Molecular Biology and Evolution, 2025, **42**, 1–18
https://doi.org/10.1093/molbev/msaf121
Advance access publication 27 June 2025
Methods

MBE OXFORD

Inference of Gene Flow between Species from Genomic Data When the Mode, Direction, and Lineages are Misspecified

Yuttapong Thawornwattana [1,2,*], **Tomáš Flouri** [2], **James Mallet** [1], **Ziheng Yang** [2,*]

[1]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA
[2]Department of Genetics, Evolution, and Environment, University College London, Gower Street, London WC1E 6BT, UK

***Corresponding authors**: E-mails: yuttapong.thawornwattana.09@ucl.ac.uk; z.yang@ucl.ac.uk.
Associate editor: Russell Corbett-Detig

Abstract

Thanks to genomic data, interspecific gene flow is increasingly recognized as a major evolutionary force that shapes biodiversity. Two models have been developed in the multispecies coalescent (MSC) framework to infer gene flow from genomic data, assuming either constant-rate continuous migration (MSC-M) or discrete introgression/hybridization (MSC-I). The extreme simplicity of these models raises concerns about their usefulness as they represent misspecified models when applied to real data. Here, we study inference of gene flow under the MSC-M model, considering mis-assignment of gene flow onto incorrect parental or daughter lineages, misspecification of the direction of gene flow, and misspecification of the mode of gene flow. Mis-assignment of gene flow to an incorrect lineage causes large biases in the estimated rates. The Bayesian test has high power for inferring both recent and ancient gene flow, between either sister lineages or nonsister lineages, although misspecification of the direction of gene flow may make it hard to distinguish early divergence with gene flow from recent complete isolation. Misspecification of the mode of gene flow (MSC-I versus MSC-M) has small local effects, and gene flow is detected with high power despite the misspecification. We analyze a genomic dataset from the purple cone spruce (*Picea* spp., Pinaceae), which putatively arose through homoploid hybrid speciation, to demonstrate practical implications of our theoretical analyses. Overall, we find that the extremely idealized models of gene flow (in particular the discrete MSC-I model) are very effective for extracting information about species divergence and gene flow from genomic data.

Keywords: gene flow, introgression, migration, multispecies coalescent, model misspecification, BPP.

Introduction

Gene flow between species is an important process that shapes biodiversity we observe today. In the past two decades, genomic data have been widely used to detect gene flow and have considerably enriched our understanding of the role of gene flow in speciation and adaptation (Mallet et al. 2016). Commonly used methods for detecting gene flow and estimating its rate from genomic data are approximate, based on species triplets (or quartets if an outgroup is used) and/or rely on summaries of sequence data, such as genome-wide site pattern counts (e.g. the *D*-statistic and HyDe, Green et al. 2010; Kubatko and Chifman 2019), estimated gene tree topologies (e.g. SNaQ, Solis-Lemus and Ane 2016; Jackson et al. 2017), or joint site frequency spectra (e.g. $\delta a \delta i$ and fastsimcoal2, Gutenkunst et al. 2009; Excoffier et al. 2021). Those methods do not make a full use of information in the data and often have reduced power to detect gene flow (Flouri et al. 2020; Ji et al. 2023; Pang and Zhang 2024; Ji et al. 2025). For example, most triplet summary methods cannot infer gene flow between sister lineages or identify the direction of gene flow (Jiao et al. 2021; Huang et al. 2022; Ji et al. 2023).

In this work, we focus on full-likelihood methods of inference under the multispecies coalescent (MSC) model, extended to account for gene flow (Rannala and Yang 2003; Jiao et al. 2021), applied to sequence data. Two approaches are commonly used for generating multilocus data, (i) sampling of short genomic fragments from sequenced genomes (e.g. Finger et al. 2022; Thawornwattana et al. 2022) and (ii) targeted sequence capture generating the so-called reduced representation data, including RADseqs (Andrews et al. 2016; Leaché and Oaks 2017), exome or transcriptome sequencing, ultraconserved elements (UCEs, Faircloth et al. 2012), anchored hybrid enrichment (AHE, Lemmon et al. 2012), conserved nonexonic elements (CNEEs, Edwards et al. 2017), and rapidly evolving long exon capture (RELEC, Karin et al. 2020). We refer to genomic fragments generated using either strategy as loci (irrespective of whether they are protein-coding). The MSC model assumes no recombination within each locus and free recombination between loci (see Zhu et al. 2022; Yan et al. 2023 for simulations that examine the impact of recombination on inference under the MSC models).

Two idealized modes of gene flow have been modeled in the MSC framework (Jiao et al. 2021). First, in the MSC-with-introgression (MSC-I) model (Flouri et al. 2020), also known as the multispecies network coalescent (MSNC; Yu et al. 2012; Wen et al. 2016; Wen and Nakhleh 2018; Zhang et al. 2018) or network multispecies coalescent (NMSC) model (Ané et al. 2024), gene flow is a discrete event and occurs in a pulse at a specific time point. The amount of gene flow is measured by the introgression probability, $\varphi_{A\to B}$, which represents the proportion of migrants in B from A at the time of introgression (Fig. 1a).

Second, the MSC-with-migration (MSC-M) model (Flouri et al. 2023) assumes that gene flow is continuous and occurs

Received: February 16, 2025. **Revised:** May 8, 2025. **Accepted:** May 14, 2025

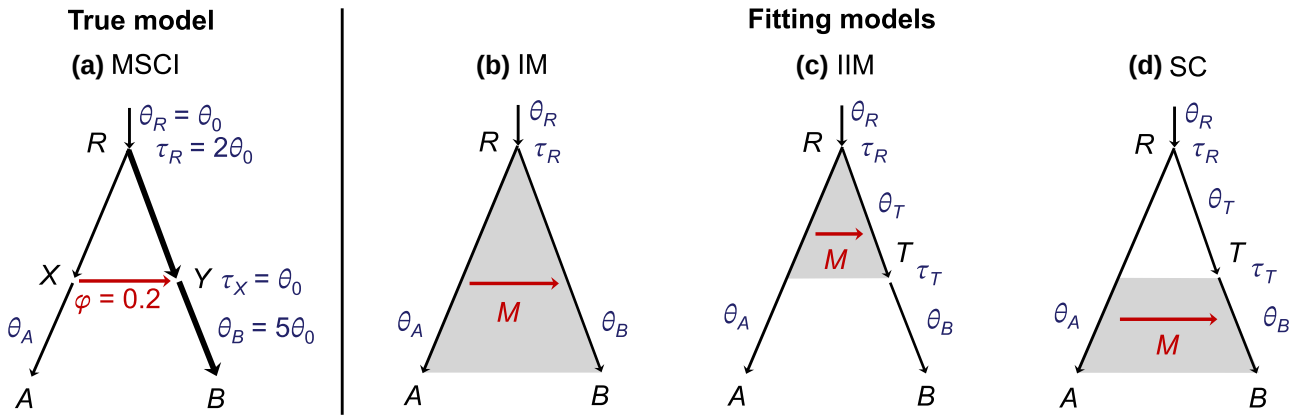

Fig. 1. a) An MSC-I model for two species (*A*, *B*), with introgression from *A* to *B* at time τ_X with introgression probability φ, used to generate data. We assume that population sizes of *A* and *B* do not change at the time of introgression (i.e. $\theta_A = \theta_X$, $\theta_B = \theta_Y$). Thus, the parameter vector is $\boldsymbol{\theta}_{\text{I}} = \{\varphi, \tau_X, \tau_R, \theta_A, \theta_B, \theta_R\}$. Each branch has an associated population size parameter $\theta = 4N\mu$, where N is the effective population size and μ is the mutation rate per site per generation. Time is measured as the expected number of mutations per site, with $\tau = T\mu$, where T is the divergence time in generations. Parameter values used in simulation are as shown: $\theta_A = \theta_X = \theta_R = \theta_0$ (thin branches), $\theta_B = \theta_Y = 5\theta_0$ (thick branches), $\tau_X = \theta_0$, $\tau_R = 2\theta_0$ and $\varphi = 0.2$, with $\theta_0 = 0.002$. b–d) Three MSC-M models used to analyze the data: IM (isolation with migration), IIM (isolation with initial migration), and SC (secondary contact). Gray shading indicates a period of continuous gene flow from *A* to *B* at rate $M_{AB} = N_B m_{AB} \equiv M$ migrants per generation, where m_{AB} is the proportion of migrants in *B* from *A* per generation. The parameter vector of the IM model is $\boldsymbol{\theta}_{\text{IM}} = \{M, \tau_R, \theta_A, \theta_B, \theta_R\}$, while those for IIM and SC are $\boldsymbol{\theta}_{\text{IIM}} = \boldsymbol{\theta}_{\text{SC}} = \{M, \tau_R, \tau_T, \theta_A, \theta_B, \theta_T, \theta_R\}$. The IIM and SC models are implemented in BPP as instances of the MSC-M model by including an unsampled ghost species that is sister to *B* and diverged with *B* at time τ_T. This creates two θ parameters for branches *RT* and *TB* as the current version of BPP does not implement the constraint $\theta_T = \theta_B$.

at a constant rate per generation over an extended time period. Gene flow from populations *A* to *B* is measured by the population migration rate $M_{A\to B} = N_B m_{A\to B}$, which is the expected number of individuals in *B* that are migrants from *A* per generation, with N_B to be the effective population size of *B*, and $m_{A\to B}$ the proportion of individuals in *B* that have migrated from *A*. Note that we use the real-world view with time running forward when defining gene-flow parameters. The MSC-M model includes the isolation-with-migration (IM) model (Nielsen and Wakeley 2001; Hey and Nielsen 2004; Hey 2010; Zhu and Yang 2012; Dalquen et al. 2017; Hey et al. 2018; Jones 2019) (Fig. 1b). Variants of MSC-M models include the isolation-with-initial-migration (IIM) model (Fig. 1c; Costa and Wilkinson-Herbots 2017) and the secondary contact (SC) model (Fig. 1d; Costa and Wilkinson-Herbots 2021). In this paper we use the terms "introgression" to refer to pulse gene flow in the MSC-I model and "migration" for continuous gene flow in the MSC-M model.

The MSC-I and MSC-M models may be viewed as extreme special cases of a general model with variable rates of gene flow over time. In the real world, the rate of gene flow may be expected to vary over time as the geographical distribution of species expand or shrink, impacting their opportunities to meet and hybridize, and as the intensity of natural selection purging introgressed alleles fluctuates over time, influenced by multiple factors including recombination and genetic drift (Martin and Jiggins 2017; Moran et al. 2021). While both the MSC-I and MSC-M models are wrong when applied to genomic data, it is interesting to know whether they produce similar inferences of gene flow (e.g. the lineages and direction of gene flow) or similar estimates of key population parameters such as species divergence times and rates of gene flow. How robust are parameter estimates to the assumed mode of gene flow? Similarly will we infer gene flow if the introgression events are incorrectly assigned to parental or daughter lineages to lineages genuinely involved in gene flow? Will we detect gene flow if the direction of gene flow is misspecified?

Here we address those questions. We investigate the impact of three kinds of model specifications on Bayesian inference of gene flow using genomic data:

1. The mode of gene flow might be incorrectly assumed. For instance, gene flow might have occurred as a single pulse of introgression or hybridization (MSC-I) but continuous gene flow (MSC-M) is assumed in data analysis.
2. Lineages involved in gene flow may be incorrectly specified. Currently, it is very challenging to assign gene-flow events to branches in a species phylogeny (Pang and Zhang 2024). For example, when gene flow is detected in many species triplets and is assigned to ancestral branches using criteria such as *f*-branch (Malinsky et al. 2018), it may be assigned incorrectly to parental or daughter branches (Suvorov et al. 2022; Ji et al. 2023; Thawornwattana et al. 2023b).
3. The direction of gene flow may be misspecified. Indeed, most summary methods that rely on gene tree topologies or site-pattern counts cannot identify the direction of gene flow (Hibbins and Hahn 2022).

In this paper, we do not consider the scenario of ghost introgression involving a source population that has gone extinct or is not sampled in the data (Huang et al. 2022; Tricou et al. 2022; Pang and Zhang 2023, 2024). Furthermore, we do not consider the search in the space of all possible models of gene flow given a set of species, either with or without the species tree fixed. Currently, inference of gene-flow models is a challenging task for both summary and full likelihood methods. While MCMC algorithms are implemented to update the MSC-I model in the programs PHYLONET/MCMC_SEQ (Wen and Nakhleh 2018) and *BEAST (Zhang et al. 2018), the implementations are computationally unfeasible except for very small datasets with <100 loci. The programs *δaδi* (Gutenkunst et al. 2009) and FASTSIMCOAL2 (Excoffier et al.

2021) can be used to compare candidate gene-flow models. These methods use data of joint site frequency spectra at SNP sites, which are genome-wide averages, and ignore information in the variation in genealogical relationships across the genome. Such data have fundamental limits in information content when used to infer the demographic history of one species (Terhorst and Song 2015; Baharian and Gravel 2018). A recent paper demonstrated that surprisingly commonly used summary methods such as the *D*-statistic (Green et al. 2010), HyDe (Kubatko and Chifman 2019), SNaQ (Solis-Lemus and Ane 2016; Jackson et al. 2017), and PhyloNet/MPL (Yu et al. 2012; Yu and Nakhleh 2015) do not have the capability to distinguish among different models (such as inflow, outflow, and ghost introgression); in other words, the different gene-flow models are unidentifiable by these methods (Pang and Zhang 2024). Currently, choice of gene-flow models is an area of active research.

Thus, we limit the scope of our study to the three kinds of model specifications identified above. We characterize the power and false positives of Bayesian tests of gene flow, and bias in estimation of the rate of gene flow under model misspecification. We use asymptotic analysis to deal with infinite data under simple scenarios and computer simulation to consider finite datasets under more complex scenarios. We corroborate our theoretical analyses by analyzing a genomic dataset from the purple cone spruce under the MSC-I and MSC-M models. We use the Bayesian program bpp because of its computational efficiency (Flouri et al. 2018, 2023), but the results should apply to other full-likelihood methods as well (e.g. IMa3, Hey et al. 2018). Because full-likelihood methods utilize all information in the data concerning the model and parameters, whereas summary methods make use of only a portion of that information, our results should also shed light on the behaviors and limitations of summary methods under similar situations. This work complements our previous studies on model misspecification where the MSC-I model was assumed in analysis of data generated under the MSC-M model (Jiao et al. 2020; Huang et al. 2022) or when the direction of introgression was misspecified under the MSC-I model (Thawornwattana et al. 2023a). We summarize results from both this study and the previous studies in the Conclusions section.

Results and Discussion

The Case of Two Species

Suppose that gene flow occurs from *A* to *B* at time τ_X with introgression probability φ (Fig. 1a) but we analyze the data under MSC-M models assuming continuous migration over an extended time period (Fig. 1b–d). What will the estimate of the migration rate (M) be like? Will we detect gene flow despite misspecification of the mode of gene flow? Similarly, what are the effects of misspecified direction of gene flow? We approach these questions using a combination of asymptotic analysis (as the number of loci approaches infinity) and computer simulation, following Jiao et al. (2020) and Huang et al. (2022). The asymptotic analysis is tractable in the simple case of two species with one sequence per species per locus, while simulations can be performed for any number of species, any number of sequences per species, and any finite number of loci.

Asymptotic Theory in the Two-species Case

We develop an asymptotic theory of maximum-likelihood (ML) estimation for the simple model of gene flow for two species, with introgression from *A* and *B* (Fig. 1a), and with data of two sequences (one from each species) per locus. The true model is MSC-I, with parameter vector $\boldsymbol{\theta}_\mathrm{i}$ (Fig. 1a), and the data are analyzed to estimate the parameter vector $\boldsymbol{\theta}_\mathrm{m}$ under three variants of the MSC-M model: isolation with migration (IM), isolation with initial migration (IIM), and SC (Fig. 1b–d). Note that when we define parameters of introgression probability or migration rate time runs forward. As there is only one sequence per species per locus, θ_B is not used in MSC-I, and θ_B and θ_T are unidentifiable under any of the MSC-M models. Thus, the parameter vector for the true MSC-I model is $\boldsymbol{\theta}_\mathrm{i} = \{\varphi, \tau_X, \tau_R, \theta_A, \theta_R\}$ (Fig. 1a), while that for the fitting MSC-M model is $\boldsymbol{\theta}_\mathrm{IM} = \{M, \tau_R, \theta_A, \theta_R\}$ or $\boldsymbol{\theta}_\mathrm{IIM} = \boldsymbol{\theta}_\mathrm{SC} = \{M, \tau_T, \tau_R, \theta_A, \theta_R\}$ (Fig. 1b–d). Later, we analyze simulated data with multiple sequences sampled per species per locus, which allow estimation of the full parameter set (Fig. 1).

Consider an infinite number of loci, each with two sequences (*a* and *b*) of length *n*. We assume the infinite-sites mutation model, so that the sequence data at each locus is summarized as *x* differences out of *n* sites. The probability for *x* is given by averaging over the unobserved coalescent time *t* between the two sequences,

$$f(x;\boldsymbol{\theta}) = \int_0^\infty f(x\,|\,t) f(t;\boldsymbol{\theta})\,\mathrm{d}t. \tag{1}$$

The density of coalescent time, $f(t;\boldsymbol{\theta})$, depends on the model of gene flow, and is given in SI text for the MSC-I model of Fig. 1a and the MSC-M models of Fig. 1b–d. Given the coalescent time *t*, the expected number of mutations at the locus is $n \times 2t$, so that $f(x\,|\,t)$ is given by the Poisson probability

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

This leads to a closed-form expression of $f(x;\boldsymbol{\theta})$, as in Huang et al. (2022).

Under the MSC models, data at different loci are independently and identically distributed (i.i.d.). When the number of loci $L \to \infty$, the MLE $\hat{\boldsymbol{\theta}}_\mathrm{m}$ under the wrong MSC-M model converges to $\boldsymbol{\theta}^*_\mathrm{m}$, which minimizes the Kullback–Leibler (KL) divergence

$$D(\boldsymbol{\theta}_\mathrm{i} \parallel \boldsymbol{\theta}_\mathrm{m}) = \sum_{x=0}^{n} f_\mathrm{i}(x;\boldsymbol{\theta}_\mathrm{i}) \log \frac{f_\mathrm{i}(x;\boldsymbol{\theta}_\mathrm{i})}{f_\mathrm{m}(x;\boldsymbol{\theta}_\mathrm{m})}. \tag{3}$$

Here the subscripts "i" and "m" specify the model under which the parameters and probabilities are defined. In effect $\boldsymbol{\theta}_\mathrm{i}$ in MSC-I is fixed and $f_\mathrm{i}(x;\boldsymbol{\theta}_\mathrm{i})$ represents the data while $f_\mathrm{m}(x;\boldsymbol{\theta}_\mathrm{m})$ represents the fitting model, and $\boldsymbol{\theta}_\mathrm{m}$ is estimated by minimizing *D*. The estimate $\boldsymbol{\theta}^*_\mathrm{m}$ is known as the *pseudo-true parameter value* or the *best-fitting parameter value* under the fitting MSC-M model. Note that when $L \to \infty$, the Bayesian estimate (the posterior mean) approaches the same limit ($\boldsymbol{\theta}^*_\mathrm{m}$) as the ML estimate (MLE). Because of model mismatch, a perfect fit is impossible, so that $D > 0$.

Of particular interest is the correspondence between the introgression probability φ in MSC-I and the migration rate M in MSC-M. Under MSC-M, the probability that a lineage from species *B* traces back to *A* (irrespective of the migration time), when one traces the genealogical history of the sampled sequences backwards in time, is

$$\varphi_0 = 1 - \mathrm{e}^{-\frac{4M}{\theta_B}\Delta\tau}, \tag{4}$$

where $\Delta\tau$ is the time period of migration (Huang et al. 2022). At the mutational time scale used here, migration occurs at the

Poisson rate of $4M/\theta_B$, and equation (4) is the cumulative distribution function of the exponential waiting time until migration. Note that the introgression probability φ in MSC-I is also the probability that a lineage from species B traces back to A (at time τ_X). Thus both φ in MSC-I and φ_0 in MSC-M measure the expected total amount of gene flow.

Inverting equation (4) gives

$$M_0 = \frac{\theta_B}{4\Delta\tau}\log\left(\frac{1}{1-\varphi}\right). \quad (5)$$

The Limiting Values of the MLEs in the Two-species Case

We use the above theory to study the asymptotic behavior of parameter estimation under the MSC-M models of Fig. 1b–d when the data are generated under the MSC-I model (Fig. 1a). We used two population sizes on the species tree, $\theta_0 = 0.002$ for the thin branches and $\theta_1 = 0.01$ for the thick branches (Fig. 1a). The species divergence time is $\tau_R = 2\theta_0$ while introgression occurs at time $\tau_X = \theta_0$. In our simulation, the divergence times (τs) and population-size parameters (θs) are proportional. This mimics the use of different types of genomic data with different neutral mutation rates (e.g. exons versus noncoding DNA). We varied the sequence length (n) and the introgression probability (φ).

The limits of the MLE ($\theta^*_{\rm m}$, equation (3)) are shown in Fig. 2. The true and best-fitting distributions of the coalescent time $t_{ab} = t$ are in supplementary fig. S1, Supplementary Material online, with the achieved KL values in supplementary fig. S2, Supplementary Material online.

Among the three MSC-M models of Fig. 1b–d, the IIM model provide the most sensible parameter estimates, with M^* tracking the introgression probability (φ) in the MSC-I model (Fig. 2) and with the lowest KL divergence (supplementary fig. S2, Supplementary Material online). The estimates τ^*_R and $\theta^*_R = \theta^*_A$ also match the true values. This means coalescence in the ancestral population (R) is correctly accounted for. The time at which migration stops (τ^*_T) is slightly younger than the actual introgression time ($\tau^*_T = 0.0015$–$0.0016 < \tau_X = 0.002$). Under the MSC-I model, gene flow results in a peak in the density of the coalescent time t_{ab} at τ_X (supplementary fig. S1, Supplementary Material online, black curve). By contrast, coalescence due to migration under the IIM model peaks in the middle of the migration period (supplementary fig. S1, Supplementary Material online, dark blue curve, middle column). Thus having $\tau^*_T < \tau_X$ gives a better fit.

The IM and SC models produce similar best-fitting parameter values, different from IIM (Fig. 2). Under the SC model, the time at which migration starts (τ^*_T) is often close to the divergence time (τ^*_R), making it similar to the IM model. At small values of φ, the estimated migration rate M^* under IM and SC matches closely the expected value M_0 (calculated using equation (5) using the migration period 0–τ_R and the true population size θ_B; Fig. 2, first column) and the species divergence time τ^*_R matches the true value as the number of sites approaches infinity. At large φ (say $\varphi > 0.5$), M is seriously underestimated, accompanied by an underestimation of τ_R (with $\tau^*_R \approx \tau_X$) and an overestimation of θ (Fig. 2 and supplementary fig. S1, Supplementary Material online).

Both IM and SC assume continuous and ongoing migration and, at a high M, predict presence of recent coalescence with $t_{ab} \approx 0$ or nearly identical sequences from the two species. In the data (generated under the MSC-I model with all migration occurring at time $\tau_X > 0$), such recent coalescence with $t_{ab} \approx 0$ is absent and nearly identical sequences between species are uncommon. The rarity of nearly identical sequences between species in the data is then hard to reconcile with the IM and SC models with a high migration rate, especially when the sequence is long (large n) (Fig. 2 and supplementary fig. S1, Supplementary Material online). As a result, the models underestimate both M and τ_R, and in effect explain recent coalescence (small t_{ab} due to introgression at time τ_X) as coalescence in the common ancestor R (with $\tau^*_R \approx \tau_X$).

Simulation Results in the Two-species Case

The asymptotic analysis assumes one sequence per species per locus. To accommodate multiple sequences per species, we used simulation. Data are simulated under the MSC-I model (Fig. 1a) and analyzed under the MSC-M models (IM, IIM, and SC; Fig. 1b–d), using BPP. The Markov chain Monte Carlo (MCMC) algorithm in BPP averages over the gene genealogy underlying the sequence alignment at each locus, similar to the integration in equation (1) averaging over the coalescent time t in the case of two sequences. We assume the JC mutation model (Jukes and Cantor 1969). In the base case, each dataset consisted of $L = 4{,}000$ loci, with $S = 4$ sequences per species per locus, and $n = 1{,}000$ sites per sequence, and the introgression probability is set at $\varphi = 0.2$. Then we varied the number of sites per sequence (n), the number of sequences per species (S), the number of loci per species (L), and the introgression probability (φ). The first three factors are related to data size while the last is a parameter that measures the amount of gene flow. We are interested in how these factors influence posterior estimates of parameters, in particular the migration rate M. The posterior means and 95% HPD CIs for parameters in the MSC-M models are summarized in Fig. 3, while the true and best-fitting distributions of the coalescent times (t_{ab}, t_{aa}, t_{bb}) are in supplementary figs. S3–S6, Supplementary Material online.

First, we consider the sequence length (n; Fig. 3, first column). This has little impact on the posterior means and highest-probability-density (HPD) credibility intervals (CIs) for parameters θ_A, θ_B, θ_R, and τ_R under all three models, or on θ_T and τ_T under the IIM model. Even with short sequences ($n = 250$), those parameters are precisely estimated. However, use of longer sequences improves the precision in the estimated migration rate M under the IIM model. At $n = 64{,}000$ sites, the estimate is $\hat{M} = 0.37$, close to the limiting value $M^* = 0.33$ from our asymptotic analysis for data of infinitely many loci of two sequences under the assumption of equal population sizes (see above). Under the IM or SC models, we obtained much smaller estimates, with $\hat{M}$ increasing from 0.05 at $n = 250$ to 0.13 at $n = 64{,}000$, accompanied by an increase in $\hat{\tau}_T$ in the SC model, which converges to τ_R at large values of n (Fig. 3; also see supplementary fig. S3, Supplementary Material online). Thus, the SC model converges to the IM model when $n \to \infty$. Those results agree with our asymptotic results (Fig. 2). The increased rate ($\hat{M}$) and duration ($\hat{\tau}_T$) of gene flow in the SC model with the increase of n may be explained by the increasingly stronger evidence of gene flow in longer sequences. The large $\hat{M}$ also improves $\hat{\theta}_T$ because migration helps explain large variation in the recipient population B caused by introgression from A in the true MSC-I model (see the improved fit to the coalescent time t_{bb} at large n in supplementary fig. S3, Supplementary Material online).

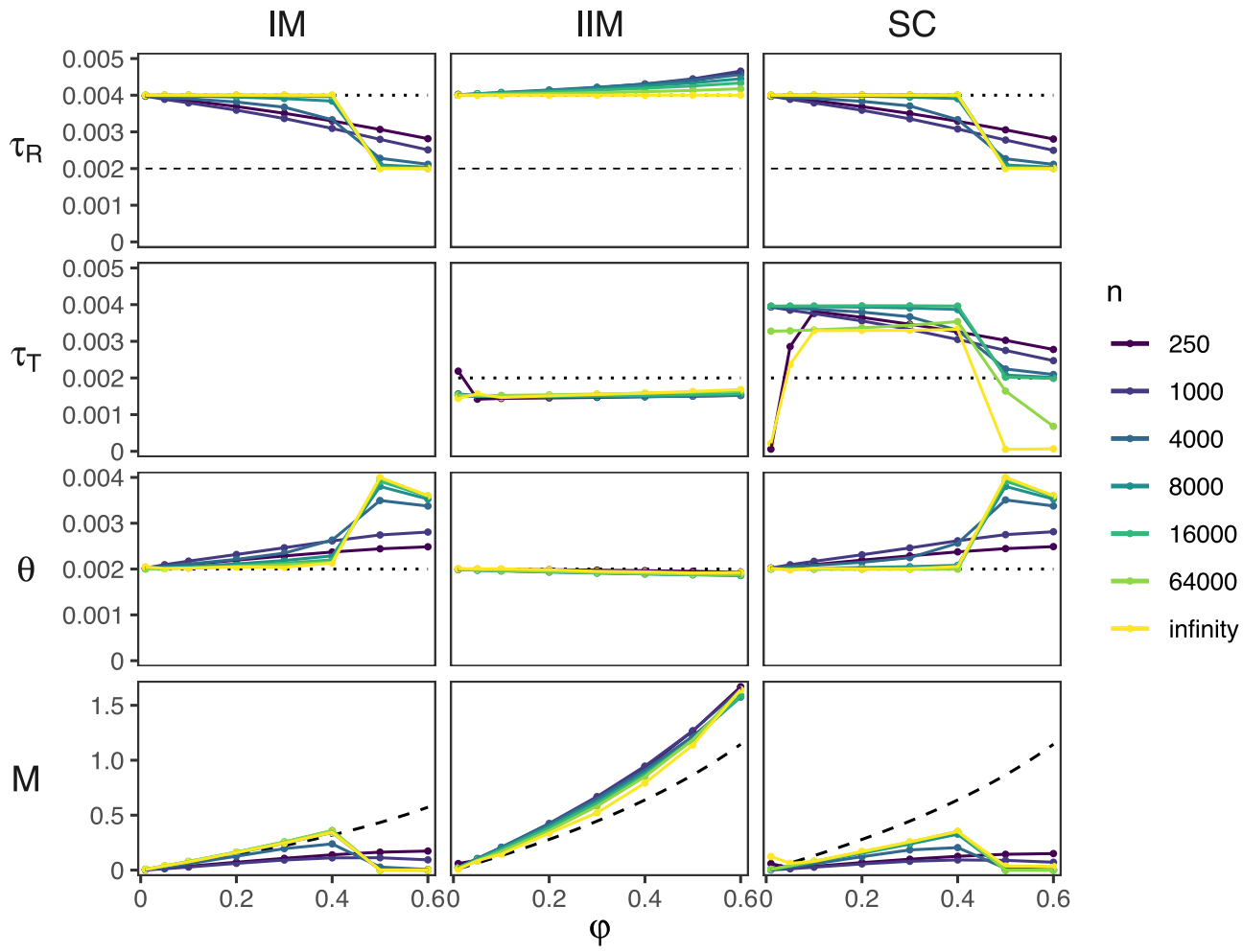

Fig. 2. Best-fitting parameter values from the asymptotic analysis under the IM, IIM and SC models (Fig. 1b–d) of data generated under the MSC-I model (Fig. 1a). In effect, the dataset consists of infinitely many loci, each with two sequences (one from each species) of n sites. The two population sizes, θ_A and θ_R, were constrained to be equal, denoted by θ. Horizontal dotted lines indicate true values. For τ_R, the dashed line indicates the introgression time τ_X in the MSC-I model. For M, dashed curves indicate the expected value M_0 based on equation (5), calculated using the true θ_B and the expected time duration of migration ($\Delta\tau$), which is τ_R for IM, $\tau_R - \tau_X$ for IIM, and τ_X for SC. The true and best-fitting distributions of the coalescent time (t) are in supplementary fig. S1, Supplementary Material online.

Under the IIM model, the time when migration ends (τ_T) is estimated to be about 0.0015, as predicted by our asymptotic analysis of both finite and infinitely long sequences (Fig. 2), while the true introgression time τ_X is 0.002. One might expect $\hat{\tau}_T$ (the time at which the migration period ends) to converge to τ_X since this is the smallest time at which sequences from A and B can coalesce (i.e. the smallest t_{ab}). We obtain $\hat{\tau}_T < \tau_X$. This may be partly attributable to the difference in how the MSC-I and the IM models account for the reduced t_{ab} due to gene flow. The probability density of t_{ab} peaks at the introgression time τ_X in the MSC-I model while it peaks in the middle of the migration period (τ_T, τ_R) in the IIM model (supplementary fig. S1, Supplementary Material online). Thus having a migration period that ends after τ_X better accommodates t_{ab} in the data. In summary, the IIM model is able to detect more gene flow and provides more precise and accurate estimates of population sizes and divergence times even with short sequences ($n = 250$) while the IM and SC models require at least 4,000 sites per locus to be able to detect substantial amounts of gene flow (Fig. 3 for M against n).

Second, we consider the effects of the number of sequences per species (S; Fig. 3, second column). Overall, estimation under all three MSC-M models benefited considerably from including multiple samples per species (with $S > 1$). With only one sequence per species ($S = 1$), no coalescent events can occur in B and T, so that θ_B in IIM and θ_T in SC are unidentifiable, and other parameters such as θ_A, θ_T, τ_T, M have large uncertainties as well, although τ_R and θ_R are well estimated. When $S > 1$, all parameters in all three MSC-M models are identifiable, and furthermore, even those parameters that are identifiable at $S = 1$ have much narrower CIs (except for θ_T in SC for the same reason as in the case of varying n). As before, the IIM model recovered more gene flow, with $\hat{M} = 0.40$ (0.35, 0.45) at $S = 16$, in comparison with $\hat{M} < 0.1$ for IM and SC.

Third, the number of loci (L; Fig. 3, third column and supplementary fig. S5, Supplementary Material online) is the sample size in the model as data at different loci are i.i.d. Increasing L led to narrower CIs. In theory, the CI width should reduce by a half as L quadruples. This holds approximately for most parameters except for θ_T in the SC model, which is poorly estimated.

Lastly, we consider the impact of the amount of gene flow in the data (φ; Fig. 3, last column). For all analysis models, the

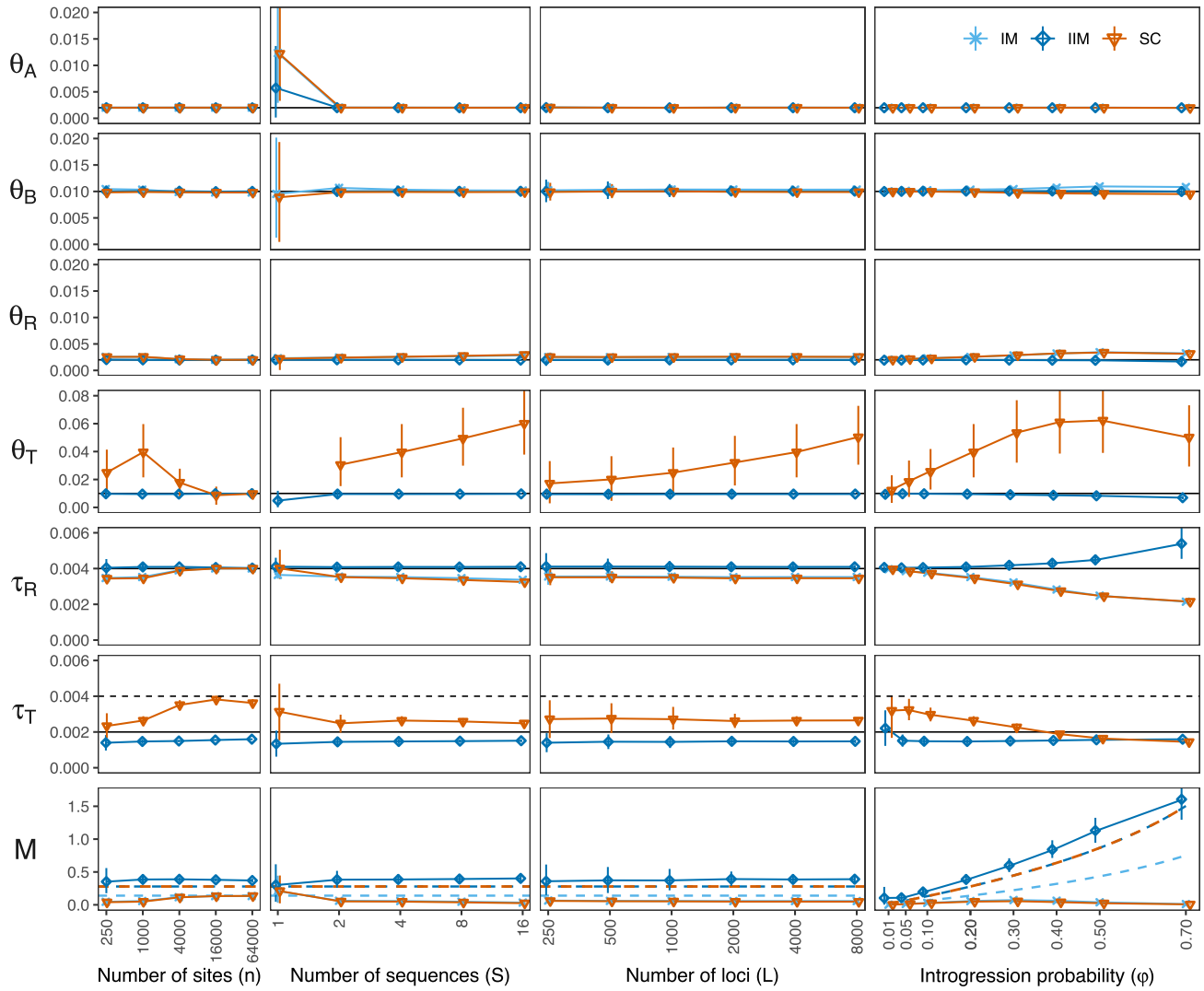

Fig. 3. Parameter estimates under the three MSC-M models (IM, IIM, and SC; Fig. 1b–d) obtained from BPP analysis of data generated under the MSC-I model (Fig. 1a), summarized as posterior means and 95% HPD CIs averaged over 30 replicate datasets. In the base case, each dataset consists of $L = 4{,}000$ loci, with $S = 4$ sequences per species at each locus and $n = 1{,}000$ sites per sequence. Parameters in the MSC-I model are given in the legend to Fig. 1. We varied four factors one at a time, keeping other factors fixed at the base case: the number of sites per sequence (n), the number of sequences per species (S), the number of loci (L), and the introgression probability (φ). The parameters θ_T and τ_T are specific to the IIM and SC models. When $S = 1$, θ_B is unidentifiable in the IIM model (Fig. 1c), and θ_T is unidentifiable in the SC model (Fig. 1d). Horizontal solid lines indicate the true values in the MSC-I model used to generate the data. For τ_T, the horizontal dotted line indicates τ_R, the upper limit of τ_T. For M, a dashed curve indicates the expected value M_0 based on equation (5), assuming the true θ_B and the expected duration of migration $\Delta\tau$; see legend to Fig. 2. The x-axes for n, S, and L are on a logarithmic scale.

extant (θ_A, θ_B) and ancestral (θ_R) population sizes are well-estimated, with the posterior mean close to the true value and with narrow CIs. Consistent with our asymptotic results, only the IIM model is able to estimate M that increases with φ. However, due to the model misspecification, as φ increases, the IIM model increasingly overestimate τ_R while $\hat{\tau}_T$ stays largely unchanged, resulting in an increasingly long period of migration ($\hat{\tau}_T$, $\hat{\tau}_R$). In effect deep coalescent events between sequences from A and B in R are being mis-interpreted as a result of migration after species divergence (supplementary fig. S6, Supplementary Material online). By contrast, the IM and SC models only detect small amounts of gene flow ($\hat{M} < 0.1$) regardless of the true value of φ (Fig. 3). In the MSC-I model, larger values of φ lead to smaller t_{ab}, with a peak at the introgression time (τ_X). The IM and SC models accommodate small t_{ab} in the data as coalescence in the ancestral population, with $\hat{\tau}_R$ gradually decreasing from τ_R to τ_X as φ increases. This reduction in $\hat{\tau}_R$ is associated with an increase in $\hat{\theta}_R$. This pattern agrees with our asymptotic analysis (Fig. 2, $n = 1{,}000$), which predicts that $\hat{\tau}_R \to \tau_X$ as $n \to \infty$ and $L \to \infty$.

Why do the IM and SC models detect much less gene flow than the IIM model across a wide range of values of φ? Those two models assume ongoing migration up to the present time and predict recent coalescent events between sequences from the two species (with $t_{ab} \approx 0$), but no such coalescence exists in the data or in the true MSC-I model. High migration rates (M) in the IM and SC models are thus incompatible with the data. Under the SC model, $\hat{\theta}_T$ is usually much larger than the true value and with a wide CI, and $\hat{M}$ is close to zero. This is because introgression from A into B in the data (generated under MSC-I) increases genetic variation in B. Here, with θ_B being well estimated and $\hat{M}$ being close to zero, having a large value of $\hat{\theta}_T$ helps explain genetic variation in B. This also explains why $\hat{\theta}_T$ peaks at $\varphi = 0.5$, where genetic variation in B is the largest under the MSC-I model (Fig. 3).

In summary, the simulation results for the case of two species agree with our asymptotic analysis of data of an infinite number of loci with one sequence per species (Fig. 2; $n = 1{,}000$). The IIM model provides the most sensible estimates and is able to recover approximately correct amounts of gene flow. However, τ_R is overestimated when φ is high. The IM and SC models are qualitatively similar: they can recover much less gene flow than the IIM model and require long sequences of at least $n = 4{,}000$ sites per locus to detect a reasonable amount of gene flow. While increasing the data size in any way (n, S, L) helps with the information content, including multiple sequences per species ($S > 1$) is particularly important.

The Bayesian test of gene flow. We also applied the Bayesian test of gene flow to analyze the simulated datasets of Fig. 3, with results shown in supplementary fig. S7, Supplementary Material online. The null hypothesis $H_0 : M_{A\to B} = 0$ and the alternative hypothesis $H_1 : M_{A\to B} > 0$ are compared using Bayes factors calculated using the Savage–Dickey density ratio (Ji et al. 2023). This formulation of the Bayes factor contrasts the prior and posterior probabilities that the migration rate is very low ($M < \epsilon = 0.001$) to assess the evidence in the data in support of gene flow. Overall, the test has high power, rejecting the null of no gene flow (with $B_{10} > 100$) in almost all datasets when the number of sites (n), the number of sequences (S) and the number of loci (L) varied around the base case (supplementary fig. S7, Supplementary Material online). At very low introgression probability ($\varphi = 0.01$, say), the test based on IM and SC has virtually no power while that based on IIM has full power. The situation is similar at very high introgression probabilities ($\varphi = 0.7$, say). Note that the MSC-I model with $\varphi = 1$ reduces to a model of complete isolation with no gene flow. Overall, the Bayesian test of gene flow, in particular the test based on the IIM model, has high power despite the misspecification of the mode of gene flow. Detecting gene flow through the test appeared to be a much easier task than estimating the amount of gene flow.

The Case of Four Species

We extend our simulation to more complex cases of four species with the phylogeny ([A, B, C], D), in which D is an outgroup, not involved in gene flow (Fig. 4a–d). Model C is an instance of the SC model while model D is an instance of the IIM model. Also gene flow is between nonsister species in model C and between sister species under model D. We simulate data under models A, B, C, or D and analyze them under models C and D. We refer to our simulation settings in the format of simulation model–analysis model. For example, in the A-C setting, data are simulated under model A and analyzed under model C (Fig. 4). Note that gene flow may be misspecified in two ways. First, the mode of gene flow may be misspecified (A-C and B–D settings). Second, gene flow may be assigned to a wrong branch (C-D and D-C settings). We also consider a combination of both kinds of misspecification (B-C and A-D settings).

The parameter values used (including species split times, introgression time, population size, and introgression probability) are in the legend to Fig. 4. As before each replicate dataset consists of $L = 250$, 1,000, or 4,000 loci, with $S = 4$ sequences per species per locus and with $n = 500$ sites in the sequence.

The posterior means and 95% HPD CIs of parameters are summarized in Fig. 4e (see also supplementary table S1, Supplementary Material online). The true and fitted distributions of coalescent times are in supplementary fig. S8, Supplementary Material online. We also conducted the Bayesian test of gene flow using the same data, with results summarized in supplementary fig. S9, Supplementary Material online. Our discussion below may refer to Huang et al. (2022, Fig. 4), where complementary results from using the MSC-I models to analyze data generated under the MSC-M models (e.g. the C-A, C-B, D-A, D-B settings) can be found.

Inference Under the Correct Model (Fig. 4, C-C and D-D)

The two cases where the analysis model matches the simulation model, C-C and D-D (Fig. 4e), represent the best-case scenarios and serve as a reference for comparison. First, we note that the Bayesian test of gene flow has full power in those two settings, even in the small datasets (supplementary fig. S9, Supplementary Material online, C-C and D-D). Second in both settings most parameters including the migration rate are well estimated, with narrow CIs covering the true parameter values (Fig. 4e, supplementary table S1, Supplementary Material online).

Parameters M, θ_S, τ_S, related to the gene-flow event, have wider CIs in the D-D setting than in the C-C setting. This may be due to two factors. First, it may be harder to estimate the rate of gene flow between sister lineages (D-D) than between nonsister lineages (C-C); for example, migration between nonsisters may cause a change to the gene-tree topology, making it easy for the method to identify migrant sequences. Second, it may be harder to estimate the rate of gene flow involving ancestral species since fewer sequences may reach the time of gene flow when one traces the history of sampled sequences backwards in time. While four sequences are from species B in the C-C setting, on average fewer than four sequences from species B and C reach the time of migration (τ_T) in the D-D setting; see, e.g. fig. S1 in Thawornwattana et al. 2023a) for time $T = 1$ or 3 coalescent units and note that here time is 2 coalescent units ($\tau_T = \theta_0$).

Inference When Gene Flow is Assigned to a Wrong Lineage (Fig. 4, C-D and D-C)

In the C-D setting, data are generated under model C with migration occurring after a period of isolation, but analyzed under model D with migration assigned to the wrong lineage of parental species. Parameters for populations far away from the migration event, such as the root divergence time and population size (τ_R, θ_R) and the outgroup population size (θ_D), are well estimated (Fig. 4, C-D). Other parameters have serious biases. Due to migration in the true model C, sequences from A are expected to be closer to those from B than to those from C, with $t_{ab} < t_{ac}$. However, the fitting model D predicts equal distances ($t_{ab} = t_{ac}$). There is a serious mismatch between the true and fitting models in the expected distributions of coalescent times (supplementary fig. S8, Supplementary Material online). Under model D, divergence time τ_S (and thus τ_T) are severely underestimated to accommodate the small t_{ab} in the data, with $\hat{\tau}_S = 0.0032$ and $\hat{\tau}_T = 0.0011$ (supplementary table S1, Supplementary Material online, C-D) while the true values are $\tau_S = 0.006$ and $\tau_T = 0.002$. With τ_S underestimated, θ_S is overestimated as well. The estimates of θ_B and θ_C are also affected, with $\hat{\theta}_B$ overestimated and $\hat{\theta}_C$ underestimated. Another conflict is that in the fitting model D, t_{ab} and t_{ac} have the same distribution while they

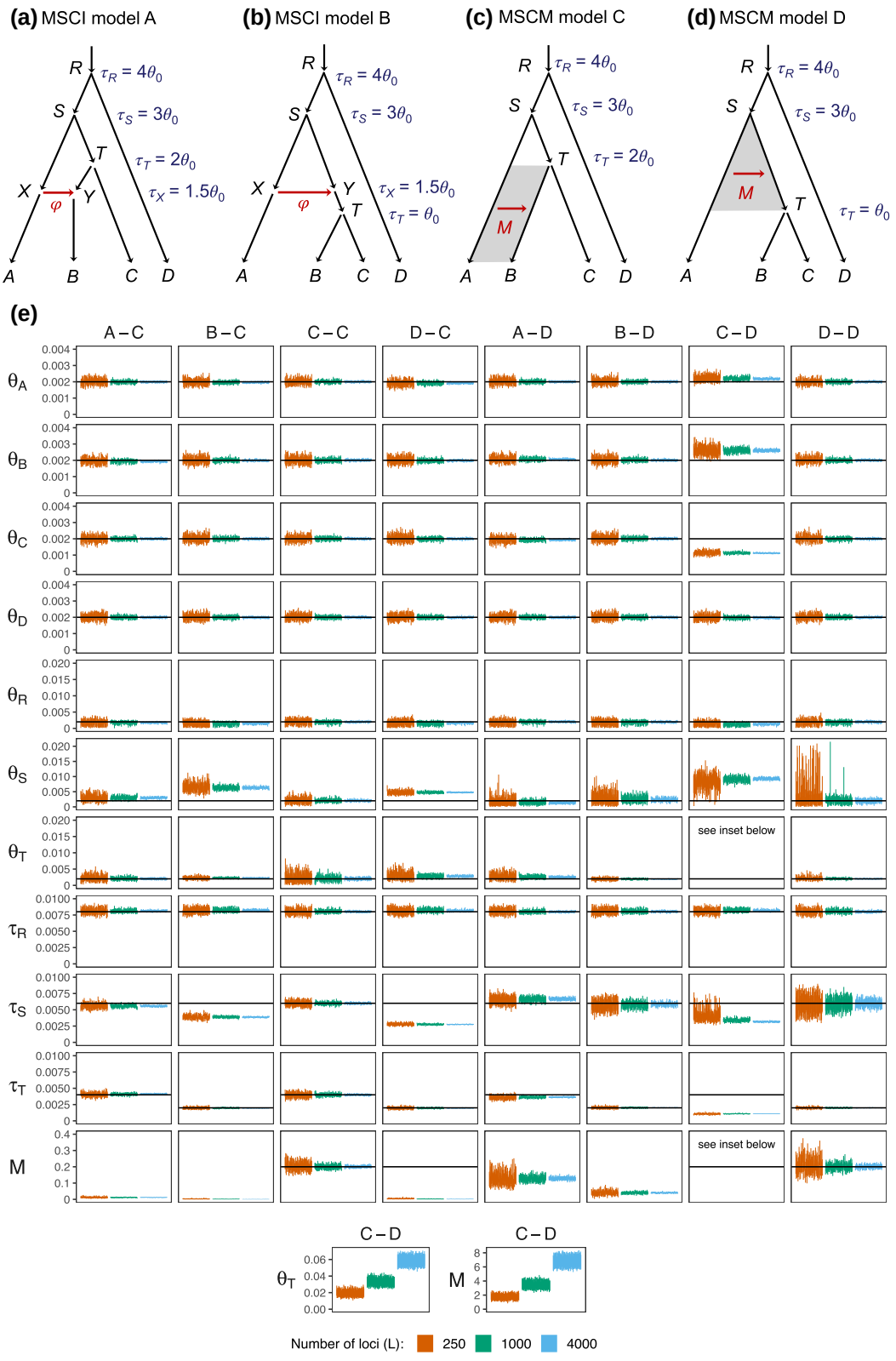

Fig. 4. a–b) Two introgression (MSC-I) models and c–d) two migration (MSC-M) models used in simulation. All branches have population size $\theta_0 = 0.002$. In MSC-I model A, the species divergence and introgression times are $\tau_R = 4\theta_0$, $\tau_S = 3\theta_0$, $\tau_T = 2\theta_0$, and $\tau_X = \tau_Y = 1.5\theta_0$. In MSC-I 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 MSC-M 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 during $(0, \tau_T)$ at rate $M = 0.2$ migrants per generation. In MSC-M model D, $\tau_R = 4\theta_0$, $\tau_S = 3\theta_0$, and $\tau_T = \theta_0$, with migration from A to T during (τ_T, τ_S) at rate $M = 0.2$. e) The 95% HPD CIs of parameters from 100 replicate datasets of $L = 250$, 1,000, and 4,000 loci. Column labels refer to the simulation model followed by the analysis model; e.g. "A-C" means the data were simulated under model A and analyzed under model C. Black solid lines indicate the true value. Estimates of θ_T and M for the C-D setting are shown separately at the bottom due to their extreme values.

are different under the true model C (supplementary fig. S8, Supplementary Material online, third row). This leads to a poor fit of t_{ac}. Estimates of the migration rate and recipient population size ($\hat{M}$ and $\hat{\theta}_T$) are unreasonably large. However, their ratio or the mutation-scaled migration rate $M/\theta_T = m/\mu$ is much better estimated, with $\hat{M}/\hat{\theta}_T = 90.4$, 105.9, and 116.3 for $L = 250$, 1,000 and 4,000, respectively, compared with the true value in model C of $M/\theta_T = 100$ (supplementary table S1, Supplementary Material online). This suggests that the proportion of migrants (m) has a greater impact on the distribution of gene trees and coalescent times than the number of migrants ($M = Nm$).

In the D-C setting, migration occurs initially after species divergence (i.e. IIM) but the analysis model assumes SC, with gene flow mis-assigned to the wrong daughter lineage. Population sizes of the modern species (θ_A, θ_B, θ_C, θ_D) are very well estimated, as are the population size and age of the root (θ_R, τ_R) and the divergence time τ_T between B and C. However, τ_S and θ_S are grossly biased (Fig. 4e). The estimated migration rate is very low ($\hat{M}^{\mathrm{D-C}} = 0.0011$; supplementary table S1, Supplementary Material online). As a result, there is an excess of coalescent times t_{ab} and t_{ac}, and a deficit in t_{bc}, over the time interval (τ_T, τ_S) in the data (caused by coalescent events between A and T in the true model) (supplementary fig. S8, Supplementary Material online). These are explained in the fitting model D by having a very recent divergence time τ_S between A and T, which is estimated to be close to $\hat{\tau}_T^{\mathrm{D-C}}$, with $\hat{\tau}_S^{\mathrm{D-C}} = 0.0028 < \tau_S^D$ and $\hat{\tau}_T^{\mathrm{D-C}} = 0.0020 = \tau_T^D$ (supplementary table S1, Supplementary Material online). Because of this underestimation of τ_S, θ_S is overestimated. Overall, the fitted distribution of t_{bc} matches the true distribution under model C reasonably well while the fitted values of t_{ab} and t_{ac} reflect the increased duration ($\hat{\tau}_S$,$\hat{\tau}_R$) of population S, with the majority of coalescent events occurring closer to $\hat{\tau}_S$ (supplementary fig. S8, Supplementary Material online).

Consistent with the large estimates of M in the C-D setting, the Bayesian test of gene flow has full power (supplementary fig. S9, Supplementary Material online, C-D). In contrast, in the D-C setting, the estimated migration rate is very low, and the Bayesian test has virtually zero power (supplementary fig. S9, Supplementary Material online, D-C). Note that in both models C and D, continuous migration (MSC-M) is assumed. Huang et al. (2022, Fig. 4) examined the A-B and B-A settings, in which the discrete introgression model (MSC-I) is used. In the A-B setting, large estimates of the introgression probability (φ) are produced, while in the B-A setting, the estimated φ is near zero.

Thus we observe the same patterns regardless of the mode of gene flow (MSC-I or MSC-M). If gene flow is between nonsister species but is mis-assigned to the parental lineage so that the assumed gene flow is between sister species (the A-B and C-D settings), we will obtain large estimates of the rate of gene flow, and the Bayesian test will infer gene flow. In contrast, if gene flow involves an ancestral branch and is between sister species but mis-assigned to a daughter branch (the B-A and D-C settings), we will obtain low estimates of the rate of gene flow and the Bayesian test may not detect gene flow. Nevertheless, in our simulations (Fig. 4 in this study and Fig. 4 in Huang et al. 2022) the impacts of mis-assigning a gene-flow event onto parental or daughter branches are local, mostly affecting parameters for lineages on the species tree involved in gene flow.

Inference When the Mode of Gene Flow is Misspecified (Fig. 4, A-C and B-D)

Next, we consider cases where the mode of gene flow is misspecified, with data generated under MSC-I and analyzed under MSC-M, but the population pair involved in gene flow is correctly specified. In the A-C setting, gene flow is between nonsister species while in the B-D setting, it is between sister species. This is an extension of our two-species analysis (Figs. 1–3) to a larger phylogeny. Previously, Huang et al. (2022, Fig. 4) examined the opposite settings, C-A and D-B, in which data were simulated under MSC-M and analyzed under MSC-I, noting that highly precise and accurate parameter estimates were obtained despite the misspecification of the mode of gene flow. In particular, in the case of SC (the C-A setting), the MSC-I model was able to recover almost all gene flow that occurred under the migration model (with $\hat{\varphi}$ under MSC-I being close to φ_0 under MSC-M).

In both the A-C and B-D settings, all population-size and divergence-time parameters are reliably estimated at comparable levels of precision to the C-C and D-D settings. Surprisingly, some parameters in the B-D setting, such as τ_S and θ_S, appear to be even more precisely estimated than in the D-D setting (Fig. 4, supplementary table S1, Supplementary Material online). A similar observation was made in the C-A setting in comparison with A-A (Huang et al. 2022). There is a slight overestimation of θ_S: $\hat{\theta}_S^{\mathrm{A-C}} = 0.003$ and $\hat{\theta}_S^{\mathrm{B-D}} = 0.0024$ while the true value is 0.002 (supplementary table S1, Supplementary Material online).

We obtain a larger estimate of the migration rate in the B-D setting than in the A-C setting even though the introgression probability is the same ($\varphi = 0.2$), with $\hat{M}^{\mathrm{B-D}} = 0.0405$ compared with $\hat{M}^{\mathrm{A-C}} = 0.0116$ (supplementary table S1, Supplementary Material online). A high migration rate (M) in model C predicts the existence of very small coalescent times t_{ab} or the existence of nearly identical sequences from A and B, and is thus incompatible with the data, which is generated under model A with gene flow at a fixed time point in the past ($\tau_X > 0$). This is the same pattern as observed in our analysis of the two-species case, where the IIM model (here, model D) recovers more gene flow and provides less biased parameter estimates than the IM or SC models (Fig. 3).

The Bayesian test of gene flow has ~ 100% power in all datasets in the A-C and B-D settings (supplementary table S9, Supplementary Material online). Gene flow is detected despite the misspecification of the mode of gene flow.

In summary, the misspecification of the mode of gene flow does not have large detrimental effects in our simulations (Fig. 4, A-C and B-D, and Huang et al. 2022, Fig. 4, C-A and D-B). If gene flow occurred in a pulse in the past but has since stopped, MSC-M models assuming ongoing gene flow (IM and SC) may underestimate the amount of gene flow, while the IIM model produced more accurate estimates. When gene flow occurs over extended time periods (as assumed in MSC-M), the MSC-I model is able to produce highly reliable parameter estimates and the test has high power for detecting gene flow.

Inference When Both the Mode of Gene Flow and the Introgression Lineage are Misspecified (Fig. 4, B-C and A-D)

Lastly, we examine the cases when both the mode of gene flow and the lineages involved in gene flow are misspecified (Fig. 4: B-C, A-D). From the discussions above, we expect the mis-assignment of lineages involved in gene flow to have more

impact than the misspecification of the mode of gene flow. The results confirm this expectation.

In the B-C setting, introgression occurs from A to T at τ_X prior to the divergence of B and C at τ_T while the fitting model assumes continuous gene flow from A to B. As model C assumes ongoing gene flow, which is absent in the data, the estimated migration rate M is close to zero ($\hat{M} < 10^{-3}$; supplementary table S1, Supplementary Material online), and the Bayesian test has zero power in detecting gene flow in any of the datasets (supplementary fig. S9, Supplementary Material online). The result is very similar to the D-C setting. Apart from the serious underestimation of the rate of gene flow (M), other effects of model misspecification are local, affecting mainly τ_S and θ_S. All modern population sizes (θ_A, θ_B, θ_C, and θ_D) as well as τ_R and θ_R are well estimated. Model C can fit coalescent times during the initial period $(0, \tau_T)$ well without requiring gene flow. Small coalescent times t_{ab} and t_{ac} due to gene flow in the data generated under model B are then explained by having a more recent divergence time τ_S. We obtain $\hat{\tau}_S = 0.0039$, which is much closer to the introgression time $\tau_X = 0.003$ than the true divergence time $\tau_S = 0.006$. The root divergence time and population size are slightly affected, with τ_R slightly overestimated and θ_R slightly underestimated.

In the A-D setting, introgression occurs from A to B, a nonsister species, while the fitting model assumes continuous gene flow from A to its sister lineage T. As in the C-D setting, the gene-flow event is mis-assigned onto the parental branch, leading to large estimates of the migration rate (M) and the Bayesian test has full power (supplementary fig. S9, Supplementary Material online). As before, the impacts on other parameters are largely local. All present-day population sizes (θ_A, θ_B, θ_C, $and\theta_D$), τ_R and θ_R are well estimated.

Inference When the Direction of Gene Flow is Misspecified on a Four-species Phylogeny

Previously, we studied the effects of misspecified direction of gene flow on parameter estimation and the Bayesian test of gene flow under the MSC-I model (Thawornwattana et al. 2023a). Here, we perform complementary analysis under the MSC-M model. We consider the two MSC-M models of Fig. 4c and d: C (recent gene flow involving nonsister species) and D (ancestral gene flow involving sister species), and, for each, consider three variants: inflow (I, $A \rightarrow B$), outflow (O, $B \rightarrow A$), and bidirectional gene flow (B, $A \leftrightarrows B$), as shown in Fig. 5a–f. Models I, O, and B make different assumptions about the direction of gene flow while both the lineages involved and the mode of gene flow (continuous migration) are correctly specified. For example, in the C:I-O setting (Fig. 5g), data are generated under the inflow model with $A \rightarrow B$ migration (Fig. 5a) but analyzed under the outflow model assuming $B \rightarrow A$ migration (Fig. 5b), so that the assumed direction is the opposite. The I-B and O-B settings represent over-parametrization rather than misspecification.

We used the same parameter values as in the previous section (Fig. 4c–d). Results are summarized in Fig. 5g and h. Estimates from the large datasets of $L = 4{,}000$ loci under models C and D are in supplementary tables S2 and S3, Supplementary Material online, respectively.

Analysis Under the SC Model (Model C, Fig. 5g)

In model C, gene flow is recent and between nonsister species A and B. It may be considered an instance of an SC model in which gene flow occurs after a period of complete isolation. As a baseline for comparison, we first consider cases where the analysis model is correctly specified, i.e. C:I-I, C:O-O, and C:B-B (Fig. 5g). In all three cases, all parameters including the migration rates are correctly estimated. The present-day population sizes (θ_A, θ_B, θ_C, $and\theta_D$) are estimated with narrow CIs, while there is more uncertainty in the ancestral population sizes (θ_T, θ_S, $and\theta_R$) (Fig. 5g). We find that the rate of inflow and outflow is estimated at similar levels of precision. We expect the rates of bidirectional gene flow to involve more uncertainties than the unidirectional rates and this is indeed the case for datasets of $L = 250$ or 1,000 loci (see M estimates under the I-I, O-O, and B-B settings in Fig. 5g), but the CIs have the same widths in large datasets of $L = 4{,}000$ loci (see estimates of M in supplementary table S2, Supplementary Material online, model C: I-I, O-O, and B-B). The Bayesian test of gene flow has full power in all datasets (supplementary fig. S10a–c, Supplementary Material online for C:I-I, C:O-O, and C:B-B).

In the other settings (I-O, O-I, B-I, and B-O; Fig. 5g), the direction of gene flow is misspecified although the I-B and O-B settings represent over-parametrization rather than misspecification. Overall, the impact of misspecification on species divergence times and population sizes is local, affecting parameters for populations involved in gene flow or their immediate ancestors. For example, the population size and divergence time at the root of the tree (τ_R and θ_R) are well estimated. Below we focus on parameters that are affected by the misspecified direction of gene flow, in particular, the migration rates.

In the C:I-B and C:O-B settings, migration is unidirectional but the analysis model allows for migration in both directions. We recover the correct migration rate in the correct direction while the migration rate in the opposite direction is estimated to be zero (Fig. 5g, supplementary table S2, Supplementary Material online). The CI widths of all parameters are comparable with those obtained from the unidirectional migration case (compare C:I-B with C:I-I and C:O-B with C:O-O), suggesting that overparameterization has a minor impact on parameter estimates. The cost of including a nonexistent migration rate in the bidirectional model (B) is thus mostly computational. The Bayesian test of gene flow supported the true migration rate with 100% power, and rejected the nonexistent gene flow with a false positive rate of 0% (supplementary fig. S10a–b, Supplementary Material online for C:I-B and C:O-B).

In the C:I-O and C:O-I settings, migration occurs in one direction but the analysis model assumes the opposite direction. The estimated migration rate in the wrong direction is not zero or negative, but is comparable with the true rate in the opposite direction. For example, in the C:I-O setting the estimate is $\hat{M}_{B\rightarrow A} = 0.176$ with the 95% CI to be (0.168, 0.184) (supplementary table S2, Supplementary Material online, C:I-O), compared with $M_{A\rightarrow B} = 0.2$ in the true model. Furthermore, the Bayesian test detected gene flow in all datasets (supplementary fig. S10a–b, Supplementary Material online for C:I-O and C:O-I). This may be considered as a false positive rate if one emphasizes the inferred wrong direction or power if one emphasizes the presence of gene flow.

Misspecification of the direction of gene flow has a local effect on estimated population-size parameters and divergence times for populations involved in gene flow (A, B, T, and the ancestor S) (Fig. 5g, C:I-O and O-I). When the true recipient population is incorrectly assumed to be a source population, we expect its population size to be overestimated to account

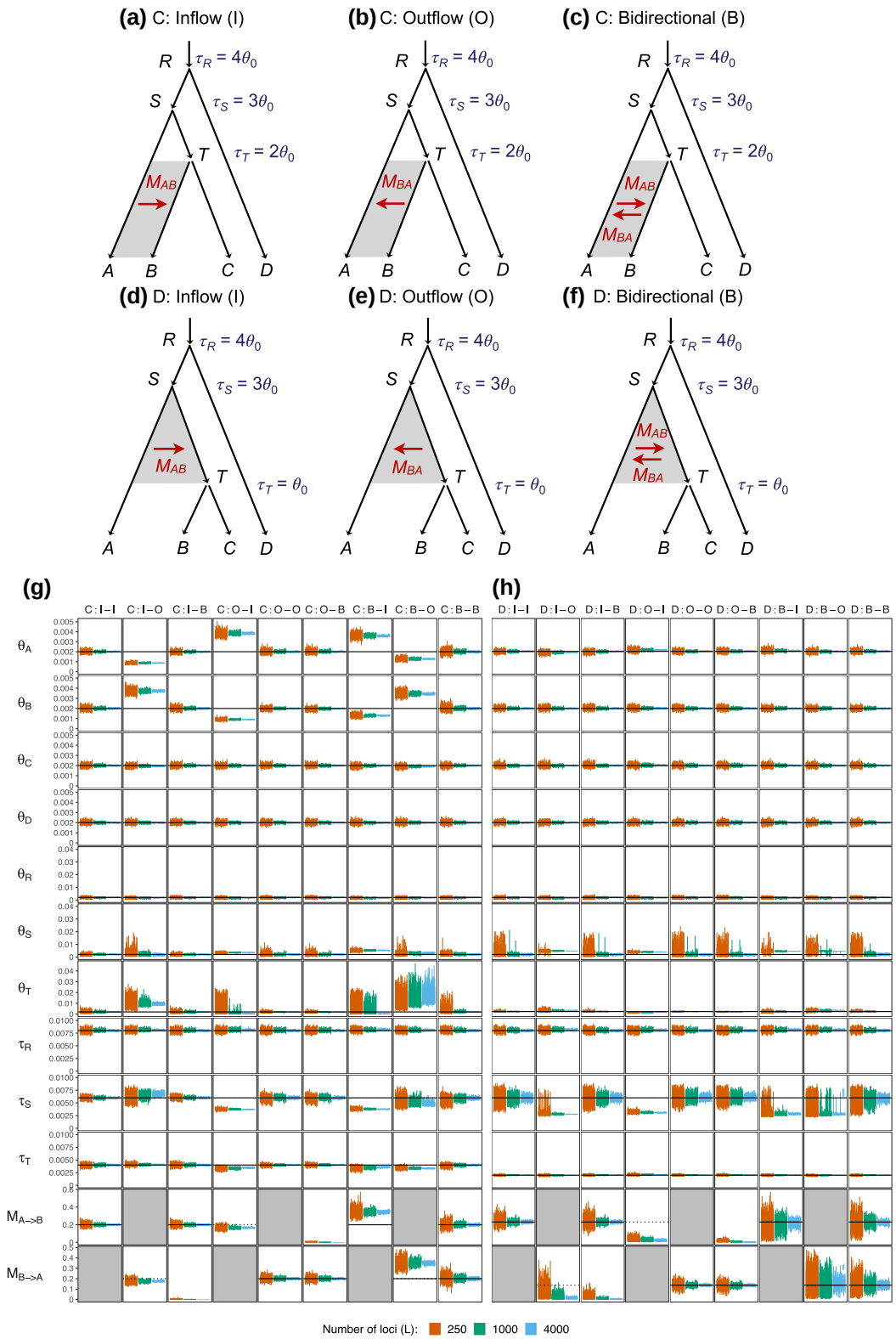

Fig. 5. a–f) MSC-M models with different directions of gene flow between nonsister lineages *A* and *B* (a–c, model C) or between sister lineages *A* and *T* (d–f, model D), with either inflow (b and d), outflow (c and e), or bidirectional gene flow (d and f). The notation C:I-O means the data were generated under model C:I (inflow) and analyzed under model C:O (outflow), etc. Parameters used to generate the data are the same as those in Fig. 4c and d. All migration rates were 0.2. For D models, the migration rate $M_{A\to T}$ is labeled $M_{A\to B}$, etc. for convenience. g–h) The 95% HPD CIs for parameters in 100 replicate datasets of $L = 250$, 1,000, and 4,000 loci for C models (g) and D models (h). Black solid line indicates the true value. Estimates for the C:I-I and D:I-I settings are identical to those for the C-C and D-D settings, respectively, in Fig. 4. A gray box indicates that the parameter does not exist in the model.

for the excess polymorphism. Conversely, a source population incorrectly assumed as a recipient should have its population size underestimated. The results confirm those expectations (Fig. 5g, supplementary table S2, Supplementary Material online). For example, in the C:I-O setting, gene flow is from $A \to B$ but assumed to be from $B \to A$. Thus θ_A is grossly

underestimated and θ_B overestimated. The large biases in population sizes may also be accompanied by biases in species divergence times (τ_T, τ_S) (Fig. 5g, C:I-O and O-I) when the model attempts to fit the coalescent times for sequences from the same species (t_{aa}, t_{bb}).

In the C:B-I and C:B-O settings, migration occurs in both directions while the analysis model incorrectly assumes unidirectional migration. Gene flow in both directions does not cancel out (unlike debts and credits), and instead shows a cumulative effect. The estimate migration rates are 0.34 and 0.35 in the C:B-I and C:B-O settings, much higher than 0.17 and 0.18 in the O-I and I-O settings, respectively. This may not be so surprising when one considers that gene flow in either direction reduces sequence divergence between the two species involved. Similarly, the Bayesian test detects gene flow with full power in all datasets for those settings (supplementary fig. S10c, Supplementary Material online, C:B-I and C:B-O). The effects on the estimation of population sizes and species divergence times are similar to the C:O-I and C:I-O settings. For example, the population size for the incorrectly assumed source population (A in C:B-I and B in C:B-O) is overestimated and that for the assumed recipient population (B in C:B-I and A in C:B-O) is underestimated (Fig. 5g). The cumulative effect of gene flow in both directions in the true model tend to reduce the coalescent time t_{ab} between sequences from species involved in gene flow (A and B), and consequently, τ_T and τ_S are underestimated (supplementary table S3, Supplementary Material online).

Analysis Under the IIM Model (Model D, Fig. 5h)

Lastly, we consider model D, which has ancestral gene flow between sister lineages A and T before T splits into two species B and C (Fig. 5d–f). This may be considered as an instance of an IIM model (Fig. 1c).

First, we note that in all settings for model D, population sizes for extant species are well estimated (Fig. 5h), because there is no gene flow during the time period (τ_T, 0) in either the true or the analysis models. This is different from the settings based on model C, in which extant species may be the source or donor populations of gene flow. Divergence time τ_T and population size θ_T are also well estimated in all settings for model D.

When the model is correctly specified (D:I-I, D:O-O, and D:B-B), the Bayesian test of gene flow has ~ 100% power, even in small datasets with 250 loci (supplementary fig. S10d–f, Supplementary Material online). All parameters are well estimated, with the CIs becoming narrower with the increase of the data size (the number of loci). Again population sizes for ancestral species (θ_S, θ_T) have far wider CIs than those for modern species. Migration rates ($M_{A\to B}$ and $M_{B\to A}$) as well as τ_S and θ_S have wider CIs under the D models (Fig. 5h) than under the corresponding C models (Fig. 5g). Inference of gene flow under model D is more challenging than under model C. This may be due to two factors. First, gene flow in D is more ancient. Second, gene flow in D is between sister lineages while that in C is between nonsisters. See our discussion above of the C-C versus D-D settings in Fig. 4e (which correspond to the C:I-I and D:I-I settings in Fig. 5g and h).

In the D:I-B and D:O-B settings, the analysis model assumes bidirectional migration while migration is in fact unidirectional. The results under D:I-B are very similar to those under D:I-I (and D:O-B to D:O-O), while the rate of migration that does not exist in the true model estimated to be zero (Fig. 5h, supplementary table S3, Supplementary Material online). The over-parametrization of the B model has no major impact on the estimation. The Bayesian test detected gene flow in the correct direction with full power while rejecting the nonexistent gene flow in the opposite direction with false positive rate of ~ 0% (supplementary fig. S10d and e, Supplementary Material online, D:I-B and D:O-B).

In the D:I-O and D:O-I settings, gene flow is unidirectional but the assumed direction is the opposite. Migration rate is estimated to be close to zero, and the divergence time τ_S is seriously underestimated, as the number of loci increases (Fig. 5h, supplementary table S3, Supplementary Material online). Furthermore, the Bayesian test often fails to detect gene flow (supplementary fig. S10d–e, Supplementary Material online). Apparently, misspecification of migration direction caused the method to misinterpret early divergence with gene flow as recent complete isolation with no gene flow. This is in contrast to the C:I-O and C:O-I settings where the estimated migration rate between nonsister species in the wrong direction is positive and close to the true value (Fig. 5g). It is also in contrast to the C:I-I and C:O-O settings where there is no model misspecification and the method is able to distinguish between early divergence with gene flow and recent complete isolation with no gene flow.

In the D:B-I and D:B-O settings, migration occurs in both directions but the analysis model allows only one direction. Migration rate in the allowed direction is estimated to be around the true rate in that direction, with wide CIs (even wider than in the D:B-B setting) (Fig. 5h, D:B-I and D:B-O). Here gene flow in the two directions does not show a cumulative effect, in contrast to the C:B-I and C:B-O settings discussed above. Consistently with parameter estimation, the Bayesian test of gene flow has only moderate power in the D:B-I and D:B-O settings, in contrast to the full power in the C:B-I and C:B-O settings (supplementary fig. S10f vs. fig. S10d, Supplementary Material online). The model underestimates the amount of gene flow between species A and T, and this is compensated by an underestimation of their split time (τ_S; Fig. 5g, D:B-I and D:B-O).

In summary, when the direction of gene flow is correctly specified, the Bayesian test has high power to detect gene flow both between sister lineages and between nonsister lineages. While ancient gene flow between sister lineages is harder to infer than recent gene flow between nonsister species, the Bayesian test easily achieves full power in both scenarios (supplementary fig. S9, Supplementary Material online, C-D and D-D), and the precision in the estimated rate of gene flow is comparable (cf: C:I-I with D:I-I, and C:O-O with D:O-O in Fig. 5g and h). When the direction of gene flow is misspecified, one can easily infer gene flow between extant nonsister lineages (with a cumulative effect if gene flow occurs in both directions). However, misspecification of the direction of gene flow between ancestral sister lineages makes it difficult to detect gene flow.

Analysis of Data from Purple Cone Spruce

To gain insights into Bayesian parameter estimation and the Bayesian test of gene flow when the true history of species divergence and gene flow is unknown, we analyzed an empirical dataset from three purple cone spruce species, *Picea wilsonii* (W), *P. likiangensis* (L), and *P. purpurea* (P) (Sun et al. 2014). The purple cone spruce is endemic to the Qinghai-Tibet Plateau, and *P. purpurea* is hypothesized to be a hybrid species, formed through homoploid hybridization between

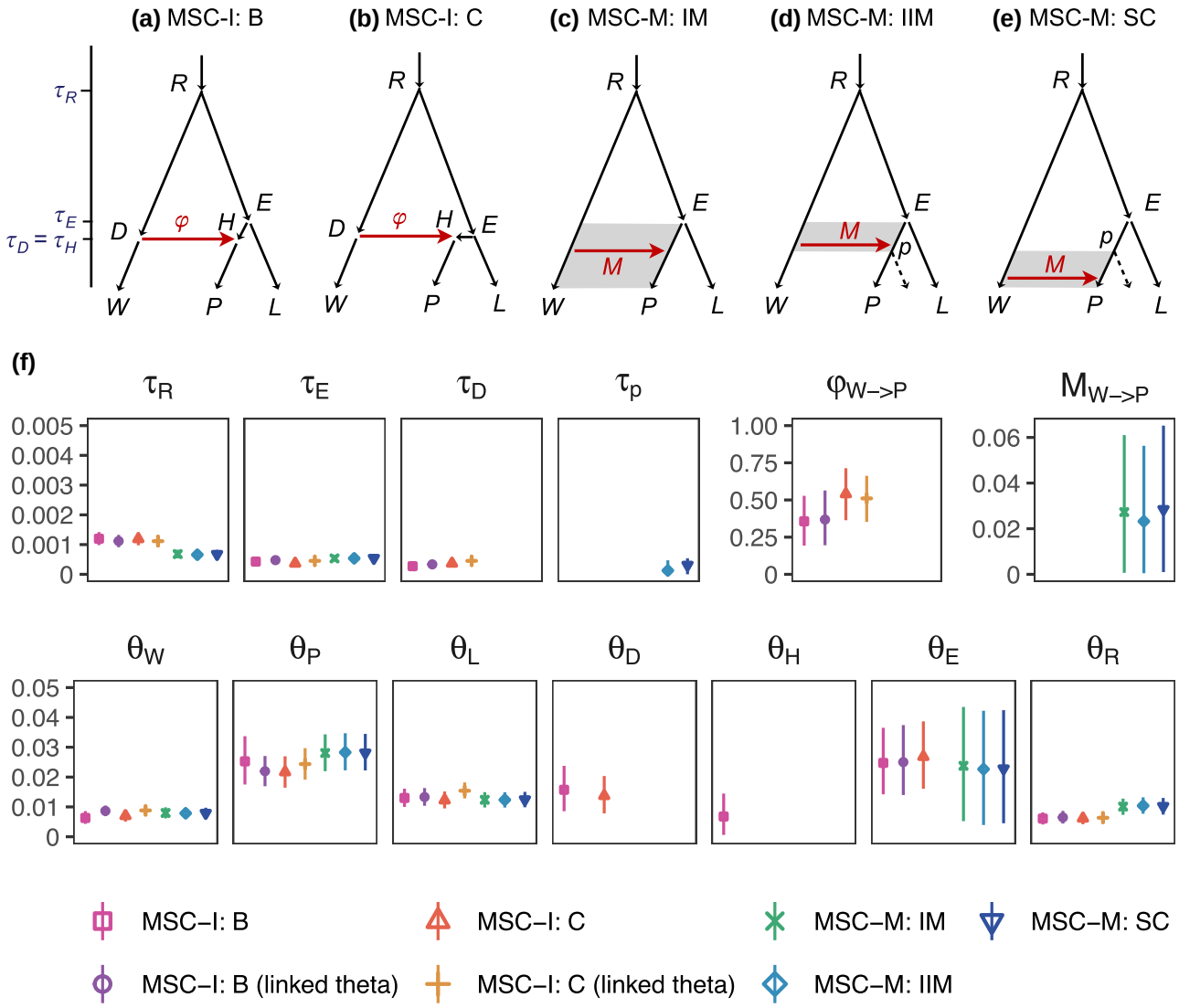

Fig. 6. a–b) Two MSC-I models (B and C; Flouri et al. 2020) and c–e) three MSC-M models (IM, IIM, and SC; Huang et al. 2022; Flouri et al. 2023) for three purple cone spruce species: *Picea wilsonii* (W), *P. purpurea* (P), and *P. likiangensis* (L). f) Posterior means and 95% HPD CIs for parameters in the models obtained in BPP analysis of genomic data. MSC-I modes B and C are implemented either with and without the linked-theta option, which forces θ to be the same before and after introgression.

P. wilsonii and *P. likiangensis* (Sun et al. 2014). Thus, the MSC-I model assuming a pulse of hybridization/introgression may be expected to be a better fit to the data than the MSC-M model. We considered two variants of the MSC-I model (Fig. 6a–b) and three variants of the MSC-M model: IM, IIM, and SC (Fig. 6c–e). Results of Bayesian model comparison are summarized in supplementary table S4, Supplementary Material online, while parameter estimates under those models are in supplementary table S5, Supplementary Material online and Fig. 6f.

Both MSC-I models B and C support the hypothesis that *P. purpurea* is an admixture or hybrid between *P. wilsonii* and *P. likiangensis*. Estimates of the introgression probability (contribution of the *P. wilsonii* parent) range over 0.35 to 0.54, close to 50% (supplementary table S5, Supplementary Material online, Fig. 6f). Estimated divergence and hybridization times are much smaller than the average coalescent time between two sequences sampled within the same species ($\theta/2$), indicating the very recent nature of those species.

Bayesian model comparison strongly favors the MSC-I models over the MSC-M models (supplementary table S4, Supplementary Material online), consistent with the hybrid origin of *P. purpurea*. We further test whether gene flow from the two parental lineages into *P. purpurea* occurred at the same time, as predicted by the hypothesis of hybrid speciation. The null hypothesis is model C with the constraint $\tau_D = \tau_E$ (Fig. 6b), while the alternative hypothesis is model B with $\tau_D < \tau_E$ (Fig. 6a). This test is inclusive, as the Bayes factor is in the range (0.01, 100) and does not strongly favor either model (supplementary table S4, Supplementary Material online).

Given that the MSC-I models fit the data better, we next examine biases in parameter estimates that result from the use of "wrong" MSC-M models. All three migration models (IM, IIM, and SC; Fig. 6c–e) produced lower rates of gene flow and more recent divergence time (τ_R). The estimated migration rate M correspond to $\varphi_0 < 1\%$, much lower than estimates of φ under the MSC-I models (Fig. 6f; supplementary

table S5, Supplementary Material online). In effect, the MSC-M models mis-interpreted early divergence with gene flow as recent divergence with very little gene flow.

This is the same pattern found in our simulations, in which the IM model may misinterpret early divergence with gene flow as more recent divergence with no or little gene flow (Fig. 2 for M in the case of two species and Fig. 4e, A-C in the case of four species). Overall, the analysis of the empirical data showed the same patterns as found in the asymptotic analysis and computer simulation. As expected from the hypothesis of hybrid speciation (Sun et al. 2014), our Bayesian model test strongly favors the pulse model of gene flow over continuous migration.

Conclusions

Here we summarize our key findings from this simulation study, by integrating with the results from previous studies, which used the MSC-I model for data analysis (Jiao et al. 2020; Huang et al. 2022; Ji et al. 2023; Thawornwattana et al. 2023a).

First, the Bayesian test of gene flow has high power in detecting both recent and ancient gene flow, either between nonsister species or between sister lineages. In previous simulations, the Bayesian test was found to have much higher power than summary methods based on genome-wide site pattern counts such as HyDe (Ji et al. 2023, Fig. 9) or on gene-tree counts such as SNaQ (Ji et al. 2025, Fig. 5).

When the mode of gene flow is misspecified (i.e. in the I-M and M-I settings), the Bayesian test is found to have high power in almost all cases we have examined. We generated data under the MSC-I model and analyzed them using variants of the MSC-M model (IM, IIM, and SC). Despite the misspecification of the mode of gene flow, the test detects gene flow in most cases, whether gene flow involved sister lineages (supplementary fig. S7, Supplementary Material online) or nonsister lineages (supplementary fig. S9 A-C and B-D and fig. S10, Supplementary Material online). This means that one is very likely to infer gene flow even if the assumed mode of gene flow is not a perfect match to reality (for example, if the rate of gene flow varies over time). We observed low power in the case of two species when φ is very low or very high (supplementary fig. S7, Supplementary Material online, φ), in which case the MSC-I model is close to the MSC model with no gene flow.

Second, misspecification of the mode of gene flow leads to underestimation of the amount of gene flow. In other words, in both the I-M and M-I settings, Bayesian estimation tends to recover less gene flow than in the true model, as measured by the total amount of gene flow as a fraction of the expected number of immigrants in the recipient population (that is, φ in MSC-I or φ_0 in MSC-M) (e.g. Fig. 3 and Fig. 4e, A-C). When the true model is MSC-I, the IM and SC models in particular produce serious underestimates of the amount of gene flow, because those models assume ongoing gene flow up to the present time and predict recent coalescent times between sequences from the two species (with t_{ab} near zero), which do not exist in the data. In such cases, the IIM model recovers a greater amount of gene flow (Fig. 3). Previously, the discrete model (MSC-I) was found to recover less than the true amount of gene flow when gene flow occurs over extended time periods according to the MSC-M model (Huang et al. 2022, Fig. 1e).

Third, while it is harder to infer gene flow between sister lineages than between nonsisters, with multiple sequences sampled per species, the Bayesian method is powerful in inferring gene flow between sister lineages: the Bayesian test based on the Bayes factor has high power to detect gene flow and Bayesian estimation produces estimates of the rate of gene flow with precision similar to the case of nonsister gene flow (e.g. Fig. 4e, D-D for the MSC-M model; see also Fig. 4, B-B in Huang et al. 2022 for the MSC-I model). Note that most summary methods (such as the D-statistic and the f-branch test) are unable to identify gene flow between sister lineages. For triplet methods based on gene trees, introgression between nonsister lineages causes changes to the gene tree topology, whereas introgression between sister lineages does not. For triplet methods based on site-pattern counts, introgression between nonsister lineages causes an asymmetry in the site-pattern counts, but introgression between sister lineages does not.

Fourth, when the direction of gene flow is correctly specified, Bayesian test has high power to detect gene flow both between sister lineages and between nonsister lineages. While ancient gene flow between sister lineages is harder to infer than recent gene flow between nonsister species, the Bayesian method can achieve similar power in the test of gene flow and similar precision in the estimated rate of gene flow (cf: C:I-I with D:I-I, and C:O-O with D:O-O in Fig. 5g and h). When the direction of gene flow is misspecified, the Bayesian test often detects gene flow. The estimated rate of gene flow may be higher or lower than the true rate in the opposite direction (e.g. Fig. 5g and h, I-O, O-I for the MSC-M model and Fig. 4 and supplementary fig. S4, Supplementary Material online in Thawornwattana et al. 2023a for MSC-I). When gene flow occurs in both directions but a unidirectional model is assumed, gene flow in the two directions does not cancel out and may instead show a cumulative effect. If gene flow is unidirectional, use of the bidirectional model leads to detection of gene flow in the correct direction, and rejection of gene flow in the wrong direction (supplementary fig. S10, Supplementary Material online, C:I-B, C:O-B, D:I-B, and D:O-B; see also (Thawornwattana et al. 2023a), supplementary fig. S2, Supplementary Material online, model B). Unlike Frequentist hypothesis testing, the Bayesian test may lead to strong rejection of the more general alternative hypothesis. Note that most summary methods cannot identify the direction of gene flow (Jiao et al. 2021; Huang et al. 2022).

Fifth, when the gene-flow event is mis-assigned to a parental or daughter branch rather than the lineage genuinely involved in gene flow, the Bayesian method may produce highly biased parameter estimates (Fig. 4e, C-D and D-C), and the Bayesian test may fail to detect the gene flow (supplementary fig. S10, Supplementary Material online, D-C). Under the MSC-I model, the inferred introgression time tends to be stuck on the species divergence time, and the estimated rate of gene flow tends to be far lower than the true rate (Huang et al. 2022, Fig. 4e, A-B and B-A).

Finally, misspecification of the mode of gene flow (MSC-I versus MSC-M) tends to have only small local effects, affecting divergence times and population sizes for species on the phylogeny around the lineages involved in the gene flow (e.g. Fig. 4e, A-C and B-D for MSC-M). The Bayesian test has high power despite the misspecification; for example, if the true model is MSC-I but data are analyzed under MSC-M, the test may still have full power (supplementary fig. S9, Supplementary Material online, A-C and B-D).

Overall, analysis of both synthetic and real datasets in this and previous studies demonstrate that the MSC-I and

MSC-M models, no doubt extreme simplifications of gene flow in the real world, are very effective for detecting gene flow and estimating its rate using genomic sequence data. The MSC-I model in particular performed well in simulations under MSC-M with gene flow over extended time periods. The MSC-M models may be most suitable to data from different populations of the same species where gene flow may be ongoing and over extended time periods. Variants of the model such as IM, IIM, and SC make different assumptions about possible gene flow following speciation, and may be useful for testing different theories of speciation (Westram et al. 2022).

Materials and Methods

Two Species Case: Asymptotic Analysis

The case of two species when the data consist of one sequence per species is simple enough to yield analytical solutions. We considered estimation of parameters under the three MSC-M models of Fig. 1b–d when data of an infinite number of loci ($L \to \infty$) were generated under the MSC-I model of Fig. 1a. We obtained the limit of the MLEs, $\boldsymbol{\theta}^*_{\mathrm{m}}$, by minimizing the KL divergence (equation (3)). As $L \to \infty$, the data are represented by the distribution of the number of differences between two sequences at the n sites, and maximizing the likelihood is equivalent to minimizing the KL divergence. Optimization was achieved using a C program that implements the BFGS algorithm from PAML (Yang 2007). The program is available at https://github.com/ythaworn/iimmsci2s. For each model and each value of φ, we ran the optimization multiple times and used the run with the lowest KL value. We excluded runs with parameter values on the optimization boundaries.

Two Species Case: Simulation

We used simulation to verify and extend our asymptotic analysis. We simulated multilocus sequence data under the MSC-I model of Fig. 1a and analyzed them under the IM, IIM, and SC models of Fig. 1b–d. We used two values of population sizes on the species tree: $\theta_A = \theta_X = \theta_R = \theta_0 = 0.002$ (thin branches) and $\theta_B = \theta_Y = \theta_1 = 0.01$ (thick branches). Introgression occurred from species A to B with probability $\varphi = 0.2$ at time $\tau_X = \theta_0$ after species divergence at time $\tau_R = 2\theta_0$. In the base case, we assumed $\varphi = 0.2$ and the data consisted of $L =$ 4,000 loci, with $S = 4$ sequences per species, and $n = 1{,}000$ sites per sequence. We varied the following four factors one at a time, keeping other parameters fixed: the number of sites per sequence (n), the number of sequences per species (S), the number of loci (L), and the introgression probability (φ). The values used were $n = 250$, 1,000, 4,000, 16,000, 64,000; $S = 1$, 2, 4, 8, 16; $L = 250$, 500, 1,000, 2,000, 4,000, 8,000; and $\varphi = 0.01$, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, and 0.7. For each setting, we simulated 30 replicate datasets. With those four factors (n, S, L, M), and 30 replicates each, there were $(5 + 5 + 6 + 8 - 3) \times 30 = 630$ datasets in total. The subtraction by 3 accounted for the fact that all four factors shared the same base case ($n = 1{,}000$, $S = 4$, $L = 4{,}000$, $\varphi = 0.2$). To generate sequence data, we first generated a gene tree with coalescent times at each locus and then simulated sequences along branches of the gene tree under the JC model (Jukes and Cantor 1969). Sequences at the tips of the tree became data at the locus. Simulation was done using the `simulate` option in BPP v4.7.0 (Flouri et al. 2018, 2020, 2023).

Each dataset was analyzed under the IM, IIM, and SC models of Fig. 1b–d to estimate parameters using BPP v4.7.0 (Flouri et al. 2023). The JC mutation model was assumed. We assigned gamma priors to population size parameters (θ), the root age (τ_R) on the species tree and the migration rate: $\theta \sim G(2, 200)$ with mean $2/200 = 0.01$, $\tau_R \sim G(4, 200)$ with mean $4/200 = 0.02$, and $M \sim G(2, 10)$ with mean $2/10 = 0.2$. For each fitting model, we performed two independent runs of MCMC, each with 32,000 iterations of burnin and 10^6 iterations of the main chain. Samples were recorded every 100 iterations. With three fitting models and two MCMC runs per dataset, there were $3 \times 2 \times 630 =$ 3, 780 MCMC runs in total. Each run of the base case took about 80 h and 2G of memory while the most expensive runs ($L = 8{,}000$ or $S = 8$) took about 200 h and 4G of memory. For datasets with $S = 16$, we allowed the MCMC run for up to 300 h and 8G of memory, which was about 4×10^5 iterations.

Four Species Case: Wrong Migration Branch and Wrong Mode of Gene Flow

Data were simulated under the four models (A-D) in a phylogeny for four species in Fig. 4 and analyzed under models C and D. Models A and B assumes discrete introgression, while models C and D assumes continuous migration. Gene flow occurred either between nonsister species (models A and C) or between sister species (models B and D). Parameters are shown in Fig. 4. All population sizes were assumed to be $\theta_0 = 0.002$. For models A and B, we used the introgression probability $\varphi = 0.2$. For models C and D, we assumed the migration rate $M = 0.2$ migrants per generation. Each dataset consisted of $S = 4$ sequences per species per locus, each of length $n = 500$ sites. We varied the number of loci: $L = 250$, 1,000, and 4,000. For each setting, we simulated 100 replicate datasets.

Each dataset was analyzed under both models C and D (Fig. 4c and d) using BPP v4.7.0 (Flouri et al. 2023). There were eight settings in total: A-C, B-C, C-C, D-C, A-D, B-D, C-D, and D-D. Settings C-C and D-D served as references for comparison as the model was correctly specified. In settings C-D and D-C, the population pair (or branches in the phylogeny) involved in migration was misspecified. In settings A-C and B-D, gene flow occurred as pulse introgression but was misspecified as continuous migration. Finally, in settings B-C and A-D an incorrect mode of gene flow was assigned to a wrong branch. Gamma priors were assigned to population sizes, root age, and migration rate as $\theta \sim G(2, 200)$ with mean 0.01, $\tau_R \sim G(4, 200)$ with mean 0.02, and $M \sim G(2, 10)$ with mean 0.2. With four data-generating models, three values of L and 100 replicates, there were $4 \times 3 \times 100 =$ 1, 200 datasets in total. We performed two independent runs of MCMC, each with 10,000 iterations of burnin and 10^6 iterations of the main chain. Samples were recorded every 100 iterations. With two fitting models (C and D) for each dataset, there were $2 \times 2 \times 1{,}200 = 4{,}800$ MCMC runs in total. The running time was about 20–30 h for datasets with $L = 250$, 80 h for $L = 1{,}000$, and 260 h for $L = 4{,}000$.

Simulation in the Four Species Case: Misspecified Direction of Gene Flow

To study the effect of incorrectly assumed direction of gene flow under the MSC-M model, we simulated sequence data

using models C and D of Fig. 4, but with three specifications concerning the direction of gene flow: inflow (I, $A \rightarrow B$), outflow (O, $B \rightarrow A$), and bidirectional gene flow (B, $A \leftrightarrows B$) (Fig. 5a–f). We used the same parameter values as before (Fig. 4c–d). Each simulated dataset was analyzed assuming the three variants of the MSC-I model (I, O, and B), generating nine settings for model C (e.g. C:I-O) and nine settings for model D (e.g. D:I-O).

We used three values of for the number of loci: $L = 250$, 1,000, and 4,000. For model C, with 3 data-generating models, three values of L and 100 replicates, there were $3 \times 3 \times 100 = 900$ datasets in total. MCMC setup was the same as before. With three fitting models (I, O, and B), two independent MCMC runs per dataset, the total number of MCMC runs was $3 \times 2 \times 900 = 5,400$. Similarly for model D, there were 900 datasets and 5,400 MCMC runs in total. We reused the datasets for C:I and D:I from the previous section.

Bayesian Test of Gene Flow

We conducted the Bayesian test of gene flow (Ji et al. 2023) to assess whether there is significant evidence in the data for gene flow. For example, for unidirectional gene flow (Figs. 1 and 4), the null hypothesis of no gene flow, $H_0 : M_{A\rightarrow B} = 0$ may be compared with the alternative hypothesis of gene flow, $H_1 : M_{A\rightarrow B} > 0$, via the Bayes factor B_{10}. As the two hypotheses are nested, B_{10} may be approximated by the Savage–Dickey density ratio, approximated by $B_{10,\varepsilon} = \frac{\mathbb{P}(\emptyset)}{\mathbb{P}(\emptyset \mid X)}$, where $\mathbb{P}(\emptyset)$ is the probability for the null interval, $\emptyset : 0 < M_{A\rightarrow B} < \varepsilon$, under the prior distribution, and $\mathbb{P}(\emptyset \mid X)$ is the corresponding posterior probability. The null interval is part of the parameter space for H_1 that represents the null hypothesis. We used $\varepsilon = 0.01$ and 0.001. $B_{10,\varepsilon}$ was calculated by processing a posterior MCMC sample under H_1 (Ji et al. 2023). $B_{10} > 100$ is considered strong evidence in favor of H_1, similar to the 1% significance level in hypothesis testing. The power of the test is defined as the proportion of replicate datasets in which $B_{10} > 100$.

Analysis of the Purple Cone Spruce Data

We analyzed a dataset for three purple cone spruce species (*Picea wilsonii*, *P. likiangensis*, and *P. purpurea*) (Sun et al. 2014). There are 11 short nuclear loci (200–600 bp per locus), with 100 sequences from *P. wilsonii*, 112 from *P. purpurea* and 120 from *P. likiangensis*. This is the "full" dataset analyzed by Flouri et al. (2020) under the MSC-I models. Here we used MSC-I and MSC-M models of Fig. 6a–e. We assigned the priors $\theta \sim G(2, 200)$ with mean 0.01, $\tau_R \sim G(2, 1,000)$ with mean 0.002, $\varphi \sim U(0, 1)$ for the MSC-I models and $w = 4M/\theta \sim G(2, 1)$ with mean 2 for the MSC-M models. For each model, we performed four independent runs of MCMC, each with 40,000 iterations of burn-in and 10^6 main iterations, sampling every 100th iteration. MCMC samples from the four replicate runs were compared to verify convergence before they were pooled to produce final posterior summaries (Fig. 6, supplementary table S5, Supplementary Material online). Each run took ~ 60 h for the MSC-I models and 160 h for the MSC-M models.

The Savage–Dickey approach to calculating the Bayes factor applies if the two compared models are nested (as in the test of gene flow). To compare nonnested models (such as the MSC-I and MSC-M models of Fig. 6), we used thermodynamic integration to calculate the marginal likelihood, using 32 Gaussian quadrature points (Lartillot and Philippe 2006; Rannala and Yang 2017) (supplementary table S4, Supplementary Material online). This involved 32 MCMC runs, with the same setup as above. We calculated adjusted Bayes factors by performing least-squares fitting of local quadratic polynomials to stabilize the estimates as described in Thawornwattana et al. (2023b). For comparison of the two variants of the MSC-I model (Fig. 6a and b), which are nested models, we also calculated Bayes factors using the Savage–Dickey density ratio (Ji et al. 2023). We tested whether the introgression time τ_D is significantly different from the species divergence time τ_E, with $H_0 : \tau_D = \tau_E$ versus $H_1 : \tau_D < \tau_E$. Bayes factors below 0.01 provide support for H_0, i.e. hybrid speciation (MSC-I: C; Fig. 6b) while values above 100 support the MSC-I: B model of introgression after divergence (Fig. 6a). Here, we calculated $B_\varepsilon = \frac{\mathbb{P}(\emptyset)}{\mathbb{P}(\emptyset \mid X)}$, where $\mathbb{P}(\emptyset)$ is the probability of the null interval $\emptyset : 0 < \tau_E - \tau_D < \varepsilon$ under the prior distribution and $\mathbb{P}(\emptyset \mid X)$ is the corresponding probability under the posterior distribution. We used $\varepsilon = 10^{-4}$ and 10^{-5}. We calculated B_ε by processing the MCMC sample for the posterior distribution under H_1 from the MSC-I: B model (Fig. 6a). The prior null probability $\mathbb{P}(\emptyset)$ was calculated from the prior distribution of $\tau_E - \tau_D$ obtained from running the MCMC without data.

Supplementary Material

Supplementary material is available at *Molecular Biology and Evolution* online.

Funding

This study has been supported by Biotechnology and Biological Sciences Research Council (BBSRC) grants (BB/T003502/1, BB/X007553/1) and a Natural Environment Research Council (NERC) grant (NE/X002071/1) to Z.Y., as well as by Harvard University.

Conflict of Interest

None declared.

Data Availability

Simulated datasets are available in Zenodo at https://doi.org/10.5281/zenodo.11182437.

References

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat Rev Genet*. 2016:17(2):81–92. https://doi.org/10.1038/nrg.2015.28.

Ané C, Fogg J, Allman ES, Baños H, Rhodes JA. Anomalous networks under the multispecies coalescent: theory and prevalence. *J Math Biol*. 2024:88(3):29. https://doi.org/10.1007/s00285-024-02050-7.

Baharian S, Gravel S. On the decidability of population size histories from finite allele frequency spectra. *Theor Popul Biol*. 2018:120(6):42–51. https://doi.org/10.1016/j.tpb.2017.12.008.

Costa RJ, Wilkinson-Herbots H. Inference of gene flow in the process of speciation: an efficient maximum-likelihood method for the isolation-with-initial-migration model. *Genetics*. 2017:205(4):1597–1618. https://doi.org/10.1534/genetics.116.188060.

Costa RJ, Wilkinson-Herbots HM. Inference of gene flow in the process of speciation: efficient maximum-likelihood implementation of a generalised isolation-with-migration model. *Theor Popul Biol*. 2021:140(6):1–15. https://doi.org/10.1016/j.tpb.2021.03.001.

Dalquen D, Zhu T, Yang Z. Maximum likelihood implementation of an isolation-with-migration model for three species. *Syst Biol*. 2017:66: 379–398. https://doi.org/10.1093/sysbio/syw063.

Edwards S, Cloutier A, Baker A. Conserved nonexonic elements: a novel class of marker for phylogenomics. *Syst Biol*. 2017:66(6): 1028–1044. https://doi.org/10.1093/sysbio/syx058.

Excoffier L, Marchi N, Marques DA, Matthey-Doret R, Gouy A, Sousa VC. fastsimcoal2: demographic inference under complex evolutionary scenarios. *Bioinformatics*. 2021:37(24):4882–4885. https://doi.org/10.1093/bioinformatics/btab468.

Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Syst Biol*. 2012:61(5):717–726. https://doi.org/10.1093/sysbio/sys004.

Finger N, Farleigh K, Bracken J, Leache A, Francois O, Yang Z, Flouri T, Charran T, Jezkova T, Williams D, *et al*. 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*. 2022:14(1): evab260. https://doi.org/10.1093/gbe/evab260.

Flouri T, Jiao X, Huang J, Rannala B, Yang Z. Efficient Bayesian inference under the multispecies coalescent with migration. *Proc Natl Acad Sci U S A*. 2023:120(44):e2310708120. https://doi.org/10.1073/pnas.2310708120.

Flouri T, Jiao X, Rannala B, Yang Z. Species tree inference with BPP using genomic sequences and the multispecies coalescent. *Mol Biol Evol*. 2018:35(10):2585–2593. https://doi.org/10.1093/molbev/msy147.

Flouri T, Jiao X, Rannala B, Yang Z. A Bayesian implementation of the multispecies coalescent model with introgression for phylogenomic analysis. *Mol Biol Evol*. 2020:37(4):1211–1223. https://doi.org/10.1093/molbev/msz296.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH-Y. A draft sequence of the Neandertal genome. *Science*. 2010:328(5979):710–722. https://doi.org/10.1126/science.1188021.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet*. 2009: 5(10):e1000695. https://doi.org/10.1371/journal.pgen.1000695.

Hey J. Isolation with migration models for more than two populations. *Mol Biol Evol*. 2010:27(4):905–920. https://doi.org/10.1093/molbev/msp296.

Hey J, Chung Y, Sethuraman A, Lachance J, Tishkoff S, Sousa VC, Wang Y. Phylogeny estimation by integration over isolation with migration models. *Mol Biol Evol*. 2018:35(11):2805–2818. https://doi.org/10.1093/molbev/msy162.

Hey J, Nielsen R. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics*. 2004:167(2):747–760. https://doi.org/10.1534/genetics.103.024182.

Hibbins MS, Hahn MW. Phylogenomic approaches to detecting and characterizing introgression. *Genetics*. 2022:220(2):iyab173. https://doi.org/10.1093/genetics/iyab173.

Huang J, Thawornwattana Y, Flouri T, Mallet J, Yang Z. Inference of gene flow between species under misspecified models. *Mol Biol Evol*. 2022:39(12):msac237. https://doi.org/10.1093/molbev/msac237.

Jackson ND, Morales AE, Carstens BC, O'Meara BC. PHRAPL: phylogeographic inference using approximate likelihoods. *Syst Biol*. 2017:66(6):1045–1053. https://doi.org/10.1093/sysbio/syx001.

Ji J, Jackson DJ, Leaché AD, Yang Z. Power of Bayesian and heuristic tests to detect cross-species introgression with reference to gene flow in the *Tamias quadrivittatus* group of North American chipmunks. *Syst Biol*. 2023:72(2):446–465. https://doi.org/10.1093/sysbio/syac077.

Ji J, Roberts T, Flouri T, Yang Z. Inference of cross-species gene flow using genomic data depends on the methods: case study of gene flow in *Drosophila*. *Syst Biol*. 2025:syaf019. https://doi.org/10.1093/sysbio/syaf019.

Jiao X, Flouri T, Rannala B, Yang Z. The impact of cross-species gene flow on species tree estimation. *Syst Biol*. 2020:69(5):830–847. https://doi.org/10.1093/sysbio/syaa001.

Jiao X, Flouri T, Yang Z. Multispecies coalescent and its applications to infer species phylogenies and cross-species gene flow. *Nat Sci Rev*. 2021:8(12):nwab127. https://doi.org/10.1093/nsr/nwab127.

Jones GR. Divergence estimation in the presence of incomplete lineage sorting and migration. *Syst Biol*. 2019:68(1):19–31. https://doi.org/10.1093/sysbio/syy041.

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

Karin BR, Gamble T, Jackman TR. Optimizing phylogenomics with rapidly evolving long exons: comparison with anchored hybrid enrichment and ultraconserved elements. *Mol Biol Evol*. 2020:37(3): 904–922. https://doi.org/10.1093/molbev/msz263.

Kubatko LS, Chifman J. An invariants-based method for efficient identification of hybrid species from large-scale genomic data. *BMC Evol Biol*. 2019:19(1):112. https://doi.org/10.1186/s12862-019-1439-7.

Lartillot N, Philippe H. Computing Bayes factors using thermodynamic integration. *Syst Biol*. 2006:55(2):195–207. https://doi.org/10.1080/10635150500433722.

Leaché AD, Oaks JR. The utility of single nucleotide polymorphism (SNP) data in phylogenetics. *Ann Rev Ecol Evol Syst*. 2017:48(1): 69–84. https://doi.org/10.1146/ecolsys.2017.48.issue-1.

Lemmon AR, Emme SA, Lemmon EM. Anchored hybrid enrichment for massively high-throughput phylogenomics. *Syst Biol*. 2012:61(5): 727–744. https://doi.org/10.1093/sysbio/sys049.

Malinsky M, Svardal H, Tyers AM, Miska EA, Genner MJ, Turner GF, Durbin R. Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. *Nat Ecol Evol*. 2018:2(12):1940–1955. https://doi.org/10.1038/s41559-018-0717-x.

Mallet J, Besansky N, Hahn MW. How reticulated are species? *Bioessays*. 2016:38(2):140–149. https://doi.org/10.1002/bies.v38.2.

Martin SH, Jiggins CD. Interpreting the genomic landscape of introgression. *Curr Opin Genet Dev*. 2017:47(Pt 3):69–74. https://doi.org/10.1016/j.gde.2017.08.007.

Moran BM, Payne C, Langdon Q, Powell DL, Brandvain Y, Schumer M. The genomic consequences of hybridization. *Elife*. 2021:10:e69016. https://doi.org/10.7554/eLife.69016.

Nielsen R, Wakeley J. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics*. 2001:158(2): 885–896. https://doi.org/10.1093/genetics/158.2.885.

Pang X-X, Zhang D-Y. Impact of ghost introgression on coalescent-based species tree inference and estimation of divergence time. *Syst Biol*. 2023:72(1):35–49. https://doi.org/10.1093/sysbio/syac047.

Pang X-X, Zhang D-Y. Detection of ghost introgression requires exploiting topological and branch length information. *Syst Biol*. 2024:73(1):207–222. https://doi.org/10.1093/sysbio/syad077.

Rannala B, Yang Z. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*. 2003:164(4):1645–1656. https://doi.org/10.1093/genetics/164.4.1645.

Rannala B, Yang Z. Efficient Bayesian species tree inference under the multispecies coalescent. *Syst Biol*. 2017:66(5):823–842. https://doi.org/10.1093/sysbio/syw119.

Solis-Lemus C, Ane C. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS Genet*. 2016: 12(3):e1005896. https://doi.org/10.1371/journal.pgen.1005896.

Sun Y, Abbott RJ, Li L, Li L, Zou J, Liu J. Evolutionary history of purple cone spruce (*Picea purpurea*) in the Qinghai–Tibet Plateau: homoploid hybrid origin and Pleistocene expansion. *Mol Ecol*. 2014:23(2):343–359. https://doi.org/10.1111/mec.2014.23.issue-2.

Suvorov A, Kim BY, Wang J, Armstrong EE, Peede D, D'Agostino ER, Price DK, Waddell PJ, Lang M, Courtier-Orgogozo V, *et al*. Widespread introgression across a phylogeny of 155 *Drosophila* genomes. *Curr Biol*. 2022:32(1):111–123.e5. https://doi.org/10.1016/j.cub.2021.10.052.

Terhorst J, Song YS. Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum. *Proc Natl Acad*

Sci U S A. 2015:112(25):7677–7682. https://doi.org/10.1073/pnas.1503717112.

Thawornwattana Y, Huang J, Flouri T, Mallet J, Yang Z. Inferring the direction of introgression using genomic sequence data. *Mol Biol Evol*. 2023a:40(8):msad178. https://doi.org/10.1093/molbev/msad178.

Thawornwattana Y, Seixas FA, Mallet J, Yang Z. 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*. 2022:71(5):1159–1177. https://doi.org/10.1093/sysbio/syac009.

Thawornwattana Y, Seixas F, Yang Z, Mallet J. Major patterns in the introgression history of *Heliconius* butterflies. *Elife*. 2023b:12:RP90656. https://doi.org/10.7554/eLife.90656.3.

Tricou T, Tannier E, de Vienne DM. Ghost lineages highly influence the interpretation of introgression tests. *Syst Biol*. 2022:71(5):1147–1158. https://doi.org/10.1093/sysbio/syac011.

Wen D, Nakhleh L. Coestimating reticulate phylogenies and gene trees from multilocus sequence data. *Syst Biol*. 2018:67(3):439–457. https://doi.org/10.1093/sysbio/syx085.

Wen D, Yu Y, Hahn MW, Nakhleh L. Reticulate evolutionary history and extensive introgression in mosquito species revealed by phylogenetic network analysis. *Mol Ecol*. 2016:25(11):2361–2372. https://doi.org/10.1111/mec.2016.25.issue-11.

Westram AM, Stankowski S, Surendranadh P, Barton N. What is reproductive isolation? *J Evol Biol*. 2022:35(9):1143–1164. https://doi.org/10.1111/jeb.14005.

Yan Z, Ogilvie HA, Nakhleh L. Comparing inference under the multispecies coalescent with and without recombination. *Mol Phylogenet Evol*. 2023:181:107724. https://doi.org/10.1016/j.ympev.2023.107724.

Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 2007:24(8):1586–1591. https://doi.org/10.1093/molbev/msm088.

Yu Y, Degnan JH, Nakhleh L. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *PLoS Genet*. 2012:8(4):e1002660. https://doi.org/10.1371/journal.pgen.1002660.

Yu Y, Nakhleh L. A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics*. 2015:16(Suppl 10):S10. https://doi.org/10.1186/1471-2164-16-S10-S10.

Zhang C, Ogilvie HA, Drummond AJ, Stadler T. Bayesian inference of species networks from multilocus sequence data. *Mol Biol Evol*. 2018:35(2):504–517. https://doi.org/10.1093/molbev/msx307.

Zhu T, Flouri T, Yang Z. A simulation study to examine the impact of recombination on phylogenomic inferences under the multispecies coalescent model. *Mol Ecol*. 2022:31(10):2814–2829. https://doi.org/10.1111/mec.v31.10.

Zhu T, Yang Z. Maximum likelihood implementation of an isolation-with-migration model with three species for testing speciation with gene flow. *Mol Biol Evol*. 2012:29(10):3131–3142. https://doi.org/10.1093/molbev/mss118.