OXFORD UNIVERSITY PRESS | Molecular Biology and Evolution

Inference of gene flow between species under misspecified models

Journal:	*Molecular Biology and Evolution*
Manuscript ID	MBE-22-0354.R1
Manuscript Type:	Article
Date Submitted by the Author:	03-Jul-2022
Complete List of Authors:	Huang, Jun; Capital Medical University, School of Biomedical Engineering Thawornwattana, Yuttapong; Harvard University, Organismic and Evolutionary Biology Flouris, Thomas; University College London, Biology Mallet, James; Harvard University, OEB Yang, Ziheng; University College London, Biology;
Key Words:	Gene flow, model misspecification, multispecies coalescent, introgression, BPP, species tree
Subject Section:	Methods

Journal (2050), Vol. 00, No. 000, pp. 1–27
DOI: 10.1000/xxx/xxx000

Inference of gene flow between species under misspecified models

Jun Huang,[1,†] Yuttapong Thawornwattana (orcid: 0000-0003-2745-163X)[2,†], Tomáš Flouri (orcid: 0000-0002-8474-9507)[3], James Mallet (orcid: 0000-0002-3370-0367)[2], and Ziheng Yang (orcid: 0000-0003-3351-7981) [1,*]
[1]School of Biomedical Engineering, Capital Medical University, Beijing, 100069, P.R. China
[2]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA
[3]Department of Genetics, Evolution and Environment, University College London, London, WC1E 6BT, UK
[†]Those authors contributed equally to this work.

Received on xxxx, revised on xxxx, accepted on xxxx

Genomic sequence data provide a rich source of information about the history of species divergences and interspecific hybridization or introgression. Despite recent advances in genomics and statistical methods, it remains challenging to infer gene flow, and as a result, one may have to estimate introgression rates and times under misspecified models. Here we use mathematical analysis and computer simulation to examine estimation bias and issues of interpretation when the model of gene flow is misspecified in analysis of genomic datasets, for example, if introgression is assigned to the wrong lineages. In the case of two species, we establish a correspondence between the migration rate in the continuous migration model and the introgression probability in the introgression model. When gene flow occurs continuously through time but in the analysis is assumed to occur at a fixed time point, common evolutionary parameters such as species divergence times are surprisingly well estimated. However, the time of introgression tends to be estimated towards the recent end of the period of continuous gene flow. When introgression events are assigned incorrectly to the parental or daughter lineages, introgression times tend to collapse onto species divergence times, with introgression probabilities underestimated. Overall, our analyses suggest that simple introgression models can be used to extract useful information concerning species divergence times and gene flow even when the model is misspecified. However, for reliable inference of gene flow it is important to include multiple samples per species, in particular, from hybridizing species.
Gene flow | model misspecification | multispecies coalescent | introgression | BPP | species tree

Introduction

Hybridization can enhance variation in recipient species, and has long been recognized as an important process in plants that can stimulate the origin of new species (e.g., Anderson, 1949; Mallet, 2007). Analyses of genomic data in the past decade has highlighted the prevalence of introgression in animals as well, including butterflies (Martin *et al.*, 2013), birds (Ellegren *et al.*, 2012), and bears (Liu *et al.*, 2014; Kumar *et al.*, 2017). Introgression may involve either sister or non-sister species and may play an important role in ecological adaptation (Mallet *et al.*, 2016; Martin and Jiggins, 2017). Introgression can be a major contributor of genealogical variation across the genome and gene tree–species tree discordance, in addition to ancestral polymorphism or delayed coalescence (Maddison, 1997; Nichols, 2001).

There is a long history of studies in population genetics of models of population subdivision and migration (Wright, 1943; Malecot, 1948; Slatkin, 1987), and a number of methods have been developed to estimate the migration rate between populations (Bahlo and Griffiths, 2000; Beerli and Felsenstein, 1999, 2001). An important limitation of models of population subdivision, when applied to data from different species or subspecies, is that they do not account for the divergence history of the populations or species. Introducing a population/species phylogeny into models of population subdivision not only improves the realism of the model but also opens up opportunities for addressing a number of interesting questions in evolutionary biology, such as estimation of species divergence times and ancestral population sizes, delineating species boundaries, and estimating the direction, rate and timing of gene flow (Jiao *et al.*, 2021).

Two classes of models of gene flow have been developed that accommodate the phylogeny of the species, both of which are extensions of the multi-species coalescent (MSC) model (Rannala and Yang, 2003). The first is the MSC-with-migration model (MSC+M, or isolation-with-migration or IM model, Hey and Nielsen, 2004; Hey, 2010; Zhu and Yang, 2012; Dalquen *et al.*, 2017; Hey *et al.*, 2018), which assumes that two species exchange migrants at a certain rate over an extended time period. The rate of

*Correspondence: z.yang@ucl.ac.uk

gene flow is measured by the proportion of migrants (m) in the receiving population per generation or by the population migration rate, $M = Nm$, the expected number of immigrants per generation, where N is the (effective) population size of the receiving population. We note that the isolation-with-initial-migration (IIM) model of Costa and Wilkinson-Herbots (2017) is an instance of the IM model, which assumes that gene flow occurs after species divergence initially but stops after a period of time, when reproductive isolation has been fully established (see below). The second class of models of gene flow is the MSC-with-introgression (MSci) model (Flouri *et al.*, 2020), also known as multispecies network coalescent model (MSNC; Wen and Nakhleh, 2018; Zhang *et al.*, 2018), which assumes that gene flow occurs at fixed time points in the past. The rate of gene flow is measured by the introgression probability (φ or γ), which is the proportion of successful immigrants in the population at the time of introgression.

In the real world, introgressed alleles may be removed by natural selection because they are involved in hybrid incompatibility and are deleterious in the genetic background of the recipient population (Dobzhansky, 1937; Muller, 1942) or because they are linked to such loci. Thus the rate of gene flow (M in IM or φ in MSci), when IM or MSci models are used, reflect the long-term effects of selection and drift as well as hybridization or introgression (Martin and Jiggins, 2017). Such an effective rate of gene flow may be expected to vary across the genome, influenced by the presence of loci in the genomic region important in ecological adaptation as well as the local recombination rate (Schumer *et al.*, 2018; Martin *et al.*, 2019; Edelman *et al.*, 2019). The rate can also vary over time, depending on geological or ecological events that cause changes in the ecology and distribution of the species, altering the chances for two species to exchange genes. One can envisage models of gene flow in which the rate varies over time and across genomic regions. For the present, such extended models are not yet available, and the feasibility of fitting such parameter-rich models to genomic datasets is unexplored. IM and MSci models implemented to date (Dalquen *et al.*, 2017; Hey *et al.*, 2018; Wen and Nakhleh, 2018; Zhang *et al.*, 2018; Flouri *et al.*, 2020) assume constant rates. They represent idealized extremes and should be considered as a first approximation when applied to analyze genomic sequence data.

Currently the most commonly used methods for inferring gene flow from multilocus sequence data are approximate methods (also known as summary or heuristic methods) based on summaries of the data that are not sufficient statistics (Fisher, 1922). They are computationally very efficient but do not make use of all the information in the data (see Degnan, 2018; Elworth *et al.*, 2019; Jiao *et al.*, 2021; Hibbins and Hahn, 2022 for recent reviews). The first class of such methods include those based on genome-wide site-pattern counts, such as the D-statistic (Green *et al.*, 2010; Durand *et al.*, 2011) and HyDe (Meng and Kubatko, 2009; Blischak *et al.*, 2018). For example, D is based on site-pattern counts for four species (three species plus an outgroup) (Green *et al.*, 2010; Durand *et al.*, 2011), and ignores information in genealogical variation across the genome (Lohse and Frantz, 2014; Shi and Yang, 2018). Both deep coalescent and gene flow create stochastic fluctuations in the genealogical history (gene tree topology and coalescent times) across the genome, with the probability distribution of the gene tree specified by the parameters in the multispecies coalescent model with gene flow. There is thus important information about those parameters in such genealogical variation, which is ignored by methods based on genome-wide site-pattern counts. See Zhu and Yang (2021, fig. 3a&c) for an illustration of the dramatic information loss in the context of species tree estimation resulting from pooling sites across loci.

Another major class of approximate methods take the two-step approach of inferring gene trees for multiple loci and using these as input data (Yu *et al.*, 2012, 2014; Yu and Nakhleh, 2015; Solis-Lemus and Ane, 2016; Wen *et al.*, 2016). Note that the reconstructed gene tree for a locus is a summary of the sequence alignment at the locus so that the two-step approach is a summary method as well. This approach typically ignores phylogenetic errors in the gene trees. Errors in estimated branch lengths (coalescent times) in gene trees are known to cause serious problems both for estimation of the species tree (DeGiorgio and Degnan, 2014) and for inference of gene flow (Wen *et al.*, 2016). Thus gene-tree branch lengths are typically ignored in these methods (Degnan, 2018), leading to considerable information loss.

Most current approximate methods are unable to identify many parameters in the coalescent model with gene flow. For example the D statistic can be used to detect gene flow between non-sister species but does not attempt to infer the direction, timing, or strength of gene flow. Yet these parameters are important for characterizing the history of species divergence and interspecific gene flow. Also current summary methods cannot infer gene flow between sister species, or estimate the rates of gene flow that occurs in both directions. Yet in nature it is likely that gene flow does occur between sister species and that two species do exchange genes in both directions. See Zhu and Yang (2021) and Yang and Flouri (2022) for recent discussions of limitations of approximate methods in analyses of genomic sequence data under the MSC model with and without gene flow.

In this paper, we focus on exact or likelihood methods of inference under the MSC-with-migration (MSC-M or IM) or with-introgression (MSci) models

using data of multilocus sequence alignments (Hey *et al.*, 2018; Wen and Nakhleh, 2018; Zhang *et al.*, 2018; Flouri *et al.*, 2020). Likelihood methods integrate over all possible gene trees underlying the sequence alignments, making efficient use of information about the model and parameters in both the gene tree topology and coalescent times while accommodating their uncertainties. They typically involve a heavy computational load. However, recent algorithmic improvements have made it possible to apply the MSci model to genome-scale datasets with >10,000 loci (Flouri *et al.*, 2020). Inferring introgression events or constructing an introgression model using genomic sequence data, however, remains a challenging task, even when a binary species tree is specified, onto which introgression events can be added (Thawornwattana *et al.*, 2022; Ji *et al.*, 2022). See Discussion for an overview of currently available methods for inferring gene flow on a species phylogeny. For these and many other reasons, the model of gene flow that we assume in our data analysis may often be incorrect. An important question is to what extent inference of gene flow, and in particular, estimation of the timing and rate of gene flow, can still be achieved when the model of gene flow is misspecified. The impact of model misspecification on estimation of other evolutionary parameters such as species divergence times is also of major concern.

Here we use mathematical analysis and computer simulation to probe estimation of the rate and time of introgression when the model is misspecified. While there are many ways in which the assumed model is wrong, we are particularly interested in a few types that are likely in real data analyses (Thawornwattana *et al.*, 2022; Finger *et al.*, 2022).

First, gene flow may be occurring continuously during a time period but an MSci model is fitted to the genomic data, which assumes that gene flow occurred at a particular time point (e.g., Wen and Nakhleh, 2018; Jiao *et al.*, 2020). We are here interested in whether species divergence times and ancestral population sizes are affected by the misspecification, and how the migration rate in the migration model (M) corresponds to the introgression probability in the MSci model (φ). The case of two species is analytically tractable. We calculate the limit of the maximum likelihood estimates (MLEs) of introgression probability and introgression time when the data size (the number of loci) approaches infinity when the data are generated under the IM model. We use computer simulation to verify and extend the analytical calculation.

Second, the introgression event may be assigned to a wrong branch on the species tree, for example, to a parental or daughter branch of the genuine introgression lineage. Alternatively, introgression may involve species that have since gone extinct or are not included in the data sample. The presence of such ghost species is known to mislead inference of the history of gene flow for the sampled species (Beerli, 2004; Tricou *et al.*, 2022). Thus we conducted simulation to examine the impact of unsampled species on the inference of gene flow. In all cases of incorrect branch specification, we use BPP to analyze multilocus sequence data simulated under the MSci model (Flouri *et al.*, 2018, 2020) to assess the impacts of model misspecification on estimation of model parameters (such as species divergence and introgression times, population sizes, and introgression probabilities). While BPP is our own Bayesian implementation of the MSci model applied to multilocus sequence data, the results in this paper should apply to all full likelihood methods.

Results

Correspondence between the IM and MSci models in the case of two species

Notation and definition of parameters

Following Jiao *et al.* (2020), we study the asymptotic behavior of Bayesian parameter estimation under the introgression (MSci) model when the data are generated under the migration (IM) model in the case of two species, with one sequence per species per locus (fig. 1). Note that here we focus on this simple case because it is analytically tractable; nevertheless, our Bayesian implementation in BPP (Flouri *et al.*, 2020) can accommodate an arbitrary number of species and an arbitrary number of sequences per species per locus. We assume an infinite number of loci, and the data at each locus consist of a pair of sequences (a, b) from the two species, with x differences at n sites. The coalescent time t for the locus is unknown and underlies the observed differences. Jiao *et al.* (2020) analyzed the IM model (fig. 1**a**) and assumed infinite sequence length ($n = \infty$) so that the true coalescent time between the two sequences (t) is known. Here we accommodate random fluctuations in the number of mutations due to finite sequence length and consider three variants of the migration model.

In the basic IM model, species A and B diverged at time τ_R and there has since been gene flow from A to B at the rate of $M_{AB} = M$ migrants per generation (fig. 1**a**). The IIM model assumes that migration occurred initially after species divergence but stopped at time $\tau_T > 0$ (Costa and Wilkinson-Herbots, 2017), and is represented by an IM model for three species including a ghost species (fig. 1**b**). Here the ghost does not necessarily represent a real species but is a mathematical device for specifying the IIM model. We also consider a secondary contact (SC) model, in which two species initially had complete isolation but came into contact at a certain time point (τ_T) with ongoing gene flow at the rate of M ever since. This is similarly specified using a ghost species at time point τ_T (fig. 1**c**). The migration model involves three types of parameters: species divergence times

(τ_R, τ_T), population sizes for extant and extinct species $(\theta_A, \theta_B, \theta_T, \theta_R)$, and the (population) migration rate M. The population size parameter for any species with (effective) population size N is defined as $\theta = 4N\mu$, where μ is the mutation rate per site per generation. We refer to a branch on the species tree by its daughter node so that branch RA is also branch A, with population size parameter θ_A. Both divergence times (τ) and population sizes (θ) are measured by the expected number of mutations per site.

Asymptotic theory

We first consider the IIM model (fig. 1**b**), where the IM model (fig. 1**a**) is a special case with $\tau_T = 0$. The backwards-in-time process of coalescent and migration in time interval (τ_T, τ_R) is described by a Markov chain with three states: AB, AA and A (Notohara, 1990). Here AB is the initial state, with two sequences in the sample, one in A and another in B; AA means both sequences are in A (in other words, sequence b is traced back into A); and A means one sequence in A (in other words, sequence b is traced back into A and has coalesced with sequence a). Note that in the Markov chain, time runs backwards, so the transition from AB to AA means migration of a sequence from A to B in the real world. With time measured by the expected number of mutations per site, the generator matrix for the Markov chain is (see Jiao *et al.*, 2020; Notohara, 1990)

$$Q = \begin{array}{c|ccc} & AB & AA & A \\ \hline AB & -w & w & 0 \\ AA & 0 & -\frac{2}{\theta_A} & \frac{2}{\theta_A} \\ A & 0 & 0 & 0 \end{array} \qquad (1)$$

where $w = m_{AB}/\mu = 4M_{AB}/\theta_B$ is the mutation-scaled migration rate, and $\frac{2}{\theta_A}$ is the coalescent rate in population A, with one time unit being the expected time taken to accumulate one mutation per site. Q has eigenvalues $\lambda_1 = 0$, $\lambda_2 = -\frac{2}{\theta_A}$, and $\lambda_3 = -w$.

Let the transition probability matrix over time t be $P(t) = \{p_{ij}(t)\} = \mathrm{e}^{Qt}$, where $p_{ij}(t)$ is the probability that the Markov chain will be in state j time t later given that it is in state i at time 0. This is

$$P(t) = \begin{bmatrix} \mathrm{e}^{-wt} & \frac{\theta_A w}{2-\theta_A w}\left(\mathrm{e}^{-wt} - \mathrm{e}^{-\frac{2}{\theta_A}t}\right) & 1 - \frac{2\mathrm{e}^{-wt} - \theta_A w \mathrm{e}^{-\frac{2}{\theta_A}t}}{2-\theta_A w} \\ 0 & \mathrm{e}^{-\frac{2}{\theta_A}t} & 1 - \mathrm{e}^{-\frac{2}{\theta_A}t} \\ 0 & 0 & 1 \end{bmatrix}. \qquad (2)$$

The probability density of coalescent time t is thus

$$f_{\mathrm{m}}(t) = \begin{cases} P_{AB,AA}(t-\tau_T)\frac{2}{\theta_A}, & \text{if } \tau_T < t < \tau_R, \\ [1 - P_{AB,A}(\tau_R - \tau_T)]\frac{2}{\theta_R}\mathrm{e}^{-\frac{2}{\theta_R}(t-\tau_R)}, & \text{if } t > \tau_R \end{cases}$$

$$= \begin{cases} \frac{2w}{2-\theta_A w}\left[\mathrm{e}^{-w(t-\tau_T)} - \mathrm{e}^{-\frac{2}{\theta_A}(t-\tau_T)}\right], & \\ & \text{if } \tau_T < t < \tau_R, \\ \left[\frac{2}{2-\theta_A w}\mathrm{e}^{-w(\tau_R-\tau_T)} - \frac{\theta_A w}{2-\theta_A w}\mathrm{e}^{-\frac{2}{\theta_A}(\tau_R-\tau_T)}\right]\frac{2}{\theta_R}\mathrm{e}^{-\frac{2}{\theta_R}(t-\tau_R)}, & \\ & \text{if } t > \tau_R. \end{cases} \qquad (3)$$

This is a function of $w = 4M_{AB}/\theta_B$ but not of M_{AB} and θ_B individually. The parameters specifying the density are thus $\Theta_{\mathrm{m}} = (w, \theta_A, \theta_R, \tau_R, \tau_T)$.

Under the secondary-contact (SC) model (fig. 1**c**), the coalescent-with-migration process over the time interval $(0, \tau_T)$ is described by the Markov chain of eq. 1. Given the parameters Θ_{m}, the probability density of coalescent time t is

$$f_{\mathrm{sc}}(t) = \begin{cases} P_{AB,AA}(t)\frac{2}{\theta_A}, & \text{if } 0 < t < \tau_T, \\ P_{AB,AA}(\tau_T)\frac{2}{\theta_A}\mathrm{e}^{-\frac{2}{\theta_A}(t-\tau_T)}, & \text{if } \tau_T < t < \tau_R, \\ \left[P_{AB,AA}(\tau_T)\frac{2}{\theta_A}\mathrm{e}^{-\frac{2}{\theta_A}(\tau_R-\tau_T)} + P_{AB,AB}(\tau_T)\right] & \\ \quad \times \frac{2}{\theta_R}\mathrm{e}^{-\frac{2}{\theta_R}(t-\tau_R)}, & \text{if } t > \tau_R \end{cases}$$

$$= \begin{cases} \frac{w\theta_A}{2-w\theta_A}\left[\mathrm{e}^{-wt} - \mathrm{e}^{-\frac{2}{\theta_A}t}\right]\frac{2}{\theta_A}, & \text{if } 0 < t < \tau_T, \\ \frac{w\theta_A}{2-w\theta_A}\left[\mathrm{e}^{-w\tau_T} - \mathrm{e}^{-\frac{2}{\theta_A}\tau_T}\right]\frac{2}{\theta_A}\mathrm{e}^{-\frac{2}{\theta_A}(t-\tau_T)}, & \text{if } \tau_T < t < \tau_R, \\ \left[\frac{w\theta_A}{2-w\theta_A}\left[\mathrm{e}^{-w\tau_T} - \mathrm{e}^{-\frac{2}{\theta_A}\tau_T}\right]\mathrm{e}^{-\frac{2}{\theta_A}(\tau_R-\tau_T)} + \mathrm{e}^{-w\tau_T}\right] & \\ \quad \times \frac{2}{\theta_R}\mathrm{e}^{-\frac{2}{\theta_R}(t-\tau_R)}, & \text{if } t > \tau_R. \end{cases} \qquad (4)$$

Similarly under the MSci model, with parameters $\Theta_{\mathrm{i}} = (\varphi, \theta_R, \theta_S, \tau_R, \tau_S)$ (fig. 1**d**), we have (Jiao *et al.*, 2020)

$$f_{\mathrm{i}}(t) = \begin{cases} \varphi\frac{2}{\theta_S}\mathrm{e}^{-\frac{2}{\theta_S}(t-\tau_S)}, & \text{if } \tau_S < t < \tau_R, \\ \left[\varphi\mathrm{e}^{-\frac{2}{\theta_S}(\tau_R-\tau_S)} + (1-\varphi)\right]\frac{2}{\theta_R}\mathrm{e}^{-\frac{2}{\theta_R}(t-\tau_R)}, & \text{if } t > \tau_R. \end{cases} \qquad (5)$$

Given the coalescent time t for a locus, the probability of observing x differences at n sites under the JC mutation model (Jukes and Cantor, 1969) is given by the binomial probability

$$f(x|t) = \left(\tfrac{3}{4} - \tfrac{3}{4}\mathrm{e}^{-\frac{8}{3}t}\right)^x \cdot \left(\tfrac{1}{4} + \tfrac{3}{4}\mathrm{e}^{-\frac{8}{3}t}\right)^{n-x}. \qquad (6)$$

The marginal probability of observing x differences at n sites, under both the migration (IM, IIM, SC) and introgression (MSci) models, is

$$f(x|\Theta) = \int_0^\infty f(x|t) f(t|\Theta)\mathrm{d}t, \qquad (7)$$

where $f(t|\Theta)$ is given by eqs. 3, 4 or 5.

For analytical tractability of the likelihood (eq. 7), we assume the infinite-sites mutation model instead of JC, and replace the binomial likelihood by a Poisson approximation

$$f(x|t) = \frac{1}{x!}(2nt)^x \mathrm{e}^{-2nt}. \qquad (8)$$

This is derived in Appendix A, as eq. A6 for the IM (with $\tau_T = 0$) and IIM (with $\tau_T > 0$) models, eq. A7 for the SC model, and eq. A9 for the MSci model.

Consider the analysis of the data under the MSci model, which are generated under any of the migration

models (IM, IIM, SC). When the number of loci $L \to \infty$, the MLE $\hat{\Theta}_{\rm i}$ under MSci will converge to $\Theta_{\rm i}^*$, which minimizes the Kullback-Leibler (KL) divergence

$$D(\Theta_{\rm m} \parallel \Theta_{\rm i}) = \sum_{x=0}^{n} f_{\rm m}(x|\Theta_{\rm m}) \log \frac{f_{\rm m}(x|\Theta_{\rm m})}{f_{\rm i}(x|\Theta_{\rm i})}, \qquad (9)$$

which is a measure of distance from the fitting introgression model to the true migration model. Here $\Theta_{\rm m}$ are fixed while $\Theta_{\rm i}$ are being estimated. The limiting values $\Theta_{\rm i}^*$ as $L \to \infty$ are also known as the *pseudo-true parameter values* for the misspecified MSci model. The BFGS optimization routine in PAML (Yang, 2007) is used to minimize eq. 9 to obtain the MLEs.

We are in particular interested in the introgression probability φ and the introgression time τ_S. Note that under the migration model, the probability that any lineage from species B traces back to A is

$$\varphi_0 = 1 - {\rm e}^{-\frac{4M}{\theta_B}\Delta\tau} = 1 - {\rm e}^{-\frac{4M}{\theta_B}(\tau_R - \tau_T)}, \qquad (10)$$

where $\Delta\tau$ is the time period of gene flow (fig. 1**a**-**c**) and where we write M_{AB} as M. Eq. 10 gives the expected proportion of migrants under the true migration model. When M is small, $\varphi_0 \approx \frac{4M}{\theta_B}\Delta\tau$, which is also given by equating the expected total number of migrants under the two models: $N_B\varphi_0 \approx m_{AB}N_B\Delta\tau/\mu$. Note that $m_{AB}N_B$ is the expected number of migrants per generation and $\Delta\tau/\mu$ is the number of generations with gene flow.

It may be noted that the theory of eq. 9 can be used to study the limiting parameter estimates (when the number of loci $L \to \infty$) in the migration model when the true model is the introgression model. One has only to flip the roles of $f_{\rm m}(x|\Theta_{\rm m})$ and $f_{\rm i}(x|\Theta_{\rm i})$ in eq. 9. This is not pursued in this paper.

Asymptotic results under the IM model

We used the asymptotic theory (eq. 9) to obtain the MLEs ($\Theta_{\rm i}^*$) under the MSci model (fig. 1**d**) when the data consist of an infinite number of loci, with one sequence of length n per species per locus, generated under the IM, IIM or SC models (fig. 1**a**-**c**). The parameter values used ($\Theta_{\rm m}$) are shown in figure 1. The MLEs are shown in figure 2 and the true and best-fitting distributions of the coalescent time t are shown in figure S1 for the IM model. The corresponding results for the IIM and SC models are in figures S2-S5, to be discussed in the next sections.

We use five methods (a-e) to fit the MSci model, with method d estimating all five parameters, while the others have some parameters fixed (fig. 2). We examined the effects of the sequence length (n) and the migration rate (M). When the data are analyzed under the MSci model, five parameters are identifiable: $\Theta_{\rm i} = (\tau_R, \tau_S, \theta_R, \theta_S, \varphi)$ (fig. 1**d**). Population sizes θ_A, θ_B, and θ_H are not identifiable because no coalescent events can occur in those populations given one sequence per species per locus. Population size θ_S is identifiable as it is possible for both sequences a and b to be traced back to population S. Nevertheless, one expects the information concerning θ_S to be very weak in datasets of two sequences per locus, especially at low migration rates. In methods c and d, θ_S and φ are estimated as free parameters. The application of the misspecified MSci model (to data generated under the IM model) led to unreasonably large estimates of θ_S (as large as 0.5 mutations per site), and the poor estimates of θ_S caused φ to be poorly estimated as well. This is partly because of our use of only one sequence per species per locus in the analytical theory. Thus we do not emphasize the analyses using methods c and d, and focus instead on methods a, b, and e, in which θ_S is fixed (at the true value θ_0 in methods a and b, or constrained to be equal to θ_R in method e). In the simulation below, we evaluate the impact of the number of sequences sampled per species.

In the IM model, migration events occur throughout the time interval $(0, \tau_R)$, at the rate of M migrants per generation (fig. 1). When such data are analyzed under the introgression model, a simple expectation might be that the introgression time τ_S should be the average $\tau_R/2$ while the introgression probability may be given by the expected proportion φ_0 of eq. 10. However, as we show below, this expectation is too simplistic. We discuss the introgression time τ_S first.

Consider the case where the true coalescent time is known (or $n = \infty$). Given the data-generating IM model, there is a strictly positive probability for the coalescent time t to be in the interval $0 < t < \varepsilon$ for any small constant $\varepsilon > 0$. In other words, there must exist loci at which the coalescent time t is arbitrarily close to 0 (fig. S1). In the MSci model, sequences a and b cannot coalesce until they are in the same population S, so that $\tau_S < t$. When the MSci model is fitted to data generated under the IM model, $\hat{\tau}_S$ is dominated by the minimum rather than the average coalescent time, and $\hat{\tau}_S \to \tau_S^* = 0$ when the number of loci $L \to \infty$ (and when the true coalescent time is known). Even though migration events occur throughout the time interval $(0, \tau_R)$, the best the MSci model can do is to lump all migration events to one time point at $\tau_S^* = 0$ (fig. 2b&e).

When the sequence length is finite ($n < \infty$), the coalescent time is not observed and is reflected in the sequence divergence or the number of mutations (x). Whatever the true coalescent time, there is a positive probability of observing no mutations between the two sequences, so that $x = 0$ may not be very strong evidence that the coalescent time is $t \approx 0$. As a result, the MLE $\hat{\tau}_S$ reflects not only the minimum coalescent time, but also the whole distribution (fig. S1). In other words, two identical sequences between species may be 'interpreted' by the MSci model as being due to random mutational fluctuations with a strictly positive coalescent time. We thus have $\tau_S^* > 0$, different from

the case where the coalescent time is known without error ($n = \infty$). Nevertheless, one expects τ_S^* to be closer to 0 than to τ_R, especially if the number of sites is large. Indeed in our calculations, $\tau_S^* \ll \frac{1}{2}\tau_R$ (fig. 2b&e).

Next we consider the introgression probability φ and again focus on methods a, b, & e (fig. 2). The estimate $\hat{\varphi}$ increases nearly linearly when M is small ($< \frac{1}{4}$, say) but tails off at large M. All estimates are smaller than φ_0 of eq. 10 but they are close at low migration rates (with $M < \frac{1}{4}$ and $\varphi < \frac{1}{4}$, say) (fig. 2a,b&e). We defer to a later section a detailed discussion of the estimation of φ, contrasting the IM, IIM, and SC models.

Finally the estimated divergence time between the two species τ_R matched the true values at low migration rates but was underestimated at high migration rates, with the ancestral population size θ_R overestimated (fig. 2). It may be tempting to interpret the underestimation of τ_R (and overestimation of θ_R) by the MSci model as being due to the difficulty of distinguishing complete isolation with recent species divergence from introgression or of distinguishing migration and coalescent events close to species divergence from ancestral polymorphism. However, this does not appear to be a correct interpretation.

We examined the true and fitted distributions of the coalescent time (fig. S1). When there is no migration ($M = 0$), the MSci model is correct, and the parameter estimates converge to the true values, with a perfect fit to the density $f_{\rm m}(t)$. At low migration rates ($M \leq 0.1$, say), the MSci model fits the density $f_{\rm m}(t)$ very well, with the discontinuity points in the true and fitting distributions coinciding; in other words, $\tau_R^* = \tau_R^{\rm m}$. At the medium migration rate of $M = 1$, the species divergence time τ_R is still correctly estimated even though the fit to the density is poor (fig. S1). At high migration rates (with $M \geq 1.4$, say), the true density has a mode in the interval $(0, \tau_R^{\rm m})$, dropping off at the discontinuity point at $\tau_R^{\rm m}$. The best fitting density starts from 0, with an exponential decay, and has a discontinuity point at τ_R with again an exponential decay. This best-fitting density is a poor fit, and the discontinuity point τ_R^* is moved to smaller values as an attempt to accommodate the migration and coalescent events in the middle of the interval $(0, \tau_R^{\rm m})$ to improve the fit (judged by the KL divergence). Thus τ_R is underestimated ($\tau_R^* < \tau_R^{\rm m}$). As a result, the population size parameter θ_R is overestimated, as those two parameters tend to be strongly negatively correlated (e.g., Burgess and Yang, 2008). In other words, the intermediate coalescent times in the interval $(0, \tau_R)$, which occur at a large proportion of loci, are accommodated or misinterpreted by the MSci model by using a more recent species divergence time (τ_R) and a larger ancestral population size (θ_R). Coalescent times in the range $\tau_R^* < t < \tau_R^{\rm m}$, which represent true migration events in the IM model, are then misinterpreted as coalescent events in the ancestral population R in the MSci model, so that the estimated introgression probability φ^* is substantially lower than the expected proportion φ_0 (eq. 10).

Asymptotic results under the IIM model

When data are generated under the IIM model (fig. 1**b**) and analyzed under the MSci model (fig. 1**d**), the results (figs. S2&S3) show similar patterns to those under the IM model discussed above. Similarly, θ_S is difficult to estimate using two sequences per locus in methods c and d, and the poor estimates of θ_S affects the estimation of the introgression probability φ. Thus we focus on methods a, b, and e, in which θ_S is fixed or constrained, and on the introgression time and introgression probability.

In the IIM model, migration events occur throughout the time interval (τ_T, τ_R) (fig. 1**b**), but the estimate of the introgression time is dominated or influenced by the minimum coalescent time, so that $\tau_S^* = \tau_T$ when the coalescent time is known (with $n = \infty$), and $\tau_S^* > \tau_T$ when n is finite. In the latter case τ_S^* was closer to τ_T than to τ_R (fig. S2).

The introgression probability φ^* grew almost linearly with the migration rate M when M was small (with $M \leq 0.2$, say), and this estimate was close to the expectation φ_0 of eq. 10 (fig. S2a, b&e). At high migration rates, eq. 10 gave a serious overestimate. This 'bias' in φ at high migration rates was accompanied by a reduction in the species divergence time (τ_R) and overestimation of the ancestral population size (θ_R). This can similarly be explained by the attempt of the MSci model to accommodate the coalescent times in the middle of the time interval (τ_T, τ_R) (fig. S3).

Asymptotic results under the SC model

Under the SC model, there was initially complete isolation after species divergence but the two species came into contact at time τ_T, with ongoing gene flow ever since (fig. 1**c**). When data of an infinite number of loci, each with two sequences, are analyzed under the MSci model, the MLEs are shown in figure S4, with fitted densities of coalescent time t shown in figure S5.

The results show patterns similar to those under the IM and IIM models discussed above. The species divergence time under the MSci model $\tau_R^* = \tau_R^{(\rm sc)}$ when the migration rate M is small but drops at very high rates (with $M > 2$). The introgression time is dominated by the minimum coalescent time, so that $\tau_S^* = 0$ when $n = \infty$, and τ_S^* is much closer to 0 than to τ_R when n is finite (fig. S4). Note that in the true model migration occurs throughout the time interval $(0, \tau_T)$.

The introgression probability φ^* grows almost linearly with the migration rate M when M is small (with $M \leq \frac{1}{4}$, say), and is close to the expectation φ_0 (eq. 10) when $M < 2$ (fig. S4a,b&e). At very high migration rates ($M > 2$), φ^* is much smaller than φ_0,

and this 'bias' is accompanied by an underestimation of τ_R and overestimation of θ_R. Similarly to the IM and IIM models discussed above, this is due to the attempt of the MSci model to accommodate the coalescent times in the middle of the interval $(0, \tau_T)$ (fig. S5).

The amount of gene flow under the IM, IIM, and SC models

While the expected total amount of gene flow, measured by φ_0 (eq. 10), is the same under the IM, IIM, and SC models of figure 1**a**-**d**, the estimates under the MSci model differ, as summarized in figure 1**e**.

At low migration rates, τ_R, θ_S and θ_R in the MSci model are accurately estimated to match those in the true model (figs. 2, S2 & S4). Consider the case of infinitely long sequences with known coalescent time. Let $\tau_R^* = \tau_R^{\mathrm{m}}$, $\theta_R^* = \theta_R^{\mathrm{m}}$, and let the introgression time be $\tau_S^* = 0$ for the IM and SC model, and $\tau_S^* = \tau_T$ for the IIM model. We also match the probability density of coalescent time t for $t > \tau_R$, so that $f_{\mathrm{i}}(t) = f_{\mathrm{m}}(t)$. With those simplifying assumptions, φ^* that minimizes the KL divergence (eq. 9) can be derived as

$$\begin{aligned}
\varphi^*_{(\mathrm{IM})} &\approx \frac{\varphi_0 - \frac{w\theta_A}{2}\left(1 - \mathrm{e}^{-\frac{2}{\theta_A}\tau_R}\right)}{\left(1 - \frac{w\theta_A}{2}\right)\left(1 - \mathrm{e}^{-\frac{2}{\theta_A}\tau_R}\right)}, \\
\varphi^*_{(\mathrm{IIM})} &\approx \frac{\varphi_0 - \frac{w\theta_A}{2}\left(1 - \mathrm{e}^{-\frac{2}{\theta_A}(\tau_R - \tau_T)}\right)}{\left(1 - \frac{w\theta_A}{2}\right)\left(1 - \mathrm{e}^{-\frac{2}{\theta_A}(\tau_R - \tau_T)}\right)}, \\
\varphi^*_{(\mathrm{SC})} &\approx \frac{\varphi_0 - \frac{w\theta_A}{2-w\theta_A}\left(\mathrm{e}^{-w\tau_T} - \mathrm{e}^{-\frac{2}{\theta_A}\tau_T}\right)\mathrm{e}^{-\frac{2}{\theta_A}(\tau_R - \tau_T)}}{1 - \mathrm{e}^{-\frac{2}{\theta_A}\tau_R}},
\end{aligned} \tag{11}$$

for the IM, IIM, and SC models, respectively. At low migration rates, eq. 11 provides accurate numerical results (methods a, b, e in figs. 2,S2&S4). From eq. 11, we have

$$\varphi_0 > \varphi^*_{(\mathrm{SC})} > \varphi^*_{(\mathrm{IIM})} = \varphi^*_{(\mathrm{IM})}. \tag{12}$$

In other words, recent gene flow (as in SC) is easier to recover by the MSci model than ancient gene flow (as in IM or IIM). Note that $\varphi^*_{(\mathrm{IIM})} = \varphi^*_{(\mathrm{IM})}$ holds only when one sequence is sampled per species; as there is no coalescent over $(0, \tau_T)$, IIM is essentially the same model as IM with a time shift (fig. 1). This will not be the case when multiple sequences per species are sampled or when the sequence length is finite.

Simulation results

As our asymptotic theory was limited to a single sequence per species per locus, we used simulation to verify and augment our analytical calculations above. We simulated data under the IM, IIM or SC models of figure 1**a**-**c**, using the same parameter values as above, and analyzed them using BPP under the MSci model (fig. 1**d**). The JC mutation model (Jukes and Cantor, 1969) was assumed. In the basic setting we used $S = 4$ sequences per species per locus, each of $n = 1000$ sites, with each dataset consisting of $L = 4000$ loci. We varied the number of sequences per species (S), the number of sites per sequence (n), the number of loci (L), and the migration rate (M) to examine their effects on parameter estimation. With multiple sequences per species, all eight parameters of the MSci model (fig. 1**d**) are identifiable.

We first note a few common features in the results (fig. 3). In nearly all cases, population sizes for extant species (θ_A, θ_B) were very well estimated, with posterior means close to the true values and with very narrow highest-probability-density (HPD) credibility intervals (CIs). The exception was parameter θ_B under the IM model (note that B is the species receiving immigrants), which was less well estimated when the dataset was small and had either short sequences ($n = 250$) or few loci ($L \leq 500$), or when the migration rate was very high. The poorer estimation of θ_B appeared to be related to the underestimation of φ and τ_R; see below. The population size for the common ancestor θ_R was mostly well estimated, although overestimated at very high migration rates. Population sizes for the ancestral species (θ_S, θ_H) are harder to estimate; indeed they had larger CIs and were influenced by misspecification of the model of gene flow. As expected from asymptotic results, the age of the root of the species tree (τ_R) was very well estimated, except at very high migration rates, when τ_R was underestimated (and θ_R overestimated).

We next examine the effects of various factors: the sequence length (n), the number of sequences per species (S), the number of loci (L), and the migration rate (M). First, the number of sites (n) had a relatively small impact on MSci parameters, when other factors were fixed (at the basic setting of $S = 4, L = 4000$, and $M = 0.2$). While $n = 250$ was small and led to large CIs for parameters such as the introgression time and probability ($\tau_S = \tau_H$ and φ), the CIs were small for all parameters when $n \geq 1000$. The introgression time τ_S decreased slightly as the sequence became longer. This is consistent with the asymptotic analysis, which suggests that τ_S^* is dominated by the smallest coalescent time or sequence divergence and should converge to 0 for the IM and SC models and to τ_T for the IIM model when $n \to \infty$ (figs. 2, S2&S4). Similarly, for the IM and SC models, $\hat{\varphi}$ increased with the increase of n when M was low, as observed in the asymptotic analysis. Under the IIM model, small datasets with short sequences ($n = 250$) produced very uncertain estimates of φ and θ_H (and, to a lesser extent, τ_S and θ_S); we discuss this effect below when we examine the impact of the number of sequences (S).

Second, we varied the number of sequences per species (S). All parameters were well estimated when multiple sequences were sampled per species ($S \geq 2$ for IM and SC or $S \geq 4$ for IIM). When only one sequence per species is in the data ($S = 1$), only five parameters $(\theta_R, \theta_S, \tau_R, \tau_S, \varphi)$ are identifiable. In the asymptotic

analysis assuming infinitely many loci, we noted that θ_S was estimated with large errors, and that the poor estimation of θ_S impacted on the estimation of φ; the parameters were grossly wrong but had no sampling errors because the data size was $L = \infty$ (figs. 2c,d, S2c,d and S4c,d). Why did $\hat{\varphi}$ not converge to 0 with narrow CIs, since the MSC model with no gene flow is nearly the correct model? We interpret this result as being due to the near unidentifiability of parameters θ_S and φ in the MSci model or very weak information concerning θ_S and φ in data of two sequences per locus even with infinitely many loci. If the true model is the simple MSC with no gene flow, the MSci model with $\varphi = 0$ will provide a perfect fit, but $\varphi > 0$ together with a large θ_S, adjusted appropriately, may provide a very good fit as well. There is a ridge in the posterior surface involving φ and θ_S, leading to wide CIs for those parameters (see parameter estimates in methods c and d in fig. S2). The simulation here, assuming a finite number of loci ($L = 4000$), confirms that interpretation. Unlike in the asymptotic analyses in where there were no uncertainties, here the two parameters had strikingly large CIs, influenced both by model misspecification and by the prior (fig. 3, θ_S with $S = 1$). The large uncertainties in parameter estimates may be considered a strength of the Bayesian analysis, as they help the investigator to avoid making incorrect inferences of a large φ when gene flow is minimal. We note that even with $S = 4$ sequences per species, estimates of φ from data generated under the IIM model involved wide CIs, with τ_S being close to τ_R, and θ_S and θ_H being very imprecise as well (fig. 3). Nevertheless, this problem of semi-unidentifiability disappeared and all parameter estimates were well-behaved in large datasets when many sequences were sampled ($S \geq 2$ for IM and SC or $S \geq 4$ for IIM; fig. 3).

Third, we examined the impact of the number of loci (L). The IIM model was hard to fit in small datasets with a small number of loci ($L \leq 1000$), generating large CIs for parameters φ and θ_H. This is the same pattern as in the case of short loci ($n = 250$) or few sequences ($S \leq 2$), discussed above. In other cases the parameters were well estimated. Note that the number of loci L is the sample size in the statistical model as data at different loci are independently and identically distributed. Theory predicts that in large datasets the variance should be proportional to $1/L$ (see O'Hagan and Forster, 2004 for the case of correctly specified models and Yang and Zhu, 2018 for the case of misspecified models), and thus the CI should decrease at the rate of $L^{-\frac{1}{2}}$. This prediction held for parameters that were well estimated (fig. 3). As discussed earlier, the introgression time τ_S is dominated by the smallest coalescent time or smallest sequence divergence. Thus increasing the number of loci led to a decrease in the estimated introgression time, and the trend was in particular apparent for the IIM model (under which $\hat{\tau}_S \to \tau_T$ when $L \to \infty$ if $n = \infty$). In all cases, the estimated introgression time ($\hat{\tau}_S$) was closer to the more recent end of the time interval for gene flow than to the midpoint (i.e., $\hat{\tau}_S < \frac{\tau_R}{2}$ for IM, $\hat{\tau}_S < \frac{\tau_R+\tau_T}{2}$ for IIM, and $\hat{\tau}_S < \frac{\tau_T}{2}$ for SC; see fig. 1**a**-**c**).

Finally, we evaluated the impact of the migration rate (M) (fig. 3). At low migration rates (M), there is a near linear relationship between the introgression probability φ and M. In general, the amount of gene flow estimated under the MSci model is less than the true amount expected under the migration model (φ_0 of eq. 10) but the two were close at low migration rates (with $M < 0.1$, say). At very high migration rates (with $M > 1.0$, say), divergence time τ_R was increasingly underestimated and the population size θ_R was overestimated. These are the same patterns as observed in the asymptotic analysis of infinite data ($L = \infty$), and are due to the attempt of the MSci model to accommodate intermediate coalescent times generated in the true migration model, as discussed earlier (see figs. 2, S2, S4). At low migration rates (with $M < 0.03$, say), the IIM model produced very uncertain estimates of φ, with τ_S, θ_S, and θ_H affected as well. This is the same pattern as observed for small datasets with short loci ($n \leq 250$), few sequences ($S \leq 2$), or few loci ($L \leq 1000$), discussed above.

In summary, our asymptotic analysis and the computer simulations suggest the following correspondence between the IM and MSci models. When gene flow occurs continuously over an extended time period (τ_T, τ_R) after divergence of two species and we fit the introgression (MSci) model, the estimated introgression time tends to be closer to the more recent end of the time period of gene flow, because the introgression time in the MSci model is dominated by the most recent coalescent time or the minimum sequence divergence between species. Indeed if the true coalescent time is known and used as data, the introgression time will converge to the time when gene flow stopped, as discussed above in the asymptotic analysis. At low migration rates ($M < \frac{1}{4}$, say), the species divergence time is correctly estimated by the MSci model, and the introgression probability φ is lower than but close to the expected proportion of migrants ($\varphi^* < \varphi_0$). The estimate is particularly close under the SC model (fig. S4). At high migration rates, the estimated introgression probability φ^* may be much less than φ_0. This is because the species divergence time τ_R is underestimated by the MSci model to account for intermediate coalescent times generated under the IM model, leading to underestimation of the introgression probability φ and overestimation of the ancestral population size θ_R.

Introgression events assigned to wrong branches

We conducted simulations to examine the bias in parameter estimates when the introgression event is

assigned on either the parental or daughter branch of the lineage genuinely involved in introgression. The data were simulated under model trees A or B and analyzed under models A or B (fig. 4**a**,**b**).

In the A-A and B-B settings (fig. 4**e**), the correct MSci model was assumed, and the performance of the method serves as a reference for comparison. Most parameters, including the species divergence times (τ_R, τ_S, τ_T, and $\tau_X = \tau_Y$) and population sizes for extant species ($\theta_A, \theta_B, \theta_C, \theta_D$), were well estimated. For well-estimated parameters, the CI width halves as the number of loci (L) quadruples, as predicted by theory. Population sizes for ancestral species (θ_R, θ_S, θ_T, θ_X, and θ_Y) were less well estimated, although performance improved with sample size: with $L =$ 4000 loci, these parameters were well estimated. Introgression probability (φ) was well estimated, but thousands of loci were necessary to obtain precise estimates with narrow CIs under the standard settings used here (4 sequences per species per locus and 500 sites per sequence).

In the A-B setting (fig. 4**e**), data were simulated under model A with $A \to B$ introgression (fig. 4**a**) but analyzed under model B with introgression incorrectly assigned to the parental branch ST. Species divergence times (τ_R, τ_S, τ_T, and $\tau_X = \tau_Y$) and population sizes for extant species (θ_A–θ_D) were all well estimated, similar to the B-B setting in which the simulation model matched the analysis model. Population sizes for ancestral species were hard to estimate, and performance was similar to that under the B-B setting. We expect τ_T in model B to be mostly determined by the smallest sequence divergence between species B and C, which should be close to $\tau_T = 2\theta_0 = 0.004$. As introgression events in the data had the effect of pushing τ_T down, τ_T had a negative bias but the bias was very small and τ_T was well estimated. The introgression time $\tau_X > \tau_T$ in model B, and as the true introgression time τ_X was smaller than τ_T, τ_X was stuck at τ_T (fig. 5**a**). There was virtually no information for θ_T as the population was estimated to have near-zero time duration with no chance for coalescent events to occur in the population. The introgression probability was seriously underestimated, converging to $\varphi^*_{\text{A-B}} \approx 0.12$ when the number of loci L increases (table 1) whereas the true value was 0.2. This smaller estimate of introgression probability is explained by the distribution of coalescent times between species in the true and fitting models (fig. S6, true model A). Under the true model A, sequences from A and B are more similar than those between A and C due to the $A \to B$ introgression, with an excess of small coalescence time t_{ab}. Under the analysis model B, t_{ab} and t_{ac} have the same distribution. Thus the true model predicts an excess of small t_{ab} in the data whereas the fitting model predicts an excess of small t_{ac}, and having a smaller φ in the fitting model helps to reduce the discrepancy.

In the B-A setting (fig. 4**e**), the simulation model (MSci model B of fig. 4**b**) assumed introgression involving the ancestral branch ST but the analysis model (model A) assigned introgression to the daughter branch TB. Again posterior means and CIs for most parameters, including species divergence times (τ_R and τ_T) and population sizes (θ_A–θ_D, and θ_R), were similar to those in the A-A setting where there was model match. Note that τ_T in the analysis model A should be mostly determined by the smallest sequence divergence between B and C, and given that this was $\tau_T = 2\theta_0 = 0.002$ in the simulation model A, τ_T was well estimated, unaffected by mis-assigned introgression event. Although the true introgression time τ_X was 0.003, it was forced to be less than τ_T by the analysis model A. As the number of loci increases, τ_X became stuck at τ_T (fig. 5**b**). However, τ_S was seriously underestimated. This may be explained as follows. In the analysis model A, τ_S was mostly determined by the shortest sequence distance between A and C. According to the simulation model B, this should be close to $\tau^{(B)}_X = 1.5\theta_0 = 0.003$, due to introgression events. Here we use the superscript to indicate that the parameter is for model B (fig. 4**b**). With mutational fluctuations in the sequences, one can expect the τ_S estimate in the B-A setting to lie between $(\tau^{(B)}_X, \tau^{(B)}_S) = (1.5\theta_0, 3\theta_0)$, but closer to $\tau^{(B)}_X$ in large datasets with many sites and/or many loci. Population size parameters θ_S and θ_Y were affected by the mis-assigned introgression events as well, as those populations are close to the introgression branches. In particular, θ_Y was very imprecise as branch TY was very short, and θ_S was overestimated because τ_S was seriously underestimated and the two parameters are negatively correlated. Finally the introgression probability (φ) was underestimated, apparently converging to $\varphi^*_{\text{B-A}} \approx 0.02$ when the number of loci increases (table 1) whereas the true value was 0.2. This greatly reduced introgression probability appeared to reflect the very poor fit of the misspecified model A to data generated under model B, apparent due to large differences between the true and fitting distributions of coalescent times (t_{ab}, t_{ac}, t_{bc}; fig. S6, second row). As τ_X and τ_S are seriously underestimation by model A, an excess of small coalescent times (t_{ab}, t_{ac}) is expected in the fitting model A but does not appear in the data, so that having a smaller φ improves the fit.

In summary, assigning introgression events to a wrong parental or daughter branch led to biased estimates of introgression times (causing the introgression events to collapse onto speciation events) and to seriously underestimated introgression probabilities.

Continuous migration versus episodic introgression

In this set of simulations, we generated data under the IM models C and D of figure 4**c**&**d** and analyzed them under the MSci models A and B, with the mode of gene flow misspecified and with gene flow assigned to either the correct branch or a wrong branch on the species tree.

In the C-A and D-B settings (fig. 4**e**), gene flow occurred continuously but the data were analyzed under the introgression model assuming gene flow at a particular time point. The mode of gene flow was misspecified, but the lineages involved were correctly identified. In the C-A setting, gene flow was between nonsister species, while in the D-B setting it was between sister species. Speciation times (τ_R, τ_S, τ_T) and population sizes (θ) were well estimated. Indeed the results for those parameters under the C-A setting were indistinguishable from those under the A-A setting, where the MSci model was used to both generate and analyze data. Similarly the results for speciation times and population sizes were extremely similar for settings D-B and B-B. Those results were consistent with earlier simulation results based on two species (fig. 3), which showed that at low migration rates, species divergence times and population sizes were well estimated under the MSci model when the data were generated under the IM model (see also Tiley *et al.*, 2022).

In the C-A setting, the introgression time τ_X appeared to converge, when the number of loci L increases, to 0.0011, which is much more recent than the average time of gene flow ($\tau_T/2 = 0.002$), and the introgression probability φ appeared to converge to $\varphi^*_{\text{C-A}} = 0.12$ (table 1), smaller than the proportion of total migrants given by eq. 10: $\varphi_0 = 1 - \mathrm{e}^{-4M\tau_T^{(C)}/\theta_B^{(C)}} = 0.148$. As discussed earlier in the case of two species, the limiting value for τ_X was nonzero, as the sequence length is finite, and the MLE $\hat{\varphi}_{\text{C-A}}$ slightly underestimates the true amount of gene flow. In the D-B setting, the introgression time τ_X appeared to converge to 0.0027, which is larger than $\tau_T = 0.002$ but much smaller than the average time of gene flow, $\frac{1}{2}(\tau_S + \tau_T) = 0.004$, and the introgression probability φ appeared to converge to $\varphi^*_{\text{D-B}} = 0.08$ (table 1), much smaller than the true proportion under the IM model, $\varphi_0 = 0.148$ (eq. 10). In both the C-A and D-B settings, the estimated introgression time was within the time interval of gene flow, but closer to the time point when gene flow stopped, while the introgression probability underestimated the amount of gene flow that actually occurred, with $\varphi^*_{\text{C-A}} < \varphi_0$ and $\varphi^*_{\text{D-B}} < \varphi_0$. Moreover, we have $\varphi^*_{\text{D-B}} < \varphi^*_{\text{C-A}}$. These patterns are consistent with our analysis of the two-species case at low migration rates (eq. 12, fig. 3), which predicts that gene flow after a period of isolation (the SC model) is easier to recover under the MSci model than gene flow that starts at speciation but stops some time afterwards (the IIM model).

In the C-B and D-A settings (fig. 4**e**), the mode of gene flow was misspecified and furthermore gene flow was assigned onto the wrong branch of the species tree. In the C-B setting, species divergence times τ_R and τ_S were well estimated, just as in the B-B setting. Divergence time τ_T was underestimated slightly, due to gene flow assigned to the wrong branch, as observed in the A-B setting. Population sizes for extant species (θ_A–θ_D) were all well estimated, as in the B-B setting. Ancestral population sizes θ_R and θ_S were as well estimated as in the B-B setting, and so was θ_X. Ancestral population sizes θ_T and θ_Y were affected by gene flow, similar to the A-B setting. Model B forces $\tau_X > \tau_T$. Thus we expect estimates of τ_X and τ_T to get stuck together, with both to be smaller than $\tau_T^{(C)} = 2\theta_0 = 0.004$; as the number of loci L increases, τ_X appears to converge to 0.0029, and φ to $\varphi^*_{\text{C-B}} = 0.10$ (table 1).

In the D-A setting, species divergence times τ_R and τ_T were well estimated, as in the A-A setting. Divergence time τ_S was underestimated, due to gene flow assigned to the wrong branch, similarly to the B-A setting. Population sizes for extant species (θ_A–θ_D) were all well estimated, as in the A-A setting. Ancestral population sizes θ_R and θ_X were as well estimated as in the A-A setting, and θ_T had a slight positive bias. Ancestral population sizes θ_S and θ_Y were affected by the gene flow, similar to the B-A setting. Introgression time and probability (τ_X and φ) did not exist in the simulation model D. Model A forces $\tau_X < \tau_T$, so we expect τ_X to be close to $\tau_T^{(D)} \approx \theta_0 = 0.002$; when the number of loci L increases, τ_X appeared to converge to 0.00186, and $\varphi_{\text{D-A}}$ to $\varphi^*_{\text{D-A}} = 0.02$ (table 1). Note that $\varphi_0 > \hat{\varphi}_{\text{C-B}} > \hat{\varphi}_{\text{D-A}}$ with $\hat{\varphi}_{\text{C-B}} < \hat{\varphi}_{\text{C-A}}$ and $\hat{\varphi}_{\text{D-A}} < \hat{\varphi}_{\text{D-B}}$. Those results are consistent with our early results for fitting the MSci model to data generated under the migration model in the two-species case (eq. 12, fig. 3), and with the results for the A-B and B-A settings that assignment of gene flow to a wrong branch reduces the estimate of φ.

In summary, the estimated introgression probabilities, at 0.12, 0.08, 0.10, and 0.02 for the C-A, D-B, C-B, and D-A settings, respectively, even though the total amount of gene flow was the same in models C and D (table 1), suggest the following general patterns. First, the introgression (MSci) model underestimates the total amount of gene flow if gene flow occurs continuously in every generation (i.e., $\hat{\varphi}_{\text{C-A}} < \varphi_0, \hat{\varphi}_{\text{D-B}} < \varphi_0$), as discussed in our analysis of the two-species case. Second, assigning gene-flow events to wrong lineages led to more serious underestimation of the strength of gene flow than the misspecification of the mode of gene flow alone (i.e., $\hat{\varphi}_{\text{C-B}} < \hat{\varphi}_{\text{C-A}}, \hat{\varphi}_{\text{D-A}} < \hat{\varphi}_{\text{D-B}}$). Third, recent gene flow in the data is more easily recovered (i.e., $\hat{\varphi}_{\text{C-A}} > \hat{\varphi}_{\text{D-B}}, \hat{\varphi}_{\text{C-B}} > \hat{\varphi}_{\text{D-A}}$).

Isolation with initial migration (IIM) model

We used the IIM model of figure 6**a** to simulate data and analyzed them under the MSci model of figure 6**b**. Species divergence times (τ_R and τ_S) and population sizes (θ_A, θ_B, θ_C, θ_R, θ_S, and even θ_X and θ_Y) were well estimated. We expect the introgression time τ_X to converge to $\tau_T = \theta_0 = 0.002$ if the sequence length is infinite and to a higher limit for finite sequence length. In our simulation $\tau_X \approx 0.00283$ at $L = 4000$ (table 1). Introgression probability φ converged to a nonzero limit, ~ 0.08 (table 1), compared with $\varphi_0 = 0.148$ by eq. 10.

The case of figure 6 is very similar to the two-species case of figure 1 except that the species tree is larger with more species, and serves to highlight the fact that the impact of the misspecification of the model of gene flow is local. The case is also very similar to the D-B setting of figure 4, with the only difference that here the hybridizing species T had only one descendent species sampled in the data whereas in figure 4 (D-B) it had two descendent species sampled. In the Bayesian analysis, this difference affects only the information content in the data. Thus estimates of parameters such as the introgression probability and introgression time were similar to those in the D-B setting of figure 4 but with wider CIs (table 1).

Ghost species

We considered two scenarios in which a species that contributed migrants to extant species has gone extinct or is otherwise unsampled in the data. Note that existence of extinct or unsampled species that *received* genetic materials from ancestors of extant species in the sample is not relevant to the analysis of the sampled data and does not constitute a model misspecification. In the first scenario, model A′ of figure 7**a**′ is used to simulate data, which assumes that species XUV contributed migrants to species B but is not included in the sample. Note that this model is equivalent to model A of figure 7**a**. When we fit model B (fig. 7**b**), the only incorrect assumption is the constraint that $\tau_X = \tau_Y$. This is a minor misspecification. Indeed all parameters shared between the simulation model and the analysis model were well estimated (fig. 7**c**). The estimates of introgression time, $\tau_X = \tau_Y \approx 0.0028$ (table 1), were close to the average of the two parameters in the true model (0.0025). Introgression probability $\varphi \approx 0.21$ (table 1) was also close to the true value (0.2). The existence of the ghost species (XUV) had very little effect on the inference.

In the second scenario (fig. 8**a**), the true model assumes continuous migration involving intermediate ancestral species that have gone extinct, and the MSci model (fig. 8**b**) was fitted to data sampled from extant species. Divergence times τ_R and τ_T were very well estimated, as were the population sizes shared between the simulation and analysis models (θ_A, θ_B, θ_C, θ_R). We expect $\hat{\tau}_T$ in model B to be dominated by the minimum coalescent time t_{ab} between sequences from A and B, and this is given by τ_T in model A. Gene flow from branches RC to SU to TB during the time period τ_T–τ_U is interpreted as introgression in the MSci model. The effective rate for this migration may be close to $M_{CU}M_{UB} = 0.04$, giving the expected amount of introgression as $\varphi_0 = 1 - e^{-4\times M_{CU}M_{UB}\times(\tau_T-\tau_U)/\theta_B} = 0.031$. The estimate was $\varphi_X \approx 0.02$ (fig. 8**c**, table 1). Introgression time τ_X should be between $(\tau_U, \tau_T) = (\theta_0, 2\theta_0) = (0.002,\ 0.004)$ and the estimate was ≈ 0.0030 (fig. 8**c**, table 1). Note that both θ_T and θ_Y were overestimated (fig. 8**c**). Branch T of figure 8**b** corresponds to branches RS and ST of figure 8**a**, which have population size $\theta_0 = 0.002$. Branch Y corresponds to a segment of branch TB over the time interval τ_U–τ_T, with population size $\theta_1 = 0.01$. Overestimation of θ_Y (and θ_T) may be because there is a shortage of t_{bb} over the time interval τ_U–τ_S in the data because of the gene flow, and the fitting MSci model, with the amount of gene flow underestimated ($\varphi < \varphi_0$), used large $\hat{\theta}_Y$ and $\hat{\theta}_T$ to compensate.

Discussion

Previous studies of introgression versus migration models

Previously Wen and Nakhleh (2018) conducted simulations to examine the inference of gene flow and estimation of introgression probability (γ, equivalent to φ here) under the introgression model when the data were generated under the migration model. The analyses using PHYLONET (Wen and Nakhleh, 2018) searched in the space of models that included the MSC model with no gene flow and the MSci model with introgression but not migration models, so that the analysis model was misspecified.

The migration model of figure 17d in Wen and Nakhleh (2018), with results reported in their table 2 (mt2 = 1), is the same as the IIM model of figure 6**a**, with unidirectional migration involving non-sister species. The parameter values used (in the notation of fig. 6**a**) are: $\theta = 0.02$ for all populations, $\tau_R = 0.025, \tau_S = 0.015$, with gene flow ceasing at time $\tau_T = 0.01$, and $M = Nm = 0.1$. According to our theory (eq. 10), the introgression probability (γ or φ) should be smaller than $\varphi_0 = 1 - e^{-(4M/\theta)\Delta\tau} = 1 - e^{-4\times 0.1/0.02\times 0.005} = 0.095$, compared with the average estimate 0.18 ± 0.05 reported by Wen and Nakhleh (2018). This result is thus in conflict with our theory. The model of table 2 (mt2 = 0) in Wen and Nakhleh (2018) is a simple IM model, equivalent to the C-A setting in figure 4. The expected introgression probability should be smaller than $\varphi_0 = 1 - e^{-(4M/\theta)\tau_S} = 0.259$, compared with the reported average estimate of 0.17 ± 0.04. In this case the

estimate seems to be approximately consistent with our theory.

The migration model of figure 17f in Wen and Nakhleh (2018) specifies unidirectional migration between two sister species. The parameter values used are $\theta = 0.02$ for all populations, $\tau = 0.015$, with $M = Nm = 0.1$. Results reported in their table 3 (mt = 1) are for the IIM model, with gene flow ceasing at time 0.01 so that the period of gene flow is $\Delta\tau = 0.005$. According to our theory (eq. 11), the introgression probability should be smaller than but very close to $\varphi^*_{(\text{IIM})} = 0.0523$. The authors did not detect gene flow under the introgression model. It is possible that the rate was so low and the datasets were so small that the MSC model with no introgression was favored over the MSci model. Note that when the data are generated under the migration model, both the MSci and MSC models are wrong. However, if the limit of the MLE $\varphi^* > 0$ in the MSci model, the MSci model will be less wrong than the MSC model, measured by the KL divergence, and is expected to win over the MSC model in large datasets Yang and Zhu (2018). Results reported in table 3 (mt = 0) of Wen and Nakhleh (2018) are for the IM model, with the average estimate of 0.11 ± 0.06. This seems consistent with our theory (eq. 11), which predicts $\varphi^*_{(\text{IM})} = 0.167$.

We note that Wen and Nakhleh (2018) used the `-s` switch for SEQ-GEN to specify the value of $\theta/2$ whereas θ should be used instead. Also the authors simulated 20 replicate datasets, each of only 200 loci, so that the datasets are likely too small to produce reliable estimates of the introgression probability. Those factors may partly explain the discrepancy of the simulation results of Wen and Nakhleh (2018) from our theoretical predictions. Note that when data are generated under the migration model and analyzed under the introgression model, the estimated introgression probability should depend on not only the migration rate per generation but also the time period during which gene flow occurs ($\Delta\tau$).

The mode of gene flow and the utility of misspecified introgression models

The asymptotic analysis has been surprisingly useful, even though based on only two species (with gene flow from A to B), with one sequence sampled per species per locus. It generated a number of insights that were confirmed and extended in our simulations, such as (a) the limiting values of introgression time when either the true coalescent times (t) or the differences between sequences (x) are given as data, (b) underestimation of the species divergence time (τ_R) by the MSci model at high migration rates, (c) underestimation of the proportion of migrants by the MSci model, and so on. Extending the theory to three sequences (e.g., a, b_1, b_2, with two sequences from the hybridizing species B; fig. 1) may remove some of the problems we encountered, such as the semi-unidentifiability of θ_S and φ at very low migration rates in figures 2c&d. The theory will be much more complex as it would have to average over the three different gene trees and the two coalescent times, rather than just one coalescent time t in the case of two sequences.

Nevertheless, our asymptotic theory and simulations (fig. 3) constitute a detailed analysis of gene flow in the two-species case. We note that the case of two species may be one of the most challenging cases for inferring gene flow. For example, approximate methods such as the D-statistic cannot detect gene flow with samples from two species at all. Furthermore, when one sequence is sampled per species per locus, the direction of gene flow is unidentifiable, as the two models assuming $A \to B$ and $B \to A$ introgressions predict the same probability distribution of the coalescent time between the two species, $f(t_{ab})$ (Yang and Flouri, 2022, fig. 10). Given the MSci model with $A \to B$ introgression (fig. 1**d**), the introgression probability φ (as well as four other parameters in the model: $\tau_R, \tau_S, \theta_R, \theta_S$) are mathematically identifiable with data of one sequence per species, but the correlation between φ and θ_S is so strong that those parameters are barely identifiable (e.g., methods c and d in fig. 2). Even with multiple sequences per species per locus, there may be a serious lack of information in the data if the dataset is small, with short loci, few sequences per species, or few loci (fig. 3, with $S \leq 4$, $L \leq 1000$, or $M \leq 0.1$ for IIM). Our results highlight the importance of sampling multiple sequences per species (in particular, from species that received immigrants) in real data analysis, besides using large datasets with hundreds or thousands of loci. Even if the model is identifiable with one sample per species, use of multiple sequences provides a major boost in information content. Many approximate methods for detecting gene flow are designed to use only one sample per species, and it has been claimed, incorrectly, that "adding more samples provides little new information with respect to introgression" (Hibbins and Hahn, 2022).

The expected lack of information to estimate parameters θ_S and φ jointly in the MSci model (fig. 1**d**) appears to apply to a recent analysis of genomic data from the *erato* group of *Heliconius* butterflies (Thawornwattana *et al.*, 2022). The estimated *H. sara* $\to$ *H. demeter* introgression probability was high with wide CIs for some chromosomal regions with a small number of loci (e.g., chromosome 21 with 4350 noncoding and 3628 coding loci, and an inversion on chromosome 15 with 149 noncoding and 167 coding loci), with the introgression time close to the species divergence time, whereas for the other large chromosomes, the estimated introgression probability was nearly zero ($\varphi < 0.01$). We believe that the true value in this case was $\varphi \approx 0$, but that the limited data from small chromosomal segments led to spurious and

poorly supported signals of introgression, as observed in our simulation (fig. 3).

While the large CIs should help one to avoid making incorrect inferences (of a high φ when the true rate of gene flow is nearly 0), a sensible approach to the problem may be to constrain the population size to be the same before and after an introgression event in the MSci model (that is, with $\theta_S = \theta_A$ and $\theta_H = \theta_B$ in fig. 1**d**). With such a constraint, implausibly large estimates for ancestral population sizes θ_S or θ_H will not occur, because population sizes θ_A and θ_B will be well estimated from data of multiple samples from the extant species, which may in turn lead to more stable estimation of the introgression probability. Sensitivity of estimates of φ to violations of the assumption about the population sizes may then be assessed.

Assigning gene flow to parental or daughter branches caused underestimation of the introgression probability, while the estimated introgression time tended to coincide with species divergence. This feature may be used to diagnose the mis-assignment in real data analysis (Ji *et al.*, 2022). A number of authors have discussed the impact of ghost species on detection of between-species gene flow (Beerli, 2004; Ottenburghs, 2020). Tricou *et al.* (2022) used simulation to demonstrate that *D*-statistics can be misled to detect false signals of introgression when it involved an unsampled ghost species. In our simulation, the impact of ghost species on Bayesian estimation of introgression rate and time was minor provided we consider the rate of gene flow in the migration and introgression models to reflect both indirect gene flow via intermediate species and direct gene flow.

In our simulations, misspecification of the mode of gene flow (continuous migration versus episodic hybridization/introgression) has relatively small and localized effects on estimation of species divergence times and population sizes around the lineages involved in gene flow, while species divergence times, population sizes for extant species and for ancestral species not involved in gene flow are largely unaffected. Even if gene flow occurs continuously over time (so that the migration model is a more realistic model), the MSci model is effective in extracting historical information about species divergence times and population sizes. Note that on the evolutionary time scale, a few hundred or thousand generations may in effect count as a fixed time point, in which case the MSci model may provide an adequate approximation.

Testing for gene flow

In this study, we fixed the model of introgression in our analyses, with all introgression events pre-identified, and then examined the effects of model misspecification. One may ask what happens if different introgression models are compared using genomic data. We note that if the true model is one of introgression, and if we infer introgression events in the MCMC algorithm, correct placement of introgression events onto the branches of the species tree will constitute the true model. Since Bayesian model selection is known to be consistent, and the true model is included in the set of models under comparison, the true model will dominate with its posterior converging to 1 when the data size (the number of loci) approaches infinity (Dawid, 2011; Yang and Zhu, 2018).

Several approaches may be taken to compare different introgression models. Both *BEAST and PHYLONET have implemented cross-model MCMC algorithms (Green, 1995), which insert and delete introgression events on the species tree, allowing the chain to move between models. The algorithms generate estimates of posterior probability for the different introgression models. Those algorithms are computationally expensive and currently the two programs can handle only small datasets. We thus did not attempt to apply them to our simulated datasets. In the BPP program, one may use the Bayes factor to compare two introgression models, using thermodynamic integration (Gelman and Meng, 1998; Lartillot and Philippe, 2006) combined with Gaussian quadrature to calculate the marginal likelihoods (Rannala and Yang, 2017). The Bayes factor for comparing two nested models (e.g., one with introgression and another without) may also be calculated through the Savage-Dickey density ratio (Dickey, 1971), which uses only a within-model MCMC run under the more general model. This has considerable computational advantages over reversible jump (Green, 1995) but works for nested models only. This approach has recently been applied in comparing and formulating introgression models in an analysis of genomic data from the *Tamias quadrivittatus* group of North American chipmunks (Ji *et al.*, 2022). Calculation of marginal likelihood or Bayes factors may be feasible if we have a small number of well-specified candidate models to evaluate but may be too tedious when there are many candidate models.

Approximate methods have also been developed to infer introgression events using summaries of the multi-locus sequence data. For example, estimated gene tree topologies may be treated as data to compare introgression models in an MCMC algorithm, as in PHYLONET/GT (Wen *et al.*, 2016). Some methods are designed to detect gene flow in a small tree with three or four species, including summary methods based on genome-wide site-pattern counts (such as *D* and HYDE discussed earlier) or on estimated gene trees (e.g., SNAQ, Solis-Lemus and Ane, 2016; Solis-Lemus *et al.*, 2017) and maximum likelihood applied to multilocus sequence alignments (e.g., 3S, Zhu and Yang, 2012; Dalquen *et al.*, 2017). The results from analyses of such species subsets then need be

combined to formulate an introgression model on the large tree for all species, which is a very challenging process (Edelman *et al.*, 2019; Thawornwattana *et al.*, 2022; Ji *et al.*, 2022).

In summary, searching in the space of species phylogenies with introgression events (or so-called phylogenetic network models) is currently a very challenging problem. There is a dire need for efficient MCMC algorithms for Bayesian inference under the MSC model with gene flow and for improvements in statistical performance of approximate methods. Both are fast-developing active research areas. We look forward to breakthroughs in the next few years.

It will also be very interesting to use the same genomic data to compare the migration (IM) and introgression (MSci) models. As discussed in the Introduction, we expect the rate of gene flow to vary over time (and across the genome), so that both kinds of models are simplistic and unrealistic. Even so it will be very useful to use the same data to evaluate which model provides a better approximation to reality. The IM and MSci models often predict very different distributions of gene trees and coalescent times (e.g., figs. S1, S3, S5; see also Jiao and Yang, 2021). This may have two implications. First, genomic data may be informative to distinguish the two kinds of models. Second, it may be too challenging to design efficient cross-model MCMC algorithms to move between them. The gene trees, which are latent variables in models of gene flow, place stringent constraints on the model. If the gene trees at all the loci are fixed when the algorithm moves from an IM model to an MSci model (or vice versa), they may constrain the model so much that changes are in effect impossible and the chain will get stuck. Consider an MCMC move from an IM model assuming $A \rightarrow B$ migration to the corresponding MSci model with $A \rightarrow B$ introgression, with all gene trees fixed. In the current IM model, the youngest sequence divergence time (or coalescent time) between the two species may be very close to zero, which means that the introgression time in the newly proposed MSci model will have to be close to zero as well. Similarly coordinated changes to gene trees during the cross-model move, as achieved in the moves that change the species divergence time (Rannala and Yang, 2003) or the species tree (Yang and Rannala, 2014; Rannala and Yang, 2017) under the MSC without gene flow, appear too difficult. Thus stochastic search in the combined space of both IM and MSci models may be infeasible. Nevertheless, one can use Bayes factors to compare the IM and MSci models.

Assumptions underlying MSC models with gene flow

Here we briefly discuss some other assumptions made in the analysis under the MSC model with gene flow, regarding recombination and selection. We have made the standard assumption that there is no recombination among sites at the same locus, and free recombination between loci. The loci here do not necessarily mean coding genes, and may represent short genomic segments generated using anchored sequencing technologies (RADseq, ultra-conserved elements, etc.), or loosely linked short genomic segments sampled from the genome (Beerli and Felsenstein, 2001; Burgess and Yang, 2008; Lohse *et al.*, 2011; Dalquen *et al.*, 2017; Hey *et al.*, 2018). A recent simulation study examined the effects of intralocus recombination on various analyses using the Bayesian program BPP, including estimation of introgression times and probabilities, and found that recombination at rates comparable to the human rate and with the sequence length at 500 bps per locus have negligible effects (Zhu *et al.*, 2022). However, excessive amounts of recombination (at rates 10 times the human rate) may cause biases in the estimated introgression probability, especially if species divergences times are comparable to coalescent times and if multiple sequences are sampled from the same species. For MSC-based analysis of genomic data from species with very high recombination rates, we suggest that shorter loci or genomic segments be used or the impact of intralocus recombination be assessed by simulation.

Natural selection is known to affect the distribution of coalescent times. Most influential will be species-specific selection or selection involved in divergent adaptation between the species. Such loci may be under very strong selection as they may contribute to species distinctness (Martin *et al.*, 2013). While identifying such gene loci may greatly enrich our understanding of the speciation process, loci and nucleotide sites under such species-specific selection are probably rare on the genome scale. In contrast, purifying selection removing deleterious mutations, such as deleterious nonsynonymous mutations in protein-coding genes, may not pose a serious concern. If the protein performs the same function in all species involved, the main effect of purifying selection will be a reduction of the neutral mutation rate. In a few cases where the noncoding and coding parts of the genome were analyzed as separate datasets, highly consistent results were obtained, with the estimated species divergence times and population sizes being nearly perfectly proportional between the two types of data (Shi and Yang, 2018; Thawornwattana *et al.*, 2018, 2022), and with the same constant of proportionality across different chromosomes (Thawornwattana *et al.*, 2022, fig. S5). Species divergence times in particular showed great consistency with $r^2 > 0.99$. We recommend analyzing non-coding loci separately from coding loci.

It is likely that natural selection (in particular, purifying selection removing deleterious mutations) may have a smaller impact on gene-tree topologies than on coalescent times between sequences. Similarly,

genome-wide averages may be very stable while genealogical variations across the genome may be caused by local variations in recombination rate, the mode and strength of selection, biased gene conversion, etc., besides coalescent and gene flow. Thus approximate methods using gene tree topologies or genome-wide averages may be more robust than likelihood-based fully parametric methods that use information from both gene tree topologies and branch lengths and that use information from both genome-wide averages and genealogical variation across the genome. It may be interesting to use computer simulation to examine the robustness of different inference methods to various model violations including the impact of selection. Applications to various genomic datasets will also provide useful empirical tests.

Conclusions

Based on our asymptotic analyses and computer simulations, we make the following observations. First, population sizes (θ) for extant species are well-estimated when two or more sequences are sampled per species per locus. For one species, θ is simply the average pairwise sequence distance. In almost all simulation settings of this paper, including those in which the model of gene flow is misspecified, θ for extant species are well estimated. We expect this to be generally true except where the species divergence is so recent that the extant species is very young. Second, estimates of species divergence or introgression times are dominated by the minimum coalescent time or the smallest sequence divergence between species (across loci and across all sequence pairs per locus). This is because in the MSC model with and without gene flow, sequence divergence time (t_{ab}) between species A and B must be greater than the species divergence time ($t_{ab} > \tau_{AB}$). Third, given species divergence times (τ), ancestral population sizes and/or introgression probabilities may be adjusted to achieve the correct amount of sequence divergence (between sequences from the same species or from different species). In particular, if a species divergence time is underestimated, the population size (θ) for the ancestral species is often overestimated to compensate. We have found that all these patterns are useful for interpreting or predicting parameter estimates under the MSci model when the model is misspecified.

The MSci models studied here are relatively new and have not been tested on many genomic datasets. Our analyses thus provide an empirical assessment of the utility of the models in various situations of model misspecification. Overall, we found that the impact of misspecification is local, so that useful inference is still possible despite the misspecification. For example, when gene flow is continuous, the MSci model which assumes that gene flow occurs at one time point gives reliable estimation of species divergence times, unlike the MSC model ignoring gene flow, which is known to lead to seriously underestimation of species divergence times (Ogilvie *et al.*, 2016; Tiley *et al.*, 2022). The estimated introgression probability may also serve as a useful guide even though this reflects both the migration rate per generation and the time duration of the period of gene flow. When gene flow is mis-assigned onto the mother or daughter branches of the genuine introgression lineage, the introgression time is pushed onto the species divergence events, but estimates of older divergence times are largely unaffected. Overall, our results suggest that the simple introgression models may be used to estimate species divergence times and quantify the intensity of gene flow even if the model is an imperfect match to reality.

Materials and Methods

Two-species simulation to establish a correspondence between the migration and the introgression models

We performed a theoretical analysis to establish the relationships between parameter estimates when the true model is the continuous migration model (IM, IIM, and SC; fig. 1**a**-**c**) but the analysis model is the episodic introgression (MSci) model (fig. 1**d**). Our theory assumes an infinite number of loci ($L = \infty$), a finite number of sites per sequence, and only one sequence from each species.

We thus conducted computer simulations to augment the theoretical analysis. Data of multi-locus sequence alignments were simulated under the IM, IIM and SC models of figure 1**a**-**c**, and then analyzed under the MSci model (fig. 1**d**). Population sizes on the species tree (fig. 1) were $\theta_0 = 0.002$ for the thin branches and $\theta_1 = 0.01$ for the thick branches. Migration occurred from species A to B after their divergence $\tau_R = \theta_0$ in the IM model, between $\tau_R = 2\theta_0$ and $\tau_T = \theta_0$ in the IIM model, and between $\tau_T = \theta_0$ and the present in the SC model. In the standard model, the migration rate was $M = 0.2$ individuals per generation. Each dataset consisted of $L = 4000$ loci, with $S = 4$ sequences per species, and $n = 1000$ sites per sequence. To aid the theoretical analysis, we conducted four sets of simulation to examine the impact of the number of sites per sequence (n), the number of sequences per species (S), the number of loci (L), and the migration rate (M). The values used were $n = 250, 1000, 4000, 16000, 64000$; $S = 1, 2, 4, 8, 16$; $L = 250, 500, 1000, 2000, 4000, 8000$; and $M = 0.01, 0.02, 0.03, 0.04, 0.05, 0.07, 0.1, 0.2, 0.3, 0.4, 0.5, 0.7, 1.0, 1.5, 2.0$. With three models (IM, IIM, and SC), four factors (n, S, L, M), and 30 replicates, a total of $3 \times (5+5+6+13) \times 30 = 2790$ datasets were simulated. Replicate datasets were simulated using BPP version 4.4.1 (Flouri *et al.*, 2018, 2020), by generating

the genealogical trees with coalescent times for each locus and then "evolving" sequences along branches of the gene tree under the JC mutation model (Jukes and Cantor, 1969). Sequences at the tips of the gene tree constituted the data at the locus.

Each dataset was analyzed using BPP under the MSci model (fig. 1**d**) to estimate the parameters. This is the so-called A00 analysis, with the model fixed (Yang, 2015). The Bayesian implementation of the MSci model in BPP accommodates gene-tree reconstruction uncertainties while making use of information in both gene tree topologies and branch lengths, and allows the estimation of the direction, the timing, and strength of introgression events (Jiao *et al.*, 2021). The JC mutation model was assumed in the analysis. Gamma priors are assigned to population size parameters (θ) and to the age of the root on the species tree; $\theta \sim G(2,400)$ and $\tau_0 \sim G(2,200)$. Note that the gamma distribution $G(a,b)$ has mean a/b and variance a/b^2, so that the shape parameter $a = 2$ means diffuse priors. Introgression probability φ was assigned the beta prior beta$(1,1)$, which is $\mathbb{U}(0,1)$.

We used 32,000 MCMC iterations as burnin, and took 2×10^5 samples, sampling every 5 iterations.

Introgression events assigned to wrong branches

In this set of simulations, the introgression event was assigned onto either the parental or a daughter branch of the branch truly involved in introgression. We used models A and B (fig. 4). The species divergence times τ are shown in the trees (fig. 4). We used $S = 4$ sequences per species per locus, with the sequence length n=500 sites. The number of loci was $L =$ 250, 1000, and 4000. We used two population sizes, with $\theta_0 = 0.002$ and $\theta_1 = 0.01$ for the thin and thick branches on the species tree, respectively. We simulated 100 replicate datasets for each parameter setting.

The data were then analyzed using BPP under both models A and B (fig. 4**a**&**b**). We assign gamma priors, $\theta \sim G(2,400)$ with mean 0.005 and $\tau_0 \sim G(2,200)$ with mean 0.01. With two trees/models, three numbers of loci, $2 \times 3 \times 100 = 600$ datasets were simulated, each analyzed twice (under each of the two models). We used 32,000 MCMC iterations as burnin, and took 2×10^5 samples, sampling every 5 iterations.

Continuous migration versus episodic introgression

The data were simulated under the IIM models of figure 4**c**&**d**, with continuous migration at the rate $M = 0.1$ per migrant per generation, and analyzed under MSci models of figure 4**a**&**b**. This is similar to the two-species case analyzed earlier using the asymptotic theory and computer simulation, but here the species tree is larger with more species. There are four combinations. In setting C-A (simulation model C and analysis model A) and D-B, gene flow is continuous in the true model but the MSci model assumes episodic introgression at a particular time point, so that the mode of gene flow is misspecified. In settings C-B and D-A, the mode of gene flow was similarly misspecified but we had in addition misassignment of gene flow to wrong branches on the species tree. Other parameter settings were the same as above. With two trees, three datasizes (L), a total of 600 datasets were generated, each analyzed under two models (fig. 4**a**&**b**).

Isolation with initial migration (IIM) model

The IIM model (Costa and Wilkinson-Herbots, 2017) assumes that there is migration after the species divergence but gene flow ceased after certain time. Suppose we sample sequence data from species A, B and C under model A of figure 6**a**. The model assumes migration between A and B over the time period τ_S–τ_T. We use the MSci model of figure 6**b** to analyze the sequence data sampled from species A, B and C. Other parameter settings are the same as above. With three values for L, we simulated 300 datasets, all analyzed under the MSci model (fig. 6**b**).

Ghost species

We simulated data under MSci model A (see fig. 1A in Flouri *et al.*, 2020) of figure 7**a**′ and analyzed them under the MSci model B of figure 7**b**, with $\tau_X = \tau_Y$ incorrectly assumed. Here introgression involved a ghost species XUV which has become extinct or is unsampled in the data. This scenario is fully represented by model A of figure 7**a**, so we simulated data on the species tree of figure 7**a**. With three values for L, 300 datasets were generated, all analyzed under the MSci model of figure 7**b**.

We also used the IIM model of figure 8**a** to generate data, with migration from species C to SU and from SU to B. Species V and W represent extinct or unsampled ghost species. The data were simulated using the species tree for five species of figure 8**a** but with no sequences sampled from species V and W, while samples from species A, B and C constitute the data used to fit the MSci model (fig. 8**b**). With three values for L and 100 replicates, 300 datasets were simulated, all analyzed under the MSci model of figure 8**b**.

Acknowledgments

This study has been supported by Biotechnology and Biological Sciences Research Council grants (BB/T003502/1, BB/R01356X/1), as well as by Harvard University.

References

Anderson, E. 1949. *Introgressive Hydridization*. John Wiley, New York.

Bahlo, M. and Griffiths, R. C. 2000. Inference from gene trees in a subdivided population. *Theor. Popul. Biol.*, 57: 79–95.

Beerli, P. 2004. Effect of unsampled populations on the estimation of population sizes and migration rates between sampled populations. *Mol. Ecol.*, 13: 827–836.

Beerli, P. and Felsenstein, J. 1999. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics*, 152: 763–773.

Beerli, P. and Felsenstein, J. 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in *n* subpopulations by using a coalescent approach. *Proc. Natl. Acad. Sci. U.S.A.*, 98: 4563–4568.

Blischak, P. D., Chifman, J., Wolfe, A. D., and Kubatko, L. S. 2018. Hyde: A Python package for genome-scale hybridization detection. *Syst. Biol.*, 67(5): 821–829.

Burgess, R. and 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.

Costa, R. J. and Wilkinson-Herbots, H. 2017. Inference of gene flow in the process of speciation: An efficient maximum-likelihood method for the isolation-with-initial-migration model. *Genetics*, 205(4): 1597–1618.

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

Dawid, A. 2011. Posterior model probabilities. In P. S. Bandyopadhyay and M. Forster, editors, *Philosophy of Statistics*, pages 607–630. Elsevier, New York.

DeGiorgio, M. and Degnan, J. H. 2014. Robustness to divergence time underestimation when inferring species trees from estimated gene trees. *Syst. Biol.*, 63(1): 66–82.

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

Dickey, J. M. 1971. The weighted likelihood ratio, linear hypotheses on normal location parameters. *Ann. Math. Statist.*, 42(1): 204–223.

Dobzhansky, T. 1937. *Genetics and the Origin of Species*. Columbia University, New York.

Durand, E. Y., Patterson, N., Reich, D., and 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., and Davey, J. e. a. 2019. Genomic architecture and introgression shape a butterfly radiation. *Science*, 366(6465): 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., and Wolf, J. B. W. 2012. The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature*, 491: 756–760.

Elworth, R. A. L., Ogilvie, H. A., Zhu, J., and Nakhleh, L. 2019. Advances in computational methods for phylogenetic networks in the presence of hybridization. *Bioinformatics and Phylogenetics*, 29: 317–360.

Finger, N., Farleigh, K., Bracken, J., Leache, A., Francois, O., Yang, Z., Flouri, T., Charran, T., Jezkova, T., Williams, D., and Blair, C. 2022. Genome-scale data reveal deep lineage divergence and a complex demographic history in the Texas horned lizard (*Phrynosoma cornutum*) throughout the southwestern and central usa. *Genome Biol. Evol.*, 14(1): 10.1093/gbe/evab260.

Fisher, R. 1922. On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc. A*, 222: 309–368.

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

Flouri, T., Jiao, X., Rannala, B., and Yang, Z. 2020. A Bayesian implementation of the multispecies coalescent model with introgression for phylogenomic analysis. *Mol. Biol. Evol.*, 37(4): 1211–1223.

Gelman, A. and Meng, X. 1998. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Stat. Sci.*, 13: 163–185.

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

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., and Stenzel, U. e. a. 2010. A draft sequence of the Neandertal genome. *Science*, 328: 710–722.

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

Hey, J. and 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., and Wang, Y. 2018. Phylogeny estimation by integration over isolation with migration models. *Mol. Biol. Evol.*, 35(11): 2805–2818.

Hibbins, M. S. and Hahn, M. W. 2022. Phylogenomic approaches to detecting and characterizing introgression. *Genetics*, 220(2): iyab173.

Ji, J., Jackson, D. J., Leache, A. D., and Yang, Z. 2022. Significant cross-species gene flow detected in the *Tamias quadrivittatus* group of North American chipmunks. *BioRxiv*, page DOI: 10.1101/2021.12.07.471567.

Jiao, X. and Yang, Z. 2021. Defining species when there is gene flow. *Syst. Biol.*, 70(1): 108—119.

Jiao, X., Flouri, T., Rannala, B., and Yang, Z. 2020. The impact of cross-species gene flow on species tree estimation. *Syst. Biol.*, 69(5): 830–847.

Jiao, X., Flouri, T., and Yang, Z. 2021. Multispecies coalescent and its applications to infer species phylogenies and cross-species gene flow. *Nat. Sci. Rev.*, 8: nwab127 (DOI: 10.1093/nsr/nwab127).

Jukes, T. and Cantor, C. 1969. Evolution of protein molecules. In H. Munro, editor, *Mammalian Protein Metabolism*, pages 21–123. Academic Press, New York.

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

Lartillot, N. and Philippe, H. 2006. Computing bayes factors using thermodynamic integration. *Syst. Biol.*, 55: 195–207.

Liu, S., Lorenzen, E. D., Fumagalli, M., Li, B., and Harris, K. e. a. 2014. Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. *Cell*, 157: 785–794.

Lohse, K. and Frantz, L. A. 2014. Neandertal admixture in Eurasia confirmed by maximum-likelihood analysis of three genomes. *Genetics*, 196(4): 1241–1251.

Lohse, K., Harrison, R., and Barton, N. 2011. A general method for calculating likelihoods under the coalescent process. *Genetics*, 189: 977–987.

Maddison, W. 1997. Gene trees in species trees. *Syst. Biol.*, 46: 523–536.

Malecot, G. 1948. *Les mathematiques de I'heredite*. Masson, Paris.

Mallet, J. 2007. Hybrid speciation. *Nature*, 446: 279–283.

Mallet, J., Besansky, N., and Hahn, M. W. 2016. How reticulated are species? *BioEssays*, 38(2): 140–149.

Martin, S. H. and Jiggins, C. D. 2017. Interpreting the genomic landscape of introgression. *Curr. Opin. Genet. Dev.*, 47: 69–74.

Martin, S. H., Dasmahapatra, K. K., Nadeau, N. J., Salazar, C., Walters, J. R., Simpson, F., Blaxter, M., Manica, A., Mallet, J., and Jiggins, C. D. 2013. Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Res.*, 23(11): 1817–1828.

Martin, S. H., Davey, J. W., Salazar, C., and Jiggins, C. D. 2019. Recombination rate variation shapes barriers to introgression across butterfly genomes. *PLoS Biol.*, 17(2): e2006288.

Meng, C. and Kubatko, L. 2009. Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. *Theo. Popul. Biol.*, 75(1): 35–45.

Muller, H. J. 1942. Isolating mechanisms, evolution, and temperature. *Biol. Symp*, 6: 71–125.

Nichols, R. 2001. Gene trees and species trees are not the same. *Trends Ecol. Evol.*, 16: 358–364.

Notohara, M. 1990. The coalescent and the genealogical process in geographically structured populations. *J. Math. Biol.*, 29: 59–75.

Ogilvie, H. A., Heled, J., Xie, D., and Drummond, A. J. 2016. Computational performance and statistical accuracy of *beast and comparisons with other methods. *Syst. Biol.*, 65: 381–396.

O'Hagan, A. and Forster, J. 2004. *Kendall's Advanced Theory of Statistics: Bayesian Inference*. Arnold, London.

Ottenburghs, J. 2020. Ghost introgression: spooky gene flow in the distant past. *Bioessays*, 42(6): e2000012.

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

Schumer, M., Xu, C., Powell, D. L., Durvasula, A., Skov, L., Holland, C., Blazier, J. C., Sankararaman, S., Andolfatto, P., Rosenthal, G. G., and Przeworski, M. 2018. Natural selection interacts with recombination to shape the evolution of hybrid genomes. *Science*, 360(6389): 656–660.

Shi, C. and 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: 159–179.

Slatkin, M. 1987. Gene flow and the geographic structure of natural populations. *Science*, 236(4803): 787–792.

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

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

Thawornwattana, Y., Dalquen, D., and Yang, Z. 2018. 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., Seixas, F. A., Mallet, J., and Yang, Z. 2022. Full-likelihood genomic analysis clarifies a complex history of species divergence and introgression: the example of the *erato-sara* group of *Heliconius* butterflies. *Syst. Biol.*

Tiley, G. P., Flouri, T., Jiao, X., Poelstra, J. P., Xu, B., Zhu, T., Rannala, B., Yoder, A. D., and Yang, Z. 2022. Estimation of species divergence times in presence of cross-species gene flow. *Syst. Biol.*

Tricou, T., Tannier, E., and de Vienne, D. M. 2022. Ghost lineages highly influence the interpretation of introgression tests. *Syst. Biol.*, page 10.1093/sysbio/syac011.

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

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

Wright, S. 1943. Isolation by distance. *Genetics*, 28: 114–138.

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. and Flouri, T. 2022. Estimation of cross-species introgression rates using genomic data despite model unidentifiability. *Mol. Biol. Evol.*

Yang, Z. and Rannala, B. 2014. Unguided species delimitation using dna sequence data from multiple loci. *Mol. Biol. Evol.*, 31(12): 3125–3135.

Yang, Z. and Zhu, T. 2018. Bayesian selection of misspecified models is overconfident and may cause spurious posterior probabilities for phylogenetic trees. *Proc. Natl. Acad. Sci. U.S.A.*, 115(8): 1854–1859.

Yu, Y. and Nakhleh, L. 2015. A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics*, 16 Suppl 10: S10.

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

Zhang, C., Ogilvie, H. A., Drummond, A. J., and Stadler, T. 2018. Bayesian inference of species networks from multilocus sequence data. *Mol. Biol. Evol.*, 35: 504–517.

Zhu, T. and 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.

Zhu, T. and Yang, Z. 2021. Complexity of the simplest species tree problem. *Mol. Biol. Evol.*, 39: 3993—-4009.

Zhu, T., Flouri, T., and Yang, Z. 2022. A simulation study to examine the impact of recombination on phylogenomic inferences under the multispecies coalescent model. *Mol. Ecol.*, page DOI: 10.1111/mec.16433.

Appendix A. Likelihood function under the IM and MSci models in the case of two species

Here we derive the likelihood function under the three continuous migration models (IM, IIM, SC) and the episodic introgression (MSci) model for two species (fig. 1**a**-**d**) when the data consist of an infinite number of loci, with two sequences sampled at each locus, one from each species. The data at each locus can be summarized as x differences at n sites. The infinite-sites mutation model is assumed so that the probability of data given the coalescent time is given by the Poisson probability (eq. 8). To calculate the likelihood, we integrate over the unknown coalescent time t, which has density $f_{\rm m}(t|\Theta_{\rm m})$ (eq. 3) under the IM or IIM model, $f_{\rm sc}(t|\Theta_{\rm m})$ (eq. 4) under the SC model, and $f_{\rm i}(t|\Theta_{\rm i})$ (eq. 5) under the MSci model.

Under the IM and IIM models (fig. 1**a**&**b**), we have from eq. 3

$$f(x|\Theta_{\rm m}) = \int_0^\infty f(x|t) f_{\rm m}(t|\Theta_{\rm m})\,{\rm d}t = \int_0^\infty \frac{1}{x!}(2nt)^x {\rm e}^{-2nt} f_{\rm m}(t|\Theta_{\rm m})\,{\rm d}t = I_1 + I_2. \tag{A1}$$

The first term is

$$\begin{aligned} I_1 &= \int_{\tau_T}^{\tau_R} \frac{1}{x!}(2nt)^x {\rm e}^{-2nt} \frac{2w}{2-w\theta_A}\left[{\rm e}^{-w(t-\tau_T)} - {\rm e}^{-\frac{2}{\theta_A}(t-\tau_T)}\right]{\rm d}t \\ &= \frac{(2n)^x}{x!}\frac{2w}{2-w\theta_A}\left[{\rm e}^{w\tau_T}\int_{\tau_T}^{\tau_R} t^x {\rm e}^{-(2n+w)t}\,{\rm d}t - {\rm e}^{\frac{2}{\theta_A}\tau_T}\int_{\tau_T}^{\tau_R} t^x {\rm e}^{-(2n+\frac{2}{\theta_A})t}\,{\rm d}t\right], \end{aligned} \tag{A2}$$

where the two integrals are

$$\begin{aligned} \int_{\tau_T}^{\tau_R} t^x {\rm e}^{-(2n+w)t}\,{\rm d}t &= \frac{1}{(2n+w)^{x+1}}\Big[\gamma(x+1,\tau_R(2n+w)) - \gamma(x+1,\tau_T(2n+w))\Big], \\ \int_{\tau_T}^{\tau_R} t^x {\rm e}^{-(2n+\frac{2}{\theta_A})t}\,{\rm d}t &= \frac{1}{(2n+\frac{2}{\theta_A})^{x+1}}\Big[\gamma(x+1,\tau_R(2n+\tfrac{2}{\theta_A})) - \gamma(x+1,\tau_T(2n+\tfrac{2}{\theta_A}))\Big], \end{aligned} \tag{A3}$$

with

$$\begin{aligned} \Gamma(a) &= \int_0^\infty t^{a-1}{\rm e}^{-t}\,{\rm d}t, \\ \gamma(a,x) &= \int_0^x t^{a-1}{\rm e}^{-t}\,{\rm d}t, \end{aligned} \tag{A4}$$

to be the gamma function and the lower incomplete gamma function, respectively, with $\gamma(a,\infty)=\Gamma(a)$.

Similarly the second term in eq. A1 is

$$\begin{aligned} I_2 &= \int_{\tau_R}^\infty \frac{1}{x!}(2nt)^x {\rm e}^{-2nt}\frac{1}{2-w\theta_A}\left[2{\rm e}^{-w(\tau_R-\tau_T)} - w\theta_A {\rm e}^{-\frac{2}{\theta_A}(\tau_R-\tau_T)}\right]\frac{2}{\theta_R}{\rm e}^{-\frac{2}{\theta_R}(t-\tau_R)}\,{\rm d}t \\ &= \frac{(2n)^x}{x!}\frac{1}{2-w\theta_A}\left[2{\rm e}^{-w(\tau_R-\tau_T)} - w\theta_T {\rm e}^{-\frac{2}{\theta_A}(\tau_R-\tau_T)}\right]\frac{2}{\theta_R}{\rm e}^{\frac{2}{\theta_R}\tau_R}\int_{\tau_R}^\infty t^x {\rm e}^{-(2n+\frac{2}{\theta_R})t}\,{\rm d}t \\ &= \frac{(2n)^x}{x!}\frac{1}{2-w\theta_A}\left[2{\rm e}^{-w(\tau_R-\tau_T)} - w\theta_T {\rm e}^{-\frac{2}{\theta_A}(\tau_R-\tau_T)}\right]\frac{2}{\theta_R}{\rm e}^{\frac{2}{\theta_R}\tau_R}\times\frac{\Gamma(x+1)-\gamma(x+1,\tau_R(2n+\frac{2}{\theta_R}))}{(2n+\frac{2}{\theta_R})^{x+1}}. \end{aligned} \tag{A5}$$

Putting everything together, we get

$$\begin{aligned} f_{\rm m}(x|\Theta_{\rm m}) = \frac{(2n)^x}{x!}\frac{2w}{2-w\theta_A}\Bigg(&\frac{{\rm e}^{w\tau_T}}{(2n+w)^{x+1}}\Big[\gamma(x+1,\tau_R(2n+w)) - \gamma(x+1,\tau_T(2n+w))\Big] \\ &- \frac{{\rm e}^{\frac{2}{\theta_A}\tau_T}}{(2n+\frac{2}{\theta_A})^{x+1}}\Big[\gamma(x+1,\tau_R(2n+\tfrac{2}{\theta_A})) - \gamma(x+1,\tau_T(2n+\tfrac{2}{\theta_A}))\Big]\Bigg) \\ &+ \frac{(2n)^x}{x!}\frac{1}{2-w\theta_A}\left[2{\rm e}^{-w(\tau_R-\tau_T)} - w\theta_A e^{-\frac{2}{\theta_A}(\tau_R-\tau_T)}\right]\frac{2}{\theta_R}{\rm e}^{\frac{2}{\theta_R}\tau_R} \\ &\times\frac{\Gamma(x+1)-\gamma(x+1,\tau_R(2n+\frac{2}{\theta_R}))}{(2n+\frac{2}{\theta_R})^{x+1}}. \end{aligned} \tag{A6}$$

Similarly, under the secondary-contact (SC) model (fig. 1**c**), the density of coalescent time t is given in eq. 4. The probability of observing x differences at n sites at a locus is

$$f_{\rm sc}(x|\Theta_{\rm m}) = \int_0^\infty f(x|t) f_{\rm sc}(t|\Theta_{\rm m})\,{\rm d}t = J_1 + J_2 + J_3, \tag{A7}$$

where

$$\begin{aligned}
J_1 &= \int_0^{\tau_T} \frac{1}{x!}(2nt)^x {\rm e}^{-2nt} \frac{w\theta_A}{2-w\theta_A}\left[{\rm e}^{-wt} - {\rm e}^{-\frac{2}{\theta_A}t}\right]\frac{2}{\theta_A}\,{\rm d}t \\
&= \frac{(2n)^x}{x!}\frac{2w}{2-w\theta_A}\left[\frac{\gamma(x+1,\tau_T(2n+w))}{(2n+w)^{x+1}} - \frac{\gamma(x+1,\tau_T(2n+\frac{2}{\theta_A}))}{(2n+\frac{2}{\theta_A})^{x+1}}\right], \\
J_2 &= \int_{\tau_T}^{\tau_R} \frac{1}{x!}(2nt)^x {\rm e}^{-2nt} \frac{w\theta_A}{2-w\theta_A}\left[{\rm e}^{-w\tau_T} - {\rm e}^{-\frac{2}{\theta_A}\tau_T}\right]\frac{2}{\theta_A}{\rm e}^{-\frac{2}{\theta_A}(t-\tau_T)}\,{\rm d}t \\
&= \frac{(2n)^x}{x!}\frac{2w}{2-w\theta_A}\left[{\rm e}^{-w\tau_T} - {\rm e}^{-\frac{2}{\theta_A}\tau_T}\right]\frac{{\rm e}^{\frac{2}{\theta_A}\tau_T}}{(2n+\frac{2}{\theta_A})^{x+1}}\left[\gamma(x+1,\tau_R(2n+\tfrac{2}{\theta_A})) - \gamma(x+1,\tau_T(2n+\tfrac{2}{\theta_A}))\right], \\
J_3 &= \int_{\tau_R}^{\infty} \frac{1}{x!}(2nt)^x {\rm e}^{-2nt}\left[\frac{w\theta_A}{2-w\theta_A}\left({\rm e}^{-w\tau_T} - {\rm e}^{-\frac{2}{\theta_A}\tau_T}\right){\rm e}^{-\frac{2}{\theta_A}(\tau_R-\tau_T)} + {\rm e}^{-w\tau_T}\right]\frac{2}{\theta_R}{\rm e}^{-\frac{2}{\theta_R}(t-\tau_R)}\,{\rm d}t \\
&= \frac{(2n)^x}{x!}\left[\frac{w\theta_A}{2-w\theta_A}\left({\rm e}^{-w\tau_T} - {\rm e}^{-\frac{2}{\theta_A}\tau_T}\right){\rm e}^{-\frac{2}{\theta_A}(\tau_R-\tau_T)} + {\rm e}^{-w\tau_T}\right]\frac{2}{\theta_R}\frac{{\rm e}^{\frac{2}{\theta_R}\tau_R}}{(2n+\frac{2}{\theta_R})^{x+1}}\left[\Gamma(x+1) - \gamma(x+1,\tau_R(2n+\tfrac{2}{\theta_R}))\right].
\end{aligned} \tag{A8}$$

Finally, under the MSci model (fig. 1**d**), the density of coalescent time t is given in eq. 5. We have

$$\begin{aligned}
f_{\rm i}(x|\Theta_{\rm i}) &= \int_0^\infty f(x|t) f_{\rm i}(t|\Theta_{\rm i})\,{\rm d}t \\
&= \int_{\tau_S}^{\tau_R} \frac{1}{x!}(2nt)^x {\rm e}^{-2nt}\,\varphi\frac{2}{\theta_S}{\rm e}^{-\frac{2}{\theta_S}(t-\tau_S)}\,{\rm d}t + \int_{\tau_R}^{\infty}\frac{1}{x!}(2nt)^x {\rm e}^{-2nt}\left[\varphi\,{\rm e}^{-\frac{2}{\theta_S}(\tau_R-\tau_S)} + (1-\varphi)\right]\frac{2}{\theta_R}{\rm e}^{-\frac{2}{\theta_R}(t-\tau_R)}\,{\rm d}t \\
&= \frac{(2n)^x}{x!}\varphi\frac{2}{\theta_S}{\rm e}^{\frac{2}{\theta_S}\tau_S}\int_{\tau_S}^{\tau_R} t^x {\rm e}^{-(2n+\frac{2}{\theta_S})t}\,{\rm d}t + \frac{(2n)^x}{x!}\left[\varphi\,{\rm e}^{-\frac{2}{\theta_S}(\tau_R-\tau_S)} + (1-\varphi)\right]\frac{2}{\theta_R}{\rm e}^{\frac{2}{\theta_R}\tau_R}\int_{\tau_R}^{\infty} t^x {\rm e}^{-(2n+\frac{2}{\theta_R})t}\,{\rm d}t \\
&= \frac{(2n)^x}{x!}\varphi\frac{2}{\theta_S}{\rm e}^{\frac{2}{\theta_S}\tau_S}\times\frac{\gamma(x+1,\tau_R(2n+\frac{2}{\theta_S})) - \gamma(x+1,\tau_S(2n+\frac{2}{\theta_S}))}{(2n+\frac{2}{\theta_S})^{x+1}} \\
&\quad + \frac{(2n)^x}{x!}\left[\varphi\,{\rm e}^{-\frac{2}{\theta_S}(\tau_R-\tau_S)} + (1-\varphi)\right]\frac{2}{\theta_R}{\rm e}^{\frac{2}{\theta_R}\tau_R}\times\frac{\Gamma(x+1) - \gamma(x+1,\tau_R(2n+\frac{2}{\theta_R}))}{(2n+\frac{2}{\theta_R})^{x+1}}.
\end{aligned} \tag{A9}$$

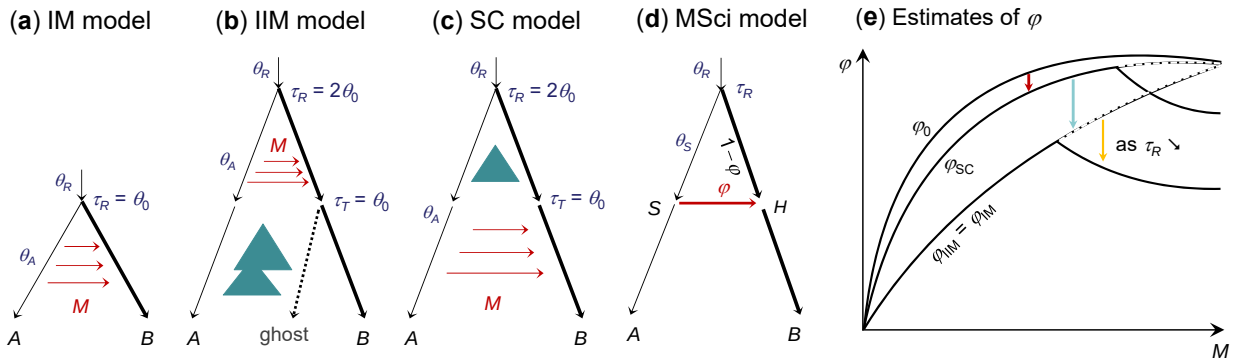


Figure 1: (**a**–**c**) Three migration models for two species A and B used to generate data: IM (isolation-with-migration), IIM (isolation-with-initial-migration), and SC (secondary-contact). The IIM model is an instance of the IM model with a ghost species at node T and with migration from species A to T (**b**). Similarly the SC model (**c**) is a case of the migration model with $\tau_T > 0$. Divergence time (τ_R) and the time point at which gene flow started or stopped (τ_T) are given next to the nodes, with population sizes $\theta_0 = 0.002$ for the thin branches and $\theta_1 = 0.01$ for the thick branches. The migration rate was $M = 0.2$ migrant individuals from A to B per generation, but other values are considered as well. Note that the time period of gene flow is $\Delta\tau = \theta_0$ in all three models, so that the total expected amount of introgression (φ_0 of eq. 10) is the same. (**d**) The introgression (MSci) model is fitted to the data of the coalescent time between two sequences (i.e., with infinite sequence length), generated under the migration models of **a**-**c**. (**e**) A schematic summary of the estimate of the introgression probability (φ) in the MSci model (**d**) when the data are generated under the models of **a**-**c**.

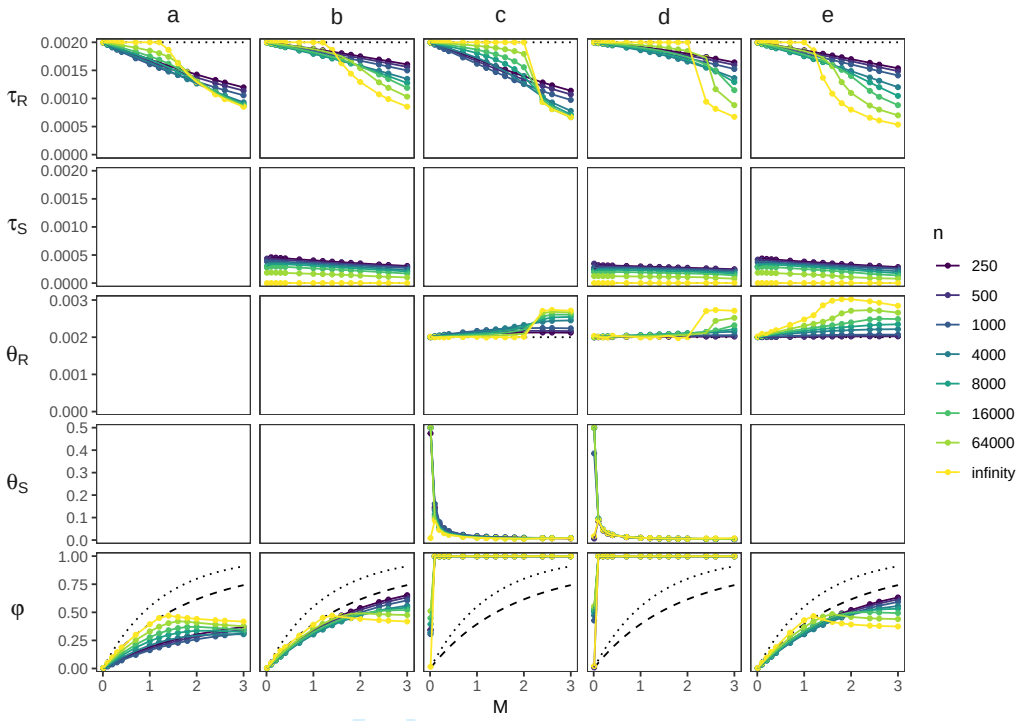


Figure 2: **(2s-IM:MLE)** Best-fitting parameter values under the MSci model of figure 1**d** when data of two sequences per locus (one per species), each of n sites, are generated under the IM model of figure 1**a**. Five methods (a-e) are used to fit the MSci model, estimating 2, 3, 4, 5, and 4 parameters, respectively, while the other parameters are fixed. In (**a**), τ_R and φ are estimated, but θ_R and θ_S are fixed at their true values in the IM model, and the introgression time $\tau_S = \tau_H$ is fixed at $\tau_T = 0$. In (b), τ_S is estimated as well. In (c), $\tau_S = 0$ is fixed, while the other four parameters are estimated. In (d), all five parameters are estimated. In (e), the constraint $\theta_R = \theta_S$ is enforced so that four free parameters are estimated. The dotted lines for φ indicate the true total amount of introgression of eq. 10. The dashed lines indicate φ^* of eq. 11. The true and best-fitting distributions of the coalescent time (t) are shown in figure S1.

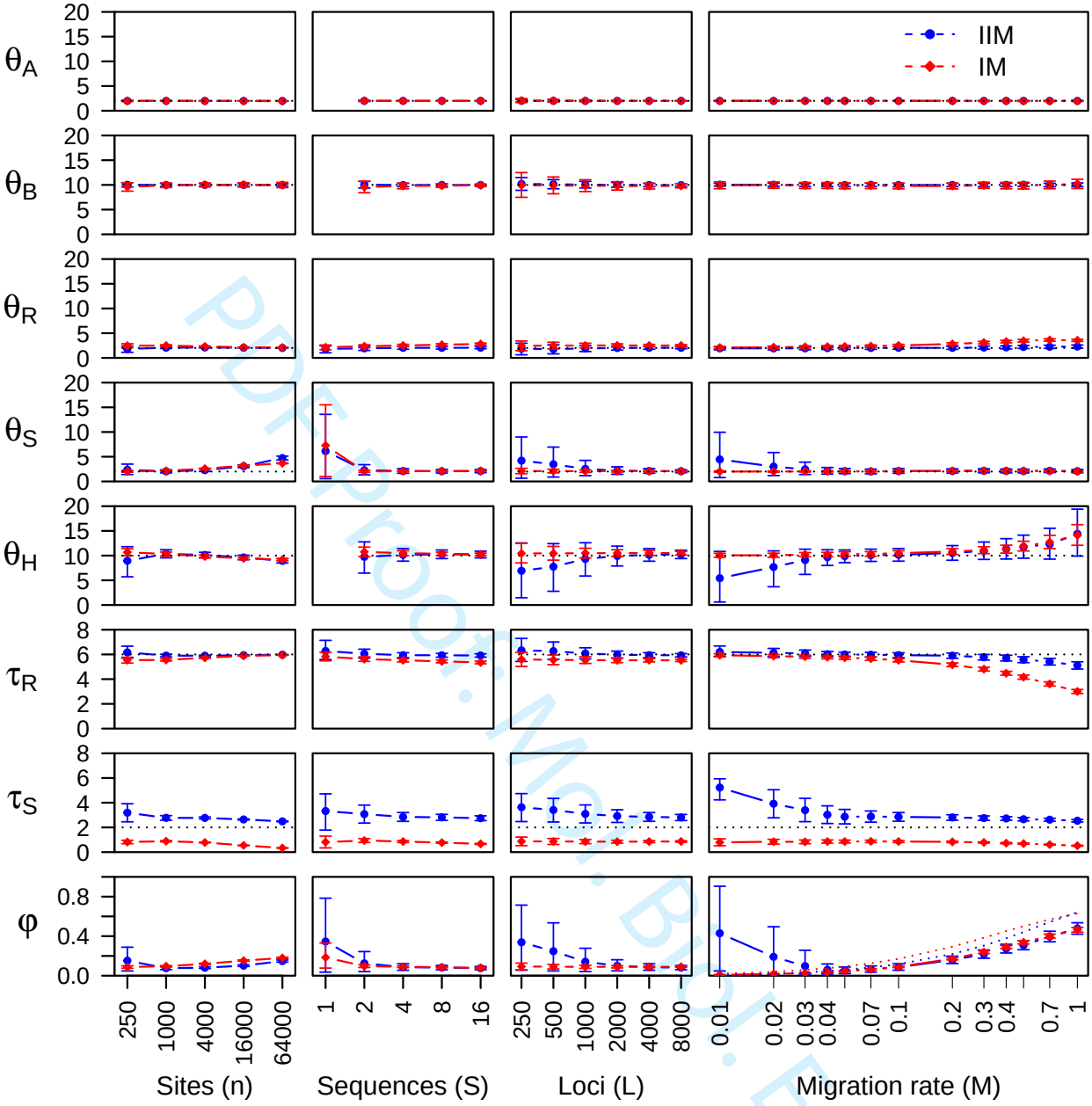


Figure 3: Average posterior means and 95% HPD CIs for parameters in the MSci model of figure 1**d** over 30 replicate datasets simulated under the migration (IM, IIM, and SC) models of figure 1**a**-**c**, plotted against the number of sites per sequence (n), the number of sequences per species (S), the number of loci (L), and the migration rate (M). Parameters in the migration model are given in the legend to figure 1. In the standard setting, each dataset consists of $L = 4000$ loci, with $S = 4$ sequences per species at each locus and $n = 1000$ sites per sequence, and the migration rate was $M = 0.2$ individuals per generation. In the four sets of simulations, one of the factors (n, S, L, M) varies while the others are fixed. When $S = 1$, population sizes θ_A, θ_B, and θ_H are unidentifiable. Estimates of τs and θs are multiplied by 10^3. Dotted lines indicate true values of identifiable parameters, except in the plot of φ against M, where it represents φ_0 of eq. 10, (which is identical for the IM, IIM, and SC models of fig. 1). Note that the n, S, L, and M axes are all on the logarithmic scale.

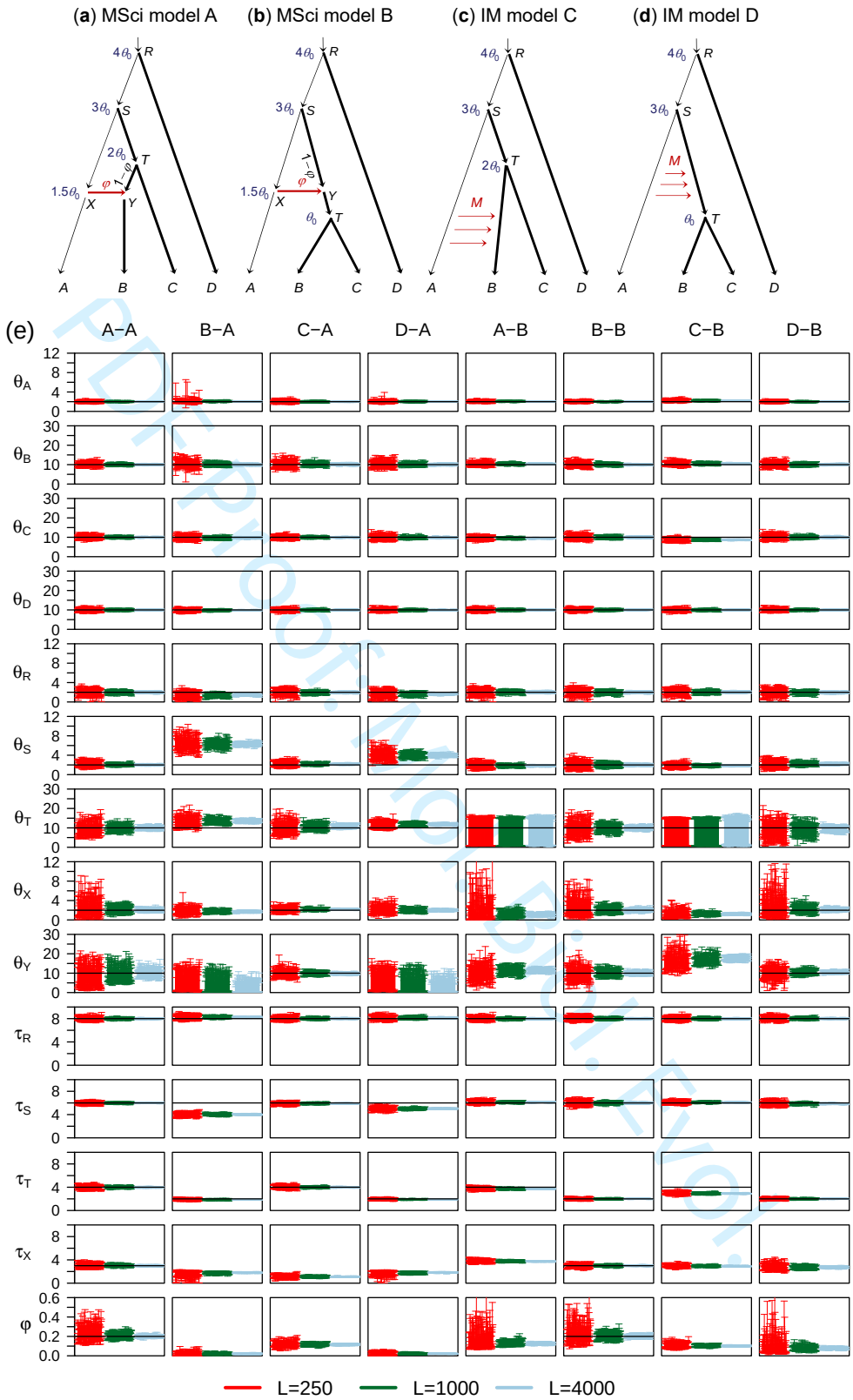


Figure 4: (**a**&**b**) Two introgression (MSci) models and (**c**&**d**) two migration (IM) models used in simulation. The thin branches have the population size $\theta_0 = 0.002$ and the thick branches have $\theta_1 = 0.01$. In MSci model A, the species divergence/introgression times are $\tau_R = 4\theta_0$, $\tau_S = 3\theta_0$, $\tau_T = 2\theta_0$, and $\tau_X = \tau_Y = 1.5\theta_0$. In MSci model B, $\tau_R = 4\theta_0$, $\tau_S = 3\theta_0$, $\tau_T = \theta_0$, and $\tau_X = \tau_Y = 1.5\theta_0$. Introgression probability is $\varphi = 0.2$. In the IM model C, $\tau_R = 4\theta_0$, $\tau_S = 3\theta_0$, and $\tau_T = 2\theta_0$, with migration occurring from species A to B over time period $(0, \tau_T)$ at the rate $M = 0.1$ migrants per generation. In the IM model D, $\tau_R = 4\theta_0$, $\tau_S = 3\theta_0$, and $\tau_T = \theta_0$, with migration from species A to ST over time period (τ_T, τ_S) at the rate $M = 0.1$. (**e**) The 95% HPD CIs for parameters in 100 replicate datasets of $L = 250, 1000$, and 4000 loci. The key is in the simulation-analysis format; i.e., 'B-A' means that data are simulated under model B and analyzed under model A. Parameters θs and τs are multiplied by 10^3. Black solid line indicates the true value.

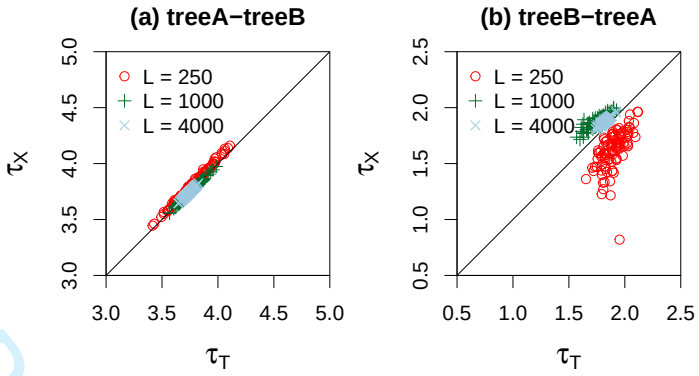


Figure 5: Posterior means of speciation/introgression times ($\times 10^{-3}$) when the introgression event is assigned to a wrong branch. In (**a**) tree A-tree B, data were simulated using species tree A (fig. 4**a**), with introgression from species A to B, but are analyzed assuming tree B, with introgression assigned incorrectly to the parental branch ST (so that $\tau_X > \tau_T$). In (**b**) tree B-tree A, data were simulated under tree B (fig. 4**b**) and analyzed assuming tree A, with introgression assigned to the daughter branch B (with $\tau_X < \tau_T$). For each datasize (with $L = 250, 1000$, or 4000 loci), 100 replicate datasets were generated and analyzed. These correspond to the A-B and B-A settings of figure 4**e**, where estimates of other parameters are shown.

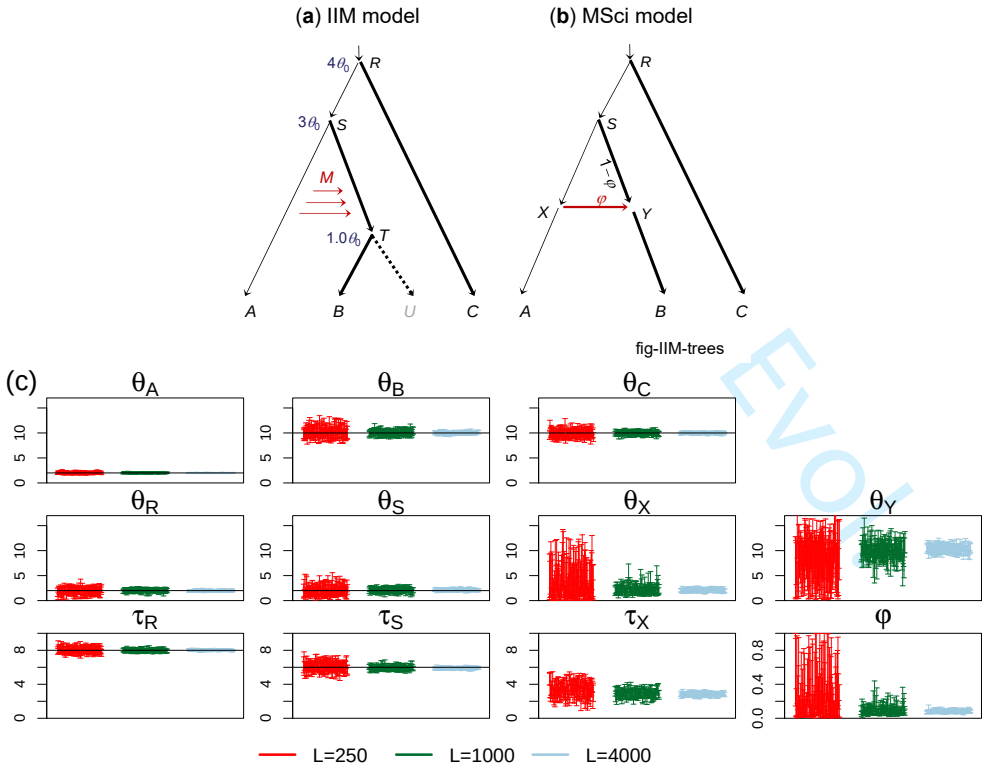


Figure 6: (**a**) An isolation-with-initial-migration (IIM) model used to simulate data. The parameter values used are $\theta_0 = 0.002$ for population sizes for the thin branches and $\theta_1 = 0.01$ for the thick branches, $\tau_R = 4\theta_0$, $\tau_S = 3\theta_0$, $\tau_T = \theta_0$ for species divergence times. The number of sequences is $S = 4$, with the sequence length $n = 500$. The migration rate is $M = 0.1$. (**b**) An MSci model used to analyze the data. (**c**) The 95% HPD CIs for parameters, with black lines indicating the true values. Estimates of θ and τ are multiplied by 10^3.

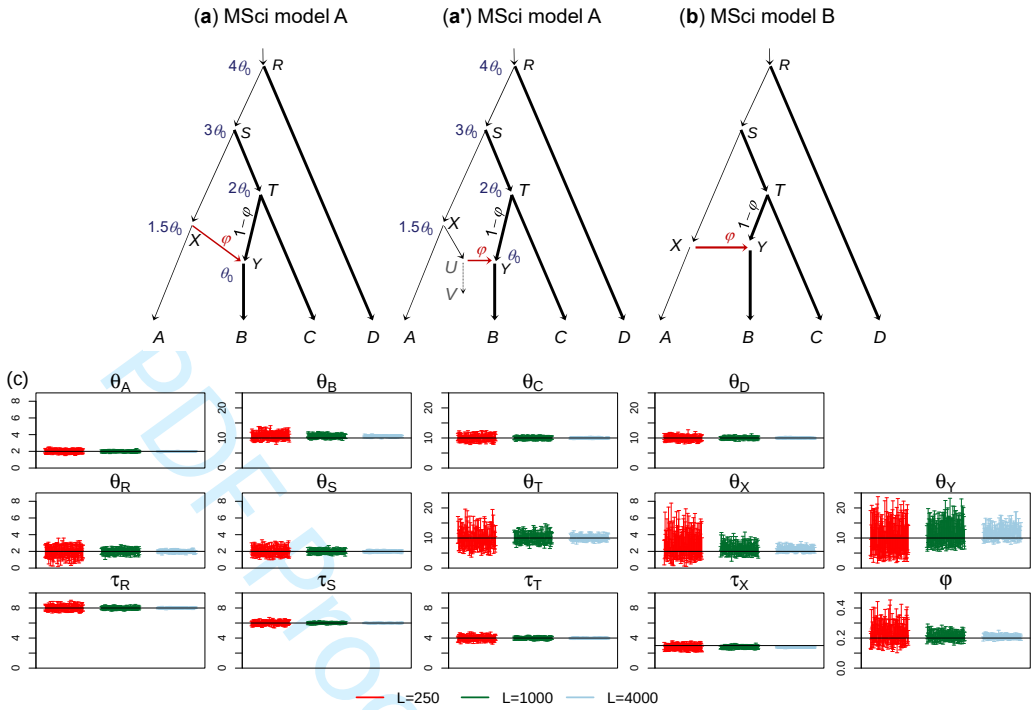


Figure 7: (**a**) MSci model A (fig. 1A in Flouri *et al.*, 2020) assumes that $\tau_X > \tau_Y$ and $\tau_T > \tau_Y$ and can represent scenario (**a**′) in which species X split into two species, and species XUV contributed migrants into species TB at time τ_Y but has since become extinct. This is assumed to simulate data, with $\theta_0 = 0.002$ for the thin branches and $\theta_1 = 0.01$ for the thick branches, $\tau_R = 4\theta_0, \tau_S = 3\theta_0, \tau_T = 2\theta_0$, $\tau_X = 1.5\theta_0$, and $\tau_Y = \theta_0$. The introgression probability is $\varphi = 0.2$. The number of sequences is $S = 4$, and the sequence length is $n = 500$. (**b**) MSci model B (fig. 1B in Flouri *et al.*, 2020) used to analyze the data, which incorrectly assumes $\tau_X = \tau_Y$. (**c**) The 95% HPD CIs for parameters, with θs and τs multiplied by 10^3 and black solid line indicating the true value.

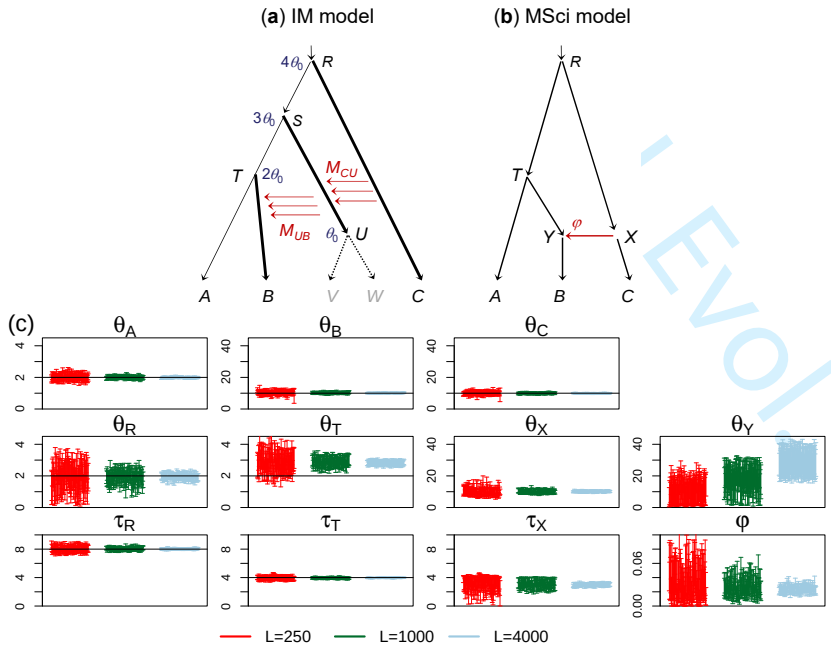


Figure 8: (**a**) Migration model involving ghost species for simulating data. The parameter values used are $\theta_0 = 0.002$ for the thin branches and $\theta_1 = 0.01$ for the thick branches, with the divergence times (τs) shown next to the internal nodes. The number of sequences is $S = 4$, and the sequence length is $n = 500$. The migration rates are $M_{CU} = M_{UB} = 0.2$ migrants per generation. (**b**) MSci model used to analyze the data. (**c**) The 95% HPD CI for parameters, with θs and τs multiplied by 10^3, and with black solid lines indicating the true values.

Table 1. Average posterior means and 95% HPD intervals (in parentheses) for introgression time (τ_X, $\times 10^3$) and introgression probability (φ_X) in the simulations

	τ_X			φ		
Analysis	$L = 250$	$L = 1000$	$L = 4000$	$L = 250$	$L = 1000$	$L = 4000$
Fig. 4 A-A	3.06 (2.63, 3.49)	3.02 (2.80, 3.24)	3.00 (2.89, 3.11)	0.23 (0.16, 0.32)	0.21 (0.17, 0.24)	0.20 (0.19, 0.22)
Fig. 4 B-A	1.62 (0.95, 2.05)	1.77 (1.54, 1.96)	1.82 (1.72, 1.91)	0.02 (0.00, 0.04)	0.02 (0.01, 0.03)	0.02 (0.02, 0.03)
Fig. 4 C-A	1.12 (0.83, 1.40)	1.11 (0.97, 1.25)	1.11 (1.04, 1.18)	0.12 (0.09, 0.15)	0.12 (0.10, 0.13)	0.12 (0.11, 0.12)
Fig. 4 D-A	1.69 (1.18, 2.07)	1.80 (1.58, 1.97)	1.86 (1.76, 1.94)	0.02 (0.01, 0.04)	0.02 (0.01, 0.03)	0.02 (0.02, 0.03)
Fig. 4 A-B	3.82 (3.53, 4.11)	3.75 (3.61, 3.90)	3.73 (3.66, 3.80)	0.18 (0.09, 0.28)	0.13 (0.11, 0.16)	0.12 (0.11, 0.14)
Fig. 4 B-B	2.98 (2.61, 3.35)	2.99 (2.80, 3.18)	3.00 (2.91, 3.10)	0.23 (0.14, 0.34)	0.20 (0.17, 0.24)	0.20 (0.18, 0.22)
Fig. 4 C-B	2.98 (2.72, 3.24)	2.93 (2.80, 3.06)	2.91 (2.85, 2.98)	0.11 (0.08, 0.14)	0.10 (0.09, 0.12)	0.10 (0.10, 0.11)
Fig. 4 D-B	2.83 (2.28, 3.38)	2.71 (2.42, 3.00)	2.73 (2.59, 2.87)	0.11 (0.04, 0.20)	0.08 (0.05, 0.10)	0.08 (0.07, 0.09)
Fig. 6 IIM	3.40 (2.38, 4.36)	2.93 (2.42, 3.43)	2.83 (2.58, 3.08)	0.24 (0.04, 0.53)	0.10 (0.05, 0.16)	0.08 (0.06, 0.10)
Fig. 7	2.81 (2.41, 3.22)	2.80 (2.60, 3.01)	2.79 (2.68, 2.89)	0.23 (0.16, 0.31)	0.21 (0.18, 0.25)	0.21 (0.19, 0.22)
Fig. 8	3.12 (1.93, 4.07)	3.05 (2.42, 3.68)	2.98 (2.73, 3.23)	0.03 (0.01, 0.06)	0.03 (0.01, 0.04)	0.02 (0.02, 0.03)

Journal (2050), Vol. 00, No. 000, pp. 1–27
DOI: 10.1000/xxx/xxx000

Inference of gene flow between species under misspecified models

Jun Huang,[1,†] Yuttapong Thawornwattana (orcid: 0000-0003-2745-163X)[2,†], Tomáš Flouri (orcid: 0000-0002-8474-9507)[3], James Mallet (orcid: 0000-0002-3370-0367)[2], and Ziheng Yang (orcid: 0000-0003-3351-7981) [1,*]
[1]School of Biomedical Engineering, Capital Medical University, Beijing, 100069, P.R. China
[2]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA
[3]Department of Genetics, Evolution and Environment, University College London, London, WC1E 6BT, UK
[†]Those authors contributed equally to this work.

Received on xxxx, revised on xxxx, accepted on xxxx

Genomic sequence data provide a rich source of information about the history of species divergences and interspecific hybridization or introgression. Despite recent advances in genomics and statistical methods, it remains challenging to infer gene flow, and as a result, one may have to estimate introgression rates and times under misspecified models. Here we use mathematical analysis and computer simulation to examine estimation bias and issues of interpretation when the model of gene flow is misspecified in analysis of genomic datasets, for example, if introgression is assigned to the wrong lineages. In the case of two species, we establish a correspondence between the migration rate in the continuous migration model and the introgression probability in the introgression model. When gene flow occurs continuously through time but in the analysis is assumed to occur at a fixed time point, common evolutionary parameters such as species divergence times are surprisingly well estimated. However, the time of introgression tends to be estimated towards the recent end of the period of continuous gene flow. When introgression events are assigned incorrectly to the parental or daughter lineages, introgression times tend to collapse onto species divergence times, with introgression probabilities underestimated. Overall, our analyses suggest that simple introgression models can be used to extract useful information concerning species divergence times and gene flow even when the model is misspecified. However, for reliable inference of gene flow it is important to include multiple samples per species, in particular, from hybridizing species.

Gene flow | model misspecification | multispecies coalescent | introgression | BPP | species tree

Introduction

Hybridization can enhance variation in recipient species, and has long been recognized as an important process in plants that can stimulate the origin of new species (e.g., Anderson, 1949; Mallet, 2007). Analyses of genomic data in the past decade has highlighted the prevalence of introgression in animals as well, including butterflies (Martin *et al.*, 2013), birds (Ellegren *et al.*, 2012), and bears (Liu *et al.*, 2014; Kumar *et al.*, 2017). Introgression may involve either sister or non-sister species and may play an important role in ecological adaptation (Mallet *et al.*, 2016; Martin and Jiggins, 2017). Introgression can be a major contributor of genealogical variation across the genome and gene tree–species tree discordance, in addition to ancestral polymorphism or delayed coalescence (Maddison, 1997; Nichols, 2001).

There is a long history of studies in population genetics of models of population subdivision and migration (Wright, 1943; Malecot, 1948; Slatkin, 1987), and a number of methods have been developed to estimate the migration rate between populations (Bahlo and Griffiths, 2000; Beerli and Felsenstein, 1999, 2001). An important limitation of models of population subdivision, when applied to data from different species or subspecies, is that they do not account for the divergence history of the populations or species. Introducing a population/species phylogeny into models of population subdivision not only improves the realism of the model but also opens up opportunities for addressing a number of interesting questions in evolutionary biology, such as estimation of species divergence times and ancestral population sizes, delineating species boundaries, and estimating the direction, rate and timing of gene flow (Jiao *et al.*, 2021).

Two classes of models of gene flow have been developed that accommodate the phylogeny of the species, both of which are extensions of the multi-species coalescent (MSC) model (Rannala and Yang, 2003). The first is the MSC-with-migration model (MSC+M, or isolation-with-migration or IM model, Hey and Nielsen, 2004; Hey, 2010; Zhu and Yang, 2012; Dalquen *et al.*, 2017; Hey *et al.*, 2018), which assumes that two species exchange migrants at a certain rate over an extended time period. The rate of

*Correspondence: z.yang@ucl.ac.uk

missing: SFS-based ABC methods

gene flow is measured by the proportion of migrants (m) in the receiving population per generation or by the population migration rate, $M = Nm$, the expected number of immigrants per generation, where N is the (effective) population size of the receiving population. We note that the isolation-with-initial-migration (IIM) model of Hey and Nielsen (2004) is an instance of the IM model, which assumes that gene flow occurs after species divergence initially but stops after a period of time, when reproductive isolation has been fully established (see below). The second class of models of gene flow is the MSC-with-introgression (MSci) model (Flouri *et al.*, 2020), also known as multispecies network coalescent model (MSNC; Wen and Nakhleh, 2018; Zhang *et al.*, 2018), which assumes that gene flow occurs at fixed time points in the past. The rate of gene flow is measured by the introgression probability (φ or γ), which is the proportion of successful immigrants in the population at the time of introgression.

I find this abbreviation inappropriate

In the real world, introgressed alleles may be removed by natural selection because they are involved in hybrid incompatibility and are deleterious in the genetic background of the recipient population (Dobzhansky, 1937; Muller, 1942) or because they are linked to such loci. Thus the rate of gene flow (M in IM or φ in MSci), when used, reflect the long-term drift as well as hybridization (Roux and Jiggins, 2017). Such gene flow may be expected to be influenced by the presence of loci in the genomic region important in ecological adaptation as well as the local recombination rate (Schumer *et al.*, 2018; Martin *et al.*, 2019; Edelman *et al.*, 2019). The rate can also vary over time, depending on geological or ecological events that cause changes in the ecology and distribution of the species, altering the chances for two species to exchange genes. One can e of gene flow in which the rate varies across genomic regions. For the presen models are not yet available, and th fitting such parameter-rich models to g is unexplored. IM and MSci mode to date (Dalquen *et al.*, 2017; Hey Wen and Nakhleh, 2018; Zhang *et a* *et al.*, 2020) assume constant rates. idealized extremes and should be c first approximation when applied to analyze genomic sequence data.

Cite theory papers establishing the concepts, not a random selection of empirical papers applying these.

This statement applies to the summary methods discussed, but not globally to ABC. Also, dies the IM model not infer gene flow between sister taxa? Did I miss something?

Currently the most commonly used methods for inferring gene flow from multilocus sequence data are approximate methods (also known as summary or heuristic methods) based on summaries of the data that are not sufficient statistics (Fisher, 1922). They are computationally very efficient but do not make use of all the information in the data (see Degnan, 2018; Elworth *et al.*, 2019; Jiao *et al.*, 2021; Hibbins and Hahn, 2022 for recent reviews). The first class of such methods include those based on genome-wide site-pattern counts, such as the D-statistic (Green *et al.*, 2010; Durand *et al.*, 2011) and HyDe (Meng and Kubatko, 2009; Blischak *et al.*, 2018). For example, D is based on site-pattern counts for four species (three species plus an outgroup) (Green *et al.*, 2010; Durand *et al.*, 2011), and ignores information in genealogical variation across the genome (Lohse and Frantz, 2014; Shi and Yang, 2018). Both deep coalescent and gene flow create stochastic fluctuations in the genealogical history (gene tree topology and coalescent times) across the genome, with the probability distribution of the gene tree specified by the parameters in the multispecies coalescent model with gene flow. There is thus important information about those parameters in such genealogical variation, which is ignored by methods based on genome-wide site-pattern counts. See Zhu and Yang (20 of the dramatic information loss in the context of species tree estimation resulting from pooling sites across loci.

second step remains unstated

Another major class of approximate methods take the two-step approach of inferring gene trees for multiple loci and using these as input data (Yu *et al.*, 2012, 2014; Yu and Nakhleh, 2015; Solis-Lemus and Ane, 2016; Wen *et al.*, 2016). Note that the reconstructed gene tree for a locus is a summary of the sequence alignment at the locus so that the two-step approach is a summary method as well. This approach typically ignores phylogenetic errors in the gene trees. Errors in estimated branch lengths (coalescent times) in gene trees are known to cause serious problems both for estimation of the species tree (DeGiorgio and Degnan, 2014) and for inference of gene flow (Wen *et al.*, 2016). Thus gene-tree branch lengths are typically ignored in these methods (Degnan, 2018) leading to consider

but see extensions by Hahn and Hibbins

Most current approximate methods are unable to identify many parameters in the coalescent model with gene flow. For example the D statistic can be used to detect gene flow between non-sister species but does not attempt to infer the direction, timing, or strength of gene flow. Yet these parameters are important for characterizing the history of species divergence and interspecific gene flow. Also current summary methods cannot infer gene flow between sister species, or estimate the rates of gene flow that occurs in both directions. Yet in nature it is likely that gene flow does occur between sister species and that two species do exchange genes in both directions. See Zhu and Yang (2021) and Yang and Flouri (2022) for recent discussions of limitations of approximate methods in analyses of genomic sequence data under the MSC model with and without gene flow.

In this paper, we focus on exact or likelihood methods of inference under the MSC-with-migration (MSC-M or IM) or with-introgression (MSci) models

The logic is not sound here: the conclusion here does not necessarily follow from the arguments above. There is the imperfection of the inference methods, and there is the potential non-identifiability of models. The focus of this paper seems to be on the latter point, but much of the argumentation confusingly evolves around the former one.

using ments (Hey *et al.*, hang *et al.*, 2018; d methods integ underlying the s ient use of infor ters in both the gene tree topology and coalescent times while accommodating their uncertainties. They typically involve a heavy computational load. However, recent algorithmic improvements have made it possible to apply the MSci model to genome-scale datasets with >10,000 loci (Flouri *et al.*, 2020). Inferring introgression events or constructing an introgression model using genomic sequence data, however, remains a challenging task, even when a binary species tree is specified, onto which introgression events can be added (Thawornwattana *et al.*, 2022; Ji *et al.*, 2022). See Discussion for an overview of currently available methods for inferring gene flow on a species phylogeny. For these and many other reasons, the model of gene flow that we assume in our data analysis may often be incorrect. An important question is to what extent inference of gene flow, and in particular, estimation of the timing and rate of gene flow, can still be achieved when the model of gene flow is misspecified. The impact of model misspecification on estimation of other evolutionary parameters such as species divergence times is also of major concern.

Here we use mathematical analysis and computer simulation to probe estimation of the rate and time of introgression when the model is misspecified. While there are many ways in which the assumed model is wrong, we are particularly interested in a few types that are likely in real data analyses (Thawornwattana *et al.*, 2022; Finger *et al.*, 2022).

First, gene flow may be occurring continuously during a time period but an MSci model is fitted to the genomic data, which assumes that gene flow occurred at a particular time point (e.g., Wen and Nakhleh, 2018; Jiao *et al.*, 2020). We are here interested in whether species divergence times and ancestral population sizes are affected by the misspecification, and how the migration rate in the migration model (M) corresponds to the introgression probability in the MSci model (φ). The case of two species is analytically tractable. We calculate the limit of the maximum likelihood estimates (MLEs) of introgression probability and introgression time when the data size (the number of loci) approaches infinity when the data are generated under the IM model. We use computer simulation to verify and extend the analytical calculation.

Second, the introgression event may be assigned to a wrong branch on the species tree, for example, to a parental or daughter branch of the genuine introgression lineage. Alternatively, introgression may involve species that have since gone extinct or are not included in the data sample. The presence of such ghost species is known to mislead inference of the history of gene flow for the sampled species (Beerli, 2004; Tricou *et al.*, 2022). Thus we conducted simulation to examine the impact of unsampled species on the inference of gene flow. In all cases of incorrect branch specification, we use BPP to analyze multilocus sequence data simulated under the MSci model (Flouri *et al.*, 2018, 2020) to assess the impacts of model misspecification on estimation of model parameters (such as species divergence and introgression times, population sizes, and introgression probabilities). While BPP is our own Bayesian implementation of the MSci model applied to multilocus sequence data, the results in this paper should apply to all full likelihood methods.

This is confusing, because the data in these cases obviously are simulated not under the MSci model.

the software

Results

Correspondence between the IM and MSci models in the case of two species

Notation and definition of parameters

Following Jiao *et al.* (2020), we study the asymptotic behavior of Bayesian parameter estimation under the introgression (MSci) model when the data are generated under the migration (IM) model in the case of two species, with one sequence per species per locus (fig. 1). Note that here we focus on this simple case because it is analytically tractable; nevertheless, our Bayesian implementation in BPP (Flouri *et al.*, 2020) can accommodate an arbitrary number of species and an arbitrary number of sequences per species per locus. We assume an infinite number of loci, and the data at each locus consist of a pair of sequences (a, b) from the two species, with x differences at n sites. The coalescent time t for the locus is unknown and underlies the observed differences. Jiao *et al.* (2020) analyzed the IM model (fig. 1**a**) and assumed infinite sequence length ($n = \infty$) so that the true coalescent time between the two sequences (t) is known. Here we accommodate random fluctuations in the number of mutations due to finite sequence length and consider three variants of the migration model.

In the basic IM model, species A and B diverged at time τ_R and there has since been gene flow from A to B at the rate of $M_{AB} = M$ migrants per generation (fig. 1**a**). The IIM model assumes that migration occurred initially after species divergence but stopped at time $\tau_T > 0$ (Costa and Wilkinson-Herbots, 2017), and is represented by an IM model for three species including a ghost species (fig. 1**b**). Here the ghost does not necessarily represent a real species but a mathematical device for specifying the IIM model. We also consider a secondary contact (SC) model, in which two species initially had complete isolation but came into contact at a certain time point (τ_T) with ongoing gene flow at the rate of M ever since. This is similarly specified using a ghost species at time point τ_T (fig. 1**c**). The migration model involves three types of parameters: species divergence times

τ_T: stop of gene flow back in time

Confusing.

Suggest to change order of questions to align with apparent order of priority implied above in lines 184 to 190.

because a ghost population is not part of the IIM model. A reference supporting what is claimed would

(τ_R, τ_T), population sizes for extant and extinct species $(\theta_A, \theta_B, \theta_T, \theta_R)$, and the (population) migration rate M. The population size parameter for any species with (effective) population size N is defined as $\theta = 4N\mu$, where μ is the mutation rate per site per generation. We refer to a branch on the species tree by its daughter node so that branch RA is also branch A, with population size parameter θ_A. Both divergence times (τ) and population sizes (θ) are measured by the expected number of mutations per site.

Asymptotic theory

We first consider the IIM model (fig. 1**b**), where the IM model (fig. 1**a**) is a special case with $\tau_T = 0$. The backwards-in-time process of coalescent and migration in time interval (τ_T, τ_R) is described by a Markov chain with three states: AB, AA and A (Notohara, 1990). Here AB is the initial state, with two sequences in the sample, one in A and another in B; AA means both sequences are in A (in other words, sequence b is traced back into A); and A means one sequence in A (in other words, sequence b is traced back into A and has coalesced with sequence a). Note that in the Markov chain, time runs backwards, so th… to AA means migration of a sequ… the real world. With time measu… number of mutations per site, the… the Markov chain is (see Jiao *et* … 1990)

$$Q = \begin{array}{c|ccc} & AB & AA & A \\ \hline AB & -w & w & 0 \\ AA & 0 & -\frac{2}{\theta_A} & \frac{2}{\theta_A} \\ A & 0 & 0 & 0 \end{array} \qquad (1)$$

where $w = m_{AB}/\mu = 4M_{AB}/\theta_B$ is the mutation-scaled migration rate, and $\frac{2}{\theta_A}$ is the coalescent rate in population A, with one time unit being the expected time taken to accumulate one mutation per site. Q has eigenvalues $\lambda_1 = 0$, $\lambda_2 = -\frac{2}{\theta_A}$, and $\lambda_3 = -w$.

Let the transition probability matrix over time t be $P(t) = \{p_{ij}(t)\} = e^{Qt}$, where $p_{ij}(t)$ is the probability that the Markov chain will be in state j time t later given that it is in state i at time 0. This is

$$P(t) = \begin{bmatrix} e^{-wt} & \frac{\theta_A w}{2-\theta_A w}\left(e^{-wt} - e^{-\frac{2}{\theta_A}t}\right) & 1 - \frac{2e^{-wt} - \theta_A w e^{-\frac{2}{\theta_A}t}}{2-\theta_A w} \\ 0 & e^{-\frac{2}{\theta_A}t} & 1 - e^{-\frac{2}{\theta_A}t} \\ 0 & 0 & 1 \end{bmatrix}. \qquad (2)$$

The probability density of coalescent time t is thus

$$f_m(t) = \begin{cases} P_{AB,AA}(t-\tau_T)\frac{2}{\theta_A}, & \text{if } \tau_T < t < \tau_R, \\ [1 - P_{AB,A}(\tau_R - \tau_T)]\frac{2}{\theta_R}e^{-\frac{2}{\theta_R}(t-\tau_R)}, & \text{if } t > \tau_R \end{cases}$$

$$= \begin{cases} \frac{2w}{2-\theta_A w}\left[e^{-w(t-\tau_T)} - e^{-\frac{2}{\theta_A}(t-\tau_T)}\right], & \text{if } \tau_T < t < \\ \left[\frac{2}{2-\theta_A w}e^{-w(\tau_R-\tau_T)} - \frac{\theta_A w}{2-\theta_A w}e^{-\frac{2}{\theta_A}(\tau_R-\tau_T)}\right]\frac{2}{\theta_R}e & \\ & \text{if } t > \tau_R. \end{cases}$$

This is a function of $w = 4M_{AB}/\theta_B$ but not of M_{AB} and θ_B individually. The parameters specifying the density are thus $\Theta_m = (w, \theta_A, \theta_R, \tau_R, \tau_T)$.

Under the secondary-contact (SC) model (fig. 1**c**), the coalescent-with-migration process over the time interval $(0, \tau_T)$ is described by the Markov chain of eq. 1. Given the parameters Θ_m, the probability density of coalescent time t is

$$f_{sc}(t) = \begin{cases} P_{AB,AA}(t)\frac{2}{\theta_A}, & \text{if } 0 < t < \tau_T, \\ P_{AB,AA}(\tau_T)\frac{2}{\theta_A}e^{-\frac{2}{\theta_A}(t-\tau_T)}, & \text{if } \tau_T < t < \tau_R, \\ \left[P_{AB,AA}(\tau_T)\frac{2}{\theta_A}e^{-\frac{2}{\theta_A}(\tau_R-\tau_T)} + P_{AB,AB}(\tau_T)\right] & \\ \quad \times \frac{2}{\theta_R}e^{-\frac{2}{\theta_R}(t-\tau_R)}, & \text{if } t > \tau_R \end{cases}$$

$$= \begin{cases} \frac{w\theta_A}{2-w\theta_A}\left[e^{-wt} - e^{-\frac{2}{\theta_A}t}\right]\frac{2}{\theta_A}, & \text{if } 0 < t < \tau_T, \\ \frac{w\theta_A}{2-w\theta_A}\left[e^{-w\tau_T} - e^{-\frac{2}{\theta_A}\tau_T}\right]\frac{2}{\theta_A}e^{-\frac{2}{\theta_A}(t-\tau_T)}, & \text{if } \tau_T < t < \tau_R, \\ \left[\frac{w\theta_A}{2-w\theta_A}\left[e^{-w\tau_T} - e^{-\frac{2}{\theta_A}\tau_T}\right]e^{-\frac{2}{\theta_A}(\tau_R-\tau_T)} + e^{-w\tau_T}\right] & \\ \quad \times \frac{2}{\theta_R}e^{-\frac{2}{\theta_R}(t-\tau_R)}, & \text{if } t > \tau_R. \end{cases} \qquad (4)$$

Similarly under the MSci model, with parameters $\Theta_i = (\varphi, \theta_R, \theta_S, \tau_R, \tau_S)$ (fig. 1**d**), we have (Jiao *et al.*, 2020)

$$f_i(t) = \begin{cases} \varphi\frac{2}{\theta_S}e^{-\frac{2}{\theta_S}(t-\tau_S)}, & \text{if } \tau_S < t < \tau_R, \\ \left[\varphi e^{-\frac{2}{\theta_S}(\tau_R-\tau_S)} + (1-\varphi)\right]\frac{2}{\theta_R}e^{-\frac{2}{\theta_R}(t-\tau_R)}, & \text{if } t > \tau_R. \end{cases} \qquad (5)$$

Given the coalescent time t for a locus, the probability of observing x differences at n sites under the JC mutation model (Jukes and Cantor, 1969) is given by the binomial probability

$$f(x|t) = \left(\tfrac{3}{4} - \tfrac{3}{4}e^{-\frac{8}{3}t}\right)^x \cdot \left(\tfrac{1}{4} + \tfrac{3}{4}e^{-\frac{8}{3}t}\right)^{n-x}. \qquad (6)$$

The marginal probability of observing x differences at n sites, under both the migration (IM, IIM, SC) and introgression (MSci) models, is

$$f(x|\Theta) = \int_0^\infty f(x|t)f(t|\Theta)dt, \qquad (7)$$

where $f(t|\Theta)$ is given by eqs. 3, 4 or 5.

For analy… (eq. 7), we assume … instead of JC, and rep… Poisson approximation

$$f(x|t) = \frac{1}{x!}(2nt)^x e^{-2n}.$$

This is derived in Appendix A, as eq. A6 for the IM (with $\tau_T = 0$) and IIM (with $\tau_T > 0$) models, eq. A7 for the SC model, and eq. A9 for the MSci model.

Consider the analysis of the data under the MSci model, which are generated under any of the migration

models (IM, IIM, SC). When the number of loci $L \to \infty$, the MLE $\hat{\Theta}_i$ under MSci will converge to Θ_i^*, which minimizes the Kullback-Leibler (KL) divergence

$$D(\Theta_m \parallel \Theta_i) = \sum_{x=0}^{n} f_m(x|\Theta_m) \log \frac{f_m(x|\Theta_m)}{f_i(x|\Theta_i)}, \qquad (9)$$

which is a measure of distance from the fitting introgression model to the true migration model. Here Θ_m are fixed while Θ_i are being estimated. The limiting values Θ_i^* as $L \to \infty$ are also known as the *pseudo-true parameter values* for the misspecified MSci model. The BFGS optimization routine in PAML (Yang, 2007) is used to minimize eq. 9 to obtain the MLEs.

We are in particular interested in the introgression probability φ and the introgression time τ_S. Note that under the migration model, the probability that any lineage from species B traces back to A is

$$\varphi_0 = 1 - e^{-\frac{4M}{\theta_B}\Delta\tau} = 1 - e^{-\frac{4M}{\theta_B}(\tau_R - \tau_T)}, \qquad (10)$$

where $\Delta\tau$ is the time period of gene flow (fig. 1**a**-c) and where we write M_{AB} as M. Eq. 10 gives the expected proportion of migrants under the true migration model. When M is small, $\varphi_0 \approx \frac{4M}{\theta_B}\Delta\tau$, which is also given by equating the expected total number of migrants under the two models: $N_B\varphi_0 \approx m_{AB}N_B\Delta\tau/\mu$. Note that $m_{AB}N_B$ is the expected number of migrants per generation and $\Delta\tau/\mu$ is the number of generations with gene flow.

It may be noted that the theory of eq. 9 can be used to study the limiting parameter e… number of loci $L \to \infty$) in the mi… the true model is the introgression … to flip the roles of $f_m(x|\Theta_m)$ and f_i… is not pursued in this paper.

Asymptotic results under the IM model

We used the asymptotic theory (eq. 9) to obtain the MLEs (Θ_i^*) under the MSci model (fig. 1**d**) when the data consist of an infinite number of loci, with one sequence of length n per species per locus, generated under the IM, IIM or SC models (fig. 1**a**-**c**). The parameter values used (Θ_m) are shown in figure 1. The MLEs are shown in figure 2 and the true and best-fitting distributions of the coalescent time t are shown in figure S1 for the IM model. The corresponding results for the IIM and SC models are in figures S2-S5, to be discussed in the next sections.

We use five methods (a-e) to fit the MSci model, with method d estimating all five parameters, while the others have some parameters fixed (fig. 2). We examined the effects of the sequence length (n) and the migration rate (M). When the data are analyzed under the MSci model, five parameters are identifiable: $\Theta_i = (\tau_R, \tau_S, \theta_R, \theta_S, \varphi)$ (fig. 1**d**). Population sizes θ_A, θ_B, and θ_H are not identifiable because no coalescent events can occur in those populations given one sequence per …pulation size θ_S is identifiable as … sequences a and b to be traced … Nevertheless, one expects the information concerning θ_S to be very weak in datasets of two sequences per locus, especially at low migration rates. In methods c and d, θ_S and φ are estimated as free parameters. The application of the misspecified MSci model (to data generated under the IM model) led to unreasonably large estimates of θ_S (as large as 0.5 mutations per site), and the poor estimates of θ_S caused φ to be poorly estimated as well. This is partly because of our use of only one sequence per species per locus in the analytical theory. Thus we do not emphasize the analyses using methods c and d, and focus instead on methods a, b, and e, in which θ_S is fixed (at the true value θ_0 in methods a and b, or constrained to be equal to θ_R in method e). In the simulation below, we evaluate the impact of the number of sequences sampled per species.

In the IM model, migration events occur throughout the time interval $(0, \tau_R)$, at the rate of M migrants per generation (fig. 1). When such data are analyzed under the introgression model, a simple expectation might be that the introgression time τ_S should be the average $\tau_R/2$ while the introgression probability may be given by the expected proportion φ_0 of eq. 10. However, as we show below, this expectation is too simplistic. We discuss the introgression time τ_S first.

Consider the case where the true coalescent time is known (or $n = \infty$). Given the data-generating IM model, there is a strictly positive probability for the coalescent time t to be in the interval $0 < t < \varepsilon$ for any small constant $\varepsilon > 0$. In other words, there must exist loci at which the coalescent time t is arbitrarily close to 0 (fig. S1). In the MSci model, sequences a and b cannot coalesce until they are in the same population S, so that $\tau_S < t$. When the MSci model is fitted to data generated under the IM model, $\hat{\tau}_S$ is dominated by the minimum rather than the average coalescent time, and $\hat{\tau}_S \to \tau_S^* = 0$ when the number of loci $L \to \infty$ (and when the true coalescent time is known). Even though migration events occur throughout the time interval $(0, \tau_R)$, the best the MSci model can do is to lump all migration events to one time point at $\tau_S^* = 0$ (fig. 2b&e).

When the sequence length is finite ($n < \infty$), the coalescent time is not observed and is reflected in the sequence divergence or the number of mutations (x). Whatever the true coalescent time, there is a positive probability of observing no mutations between the two sequences, so that $x = 0$ may not be very strong evidence that the coalescent time is $t \approx 0$. As a result, the MLE $\hat{\tau}_S$ reflects not only the minimum coalescent time, but also the whole distribution (fig. S1). In other words, two identical sequences between species may be 'interpreted' by the MSci model as being due to random mutational fluctuations with a strictly positive coalescent time. We thus have $\tau_S^* > 0$, different from

I agree, but there is potentially a more substantial issue with Eq. (10).

the case where the coalescent time is known without error ($n = \infty$). Nevertheless, one expects τ_S^* to be closer to 0 than to τ_R, especially if the number of sites is large. Indeed in our calculations, $\tau_S^* \ll \frac{1}{2}\tau_R$ (fig. 2b&e).

Next we consider the introgression probability φ and again focus on methods a, b, & e (fig. 2). The estimate $\hat{\varphi}$ increases nearly linearly when M is small ($< \frac{1}{4}$, say) but tails off at large M. All estimates are smaller than φ_0 of eq. 10 but they are close at low migration rates (with $M < \frac{1}{4}$ and $\varphi < \frac{1}{4}$, say) (fig. 2a,b&e). We defer to a later section a detailed discussion of the estimation of φ, contrasting the IM, IIM, and SC models.

Finally the estimated divergence time between the two species τ_R matched the true values at low migration rates but was underestimated at high migration rates, with the ancestral population size θ_R overestimated (fig. 2). It may be tempting to interpret the underestimation of τ_R (and overestimation of θ_R) by the MSci model as being due to the difficulty of distinguishing complete isolation with recent species divergence from introgression or of distinguishing migration and coalescent events close to species divergence from ancestral polymorphism. However, this does not appear to be a correct interpretation.

We examined the true and fitted distributions of the coalescent time (fig. S1). When there is no migration ($M = 0$), the MSci model is correct, and the parameter estimates converge to the true values, with a perfect fit to the density $f_m(t)$. At low migration rates ($M \leq 0.1$, say), the MSci model fits the density $f_m(t)$ very well, with the discontinuity points in the true and fitting distributions coinciding; in other words, $\tau_R^* = \tau_R^m$. At the medium migration rate of $M = 1$, the species divergence time τ_R is still correctly estimated even though the fit to the density is poor (fig. S1). At high migration rates (with $M \geq 1.4$, say), the true density has a mode in the interval $(0, \tau_R^m)$, dropping off at the discontinuity point at τ_R^m. The best fitting density starts from 0, with an exponential decay, and has a discontinuity point at τ_R with again an exponential decay. This best-fitting density is a poor fit, and the discontinuity point τ_R^* is moved to smaller values as an attempt to accommodate the migration and coalescent events in the middle of the interval $(0, \tau_R^m)$ to improve the fit (judged by the KL divergence). Thus τ_R is underestimated ($\tau_R^* < \tau_R^m$). As a result, the population size parameter θ_R is overestimated, as those two parameters tend to be strongly negatively correlated (e.g., Burgess and Yang, 2008). In other words, the intermediate coalescent times in the interval $(0, \tau_R)$, which occur at a large proportion of loci, are accommodated or misinterpreted by the MSci model by using a more recent species divergence time (τ_R) and a larger ancestral population size (θ_R). Coalescent times in the range $\tau_R^* < t < \tau_R^m$, which represent true migration events in the IM model, are then misinterpreted as coalescent events in the ancestral population R in the MSci model, so that the estimated introgression probability φ^* is substantially lower than the expected proportion φ_0 (eq. 10).

Asymptotic results under the IIM model

When data are generated under the IIM model (fig. 1**b**) and analyzed under the MSci model (fig. 1**d**), the results (figs. S2&S3) show similar patterns to those under the IM model discussed above. Similarly, θ_S is difficult to estimate using two sequences per locus in methods c and d, and the poor estimates of θ_S affects the estimation of the introgression probability φ. Thus we focus on methods a, b, and e, in which θ_S is fixed or constrained, and on the introgression time and introgression probability.

In the IIM model, migration events occur throughout the time interval (τ_T, τ_R) (fig. 1**b**), but the estimate of the introgression time is dominated or influenced by the minimum coalescent time, so that $\tau_S^* = \tau_T$ when the coalescent time is known (with $n = \infty$), and $\tau_S^* > \tau_T$ when n is finite. In the latter case τ_S^* was closer to τ_T than to τ_R (fig. S2).

The introgression probability φ^* grew almost linearly with the migration rate M when M was small (with $M \leq 0.2$, say), and this estimate was close to the expectation φ_0 of eq. 10 (fig. S2a, b&e). At high migration rates, eq. 10 gave a serious overestimate. This 'bias' in φ at high migration rates was accompanied by a reduction in the species divergence time (τ_R) and overestimation of the ancestral population size (θ_R). This can similarly be explained by the attempt of the MSci model to accommodate the coalescent times in the middle of the time interval (τ_T, τ_R) (fig. S3).

Asymptotic results under the SC model

Under the SC model, there was initially complete isolation after species divergence but the two species came into contact at time τ_T, with ongoing gene flow ever since (fig. 1**c**). When data of an infinite number of loci, each with two sequences, are analyzed under the MSci model, the MLEs are shown in figure S4, with fitted densities of coalescent time t shown in figure S5.

The results show patterns similar to those under the IM and IIM models discussed above. The species divergence time under the MSci model $\tau_R^* = \tau_R^{(sc)}$ when the migration rate M is small but drops at very high rates (with $M > 2$). The introgression time is dominated by the minimum coalescent time, so that $\tau_S^* = 0$ when $n = \infty$, and τ_S^* is much closer to 0 than to τ_R when n is finite (fig. S4). Note that in the true model migration occurs throughout the time interval $(0, \tau_T)$.

The introgression probability φ^* grows almost linearly with the migration rate M when M is small (with $M \leq \frac{1}{4}$, say), and is close to the expectation φ_0 (eq. 10) when $M < 2$ (fig. S4a,b&e). At very high migration rates ($M > 2$), φ^* is much smaller than φ_0,

Is this a finding of this study or known from before? Please clarify. I think what is meant here is that thetas A, B, and H now are identifiable.

Read about BPP - between-pop only?

and this 'bias' is accompanied by an und… of τ_R and overestimation of θ_R. Similarly to the IM and IIM models discussed above, this is due to the attempt of the MSci model to accommodate the coalescent times in the middle of the interval $(0, \tau_T)$ (fig. S5).

The amount of gene flow under the IM, IIM, and SC models

While the expected total amount of gene flow, measured by φ_0 (eq. 10), is the same under the IM, IIM, and SC models of figure 1**a**-**d**, the estimates under the MSci model differ, as summarized in figure 1**e**.

At low migration rates, τ_R, θ_S and θ_R in the MSci model are accurately estimated to match those in the true model (figs. 2, S2 & S4). Consider the case of infinitely long sequences with known coalescent time. Let $\tau_R^* = \tau_R^m$, $\theta_R^* = \theta_R^m$, and let the introgression time be $\tau_S^* = 0$ for the IM and SC model, and $\tau_S^* = \tau_T$ for the IIM model. We also match the probability density of coalescent time t for $t > \tau_R$, so that $f_i(t) = f_m(t)$. With those simplifying assumptions, φ^* that minimizes the KL divergence (eq. 9) can be derived as

Reproduce this step.

$$\varphi^*_{(\mathrm{IM})} \approx \frac{\varphi_0 - \frac{w\theta_A}{2}\left(1 - e^{-\frac{2}{\theta_A}\tau_R}\right)}{\left(1 - \frac{w\theta_A}{2}\right)\left(1 - e^{-\frac{2}{\theta_A}\tau_R}\right)},$$

$$\varphi^*_{(\mathrm{IIM})} \approx \frac{\varphi_0 - \frac{w\theta_A}{2}\left(1 - e^{-\frac{2}{\theta_A}(\tau_R - \tau_T)}\right)}{\left(1 - \frac{w\theta_A}{2}\right)\left(1 - e^{-\frac{2}{\theta_A}(\tau_R - \tau_T)}\right)}, \quad (11)$$

$$\varphi^*_{(\mathrm{SC})} \approx \frac{\varphi_0 - \frac{w\theta_A}{2-w\theta_A}\left(e^{-w\tau_T} - e^{-\frac{2}{\theta_A}\tau_T}\right)e^{-\frac{2}{\theta_A}(\tau_R - \tau_T)}}{1 - e^{-\frac{2}{\theta_A}\tau_R}},$$

for the IM, IIM, and SC models, respectively. At low migration rates, eq. 11 provides accurate numerical results (methods a, b, e in figs. 2,S2&S4). From eq. 11, we have

$$\varphi_0 > \varphi^*_{(\mathrm{SC})} > \varphi^*_{(\mathrm{IIM})} = \varphi^*_{(\mathrm{IM})}. \quad (12)$$

In other words, recent gene flow (as in SC) is easier to recover by the MSci model than ancient gene flow (as in IM or IIM). Note that $\varphi^*_{(\mathrm{IIM})} = \varphi^*_{(\mathrm{IM})}$ holds only when one sequence is sampled per species; as there is no coalescent over $(0, \tau_T)$, IIM is essentially the same model as IM with a time shift (fig. 1). This will not be the case when multiple sequences per species are sampled or when the sequence length is finite.

Simulation results

As our asymptotic theory was limited to a single sequence per species per locus, we used simulation to verify and augment our analytical calculations above. We simulated data under the IM, IIM or SC models of figure 1**a**-**c**, using the same parameter values as above, and analyzed them using BPP under the MSci model (fig. 1**d**). The JC mutation model (Jukes and Cantor, 1969) was assumed. In the basic setting we used $S = 4$ sequences per species per locus, each of $n =$ …s, with each dataset consisting of $L = 4000$ loci. We varied the number of sequences per species (S), the number of sites per sequence (n), the number of loci (L), and the migration rate (M) to examine their effects on parameter estimation. With multiple sequences per species, all eight parameters of the MSci model (fig. 1**d**) are identifiable.

We first note a few common features in the results (fig. 3). In nearly all cases, population sizes for extant species (θ_A, θ_B) were very well estimated, with posterior means close to the true values and with very narrow highest-probability-density (HPD) credibility intervals (CIs). The exception was parameter θ_B under the IM model (note that B is the species receiving immigrants), which was less well estimated when the dataset was small and had either short sequences ($n = 250$) or few loci ($L \leq 500$), or when the migration rate was very high. The poorer estimation of θ_B appeared to be related to the underestimation of φ and τ_R; see below. The population size for the common ancestor θ_R was mostly well estimated, although overestimated at very high migration rates. Population sizes for the ancestral species (θ_S, θ_H) are harder to estimate; indeed they had larger CIs and were influenced by misspecification of the model of gene flow. As expected from asymptotic results, the age of the root of the species tree (τ_R) was very well estimated, except at very high migration rates, when τ_R was underestimated (and θ_R overestimated).

We next examine the effects of various factors: the sequence length (n), the number of sequences per species (S), the number of loci (L), and the migration rate (M). First, the number of sites (n) had a relatively small impact on MSci parameters, when other factors were fixed (at the basic setting of $S = 4, L = 4000$, and $M = 0.2$). While $n = 250$ was small and led to large CIs for parameters such as the introgression time and probability ($\tau_S = \tau_H$ and φ), the CIs were small for all parameters when $n \geq 1000$. The introgression time τ_S decreased slightly as the sequence became longer. This is consistent with the asymptotic analysis, which suggests that τ_S^* is dominated by the smallest coalescent time or sequence divergence and should converge to 0 for the IM and SC models and to τ_T for the IIM model when $n \to \infty$ (figs. 2, S2&S4). Similarly, for the IM and SC models, $\hat{\varphi}$ increased with the increase of n when M was low, as observed in the asymptotic analysis. Under the IIM model, small datasets with short sequences ($n = 250$) produced very uncertain estimates of φ and θ_H (and, to a lesser extent, τ_S and θ_S); we discuss this effect below when we examine the impact of the number of sequences (S).

Second, we varied the number of sequences per species (S). All parameters were well estimated when multiple sequences were sampled per species ($S \geq 2$ for IM and SC or $S \geq 4$ for IIM). When only one sequence per species is in the data ($S = 1$), only five parameters $(\theta_R, \theta_S, \tau_R, \tau_S, \varphi)$ are identifiable. In the asymptotic

... had an impact on ...
or
... affected ...

point estimates

I missed why a model without gene flow is (nearly) the correct model given the parameters in the caption of Fig. 1. Please clarify.

I could not find results for the SC model in Fig. 3.

analysis assuming infinitely many loci, we noted that θ_S was estimated with large errors, and that the poor estimation of θ_S impacted on the estimation of φ; the parameters were grossly wrong but had no sampling errors because the data size was $L = \infty$ (figs. 2c,d, S2c,d and S4c,d). Why did $\hat{\varphi}$ not converge to 0 with narrow CIs, since the MSC model with no gene flow is nearly the correct model? We interpret this result as being due to the r
and φ in the MS
concerning θ_S an
even with infinite
simple MSC with
$\varphi = 0$ will provi
with a large θ_S, a
very good fit as v
surface involving φ and θ_S, leading to w
thos ter estimates i
and on here, assur
num nfirms that int
Unl lyses in where
no u parameters ha
larg y model miss
and with $S = 1$).
uncertainties in parameter estimates may be
a strength of the Bayesian analysis, as they help the investigator to avoid making incorrect inferences of a large φ when gene flow is minimal. We note that even with $S = 4$ sequences per species, estimates of φ from data generated under the IIM model involved wide CIs, with τ_S being close to τ_R, and θ_S and θ_H being very imprecise as well (fig. 3). Nevertheless, this problem of semi-unidentifiability disappeared and all parameter estimates were well-behaved in large datasets when many sequences were sampled ($S \geq 2$ for IM and SC or $S \geq 4$ for IIM; fig. 3).

Wording: the simulation results shown are compatible with the existence/interpretation of a ridge (in the sense of a necessary condition), but they do not *confirm* anything the sense of a sufficient

I do not think that the ability to reflect the relevant uncertainty about the model is exclusive to the Bayesian approach. I actually think that the additional uncertainty about the appropriateness if the prior confounds the focal uncertainty, which I consider a disadvantage.

Could this be improved by including information from lineages sampled within populations? Is this perhaps already done in PPB?

Third, we examined the impact of the number of loci (L). The IIM model was hard to fit in small datasets with a small number of loci ($L \leq 1000$), generating large CIs for parameters φ and θ_H. This is the same pattern as in the case of short loci ($n = 250$) or few sequences ($S \leq 2$), discussed above. In other cases the parameters were well estimated. Note that the number of loci L is the sample size in the statistical model as data at different loci are independently and identically distributed. Theory predicts that in large datasets the variance should be proportional to $1/L$ (see O'Hagan and Forster, 2004 for the case of correctly specified models and Yang and Zhu, 2018 for the case of misspecified models), and thus the CI should decrease at the rate of $L^{-\frac{1}{2}}$. This prediction held for parameters that were well estimated (fig. 3). As discussed earlier, the introgression time τ_S is dominated by the smallest coalescent time or smallest sequence divergence. Thus increasing the number of loci led to a decrease in the estimated introgression time, and the trend was in particular apparent for the IIM model (under which $\hat{\tau}_S \to \tau_T$ when $L \to \infty$ if $n = \infty$). In all cases, the estimated introgression time ($\hat{\tau}_S$) was closer to the more recent end of the time interval for gene flow than to the midpoint (i.e., $\hat{\tau}_S < \frac{\tau_R}{2}$ for IM, $\hat{\tau}_S < \frac{\tau_R+\tau_T}{2}$ for IIM, and $\hat{\tau}_S < \frac{\tau_T}{2}$ for SC; see fig. 1**a**-**c**).

Finally, we evaluated the impact of the migration rate (M) (fig. 3). At low migration rates (M), there is a near linear relationship between the introgression probability φ and M. In general, the amount of gene flow estimated under the MSci model is less than the true amount expected under the migration model (φ_0 of eq. 10) but the two were close at low migration rates (with $M < 0.1$, say). At very high migration 1.0, say), divergence time τ_R was restimated and the population size ated. These are the same patterns asymptotic analysis of infinite data due to the attempt of the MSci odate intermediate coalescent times rue migration model, as discussed 2, S2, S4). At low migration rates say), the IIM model produced very s of φ, with τ_S, θ_S, and θ_H affected same pattern as observed for small datasets with short loci ($n \leq 250$), few sequences ($S \leq 2$), or few loci ($L \leq 1000$), discussed above.

In summary, our asymptotic analysis and the computer simulations sug espondence between the IM a gene flow occurs continuous ime period (τ_T, τ_R) after divergence of two species and we fit the introgression (MSci) model, the estimated introgression time tends to be closer to the more recent end of the time period of gene flow, because the introgression time in the MSci model is dominated by the most recent coalescent time or the minimum sequence divergence between species. Indeed if the true coalescent time is known and used as data, the introgression time will converge to the time when gene flow stopped, as discussed above in the asymptotic analysis. At low migration rates ($M < \frac{1}{4}$, say), the species divergence time is correctly estimated by the MSci model, and the introgression probability φ is lower than but close to the expected proportion of migrants ($\varphi^* < \varphi_0$). The particularly close under the SC model (fig. migration rates, the estimated introgressio φ^* may be much less than φ_0. This is species divergence time τ_R is underestin MSci model to account for intermediat times generated under the IM model, underestimation of the introgression probability φ and overestimation of the ancestral population size θ_R.

I think this should say 'estimator of...'.

I think this is only part of the explanation, as I suspect that the analytical expression in Eq. (10) is wrong

Introgression events assigned to wrong branches

We conducted simulations to examine the bias in parameter estimates when the introgression event is

assigned on either the parental or daughter branch of the lineage genuinely involved in introgression. The data were simulated under model trees A or B and analyzed under models A or B (fig. 4**a**,**b**).

In the A-A and B-B settings (fig. 4**e**), the correct MSci model was assumed, and the performance of the method serves as a reference for comparison. Most parameters, including the species divergence times (τ_R, τ_S, τ_T, and $\tau_X = \tau_Y$) and population sizes for extant species ($\theta_A, \theta_B, \theta_C, \theta_D$), were well estimated. For well-estimated parameters, the CI width halves as the number of theory. Population θ_T, θ_X, and θ_Y) performance impr 4000 ere well estimated. Introgr not verify this thousa ent. Performance estima in the B-B setting seemed used here (4 sequences p superior. This comment sites per sequence). applies to here and other places in the manuscript.

In the A-B setting (fig. l under model A with $A \rightarrow$ B introgression (fig. 4**a**) but analyzed under model B with introgression incorrectly assigned to the parental branch *ST*. Species divergence times (τ_R, τ_S, τ_T, and $\tau_X = \tau_Y$) and population sizes for extant species (θ_A–θ_D) were all well estimated, similar to the B-B setting in which the simulation model matched the analysis model. Population sizes for ancestral species were hard to estimate, and performance was similar to that under the B-B setting. We expect τ_T in model B t rmined by the smallest sequence dive species *B* and *C*, which should be c $= 0.004$. As introgression events in e effect of pushing τ_T down, τ_T had a negative bias but the bias was very small and τ_T was well estimated. The introgression time $\tau_X > \tau_T$ in model B, and as the true introgression time τ_X was smaller than τ_T, τ_X was stuck at τ_T (fig. 5**a**). There was virtually no information for θ_T as the population was estimated to have near-zero time duration with no chance for coalescent events to occur in the population. The introgression probability was seriously underestimated, converging to $\varphi^*_{\text{A-B}} \approx 0.12$ when the number of loci L increases (table 1) whereas the true value was 0.2. This smaller estimate of introgression probability is explained by the distribution of coalescent times between species in the true and fitting models (fig. S6, true model A). Under the true model A, sequences from *A* and *B* are more similar than those between *A* and *C* due to the $A \rightarrow B$ introgression, with an excess of small coalescence time t_{ab}. Under the analysis model B, t_{ab} and t_{ac} have the same distribution. Thus the true model predicts an excess of small t_{ab} in the data whereas the fitting model predicts an excess of small t_{ac}, and having a smaller φ in the fitting model helps to reduce the discrepancy.

In the B-A setting (fig. 4**e**), the simulation model (MSci model B of fig. 4**b**) assumed introgression involving the ancestral branch *ST* but the analysis model (model A) assigned introgression to the daughter branch *TB*. Again posterior means and CIs for most parameters, including species divergence times (τ_R and τ_T) and population sizes (θ_A–θ_D, and θ_R), were similar to those in the A-A setting where there was model match. Note that τ_T in the analysis model A should be mostly determined by the smallest sequence divergence between *B* and *C*, and given that this was $\tau_T = 2\theta_0 = 0.002$ in the simulation model A, τ_T was well estimated, unaffected by mis-assigned introgression event. Although the true introgression time τ_X was 0.003, it was forced to be less than τ_T by the analysis model A. As the number of loci increases, τ_X became stuck at τ_T (fig. 5**b**). However, τ_S was seriously underestimated. This may be explained as follows. In the analysis model A, τ_S was mostly determined by the shortest sequence distance between *A* and *C*. According to the simulation model B, this should be close to $\tau_X^{(B)} = 1.5\theta_0 = 0.003$, due to introgression events. Here we use the superscript to indicate that the parameter is for model B (fig. 4**b**). With mutational fluctuations in the sequences, one can expect the τ_S estimate in the B-A setting to lie between $(\tau_X^{(B)}, \tau_S^{(B)}) = (1.5\theta_0, 3\theta_0)$, but closer to $\tau_X^{(B)}$ in large datasets with many sites and/or many loci. Population size parameters θ_S and θ_Y were affected by the mis-assigned introgression events as well, as those populations are close to the introgression branches. In particular, θ_Y was very imprecise as branch *TY* was very short, and θ_S was overestimated because τ_S was seriously underestimated and the two parameters are negatively correlated. Finally the introgression probability (φ) was underestimated, apparently converging to $\varphi^*_{\text{B-A}} \approx 0.02$ when the number of loci increases (table 1) whereas the true value was 0.2. This greatly reduced introgression probability appeared to reflect the very poor fit of the misspecified model A to data generated under model B, apparent due to large differences between the true and fitting distributions of coalescent times (t_{ab}, t_{ac}, t_{bc}; fig. S6, second row). As τ_X and τ_S are seriously underestimation by model A, an excess of small coalescent times (t_{ab}, t_{ac}) is expected in the fitting model A but does not appear in the data, so that having a smaller φ improves the fit.

In summary, assigning introgression events to a wrong parental or daughter branch led to biased estimates of introgression times (causing the introgression events to collapse onto speciation events) and to seriously underestimated introgression probabilities.

Simon: Find out if only between-population coalescences are used or also within-population coalescences.

Refer to the estimate of the parameter, not the parameter.

Refer to the estimate, not the parameter.

Do the authors mean tau_bc?

Continuous migration versus episodic introgression

In this set of simulations, we generated data under the IM models C and D of figure 4**c**&**d** and analyzed them under the MSci models A and B, with the mode of gene flow misspecified and with gene flow assigned to either the correct branch or a wrong branch on the species tree.

In the C-A and D-B settings (fig. 4**e**), gene flow occurred continuously but the data were analyzed under the introgression model assuming gene flow at a particular time point. The mode of gene flow was misspecified, but the lineages involved were correctly identified. In the C-A setting, gene flow was between nonsist tting it was between (τ_R, τ_S, τ_T) and pop ated. Indeed the resu e C-A setting were i r the A-A setting, where the MSci model was used to both generate and analyze data. Similarly the results for speciation times and population sizes were extremely similar for settings D-B and B-B. Those results were consistent with earlier simulation results based on two species (fig. 3), which showed that rates, species divergence times and were well estimated under the MSci model when the data were generated under the IM model (see also Tiley *et al.*, 2022).

In the C-A setting, the introgression time $\hat{\tau}_X$ appeared to converge, when the number of loci L increases, to 0.0011, which is much more recent than the average time of gene flow ($\tau_T/2 = 0.002$), and the introgression probability φ appeared to converge to $\varphi^*_{\text{C-A}} = 0.12$ (table 1), smaller than the proportion of total migrants given by eq. 10: $\varphi_0 = 1 - e^{-4M\tau_T^{(C)}/\theta_B^{(C)}} = 0.148$. As discussed earlier in the case of two species, the limiting value for τ_X was nonzero, as the sequence length is finite, and the MLE $\hat{\varphi}_{\text{C-A}}$ slightly underestimates the true amount of gene flow. In the D-B setting, the introgression time τ_X appeared to converge to 0.0027, which is larger than $\tau_T = 0.002$ but much smaller than the average time of gene flow, $\frac{1}{2}(\tau_S + \tau_T) = 0.004$, and the introgression probability φ appeared to converge to $\varphi^*_{\text{D-B}} = 0.08$ (table 1), much smaller than the true proportion under the IM model, $\varphi_0 = 0.148$ (eq. 10). In both the C-A and D-B settings, the estimated introgression time was within interval of gene flow, but closer to the time point when gene flow stopped, while the introgression probability underestimated the amount of gene flow that actually occurred, with $\varphi^*_{\text{C-A}} < \varphi_0$ and $\varphi^*_{\text{D-B}} < \varphi_0$. Moreover, we have $\varphi^*_{\text{D-B}} < \varphi^*_{\text{C-A}}$. These patterns are consistent with our analysis of the two-species case at low migration rates (eq. 12, fig. 3), which predicts that gene flow after a period of isolation (the SC model) is easier to recover under the MSci model than gene flow that starts at speciation but stops some time afterwards (the IIM model).

In the C-B and D-A settings (fig. 4**e**), the mode of gene flow was misspecified and furthermore gene flow was assigned onto the wrong branch of the species tree. In the C-B setting, species divergence times τ_R and τ_S were well estimated, just as in the B-B setting. Divergence time τ_T was underestimated slightly, due to gene flow assigned to the wrong branch, as observed in the A-B setting. Population sizes for extant species (θ_A–θ_D) were all well estimated, as in the B-B setting. Ancestral population sizes θ_R and θ_S were as well estimated as in the B-B setting, and so was θ_X. Ancestral population sizes θ_T and θ_Y were affected by gene flow, similar to the A-B setting. Model B forces $\tau_X > \tau_T$. Thus we expect estimates of τ_X and τ_T to get stuck together, with both to be smaller than $\tau_T^{(C)} = 2\theta_0 = 0.004$; as the number of loci L increases, τ_X appears to converge to 0.0029, and φ to $\varphi^*_{\text{C-B}} = 0.10$ (table 1).

In the D-A setting, species divergen and τ_T were well estimated, as in the Divergence time τ_S was underestimated, flow assigned to the wrong branch, sim B-A setting. Population sizes for extant θ_D) were all well estimated, as in the Ancestral population sizes θ_R and θ_X w estimated as in the A-A setting, and θ_T positive bias. Ancestral population sizes were affected by the gene flow, similar to the B-A setting. Introgression time and probability (τ_X and φ) did not exist in the simulation model D. Model A forces $\tau_X < \tau_T$, so we expect $\hat{\tau}_X$ to be close to $\tau_T^{(D)} \approx \theta_0 = 0.002$; when the number of loci L increases, τ_X appeared to converge to 0.00186, and $\hat{\varphi}_{\text{D-A}}$ to $\varphi^*_{\text{D-A}} = 0.02$ (table 1). Note that $\varphi_0 > \hat{\varphi}_{\text{C-B}} > \hat{\varphi}_{\text{D-A}}$ with $\hat{\varphi}_{\text{C-B}} < \hat{\varphi}_{\text{C-A}}$ and $\hat{\varphi}_{\text{D-A}} < \hat{\varphi}_{\text{D-B}}$. Those results are c with our early results for fitting the MSci mod generated under the migration model in the two case (eq. 12, fig. 3), and with the results for the A-B and B-A settings that assignment of gene flow to a wrong branch reduces the estimate of φ.

In summary, the estimated introgression probabilities, at 0.12, 0.08, 0.10, and 0.02 for the C-A, D-B, C-B, and D-A settings, respectively, even though the total amount of gene flow was the same in models C and D (table 1), suggest the following general patterns. First, the introgression (MSci) model underestimates the total amount of gene flow if gene flow occurs continuously in every generation (i.e., $\hat{\varphi}_{\text{C-A}} < \varphi_0, \hat{\varphi}_{\text{D-B}} < \varphi_0$), as discussed in our analysis of the two-species case. Second, assigning gene-flow events to wrong lineages led to more serious underestimation of the strength of gene flow than the misspecification of the mode of gene flow alone (i.e., $\hat{\varphi}_{\text{C-B}} < \hat{\varphi}_{\text{C-A}}, \hat{\varphi}_{\text{D-A}} < \hat{\varphi}_{\text{D-B}}$). Third, recent gene flow in the data is more easily recovered (i.e., $\hat{\varphi}_{\text{C-A}} > \hat{\varphi}_{\text{D-B}}, \hat{\varphi}_{\text{C-B}} > \hat{\varphi}_{\text{D-A}}$).

I could not locate this parameter in Fig. 4 first because I was confused about the subscript index. Omit them!

No need to restate the formula.

estimated

suggested

If this approx. sign is correct or not. Whether it is depends on whether the authors mean the parameter or estimate

No need to reintroduce the abbreviation

Please motivate this analysis in a topic sentence. What point/aspect is being investigated?

Isolation with initial migration (IIM) model

We used the IIM model of figure 6**a** to simulate data and analyzed them under the MSci model of figure 6**b**. Species divergence times (τ_R and τ_S) and population sizes (θ_A, θ_B, θ_C, θ_R, θ_S, and even θ_X and θ_Y) were well estimated. We expect the introgression time τ_X to converge to $\tau_T = \theta_0 = 0.002$ if the sequence length is infinite and to a higher limit for finite sequence length. In our simulation $\tau_X \approx 0.00283$ at $L = 4000$ (table 1). Introgression probability φ converged to a nonzero limit, ~ 0.08 (table 1), compared with $\varphi_0 = 0.148$ by eq. 10.

The case of figure 6 is very similar to the two-species case of figure 1 except that the species tree is larger with more species, and serves to highlight the fact that the impact of the misspecification of t flow is local. The case is also very s B setting of figure 4, with the only here the hybridizing species T had only one descendent species sampled in the data whereas in figure 4 (D-B) it had two descendent species sampled. In the Bayesian analysis, this difference affects only the information content in the data. Thus estimates of parameters such as the introgression probability and introgression time were similar to those in the D-B setting of figure 4 but with wider CIs (table 1).

I do not understand what the authors aim to say here.

I suggest to stick to conventional panel labels.

Ghost species

We considered two scenarios in which a species that contributed migrants to extant species has gone extinct or is otherwise unsampled in the data. ~~Note that existence of extinct or unsampled species that *received* genetic materials from ancestors of extant species in the sample is not relevant to the analysis of the sampled data and does not constitute a model misspecification~~. In the first scenario, model A′ of figure 7**a**′ is used to simulate data, which assumes that species XUV contributed migrants to species B but is not included in the sample. Note that this model is equivalent to model A of figure 7**a**. When we fit model B (fig. 7**b**), the only incorrect assumption is the constraint that $\tau_X = \tau_Y$. This is a minor misspecification. Indeed all parameters shared between the simulation model and the analysis model were well estimated (fig. 7**c**). The estimates of introgression time, $\tau_X = \tau_Y \approx 0.0028$ (table 1), were close to the average of the two parameters in the true model (0.0025). Introgression probability $\varphi \approx 0.21$ (table 1) was also close to the true value (0.2). The existence of the ghost species (XUV) had very little effect on the inference.

In the second scenario (fig. 8**a**), the true model assumes continuous migration involving intermediate ancestral species that have gone extinct, and the MSci model (fig. 8**b**) was fitted to data sampled from extant species. Divergence times τ_R and τ_T were very well estimated, as were the population sizes shared between the simulation and analysis models (θ_A, θ_B, θ_C, θ_R). We expect $\hat{\tau}_T$ in model B to be dominated by the minimum coalescent time t_{ab} between sequences from A and B, and this is given by τ_T in model A. Gene flow from branches RC to SU to TB during the time period τ_T–τ_U is interpreted as introgression in the MSci model. The effective rate for this migration may be close to $M_{CU}M_{UB} = 0.04$, giving the expected amount of introgression as $\varphi_0 = 1 - e^{-4\times M_{CU}M_{UB}\times(\tau_T-\tau_U)/\theta_B} = 0.031$. The estimate was $\hat{\varphi}_X \approx 0.02$ (fig. 8**c**, table 1). Introgression time τ_X should be between $(\tau_U, \tau_T) = (\theta_0, 2\theta_0) = (0.002,\ 0.004)$ and the estimate was ≈ 0.0030 (fig. 8**c**, table 1). Note that both θ_T and θ_Y were overestimated (fig. 8**c**). Branch T of figure 8**b** corresponds to branches RS and ST of figure 8**a**, which ave population size $\theta_0 = 0.002$. Branch Y corresponds o a segment of branch TB over the time interval τ_U–T, with population size $\theta_1 = 0.01$. Overestimation of θ_Y (and θ_T) may be because there is a shortage of t_{bb} over the time interval τ_U–τ_S in the data because of the gene flow, and the fitting MSci model, with the amount of gene flow underestimated ($\hat{\varphi} < \varphi_0$), used large $\hat{\theta}_Y$ and $\hat{\theta}_T$ to compensate.

Discussion

Previous studies of introgression versus migration models

Previously Wen and Nakhleh (2018) conducted simulations to examine the inference of gene flow and estimation of introgression probability (γ, equivalent to φ here) under the introgression model when the data were generated under the migration model. The analyses using PHYLONET (Wen and Nakhleh, 2018) searched in the space of models that included the MSC model with no gene flow and the MSci model with introgression, but not migration models, ~~so that~~ the analysis model was misspecified.

The migration model of figure 17d in Wen and Nakhleh (2018), with results reported in their table 2 (mt2 = 1), is the same as the IIM model of figure 6**a**, with unidirectional migration involving non-sister species. The parameter values used (in the notation of fig. 6**a**) are: $\theta = 0.02$ for all populations, $\tau_R = 0.025, \tau_S = 0.015$, with gene flow ceasing at time $\tau_T = 0.01$, and $M = Nm = 0.1$. According to our theory (eq. 10), the introgression probability (γ or φ) should be smaller than $\varphi_0 = 1 - e^{-(4M/\theta)\Delta\tau} = 1 - e^{-4\times 0.1/0.02\times 0.005} = 0.095$, compared with the average estimate 0.18 ± 0.05 reported by Wen and Nakhleh (2018). This result is thus in conflict with our theory. The model of table 2 (mt2 = 0) in Wen and Nakhleh (2018) is a simple IM model, equivalent to the C-A setting in figure 4. The expected introgression probability should be smaller than $\varphi_0 = 1 - e^{-(4M/\theta)\tau_S} = 0.259$, compared with the reported average estimate of 0.17 ± 0.04. In this case the

estimate seems to be approximately consistent with our theory.

The migration model of figure 17f in Wen and Nakhleh (2018) specifies unidirectional migration between two sister species. The parameter values used are $\theta = 0.02$ for all populations, $\tau = 0.015$, with $M = Nm = 0.1$. Results reported in their table 3 (mt = 1) are for the IIM model, with gene flow ceasing at time 0.01 so that the period of gene flow is $\Delta\tau = 0.005$. According to our theory (eq. 11), the introgression probability should be smaller than but very close to $\varphi^*_{(\text{IIM})} = 0.0523$. The authors did not detect gene flow under the introgression model. It is possible that the rate was so low and the datasets were so small that the MSC model with no introgression was favored over the MSci model. Note that when the data are generated under the migration model, both the MSci and MSC models are wrong. However, if the limit of the MLE $\varphi^* > 0$ in the MSci model, the MSci model will be less wrong than the MSC model, measured by the KL divergence, and is expected to win over the MSC model in large datasets Yang and Zhu (2018). Results reported in table 3 (mt = 0) of Wen and Nakhleh (2018) are for the IM model, with the average estimate of 0.11 ± 0.06. This seems consistent with our theory (eq. 11), which predicts $\varphi^*_{(\text{IM})} = 0.167$.

We note that Wen and Nakhleh (2018) used the `-s` switch for SEQ-GEN to specify the value of $\theta/2$ whereas θ should be used instead. Also the authors simulated 20 replicate datasets, each of only 200 loci, so that the datasets are likely too small to produce reliable estimates of the introgression probability. Those factors may partly explain the discrepancy of the simulation results of Wen and Nakhleh (2018) from our theoretical predictions. Note that when data are generated under the migration model and analyzed under the introgression model, the estimated introgression probability should depend on not only the migration rate per generation but also the time period during which gene flow occurs ($\Delta\tau$).

The mode of gene flow and the utility of misspecified introgression models

The asymptotic analysis has been surprisingly useful, even though based on only two species (with gene flow from A to B), with one sequence sampled per species per locus. It generated a number of insights that were confirmed and extended in our simulations, such as (a) the limiting values of introgression time when either the true coalescent times (t) or the differences between sequences (x) are given as data, (b) underestimation of the species divergence time (τ_R) by the MSci model at high migration rates, (c) underestimation of the proportion of migrants by the MSci model, and so on. Extending the theory to three sequences (e.g., a, b_1, b_2, with two sequences from the hybridizing species B; fig. 1) may remove some of the problems we encountered, such as the semi-unidentifiability of θ_S and φ at very low migration rates in figures 2c&d. The theory will be much more complex as it would have to average over the three different gene trees and the two coalescent times, rather than just one coalescent time t in the case of two sequences.

Nevertheless, our asymptotic theory and simulations (fig. 3) constitute a detailed analysis of gene flow in the two-species case. We note that the case of two species may be one of the most challenging cases for inferring gene flow. For example, approximate methods such as the D-statistic cannot detect gene flow with samples from two species at all. Furthermore, when one sequence is sampled per species per locus, the direction of gene flow is unidentifiable, as the two models assuming $A \to B$ and $B \to A$ introgressions predict the same probability distribution of the coalescent time between the two species, $f(t_{ab})$ (Yang and Flouri, 2022, fig. 10). Given the MSci model with $A \to B$ introgression (fig. 1**d**), the introgression probability φ (as well as four other parameters in the model: $\tau_R, \tau_S, \theta_R, \theta_S$) are mathematically identifiable with data of one sequence per species, but the correlation between φ and θ_S is so strong that those parameters are barely identifiable (e.g., methods c and d in fig. 2). Even with multiple sequences per species per locus, there may be a serious lack of information in the data if the dataset is small, with short loci, few sequences per species, or few loci (fig. 3, with $S \leq 4$, $L \leq 1000$, or $M \leq 0.1$ for IIM). Our results highlight the importance of sampling multiple sequences per species (in particular, from species that received immigrants) in real data analysis, besides using large datasets with hundreds or thousands of loci. Even if the model is identifiable with one sample per species, use of multiple sequences provides a major boost in information content. Many approximate methods for detecting gene flow are designed to use only one sample per species, and it has been claimed, incorrectly, that "adding more samples provides little new information with respect to introgression" (Hibbins and Hahn, 2022).

The expected lack of information to estimate parameters θ_S and φ jointly in the MSci model (fig. 1**d**) appears to apply to a recent analysis of genomic data from the *erato* group of *Heliconius* butterflies (Thawornwattana *et al.*, 2022). The estimated *H. sara* → *H. demeter* introgression probability was high with wide CIs for some chromosomal regions with a small number of loci (e.g., chromosome 21 with 4350 noncoding and 3628 coding loci, and an inversion on chromosome 15 with 149 noncoding and 167 coding loci), with the introgression time close to the species divergence time, whereas for the other large chromosomes, the estimated introgression probability was nearly zero ($\varphi < 0.01$). We believe that the true value in this case was $\varphi \approx 0$, but that the limited data from small chromosomal segments led to spurious and

poorly supported signals of introgression, as observed in our simulation (fig. 3).

While the large CIs should help one to avoid making incorrect inferences (of a high φ when the true rate of gene flow is nearly 0), a sensible approach to the problem may be to constrain the population size to be the same before and after an introgression event in the MSci model (that is, with $\theta_S = \theta_A$ and $\theta_H = \theta_B$ in fig. 1**d**). With such [illegible] estimates for ancestral [illegible] not occur, because popu[illegible] well estimated from dat[illegible] extant species, which n[illegible] estimation of the introg[illegible] of estimates of φ to vio[illegible] the population sizes ma[illegible]

Assigning gene flo[illegible] branches caused underestimation of the introgression probability, while the estimated introgression time tended to coincide with species divergence. This feature m[illegible]e mis-assignment in real data analysis (Ji *et al.*, 2022). A number of authors have discussed the impact of ghost species on detection of between-species gene flow (Beerli, 2004; Ottenburghs, 2020). Tricou *et al.* (2022) used simulation to demonstrate that *D*-statistics can be misled to detect false signals of introgression when it involved an unsampled ghost species. In our simulation, the impact of ghost species on Bayesian estimation of introgression rate and time was minor provided we consider the rate of gene flow in the migration and introgression models to reflect both indirect gene flow via intermediate species and direct gene flow.

In our simulations, misspecification of the mode of gene flow (continuous migration versus episodic hybridization/introgression) has relatively small and localized effects on estimation of species divergence times and population sizes around the lineages involved in gene flow, while species divergence times, population sizes for extant species and for ancestral species not involved in gene flow are largely unaffected. Even if gene flow occurs continuously over time (so that the migration model is a more realistic model), the MSci model is effective in extracting historical information about species divergence times and population sizes. Note that on the evolutionary time scale, a few hundred or thousand generations may in effect count as a fixed time point, in which case the MSci model may provide an adequate approximation.

Testing for gene flow

In this study, we fixed the model of introgression in our analyses, with all introgression events pre-identified, and then examined the effects of model misspecification. One may ask what happens if different introgression models are compared using genomic data. We note that if the true model is one of introgression, and if we infer introgression events in the MCMC algorithm, correct placement of introgression events onto the branches of the species tree will constitute the true model. Since Bayesian model selection is known to be consistent, and the true model is included in the set of models under comparison, the true model will dominate with its posterior converging to 1 when the data size (the number of loci) approaches infinity (Dawid, 2011; Yang and Zhu, 2018).

Several approaches may be taken to compare different introgression models. Both *BEAST and PHYLONET have implemented cross-model MCMC algorithms (Green, 1995), which insert and delete introgression events on the species tree, allowing the chain to move between models. The algorithms generate estimates of posterior probability for the different introgression models. Those algorithms are computationally expensive and currently the two programs can handle only small datasets. We thus did not attempt to apply them to our simulated datasets. In the BPP program, one may use the Bayes factor to compare two introgression models, using thermodynamic integration (Gelman and Meng, 1998; Lartillot and Philippe, 2006) combined with Gaussian quadrature to calculate the marginal likelihoods (Rannala and Yang, 2017). The Bayes factor for comparing two nested models (e.g., one with introgression and another without) may also be calculated through the Savage-Dickey density ratio (Dickey, 1971), which uses only a within-model MCMC run under the more general model. This has considerable computational advantages over reversible jump (Green, 1995) but works for nested models only. This approach has recently been applied in comparing and formulating introgression models in an analysis of genomic data from the *Tamias quadrivittatus* group of North American chipmunks (Ji *et al.*, 2022). Calculation of marginal likelihood or Bayes factors may be feasible if we have a small number of well-specified candidate models to evaluate but may be too tedious when there are many candidate models.

Approximate methods have also been developed to infer introgression events using summaries of the multi-locus sequence data. For example, estimated gene tree topologies may be treated as data to compare introgression models in an MCMC algorithm, as in PHYLONET/GT (Wen *et al.*, 2016). Some methods are designed to detect gene flow in a small tree with three or four species, including summary methods based on genome-wide site-pattern counts (such as *D* and HYDE discussed earlier) or on estimated gene trees (e.g., SNAQ, Solis-Lemus and Ane, 2016; Solis-Lemus *et al.*, 2017) and maximum likelihood applied to multilocus sequence alignments (e.g., 3S, Zhu and Yang, 2012; Dalquen *et al.*, 2017). The results from analyses of such species subsets then need be

It was unclear to me what exactly this meant. I suggest the authors stress that their observation is restricted to a small set of scenarios examined; other scenarios may cause a stronger bias.

Unclear what it refers to.

This paragraph relates to other literature and should come later

future perspective

This is more appropriate for the Discussion section.

combined to formulate an introgression model on the large tree for all species, which is a very challenging process (Edelman *et al.*, 2019; Thawornwattana *et al.*, 2022; Ji *et al.*, 2022).

In summary, searching in the space of species phylogenies with introgression events (or so-called phylogenetic network models) is currently a very challenging problem. There is a dire need for efficient MCMC algorithms for Bayesian inference under the MSC model with gene flow and for improvements in statistical performance of approximate methods. Both are fast-developing active research areas. We look forward to breakthroughs in the next few years.

It will also be very interesting to use the same genomic data to compare the migration (IM) and introgression (MSci) models. As discussed in the Introduction, we expect the rate of gene flow to vary over time (and across the genome), so that both kinds of models are simplistic and unrealistic. Even so it will be very useful to use the same data to evaluate which model provides a better approximation to reality. The IM and MSci models often predict very different distributions of gene trees and coalescent times (e.g., figs. S1, S3, S5; see also Jiao and Yang, 2021). This may have two implications. First, genomic data may be informative to distinguish the two kinds of models. Second, it may be too challenging to design efficient cross-model MCMC algorithms to move between them. The gene trees, which are latent variables in models of gene flow, place stringent constraints on the model. If the gene trees at all the loci are fixed when the algorithm moves from an IM model to an MSci model (or vice versa), they may constrain the model so much that changes are in effect impossible and the chain will get stuck. Consider an MCMC move from an IM model assuming $A \rightarrow B$ migration to the corresponding MSci model with $A \rightarrow B$ introgression, with all gene trees fixed. In the current IM model, the youngest sequence divergence time (or coalescent time) between the two species may be very close to zero, which means that the introgression time in the newly proposed MSci model will have to be close to zero as well. Similarly coordinated changes to gene trees during the cross-model move, as achieved in the moves that change the species divergence time (Rannala and Yang, 2003) or the species tree (Yang and Rannala, 2014; Rannala and Yang, 2017) under the MSC without gene flow, appear too difficult. Thus stochastic search in the combined space of both IM and MSci models may be infeasible. Nevertheless, one can use Bayes factors to compare the IM and MSci models.

Assumptions underlying MSC models with gene flow

Here we briefly discuss some other assumptions made in the analysis under the MSC model with gene flow, regarding recombination and selection. We have made the standard assumption that there is no recombination among sites at the same locus, and free recombination between loci. The loci here do not necessarily mean coding genes, and may represent short genomic segments generated using anchored sequencing technologies (RADseq, ultra-conserved elements, etc.), or loosely linked short genomic segments sampled from the genome (Beerli and Felsenstein, 2001; Burgess and Yang, 2008; Lohse *et al.*, 2011; Dalquen *et al.*, 2017; Hey *et al.*, 2018). A recent simulation study examined the effects of intralocus recombination on various analyses using the Bayesian program BPP, including estimation of introgression times and probabilities, and found that recombination at rates comparable to the human rate and with the sequence length at 500 bps per locus have negligible effects (Zhu *et al.*, 2022). However, excessive amounts of recombination (at rates 10 times the human rate) may cause biases in the estimated introgression probability, especially if species divergences times are comparable to coalescent times and if multiple sequences are sampled from the same species. For MSC-based analysis of genomic data from species with very high recombination rates, we suggest that shorter loci or genomic segments be used or the impact of intralocus recombination be assessed by simulation.

Natural selection is known to affect the distribution of coalescent times. Most influential will be species-specific selection or selection involved in divergent adaptation between the species. Such loci may be under very strong selection as they may contribute to species distinctness (Martin *et al.*, 2013). While identifying such gene loci may greatly enrich our understanding of the speciation process, loci and nucleotide sites under such species-specific selection are probably rare on the genome scale. In contrast, purifying selection removing deleterious mutations such as deleterious nonsynonymous mutations in protein-coding genes, may not pose a serious problem. If the protein performs the same function in the species involved, the main effect of purifying selection may be a reduction of the neutral mutation rate. In cases where the noncoding and coding sites in the genome were analyzed as separate datasets, highly consistent results were obtained, with the estimated species divergence times and population sizes being nearly perfectly proportional between the two types of data (Shi and Yang, 2018; Thawornwattana *et al.*, 2018, 2022), and with the same constant of proportionality across different chromosomes (Thawornwattana *et al.*, 2022, fig. S5). Species divergence times in particular showed great consistency with $r^2 > 0.99$. We recommend analyzing non-coding loci separately from coding loci.

It is likely that natural selection (in particular, purifying selection removing deleterious mutations) may have a smaller impact on gene-tree topologies than on coalescent times between sequences. Similarly,

genome-wide averages may be very stable while genealogical variations across the genome may be caused by local variations in recombination rate, the mode and strength of selection, biased gene conversion, etc., besides coalescent and gene flow. Thus approximate methods using gene tree topologies or genome-wide averages may be more robust than likelihood-based fully parametric methods that use information from both gene tree topologies and branch lengths and that use information from both genome-wide averages and genealogical variation across the genome. It may be interesting to use computer simulation to examine the robustness of different inference methods to various model violations including the impact of selection. Applications to various genomic datasets will also provide useful empirical tests.

Conclusions

Based on our asymptotic analyses and computer simulations, we make the following observations. First, population sizes (θ) for extant species are well-estimated when two or more sequences are sampled per species per locus. For one species, θ is simply the average pairwise sequence distance. In almost all simulation settings of this paper, including those in which the model of gene flow is misspecified, θ for extant species are well esti… this to be generally true except … divergence is so recent that the ext… young. Second, estimates of spe… introgression times are dominated … coalescent time or the smallest se… between species (across loci and a… pairs per locus). This is because i… with and without gene flow, sequen… (t_{ab}) between species A and B mu… the species divergence time ($t_{ab} > t_{AB}$). Third, given species divergence times (τ), ancestral population sizes and/or introgression probabilities may be adjusted to achieve the correct amount of sequence divergence (between sequences from the same species or from different species). In particular, if a species divergence time is underestimated, the population size (θ) for the ancestral species is often overestimated to compensate. We have found that all these patterns are useful for interpreting or predicting para… mates under the MSci model when the model is misspecified.

The MSci models studied here are relatively new and have not been tested on many genomic datasets. Our analyses thus provide an empirical assessment of the utility of the models in various situations of model misspecification. Overall we found that the impact of misspecification is local, so that useful inference is still possible despite the misspecification. For example, when gene flow is continuous, the MSci model which assumes that gene flow occurs at one time point gives reliable estimation of species divergence times, unlike the MSC model ignoring gene flow, which is known to lead to seriously underestimation of species divergence times (Ogilvie *et al.*, 2016; Tiley *et al.*, 2022). The estimated introgression probability may also serve as a useful guide even though this reflects both the migration rate per generation and the time duration of the period of gene flow. When gene flow is mis-assigned onto the mother or daughter branches of the genuine introgression lineage, the introgression time is pushed onto the species divergence events, but estimates of older divergence times are largely unaffected. Overall, our results suggest that the simple introgression models may be used to estimate species divergence times and quantify the intensity of gene flow even if the model is an imperfect match to reality.

Materials and Methods

Two-species simulation to establish a correspondence between the migration and the introgression models

We performed a theoretical analysis to establish the relationships between parameter estimates when the true model is the continuous migration model (IM, IIM, and SC; fig. 1**a**-**c**) but the analysis model is the episodic introgression (MSci) model (fig. 1**d**). Our theory assumes an infinite number of loci ($L = \infty$), a finite number of sites per sequence, and only one sequence from each species.

We thus conducted computer simulations to augment the theoretical analysis. Data of multi-locus sequence alignments were simulated under the IM, IIM and SC models of figure 1**a**-**c**, and then analyzed under the MSci model (fig. 1**d**). Population sizes on the species tree (fig. 1) were $\theta_0 = 0.002$ for the thin branches and $\theta_1 = 0.01$ for the thick branches. Migration occurred from species A to B after their divergence $\tau_R = \theta_0$ in the IM model, between $\tau_R = 2\theta_0$ and $\tau_T = \theta_0$ in the IIM model, and between $\tau_T = \theta_0$ and the present in the SC model. In the standard model, the migration rate was $M = 0.2$ individuals per generation. Each dataset consisted of $L = 4000$ loci, with $S = 4$ sequences per species, and $n = 1000$ sites per sequence. To aid the theoretical analysis, we conducted four sets of simulation to examine the impact of the number of sites per sequence (n), the number of sequences per species (S), the number of loci (L), and the mi… The values used were $n = 250, 10$…0; $S = 1, 2, 4, 8, 16$; $L = 250, 500, 1000, 2000, 4000, 8000$; and $M = 0.01, 0.02, 0.03, 0.04, 0.05, 0.07, 0.1, 0.2, 0.3, 0.4, 0.5, 0.7, 1.0, 1.5, 2.0$. With three models (IM, IIM, and SC), four factors (n, S, L, M), and 30 replicates, a total of $3 \times (5 + 3 + 6 + 13) \times 30 = 2790$ datasets were simulated. Replicate datasets were simulated using BPP version 4.4.1 (Flouri *et al.*, 2018, 2020), by generating

Unclear who the subject of this sentence is. Are the authors referring to the behaviour of an inferential model or an a posteriori action by the user?

Unclear.

Confusing, so omit.

redundant to figure captions

At some point please state if BPP only operates with between-population coalescences or also within-population times.

the genealogical trees with c locus and then "evolving" sequences along branches of the gene tree under the JC mutation model (Jukes and Cantor, 1969). Sequences at the tips of the gene tree constituted the data at the locus.

Each dataset was analyzed using BPP under the MSci model (fig. 1**d**) to estimate the parameters. This is the so-called A00 analysis, with the model fixed (Yang, 2015). The Bayesian implementation of the MSci model in BPP accommodates gene-tree reconstruction uncertainties while making use of information in both gene tree topologies and branch lengths, and allows the estimation of the direction, the timing, and strength of introgression events (Jiao *et al.*, 2021). The JC mutation model was assumed in the analysis. Gamma priors are assigned to population size parameters (θ) and to the age of the root on the species tree; $\theta \sim G(2,400)$ and $\tau_0 \sim G(2,200)$. Note that the gamma distribution $G(a,b)$ has mean a/b and variance a/b^2, so that the shape parameter $a = 2$ means diffuse priors. Introgression probability φ was assigned the beta prior beta$(1,1)$, which is $\mathbb{U}(0,1)$.

We used 32,000 MCMC iterations as burnin, and took 2×10^5 samples, sampling every 5 iterations.

Inconsistent way of reporting param values used.

Introgression events assigned to w

In this set of simulations, the introgression event was assigned onto either the parental or a daughter branch of the branch truly involved in introgression. We used models A and B (fig. 4). The species divergence times τ are shown in the trees (fig. 4). We used $S = 4$ sequences per species per locus, with the sequence length n=500 sites. The number of loci was $L =$ 250, 1000, and 4000. We used two population sizes, with $\theta_0 = 0.002$ and $\theta_1 = 0.01$ for the thin and thick branches on the species tree, respectively. We simulated 100 replicate datasets for each parameter setting.

The data were then analyzed using BPP under both models A and B (fig. 4**a**&**b**). We assign gamma priors, $\theta \sim G(2,400)$ with mean 0.005 and $\tau_0 \sim G(2,200)$ with mean 0.01. With two trees/models, three numbers of loci, $2 \times 3 \times 100 = 600$ datasets were simulated, each analyzed twice (under each of the two models). We used 32,000 MCMC iterations as burnin, and took 2×10^5 samples, sampling every 5 iterations.

Continuous migration versus episodic introgression

The data were simulated under the IIM models of figure 4**c**&**d**, with continuous migration at the rate $M = 0.1$ per migrant per generation, and analyzed under MSci models of figure 4**a**&**b**. This is similar to the two-species case analyzed earlier using the asymptotic theory and computer simulation, but here the species tree is larger with more species. There are four combinations. In setting C-A (simulation model C and analysis model A) and D-B, gene flow is continuous in the true model but the MSci model assumes episodic introgression at a particular time point, so that the mode of gene flow is misspecified. In settings C-B and D-A, the mode of gene flow was similarly misspecified but we had in addition misassignment of gene flow to wrong branches on the species tree. Other parameter settings were the same as above. With two trees, three datasizes (L), a total of 600 datasets were generated, each analyzed under two models (fig. 4**a**&**b**).

Isolation with initial migration (IIM) model

The IIM model (Costa and Wilkinson-Herbots, 2017) assumes that there is migration after the species divergence but gene flow ceased after certain time. Suppose we sample sequence data from species A, B and C under model A of figure 6**a**. The model assumes migration between A and B over the time period τ_S–τ_T. We use the MSci model of figure 6**b** to analyze the sequence data sampled from species A, B and C. Other parameter settings are the same as above. With three values for L, we simulated 300 datasets, all analyzed under the MSci model (fig. 6**b**).

confusing

Ghost species

We simulated data under MSci model A (see fig. 1A in Flouri *et al.*, 2020) of figure 7**a**′ and analyzed them under the MSci model B of figure 7**b**, with $\tau_X = \tau_Y$ incorrectly assumed. Here introgression involved a ghost species XUV which has become extinct or is unsampled in the data. This scenario is fully represented by model A of figure 7**a**, so we simulated data on the species tree of figure 7**a**. With three values for L, 300 datasets were generated, all analyzed under the MSci model of figure 7**b**.

We also used the IIM model of figure 8**a** to generate data, with migration from species C to SU and from SU to B. Species V and W represent extinct or unsampled ghost species. The data were simulated using the species tree for five species of figure 8**a** but with no sequences sampled from species V and W, while samples from species A, B and C constitute the data used to fit the MSci model (fig. 8**b**). With three values for L and 100 replicates, 300 datasets were simulated, all analyzed under the MSci model of figure 8**b**.

Acknowledgments

This study has been supported by Biotechnology and Biological Sciences Research Council grants (BB/T003502/1, BB/R01356X/1), as well as by Harvard University.

References

Anderson, E. 1949. *Introgressive Hydridization*. John Wiley, New York.

Bahlo, M. and Griffiths, R. C. 2000. Inference from gene trees in a subdivided population. *Theor. Popul. Biol.*, 57: 79–95.

Beerli, P. 2004. Effect of unsampled populations on the estimation of population sizes and migration rates between sampled populations. *Mol. Ecol.*, 13: 827–836.

Beerli, P. and Felsenstein, J. 1999. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics*, 152: 763–773.

Beerli, P. and Felsenstein, J. 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in *n* subpopulations by using a coalescent approach. *Proc. Natl. Acad. Sci. U.S.A.*, 98: 4563–4568.

Blischak, P. D., Chifman, J., Wolfe, A. D., and Kubatko, L. S. 2018. Hyde: A Python package for genome-scale hybridization detection. *Syst. Biol.*, 67(5): 821–829.

Burgess, R. and 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.

Costa, R. J. and Wilkinson-Herbots, H. 2017. Inference of gene flow in the process of speciation: An efficient maximum-likelihood method for the isolation-with-initial-migration model. *Genetics*, 205(4): 1597–1618.

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

Dawid, A. 2011. Posterior model probabilities. In P. S. Bandyopadhyay and M. Forster, editors, *Philosophy of Statistics*, pages 607–630. Elsevier, New York.

DeGiorgio, M. and Degnan, J. H. 2014. Robustness to divergence time underestimation when inferring species trees from estimated gene trees. *Syst. Biol.*, 63(1): 66–82.

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

Dickey, J. M. 1971. The weighted likelihood ratio, linear hypotheses on normal location parameters. *Ann. Math. Statist.*, 42(1): 204–223.

Dobzhansky, T. 1937. *Genetics and the Origin of Species*. Columbia University, New York.

Durand, E. Y., Patterson, N., Reich, D., and 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., and Davey, J. e. a. 2019. Genomic architecture and introgression shape a butterfly radiation. *Science*, 366(6465): 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., and Wolf, J. B. W. 2012. The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature*, 491: 756–760.

Elworth, R. A. L., Ogilvie, H. A., Zhu, J., and Nakhleh, L. 2019. Advances in computational methods for phylogenetic networks in the presence of hybridization. *Bioinformatics and Phylogenetics*, 29: 317–360.

Finger, N., Farleigh, K., Bracken, J., Leache, A., Francois, O., Yang, Z., Flouri, T., Charran, T., Jezkova, T., Williams, D., and Blair, C. 2022. Genome-scale data reveal deep lineage divergence and a complex demographic history in the Texas horned lizard (*Phrynosoma cornutum*) throughout the southwestern and central usa. *Genome Biol. Evol.*, 14(1): 10.1093/gbe/evab260.

Fisher, R. 1922. On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc. A*, 222: 309–368.

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

Flouri, T., Jiao, X., Rannala, B., and Yang, Z. 2020. A Bayesian implementation of the multispecies coalescent model with introgression for phylogenomic analysis. *Mol. Biol. Evol.*, 37(4): 1211–1223.

Gelman, A. and Meng, X. 1998. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Stat. Sci.*, 13: 163–185.

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

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., and Stenzel, U. e. a. 2010. A draft sequence of the Neandertal genome. *Science*, 328: 710–722.

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

Hey, J. and 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., and Wang, Y. 2018. Phylogeny estimation by integration over isolation with migration models. *Mol. Biol. Evol.*, 35(11): 2805–2818.

Hibbins, M. S. and Hahn, M. W. 2022. Phylogenomic approaches to detecting and characterizing introgression. *Genetics*, 220(2): iyab173.

Ji, J., Jackson, D. J., Leache, A. D., and Yang, Z. 2022. Significant cross-species gene flow detected in the *Tamias quadrivittatus* group of North American chipmunks. *BioRxiv*, page DOI: 10.1101/2021.12.07.471567.

Jiao, X. and Yang, Z. 2021. Defining species when there is gene flow. *Syst. Biol.*, 70(1): 108—119.

Jiao, X., Flouri, T., Rannala, B., and Yang, Z. 2020. The impact of cross-species gene flow on species tree estimation. *Syst. Biol.*, 69(5): 830–847.

Jiao, X., Flouri, T., and Yang, Z. 2021. Multispecies coalescent and its applications to infer species phylogenies and cross-species gene flow. *Nat. Sci. Rev.*, 8: nwab127 (DOI: 10.1093/nsr/nwab127).

Jukes, T. and Cantor, C. 1969. Evolution of protein molecules. In H. Munro, editor, *Mammalian Protein Metabolism*, pages 21–123. Academic Press, New York.

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

Lartillot, N. and Philippe, H. 2006. Computing bayes factors using thermodynamic integration. *Syst. Biol.*, 55: 195–207.

Liu, S., Lorenzen, E. D., Fumagalli, M., Li, B., and Harris, K. e. a. 2014. Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. *Cell*, 157: 785–794.

Lohse, K. and Frantz, L. A. 2014. Neandertal admixture in Eurasia confirmed by maximum-likelihood analysis of three genomes. *Genetics*, 196(4): 1241–1251.

Lohse, K., Harrison, R., and Barton, N. 2011. A general method for calculating likelihoods under the coalescent process. *Genetics*, 189: 977–987.

Maddison, W. 1997. Gene trees in species trees. *Syst. Biol.*, 46: 523–536.

Malecot, G. 1948. *Les mathematiques de l'heredite*. Masson, Paris.

Mallet, J. 2007. Hybrid speciation. *Nature*, 446: 279–283.

Mallet, J., Besansky, N., and Hahn, M. W. 2016. How reticulated are species? *BioEssays*, 38(2): 140–149.

Martin, S. H. and Jiggins, C. D. 2017. Interpreting the genomic landscape of introgression. *Curr. Opin. Genet. Dev.*, 47: 69–74.

Martin, S. H., Dasmahapatra, K. K., Nadeau, N. J., Salazar, C., Walters, J. R., Simpson, F., Blaxter, M., Manica, A., Mallet, J., and Jiggins, C. D. 2013. Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Res.*, 23(11): 1817–1828.

Martin, S. H., Davey, J. W., Salazar, C., and Jiggins, C. D. 2019. Recombination rate variation shapes barriers to introgression across butterfly genomes. *PLoS Biol.*, 17(2): e2006288.

Meng, C. and Kubatko, L. 2009. Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. *Theo. Popul. Biol.*, 75(1): 35–45.

Muller, H. J. 1942. Isolating mechanisms, evolution, and temperature. *Biol. Symp*, 6: 71–125.

Nichols, R. 2001. Gene trees and species trees are not the same. *Trends Ecol. Evol.*, 16: 358–364.

Notohara, M. 1990. The coalescent and the genealogical process in geographically structured populations. *J. Math. Biol.*, 29: 59–75.

Ogilvie, H. A., Heled, J., Xie, D., and Drummond, A. J. 2016. Computational performance and statistical accuracy of *beast and comparisons with other methods. *Syst. Biol.*, 65: 381–396.

O'Hagan, A. and Forster, J. 2004. *Kendall's Advanced Theory of Statistics: Bayesian Inference*. Arnold, London.

Ottenburghs, J. 2020. Ghost introgression: spooky gene flow in the distant past. *Bioessays*, 42(6): e2000012.

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

Schumer, M., Xu, C., Powell, D. L., Durvasula, A., Skov, L., Holland, C., Blazier, J. C., Sankararaman, S., Andolfatto, P., Rosenthal, G. G., and Przeworski, M. 2018. Natural selection interacts with recombination to shape the evolution of hybrid genomes. *Science*, 360(6389): 656–660.

Shi, C. and 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: 159–179.

Slatkin, M. 1987. Gene flow and the geographic structure of natural populations. *Science*, 236(4803): 787–792.

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

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

Thawornwattana, Y., Dalquen, D., and Yang, Z. 2018. 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., Seixas, F. A., Mallet, J., and Yang, Z. 2022. Full-likelihood genomic analysis clarifies a complex history of species divergence and introgression: the example of the *erato-sara* group of *Heliconius* butterflies. *Syst. Biol.*

Tiley, G. P., Flouri, T., Jiao, X., Poelstra, J. P., Xu, B., Zhu, T., Rannala, B., Yoder, A. D., and Yang, Z. 2022. Estimation of species divergence times in presence of cross-species gene flow. *Syst. Biol.*

Tricou, T., Tannier, E., and de Vienne, D. M. 2022. Ghost lineages highly influence the interpretation of introgression tests. *Syst. Biol.*, page 10.1093/sysbio/syac011.

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

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

Wright, S. 1943. Isolation by distance. *Genetics*, 28: 114–138.

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. and Flouri, T. 2022. Estimation of cross-species introgression rates using genomic data despite model unidentifiability. *Mol. Biol. Evol.*

Yang, Z. and Rannala, B. 2014. Unguided species delimitation using dna sequence data from multiple loci. *Mol. Biol. Evol.*, 31(12): 3125–3135.

Yang, Z. and Zhu, T. 2018. Bayesian selection of misspecified models is overconfident and may cause spurious posterior probabilities for phylogenetic trees. *Proc. Natl. Acad. Sci. U.S.A.*, 115(8): 1854–1859.

Yu, Y. and Nakhleh, L. 2015. A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics*, 16 Suppl 10: S10.

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

Zhang, C., Ogilvie, H. A., Drummond, A. J., and Stadler, T. 2018. Bayesian inference of species networks from multilocus sequence data. *Mol. Biol. Evol.*, 35: 504–517.

Zhu, T. and 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.

Zhu, T. and Yang, Z. 2021. Complexity of the simplest species tree problem. *Mol. Biol. Evol.*, 39: 3993—-4009.

Zhu, T., Flouri, T., and Yang, Z. 2022. A simulation study to examine the impact of recombination on phylogenomic inferences under the multispecies coalescent model. *Mol. Ecol.*, page DOI: 10.1111/mec.16433.

Appendix A. Likelihood function under the IM and MSci models in the case of two species

Here we derive the likelihood function under the three continuous migration models (IM, IIM, SC) and the episodic introgression (MSci) model for two species (fig. 1**a**-**d**) when the data consist of an infinite number of loci, with two sequences sampled at each locus, one from each species. The data at each locus can be summarized as x differences at n sites. The infinite-sites mutation model is assumed so that the probability of data given the coalescent time is given by the Poisson probability (eq. 8). To calculate the likelihood, we integrate over the unknown coalescent time t, which has density $f_{\rm m}(t|\Theta_{\rm m})$ (eq. 3) under the IM or IIM model, $f_{\rm sc}(t|\Theta_{\rm m})$ (eq. 4) under the SC model, and $f_{\rm i}(t|\Theta_{\rm i})$ (eq. 5) under the MSci model.

Under the IM and IIM models (fig. 1**a**&**b**), we have from eq. 3

$$f(x|\Theta_{\rm m}) = \int_0^\infty f(x|t) f_{\rm m}(t|\Theta_{\rm m})\,{\rm d}t = \int_0^\infty \frac{1}{x!}(2nt)^x {\rm e}^{-2nt} f_{\rm m}(t|\Theta_{\rm m})\,{\rm d}t = I_1 + I_2.$$

For the IM, IIM, and SCM, the likelihood functions were derived by Wilkinson-Herbots and Costa, if I am not mistaken.

The first term is

$$\begin{aligned} I_1 &= \int_{\tau_T}^{\tau_R} \frac{1}{x!}(2nt)^x {\rm e}^{-2nt} \frac{2w}{2-w\theta_A}\left[{\rm e}^{-w(t-\tau_T)} - {\rm e}^{-\frac{2}{\theta_A}(t-\tau_T)}\right]{\rm d}t \\ &= \frac{(2n)^x}{x!}\frac{2w}{2-w\theta_A}\left[{\rm e}^{w\tau_T}\int_{\tau_T}^{\tau_R} t^x {\rm e}^{-(2n+w)t}\,{\rm d}t - {\rm e}^{\frac{2}{\theta_A}\tau_T}\int_{\tau_T}^{\tau_R} t^x {\rm e}^{-(2n+\frac{2}{\theta_A})t}\,{\rm d}t\right], \end{aligned} \tag{A2}$$

where the two integrals are

$$\begin{aligned} \int_{\tau_T}^{\tau_R} t^x {\rm e}^{-(2n+w)t}\,{\rm d}t &= \frac{1}{(2n+w)^{x+1}}\left[\gamma(x+1,\tau_R(2n+w)) - \gamma(x+1,\tau_T(2n+w))\right], \\ \int_{\tau_T}^{\tau_R} t^x {\rm e}^{-(2n+\frac{2}{\theta_A})t}\,{\rm d}t &= \frac{1}{(2n+\frac{2}{\theta_A})^{x+1}}\left[\gamma(x+1,\tau_R(2n+\tfrac{2}{\theta_A})) - \gamma(x+1,\tau_T(2n+\tfrac{2}{\theta_A}))\right], \end{aligned} \tag{A3}$$

with

$$\begin{aligned} \Gamma(a) &= \int_0^\infty t^{a-1}{\rm e}^{-t}\,{\rm d}t, \\ \gamma(a,x) &= \int_0^x t^{a-1}{\rm e}^{-t}\,{\rm d}t, \end{aligned} \tag{A4}$$

to be the gamma function and the lower incomplete gamma function, respectively, with $\gamma(a,\infty) = \Gamma(a)$.

Similarly the second term in eq. A1 is

$$\begin{aligned} I_2 &= \int_{\tau_R}^\infty \frac{1}{x!}(2nt)^x {\rm e}^{-2nt}\frac{1}{2-w\theta_A}\left[2{\rm e}^{-w(\tau_R-\tau_T)} - w\theta_A {\rm e}^{-\frac{2}{\theta_A}(\tau_R-\tau_T)}\right]\frac{2}{\theta_R}{\rm e}^{-\frac{2}{\theta_R}(t-\tau_R)}\,{\rm d}t \\ &= \frac{(2n)^x}{x!}\frac{1}{2-w\theta_A}\left[2{\rm e}^{-w(\tau_R-\tau_T)} - w\theta_T {\rm e}^{-\frac{2}{\theta_A}(\tau_R-\tau_T)}\right]\frac{2}{\theta_R}{\rm e}^{\frac{2}{\theta_R}\tau_R}\int_{\tau_R}^\infty t^x {\rm e}^{-(2n+\frac{2}{\theta_R})t}\,{\rm d}t \\ &= \frac{(2n)^x}{x!}\frac{1}{2-w\theta_A}\left[2{\rm e}^{-w(\tau_R-\tau_T)} - w\theta_T {\rm e}^{-\frac{2}{\theta_A}(\tau_R-\tau_T)}\right]\frac{2}{\theta_R}{\rm e}^{\frac{2}{\theta_R}\tau_R}\times\frac{\Gamma(x+1)-\gamma(x+1,\tau_R(2n+\frac{2}{\theta_R}))}{(2n+\frac{2}{\theta_R})^{x+1}}. \end{aligned} \tag{A5}$$

Putting everything together, we get

$$\begin{aligned} f_{\rm m}(x|\Theta_{\rm m}) = \frac{(2n)^x}{x!}\frac{2w}{2-w\theta_A}\Bigg(&\frac{{\rm e}^{w\tau_T}}{(2n+w)^{x+1}}\left[\gamma(x+1,\tau_R(2n+w)) - \gamma(x+1,\tau_T(2n+w))\right] \\ &- \frac{{\rm e}^{\frac{2}{\theta_A}\tau_T}}{(2n+\frac{2}{\theta_A})^{x+1}}\left[\gamma(x+1,\tau_R(2n+\tfrac{2}{\theta_A})) - \gamma(x+1,\tau_T(2n+\tfrac{2}{\theta_A}))\right]\Bigg) \\ &+ \frac{(2n)^x}{x!}\frac{1}{2-w\theta_A}\left[2{\rm e}^{-w(\tau_R-\tau_T)} - w\theta_A e^{-\frac{2}{\theta_A}(\tau_R-\tau_T)}\right]\frac{2}{\theta_R}{\rm e}^{\frac{2}{\theta_R}\tau_R} \\ &\times\frac{\Gamma(x+1)-\gamma(x+1,\tau_R(2n+\frac{2}{\theta_R}))}{(2n+\frac{2}{\theta_R})^{x+1}}. \end{aligned} \tag{A6}$$

Similarly, under the secondary-contact (SC) model (fig. 1**c**), the density of coalescent time t is given in eq. 4. The probability of observing x differences at n sites at a locus is

$$f_{\rm sc}(x|\Theta_{\rm m}) = \int_0^\infty f(x|t) f_{\rm sc}(t|\Theta_{\rm m})\,{\rm d}t = J_1 + J_2 + J_3, \tag{A7}$$

where

$$\begin{aligned}
J_1 &= \int_0^{\tau_T} \frac{1}{x!}(2nt)^x {\rm e}^{-2nt} \frac{w\theta_A}{2-w\theta_A}\left[{\rm e}^{-wt} - {\rm e}^{-\frac{2}{\theta_A}t}\right]\frac{2}{\theta_A}\,{\rm d}t \\
&= \frac{(2n)^x}{x!}\frac{2w}{2-w\theta_A}\left[\frac{\gamma(x+1,\tau_T(2n+w))}{(2n+w)^{x+1}} - \frac{\gamma(x+1,\tau_T(2n+\frac{2}{\theta_A}))}{(2n+\frac{2}{\theta_A})^{x+1}}\right], \\
J_2 &= \int_{\tau_T}^{\tau_R} \frac{1}{x!}(2nt)^x {\rm e}^{-2nt} \frac{w\theta_A}{2-w\theta_A}\left[{\rm e}^{-w\tau_T} - {\rm e}^{-\frac{2}{\theta_A}\tau_T}\right]\frac{2}{\theta_A}{\rm e}^{-\frac{2}{\theta_A}(t-\tau_T)}\,{\rm d}t \\
&= \frac{(2n)^x}{x!}\frac{2w}{2-w\theta_A}\left[{\rm e}^{-w\tau_T} - {\rm e}^{-\frac{2}{\theta_A}\tau_T}\right]\frac{{\rm e}^{\frac{2}{\theta_A}\tau_T}}{(2n+\frac{2}{\theta_A})^{x+1}}\left[\gamma(x+1,\tau_R(2n+\tfrac{2}{\theta_A})) - \gamma(x+1,\tau_T(2n+\tfrac{2}{\theta_A}))\right], \\
J_3 &= \int_{\tau_R}^{\infty} \frac{1}{x!}(2nt)^x {\rm e}^{-2nt}\left[\frac{w\theta_A}{2-w\theta_A}\left({\rm e}^{-w\tau_T} - {\rm e}^{-\frac{2}{\theta_A}\tau_T}\right){\rm e}^{-\frac{2}{\theta_A}(\tau_R-\tau_T)} + {\rm e}^{-w\tau_T}\right]\frac{2}{\theta_R}{\rm e}^{-\frac{2}{\theta_R}(t-\tau_R)}\,{\rm d}t \\
&= \frac{(2n)^x}{x!}\left[\frac{w\theta_A}{2-w\theta_A}\left({\rm e}^{-w\tau_T} - {\rm e}^{-\frac{2}{\theta_A}\tau_T}\right){\rm e}^{-\frac{2}{\theta_A}(\tau_R-\tau_T)} + {\rm e}^{-w\tau_T}\right]\frac{2}{\theta_R}\frac{{\rm e}^{\frac{2}{\theta_R}\tau_R}}{(2n+\frac{2}{\theta_R})^{x+1}}\left[\Gamma(x+1) - \gamma(x+1,\tau_R(2n+\tfrac{2}{\theta_R}))\right].
\end{aligned} \tag{A8}$$

Finally, under the MSci model (fig. 1**d**), the density of coalescent time t is given in eq. 5. We have

$$\begin{aligned}
f_{\rm i}(x|\Theta_{\rm i}) &= \int_0^\infty f(x|t) f_{\rm i}(t|\Theta_{\rm i})\,{\rm d}t \\
&= \int_{\tau_S}^{\tau_R} \frac{1}{x!}(2nt)^x {\rm e}^{-2nt}\,\varphi\frac{2}{\theta_S}{\rm e}^{-\frac{2}{\theta_S}(t-\tau_S)}\,{\rm d}t + \int_{\tau_R}^{\infty}\frac{1}{x!}(2nt)^x {\rm e}^{-2nt}\left[\varphi\,{\rm e}^{-\frac{2}{\theta_S}(\tau_R-\tau_S)} + (1-\varphi)\right]\frac{2}{\theta_R}{\rm e}^{-\frac{2}{\theta_R}(t-\tau_R)}\,{\rm d}t \\
&= \frac{(2n)^x}{x!}\varphi\frac{2}{\theta_S}{\rm e}^{\frac{2}{\theta_S}\tau_S}\int_{\tau_S}^{\tau_R} t^x {\rm e}^{-(2n+\frac{2}{\theta_S})t}\,{\rm d}t + \frac{(2n)^x}{x!}\left[\varphi\,{\rm e}^{-\frac{2}{\theta_S}(\tau_R-\tau_S)} + (1-\varphi)\right]\frac{2}{\theta_R}{\rm e}^{\frac{2}{\theta_R}\tau_R}\int_{\tau_R}^{\infty} t^x {\rm e}^{-(2n+\frac{2}{\theta_R})t}\,{\rm d}t \\
&= \frac{(2n)^x}{x!}\varphi\frac{2}{\theta_S}{\rm e}^{\frac{2}{\theta_S}\tau_S}\times\frac{\gamma(x+1,\tau_R(2n+\frac{2}{\theta_S})) - \gamma(x+1,\tau_S(2n+\frac{2}{\theta_S}))}{(2n+\frac{2}{\theta_S})^{x+1}} \\
&\quad + \frac{(2n)^x}{x!}\left[\varphi\,{\rm e}^{-\frac{2}{\theta_S}(\tau_R-\tau_S)} + (1-\varphi)\right]\frac{2}{\theta_R}{\rm e}^{\frac{2}{\theta_R}\tau_R}\times\frac{\Gamma(x+1) - \gamma(x+1,\tau_R(2n+\frac{2}{\theta_R}))}{(2n+\frac{2}{\theta_R})^{x+1}}.
\end{aligned} \tag{A9}$$

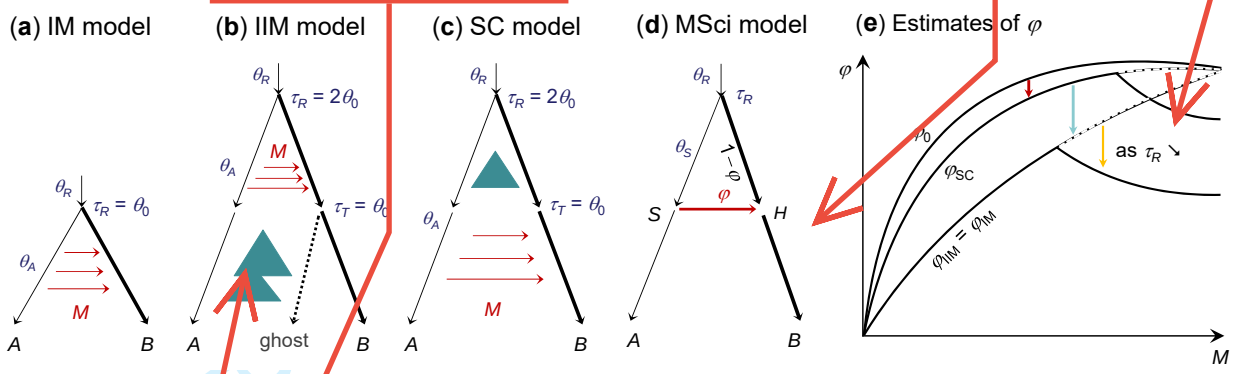


Figure 1: (**a**–**c**) Three migration models for two species A and B used to generate data: IM (isolation-with-migration), IIM (isolation-with-initial-migration), and SC (secondary-contact). The IIM model is an instance of the IM model with a ghost species at node T and with migration from species A to T (**b**). Similarly the SC model (**c**) is a case of the migration model with $\tau_T > 0$. Divergence time (τ_R) and the time point at which gene flow started or stopped (τ_T) are given next to the nodes, with population sizes $\theta_0 = 0.002$ for the thin branches and $\theta_1 = 0.01$ for the thick branches. The migration rate was $M = 0.2$ migrant individuals from A to B per generation, but other values are considered as well. Note that the time period of gene flow is $\Delta\tau = \theta_0$ in all three models, so that the total expected amount of introgression (φ_0 of eq. 10) is the same. (**d**) The introgression (MSci) model is fitted to the data of the coalescent time between two sequences (i.e., with infinite sequence length), generated under the migration models of **a**-**c**. (**e**) A schematic summary of the estimate of the introgression probability (φ) in the MSci model (**d**) when the data are generated under the models of **a**-**c**.

The IIM model as such has no ghost species. Do the authors want to say that this is the model to which they added a ghost population?

State time tau_S

unclear what this is meant to

Unclear in two ways. Do the authors mean the IM model? tau_T has not been defined. A definition of the SC model in terms of tau_T only cannot be sufficient given the current set of parameters.

I advise against assigning specific point values here, as this blurs the distinction between model description and model instances chosen by the authors

Please state which effective size M is scaled by, N1 or N2.

I find the green triangles unnecessary, if not confusing. Simply having no red arrows should do. Compare to Fig. 4, which appears clear w/o green triangles.

$\theta_A = \theta_S = \theta_R = \theta_0 = 0.002$

$\theta_B = \theta_H = \theta_1 = 0.01$

$\Delta\tau = \theta_0 = 0.002$

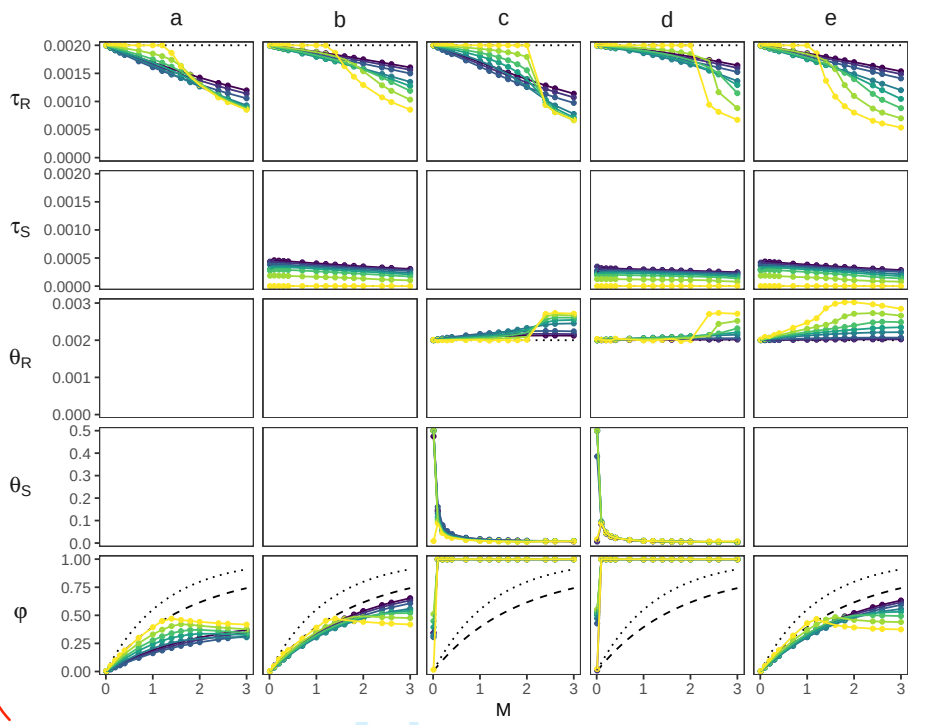


Figure 2: (**2s-IM:MLE**) Best-fitting parameter values under the MSci model of figure 1**d** when data of two sequences per locus (one per species), each of n sites, are generated under the IM model of figure 1**a**. Five methods (a-e) are used to fit the MSci model, estimating 2, 3, 4, 5, and 4 parameters, respectively, while the other parameters are fixed. In (**a**), τ_R and φ are estimated, but θ_R and θ_S are fixed at their true values in the IM model, and the introgression time $\tau_S = \tau_H$ is fixed at $\tau_T = 0$. In (b), τ_S is estimated as well. In (c), $\tau_S = 0$ is fixed, while the other four parameters are estimated. In (d), all five parameters are estimated. In (e), the constraint $\theta_R = \theta_S$ is enforced so that four free parameters are estimated. The dotted lines for φ indicate the true total amount of introgression of eq. 10. The dashed lines indicate φ^* of eq. 11. The true and best-fitting distributions of the coalescent time (t) are shown in figure S1.

Introgression at the most recent point in time (!)

Please clarify why all three models are stated here, but only data for two of them are shown. The corresponding Results text confusingly also explains results related to the SCM, even though no results are apparently shown for the SCM in Fig. 3.

I think it would be preferable to show the rel. error

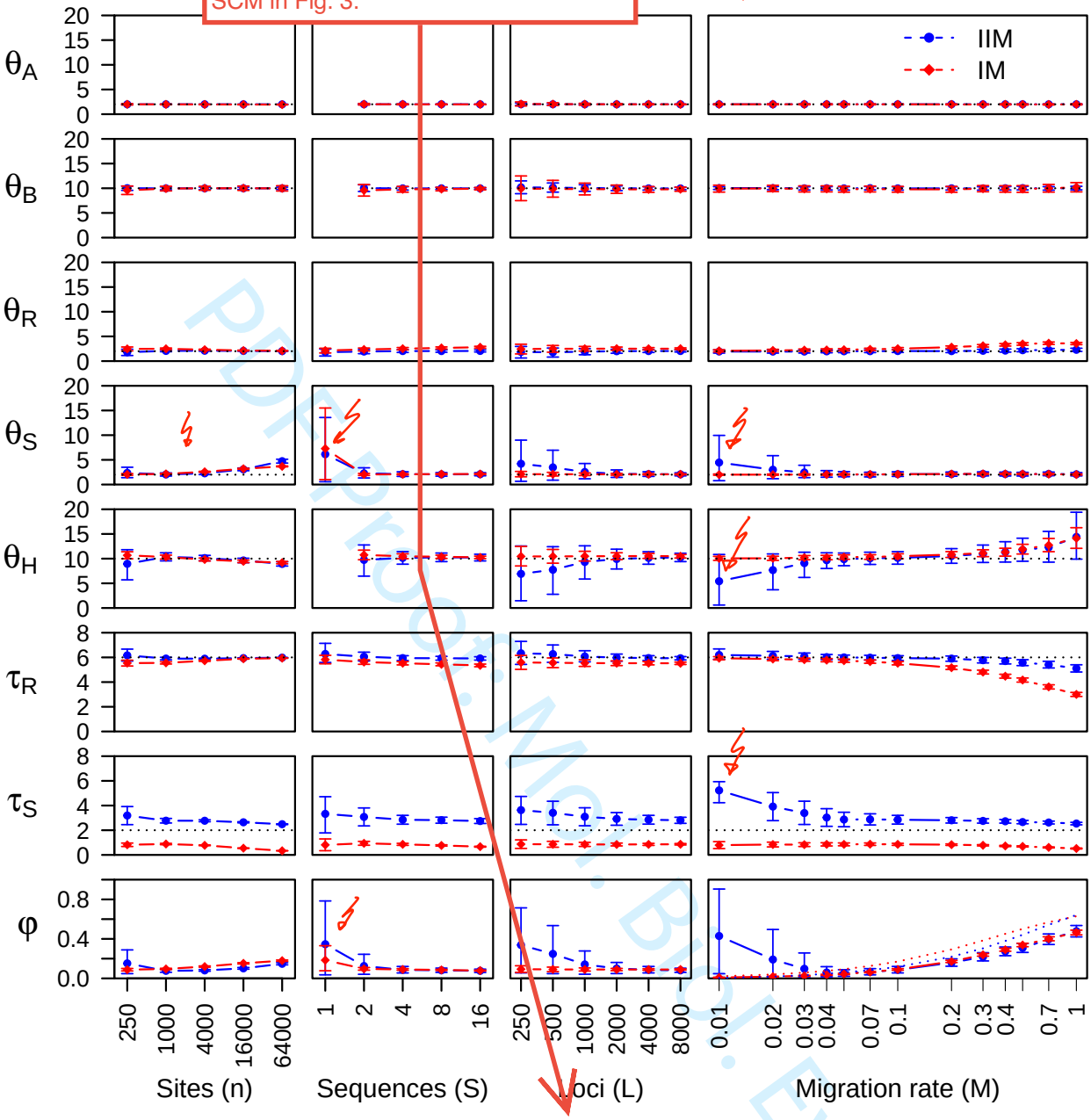
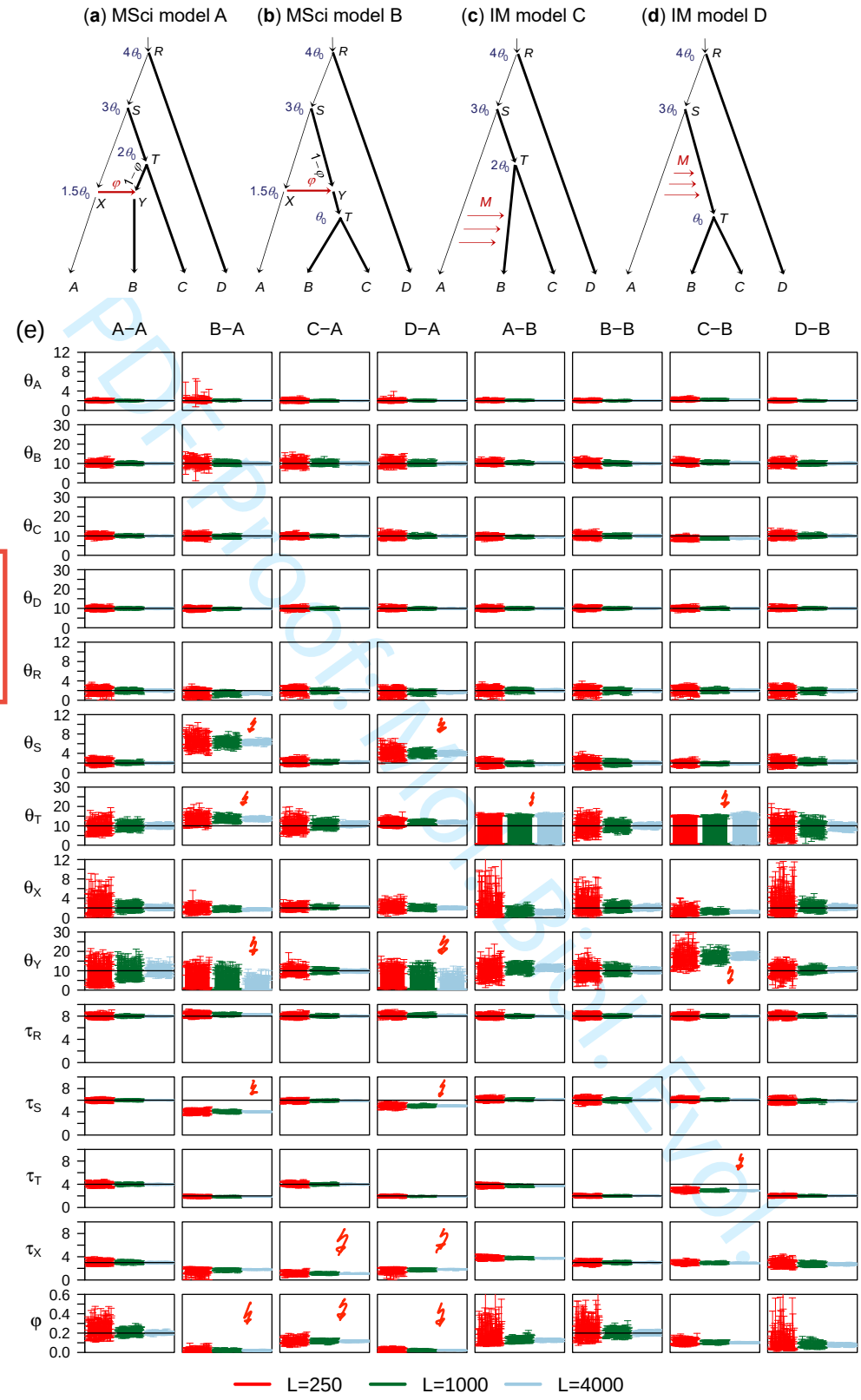


Figure 3: Average posterior means and 95% HPD CIs for parameters in the MSci model of figure 1**d** over 30 replicate datasets simulated under the migration (IM, IIM, and SC) models of figure 1**a**-**c**, plotted against the number of sites per sequence (n), the number of sequences per species (S), the number of loci (L), and the migration rate (M). Parameters in the migration model are given in the legend to figure 1. In the standard setting, each dataset consists of $L = 4000$ loci, with $S = 4$ sequences per species at each locus and $n = 1000$ sites per sequence, and the migration rate was $M = 0.2$ individuals per generation. In the four sets of simulations, one of the factors (n, S, L, M) varies while the others are fixed. When $S = 1$, population sizes θ_A, θ_B, and θ_H are unidentifiable. Estimates of τs and θs are multiplied by 10^3. Dotted lines indicate true values of identifiable parameters, except in the plot of φ against M, where it represents φ_0 of eq. 10, (which is identical for the IM, IIM, and SC models of fig. 1). Note that the n, S, L, and M axes are all on the logarithmic scale.

I think other parameters must also have been scaled, right?

Please relocate the provision of these values to a position separate from the caption of Fig. 1, (e.g. the one in Materials and Methods) and then adjust the reference here.



The tick labels of the Y axes do not correspond to the supposed scale, except for tau_S.

Figure 4: (**a**&**b**) Two introgression (MSci) models and (**c**&**d**) two migration (IM) models used in simulation. The thin branches have the population size $\theta_0 = 0.002$ and the thick branches have $\theta_1 = 0.01$. In MSci model A, the species divergence/introgression times are $\tau_R = 4\theta_0$, $\tau_S = 3\theta_0$, $\tau_T = 2\theta_0$, and $\tau_X = \tau_Y = 1.5\theta_0$. In MSci model B, $\tau_R = 4\theta_0$, $\tau_S = 3\theta_0$, $\tau_T = \theta_0$, and $\tau_X = \tau_Y = 1.5\theta_0$. Introgression probability is $\varphi = 0.2$. In the IM model C, $\tau_R = 4\theta_0$, $\tau_S = 3\theta_0$, and $\tau_T = 2\theta_0$, with migration occurring from species A to B over time period $(0, \tau_T)$ at the rate $M = 0.1$ migrants per generation. In the IM model D, $\tau_R = 4\theta_0$, $\tau_S = 3\theta_0$, and $\tau_T = \theta_0$, with migration from species A to ST over time period (τ_T, τ_S) at the rate $M = 0.1$. (**e**) The 95% HPD CIs for parameters in 100 replicate datasets of $L = 250, 1000$, and 4000 loci. The key is in the simulation-analysis format; i.e., 'B-A' means that data are simulated under model B and analyzed under model A. Parameters θs and τs are multiplied by 10^3. Black solid line indicates the true value.

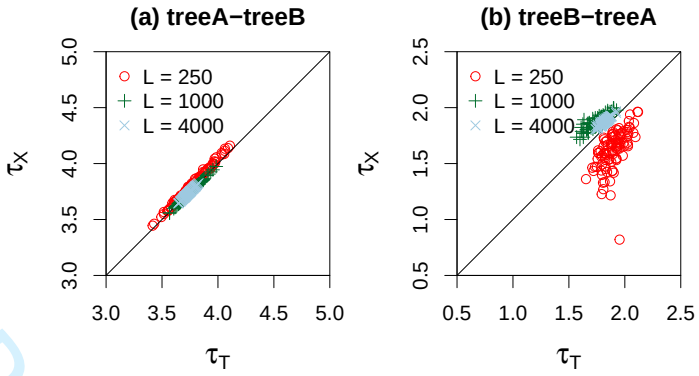


Figure 5: Posterior means of speciation/introgression times ($\times 10^{-3}$) when the introgression event is assigned to a wrong branch. In (**a**) tree A-tree B, data were simulated using species tree A (fig. 4**a**), with introgression from species A to B, but are analyzed assuming tree B, with introgression assigned incorrectly to the parental branch ST (so that $\tau_X > \tau_T$). In (**b**) tree B-tree A, data were simulated under tree B (fig. 4**b**) and analyzed assuming tree A, with introgression assigned to the daughter branch B (with $\tau_X < \tau_T$). For each datasize (with $L = 250, 1000$, or 4000 loci), 100 replicate datasets were generated and analyzed. These correspond to the A-B and B-A settings of figure 4**e**, where estimates of other parameters are shown.

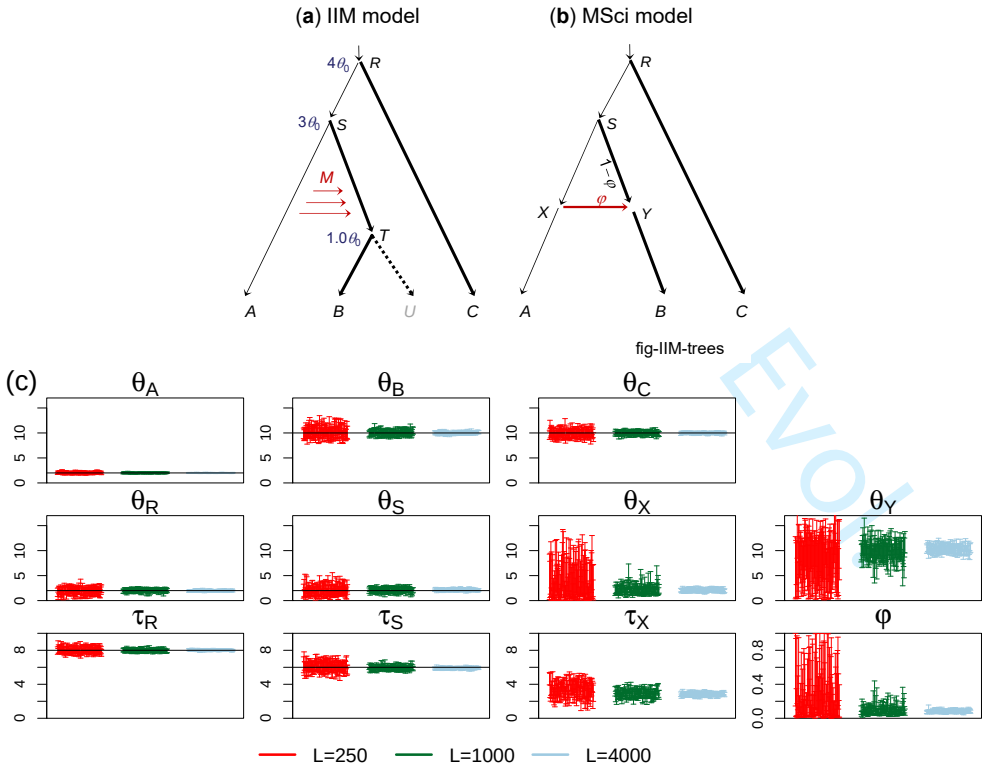


Figure 6: (**a**) An isolation-with-initial-migration (IIM) model used to simulate data. The parameter values used are $\theta_0 = 0.002$ for population sizes for the thin branches and $\theta_1 = 0.01$ for the thick branches, $\tau_R = 4\theta_0$, $\tau_S = 3\theta_0$, $\tau_T = \theta_0$ for species divergence times. The number of sequences is $S = 4$, with the sequence length $n = 500$. The migration rate is $M = 0.1$. (**b**) An MSci model used to analyze the data. (**c**) The 95% HPD CIs for parameters, with black lines indicating the true values. Estimates of θ and τ are multiplied by 10^3.

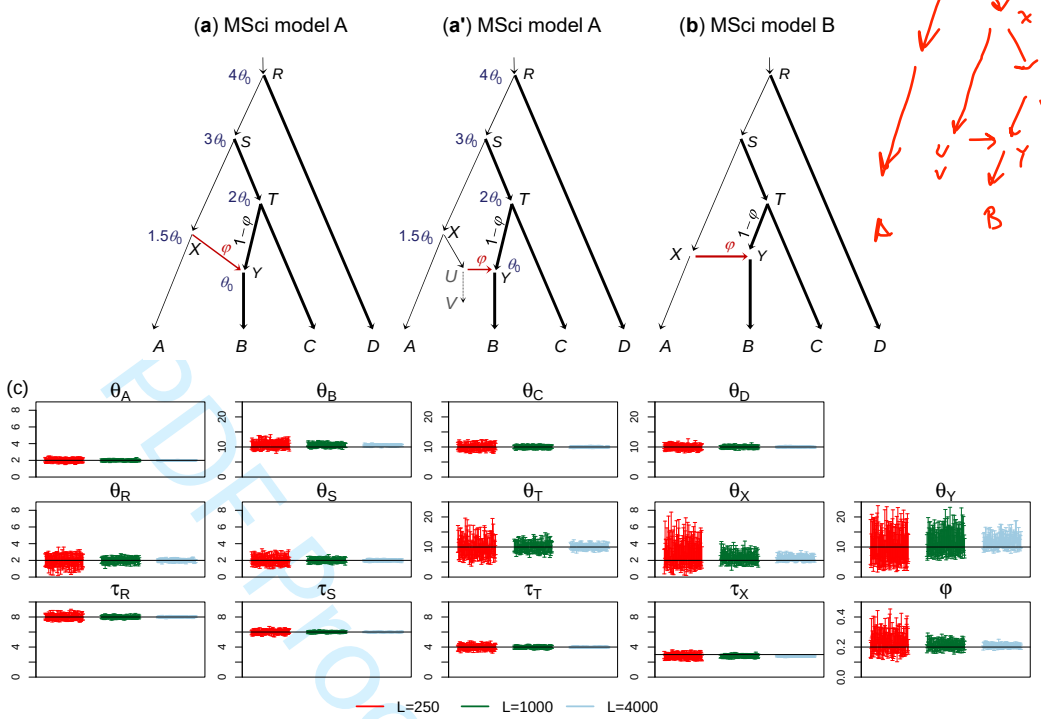


Figure 7: (**a**) MSci model A (fig. 1A in Flouri *et al.*, 2020) assumes that $\tau_X > \tau_Y$ and $\tau_T > \tau_Y$ and can represent scenario (**a**′) in which species X split into two species, and species XUV contributed migrants into species TB at time τ_Y but has since become extinct. This is assumed to simulate data, with $\theta_0 = 0.002$ for the thin branches and $\theta_1 = 0.01$ for the thick branches, $\tau_R = 4\theta_0, \tau_S = 3\theta_0, \tau_T = 2\theta_0$, $\tau_X = 1.5\theta_0$, and $\tau_Y = \theta_0$. The introgression probability is $\varphi = 0.2$. The number of sequences is $S = 4$, and the sequence length is $n = 500$. (**b**) MSci model B (fig. 1B in Flouri *et al.*, 2020) used to analyze the data, which incorrectly assumes $\tau_X = \tau_Y$. (**c**) The 95% HPD CIs for parameters, with θs and τs multiplied by 10^3 and black solid line indicating the true value.

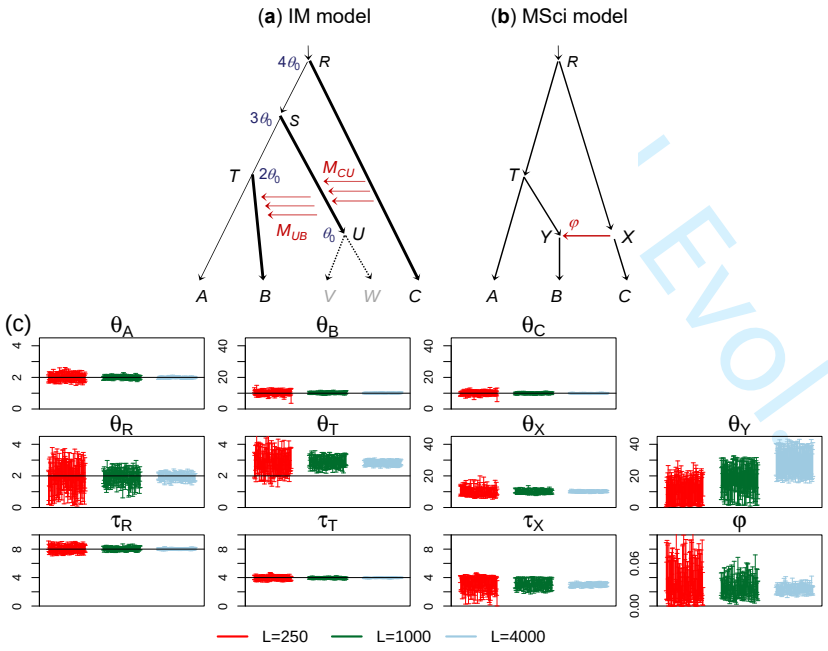


Figure 8: (**a**) Migration model involving ghost species for simulating data. The parameter values used are $\theta_0 = 0.002$ for the thin branches and $\theta_1 = 0.01$ for the thick branches, with the divergence times (τs) shown next to the internal nodes. The number of sequences is $S = 4$, and the sequence length is $n = 500$. The migration rates are $M_{CU} = M_{UB} = 0.2$ migrants per generation. (**b**) MSci model used to analyze the data. (**c**) The 95% HPD CI for parameters, with θs and τs multiplied by 10^3, and with black solid lines indicating the true values.

Table 1. Average posterior means and 95% HPD intervals (in parentheses) for introgression time (τ_X, $\times 10^3$) and introgression probability (φ_X) in the simulations

	τ_X			φ		
Analysis	$L = 250$	$L = 1000$	$L = 4000$	$L = 250$	$L = 1000$	$L = 4000$
Fig. 4 A-A	3.06 (2.63, 3.49)	3.02 (2.80, 3.24)	3.00 (2.89, 3.11)	0.23 (0.16, 0.32)	0.21 (0.17, 0.24)	0.20 (0.19, 0.22)
Fig. 4 B-A	1.62 (0.95, 2.05)	1.77 (1.54, 1.96)	1.82 (1.72, 1.91)	0.02 (0.00, 0.04)	0.02 (0.01, 0.03)	0.02 (0.02, 0.03)
Fig. 4 C-A	1.12 (0.83, 1.40)	1.11 (0.97, 1.25)	1.11 (1.04, 1.18)	0.12 (0.09, 0.15)	0.12 (0.10, 0.13)	0.12 (0.11, 0.12)
Fig. 4 D-A	1.69 (1.18, 2.07)	1.80 (1.58, 1.97)	1.86 (1.76, 1.94)	0.02 (0.01, 0.04)	0.02 (0.01, 0.03)	0.02 (0.02, 0.03)
Fig. 4 A-B	3.82 (3.53, 4.11)	3.75 (3.61, 3.90)	3.73 (3.66, 3.80)	0.18 (0.09, 0.28)	0.13 (0.11, 0.16)	0.12 (0.11, 0.14)
Fig. 4 B-B	2.98 (2.61, 3.35)	2.99 (2.80, 3.18)	3.00 (2.91, 3.10)	0.23 (0.14, 0.34)	0.20 (0.17, 0.24)	0.20 (0.18, 0.22)
Fig. 4 C-B	2.98 (2.72, 3.24)	2.93 (2.80, 3.06)	2.91 (2.85, 2.98)	0.11 (0.08, 0.14)	0.10 (0.09, 0.12)	0.10 (0.10, 0.11)
Fig. 4 D-B	2.83 (2.28, 3.38)	2.71 (2.42, 3.00)	2.73 (2.59, 2.87)	0.11 (0.04, 0.20)	0.08 (0.05, 0.10)	0.08 (0.07, 0.09)
Fig. 6 IIM	3.40 (2.38, 4.36)	2.93 (2.42, 3.43)	2.83 (2.58, 3.08)	0.24 (0.04, 0.53)	0.10 (0.05, 0.16)	0.08 (0.06, 0.10)
Fig. 7	2.81 (2.41, 3.22)	2.80 (2.60, 3.01)	2.79 (2.68, 2.89)	0.23 (0.16, 0.31)	0.21 (0.18, 0.25)	0.21 (0.19, 0.22)
Fig. 8	3.12 (1.93, 4.07)	3.05 (2.42, 3.68)	2.98 (2.73, 3.23)	0.03 (0.01, 0.06)	0.03 (0.01, 0.04)	0.02 (0.02, 0.03)