2019

Phylogenetic trees and networks can serve as powerful and complementary approaches for analysis of genomic data

Christopher Blair
CUNY New York City College of Technology

Cécile Ané
University of Wisconsin-Madison

Point-of-view/Opinion

Phylogenetic trees and networks can serve as powerful and complementary approaches for analysis of genomic data

Christopher Blair[1,2], Cécile Ané[3]

[1]Department of Biological Sciences, New York City College of Technology, The City University of New York, 285 Jay Street, Brooklyn, NY 11201

[2]Biology PhD Program, CUNY Graduate Center, 365 5th Ave., New York, NY 10016

[3]Departments of Botany and of Statistics, University of Wisconsin – Madison, 1300 University Ave, Madison WI 53706

Correspondence: CBlair@citytech.cuny.edu; 718-260-5342

Abstract

Genomic data have had a profound impact on nearly every biological discipline. In systematics and phylogenetics, the thousands of loci that are now being sequenced can be analyzed under the multispecies coalescent model (MSC) to explicitly account for gene tree discordance due to incomplete lineage sorting (ILS). However, the MSC assumes no gene flow post divergence, calling for additional methods that can accommodate this limitation. Explicit phylogenetic network methods have emerged, which can simultaneously account for ILS and gene flow by representing evolutionary history as a directed acyclic graph. In this point-of-view we highlight some of the strengths and limitations of phylogenetic networks and argue that tree-based inference should not be blindly abandoned in favor of networks simply because they represent more parameter rich models. Attention should be given to model selection of reticulation complexity, and the most robust conclusions regarding evolutionary history are likely obtained when combining tree- and network-based inference.

Key words: coalescence, phylogenomics, reticulations, gene flow, species tree, species network

Gene tree discordance

The acquisition of genomic data from non-model organisms for evolutionary inference continues to fundamentally change the theory and practice of systematics. For years, empiricists were limited to data sets consisting of one to a few loci to reconstruct phylogenetic relationships. With the advent of next-generation sequencing (NGS) technologies in evolutionary biology, empiricists were bombarded with an enormous quantity of data. These genomic data sets, coupled with the development of new models such as the multispecies coalescent (MSC; Hudson 1983; Tajima 1983; Rannala and Yang 2003), paved the way for a new era in molecular systematics (Edwards 2009; Edwards et al. 2016; Bravo et al. 2019). The MSC was a particularly valuable addition to the systematist's toolbox, because it was able to explicitly accommodate gene tree/species tree discordance due to incomplete lineage sorting (ILS). Theory suggests that the frequency of ILS in a data set is directly related to effective population sizes and the number of generations separating speciation events (Maddison 1997; Degnan and Rosenberg 2009). Thus, ILS is likely a prominent cause of gene tree discordance in studies of adaptive radiations, old or recent. ILS is even more problematic at the population/phylogeographic level (i.e. when interested in estimating a population tree).

A multitude of theoretical and empirical studies have suggested that concatenating loci (i.e. the supermatrix approach) can lead to incorrect estimates of the species / population tree in the presence of ILS, particularly in shallow phylogenies at or below the species level where ILS is very strong (e.g. Kubatko and Degnan 2007; Mirarab et al. 2014; Roch and Steel 2015). There are now several different approaches to estimating species trees that are statistically consistent under the MSC (as reviewed in Edwards 2016). These include fully probabilistic Bayesian methods such as BPP (Flouri et al. 2018) and StarBEAST2 (Ogilvie et al. 2017), summary statistic methods such as ASTRAL (Zhang et al. 2018b), MP-EST (Liu et al. 2010), NJst (Liu and Yu 2011), and methods based on site pattern frequencies and algebraic statistics (SVDquartets; Chifman and Kubatko 2014; Wascher and Kubatko 2019). A complete overview

of species tree methods is beyond the scope of this paper, and each has its strengths and weaknesses. For example, fully parametric methods are able to estimate additional parameters of interest such as divergence times and effective population sizes. However, these methods are far more computationally demanding compared to summary methods, which estimate a species tree from gene tree topologies. Unfortunately, recent theory shows that summary approaches (and fully partitioned concatenation) may in fact be inconsistent estimators of species trees due to a finite number of sites per locus (Roch et al. 2019).

Impact of gene flow on species tree inference

As alluded to above, the MSC has helped usher in a new era of molecular systematics. However, one major assumption of the MSC is the absence of gene flow post divergence. As several studies show, gene flow can be ubiquitous during the process of speciation (Mallet 2005, 2007, 2008; Cui et al. 2013; Jónsson et al. 2014). This means a fundamental violation of the MSC. The question is, then, is this violation significant enough to result in inaccurate phylogenetic inference and/or species delimitation? Studies have shown that the MSC implemented in *BEAST and other commonly used software may estimate inaccurate species trees when gene flow occurs between non-sister taxa (Eckert and Carstens 2008; Chung and Ané 2011; Leaché et al. 2014). Gene flow and introgression can also negatively impact the estimation of divergence times and population sizes when using fully parametric methods (Leaché et al. 2014), with gene flow resulting in an overestimation of ILS through overestimated population sizes and underestimated speciation times. However, if significant gene flow is present in a data set, it is far more likely to occur in sister taxa due to genetic and geographic constraints, which might not bias the analysis, at least with respect to the species tree topology. In fact, gene flow between sister species may have a positive impact on species tree reconstruction because it homogenizes alleles across species boundaries (Leaché et al. 2014), so long as speciation is not too fast. However, under rapid speciation (e.g. strong ILS), gene

flow between sister species negatively influences species trees inferred through concatenation and summary-based methods, due to extra anomalous gene trees caused by gene flow.

Without gene flow, the topology of the species tree on three taxa is always the most frequent among rooted gene trees: there are no three-taxon anomalous rooted gene trees. With gene flow, however, the three-taxon topology matching the species trees might be less frequent than some other discordant topology (called “anomalous gene tree”), when speciation is quickly followed by gene flow and by rapid speciation again, shown on Fig. 1. In this example, the tree-like history obtained by removing the horizontal gene flow arrow is $S_1S_2|O$, where the two species S_1 and S_2 that diverged after gene flow are sister to each other. Alternatively, if we keep the horizontal gene flow arrow but remove its partner edge (i.e. remove the vertical inheritance and keep gene flow inheritance only), we still get the same displayed tree: $S_1S_2|O$. Therefore, in this scenario it is reasonable to consider $S_1S_2|O$ as being the true species tree that tree-based methods should recover. Without gene flow, S_1 and S_2 are sister in gene trees more often than either is sister to an outgroup O: the matching gene tree $S_1S_2|O$ is more frequent than either discordant gene tree $S_1O|S_2$ and $S_2O|S_1$. With gene flow (as in Fig. 1), alleles sampled from each of S_1 and S_2 may not have time to coalesce in the ancestral admixed species, in which case they are likely to be non-sister in their gene tree because one of S_1 or S_2 might trace its ancestry through gene flow (Fig. 1). If speciation and gene flow events occur sufficiently rapidly, then both discordant gene trees with S_1 and S_2 non-sister are anomalous. That is, both are more frequent than the matching gene tree $S_1S_2|O$, contrary to the expectation under ILS only. In this scenario, ASTRAL, NJst, and MP-EST have been shown to be inconsistent (Solís-Lemus et al, 2016, Long and Kubatko 2018). This negative influence of gene flow does not seem to affect the SVDquartets method, which shows moderate accuracy with large data set sizes (Long and Kubatko 2018). It is hard to make broad generalizations regarding the overall impact of gene flow on species tree inference, as additional nuances of the data such as marker choice,

the strength of gene flow, and proportion of sampled individuals exchanging genes may exert different outcomes.

Accommodating gene flow

Due to the apparent violation of the no-gene-flow assumption of the MSC in an increasing number of studies, a suite of recent methods has attempted to address this concern from an analytical perspective. For example, the AIM addition to StarBEAST2 (Mueller et al. 2018) and IMa3 (Hey et al. 2018) can explicitly accommodate gene flow when estimating a species tree. In these methods, gene flow is modelled by a migration rate for each pair of coexisting species. Similarly, recent additions to BPP (Flouri et al. 2018) utilize the multispecies coalescent with introgression (MSCi) model to allow for introgression when estimating evolutionary parameters such as divergence times and effective population sizes. A different class of methods aims to infer explicit phylogenetic networks (Solís-Lemus et al. 2017; Hejase et al. 2018; Wen et al. 2018; Zhang et al. 2018a). In these models, the idea of a strictly bifurcating tree is abandoned in favor of a network (a directed acyclic graph) that allows reticulation edges. Each reticulation edge summarizes gene flow that might have occurred over a period of time into a single instantaneous event with an associated genomic weight corresponding to the proportion of alleles in the recipient population that were inherited from the donor population as a result of the entire period of gene flow. These phylogenetic networks offer an elegant simplifying framework for modelling (supposedly) discrete populations, and for summarizing a number of biological processes by which alleles transfer between one population and another. The techniques based on these network models can be used to examine the history of introgression across the entire phylogeny. An attractive feature of these methods is the ability to simultaneously account for both ILS and gene flow when reconstructing the evolutionary history of a clade. For reviews, see Elworth et al. (2018) and Degnan (2018). We shortly describe the most commonly used methods below.

A variety of phylogenetic network methods are implemented in the program PhyloNet (Than et al. 2008; Wen et al. 2018). An attractive feature of PhyloNet relates to flexibility, as there are different analytical techniques depending on the characteristics of the data. For example, there are parsimony, likelihood, pseudolikelihood, and Bayesian methods that use gene trees as input, Bayesian methods that use multilocus sequences, and parametric methods that use bi-allelic markers. The SpeciesNetwork package for BEAST2 (Zhang et al 2018a) uses multilocus sequences in a full Bayesian framework, a birth-hybridization process prior, and can accommodate rate variation across genes and across lineages. Unfortunately, SpeciesNetwork and the majority of network methods available in PhyloNet are very computationally demanding, because the network coalescent needs to track an explosive number of coalescent histories, as the number *or* depth of reticulations increases (Elworth et al. 2018). This computational burden limits the utility of full-likelihood methods to data sets consisting of few loci, taxa or hybrid edges (Hejase and Liu 2016).

One network method that is gaining in popularity is SNaQ (Solís-Lemus and Ané 2016) implemented in the PhyloNetworks package (Solís-Lemus et al. 2017). It uses a pseudolikelihood on four taxa to significantly increase computational tractability. SNaQ allows for rate variation across lineages and across genes, and does not require that gene trees be rooted with an outgroup. This flexibility is at the expense of discarding potentially useful data from branch lengths. Unlike other methods, it estimates a semi-rooted network, which can later be rooted with an outgroup.

Data sub-sampling to deal with computational complexity

All methods that explicitly accommodate both gene flow and ILS are more computationally demanding than those that model ILS alone (Hejase and Liu 2016). Given the limitations of computational capabilities, researchers may be left with substantially pruning their data in order to make their network analysis feasible. Reducing the number of species is the most effective

way to reduce the computation time. For some methods (e.g. fully Bayesian methods), reducing the number of genes may also be necessary. Pruning is even more necessary with methods that estimate a network directly from the sequence data. Thus, when the backbone tree topology is of primary interest, is it a better approach to analyze more data using tree-based methods, or subsample to estimate a network on a smaller taxon set? One recent study found that network methods provided more accurate species trees even with relatively few genes (Solís-Lemus et al. 2016).

In contrast with summary-based species tree methods like ASTRAL, it is presently uncertain how gene tree error influences network estimation and whether or not (or how) the data should be subsampled based on some criterion (e.g. locus informativeness, taxon coverage) prior to analysis. Some early analyses suggest that a moderate number of genes is required to accurately recover the major tree in the network, but a relatively large number of genes is required to determine the appropriate number and location of reticulations (Solís-Lemus and Ané 2016). Elucidating optimal sampling regimes for phylogenetic networks is an active area of study that will likely attract attention from both theoretical and empirical systematists. However, as methods like SNaQ use gene trees as input, optimal sampling protocols for summary-based species tree methods (e.g. phylogenetically informative, non-fragmentary loci, RAxML for gene tree inference) may likely translate to networks (Mirarab et al. 2014; Xi et al. 2015; Hosner et al. 2016; Meiklejohn et al. 2016).

Model selection for the presence of gene flow and for the number of reticulations

Another major challenge with networks is the selection of an appropriate number of reticulations. As a start, an important question that researchers should ask is whether or not they think that introgression was important during diversification of the study group. In other words, is the added complexity of a network really necessary to reconstruct evolutionary history? This amounts to choosing between a species tree and a reticulate species network, that

is, between zero reticulations versus one or more reticulations. In addition to examination of external information, exploratory population-level analyses (e.g. STRUCTURE; Pritchard et al. 2000) could be performed if the data contain multiple samples per species. Signs of significant admixture may then suggest the use of network methods. Unfortunately, sampling a single individual per species precludes the use of population genetic clustering methods, making it difficult to determine if a network is warranted. Further, the interpretation of admixture plots from clustering analyses can be problematic due to assumptions regarding population sampling and historical demography (Lawson et al. 2018). In these cases, other exploratory analyses such as 3s (Zhu and Yang 2012; Dalquen et al. 2017) may be used to test for gene flow between two species. Although limited to two species (plus an outgroup) and three alleles per locus, 3s can easily accommodate thousands of loci and offers a rigorous likelihood ratio test. The ABBA-BABA test (Green et al. 2010; Durand et al. 2011), D_{FOIL} (Pease and Hahn 2015) and HyDe (Blischak et al. 2018) are similar approaches, working on a subset of three or four ingroup species (plus an outgroup), which also use the sequence data directly. These tests are extremely fast, and can be applied to all adequate subsets of taxa to address specific questions (e.g. ExD_{FOIL}; Lambert et al. in press, to apply the D_{FOIL} test to all five-taxon subsets with the five-taxon topology required by D_{FOIL}). These approaches offer rigorous ways to do model selection and weigh the evidence in favor of a tree versus a reticulated history. However, they can result in thousands of tests (each on a small subset of taxa), creating challenges with multiple testing and with possible contradictory results across subsets sampled from the same groups.

Hypothesis testing on full networks

Rigorous model selection of the full network, with all taxa, is much more challenging. Under a likelihood framework, adding hybrid edges will increase the likelihood or pseudolikelihood score, and lead to overparameterized models if one does not penalize for the added model complexity.

The TreeMix method, which takes allele frequencies as input, uses a pre-specified number of migration edges to ease interpretability, and mentions multiple testing as a serious impediment to deriving a rigorous framework for selecting the appropriate number of reticulation events (Pickrell and Pritchard 2012). For example, when comparing a tree model to a network model with one reticulation, there are about $4n^2$ network models that can augment each tree model, where n is the number of species. This is because there are about $2n$ edges in the tree, and a reticulation is defined by a pair of edges: the donor and the recipient of gene flow. Therefore, the hypothesis of "one reticulation" contains vastly more models than the hypothesis "zero reticulations". More generally, the hypothesis of h+1 reticulations also contains vastly more models that the hypothesis of h reticulations, such that model selection criteria like AIC and BIC are inadequate for selecting the appropriate number of reticulations. The penalty of a given hypothesis needs to account for the number of models that make this hypothesis, in addition to the number of parameters in these models (Barron et al. 1999; Baraud et al. 2009 for Gaussian models). The penalties used by AIC and BIC only penalize the number of parameters such as branch lengths and inheritance parameters. AIC and BIC fail to account for the number of networks within a fixed number of reticulations (parameters), which grows very fast with the number of reticulations. A similar issue arises in phylogenetic comparative methods, for selecting among models with an unknown number of events along a phylogeny, like shifts in selection regime or diversification rate. The growing number of models within a given number of shifts needs to be accounted for by the model selection procedure, otherwise the traditional AIC and BIC fail to control the risk of false shift detection (Khabbazian et al. 2016; Bastide et al. 2018). With more computationally feasible pseudolikelihood approaches like SNaQ, the problem is even more complex because pseudolikelihoods (also called composite likelihoods) cannot even be used for model selection: a full likelihood is required to perform a likelihood ratio test, or to perform model selection with information criteria (AIC or BIC). One solution is to plot the maximum number of reticulations versus the pseudolikelihood score, to determine when a

plateau begins to appear. Such data-driven methods to calibrate model selection penalties have theoretical foundations in some situations (Baudry et al. 2012; implemented in the R package capusche).

Bayesian model selection and Bayes factors

Bayesian frameworks naturally account for the number of networks of a given reticulation complexity, through a prior distribution on the number of reticulations. In PhyloNet (Wen and Nakhleh 2018), the number of reticulations is given a Poisson distribution, censored at an allowed maximum. In BEAST2 (Zhang et al. 2018a) the network is given a model-based birth-hybridization prior, where each lineage speciates with rate λ, and each pair of lineages merge with rate v. In the AIM extension to StarBEAST2 (Mueller et al. 2018), the number of non-zero migration rates is given a Poisson prior. In these approaches, informative priors are used to ensure that inferred networks have a manageable number of reticulations (e.g. by constraining the mean of the Poisson distribution, or by ensuring hybridization rates lower than birth rates). It is unclear how the prior average number of reticulations affects the posterior estimated number of reticulations, but Bayes factors could be used to test specific hypotheses (ratio of marginal likelihoods of two models). For example, the Bayes factor for one or more reticulations could be calculated by comparing the posterior with the prior probabilities for no reticulation and for one or more reticulations overall. One could also calculate the Bayes factor for the presence / absence of gene flow between a particular pair of lineages, although the large number of pairs of lineages bring up multiple comparison issues. Alternatively, the posterior credibility interval for an inheritance value or for a migration rate can be used to assess the importance of a migration route, as done in G-PhoCS (Gronau et al. 2011), where the user needs to specify migration bands on a constraint tree. Overall model comparison is much more difficult, because marginal likelihoods are very difficult to calculate. The easy harmonic mean estimator is unstable and should not be used (Lartillot and Philippe 2006). Reliable alternatives, such as stepping stone or

path sampling, require extra sampling procedures and specific implementation, not currently available for Bayesian network estimation. The main downside of these rigorous Bayesian approaches is that they do not scale to modern data sets with many loci, or more than a handful of taxa.

Simulation-based assessment of candidate hypotheses

In cases when a handful of reticulation hypotheses can be formulated, the model selection task is much simpler. Simulations followed by Approximate Bayesian Computation (ABC) or machine learning can be used. For instance, Burbrink and Gehara (2018) compared three hypotheses: a bifurcation hypothesis with no gene flow, a unidirectional migration hypothesis with continuous gene flow between two particular lineages, and a hybridization hypothesis between two specific lineages. They used neural networks to estimate the posterior probability of each of the three models. The idea is to simulate a large number of data sets under each hypothesis and calculate various summary statistics on each data set. These simulated data can then serve as input to train a machine learning method to predict the generating hypothesis given an observed set of summary statistics. Pudlo et al. (2016) used this approach with random forests and presented a rigorous method to estimate model posterior probabilities. Alternatively, ABC approaches estimate model posterior probabilities by retaining a fraction of all the simulated data sets that fall closest to the observed data according to summary statistics. For example, Nater et al. (2015) used fastsimcoal2 (Excoffier et al. 2013) and ms (Hudson 2002) to simulate data from each of four scenarios of divergence with gene flow between four species of flycatchers, then used the R package “abc” (Csilléry et al. 2012) to estimate the posterior probabilities of these four models. The main hurdle of these ABC and machine learning approaches is the need to narrow down the problem to a small set of hypotheses, and a good choice of summary statistics.

In conclusion, objective means of selecting the minimum number of hybrid edges to best explain the data are generally lacking, which can have deleterious downstream consequences. For example, suppose a researcher decides that 10 hybrid edges best represent the data, when in fact, hybridization/introgression has been rare within the clade. These results would suggest that multiple species boundaries may need re-evaluation, which could then have negative consequences from a conservation perspective. Model selection is now commonly used in phylogeography studies to select among competing demographic models with or without gene flow (Gutenkunst et al. 2009; Jouganous et al. 2017), but analogous methods at and above the species level are lacking. Full Bayesian methods show promise, but their extreme computational expense precludes analyses of genomic data and/or a large number of species, the former of which is often needed to accurately detect introgression (Solís-Lemus and Ané 2016). Using Bayes factors to select a model with zero reticulations (species tree) versus one or more reticulations (species network) seems like a useful approach that should be explored further.

Model violations may lead to spurious or extra reticulations

Modelling gene flow is challenging because the patterns created by gene flow reside in the *variability* of phylogenetic relationships, and not in the *average* phylogenetic signal. For example, no variation corresponds to complete agreement between gene trees, and no gene flow. If instead genes give a mixed signal with two trees, say either A and B sister (AB,C) or B and C sister (A,BC), then variation is most extreme when each tree is supported by half of the genes. This case would provide the strongest evidence for gene flow, with B suspected of hybrid origin (Fig. 2a-2b), or A or C suspected of hybrid origin depending on branch lengths in gene trees (Fig. 2c-2d). In this three-taxon example, variation in gene trees and variation in branch lengths are both informative about reticulation. With two species only, variation in gene tree branch lengths, that is, genetic distance, is informative about reticulation (Fig. 3), although linked selection and variation in mutation rate could mask this information. Bimodal and

multimodal variation in divergence times has been used by Wen and Nakhleh (2018) to detect repeated bouts of gene flow between the same pair of lineages, assuming no rate variation.

Since gene flow signals itself through phylogenetic variation, it is imperative to account for ILS also, and detect gene flow from variation not explained by ILS. As a crude analogy, estimating a tree is like estimating an average, which is the idea underlying the concatenation of genes. In contrast, estimating reticulation edges is like estimating a variance, from how gene trees vary away from the species tree. In the very different context of linear regression, It is well known that testing hypotheses about variances is a lot harder than testing hypotheses about means, and most tests about variances are extremely sensitive to violations of their model assumptions (e.g. Box 1953). One can worry, therefore, that violations of model assumptions could affect network inference much more dramatically than tree inference. For example, undetected paralogy for a few genes might not affect the "average" signal much (although see Brown and Thomson 2017), but could affect the "discordance" signal enough to cause an extra reticulation edge during network estimation. A few outlier genes might mislead the estimation of networks, by using spurious reticulations to explain the observed gene tree variation. More generally, other sources of noise could negatively affect the number and position of inferred reticulations. Outliers and model violations could be due to systematic biases, undetected paralogy wrongly interpreted as allelic variation, or undetected allelic variation. In fact, Lambert et al. (in press) showed that batch-specific errors in RADseq data causes biased inference of introgression with D_{FOIL} tests (Pease and Hahn 2015), the 5-taxon extension of the ABBA-BABA test (Durand et al. 2011), with spurious introgression inferred to explain the similarity between samples in the same batch.

To assess these speculations, we considered the constant-rate assumption made by most methods that use aligned sequences directly, or branch lengths in gene trees. The assumption that substitution rates did not vary across genes and did not vary across lineages might be reasonable in recent species complexes or for phylogeography studies, but doubtful in

most other groups. To illustrate a violation of a molecular clock, we simulated loci under the coalescent on the four-taxon species tree assumed by the ABBA-BABA test. Gene tree branch lengths were rescaled by different factors to obtain substitutions per sites and simulate an increase in the rate of evolution in two of the four taxa, thus violating the clock assumption. We analyzed the data with the ABBA-BABA test, because this fast method is commonly used as a rigorous test of reticulation on large concatenated alignments. We also ran the Bayesian method for unlinked bi-allelic sites in PhyloNet, an explicit network approach using the same data type as the ABBA-BABA test, and which also assumes no rate variation across lineages. Rate heterogeneity caused both methods to incorrectly infer introgression, which would incorrectly suggest the presence of gene flow and potentially lead to overparameterization (Fig. 4).

This example illustrates that extra introgression events may be wrongly inferred by network approaches, as an artifact to fit residual variation not already explained by a tree model, even if this residual variation might be due to model violations. Researchers need to seek a balance between methods that have relaxed assumptions about gene flow but make strong assumptions about the substitution process, and traditional tree-based methods that make a strong assumption of no gene flow, but with relaxed assumptions about the substitution process. Our limited simulation suggests that researchers should use multiple analytical tools that make different assumptions to help determine the propensity of introgression. Further, if a network is indicated, assessing topological congruence across alternative inference methods and across various taxon or gene subsampling strategies may provide more confidence in the location of reticulation.

It is important to note that far fewer studies have examined properties of phylogenetic networks in detail as compared to tree-based methods. Thus, more simulation and empirical studies are needed to better elucidate strengths and weaknesses of the variety of network methods now available, under various model violations and sampling strategies. Tools for model

adequacy, not simply model selection, should also be developed to help diagnose cases when assumption violations are responsible for favoring extra reticulations (Brown 2014).

Multi-pronged approach

Given the realization that gene flow has likely impacted the evolutionary history of numerous groups, along with increasing ease with which large molecular data sets can be collected, interest in network methods will continue. However, it is important for researchers to realize that these methods are best used in addition to, not in place of, more conventional tree-based methods. For many biological systems it is likely that choosing either tree or network approaches in isolation would be less than ideal. Tree-based methods can generally accommodate more data and can estimate additional evolutionary parameters. They have received decades of work. These methods can now accommodate flexible assumptions about the substitution process and evolutionary rate variation, with rigorous tools for model selection, data partitioning, etc. They can combine different data types such as molecules, morphology, geography and fossil dates, enabling integrated analyses to answer complex questions on divergence times, on the diversification process, or on the effect of traits on diversification (see for instance Gavryushkina et al. 2017 for an integrated model of molecular evolution, morphological evolution, diversification process and fossilization process). However, tree-based methods can possibly lead to incorrect inferences if there is widespread introgression. Given their speed, tree-based methods are still much easier to use than network-based methods, to assess violations of assumptions. For example, it is computationally feasible to repeatedly delete one taxon at a time or one gene at a time and infer a tree each time, to detect outlier genes that might need further scrutiny (e.g. Brown and Thomson 2017). Conversely, networks are attractive because they relax the assumption of no gene flow. However, they are only applicable to small taxonomic subsets due to their computational demands. They may also provide an incorrect backbone topology simply due to lack of phylogenetic signal when

analyzing relatively small data sets, or due to long-branch attraction having a stronger negative impact with restricted taxon sampling (Heath et al. 2008; Roch et al. 2019). Networks may also provide spurious inferences due to violations of model assumptions. Therefore, we advocate for leveraging the strength of both tree and network approaches, which complement each other. For instance, Burbrink and Gehara (2018) discovered ancient reticulation in New World kingsnakes using species networks, and used tree-based methods to estimate divergence times and ancestral geographical areas. Similarly, using thousands of UCE loci, Blair et al. (2019) combined tree-based approaches to delimit species of rattlesnakes and estimate species trees and divergence times with a network approach to estimate the history of deep reticulation. By combining both types of analyses these studies were able to provide a more comprehensive picture into the evolutionary history of these groups.

Phylogenetic trees are commonly used for hypothesis testing in a variety of fields from medicine and epidemiology to ecology, evolution, biogeography and conservation. Reconstruction of accurate historical relationships remains paramount, results of which can directly and indirectly impact human health and ecosystem function. Prior to the adoption of the MSC, a traditional systematist typically inferred a tree using several optimality criteria (e.g. parsimony, likelihood, Bayesian). Nowadays, it is common to perform a suite of coalescent-based species tree analyses in addition to concatenation, the latter of which is widely used to test for species monophyly and help assign individuals to species prior to species tree inference (Blair et al. 2019). If gene flow is expected or present, adding networks to the list of tools seems like a natural extension to the field as a whole, and will no doubt provide novel insight into the propensity for reticulation/hybridization throughout the Tree of Life.

Acknowledgments

We would like to thank B. Carstens, L. Kubatko, M. Hahn, and two anonymous reviewers for their comments, which helped improve the overall quality of the manuscript.

References

Baraud Y., Giraud C., Huet S. 2009. Gaussian model selection with an unknown variance. Ann. Stat. 37:630–672.

Barron A., Birgé L., Massart P. 1999. Risk bounds for model selection via penalization. Probab. Theory Relat. Fields. 113:301–413.

Bastide P., Ané C., Robin S., Mariadassou M. 2018. Inference of adaptive shifts for multivariate correlated traits. Syst. Biol. 67:662–680.

Baudry J.-P., Maugis C., Michel B. 2012. Slope heuristics: overview and implementation. Stat. Comput. 22:455–470.

Blair C., Bryson R.W., Linkem C.W., Lazcano D., Klicka J., McCormack J.E. 2019. Cryptic diversity in the Mexican highlands: Thousands of UCE loci help illuminate phylogenetic relationships, species limits and divergence times of montane rattlesnakes (Viperidae: *Crotalus*). Mol. Ecol. Resour. 19:349–365.

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

Box G.E.P. 1953. Non-normality and tests on variances. Biometrika. 40:318–335.

Bravo G.A., Antonelli A., Bacon C.D., Bartoszek K., Blom M.P.K., Huynh S., Jones G., Knowles L.L., Lamichhaney S., Marcussen T., Morlon H., Nakhleh L.K., Oxelman B., Pfeil B., Schliep A., Wahlberg N., Werneck F.P., Wiedenhoeft J., Willows-Munro S., Edwards S.V. 2019. Embracing heterogeneity: coalescing the Tree of Life and the future of phylogenomics. PeerJ. 7:e6399.

Brown J.M. 2014. Detection of implausible phylogenetic inferences using posterior predictive assessment of model fit. Syst. Biol. 63:334–348.

Brown J.M., Thomson R.C. 2017. Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses. Syst. Biol. 66:517–530.

Burbrink F.T., Gehara M. 2018. The biogeography of deep time phylogenetic reticulation. Syst. Biol. 67:743–755.

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

Chung Y., Ané C. 2011. Comparing two Bayesian methods for gene tree/species tree reconstruction: simulations with incomplete lineage sorting and horizontal gene transfer. Syst. Biol. 60:261–275.

Csilléry K., François O., Blum M.G.B. 2012. abc: an R package for approximate Bayesian computation (ABC). Methods Ecol. Evol. 3:475–479.

Cui R., Schumer M., Kruesi K., Walter R., Andolfatto P., Rosenthal G.G. 2013. Phylogenomics reveals extensive reticulate evolution in *Xiphophorus* fishes. Evol. Int. J. Org. Evol. 67:2166–2179.

Dalquen D.A., Zhu T., Yang Z. 2017. Maximum likelihood implementation of an isolation-with-migration model for three species. Syst. Biol. 66:379–398.

Degnan J.H. 2018. Modeling hybridization under the network multispecies coalescent. Syst. Biol. 67:786–799.

Degnan J.H., Rosenberg N.A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24:332–340.

Durand E.Y., Patterson N., Reich D., Slatkin M. 2011. Testing for ancient admixture between closely related populations. Mol. Biol. Evol. 28:2239–2252.

Eckert A.J., Carstens B.C. 2008. Does gene flow destroy phylogenetic signal? The performance of three methods for estimating species phylogenies in the presence of gene flow. Mol. Phylogenet. Evol. 49:832–842.

Edwards S.V. 2009. Is a new and general theory of molecular systematics emerging? Evolution. 63:1–19.

Edwards S.V. 2016. Inferring species trees. In: Kliman R., editor. Encyclopedia of Evolutionary Biology. New York: Elsevier Inc.

Edwards S.V., Xi Z., Janke A., Faircloth B.C., McCormack J.E., Glenn T.C., Zhong B., Wu S., Lemmon E.M., Lemmon A.R., Leaché A.D., Liu L., Davis C.C. 2016. Implementing and testing the multispecies coalescent model: a valuable paradigm for phylogenomics. Mol. Phylogenet. Evol. 94:447–462.

Elworth R. a. L., Ogilvie H.A., Zhu J., Nakhleh L. 2018. Advances in computational methods for phylogenetic networks in the presence of hybridization. Available from /paper/Advances-in-Computational-Methods-for-Phylogenetic-Elworth-Ogilvie/efba6207c5aab3a63d3e9ed47780bc36d46a7eb0.

Excoffier L., Dupanloup I., Huerta-Sánchez E., Sousa V.C., Foll M. 2013. Robust demographic inference from genomic and SNP data. PLoS Genet. 9:e1003905.

Flouri T., Jiao X., Rannala B., Yang Z. 2018. Species tree inference with BPP using genomic sequences and the multispecies coalescent. Mol. Biol. Evol. 35:2585–2593.

Gavryushkina A., Heath T.A., Ksepka D.T., Stadler T., Welch D., Drummond A.J. 2017. Bayesian total-evidence dating reveals the recent crown radiation of penguins. Syst. Biol. 66:57–73.

Green R.E., Krause J., Briggs A.W., Maricic T., Stenzel U., Kircher M., Patterson N., Li H., Zhai W., Fritz M.H.-Y., Hansen N.F., Durand E.Y., Malaspinas A.-S., Jensen J.D., Marques-Bonet T., Alkan C., Prüfer K., Meyer M., Burbano H.A., Good J.M., Schultz R., Aximu-Petri A., Butthof A., Höber B., Höffner B., Siegemund M., Weihmann A., Nusbaum C., Lander E.S., Russ C., Novod N., Affourtit J., Egholm M., Verna C., Rudan P., Brajkovic D., Kucan Ž., Gušic I., Doronichev V.B., Golovanova L.V., Lalueza-Fox C., Rasilla M. de la, Fortea J., Rosas A., Schmitz R.W., Johnson P.L.F., Eichler E.E., Falush D., Birney E., Mullikin J.C., Slatkin M., Nielsen R., Kelso J., Lachmann M., Reich D., Pääbo S. 2010. A draft sequence of the Neandertal genome. Science. 328:710–722.

Gronau I., Hubisz M.J., Gulko B., Danko C.G., Siepel A. 2011. Bayesian inference of ancient human demography from individual genome sequences. Nat. Genet. 43:1031–1034.

Gutenkunst R.N., Hernandez R.D., Williamson S.H., Bustamante C.D. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLOS Genet. 5:e1000695.

Heath T.A., Hedtke S.M., Hillis D.M. 2008. Taxon sampling and the accuracy of phylogenetic analyses. J. Syst. Evol. 46:239–257.

Hejase H.A., Liu K.J. 2016. A scalability study of phylogenetic network inference methods using empirical datasets and simulations involving a single reticulation. BMC Bioinformatics. 17:422.

Hejase H.A., VandePol N., Bonito G.M., Liu K.J. 2018. FastNet: fast and accurate statistical inference of phylogenetic networks using large-scale genomic sequence data. In: Blanchette M., Ouangraoua A. (eds) Comparative Genomics. RECOMB-CG 2018. Lecture Notes in Computer Science, vol 11183. Springer, Cham. pp. 242–259.

Hey J., Chung Y., Sethuraman A., Lachance J., Tishkoff S., Sousa V.C., Wang Y. 2018. Phylogeny estimation by integration over isolation with migration models. Mol. Biol. Evol. 35:2805–2818.

Hosner P.A., Faircloth B.C., Glenn T.C., Braun E.L., Kimball R.T. 2016. Avoiding missing data biases in phylogenomic inference: an empirical study in the Landfowl (Aves: Galliformes). Mol. Biol. Evol. 33:1110–1125.

Hudson R.R. 1983. Testing the constant-rate neutral allele model with protein sequence data. Evolution. 37:203–217.

Hudson R.R. 2002. Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics. 18:337–338.

Jónsson H., Schubert M., Seguin-Orlando A., Ginolhac A., Petersen L., Fumagalli M., Albrechtsen A., Petersen B., Korneliussen T.S., Vilstrup J.T., Lear T., Myka J.L., Lundquist J., Miller D.C., Alfarhan A.H., Alquraishi S.A., Al-Rasheid K.A.S., Stagegaard J., Strauss G., Bertelsen M.F., Sicheritz-Ponten T., Antczak D.F., Bailey E., Nielsen R., Willerslev E., Orlando L. 2014. Speciation with gene flow in equids despite extensive chromosomal plasticity. Proc. Natl. Acad. Sci. 111:18655–18660.

Jouganous J., Long W., Ragsdale A.P., Gravel S. 2017. Inferring the joint demographic history of multiple populations: beyond the fiffusion approximation. Genetics. 206:1549–1567.

Khabbazian M., Kriebel R., Rohe K., Ané C. 2016. Fast and accurate detection of evolutionary shifts in Ornstein–Uhlenbeck models. Methods Ecol. Evol. 7:811–824.

Kubatko L.S., Degnan J.H. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst. Biol. 56:17–24.

Lambert S.M., Streicher J.W., Fisher-Reid M.C., Cruz F.R.M. de la, Martínez-Méndez N., Vázquez U.O.G., Oca A.N.M. de, Wiens J.J. 2019. Inferring introgression using RADseq and DFOIL: power and pitfalls revealed in a case study of spiny lizards (*Sceloporus*). Mol. Ecol. Resour. 19:818–837.

Lartillot N., Philippe H. 2006. Computing Bayes Factors using thermodynamic integration. Syst. Biol. 55:195–207.

Lawson D.J., Dorp L. van, Falush D. 2018. A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. Nat. Commun. 9:3258.

Leaché A.D., Harris R.B., Rannala B., Yang Z. 2014. The influence of gene flow on species tree estimation: a simulation study. Syst. Biol. 63:17–30.

Liu L., Yu L. 2011. Estimating species trees from unrooted gene trees. Syst. Biol. 60:661–667.

Liu L., Yu L., Edwards S.V. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol. Biol. 10:302.

Long C., Kubatko L. 2018. The effect of gene flow on coalescent-based species-tree inference. Syst. Biol. 67:770–785.

Maddison W.P. 1997. Gene trees in species trees. Syst. Biol. 46:523–536.

Mallet J. 2005. Hybridization as an invasion of the genome. Trends Ecol. Evol. 20:229–237.

Mallet J. 2007. Hybrid speciation. Nature. 446:279–283.

Mallet J. 2008. Hybridization, ecological races and the nature of species: empirical evidence for the ease of speciation. Philos. Trans. R. Soc. B Biol. Sci. 363:2971–2986.

Meiklejohn K.A., Faircloth B.C., Glenn T.C., Kimball R.T., Braun E.L. 2016. Analysis of a rapid evolutionary radiation using ultraconserved elements: evidence for a bias in some multispecies coalescent methods. Syst. Biol. 65:612–627.

Mirarab S., Reaz R., Bayzid M.S., Zimmermann T., Swenson M.S., Warnow T. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinforma. Oxf. Engl. 30:i541-548.

Mueller N.F., Ogilvie H., Zhang C., Drummond A., Stadler T. 2018. Inference of species histories in the presence of gene flow. bioRxiv.:348391.

Nater A., Burri R., Kawakami T., Smeds L., Ellegren H. 2015. Resolving evolutionary relationships in closely related species with whole-genome sequencing data. Syst. Biol. 64:1000–1017.

Ogilvie H.A., Bouckaert R.R., Drummond A.J. 2017. StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. Mol. Biol. Evol. 34:2101–2114.

Pease J.B., Hahn M.W. 2015. Detection and polarization of introgression in a five-taxon phylogeny. Syst. Biol. 64:651–662.

Pickrell J.K., Pritchard J.K. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. PLOS Genet. 8:e1002967.

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

Pudlo P., Marin J.-M., Estoup A., Cornuet J.-M., Gautier M., Robert C.P. 2016. Reliable ABC model choice via random forests. Bioinformatics. 32:859–866.

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

Roch S., Nute M., Warnow T. 2019. Long-branch attraction in species tree estimation: inconsistency of partitioned likelihood and topology-based summary methods. Syst. Biol. 68:281–297.

Roch S., Steel M. 2015. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. Theor. Popul. Biol. 100C:56–62.

Solís-Lemus C., Ané C. 2016. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. PLOS Genet. 12:e1005896.

Solís-Lemus C., Bastide P., Ané C. 2017. PhyloNetworks: a package for phylogenetic networks. Mol. Biol. Evol. 34:3292–3298.

Solís-Lemus C., Yang M., Ané C. 2016. Inconsistency of species tree methods under gene flow. Syst. Biol. 65:843–851.

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics. 105:437–460.

Than C., Ruths D., Nakhleh L. 2008. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. BMC Bioinformatics. 9:322.

Wascher M., Kubatko L. 2019. Consistency of SVDQuartets and maximum likelihood for coalescent-based species tree estimation. bioRxiv.:523050.

Wen D., Nakhleh L. 2018. Coestimating reticulate phylogenies and gene trees from multilocus sequence data. Syst. Biol. 67:439–457.

Wen D., Yu Y., Zhu J., Nakhleh L. 2018. Inferring phylogenetic networks using PhyloNet. Syst. Biol. 67:735–740.

Xi Z., Liu L., Davis C.C. 2015. Genes with minimal phylogenetic information are problematic for coalescent analyses when gene tree estimation is biased. Mol. Phylogenet. Evol. 92:63–71.

Zhang C., Ogilvie H.A., Drummond A.J., Stadler T. 2018a. Bayesian inference of species networks from multilocus sequence data. Mol. Biol. Evol. 35:504–517.

Zhang C., Rabiee M., Sayyari E., Mirarab S. 2018b. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics. 19:153.

Zhu T., Yang Z. 2012. Maximum likelihood implementation of an isolation-with-migration model with three species for testing speciation with gene flow. Mol. Biol. Evol. 29:3131–3142.

Figure legends

Figure 1: Anomalous gene trees are possible when speciation is quickly followed by gene flow and another speciation. The population history is depicted by thick lines (light grey). The history of three unlinked genes are shown with thin lines, embedded within the population history. If speciation is rapid and gene flow is strong, the matching gene tree $S_1S_2|O$ may be less frequent than either discordant gene tree $S_1O|S_2$ and $S_2O|S_1$.

Figure 2: Example of evidence of reticulation from variation in gene trees across the genome. In all panels, numbers on edges represent average numbers of substitutions per site. a) Variation in gene trees; with 50% of genes having the first gene tree AB|C and 50% of genes having the second gene tree A|BC, where the genetic distance between A and C is invariant across gene trees. b) Species network that could give rise to the gene trees in a), in which B is of hybrid origin. c) Variation in gene trees with 50% of genes having tree AB|C and 50% of genes having A|BC, where the genetic distance between A and B is invariant across gene trees. d) Species network that could give rise to the gene trees in c), in which C is of hybrid origin between B and another parent unsampled or extinct, denoted as a cross in the first network. The second network is obtained by suppressing the unsampled parent taxon in the first network, and by simplifying nodes of degree 2. A third scenario, not shown, corresponds to the case where the genetic distance between B and C is invariant and A is of hybrid origin between B and an unsampled or extinct species. Note that ILS needs to be absent to explain that there would be only two possible gene trees in a) and c). That is, networks in b) and d) need to have small population sizes, i.e. long branch lengths in coalescent units. The presence of ILS would cause extra variation in gene trees: in branch lengths, and in topology with some genes having tree AC|B.

Figure 3: Example of evidence or reticulation from variation in genetic distance across the genome. a) Variation in gene trees, with some proportion of genes having short branch lengths, intermediate branch lengths, or long branch lengths. b) Species network that could explain this tri-modal pattern of genetic distance variation, in which a species experienced a split into separate populations followed by population merge, twice in its history. The species network is shown with thick lines (light grey), with gene trees embedded inside as thin lines. Species split and merge might have been caused by glaciation fragmenting the species' habitat, followed by population expansion after glacier retreat. Unlike in Fig. 2, ILS is necessary for gene trees to vary: the genes with intermediate or long branch lengths are those that failed to coalesce during one or both periods when there was a single population.

Figure 4: Probability of falsely detecting reticulation increases with variation in substitution rates by methods assuming no rate variation. Data were simulated under the four-taxon tree used by the ABBA-BABA test, with an internal branch of one coalescent unit: $(((P_1{:}1,P_2{:}1){:}1,P_3{:}2){:}0.1,O{:}2.1)$. Branch lengths in gene trees were converted to substitutions per site by multiplying them by 0.04, except for the two external branches to P_2 and P_3, which were multiplied by $0.04r$, r =1 to 5 (rate ratio). A molecular clock holds when r =1, and the clock is violated otherwise. Aligned sequences were generated from gene trees using the JC model. Each simulated data set consisted of 5000 unlinked sites, as is assumed by the PhyloNet methods for SNP data. **Gray**: The MCMC_BiMarkers method in PhyloNet was used to estimate a network after removing sites with more than two alleles using default parameters (such as a prior mean of one reticulation), except that the maximum number of reticulations was set to one to increase the relative weight of a species tree, and the population size (scaled by the mutation

rate) was allowed to vary across branches (options -varytheta -esptheta). The analysis used two million iterations, discarding $4x10^5$ for burnin, and sampled every 800. The network with gene flow from P_3 to P_2 was recovered with high posterior probability in all cases when the clock was violated. **Black**: The alignment was analyzed with the ABBA-BABA test. Because sites were unlinked, p-values were obtained with a chi-square test comparing the proportions of ABBA and BABA sites to 0.5. The curve shows the proportion of times that the p-value was less than 0.05, out of 200 replicates. The D statistic tended to be significantly positive: there was an excess of ABBA sites where taxa P_2 and P_3 have a shared state, falsely detecting gene flow between P_2 and P_3 when the clock was violated. Intuitively, the problem is that an elevated mutation rate in taxa P_2 and P_3 causes a non-negligible excess of homoplasious ABBA sites with independent substitutions in P_2 and in P_3, and an asymmetry in ABBA versus BABA sites. Yet, symmetry is expected under ILS alone and under a clock, because the species tree is unchanged if we swap P_1 with P_2, so long as the external branch lengths to P_1 with P_2. are equal, as is the case under a molecular clock. Therefore, reticulation between P_2 and P_3 is inferred by methods assuming a clock, to explain the excess number of ABBA sites.