Syst. Biol. 69(5):830–847, 2020

DOI:10.1093/sysbio/syaa001
Advance Access publication January 24, 2020

The Impact of Cross-Species Gene Flow on Species Tree Estimation

XIYUN JIAO[1], TOMÁŠ FLOURI[1], BRUCE RANNALA[2], AND ZIHENG YANG[1,*]

[1]*Department of Genetics, University College London, Gower Street, London WC1E 6BT, UK and* [2]*Department of Evolution and Ecology, University of California, Davis, CA 95616, USA*

[*]*Correspondence to be sent to: Department of Genetics, University College London, Gower Street, London WC1E 6BT, UK; E-mail: z.yang@ucl.ac.uk.*

Received 30 June 2018; reviews returned 12 November 2019; accepted 15 January 2020
Associate Editor: Peter Foster

Abstract.—Recent analyses of genomic sequence data suggest cross-species gene flow is common in both plants and animals, posing challenges to species tree estimation. We examine the levels of gene flow needed to mislead species tree estimation with three species and either episodic introgressive hybridization or continuous migration between an outgroup and one ingroup species. Several species tree estimation methods are examined, including the majority-vote method based on the most common gene tree topology (with either the true or reconstructed gene trees used), the UPGMA method based on the average sequence distances (or average coalescent times) between species, and the full-likelihood method based on multilocus sequence data. Our results suggest that the majority-vote method based on gene tree topologies is more robust to gene flow than the UPGMA method based on coalescent times and both are more robust than likelihood assuming a multispecies coalescent (MSC) model with no cross-species gene flow. Comparison of the continuous migration model with the episodic introgression model suggests that a small amount of gene flow per generation can cause drastic changes to the genetic history of the species and mislead species tree methods, especially if the species diverged through radiative speciation events. Estimates of parameters under the MSC with gene flow suggest that African mosquito species in the *Anopheles gambiae* species complex constitute such an example of extreme impact of gene flow on species phylogeny. [IM; introgression; migration; MSci; multispecies coalescent; species tree.]

Cross-species hybridization or introgression has long been recognized as an important process that generates biological diversity in plants (e.g., Anderson 1949; Mallet 2007). Analyses of genomic data in the past few years suggest that introgression is also common in animals (Ellegren et al. 2012; Chan et al. 2013; Kumar et al. 2017; Mao et al. 2018; Wu et al. 2018), including humans and their close relatives (Nielsen et al. 2017). Introgression may involve either sister or nonsister species (e.g., Mallet et al. 2016) and may play an important role in the speciation process (Mallet et al. 2016; Martin and Jiggins 2017). Introgression, together with deep coalescence (also known as incomplete lineage sorting), may cause difficulties for species tree reconstruction (Maddison 1997; Nichols 2001). In extreme cases, the whole genome, and in particular the autosomes, are affected by such pervasive gene flow that they do not reflect the species phylogeny anymore. This appears, for example, to be the case with the *Anopheles gambiae* species complex, in which the autosomes suggest different species relationships from the X chromosome, which, being apparently enriched with sterility genes and resistant to cross-species gene flow, reflects the true history of species divergences (Fontaine et al. 2015). The *Heliconius* butterflies appear to be another such case, with the Z chromosome favoring different phylogenies from the autosomes (Edelman et al. 2019). In both examples, the species arose through a rapid succession of speciation events, generating species phylogenies with very short interior branches, which are challenging to reconstruct even without cross-species gene flow.

A number of methods have been developed to detect gene flow across species using genetic sequence data, including population genetic methods based on F_{st} and summary methods that make use of observed site patterns (Green et al. 2010; Durand et al. 2011) or estimated gene tree topologies (Yu et al. 2012, 2014; Yu and Nakhleh 2015; Solis-Lemus and Ane 2016; Wen et al. 2016). Full-likelihood methods based on sequence alignments (Hey and Nielsen 2004; Hey et al. 2018; Wen and Nakhleh 2018; Zhang et al. 2018) are also being actively developed. See Degnan (2018) and Folk et al. (2018) for recent reviews.

Here, we consider the question of how much gene flow is sufficient to mislead species tree estimation methods that accommodate the coalescent process but not gene flow. As discussed by Folk et al. (2018), the impact of gene flow on the tree of life is an important topic worth serious study. We focus on the case of three species with sequences evolving under the molecular clock and are specifically interested in closely related species, for which gene flow may be a major concern. We consider two distinct models of gene flow, both of which accommodate the multispecies coalescent (MSC). The model of isolation with migration (IM) assumes continuous migration, with the species exchanging migrants at a certain rate in every generation (see Fig. 1b, d for examples) (Hey and Nielsen 2004; Hey 2010). The model of multispecies coalescent with introgression (MSci) assumes episodic introgression or hybridization, with introgressions occurring at fixed time points in the past (e.g., Fig. 1a, c) (Yu et al. 2012, 2014; Wen and Nakhleh 2018). The MSci model was called the "multispecies network coalescent" by Wen and Nakhleh (2018). As the term "network" has been used to refer to both biological and nonbiological processes (Solis-Lemus and Ane 2016; Degnan 2018), we follow Degnan (2018) and use the more-expressive term "MSC with

introgression" to highlight the two major factors in the model: coalescent and introgression. We consider two scenarios of gene flow on the species tree $(A,(B,C))$. Inflow is introgression (or migration) from the outgroup species A to the ingroup species C, and outflow is in the reverse direction. In both scenarios, gene flow makes species A and C look similar, potentially misleading species tree methods to infer the incorrect tree $(B,(C,A))$.

Previously Leaché et al. (2014) used computer simulation to examine the impact of continuous migration on species tree inference, finding that the effects depend on the species tree topology and the mode of gene flow. For example, while gene flow between nonsister species might mislead species tree inference, gene flow between sister species actually made it easier to infer the species tree. Solis-Lemus et al. (2016) used simulation to study the statistical consistency of several species tree methods that ignore gene flow, including ASTRAL (Mirarab et al. 2014) and NJST (Liu and Yu 2011), when the model assumes cross-species introgression. A method was considered inconsistent if the probability of retrieving the correct species tree fails to increase when the number of gene trees (the number of loci) increases. The approach we take here is largely analytical, not affected by sampling errors in simulation. Hahn and Nakhleh (2016, Fig. 3) calculated gene tree probabilities under an introgression model, arguing that the concept of a species tree is poorly defined when there is gene flow. Long and Kubatko (2018) studied the probabilities of gene tree topologies under a model of isolation with initial migration (Wilkinson-Herbots 2012) for three species, in which there is initial gene flow between sister species after their divergence. One might expect gene flow between sister species to make them appear more similar, making it easier to infer the species tree, but surprisingly with different population sizes, gene flow between sister species can cause the most probable gene tree topology to differ from the species tree, leading to so-called "anomalous gene trees".

Here we study the asymptotic behavior of several species tree inference methods when there is gene flow affecting nonsister species. All the methods ignore gene flow but are statistically consistent in the case of three species when there is deep coalescence but no gene flow. We consider both continuous migration and episodic introgression. The first method is the majority-vote method of using the most common (true) gene tree topology as the estimate of the species tree. This is known to be statistically consistent in the case of three species and three sequences per locus when there is no gene flow, with the estimated species tree approaching the true tree when the number of loci approaches infinity (Hudson 1983). We derive the distribution of the *true* gene tree topologies under each model of gene flow, and examine the impact of phylogenetic reconstruction errors when the *estimated* gene trees are used to estimate the species tree. Use of estimated gene trees is known to be consistent in the case of three species when the model involves coalescent but no gene flow (Yang 2002). Next we consider the UPGMA method, which uses the average sequence distance between species (or the average coalescent time between species since we assume the molecular clock) to infer the species tree (Liu et al. 2009). This is equivalent to calculating sequence distances using concatenated data followed by UPGMA reconstruction of the species tree. This method is known to be consistent in the case of three species (Liu et al. 2009). Finally, we consider the maximum likelihood (ML) method of species tree estimation, which averages over the gene trees and branch lengths and thus accounts for phylogenetic reconstruction errors (Yang 2002; Xu and Yang 2016). While full-likelihood methods of species tree estimation under the MSC applied to multilocus sequence alignments, including ML (Yang 2002; Zhu and Yang 2012) and Bayesian inference (Liu and Pearl 2007; Heled and Drummond 2010; Yang and Rannala 2014), are analytically intractable, the 3s program has an efficient ML implementation that can handle thousands of loci (Yang 2002; Dalquen et al. 2017). We thus use 3s to analyze tens of thousands of loci to approximate the case of infinite data. Note that our interest is in the consistency or inconsistency of each species tree estimation method in the face of gene flow as the number of loci approaches infinity. To our knowledge, this represents the first analysis of full-likelihood methods based on sequence alignments, which may be expected to make the most efficient use of information in the sequence data and to have statistically optimal performance when the model is correct. The performance of likelihood methods when the model is mis-specified is unknown. As pointed out by Solis-Lemus et al. (2016), statistical consistency is conventionally defined under the true model, but here we follow the tradition in statistical phylogenetics to examine the impact of model violations.

We note that with cross-species gene flow, the concept of the true species tree may be ambiguous (Hahn and Nakhleh 2016). One strategy, adopted in PhyloNetworks (Solis-Lemus et al. 2017), is to use the introgression probability to define the *major species tree*, which is the species tree represented by parental branches at the hybridization nodes with contribution probabilities $> \frac{1}{2}$. For example, in Figure 1a, the true species tree is $(A(CB))$ if $\varphi < \frac{1}{2}$ and $((AC)B)$ if $\varphi > \frac{1}{2}$. With this definition, the true species tree changes when the introgression probability increases from below 0.5 to above 0.5, with the truth always on the winning side. Another strategy is to start with a species tree and add gene flow onto it, with the true species tree to be always the starting species tree. In this article, we use the second strategy, motivated by the inferred pattern of gene flow in the *Anopheles gambiae* species group (see Discussion section).

THEORY

We consider two gene flow scenarios: "inflow" where there is gene flow (introgression or migration) from the outgroup species A to the ingroup species C on

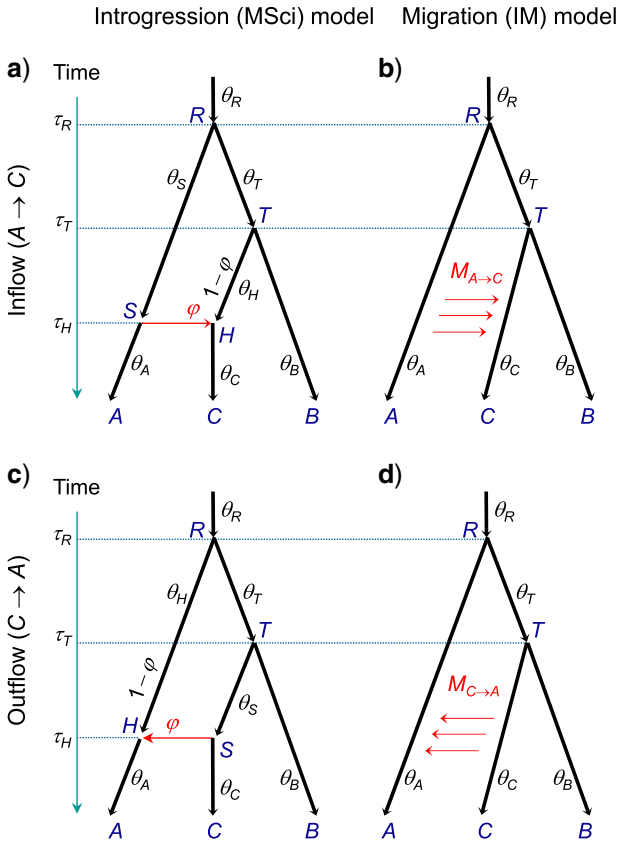

FIGURE 1. Species tree $(A,(B,C))$, with introgression (a and c) or migration (b and d) from the outgroup species A to the ingroup species C (inflow, a and b) or in the reverse direction (outflow, c and d). In the MSci models, an introgression or hybridization event occurs at time $\tau_H = \tau_S$, and the hybrid species H consists of a proportion φ of introgressed individuals. In the IM model, gene flow occurs continuously (i.e., in every generation) over the time period $(0, \tau_T)$. The migration rate in (b) is defined as $M_{A\to C} = N_C m_{A\to C}$, where N_C is (effective) population size of species C and $m_{A\to C}$ is the proportion of immigrants (from species A) in the receiving population C, so that $M_{A\to C}$ is the expected number of migrants from A to C per generation. Migration rate in (d) is defined similarly as $M_{C\to A} = N_A m_{C\to A}$. In this article, the backbone tree represented by black thick branches is considered the true species tree, irrespective of the strength of gene flow (or the values of φ or M).

the species tree $(A,(B,C))$ (Fig. 1a,b) and "outflow" in the opposite direction (Fig. 1c,d). Both the species divergence times (τ_R, τ_S, and τ_T) and the population size parameters (θs) in the model are measured by the expected number of mutations/substitutions per site. The data consist of multiple loci, with three sequences—one from each of the three species at each locus (a,b,c). The possible gene trees at each locus are $G_1 = (a,(b,c))$, which matches the species tree; $G_2 = (b,(c,a))$, which reflects the introgression or migration pattern; and $G_3 = (c,(a,b))$ which matches neither. Sequence data are then analyzed to infer the species tree under the MSC model (assuming no gene flow). We derive analytical results for two simple methods of species tree estimation assuming no gene flow: 1) the majority-vote method for which the correct species tree is inferred if $\mathbb{P}(G_1) > \mathbb{P}(G_2)$ and 2) the UPGMA method using the average coalescent times across loci (Liu et al. 2009) for which the correct species tree is inferred if $\mathbb{E}(t_{bc}) < \mathbb{E}(t_{ac})$. Initially, we ignore sampling errors in the gene trees or the estimated sequence distances. We study the asymptotic behavior of the two methods as the number of loci approaches infinity. We derive results for the two methods under both the MSci and IM models.

We note that the analytical results derived here apply to the specific scenarios of species tree and gene flow. Even for the case of three species, we do not consider gene flow between sister species (Dalquen et al. 2017; Long and Kubatko 2018) or bidirectional gene flow ($A \leftrightarrow C$). Larger species trees with more than three species and more complex gene flow scenarios involving more than one pair of species will add much complexity to the analysis (Zhu et al. 2016; Zhu and Degnan 2017).

Gene Flow under the MSci Model

Here, we consider properties of the majority-vote and UPGMA methods for data arising under an MSci model of admixture. Referring to Figure 1, we define

$$P_S = e^{-\frac{2}{\theta_S}(\tau_R - \tau_S)} \text{ and } P_T = e^{-\frac{2}{\theta_T}(\tau_R - \tau_T)}$$

to be the probability that two sequences entering either species S, or species T, do not coalesce in that species and instead both enter the ancestral species (species R for both). Note that $\frac{2}{\theta_S}(\tau_R - \tau_S)$ and $\frac{2}{\theta_T}(\tau_R - \tau_T)$ are known as the branch lengths in coalescent units in Figure 1.

We first consider properties of the methods with data generated by inflow.

UPGMA method with inflow.— Here, we derive results for a model assuming instantaneous introgression from A to C (inflow), with the introgression probability φ. Sequences a and b can coalesce in R only, and the coalescent time has the exponential density $f(t_{ab}) = \frac{2}{\theta_R} e^{-\frac{2}{\theta_R}(t_{ab} - \tau_R)}$ with expectation $\mathbb{E}(t_{ab}) = \tau_R + \frac{\theta_R}{2}$. Sequences a and c can coalesce in population S as well as R, while sequences b and c can coalesce in species T as well as R. This means that $\mathbb{E}(t_{ab}) > \max\{\mathbb{E}(t_{ac}), \mathbb{E}(t_{bc})\}$. The probability density of the coalescent time t_{ac} is

$$f(t_{ac}) = \begin{cases} \varphi \frac{2}{\theta_S} e^{-\frac{2}{\theta_S}(t_{ac} - \tau_S)}, & \text{if } \tau_S < t_{ac} < \tau_R, \\ \varphi P_S \frac{2}{\theta_R} e^{-\frac{2}{\theta_R}(t_{ac} - \tau_R)} & \\ \quad + (1-\varphi) \frac{2}{\theta_R} e^{-\frac{2}{\theta_R}(t_{ac} - \tau_R)}, & \text{if } t_{ac} > \tau_R, \end{cases} \quad (1)$$

with expectation

$$\mathbb{E}(t_{ac}) = \varphi \left[\tau_S + \frac{\theta_S}{2} + P_S \left(\frac{\theta_R}{2} - \frac{\theta_S}{2} \right) \right] + (1-\varphi) \left(\tau_R + \frac{\theta_R}{2} \right). \quad (2)$$

Similarly, sequences b and c can coalesce in species T and R, so we have the density

$$f(t_{bc})=\begin{cases}(1-\varphi)\frac{2}{\theta_T}e^{-\frac{2}{\theta_T}(t_{bc}-\tau_T)}, & \text{if } \tau_T<t_{bc}<\tau_R,\\ (1-\varphi)P_T\frac{2}{\theta_R}e^{-\frac{2}{\theta_R}(t_{bc}-\tau_R)} & \\ +\varphi\frac{2}{\theta_R}e^{-\frac{2}{\theta_R}(t_{bc}-\tau_R)}, & \text{if } t_{bc}>\tau_R,\end{cases} \quad (3)$$

with expectation

$$\mathbb{E}(t_{bc})=(1-\varphi)\left[\tau_T+\frac{\theta_T}{2}+P_T\left(\frac{\theta_R}{2}-\frac{\theta_T}{2}\right)\right]+\varphi\left[\tau_R+\frac{\theta_R}{2}\right]. \quad (4)$$

Then $\mathbb{E}(t_{bc})>\mathbb{E}(t_{ac})$ and the UPGMA method based on average coalescent times infers an incorrect species tree if and only if

$$\varphi>\frac{1}{1+\frac{\tau_R-\tau_S+\frac{1}{2}(1-P_S)(\theta_R-\theta_S)}{\tau_R-\tau_T+\frac{1}{2}(1-P_T)(\theta_R-\theta_T)}}. \quad (5)$$

Majority-vote method with inflow—. If sequences b and c coalesce in population T, the gene tree will be $G_1=(a,(b,c))$, while if a and c coalesce in population S, the gene tree will be the $G_2=(b,(c,a))$. If neither of those events occurs, both coalescent events for the three sequences will occur in species R and the three gene trees will occur with equal probabilities. Thus $\mathbb{P}(G_3)<\min\{\mathbb{P}(G_1),\mathbb{P}(G_2)\}$. We have

$$\begin{aligned}\mathbb{P}(G_1)&=\frac{1}{3}\varphi P_S+(1-\varphi)\left(1-P_T+\frac{1}{3}P_T\right),\\ \mathbb{P}(G_2)&=\varphi\left(1-P_S+\frac{1}{3}P_S\right)+\frac{1}{3}(1-\varphi)P_T,\\ \mathbb{P}(G_3)&=\frac{1}{3}[\varphi P_S+(1-\varphi)P_T]=1-\mathbb{P}(G_1)-\mathbb{P}(G_2).\end{aligned} \quad (6)$$

For example, gene tree G_1 results from sequences b and c coalescing first. If sequence c enters species S (which happens with probability φ), this can occur only if sequences c and a do not coalesce in species S. This is the first term, $\varphi P_S\cdot\frac{1}{3}$, in $\mathbb{P}(G_1)$. If sequence c enters species H (which happens with probability $1-\varphi$), sequences b and c can coalesce in species T or R. Hence the second term $(1-\varphi)\left(1-P_T+\frac{1}{3}P_T\right)$.

Thus $\mathbb{P}(G_1)<\mathbb{P}(G_2)$ and the majority-vote method based on the most common gene tree infers an incorrect species tree if and only if

$$\varphi>\frac{1-P_T}{2-P_S-P_T}=\frac{1}{1+\frac{1-P_S}{1-P_T}}. \quad (7)$$

This can also be obtained by noting that $\mathbb{P}(G_1)<\mathbb{P}(G_2)$ if and only if the probability that sequences b and c coalesce in population T is smaller than the probability that sequences a and c coalesce in population S: that is, if $(1-\varphi)(1-P_T)<\varphi(1-P_S)$.

Note that the gene tree probabilities (Equation 6) and the φ limit based on them (Equation 7) depend on only the internal branch lengths (in coalescent units) on the species tree, but not the species divergence times and population sizes. The φ limits for both gene tree probabilities and the average coalescent times (Equations 5 and 7) depend on the gene tree topologies and coalescent times but not the mutation rate: they are functions of τ/θ ratios and not of τs and θs individually.

UPGMA method with outflow.— Here, we derive results for a model assuming instantaneous introgression from C to A (outflow) with introgression probability φ. Referring to Figure 1c we define

$$P_S^*=e^{-\frac{2}{\theta_S}(\tau_T-\tau_S)}$$

to be the probability that two sequences entering population S do not coalesce in that population and instead enter its ancestor (population T) (Fig. 1c). Sequences a and c can coalesce in populations S, T, and R, and sequences b and c or a and b can coalesce in species T and R. The probability density of the coalescent time t_{ac} is

$$f(t_{ac})=\begin{cases}\varphi\frac{2}{\theta_S}e^{-\frac{2}{\theta_S}(t_{ac}-\tau_S)}, & \text{if } \tau_S<t_{ac}<\tau_T,\\ \varphi P_S^*\frac{2}{\theta_T}e^{-\frac{2}{\theta_T}(t_{ac}-\tau_T)}, & \text{if } \tau_T<t_{ac}<\tau_R,\\ \left[\varphi P_S^*P_T+(1-\varphi)\right]\frac{2}{\theta_R}e^{-\frac{2}{\theta_R}(t_{ac}-\tau_R)}, & \text{if } t_{ac}>\tau_R,\end{cases} \quad (8)$$

with expectation

$$\begin{aligned}\mathbb{E}(t_{ac})=\ &\varphi\left[\tau_S+\frac{\theta_S}{2}+P_S^\star\left(\frac{\theta_T}{2}-\frac{\theta_S}{2}\right)+P_S^\star P_T\left(\frac{\theta_R}{2}-\frac{\theta_T}{2}\right)\right]\\ &+(1-\varphi)\left[\tau_R+\frac{\theta_R}{2}\right].\end{aligned} \quad (9)$$

The coalescent time between sequences b and c has the density

$$f(t_{bc})=\begin{cases}\frac{2}{\theta_T}e^{-\frac{2}{\theta_T}(t_{bc}-\tau_T)}, & \text{if } \tau_T<t_{bc}<\tau_R,\\ P_T\frac{2}{\theta_R}e^{-\frac{2}{\theta_R}(t_{bc}-\tau_R)}, & \text{if } t_{bc}>\tau_R,\end{cases} \quad (10)$$

with expectation

$$\mathbb{E}(t_{bc})=\tau_T+\frac{\theta_T}{2}+P_T\left(\frac{\theta_R}{2}-\frac{\theta_T}{2}\right). \quad (11)$$

The probability density of the coalescent time t_{ab} is

$$f(t_{ab})=\begin{cases}\varphi\frac{2}{\theta_T}e^{-\frac{2}{\theta_T}(t_{ab}-\tau_T)}, & \text{if } \tau_T<t_{ab}<\tau_R,\\ \left[\varphi P_T+(1-\varphi)\right]\frac{2}{\theta_R}e^{-\frac{2}{\theta_R}(t_{ab}-\tau_R)}, & \text{if } t_{ab}>\tau_R,\end{cases} \quad (12)$$

with the expectation

$$\mathbb{E}(t_{ab})=\varphi\left[\tau_T+\frac{\theta_T}{2}+P_T\left(\frac{\theta_R}{2}-\frac{\theta_T}{2}\right)\right]+(1-\varphi)\left(\tau_R+\frac{\theta_R}{2}\right). \quad (13)$$

It is easy to see that $\mathbb{E}(t_{ab})>\max\{\mathbb{E}(t_{ac}),\mathbb{E}(t_{bc})\}$. Then $\mathbb{E}(t_{bc})>\mathbb{E}(t_{ac})$ and the UPGMA method infers an

incorrect species tree if and only if

$$\varphi > \frac{\tau_R - \tau_T + \frac{1}{2}(1-P_T)(\theta_R - \theta_T)}{\tau_R - \tau_S + \frac{1}{2}(\theta_R - \theta_S) - \frac{1}{2}P_S^\star(\theta_T - \theta_S) - \frac{1}{2}P_T P_S^\star(\theta_R - \theta_T)}. \quad (14)$$

Majority-vote method with outflow—. When we trace the genealogy of sequences a, b, and c backwards in time, sequence a may enter species S (with probability φ) or species H (with probability $1-\varphi$). Consider the first case, of sequence a entering species S. If a and c coalesce in population S, the gene tree will be G_2. Otherwise, the two coalescent events for the three sequences can occur in either T or R and the three gene trees occur with equal probabilities. In the second case, sequence a enters species H. Then if b and c coalesce in population T, the gene tree will be G_1, and otherwise, both coalescent events for the three sequences will occur in R with equal probabilities for the three gene trees. This means that G_3 is the least frequent gene tree, with $\mathbb{P}(G_3) < \min\{\mathbb{P}(G_1), \mathbb{P}(G_2)\}$. We have

$$\begin{aligned}
\mathbb{P}(G_1) &= \frac{1}{3}\varphi P_S^\star + (1-\varphi)\left(1 - P_T + \frac{1}{3}P_T\right),\\
\mathbb{P}(G_2) &= \varphi\left(1 - P_S^\star + \frac{1}{3}P_S^\star\right) + \frac{1}{3}(1-\varphi)P_T,\\
\mathbb{P}(G_3) &= \frac{1}{3}\left[\varphi P_S^\star + (1-\varphi)P_T\right] = 1 - \mathbb{P}(G_1) - \mathbb{P}(G_2).
\end{aligned} \quad (15)$$

Thus $\mathbb{P}(G_1) < \mathbb{P}(G_2)$, i.e., the majority-vote method infers an incorrect species tree, if and only if

$$\varphi > \frac{1-P_T}{2-P_S^\star - P_T} = \frac{1}{1 + \frac{1-P_S^\star}{1-P_T}}. \quad (16)$$

Gene Flow under the IM Model

UPGMA method with inflow—. We first consider an IM model with inflow. Define the migration rate (in forward time) from species A to C under the IM model to be $M_{AC} = m_{AC}N_C$ migrants per generation (Fig. 1b). When we trace the genealogy of the sample backwards in time, the process of coalescence and migration during time interval $(0, \tau_T)$ can be described by a Markov chain with three states: S_{abc}, S_{aab}, and S_{ab} (Hobolth et al. 2011; Zhu and Yang 2012). Here, S_{abc} is the initial state with the three sequences in the three populations, S_{aab} is the state after sequence c enters species A (tracing the genealogy backwards in time), with two sequences in A and a third in B, and S_{ab} is the state after sequence c enters A and coalesces with sequence a, so that two sequences remain in the sample.

As the model assumes unidirectional migration, the only migration possible is from C to A (when time runs backwards), at the rate m_{AC} per generation. The only coalescent possible during the time epoch $(0, \tau_T)$ is between a and c in population A, at the rate $1/(2N_A)$ per generation (Fig. 1b). Divide both rates by the mutation rate μ per generation so that one time unit is the expected time taken to accumulate one mutation per site. Then the migration rate becomes $w_{AC} = m_{AC}/\mu = 4M_{AC}/\theta_C$, and the coalescent rate in species A becomes $2/\theta_A$. Thus the backward process of migration and coalescence can be described by a Markov chain with the generator matrix Q

	S_{abc}	S_{aab}	S_{ab}
S_{abc}	$-w_{AC}$	w_{AC}	0
S_{aab}	0	$-\frac{2}{\theta_A}$	$\frac{2}{\theta_A}$
S_{ab}	0	0	0

The eigenvalues of Q are $\lambda_1 = 0$, $\lambda_2 = -\frac{2}{\theta_A}$, and $\lambda_3 = -w_{AC}$. The transition probability matrix, $P(t) = \exp(Qt)$, is

	S_{abc}	S_{aab}	S_{ab}
S_{abc}	$e^{-w_{AC}t}$	$\frac{\theta_A w_{AC}}{2-\theta_A w_{AC}} \times \left(e^{-w_{AC}t} - e^{-\frac{2}{\theta_A}t}\right)$	$1 - \frac{2}{2-\theta_A w_{AC}}e^{-w_{AC}t} + \frac{\theta_A w_{AC}}{2-\theta_A w_{AC}}e^{-\frac{2}{\theta_A}t}$
S_{aab}	0	$e^{-\frac{2}{\theta_A}t}$	$1 - e^{-\frac{2}{\theta_A}t}$
S_{ab}	0	0	1

Sequences a and b can coalesce in population R only, with the coalescent time having an exponential distribution $f(t_{ab}) = \frac{2}{\theta_R}e^{-\frac{2}{\theta_R}(t_{ab}-\tau_R)}$, $t_{ab} > \tau_R$, with expectation $\mathbb{E}(t_{ab}) = \tau_R + \frac{\theta_R}{2}$. Sequences a and c can coalesce in both A and R, while sequences b and c can coalesce in both T and R. Thus $\mathbb{E}(t_{ab}) > \max\{\mathbb{E}(t_{ac}), \mathbb{E}(t_{bc})\}$.

The probability density of coalescent time t_{ac} is

$$f(t_{ac}) = \begin{cases} P_{abc,aab}(t_{ac})\frac{2}{\theta_A}, & \text{if } 0 < t_{ac} < \tau_T,\\ P_{abc,aab}(\tau_T)\frac{2}{\theta_A}e^{-\frac{2}{\theta_A}(t_{ac}-\tau_T)}, & \text{if } \tau_T < t_{ac} < \tau_R,\\ \left(P_{abc,aab}(\tau_T)P_A + P_{abc,abc}(\tau_T)\right) & \\ \times \frac{2}{\theta_R}e^{-\frac{2}{\theta_R}(t_{ac}-\tau_R)}, & \text{if } t_{ac} > \tau_R, \end{cases} \quad (17)$$

where $P_A = e^{-\frac{2}{\theta_A}(\tau_R - \tau_T)}$ is the probability that two sequences entering population A at time τ_T do not coalesce before they reach time τ_R. The expectation of t_{ac} is

$$\begin{aligned}
\mathbb{E}(t_{ac}) = &\left(\frac{1}{w_{AC}} + \frac{\theta_A}{2}\right) + \frac{\theta_A w_{AC}}{2-\theta_A w_{AC}}\left(\tau_T + \frac{\theta_A}{2}\right)e^{-\frac{2}{\theta_A}\tau_T}\\
&- \frac{2}{2-\theta_A w_{AC}}\left(\tau_T + \frac{1}{w_{AC}}\right)e^{-w_{AC}\tau_T}\\
&+ P_{abc,aab}(\tau_T)\left[\tau_T + \frac{\theta_A}{2} + P_A\left(\frac{\theta_R}{2} - \frac{\theta_A}{2}\right)\right]\\
&+ P_{abc,abc}(\tau_T)\left(\tau_R + \frac{\theta_R}{2}\right).
\end{aligned} \quad (18)$$

The coalescent time t_{bc} has the density

$$f(t_{bc})=\begin{cases} P_{abc,abc}(\tau_T)\frac{2}{\theta_T}e^{-\frac{2}{\theta_T}(t_{bc}-\tau_T)}, & \text{if } \tau_T<t_{bc}<\tau_R,\\ [P_{abc,abc}(\tau_T)P_T & \\ +(1-P_{abc,abc}(\tau_T))] & \\ \frac{2}{\theta_R}e^{-\frac{2}{\theta_R}(t_{bc}-\tau_R)}, & \text{if } t_{bc}>\tau_R, \end{cases} \tag{19}$$

with expectation

$$\mathbb{E}(t_{bc}) = P_{abc,abc}(\tau_T)\left[\tau_T+\frac{\theta_T}{2}+P_T\left(\frac{\theta_R}{2}-\frac{\theta_T}{2}\right)\right] +(1-P_{abc,abc}(\tau_T))\left(\tau_R+\frac{\theta_R}{2}\right). \tag{20}$$

Determining the threshold value of w_{AC} or M_{AC} for which $\mathbb{E}(t_{bc})>\mathbb{E}(t_{ac})$, so that the UPGMA method is inconsistent, is not analytically tractable but the threshold can be calculated numerically through a linear search for given values of parameters (θs and τs).

Majority-vote method with inflow—. Using the Markov chain characterization of the backward process of coalescent and migration described above, the probabilities of three gene trees can be easily derived (Fig. 1b). We have

$$\begin{aligned} \mathbb{P}(G_1)&=P_{abc,abc}(\tau_T)(1-P_T+\frac{1}{3}P_T)+P_{abc,aab}(\tau_T)P_A\cdot\frac{1}{3},\\ \mathbb{P}(G_2)&=P_{abc,ab}(\tau_T)+P_{abc,aab}(\tau_T)(1-P_A\\ &\quad+\frac{1}{3}P_A)+\frac{1}{3}P_{abc,abc}(\tau_T)P_T,\\ \mathbb{P}(G_3)&=\frac{1}{3}P_{abc,abc}(\tau_T)P_T+\frac{1}{3}P_{abc,aab}(\tau_T)P_A, \end{aligned} \tag{21}$$

where $P_A=e^{-\frac{2}{\theta_A}(\tau_R-\tau_T)}$. For example, gene tree G_1 occurs if sequences b and c coalesce first. This can only occur if there is no coalescent (between sequences a and c) in the time epoch $(0,\tau_T)$, so that the state of the Markov chain at time τ_T must be either S_{abc} or S_{aab}. In the former case (state S_{abc} at time τ_T), sequences b and c may coalesce first, in either species T (with probability $1-P_T$) or species R (with probability $P_T\cdot\frac{1}{3}$). In the latter case (state S_{aab} at time τ_T), sequences b and c may coalesce first only if there is no coalescence (between sequences a and c) in species A between τ_T and τ_R (with probability $P_A\cdot\frac{1}{3}$).

It is easy to see that $\mathbb{P}(G_3)<\min\{\mathbb{P}(G_1),\mathbb{P}(G_2)\}$. We have $\mathbb{P}(G_1)<\mathbb{P}(G_2)$ if and only if $P_{abc,abc}(\tau_T)(2-P_T)+P_{abc,aab}(\tau_T)P_A<1$, or if and only if

$$e^{-w_{AC}\tau_T}(2-P_T)+\frac{\theta_A w_{AC}\left(e^{-w_{AC}\tau_T}-e^{-\frac{2}{\theta_A}\tau_T}\right)}{2-\theta_A w_{AC}}P_A<1. \tag{22}$$

Again the threshold value of w_{AC} for this condition to be satisfied, so that the majority-vote method is inconsistent, can be calculated numerically through a linear search.

UPGMA method with outflow—. Similar to the IM model of Figure 1b, we use a Markov chain to characterize the coalescent-migration process in the time interval $(0,\tau_T)$ (Fig. 1d). The migration rate from C to A (in forward time) is M_{CA} migrants per generation or $w_{CA}=m_{CA}/\mu=4M_{CA}/\theta_A$ when time is scaled by the mutation rate. The coalescent rate (in population C after sequence a moves into C) is $1/(2N_C)$ per generation or $2/\theta_C$ on the mutational time scale. For a sample of three sequences (a,b,c), the three states of the Markov chain are S_{abc}, S_{ccb}, and S_{cb}, where S_{ccb} indicates three sequences exist with two sequences in species C and one in B, and S_{cb} indicates two sequences exist with one in species C and one in B. The rate matrix Q is

	S_{abc}	S_{ccb}	S_{cb}
S_{abc}	$-w_{CA}$	w_{CA}	0
S_{ccb}	0	$-\frac{2}{\theta_C}$	$\frac{2}{\theta_C}$
S_{cb}	0	0	0

This has the eigenvalues $\lambda_1=0$, $\lambda_2=-\frac{2}{\theta_C}$, and $\lambda_3=-w_{CA}$. The transition probability matrix $P(t)=\exp(Qt)$ is

	S_{abc}	S_{ccb}	S_{cb}
S_{abc}	$e^{-w_{CA}t}$	$\frac{\theta_C w_{CA}}{2-\theta_C w_{CA}}\times\left(e^{-w_{CA}t}-e^{-\frac{2}{\theta_C}t}\right)$	$1-\frac{2}{2-\theta_C w_{CA}}e^{-w_{CA}t}+\frac{\theta_C w_{CA}}{2-\theta_C w_{CA}}e^{-\frac{2}{\theta_C}t}$
S_{ccb}	0	$e^{-\frac{2}{\theta_C}t}$	$1-e^{-\frac{2}{\theta_C}t}$
S_{cb}	0	0	1

Sequences a and c can coalesce in populations C, T, and R. Sequences b and c or a and b can coalesce in T and R. The probability density of coalescent time t_{ac} is

$$f(t_{ac})=\begin{cases} P_{abc,ccb}(t_{ac})\frac{2}{\theta_C}, & \text{if } 0<t_{ac}<\tau_T,\\ P_{abc,ccb}(\tau_T)\frac{2}{\theta_T}e^{-\frac{2}{\theta_T}(t_{ac}-\tau_T)}, & \text{if } \tau_T<t_{ac}<\tau_R,\\ (P_{abc,ccb}(\tau_T)P_T+P_{abc,abc}(\tau_T)) & \\ \times\frac{2}{\theta_R}e^{-\frac{2}{\theta_R}(t_{ac}-\tau_R)}, & \text{if } t_{ac}>\tau_R, \end{cases} \tag{23}$$

with expectation

$$\begin{aligned} \mathbb{E}(t_{ac})&=\left(\frac{1}{w_{CA}}+\frac{\theta_C}{2}\right)+\frac{\theta_C w_{CA}}{2-\theta_C w_{CA}}\left(\tau_T+\frac{\theta_C}{2}\right)e^{-\frac{2}{\theta_C}\tau_T}\\ &\quad-\frac{2}{2-\theta_C w_{CA}}\left(\tau_T+\frac{1}{w_{CA}}\right)e^{-w_{CA}\tau_T}\\ &\quad+P_{abc,ccb}(\tau_T)\left[\tau_T+\frac{\theta_T}{2}+P_T\left(\frac{\theta_R}{2}-\frac{\theta_T}{2}\right)\right]\\ &\quad+P_{abc,abc}(\tau_T)\left(\tau_R+\frac{\theta_R}{2}\right). \end{aligned} \tag{24}$$

The coalescent time t_{bc} has the density

$$f(t_{bc})=\begin{cases}\frac{2}{\theta_T}e^{-\frac{2}{\theta_T}(t_{bc}-\tau_T)}, & \text{if } \tau_T<t_{bc}<\tau_R,\\ P_T\frac{2}{\theta_R}e^{-\frac{2}{\theta_R}(t_{bc}-\tau_R)}, & \text{if } t_{bc}>\tau_R,\end{cases} \quad (25)$$

with expectation

$$\mathbb{E}(t_{bc})=\tau_T+\frac{\theta_T}{2}+P_T\left(\frac{\theta_R}{2}-\frac{\theta_T}{2}\right). \quad (26)$$

The probability density of the coalescent time t_{ab} is

$$f(t_{ab})=\begin{cases}P_{abc,ccb}(\tau_T)\frac{2}{\theta_T}e^{-\frac{2}{\theta_T}(t_{ab}-\tau_T)}, & \text{if } \tau_T<t_{ab}<\tau_R,\\ \left[P_{abc,ccb}(\tau_T)P_T+(1-P_{abc,ccb}\right. & \\ \left.(\tau_T))\right]\frac{2}{\theta_R}e^{-\frac{2}{\theta_R}(t_{ab}-\tau_R)}, & \text{if } t_{ab}>\tau_R,\end{cases} \quad (27)$$

with expectation

$$\mathbb{E}(t_{ab}) = P_{abc,ccb}(\tau_T)\left[\tau_T+\frac{\theta_T}{2}+P_T\left(\frac{\theta_R}{2}-\frac{\theta_T}{2}\right)\right] +(1-P_{abc,ccb}(\tau_T))\left(\tau_R+\frac{\theta_R}{2}\right). \quad (28)$$

It is easy to see that $\mathbb{E}(t_{ab})>\max\{\mathbb{E}(t_{ac}),\mathbb{E}(t_{bc})\}$. Again the threshold value of w_{CA} or M_{CA} for which $\mathbb{E}(t_{bc})>\mathbb{E}(t_{ac})$, so that the UPGMA method is inconsistent, can be calculated numerically through a linear search.

Majority-vote method with outflow—. The procedure we used previously to analyze the IM model with inflow is also applied here (Fig. 1d). We have

$$\mathbb{P}(G_1)=P_{abc,abc}(\tau_T)(1-P_T+\frac{1}{3}P_T)+\frac{1}{3}P_{abc,ccb}(\tau_T),$$

$$\mathbb{P}(G_2)=P_{abc,cb}(\tau_T)+\frac{1}{3}P_{abc,abc}(\tau_T)P_T+\frac{1}{3}P_{abc,ccb}(\tau_T),$$

$$\mathbb{P}(G_3)=\frac{1}{3}P_{abc,abc}(\tau_T)P_T+\frac{1}{3}P_{abc,ccb}(\tau_T),$$

It is easy to see that G_3 is the least probable gene tree, with $\mathbb{P}(G_3)<\min\{\mathbb{P}(G_1),\mathbb{P}(G_2)\}$. Furthermore, $\mathbb{P}(G_1)<\mathbb{P}(G_2)$ if and only if $P_{abc,abc}(\tau_T)(1-P_T)<P_{abc,cb}(\tau_T)$, or if and only if

$$e^{-w_{CA}\tau_T}(1-P_T)+\frac{2e^{-w_{CA}\tau_T}-(\theta_C w_{CA})e^{-\frac{2}{\theta_C}\tau_T}}{2-\theta_C w_{CA}}<1. \quad (29)$$

Again a linear-search algorithm can be used to determine the value of w_{CA} for which this condition is satisfied and the majority-vote method is inconsistent.

Results

The majority-vote and UPGMA methods

Here, we apply the theoretical results developed above to determine the amount of gene flow, as measured by the migration rate M in the IM model and the introgression probability φ in the MSci model, that is sufficient to mislead species tree estimation. These thresholds define the boundary of the zone of inconsistency for each gene flow model and each inference method. In Figure 2, the threshold values of φ under the MSci model and of M under the IM model are plotted against τ_T/τ_R. All populations are assumed to have the same size (θ), and $\tau_R=5\theta$ or 10θ is fixed. In the case of the MSci model, there is an introgression event at time $\tau_H=\tau_S=\tau_R/5$ (see Fig. 1a,c). Note that a larger τ_T/τ_R (or τ_T since τ_R is fixed) means that the internal branch in the species tree is shorter and the species tree is more challenging to recover, so that all methods are expected to be more sensitive to gene flow.

We focus on hard species trees with a short internal branch, that is, with τ_T/τ_R larger than $\frac{1}{2}$ or even close to 1. Note that the plotted threshold value is the point at which the two species trees are equally good. For example, in the case of the inflow introgression model and the majority-vote method, with $\tau_R=5\theta$ and $\tau_T/\tau_R=0.95$, the threshold is $\varphi_{\text{lim}}=0.282$ (Fig. 2a). Thus, if and only if the introgression probability is higher than 28.2% will the mismatching gene tree G_2 be more frequent than the matching gene tree G_1, and the majority-vote method infer the incorrect species tree (and be inconsistent). The threshold of $\varphi_{\text{lim}}=0.0588$ for the UPGMA method (Fig. 2a) is much lower, suggesting that the UPGMA method is much more sensitive to gene flow than the majority-vote method. The results for outflow introgression are similar, with the majority-vote method being more robust to gene flow than the UPGMA method. Indeed when $\tau_T>0.9\tau_R$, the φ thresholds are nearly identical for inflow versus outflow introgressions.

For the inflow migration model (Fig. 2b) and with $\tau_R=5\theta$ and $\tau_T/\tau_R=0.95$, the lower limit for migration rate is $M_{AC}=0.0183$ for the majority-vote method and $M_{AC}=0.00473$ for UPGMA. The limiting M_{CA} values for the outflow model are similar. Again the majority-vote method is much more robust to gene flow than UPGMA. Note that in population genetic models of subdivision, gene flow of rates $M\approx0.1$ immigrants per generation is considered low enough so that strong population differentiation will not occur, yet such low levels can still lead to inconsistency when the species tree is difficult to reconstruct due to short internal branches.

UPGMA using Reconstructed Gene Trees

The gene tree probabilities we derived above are for the true gene trees. In analyses of real data, gene trees estimated from sequence alignments may differ from the true gene trees due to inference errors. Here, we study properties of the majority-vote method when it is applied to estimated gene trees. We expect random sampling errors to be unimportant for the UPGMA method based on average sequence distances between species because the number of sites in multilocus data sets is huge.

The impact of phylogenetic errors under the MSC model without gene flow and in the case of three species and Jukes–Cantor (JC) model (Jukes and Cantor,

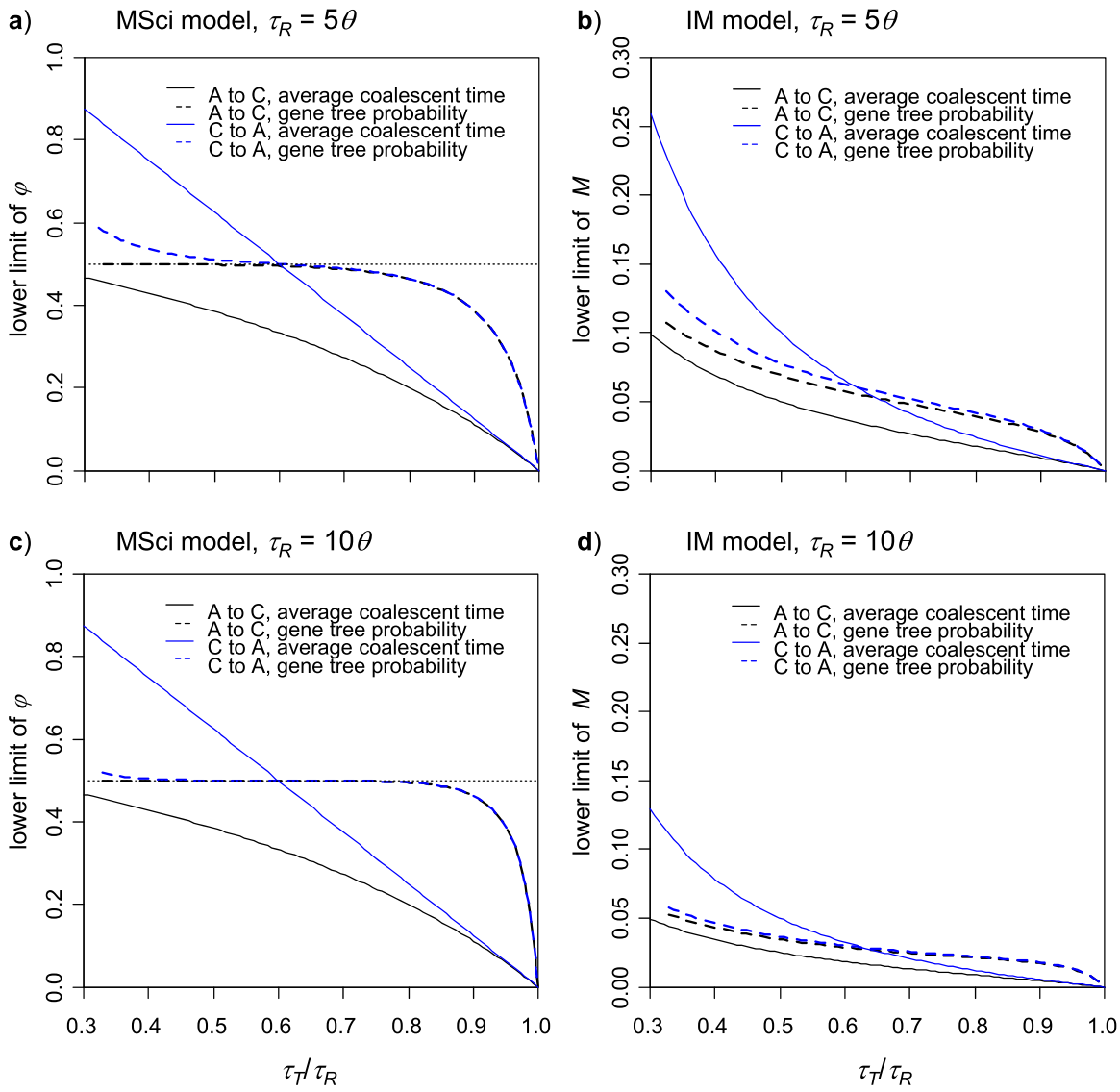

FIGURE 2. The lower limit of the introgression probability φ in the MSci model and of the migration rate M in the IM model that is necessary and sufficient to mislead species tree estimation methods UPGMA (based on average coalescent times) and majority-vote (based on gene tree probabilities), plotted against τ_T/τ_R. All populations are assumed to have the same size parameter (θ), while the age of the root is $\tau_R = 5\theta$ in (a) and (b) and $\tau_R = 10\theta$ in (c) and (d). In the MSci model, we use $\tau_H = \tau_S = \tau_R/5$. Note that the φ and M limits do not depend on the precise value of θ.

1969) was studied by Yang (2002). Without phylogenetic errors, the probabilities of the gene trees satisfy $\mathbb{P}(G_1) > \mathbb{P}(G_2) = \mathbb{P}(G_3)$ (Hudson 1983). Let $\mathbb{P}_n(G_k), k = 1,2,3$, be the probability that the estimated gene tree (the ML gene tree, for example) is G_k at a locus with sequences of n sites, with $\mathbb{P}_\infty(G_k) = \mathbb{P}(G_k)$. While phylogenetic reconstruction errors may cause the true matching gene tree (G_1) to be reconstructed as a mismatching gene tree (G_2 or G_3) and vice versa, reconstruction errors on balance always inflate the gene tree-species tree mismatch probability, but do not change the order of the gene trees. In other words, $\mathbb{P}_n(G_1) < \mathbb{P}(G_1)$ and $\mathbb{P}_n(G_2) > \mathbb{P}(G_2)$, but the relationship $\mathbb{P}_n(G_1) > \mathbb{P}_n(G_2) = \mathbb{P}_n(G_3)$ still holds (Yang 2002). Thus the majority-vote method, when applied to estimated gene trees, is consistent, and the probability of inferring the correct species tree will approach one with the increase in the number of loci or gene trees, even if the gene trees are estimated with sampling errors. The internal branch length in the species tree (in coalescent units) is nevertheless inconsistently estimated and underestimated, because the gene tree probabilities are distorted by phylogenetic errors.

The effects of phylogenetic errors on the gene tree probabilities when there is gene flow (under either the IM or MSci models) are unknown. Here, we use simulation to explore the issue. The ML gene tree for the case of three sequences and JC model with the molecular clock is analytically tractable (Yang 1994, 2000). The sequence alignment at each locus can be summarized

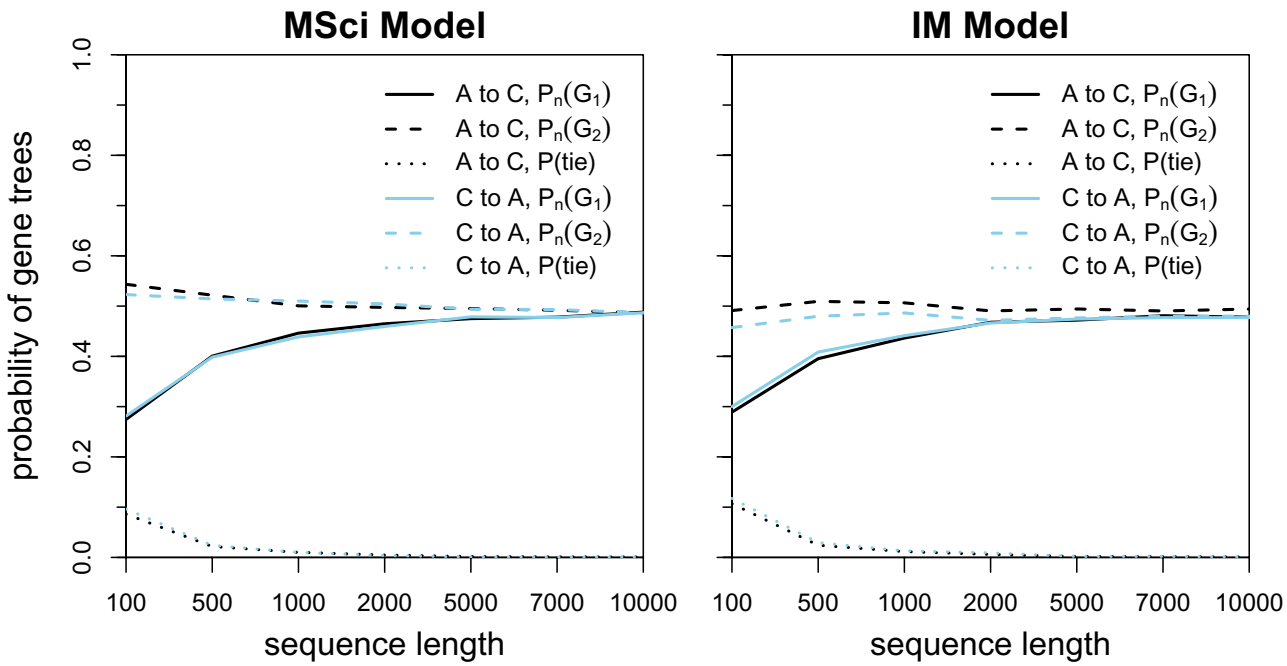

FIGURE 3. Probabilities of estimated gene trees, $\mathbb{P}_n(G_1)$ and $\mathbb{P}_n(G_2)$, as a function of sequence length (n), under the MSci and IM models. The following parameter values are used: $\theta = 0.01$ for all populations, $\tau_T = 0.04$, and $\tau_R = 0.05$; and for the MSci model, $\tau_H = 0.01$ and $\varphi = 0.4638$ for inflow introgression ($A \to C$) and 0.4643 for outflow introgression ($C \to A$), while for the IM model, $M = 0.0393$ for inflow migration and 0.0419 for outflow immigration. At those parameter values, $\mathbb{P}(G_1) = \mathbb{P}(G_2)$ when the sequence length is infinity (so that there are no phylogenetic reconstruction errors). $\mathbb{P}$(tie) is the proportion of data sets in which two or three gene trees are equally best.

as five site pattern counts (n_0–n_4), for xxx, xxy, xyx, yxx, and xyz, where x, y, z are any three distinct nucleotides, and the gene trees G_1, G_2, or G_3 is the ML tree if n_1, n_2, or n_3 is the greatest among the three. There is then no need for ML iteration to estimate the gene tree and branch lengths at each locus. We used this approach to calculate the probabilities of estimated gene trees $\mathbb{P}_n(G_1), \mathbb{P}_n(G_2)$, and $\mathbb{P}_n(G_3)$ by simulating 10^5 data sets or loci using BPP4 under the MSci or IM models (Yang 2015; Flouris et al. 2018). In Figure 3, we chose a set of parameter values for the MSci and IM models from Figure 2a, b for which $\mathbb{P}(G_1) = \mathbb{P}(G_2)$, and simulated data at different sequence lengths. In this case, phylogenetic reconstruction errors are seen to inflate the gene tree-species tree mismatch probability, with $\mathbb{P}_n(G_1) < \mathbb{P}(G_1)$.

We then used a linear search to find the minimum φ in the MSci model and minimum M in the IM model at which $\mathbb{P}_n(G_1) = \mathbb{P}_n(G_2)$, with the gene tree probabilities determined by simulating 10^5 loci and for each locus by determining the ML tree using the observed site pattern counts. The results for the MSci model are shown in Figure 4a, c. We focus on hard species trees and a small amount of introgression, with $\varphi \ll \frac{1}{2}$. In such cases, phylogenetic reconstruction errors inflate gene tree-species tree conflicts, with $\mathbb{P}_n(G_1) < \mathbb{P}(G_1)$. As a result, the low limit of φ necessary to mislead the majority-vote method of species tree estimation is lower than when true gene trees are used (Fig. 2). In other words, species tree estimation is more sensitive to introgression when gene trees are reconstructed from sequence data than when true gene trees are used. The patterns are the same for the inflow ($A \to C$) and outflow ($C \to A$) introgressions.

The results under the IM model are similar (Fig. 4b, d). For hard species trees with short internal branches, phylogenetic reconstruction errors inflate the gene tree-species tree conflicts, making the estimation of the species tree even harder.

Full-Likelihood Methods

Full-likelihood methods applied to multilocus sequence alignments, including ML and Bayesian methods, integrate over the gene tree topologies and coalescent times and naturally accommodate phylogenetic reconstruction errors due to limited number of sites at each locus (see, for a review, Xu and Yang 2016). They are not tractable analytically. Nevertheless, in the case of three species and three sequences per locus, an efficient ML implementation of the MSC model exists in the 3s program (Yang, 2002; Dalquen et al., 2017). Here, we use 3s to analyze data sets of 10,000 loci (with sequence length $n = 1000$), assuming that at such large data size, the estimates are close to the infinite-data limits. In Figure 5, we conducted similar calculations to Figure 4, but with ML (the 3s program) replacing majority-vote. Consider Figure 5a, where the true model is the MSci model. For each value of τ_T, a linear search (bisection) is used to find the lowest value of φ in the true MSci model at which the two species trees

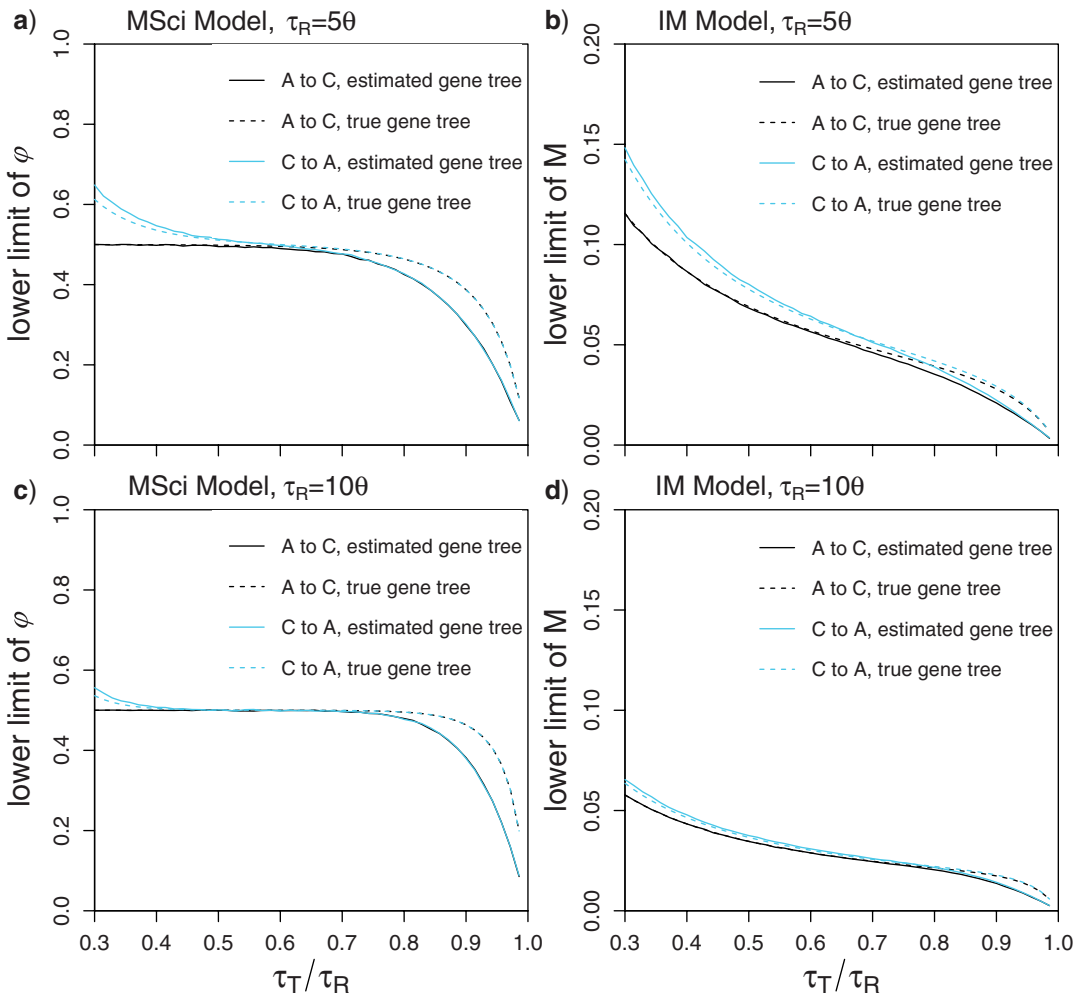

FIGURE 4. The lower limit of the introgression probability φ in the MSci model and the migration rate M in the IM model necessary to mislead the majority-vote method of species tree estimation, plotted against τ_T/τ_R, when either the true or estimated gene trees are used. For estimated gene trees the sequence length is $n=1000$. At the φ or M values shown, $\mathbb{P}(G_1)=\mathbb{P}(G_2)$ when the true gene trees are used or $\mathbb{P}_n(G_1)=\mathbb{P}_n(G_2)$ when estimated gene trees are used. The estimated gene tree is the ML tree from the sequence alignment of $n=1000$ sites, determined by using the site pattern counts at the locus. Monte Carlo simulation, with 10^5 replicates (loci or sequence alignments), was used to estimate the probabilities of the ML gene trees: $\mathbb{P}_n(G_1)$ and $\mathbb{P}_n(G_2)$. Parameters other than φ or M are fixed at $\theta=0.01$ for all populations, $\tau_H=\tau_S=\tau_R/5$ for the MSci model (a and c), while $\tau_R=5\theta$ in (a) and (b) and $\tau_R=10\theta$ in (c) and (d). Results for the true gene trees are from Figure 2, shown here for comparison.

have the same log likelihood under the MSC model, $\ell_1=\ell_2$, where the log likelihood is calculated under the JC model and the molecular clock by averaging over the three gene trees and integrating over the two coalescent times in each gene tree (Yang 2002). For each φ (and τ_T as well as other parameters in the MSci model), a data set of 10,000 loci is simulated and analyzed using 3s to determine whether $\ell_1>\ell_2$ (e.g., whether φ is too small). Each round of bisection reduces the interval of uncertainty by a half. The scatter-points in the plots (Fig. 5) show some fluctuations, due to the finite nature of the data sets, and are used to fit a smoothed curve.

The lower limits of φ in the MSci model and of M in the IM model for the ML/3s method of Figure 5 are much lower than the corresponding values for the majority-vote and the UPGMA methods of Figures 2 and 4. The ML/3s method assuming MSC without gene flow infers the incorrect species tree at much lower levels of gene flow (and is less robust to gene flow) than the majority-vote or UPGMA methods.

Differences between the IM and MSci Models

The MSci and IM models are two idealized models that accommodate gene flow between species. The MSci model assumes episodic introgression or hybridization events that occur at fixed time points in the past, while the IM model assumes continuous-time migration with migrants occurring at a certain rate in every generation. The MSci and IM models represent two extremes and in reality a combination of the two processes may be more realistic. When the two models are applied to the same data, a frequently asked question is how the important parameters of the two models (φ in MSci and M in IM

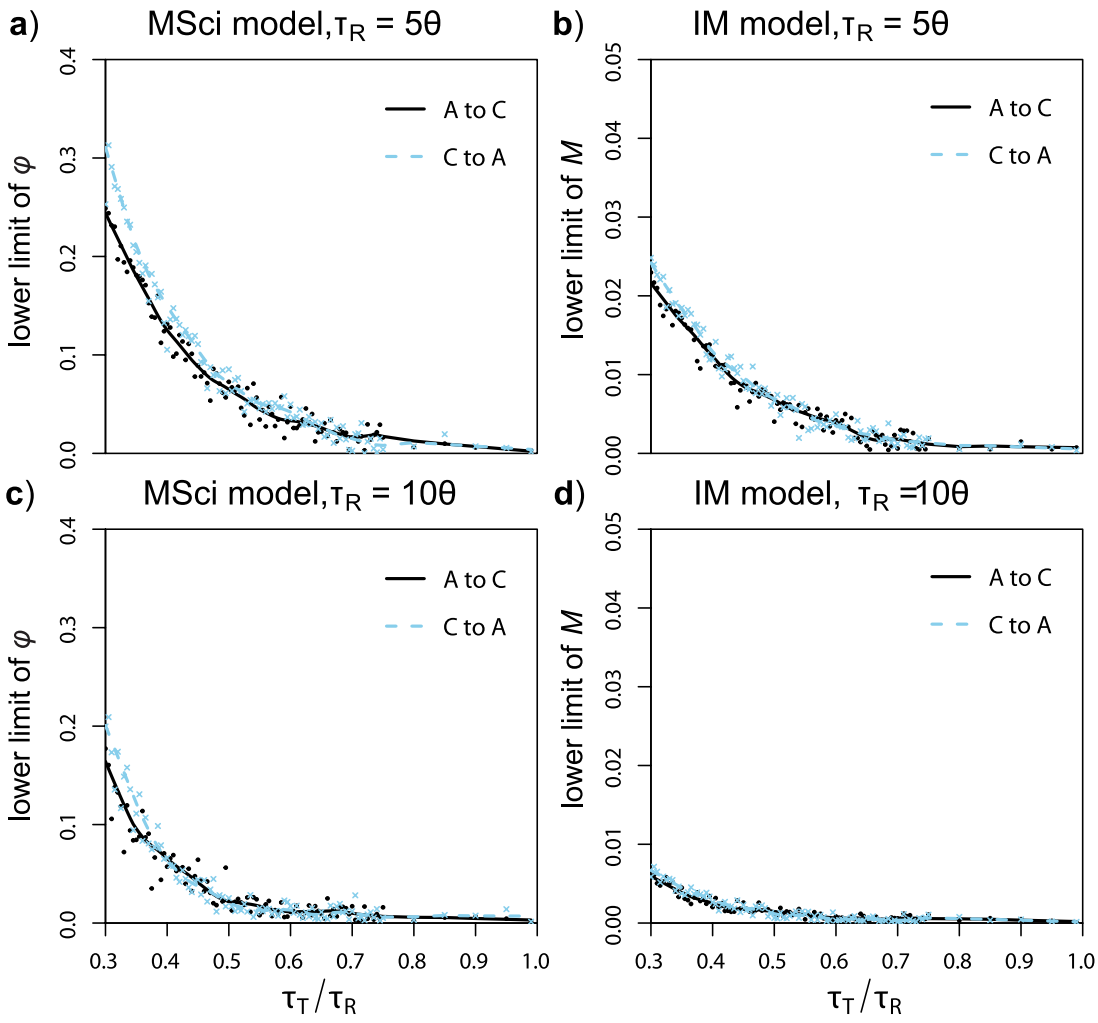

FIGURE 5. The lower limit of φ in the MSci model and M in the IM model necessary to mislead the ML/3s method of species tree estimation under the MSC, plotted against τ_T/τ_R. The sequence length is 1000. Data of 10,000 loci were analyzed using 3s to determine the ML species tree. Parameter values used are the same as in Figure 4.

models, say) correspond to each other. One may expect a higher migration rate M should correspond to a larger introgression probability φ, but the precise relationship may depend on the values of other parameters in the models, as well as the data configurations (the number of loci, the number of sequences per locus, and the sequence length). Bearing in mind those caveats we conducted three analyses to address this question.

Firstly, we may compare the limiting values of φ under the MSci model and M in the IM model for the same τ_T/τ_R ratio in Figure 2. For example, for the inflow migration or introgression ($A \to C$), with $\tau_R = 5\theta$ and $\tau_T/\tau_R = 0.95$, the limiting values to achieve equal gene tree probabilities, $\mathbb{P}(G_1) = \mathbb{P}(G_2)$, are $\varphi_{\text{lim}} = 0.282$ for the MSci model and $M_{AC} = 0.0183$ for the IM model. The limiting values for achieving equal average coalescent times, $\mathbb{E}(t_{bc}) = \mathbb{E}(t_{ac})$, are $\varphi_{\text{lim}} = 0.0588$ for the MSci model and $M_{AC} = 0.00473$ for the IM model. Here, we are matching certain summaries of the data to establish a correspondence between φ and M in the two models.

Secondly, we used the IM model to simulate large data sets of 10,000 loci, with two sequences per species per locus and with sequence length of 100 or 1000 sites, and then used BPP to analyze the data under the MSci model (Flouris et al. 2019). The rationale is that the data sets are so large that random sampling errors in the estimates are negligible, as are the impact of the prior or the differences between MLEs and Bayesian estimates. In other words, the posterior means of parameters should be close to the *pseudotrue parameter values*, which are the limits of the MLEs when the data size approaches infinity. This approach also allows us to examine the estimates of other parameters in the model besides the rate of gene flow. The results are listed in Table 1.

In the case of inflow migration ($A \to C$), population sizes (θ) for most populations (in particular, the extant

TABLE 1. Posterior means and 95% highest probability density credible intervals (below) for parameters in the MSci model when large data sets of 10,000 loci simulated under the IM model are analyzed using BPP under the MSci model

θ_A	θ_B	θ_C	θ_R	θ_T	θ_S	θ_H	τ_R	τ_T	τ_H	φ
Inflow migration ($A \to C$) with $M_{AC} = 0.0393$, $n = 100$ sites										
0.0096	0.0095	0.0111	0.0121	0.0185	0.0149	0.0106	0.0478	0.0386	0.0075	0.3319
0.0093	0.0093	0.0107	0.0109	0.0131	0.0140	0.0092	0.0471	0.0372	0.0072	0.3208
0.0099	0.0098	0.0115	0.0134	0.0241	0.0157	0.0119	0.0485	0.0401	0.0077	0.3435
Inflow migration ($A \to C$) with $M_{AC} = 0.0393$, $n = 1000$ sites										
0.0091	0.0099	0.0100	0.0106	0.0115	0.0250	0.0100	0.0496	0.0398	0.0050	0.4393
0.0089	0.0097	0.0098	0.0102	0.0107	0.0243	0.0095	0.0495	0.0395	0.0049	0.4300
0.0094	0.0101	0.0103	0.0109	0.0122	0.0256	0.0106	0.0498	0.0400	0.0051	0.4485
Outflow migration ($C \to A$) with $M_{CA} = 0.0419$, $n = 100$ sites										
0.0108	0.0098	0.0096	0.0154	0.0032	0.0122	0.0094	0.0437	0.0417	0.0068	0.3012
0.0104	0.0095	0.0092	0.0145	0.0020	0.0115	0.0083	0.0431	0.0410	0.0066	0.2907
0.0112	0.0101	0.0099	0.0163	0.0049	0.0129	0.0106	0.0442	0.0424	0.0071	0.3122
Outflow migration ($C \to A$) with $M_{CA} = 0.0419$, $n = 1000$ sites										
0.0102	0.0099	0.0094	0.0111	0.0094	0.0243	0.0098	0.0485	0.0401	0.0051	0.4437
0.0099	0.0097	0.0091	0.0107	0.0089	0.0237	0.0093	0.0483	0.0399	0.0050	0.4342
0.0105	0.0101	0.0096	0.0115	0.0099	0.0250	0.0104	0.0488	0.0403	0.0052	0.4533

Note: The true model is IM, with parameter values $\theta_A = \theta_B = \theta_C = \theta_T = \theta_R = 0.01$, $\tau_R = 0.05$ and $\tau_T = 0.04$, and with $M_{AC} = 0.0393$ and $M_{CA} = 0.0419$. These are the lower limits of M in the IM model, at which $\mathbb{P}(G_1) = \mathbb{P}(G_2)$ (Fig. 2). Large data sets of 10,000 loci, with two sequences from each species per locus and with the sequence length $n = 100$ or 1000 sites, were simulated under the IM model, and analyzed using BPP under the MSci model. Inverse-gamma priors are assigned on θ and τ_0: $\theta \sim \text{IG}(3, 0.02)$ with mean $0.02/(3-1) = 0.01$ for all θs, and $\tau_0 \sim \text{IG}(3, 0.1)$ with mean 0.05 for the age of the species tree root, while the introgression probability φ is assigned the $\mathbb{U}(0, 1)$ prior.

species) are well estimated, especially at $n = 1000$. Species divergence times τ_R and τ_T estimated under MSci are very similar to the true values under the IM model (0.05 and 0.04, respectively). The age of the hybridization node τ_H ($= \tau_S$) is very small, especially at the larger sequence length. Under the IM model, migration occurs over the whole time interval $(0, \tau_T)$, and one might naively expect τ_H under MSci to be close to the mid value. Instead the MSci estimate of τ_H is near the lower limit. Indeed the estimate should be smaller for longer sequences and/or more loci, because under the MSci model, the hybridization time must be smaller than the between-species sequence divergence time, with $\tau_H < t_{ac}$. Thus, longer sequences or more loci should provide stronger evidence that the minimum sequence divergence t_{ac} generated under the IM model can be arbitrarily small, leading to reduced estimates of τ_H under the MSci model. The migration rate $M_{AC} = 0.0393$ in the IM model is relatively small, while the estimates of φ are substantial, at 33–44%. The results for the case of outflow migration ($C \to A$) are similar to those of inflow migration (Table 1). Again, species divergence times τ_R and τ_T are well estimated, as are population sizes, even though the model is incorrect. The estimated hybridization time $\tau_H (= \tau_S)$ is very small. Although the migration rate is only $M_{AC} = 0.0419$, the estimates of φ in the MSci model are substantial, at 30–44%. Overall, very low migration rates, on the order of $M = 1$–5% may correspond to substantial introgression probabilities close to 50%.

Thirdly, we analyzed the case of two species (A and B) with one sequence per locus for each species when the true sequence distance (or the coalescent time between the two sequences) is known (Fig. 6a, b). Using the theory for the IM model developed earlier, we have the probability density of coalescent time t between the two sequences to be

$$f_M(t) = \begin{cases} \frac{2w}{2-\theta_A w}\left(e^{-wt} - e^{-\frac{2}{\theta_A}t}\right), & \text{if } 0 < t < \tau_R, \\ \left(\frac{2}{2-\theta_A w}e^{-w\tau_R} - \frac{\theta_A w}{2-\theta_A w}e^{-\frac{2}{\theta_A}\tau_R}\right)\frac{2}{\theta_R}e^{-\frac{2}{\theta_R}(t-\tau_R)}, & \text{if } t > \tau_R, \end{cases} \tag{30}$$

where $w = m_{AB}/\mu = 4M_{AB}/\theta_B$ is the mutation-scaled migration rate from A to B. The density depends on w but not on M_{AB} and θ_B individually, so that the parameters specifying the density for the migration model are $\boldsymbol{\theta}^{(M)} = \{w, \theta_A, \theta_R, \tau_R\}$. Similarly the density under the introgression model is specified by parameters $\boldsymbol{\theta}^{(I)} = \{\varphi, \theta_S, \theta_R, \tau_H, \tau_R\}$ and is

$$f_I(t) = \begin{cases} \varphi\frac{2}{\theta_S}e^{-\frac{2}{\theta_S}(t-\tau_S)}, & \text{if } \tau_S < t < \tau_R, \\ \left[\varphi e^{-\frac{2}{\theta_S}(\tau_R-\tau_S)} + (1-\varphi)\right] & \\ \quad \frac{2}{\theta_R}e^{-\frac{2}{\theta_R}(t-\tau_R)}, & \text{if } t > \tau_R. \end{cases} \tag{31}$$

Note that in the MSci model (Fig. 6b), $\tau_H = \tau_S$ so we use the two interchangeably. Some parameters (such as τ_R and θ_R) are common between the two models but they may take different values: when the parameter definition may be unclear from the context, we use $\tau_R^{(M)}$, with the superscript "(M)" or "(I)" to indicate the model involved.

The Kullback–Leibler (KL) divergence

$$D(\boldsymbol{\theta}^{(M)} || \boldsymbol{\theta}^{(I)}) = \int_0^\infty f_M(t) \log \frac{f_M(t)}{f_I(t)} dt, \tag{32}$$

is a measure of distance from the fitting introgression model to the true migration model. By minimizing D, we obtain the pseudotrue parameter values under MSci, $\boldsymbol{\theta}^{(I)*}$, when the true model is the IM model with parameters $\boldsymbol{\theta}^{(M)}$. Here $\boldsymbol{\theta}^{(M)}$ are fixed while $\boldsymbol{\theta}^{(I)}$ are being optimized. Because the IM model allows arbitrarily

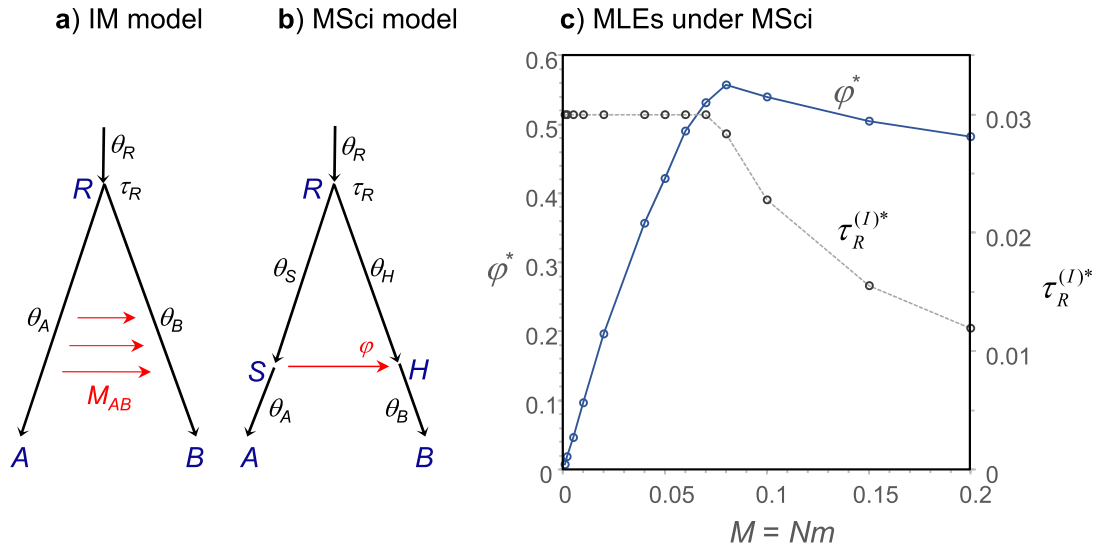

FIGURE 6. The a) migration and b) introgression models for two species (A, B), and c) the pseudotrue parameter values of φ and $\tau_R^{(I)}$ in the introgression model plotted against the migration rate $M_{AB}(=N_B m_{AB}=w\theta_B/4)$. The pseudotrue parameter values are the parameter values under the introgression model $\boldsymbol{\theta}^{(I)}$ that minimizes the KL divergence of equation 32: they are the limiting values of the MLEs of $\boldsymbol{\theta}^{(I)}$, when the number of loci $L\to\infty$, when the MSci model is fitted to data of L loci generated under the IM model, with two sequences (one from each species) of infinite length per locus. Other parameters in the migration model are fixed at $\tau_R^{(M)}=0.03$ and $\theta=0.01$ (for all populations). In the MSci model, $\theta=0.01$ is assumed for all populations, while φ and $\tau_R^{(I)}$ are estimated by minimizing the KL divergence (equation 32).

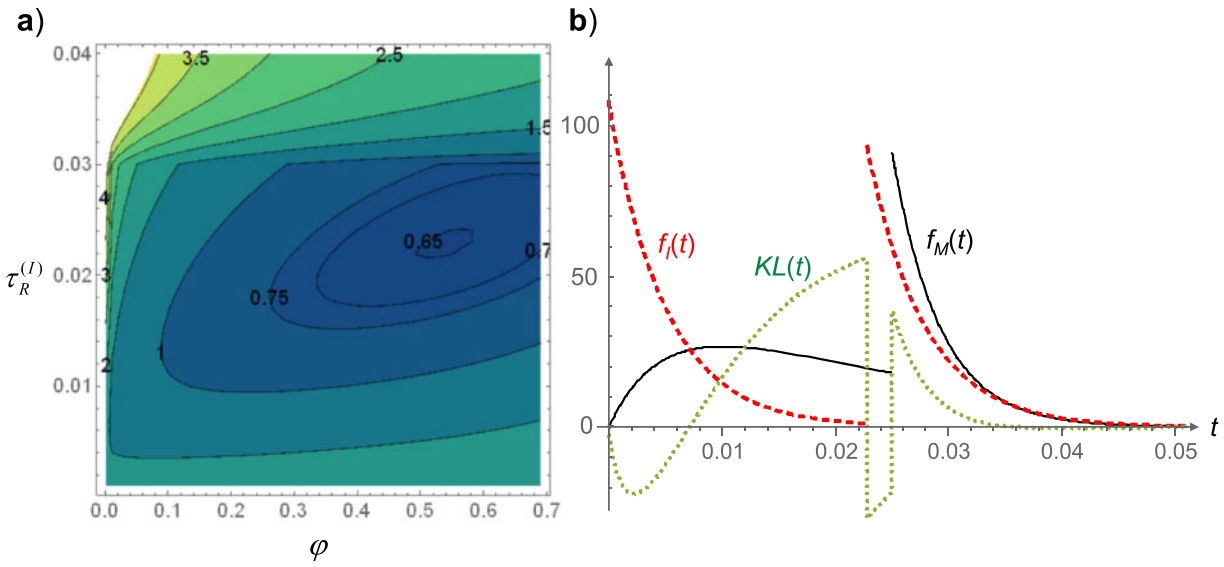

FIGURE 7. a) Contour plot of $D(\boldsymbol{\theta}_M||\boldsymbol{\theta}_I)$ (equation 32) as a function of φ and $\tau_R^{(I)}$ in the MSci model for the case of $M=0.1$ or $w=40$. Parameters in the IM model are $\tau_R^{(M)}=0.03$ and $\theta=0.01$ for all populations. The MLE under the MSci model, which minimizes D, is at $\varphi^*=0.5395$ and $\tau_R^{(I)*}=0.02280$. Note that in the MSci model, $\theta=0.01$ is fixed for all populations and only φ and $\tau_R^{(I)}$ are optimized as $\tau_H=0$. The D surface is not smooth at $\tau_R^{(I)}=\tau_R^{(M)}=0.03$. b) The densities $f_M(t)$ (equation 30) and $f_I(t)$ (equation 31) as well as the integrand, $f_M(t)\log\frac{f_M(t)}{f_I(t)}$, of equation 32, plotted at the MLEs of (a). In other words, the red dashed curve $f_I(t)$ is the best fit of the MSci model to the infinite-sized data under the IM model represented by the black solid curve $f_M(t)$. Note that $f_M(t)$ and $f_I(t)$ have one discontinuity point, while the integrand has two.

small coalescent time t while $\tau_H<t$ under the MSci model, we have $\tau_H^*=0$. As the integrand of equation 32 has one or two discontinuity points, depending on whether $\tau_R^{(M)}$ and $\tau_R^{(I)}$ are identical (Fig. 7), we applied Gaussian quadrature, with 64 points, to each continuous segment to calculate the integral of equation 32. The BFGS optimization routine in PAML (Yang 2007) is used to produce the MLEs. One such model fitting is illustrated in Figure 7.

Figure 6c shows the estimates of φ and $\tau_R^{(I)}$ under the assumption that θ is the same for all populations and also between the two models; thus $\theta=0.01$ is fixed and only

φ and $\tau_R^{(I)}$ are estimated by minimizing D in equation 32. When M in the IM model is small, φ increases nearly linearly with the increase of M (Fig. 6c). However, when M is large (>0.08, say, corresponding to $\varphi = 0.56$) and further increases, φ decreases.

As a summary of all three analyses above, we note that small values of the migration rate, in the order of 0.01–0.1 migrants per generation, may correspond to large introgression probabilities and have a large effect on the genetic history of species divergences represented by the gene tree probabilities. Furthermore, the IM and MSci models make very different predictions of the distribution of coalescent times, so that the two models may be easily distinguishable using genomic sequence data when both models are implemented in the same program.

Discussion

The Impact of Gene Flow on Species Tree Estimation

The impact of gene flow, either in the form of episodic introgressive hybridization or continuous migration, on species tree estimation clearly depends on how challenging the species tree is. Our analyses suggest that when the species tree is hard with very short internal branch lengths, even a small amount of gene flow can cause species tree estimation to become inconsistent. We found that the limiting φ and M values for the gene tree probabilities are much higher than the values for the sequence distances (Figs. 2 and 4), indicating that the majority-vote method based on gene tree topologies is more robust to gene flow than the UPGMA method based on average sequence distances. The full-likelihood method making use of information in both gene tree topologies and branch lengths is even more sensitive. This difference in sensitivity may be explained by the fact that a small amount of gene flow may easily affect the branch lengths or sequence distances but may not alter the distribution of gene tree topologies. For hard species trees, phylogenetic reconstruction errors tend to inflate the gene tree-species tree conflicts, adding further challenges to correct inference of the species tree.

When the effects of gene flow is a concern, it is important to use species tree methods that account for both the coalescent process and cross-species gene flow. Such methods are under active development. A number of methods have been developed to detect gene flow using sequence data (Green et al. 2010; Durand et al. 2011; Solis-Lemus and Ane 2016; Dalquen et al. 2017), and furthermore, a number of methods have been developed to infer the species tree with reticulation events, including summary methods based on estimated gene tree topologies (Yu et al. 2014; Yu and Nakhleh 2015; Solis-Lemus and Ane 2016; Wen et al. 2016; Allman et al. 2019) and full-likelihood methods applied to sequence alignments (Hey et al. 2018; Wen and Nakhleh 2018; Zhang et al. 2018). Furthermore, we have in this article examined the effects of gene flow on the inference of species tree topology only. The impact of gene flow on evolutionary parameters such as species divergence times and ancestral population sizes merits detailed study (Dalquen et al. 2017; Wen and Nakhleh 2018).

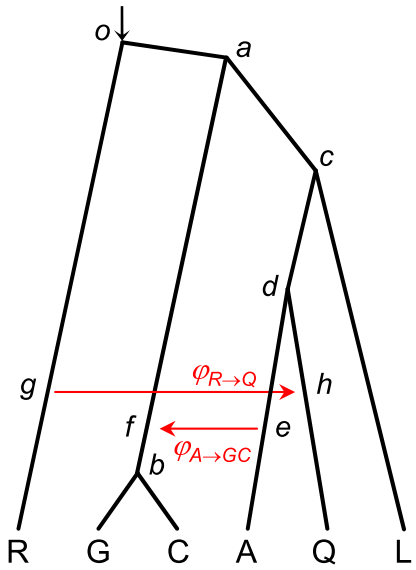

FIGURE 8. The species tree for the *Anopheles gambiae* species complex inferred by Thawornwattana et al. (2018) from the Xag region of the X chromosome, with two migration events for the autosomes. Redrawn following (Thawornwattana et al. 2018, Fig. 6).

Introgression in the Anopheles gambiae Species Complex

The *Anopheles gambiae* species complex is comprised of eight recognized species and includes major malaria vectors in Africa. Genome sequence data from six of the species, *A. gambiae* (G), *Anopheles coluzzii* (C), *Anopheles arabiensis* (A), *Anopheles melas* (L), *Anopheles merus* (R), and *Anopheles quadriannulatus* (Q), have been analyzed to estimate the species phylogeny and to infer the direction and intensity of gene flow across species (Fontaine et al. 2015; Thawornwattana et al. 2018). *Anopheles gambiae*, *A. coluzzii*, and *A. arabiensis* have large overlapping geographical distributions across sub-Saharan Africa and are major malaria vectors (Wiebe et al. 2017). *Anopheles melas* and *A. merus* are found in coastal waters of eastern and western Africa, respectively, and are minor vectors. *Anopheles quadriannulatus* does not bite humans and is not a malaria vector.

Fontaine et al. (2015) analyzed the genomic sequence data for the six species using sliding windows and suggested that gene flow is so widespread that the predominant gene trees for the autosomes are different from the species phylogeny, and that the X chromosome, which is apparently not affected by gene flow, reflects the true species history. This conclusion is supported in the reanalysis of Thawornwattana et al. (2018), although the inferred species trees were different. The species tree obtained by Thawornwattana et al. (2018) for the X-chromosomal data is shown in Figure 8. Besides being inferred using coalescent-based methods that

TABLE 2. Estimates and limiting values of φ for different chromosomal arms obtained from the genomic data of the *Anopheles gambiae* species complex

	$A \to G$			$R \to Q$		
Data set	$\hat{\varphi}_{A\to G}$	φ_{lim} (UPGMA)	φ_{lim} (M-V)	$\hat{\varphi}_{R\to Q}$	φ_{lim} (UPGMA)	φ_{lim} (M-V)
2L1+2 coding	0.94	0.89	0.74	0.28	0.77	0.76
2L1+2 noncoding	0.98	0.60	0.56	0.00	0.25	0.35
2La coding	0.73	0.14	0.16	0.01	0.30	0.42
2La noncoding	0.64	0.00	0.00	0.00	0.18	0.29
2R coding	0.97	0.90	0.81	0.34	0.80	0.79
2R noncoding	0.97	0.68	0.63	0.22	0.82	0.80
3L1+2 coding	0.94	0.76	0.66	0.32	0.70	0.69
3L1+2 noncoding	0.96	0.73	0.64	0.33	0.83	0.82
3La coding	0.93	0.87	0.78	0.65	0.56	0.56
3La noncoding	0.98	0.77	0.66	0.54	0.57	0.55
3R coding	0.95	0.98	0.95	0.43	0.65	0.64
3R noncoding	0.98	0.64	0.61	0.03	0.30	0.38

Note: The estimates ($\hat{\varphi}_{A\to G}$ and $\hat{\varphi}_{R\to Q}$) are posterior means in the BPP analysis of genomic data from all six species (Flouris et al. 2019, Table 1), while the lower limits for $\varphi_{A\to G}$ and $\varphi_{R\to Q}$ are calculated using triplet data (AQG and RQA) using equations for UPGMA and majority-vote (M-V), respectively.

accommodate deep coalescence, this tree is supported by additional sources of evidence, such as chromosome inversion data and well-known evidence of introgression from *A. arabiensis* → *A. gambiae* + *A. coluzzii* (Slotman et al. 2005). Moreover, computer simulation confirmed that the species trees inferred by Fontaine et al. (2015) were artifactual and reflected systematic errors of the sliding-windows analyses (Thawornwattana et al. 2018). Here, we apply the theory of this article to the species tree of Figure 8 and the parameter estimates under MSci from Flouris et al. (2019, Table 1) to confirm that the introgression rates affecting the autosomes are high enough to mislead species tree estimation.

For the A → G introgression, we used parameter estimates (τs and θs) for the triplet AQG to determine the low limit of φ in the MSci model. We retrieved the BPP estimates of parameters τs and θs for nodes a, d, e, and f in the species tree of Figure 8 from Flouris et al. (2019, Table 1) to calculate φ_{lim} for outflow (equations 14 and 16). The results are shown in Table 2. For both the coding and noncoding data and for almost all chromosomal arms, the estimates ($\hat{\varphi}$) are larger than the limiting values (φ_{lim}). Thus, both majority-vote (based on gene tree topologies) and UPGMA (based on average sequence divergences) are inconsistent and are expected to infer an incorrect species tree.

For the R → Q introgression, we used the RQA triplet and the inflow equations 5 and 7. All estimates of $\varphi_{R\to Q}$ are smaller than φ_{lim} for both the majority-vote and UPGMA methods except for the 3La coding region (Table 2). Thus, for most of the autosomal arms introgression from R to Q did not reach a sufficient intensity to mislead species tree estimation.

Migration, introgression, and the concept of species tree

The concept of the true species tree in the presence of cross-species gene flow may be poorly defined, especially if the models of gene flow are too simple to capture the major features of species divergence history. As mentioned in Introduction section, we have assumed in this article the backbone species tree to be the true species tree, whether the estimated introgression probability is below or above 50%. If the model of introgression/hybridization applies, it is reasonable to use the introgression probability at each hybridization node to define the "major hybridization branch" and the "major species tree," as in Solis-Lemus et al. (2017). The interpretations of our results will then have to be adjusted accordingly. For example, Figure 2a shows that at $\tau_T/\tau_R = 0.4$ under the outflow model, $\mathbb{P}(G_1) < \mathbb{P}(G_2)$ if and only if $\varphi > 0.536$. With the true species tree defined to be the major species tree, majority-vote will be consistent when $\varphi > 0.536$ (as well as when $\varphi < 0.5$); it is inconsistent only when $0.5 < \varphi < 0.536$. Note that the results in Figures 2, 3, 4, and 5 concerning the φ and M limits are all valid, but the change of the definition of the true species tree may change the consistency or inconsistency of the methods.

Our analysis in this article has been motivated by the inferred patterns of gene flow and species divergences in the *Anopheles gambiae* species complex, where the introgression or hybridization model appears to be a poor fit, and the major species tree for the autosomes does not appear to reflect the true history of species divergence and gene flow (Fontaine et al. 2015; Thawornwattana et al. 2018). Gene flow from *A. arabiensis* to *A. gambiae* (or *A. coluzzii*) affecting the autosomes appears to be ongoing (Slotman et al. 2005), and the continuous migration model may be a more realistic description than the episodic introgression model. Estimates of the migration rate $M_{A\to G}$ from the 3s program vary among the chromosomal arms with the estimate from the combined autosomal data to be 0.22 migrants per generation. In contrast, the estimated rate in the opposite direction is $M_{G\to A} = 0$ for all chromosomal arms (Thawornwattana et al. 2018, Table S3), consistent with the experiment of Slotman et al. (2005), which was unsuccessful in introducing *A. gambiae* into the *A. arabiensis* host. Here we use the estimate

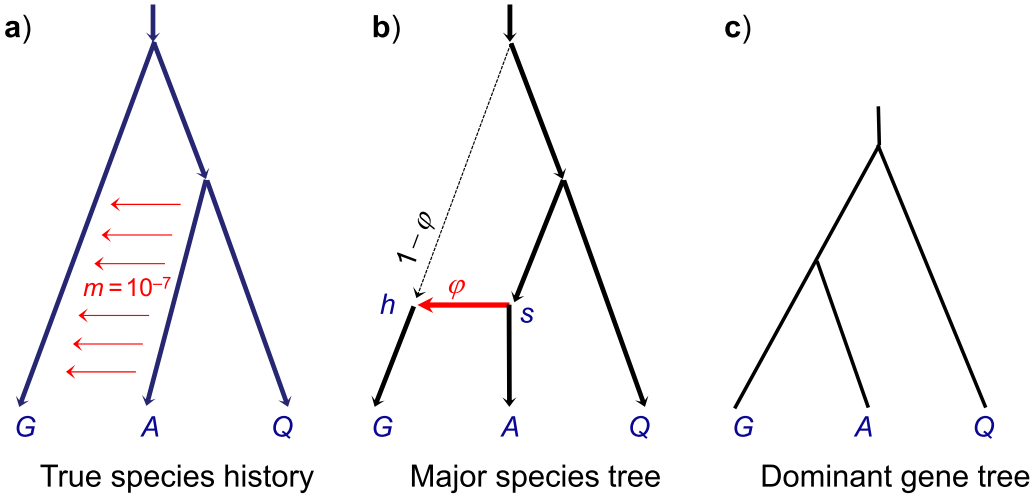

FIGURE 9. a) A plausible species history for *A. arabiensis* (A), *A. quadriannulatus* (Q), and *A. gambiae* (G), with continuous gene flow at the migration rate of $m_{A\to G} \approx 10^{-7}$. b) The inferred history when the introgression model is fitted to the genomic (autosomal) sequence data, with the introgression probability $\varphi_{A\to G} > 90\%$ (Table 1). The *major species tree* is represented by the thick branches. b) The dominant gene tree. Parameter values $m_{A\to G}$ and $\varphi_{A\to G}$ are based on rough calculations using the estimates from the genomic sequence data (see text for discussions).

$M_{A\to G} = 0.22$ as well as the scaled population size $\theta_G = 4N_G\mu = 0.027$ to calculate the migration rate or the proportion of migrants per generation, $m_{A\to G}$. Suppose the mutation rate is 2×10^{-9} per site per generation based on fruit flies (Keightley et al. 2014). Then we have $N_G = 0.027/(4\times2\times10^{-9}) = 3.4\times10^6$, and $m_{A\to G} = M_{A\to G}/N_G = 6.5\times10^{-8}$ per generation. We note that the migration rates m and M are averages over evolutionary time scales and reflect the action of natural selection removing introgressed alleles or chromosomes (Martin and Jiggins 2017): for example, m is very likely to be much lower than the proportion of hybrids in the population. Low levels of gene flow are consistent with the distinct species status, given *A. gambiae* and *A. arabiensis* have extensively overlapping geographical distributions, as is the fact that hybridization experiments failed to produce introgression in the reverse *A. gambiae* → *A. arabiensis* direction (Slotman et al. 2005). Nevertheless, as predicted by our theoretical calculations, even such low levels of migration can produce very high introgression probabilities when the data are analyzed under the MSci model. Indeed the estimates of $\varphi_{A\to G}$ from the same genomic data are greater than 90% for every autosomal arm except for the 2La inversion which has a different history (Table 2) (Flouris et al. 2019). The evolutionary history in this species complex may be close to the one depicted in Figure 9a (see Mallet et al. 2016, Fig. 3). A very low migration rate, with the proportion of migrants below one in 10 million or with far less than one migrant individual per generation, can have a huge impact on the genetic history of the species, such that a sequence sampled today from *G* will most likely trace its history to the introgressing parental species *A. arabiensis*. Both the major species tree and the dominant gene tree (Fig. 9b, c) reflect cross-species migration, even though the species phylogeny describing the order of species divergences is undoubtedly the one in Figure 9a.

One may imagine a situation in which the whole genome, including sex chromosomes, are affected by continuous migration, and the whole genome may consistently support the incorrect species phylogeny (Mallet et al. 2016, Fig. 3). It appears that even in such cases, the true history of species divergences including introgression or migration events may still be recoverable using genomic sequence data, given that the IM and MSci models (such as the models in Fig. 9a,b) make very different predictions on the distribution of gene trees (topologies and coalescent times).

ACKNOWLEDGMENTS

We thank three anonymous reviewers, the Associate Editor, and Editor for a number of constructive comments.

FUNDING

This work was 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).

REFERENCES

Allman E.S., Bannos H., Rhodes J.A. 2019. Nanuq: a method for inferring species networks from gene trees under the coalescent model. arXiv:1905.07050.

Anderson E. 1949. Introgressive hydridization. New York: John Wiley.

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

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

Degnan J.H. 2018. Modeling hybridization under the network multispecies coalescent. Syst. Biol. 67:786–799.
Durand E.Y., Patterson N., Reich D., Slatkin M. 2011. Testing for ancient admixture between closely related populations. Mol. Biol. Evol. 28:2239–2252.
Edelman N.B., Frandsen P.B., Miyagi M., Clavijo B., Davey J., Dikow R.B., Garcia-Accinelli G., Van Belleghem S.M., Patterson N., Neafsey D.E., Challis R., Kumar S., Moreira G. R.P., Salazar C., Chouteau M., Counterman B.A., Papa R., Blaxter M., Reed R.D., Dasmahapatra K.K., Kronforst M., Joron M., Jiggins C.D., McMillan W.O., Di Palma F., Blumberg A. J., Wakeley J., Jaffe D., Mallet J. 2019. Genomic architecture and introgression shape a butterfly radiation. Science 366:594–599.
Ellegren H., Smeds L., Burri R., Olason P.I., Backstrom N., Kawakami T., Kunstner A., Makinen H., Nadachowska-Brzyska K., Qvarnstrom A., Uebbing S., Wolf J.B.W. 2012. The genomic landscape of species divergence in *Ficedula* flycatchers. Nature 491:756–760.
Flouris T., Jiao X., Rannala B., Yang Z. 2018. Species tree inference with BPP using genomic sequences and the multispecies coalescent. Mol. Biol. Evol. 35:2585–2593.
Flouris T., Jiao X., Rannala B., Yang Z. 2019. A Bayesian implementation of the multispecies coalescent model with introgression for comparative genomic analysis. Mol. Biol. Evol. doi: 10.1093/molbev/msz296.
Folk R.A., Soltis P.S., Soltis D.E., Guralnick R. 2018. New prospects in the detection and comparative analysis of hybridization in the tree of life. Am. J. Bot. 105:364–375.
Fontaine M.C., Pease J.B., Steele A., Waterhouse R.M., Neafsey D.E., Sharakhov I.V., Jiang X., Hall A.B., Catteruccia F., Kakani E., Mitchell S.N., Wu Y.C., Smith H.A., Love R.R., Lawniczak M.K., Slotman M.A., Emrich S.J., Hahn M.W., Besansky N.J. 2015. Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science, 347:1258524.
Green R.E., Krause J., Briggs A.W., Maricic T., Stenzel U., Kircher M., Patterson N., Li H., Zhai W., Fritz M.H., Hansen N.F., Durand E.Y., Malaspinas A.S., Jensen J.D., Marques-Bonet T., Alkan C., Prufer K., Meyer M., Burbano H.A., Good J.M., Schultz R., Aximu-Petri A., Butthof A., Hober B., Hoffner B., Siegemund M., Weihmann A., Nusbaum C., Lander E.S., Russ C., Novod N., Affourtit J., Egholm M., Verna C., Rudan P., Brajkovic D., Kucan Z., Gusic I., Doronichev V.B., Golovanova L.V., Lalueza-Fox C., de la Rasilla M., Fortea J., Rosas A., Schmitz R.W., Johnson P.L., Eichler E.E., Falush D., Birney E., Mullikin J.C., Slatkin M., Nielsen R., Kelso J., Lachmann M., Reich D., Paabo S. 2010. A draft sequence of the Neandertal genome. Science 328:710–722.
Hahn M.W., Nakhleh L. 2016. Irrational exuberance for resolved species trees. Evolution 70:7–17.
Heled J., Drummond A.J. 2010. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27:570–580.
Hey J. 2010. Isolation with migration models for more than two populations. Mol. Biol. Evol. 27:905–920.
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 167:747–760.
Hey J., Chung Y., Sethuraman A., Lachance J., Tishkoff S., Sousa V.C., Wang Y. 2018. Phylogeny estimation by integration over isolation with migration models. Mol. Biol. Evol. 35:2805–2818.
Hobolth A., Andersen L., Mailund T. 2011. On computing the coalescence time density in an isolation-with-migration model with few samples. Genetics 187:1241–1243.
Hudson R. 1983. Testing the constant-rate neutral allele model with protein sequence data. Evolution 37:203–217.
Jukes T., Cantor C. 1969. Evolution of protein molecules. In: Munro H., editor. Mammalian protein metabolism. New York: Academic Press, p. 21–123.
Keightley, P. D., Ness, R. W., Halligan, D. L., and Haddrill, P. R. 2014. Estimation of the spontaneous mutation rate per nucleotide site in a *Drosophila melanogaster* full-sib family. Genetics 196: 313–320.
Kumar V., Lammers F., Bidon T., Pfenninger M., Kolter L., Nilsson M.A., Janke A. 2017. The evolutionary history of bears is characterized by gene flow across species. Sci. Rep. 7:46487.
Leaché A.D., Harris R.B., Rannala B., Yang Z. 2014. The influence of gene flow on Bayesian species tree estimation: a simulation study. Syst. Biol. 63:17–30.
Liu L., Pearl D.K. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst. Biol. 56:504–514.
Liu L., Yu L. 2011. Estimating species trees from unrooted gene trees. Syst. Biol. 60:661–667.
Liu L., Yu L., Pearl D.K., Edwards S.V. 2009. Estimating species phylogenies using coalescence times among sequences. Syst. Biol. 58:468–477.
Long C. and Kubatko L. 2018. The effect of gene flow on coalescent-based species-tree inference. Syst. Biol. 67:770–785.
Maddison W. 1997. Gene trees in species trees. Syst. Biol. 46:523–536.
Mallet J. 2007. Hybrid speciation. Nature 446:279–283.
Mallet J., Besansky N., Hahn M.W. 2016. How reticulated are species? BioEssays 38:140–149.
Mao Y., Economo E.P., Satoh N. 2018. The roles of introgression and climate change in the rise to dominance of *Acropora* corals. Curr. Biol. 28:3373–3382e5.
Martin S.H., Jiggins C.D. 2017. Interpreting the genomic landscape of introgression. Curr. Opin. Genet. Dev. 47:69–74.
Mirarab S., Reaz R., Bayzid M.S., Zimmermann T., Swenson M.S., Warnow T. 2014. Astral: genome-scale coalescent-based species tree estimation. Bioinformatics 30:i541–i548.
Nichols R. 2001. Gene trees and species trees are not the same. Trends Ecol. Evol. 16:358–364.
Nielsen R., Akey J.M., Jakobsson M., Pritchard J.K., Tishkoff S., Willerslev E. 2017. Tracing the peopling of the world through genomics. Nature 541:302.
Slotman M.A., Della Torre A., Calzetta M., Powell J.R. 2005. Differential introgression of chromosomal regions between *Anopheles gambiae* and *An. arabiensis*. Am. J. Trop. Med. Hyg. 73:326–335.
Solis-Lemus C., Ane C. 2016. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. PLoS Genet. 12:e1005896.
Solis-Lemus C., Yang M., Ane C. 2016. Inconsistency of species tree methods under gene flow. Syst. Biol. 65:843–851.
Solis-Lemus C., Bastide P., Ane C. 2017. Phylonetworks: a package for phylogenetic networks. Mol. Biol. Evol. 34:3292–3298.
Thawornwattana Y., Dalquen D., Yang Z. 2018. Coalescent analysis of phylogenomic data confidently resolves the species relationships in the *Anopheles gambiae* species complex. Mol. Biol. Evol. 35:2512–2527.
Wen D., Nakhleh L. 2018. Coestimating reticulate phylogenies and gene trees from multilocus sequence data. Syst. Biol. 67:439–457.
Wen D., Yu Y., Nakhleh L. 2016. Bayesian inference of reticulate phylogenies under the multispecies network coalescent. PLoS Genet. 12:e1006006.
Wiebe A., Longbottom J., Gleave K., Shearer F.M., Sinka M.E., Massey N.C., Cameron E., Bhatt S., Gething P.W., Hemingway J., Smith D.L., Coleman M., Moyes C.L. 2017. Geographical distributions of African malaria vector sibling species and evidence for insecticide resistance. Malar. J. 16:85.
Wilkinson-Herbots H.M. 2012. The distribution of the coalescence time and the number of pairwise nucleotide differences in a model of population divergence or speciation with an initial period of gene flow. Theor. Popul. Biol. 82:92–108.
Wu D.-D., Ding X.-D., Wang S., Wojcik J.M., Zhang Y., Tokarska M., Li Y., Wang M.-S., Faruque O., Nielsen R., Zhang Q., Zhang Y.-P. 2018. Pervasive introgression facilitated domestication and adaptation in the Bos species complex. Nat. Ecol. Evol. 2:1139–1145.
Xu B., Yang Z. 2016. Challenges in species tree estimation under the multispecies coalescent model. Genetics 204:1353–1368.
Yang Z. 1994. Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. Syst. Biol. 43:329–342.
Yang Z. 2000. Complexity of the simplest phylogenetic estimation problem. Proc. R. Soc. B: Biol. Sci. 267:109–116.
Yang Z. 2002. Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. Genetics 162:1811–1823.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591.

Yang Z. 2015. The BPP program for species tree estimation and species delimitation. Curr. Zool. 61:854–865.
Yang Z., Rannala B. 2014. Unguided species delimitation using DNA sequence data from multiple loci. Mol. Biol. Evol. 31:3125–3135.
Yu Y., Nakhleh L. 2015. A maximum pseudo-likelihood approach for phylogenetic networks. BMC Genomics 16 Suppl 10:S10.
Yu Y., Degnan J.H., Nakhleh L. 2012. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. PLoS Genet. 8:e1002660.
Yu Y., Dong J., Liu K.J., Nakhleh L. 2014. Maximum likelihood inference of reticulate evolutionary histories. Proc. Natl. Acad. Sci. USA 111:16448–16453.
Zhang C., Ogilvie H.A., Drummond A.J., Stadler T. 2018. Bayesian inference of species networks from multilocus sequence data. Mol. Biol. Evol. 35:504–517.
Zhu J., Yu Y., Nakhleh L. 2016. In the light of deep coalescence: revisiting trees within networks. BMC Bioinformatics 17:415.
Zhu S., Degnan J.H. 2017. Displayed trees do not determine distinguishability under the network multispecies coalescent. Syst. Biol. 66:283–298.
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. 29:3131–3142.

Corrigendum. Jiao X, Flouri T, Rannala B, Yang Z. 2020. The impact of cross-species gene flow on species tree estimation. *Systematic Biology* 69:830-847.

A coding error affected the numerical correctness of figures 6c and 7a&b in the paper. In the legend to figure 7, $\varphi^* = 0.5395$ and $\tau_R^{(I)*} = 0.02280$ should read $\varphi^* = 0.57158$ and $\tau_R^{(I)*} = 0.02653$. The conclusions of the paper are not affected. The correct figures with legends are as follows. We thank Yuttapong Thawornwattana for pointing out the errors.

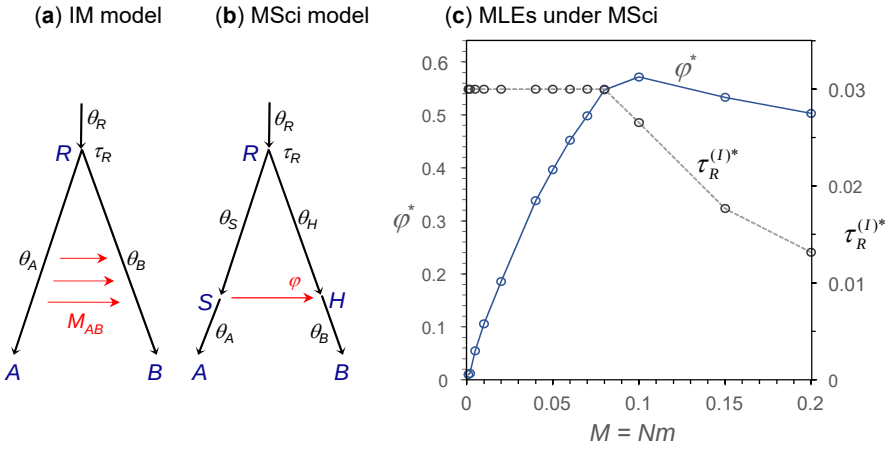

Figure 6: The (**a**) migration and (**b**) introgression models for two species (A,B), and (**c**) the pseudo-true parameter values of φ and $\tau_R^{(I)}$ in the introgression model plotted against the migration rate $M_{AB}(= N_B m_{AB} = w\theta_B/4)$. The pseudo-true parameter values are parameter values under the introgression model $\boldsymbol{\theta}^{(I)}$ that minimizes the KL divergence of equation 32: they are the limiting values of the MLEs of $\boldsymbol{\theta}^{(I)}$, when the number of loci $L \to \infty$, when the MSci model is fitted to data of L loci generated under the IM model, with two sequences (one from each species) of infinite length per locus. Other parameters in the migration model are fixed at $\tau_R^{(M)} = 0.03$ and $\theta = 0.01$ (for all populations). In the MSci model, $\theta = 0.01$ is assumed for all populations, while φ and $\tau_R^{(I)}$ are estimated by minimizing the KL divergence (equation 32).

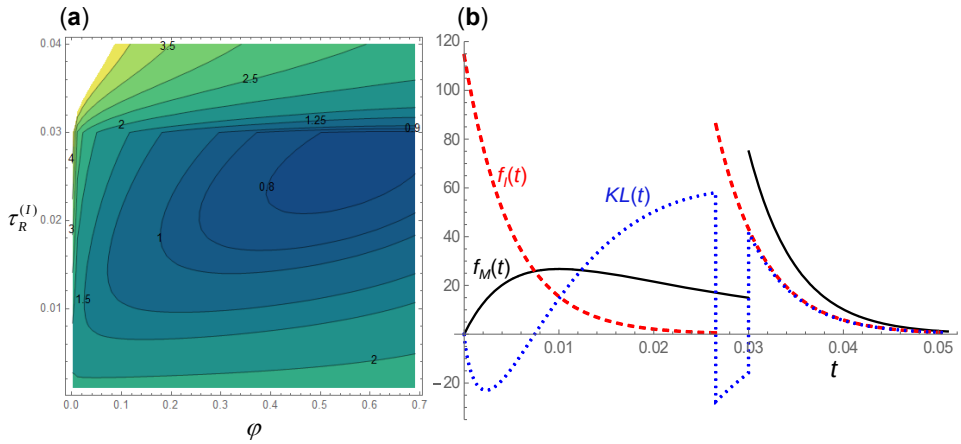

Figure 7: (**a**) Contour plot of $D(\boldsymbol{\theta}_M||\boldsymbol{\theta}_I)$ (equation 32) as a function of φ and $\tau_R^{(I)}$ in the MSci model for the case of $M = 0.1$ or $w = 40$. Parameters in the IM model are $\tau_R^{(M)} = 0.03$ and $\theta = 0.01$ for all populations. The MLE under the MSci model, which minimizes D, is at $\varphi^* = 0.57158$ and $\tau_R^{(I)*} = 0.02653$. Note that in the MSci model, $\theta = 0.01$ is fixed for all populations and only φ and $\tau_R^{(I)}$ are optimized as $\tau_H = 0$. The D surface is not smooth at $\tau_R^{(I)} = \tau_R^{(M)} = 0.03$. (**b**) The densities $f_M(t)$ (equation 30) and $f_I(t)$ (equation 31) as well as the integrand, $f_M(t)\log\frac{f_M(t)}{f_I(t)}$, of equation 32, plotted at the MLEs of (**a**). In other words, the red dashed curve $f_I(t)$ is the best fit of the MSci model to the infinite-sized data under the IM model represented by the black solid curve $f_M(t)$. Note that $f_M(t)$ and $f_I(t)$ have one discontinuity point, while the integrand has two.