Syst. Biol. 65(5):843–851, 2016

DOI:10.1093/sysbio/syw030
Advance Access publication May 5, 2016

Inconsistency of Species Tree Methods under Gene Flow

CLAUDIA SOLÍS-LEMUS[1,*], MENGYAO YANG[1], AND CÉCILE ANÉ[1,2]

[1]*Department of Statistics, University of Wisconsin, Madison, WI, 53706, USA;* [2]*Department of Botany, University of Wisconsin, Madison, WI, 53706, USA*
[*]*Correspondence to be sent to: Claudia Solís-Lemus, Department of Statistics, University of Wisconsin-Madison, 1220 Medical Sciences Center, 1300 University Avenue, Madison, WI, 53706, USA; E-mail: claudia@stat.wisc.edu.*

Received 20 October 2015; reviews returned 30 March 2016; accepted 1 April 2016
Associate Editor: Peter Foster

Abstract.—Coalescent-based methods are now broadly used to infer evolutionary relationships between groups of organisms under the assumption that incomplete lineage sorting (ILS) is the only source of gene tree discordance. Many of these methods are known to consistently estimate the species tree when all their assumptions are met. Nonetheless, little work has been done to test the robustness of such methods to violations of their assumptions. Here, we study the performance of two of the most efficient coalescent-based methods, ASTRAL and NJst, in the presence of gene flow. Gene flow violates the assumption that ILS is the sole source of gene tree conflict. We find anomalous gene trees on three-taxon rooted trees and on four-taxon unrooted trees. These anomalous trees do not exist under ILS only, but appear because of gene flow. Our simulations show that species tree methods (and concatenation) may reconstruct the wrong evolutionary history, even from a very large number of well-reconstructed gene trees. In other words, species tree methods can be inconsistent under gene flow. Our results underline the need for methods like PhyloNet, to account simultaneously for ILS and gene flow in a unified framework. Although much slower, PhyloNet had better accuracy and remained consistent at high levels of gene flow. [Anomalous gene trees; ASTRAL; coalescent; concatenation; hybridization; network; NJst; PhyloNet]

Methods based on the coalescent, here called "species-tree" methods, are now widely used to reconstruct the evolutionary history of species in the presence of gene tree discordance. The multispecies coalescent process (Knowles and Kubatko 2010) models incomplete lineage sorting (ILS), one of the main sources of gene tree conflict, so it provides a powerful probabilistic framework to infer species relationships from molecular data. Among species tree methods, ASTRAL (Mirarab et al. 2014c) and NJst (Liu and Yu 2011) are extensively used because they are fast enough to handle large genomic data, and both have been shown to be among the most accurate (Whelan 2011; Mirarab et al. 2014b, 2014c). Compared to other coalescent-based methods, ASTRAL and NJst both have the advantage of using unrooted gene tree topologies as input, so they are robust to rooting and branch length errors in gene trees. Branch length errors can result from a paucity of informative sites, or from assumption violations in the substitution model, such as rate variation across genes and/or across lineages.

Summary methods like ASTRAL or NJst tend to be robust to violations of assumptions. However, little work has been done to study the robustness of coalescent-based methods to the violation of their main assumption: that all gene tree discordance is explained solely by ILS (but see Chung and Ané (2011) for robustness to diffuse gene flow affecting all populations, and Lanier and Knowles (2012) for robustness to within-gene recombination). A fast alternative to species tree methods is concatenation: where all loci are assumed to share the same tree *a priori* and gene tree conflict is ignored. Although this strategy is fast and can recover the species tree accurately in some cases (Mirarab et al. 2014b) it can also be misleading and inconsistent in other situations (Kubatko and Degnan 2007; Roch and Steel 2015), lacking robustness to the violation of its main assumption: that all loci share the same tree topology. In other words, concatenation is not robust to the presence of ILS. Note that this is not a consistency issue in the statistical sense (which would require studying the behavior of a method when its own assumptions are met). We will keep with the historical and liberal use of the term "inconsistent," but will specify assumptions. While most species tree methods are consistent under their own assumptions (e.g., Warnow 2015), here we raise the question of whether species tree methods are consistent when ILS is not the only source of gene tree conflict.

We consider here two sources of discordance, ILS and gene flow. We chose to study ASTRAL and NJst, both coalescent-based and widely used because they combine accuracy with speed. ASTRAL only relies on each quartet being consistently estimated. That is, it requires that for each four-taxon set, the quartet that agrees with the species tree (major quartet) has greater frequency than the other two quartets (minor quartets). Degnan (2013) showed that this is true when ILS is the only source of discordance: there are no anomalous unrooted gene trees (AUGTs) on four taxa. Thus, ASTRAL will reconstruct the correct species tree given enough genes (Warnow 2015). Here, we show that the presence of gene flow can create AUGTs even on four taxa. Thus, ASTRAL may no longer be consistent. More specifically, we present a scenario where the two minor quartets, those in disagreement with the species tree, are each supported by more genes than the quartet matching the species tree. In our simulations, we found that ASTRAL was inconsistent: given more and more well-reconstructed gene trees, it does not recover the tree with the major vertical signal. Instead, it reconstructs a wrong topology signaled by the minor quartets. To verify whether the reason for inconsistency was restricted to anomalous quartets we also studied the accuracy of NJst, which

is not quartet-based. NJst uses the complete gene trees as input, computes pairwise distances as the average number of nodes that separate two taxa in the gene trees, and then uses these distances to infer the species tree. Our simulations found that NJst was also inconsistent.

Our work emphasizes the problem of estimating a species tree when both ILS and gene flow play key roles in the discord between gene trees. There is an urgent need to use probabilistic methods that account for both sources of discordance (e.g., Kubatko 2009; Meng and Kubatko 2009; Yu et al. 2012, 2014). We included one such method in our simulations, PhyloNet (Yu et al. 2012, 2014). It uses a network to model gene flow and performs maximum likelihood on a set of rooted input gene trees, given a user-defined number of reticulations in the network. In our simulations, PhyloNet showed a consistent recovery of the true species tree, even under strong gene flow. This gain in accuracy comes at a computational cost, however: maximum likelihood in PhyloNet is much slower than ASTRAL or NJst and does not scale to many taxa. Our work shows the importance of modeling gene flow in addition to ILS, and a need for methods that scale to genomic data sets (Yu and Nakhleh 2015; Solís-Lemus and Ané 2016).

Gene Tree Model with ILS and Gene Flow

The multispecies coalescent model has already been utilized to simultaneously account for ILS and gene flow (Kubatko 2009; Meng and Kubatko 2009; Yu et al. 2012, 2014; Yu and Nakhleh 2015; Solís-Lemus and Ané 2016). In these papers, the models do not discriminate between gene flow, hybridization, or horizontal gene transfer (HGT), as the mathematical model is appropriate for any of these biological realities, although they assume that each event is restricted in time. To compute the probability of a gene tree given a species network, the coalescent model is considered inside each branch of the species network, just as in a species tree. Each reticulation node represents a gene flow event, at which point a gene lineage inherits one of the two parents' genetic material with inheritance probabilities γ and $1-\gamma$ (Fig. 1, left). In other words, γ summarizes gene flow across the genome, as the proportion of genes inherited through reticulation. The main underlying tree is obtained by suppressing edges with $\gamma < 0.5$ (Fig. 1, center and right).

It is worth noting that the model in Yu et al. (2012, 2014) is slightly different from that in Kubatko (2009), a distinction that has not been made quite explicit in the literature. In Kubatko (2009), all alleles at a given locus must be inherited from the same parent. We focus here on the more flexible model in Yu et al. (2012, 2014), in which each allele originates from a reticulation edge independently of all other alleles at the same locus and at other loci (see "Discussion" section for more details).

AUGTs on Four Taxa

Figure 1 shows a four-taxon tree with a gene flow event, forming a four-taxon network with

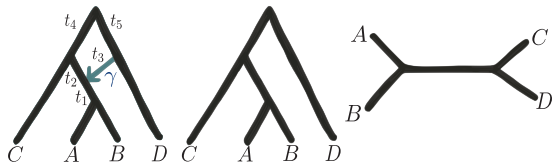

FIGURE 1. Four-taxon network with one hybridization event (left). If $\gamma < 0.5$, the species tree displaying the major vertical inheritance signal has D sister to ABC, and C sister to AB (center). The other tree displayed by this network is obtained by keeping the gene flow arrow and suppressing the edge of length t_2. Both trees have the same unrooted topology: AB|CD (right), thus called the major quartet (for all γ). For some branch lengths and γ values, this major quartet can, in fact, be supported by fewer genes than either minor quartets CA|BD and CB|AD, causing species tree reconstruction methods to favor grouping D with A or B instead of grouping A and B sister to each other. An edge length of $t_3 = 0$ corresponds to gene flow between contemporary species, if both have descendants in the sample. If the donor population became extinct or if none of its descendant species were sampled, then $t_3 > 0$.

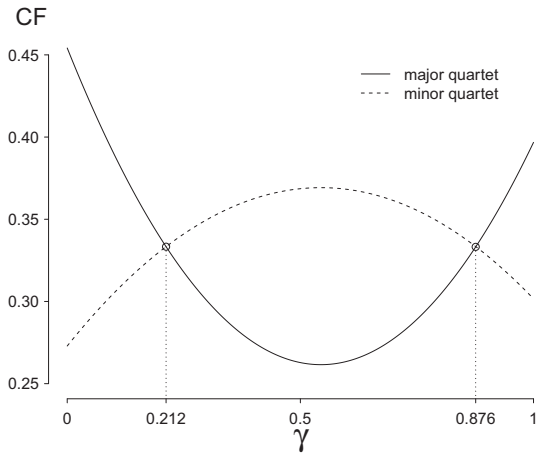

FIGURE 2. Anomaly zone for unrooted quartets, when $t_1 = t_2 = 0.1, t_3 = 0.0, t_4 = 2.0, t_5 = 2.1$ in Figure 1. When γ is in $(0.212, 0.867)$, both minor quartets (CA|BD and CB|AD) have greater probability than the major quartet, AB|CD.

six numerical parameters: $(\gamma, t_1, t_2, t_3, t_4, t_5)$. Some parameter combinations yield what Degnan (2013) denotes as (AUGTs). AUGTs are unrooted gene trees that do not match the species tree, yet have a higher probability than the topology matching the species tree. The underlying species tree (Fig. 1) is obtained by removing the gene flow arrow, assuming that gene flow affected less than 50% of genes ($\gamma < 0.5$). The unrooted gene tree matching this species tree is AB|CD. Its probability is the expected frequency of genes with this topology, also called *concordance factor* (CF) (Table 1). The CF of the two quartets that disagree with the "major" quartet displayed by the species tree increases with γ, and both are greater than the CF of the major quartet when $0.212 < \gamma < 0.876$, creating an *anomaly zone* (Fig. 2). These quartet CFs are given by (see Solís-Lemus

TABLE 1. Concordance factors (proportion of gene trees) of the three unrooted quartets under the gene flow model in Figure 1 and two sets of branch lengths: $t_1=t_2=0.1, t_3=0.0, t_4=2.0, t_5=2.1$ (set 1) and $t_1=t_2=0.01, t_3=t_4=t_5=1.0$ (set 2), for $\gamma=0.1$ or 0.3

Quartet	Resolution	Set 1		Set 2	
		$\gamma=0.1$	$\gamma=0.3$	$\gamma=0.1$	$\gamma=0.3$
AB|CD	major[a]	0.390	0.300	0.298	0.260
CA|BD	minor[b]	0.305	0.350	0.351	0.370
CB|AD	minor[b]	0.305	0.350	0.351	0.370

Notes: [a] The major quartet is the one that agrees with the species tree, which depicts the major vertical inheritance signal.
[b] The two minor CFs are actually greater than the major CF, except when branch lengths are long enough (set 1) and gene flow is less severe ($\gamma=0.1$).

TABLE 2. Concordance factors (proportion of gene trees) of the three rooted triplets under the gene flow model in Figure 1 and two sets of branch lengths (as in Table 1), when we only consider taxa A, B, and C

Triplet	Resolution	Set 1		Set 2	
		$\gamma=0.1$	$\gamma=0.3$	$\gamma=0.1$	$\gamma=0.3$
AB|C	major[a]	0.401	0.363	0.313	0.307
CA|B	minor[b]	0.299	0.319	0.343	0.347
CB|A	minor[b]	0.299	0.319	0.343	0.347

Notes: [a] The major triplet is the one that agrees with the species tree, which depicts the major vertical inheritance signal.
[b] The two minor CFs are actually greater than the major CF when branches are short enough (set 2).

and Ané 2016):

$$\mathrm{CF}_{\mathrm{AB|CD}} = (1-\gamma)^2(1-2/3e^{-t_1-t_2})+2\gamma(1-\gamma)(1-e^{-t_1}+1/3e^{-t_1-t_4-t_5})+\gamma^2(1-2/3e^{-t_1-t_3})$$

and $\mathrm{CF}_{\mathrm{CA|BD}}=\mathrm{CF}_{\mathrm{CB|AD}}=(1-\mathrm{CF}_{\mathrm{AB|CD}})/2$.

Anomalous rooted gene trees on three Taxa

The presence of gene flow can also create anomalous rooted gene trees (Table 2). Under the model in Figure 1, the rooted gene trees on species A, B, and C can be one of three rooted triplets: the major triplet displayed by the species tree, C|AB, or the minor triplets conflicting with the species tree, CA|B or AB|A. Their expected frequencies are

$$\mathrm{CF}_{\mathrm{C|AB}} = (1-\gamma)^2(1-2/3e^{-t_1-t_2})+2\gamma(1-\gamma)(1-e^{-t_1}+1/3e^{-t_1-t_4})+\gamma^2(1-2/3e^{-t_1-t_3-t_5})$$

and $\mathrm{CF}_{\mathrm{CA|B}}=\mathrm{CF}_{\mathrm{CB|A}}=(1-\mathrm{CF}_{\mathrm{C|AB}})/2$. The anomaly zone for rooted gene trees on three taxa does not appear to be as severe as that for unrooted trees on four taxa (Table 2). However, future work will be needed for fully characterize this anomaly zone and how it depends on the gene flow network topology.

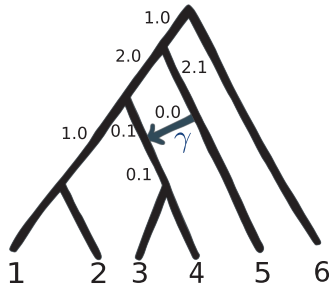

FIGURE 3. Species tree with gene flow used for simulations. The true species tree is obtained by ignoring the reticulation edge (labeled by γ), for $\gamma<0.5$. Branch lengths are in coalescent units, and extend the first set of branch lengths ("set 1") in Tables 1 and 2. Simulations used $\gamma=0.1$ and 0.3.

INCONSISTENCY WHEN THE ILS-ONLY MODEL IS VIOLATED

From True Gene Trees

We simulated gene trees with ms (Hudson 2002) under a six-taxon tree expanded by one gene flow event (Fig. 3) to see if species tree methods were robust to the presence of gene flow and could still reconstruct the true underlying species tree (obtained by removing the edge annotated by γ) despite gene flow. We chose a situation where some quartets are anomalous (Fig. 3) with branch lengths around the gene flow event that correspond to the less severe case in Table 1 (set 1), in which AUGTs only appear with $\gamma=0.3$, not with $\gamma=0.1$. For instance, to simulate 50 genes with $\gamma=0.1$, we used `ms 6 50 -T -I 6 1 1 1 1 1 1 -ej 0.1 1 2 -ej 0.5 3 4 -es 0.55 4 0.9 -ej 0.55 7 5 -ej 0.6 2 4 -ej 1.6 4 5 -ej 2.1 5 6`. ASTRAL and NJst were given the true gene trees as input, as opposed to gene trees estimated from molecular sequences. This choice was to consider the best-case scenario, in which error in the reconstructed species tree is solely due to the violation of the model of gene tree discordance, not to gene tree error reconstruction (mutational variance, see Huang et al. 2010). For each method, we calculated the average Robinson–Foulds (RF) distance between the estimated species tree and the true tree, and the number of times that the true species tree was correctly recovered. We also recorded the frequency of recovering other trees displaying the horizontal inheritance signal.

To compare the results between gene flow levels that do or do not lead to AUGTs, we chose $\gamma=0.1, 0.3$. For each γ value, we varied the number of gene trees from 10 to 10,000 and replicated each scenario 100 times. Ideally, as the number of genes increases, the frequency with which the true species tree is recovered should increase toward 100% and the mean RF distance should decrease to zero. With low gene flow ($\gamma=0.1$), both ASTRAL and NJst were able to estimate the true species tree with more and more accuracy as the number of genes increased (Fig. 4, solid lines and Fig. 5, left). However, at a higher level of gene flow ($\gamma=0.3$), the mean RF distance did not converge to zero with more and more genes (Fig. 4, dashed lines). Both methods reconstructed the wrong species tree with high probability, even with 10,000

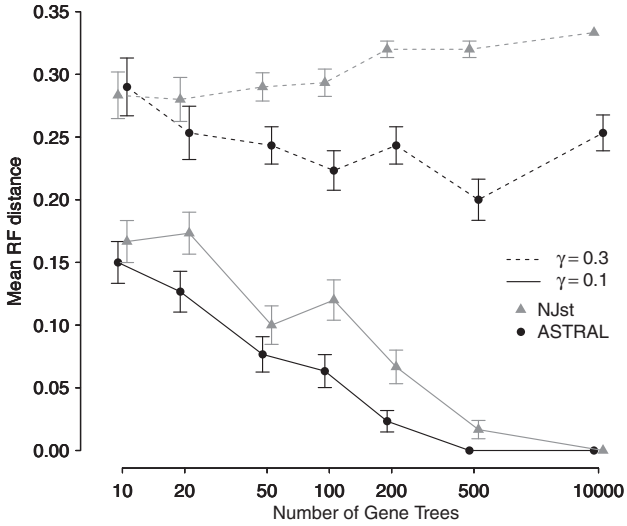

FIGURE 4. Accuracy of ASTRAL and NJst from true gene trees under ILS and gene flow (Fig. 3), measured by the mean RF distance between the true species tree and the estimated species tree. Bars show one standard error around the mean, each based on 100 replicates. Points were jittered horizontally to avoid clutter.

genes. In fact, both methods tended to reconstruct a tree displaying one of the anomalous quartets identified above: with either taxon 3 or 4 sister to a clade formed by taxa 1 and 2 (trees of types a and b in Fig. 5), especially from many genes. The frequency with which NJst inferred an incorrect tree of type a clearly converged to 100%, and was above 70% with as few as 20 genes. ASTRAL fared a little better, with an incorrect species tree reconstruction (of type a) in 69.8% cases from 50 or more genes, on average. Surprisingly, the alternate tree displayed in the network, where taxa 3 and 4 form a clade sister to taxon 5, was almost never recovered.

From Sequence Alignments

We used the previously simulated gene trees to simulate sequences of length 500 under HKY with Seq-Gen (Rambaut and Grassly 1997). For each gene, κ was drawn uniformly in $(1,4)$ and θ was drawn uniformly in $(0.025, 0.05)$. The base frequencies were each drawn uniformly in $(0.15, 0.35)$ then normalized to sum up to 1. Rate variation across sites was also simulated by drawing α uniformly in $(0.3, 3)$ for each gene. This was meant to mimic realistic conditions with variation between genes.

RAxML (Stamatakis 2014) was used with HKY and rate variation across sites to estimate a species tree from the concatenated alignment and to estimate individual gene trees, which were then used as input for ASTRAL, NJst, and PhyloNet (Yu et al. 2014). Unlike ASTRAL and NJst, PhyloNet requires rooted input trees. Estimated gene trees were rooted using the outgroup in the species tree (6). This rooting may have been erroneous in a few gene trees with deep coalescences in the most ancestral population. We assumed one reticulation in PhyloNet. Bootstrapping was also conducted for concatenation, ASTRAL and NJst (100 replicates) but not for PhyloNet because of its computational burden (see below).

Concatenation performed poorly in the presence of ILS and gene flow (Fig. 6). ASTRAL and NJst performed accurately with low levels of gene flow ($\gamma = 0.1$), but failed to reconstruct the correct species tree with $\gamma = 0.3$, as before. PhyloNet, on the contrary, recovered the correct species phylogeny accurately, by accounting for both ILS and gene flow in its underlying model. These results highlight the importance of using coalescent-based network models to reconstruct the evolutionary history of species when there is suspicion of gene flow.

To see if concatenation, ASTRAL or NJst gave high support for an incorrect tree, we computed the average bootstrap support for various bipartitions (Fig. 7). Regardless of γ, concatenation gave equivocal support for either relationship. ASTRAL and NJst gave high support for the true bipartition at low levels of gene flow ($\gamma = 0.1$). However, at higher levels of gene flow ($\gamma = 0.3$) both ASTRAL and NJst gave low or no bootstrap support for the true bipartition and increasingly high support for an incorrect relationship instead.

DISCUSSION

Gene Flow can Cause Anomalous Gene Trees

We identified a situation where the presence of gene flow causes the appearance of AUGTs on four taxa, and anomalous rooted gene trees on three taxa. In this situation with only four (or three) taxa, all species tree methods are necessarily inconsistent: a coalescent model ignoring gene flow would necessarily favor one of the two incorrect four-taxon unrooted trees (or three-taxon rooted trees), when the true species tree is the one that is supported by the least proportion of genes. We further considered a situation with more taxa, and tested the accuracy of two widely used and fast species tree methods—one based on quartets (ASTRAL) and one using full unrooted gene trees (NJst). Both were found to be inconsistent if gene flow was severe enough. Concatenation was inconsistent as well.

Qualitatively, we identified an anomaly zone when a speciation event is very rapidly followed by directional gene flow into only one of the two descendant populations, which then again speciates very shortly after into two sister species, A and B. If this second speciation occurs very rapidly, alleles do not have time to sort after gene flow, so two alleles sampled from A and B may frequently originate from different parental populations: one inherited vertically and the other allele inherited by gene flow. The combination of gene flow followed by ILS can cause A and B to be non-monophyletic in gene trees. This discordance is exacerbated if gene flow occurred rapidly after a first speciation, as ILS would also affect gene trees in which both alleles originated from the same parental population. If this pattern occurs across many genes, species tree methods tend to infer a species tree in which

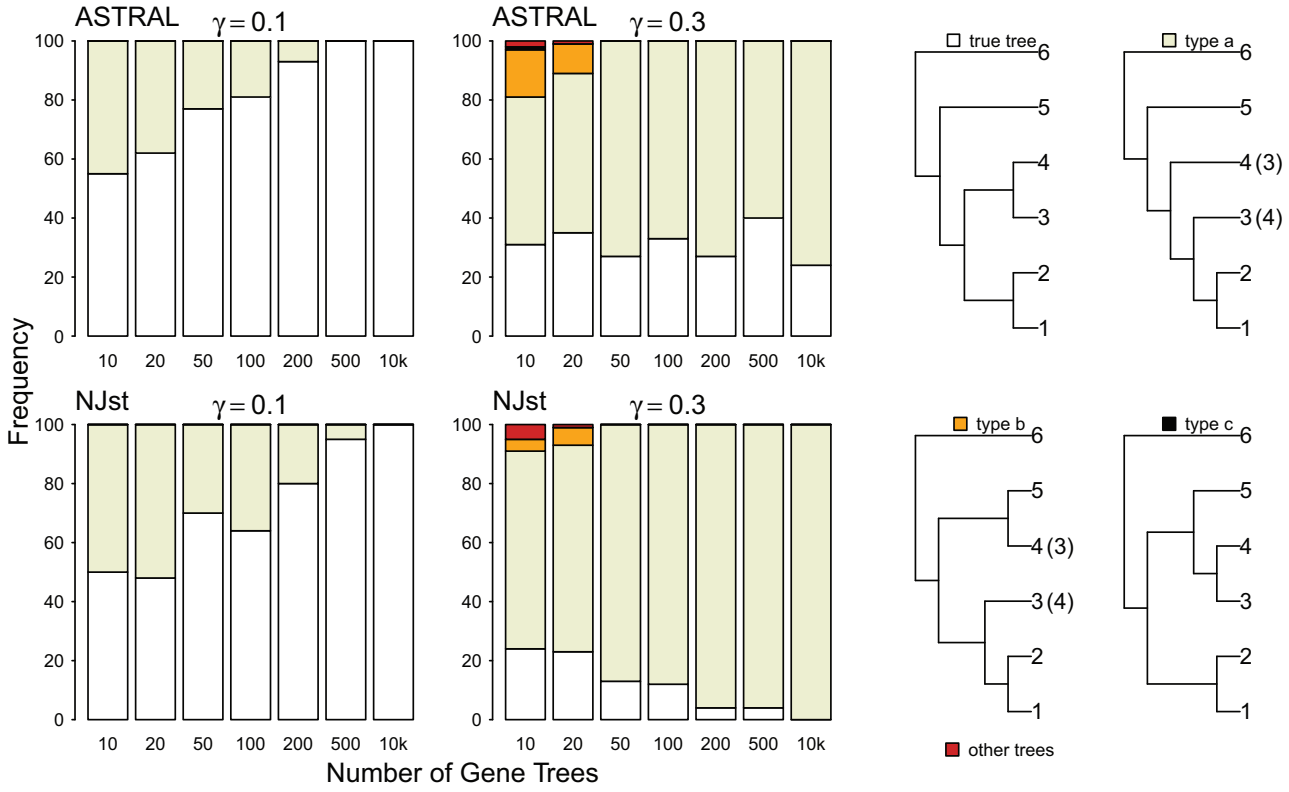

FIGURE 5. Trees inferred by ASTRAL and NJst from true gene trees under ILS and gene flow (Fig. 3). Both methods estimate an unrooted species tree, which was rooted by the outgroup (6). Inferred trees were classified into five categories: the true species tree, trees showing an anomalous quartet grouping species 3 or 4 with clade (1, 2) (trees of type a and b), the alternate tree displayed by the network in Figure 3 with (3, 4) sister to species 5, and all other trees. Bars show the proportion of replicates in which the inferred species tree belonged in each category. For both ASTRAL and NJst, the majority of estimated trees were of type a for $\gamma = 0.3$.

the sister species that radiated shortly after gene flow are not monophyletic.

Networks are Needed to Account for Gene Flow

Coalescent-based methods have been widely used to reconstruct species trees from a set of discordant gene trees. This discordance is modeled probabilistically by the coalescent to account for ILS. However, the assumption that ILS alone caused discordance in the underlying gene trees is very restrictive. The presence of gene flow is now supported by a large body of evidence, at all levels in the tree of life (e.g., Cui et al. 2013; Jónsson et al. 2014; Clark and Messer 2015; Fontaine et al. 2015; Gallus et al. 2015).

Just as concatenation was found to be inconsistent to the presence of ILS, species tree methods, likewise, are shown here to be inconsistent when the assumption of only one source of gene tree discordance (ILS) is violated. More work is needed to fully characterize the region of inconsistency, however. In our simulation settings, species tree methods were accurate under low levels of gene flow. Inconsistency was only observed under high levels of gene flow. In empirical studies, discrepancies between trees obtained by concatenation and by coalescent-based methods have historically been

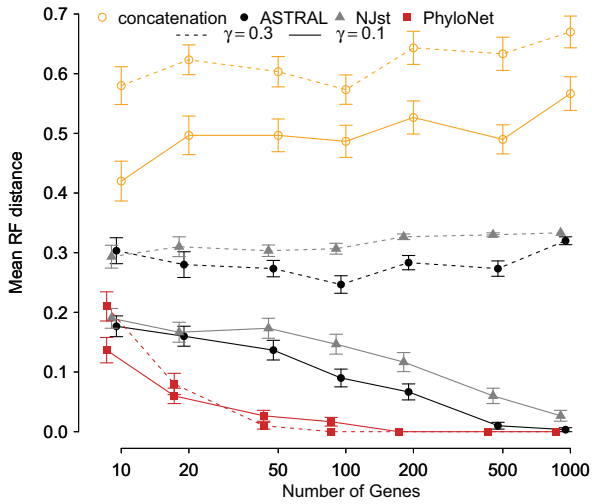

FIGURE 6. Accuracy of concatenation (with RAxML), ASTRAL, NJst, and PhyloNet as in Figure 4, but from sequence alignments. For PhyloNet, the estimated species tree was extracted from the estimated network by keeping the major hybrid edge (with $\hat{\gamma} > 0.5$) and by suppressing the minor hybrid edge (with $\hat{\gamma} < 0.5$).

explained by the presence of ILS (Song et al. 2012; Mirarab et al. 2014c; Warnow 2015). However, these observed discrepancies might also be caused by other

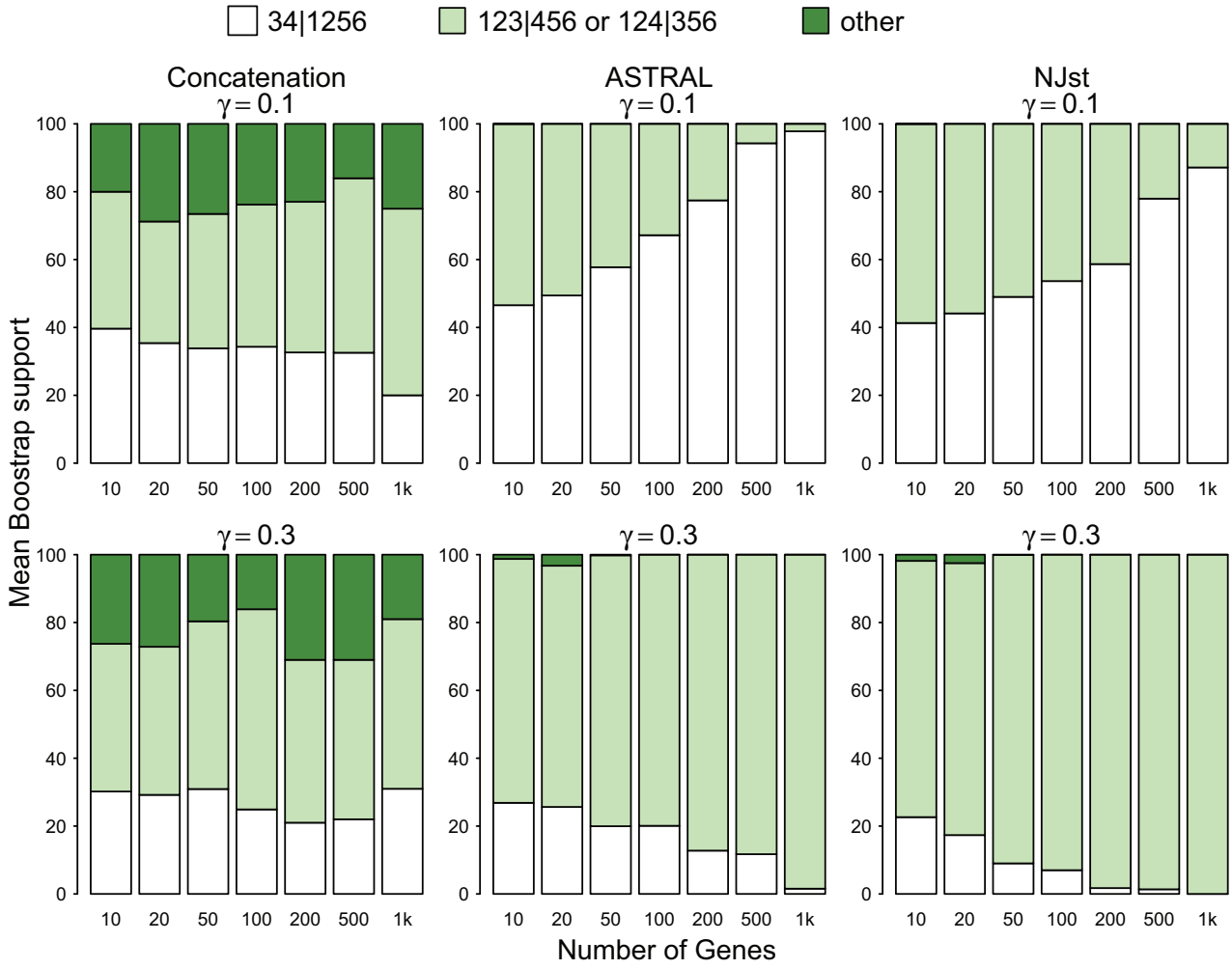

FIGURE 7. Bootstrap support of clades in species trees inferred by concatenation, ASTRAL and NJst from sequence alignments under ILS and gene flow (Fig. 3), averaged across simulation replicates. Estimated species trees were rooted with the outgroup (6). Bootstrap support was calculated for the true bipartition (34|1256) and for two alternate bipartitions splitting 3 and 4 (123|456 and 124|356), as in trees of types a and b in Figure 5.

processes including gene tree estimation error (Mirarab et al. 2014a; Springer and Gatesy 2016) or by gene flow. More theoretical work is needed to determine how much gene flow is necessary to cause inconsistency, but that is beyond the scope of the present work (but see Daskalakis and Roch 2015).

The inconsistency of species tree methods is a strong reason to use methods that explicitly account for gene flow. To do so, we need to shift to a network paradigm. Networks explicitly add gene flow events onto a species tree. This paradigm shift is not easy, because a network displays several trees. We propose here to consider networks with inheritance probabilities, so that each reticulate node can be attributed a "major" parent edge from which more than 50% of genes originated. These edges identify one "major tree" displaying the major vertical inheritance pattern, true for a majority of the genome. Awareness about the need for explicit phylogenetic networks has increased recently, to better explain evolutionary histories at various levels (organismal vs. molecular lineages) (Bapteste et al. 2013; Mindell 2013; Morrison 2014). The present work shows that, even for the purpose of finding the major tree-like pattern, accounting for gene flow can be necessary.

In recent years, there has been an explosion of methods to reconstruct phylogenetic networks from different sources of data (e.g., Than et al. 2008; Kubatko 2009; Meng and Kubatko 2009; Huson et al. 2010; Yu et al. 2012, 2014; Grünewald et al. 2013; Yang et al. 2014; Yu and Nakhleh 2015; Solís-Lemus and Ané 2016). However, for accurate estimation it is best to utilize probabilistic methods that account for ILS and gene tree estimation error, otherwise extra gene flow events need to be invoked to explain discordance caused by gene tree error or ILS. In our simulations, we used the maximum likelihood method in PhyloNet, modeling both ILS and gene flow through reticulation edges. PhyloNet was more accurate than species tree methods, even with only a few genes, and especially at high levels of gene flow. Unfortunately, modeling the coalescent

with gene flow has a heavy computational cost. On 6 taxa and 1000 genes, ASTRAL and NJst took 0.22 and 0.4 s on average, compared to over 9 min for PhyloNet, making the bootstrap challenging. Unlike concatenation, ASTRAL or NJst, PhyloNet's running time explodes very quickly with more taxa, taking over 170 h on average, for instance, on 10 taxa, 100 genes and 2 reticulations (Solís-Lemus and Ané 2016). It is also imperative to use explicit networks as opposed to implicit networks (like split networks), despite the computational advantage offered by fast implicit network approaches (Than et al. 2008; Huson et al. 2010; Grünewald et al. 2013; Yang et al. 2014). In implicit networks, nodes do not represent ancestral species, making biological interpretation difficult. Therefore, shifting from a species tree to a species network paradigm is not easy in practice. The computational cost of accurate network estimation methods has hampered their adoption.

Explicit Network Models

Kubatko (2009; see also Meng and Kubatko 2009; Gerard et al. 2011) and Yu et al. (2014; see also Yu et al. 2011, 2012) propose two different probability models to simultaneously account for ILS and gene flow. Both models are based on the multispecies coalescent so they share assumptions with the standard species tree methods, such as unlinked loci, no recombination within loci, and constant ancestral population sizes for methods using branch lengths in gene trees.

These two network models differ on one key assumption: how multiple alleles at a given locus trace back to one or the other parent of a reticulation node. In Kubatko (2009) (and implemented in STEM-hy), all alleles at a given locus are assumed to originate from the same parental species. In other words, all gene lineages at a given locus evolve on the same "parental" species tree, obtained by choosing one reticulate edge from the network and dropping the alternate parental edge, at each reticulation node. Different loci may evolve along different parental species trees, independently of each other, and with probabilities determined by the γ inheritance values. Thus, the likelihood of a species network can be expressed as a linear combination of likelihoods from all the possible parental species trees under the coalescent. In contrast, Yu et al. (2014) consider a model (implemented in PhyloNet) where all alleles at a given locus need not originate from the same parental species. To allow for this flexibility, different alleles are assumed to trace back through a given parent edge with the edge's inheritance probability γ, and independently of each other. This assumption complicates the likelihood calculation of the network, which cannot be obtained from that of parental species trees. Instead, Yu et al. (2012) use the coalescent on trees having multiple leaves labeled by the same species ("MUL-trees") and a mapping of alleles onto this tree.

Because this second model (Yu et al. 2014) allows for more flexibility, it appears to be more relevant biologically. The difference between the two models is relevant only when alleles at a particular locus do not coalesce more recently than the introgression. This can happen if multiple individuals are sampled from a species that received genetic material from a donor population or if gene flow was followed by speciation, and if the different alleles from the different daughter species or individuals did not have sufficient time to sort. Following these alleles back in time, they have not had time to coalesce until the gene flow event. At this point, it is then natural to assume that these alleles were sampled at random from the ancestral, admixed population. If γ is the proportion of migrant individuals from the donor population, then it is reasonable to assume that each allele comes from the donor population with probability γ, independently of the other alleles. Non-independence between allele origins might result from selection, but this would probably affect only a small proportion of loci. The appearance of anomalous gene trees on four taxa is contingent on this model with independent parental origins of multiple alleles at the same locus. Hence we find it important to draw attention to the biological interpretation of this coalescent network model.

Further Challenges

There are many other biological processes that lead to gene tree variation, in addition to ILS and processes modeled by a network (like gene flow, introgression hybridization, or HGT). Gene duplication and loss, for example, are typically ignored by coalescent-based methods (but see Rasmussen and Kellis 2012; To and Scornavacca 2015). Population structure prior to speciation can also lead to anomalous rooted gene trees on three taxa (Slatkin and Pollack 2008), but population structure is ignored by current phylogenetic network models. Also, HGT might be too frequent and too widespread to be efficiently modeled by a network. Bacterial networks, for example, might be so complex in reality that they we might not be able to infer them accurately and their full complexity might not even be identifiable (Pardi and Scornavacca 2015).

A network model may provide a good representation of highways of gene transfer (Beiko et al. 2005; Bansal et al. 2013), but an additional process might also be needed to model diffuse HGT events, that each affected only a handful of genes and that collectively affect all parts of the species tree or species network. Szöllosi et al. (2012) describe such a model with branch-specific rates of gene transfer applying to sampled species, and with global rates applying to sampled and unsampled species in Szöllosi et al. (2013). As reviewed in Szöllosi et al. (2015), simultaneously accounting for several of these biological processes is extremely challenging. Nonetheless, our work shows that doing so is necessary. Coalescent-based network methods accounting for both ILS and gene flow are a great step toward a unified, more robust framework. A current challenge is to make these methods more scalable to larger data sets, and to incorporate other biological processes.

Funding

This work was supported in part by the National Science Foundation [DEB 0936214 and DEB 1354793].

Acknowledgments

We want to thank David Baum for insightful discussions on the biological realism of gene flow models.

References

Bansal M.S., Banay G., Harlow T.J., Gogarten J.P., Shamir R. 2013. Systematic inference of highways of horizontal gene transfer in prokaryotes. Bioinformatics 29:571–579.

Bapteste E., van Iersel L., Janke A., Kelchner S., Kelk S., McInerney J.O., Morrison D.A., Nakhleh L., Steel M., Stougie L., Whitfield J. 2013. Networks: expanding evolutionary thinking. Trends Genet. 29:439–441.

Beiko R.G., Harlow T.J., Ragan M.A. 2005. Highways of gene sharing in prokaryotes. Proc. Natl Acad. Sci. USA 102:14332–14337.

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.

Clark A., Messer P. 2015. Conundrum of jumbled mosquito genomes. Science 347:27–28.

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

Daskalakis C., Roch S. 2015. Species trees from gene trees despite a high rate of lateral genetic transfer: a tight bound. arXiv:1508.01962.

Degnan J.H. 2013. Anomalous unrooted gene trees. Syst. Biol. 62: 574–590.

Fontaine M., Pease J., Steele A., Waterhouse R., Neafsey D., Sharakhov I., Jiang X., Hall A., Catteruccia F., Kakani E., Mitchell S., Wu Y.-C., Smith H., Love R., Lawniczak M., Slotman M., Emrich S., Hahn M., Besansky N. 2015. Highly evolvable malaria vectors: the genomes of 16 anopheles mosquitoes. Science 347:12585241–12585246.

Gallus S., Janke A., Kumar V., Nilsson M.A. 2015. Disentangling the relationship of the Australian marsupial orders using retrotransposon and evolutionary network analyses. Genome Biol. Evol. 7:985–992.

Gerard D., Gibbs H.L., Kubatko L. 2011. Estimating hybridization in the presence of coalescence using phylogenetic intraspecific sampling. BMC Evol. Biol. 11:291.

Grünewald S., Spillner A., Bastkowski S., Bögershausen A., Moulton V., Grünewald S., Bögershausen A. 2013. SuperQ: computing supernetworks from quartets. IEEE/ACM Trans. Comput. Biol. Bioinform. 10:151–160.

Huang H., He Q., Kubatko L.S., Knowles L.L. 2010. Sources of error inherent in species-tree estimation: impact of mutational and coalescent effects on accuracy and implications for choosing among different methods. Syst. Biol. 59:573–583.

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

Huson D., Rupp R., Scornavacca C. 2010. Phylogenetic networks. 1st ed. New York: Cambridge University Press.

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. USA 111:18655–18660.

Knowles L.L., Kubatko L.S. 2010. Estimating species trees: practical and theoretical aspects. 1st ed. Hoboken, NJ: Wiley-Blackwell.

Kubatko L.S. 2009. Identifying hybridization events in the presence of coalescence via model selection. Syst. Biol. 58:478–488.

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

Lanier H.C., Knowles L.L. 2012. Is recombination a problem for species-tree analyses? Syst. Biol. 61:691–701.

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

Meng C., Kubatko L.S. 2009. Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. Theoret. Popul. Biol. 75:35–45.

Mindell D.P. 2013. The tree of life: metaphor, model, and heuristic device. Syst. Biol. 62:479–489.

Mirarab S., Bayzid M.S., Boussau B., Warnow T., Bayzid S., Boussau B., Warrow T. 2014a. Statistical binning enables an accurate coalescent-based estimation of the avian tree. Science 346:12504631–1250463–9.

Mirarab S., Bayzid M.S., Warnow T. 2014b. Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. Syst. Biol. 0:1–15.

Mirarab S., Reaz R., Bayzid M.S., Zimmermann T., Swenson M.S., Warnow T. 2014c. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30:i541–i548.

Morrison D.A. 2014. Is the tree of life the best metaphor, model, or heuristic for phylogenetics? Syst. Biol. 63:628–638.

Pardi F., Scornavacca C. 2015. Reconstructible phylogenetic networks: do not distinguish the indistinguishable. PLoS Comput. Biol. 11:e1004135.

Rambaut A., Grassly N. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13:235–238.

Rasmussen M.D., Kellis M. 2012. Unified modeling of gene duplication, loss, and coalescence using a locus tree. Genome Res. 22: 755–765.

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

Slatkin M., Pollack J.L. 2008. Subdivision in an ancestral species creates asymmetry in gene trees. Mol. Biol. Evol. 25:2241–2246.

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

Song S., Liu L., Edwards S.V., Wu. S.2012. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proc. Natl Acad. Sci. USA 109: 14942–14947.

Springer M.S., Gatesy J. 2016. The gene tree delusion. Mol. Phylogenet. Evol. 94:1–33.

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

Szöllosi G.J., Boussau B., Abby S.S., Tannier E., Daubin V. 2012. Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. Proc. Natl Acad. Sci. USA 109:17513–17518.

Szöllosi G.J., Tannier E., Daubin V., Boussau B. 2015. The inference of gene trees with species trees. Syst. Biol. 64:e42–e62.

Szöllosi G.J., Tannier E., Lartillot N., Daubin V. 2013. Lateral gene transfer from the dead. Syst. Biol. 62:386–397.

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

To T.-H., Scornavacca C. 2015. Efficient algorithms for reconciling gene trees and species networks via duplication and loss events. BMC Genom. 16:1–14.

Warnow T. 2015. Concatenation analyses in the presence of incomplete lineage sorting. PLoS Curr. 7.

Whelan N. 2011. Species tree inference in the age of genomics. Trends Evol. Biol. 3:5.

Yang J., Grünewald S., Xu Y., Wan X.-F. 2014. Quartet-based methods to reconstruct phylogenetic networks. BMC Syst. Biol. 8:21.

Yu Y., Degnan L., Nakhleh L. 2012. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. PLoS Genet. 8:e1002660.

Yu Y., Dong J., Liu K.J., Nakhleh L. 2014. Maximum Likelihood Inference of Reticulate Evolutionary Histories. Proc. Natl Acad. Sci. USA 111:16448–16453.

Yu Y., Nakhleh L. 2015. A maximum pseudo-likelihood approach for phylogenetic networks. BMC Genom. 16:S10.

Yu Y., Than C., Degnan J.H., Nakhleh L. 2011. Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. Syst. Biol. 60: 138–149.