Co-estimating Reticulate Phylogenies and Gene Trees from Multi-locus Sequence Data

Dingqiao Wen and Luay Nakhleh[1]
Department of Computer Science, Rice University, 6100 Main Street, Houston, TX 77005, USA.
[1]To whom correspondence should be addressed; E-mail: nakhleh@rice.edu.

Abstract

The multispecies network coalescent (MSNC) is a stochastic process that captures how gene trees grow within the branches of a phylogenetic network. Coupling the MSNC with a stochastic mutational process that operates along the branches of the gene trees gives rise to a generative model of how multiple loci from within and across species evolve in the presence of both incomplete lineage sorting (ILS) and gene flow. We report on the first Bayesian method for sampling the parameters of this generative model, including the species phylogeny, gene trees, divergence times, and population sizes, from DNA sequences of multiple independent loci. We demonstrate the utility of our method by analyzing simulated data and reanalyzing three biological data sets. Our results demonstrate the significance of not only co-estimating species phylogenies and gene trees, but also accounting for gene flow and ILS simultaneously. In particular, we show that when gene flow occurs, our method accurately estimates the evolutionary histories, coalescence times, and divergence times. Methods that do not account for gene flow, on the other hand, underestimate divergence times and overestimate coalescence times.

The availability of sequence data from multiple loci across the genomes of species and individuals within species is enabling accurate estimates of gene and species evolutionary histories, as well as parameters such as divergence times and ancestral population sizes (Rannala and Yang, 2003). Several statistical methods have been developed for obtaining such estimates (Rannala and Yang, 2003; Edwards et al., 2007; Heled and Drummond, 2010; Bouckaert et al., 2014). All these methods employ the *multispecies coalescent* (Degnan and Rosenberg, 2009) as the stochastic process that captures the relationship between species trees and gene genealogies.

As evidence of hybridization (gene flow between different populations of the same species or across different species) continues to accumulate (Rieseberg, 1997; Arnold, 1997; Barton, 2001; Koonin et al., 2001; Gogarten et al., 2002; Mallet, 2005, 2007), there is a pressing need for statistical methods that infer species phylogenies, gene trees, and their associated parameters in the presence of hybridization. We recently introduced for this purpose the *multispecies network coalescent* (MSNC) along with a maximum likelihood search heuristic (Yu et al., 2014) and a Bayesian sampling technique (Wen et al., 2016b). However, these methods use gene tree estimates as input. Using these estimates, instead of using the sequence data directly, has at least three drawbacks. First, the sequence data allows for learning more about the model than gene tree estimates (Rannala and Yang, 2003). Second, gene tree estimates could well include erroneous information, resulting in wrong inferences (DeGiorgio and Degnan, 2014; Wen et al., 2016b). Third, co-estimating the species phylogeny and gene trees results in better estimates of the gene trees themselves (DeGiorgio and Degnan, 2014; Zimmermann et al., 2014).

We report here on the first statistical method for co-estimating species (or, population) phylogenies and gene trees along with parameters such as ancestral population sizes and divergence times using DNA sequence alignments from multiple independent loci. Our method utilizes a two-step generative process (Fig. 1) that links, via latent variables that correspond to local gene genealogies, the sequences of multiple, unlinked loci from across a set of genomes to the phylogenetic network Nakhleh (2010) that models the evolution of the genomes themselves.

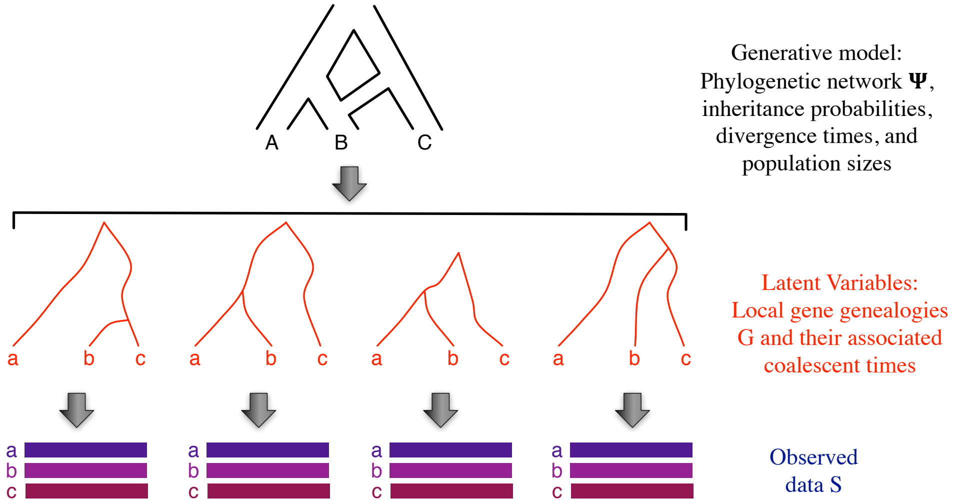

Figure 1: From a phylogenetic network to multi-locus sequences via latent gene genealogies. The multi-species network coalescent (Yu et al., 2014) is a stochastic process that defines a probability distribution on gene genealogies along with their coalescent times. The parameters of the process consist of a phylogenetic network topology, inheritance probabilities, divergence times, and population sizes. Each gene genealogy, when coupled with model of sequence evolution, defines a probability distribution on sequence alignments.

Our method consists of a reversible-jump Markov chain Monte Carlo (RJMCMC) sampler of the posterior of this generative process. In particular, our method co-estimates, in the form of posterior samples, the phylogenetic network and its associated parameters for the genomes as well as the local genealogies for the individual loci. We demonstrate the performance of our method on simulated data. Furthermore, we analyze three biological data sets, and discuss the insights afforded by our method. In particular, we find that methods that do not account, wrongly, for gene flow in the data tend to underestimate divergence times of the species or populations and overestimate the coalescent times of individual gene genealogies. Our method, on the other hand, estimates both the divergence times and coalescent times with high accuracy. Furthermore, we demonstrate that coalescent times are much more accurately estimated when the estimation is done simultaneously with the phylogenetic network than when the estimation is done in isolation.

As the model underlying out method extends the multispecies coalescent to cases that include gene flow, our method is applicable to data from different sub-populations, not only different species, and to data where more than one individual per species or sub-population is sampled. The method is implemented and publicly available in the PhyloNet software package (Than et al., 2008).

Model

The data in our case is a set $\mathscr{S} = \{S_1, \ldots, S_m\}$ where S_i is a DNA sequence alignment from locus i (the bottom part in Fig. 1). A major assumption is that there is no recombination within any of the m loci, yet there is free recombination between loci. The model $\mathscr{M}$ consists of a phylogenetic network Ψ (the topology, divergence times, and population sizes) and a vector of inheritance probabilities Γ (the top part in Fig. 1). The topology of a phylogenetic network is a rooted, directed, acyclic graph, whose leaves are labeled by the taxa under study. Every node in the network has at most two parents, and nodes with two parents are called reticulation nodes. Associated with every internal node of the phylogenetic network is a divergence time parameter (the leaves are all assumed to be at time 0). Associated with every branch of the network, including one incident into the root, is a population size parameter. Furthermore, associated with the branches coming into reticulation nodes are the inheritance probabilities given by Γ. See Methods for a formal definition.

The posterior of the model is given by

$$p(\mathscr{M}|\mathscr{S}) \propto p(\mathscr{S}|\mathscr{M})p(\mathscr{M}) = p(\mathscr{M}) \prod_{i=1}^{m} \int_G p(S_i|g)p(g|\mathscr{M})dg, \tag{1}$$

where the integration is taken over all possible gene trees (the middle part in Fig. 1). The term $p(S_i|g)$ gives the gene tree likelihood, which is computed using Felsenstein's algorithm (Felsenstein, 1981) assuming a model of sequence evolution, and $p(g|\mathscr{M})$ is the probability density function for the gene trees, which was derived for the cases of species tree and species network in (Rannala and Yang, 2003) and (Yu et al., 2014), respectively; see Methods.

The integration in Eq. (1) is computationally infeasible except for very small data sets. Furthermore, in many analyses, the gene trees for the individual loci are themselves a quantity of interest. Therefore, to obtain gene trees, we sample from the posterior as given by

$$p(\Psi, \Gamma, G|S) \propto p(\mathscr{M}) \prod_{i=1}^{m} p(S_i|g_i)p(g_i|\mathscr{M}) = p(\Psi)p(\Gamma) \prod_{i=1}^{m} p(S_i|g_i)p(g_i|\Psi, \Gamma), \tag{2}$$

where $G = (g_1, \ldots, g_m)$ is a vector of gene trees, one for each of the m loci. This co-estimation approach is adopted by the two popular Bayesian methods $*$BEAST (Heled and Drummond, 2010) and BEST (Liu, 2008), both of which co-estimate species trees (hybridization is not accounted for) and gene trees.

To fully specify the co-estimation given by Eq. (2), two priors $p(\Psi)$ and $p(\Gamma)$ need to be defined. For phylogenetic network Ψ, we denote by Ψ_{ret}, Ψ_{top}, Ψ_τ, and Ψ_θ its number of reticulation nodes, topology, divergence times, and population sizes, respectively. We have

$$p(\Psi|\nu, \delta, \eta, \psi) = p(\Psi_{ret}|\nu) \times p(\Psi_{top}|\Psi_{ret}, \Psi_\tau, \eta) \times p(\Psi_\tau|\delta) \times p(\Psi_\theta|\psi), \tag{3}$$

where ν, δ, η, and ψ are hyper-parameters. For the inheritance probabilities Γ, we use a uniform prior on $[0, 1]$, though a Beta distribution would also be appropriate in general cases; see Methods for full details.

As computing the posterior distribution given by Eq. (2) is computationally intractable, we implement a Markov chain Monte Carlo (MCMC) sampling procedure based on the Metropolis-Hastings algorithm. In each iteration of the sampling, a new state (Ψ', Γ', G') is proposed and either accepted or rejected based on the Metropolis-Hastings ratio r that is composed of the likelihood, prior, and Hastings ratios. When the proposal changes the dimensionality of the sample by adding a new reticulation to or removing an existing reticulation from the phylogenetic network, the absolute value of the determinant of the Jacobian matrix is also taken into account, which results in a reversible-jump MCMC, or RJMCMC (Green, 1995, 2003).

Our sampling algorithm employs three categories of moves: One for sampling the phylogenetic network and its parameters, one for sampling the inheritance probabilities, and one for sampling the gene trees. To propose a new state of the Markov chain, one element from $(\Psi, \gamma_1, \ldots, \gamma_{\Psi_{ret}}, g_1, \ldots, g_m)$ is selected at random, then a move from the corresponding category is applied. The workflow, design and full derivation of the Hastings ratios of the moves are given in the SI.

Results

We implemented our method in PhyloNet (Than et al., 2008), a publicly available, open-source software package for phylogenetic network inference and analysis. We studied the performance of the method on synthetic data and revisited the analyses of three biological data sets, as we now describe.

Performance on simulated data

For the synthetic data, we generated sequence alignments of varying numbers of loci by simulating their evolution within the branches of the phylogenetic network of Fig. 2**A**. The network topology and associated parameters are inspired by the phylogenetic network of the mosquito data set in (Wen et al., 2016a); full details of the simulation setup are given in the SI. A few observations are in order. First, while ∗BEAST is not designed to deal with hybridization, it inferred the tree topology (Fig. 2**B**) that is obtainable by removing the two hybridization events (the two arrows) from the true phylogenetic network. Second, our method identified the true phylogenetic network as the one with the highest posterior (2**C**). Furthermore, the estimated inheritance probabilities are very close to the true ones. Third, since ∗BEAST does not account for hybridization, it accounts for all gene tree heterogeneity as being caused by incomplete lineage sorting (ILS) by underestimating all branch lengths. Indeed, 2**D** shows that the minimum coalescent times of the co-estimated gene trees by ∗BEAST force the divergence times in the inferred species tree to be very

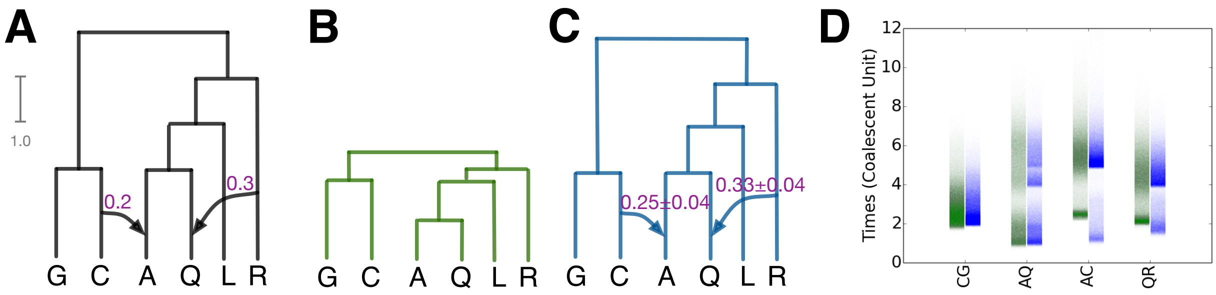

Figure 2: Results of ∗BEAST and our method on simulated data of 128 loci each of length 500 nucleotides. (**A**) The true phylogenetic network with the shown inheritance probabilities. (**B**) The MPP (maximum a posteriori probability) species tree estimated by ∗BEAST (94%) with the average divergence times. (**C**) The MPP phylogenetic network along with the inheritance probabilities estimated by our method. The scale bar of divergence times represents 1 coalescent unit for (**A-C**). (**D**) The coalescent times of the MRCAs of (C,G), (A,Q), (A,C), (Q,R) from co-estimated gene trees inferred by ∗BEAST (green) and our method (blue).

low. Our method, on the other hand, accurately estimates the branch lengths of the inferred phylogenetic network since networks differentiate between divergence and migration times. For example, 2**D** shows that the coalescent times of clade (C,G) across all co-estimated gene trees is a continuum with a minimum value around 2, which defines the divergence time of these two taxa in the phylogenetic network. Our method clearly identifies two groups of coalescent times for each of the two clades (A,C) and (Q,R): The lower group of coalescent times correspond to hybridization, while the upper group of coalescent times correspond to the coalescences above the respective MRCAs of the clades. We also note that the minimum value of coalescent times corresponding to (Q,R) is larger than that corresponding to (A,Q), which correctly reflects the fact that hybridization from R to Q happened before hybridization from C to A, as indicated in the true phylogenetic network. Finally, for clade (A,Q), three groups of coalescence times are identified by our method, which makes sense since there are three common ancestors of A and Q in the network: at the MRCA of (A,Q) in the case of no hybridization involving either of the two taxa, at the MRCA of (A,Q,L,R) in the case of the hybridization involving Q, and at the root of the network in the case of the hybridization involving A.

More generally, we inspected the trace plots of our method on the posterior values to ensure that the MCMC chains converged and mixed well. We also evaluated the accuracy of the estimated gene trees by computing their Robinson-Foulds (RF) distances (Robinson and Foulds, 1981) and Normalized Rooted Branch Scores (nrBS) (Heled and Drummond, 2010; Kuhner and Felsenstein, 1994) against the true gene trees. Recent studies (DeGiorgio and Degnan, 2014; Zimmermann et al., 2014) showed that simultaneous inference of species trees and gene trees leads to more accurate gene tree estimates. The average RF distances between true gene trees and gene trees estimated by our method, $*$BEAST, and RAxML (Stamatakis, 2014) (in the case of RAxML, only gene trees are inferred) are 0.7, 0.9, and 1.0, respectively, further demonstrating the gains in gene tree topological accuracy from co-estimation. Furthermore, the nrBS values of our method were lower than those of *BEAST. The accuracy of inferences made by our method improved as more loci were used. It is important to note here that this analysis is not aimed at establishing the superiority of one method over the other. Our method and *BEAST make different assumptions about the processes at play and employ different models. The goal here is to demonstrate the utility of our method when the evolutionary history involves hybridization. When the evolutionary history is not reticulate, the performance of both methods is very similar. Full details of the simulation setup and results are given in SI.

Comparing our method to existing phylogenetic network inference methods (Yu et al., 2014; Wen et al., 2016b) that use gene tree estimates as input, our method not only estimates more parameters, such as divergence times and population sizes, but also estimates more accurate gene trees and phylogenetic networks. For example, when we fed the true gene trees simulated within the network of Fig. 2**A** on 16, 32, 64, and 128 loci to the Bayesian method of (Wen et al., 2016b), the proportions of the true phylogenetic network being sampled were 0%, 39%, 45%, and 60% for these numbers of loci, respectively. In other words, even when the true gene trees were used as input, inference from gene tree data requires larger numbers of loci to obtain accurate inferences than would be required for inference from sequence data.

In addition to the synthetic data, we analyzed a bread wheat genome data set from (Marcussen et al., 2014), a data set of seven Saccharomyces species from (Rokas et al., 2003), and an Anopheles mosquitoes (*An. gambiae* complex) data set from (Fontaine et al., 2015).

Analysis of a bread wheat data set

The bread wheat data set consists of three subgenomes of *Triticum aestivum*, TaA (A subgenome), TaB (B subgenome) and TaD (D subgenome), and five diploid relatives Tm (*T. monococcum*), Tu (*T. urartu*), Ash (*Ae. sharonensis*), Asp (*Ae. speltoides*) and At (*Ae. tauschii*). Marcussen *et al.* found that each of the A and

B lineages is more closely related to D than to each other, as represented by the phylogenetic network in Fig. 3**A** inferred using the parsimony approach of (Yu et al., 2011) given gene tree topologies of TaA, TaB, and TaD. Based on this network, they proposed an evolutionary history of *Triticum aestivum*, where about 7 million years ago the A and B genomes diverged from a common ancestor and 1∼2 million years later these genomes gave rise to the D genome through homoploid hybrid speciation (Marcussen et al., 2014).

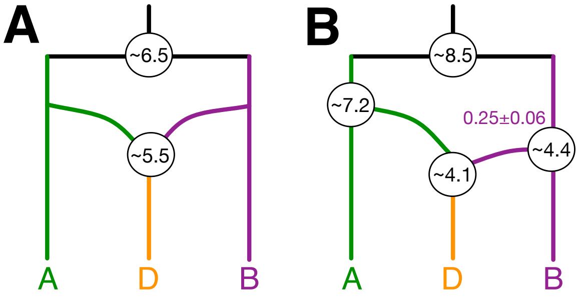

Figure 3: Phylogenetic history of the bread wheat data of (Marcussen et al., 2014). (**A**) The phylogenetic network inferred using the parsimony approach of (Yu et al., 2011) given gene tree topologies of TaA, TaB, and TaD. The times are estimated by gene tree analyses. (**B**) The phylogenetic network inferred by our method given 68 loci of eight genomes. The times at the internal nodes are in millions of years.

We fed 68 loci of the eight genomes into our method. The only network in the 95% credible set is identical to the one in (Marcussen et al., 2014), shown in Fig. 3**B** where A, B, and D represent ((TaA,Tu),Tm), (TaB,Asp), and ((TaD,At),Ash), respectively. Assuming a mutation rate of 1×10^{-9} per-site per-generation and 1 year per generation, a plausible evolutionary history posits that a common ancestor of A, B, and D started differentiation ∼8.5 Ma into(A,D) and B genome lineages. Subsequently, (A,D) speciated at ∼7 Ma into A and D lineages. The hybridization occurred around 4-4.5 Ma from B to D genome lineages. Although both proposed evolutionary histories contain one hybridization, phylogenetic networks with two or more reticulations were inferred on larger data sets by our method and by the authors of the original study (Marcussen et al., 2014). See SI for full details.

Analysis of a yeast data set

The yeast data set of (Rokas et al., 2003) consists of 106 loci from seven Saccharomyces species, *S. cerevisiae* (Scer), *S. paradoxus* (Spar), *S. mikatae* (Smik), *S. kudriavzevii* (Skud), *S. bayanus* (Sbay), *S. castellii* (Scas), *S. kluyveri* (Sklu). Rokas *et al.* (Rokas et al., 2003) reported on extensive incongruence of single-gene phylogenies and revealed the species tree from concatenation method (Fig. 4**A**). Edwards *et al.* (Edwards et al., 2007) reported as the two main species trees and gene tree topologies sampled from BEST (Liu, 2008) the two trees shown in Fig. 4**A-B**. The other gene tree topologies (Fig. 4**C**) exhibited weak phylogenetic signals among Sklu, Scas and the other species. Bloomquist and Suchard (Bloomquist and Suchard, 2010) reanalyzed the data set without Sklu since it added too much noise to their analysis. Their analysis resulted in many horizontal events between Scas and the rest of the species because the Scas lineage-specific rate variation is much stronger than that of the other species. Yu *et al.* (Yu et al., 2013a) analyzed the 106-locus data set restricted to the five species Scer, Spar, Smik, Skud, and Sbay and identified a maximum parsimony network that supports a hybridization from Skud to Sbay with inheritance probability of 0.38.

Analyzing the 106-locus data set using our method, the 95% credible set contains many topologies with

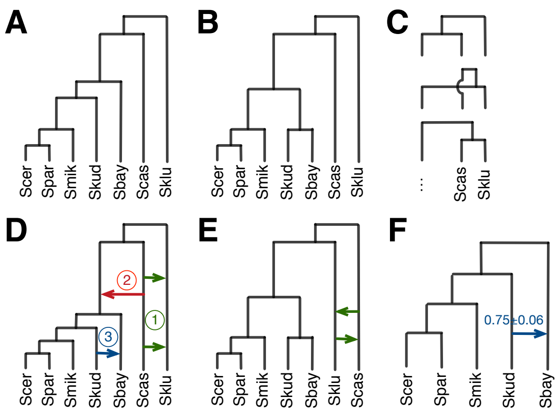

Figure 4: Results on the yeast data set of (Rokas et al., 2003). (**A**) The species tree inferred using the concatenation method (Rokas et al., 2003) and the main species tree and gene tree topology sampled using BEST (Edwards et al., 2007). (**B**) The second most frequently sampled species and gene tree topology by BEST (Edwards et al., 2007). (**C**) Many other gene tree topologies were sampled by BEST (Edwards et al., 2007), indicating weak phylogenetic signals among Sklu, Scas, and the rest of the species. (**D**) A representative phylogenetic network inferred by our method on all 106 loci. (**E**) A representative phylogenetic network inferred by our method on the 28 loci with strong phylogenetic signal (see SI). (**F**) The single phylogenetic network inferred using all 106 loci from the five species Scer, Spar, Smik, Skud, Sbay.

similar hybridization patterns; the representative network is shown in Fig. 4**D**. All the previous findings are encompassed by the networks inferred by our method. The two hybridizations between Sklu and Scas (green edges in 4**D**) indicate the weak phylogenetic signals among Sklu, Scas and the rest of the species. The hybridization from Scas to the other species except for Sklu (red edge in 4**D**) captures the stronger lineage-specific rate variation in Scas. Finally, the hybridization from Skud to Sbay (blue edge in 4**D**) resolves the incongruence between the two main species tree topologies in 4**A-B**.

We further investigated the phylogenetic signal in each locus by counting the number of internal branches in the 70% majority-rule consensus of 100 maximum likelihood bootstrap trees. We found that only 28 out of the 106 loci contain four internal branches, and no locus had a consensus tree with all five internal branches. A representative phylogenetic network in the 95% credible set given these 28 loci, shown in Fig. 4**E**, indicates the weak phylogenetic signals among Sklu, Scas and the rest of the species. We then analyzed the 106-locus data set restricted to the five species Scer, Spar, Smik, Skud, and Sbay. The phylogenetic signal in this data set is very strong—the consensus trees of 99 out of the 106 loci contain two internal branches. The MPP phylogenetic network in Fig. 4**F** contains the hybridization from Skud to Sbay, which is identical to the sub-network in 4**D**. See SI for full details. In summary, analysis of the yeast data set demonstrates the effect of phylogenetic signal in the individual loci on the inference and the care that must be taken when selecting loci of analysis of reticulate evolutionary histories.

Analysis of a mosquitoes data set

The Anopheles mosquitoes (*An. gambiae* complex) data set of (Fontaine et al., 2015) consists of genome alignment of *An. gambiae* (G), *An. coluzzii* (C), *An. arabiensis* (A), *An. quadriannulatus* (Q), *An. merus* (R) and *An. melas* (L). Fontaine *et al.* reported on extensive introgressions in the *An. gambiae* complex

(Fontaine et al., 2015). Gene tree analyses were performed to detect the donor, recipient and migration times of the reticulation edges. Three major introgressions were added to the species tree backbone recovered from the X chromosome, resulting in a plausible phylogenetic network. More recently, Wen *et al.* (Wen et al., 2016a) reanalyzed the data set using the maximum likelihood method of (Yu et al., 2014) and then using the Bayesian method of (Wen et al., 2016b) and provided new insights into the evolutionary history of the *An. gambiae* complex.

Fontaine et al. inferred gene trees on 50-kb genomic windows using maximum likelihood and tabulated and analyzed the frequencies of distinct gene tree topologies across the chromosomes. However, such large genomic windows are very likely to include recombination. Indeed, a simple comparison of the 70% majority-rule consensus of 100 maximum likelihood bootstrap trees on the entire window against individual trees inferred from smaller regions of the same window highlight this issue (see SI for details).

To avoid using such large genomic windows, we randomly sampled 228 1-kb regions from the X chromosome. We fed the 228-locus data set into ∗BEAST and our method. Our method produces a phylogenetic network with many reticulations on this data set. We assessed the phylogenetic signal in each locus by computing the number of internal branches in the 70% majority-rule consensus of 100 maximum likelihood bootstrap trees. We found that only 59 out of the 228 loci contain three internal branches, and no locus had a consensus tree with all four internal branches. Analyzing those 59 loci data set using our method, the 95% credible set contains three topologies grouping (C,G) with R and positing hybridization from A, Q, or (A,Q) to (C,G) with inheritance probability 0.33 (Fig. 5**B**). The MPP species tree inferred by ∗BEAST (Fig. 5**A**) groups (A,Q) with (C,G) to account for heterogeneity across loci by means of ILS alone. The divergence times of the MRCAs of (C,G), (A,Q), (A,Q,C,G), and (R,C,G) inferred by ∗BEAST are similar to those inferred by our method. The minimum coalescent times of clades (C,G), (A,Q), (A,Q,C,G) and (R,C,G) co-estimated by ∗BEAST (green) and our method (blue) in Fig. 5**C** further confirm this statement. However, ∗BEAST underestimates the divergence times of the MRCA of (R,L) (or the root) to reconcile the divergence times of (R,C,G). BEAST, which infers gene trees from sequences without regard to a species tree, significantly underestimates all the coalescent times (sandy brown bars in Fig. 5**C**).

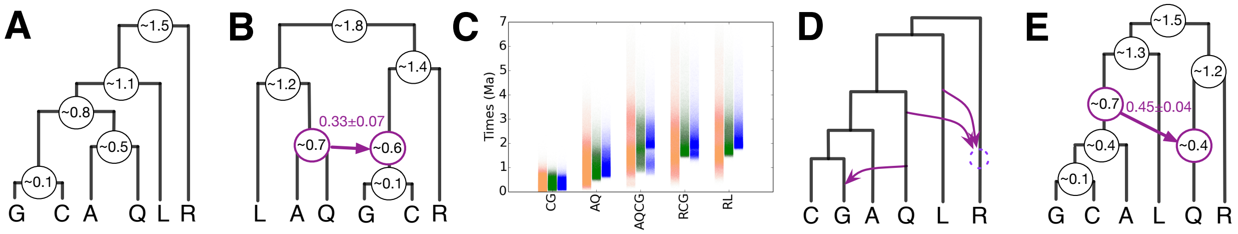

Figure 5: Results on the *An. gambiae* complex data of (Fontaine et al., 2015). (**A**) The MPP species tree inferred by ∗BEAST on regions with strong phylogenetic signal from the X chromosome. (**B**) The phylogenetic network inferred by our method on the same regions as in **A**. (**C**) The coalescent times of the MRCAs of (C,G), (A,Q), (A,Q,C,G), (R,C,G) and (R,L) from gene trees inferred by BEAST (sandy brown), ∗BEAST (green), and our method (blue) on the same regions as in **A**. (**D**) The phylogenetic network inferred by the Bayesian method of (Wen et al., 2016b) given gene tree topologies from 2791 regions with varying lengths of 1∼20-kb from the autosomes. (**E**) The phylogenetic network inferred by our method given 382 1-kb regions with strong phylogenetic signals from the autosomes.

For the autosomes, we randomly sampled 382 1-kb regions with strong phylogenetic signal and fed the data set into our method. The MPP phylogenetic network shown in Fig. 5**E** groups Q with R, and reveals hybridization from (A,C,G) to Q. This network can be embedded in the phylogenetic network inferred by the Bayesian method of (Wen et al., 2016b) given gene tree topologies from 2791 regions with varying lengths of 1$\sim$20-kb from the autosomes (Fig. 5**D**). Using data with strong phylogenetic signal would significantly reduce the complexity of the model. See SI for details.

Discussion

To conclude, we have devised a Bayesian framework for sampling the parameters of the MSNC model, including the species phylogeny, gene trees, divergence times, and population sizes, from sequences of multiple independent loci. Our work provides the first general framework for Bayesian phylogenomic inference from sequence data in the presence of hybridization. The method is publicly available in the open-source software package PhyloNet (Than et al., 2008). We demonstrate the utility of our method on simulated data and three biological data sets. Our results demonstrate several important aspects. First, ignoring hybridization when it had occurred results in underestimating the divergence times of species and overestimating the coalescent times of individual loci. Second, co-estimation of species phylogeny and gene trees results in more accurate gene tree estimates than the inferences of gene trees from sequences directly. Third, comparing to existing phylogenetic network inference methods (Yu et al., 2014; Wen et al., 2016b) that use gene tree estimates as input, our method not only estimates more parameters, such as divergence times and population sizes, but also estimates more accurate phylogenetic networks. Last but not the least, the phylogenetic signal in the individual loci on the inference must be taken into consideration when selecting loci of analysis of reticulate evolutionary histories. In particular, when there is low phylogenetic signal in the data, tree inference methods tend to result in unresolved trees. In the case of network methods, the counterpart to an unresolved tree is an overly complex network. In other words, while low signal is captured by a soft polytomy in trees, it is captured by multiple reticulations in networks. Therefore, it is very important that the signal in individual loci is carefully assessed in network inference, and indeed, in phylogenomics in general.

Finally, we identify several directions for further improvements of our proposed approach. First, while priors on species trees, such as the birth-death model, have been developed and employed by inference methods, similar prior distributions on phylogenetic networks are currently lacking. Second, while techniques such as the majority-rule consensus exist for summarizing the trees sampled from the posterior distribution, principled methods for summarizing sampled networks are needed. Last but not least, the sequence data used here, and in almost all phylogenomic analyses, consist of haploid sequences of randomly phased diploid genomes. The effect of random phasing on inferences in general needs to be studied in detail. Furthermore, the model could be extended to work directly on unphased data by integrating over possible phasings (Gronau et al., 2011).

Methods

Phylogenetic networks and their parameters

A *phylogenetic $\mathscr{X}$-network*, or $\mathscr{X}$-network for short, Ψ is a directed, acyclic graph (DAG) with $V(\Psi) = \{s, r\} \cup V_L \cup V_T \cup V_N$, where

- $indeg(s) = 0$ and $outdeg(s) = 1$ (s is a special node, that is the parent of the root node, r);
- $indeg(r) = 1$ and $outdeg(r) = 2$ (r is the *root* of Ψ);

- $\forall v \in V_L$, $indeg(v) = 1$ and $outdeg(v) = 0$ (V_L are the *external tree nodes*, or *leaves*, of Ψ);
- $\forall v \in V_T$, $indeg(v) = 1$ and $outdeg(v) \geq 2$ (V_T are the *internal tree nodes* of Ψ); and,
- $\forall v \in V_N$, $indeg(v) = 2$ and $outdeg(v) = 1$ (V_N are the *reticulation nodes* of Ψ).

The network's edges, $E(\Psi) \subseteq V \times V$, consist of *reticulation edges*, whose heads are reticulation nodes, *tree edges*, whose heads are tree nodes, and special edge $(s, r) \in E$. Furthermore, $\ell : V_L \to \mathscr{X}$ is the *leaf-labeling* function, which is a bijection from V_L to $\mathscr{X}$. Each node in $V(\Psi)$ has a species divergence time parameter and each edge in $E(\Psi)$ has an associated population size parameter. The edge $er(\Psi) = (s, r)$ is infinite in length so that all lineages that enter it coalesce on it eventually. Finally, for every pair of reticulation edges e_1 and e_2 that share the same reticulation node, we associate an inheritance probability, γ, such that $\gamma_{e_1}, \gamma_{e_2} \in [0, 1]$ with $\gamma_{e_1} + \gamma_{e_2} = 1$. We denote by Γ the vector of inheritance probabilities corresponding to all the reticulation nodes in the phylogenetic network (for each reticulation node, Γ has the value for one of the two incoming edge only).

Given a phylogenetic network Ψ, we use the following notation:

- Ψ_{top}: The leaf-labeled topology of Ψ; that is, the pair (V, E) along with the leaf-labeling ℓ.
- Ψ_{ret}: The number of reticulation nodes in Ψ. $\Psi_{ret} = 0$ when Ψ is a phylogenetic tree.
- Ψ_τ: The species divergence time parameters of Ψ. $\Psi_\tau \in (\mathbb{R}^+)^{|V(\Psi)|}$.
- Ψ_θ: The population size parameters of Ψ. $\Psi_\theta \in (\mathbb{R}^+)^{|E(\Psi)|}$

We use Ψ to refer to the topology, species divergence times and population size parameters of the phylogenetic network.

It is often the case that divergence times associated with nodes in the phylogenetic network are measured in units of years, generations, or coalescent units. On the other hand, branch lengths in gene trees are often in units of expected number of mutations per site. We convert estimates back and forth between units as follows:

- Given divergence time in units of expected number of mutations per site τ, mutation rate per site per generation μ and the number of generations per year g, $\tau/\mu g$ represents divergence times in units of years.
- Given population size parameter in units of population mutation rate per site θ, $2\tau/\theta$ represents divergence times in coalescent units.

The likelihood function

Felsenstein (Felsenstein, 1981) introduced a pruning algorithm that efficiently calculates the likelihood of gene tree g and DNA evolution model parameters Φ as

$$p(S|g, \Phi) = \prod_{i=1}^{l} p(s_i|g, \Phi),$$

where s_i is i-th site in S, and

$$p(s_i|g, \Phi) = p(s_i|g_{top}, g_\tau, \pi, q, \mu).$$

Here, g_{top} is the tree topology, g_τ is the divergence times of the gene tree, $\pi = \{\pi_A, \pi_T, \pi_C, \pi_G\}$ is a vector of equilibrium frequencies of the four nucleotides, $q = \{q_{AT}, q_{AC}, q_{AG}, q_{TC}, q_{TG}, q_{CG}\}$ is a vector of substitution rates between pairs of nucleotides, and μ is the mutation rate. Over a branch j whose length (in expected number of mutations per site) is t_j, the transition probability is calculated as $e^{\mu q t_j}$. In the implementation, we use the BEAGLE library (Ayres et al., 2011) for more efficient implementation of Felsenstein's algorithm.

Yu *et al.* (Yu et al., 2012, 2013b, 2014) fully derived the mass and density functions of gene trees under the multispecies network coalescence, where the lengths of a phylogenetic network's branches are given in coalescent units. Here, we derive the probability density function (pdf) of gene trees for a phylogenetic network given by its topology, divergence/migration times and population size parameters following (Rannala and Yang, 2003; Yu et al., 2014). Coalescence times in the (sampled) gene trees posit temporal constraints on the divergence and migration times of the phylogenetic network.

We use $\tau_\Psi(v)$ to denote the divergence time of node v in phylogeny Ψ (tree or network). Given a gene tree g whose coalescence times are given by τ' and a phylogenetic network Ψ whose divergence times are given by τ, we define a coalescent history with respect to times to be a function $h : V(g) \rightarrow E(\Psi)$, such that the following condition holds:

- if $(x, y) \in E(\Psi)$ and $\tau_\Psi(x) > \tau_g(v) \geq \tau_\Psi(y)$, then $h(v) = (x, y)$.
- if r is the root of Ψ and $\tau_g(v) \geq \tau_\Psi(r)$, then $h(v) = er(\Psi)$.

The quantity $\tau_g(v)$ indicates at which point of branch (x, y) coalescent event v happens. We denote the set of coalescent histories with respect to coalescence times for gene tree g and phylogenetic network Ψ by $H_\Psi(g)$.

Given a phylogenetic network Ψ, the pdf of the gene tree random variable is given by

$$p(g|\Psi, \Gamma) = \sum_{h \in H_\Psi(g)} p(h|\Psi, \Gamma), \tag{4}$$

where $p(h|\Psi, \Gamma)$ gives the pdf of the coalescent history (with respect to divergence times) random variable.

Consider gene tree g for locus j and an arbitrary $h \in H_\Psi(g)$. For an edge $b = (x, y) \in E(\Psi)$, we define $T_b(h)$ to be a vector of the elements in the set $\{\tau_g(w) : w \in h^{-1}(b)\} \cup \{\tau_\Psi(y)\}$ in increasing order. We denote by $T_b(h)[i]$ the i-th element of the vector. Furthermore, we denote by $u_b(h)$ the number of gene lineages entering edge b and $v_b(h)$ the number of gene lineages leaving edge b under h. Then we have

$$p(h|\Psi, \Gamma) = \prod_{b \in E(\Psi)} \left[\prod_{i=1}^{|T_b(h)|-1} \frac{2}{\theta_b} e^{-(\frac{2}{\theta_b})\binom{u_b(h)-i+1}{2}(T_b(h)_{i+1} - T_b(h)_i)} \right] \times e^{-(\frac{2}{\theta_b})\binom{v_b(h)}{2}(\tau_\Psi(b) - T_b(h)_{|T_b(h)|})} \times \Gamma_b^{u_b(h)}, \tag{5}$$

where $\theta_b = 4N_b\mu$ and N_b is the population size corresponding to branch b, μ is the mutation rate per-site per-generation, and Γ_b is the inheritance probability associated with branch b.

Prior distributions

We extended the prior of phylogenetic network composed of topology and branch lengths in (Wen et al., 2016b) to phylogenetic networks composed of topology, divergence times and population sizes, as given by Eq. (3), where $p(\Psi_{ret}|\nu)$, the prior on the number of reticulation nodes, and $p(\Psi_{top}|\Psi_{ret}, \Psi_\tau, \eta)$, the prior on the diameters of reticulation nodes, were defined in (Wen et al., 2016b).

It is important to note here that if Ψ_{top} does not follow the phylogenetic network definition, then $p(\Psi|\nu, \delta, \eta, \psi) = 0$. This is crucial since, in the MCMC kernels we describe below, we allow the moves to produce directed graphs that slightly deviate from the definition; in this case, having the prior be 0 guarantees that the proposal is rejected. Using the strategy, rather than defining only "legal" moves simplifies the calculation of the Hastings ratios. See more details below.

Rannala and Yang used independent Gamma distributions for time intervals (branch lengths) instead of divergence times. However, in the absence of any information on the number of edges of the species network as well as the time intervals, it is computationally intensive to infer the hyperparameters of independent Gamma distributions. Currently, we support two kinds of priors, a uniform distribution (as in BEST (Liu, 2008)), and an exponential distribution $\tau_v \sim \mathrm{Exp}(\delta)$.

We assume one population size per edge, including the edge above the root. Population size parameters are Gamma distributed, $\theta_b \sim \Gamma(2, \psi)$, with a mean 2ψ and a shape parameter of 2. In the absence of any information on the population size, we use the noninformative prior $P_\psi(x) = 1/x$ for hyperparameter ψ (Heled and Drummond, 2010). The number of elements in θ is $|E(\Psi)| + 1$. To simplify inference, our implementation also supports a constant population size across all branches, in which case θ contains only one element.

For the prior on the inheritance probabilities, we use $\Gamma_b \sim \mathrm{Beta}(\alpha, \beta)$. Unless there is some specific knowledge on the inheritance probabilities, a uniform prior on $[0, 1]$ is adopted by setting $\alpha = \beta = 1$. If the amount of introgressed genomic data is suspected to be small in the genome, the hyper-parameters α and β can be appropriately set to bias the inheritance probabilities to values close to 0 and 1 (a U-shaped distribution).

Data access

All methods describe in this work have been implemented in the PhyloNet software (Than et al., 2008), which is freely available for download in open source at http://bioinfo.cs.rice.edu/phylonet.

Acknowledgments

LN was supported in part by NSF grants CCF-1302179 and DBI-1062463. The work was also supported in part by NSF grants OCI-0959097 (Data Analysis and Visualization Cyberinfrastructure), and CNS-1338099 (Big-Data Private-Cloud Research Cyberinfrastructure).

References

Arnold, M. L. 1997. Natural Hybridization and Evolution. Oxford University Press, Oxford.

Ayres, D. L., A. Darling, D. J. Zwickl, P. Beerli, M. T. Holder, P. O. Lewis, J. P. Huelsenbeck, F. Ronquist, D. L. Swofford, M. P. Cummings, et al. 2011. Beagle: an application programming interface and high-performance computing library for statistical phylogenetics. Systematic biology 61:170–173.

Barton, N. 2001. The role of hybridization in evolution. Molecular Ecology 10:551–568.

Bloomquist, E. and M. Suchard. 2010. Unifying vertical and nonvertical evolution: A stochastic ARG-based framework. Systematic Biology 59:27–41.

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

DeGiorgio, M. and J. H. Degnan. 2014. Robustness to divergence time underestimation when inferring species trees from estimated gene trees. Systematic biology 63:66–82.

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

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

Felsenstein, J. 1981. Evolutionary trees from dna sequences: a maximum likelihood approach. J Mol Evol 17:368–376.

Fontaine, M. C., J. B. Pease, A. Steele, R. M. Waterhouse, D. E. Neafsey, I. V. Sharakhov, X. Jiang, A. B. Hall, F. Catteruccia, E. Kakani, et al. 2015. Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science 347:1258524.

Gogarten, J. P., W. F. Doolittle, and J. G. Lawrence. 2002. Prokaryotic evolution in light of gene transfer. Molecular biology and evolution 19:2226–2238.

Green, P. J. 1995. Reversible jump markov chain monte carlo computation and bayesian model determination. Biometrika 82:711–732.

Green, P. J. 2003. Trans-dimensional Markov chain Monte Carlo. Pages 179–198 *in* Highly Structured Stochastic Processes (P. Green, N. Hjort, and S. Richardson, eds.). Oxford University Press, Oxford, UK.

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

Heled, J. and A. J. Drummond. 2010. Bayesian inference of species trees from multilocus data. Molecular biology and evolution 27:570–580.

Koonin, E. V., K. S. Makarova, and L. Aravind. 2001. Horizontal gene transfer in prokaryotes: quantification and classification 1. Annual Reviews in Microbiology 55:709–742.

Kuhner, M. K. and J. Felsenstein. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Molecular Biology and Evolution 11:459–468.

Liu, L. 2008. Best: Bayesian estimation of species trees under the coalescent model. Bioinformatics 24:2542–2543.

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.

Marcussen, T., S. R. Sandve, L. Heier, M. Spannagl, M. Pfeifer, K. S. Jakobsen, B. B. Wulff, B. Steuernagel, K. F. Mayer, O.-A. Olsen, et al. 2014. Ancient hybridizations among the ancestral genomes of bread wheat. Science 345:1250092.

Nakhleh, L. 2010. Evolutionary phylogenetic networks: models and issues. Pages 125–158 *in* The Problem Solving Handbook for Computational Biology and Bioinformatics (L. Heath and N. Ramakrishnan, eds.). Springer, New York.

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

Rieseberg, L. 1997. Hybrid origins of plant species. Annu. Rev. Ecol. Syst. 28:359–389.

Robinson, D. and L. Foulds. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.

Rokas, A., B. L. Williams, N. King, and S. B. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804.

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

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

Wen, D., Y. Yu, M. W. Hahn, and L. Nakhleh. 2016a. Reticulate evolutionary history and extensive introgression in mosquito species revealed by phylogenetic network analysis. Molecular Ecology 25:2361–2372.

Wen, D., Y. Yu, and L. Nakhleh. 2016b. Bayesian inference of reticulate phylogenies under the multispecies network coalescent. PLoS genetics 12:e1006006.

Yu, Y., R. M. Barnett, and L. Nakhleh. 2013a. Parsimonious inference of hybridization in the presence of incomplete lineage sorting. Systematic Biology 62:738–751.

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

Yu, Y., J. Dong, K. J. Liu, and L. Nakhleh. 2014. Maximum likelihood inference of reticulate evolutionary histories. Proceedings of the National Academy of Sciences 111:16448–16453.

Yu, Y., N. Ristic, and L. Nakhleh. 2013b. Fast algorithms and heuristics for phylogenomics under ILS and hybridization. BMC Bioinformatics 14:S6.

Yu, Y., T. Warnow, and L. Nakhleh. 2011. Algorithms for mdc-based multi-locus phylogeny inference: beyond rooted binary gene trees on single alleles. Journal of Computational Biology 18:1543–1559.

Zimmermann, T., S. Mirarab, and T. Warnow. 2014. BBCA: Improving the scalability of *BEAST using random binning. BMC genomics 15:S11.