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. 2014b, 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 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., Degnan 2013), here we raise the question of whether species tree methods are consistent when ILS is 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 \( y \) and \( 1 - y \) (Fig. 1, left). In other words, \( y \) summarizes gene flow across the genome, as the proportion of genes inherited through reticulation. The main underlying tree is obtained by suppressing edges with \( y < 0.5 \) (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 six numerical parameters: \( (y, t_1, t_2, t_3, t_4, t_5) \). 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 \( (y < 0.5) \). 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 \( y \) and both are greater than the CF of the major quartet when \( 0.212 < y < 0.876 \), creating an anomaly zone (Fig. 2). These quartet CFs are given by (see Solís-Lemus
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, CA\(B\), or the minor triplets conflicting with the species tree, CA\(B\) or AB\(A\). Their expected frequencies are

\[
\text{CF}_{\text{CA}B} = (1-\gamma)^2(1-2/3e^{-\gamma})+2\gamma(1-\gamma)(1-e^{-\gamma})+1/3e^{-\gamma} - e^{-\gamma})+2\gamma(1-2/3e^{-\gamma} - e^{-\gamma})
\]

and

\[
\text{CF}_{\text{CB}A} = \text{CF}_{\text{CA}B} = (1-\gamma)^2(2/3e^{-\gamma})+2\gamma(1-\gamma)(1-e^{-\gamma})+1/3e^{-\gamma} - e^{-\gamma})+2\gamma(1-2/3e^{-\gamma} - e^{-\gamma})
\]

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 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.

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 more genes, on average. Surprisingly, the alternate tree (Fig. 6) fared a little better, with an incorrect species tree reconstruction (of type b) 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 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, $x$ was drawn uniformly in (1, 4) and $\theta$ 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 $\alpha$ 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,

**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-taxa unrooted trees (or three-taxa 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
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; Jonsson 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 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
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 (organisinal 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 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. Szöllösí 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 Szöllösí et al. (2013). As reviewed in Szöllösí 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.
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


