A Bayesian Implementation of the Multispecies Coalescent Model with Introgression for Phylogenomic Analysis

Tomáš Flouri,[1] Xiyun Jiao,[1] Bruce Rannala,[2] and Ziheng Yang*,[1]

[1]Department of Genetics, Evolution and Environment, University College London, London, United Kingdom

[2]Department of Evolution and Ecology, University of California, Davis, Davis, CA

*Corresponding author: E-mail: z.yang@ucl.ac.uk.

Associate editor: Michael Rosenberg

Abstract

Recent analyses suggest that cross-species gene flow or introgression is common in nature, especially during species divergences. Genomic sequence data can be used to infer introgression events and to estimate the timing and intensity of introgression, providing an important means to advance our understanding of the role of gene flow in speciation. Here, we implement the multispecies-coalescent-with-introgression model, an extension of the multispecies-coalescent model to incorporate introgression, in our Bayesian Markov chain Monte Carlo program BPP. The multispecies-coalescent-with-introgression model accommodates deep coalescence (or incomplete lineage sorting) and introgression and provides a natural framework for inference using genomic sequence data. Computer simulation confirms the good statistical properties of the method, although hundreds or thousands of loci are typically needed to estimate introgression probabilities reliably. Reanalysis of data sets from the purple cone spruce confirms the hypothesis of homoploid hybrid speciation. We estimated the introgression probability using the genomic sequence data from six mosquito species in the *Anopheles gambiae* species complex, which varies considerably across the genome, likely driven by differential selection against introgressed alleles.

Key words: **Bayesian inference, BPP, introgression, multispecies coalescent with introgression, MSci, MCMC.**

Introduction

A number of recent studies have revealed cross-species hybridization/introgression in a variety of species ranging from *Arabidopsis* (Arnold et al. 2016), butterflies (Martin et al. 2013), corals (Mao et al. 2018), and birds (Ellegren et al. 2012) to mammals such as bears (Liu et al. 2014; Kumar et al. 2017), cattle (Wu et al. 2018), gibbons (Chan et al. 2013; Shi and Yang 2018), and hominins (Nielsen et al. 2017). Introgression may play an important role in speciation (Harrison and Larson 2014; Mallet et al. 2016; Martin and Jiggins 2017). Inference of introgression and estimation of migration rates can contribute to our understanding of the speciation process (Mallet et al. 2016; Martin and Jiggins 2017). Furthermore, introgression and deep coalescence (or incomplete lineage sorting) are two major challenges for species tree reconstruction (Martin et al. 2013; Liu et al. 2014; Fontaine et al. 2015).

There is a large body of literature on the use of networks to model non-treelike evolution (Huson et al. 2011) and a number of methods have been developed to detect cross-species gene flow. Most use summaries of the multilocus sequence data such as the estimated gene trees (Solis-Lemus and Ane 2016; Wen et al. 2016; Solis-Lemus et al. 2017; Cao et al. 2019) or the counts of parsimony-informative site patterns (Green et al. 2010; Durand et al. 2011; Blischak et al. 2018). See Degnan (2018) and Folk et al. (2018) for recent reviews. We focus on coalescent-based full-likelihood models applied to multilocus sequence alignments from closely related species. These come in two forms. The isolation-with-migration (IM) model assumes continuous migration, with species exchanging migrants at certain rates every generation (Hey and Nielsen 2004; Hey 2010), whereas the multispecies-coalescent-with-introgression (MSci) model assumes episodic introgression/hybridization (Yu et al. 2014). Although the probability density of the gene trees under the IM (Hey 2010) and MSci (Yu et al. 2014) models is straightforward to compute, developing a Bayesian Markov chain Monte Carlo (MCMC) program that is feasible for use with genome-scale data sets has been challenging. The space of unknown genealogical histories (including the migration/introgression histories) is large, and constraints between the species tree and the gene trees make it difficult to traverse the parameter space in the posterior. Current implementations of full-likelihood methods through MCMC include IMa3 (Hey 2010; Hey et al. 2018) for the IM model, and *BEAST (Zhang et al. 2018; Jones 2019) and PHYLONET/MCMC_SEQ (Wen and Nakhleh 2018; Wen et al. 2018) for the MSci model. It does not appear computationally feasible to apply those programs to realistically sized data sets, with more than 200 loci, say.

In this article, we extend the multispecies-coalescent (MSC) model in the BPP program (Rannala and Yang 2003; Burgess and Yang 2008; Yang 2015) to accommodate introgression, resulting in the MSci model (Degnan 2018). The MSci model can be used to estimate species divergence times

Article

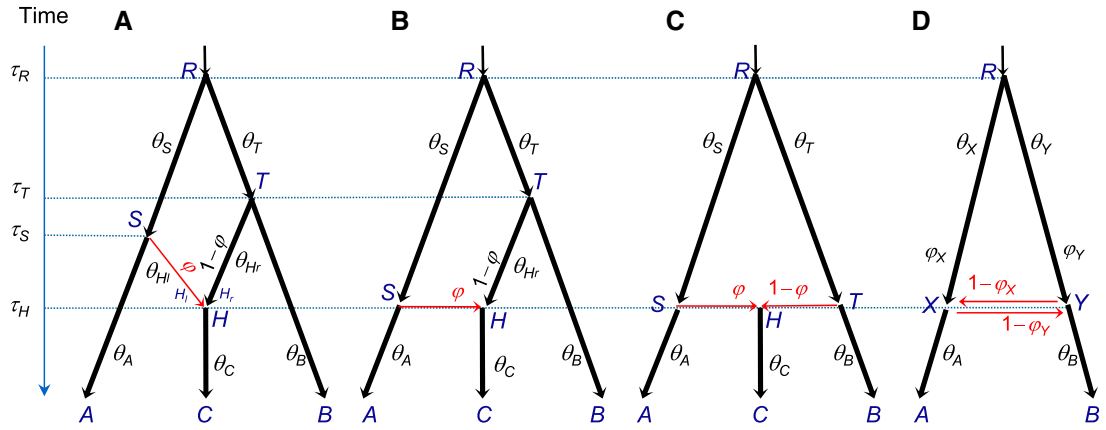

Fig. 1. The MSci model with four different types of hybridization (introgression or admixture) events. In (*A*), two parental species *SH* and *TH* merge to form a hybrid species *H*, at time τ_H, leading to extinction of the parental species. In (*B*), there is introgression from species *RSA* to species *THC* at time $\tau_H = \tau_S$, with introgression probability φ. In (*C*), species *RSA* and *RTB* come into contact to form hybrid species *H* at time $\tau_S = \tau_H = \tau_T$, which evolves into species *C*, while the two parent species become *A* and *B*. In (*D*), bidirectional introgression occurs between species *RXA* and *RYB* at time $\tau_X = \tau_Y$, with introgression probabilities φ_X and φ_Y. Parameters in the model include speciation/hybridization times (τs), population sizes (θs), and introgression probabilities (φs). The models are represented using the extended Newick notation (see Appendix) (Cardona et al. 2008), as (*A*–*C*): $((A, (C)H)S, (H, B)T)R$ and (*D*): $((A, Y)X, (X, B)Y)R$. Arrows indicate the direction of time from parent to child or from source to target populations.

and the number, timings, and intensities of introgression events. By accommodating gene flow and providing more reliable estimates of evolutionary parameters, the model may also be used in heuristic species delimitation (Jackson et al. 2017; Leaché et al. 2019). We conduct simulation to examine the statistical properties of the method, in comparison with two summary methods, SNaQ (Solis-Lemus and Ane 2016; Solis-Lemus et al. 2017) and HyDe (Blischak et al. 2018). We apply the new method to data sets of purple cone spruce (Sun et al. 2014; Zhang et al. 2018), budding yeast (Rokas et al. 2003; Wen and Nakhleh 2018), and *Anopheles* mosquito genomes (Fontaine et al. 2015; Thawornwattana et al. 2018a), to examine the computational efficiency of our algorithms in comparison with previous implementations (Wen and Nakhleh 2018; Zhang et al. 2018) and to estimate the introgression probability and to study its variation across the genome.

Results

The MSci Model

We extend the MSC model (Rannala and Yang 2003) to accommodate cross-species hybridization (or introgression) by introducing hybridization (or *H*) nodes (fig. 1). Each *H* node has two parents (H_l and H_r, for left and right) and one daughter, although the *H* node and its parents may have the same age when there is an admixture or horizontal gene transfer (fig. 1*B* and *C*). In model A, both parental species become extinct after hybridization, whereas model B represents an introgression from species *RSA* into *THC*. Model C represents hybrid speciation, whereas model D represents bidirectional introgression (Kubatko 2009).

When we trace a lineage backward in time and reach an *H* event, the lineage may traverse either the left or the right parental species, according to the introgression probability (φ or $1-\varphi$). This probability is equivalent to the "inheritance probability" γ of Yu et al. (2014) and the "heritability" of Solis-Lemus and Ane (2016). The MSci model includes three sets of parameters: the speciation and introgression times ($\boldsymbol{\tau}$); the population size parameters ($\boldsymbol{\theta}$), with each $\theta = 4N\mu$, where *N* is the effective population size and μ is the mutation rate per generation per site; and the introgression probabilities ($\boldsymbol{\varphi}$). Both τs and θs are measured by the expected number of mutations per site. Here, we assume that the MSci model is fixed; cross-model MCMC moves will be developed in future work.

Let $\boldsymbol{G} = \{G_i\}$ be the set of gene trees for the *L* loci. For each locus *i*, G_i represents the gene-tree topology, the branch lengths (coalescent times), and the paths taken at the *H* nodes, indicated by a set of flags for each gene-tree branch, with l for left, r for right and $\varnothing$ for null (meaning that the branch does not pass the *H* node). The data $\boldsymbol{X} = \{X_i\}$ are the sequence alignments at the *L* loci. Sites within the same locus are assumed to share the same genealogical history, whereas the gene trees and coalescent times are assumed to be independent among loci given the species tree and parameters. The ideal data for this kind of analysis are loosely linked short genomic segments (called loci), so that recombination within a locus is unimportant, whereas different loci are largely independent (Burgess and Yang 2008; Lohse et al. 2016; Hey et al. 2018). The Bayesian formulation consists of two components. The first is the probability density of gene trees given the species tree under the MSci model, $f(G_i|\boldsymbol{\tau}, \boldsymbol{\theta}, \boldsymbol{\varphi})$, given in Yu et al. (2014) although here G_i includes the flags for hybrid nodes. Note that this differs from the density given by Kubatko (2009), as pointed out by Solis-Lemus and Ane (2016). The second component is the likelihood of the sequence data at each locus *i* given the gene tree, $f(X_i|G_i)$ (Felsenstein 1981). The posterior probability density of the parameters on the species tree given sequence data is then

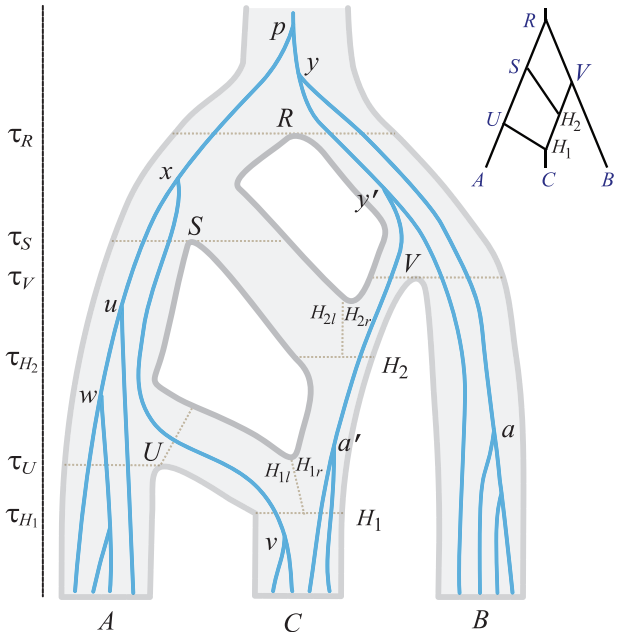

FIG. 2. A species tree for three species (*A*, *B*, *C*) with a gene tree for 12 sequences running inside it to illustrate the gene-tree node-age move and the gene-tree SPR move. There are four speciation nodes (*R*, *S*, *U*, *V*) and two hybridization nodes (H_1 and H_2). This MSci model is also used in simulation, where the model is referred to as "2H".

$$f(\boldsymbol{\tau}, \boldsymbol{\theta}, \boldsymbol{\varphi}|\boldsymbol{X}) \propto f(\boldsymbol{\tau}, \boldsymbol{\theta}, \boldsymbol{\varphi}) \prod_{i=1}^{L} \int_{G_i} f(G_i|\boldsymbol{\tau}, \boldsymbol{\theta}, \boldsymbol{\varphi}) f(X_i|G_i)\, \mathrm{d}G_i, \tag{1}$$

where $f(\boldsymbol{\tau}, \boldsymbol{\theta}, \boldsymbol{\varphi})$ is the prior on parameters. We assign inverse-gamma priors on θs and τs and a beta prior on φ.

We have implemented six MCMC proposals to average over the gene trees (G_i) and sample from the posterior (eq. 1). Those proposals 1) change node ages on gene trees, 2) change gene-tree topologies using subtree pruning and regrafting (SPR), 3) change θs on the species tree using sliding windows, 4) change τs on the species tree using a variant of the rubber-band algorithm (Rannala and Yang 2003), 5) changing all node ages on the species tree and gene trees using a multiplier, and 6) change the introgression probabilities φs using sliding windows. The proposals are detailed in Materials and Methods using the example species tree model of figure 2.

Simulation Study

We conducted three sets of simulation to examine the performance of BPP in different situations.

The first set includes multiple sequences from each species and examines BPP estimation of parameters in the MSci model and the impact of factors such as the number of loci, the introgression probability φ, and the species tree model. We used models A and C of figure 1. Each data set consisted of 10, 100, or 1,000 loci, with 10 sequences from each species per locus (and 30 sequences in total). We used two values for φ (0.1 and 0.5) and two values for θ (0.001 and 0.01). Either a JC (Jukes and Cantor 1969) or a GTR + Γ (Yang 1994a, 1994b) substitution model was used to simulate data, but JC was always used to analyze them. The results may be summarized as follows (supplementary figs. S1–S8, Supplementary Material online).

- First, there were large variations in estimation precision and accuracy among the different parameters. For example, estimates of θs for modern species were accurate even in small data sets in all combinations of trees, models, and θ values. In contrast, θs for some ancestral species (such as θ_S, θ_T, θ_{H_l}, and θ_{H_r} in model A when the true $\theta = 0.001$) were poorly estimated, with the posterior dominated by the prior even with 100 or 1,000 loci. Those parameters were hard to estimate as very few sequences enter or coalesce in the populations. These same parameters were much better estimated when the true $\theta = 0.01$ as then many more sequences could enter and coalesce in the ancestral species. For similar reasons, the ages of ancestral nodes such as τ_H and τ_S were much better estimated when the true $\theta = 0.01$ than when $\theta = 0.001$.
- Second, parameter estimates under model C were more precise than under model A, because the former has 9 parameters, whereas the latter 13.
- Third, there were virtually no differences in the results whether the data were simulated under JC or GTR + Γ. As the role of the mutation model in BPP is to correct for multiple hits at the same site and as the simulated sequences are highly similar, the choice of the mutation model is unimportant. Similar observations were made in previous simulations examining species tree estimation without introgression (Shi and Yang 2018).
- Last, the data size (the number of loci) had a huge impact on the precision and accuracy of estimation. In particular, data of only 10 or 100 loci did not produce reliable estimates of φ, whereas estimates from 1,000 loci were both precise (with narrow intervals) and accurate (close to the true values). Because the MSci models are parameter rich, large data sets in the order of 1,000 loci are necessary for reliable inference.

In the second set of simulations we compared BPP with two summary methods: SNAQ (Solis-Lemus and Ane 2016; Solis-Lemus et al. 2017) and HYDE (Blischak et al. 2018), using one sequence per species. We simulated data under model A, with three ingroup species (*A*, *B*, and *C*), as well as two outgroup species *D* and *E*, as required by SNAQ (Solis-Lemus et al. 2017). One sequence was sampled per species per locus. The data were then analyzed using the three programs to estimate φ (supplementary fig. S9, Supplementary Material online). Data size had a large impact on the precision and accuracy of the estimates. All three methods performed poorly with 10 or 100 loci (or gene trees), but the estimates were close to the true values with 1,000 loci. Overall, the three methods had similar performance in estimating φ. In some small data sets, SNAQ and HYDE had extreme estimates of 0, whereas BPP always produced nonzero estimates, due to Bayesian shrinkage through the prior.

Note that the problem examined here is a conventional parameter estimation problem under a well-specified model, so that standard statistical theory applies, which states that the Bayesian method has optimal large-sample properties (O'Hagan and Forster 2004). The small differences among the methods suggest that information in the data concerning φ mostly lies in the proportions of gene trees, which may be reliably estimated even if phylogenetic information content at each individual locus is low. We note that BPP has several advantages. 1) BPP accommodates the uncertainties in the data appropriately and produces posterior credible intervals (CIs), whereas SNaQ and HyDe generate point estimates only. 2) BPP estimates all 13 parameters in the model, whereas SNaQ and HyDe estimate only 2 (φ and the internal branch length) with the others unidentifiable. Estimates of ancestral population sizes (θs) and species divergence and introgression times (τs) may be useful for understanding the evolutionary history of the species. 3) BPP can use loci of any data configuration, including loci with sequences from only one or two species, which are informative for BPP but carry no information about gene trees. 4) Some introgression models or biologically important scenarios are unidentifiable using SNaQ and HyDe but can be analyzed using BPP (see below). In contrast, SNaQ and HyDe have a huge computational advantage over BPP and may be very useful for exploratory analysis in large data sets.

The third set of simulations explored the performance of BPP under models that are unidentifiable using SNaQ and HyDe. We used model D of figure 1 and model 2H of figure 2, with results in supplementary figure S10, Supplementary Material online. Model D represents bidirectional introgression between two species. Population size parameters (θs) for modern species *A* and *B* were well estimated even with 100 loci, as was θ_R for the root, but θ_X for species *X* (branch *X*-*R*) and θ_Y (for branch *Y*-*R*) were more poorly estimated (supplementary fig. S10*A*, Supplementary Material online). Both τ parameters were well estimated. The introgression probabilities φ_X and φ_Y were poorly estimated in small data sets of 10 or 100 loci but were fairly accurate with 1,000 loci.

Model 2H (fig. 2) involves two introgression events on a species tree of three species. There were large differences in information content for different parameters (supplementary fig. S10*B*, Supplementary Material online). Parameters θs for modern species were well estimated even in small data sets, but θs for most ancestral species were poorly estimated because of lack of coalescent events in those populations. Parameter $\theta_{H_{1r}}$ was more accurately estimated than $\theta_{H_{1l}}$ because more sequences passed node H_1 from the right (with probability $1-\varphi=0.9$) than from the left (with $\varphi=0.1$), and $\theta_{H_{1l}}$ and $\theta_{H_{1r}}$ were more reliably estimated than $\theta_{H_{2l}}$ and $\theta_{H_{2r}}$ because more sequences passed node H_1 than node H_2. Similarly τ_{H_1} was better estimated than τ_{H_2}. With 1,000 loci, all six node ages (for R, S, U, V, H_1, and H_2) were well estimated. The two introgression probabilities (φ_{H_1} and φ_{H_2}) were poorly estimated with 10 or 100 loci but were reliably estimated when 1,000 loci were used.

In summary, in both introgression scenarios of models D and 2H, where SNaQ and HyDe are inapplicable, BPP appears to be a well-behaved method, providing reliable estimates of introgression probabilities as well as species divergence and introgression times.

Analysis of the Purple Cone Spruce Data

We analyzed three data sets concerning the origin of the purple cone spruce in the Qinghai–Tibet Plateau, *Picea purpurea*, hypothesized to be a hybrid species, formed through homoploid hybridization between *P. wilsonii* (*W*) and *P. likiangensis* (*L*) (Sun et al. 2014). Two small data sets were previously analyzed using *BEAST under model A of figure 3, whereas the third one (the "Full" data) is a much larger data set from which the first two were sampled. We attempted to apply PhyloNet/MCMC_SEQ (Wen and Nakhleh 2018) to analyze any of those data sets but were unsuccessful. The program used all 144 cores on our server and did not produce any output after 5 days. The data sets appeared to be too large for the program.

Parameter estimates under model A from the two small data sets were similar to those in Zhang et al. (2018), with φ estimates between 0.32 and 0.44, although the estimates involved large uncertainties (supplementary table S1, Supplementary Material online). The uncertainty is apparently due to the use of only 11 loci (although many sequences are available at each locus) and the shallowness of the species tree, with species divergence times being comparable with coalescent times (or with similar τs and θs). Accommodating rate variation among loci had very small effects. The full data produced similar parameter estimates to the two small data sets, but φ is larger, at 0.47–0.49. We also applied model C (fig. 3) to the full data, which produced more precise estimates because of the smaller number of parameters (fig. 3 and supplementary table S1, Supplementary Material online). The φ estimate under model C was 0.53, with the 95% highest posterior density (HPD) CI to be 0.36–0.71.

Note that the models represent different biological scenarios. Model A assumes the existence, and subsequent extinction at the time of hybridization, of species *DH* and *EH* (fig. 3). This is not a very plausible model (Sun et al. 2014). Model C represents speciation through homoploid hybridization, with species *RDW* and *REL* coming into contact and forming a hybrid species (*H*) at time τ_H. A possible scenario is that changes in species distribution may have led to habitat overlap between *P. wilsonii* and *P. likiangensis* during the Quaternary glaciation in the central Qinghai–Tibet Plateau (Sun et al. 2014). We calculated marginal likelihoods (Bayes factors) to compare models A–C (fig. 1). The log marginal likelihood was −18,361 for model A, −18,359 and −18,361 for two cases of model B (with $\tau_H=\tau_D$ and $\tau_H=\tau_E$, respectively), and −18,362 for model C, suggesting the fit of the models to data is similar. The marginal likelihoods are thus indecisive. We suggest that model C should be preferred, because of its biological plausibility.

Analysis of the Budding Yeast Data Set

We fitted the MSci model of figure 4 to the data of 106 loci from 5 species of budding yeast. This model had a posterior probability of >95% in the PhyloNet/MCMC_SEQ analysis of the

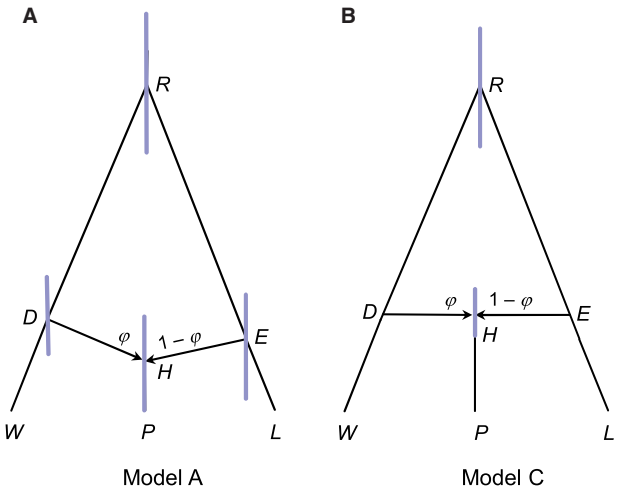

Fig. 3. Two species trees for the purple cone spruce *Picea purpurea* (P) from the Qinghai–Tibet Plateau, and two parental species *P. wilsonii* (W) and *P. likiangensis* (L). These correspond to (A) Model A and (B) Model C of figure 1. The branch lengths represent the posterior means of divergence times (τs) estimated from the "Full" data set, with node bars showing the 95% HPD intervals (see supplementary table S1, Supplementary Material online).

same data by Wen and Nakhleh (2018). The φ estimate from BPP was 0.70 (with the 95% HPD CI 0.56–0.83), compared with 0.75 ± 0.06 in Wen and Nakhleh (2018). The small differences may be due to the use of different priors and the assumption of a constant θ across all populations in the PHYLONET analysis. The results confirm the expectation that full-likelihood programs, if computationally feasible, should produce similar results. Running time for achieving an effective sample size (ESS) of 1,000 for φ was ~3 min for BPP using all 8 threads on a notebook, compared with ~17 h for PHYLONET using 32 threads on a computer server (Wen and Nakhleh 2018). If we make a 10-fold allowance for the fact that the model is fixed in BPP while PHYLONET spent computational efforts attempting changes to the model, this very roughly translates into a 100-fold difference in mixing/computational efficiency between the two programs ($17 \times 60/3 \times 4/10 = 136$).

Variable Introgression across the Genome in the *Anopheles gambiae* Species Complex

To examine the variation in introgression intensity across the *Anopheles* genome, we analyzed blocks of 100 loci, assuming the species tree of figure 5. Estimates of $\varphi_{A \to GC}$ for the *A. arabiensis* → *A. gambiae* + *A. coluzzii* introgression and $\varphi_{R \to Q}$ for the *A. merus* → *A. quadriannulatus* introgression vary considerably across genomic regions or chromosomal arms (fig. 6). The probability $\varphi_{A \to GC}$ is high (>0.5) in most blocks, whereas $\varphi_{R \to Q}$ is high in 3La and 3R.

We then merged the loci on the same chromosomal arms/regions to form 12 large coding and noncoding data sets, and analyzed them under the model of fig. 5 (table 1). We also sampled three sequences per locus to form data triplets for analysis using the maximum likelihood (ML) program 3S (Thawornwattana et al. 2018a, supplementary table S3, Supplementary Material online, GAR and RQO). For all

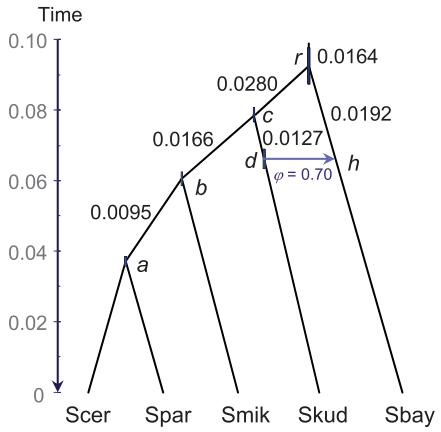

Fig. 4. The species tree for five species of budding yeast, with one introgression event. The branches were drawn to reflect the posterior means of divergence/introgression times (τs) from BPP, with node bars showing the 95% HPD intervals, whereas posterior means of population sizes (θs) are shown along the branches.

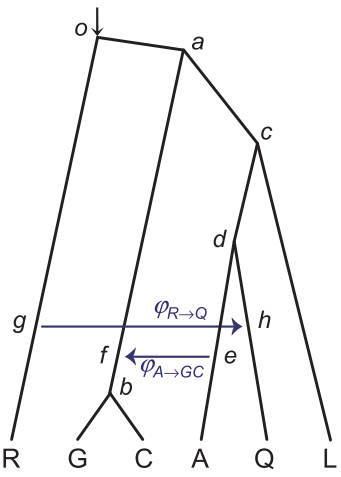

Fig. 5. A species tree with two introgression events for the *Anopheles gambiae* species complex.

autosomes, the introgression/migration rate from *A. arabiensis* to *A. gambiae* + *A. coluzzii* is very high (with $\varphi_{A \to GC} > 0.5$), whereas $M_{A \to G}$ ranges from 0.12 to 1.12. To reconcile the estimates from the two models, note that M is the expected number of migrants per generation, so that even a small M may mean a large number of migrants accumulated over many generations. As noted previously (Fontaine et al. 2015; Thawornwattana et al. 2018a), the autosomes are overwhelmed by the $A \to GC$ introgression so that all species tree methods that ignore gene flow infer incorrect species trees. In population genetic models of population subdivision, migration rates of $M \ll 1$ do not lead to substantial population subdivision. However, here M as low as 0.1 may have a significant impact on the species phylogeny if the species arose through radiative speciation events and the ancestral species had large sizes.

Parameter $\varphi_{R \to Q}$ varied across chromosomal regions and was high for the inversion region 3La. Coding and noncoding loci produced highly consistent estimates of species trees,

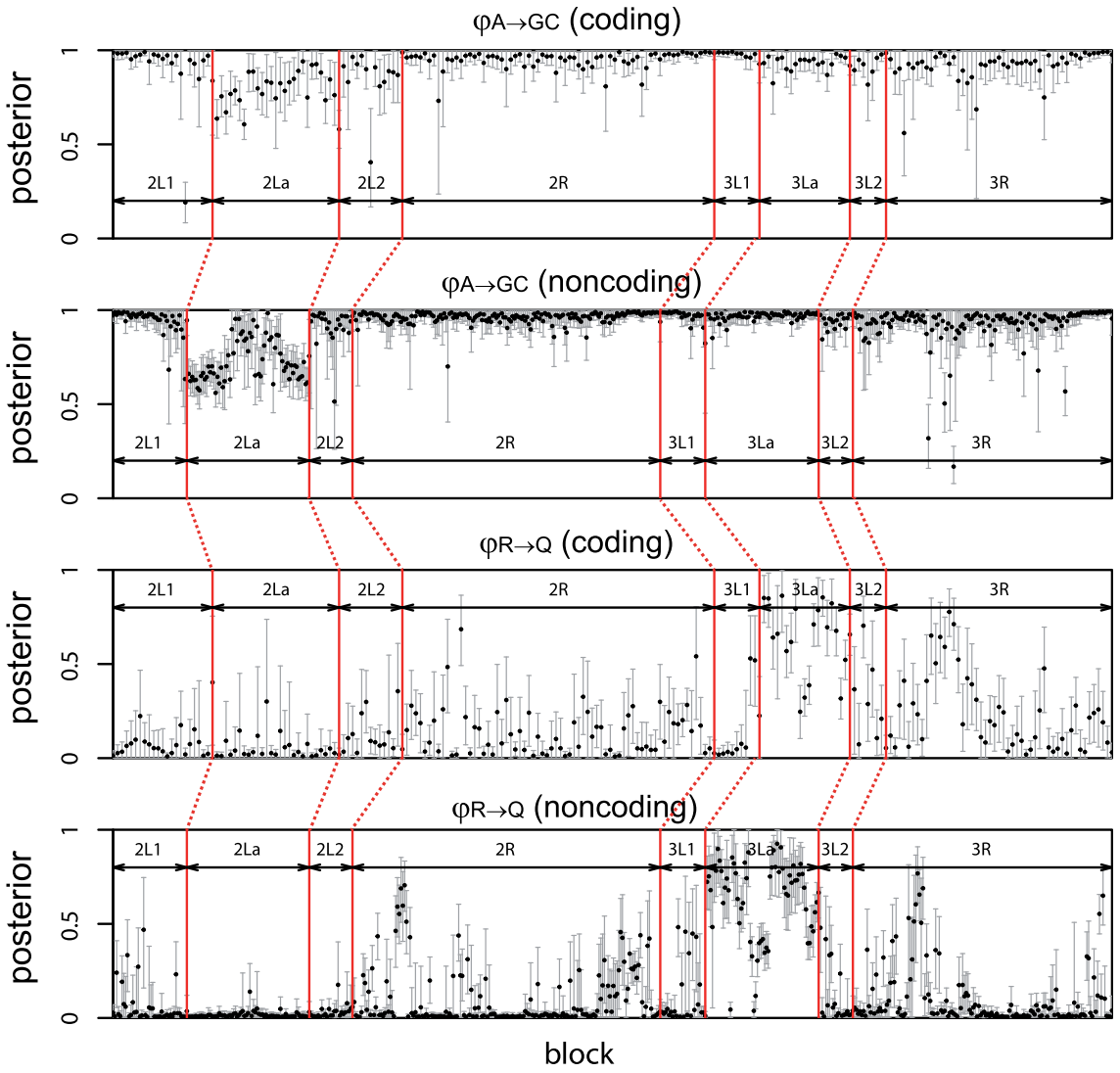

FIG. 6. Posterior means and 95% HPD intervals of the introgression probabilities $\varphi_{A\to GC}$ and $\varphi_{R\to Q}$ for the *Anopheles gambiae* species complex in BPP analysis of the blocks. Each block consists of 100 loci, which are assumed to have the same φ at each hybridization node.

species divergence times, and population sizes (supplementary table S2, Supplementary Material online) (see also Thawornwattana et al. 2018a), but estimates of migration rate/introgression probability differed between the two data sets. The higher introgression rates for coding than noncoding loci in regions 3La, 2L1 + 2, and 3R suggest the intriguing possibility that the introgressed genes may have brought adaptive advantages, so that introgression is aided by natural selection. The functions of coding genes or exons that are most likely transferred across the species barriers could be examined to explore this hypothesis. There is overall consistency between estimates from the IM model in 3s and the MSci model in BPP in that regions with high φ tend to have high M as well. Note that both φ and M reflect long-term effective gene flow, after the filtering of introgressed alleles by natural selection.

Discussion

Identifiability of MSci Models

If the probability distributions of the data are identical for two sets of parameter values (Θ and Θ'), with $f(X|\Theta) = f(X|\Theta')$ for all possible data X, then Θ is unidentifiable given data X. Previous studies of identifiability have mostly focused on the use of gene-tree topologies as data (Zhu and Degnan 2017; Degnan 2018). Note that a model unidentifiable given gene-tree topologies alone may be identifiable given gene trees with branch lengths or coalescent times, and that a model unidentifiable when one sequence is sampled per species may be identifiable when multiple samples per species are available (Yu et al. 2012; Pardi and Scornavacca 2015; Zhu and Degnan 2017).

A comprehensive examination of the identifiability issue under MSci is beyond the scope of this article. Here, we consider a few simple cases. First, the population size parameter θ is unidentifiable if at most one sequence per locus is sampled from that species or its descendant species and τs associated with a hybridization event may also be unidentifiable. Consider the model of figure 2 and suppose the data consist of one sequence from each species. Then θ_A, θ_B, and θ_C, as well as $\theta_{H_{1l}}$, $\theta_{H_{1r}}$, $\theta_{H_{2l}}$, and $\theta_{H_{2r}}$, are unidentifiable. In addition, τ_{H_1} and τ_{H_2} are unidentifiable. Parameters τ_U, τ_V, τ_S, τ_R, and

Table 1. Maximum Likelihood (3s) Estimates of Migration Rate ($M = Nm$) and Bayesian (BPP) Estimates of Introgression Probability (φ) from the *Anopheles* Genomic Data.

Data Set	Loci	*A. arabiensis* → *A. gambiae* + *A. coluzzii*		*A. merus* → *A. quadriannulatus*	
		$\widehat{M}_{A\to G}$ (3s, GAR)	$\varphi_{A\to GC}$ (BPP)	$\widehat{M}_{R\to Q}$ (3s, RQO)	$\varphi_{R\to Q}$ (BPP)
2L1 + 2 coding	3,585	0.319 0.372	0.943 (0.924, 0.963)	0.025 0.020	0.281 (0.242, 0.322)
2L1 + 2 noncoding	6,434	0.239 0.209	0.977 (0.967, 0.985)	0.000 0.000	0.002 (0.000, 0.003)
2La coding	2,776	0.915 1.120	0.731 (0.697, 0.766)	0.052 0.036	0.006 (0.002, 0.009)
2La noncoding	6,732	0.989 0.935	0.640 (0.627, 0.653)	0.000 0.000	0.001 (0.000, 0.002)
2R coding	6,849	0.511 0.477	0.966 (0.958, 0.975)	0.030 0.024	0.340 (0.295, 0.391)
2R noncoding	17,027	0.357 0.297	0.971 (0.961, 0.986)	0.007 0.007	0.222 (0.122, 0.342)
3L1 + 2 coding	1,747	0.168 0.340	0.939 (0.923, 0.956)	0.000 0.024	0.321 (0.271, 0.369)
3L1 + 2 noncoding	4,319	0.267 0.267	0.959 (0.951, 0.968)	0.000 0.003	0.330 (0.304, 0.356)
3La coding	1,998	0.866 0.790	0.931 (0.914, 0.947)	0.127 0.105	0.650 (0.619, 0.680)
3La noncoding	6,208	0.692 0.671	0.977 (0.970, 0.984)	0.021 0.020	0.544 (0.456, 0.700)
3R coding	4,977	0.393 0.365	0.945 (0.932, 0.958)	0.042 0.042	0.430 (0.371, 0.488)
3R noncoding	14,323	0.334 0.281	0.977 (0.971, 0.984)	0.009 0.009	0.030 (0.012, 0.062)

NOTE.—ML estimates from 3s were obtained from two random samples of the GAR and RQO triplets, whereas the BPP estimates (posterior means and 95% HPD intervals) used all 12 sequences at each locus.

φ_{H_1} and φ_{H_2} are identifiable, as are $\theta_R, \theta_S, \theta_U$, and θ_V. In this case, the gene tree at any locus depends on whether sequence c takes the left path at H_1 and enters species U (which happens with probability φ_{H_1}), or it takes the right path and enters species H_2 (which happens with probability $1 - \varphi_{H_1}$), but not on the age of the H_1 node. The same applies to the path taken by sequence c at H_2.

The most interesting case for the MSci model implemented here is where multiple sequences are sampled from each species at each locus, with multiple sites per locus. We speculate that the MSci model is identifiable on such data of sequence alignments as long as it is identifiable when the data consist of gene trees with coalescent times: Θ is identifiable using multilocus data X if and only if $f(G, \boldsymbol{t}|\Theta) \neq f(G, \boldsymbol{t}|\Theta')$ for some G and $\boldsymbol{t}$. Note that identifiability implies statistical consistency for a full-likelihood method as implemented here. If the model is identifiable, the Bayesian parameter estimates will approach the true values when the number of loci approaches infinity.

Here, we note an interesting unidentifiability issue with model D of figure 1. Let $\Theta = (\theta_A, \theta_B, \theta_R, \theta_X, \theta_Y, \tau_R, \tau_X, \varphi_X, \varphi_Y)$ be the parameters of the model, and let Θ' have the same parameter values as Θ except that $\theta'_X = \theta_Y$, $\theta'_Y = \theta_X$, $\varphi'_X = 1 - \varphi_X$, and $\varphi'_Y = 1 - \varphi_Y$. Then $f(G|\Theta) = f(G|\Theta')$ for any Θ, G, and data configuration (with n_A and n_B sequences from A and B, respectively, say). Thus for every point Θ in the parameter space, there is a "mirror" point Θ' with exactly the same likelihood. With Θ, a certain number of A sequences may take the left (upper) path at X (with probability φ_X) and enter population XR, coalescing at the rate $2/\theta_X$, whereas with Θ', the same A sequences may take the right (horizontal) path (with probability $1 - \varphi'_X = \varphi_X$) and enter population YR, coalescing at the rate $2/\theta'_Y = 2/\theta_X$. The differences between the two scenarios are in the labeling only, with "left" and X under Θ corresponding to "right" and Y under Θ', but the probabilities involved are exactly the same. The same argument applies to sequences from B going through node Y, and to sequences from A and B considered jointly. This is a case of the label-switching problem. Arguably Θ and Θ' have the same biological interpretations concerning the relatedness of the sequences sampled from A and B. If the priors on φ_X and φ_Y are symmetrical, say beta(α, α), the posterior density will satisfy $f(\Theta|\boldsymbol{X}) = f(\Theta'|\boldsymbol{X})$ for all $\boldsymbol{X}$. Otherwise, the "twin towers" may not have exactly the same height.

Note that the label-switching kind of unidentifiability does not hinder the utility of the model. One can apply an identifiability constraint, such as $\varphi < 0.5$, to remove the unidentifiability. However, in the general case of multiple bidirectional introgression events or multiple species on the species tree, it may be complicated to decide on the identifiability of the model.

Finally, we point out that there are many scenarios of data configurations and parameter settings in which some parameters are only weakly identifiable and very hard to estimate. For example, if θ_C is very small relative to τ_{H_1} in figure 2, sequences from C will have coalesced before reaching node H_1, so that only one C sequence passes H_1 and the data will have little information about $\tau_{H_1}, \theta_{H_{1l}}, \theta_{H_{1r}}, \tau_{H_2}, \theta_{H_{2l}}$, and $\theta_{H_{2r}}$.

Full-Likelihood and Summary Methods to Accommodate Introgression/Migration

Although biologically simplistic, the MSci and IM models offer powerful tools for analysis of genomic sequence data from closely related species, when cross-species gene flow appears to be the norm (Mallet et al. 2016; Martin and Jiggins 2017). Full-likelihood implementations of those models, including the ML (Zhu and Yang 2012; Dalquen et al. 2017) and Bayesian MCMC methods (Hey et al. 2018; Wen and Nakhleh 2018; Zhang et al. 2018), make efficient use of the information in the data and naturally accommodate phylogenetic uncertainties at individual loci caused by high sequence similarities (Edwards et al. 2016; Xu and Yang 2016). The complexity of those models means that large data sets with hundreds or thousands of loci may be necessary to obtain reliable parameter estimates, as indicated by our analyses of both simulated and real data. In this article, we have developed new MCMC proposal algorithms for MSci models (of

types A–D of fig. 1) and have successfully applied them to analyze large data sets of over 10,000 loci (table 1). The algorithms appear to have good mixing efficiency. We suggest that this is a promising start, from which further improvements to the algorithms may be possible. Future work will include implementation of efficient MCMC proposals to move between MSci models, and a systematic examination of identifiability issues.

We note that the computational load for BPP increases with an increase in the number of species, the number of hybridization events, the number of loci, the number of sequences per locus, or the number of sites per sequence. Increasing the number of species or hybridization events increases the number of parameters (as well as the number of models when the model is changing in the MCMC) so that the parameter space becomes much larger. Increasing the number of loci also increases the posterior search space since the MCMC has to sample in the space of gene trees for each locus (Flouri et al. 2018). In comparison, the number of sites per sequence has the least impact on the amount of computation. We found it helpful to make a distinction between computational efficiency of an MCMC algorithm, which reflects the computational time for each MCMC iteration, and mixing efficiency, which is measured by the ESS in parameter estimates for a given number of MCMC iterations. When the data set gets larger, particularly with more loci in the data set, the posterior for parameters becomes spiky, which in general leads to a deterioration of MCMC mixing efficiency, so that a greater number of MCMC iterations become necessary to produce estimates with acceptable precision. We conjecture that poor mixing is a more serious problem than poor computational efficiency for most MCMC algorithms in phylogenomics.

Our simulation suggests that under simple introgression scenarios, summary methods such as SNAQ and HYDE can produce as reliable estimates of the introgression probability as BPP. However, full-likelihood methods provide measures of uncertainties and are applicable to complex introgression scenarios which are unidentifiable using summary methods. Summary methods are simple to implement, computationally efficient, and useful for analyzing large data sets. They can be used to generate hypotheses for further testing and estimation using BPP. Furthermore, the current implementation of MSci in BPP assumes the molecular clock and is unsuitable for distantly related species. Summary methods such as SNAQ use outgroups to root the tree without the need for the molecular clock.

Variable Introgression Probability across the Genome

The models implemented here (fig. 1A–C) assume that the introgression probability or migration rate is constant among loci or across the genome. However, the impact of introgressed alleles on the fitness of the individual may strongly depend on the function of the genes in the introgressed region. Genes involved in cross-species incompatibilities are unlikely to be accepted in the recipient species. For example, crossing experiments between *A. arabiensis* and *A. gambiae* highlighted large differences between the chromosomes, with the X chromosome being most resistant to introgression, presumably because it harbors genes involved in cross-species sterility and inviability (Slotman et al. 2005). Differential selection across the genome means that the φ parameter should vary among loci. Note that φ in our models when estimated from genetic sequence data reflects the long-term combined effects of migration, recombination, and natural selection. It may be very different from the per-generation hybridization rate, which should apply to the whole genome.

In our analysis of the *Anopheles* genomic data, we used blocks of 100 loci to partially accommodate the variation in migration rate or introgression probability across chromosomal regions (fig. 6). We leave it to future work to implement MSci models with φ varying among loci. We note that many sequences per locus may be necessary to estimate locus-specific migration rates or introgression probabilities.

Materials and Methods

MCMC Proposals

We have adapted the five proposals in Rannala and Yang (2003) to accommodate hybridization nodes on the species tree and added another move to update the φ parameters.

Step 1. Change node ages on gene trees using a sliding window. Suppose the concerned gene-tree node is node x with age t_x in population X, with parent node p in population P and two daughter nodes u and v in populations U and V, respectively. To propose a new node age t_x^* first determine the bounds, $t_L < t_x^* < t_U$, with t_U determined by the age of the parent node (t_p) and t_L by the age of the oldest daughter node: $t_L \geq \max(t_u, t_v)$. In addition, if the two daughter nodes are in different populations (with $U \neq V$), t_L must be older than the youngest common ancestor of populations U and V on the species tree.

Generate the new age t_x^* by sampling around t_x, reflected into the interval (t_L, t_U). The new node x^* has to reside in a population that is descendant to the parent population P and ancestral to the child populations U and V. Among those target populations, we sample one uniformly. Given the sampled population for x^*, we sample the flags for the three branches: p–x^*, x^*–u, and x^*–v. In each case, the two ends of the branch are already assigned a population. This move may cause large changes to the flags even though it does not change the gene-tree topology. For example, consider the change of t_x in figure 2. Node x is in population S, with branch x–v having the flags $l\varnothing$, since the branch passes H_1 from the left and does not pass H_2. Suppose the new age, generated in the interval (t_u, t_p), is $t_x^* > \tau_R$ so that the new node x^* resides in R. The resampled flags for branch x^*–v may be rr, if the new branch passes both H_1 and H_2 from the right. The proposal ratio is given by the probabilities of sampling the flags at the H nodes.

Step 2. SPR move to change the gene-tree topology. This move cycles through the nonroot nodes on the gene trees. Suppose the node is a. We prune off its parent node y. The remaining part of the gene tree is called the backbone. We sample the new age (t_y^*) before reattaching the subtree y–a onto the backbone. This move always changes the node age t_y but may not change the gene-tree topology.

First, we determine the bounds on the age of reattachment point: $t_L < t_y^* < t_U$. The maximum age is unbounded, whereas the minimum is $t_L \geq t_a$. However, if there are no branches on the backbone passing the population of node *a*, the reattachment point has to be in an ancestral species (in which there exists at least one branch on the backbone) and t_L has to be greater. For example, clade *a* in figure 2 resides in population *B*, and if we prune off clade *a*, there will still be branches in *B* on the backbone for reattachment. In contrast, clade a' resides in population H_{1r}, but if we prune off y'–a', there will be no branches in population H_{1r} for reattachment and the youngest ancestor of H_{1r} with branches on the backbone is *V*, so that $t_L = \tau_V$.

We generate a new age t_y^* around the current age (t_y), reflected into the interval (t_L, t_U) if necessary, and then reattach *y* and clade *a* to a branch on the backbone at time t_y^*. A feasible target branch should cover t_y^* and should at time t_y^* be in a population ancestral to the population of *a*. We sample a target branch at random, either uniformly or with weights determined using local likelihoods.

In case the new branch y^*–*a* passes hybridization nodes, we sample the flags at each hybridization node, as in step 1. Suppose we prune off branch y'–a' in figure 2 and the new age is $t_y'^* > \tau_R$. Then, we let branch y'^*–a' go through H_{2l} or H_{2r} according to their probabilities. The proposal ratio is given by the number of target branches for reattachment and the probabilities for sampling the flags.

Step 3. Change θs on the species tree using a sliding window. This step is the same as in Rannala and Yang (2003).

Step 4. Change τs on the species tree using a variant of the rubber-band algorithm (Rannala and Yang 2003). We generate a new age (τ^*) around the current age, reflected into the interval (τ_L, τ_U), determined using the ages of the parent nodes and daughter nodes on the species tree. Next we change the ages of the affected nodes on the gene trees using the rubber-band algorithm. An affected node has age in the interval (τ_L, τ_U) and resides in the current population (with age τ) or the two daughter populations (if a speciation node is changed), or in the two current populations (H_l and H_r) and the daughter population (if an *H* node is changed). For example, to change τ_S, the bounds are (τ_{H_2}, τ_R), and the affected nodes on the gene tree of figure 2 are in species *S*, *U*, and H_2. These are *x*, *u*, and *w*. To change τ_{H_2} the bounds are (τ_{H_1}, τ_V) and the affected nodes are in species H_{2l}, H_{2r}, and H_{1r}, and are a'. The proposal for changing node ages on the gene tree given the bounds is as in (Rannala and Yang 2003, eqs. A7 and A8).

Step 5. Rescale all node ages on the species tree and on the gene trees using a mixing step (a multiplier) (Rannala and Yang 2003).

Step 6. Change the introgression probability φ for each introgression event using a sliding window. This step affects the gene-tree density, but not the sequence likelihood.

The sliding window used in BPP is the Bactrian move with the triangle kernel (Yang and Rodriguez 2013; Thawornwattana et al. 2018b). Step lengths are adjusted automatically during the burn-in, to achieve an acceptance rate of ~30% (Yang and Rodriguez 2013).

Simulation Study

We conducted three sets of simulations. The first set includes multiple sequences from each species and examines BPP estimation of parameters in the MSci model and the impact of the number of loci, the introgression probability φ, and the species tree model. The second set compares BPP with two summary methods: SNAQ (Solis-Lemus and Ane 2016; Solis-Lemus et al. 2017) and HYDE (Blischak et al. 2018), using one sequence per species. The third set explores the performance of BPP when the model is unidentifiable using SNAQ and HYDE.

For the first set of simulations, multilocus data sets were simulated under the MSci models A and C of figure 1 and then analyzed using BPP to examine the precision and accuracy of parameter estimation. For model A, we used $\tau_R = 0.03$, $\tau_S = 0.02$, $\tau_T = 0.02$, and $\tau_H = 0.01$. For model C, we used $\tau_R = 0.03$ and $\tau_S = \tau_T = \tau_H = 0.01$. We used two values of φ (0.1 and 0.5) and two values of θ (0.001 and 0.01), applied to all populations. Each data set consisted of 10, 100, or 1,000 loci, and at each locus, 10 sequences were sampled from each species (with 30 sequences in total). The sequence length was 500 sites.

Data were generated using the "simulate" option of BPP. Gene trees with branch lengths (coalescent times) were simulated under the MSci model. Then, sequences were "evolved" along the branches of the gene tree according to either the JC (Jukes and Cantor 1969) or the GTR + Γ (Yang 1994a, 1994b) models, and the sequences at the tips of the gene tree constituted the data at the locus. In the GTR + Γ model, the GTR parameters varied among loci according to estimates obtained for chromosomal arm 2L from the *A. gambiae* species complex (Thawornwattana et al. 2018a). The base-frequency parameters were generated from a Dirichlet distribution $(\pi_T, \pi_C, \pi_A, \pi_G) \sim \text{Dir}(25.18, 20.50, 25.22, 20.38)$. The GTR exchangeability parameters (Yang 1994a) were $(a, b, c, d, e, f) \sim \text{Dir}(7.59, 3.23, 2.95, 2.93, 2.93, 7.57)$. The overall rates for loci varied according to a gamma distribution *G*(5, 5), whereas the rates for sites at the same locus varied according to the gamma distribution with mean one, $G(\alpha, \alpha)$ (Yang 1994b), with the shape parameter α sampled from *G*(20, 4).

The number of replicates was 10. Thus, with two trees (*A* and *C*), two φ values (0.1 and 0.5), two θ values (0.001 and 0.01), two mutation models (JC and GTR + Γ), and three data sizes (10, 100, and 1,000 loci), a total of $480 = 2 \times 2 \times 2 \times 2 \times 3 \times 10$ replicate data sets were generated.

Each data set was analyzed using BPP. The JC model was always assumed whether the data were simulated under JC or GTR + Γ. Inverse-gamma priors were assigned on parameters θ and τ_0 (the root age), with the shape parameter 3 and the prior mean equal to the true value: IG(3, 0.02) for $\theta = 0.01$ and IG(3, 0.002) for $\theta = 0.001$, and $\tau_0 \sim$ IG(3, 0.06). The inverse-gamma distribution with shape parameter $\alpha = 3$ has the coefficient of variation 1 and constitutes a diffuse prior. The uniform prior $\mathbb{U}(0, 1)$ was used for φ.

Pilot runs were used to determine the suitable chain length, and then the same settings (such as the burn-in, the number of MCMC iterations, and the sampling frequency) were used to analyze all replicates. Convergence was assessed

by running the same analysis multiple times and confirming consistency between runs (Yang 2015; Flouri et al. 2018).

The second set of simulation was to compare BPP with summary methods. Most methods are designed to test for the presence of gene flow (hybridization or migration) (Degnan 2018). Here, we used two methods that can estimate the introgression probability under a fixed introgression model: SNaQ (Solis-Lemus and Ane 2016) implemented in the program PhyloNetworks (Solis-Lemus et al. 2017) and HyDe (Blischak et al. 2018). The basic algorithms for SNaQ and HyDe are formulated for the case of three species with one or two outgroup species used to root the tree. SNaQ uses the proportions of the three gene-tree topologies, based on the observation that the probabilities for the two mismatching gene trees (which have different topologies from the species tree) are the same if there is deep coalescent but no gene flow while they are different if there is gene flow as well (Yu et al. 2014). HyDe uses the proportions of the three parsimony-informative site patterns pooled across loci or genomic regions (*xxyy*, *xyxy*, and *xyyx*), based on the observation that the probabilities for the two "mismatching" site patterns (*xyxy* and *xyyx*) are the same if there is deep coalescent but no gene flow while these are different if there is gene flow as well (Green et al. 2010).

We used model A of figure 1, plus two outgroup species *D* and *E*, to simulate $L = 10$, 100, or 1,000 loci, with one sequence per species per locus. The data were then analyzed using SNaQ and HyDe, as well as BPP. The JC model was used both to simulate and to analyze the data. For SNaQ, gene trees were inferred using RAxML (Stamatakis et al. 2012). For BPP, the point estimates (posterior means) of φ were used for comparison even though estimates for all parameters, with CIs, were produced.

The third set of simulation explores the performance of BPP under models that are unidentifiable using SNaQ and HyDe. An MSci model may be identifiable given the gene trees with coalescent times but unidentifiable given gene-tree topologies only (Degnan 2018). We simulated and analyzed data using BPP under two models: model D of figure 1 with bidirectional introgression between species *A* and *B* and the model of figure 2 (referred to as model 2H), with three species and two introgression events. Under model D, there is only one gene tree between two species so that its frequency is uninformative and SNaQ is not applicable, and nor is HyDe. Under model 2H, frequencies of three gene trees or three site patterns cannot be used to estimate two introgression probabilities and two internal branch lengths: It is thus impossible to apply SNaQ and HyDe to such data.

For model D, we used the following parameter values: $\tau_R = 0.01, \tau_X = \tau_Y = 0.005, \varphi_X = 0.1, \varphi_Y = 0.3$, and $\theta = 0.01$ for all populations. We simulated 10 replicate data sets, each of 10, 100, or 1,000 loci. At each locus, we sampled 10 sequences per species (20 sequences in total), with the sequence length to be 500. The JC mutation model was used both to simulate and to analyze data by BPP. Note that there is an interesting identifiability issue (or label-switching issue) with model D, such that the two sets of parameters $\Theta = (\theta_A, \theta_B, \theta_R, \theta_X, \theta_Y, \tau_R, \tau_X, \varphi_X, \varphi_Y)$ and $\Theta' = (\theta_A, \theta_B, \theta_R, \theta_Y, \theta_X, \tau_R, \tau_X, 1-\varphi_X, 1-\varphi_Y)$ are unidentifiable (see Discussion). Thus, an identifiability constraint should be applied, such as $\varphi_X < 0.5$. We ran the MCMC without any constraint, but the MCMC sample was preprocessed, with Θ replaced by Θ' if the sampled value for $\varphi_X > 0.5$, before the posterior summary was generated.

For model 2H (fig. 2), the following parameter values were used: $\tau_R = 0.04, \tau_S = 0.03, \tau_U = 0.02, \tau_V = 0.03, \tau_{H_1} = 0.01, \tau_{H_2} = 0.02, \varphi_{H_1} = 0.1, \varphi_{H_2} = 0.5$, and $\theta = 0.01$ for all populations. As above, 10 sequences per species were generated, with 30 sequences per locus. The sequence length was 500. The JC model was used both to simulate and to analyze the data.

Analysis of the Purple Cone Spruce Data Sets

We reanalyzed sequence data concerning the origin of the purple cone spruce *Picea purpurea* (P) from the Qinghai–Tibet Plateau, hypothesized to have originated through homoploid hybridization between *P. wilsonii* (W) and *P. likiangensis* (L) (Sun et al. 2014). The data were generated by Sun et al. (2014). To make the computation feasible for the *BEAST program, Zhang et al. (2018) constructed and analyzed two nonoverlapping data subsets (data sets 1 and 2), each with 40, 30, and 30 phased sequences for P, W, and L, respectively, at 11 autosomal loci. We analyzed these data sets for comparison with the analysis of Zhang et al. (2018) using *BEAST. We also used BPP to analyze the "Full" data set from which data sets 1 and 2 were sampled, with 112, 100, and 120 sequences per locus for the same 11 loci.

The species tree of figure 3 was assumed (Sun et al. 2014). The priors were $\tau_0 \sim \text{IG}(3, 0.004)$, $\theta \sim \text{IG}(3, 0.003)$, and $\varphi \sim \mathbb{U}(0, 1)$. Rates for loci were either constant or had a Dirichlet distribution with $\alpha = 2$ (Burgess and Yang 2008). We used a burn-in of 32,000 iterations and took 10^5 samples, sampling every 10 iterations. The program was run at least twice for each analysis, to check for consistency between runs. Each run (on a single core) took ~5 days.

Marginal likelihood for models A–C (fig. 1) was calculated using thermodynamic integration with Gaussian quadrature (Lartillot and Philippe 2006; Rannala and Yang 2017), with 16 quadrature points.

Analysis of the Budding Yeast Data Set

We analyzed a budding yeast data set with 106 loci and five species: *Saccharomyces cerevisiae* (Scer), *Saccharomyces paradoxus* (Spar), *Saccharomyces mikatae* (Smik), *Saccharomyces kudriavzevii* (Skud), and *Saccharomyces bayanus* (Sbay). This is a subset of the data set published by Rokas et al. (2003) and previously analyzed by Wen and Nakhleh (2018). The species tree or MSci model is shown in figure 4, with a Skud → Sbay introgression. We used inverse-gamma priors IG(3, 0.04) for θs and IG(3, 0.2) for τ_0, and $\varphi \sim \mathbb{U}(0, 1)$.

Analysis of the Genomic Data from the *A. gambiae* Species Complex

We used the coding and noncoding loci compiled by Thawornwattana et al. (2018a) from the genomic sequences for six species in the *A. gambiae* species complex: *A. gambiae*

(G), *A. coluzzii* (C), *A. arabiensis* (A), *A. melas* (L), *A. merus* (R), and *A. quadriannulatus* (Q) (Fontaine et al. 2015). There are 12 sequences per locus, with two sequences per species.

We analyzed blocks of 100 loci, as in Thawornwattana et al. (2018a), and then combined loci for each of the eight chromosomal arms/regions: 2L1, 2La (the inversion region on 2L), 2L2, 2R, 3L1, 3La (the inversion region on 3L), 3L2, and 3R. Since our objective was to estimate the introgression probability for the autosomes, the X chromosome was not used. The species tree is in figure 5, from Thawornwattana et al. (2018a, figure 6). The priors were $\tau_0 \sim \text{IG}(3, 0.2)$ with mean 0.1 for the age of the root, $\theta \sim \text{IG}(3, 0.04)$ with mean 0.02, and $\varphi \sim \mathbb{U}(0, 1)$. We used a burn-in of 16,000 iterations, and took 5×10^5 samples, sampling every 2 iterations. Pilot runs suggest that this generates ESS $>$ 1,000. Each analysis of the block took a few hours, whereas the analysis of the 12 large combined data sets of table 1 each took 1–2 weeks.

For comparison, we used the ML program 3s (Zhu and Yang 2012; Dalquen et al. 2017) to estimate the migration rate $M = Nm$ under the IM model. The implementation assumes three species (1, 2, and 3, say), with three sequences per locus. We sampled three sequences, with half of the loci having the "123" configuration, a quarter with "113," and another quarter with "223" (Thawornwattana et al. 2018a). We generated two replicate data sets by sampling the GAR and RQO triplets to estimate the migration rate $M_{A\to G}$ and $M_{R\to Q}$ (fig. 5). Although limited to three sequences, 3s can use tens of thousands of loci and each run took a few minutes.

Software Availability

The MCMC algorithms described in the article are implemented in BPP Version 4 (Yang 2015; Flouri et al. 2018), available at https://github.com/bpp. The python3 code and scripts for simulating and analyzing the sequence data and for making plots using ggplot are available at https://github.com/brannala/NetworkMSCSimulations.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank Luay Nakhleh, Melisa Olave, Axel Meyer, and two anonymous reviewers for constructive comments. This study has been supported by Biotechnological and Biological Sciences Research Council grants (BB/N000609/1 and BB/P006493/1) to Z.Y. and a BBSRC equipment grant (BB/R01356X/1).

Appendix: Extended Newick Notation for the MSci Model

We use the extended Newick notation (Cardona et al. 2008) to represent the MSci model in the BPP program. The parenthesis notation "$(A, B)S$" specifies two branches from the speciation node S to two daughter species A and B, whereas "$(A)H$" specifies one branch from H to A. Every branch is represented once. Each tip species occurs once. Internal nodes for speciation nodes may and may not be labeled, but hybridization (H) nodes must be labeled. In models A–C of figure 1, each H node occurs twice in the notation, once as a label for an ancestral node and another time as a tip, and the introgression probability φ is identified with the ancestral node (whereas its "mirror" tip node has $1 - \varphi$). Thus, models A–C of figure 1 are represented as $((A, (C)H)S, (H, B)T)R$, with parameter φ assigned to the SH branch and $1 - \varphi$ to the TH branch. The extended Newick notation is not unique. The representation $((A, H)S, ((C)H, B)T)R$ specifies an equivalent model, with parameter φ assigned to branch TH and $1 - \varphi$ to branch SH.

The three types of models in figure 1 (A–C) are distinguished using the metadata variable "tau-parent," which is assigned the value "yes" or "no" depending on whether the parent node has an age (τ) distinct from that of the hybridization node. Thus, models A–C of figure 1 are represented as

$$\begin{aligned}(\text{A}) &: ((A, (C)H[\&\text{tau-parent} = \text{yes}])\text{S},\\ &\quad (\text{H}[\&\text{tau-parent} = \text{yes}], \text{B})\text{T})\text{R}.\end{aligned}$$

$$\begin{aligned}(\text{B}) &: ((A, (C)H[\&\text{tau-parent} = \text{no}])\text{S},\\ &\quad (\text{H}[\&\text{tau-parent} = \text{yes}], \text{B})\text{T})\text{R}.\end{aligned}$$

$$\begin{aligned}(\text{C}) &: ((A, (C)H[\&\text{tau-parent} = \text{no}])\text{S},\\ &\quad (\text{H}[\&\text{tau-parent} = \text{no}], \text{B})\text{T})\text{R}.\end{aligned}$$

Model D (bidirectional introgression) differs from models A–C in that each of nodes X and Y has two parent nodes and two daughter nodes. The model is represented as $((A, (B)Y)X, (X)Y)R$. The notation is again not unique, and equivalent representations include $(((A)X, B)Y, (Y)X)R$ or more concisely, $((X, B)Y, (A, Y)X)R$. In both notations, the φ parameter is assigned to the branch with an older parent, whereas the horizontal branch has $1 - \varphi$.

As a more complex example, the species graph for the *Anopheles* mosquitoes of figure 5 is represented as

$$\begin{aligned}&((R, (Q)h[\&\text{tau-parent} = \text{no}])g, (f[\&\text{tau-parent} = \text{yes}],\\ &\quad (((((G, C)b)f[\&\text{tau-parent} = \text{no}], A)e,\\ &\quad h[\&\text{tau-parent} = \text{yes}])d, L)c)a)o.\end{aligned}$$

References

Arnold BJ, Lahner B, DaCosta JM, Weisman CM, Hollister JD, Salt DE, Bomblies K, Yant L. 2016. Borrowed alleles and convergence in serpentine adaptation. *Proc Natl Acad Sci U S A.* 113(29):8320–8325.

Blischak PD, Chifman J, Wolfe AD, Kubatko LS. 2018. Hyde: a python package for genome-scale hybridization detection. *Syst Biol.* 67(5):821–829.

Burgess R, Yang Z. 2008. Estimation of hominoid ancestral population sizes under Bayesian coalescent models incorporating mutation rate variation and sequencing errors. *Mol Biol Evol.* 25(9):1979–1994.

Cao Z, Liu X, Ogilvie HA, Yan Z, Nakhleh L. 2019. Practical aspects of phylogenetic network analysis using phylonet. BioRxiv. DOI: dx.doi.org/10.1101/746362

Cardona G, Rossello F, Valiente G. 2008. Extended Newick: it is time for a standard representation of phylogenetic networks. *BMC Bioinformatics* 9(1):532.

Chan YC, Roos C, Inoue-Murayama M, Inoue E, Shih CC, Pei KJ, Vigilant L. 2013. Inferring the evolutionary histories of divergences in *Hylobates* and *Nomascus* gibbons through multilocus sequence data. *BMC Evol Biol*. 13(1):82.

Dalquen DA, Zhu T, Yang Z. 2017. Maximum likelihood implementation of an isolation-with-migration model for three species. *Syst Biol*. 66(3):379–398.

Degnan JH. 2018. Modeling hybridization under the network multispecies coalescent. *Syst Biol*. 67(5):786–799.

Durand EY, Patterson N, Reich D, Slatkin M. 2011. Testing for ancient admixture between closely related populations. *Mol Biol Evol*. 28(8):2239–2252.

Edwards SV, Xi Z, Janke A, Faircloth BC, McCormack JE, Glenn TC, Zhong B, Wu S, Lemmon EM, Lemmon AR, et al. 2016. Implementing and testing the multispecies coalescent model: a valuable paradigm for phylogenomics. *Mol Phylogenet Evol*. 94(Pt A): 447–462.

Ellegren H, Smeds L, Burri R, Olason PI, Backstrom N, Kawakami T, Kunstner A, Makinen H, Nadachowska-Brzyska K, Qvarnstrom A, et al. 2012. The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature* 491(7426):756–760.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*. 17(6):368–376.

Flouri T, Jiao X, Rannala B, Yang Z. 2018. Species tree inference with BPP using genomic sequences and the multispecies coalescent. *Mol Biol Evol*. 35(10):2585–2593.

Folk RA, Soltis PS, Soltis DE, Guralnick R. 2018. New prospects in the detection and comparative analysis of hybridization in the tree of life. *Am J Bot*. 105(3):364–375.

Fontaine MC, Pease JB, Steele A, Waterhouse RM, Neafsey DE, Sharakhov IV, Jiang X, Hall AB, Catteruccia F, Kakani E, et al. 2015. Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science* 347(6217):1258524.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH, et al. 2010. A draft sequence of the Neandertal genome. *Science* 328(5979):710–722.

Harrison RG, Larson EL. 2014. Hybridization, introgression, and the nature of species boundaries. *J Hered*. 105(S1):795–809.

Hey J. 2010. Isolation with migration models for more than two populations. *Mol Biol Evol*. 27(4):905–920.

Hey J, Chung Y, Sethuraman A, Lachance J, Tishkoff S, Sousa VC, Wang Y. 2018. Phylogeny estimation by integration over isolation with migration models. *Mol Biol Evol*. 35(11):2805–2818.

Hey J, Nielsen R. 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* 167(2):747–760.

Huson DH, Rupp R, Cornavacca C. 2011. Phylogenetic networks: concepts, algorithms and applications. Cambridge: Cambridge University Press.

Jackson ND, Carstens BC, Morales AE, O'Meara BC. 2017. Species delimitation with gene flow. *Syst Biol*. 66(5):799–812.

Jones GR. 2019. Divergence estimation in the presence of incomplete lineage sorting and migration. *Syst Biol*. 68(1):19–31.

Jukes T, Cantor C. 1969. Evolution of protein molecules. In: Munro H, editor. Mammalian protein metabolism. New York: Academic Press. p. 21–123.

Kubatko LS. 2009. Identifying hybridization events in the presence of coalescence via model selection. *Syst Biol*. 58(5):478–488.

Kumar V, Lammers F, Bidon T, Pfenninger M, Kolter L, Nilsson MA, Janke A. 2017. The evolutionary history of bears is characterized by gene flow across species. *Sci Rep*. 7:46487.

Lartillot N, Philippe H. 2006. Computing Bayes factors using thermodynamic integration. *Syst Biol*. 55(2):195–207.

Leaché AD, Zhu T, Rannala B, Yang Z. 2019. The spectre of too many species. *Syst Biol*. 68(1):168–181.

Liu S, Lorenzen ED, Fumagalli M, Li B, Harris K, Xiong Z, Zhou L, Korneliussen TS, Somel M, Babbitt C, et al. 2014. Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. *Cell* 157(4):785–794.

Lohse K, Chmelik M, Martin SH, Barton NH. 2016. Efficient strategies for calculating blockwise likelihoods under the coalescent. *Genetics* 202(2):775–786.

Mallet J, Besansky N, Hahn MW. 2016. How reticulated are species? *Bioessays* 38(2):140–149.

Mao Y, Economo EP, Satoh N. 2018. The roles of introgression and climate change in the rise to dominance of *Acropora* corals. *Curr Biol*. 28(21):3373–3382.e5.

Martin SH, Dasmahapatra KK, Nadeau NJ, Salazar C, Walters JR, Simpson F, Blaxter M, Manica A, Mallet J, Jiggins CD. 2013. Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Res*. 23(11):1817–1828.

Martin SH, Jiggins CD. 2017. Interpreting the genomic landscape of introgression. *Curr Opin Genet Dev*. 47:69–74.

Nielsen R, Akey JM, Jakobsson M, Pritchard JK, Tishkoff S, Willerslev E. 2017. Tracing the peopling of the world through genomics. *Nature* 541(7637):302–310.

O'Hagan A, Forster J. 2004. Kendall's advanced theory of statistics: Bayesian inference. London: Arnold.

Pardi F, Scornavacca C. 2015. Reconstructible phylogenetic networks: do not distinguish the indistinguishable. *PLoS Comput Biol*. 11(4):e1004135.

Rannala B, Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* 164(4):1645–1656.

Rannala B, Yang Z. 2017. Efficient Bayesian species tree inference under the multispecies coalescent. *Syst Biol*. 66(5):823–842.

Rokas A, Williams B, King N, Carroll S. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425(6960):798–804.

Shi C, Yang Z. 2018. Coalescent-based analyses of genomic sequence data provide a robust resolution of phylogenetic relationships among major groups of gibbons. *Mol Biol Evol*. 35(1):159–179.

Slotman MA, Della Torre A, Calzetta M, Powell JR. 2005. Differential introgression of chromosomal regions between *Anopheles gambiae* and *An. arabiensis*. *Am J Trop Med Hyg*. 73(2):326–335.

Solis-Lemus C, Ane C. 2016. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS Genet*. 12(3):e1005896.

Solis-Lemus C, Bastide P, Ane C. 2017. PhyloNetworks: a package for phylogenetic networks. *Mol Biol Evol*. 34(12):3292–3298.

Stamatakis A, Aberer A, Goll C, Smith S, Berger S, Izquierdo-Carrasco F. 2012. RAxML-Light: a tool for computing terabyte phylogenies. *Bioinformatics* 28(15):2064–2066.

Sun Y, Abbott RJ, Li L, Li L, Zou J, Liu J. 2014. Evolutionary history of purple cone spruce (*Picea purpurea*) in the Qinghai–Tibet Plateau: homoploid hybrid origin and Pleistocene expansion. *Mol Ecol*. 23(2):343–359.

Thawornwattana Y, Dalquen D, Yang Z. 2018a. Coalescent analysis of phylogenomic data confidently resolves the species relationships in the *Anopheles gambiae* species complex. *Mol Biol Evol*. 35(10):2512–2527.

Thawornwattana Y, Dalquen D, Yang Z. 2018b. Designing simple and efficient Markov chain Monte Carlo proposal kernels. *Bayesian Anal*. 13(4):1037–1059.

Wen D, Nakhleh L. 2018. Coestimating reticulate phylogenies and gene trees from multilocus sequence data. *Syst Biol*. 67(3):439–457.

Wen D, Yu Y, Nakhleh L. 2016. Bayesian inference of reticulate phylogenies under the multispecies network coalescent. *PLoS Genet*. 12(5):e1006006.

Wen D, Yu Y, Zhu J, Nakhleh L. 2018. Inferring phylogenetic networks using phylonet. *Syst Biol*. 67(4):735–740.

Wu D-D, Ding X-D, Wang S, Wojcik JM, Zhang Y, Tokarska M, Li Y, Wang M-S, Faruque O, Nielsen R, et al. 2018. Pervasive introgression

facilitated domestication and adaptation in the *Bos* species complex. *Nat Ecol Evol*. 2(7):1139–1145.

Xu B, Yang Z. 2016. Challenges in species tree estimation under the multispecies coalescent model. *Genetics* 204(4): 1353–1368.

Yang Z. 1994a. Estimating the pattern of nucleotide substitution. *J Mol Evol*. 39(1):105–111.

Yang Z. 1994b. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol*. 39(3):306–314.

Yang Z. 2015. The BPP program for species tree estimation and species delimitation. *Curr Zool*. 61(5):854–865.

Yang Z, Rodriguez CE. 2013. Searching for efficient Markov chain Monte Carlo proposal kernels. *Proc Natl Acad Sci U S A*. 110(48):19307–19312.

Yu Y, Degnan JH, Nakhleh L. 2012. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *PLoS Genet*. 8(4):e1002660.

Yu Y, Dong J, Liu KJ, Nakhleh L. 2014. Maximum likelihood inference of reticulate evolutionary histories. *Proc Natl Acad Sci U S A*. 111(46):16448–16453.

Zhang C, Ogilvie HA, Drummond AJ, Stadler T. 2018. Bayesian inference of species networks from multilocus sequence data. *Mol Biol Evol*. 35(2):504–517.

Zhu S, Degnan JH. 2017. Displayed trees do not determine distinguishability under the network multispecies coalescent. *Syst Biol*. 66(2):283–298.

Zhu T, Yang Z. 2012. Maximum likelihood implementation of an isolation-with-migration model with three species for testing speciation with gene flow. *Mol Biol Evol*. 29(10):3131–3142.