Multispecies Coalescent and its Applications to Infer Species Phylogenies and Cross-Species Gene Flow

Xiyun Jiao,1,2 Tomáš Flouri1 and Ziheng Yang1,*

1Department of Genetics, Evolution and Environment, University College London, UK
2Department of Statistics and Data Science, Southern University of Science and Technology, China

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

Abstract
Multispecies coalescent (MSC) is the extension of the single-population coalescent model to multiple species. It integrates the phylogetic process of species divergences and the population genetic process of coalescent, and provides a powerful framework for a number of inference problems using genomic sequence data from multiple species, including estimation of species divergence times and population sizes, estimation of species trees accommodating discordant gene trees, inference of cross-species gene flow, and species delimitation. In this review, we introduce the major features of the MSC model, discuss full-likelihood and heuristic methods of species tree estimation, and summarize recent methodological advances in inference of cross-species gene flow. We discuss the statistical and computational challenges in the field and research directions where breakthroughs may be likely in the next few years.

Key words: anomaly zone, BPP, deep coalescence, gene flow, Markov chain Monte Carlo, multispecies coalescent, species tree

Introduction
Developed in the 1980s, the coalescent is a stochastic process that describes the genealogical history of a sample of DNA sequences taken from a population [1–3]. Whereas traditional population genetic models of drift and mutation describe changes in allele frequencies over generations in the population, the coalescent focuses on the sample and traces the genealogical history of lineage joining of the sampled sequences backwards in time. The coalescent model is in particular suited to inference using genetic sequence data [4–7].

The multispecies coalescent (MSC) is an extension of the single-population coalescent to the case of multiple species [8]. It integrates the process of species divergences and the within-population process of drift and mutation. Placing the coalescent in the context of a species phylogeny makes it possible to use the ever-increasing genomic sequence data from multiple species to address a number of important biological questions, and in the past two decades, the MSC has emerged as the natural framework for such inferences. These include estimation of population parameters (such as species divergence times, population sizes for extant species and extinct ancestors, and rates of cross-species gene flow), estimation of species phylogeny accommodating heterogeneous gene genealogies across the genome, and delineation of species boundaries (species delimitation) [9–12]. In molecular phylogenetics, incorporation of the MSC to accommodate the so-called gene-tree–species-tree conflicts has been heralded as a “paradigm shift” [13]. Stochastic fluctuation in genealogical history of sequences across the genome, when accommodated in the model, is not a ‘conflict’ or ‘problem’, but rather a source of information for important evolutionary parameters such as ancestral population sizes [14–16] and rates of cross-species gene flow [17, 18].

The past decade has seen exciting advancements in the implementation and extension of the MSC model for inference using genomic sequence data. The data we consider in this review are sequence alignments at hundreds or thousands of loci, with the different loci having independent coalescent histories while all sites in the sequence at the same locus share the same history. Ideal data for such analysis are short segments sampled from the genome that are far apart [16]. While we use the term gene or locus, the data should ideally be noncoding DNA, although exonic data have been successfully used in such analyses [19, 20]. We describe the major features of the MSC model (in particular, the probability distribution of gene trees and coalescent times), and discuss its applications in two major areas: the estimation of the species phylogeny and the inference of cross-species gene flow. We focus on full-likelihood methods (maximum likelihood or ML and Bayesian...
The basic coalescent model has two characterizations of the same process, the forward Fisher-Wright model and the backward coalescent approach of focusing on the sample backwards in time, lineages join or coalesce when we reach their common ancestors. While the sample backwards in time, lineages join or coalesce (fig. 1b). When we trace the genealogical history of the next generation are generated by random sampling of parental sequences to choose from, the probability that the two sequences pick the same parent (that is, they coalesce) in the previous generation is \( \frac{1}{N} \). In other words, coalescent occurs as a Poisson process at the rate of \( \frac{1}{2N} \), faster in smaller populations, and the coalescent time (the waiting time until the two sequences find their common ancestor) has a geometric distribution with the mean of \( 2N \) generations. Thus two sequences sampled at random are on average separated by \( 2N \) generations or \( \theta = 4N \mu \) mutations per site, where \( \mu \) is the mutation rate per site per generation. Parameter \( \theta \), known as the population size parameter, is the average distance between two sequences sampled at random from the population. It is also known as heterozygosity and can vary hugely even between close species. Typical values include \( \theta \approx 0.1\% \) for the humans [28] and 0.1-5\% for Heliconius butterflies [29].

In analysis of sequence data, it is convenient to measure time by the mutational distance so that one time unit is the expected time to accumulate one mutation per site. With this time unit, the coalescent waiting time for two sequences (\( t_2 \)) is approximately exponential with the mean \( \frac{\theta}{2} \) and density

\[
f(t_2) = \frac{\theta}{2} e^{-\frac{\theta}{2} t_2}.
\]

If there are \( n > 2 \) sequences in the sample, there will be \( \binom{n}{2} \) pairs and each pair coalesce at the rate of \( \frac{\theta}{2n} \), with the total rate \( \binom{n}{2} \cdot \frac{\theta}{2} \). The time until the next coalescent event has an exponential distribution with mean \( \frac{2}{\binom{n}{2}} / (\binom{n}{2} / (\binom{n}{2} - 1)) \). When a coalescent occurs, each of the \( \binom{n}{2} \) pairs has the same probability to join. The number of lineages is then reduced from \( n \) to \( n-1 \), and the process repeats, until the most recent common ancestor (MRCA) is reached (fig. 1b).

The \( n-1 \) successive coalescent events generate a genealogical tree \( (G) \) of the sequences in the sample. This is a rooted tree with the internal nodes ranked by age, and is called the ranked tree or labelled history [30] (fig. 1b). The number of possible labelled histories for a sample of size \( n \) is \( H_n = \prod_{i=2}^{n} \binom{i}{2} = \frac{n(n-1)!}{2} \), and each of them occurs with equal probability, \( f(G) = 1/H_n \). Furthermore, the \( n-1 \) coalescent times \( t = (t_n, t_{n-1}, \ldots, t_2) \) are independent exponential variables, with means \( E(t_i) = \frac{\theta}{2} / (\binom{i}{2}) \). The joint probability density of the gene
independently in different populations, with the coalescent size parameters (or equivalently the gene tree fits inside the species tree species divergence time: sequences split before species divergence time:

This intrinsic constraint between the species tree and the gene trees is the source of computational challenges in Bayesian implementations of the MSC model. Note that the first two histories correspond to the same rooted gene tree $G_1$, and there are three gene trees: $G_1, G_2$, and $G_3$.

tree and coalescent times is thus

$$f(G,t) = \frac{1}{\prod_{i=2}^{n}(\frac{i}{g})} \prod_{i=2}^{n} \left( \frac{\theta_i}{\theta} \right) \exp \left\{ -\frac{(i-1)}{\theta} t_i \right\}$$  \hspace{1cm} (2)

Multispecies coalescent: basic features

The extension of the single-population coalescent to multiple species has been called the interspecific coalescent [31] or censored coalescent [8], and is now commonly known as the multispecies coalescent (MSC) [32]. Suppose there are $s$ species, which are related through a species phylogeny. Instead of a single parameter $\theta$, the model now involves two sets of parameters: $s-1$ species divergence times ($\tau$s) and $2s-1$ population size parameters ($\theta$s), with a total of $3s-2$ parameters (fig. 2). Both $\tau$s and $\theta$s are measured in the expected number of mutations per site.

Given the species tree, coalescent events occur independently in different populations, with the coalescent rate ($\frac{1}{\theta}$) given by the population size. When we trace the history of the sequences at a locus backwards in time and reach a speciation event, the coalescent process and rate are reset, because of the change in population size and because of sequences coming from the sibling species. For example, in figure 3, sequences $c_1$ and $c_2$ coalesce at the rate $\frac{1}{\theta}$ in species $C$. When they enter species $BC$ at time $\tau_{BC}$, the coalescent rate (for each pair) is reset to $\frac{1}{\theta_{BC}}$ and the number of lineages becomes 3. Furthermore, we assume that gene trees at different loci are independent. One important feature of the MSC model is that the divergence time between sequences from two species must be greater than the species divergence time: sequences split before species or equivalently the gene tree fits inside the species tree. This intrinsic constraint between the species tree and the gene trees is the source of computational challenges in Bayesian implementations of the MSC model.

There are two important probability distributions under the MSC model: the (marginal) probabilities of gene tree topologies [21, 33, 34] and the joint distribution of the gene tree topology and coalescent times [8]. The former is useful for two-step methods of species tree estimation, which use reconstructed gene tree topologies as data, while the latter is used in full-likelihood methods, which use information in gene-tree branch lengths (coalescent times) as well.

Probabilities of gene tree topologies

Under the MSC model, the gene tree topologies and coalescent times have a joint probability distribution given the species tree and parameters. For small species trees, it is easy to derive the marginal probability of gene tree topologies [2, 33, 35]. This line of work typically assumes one sequence sampled per species at every locus, so that there is no coalescent in modern species at the tips of the species tree. The case of three species is considered in [2]. Let the three species be $A, B$ and $C$, with the phylogeny $S = (A, (B, C))$ (fig. 2). Let the divergence times be $\tau = (\tau_{BC}, \tau_{ABC})$ and the population sizes be $\theta = (\theta_{BC}, \theta_{ABC})$. Suppose three sequences are sampled from the three species ($a, b, c$). There are three possible gene tree topologies: $G_1 = (a, (b, c))$ matches the species tree, while $G_2 = (b, (c, a))$ and $G_3 = (c, (a, b))$ are the mismatching gene trees.

When we trace the genealogy of the three sequences, sequences $b$ and $c$ may coalesce in population $BC$ as a Poisson event at the rate of $\frac{\lambda}{\theta_{BC}}$ just as in the single-population coalescent. Note that the probability that a Poisson event of rate $\lambda$ does not occur in a time interval $t$ is $e^{-\lambda t}$. Thus the probability that sequences $b$ and $c$ do not coalesce in population $BC$ or over the time interval $\Delta \tau = \tau_{ABC} - \tau_{BC}$ is

$$\phi = e^{-2\Delta \tau / \theta_{BC}} = e^{-2(\tau_{ABC} - \tau_{BC}) / \theta_{BC}}.$$  \hspace{1cm} (3)

Here $\Delta \tau / (\theta_{BC})$ is known as the internal branch length in coalescent units—one coalescent unit in population $BC$ is $2N_{BC}$ generations or $\theta_{BC}/2$ mutations per site. If $b$ and $c$ coalesce in population $BC$, the gene tree must be $G_1$. Otherwise all three sequences enter species $ABC$ and coalesce in random order so that the three gene trees
occur with equal probability. Thus the probabilities for the three gene trees \((G_1, G_2, G_3)\) are
\[
\begin{align*}
P(G_1) &= (1 - \phi) + \frac{1}{2} \phi = 1 - \frac{2}{3} \phi, \\
P(G_2) &= P(G_3) = \frac{1}{3} \phi.
\end{align*}
\]

For certain species trees and parameter values, a mismatching gene tree may be more probable than the matching gene tree. The species tree is then said to be in the anomaly zone [33, 34]. The anomaly zone does not exist for species trees of three species — as \(P(G_1) > P(G_2) = P(G_3)\) in eq. 4, but can occur for asymmetrical species trees of four species, and for any species tree of five or more species [34].

Consider the asymmetrical species tree for four species \(S = (A, (B, (C, D)))\) of figure 4, and suppose the three divergence times are very close, with \(\tau_{ABCD} \approx \tau_{BCD} \approx \tau_{CD} \approx \tau_{D}\). Then all three coalescent events for the four sequences \((a, b, c, d)\) will most likely occur in the root population \(ABCD\), so that the \(18 = \frac{1}{2} \cdot \frac{1}{3} \cdot \frac{1}{3}\) labelled histories will have nearly equal probability \(\frac{1}{18}\). There are 15 possible rooted gene trees, 12 asymmetrical and 3 symmetrical. Each symmetrical gene tree (e.g., \(G_2\) in fig. 4) corresponds to two labelled histories \((G_{2a}, G_{2b})\), so that its probability is \(\sim \frac{1}{18}\). Each of the 12 asymmetrical gene trees (e.g., \(G_1\) in fig. 4) is compatible with only one labelled history, with probability \(\sim \frac{1}{18}\). Thus \(P(G_2) \approx 2P(G_1)\). When the divergence times \((\tau_s)\) are unequal but the internal branches are short enough, it is possible for the symmetrical mismatching gene tree \(G_2\) to have a higher probability than the matching asymmetrical gene tree \(G_1\), in which case the species tree is in the anomaly zone.

If the species tree is in the anomaly zone, the simple majority-vote approach of using the most commonly observed gene tree as the estimate of the species tree is statistically inconsistent: the more gene trees there are, the more certain that the species-tree estimate will be incorrect. Note that the existence of the anomaly zone is not an intrinsic difficulty for species tree estimation; it instead highlights the importance of adopting a proper statistical inference framework. Full-likelihood methods are consistent for all species trees both in and outside the anomaly zone, as they accommodate the probability distribution of the gene trees under the MSC appropriately. The discussion of the anomaly zone typically assumes true gene trees and ignores phylogenetic reconstruction errors in estimated gene trees. There have been only a handful of empirical examples of the anomaly zone, in African Anopheles mosquitoes [20], skinks [36], flightless birds [32], and gibbons [19].

The probabilities of gene tree topologies can be used to calculate the likelihood function for estimating the species tree using (reconstructed) gene trees as input data, as in the STELLS program [38]. However, popular heuristic methods such as MP-EST [39] and ASTRAL [40] do not use this theory and are instead based on species triplets or quartets. Furthermore, calculation of the probabilities of gene tree topologies, which involves summing over all coalescent histories that are compatible with each gene tree, becomes expensive when the number of species increases [21].

Joint probability distribution of gene trees and coalescent times

While the marginal probability of the gene tree topology may be challenging to compute, it is straightforward to derive the joint distribution of gene tree topologies and coalescent times. The general form, for an arbitrary species tree and an arbitrary number of sequences, is given in [8].

The joint density of gene trees and coalescent times is a product over the populations on the species tree, and as a result we focus on the contribution from one population. A population is represented by a branch on the species tree (say \(XY\)) or by the daughter node of the branch (say \(X\)). Let \(\tau_X\) and \(\tau_Y\) be node ages or divergence times, and \(\theta_X\) be the population size.

Suppose \(m\) sequences enter the population at time \(\tau_X\) and \(m - 1\) sequences leave the population at time \(\tau_Y\), with \(1 \leq l \leq m\). For example, in the gene tree of figure 3, \(m = 3\) lineages enter population \(BC\) while \(l = 2\) lineages leave it. Unlike the single-population coalescent, under the MSC, lineages entering a population do not necessarily find their common ancestor in that population, and the coalescent process may be 'censored' [8]. Note that if \(X\) is the root of the species tree, \(l = 1\) must be 1.

The MSC density for the part of the gene tree residing in population \(XY\) is the product of three components. The first is the joint density of the \(m - l\) independent exponential coalescent waiting times \(\{t_{X1}, \ldots, t_{1}^{X}\}\). The second component is for the gene tree topology in \(XY\), and is a product of \(m - l\) probabilities, each being the probability, \(1/(\binom{m}{l})\), of choosing two out of \(i\) lineages to join, for \(i = m, m - 1, \ldots, l + 1\). These two components are the same as in the single-population coalescent. The third component is the probability that no coalescent events occur in the last time interval before reaching \(\tau_Y\). Multiplying the three components, we obtain the MSC.
density of the gene tree in $XY$ as
\[
\left(\frac{2}{\theta X}\right)^m \exp\left\{ -\sum_{i=1}^m \frac{i(i-1)}{\theta X} Y - \frac{i-1}{\theta X} X - \sum_{i=1}^m \frac{i}{\theta X} Y \right\}.
\]

For example, the contribution of species $BC$ to the MSC density of the gene tree in figure 3 is
\[
\frac{2}{\theta BC} \exp\left\{ -\frac{6}{\theta BC} Y - \frac{2}{\theta BC} \left( \tau_{BC} - t_{BC} - t_{BC} \right) \right\}.
\]

As coalescent processes in different populations operate independently, the MSC density for the whole gene tree at a locus is the product of the contributions across all populations. For the gene tree of figure 3, this is
\[
f(G_j|S, \Theta) = \left[ \frac{2}{\theta A} e^{-\frac{2}{\theta A} t^A} \right] \times \left[ e^{-\frac{2}{\theta BC} t_{BC}} \right] \times \left[ \frac{2}{\theta BC} e^{\frac{2}{\theta BC} t_{BC}} \right] \times \left[ e^{-\frac{6}{\theta ABC} t_{ABC} - \frac{2}{\theta ABC} t_{ABC}} \right].
\]

The four pairs of brackets correspond to species $A$, $C$, $BC$, and $ABC$, respectively. Coalescent is not possible in species $B$ as only one sequence is sampled from that species.

With multiple loci in the data, the joint MSC density of the gene trees is a product across all loci, because the genealogical histories at different loci are assumed to be independent. The formulation allows the loci to have different sampling configurations. For example, the number of sequences from each species may vary among loci and some species may be missing at some loci.

**Species Tree Inference under the MSC**

**Species-tree – gene-tree conflicts**

The gene tree representing the coalescent history of the sequences at a locus may not match the species tree. Such a discordance may occur because when we trace the history of the sample backwards in time, sequences from different species may not coalesce as soon as they reach the most recent common ancestor on the species tree but instead coalesce in more ancient ancestors (e.g., gene trees $G_{1b}$, $G_{2a}$, $G_{3a}$ in fig. 2). This *delayed coalescence* or *deep coalescence* is also known as *incomplete lineage sorting* (ILS). While several biological processes, including gene duplication followed by gene loss or horizontal gene transfer [41, 42], can cause the gene tree to differ from the species tree as well, deep coalescence is more fundamental because coalescent is simply biological reproduction and drift and thus may affect every species. Deep coalescence is more common when multiple species arose through a rapid succession of speciation events resulting in very short internal branches on the species tree relative to the coalescent waiting time (note that $\phi$ in eq. 3 is greater for smaller $\Delta t$ and larger $\theta_{BC}$). The existence of the anomaly zone is an extreme case of deep coalescence. Deep coalescence is related to how short the internal branches are, rather than how deep they are on the species tree, and may thus occur in both shallow and deep species trees [43].

**Full likelihood methods**

ML methods [44, 45] and Bayesian inference [46–49] use the joint distribution of gene trees and coalescent times [8] and operate on multilocus sequence data directly. Let the sequence data be $X = \{X_j\}$, where $X_j$ is the alignment of $n_j$ sequences at the $j$th locus, for $j = 1, 2, …, L$. Let $S$ be the species tree and $\Theta = \{\tau, \it{\theta}, \eta\}$ be the vector of parameters, including species divergence times ($\tau$), population sizes ($\it{\theta}$), and parameters in the mutation model ($\eta$). The likelihood of the sequence data given the MSC model has the form
\[
f(X|S, \Theta) = \prod_{j=1}^L \int f(X_j|G_j, t_j, \eta) f(G_j, t_j|S, \Theta) dt_j,
\]
where $f(X_j|G_j, t_j, \eta)$ is the phylogenetic likelihood given the gene tree $G_j$ and branch lengths $t_j$ at locus $j$ [50], while $f(G_j, t_j|S, \Theta)$ is the MSC density of the gene tree described above [8]. As the genealogical histories at different loci are independent, the likelihood of the sequence data is a product across all loci. The summation in eq. 8 is over all possible gene tree topologies for the sequences, and the integral is $(n_j - 1)$-dimensional, over the $n_j - 1$ coalescent times on each gene tree. The gene trees and coalescent times are not observed, and the likelihood function averages over them, accommodating their uncertainties.

The species tree $S$ and the MSC parameters $\Theta$ can be estimated using ML by maximizing eq. 8. Both the phylogenetic likelihood $f(X_j|G_j, t_j, \eta)$ and MSC density $f(G_j, t_j|S, \Theta)$ are straightforward to calculate, but averaging over all possible gene tree topologies and coalescent times at each locus is computationally infeasible except for small data sets. The only ML implementation available is the 3S program [44, 45], which enumerates the gene trees and uses numerical integration (Gaussian quadrature) to calculate the integrals. Although limited to three species and three sequences per locus, 3S can handle tens of thousands of loci.

With more than three species, the Bayesian method has a computational advantage over ML, with the Markov chain Monte Carlo (MCMC) algorithm averaging over the gene trees and coalescent times. We assign prior distributions to the species tree and model parameters. For example, the species tree can be assigned a uniform prior over all rooted trees, while the population-size parameters ($\it{\theta}$s) can be assigned gamma or inverse-gamma priors. The inverse-gamma priors for $\it{\theta}$s are conjugate (so that both the prior and posterior for $\it{\theta}$s are inverse-gamma), allowing the $\it{\theta}$s to be integrated out analytically [51], which helps with MCMC mixing. The age of the species-tree root can be assigned a gamma or inverse-gamma prior, while the other node ages can be constructed using a Dirichlet distribution [52].
The MCMC algorithm samples from the joint posterior distribution of the species tree, the MSC parameters, and the gene trees at all loci:

\[ f(S, \Theta, G, t|X) \]

\[ \propto f(S, \Theta) \prod_{j=1}^{L} f(X_j|G_j, t_j, \eta_j)f(G_j|t_j, S, \Theta). \quad (9) \]

In particular the samples of \((S, \Theta)\) generated by the algorithm are from the marginal posterior \(f(S, \Theta|X)\), and the frequency at which a species tree is visited is an estimate of its posterior probability. In this way, MCMC averages out the gene trees and coalescent times numerically.

The first implementation of the Bayesian approach is the program BEST [53]. This uses the samples of gene trees with branch lengths produced by MRBAYES [54] and applies an importance-sampling correction because MRBAYES does not assume that the gene trees are distributed according to the MSC density. This strategy does not work well, as the species tree and the gene trees place tight constraints on each other in the MSC model. Currently two Bayesian programs under the MSC are in common use: *BEAST [46] and BPP [47–49], both of which explicitly use the MSC model. The algorithm in BPP for species tree inference goes through several proposal steps in each MCMC iteration, as follows.

1. Update the coalescent times \(t_j\) on the gene tree at each locus \(j\);
2. Update the gene tree topology \(G_j\) at each locus \(j\) through a subtree-pruning-and-regrafting (SPR) algorithm;
3. Update the population sizes \((\Theta_s)\);
4. Update the species divergence times \((\tau_s)\);
5. Update the species tree topology \(S\) through an NNI (nearest-neighbour interchange) or SPR move, which may change the gene trees to avoid conflicts;
6. Use a multiplier to rescale all node ages on the species tree and on all gene trees.

Perhaps the greatest challenge in such MCMC algorithms comes from the constraint between the species tree and the gene trees. Consider step 4 for changing species divergence time \(\tau_{AB}\), the age of the ancestral node for two sister species/clades \(A\) and \(B\). Let \(t_{ab}\) be the sequence divergence time for two sequences from \(A\) and \(B\). Then \(\tau_{AB} \leq t_{ab}\). If the dataset includes thousands of loci and many sequences from \(A\) and \(B\) at each locus, the smallest of \(t_{ab}\) among all loci may be almost identical to the current \(\tau_{AB}\). Then when we use a sliding window to change \(\tau_{AB}\), the window size will have a width near zero, and the MCMC is virtually stuck. A "rubber-band" algorithm was proposed in [8], which changes \(\tau\) and the affected node ages on gene trees jointly. Similarly, in step 5, it is very inefficient to change the species tree when all gene trees are fixed. A breakthrough was to make coordinated changes to the gene trees when an NNI algorithm is used to change the species tree [47]. The algorithm has since been extended to SPR [48, 55] and ported to *BEAST as well [55, 56]. Those improvements have pushed the limit of datasets that can be analyzed using Bayesian MCMC programs from ~100 to ~10,000 loci [19, 20].

Heuristic or summary methods

Many heuristic methods for species tree estimation have been developed, which use summaries of the data rather than the original multilocus sequence alignments. For extensive reviews, see [9, 10, 22, 23]. Here we mention four commonly used ones: MP-EST [39], ASTRAL [40], NJ-ST [57], and SVDQUARTETS [58].

MP-EST [39] estimates triplet gene trees under the molecular clock (rate constancy among lineages), and then uses a composite likelihood function, treating the frequencies of the triplet gene trees as input data from a trinomial distribution (with probabilities given in eq. 4). A composite or pseudo-likelihood is constructed by multiplying those probabilities for all possible triplets, ignoring lack of independence among them. This composite likelihood is maximized to estimate the species tree.

ASTRAL [40] uses a phylogenetic method to infer unrooted gene trees, and extracts the quartets from them. It then finds the species tree that is most compatible with the quartets in the set. A procedure has also been developed to attach local support values for nodes on the inferred species tree [59].

NJ-ST [57] uses a distance method to estimate an unrooted species tree from a collection of unrooted gene trees. The species tree estimate is the neighbor-joining tree built from a distance matrix where the distance between two species is defined as the average number of internal nodes on the gene tree between the species.

All those three methods are two-step methods, treating estimated gene tree topologies as data. They are consistent, with the probability to recover the correct species tree approaching one when the number of gene trees increases. As discussed above, the anomaly zone does not exist for rooted triplets or equivalently for unrooted quartets. However, the argument for consistency is based on the assumption that the input gene trees are known without error. Phylogenetic reconstruction errors are known to affect the performance of two-step methods [60]. Furthermore, as those two-step methods uses gene tree topologies but not branch lengths or coalescent times, they suffer from unidentifiability issues. They can estimate the species tree topology but not all parameters in the MSC model.

Another summary method is called SVDQUARTETS [58]. This is a quartet method, designed for data of coalescent-independent sites, sites that have independent
simulate were simulated using the BPP and MP-EST. The number of replicates is 100 for BPP and 500 for the other methods.

Comparison between full-likelihood and heuristic methods
Figure 5 shows results from a small simulation to illustrate the different performance of a full likelihood method (BPP), two summary methods (ASTRAL and MP-EST), and ML analysis of concatenated data. The species tree is challenging with short internal branches in both sets of simulations. BPP recovered the true species tree with higher probability than the two summary methods and concatenation. For set 1, all four methods are consistent, with the probability of recovering the true species tree approaching 1 for every method when the number of loci increases. For set 2 the species tree is in the anomaly zone, and concatenation/ML is inconsistent, with the probability for the mismatching balanced tree approaching 1, while the other three methods are consistent. Note that the ML method applied to concatenated data assumes one tree and one set of divergence times for all loci and can be inconsistent [67].

Heuristic methods based on data summaries have a huge computational advantage over full-likelihood methods. For large datasets with hundreds or thousands of species and thousands of loci, they may be the only methods that are currently feasible computationally. Heuristic methods have poorer statistical performance than full-likelihood methods, and the difference can be large for challenging species trees with short internal branches [9, 19, 62, 64, 68]. As two-step methods typically ignore phylogenetic reconstruction errors in gene trees, their performance may suffer from uncertainties in the gene trees [60, 64]: for those methods, species trees are only as good as the gene trees on which they are built [9, 23].

An important strength of full-likelihood methods is that they can provide estimates of parameters in the MSC
model when the species tree is fixed [16, 56]. The MSC model for a species tree of \( s \) species has \( s - 1 \) divergence times (\( \tau_s \)) and \( 2s - 1 \) population sizes (\( \theta_s \)) (fig. 2), all of which can be identified and estimated by full likelihood methods using multilocus sequence data. In contrast, summary methods use only a portion of information in the data and are unable to identify all parameters in the model. For example, in the case of three species, the MSC model involves seven parameters (two \( \tau_s \) and five \( \theta_s \)), but there are only two distinct frequencies of gene trees (eq. 4), so that two-step methods using gene tree topologies alone can identify only the internal branch length in coalescent units: \( \phi \) or \( 2\Delta \tau / \theta_{BC} \) of eq. 3. For large datasets for which species tree estimation using full-likelihood methods is too expensive, it may be advisable to use summary methods to infer the species tree, and then full likelihood methods to estimate the population parameters on the species tree.

**Multispecies Coalescent with Migration or Introgression**

In the past two decades, analyses of genomic data have highlighted the prevalence of cross-species gene flow [69–71]. Ancient gene flow has been detected in a variety of species, from mosquitoes [20, 72] and butterflies [73] to hominins [74]. Like deep coalescence, gene flow causes genealogical fluctuations across the genome, posing challenges to species tree estimation [75–78]. Perhaps more importantly, hybridization can lead to rapid genomic changes, leading to beneficial new phenotypes and ecological adaptations. Inferring the mode and timing of gene flow may help us to achieve a better and richer understanding of the process of speciation and adaptation [70, 71].

Two types of models of gene flow have been developed, both as extensions to the MSC model. The first is the migration model (MSC+M), also known as the isolation-with-migration (IM) model [17, 79], which assumes that gene flow occurs at a certain rate every generation. The second is the hybridization/introgression model (MSC+I or MSci) [80, 81], in which hybridization occurs at a fixed time point in the past. Here we discuss the distribution of the gene trees under these models of gene flow. ML and Bayesian methods of inference proceed as before (eqs. 8 & 9), except that the model may involve parameters that measure the timing and strength of gene flow and the gene tree may include the migration or introgression history, as well as the tree topology and coalescent times. We will also mention a few heuristic methods for testing the presence of gene flow and estimating its rate.

**Isolation-with-migration (IM)**

Consider two populations \( A \) and \( B \) with population sizes \( \theta_A \) and \( \theta_B \) that have been exchanging migrants at the rates of \( m_{AB} \) and \( m_{BA} \) since their divergence at time \( \tau_R \) (fig. 6a). The parameter vector in the IM model for two species is thus \( \Theta = \{ \theta_A, \theta_B, \theta_R, \tau_R, M_{AB}, M_{BA} \} \). Here the population migration rate \( M_{AB} = m_{AB}/N_B \) is the expected number of migrants from \( A \) to \( B \) (in the real world with time running forward) per generation, with \( m_{AB} \) to be the proportion of individuals in population \( B \) that are immigrants from population \( A \). The rate \( M_{BA} = m_{BA}/N_A \) is defined similarly. Note that migration rates in the IM model reflect the long-term effects of migration, genetic drift, recombination, as well as natural selection purging introduced alleles [71]. We consider the probability density of the gene trees under the IM model. There are two formulations, depending on whether the gene tree at a locus includes the migration history.

In the first formulation, the gene tree includes the tree topology and coalescent times, but not the migration history (or with the migration history integrated out). This relies on the theory developed in the *structured coalescent* framework in which the backwards-in-time process of coalescence and migration is described using a continuous-time Markov chain [82–84]. The state of the chain is specified by the number of sequences in the sample and their population IDs [18, 45, 61]. Consider the IM model for two species (\( A \) and \( B \)) of figure 6a and suppose two sequences (\( a \) and \( b \)) are sampled at locus \( j \) (fig. 6b), so that the gene tree is just the sequence divergence time \( t_j \) (we suppress the subscript and write \( t_j \) as \( t \) hence). When we trace the genealogy of the two sequences backwards in time, the sequences may move between populations and may coalesce. The possible states are \( s_{AA}, s_{AB}, s_{BB}, s_A \) and \( s_B \). Here \( s_{AA} \) means that both sequences are in population \( A \), \( s_{BB} \) means both are in \( B \), while \( s_{AB} \) means one is in \( A \) and the other in \( B \). With only two sequences in the sample, there is no need to distinguish \( s_{AB} \) and \( s_{BA} \). If the two sequences have coalesced, the state becomes \( s_A \) or \( s_B \), and these are lumped into one artificial absorbing state, \( s_{AB} \), since there is no need to trace the history any further. Let \( Q = \{ q_{uv} \} \) be the generator matrix for the Markov chain over the time interval \( (0, \tau_R) \), where \( q_{uv} \) is the instantaneous rate of transition from states \( u \) to \( v \). This is

\[
Q = \begin{pmatrix}
 s_{AA} & s_{AB} & s_{BB} & s_{AB}
\end{pmatrix}
\begin{pmatrix}
 -2(m_1 + \frac{1}{2\mu}) & 2m_1 & 0 & 0 \\
 m_2 & -(m_1 + m_2) & m_1 & 0 \\
 0 & 2m_2 & -(m_1 + \frac{1}{2\mu}) & \frac{1}{2} \\
 0 & 0 & 0 & 0
\end{pmatrix}
\]

Here the time unit is one mutation per site, \( m_1 = \frac{4M_{AB}}{\theta_B} = \frac{m_{AB}}{\mu} \) is the mutation-scaled migration rate into species \( A \) and \( m_2 = \frac{4M_{BA}}{\theta_A} = \frac{m_{BA}}{\mu} \) is the rate into \( B \). Note that the Markov chain runs backwards in time while the migration rates (e.g., \( M_{AB} \) and \( m_2 \)) are defined under the real-world forward-in-time view. For example, in the first row, the transition from \( s_{AA} \) to \( s_{AB} \) represents migration from \( A \) to \( B \) in the real world, and either sequence in \( A \) can be the migrant, so that the rate is \( 2m_{BA} \) per generation or \( 2m_{BA}/\mu = m_1 \) per mutational

---

**Downloaded from https://academic.oup.com/nsr/advance-article/doi/10.1093/nsr/nwab127/6321855 by guest on 22 July 2021**
FIG. 6. (a) Migration (MSC+M) or isolation-with-migration (IM) model for two species (A and B) showing the parameters. (b) A gene tree for two sequences (a and b) with divergence time \( t \) and four migration events, with \( t = \sum_{k=1}^{4} \tau_k \). The migration rates (per mutational time unit) are shown beneath the horizontal lines representing migration events. Note that time runs forward in (a) when we define migration rates (\( M_{AB} \) or \( m_2 \)) and backward in (b) when we trace the genealogical history at the locus.

time unit. The transition from \( s_{AA} \) to \( s_{AB} \) means that the two sequences coalesce in A, with rate \( \frac{2 \theta}{\tau} \). State \( s_{BB} \) is not reachable from \( s_{AA} \) instantaneously.

The transition probability matrix over any time \( 0 < t < \tau \) is then

\[
Q(t) = \begin{pmatrix}
\frac{2 \theta}{\tau} & \frac{2 \theta}{\tau} \\
\theta & \theta + \frac{2 \theta}{\tau}
\end{pmatrix}
\]

where \( \theta = \frac{2 \theta}{\tau} \) is the probability that both sequences are in state \( s_{BB} \), \( \tau \) is the probability that both sequences are in state \( s_{BB} \) or \( s_{AB} \), and \( \theta = \theta_B \) [18].

In the case of \( s = 2 \), the two sequences coalesce during \( \Delta t \) and both enter the ancestral population at time \( \Delta t \). In the gene tree of figure 6b, the time period \( (0, t) \) is broken into six time segments by the coalescent, migration, and speciation events, and within each segment, the number of lineages is constant, as are the coalescent and migration rates.

Then the probability density of the gene tree \( G \) is given by the rates for the coalescent and migration events times the probability of no events over the whole time period

\[
f(G) = \left[ m_2^2 \exp \left( -\frac{2}{\theta} \sum_{i=1}^{3} m_i (s_1 + 2 s_2 + s_3 + s_5) \right) \right] \\
\times \left[ m_2^2 \exp \left( -\frac{2}{\theta} \sum_{i=1}^{3} m_i (s_1 + s_2 + s_3 + 2 s_4 + s_7) \right) \times \left[ \frac{2}{\theta} \exp \left( -\frac{2}{\theta} \sum_{i=1}^{3} m_i (s_1 + 2 s_2 + s_3 + s_5) \right) \right] \right)
\]
but not migration events. The likelihood depends on the gene tree and coalescent times over the migration history at every locus and becomes inefficient at high migration rates, as there will be many migration events to average over. Note that the sequence occurs after the waiting time \( s_6 = t - \tau \), so the rate is \( \frac{2}{6\theta} \) and the probability of no event is \( e^{-\frac{2}{6\theta} (t-\tau)} \).

Unlike eq. 11 in which the gene tree means divergence time \( t \), here \( G \) represents the full coalescent and migration history at the locus, such as the (backwards-in-time) transitions of sequence \( b \) from \( B \) into \( A \) at time \( s_1 \) and back to \( B \) at time \( s_1 + s_2 \), and so on. If we sum over all possible histories which have divergence time \( t \) (one of which is that of fig. 6b), the marginal density \( f(t) \) will be given by eq. 11.

Eq. 12 is easily generalizable to more species and sequences. For a general gene tree, one can break the time period from the present time to the root of the gene tree into time segments by the coalescent and migration events at the locus and by the speciation events. Then the probability density of the gene tree is simply given as the product of rates for the coalescent and migration events that occurred times the probability of no events over the whole time period.

This formulation is used in Bayesian implementations of the IM model such as IMA [88, 89] and G-PHOCS [90]. The posterior is given by eq. 9 except that the gene tree \( G_i \) includes the migration history. G-PHOCS is an extension of an earlier version of BPP [8,16] and is computationally more efficient than IMA and can deal with a few thousand loci. The algorithm averages over the migration history at every locus and becomes inefficient at high migration rates, as there will be many migration events to average over. Note that the sequence likelihood depends on the gene tree and coalescent times but not migration events.

**Multispecies coalescent with introgression (MSci)**

The introgression or MSci model assumes that gene flow occurs between species at fixed time points in the past (fig. 7). There are two types of nodes on the species tree: speciation nodes and hybridization nodes. While a speciation node (if it is not the root) has one parent, a hybridization node has two parents, with their contributions to the hybrid species represented by probabilities \( \varphi \) and \( 1 - \varphi \). When we trace the history of sequences backwards in time and meet a hybridization node, each sequence picks one of the two parents according to probabilities \( \varphi \) and \( 1 - \varphi \). The parameters in the model includes the introgression probabilities as well as the species divergence/introgression times \( (\tau_s) \) and population sizes \( (8\theta_s) \), with \( \Theta = \{ \tau, \theta, \varphi \} \). The introgression probability \( \varphi \), also written as \( \gamma \), has been called (inappropriately) 'inheritance probability' or 'heritability'. Like the migration rate in the IM model, the introgression probability reflects the long-term effects of drift and selection on introgressed alleles. The MSci model has been referred to as the network multispecies coalescent (NMSC) [91, 92] or multispecies network coalescent (MSNC) [93, 94]. We avoid the term 'network' as it has been used to refer to a variety of processes including gene-tree reconstruction errors [95].

Four types of MSci models are implemented in BPP (fig. 7) [81]. In model A, two species \( SH \) and \( TH \) merge to form a hybrid species \( HC \). This scenario may be rare, but the model can be used to accommodate introductions involving ghost or unsampled species (fig. 8a&b). Model B assumes introgression from species \( RA \) to \( TC \) at time \( \tau_S = \tau_H \). This is distinguishable using genetic data from the alternative model in which there is introgression from \( RB \) to \( SC \) (B2, fig. 8d). Model C (fig. 7c) is a case of hybrid speciation. Model D assumes that two species \( RA \) and \( RB \) came into contact at time \( \tau_X = \tau_Y \) and exchanged migrants.

The two parental branches are sometimes called the ‘major hybrid edges’ and ‘minor hybrid edges’, according as \( \varphi > \frac{1}{2} \), and the binary species tree that remains after all minor hybrid branches are removed is called the ‘major species tree’ [95]. This characterization is useful if gene flow occurs in pulses as assumed by the MSci model but may be misleading if gene flow is continuous. For example, continuous migration at a low rate per generation can drastically change the gene tree distribution so that when the MSci model is fitted to the data, the major species tree may reflect gene flow, rather than species divergences [20, 72, 78].

![FIG. 7. (a) MSci models A, B, C, and D implemented in BPP [81], showing the parameters. In model A, two parental species \( SH \) and \( TH \) merge to form a hybrid species \( H \) at time \( \tau_H \), but both parental species become extinct (see fig. 8a&b for alternative interpretations). In model B, there is introgression from species \( RA \) to \( TC \) at time \( \tau_S = \tau_H \). In model C, species \( RA \) and \( RB \) come into contact to form hybrid species \( HC \) at time \( \tau_S = \tau_H = \tau_Y \). Model D assumes bidirectional introgression between species \( RA \) and \( RB \) at time \( \tau_X = \tau_Y \). Here the introgression probability \( \varphi \) is assigned to the horizontal (introgression) branch at each hybridization node whereas in [81] it is sometimes assigned to the vertical branch.](https://academic.oup.com/nsr/advance-article/doi/10.1093/nsr/nwab127/6321855)

![FIG. 8. (a&b) Two interpretations of model A, alternative to figure 7a, involving a ghost species \( X \). In model A2, species \( SUX \) contributes migrants to species \( TH \) at time \( \tau_H \) and has since become extinct or unsampled in the data, while in model A3, \( TUX \) is the ghost species. Models A1 (fig. 7a), A2, and A3 are indistinguishable using genetic data. (c&d) Two versions of model B, which are identifiable using genetic data.](https://academic.oup.com/nsr/advance-article/doi/10.1093/nsr/nwab127/6321855)
This is used in full-likelihood implementations of the MSci model. This joint density is very similar to that under the MSC without gene flow (eq. 7), with the only modification that each time a sequence passes a hybridization node, there is a probability \( \varphi \) or \( 1 - \varphi \) depending on the parental path taken. Thus

\[
\begin{align*}
 f(G_j, t_j | S, \Theta) &= \left[ e^{-\frac{2}{\theta} (q_j - r_j)} \right] \times \left[ \frac{\varphi}{2 \sigma_0} e^{-\frac{2}{\theta} (q_j - r_j)} \right] \\
&\quad \times \left[ 1 - \varphi \right] \times \left[ \frac{2}{\theta} e^{-\frac{2}{\theta} (q_j - r_j)} \right] \times \\
&\quad \times \left[ \frac{2}{\theta} e^{-\frac{2}{\theta} (q_j - r_j)} \right].
\end{align*}
\]

The five pairs of brackets correspond to species \( C, S, H, T \) and \( R \) (fig. 9b). For species \( S \) (i.e., \( SR \)), sequence \( c \) picks parental path \( S \) and coalesces with sequence \( a \) at time \( t_2 \), so that the contribution to the gene-tree density from \( S \) is \( \varphi \frac{2}{\theta} e^{-\frac{2}{\theta} (q_j - r_j)} \) Introggression is counted as an event in the receiving population (rather than the source population) when we trace the lineages backwards in time and reach a hybridization node.

Bayesian implementations of the introgression model can then proceed as before, with the joint posterior of the MSci model and parameters given by eq. 9, except that \( S \) now represents the MSci model, the parameter vector \( \Theta \) includes the introgression probabilities (\( \varphi \)) as well as the divergence/introgression times (\( \tau \)) and population sizes (\( \theta \)). and the gene tree \( G_j \) includes the introgression history at the locus. There are currently three Bayesian MCMC implementations of the MSci model: PHYLONET/MCMC-SEQ [93], *BEAST [94, 97], and BPP [81] (table 1). PHYLONET and *BEAST can allow changes to hybridization events in the MCMC and can infer the introgression model from the data. Those programs appear to reach their limits with \( \sim 100 \) loci. BPP assumes that the MSci model is specified and fixed and the program estimates the parameters under the model. It has been applied to datasets of over 10,000 loci [29, 81]. Also BPP implements four different types of introgression models (fig. 7), while only model A is available in PHYLONET and *BEAST.

Binary species trees generated by taking different parental paths at hybridization nodes are called “displayed species trees” [92] or “parental species trees”. An interesting formulation of the MSci model specifies the distribution of the gene trees as a mixture over the displayed species trees, with the mixing probabilities given by the introgression probabilities at the hybridization nodes (fig. 10) [e.g., 98, 99]. To simulate a gene tree, one would sample a displayed species tree first and then generate the gene tree according to the simple MSC model. This is in general incorrect as it forces all sequences at the locus to take the same parental path at each hybridization node, whereas correctly there should be a binomial sampling process when two or more sequences reach a hybridization node. In the model of figure 10, if sequences \( b \) and \( c \) reach species \( Y \), it should be possible
matches the species tree, while ABBA or genome is available from each species [102]. There based on the counts of site patterns when one sequence species (A, B, and C) or genome is available from each species. There are three parsimony-informative site patterns: AABB matches the species tree, while ABBA and BABA are the mismatching patterns, where A and B are any two distinct nucleotides. The probabilities for the two mismatching site patterns ABBA and BABA should be equal if there exists deep coalescence but no gene flow, but they are different if there is gene flow between the non-sister species (A and C or B and C) in addition to deep coalescence. Thus, gene flow can be tested by using the site-pattern frequencies to examine the deviation of

\[ D = \frac{f_{ABBA} - f_{BABA}}{f_{ABBA} + f_{BABA}} \]

from 0. The D-statistic has been extended to the case of five species, assuming a symmetric species tree in the so-called DFOIL test [103]. The site pattern frequencies can also be used to estimate the introgression probability, as in the program HYDE [104, 105]. From

\[ f_1 \frac{p_{ABBA} - p_{ABAB}}{p_{ABBA} - p_{ABAB}} = \frac{1}{1 - \phi} \tag{16} \]

one gets the estimate

\[ \phi = \frac{f_1}{f_1 + f_2} \tag{17} \]

This is based on the hybridization model with \( \tau_2 = \tau_1 \) and \( \Theta_2 = \Theta_1 \) (fig. 7c). The estimate should be biased if this symmetry does not hold.

A similar argument may be applied to gene tree topologies instead of site patterns. The probabilities of the two mismatching gene trees \((b, c, a)\) and \((c, a, b)\) are equal if there exists deep coalescence but no gene flow, but different if there is in addition gene flow between the non-sister species (A and C or B and C). Thus the observed frequencies of gene tree topologies can be used to estimate the introgression probability, as in the SNAQ method [95, 106]. Assume \( \Theta_5 = \Theta_4 \) and \( \Theta_3 = \Theta_2 \) (fig. 7c). The estimate should be biased if this symmetry does not hold.

The D-statistic cannot be used to detect gene flow between sister species or to estimate the time of introgression. Such unidentifiability issues also exist in other methods which detect hybridization events using genome-wide averages, such as the average interspecies sequence divergence [107] or the joint allele frequency spectrum [108].

Unidentifiability, low information content, and challenges of identifying the mode of gene flow

One area where more research is urgently needed is the identifiability of introgression models. If the probability distributions of the data are identical for two sets of parameter values (\( \Theta \) and \( \Theta' \)), with \( f(X|\Theta) = f(X|\Theta') \) for essentially every dataset \( X \), then \( \Theta \) is unidentifiable given data \( X \). Several studies have examined identifiability issues of summary methods that use gene tree topologies as data [76, 80, 91, 109], but little research has been done on full likelihood methods.

### Table 1. A partial list of computer programs implementing the MSC model with and without gene flow

| Method            | MSC         | IM & MSci |
|-------------------|-------------|------------|
| Full likelihood   | 3S          | 3S         |
| BPP               | IMA3        |            |
| *BEAST            | G-PHOCS     | BPP        |
|                   |             | *BEAST     |
|                   |             | PHYLONET   |
| Two-step          | ASTRAL      | PHYLONET   |
|                   | MP-EST      |            |
|                   | NJ-ST       |            |
|                   |             |            |

Downloaded from https://academic.oup.com/nsr/advance-article/doi/10.1093/nsr/nwab127/6321855 by guest on 22 July 2021
Some cases of unidentifiability are easy to identify. If it is impossible for two or more sequences to be in one species when we trace the genealogical history of the sample backwards in time, the population size (θ) for that species will be unidentifiable, since it takes two sequences to define a distance. For example, in the MSC model with no gene flow (fig. 2), the population sizes for the extant species are unidentifiable if only one sequence is sampled from each species per locus, but this unidentifiability disappears when multiple sequences are available from each species. Furthermore parameters or models that are unidentifiable using gene tree topologies alone may become identifiable when both gene trees and branch lengths (coalescent times) are used. In the case of three species, there are only three gene trees, so that use of gene tree topologies can identify only one (under the MSC model) or two (under the MSci model) parameters, whereas there are 7 (fig. 2) and 13 (fig. 7a) parameters in those two models respectively, which are all identifiable when information from both gene trees and coalescent times is used.

The identifiability of full-likelihood methods applied to data of multilocus sequence alignments, with multiple sequences per species, is the most interesting case, because full-likelihood methods are expected to be optimal from a statistical point of view and because multilocus alignments are the dominating data form in such analyses. Flouri et al. [81] conjectured that the MSci model is identifiable on multilocus sequence alignments as long as it is identifiable on data of gene trees with coalescent times. Given this, the problem of identifiability can be studied by considering the gene trees with coalescent times (G_j and t_j) as the input data.

It is noted that MSci model D (fig. 7d) has an identifiability issue of the label-switching type [81] (fig. 11). For every set of parameters, \( \Theta = (\theta_R, \theta_A, \theta_B, \theta_X, \theta_Y, t_A, t_B, \phi_X, \phi_Y) \), there is a “mirror” point \( \Theta' \), which has identical parameter values as \( \Theta \) except that \( \theta'_R = \theta_R, \theta'_A = \theta_A, \theta'_B = \theta_B, \theta'_X = 1 - \phi_X \) and \( \theta'_Y = 1 - \phi_Y \). Both \( \Theta \) and \( \Theta' \) have exactly the same likelihood, \( f(X|S, \Theta) = f(X|S, \Theta') \) for all possible data \( X \). This is a label-switching issue, and does not affect the utility of the model: one may apply a constraint such as \( \phi_X > \frac{1}{2} \) to remove the unidentifiability or apply more sophisticated post-processing of the MCMC sample if the “twin towers” are not well separated [110]. The cases where the bidirectional introgression involves non-sister species or where there are multiple introgression events are yet to be studied.

Even if all parameters are identifiable, typical datasets may lack information for their reliable estimation. For example, typical datasets may be highly informative about species divergence times, but not about population sizes for ancestral species, especially if those species correspond to very short branches on the species tree [111]. In the case of three species both gene flow between non-sister species and population structure in the ancestral species can cause the asymmetry in the probabilities of the two mismatching gene trees [112], so that the two models are unidentifiable using gene tree topologies alone. In general it may be hard to distinguish the different models of gene flow, such as the complete isolation model (MSC with no gene flow), the migration (IM) model, the isolation-with-initial-migration (IIM) model [113], and the introgression (MSci) model. Simulation may be useful to evaluate the power to distinguish such models using genomic datasets.

Conclusion

The multispecies coalescent model provides a powerful framework for analysis of genomic sequences sampled from multiple species to extract the rich information about the evolutionary history of the species. Incorporating species phylogeny in population genetic models of population subdivision opens up opportunities for addressing many exciting questions in evolutionary biology, such as detecting gene flow during and after species formation and delineating species boundaries, as well as inferring demographic changes and estimating population sizes for extinct ancestral species. As discussed in [92], the basic MSC model accommodating species phylogeny and coalescent is in effect a null model, which can be extended to include other important biological processes, leading to models such as:

- \( H_0 \) : MSC (null model)
- \( H_1 \) : MSC + migration (MSC+M or IM model)
- \( H_2 \) : MSC + introgression (MSC+I or MSci model)
- \( H_3 \) : MSC + population structure
- \( H_4 \) : MSC + recombination
- etc.

Currently, large differences exist between full-likelihood methods and heuristic methods. The former have higher statistical efficiency while the latter are orders-of-magnitude faster computationally. There is thus much room for improvement for both classes of methods. For the present, a pragmatic approach to analyzing large datasets may be to use summary methods.
to estimate the species tree and then full likelihood methods to estimate the parameters.

Analysis of the simple three-species case [62] suggests that there is rich historical information both in gene-tree branch lengths (which 2-step methods such as ASTRAL, MP-EST, and SNAQ ignore) and in the stochastic fluctuation of genealogical history across loci (which genome-averaging approaches such as SVDQUARTETS and D-statistic ignore). Heuristic methods that make use of both kinds of information may thus have much improved power. For Bayesian implementations of the MSC model, mixing inefficiency of the MCMC algorithm appears to be a far more serious problem than the increase in computational cost for each MCMC iteration [48]. Developing smart proposal algorithms that respect and accommodate the mutual constraints between the species tree and the gene trees is likely to bring dramatic improvement to the capacity of the full-likelihood methods. To empirical biologists the MSC framework makes it possible to ask exciting evolutionary questions; to method developers it offers rich opportunities for testing cutting-edge algorithms in computational statistics (in particular, trans-model MCMC algorithms).

With the advancements of sequencing technologies and rapid accumulation of genomic sequence data as the driving force, the field will in all likelihood continue to be a research hotspot in the coming years.

Acknowledgments

This study has been supported by Biotechnology and Biological Sciences Research Council grants (BB/P006493/1 and BB/T003502/1) to Z.Y.

References

[1] J.F.C. Kingman. The coalescent. Stochastic Processes Appl., 13:235–248, 1982.

[2] R.R. Hudson. Testing the constant-rate neutral allele model with protein sequence data. Evolution, 37:203–217, 1983.

[3] F. Tajima. Evolutionary relationship of DNA sequences in finite populations. Genetics, 105(2):437–460, 1983.

[4] R.R. Hudson. Genealogies and the coalescent process. In D.J. Futuyma and J.D. Antonovics, editors, Oxford Surveys in Evolutionary Biology, pages 1–44. Oxford University Press, New York, 1990.

[5] J. Hein, M.H. Schierup, and C. Wiuf. Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory. Oxford University Press, Oxford, 2005.

[6] M. Nordborg. Coalescent theory. In D. Balding, M. Bishop, and C. Cannings, editors, Handbook of Statistical Genetics, pages 843–877. Wiley, San Francisco, 2007.

[7] J. Wakeley. Coalescent Theory: An Introduction. Roberts & Company, Greenwood Village, Colorado, 2009.

[8] B. Rannala and Z. Yang. Challenges in species tree estimation using DNA sequences from multiple loci. Genetics, 164(4):1645–1658, 2003.

[9] J.H. Degnan and N.A. Rosenberg. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol., 24:332–340, 2009.

[10] L. Kubatko. The multispecies coalescent. In D. Balding, I. Moltke, and C. Marioni, editors, Handbook of Statistical Genomics, pages 219–245. Wiley, New York, 4th edition, 2019.

[11] B. Rannala, S. Edwards, A.D. Leaché, and Z. Yang. The multispecies coalescent model and species tree inference. In C. Scornavacca, F. Delson, and N. Galtier, editors, Phylogenetics in Genomic Era, book section 3.3, pages 3.3:1–20. No Commercial Publisher, 2020.

[12] B. Rannala and Z. Yang. Species delimitation. In C. Scornavacca, F. Delson, and N. Galtier, editors, Phylogenetics in Genomic Era, book section 5.5, pages 5.5:1–18. No Commercial Publisher, 2020.

[13] S.V. Edwards. Is a new and general theory of molecular systematics emerging? Evolution, 63(1):1–19, 2009.

[14] J.H. Gillespie and C.H. Langley. Are evolutionary rates really variable. J. Mol. Evol., 13:27–34, 1979.

[15] N. Takahata. An attempt to estimate the effective size of the ancestral species common to two extant species from which homologous genes are sequenced. Genet. Res., 48:187–190, 1986.

[16] R. Burgess and Z. Yang. Estimation of hominoid ancestral population sizes under bayesian coalescent models incorporating mutation rate variation and sequencing errors. Mol. Biol. Evol., 25:1979–1994, 2008.

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

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

[19] C.M. Shi and Z. Yang. Coalescent-based analyses of genomic sequence data provide a robust resolution of phylogenetic relationships among major groups of gibbons. Mol. Biol. Evol., 35:159–179, 2018.

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

[21] J.H. Degnan and N.A. Rosenberg. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol., 24:332–340, 2009.

[22] Z. Yang. Molecular Evolution A Statistical Approach. Oxford University Press, Oxford, England, 2014.

[23] L. Liu, Z. Xi, S. Wu, C. Davis, and S.V. Edwards. Estimating phylogenetic trees from genome-scale data. Ann. NY. Acad. Sci., page doi: 10.1111/nyas.12747, 2015.

[24] S. Sankaranraman. Methods for detecting introgressed archaic sequences. Curr. Opin. Genet. Dev., 62:85–90, 2020. doi: 10.1016/j.gde.2020.05.026.

[25] K.L. Korunes and A. Goldberg. Human genetic admixture. PloS Genet., 17(3):e1009374, 2021. doi: 10.1371/journal.pgen.1009374.

[26] J. Wright. Evolution in Mendelian populations. Genetics, 16:97–159, 1931.

[27] N. Yu, F.C. Chen, S. Ota, L.B. Jorde, P. Pamilo, L. Patthy, M. Ramsay, T. Jenkins, S.K. Shyne, and W.H. Li. Larger genetic differences within africans than between africans and eurasians. Genetics, 161:269–274, 2002.

[28] Y. Thawornwattana, J. Mallet, and Z. Yang. Complex introgression history of the erato-sara clade of Heliconius butterflies. bioRxiv, 2021. doi: 10.1101/2021.02.10.430600.

[29] A.W.F. Edwards. Estimation of the branch points of a branching diffusion process. J. R. Stat. Soc. B, 32(2):155–164, 1970.
and C.D. Jiggins. Genome-wide evidence for speciation with gene flow in Heliconius butterflies. *Genome Res.*, 23 (11):1817–1828, 2013.

[74] R. Nielsen, M.M. Akery, M. Jakobsson, J.K. Pritchard, S. Tishkoff, and E. Willerslev. Tracing the peopling of the world through genomics. *Nature*, 541:302, 2017.

[75] A.D. Leaché, R.B. Harris, B. Rannala, and Z. Yang. The influence of gene flow on Bayesian species tree estimation: a simulation study. *Syst. Biol.*, 63(1):17–30, 2014.

[76] C. Solis-Lemus and J.F. Yang. Inconsistency of species tree methods under gene flow. *Syst. Biol.*, 65(5):843–851, 2016. doi: 10.1093/sysbio/syw030.

[77] C. Long and L.S. Kubatko. The effect of gene flow on coalescent-based species-tree inference. *Syst. Biol.*, 67:770–785, 2018.

[78] X. Jiao, T. Flouri, B. Rannala, and Z. Yang. The impact of cross-species gene flow on species tree estimation. *Syst. Biol.*, 69:830–847, 2020. doi: 10.1093/sysbio/sya001.

[79] R. Nielsen and J. Wakeley. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics*, 158:885–896, 2001.

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

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

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

[83] H.B. Nath and R.C. Griffiths. The coalescent in two colonies with symmetric migration. *J. Math. Biol.*, 31:841–852, 1993.

[84] H.M. Wilkinson-Herbots. Genealogy and subpopulation differentiation under various models of population structure. *J. Math. Biol.*, 37:535–585, 1998.

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

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

[87] Y. Wang and J. Hey. Estimating divergence parameters with small samples from a large number of loci. *Genetica*, 184:363–379, 2010.

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

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

[90] I. Gronau, M.J. Hubisz, B. Gatto, C.G. Danko, and A. Siepel. Bayesian inference of ancient human demography from individual genome sequences. *Nature Genet.*, 43:1031–1034, 2011.

[91] S. Zhu and J. Degnan. Displayed trees do not determine distinguishability under the network multispecies coalescent. *Syst. Biol.*, 66:283–298, 2017.

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

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

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

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

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

[97] G.R. Jones. Divergence estimation in the presence of incomplete lineage sorting and migration. *Syst. Biol.*, 68 (1):19–31, 2019.

[98] L.S. Kubatko. Identifying hybridization events in the presence of coalescence via model selection. *Syst. Biol.*, 58(5):478–488, 2009. doi: 10.1093/sysbio/syp055.

[99] C. Meng and L.S. Kubatko. Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. *Theor. Popul. Biol.*, 75:35–45, 2009.

[100] J. Zhu, Y. Yu, and L. Nakhleh. In the light of deep coalescence: revisiting trees within networks. *BMC Bioinform.*, 17:415, 2016.

[101] E.Y. Durand, N. Patterson, D. Reich, and M. Slatkin. Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.*, 28:2239–2252, 2011.

[102] N. Patterson, P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, and D. Reich. Ancient admixture in human history. *Genetics*, 192:1065–1093, 2012.

[103] J.B. Pease and M.W. Hahn. Detection and polarization of introgression in a living primate phylogeny. *Syst. Biol.*, 64:651–662, 2015.

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

[105] L.S. Kubatko and J. Chifman. An invariants-based method for efficient identification of hybrid species from largescale genomic data. *BMC Evol. Biol.*, 19(1):112, 2019. doi: 10.1186/s12862-019-1439-7.

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

[107] S. Aeschbacher, J.P. Selby, J.H. Willis, and G. Coop. Population-genomic inference of the strength and timing of selection against gene flow. *Proc. Natl. Acad. Sci. U.S.A.*, 114:7061–7066, 2017.

[108] A.D. Kern and J. Hey. Exact calculation of the joint allele frequency spectrum for isolation with migration models. *Genetics*, 207:241–253, 2017.

[109] F. Pardi and C. Scornavacca. Reconstructible phylogenetic networks: Do not distinguish the indistinguishable. *PLoS Comput. Biol.*, 11(4):e1004135, 2015. doi: 10.1371/journal.pcbi.1004135.

[110] M. Stephens. Dealing with label switching in mixture models. *J. R. Statist. Soc. B.*, 62:795–809, 2000.

[111] J. Huang, T. Flouri, and Z. Yang. A simulation study to examine the information content in phylogenomic datasets under the multispecies coalescent model. *Mol. Biol. Evol.*, 37(11):3211–3224, 2020. doi: 10.1093/molbev/msaa166.

[112] M. Slatkin and J.L. Pollack. Subdivision in an ancestral world through genomics. *Syst. Biol.*, 69:830–847, 2020. doi: 10.1093/sysbio/sya001.

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