*Syst. Biol.* 69(5):848–862, 2020

DOI:10.1093/sysbio/syaa005
Advance Access publication February 3, 2020

# Implementing Large Genomic Single Nucleotide Polymorphism Data Sets in Phylogenetic Network Reconstructions: A Case Study of Particularly Rapid Radiations of Cichlid Fish

MELISA OLAVE[1,2] AND AXEL MEYER[1]*

[1]*Department of Biology, University of Konstanz, 78457 Konstanz, Germany and* [2]*Current: Instituto Argentino de Investigaciones de Zonas Arídas, Consejo Nacional de Investigaciones Cientificas y Técnicas (IADIZA-CONICET), 5500 Mendoza, Argentina*
**Correspondence to be sent to: Department of Biology, University of Konstanz, 78457 Konstanz, Germany; E-mail: axel.meyer@uni-konstanz.de*

*Received 15 July 2019; reviews returned 9 January 2020; accepted 23 January 2020*
*Associate Editor: Claudia Solis-Lemus*

*Abstract*.—The Midas cichlids of the *Amphilophus citrinellus* spp. species complex from Nicaragua (13 species) are an extraordinary example of adaptive and rapid radiation (<24,000 years old). These cichlids are a very challenging group to infer its evolutionary history in phylogenetic analyses, due to the apparent prevalence of incomplete lineage sorting (ILS), as well as past and current gene flow. Assuming solely a vertical transfer of genetic material from an ancestral lineage to new lineages is not appropriate in many cases of genes transferred horizontally in nature. Recently developed methods to infer phylogenetic networks under such circumstances might be able to circumvent these problems. These models accommodate not just ILS, but also gene flow, under the multispecies network coalescent (MSNC) model, processes that are at work in young, hybridizing, and/or rapidly diversifying lineages. There are currently only a few programs available that implement MSNC for estimating phylogenetic networks. Here, we present a novel way to incorporate single nucleotide polymorphism (SNP) data into the currently available PhyloNetworks program. Based on simulations, we demonstrate that SNPs can provide enough power to recover the true phylogenetic network. We also show that it can accurately infer the true network more often than other similar SNP-based programs (PhyloNet and HyDe). Moreover, our approach results in a faster algorithm compared to the original pipeline in PhyloNetworks, without losing power. We also applied our new approach to infer the phylogenetic network of Midas cichlid radiation. We implemented the most comprehensive genomic data set to date (RADseq data set of 679 individuals and >37K SNPs from 19 ingroup lineages) and present estimated phylogenetic networks for this extremely young and fast-evolving radiation of cichlid fish. We demonstrate that the MSNC is more appropriate than the multispecies coalescent alone for the analysis of this rapid radiation. [Genomics; multispecies network coalescent; phylogenetic networks; phylogenomics; RADseq; SNPs.]

Rapidly evolved and young adaptive radiations have long fascinated biologists and serve as model systems to explore mechanisms of lineage diversification (Schluter 2000; Mayr 2001; Coyne and Orr 2004; Elmer et al. 2010a, 2010c). The Midas cichlids of the *Amphilophus citrinellus* spp. from Nicaragua are an extraordinary example of a very recent radiation. At least 13 currently described species have diversified in less than 24,000 years. An ancestral population of the old great Lake Managua and Nicaragua (Fig. 1) independently colonized seven young crater lakes. Six described endemic species live in the oldest crater lake, Apoyo (24,000 years; Kutterolf et al. 2007), and four endemic described species inhabit crater Lake Xiloá (6000 years; Kutterolf et al. 2007). All of them originated through sympatric speciation and likely in the presence of gene flow (Wilson et al. 2000; Barluenga et al. 2004, 2006; Geiger et al. 2010; Kautt et al. 2012, 2016; Elmer et al. 2014). Other Midas cichlid lineages that have diverged genetically and morphologically from the great lake source lineages after colonization (Barluenga et al. 2006; Elmer et al. 2010a; Kautt et al. 2018) live in more recently originated crater lakes (Fig. 1; here called *A.* cf. *citrinellus* as "species candidates"). This cichlid adaptive radiation is a very interesting model system to test alternative phylogenetic models. Different phylogenies based on mitochondrial (Barluenga et al. 2006), microsatellite (Barluenga and Meyer 2010), amplified fragment length polymorphism (AFLP) (Geiger et al. 2010; Kautt et al. 2012), and RADseq data (Kautt et al. 2016) have been published, but none of them includes samples from all geographic lineages based on a genome-wide data set. However, these cichlids are a very challenging group to infer its evolutionary history in phylogenetic analyses, due to the apparent prevalence of ILS, as well as past and current gene flow (Kautt et al. 2016, 2018). It is known that not all of these species and species candidates have evolved complete pre- or postzygotic intrinsic reproductive barriers, because they can be hybridized relatively easily in the laboratory (e.g., Franchini et al. 2014; Machado-Schiaffino et al. 2014).

Phylogenies are basic knowledge and a first requirement for understanding the diversification and dynamics during the evolution of rapid radiations. When using genomic scale data sets for phylogenetic inference, the discord among individual gene trees is clear (Bravo et al. 2019). The gene tree–species tree discordance is a particularly difficult problem in extremely young adaptive radiations. In the last decade, several new methods based on the multispecies coalescent (MSC) were developed that notably improved the phylogenetic estimations (Knowles and Kubatko 2010; Leaché and Rannala 2011; Xu and Yang 2016). The MSC model accommodates incomplete lineage sorting (ILS) as the source of gene tree–species tree discordance, but it ignores other possible sources (e.g., hybridization, introgression, and horizontal gene transfer). Violations of this assumption bias results, hindering the discovery of the "true" phylogeny (Leaché et al. 2013; Solís-Lemus et al. 2016; Wen and Nakhleh 2017;

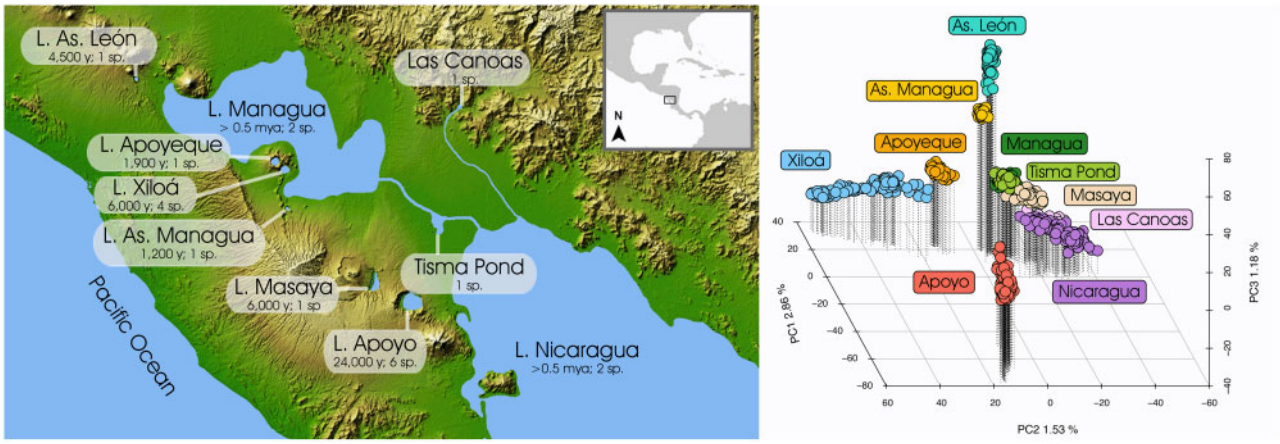


FIGURE 1. (Left) Sampling map showing the distribution of the different described species and species candidates, including the known age of lakes (Lake = L.) and number of species. (Right) Principal component analysis (PCA) of genetic data color-coded by geographic location.

Long and Kubatko 2018; Jiao et al. 2019). Assuming a strict vertical transfer of genes from ancestral lineages to new lineages is not appropriate in many cases of genes transferred horizontally in nature (Mallet 2005, 2007; Abbott et al. 2013; Kagawa and Takimoto 2018; Blair and Ané 2019), and it is too simple to approach many young adaptive radiations (e.g., Kozak et al. 2015; Pease et al. 2016; Meier et al. 2017; Irisarri et al. 2018; Malinsky et al. 2018). The recently proposed multispecies network coalescent (MSNC; Yu et al. 2014; Degnan 2018; Blair and Ané 2019) model can account for the diversity of processes that are at work in young, hybridizing, and/or rapidly diversifying lineages, by accommodating both ILS and gene flow. However, there are currently only a few programs available that implement MSNC (Yu et al. 2014; Solís-Lemus et al. 2017; Zhang et al. 2017; Zhu et al. 2018; Jiao et al. 2019; reviewed in Blair and Ané 2019). Specifically, a very efficient strategy was developed by Solís-Lemus et al. (2017), implemented in the program PhyloNetworks. The program uses a pseudolikelihood approach and allows the estimation of bigger and more complex networks while maintaining accuracy, and it shows clear advantages regarding CPU time (Solís-Lemus and Ané 2016). The method gains its efficiency by focusing on quartets (i.e., unrooted trees of four leaves) and by calculating the observed quartet concordance factor (CF). The CF of a given quartet (or split) is the proportion of genes whose true tree supports that particular quartet (Fig. 2a; Baum 2007). Given the fact that these calculations are exclusively informed by upstream reconstructed gene trees, the type of data that can be used is limited. Specifically, gene tree reconstruction requires long DNA sequences, therefore, single nucleotide polymorphism (SNPs) or AFLP data are not suitable. This same limitation is shared by most of the currently available MSNC programs, with the only exception of the program PhyloNet (Than et al. 2008; Wen et al. 2018), that can also take biallelic markers for phylogenetic network inferences (Zhu and Nakhleh 2018; Zhu et al. 2018). Overcoming this limitation in PhyloNetworks would be very useful for the many short-read data sets that are currently generated by next-generation sequencing (NGS) technologies and increasingly applied to phylogenomic studies (see Leaché and Oaks 2017).

Here, we present the first calculations to obtain CF from a genome-wide SNP data set and apply MSNC to reconstruct the evolutionary history of the rapid radiation of Midas cichlids. Our approach solves the data type limitation, and now allows us to explore large empirical SNP data sets from >37,000 SNPs for 679 individuals of Midas cichlids, representing a total of 19 ingroup lineages + 1 outgroup.

## MATERIALS AND METHODS

### *Quartet CF Calculations from SNP Data Sets*

The current PhyloNetwork pipeline involves two possible ways for obtaining a phylogenetic network: i) it takes upstream quartet CFs estimated by the program BUCKy (Larget et al. 2010) using posterior samples of gene trees as inputs; or ii) it directly takes the upstream reconstructed gene trees and calculates CFs (Solís-Lemus and Ané 2016; Solís-Lemus et al. 2017). The rationale implies that splits between closely related taxa are recovered with higher CFs, but the presence of a hybrid would be reflected in deviation of the two remaining CFs (Solís-Lemus and Ané 2016). For example, consider Figure 2a and assuming a sampling of 10 gene trees in which there are 5 gene trees representing the "topology 1" describing the association of AB_CD, plus a total of 3 gene trees displaying the alternative "topology 2" = AC_BD, and 2 for the case of the "topology 3" = AD_BC; then, the CF for $CF_{AB\_CD}$ = 0.50; $CF_{AC\_BD}$= 0.30; and $CF_{AD\_BC}$ = 0.20. Then, the most common split $CF_{AB\_CD}$ (=0.50) is likely reflecting a natural evolutionary group of organism, while in absence of postdivergence gene flow the expectations are that $CF_{AC\_BD}$ and $CF_{AD\_BC}$ should be approximately equal.

Following the same logical description as above, here we introduce a novel calculation of CFs from SNP data.

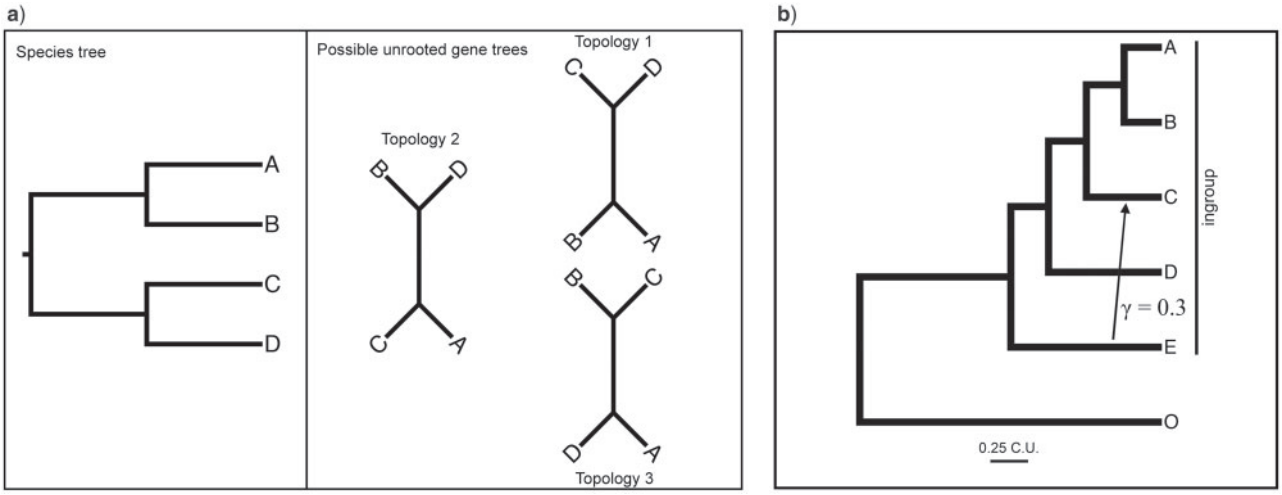


FIGURE 2. a) Example of a four-taxon species tree (left) and the three possible unrooted gene trees for a given species quartet (right). The CFs can be calculated as the proportion of observed gene trees supporting each topology. Thus, assuming a sampling of 10 gene trees, and that there are 5 gene trees recovering the "topology 1," describing the association of AB_CD, a total of 3 gene trees displaying the alternative "topology 2" AC_BD, and 1 following the "topology 3" AD_BC; then, the CF for $CF_{AB\_CD} = 0.50$; $CF_{AC\_BD} = 0.30$; and $CF_{AD\_BC} = 0.20$. b) A general phylogenetic network used to generate simulated data. We fixed $\gamma = 0.3$ and explored three total depths 0.5N, 1N, and 2N for the ingroup only, as well as total depths 1N, 2N, and 4N including the outgroup, respectively. Branch length shown here (in coalescent units [c.u.]) was re-scaled multiplying 0.5 and 2.

Each species quartet is considered, and the proportion of sites supporting each of the three possible alternatives is obtained from a sample of biallelic SNPs. For example, consider the following matrix containing 10 SNPs:

| Locus → | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Sp A | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 |
| Sp B | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 |
| Sp C | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 |
| Sp D | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |

The example above shows the number of times that the split AB_CD appears is equal to five (locus 1–5). Then, the split AC_BD is shown three times (locus 6–8), and AD_BC appears twice (locus 9 and 10). Then, we can obtain their proportion by dividing the total number of SNPs in the matrix (=10 in this case), then: $CF_{AB\_CD} = 0.50$, $CF_{AC\_BD} = 0.30$, and $CF_{AD\_BC} = 0.20$.

We developed the R function "SNPs2CF" (available on www.github.com/melisaolave/SNPs2CF), that takes a concatenated SNP matrix in phylip format, selects all the possible species quartets combinations, and calculates the CFs. The algorithm only considers biallelelic SNPs and ignores singletons, ambiguities, and missing data. Indels are ignored by default, but it is possible to consider them as a fifth state (setting indels.as.fifth.state=TRUE). Given that ambiguities are ignored, data have to be phased (no IUPAC ambiguities accepted). The input matrix can be coded with upper case or lower case letters, including only [A, C, G, T] or [a, c, g, t] or numbers [0, 1, 2, 3]. Note that the number of SNPs that satisfy these conditions are commonly fewer than the full matrix, and that the number of SNPs used is likely to differ among species quartets (see also Supplementary Material available on Dryad at http://dx.doi.org/10.5061/dryad.p5hqbzkkt for further details). Our approach assumes data sets evolving under an infinite sites model (Kimura 1969), neutrality (no selection), and SNP independence (unlinked). Finally, our function can approximate uncertainty by creating pseudoreplicates and obtain a confidence interval. Once the CF table is obtained, it can be loaded into PhyloNetworks and its current pipeline can be used for phylogenetic network estimation. Our SNPs2CF() function can be run in multiple cores in parallel, making it highly efficient with large genomic data sets.

### *Simulation Study: CF from SNP Data and the Power of Our Method*

*SNPs versus gene trees for phylogenetic network reconstruction.*—We conducted a simulation study to contrast the CF calculations based on our new SNP-based pipeline compared to using gene trees. We used the same phylogenetic topology proposed in the simulation study by Wen and Nakhleh (2017; Fig. 2b when considering the ingroup species only: A–E). We used the inheritance probability gamma ($\gamma$) of 0.3 and focused on three total depths for the tree: 0.5N, 1N, and 2N (in coalescent units). A total of 50 independent data sets were simulated using the ms program (Hudson 2002). We simulated 500bp DNA sequences using Seq-Gen program (Rambaut and Grassly 1997), under a Jukes and Cantor model (Jukes and Cantor 1969) with theta = 0.008 (option –s). Then, we compared the CF calculations using batches of 500, 1000, 1500 and 2000 genes/SNPs based on i) the true gene trees (i.e., the ms program output), ii) reconstructed gene

trees, and iii) SNPs. There could be errors associated with gene tree inference, thus we included both, true gene trees and estimated gene trees for comparisons. Note that using exactly the same batches of simulated sequences between SNPs and inferred gene trees is not possible for this simulation study. This is because after discomposing the different quartet combinations, invariable sites and singletons commonly appear within quartets. These sites are ignored by our function SNPs2CF (see Supplementary Material available on Dryad for further explanations). To ensure that at least 2000 SNPs were informing each species quartet, we simulated a larger number of sequences than required (=30,000 loci of length 500 bp), extracted one random SNP per locus and constructed a concatenated SNP matrix. Then, we included a max.SNPs object in our SNPs2CF() function to restrict the number of SNPs per quartet to 500, 1000, 1500, and 2000. This ensures that contrasts between CFs calculated from gene trees and SNPs are comparable.

Gene trees were estimated using RAxML v8 (Stamatakis 2014) using the rapid bootstrap analysis and search of best-scoring maximum likelihood tree (option a) with 100 bootstrap replicates. Finally, the CF was calculated using the function readTrees2CF() in PhyloNetworks' julia package. The time it took RAxML for each gene tree inference plus the time required for CF calculations was recorded, for comparison with the time required by our function SNPs2CF().

The phylogenetic networks were reconstructed using the function snaq!() in PhyloNetworks, by taking the previously calculated CF table and running 10 independent analyses. Then, we used PhyloNet program (Than et al. 2008) to compare the tree-based distances (cmpnets command) between the rooted inferred network with respect to the true network (Fig. 2b). This distance method is designed to handle networks, by considering the topology and the hybridizing edges. Distances equal to 0 represent a perfect reconstruction, while distances >0 represent errors in the inference.

*Program performance comparisons: PhyloNetworks, PhyloNet, and HyDe.*—We included a second set of simulations to compare the performance of PhyloNetworks program using both gene trees and SNPs, as well as we used the same SNP matrix in PhyloNet program (Than et al. 2008) and the Hybrid Detection program (HyDe; Blischak et al. 2018). Because HyDe requires an outgroup, here we simulated data sets under the network including the outgroup (Fig. 2b). Although the ingroup lineages were kept with identical depth as described above (0.5N, 1N, and 2N), for this set of simulations the outgroup pushed the total depth to 1N, 2N, and 4N, respectively. A total of 50 independent data sets were simulated using the same programs (i.e., ms and seq-gen) and settings as described above (see Supplementary Material available on Dryad for specific commands). We constructed matrices including 1000, 5000, and 10,000 SNPs and true gene trees. PhyloNetworks was implemented following the same steps described above, with the only difference that here we did not force the number of SNPs informing a species quartet to be the same than the number of gene trees, then this part of the simulation represents better the cases of empirical studies based on SNP data. For phylogenetic network inference in PhyloNet, we used the Bayesian algorithm (command MCMC_BiMarkers; Zhu et al. 2018). Maximum reticulations were set to one (-mr) and all sites were considered as polymorphic (-op). Bayesian analyses were run during 500,000 steps of the Markov Chain Monte Carlo (-cl), sampling every 500 (-sf), and burnin 200,000 (-bl). All remaining settings were kept as default. Convergence was assessed in all cases by considering effective samples size >200. From all samples, we considered the maximum a posteriori network (MAP) as the best network obtained for comparison with PhyloNetworks result.

HyDe does not reconstruct explicit phylogenetic networks, but it takes SNP matrices to test for hybrid lineages given triplets of species and it estimates the $\gamma$ parameter. Then, we implemented the same SNP matrices in HyDe program for comparison. We specified the outgroup lineage (-o), the number of sites (-s), total species (-t), and individuals (-n). We evaluated all triplets that included the species "C" (Fig. 2b) as a hybrid lineage (six triplets). We considered the proportion of replicates detecting "true positives" as all triplets involving the lineage "E" as one of the parental species, as well as "false parental inference" to significant results in triplets including two other species inferred as parental species of the hybrid lineage "C."

All simulated data sets are available in Dryad doi:10.5061/dryad.p5hqbzkkt.

### *Empirical RADseq Data Set of Midas Cichlids*

Based on an empirical RADseq data set of Midas cichlids (accessions: PRJEB12689 and PRJEB27345; Kautt et al. 2016, 2018), we reconstructed phylogenetic networks under the MSNC. The data set consists of a total of 14 genomic libraries for a total of 679 individuals, representing 12 out of the 13 described species (only *Amphilophus supercilious* is missing), plus five candidate species and one outgroup, *Archocentrus centrachus* (see Supplementary Appendix A available on Dryad for full details of individuals included). Individual-species classifications were taken from Kautt et al. (2016, 2018). Specifically, allopatric species (or species candidates) were classified based on their sampling location. Sympatric individuals were first classified based on their morphology, then contrasted with the genetic data. The raw data consist of an average number of 111.8M reads per sample (see Supplementary Tables S1 and S2 available on Dryad for details). The pipeline STACKS version 1.41 (Catchen et al. 2013) was used to demultiplex and process the genomic sequences. One mismatch in the adapter sequence (–adapter_mm) and a barcode distance of two was used in *process*

*radtags* to allow barcode rescue (–barcode_dist). After removing ambiguous barcodes and ambiguous RAD-tags, the average of total retained reads was 85.88 M per sample. We used bwa mem (Li 2012) to map reads along our reference genome for the Midas cichlid (*A. citrinellus*; Elmer et al. 2014). Then PSTACKS was run with minimum depth coverage of five (-m 5), to extract stacks and align them with the reference genome (mean of genome matches = 0.9605; standard deviation [SD] = 0.01). The catalog of genomic sequences was built in CSTACKS, allowing for two mismatches between sample tags when building the catalog and loci for each individual identified using SSTACKS under default options. From SSTACKS output, we directly ran the POPULATIONS module with low filtering (-m 5), and generated a vcf file. The resulting output was processed in R version 3.2.2 (R Core Team 2016) using packages plyr (Wickham 2011) and pegas (Paradis 2010) to read STACKS output, to manually eliminate SNPs that may represent sequencing errors. Specifically, we removed the five last base pairs in the 3′ end of all loci and loci with high theta values (i.e., $\theta$ that represents the upper 95% quantile), since such high values are suggestive of sequencing and assembly errors. Then, missing data were checked, and 10 individuals were removed due to large missing data proportion (>63% loci missing in an individual). We generated a whitelist and ran POPULATIONS with filters for missing data (-p 4 and -r 0.5) and kept the minimum allowed depth to five (-m 5). The resulting matrix has an average theta of 0.003 (sd = 0.0016; Supplementary Fig. S1 available on Dryad) with a uniform distribution of segregating sites along loci (25% missing data; Supplementary Fig. S2 available on Dryad). We extracted one random SNP per locus. The final matrix consists of 37,180 SNPs for 1358 alleles (i.e., two alleles per individual). Several exploratory analyses were conducted, including a principal component analysis (PCA) shown in Figure 1, using the function dudi.pca() of the ade4 package (Dray and Dufour 2007).

### *Bifurcating Species Tree Reconstruction for the Midas Cichlid Data Set*

We reconstructed a species tree under the MSC, using the SVDquartets program (Chifman and Kubatko 2014) including the full matrix of >37K SNPs and 1358 alleles. Because including all 1358 alleles represents a total of $1.41 \times 10^{11}$ possible quartets combinations, it was not possible to explore the full number of quartets, and, thus, we subsampled 1 billion quartets. To explore the consistency of the subsampled quartets, the analysis was run two independent times, with identical results. To assess uncertainty, we performed 100 bootstrap replicates.

### *Phylogenetic Network Inference from SNP Data of the Midas Cichlid Radiation*

The full number of quartets is too large to be processed by the PhyloNetworks program. Thus, we developed a strategy to subsample the whole space of possible quartets, while ensuring a good representation. We included the n.quartets object in our SNPs2CF() function, that allows to subsample the number of individuals for a species quartet (for further details see manual available on www.github.com/melisaolave/SNPs2CF). Then, we sampled one allele per species quartet by changing the default value "all" to = 1 (i.e., 20 lineages = 4845 quartets). To explore consistency among results, we generated two independent CF tables by randomly subsampling different individuals and running two independent phylogenetic network inferences (both CF tables led to the same network estimation, thus we only show one of them here).

Next, we focused on crater lakes Xiloá and Apoyo since these two crater lakes contain several endemic sympatric species (four in Xiloá and six in Apoyo [five of them included here]) and, thus, current and past gene flow is plausible to occur in these lineages (Kautt et al. 2016). In addition, gene flow between species in Lake Xiloá and Lake Apoyeque was detected previously (Kautt et al. 2018), then we also included samples of *A.* cf. *citrinellus* from this lake. In all other crater lakes not more than one genetic cluster has been identified. *Amphilophus labiatus* and *A. citrinellus* from the two sources of Lakes Managua and Nicaragua were included as outgroups for the estimations of Xiloá and Apoyo networks, respectively. We then reconstructed two phylogenetic networks for Lake Apoyo and Xiloá independently, subsampling 200 alleles per species quartet (n.quartets = 200). Thus, the total number of sampled quartets for seven lineages = $\binom{7}{4} \times 200 = 7000$. We used the SVDquartet tree as starting topology, and then inferred the networks using hmax = 0 (no hybridization), 1, 2, 3, and 4. Following the author's recommendations, we provided the hmax -1 network to the next inference. In each case, we used 20 independents runs with random seeds. In all analyses, the best number of hybridization parameters was selected by plotting the likelihood scores and observing the tipping points in the distributions, that is, a sharp improvement is expected until the number of hybridizing edges reaches the best value and a slower linear improvement thereafter. Finally, to compare the efficiency of our SNPs2CF() function, we recorded the time taken when repeating the CF calculations for the empirical data set when running in a single core and in parallel using 5, 10, 15, and 20 cores and allowing a maximum of 4GB of RAM memory per core on the scientific computing cluster of the University of Konstanz (only for the case of the large network including representation of all lakes).

## Results and Discussion

### *Simulation Results: CF from SNP Data*

For each scenario of different total depths (0.5N, 1N, and 2N), we worked using 50 replicated data sets, and

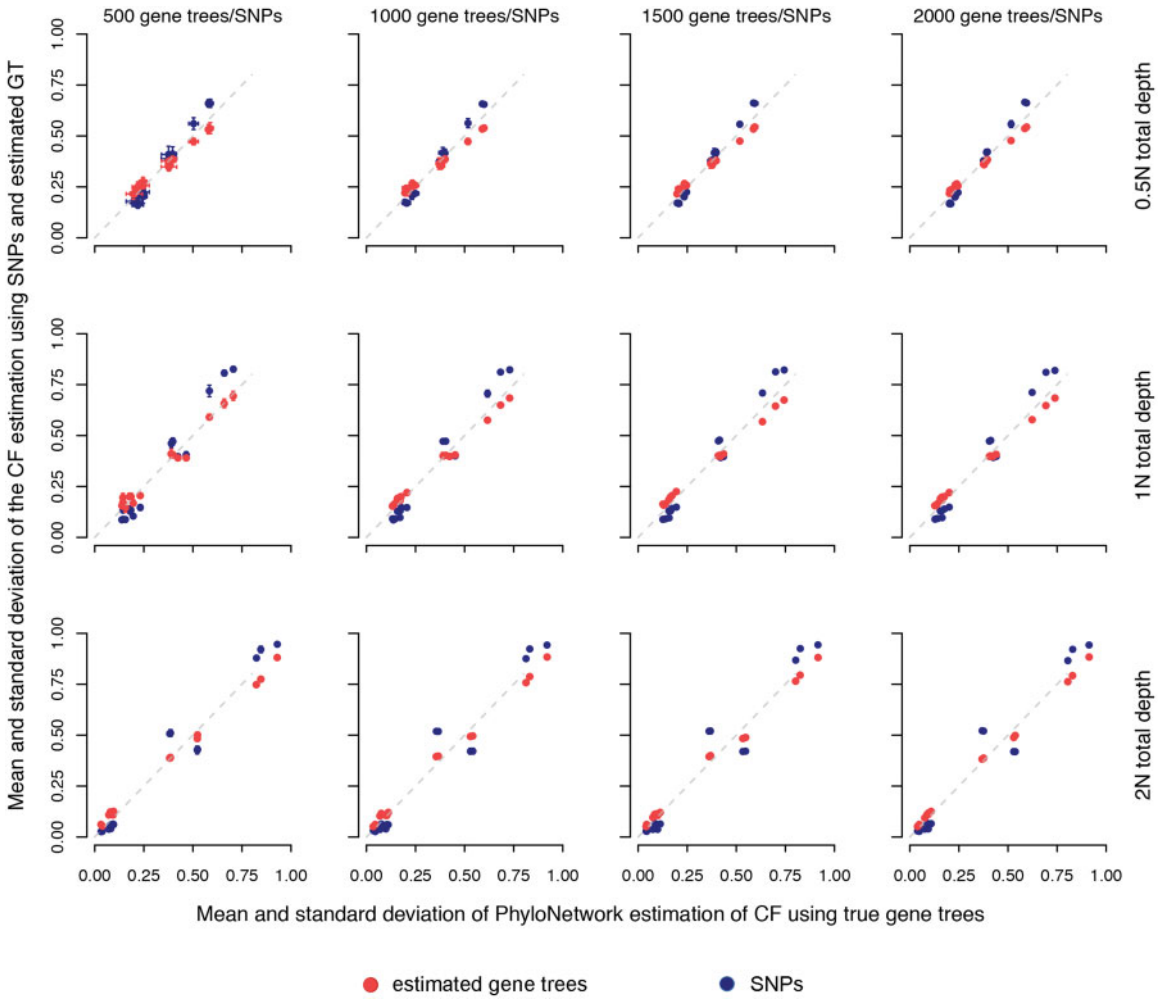


FIGURE 3. Scatter plot of mean (dots) and standard deviation (lines) of CF calculations among 50 simulated data sets based on our new method using SNPs (blue) and when using PhyloNetworks based on estimated gene trees (GT; red) versus the CF calculations using the true gene trees, using simulations based on ingroup species only (Fig. 2b). Each dot represents the mean among 50 replicates of CF calculation for each of the three possible combinations in a given quartet. Perfect match is expected if points are overlapping the 45° dotted gray line drawn as reference. Each row corresponds to one different scenario (0.5N, 1N, and 2N total depth). From the left to the right, calculations were obtained for 500, 1000, 1500, and 2000 genes or SNPs.

obtained the calculations of each species quartet i) using the true gene trees, ii) from estimated gene trees by RAxML program, and iii) from SNP data based on our new calculations in PhyloNetwork package. There is a high correspondence between the calculations based on SNPs and estimated gene trees with respect to the CF obtained based on the true gene trees (Fig. 3 and Supplementary Fig. S3 available on Dryad). Increasing the number of genes or SNPs for each quartet results in reducing dispersion of the points (i.e., SD close to zero). Increasing the total depth (1N–2N) of the true phylogenetic network results in the dispersion of the CF calculations closer to the extremes 0 and 1, while the calculation under the scenario of 0.5N clearly shows an increase in conflict of gene trees/SNPs (Fig. 3).

### *Simulation Results: SNPs Provide Enough Power for Phylogenetic Network Inference*

We reconstructed the inferred networks using SNP-based calculations and the inferred and true gene trees, and obtained the distance between the reconstructed network and the true network among the replicated data sets. Networks obtained under the most challenging scenario explored here (0.5N) and the smallest data set of 500 gene trees/SNPs per quartet results in the highest distances obtained (Supplementary Fig. S4 available on Dryad). However, given enough data (2000 loci), the true network is recovered in >90% of the replicates when using SNPs or the true gene trees (Fig. 4). Power provided by the estimated gene tree approach increases when reducing the level of ILS (1N and 2N; Fig. 4). Since PhyloNetwork was originally written

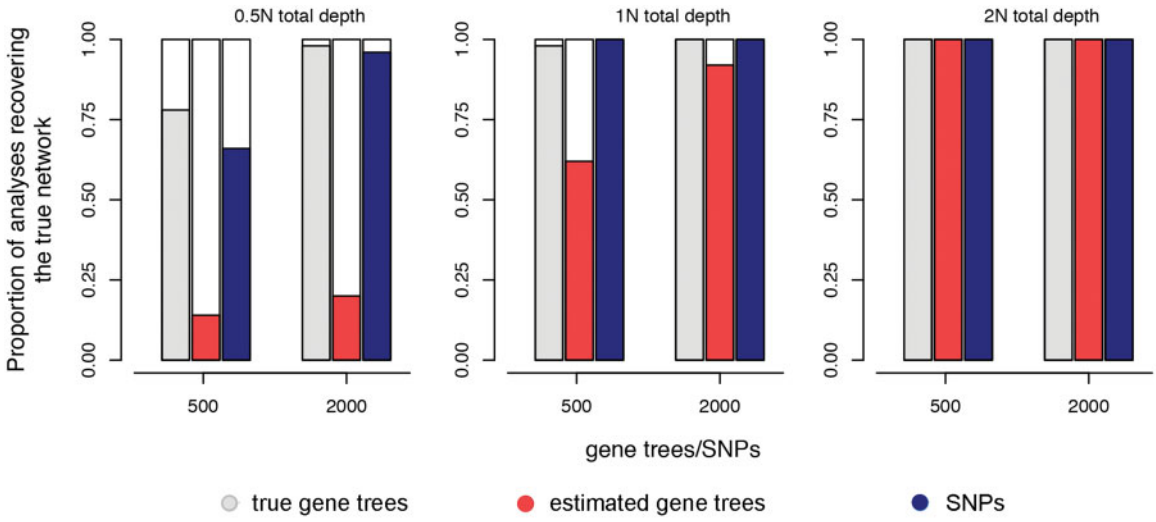


FIGURE 4. Proportion of true networks obtained among the 50 replicated analyses performed for each scenario (0.5N, 1N, and 2N) using the true gene trees (gray), the estimated gene trees (red) and SNPs (blue), using simulations based on ingroup species only (Fig. 2b). Bars reaching 1 represent perfect match in all replicates.

following theory designed to study gene trees, it is unexpected that SNP data are apparently providing a greater power than the inferred gene trees (Fig. 4). The most plausible explanation is that this is associated with errors in upstream gene tree reconstruction, since using the true gene trees increases power (Fig. 4). Low mutation rates might lead to poor gene tree reconstruction (we used a low theta = 0.008 in our simulations; see Materials and Methods section). Also, we used the popular RAxML likelihood-based program for gene tree estimation, which performs fast, and makes a simulation study possible in a realistic time frame (a total of 300,000 gene trees were estimated with RAxML in the simulation study). However, the gene tree inference could be improved by performing a more complete analysis including other programs for phylogenetic inference (e.g., BEAST; Drummond et al. 2012; or MrBayes; Ronquist and Huelsenbeck 2003), as well as conducting more exhaustive analyses and/or providing correct priors. Moreover, using the program BUCKy (Larget et al. 2010) to calculate the CF from gene trees can provide an advantage, since it also considers uncertainty. On the other hand, the true gene trees reflect the maximum power of PhyloNetwork when the gene trees are perfectly inferred and they lead to similar results than using SNPs in our simulations study (Supplementary Fig. S3 available on Dryad and Fig. 4; below we discuss advantages of implementing SNPs for phylogenetic inferences in comparison to gene trees).

It would be interesting in future studies to explore deeper phylogenies and compare the power provided by SNPs versus gene trees in such circumstances, since it is possible that gene trees may provide an advantage over SNPs. For example, most programs for gene tree estimations allow the implementation of complex substitution models while, currently, our approach only considers an equal rate for transitions-transversion substitution rates. In addition, exploring the impact of sampling multiple individuals versus one individual per species needs to be explored. As an advantage, having multiple individuals per species allows to estimate the population size of extant populations (i.e., the lengths of external branches in coalescent units). Intuitively, multiple individuals per species would likely improve performance in PhyloNetworks; however, this feature has not been tested with simulations and deserves further studies (Solís-Lemus and Ané 2016).

### *Comparisons with Other Programs: PhyloNetworks, PhyloNet, and HyDe*

Our second set of simulations (Fig. 5) was implemented to compare the performance of PhyloNetworks using true gene trees and SNPs (here, without forcing the number of SNPs informing a species quartet to be equal to the number of gene trees), as well as for comparison with PhyloNet and HyDe programs. Our results show an excellent accuracy of PhyloNetworks based on the true gene tree data, to recover the true network for all scenarios explored here (Fig. 5). Given the fact that the number of SNPs to inform a species quartet is smaller than the total number of SNPs in the full matrix (Table 1), our approach requires at least 5000 SNPs in the full matrix to confidently infer the true network in the most challenging scenarios with total depth = 1N and 2N (Fig. 5). PhyloNet has a clear tendency in recovering the inverted gene flow direction (from C → E). PhyloNet poor performance found by our simulation study might also be related to poor chain mixing during the heuristic analyses. Running a longer chain (or multiple chains) for the range of scenarios explored here would make our simulation study not feasible in realistic time frame; however, empirical studies could address this by performing

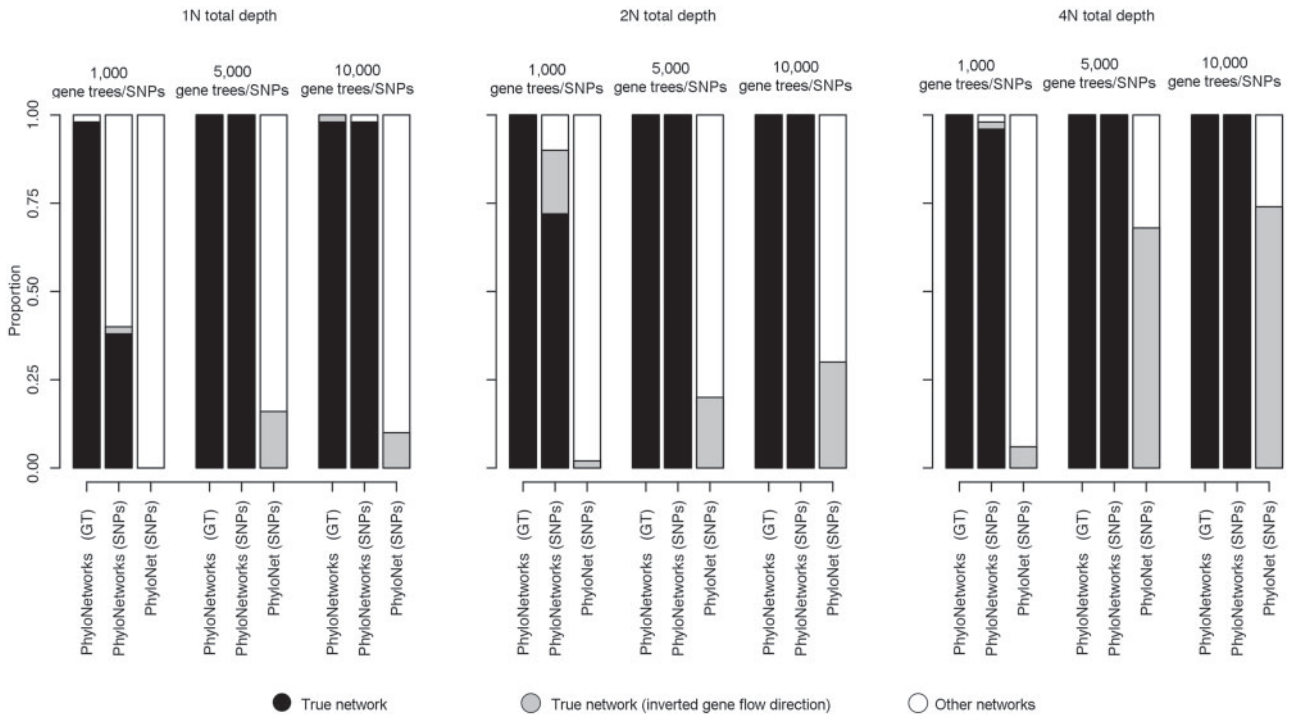


FIGURE 5. Proportion of true networks (black), true network topology with inverted gene flow direction (gray) and other possible networks inferred by PhyloNetworks and PhyloNet, using simulated data sets based on the full network (including outgroup; Fig. 2b). PhyloNetworks was run with the true gene trees (GT) and also using SNPs. PhyloNet was implemented using SNPs and the maximum a posterior (MAP) network was taken to compute the accuracy. A total of 50 replicated analyses were performed for each scenario (1N, 2N, and 4N), including 1000, 5000, and 10,000 gene trees/SNPs.

TABLE 1. Average number and the 95% confidence interval (CI) of SNPs informing the CF calculated per a species quartet in our simulations study based on six species (true network in Fig. 2b)

| | 1N total depth | | 2N total depth | | 4N total depth | |
|---|---|---|---|---|---|---|
| SNPs/scenario | Average SNPs | 95% CI | Average SNPs | 95% CI | Average SNPs | 95% CI |
| 1000 | 89 | 73–106 | 73 | 56–95 | 63.54 | 43–94 |
| 5000 | 457 | 405–520 | 367 | 302–470 | 323.64 | 238–492 |
| 10,000 | 915 | 832–1034 | 725 | 604–944 | 645.1 | 481–992 |

more exhaustive analyses. Here, our simulation study considers a scenario of six species, and we consider the MAP network as the best-inferred network. Based on different simulated conditions, including a simpler network of five lineages, it was already shown in Zhu et al. (2018) that at least 10,000 SNPs are required to accurately recover the true network among posterior samples after burnin (see Fig. 5 in Zhu et al. 2018). On the other hand, only a small proportion of true positives ($P < 0.05$) were detected when applying the same data sets into the HyDe program (Table 2). Our results are not surprising, since it was already shown that at least 50,000 SNPs were required to successfully detect hybridization (Blischak et al. 2018). Moreover, our results also show that there is a relatively high rate of false parental inference (0.08–0.24 among scenarios) with these small data sets (Table 2). Thus, at least under our simulated conditions, it is promising that incorporating smaller SNP data sets in PhyloNetworks can potentially perform better than using both, PhyloNet and HyDe.

### *Advantages of Using SNPs in Phylogenomics: Time Efficiency and Increased Flexibility in Data Types*

There are several advantages of using SNPs in phylogenomics. We have already discussed about the advantages of avoiding potential gene tree estimate errors (first Results and Discussion section above), that seems to be the main explanation of increased power in SNPs (Fig. 4). In addition, there is an improvement in the time consumed to obtain CF from SNPs using our new method, compared to the currently available pipeline that requires upstream gene tree estimation (Table 3). Note that RAxML can be run in parallel (as well as our SNPs2CF() function), but for practical comparison all analyses were run on a single core. To account for the time taken in gene tree reconstruction, we calculated the total time required for the summation of trees (500, 1000, 1500, and 2000). Our algorithm is highly efficient in obtaining the CFs from a SNP matrix in phylip format, taking as little as 2 s with the smallest data set of 500 SNPs, and rising to 21–30 s for processing 2000 SNPs per species quartet (Table 3). When considering the time

TABLE 2. Results of HyDe program applied to evaluate all possible triplets involving the hybrid species "C" among the different scenarios.

| | 1N | | 2N | | 4N | |
|---|---|---|---|---|---|---|
| SNPs/scenario | True positives | False parental inference | True positives | False parental inference | True positives | False parental inference |
| 1000 | 0.10 | 0.18 | 0.20 | 0.10 | 0.16 | 0.10 |
| 5000 | 0.12 | 0.14 | 0.22 | 0.10 | 0.16 | 0.12 |
| 10,000 | 0.12 | 0.10 | 0.20 | 0.24 | 0.18 | 0.08 |

*Notes*: Triplets involving species "E" as one parental were consider as true positives, while cases involving other two parental species were considered as false parental inference.

TABLE 3. Average time consumption (in seconds) and standard deviation (between brackets) for obtaining CF based on SNP data sets, in comparison to implementing the current pipeline of CF requiring upstream gene tree (GT) estimation, for each scenario of total depth (t.d.) = 0.5N, 1N, and 2N.

| Scenario (t.d.) | Methods | Mean (±SD) time taken (in seconds) per gene trees/SNPs | | | |
|---|---|---|---|---|---|
| | | 500 | 1000 | 1500 | 2000 |
| 0.5N | SNPs | 2.06 (±0.20) | 6.01 (±0.2) | 12.70 (±0.45) | 21.50 (±0.81) |
| | GT | 2182.23 (±708.70) | 4436.70 (±1492.37) | 6637.41 (±2226.27) | 8743.46 (±2887.18 |
| 1N | SNPs | 2.53 (±0.20) | 8.08 (±0.32) | 17.27 (±0.83) | 29.70 (±1.61) |
| | GT | 2152.92 (±688.94) | 4422.83 (±1308.29) | 6857.55 (±2050.18) | 9258.93 (±2860.21) |
| 2N | SNPs | 2.59 (±0.12) | 8.84 (±0.35) | (±0.43) 19.02 | 33.22 (±1.09) |
| | GT | 1667.07 (±338.84) | 3235.81 (±393.73) | 4798.66 (±432.07) | 6379.55 (±519.90) |

*Notes*: Here, the gene trees were estimated using RAxML program. All analyses were performed in a single core.

taken for gene tree reconstruction, the time consumption rises up to 2.5 h for 2000 loci.

We also tested the improvement in efficiency when running our SNPs2CF() function in multiple cores in parallel (Supplementary Fig. S5 available on Dryad). This was done using the empirical Midas cichlid data set by subsampling 4845 quartets. The time required to calculate CF based on 37K SNPs is ∼200 min (∼3.3 h), while when using more than five cores the time taken decreased to less than 50 min (Supplementary Fig. S5 available on Dryad). Thus, we have shown that there is a clear advantage in CPU time when focusing on a SNP matrix, in contrast to individual gene tree reconstruction (Table 3). Our method does not require previous treatment and, thus, a matrix in a simple phylip format can directly be used to calculate CF. Efficiency in time consumption allows to incorporate a larger number of markers to represent the genomic polymorphism. As shown here, our approach can easily consider thousands of loci in a realistic time frame (Supplementary Fig. S5 available on Dryad).

As a second point, using SNP data for phylogenetic network reconstruction provides advantages over gene trees given it increases flexibility in the type of data that can be used. Although SNPs can be extracted from DNA sequences obtained by virtually any NGS sequencing technology, only those capable of assembling long sequences are suitable for reconstructing gene trees. Specifically, a suitable amount of mutations presents in each locus is required (the minimum number depends on the number of taxa included, among other factors). This is usually a problem when having short DNA sequences, such as those produced by genotyping by sequencing or RADseq technologies (up to ∼150 bp). Additionally, there are many cases where even long sequences might not be suitable for gene tree reconstruction, particularly if the number of mutations present is too low in recent speciation scenarios (e.g., ultraconserved elements or hybrid enrichment). Few mutations may not provide the power and then gene trees cannot be (or are poorly) reconstructed. Alternatively, SNP data are becoming increasingly common in phylogenetics and even ecology and evolutionary biology, and recently their great potential for phylogenomics was lauded (Leaché and Oaks 2017).

As a third point, phylogenetic tree/network programs commonly assume no within locus recombination. This is an assumption that cannot always be satisfied and, while it is usually desirable to have longer DNA sequences (to capture more mutations), the probability of within locus recombination increases as well. However, recombination issues are discarded when extracting one SNP per locus, given that there is no possibility of intragenic recombination in a single nucleotide.

### *The Phylogenetic Network of the Midas Cichlid Rapid Radiation*

The best phylogenetic network that describes the evolution of the Midas cichlid species flocks has three hybridizing edges (Fig. 6) and a well-resolved topology with a good concordance with the strict

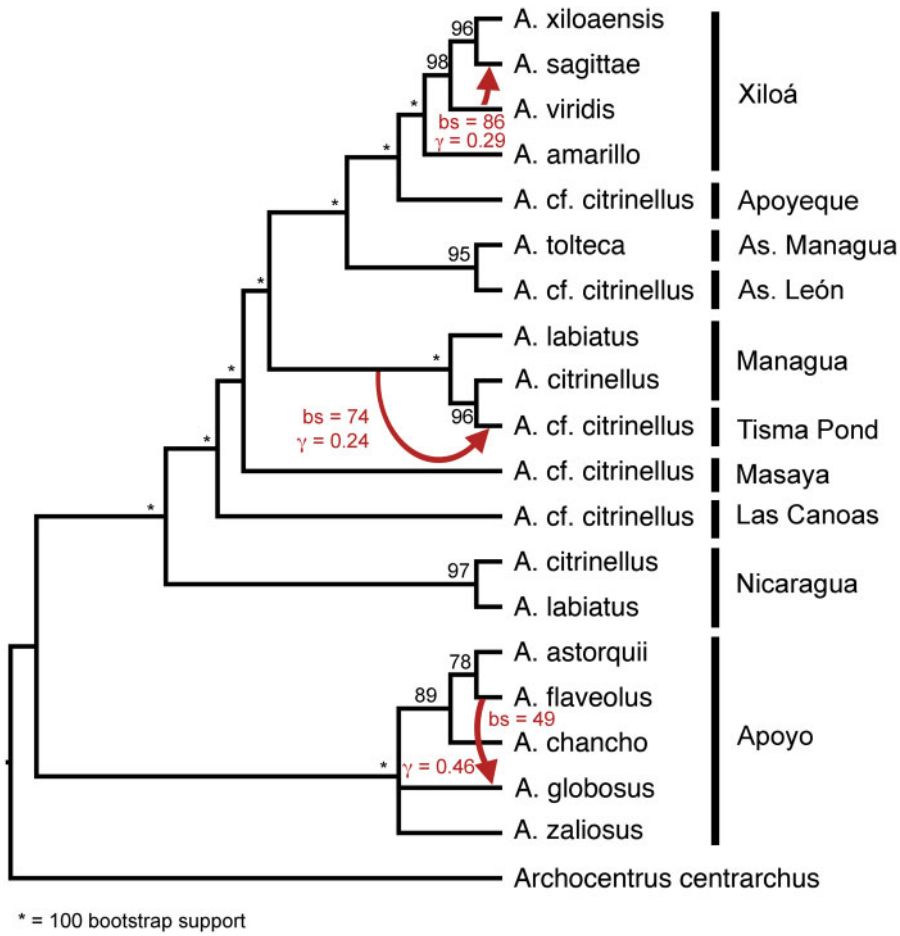


FIGURE 6. Phylogenetic network reconstructed for the Midas cichlid fish rapid radiation, using 37,180 SNPs and subsampling a total of 4845 quartets. The single node with <50 bootstrap support was collapsed. Red arrows represent direction of gene flow and the associated γ estimation representing the proportion of introgressed genome, and its bootstrap support (bs). For simpler visualization, branch lengths were converted into a cladogram.

coalescent-based species tree (Supplementary Fig. S4 available on Dryad), although not identical. Specifically, topological differences are found within the clade comprising species from Lake Apoyo. This network supports hybridization ($\gamma > 0$): i) from *Amphilophus flaveolus* to *Amphilophus globosus* (from Lake Apoyo; $\gamma$ = 0.46 and bootstrap support = 49); ii) bet ween an ancestral lineage of the clade Managua and *A.* cf. *citrinellus* (from Tisma Pond; $\gamma$ = 0.24 and bootstrap support = 74); and iii) from *Amphilophus viridis* and *Amphilophus sagittae* (from Lake Xiloá; $\gamma$ = 0.29 and bootstrap support = 86). However, the phylogenetic networks including only sympatric species from Xiloá and the geographically proximal Lake Apoeyeque (Fig. 7) does not detect hybridization between *A. viridis* and *A. sagittae*, but from *A. sagittae* from Lake Xiloá and *A.* cf. *citrinellus* from Lake Apoyeque ($\gamma$ = 0.11 and bootstrap support = 72). Finally, the phylogenetic network reconstructed for the sympatric species from Lake Apoyo (Fig. 8) also differs from the relationships previously inferred (Fig. 6). This network has two hybridizing edges: i) from *Amphilophus chancho* to an ancestor of the clade ([*A. globosus*, *A. flaveolus*], *Amphilophus astorquii*) ($\gamma$ = 0.0452 and bootstrap support = 82) and ii) from *A. flaveolus* to *A. astorquii* ($\gamma$ = 0.0662 and bootstrap support = 70). Apoyo is the oldest of all crater lakes (24,000 years), is clearly monophyletic and likely the one that was colonized first, thus, its position as sister to the rest of the Midas complex (Fig. 6 and Supplementary Fig. S4 available on Dryad) is reasonable. The Apoyo clade is recovered in our three phylogenetic analyses with different results (i.e., Figs. 6 and 8 and Supplementary Fig. S4 available on Dryad), suggesting that this is likely the most complicated scenario for phylogenetic reconstruction. Uncertainty is also reflected on the CF calculations, with groups of sympatric species having very similar frequency for the three alternative splits (histograms in Fig. 8 and Supplementary Fig. S8 available on Dryad). Thus, the two possible phylogenetic networks either explain the genetic discordance due to high ILS (sister lineages) or through a larger $\gamma$ parameter (hybridizing edges). Previous studies have also reported different topological resolution for species in Apoyo, based on strict coalescent methods (Kautt et al. 2016).

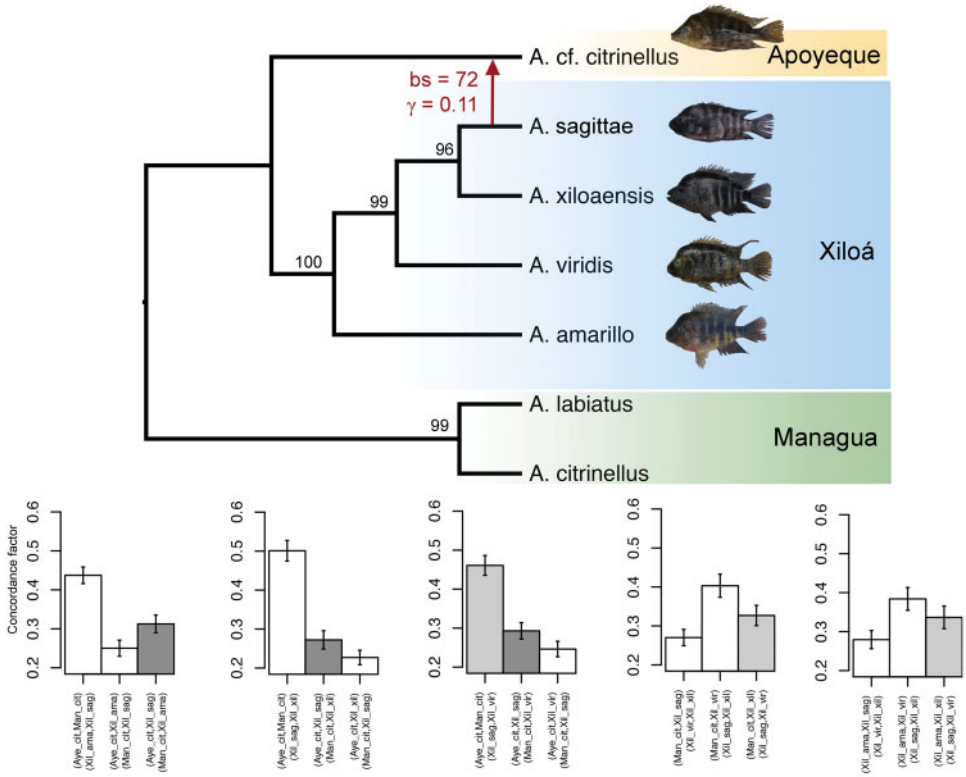


FIGURE 7. (Top) Phylogenetic network reconstructed for the sympatric species in lake Xiloá by subsampling a total of 7000 quartets. Red arrows represent direction of gene flow, and the associated γ estimation representing the proportion of introgressed genome, and their corresponding bootstrap support (bs). For simpler visualization, branch lengths were converted into a cladogram. (Bottom) Average and standard deviation of CF for the three possible splits within each species quartet of interest. Dark gray bars correspond to the comparisons including Aye_cit and Xil_sagittae (i.e., *A. sagittae* and *A.* cf. *citrinellus* [Apoyeque]), and light gray correspond to comparisons including Xil_sag and Xil_vir (i.e., *A. sagittae* and *A. viridis*). See also Supplementary Figure S7 available on Dryad for remaining histograms of CFs. References: Aye_cit: *A.* cf. *citrinellus* (Apoyeque); Xil_ama: *Amphilophus amarillo* (Xiloá); Xil_sag: *A. sagittae* (Xiloá); Xil_xil: *Amphilophus xiloaensis* (Xiloá); Xil_vir: *A. viridis* (Xiloá); Man_cit: *A. citrinellus* (Managua).

The different explanations are probably emerging from a complex scenario involving extremely rapid speciation (<24,000 years) with different species pairs evolving in the presence of gene flow. The two different networks are biologically plausible. However, since the network in Figure 8 was obtained exploring a larger number of quartets (=7000), we believe this second network might be a better representation of the clade's diversification. Furthermore, if γ is closer to 0.5, it involves a range of different and more complex scenarios. For example, it could suggest a case of homoploid hybrid speciation, which is known to be extremely rare in animals (e.g., Schwarz et al. 2005; Mavárez et al. 2006; Lamichhaney et al. 2018). Alternatively, a very large proportion of individuals may have admixed (approximately equal $N_e$ introgressing from a parental lineage to the hybrid lineage). Another possibility involves demographic changes after hybridization and/or positive natural selection acting to increase the frequency of the introgressed genes. All the alternatives are plausible, but the reduction of γ from 0.46 (almost half of the genome) to 0.0662 receives a higher bootstrap support, and it is probably a more parsimonious explanation for their relationships.

We also recovered differences in the hybridizing edges inferred for Xiloá species (Figs. 6 and 7). The level of conflict among Xiloá species is lower (compared to Apoyo species), given that the topological relationships are identical among our three phylogenetic reconstructions (Figs. 6 and 7 and Supplementary Fig. S4 available on Dryad), and with previous MSC inferences (Kautt et al. 2016). However, when increasing the number of subsampled quartets (Fig. 7), we do not find evidence supporting the reticulation of *A. viridis* and *A. sagittae*, even though there is a clear increment for the CFs of the splits including these two species (light gray bars on Fig. 7 and Supplementary Fig. S7 available on Dryad). Instead, this network reconstruction detects allopatric gene flow between *A. sagittae* and *A.* cf. *citrinellus* from Apoyeque. Lakes Xiloá and Apoyeque are geographically close, and evidence of gene flow between them was already reported (Kautt et al. 2018). However, a novel finding from our work is the evidence of gene flow from (specifically) *A. sagittae* to *A.* cf. *citrinellus* in Lake Apoyeque. CF calculations for Xiloa-Apoyeque species are clearly showing a higher frequency favoring the split including *A. sagittae* and *A.* cf. *citrinellus* (dark gray bars on Fig. 7 and Supplementary

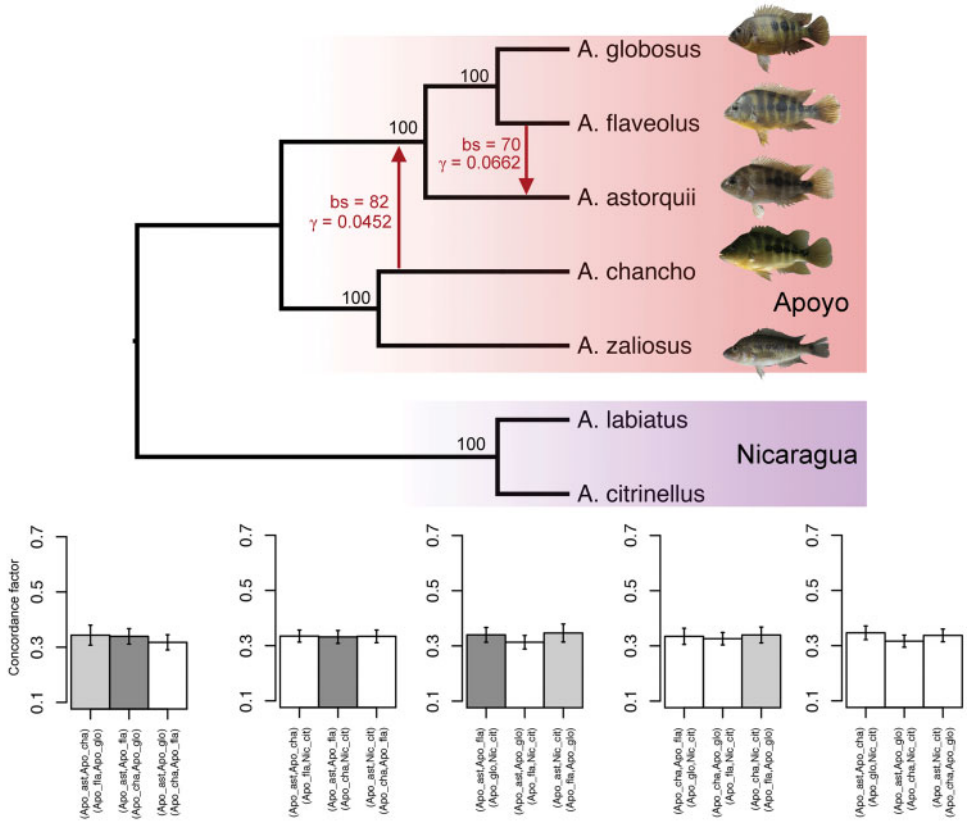


FIGURE 8. (Top) Phylogenetic network reconstructed for the sympatric species in lake Apoyo by subsampling a total of 7000 quartets. Red arrows represent direction of gene flow and the associated γ estimation representing the proportion of introgressed genome, and their corresponding bootstrap support (bs). For simpler visualization, branch lengths were converted into a cladogram. (Bottom) Average and standard deviation of CF for the three possible splits within each species quartet of interest. Dark gray bars correspond to the comparisons including Apo_fla and Apo_ast (i.e., *A. flaveolus* and *A. astorquii*), and light gray correspond to comparisons including Apo_fla and Apo_glo (i.e., *A. flaveolus* and *A. globosus*). See also Supplementary Figure S8 available on Dryad for remaining histograms of CFs. References: Apo_fla: *A. flaveolus* (Apoyo); Apo_glo: *A. globosus* (Apoyo); Apo_ast: *A. astorquii* (Apoyo); Apo_cha: *A. chancho* (Apoyo); Nic_cit: *A. citrinellus* (Nicaragua).

Fig. S7 available on Dryad). Our result implies that about 11% of the genome in Apoyeque lineage has introgressed from *A. sagittae* (Fig. 7). Lake Apoyeque is surrounded by a highly elevated landscape, and connection between the two lakes due to an increment in water level seems unreasonable. However, underground connections between them have been suggested (c.f., Elmer et al. 2010b, 2013), as well as human actions to stock fish for fishing purposes have been suggested as well (Villa 1976). The latter hypothesis seems unlikely, since the access to Apoyeque is difficult and currently there are no human habitations close to it.

The last hybridizing edge detected in our analyses involves fishes from Tisma Pond and Lake Managua (Fig. 6). This is not a surprising result since both bodies of water are connected through the Tipitapa River (Fig. 1) that flows from Lake Managua into Lake Nicaragua. Thus, admixture is possible between them. Surprisingly, gene flow between species of the two great lakes (Nicaragua and Managua) is not detected in our analyses, despite the connection through this same river. This result highlights possible taxonomic issues since there is no evidence of gene flow between lineages currently named as the same species (see also taxonomic implication section).

Traditionally, it is known that ancestral lineages in the two source lakes have colonized other crater lakes (Barluenga et al. 2006; Kautt et al. 2016, 2018). Although our network fits well this story, we note that colonization routes are not that straight forward to see in the network (Fig. 6) and that this network can fit well multiple colonization route scenarios. For example, recovering *A.* cf. *citrinellus* fishes from Crater Lake Asososca León as sister lineage of the described *Amphilophus tolteca* species from Lake As. Managua (Fig. 6 and Supplementary Fig. S6 available on Dryad) is surprising since these isolated crater lakes are on the opposite sides of Lake Managua (Fig. 1), and it seems biologically unlikely that migration between the two could have happened. Historically, it has been thought that both lakes were colonized from ancestral lineages from Lake Managua. A simple explanation for this pattern is that both lakes could have been colonized during similar times from the same ancestral lineages, probably, from the contiguous Lake Managua. The same logic might apply to the inferred sister lineages from Lake Xiloá and Lake Apoyeque. Only proper demographic analyses are

appropriate to tell apart the different colonization routes (ideally including geographic information of ancestral lineages).

Finally, our work demonstrates that the young Midas cichlid radiation has a complex evolutionary history of diversification, not only due to high ILS (Kautt et al. 2016, 2018), but also due to hybridization (Figs. 6–8). Although the strict-coalescent tree reveals conflicting signs reflected among some low-supported nodes (Supplementary Fig. S6 available on Dryad), the phylogenetic network improved support (i.e., most nodes > 89 bootstrap support; Fig. 6).

### *Taxonomic Implications*

Large paraphyly is recovered for the *A.* cf. *citrinellus* lineages (i.e., "species candidates" from lakes: Apoyeque, As. León, Tisma Pond, Masaya and Las Canoas), as well as the *A. citrinellus* and *A. labitatus* species from Lakes Managua and Nicaragua. Our results support the independent evolution of several fish lineages living in allopatry. Although it was previously shown that each lake has its morphometrically unique and different genetic lineages (Elmer et al. 2010; Machado-Schiaffino et al. 2017; Kautt et al. 2018), here we contributed to this issue by suggesting that the different *A. citrinellus* and *A. labiatus* lineages might in fact be different species (Fig. 6 and Supplementary Fig. S6 available on Dryad). Interestingly, our results support previous findings (Machado-Schiaffino et al. 2017; Kautt et al. 2018) of *A. citrinellus* and *A. labiatus* from the two great lakes Managua and Nicaragua as more closely related with their heterospecifics (i.e., *A. citrinellus* and *A. labiatus* are more closely related to other lineages within each lake; Fig. 6 and Supplementary Fig. S6 available on Dryad). Except for the case of fishes from Managua—Tisma Pond and Xiloá—Apoyeque, fishes from the remaining bodies of water seem to have been isolated since their colonization, because we did not find evidence of gene flow among most of the allopatric lakes. Although these are very young lineages (lakes origins range from 24,000 and 2000 years old; reviewed by Elmer et al. 2010a), in such a short time they did not just accumulate strong genetic differences but they are also phenotypically divergent (Elmer et al. 2010a, Elmer et al. 2014; Kautt et al. 2018). The role of genetic drift would appear to be clearly important here. Very few colonizers could have led to a strong founder effect (Kautt et al. 2018), probably leading to fixed differences in a small number of generations. Given the long historical and ongoing discussion about species concepts (e.g., De Queiroz 2007; Zachos 2016) and problems related to species delimitation (Carstens et al. 2013; Olave et al. 2014; Sukumaran and Knowles 2017), the extremely young Midas adaptive radiation is an interesting test case which with one can approach taxonomic questions in light of the speciation continuum based on genome-wide genetic variation. Further specific studies are needed to evaluate whether these several paraphyletic lineages already qualify as different species.

## Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.p5hqbzkkt.

## Funding

This work was supported by the Alexander von Humboldt Foundation and received financial support from the Zukunftskolleg Institute for Advanced Study (IAS) of the University of Konstanz to M.O.; An European Research Council advanced grant [ERC "GenAdap" 293700] and the Deutsche Forschungsgemeinschaft [DFG ME 1725/21-1] to A.M.

## Acknowledgments

We are grateful to Andreas Kautt and Gonzalo Machado-Schiaffino of the Meyer-lab for providing the RADseq libraries used here. We thank the AE Claudia Solís-Lemus for her useful comments on PhyloNetworks usages and CF calculations, as well as her recommendations in an earlier version of this article. We also thank Brian Carstens and three anonymous reviewers for their useful comments and recommendations. This article also received valuable input from Andreas Kautt and Darrin Husley. Analyses were performed on the scientific computing cluster of the University of Konstanz.

## References

Abbott R., Albach D., Ansell S., Arntzen J.W., Baird S.J., Bierne N., Boughman J., Brelsford A., Buerkle C.A., Buggs R., Butlin R.K., Dieckmann U., Eroukhmanoff F., Grill A., Cahan S.H., Hermansen J.S., Hewitt G., Hudson A.G., Jiggins C., Jones J., Keller B., Marczewski T., Mallet J., Martinez-Rodriguez P., Möst M., Mullen S., Nichols R., Nolte A.W., Parisod C., Pfennig K., Rice A.M., Ritchie M.G., Seifert B., Smadja C.M., Stelkens R., Szymura J.M., Väinölä R., Wolf J.B., Zinner D. 2013. Hybridization and speciation. J. Evol. Biol. 26(2):229–246.

Barluenga M., Meyer A. 2004. The Midas cichlid species complex: incipient sympatric speciation in Nicaraguan cichlid fishes? Mol. Ecol. 13(7):2061–2076.

Barluenga M., Meyer A. 2010. Phylogeography, colonization and population history of the Midas cichlid species complex (*Amphilophus* spp.) in the Nicaraguan crater lakes. BMC Evol. Biol. 10(1):326.

Barluenga M., Stölting K.N., Salzburger W., Muschick M., Meyer A. 2006. Sympatric speciation in Nicaraguan crater lake cichlid fish. Nature 439(7077):719.

Baum D.A. 2007. Concordance trees, concordance factors, and the exploration of reticulate genealogy. Taxon 56:417–426.

Blair C., Ané C. 2019. Phylogenetic trees and networks can serve as powerful and complementary approaches for analysis of genomic data. Syst. Biol. 69(3):593–601.

Blischak P.D., Chifman J., Wolfe A.D., Kubatko L.S. 2018. HyDe: a Python package for genome-scale hybridization detection. Syst. Biol. 67(5):821–829.

Bravo G.A., Antonelli A., Bacon C.D., Bartoszek K., Blom M.P., Huynh S., Edwards S. 2019. Embracing heterogeneity: coalescing the Tree of Life and the future of phylogenomics. PeerJ 7:e6399.

Carstens B.C., Pelletier T.A., Reid N.M., Satler J.D. 2013. How to fail at species delimitation. Mol. Ecol. 22(17):4369–4383.
Catchen J., Hohenlohe P.A., Bassham S., Amores A., Cresko W.A. 2013. Stacks: an analysis tool set for population genomics. Mol. Ecol. 22(11):3124–3140.
Chifman J., Kubatko L. 2014. Quartet inference from SNP data under the coalescent model. Bioinformatics 30(23):3317–3324.
Coyne J.A., Orr H.A. 2004. Speciation. Sunderland, MA.
Degnan J.H. 2018. Modeling hybridization under the network multispecies coalescent. Syst. Biol. 67(5):786–799.
De Queiroz, K. 2007. Species concepts and species delimitation. Syst. Biol. 56(6):879–886.
Dray S., Dufour A.B. 2007. The ade4 package: implementing the duality diagram for ecologists. J. Stat. Softw. 22(4):1–20.
Drummond A.J., Suchard M.A., Xie D., Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29(8):1969–1973.
Elmer K.R., Kusche H., Lehtonen T.K., Meyer A. 2010a. Local variation and parallel evolution: morphological and genetic diversity across a species complex of neotropical crater lake cichlid fishes. Philos. Trans. R. Soc. B 365(1547):1763–1782.
Elmer K.R., Fan S., Gunter H.M., Jones J.C., Boekhoff S., Kuraku S., Meyer A. 2010b. Rapid evolution and selection inferred from the transcriptomes of sympatric crater lake cichlid fishes. Mol. Ecol. 19:197–211.
Elmer K.R., Lehtonen T.K., Kautt A.F., Harrod C., Meyer A. 2010c. Rapid sympatric ecological differentiation of crater lake cichlid fishes within historic times. BMC Biol. 8(1):60.
Elmer K.R., Lehtonen T.K., Fan S., Meyer A. 2013. Crater lake colonization by Neotropical cichlid fishes. Evolution 67(1):281–288.
Elmer K.R., Fan S., Kusche H., Spreitzer M.L., Kautt A.F., Franchini P., Meyer A. 2014. Parallel evolution of Nicaraguan crater lake cichlid fishes via non-parallel routes. Nat. Commun. 5:5168.
Franchini P., Fruciano C., Spreitzer M.L., Jones J.C., Elmer K.R., Henning F., Meyer A. 2014. Genomic architecture of ecologically divergent body shape in a pair of sympatric crater lake cichlid fishes. Mol. Ecol. 23(7):1828–1845.
Geiger M.F., McCrary J.K., Schliewen U.K. 2010. Not a simple case—a first comprehensive phylogenetic hypothesis for the Midas cichlid complex in Nicaragua (Teleostei: Cichlidae: Amphilophus). Mol. Phylogenet. Evol. 56(3):1011–1024.
Irisarri I., Singh P., Koblmüller S., Torres-Dowdall J., Henning F., Franchini P., Sturmbauer C. 2018. Phylogenomics uncovers early hybridization and adaptive loci shaping the radiation of Lake Tanganyika cichlid fishes. Nat. Commun. 9(1):3159.
Jiao X., Flouris T., Rannala B., Yang Z. 2019. The impact of cross-species gene flow on species tree estimation. Syst. Biol. 69(5): 830–847.
Jukes T.H., Cantor C.R. 1969. Evolution of protein molecules. In: Munro H.N., editor. Mammalian protein metabolism. New York: Academic Press. p. 21–132.
Kagawa K., Takimoto G. 2018. Hybridization can promote adaptive radiation by means of transgressive segregation. Ecol. Lett. 21(2):264–274.
Kautt A.F., Elmer K.R., Meyer A. 2012. Genomic signatures of divergent selection and speciation patterns in a 'natural experiment', the young parallel radiations of Nicaraguan crater lake cichlid fishes. Mol. Ecol. 21(19):4770–4786.
Kautt A.F., Machado-Schiaffino G., Meyer A. 2016. Multispecies outcomes of sympatric speciation after admixture with the source population in two radiations of Nicaraguan crater lake cichlids. PLoS Genet. 12(6):e1006157.
Kautt A.F., Machado-Schiaffino G., Meyer A. 2018. Lessons from a natural experiment: Allopatric morphological divergence and sympatric diversification in the Midas cichlid species complex are largely influenced by ecology in a deterministic way. Evol. Lett. 2(4):323–340.
Kimura M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics 61(4):893.
Knowles L.L., Kubatko L.S., editors. 2010. Estimating species trees: an introduction to concepts and models. In: Estimating species trees: practical and theoretical aspects. USA: Wiley-Blackwell. p. 1–14.
Kozak K.M., Wahlberg N., Neild A.F., Dasmahapatra K.K., Mallet J., Jiggins C.D. 2015. Multilocus species trees show the recent adaptive radiation of the mimetic Heliconius butterflies. Syst. Biol. 64(3):505–524.
Kutterolf S., Freundt A., Perez W., Wehrmann H., Schmincke H.U. 2007. Late Pleistocene to Holocene temporal succession and magnitudes of highly-explosive volcanic eruptions in west-central Nicaragua. J. Volcanol. Geotherm. Res. 163(1–4):55–82.
Lamichhaney S., Han F., Webster M.T., Andersson L., Grant B.R., Grant P.R. 2018. Rapid hybrid speciation in Darwin's finches. Science 359(6372):224–228.
Leaché A.D., Rannala B. 2011. The accuracy of species tree estimation under simulation: a comparison of methods. Syst. Biol. 60(2):126–137.
Leaché A.D., Harris R.B., Rannala B., Yang Z. 2013. The influence of gene flow on species tree estimation: a simulation study. Syst. Biol. 63(1):17–30.
Leaché A.D., Oaks J.R. 2017. The utility of single nucleotide polymorphism (SNP) data in phylogenetics. Annu. Rev. Ecol. Evol. Syst. 48:69–84.
Larget B.R., Kotha S.K., Dewey C.N., Ané C. 2010. BUCKy: gene tree/species tree reconciliation with Bayesian concordance analysis. Bioinformatics 26(22):2910–2911.
Li H. 2012 Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. Bioinformatics 28:1838–1844.
Long C., Kubatko L. 2018. The effect of gene flow on coalescent-based species-tree inference. Generations 1:2N.
Machado-Schiaffino G., Kautt A.F., Torres-Dowdall J., Baumgarten L., Henning F., Meyer A. 2017. Incipient speciation driven by hypertrophied lips in Midas cichlid fishes? Mol. Ecol. 26(8):2348–2362.
Machado-Schiaffino G., Henning F., Meyer A. 2014. Species-specific differences in adaptive phenotypic plasticity in an ecologically relevant trophic trait: hypertrophic lips in Midas cichlid fishes. Evolution 68(7):2086–2091.
Malinsky M., Svardal H., Tyers A.M., Miska E.A., Genner M.J., Turner G.F., Durbin R. 2018. Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. Nat. Ecol. Evol. 2(12):1940.
Mallet J. 2005. Hybridization as an invasion of the genome. Trends Ecol. Evol. 20(5):229–237.
Mallet J. 2007. Hybrid speciation. Nature 446(7133):279.
Mavárez J., Salazar C.A., Bermingham E., Salcedo C., Jiggins C.D., Linares M. 2006. Speciation by hybridization in Heliconius butterflies. Nature 441(7095):868.
Mayr E. 2001. What evolution is. (Science Masters Series). USA: Basic Books.
Meier J.I., Marques D.A., Mwaiko S., Wagner C.E., Excoffier L., Seehausen O. 2017. Ancient hybridization fuels rapid cichlid fish adaptive radiations. Nat. Commun. 8:14363.
Olave M., Sola E., Knowles L.L. 2014. Upstream analyses create problems with DNA-based species delimitation. Syst. Biol. 63(2):263–271.
Paradis E. 2010. pegas: an R package for population genetics with an integrated-modular approach. Bioinformatics 26:419–420.
Pease J.B., Haak D.C., Hahn M.W., Moyle L.C. 2016. Phylogenomics reveals three sources of adaptive variation during a rapid radiation. PLoS Biol. 14(2):e1002379.
Rambaut A., Grass N.C. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Bioinformatics 13(3):235–238.
Ronquist F., Huelsenbeck J.P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19(12):1572–1574.
Schluter D. 2000. The ecology of adaptive radiation. Oxford: OUP.
Schwarz D., Matta B.M., Shakir-Botteri N.L., McPheron B.A. 2005. Host shift to an invasive plant triggers rapid animal hybrid speciation. Nature 436(7050):546.
Solís-Lemus C., Yang M., Ané C. 2016. Inconsistency of species tree methods under gene flow. Syst. Biol. 65(5):843–851.

Solís-Lemus C., Ané C. 2016. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. PLoS Genet. 12(3):e1005896.
Solís-Lemus C., Bastide P., Ané C. 2017. PhyloNetworks: a package for phylogenetic networks. Mol. Biol. Evol. 34(12):3292–3298.
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30(9):1312–1313.
Sukumaran J., Knowles L.L. 2017. Multispecies coalescent delimits structure, not species. Proc. Natl. Acad. Sci. USA 114(7):1607–1612.
Than C., Ruths D., Nakhleh L. 2008. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. BMC Bioinformatics 9(1):322.
Villa J. 1976. Ichthyology of the lakes of Nicaragua: historical perspective.
Wen D., Nakhleh L. 2017. Coestimating reticulate phylogenies and gene trees from multilocus sequence data. Syst. Biol. 67(3):439–457.
Wen D., Yu Y., Zhu J., Nakhleh L. 2018. Inferring phylogenetic networks using PhyloNet. Syst. Biol. 67(4):735–740.
Wickham H. 2011. The split-apply-combine strategy for data analysis. J. Stat. Softw. 40(1):1–29. URL http://www.jstatsoft.org/v40/i01/.
Wilson A.B., Noack-Kunnmann K., Meyer A. 2000. Incipient speciation in sympatric Nicaraguan crater lake cichlid fishes: sexual selection versus ecological diversification. Proc. R. Soc. Lond. Ser. B 267(1458):2133–2141.
Xu B., Yang Z. 2016. Challenges in species tree estimation under the multispecies coalescent model. Genetics 204(4):1353–1368.
Yu Y., Dong J., Liu K.J., Nakhleh L. 2014. Maximum likelihood inference of reticulate evolutionary histories. Proc. Natl. Acad. Sci. USA 111(46):16448–16453.
Yu Y., Nakhleh L. 2015. A maximum pseudo-likelihood approach for phylogenetic networks. BMC Genomics 16(10):S10.
Zachos F.E. 2016. Species concepts in biology: historical development, theoretical foundations and practical relevance. Switzerland: Springer.
Zhang C., Ogilvie H.A., Drummond A.J., Stadler T. 2017. Bayesian inference of species networks from multilocus sequence data. Mol. Biol. Evol. 35(2):504–517.
Zhu J., Nakhleh L. 2018. Inference of species phylogenies from bi-allelic markers using pseudo-likelihood. Bioinformatics 34(13):i376-i385.
Zhu J., Wen D., Yu Y., Meudt H.M., Nakhleh L. 2018. Bayesian inference of phylogenetic networks from bi-allelic genetic markers. PLoS Comput. Biol. 14(1):e1005932.