River delta and the other southern *T. pella* samples (8 and 9). The deep split between northern and southern samples within *T. pyra* on the other hand, which we find in the mitochondrial tree (Fig. 3), is not well-supported by the multilocus MSC analyses. However, the analysis of the



(a) Heterozygous position picked up by allele phasing

(b) Erroneous insertion of IUPAC ambiguity by ABYSS

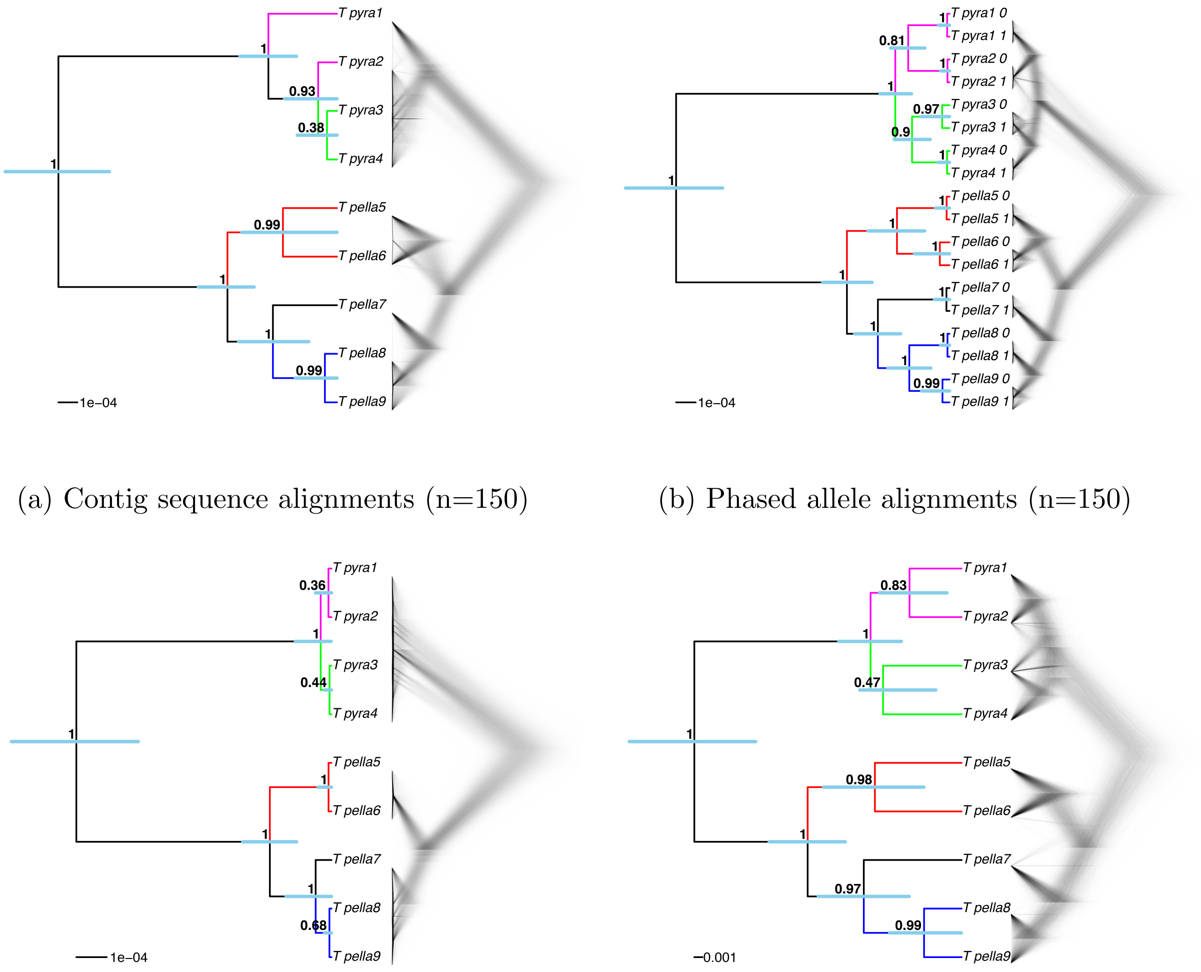
Figure 4: Detection of heterozygous sites in FASTQ reads. The figure shows two UCE loci for sample 5 (*T. pella*). Displayed in both cases are the FASTQ reads, the ABYSS contig sequence, the two phased allele sequences and the correct IUPAC consensus sequence generated from our phased allele sequences. (a) An example of true heterozygous sites, which are correctly represented in the phased allele sequences sequences but are not coded as IUPAC ambiguities in resulting ABYSS contig. Instead ABYSS makes a majority call for this position, thereby masking the heterozygous site by eliminating one of the two variants. This is the case for all heterozygous sites that were picked up by the allele sequences in our data. (b) An example of a UCE locus that contains IUPAC ambiguity codes in the ABYSS contig sequence. Contrary to expectations, the ambiguity calls at these positions are not supported by the FASTQ reads and appear to be inserted by ABYSS at random positions. Our phased allele sequences, on the other hand, represent the FASTQ reads correctly and do not call this position as heterozygous. We observed this same patters across all 26 loci in our data with ABYSS-inserted IUPAC ambiguity codes.

allele dataset returns a phylogenetic signal, possibly tracking a genetic divergence between these two clades, but their monophyly is not very strongly supported (Fig. 5b).

MSC Results of Simulated Data

Species tree topology.— For the simulated data, we analyzed six different datasets under the MSC model: contig sequence MSAs (n=150, STACEY), allele sequence MSAs (n=150, STACEY), IUPAC consensus MSAs (n=150, STACEY), chimeric allele MSAs (n=150, STACEY), reduced SNP data (n=150, SNAPP), and the complete SNP dataset (n=820, SNAPP). All resulting species trees (Figs. 6a to 6f) correctly return the topology of the species tree that was used to simulate the data (Fig. 6g). All central nodes in the species trees are supported by $\geq$90% PP in all analyses, with the exception of the species tree resulting from the reduced SNP dataset, which shows very weak support for two nodes and has a large uncertainty interval around the root-height (Fig. 6e). However, these shortcomings disappeared when we added more (unlinked) SNPs to the dataset (Fig. 6f). The full SNP dataset (n=820) produced the correct species tree topology with high node support consistently throughout ten independently simulated datasets (Supplementary Fig. S4 available on Dryad). The SNAPP species tree topology appeared to be unaffected by the chosen clade assignment model; while we allowed every sequence to be its own taxon in Figs. 6e and 6f, we also applied the correct species assignment (Fig. 6g) in two additional analyses (reduced and complete SNP data) that returned the same tree topology (Supplementary Figs. S5 and S6 available on Dryad).

Species delimitation.— Although the inferred species tree topology was consistent among all four sequence-based MSC analyses (Figs. 6a to 6d), the inferred node heights varied considerably between the species trees resulting from the different data processing schemes. For the contig sequence data (Fig. 6a) and the chimeric allele data (Fig. 6d), the node



(a) Contig sequence alignments (n=150)

(b) Phased allele alignments (n=150)

(c) IUPAC consensus alignments (n=150)

(d) SNPs (n=598)

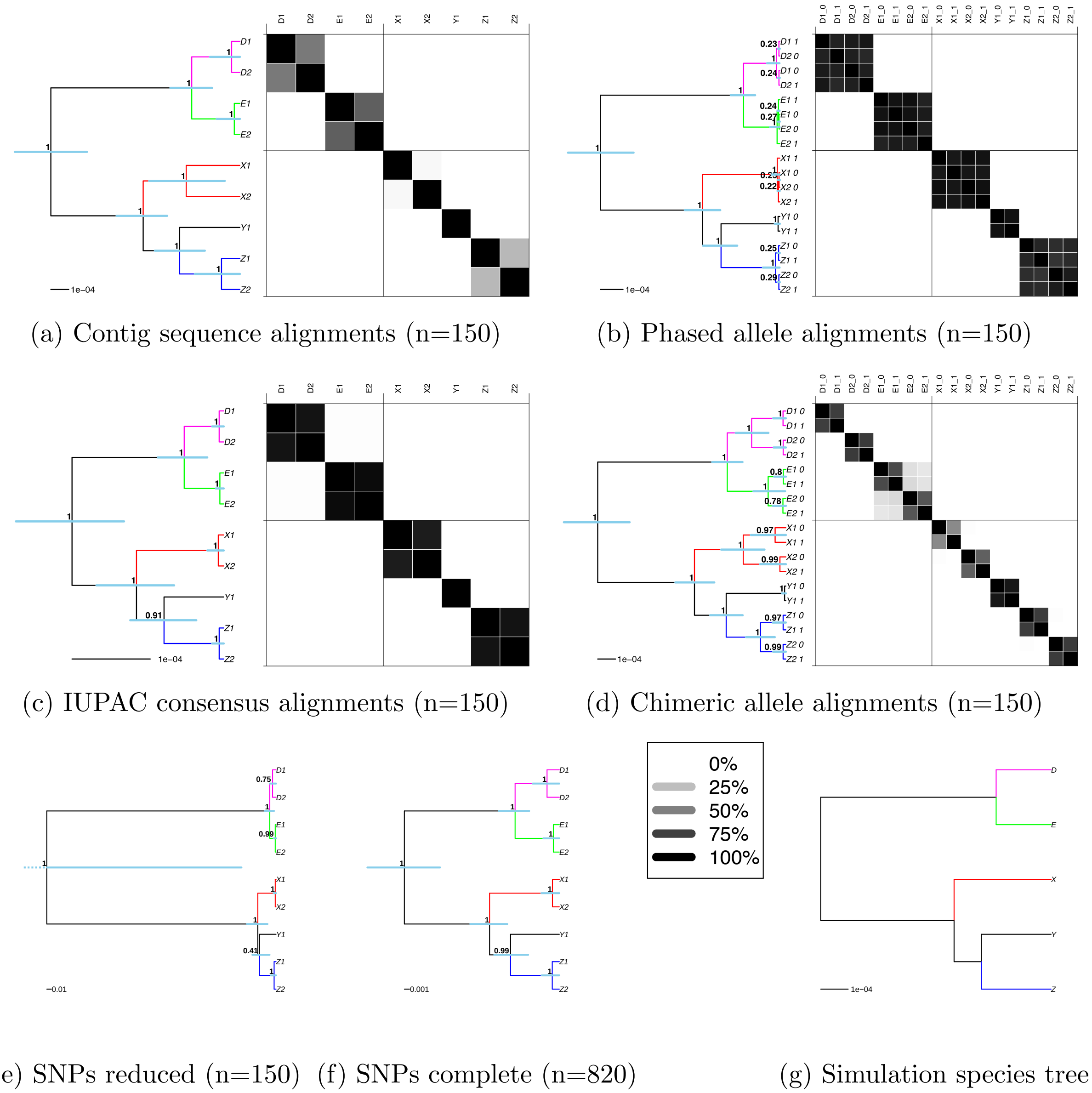
Figure 5: MSC species trees for the empirical *Topaza* data, based on four different data types used in this study: contig sequence MSAs, phased allele sequence MSAs, IUPAC consensus sequence MSAs and SNP data. (a) STACEY species tree from UCE contig alignments (n=150), (b) STACEY species tree from UCE allele alignments (n=150), (c) STACEY species tree from UCE IUPAC consensus alignments (n=150) and (d) SNAPP species tree from SNP data (1 SNP per locus if present, n=598). Shown are the maximum clade credibility tree (node values = PP, error-bars = 95% HPD of divergence times) and a plot of the complete posterior species tree distribution (excluding burn-in).

heights within the five simulated species (D,E,X,Y,Z) were too high, which led to an overestimation of the number of coalescent species in the dataset (see similarity matrices). Conversely, the phased allele data (Fig. 6b) and the IUPAC consensus data (Fig. 6c) correctly delimited the five coalescent species from the simulation input tree (Fig. 6g). The STACEY results showed the same pattern in all ten simulation replicates (Supplementary Fig. S7 available on Dryad).

Accuracy of divergence time estimation.— For all four sequence-based analyses (Figs. 6a to 6d) the average substitution rate across all loci was set to '1'. Under these settings, we expected the absolute values of the sequence-based analyses to return the node height values of the simulation input tree, which used substitution rates scaled in the same manner. The phased allele MSAs produced the most accurate estimation of divergence times out of all tested datasets (see proximity of estimates to simulation input value, represented by green line in Fig. 7). This was the case for all nodes in the species tree, namely (D,E), (Y,Z), (X,(Y,Z)), and ((D,E)(X,(Y,Z))). The divergence time estimates resulting from the phased allele data accurately recovered the true values and did not show any bias throughout ten simulation replicates (Supplementary Fig. S8 available on Dryad). This contrasts with the contig MSAs and the chimeric allele MSAs that consistently overestimated the height of all nodes and the IUPAC consensus MSAs which consistently underestimated the height of all nodes (Fig. 7, Supplementary Fig. S8).

Additional Analyses

We ran additional analyses of the contig and the phased allele MSAs for both the empirical and simulated data using a summary coalescent approach as implemented in MP-EST (Yu et al. 2007), which can be found in online Appendix 2 and Supplementary Figs. S9 to S11 (available on Dryad).



(a) Contig sequence alignments (n=150)

(b) Phased allele alignments (n=150)

(c) IUPAC consensus alignments (n=150)

(d) Chimeric allele alignments (n=150)

(e) SNPs reduced (n=150)

(f) SNPs complete (n=820)

(g) Simulation species tree

Figure 6: MSC species tree results for different data processing schemes of simulated data. (a) to (d) show the STACEY results of the four different types of MSAs analyzed in this study (see sub-figure captions). Displayed in these panels are the maximum clade credibility trees and the similarity matrices depicting the posterior probability of two samples belonging to the same clade, as calculated with SpeciesDelimitationAnalyser. Dark panels depict a high pairwise similarity, whereas light panels depict low similarity scores (see legend). (e) and (f) show the maximum clade credibility trees resulting from SNAPP for our two SNP datasets, (reduced and complete). (g) shows the species tree under which the sequence data were simulated in this study. Node support values in PP, blue bars representing 95% HPD confidence intervals.

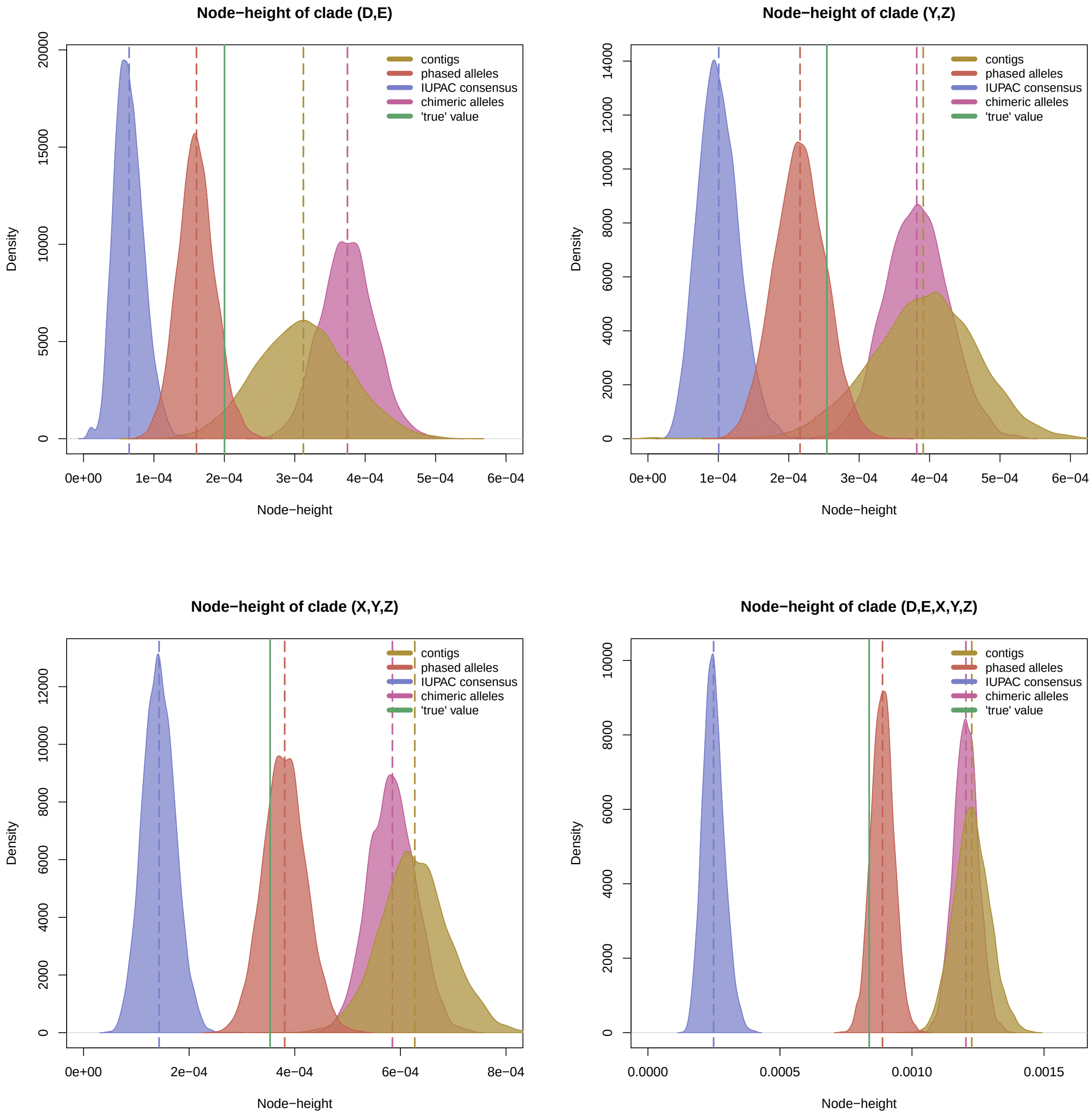


Figure 7: Posterior distributions of divergence times estimated with STACEY. Each subfigure shows several density plots of node-height estimates for a different node in the STACEY species tree (see sub-figure titles). The four density plots in each sub-figure are approximated from all node height values in the posterior sample (excl. 10% burnin), as estimated by STACEY for the four different data processing schemes tested in this study: contig sequences (yellow), phased allele sequences (red), IUPAC consensus sequences (blue) and chimeric allele sequences (pink). The dotted lines show the means of these posterior distributions. The solid green line shows the true node height value, which is the node height for the respective clade in the input species tree, under which the sequence alignments were simulated.

Discussion

Allele Phasing is the Preferred Data Processing Scheme

We tested whether phylogenetic inference improves by phasing sequence capture data into allele sequences, in comparison to the standard workflow of analyzing contig sequences (Faircloth et al. 2012; McCormack et al. 2012; Smith et al. 2014; Faircloth 2015). The answer is yes. We find that phased allele data outperform contig sequences in terms of species delimitation (Fig. 6) and the estimation of divergence times (Fig. 7). Contig sequence MSAs lead to a consistent overestimation of divergence times (Fig. 7), which in turn lead to an overestimation of the number of coalescent species in our simulated data (Fig. 6a). These results support earlier work by Lischer et al. (2014), who concluded that consensus sequences introduce a bias towards older node heights.

Besides the qualitative advantages of using phased allele sequences for phylogenetic analyses, there are further theoretical arguments for compiling and analyzing allele sequence MSAs from sequence capture datasets. First, allele sequences represent the smallest evolutionary unit on which selection and other evolutionary processes act. Therefore, the coalescent models that underly our phylogenetic methods, including the MSC model Degnan and Rosenberg (2009), have been developed for allele sequences. Contig sequences, on the other hand, represent an artificial and possibly chimeric sequence construct that arises from merging all read variation at a given locus into a single sequence. This process masks information by eliminating one of the two variants at a heterozygous site (Fig. 4). This shortcoming of the most common assemblers (e.g. ABYSS, Trinity and Velvet) is due to the fact that they were designed to assemble haploid sequences and are not optimized for heterozygous sequences or genomes (Bodily et al. 2015). Second, not only are allele sequences the more appropriate data type, but phasing sequence capture data also leads to a doubling of the effective sample size, since two sequences are compiled for a diploid individual, in contrast to the single sequence per individual that is recovered when taking the contig approach. Our results demonstrate how these sequences can be treated as independent samples from a population by using the assignment-free BirthDeathCollapse model as implemented in STACEY. Because STACEY requires no *a priori* assignment of sequences to taxon, this avoids a violation of the MSC that would occur when analyzing allele sequences as separate taxa in *BEAST, because *BEAST assumes each taxon constitutes a separate coalescent species. Third, sequence capture datasets such as UCEs are optimal for allele phasing because they contain high read coverage collected across short genomic intervals that are optimal for read-connectivity based phasing. The workflow developed in this study is now fully integrated into the PHYLUCE pipeline, making allele phasing and SNP extraction for sequence capture data easily available to a broad user group. Given these advantages of allele sequences over contig sequences and given the easy availability of the processing workflow, we recommend that allele phasing be considered as a standard practice for future sequence capture studies.

Phasing of Heterozygous Sites Matters

Several studies have accounted for heterozygosity by inserting IUPAC ambiguity codes into their sequences at variable positions (Potts et al. 2014; Schrempf et al. 2016), rather than phasing SNPs to produce separate allele sequences. Here, we directly compared these two approaches, and found that the IUPAC consensus sequences performed equally well to the phased allele sequences for estimating the species tree topology (Fig. 6). However, IUPAC consensus sequence data led to a consistent underestimation of the divergence times of all nodes in the species tree (Fig. 7). Our results contrast with those of (Lischer et al. 2014), who reported an overestimation of divergence times for alignments containing IUPAC ambiguity codes. The differences between our results may simply be caused by the different tree inference programs that we used. Lischer et al. (2014) applied a Neighbour Joining (NJ) tree algorithm as implemented in the software PHYLIP (Felsenstein 2005) that treats two sequences containing the same ambiguity codes as identical. In effect, the approach used by Lischer et al. (2014) did not truly investigate the effect of IUPAC ambiguity codes on phylogenetic estimates but rather the effect of removing heterozygous sites. Our approach of analyzing IUPAC consensus sequences under the MSC in STACEY, on the other hand, properly integrates these IUPAC ambiguity codes into the calculation of the gene tree likelihoods. Thus, we conclude that IUPAC ambiguity codes introduce a bias towards younger divergence times, even when properly integrating IUPAC ambiguities into the phylogenetic model. The underlying cause of this discrepancy should be further investigated in future studies.

We also tested whether the improved performance of phased allele sequence data may merely be an effect of doubling the number of sequences in the MSAs, since we are producing two allele sequences for each individual rather than one contig sequence. Therefore, we generated a dataset of chimeric allele sequences that contains the same number of sequences as the phased allele data, but we randomly shuffled all heterozygous positions within an individual between the two allele sequences. As with the contig data, the chimeric allele data led to an overestimation of the number of coalescent species (Fig. 6d) and to a biased estimation towards older divergence times (Fig. 7). The fact that contig sequences and chimeric allele sequences produce very similar results in our analyses is not surprising, because contigs, themselves, represent chimeric consensus sequences of the variation found at a locus within an individual. The similarity of the results between contig MSAs and chimeric allele MSAs also shows that the number of sequences being analyzed does not affect our topology, species delimitation and divergence time estimates (Figs. 6 and 7).

Based on these findings, we conclude that proper phasing of heterozygous positions is clearly preferable to the alternative of coding heterozygous sites as IUPAC ambiguity codes, particularly when the estimation of divergence times is of interest. Further, allele sequences are theoretically more appropriate input for coalescent models and should be the preferred data type input to these models. The scalability of this approach to larger sample sizes and the applicability of our results to studies of older divergences are questions that should be investigated in future studies.

UCEs as source for SNP data

Due to the size (number of loci) of many sequence capture datasets, it is often unfeasible to analyze all MSAs jointly in one MSC analysis (Smith et al. 2014; Manthey et al. 2016) because of computational limitations. For all sequence-based MSC analyses in this study, we reduced the UCE dataset from 820 loci to 150 loci in order to reach convergence of the MCMC within a reasonable time frame (three to four days, single core on a Mac Pro, Late 2013, 3.5 GHz 6-Core Intel Xeon E5 processor). However, a viable approach to data reduction, while keeping the multilocus information of all loci, is to analyze only a single polymorphic position per MSA using SNAPP (Bryant et al. 2012). In our study, this approach produces the correct species tree topology and also estimated the relative node-heights correctly (Fig. 6f). However, SNAPP can only estimate relative and not absolute values for divergence times (Bryant et al. 2012), in contrast to the sequence-based analyses Figs. 6a to 6d that deliver absolute divergence time estimates.

Sequence capture datasets such as UCEs provide a suitable data source to extract both full sequence alignments and SNP datasets of sufficient size for robust species tree estimation. Even though sequence capture data are not commonly thought of as a source of SNPs, they can, in many cases, be preferable to other sequencing techniques, such as RAD sequencing, for producing SNP data. This is because sequence capture data yields a sizable, complete SNP matrix (SNPs recovered for all individuals), due to targeted sequence enrichment. In this study, the complete matrix of unlinked SNPs in the empirical data consisted of 598 positions, which were present and sufficiently supported (>three high-quality reads per haplotype) in all taxa. Particularly when evolutionary distances between individuals are large, RAD sequencing and other restriction-site based sequencing techniques are not expected to yield many loci shared by all individuals, whereas UCE data are less sensitive to large evolutionary distances (Harvey et al. 2016). In these cases, the size of the complete SNP matrix resulting from UCE data can exceed that resulting from RAD sequencing. Additionally, UCE data provide hundreds to thousands of full sequence MSAs as well as the complete mitochondrial genome as a byproduct of the sequence enrichment. The mitochondrial genome provides an excellent marker for estimating absolute divergence times (Fig. 3), based on substitution rates of mitochondrial markers which are known for birds (Lerner et al. 2011), and thus remains a valuable source of phylogenetic information.

In this study, we present and make available a new SNP calling pipeline for sequence capture data. In contrast to other SNP calling software such as GATK (McKenna et al. 2010) that uses BAM files, our approach uses full sequence MSAs as input (see Fig. 2), in order to identify and extract sites in the alignments that show variation between any user-defined group of sequences. Although our SNP calling script can be applied to any type of sequence alignments (i.e. allele or contig sequence alignments), we recommend using SNPs extracted from phased allele alignments for phylogenetic analyses, because they represent the true heterozygous information. The user can choose whether or not to allow missing data or ambiguities in the extracted positions, whether to extract them in binary format (as e.g. required by SNAPP) or as nucleotides, and if only a single SNP per locus or all SNPs should be extracted. Thus our SNP calling mechanism is an easy, open-source and straightforward tool to derive SNP data from any set of multiple sequence alignments.

Phylogenetic relationships in Topaza

One or two species?.— Our results show a separation of two lineages within the genus *Topaza* that is dated at ca. 2.4 Ma in the mitochondrial tree (Fig. 3). These lineages are consistent with the previously described morphospecies *T. pyra* (Gould, 1846) and *T. pella* (Linnaeus, 1758) that are generally accepted in the ornithological community (Hu et al. 2000; del Hoyo et al. 2016a). However, the species status of *T. pyra* has been challenged by some authors (Ornés-Schmitz and Schuchmann 2011; Schuchmann 1999). These authors concluded that *Topaza* is a monotypic genus with *T. pyra* being a subspecies of *T. pella*, which they refer to as *T. pella pyra*. Their findings are based on the analyses of plumage coloration, in which they found an "east-west clinal trend of characters" (Ornés-Schmitz and Schuchmann 2011). In contrast, we do not find such an east-west clinal trend in the genetic data. Instead, *T. pyra* is consistently supported as a separate lineage across all analyses, lending no support for the conspecificity of these two taxa (Figs. 3 and 5).

One aim of this study was to evaluate the genetic structure within these two morphospecies, *T. pyra* and *T. pella*. The mitochondrial tree shows two divergent clades within *T. pyra* (Fig. 3), but these clades are not strongly supported by the UCE data (Fig. 5), even though the allele sequence data are picking up a signal that possibly indicates two clades are in the process of diversifying (Fig. 5b). For *T. pella*, on the other hand, we consistently find the same clades throughout all multilocus MSC analyses (Fig. 5), leading us to distinguish between the following populations that are congruent with previous morphological subspecies descriptions:

Northern T. pella population: T. pella pella.— For the mitochondrial tree and all MSC species trees, we find the northern *T. pella* samples 5 and 6 to be sister taxa with high support values (98-100% PP, Figs. 3 and 5). Particularly in the mitochondrial tree (Fig. 3), these two samples appear as close sister taxa, separated by only very short terminal branches. Their close position in the mitochondrial tree shows that, even though geographically far apart, samples 5 and 6 share a relatively recent MRCA in the mitochondrial genealogy, indicating some rather recent gene flow. The sampling locality of sample 5 is within the range of the subspecies *T. pella pella*, which extends mainly across the Guiana shield (Peters 1945; Schuchmann 1999; Hu et al. 2000; Ornés-Schmitz and Schuchmann 2011). Given the sampling location of genetically related sample 6, which also has been morphologically identified as *T. pella pella* (Table 1), we propose that the distribution range of *T. pella pella* extends from the Guiana shield all the way south to the northern Amazon River bank (see map in Fig. 3).

Southern T. pella population: T. pella microrhyncha.— In the same manner as for the northern population *T. pella pella*, we also consistently find the southern *T. pella* samples 8 and 9 to be sister taxa (99-100% PP, Figs. 3 and 5). The sampling locations of these two samples are included in the distribution range of the previously recognized subspecies *T. pella microrhyncha*, extending from the southern bank of the Amazon River as far South as Porto Velho (Brazil) at the Madeira River, close to the border to Bolivia (Peters 1945; Schuchmann 1999; Ornés-Schmitz and Schuchmann 2011). This southernmost boundary of *T. pella microrhyncha* is not accepted by Hu et al. (2000), who instead conclude that this southernmost population belongs to *T. pella pella*. In contrast to the findings by Hu et al., our genetic data clearly support the southernmost sample 9 belonging to the same population as sample 8, which was morphologically identified as *T. pella microrhyncha*. This leads us to propose that the distribution range of *T. pella microrhyncha* is in fact as shown in Fig. 3, in agreement with the findings by Peters (1945),Schuchmann (1999), and Ornés-Schmitz and Schuchmann (2011).

Estuary region of Amazon River: T. pella smaragdula.— Our results show a mixed signal concerning the phylogenetic placement of sample 7, which was collected from the southern estuary region of the Amazon River and morphologically identified as *T. pella smaragdula*. The sampling locality also falls into the range of the subspecies *T. pella smaragdula* (Peters 1945; Hu et al. 2000; Ornés-Schmitz and Schuchmann 2011), with a distribution including the Amazon River estuary and extending north along the coast to French Guiana. All MSC analyses of the UCE sequence and SNP data place sample 7 with high confidence (97-100% PP) as sister to the southern clade *T. pella microrhyncha* (Fig. 5), whereas in the mitochondrial phylogeny this sample is placed as sister to *T. pella pella* in the North.

The discordance between a gene tree and the species tree in a scenario such as this could be the effect of incomplete lineage sorting, which is most likely if the species or clades in question have diverged rather recently and if population sizes are large. Given that the divergence between *T. pella pella* and *T. pella microrhyncha* appears to be considerably deep based on the multilocus data (crown height of *T. pella*) see Fig. 5) and given that mitochondria are generally considered to have only 25% of the population size of nuclear loci, it is rather unlikely that the position of sample 7 in the mitochondrial tree is the result of incomplete lineage sorting in this case. It seems more likely that the separate position of sample 7 in the mitochondrial tree is the result of introgression of the mitochondrial genome from *T. pella pella* into the gene pool of *T. pella smaragdula*. However, a denser taxon sampling would be necessary to further evaluate the evolutionary history of this particular population. The case of sample 7 highlights that the mitochondrial tree presents a single gene tree phylogeny that only shows one of many genealogies and therefore must not be equated with a species tree phylogeny. Hence it is important to generate multilocus data for an informed inference of the species tree phylogeny.

Summarizing biogeographic remarks.— The presence of genetically similar individuals sampled at great geographic distances (e.g. samples 5 and 6) suggests that *Topaza* hummingbirds maintain high levels of gene flow across vast distances of rainforest habitat.

At the same time, we find indicators of phylogenetic structure within species, distinguishing samples that are separated by only a small geographic distance (see e.g. samples 6 and 8). These samples are however separated by the Amazon River, which has been found to constitute a dispersal barrier for various species of birds and many other animals (Remsen and Parker 1983; Clair 2003; Hayes and Sewlal 2004; Moore et al. 2008; Fernandes et al. 2012; Ribas et al. 2012; Thom and Aleixo 2015). Even though some hummingbird species are known to disperse across large distances (Wyman et al. 2004; Russell et al. 1994), the Amazon River and its associated habitats (such as seasonally flooded forests) may be part of a complex network of factors that inhibit gene flow among populations of *Topaza* hummingbirds.

Conclusions

In this study, we demonstrate that properly phasing allele sequences produces the most suitable dataset for phylogenetic analyses, particularly when these allele sequences are treated as independent sequences under the MSC. Contig sequences, on the other hand, which are commonly used for phylogenetic inference, lead to biases in the estimation of divergence times and may cause problems for certain types of phylogenetic analyses. Additionally, phased allele sequences provide a useful template for the extraction of SNPs, and we argue that sequence capture data can provide sizable SNP datasets that can be also used for phylogenetic analyses. Our empirical results suggest the separation of two species within the genus *Topaza*, and we further find genetic structure within both of these species, justifying the definition of separate subspecies. Based on our empirical and simulated results, we conclude that allele phasing should be considered as one “best practice” for processing sequence capture data, although the sample-size, time, and analytical limitations of this approach have not been well-established.

Supplementary Material

Supplemental Figs. S1-S11, Supplemental Tables S1 and S2, online Appendices 1 and 2 and all scripts, data and setup-files relevant to analyses and figures in the manuscript are available from the Dryad Digital Repository:

Availability

We integrated all scripts and documentation necessary for phasing and SNP extraction as open-source into the PHYLUCE pipeline (`http://https://github.com/faircloth-lab/phyluce/blob/working/bin/snps/`). All data processing and analyses steps executed on the data are stored in bash-scripts on our project GitHub page at `https://github.com/tobiashofmann88/topaza_uce`. We further provide a documented workflow of processing the raw reads into UCE contig alignments at `https://github.com/tobiashofmann88/UCE-data-management/wiki`.

Acknowledgments

We wish to thank all those ornithologists who have dedicated their time to collecting samples in Amazonia; museum curators for providing us with samples for this study; Brazilian authorities for issuing the permits needed for this work; our lab engineer Anna Ansebo for laboratory assistance; Alexander Zizka for assistance in creating the sampling and range maps; HBW Alive for providing the *Topaza* illustrations; and colleagues at our labs for discussions and feedback. Computational analyses were performed on the bioinformatics computer cluster Albiorix at the Department of Biological and Environmental Sciences, University of Gothenburg.

Funding

This work was funded by the Swedish Research Council to A. Antonelli (B0569601) and B. Oxelman (2012-3917); the CNPq (grants 310593/2009-3; 'INCT em Biodiversidade e Uso da Terra da Amazônia' 574008/2008-0; 563236/2010-8; and 471342/2011-4), FAPESPA (ICAAF 023/2011), and NSF-FAPESP (grant 1241066 - Dimensions US-BIOTA-São Paulo: Assembly and evolution of the Amazonian biota and its environment: an integrated approach) to A. Aleixo; the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013, ERC Grant Agreement n. 331024), the Swedish Foundation for Strategic Research and a Wallenberg Academy Fellowship to A. Antonelli.

*

References

Bodily, P. M., M. Fujimoto, C. Ortega, N. Okuda, J. C. Price, M. J. Clement, and Q. Snell. 2015. Heterozygous genome assembly via binary classification of homologous sequence. BMC Bioinformatics 16:S5.

Bolger, A. M., M. Lohse, and B. Usadel. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–20.

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

Bryant, D., R. Bouckaert, J. Felsenstein, N. A. Rosenberg, and A. RoyChoudhury. 2012. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. Molecular Biology and Evolution 29:1917–32.

Clair, C. C. S. 2003. Comparative permeability of roads, rivers, and meadows to songbirds in Banff national park. Conservation Biology 17:1151–1160.

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

del Hoyo, J., N. Collar, G. Kirwan, and P. Boesman. 2016a. Fiery Topaz (*Topaza pyra*). *in* Handbook of the Birds of the World Alive (J. del Hoyo, A. Elliott, J. Sargatal, D. Christie, and E. de Juana, eds.). Lynx Edicions, Barcelona, Spain.

del Hoyo, J., A. Elliott, J. Sargatal, D. Christie, and E. de Juana. 2016b. Handbook of the Birds of the World Alive. Lynx Edicions, Barcelona, Spain.

Drummond, A. J., M. A. Suchard, D. Xie, and A. Rambaut. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Molecular Biology and Evolution 29:1969–73.

Edwards, S. V., L. Liu, and D. K. Pearl. 2007. High-resolution species trees without concatenation. Proceedings of the National Academy of Sciences 104:5936–5941.

Eriksson, J. S., J. L. Blanco-Pastor, F. Sousa, Y. J. Bertrand, and B. E. Pfeil. 2017. A cryptic species produced by autopolyploidy and subsequent introgression involving Medicago prostrata (Fabaceae). Molecular Phylogenetics and Evolution 107:367–381.

Faircloth, B. C. 2015. PHYLUCE is a software package for the analysis of conserved genomic loci. Bioinformatics 32:786–788.

Faircloth, B. C., J. E. McCormack, N. G. Crawford, M. G. Harvey, R. T. Brumfield, and T. C. Glenn. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Systematic Biology 61:717–26.

Faircloth, B. C., L. Sorenson, F. Santini, and M. E. Alfaro. 2013. A phylogenomic perspective on the radiation of ray-finned fishes based upon targeted sequencing of ultraconserved elements (UCEs). PLoS ONE 8:e65923.

Felsenstein, J. 2005. Phylip (phylogeny inference package) version 3.6. distributed by the author. dep genome sci univ washington, seattle.

Fernandes, A. M., M. Wink, and A. Aleixo. 2012. Phylogeography of the chestnut-tailed antbird (*Myrmeciza hemimelaena*) clarifies the role of rivers in Amazonian biogeography. Journal of Biogeography 39:1524–1535.

Garrick, R. C., P. Sunnucks, and R. J. Dyer. 2010. Nuclear gene phylogeography using PHASE: dealing with unresolved genotypes, lost alleles, and systematic bias in parameter estimation. BMC Evolutionary Biology 10:118.

Giarla, T. C. and J. A. Esselstyn. 2015. The challenges of resolving a rapid, recent radiation: empirical and simulated phylogenomics of philippine shrews. Systematic Biology 64:727–740.

Gnirke, A., A. Melnikov, J. Maguire, P. Rogov, E. M. LeProust, W. Brockman, T. Fennell, G. Giannoukos, S. Fisher, C. Russ, S. Gabriel, D. B. Jaffe, E. S. Lander, and C. Nusbaum. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nature Biotechnology 27:182–189.

Harvey, M. G., B. T. Smith, T. C. Glenn, B. C. Faircloth, and R. T. Brumfield. 2016. Sequence capture versus restriction site associated DNA sequencing for shallow systematics. Systematic Biology Advance Access syw036.

Hayes, F. E. and J. A. N. Sewlal. 2004. The Amazon River as a dispersal barrier to passerine birds: effects of river width, habitat and taxonomy. Journal of Biogeography 31:1809–1818.

He, D., A. Choi, K. Pipatsrisawat, A. Darwiche, and E. Eskin. 2010. Optimal algorithms for haplotype assembly from whole-genome sequence data. Bioinformatics 26:i183–i190.

Hu, D.-S., L. Joseph, and D. J. Agro. 2000. Distribution, variation, and taxonomy of *Topaza* Hummingbirds (Aves: Trochilidae). Ornitologia Neotropical 11:123–142.

Iqbal, Z., M. Caccamo, I. Turner, P. Flicek, and G. McVean. 2012. De novo assembly and genotyping of variants using colored de Bruijn graphs. Nature Genetics 44:226–232.

Jones, G. 2017. Algorithmic improvements to species delimitation and phylogeny estimation under the multispecies coalescent. Journal of Mathematical Biology 74:447–467.

Jones, G., Z. Aydin, and B. Oxelman. 2014. DISSECT: an assignment-free Bayesian discovery method for species delimitation under the multispecies coalescent. Bioinformatics 31:991–998.

Katoh, K., G. Asimenos, and H. Toh. 2009. Multiple alignment of DNA sequences with MAFFT. Methods in Molecular Biology 537:39–64.

Kolaczkowski, B. and J. W. Thornton. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature 431:980–984.

Kubatko, L. S. and J. H. Degnan. 2007. Inconsistency of Phylogenetic Estimates from Concatenated Data under Coalescence. Systematic Biology 56:17–24.

Leaché, A. D., B. L. Banbury, J. Felsenstein, A. N. M. De Oca, and A. Stamatakis. 2015. Short tree, long tree, right tree, wrong tree: New acquisition bias corrections for inferring SNP phylogenies. Systematic Biology 64:1032–1047.

Leaché, A. D. and J. R. Oaks. 2017. The Utility of Single Nucleotide Polymorphism (SNP) Data in Phylogenetics. Annual Review of Ecology, Evolution, and Systematics 48:69–84.

Lerner, H. R., M. Meyer, H. F. James, M. Hofreiter, and R. C. Fleischer. 2011. Multilocus resolution of phylogeny and timescale in the extant adaptive radiation of Hawaiian honeycreepers. Current Biology 21:1838–1844.

Li, H. and R. Durbin. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26:589–595.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–9.

Lischer, H. E., L. Excoffier, and G. Heckel. 2014. Ignoring heterozygous sites biases phylogenomic estimates of divergence times: Implications for the evolutionary history of microtus voles. Molecular Biology and Evolution 31:817–831.

Manthey, J. D., L. C. Campillo, K. J. Burns, and R. G. Moyle. 2016. Comparison of target-capture and restriction-site associated DNA sequencing for phylogenomics: a test in cardinalid tanagers (Aves, Genus: *Piranga*). Systematic Biology Advance Access syw005.

McCormack, J. E., B. C. Faircloth, N. G. Crawford, P. A. Gowaty, R. T. Brumfield, and T. C. Glenn. 2012. Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. Genome Research 22:746–754.

McGuire, J., C. C. Witt, J. V. Remsen, A. Corl, D. L. Rabosky, D. L. Altshuler, and

R. Dudley. 2014. Molecular phylogenetics and the diversification of hummingbirds. Current Biology 24:910–916.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research 20:1297–303.

Meiklejohn, K. A., B. C. Faircloth, T. C. Glenn, R. T. Kimball, and E. L. Braun. 2016. Analysis of a rapid evolutionary radiation using ultraconserved elements (UCEs): Evidence for a bias in some multispecies coalescent methods. Systematic Biology Advance Access syw014.

Milne, I., G. Stephen, M. Bayer, P. J. A. Cock, L. Pritchard, L. Cardle, P. D. Shaw, and D. Marshall. 2013. Using Tablet for visual exploration of second-generation sequencing data. Briefings in Bioinformatics 14:193–202.

Mirarab, S., R. Reaz, M. S. Bayzid, T. Zimmermann, M. S. Swenson, and T. Warnow. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30:541–548.

Moore, R. P., W. D. Robinson, I. J. Lovette, and T. R. Robinson. 2008. Experimental evidence for extreme dispersal limitation in tropical forest birds. Ecology Letters 11:960–968.

Mossel, E. and E. Vigoda. 2005. Phylogenetic MCMC algorithms are misleading on mixtures of trees. Science 309:2207–9.

Ornés-Schmitz, A. and K. L. Schuchmann. 2011. Taxonomic review and phylogeny of the hummingbird genus *Topaza* (Gray, 1840) using plumage color spectral information. Ornitologia Neotropical Pages 25–38.

Peters, J. L. 1945. Check-list of birds of the world. Volume 5 ed. Harvard Univ. Press, Cambridge, Massachusetts.

Potts, A. J., T. A. Hedderson, and G. W. Grimm. 2014. Constructing Phylogenies in the Presence Of Intra-Individual Site Polymorphisms (2ISPs) with a Focus on the Nuclear Ribosomal Cistron. Systematic Biology 63:1–16.

Rambaut, A., M. A. Suchard, W. Xie, and A. Drummond. 2013. Tracer v1.6.

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

Remsen, J. V. and T. A. Parker. 1983. Contribution of river-created habitats to bird species richness in Amazonia. Biotropica 15:223–231.

Ribas, C. C., a. Aleixo, a. C. R. Nogueira, C. Y. Miyaki, and J. Cracraft. 2012. A palaeobiogeographic model for biotic diversification within Amazonia over the past three million years. Proceedings of the Royal Society B: Biological Sciences 279:681–689.

Russell, R. W., F. L. Carpenter, M. A. Hixon, and D. C. Paton. 1994. The impact of variation in stopover habitat quality on migrant rufous hummingbirds. Conservation Biology 8:483–490.

Schrempf, D., B. Q. Minh, N. De Maio, A. von Haeseler, and C. Kosiol. 2016. Reversible polymorphism-aware phylogenetic models and their application to tree inference. Journal of Theoretical Biology 407:362–370.

Schuchmann, K., G. Kirwan, and P. Boesman. 2016. Crimson Topaz (*Topaza pella*). *in* Handbook of the Birds of the World Alive (J. del Hoyo, A. Elliott, J. Sargatal, D. Christie, and E. de Juana, eds.). Lynx Edicions, Barcelona, Spain.

Schuchmann, K. L. 1999. Family Trochilidae (hummingbirds). Pages 468–680 *in* Handbook of the Birds of the World Alive (J. del Hoyo, A. Elliott, and J. Sargatal, eds.) volume 5 ed. Lynx Edicions, Barcelona, Spain.

Simpson, J. T., K. Wong, S. D. Jackman, J. E. Schein, S. J. M. Jones, and I. Birol. 2009. ABySS: a parallel assembler for short read sequence data. Genome Research 19:1117–23.

Smith, B. T., M. G. Harvey, B. C. Faircloth, T. C. Glenn, and R. T. Brumfield. 2014. Target capture and massively parallel sequencing of ultraconserved elements for comparative studies at shallow evolutionary time scales. Systematic Biology 63:83–95.

Sullivan, B. L., C. L. Wood, M. J. Iliff, R. E. Bonney, D. Fink, and S. Kelling. 2009. eBird: A citizen-based bird observation network in the biological sciences. Biological Conservation 142:2282–2292.

Thom, G. and A. Aleixo. 2015. Cryptic speciation in the white-shouldered antshrike (*Thamnophilus aethiops*, Aves - Thamnophilidae): The tale of a transcontinental radiation across rivers in lowland Amazonia and the northeastern Atlantic Forest. Molecular Phylogenetics and Evolution 82:95–110.

Wyman, S. K., R. K. Jansen, and J. L. Boore. 2004. Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20:3252–5.

Yang, Z. 2015. The BPP program for species tree estimation and species delimitation. Current Zoology 61:854–865.

Yu, L., Y.-W. Li, O. a. Ryder, and Y.-P. Zhang. 2007. Analysis of complete mitochondrial genome sequences increases phylogenetic resolution of bears (Ursidae), a mammalian family that experienced rapid speciation. BMC Evolutionary Biology 7:198.