



HAL
open science

The Multi-species Coalescent Model and Species Tree Inference

Bruce Rannala, Scott V. S. V. Edwards, Adam Leaché, Ziheng Yang

► **To cite this version:**

Bruce Rannala, Scott V. S. V. Edwards, Adam Leaché, Ziheng Yang. The Multi-species Coalescent Model and Species Tree Inference. Scornavacca, Celine; Delsuc, Frédéric; Galtier, Nicolas. Phylogenetics in the Genomic Era, No commercial publisher | Authors open access book, pp.3.3:1–3.3:21, 2020. hal-02535622

HAL Id: hal-02535622

<https://hal.science/hal-02535622>

Submitted on 10 Apr 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.


L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.




Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Chapter 3.3 The Multi-species Coalescent Model and Species Tree Inference


Bruce Rannala

Department of Evolution and Ecology, University of California Davis
One Shields Avenue, Davis CA USA
brannala@ucdavis.edu
 <https://orcid.org/0000-0002-8355-9955>


Scott V. Edwards

Department of Organismic and Evolutionary Biology and Museum of Comparative Zoology, Harvard University
Cambridge, MA 02138, USA
sedwards@fas.harvard.edu
 <https://orcid.org/0000-0003-2535-6217>

Adam Leaché

Department of Biology & Burke Museum of Natural History and Culture, University of Washington
Seattle, WA 98195-1800, USA
leache@uw.edu
 <https://orcid.org/0000-0001-8929-6300>

Ziheng Yang¹

Department of Genetics, Evolution and Environment, University College London
London WC1E 6BT, United Kingdom
z.yang@ucl.ac.uk
 <https://orcid.org/0000-0003-3351-7981>

Abstract

The multispecies coalescent (MSC) is an extension of the single-population coalescent model of population genetics to the case of multiple species. The MSC naturally accommodates speciation events (with subsequent genetic isolation between species) and the coalescent process within each species. It provides a framework for analysis of multilocus genomic sequence data from multiple species in a number of inference problems including species tree estimation, accounting for ancestral polymorphism and deep coalescence. Within this framework, the genealogical fluctuations across genes or genomic regions (and the gene tree/species tree conflicts that may result) are not seen as a problem but rather as a source of information for estimating important parameters such as species divergence times, ancestral population sizes, and the timings, directions, and intensities of cross-species introgression or hybridisation events. This chapter outlines the basic theory of the MSC and its important applications in analysis of genomic sequence data, describing the most widely-used full-likelihood and heuristic methods of species tree estimation. We discuss several active areas of research in which we predict future developments will occur, including inference of introgression events on a species phylogeny.

How to cite: Bruce Rannala, Scott V. Edwards, Adam Leaché, and Ziheng Yang (2020). The Multi-species Coalescent Model and Species Tree Inference. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter No. 3.3, pp. 3.3:1–3.3:21. No commercial publisher | Authors open access book. The book is freely available at <https://hal.inria.fr/PGE>.


¹ Z.Y. is supported by a Biotechnological and Biological Sciences Research Council grant (BB/P006493/1).




© Bruce Rannala, Scott V. Edwards, Adam Leaché and Ziheng Yang.
Licensed under Creative Commons License CC-BY-NC-ND 4.0.

Phylogenetics in the genomic era.

Editors: Celine Scornavacca, Frédéric Delsuc and Nicolas Galtier; chapter No. 3.3; pp. 3.3:1–3.3:21

 A book completely handled by researchers.

 No publisher has been paid.

3.3:2 Species Tree Inference

1 Introduction

The need by scientists and the public for robust phylogenies and a complete Tree of Life grows every day. Phylogenies are a fundamental building block of evolutionary biology. They provide a detailed genealogical “map” that has applications in a variety of fields such as biogeography, molecular evolution, pathogen evolution, and comparative genomics. In recent years, our ability to infer phylogenies has grown dramatically, not only because of technical advances in high-throughput DNA sequencing, but also through theoretical advances (Boussau et al., 2013; Bravo et al., 2019; Liu et al., 2009, 2015; Rannala and Yang, 2017, 2008). Of the many types of theoretical advances that have been made in the last 20 years, this chapter will focus on the application of the multispecies coalescent model (MSC) to phylogenetic inference. We regard this as one of the most important new directions for phylogenetics since DNA sequencing became more widespread among systematists in the late 1980s.

When the polymerase chain reaction (PCR) became widely available in the late 1980s, population geneticists and evolutionary biologists immediately began estimating phylogenetic trees with DNA sequences (Kocher et al., 1989). Molecular systematics of course goes back even further, but molecular cloning of individual genes was laborious. Within-species studies of gene trees first became possible with the advent of restriction enzymes and their application to DNA diversity in the late 1970s (Brown et al., 1982; Wilson et al., 1985). The shift from allozymes and protein polymorphisms to DNA differences had profound impacts on the evolutionary biology community, not only technically but also because of the insights that were provided to empiricists and theoreticians (Avice, 1994). Whereas allozyme electrophoresis could allow one to tell different alleles in a given species apart, DNA differences allowed one to measure the evolutionary or genetic distance between alleles (Avice et al., 1979). This improved precision led population genetics and phylogenetics into a wholly new territory.

Early investigations of gene trees in closely related populations and species quickly revealed that the gene tree of alleles from different populations did not always correspond to the species tree (Avice et al., 1987). One of the most common reason for this discordance is now well understood – the failure of alleles to coalesce as one moves backward in time toward successive speciation events, or, thinking forward in time, the failure of genetic drift to “sort” alleles into their descendant populations fast enough before the next speciation event. Adopting a forward-time definition, this phenomenon was dubbed “incomplete lineage sorting” by Avice and, taking a backward time perspective, “deep coalescence” by Maddison (1997). Gene tree-species tree discordance can be caused by other biological processes such as gene duplication, introgression or horizontal gene transfer (Nichols, 2001; Edwards, 2009; Szollosi et al., 2015), but these are not inherent to population divergences in the same fundamental way that the coalescent process is because the coalescent operates in all finite populations whereas the other processes are not always present.

Avice also formalized the distinction between a gene tree and a species tree. The concept of a species tree is synonymous with phylogeny and had, of course, been fundamental to evolutionary biology since Darwin’s *On the Origin of Species* was published in 1859 (Darwin, 1859). However, it was empirical studies of gene trees in natural populations that drove home the distinction between a gene tree and the species tree that generated it (Hare, 2001). In the early days of DNA sequencing, and frequently even today, researchers refer to the gene tree as the species tree, or use methods, such as concatenation, that assume that the two are the same (see Chapter 3.4 [Bryant and Hahn 2020]). Although the distinction between gene trees and the species tree has been appreciated for decades, computational methods for estimating the species tree accommodating gene tree discordance have only been available since about 2006.

The gene tree-species tree mismatch probability in the case of three species was derived by

Hudson (1983). The mismatch probability was used to estimate the population sizes for the human-chimpanzee common ancestor (Takahata et al., 1995). The probabilities of gene tree topologies (typically assuming one sequence from each species) given a species tree was further studied by Pamilo and Nei (1988) and more recently by Rosenberg (2002), Degnan and Salter (2005), Degnan and Rosenberg (2006) and Wu (2012, 2016), who developed algorithms for automatic calculation of such probabilities. The most well-known result from this line of research is the existence of the so-called anomaly zone, the zone of species tree and parameter values for which the most probable gene tree has a different topology from the species tree. The full probability distribution of gene trees with branch lengths (coalescent times) for an arbitrary species tree – the multispecies coalescent model – was first fully described by Rannala and Yang (2003). This forms the basis for exact or full-likelihood methods of species tree inference – those that use the observable DNA sequence data directly rather than data summaries such as the collection of inferred gene tree topologies.

2 The Multispecies Coalescent

The multispecies coalescent (MSC) describes the probability distribution of the gene tree, G , underlying a sample of DNA sequences from two or more species (or genetically isolated populations). The MSC is an extension of the coalescent theory for a single randomly mating population. Thus, we begin with a description of the single-population coalescent process.

2.1 The single-population coalescent process

The coalescent theory of population genetics (Kingman, 1982; Hudson, 1983; Tajima, 1983) provides the probability distribution of the gene genealogical history (or gene tree) for a random sample of n sequences at a neutral non-recombining locus. The process is usually formulated in terms of a single parameter

$$\theta = 4N\mu, \quad (1)$$

where N is the effective population size (the population size for an idealized Fisher-Wright model) and μ is the mutation rate per-site per generation. While in classical population genetics, θ is defined using a per-locus mutation rate, the per-site rate used here is far more convenient in analysis of genomic sequence data. Thus θ is the heterozygosity or the average number of mutations per site between two randomly sampled sequences from the population.

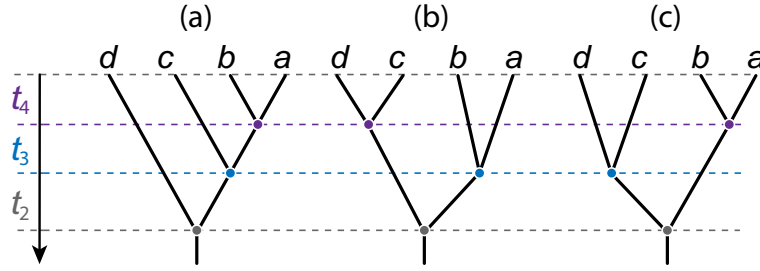
The coalescent process tracks the genealogical history of the sequences going backwards in time from the present into the past. The n sequences in the sample go through $n - 1$ coalescent events, each time reducing the number of sequences by one, until the most recent common ancestor of the whole sample. With j sequences in the sample, each pair coalesce at the rate $2/\theta$, so that the total rate for $\binom{j}{2}$ pairs is $\binom{j}{2} \frac{2}{\theta}$. The coalescent waiting time until the next coalescence event (which reduces the number of lineages from j to $j - 1$) is t_j , is an exponential variable with density

$$f(t_j|\theta) = \binom{j}{2} \frac{2}{\theta} e^{-\binom{j}{2} \frac{2}{\theta} t_j}, \quad (2)$$

This has expectation $\theta/[j(j-1)]$. The coalescence times $\mathbf{t} = \{t_n, t_{n-1}, \dots, t_2\}$ are independent random variables with joint probability density

$$f(\mathbf{t}|\theta) = \prod_{j=2}^n f(t_j|\theta) = \prod_{j=2}^n \left[\binom{j}{2} \frac{2}{\theta} e^{-\sum_{j=2}^n \binom{j}{2} \frac{2}{\theta} t_j} \right], \quad (3)$$

3.3:4 Species Tree Inference



■ **Figure 1** Examples of labelled histories (gene trees with internal nodes rank-ordered according to age) for 4 sequences (a, b, c, d) generated under a coalescent process. There is only one labelled history for a gene tree with the topology $((a, b), c), d$ shown in (a) while (b) and (c) are the two alternative labelled histories of the topology $((a, b), (c, d))$ obtained by interchanging the rank order of ages associated with internal nodes.

where time is scaled in units of expected mutations per site. Note that with DNA sequence data, coalescence time (or population size) and mutation rate are not separately identifiable, so that the estimable parameter is $\theta = 4N\mu$, not N and μ separately.

The coalescent process also imposes a probability distribution on gene tree topologies. A “labelled history” (Edwards, 1970) is an ultrametric rooted binary tree with tips labelled and internal nodes rank-ordered according to time or age (see Figure 1). The rank order is completely determined for a fully asymmetrical tree (Figure 1a) but for more symmetrical trees there may be two or more possible rank orderings of the internal nodes (Figure 1b&c). Under the coalescent process all distinct labelled histories have equal probabilities. Degnan and Salter (2005) used a different terminology referring to the alternative orderings of a labelled history as different “instantiations” of the same history. Equation 3 gives the probability density of the times averaged across possible labelled histories. The number of possible labelled histories for n sequences is

$$H_n = \binom{n}{2} \binom{n-1}{2} \dots \binom{2}{2} = \frac{n!(n-1)!}{2^{n-1}}, \quad (4)$$

Because all labelled histories have equal probability under the process, the probability of the gene tree, $G = \{t, T\}$, defined by a set of coalescence times, t , and a labelled history, T , is

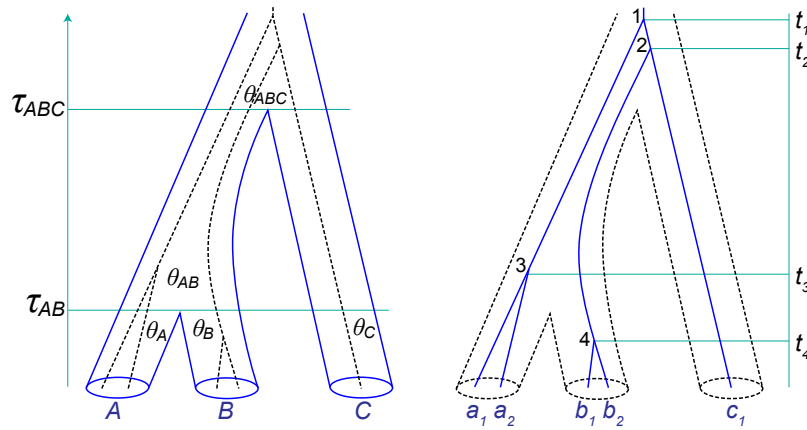
$$f(G|\theta) = f(t|\theta) \times \frac{2^{n-1}}{n!(n-1)!} = \left(\frac{2}{\theta}\right)^{n-1} e^{-\frac{2}{\theta} \sum_{j=2}^n \binom{j}{2} t_j}. \quad (5)$$

This probability density applies to a sample from a single panmictic population conforming to a neutral Fisher-Wright model as well as from other neutral models with exchangeable offspring distributions (Kingman, 1982). It can be used within a Bayesian framework for inferring θ using sampled sequences.

The above introduction to the coalescent has focussed on the distribution of the gene trees (topologies and coalescent times) under the model. Many other aspects of the coalescent can be studied as well. Furthermore, the basic neutral coalescent model has been extended to allow for multiple biological processes, including demographic changes over time, recombination (Hudson, 1983; Hudson and Kaplan, 1985; Griffiths and Marjoram, 1996), and selection (Krone and Neuhauser, 1997). The reader may consult Hudson (1990), Nordborg (2007) and Wakeley (2009) for reviews.

2.2 The MSC process

The coalescent process model has been extended to the case of multiple species, which are related through a phylogenetic tree, with one or more sequences sampled from each species. A species



■ **Figure 2** A species tree for three species (A, B, C) with a gene tree for five sequences embedded inside, to illustrate the parameters in the MSC model, $\theta = (\tau_{AB}, \tau_{ABC}, \theta_A, \theta_B, \theta_C, \theta_{AB}, \theta_{ABC})$ and the gene tree density under the model.

tree of s species have $2s - 1$ nodes, of which s represent contemporary species and $s - 1$ represent ancestral species. The MSC model on a species tree of s species thus has $s - 1$ divergence times (τ s) and $2s - 1$ population size parameters (θ s). Both divergence times and population sizes are scaled by mutation rate, so that both τ s and θ s are measured by expected number of mutations per site. The parameters for a species tree for $s = 3$ species are shown in Figure 2. Each population operates as an independent coalescent process during its existence, with population i having a scaled coalescence rate of $\theta_i = 4N_i\mu$. All populations (except the one at the root of the species tree) exist for a finite period of time determined by the species divergence times.

2.2.1 Probability density of gene trees within a species tree

The probability density of an arbitrary gene tree at a locus given the MSC model (species phylogeny with the associated parameters) has been determined by Rannala and Yang (2003). Given the species tree, the gene trees are assumed to be independent among loci. At each locus, the coalescent process is independent among populations on the species tree. Thus we focus on the part of the gene tree residing in one population, say, species X , with parental species P . Let τ_X and τ_P be the age of the two nodes in the species tree. X may be a contemporary species (in which case $\tau_X = 0$) or an ancestral species. Going backwards in time, let m be the number of sequences that enter population X at time τ_X and let $n \geq 1$ be the number of sequences that remain at the end of the population at time τ_P . For example, in figure 2 the species AB (with age τ_{AB}) has parental species ABC (with age τ_{ABC}). In the gene tree in figure 2, $m = 3$ lineages enter species AB and $n = 2$ lineages leave it. The probability density for the $m - n$ coalescent waiting times between coalescence events is

$$\prod_{j=n+1}^m \left[\frac{2}{\theta} \exp \left\{ -\frac{j(j-1)}{2} \frac{2}{\theta} t_j \right\} \right] = \left(\frac{2}{\theta} \right)^{m-n} \exp \left\{ -\sum_{j=n+1}^m \frac{j(j-1)}{\theta} t_j \right\}. \quad (6)$$

An important difference of the MSC from the single-population coalescent is that it is possible for $n \geq 1$ lineages to remain at the end of the population at time τ_P . We have to account for the probability that the n sequences do not coalesce in the remaining period of population existence which has

3.3:6 Species Tree Inference

duration $\left(\tau_P - \tau_X - \sum_{j=n+1}^m t_j\right)$. This probability of no events is

$$\exp\left\{-\frac{n(n-1)}{\theta}\left(\tau_P - \tau_X - \sum_{j=n+1}^m t_j\right)\right\}. \quad (7)$$

If population X is the root of the species tree, then n must be 1 and this term disappears. Combining the two components gives the probability density for the part of the gene tree in population X

$$\left(\frac{2}{\theta}\right)^{m-n} \exp\left\{-\sum_{j=n+1}^m \left(\frac{j(j-1)}{\theta} t_j\right) - \frac{n(n-1)}{\theta}\left(\tau_P - \tau_X - \sum_{j=n+1}^m t_j\right)\right\}. \quad (8)$$

The probability density for the whole gene tree at the locus is the product of the probabilities across all populations on the species phylogeny.

For example, given the MSC model for three species in figure 2, the gene tree for the five sampled sequences has the density

$$\begin{aligned} f(G|\boldsymbol{\theta}) &= \left[e^{-\frac{2}{\theta_A} \tau_{AB}} \right] \times \left[\frac{2}{\theta_B} e^{-\frac{2}{\theta_B} t_4} \right] \times \left[\frac{2}{\theta_{AB}} e^{-\frac{3 \times 2}{\theta_{AB}} (t_3 - \tau_{AB})} \cdot e^{-\frac{2}{\theta_{AB}} (\tau_{ABC} - t_3)} \right] \\ &\times \left[\frac{2}{\theta_{AB}} \cdot \frac{2}{\theta_{AB}} e^{-\frac{3 \times 2}{\theta_{ABC}} (t_2 - \tau_{ABC})} \cdot e^{-\frac{2}{\theta_{ABC}} (t_1 - t_2)} \right]. \end{aligned} \quad (9)$$

The terms in the four pairs of brackets correspond to four species A, B, AB , and ABC , respectively. There is no possibility for coalescent in species C when only one sequence is sampled from the species.

With multiple loci in the data, the probability density for all gene trees is a product over the loci. The formulation allows different sampling configurations at different loci; for example, the number of sequences for each species may vary among loci.

The coalescent is a fundamental process that is operating regardless of whether the species are recently divergent or distantly related, and whether or not the species arose through rapid speciation events so that incomplete lineage sorting is commonplace (Edwards et al., 2016; Degnan, 2018). In cases where species divergences are far apart relative to population sizes, the species tree will have long internal branches and there will be little ILS or gene tree-species tree discordance, but this is exactly as predicted by the MSC model. As discussed by Degnan (2018), the MSC should be considered a null model, and other biological processes, such as recombination, population structure, gene flow, etc. may be incorporated in the model in addition, leading to models such as MSC with recombination, MSC with demographic changes, MSC with migration (which is the IM model Hey, 2010; Hey et al., 2018), MSC with introgression (Yu et al., 2014; Zhang et al., 2018; Wen and Nakhleh, 2018), and so on. Many of these models are not yet implemented because of their complexity, but conceptually they should be possible.

2.2.2 Probabilities of gene tree topologies

Another aspect of the MSC that has been of interest is the marginal probabilities of gene tree topologies conditioned on a particular species tree and branch lengths and, in particular, the probability that the gene tree topology matches that of the species tree (Pamilo and Nei, 1988; Rosenberg, 2002; Degnan and Rosenberg, 2009). As noted above, the labelled histories have equal probabilities under the single population coalescent process. However, this is not the case for the MSC.

The simplest case concerns three species A, B , and C , with three sequences (a, b, c) , with one sequence sampled from each species. The probabilities of the three rooted gene tree topologies $G_1 = ((a, b), c)$, $G_2 = ((c, a), b)$ and $G_3 = ((b, c), a)$, given the species tree $S = ((A, B), C)$, were

derived by Hudson (1983). Let the species tree be $((A, B), C)$, the divergence times be τ_{AB} and τ_{ABC} , and the ancestral population size parameters be θ_{AB} and θ_{ABC} . The probability that sequences a and b coalesce in the ancestral population AB (in which case the gene tree must be G_1) is $1 - e^{-x} = 1 - e^{-(\tau_{ABC} - \tau_{AB})/(\frac{\theta_{AB}}{2})}$, and the probability that sequences a and b do not coalesce in population AB (in which case all three sequences enter the ancestor ABC and the three gene trees occur with equal probability) is e^{-x} . Here $x = 2(\tau_{ABC} - \tau_{AB})/\theta_{AB}$ is known as the internal branch length in coalescent units: one coalescent unit in population AB is $2N_{AB}$ generations or $\theta_{AB}/2$ mutations per site. Thus the probabilities for the three gene tree topologies are

$$P(G_1|S) = (1 - e^{-x}) + \frac{1}{3}e^{-x}, \quad (10)$$

$$P(G_2|S) = P(G_3|S) = \frac{1}{3}e^{-x}. \quad (11)$$

The probabilities that the gene tree matches (or mismatches) the species tree are then

$$P_{\text{match}} = P(G_1|S) = 1 - \frac{2}{3}e^{-x}, \quad (12)$$

$$P_{\text{mismatch}} = P(G_2|S) + P(G_3|S) = \frac{2}{3}e^{-x}. \quad (13)$$

In the limit as $x \rightarrow \infty$, the probabilities $P_{\text{match}} \rightarrow 1$ and $P_{\text{mismatch}} \rightarrow 0$ while as $x \rightarrow 0$, $P_{\text{match}} \rightarrow 1/3$ and $P_{\text{mismatch}} \rightarrow 2/3$. Thus, the most difficult species trees to infer using gene trees are those with short internal branches.

Degnan and Salter (2005) developed algorithms for calculating the gene tree probabilities given an arbitrary number of species and arbitrary species tree, with one sequence sampled from each species. The algorithms are computationally expensive owing to the explosive growth in the number of tree topologies with increasing numbers of species and sequences. The gene tree probabilities can be used to estimate the species tree by maximum likelihood, treating the gene tree topologies as data (Wu, 2012, 2016). These are the so-called two-step methods of species tree inference. In practice, almost all two-step methods are based on triplets or quartets, using rooted trees for three species or unrooted trees for four species, and then assembling the results to produce a species tree estimate for all species.

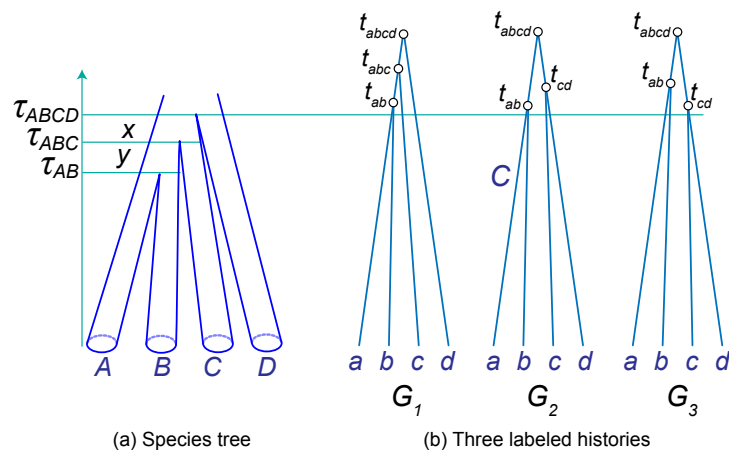
2.2.3 The anomaly zone

The most well-known result from the calculations of gene tree probabilities is the existence of so-called anomaly zone, defined as the zone of species tree and parameter values under which the most probable gene tree has a topology different from the species tree topology (Degnan and Rosenberg, 2006). There is no anomaly zone for three species, but anomaly zone may exist for rooted species trees of four or more species. In the anomaly zone, a ‘‘majority-vote’’ method that uses the most frequent gene tree as the species tree estimate will be inconsistent. Such gene trees are called anomalous gene trees. The anomaly zone exists because the coalescent process generates a uniform distribution on labelled histories but not on rooted tree topologies. As illustrated in Figure 1 asymmetrical topologies have only one labelled history whereas symmetrical topologies can have two or more. This may result in an even greater probability for symmetrical gene-tree topologies even if the species tree has an asymmetrical topology.

Consider the case of four species, related through the asymmetrical phylogeny $((A, B), C), D)$, and the gene trees for four sequences (a, b, c, d) , with one sequence from each species (figure 3). Let the internal branch lengths in coalescent units in the species tree be $x = 2(\tau_{ABCD} - \tau_{ABC})/\theta_{ABC}$ and $y = 2(\tau_{ABC} - \tau_{AB})/\theta_{AB}$. Consider the limit as $x \rightarrow 0$ and $y \rightarrow 0$. In this case, the probability of a coalescence in either ancestral species AB or ABC approaches zero and all coalescence events will occur in the root species $ABCD$. The coalescent process in $ABCD$ is equivalent to a single

3.3:8 Species Tree Inference

population coalescent with four sequences so applying Equation 4 we have 18 possible labelled histories, each with equal probability $1/18$. Of these, 12 are fully asymmetrical (such as labelled histories G_2 and G_3 in figure 3) and 6 are symmetrical (such as labelled history G_1 in figure 3). Each asymmetrical labelled history corresponds to one unique tree topology because there is only one possible way to order the internal nodes. The 6 symmetrical labelled histories form 3 pairs, with each pair corresponding to one tree topology with two possible node orderings (e.g., G_2 and G_3 in figure 3). Thus, each symmetrical rooted tree topology receives probability $1/18 + 1/18 = 2/18$ whereas each asymmetrical rooted tree topology receives probability $1/18$. When the branch lengths x and y are nonzero but very small, the symmetrical and mismatching gene tree (corresponding to G_2 and G_3 in figure 3) may still be more frequent than the asymmetrical and matching gene tree G_1 , even if not twice as frequent. Consequently, the most common gene tree will have a mismatching symmetrical topology that is different from the species tree, and this combination of species tree topology and branch lengths is in the anomaly zone.



■ **Figure 3** A species tree for four species (A, B, C, D) with very short internal branches and three labelled histories for four sequences (a, b, c, d) to illustrate the existence of the anomaly zone.

The anomaly zone has been shown to affect empirical dataset from lizards (Linkem et al., 2016), flightless birds (Cloutier et al., 2019), gibbons (Shi and Yang, 2018), and African mosquitoes (Thawornwattana et al., 2018). The anomaly zone can be identified by estimating parameters in the MSC model using Bayesian inference programs such as *bpp*, and then simulating gene trees using those parameters to estimate gene tree probabilities –to confirm that the most probable gene tree does not match the species tree (Shi and Yang, 2018).

While it is well-known that species phylogenies with very short internal branches are hard to recover, the importance of the anomaly zone may have been exaggerated in the literature. Note that the anomaly zone is the zone of inconsistency for the simple “majority vote” method only. Other methods may, or may not, be inconsistent in the anomaly zone. In particular, methods based on the likelihood function for the sequence data, including maximum likelihood and Bayesian methods (see below), are consistent both inside and outside the anomaly zone; indeed they are consistent over the entire space of species trees (Xu and Yang, 2016).

3 Species Tree Inference Methods

The MSC provides a framework for developing parametric multi-locus statistical methods for species tree inference. Such methods allow gene trees to differ from species trees due to ILS and provide

estimates of ancestral demographic parameters. Because the MSC operates in all finite populations it is the canonical model for species tree inference. We begin by describing the maximum likelihood and Bayesian methods that have been developed for species tree inference. These are often referred to as full-likelihood methods because they use an exact likelihood function. Full-likelihood methods are known to possess optimal statistical properties such as consistency and efficiency. We then consider two of the most widely used approximate methods: MP-EST (Liu et al., 2010) and ASTRAL (Mirarab et al., 2014; Mirarab and Warnow, 2015). These programs are examples of “super-tree” methods which infer larger trees by combining estimates of smaller trees. One of the methods (MP-EST) approximates the MSC using pseudo-likelihood while the other (ASTRAL) uses a simple heuristic that may provide estimates that are statistically consistent when gene trees arise under the MSC. The statistical properties of heuristic methods often can only be studied by computer simulation. See Yang and Rannala (2014), Edwards (2016) and Xu and Yang (2016) for an overview of other approximate methods.

Next, we consider another class of approximate methods, so-called concatenation methods. These methods combine all the loci into a single matrix of sequences and are examples of “super-gene” or “super-matrix” approaches to species tree inference that implicitly assume no ILS. Gatesy and Springer (2013) and Edwards et al. (2016) review the extensive discussions concerning relative strengths and weaknesses of concatenation versus two-step and coalescent methods for species tree inference. We briefly summarize several of the problems that can arise with approximate inference methods that use concatenation. The reader may consult Chapter 3.4 (Bryant and Hahn 2020) for a different perspective. Finally, we discuss some criticisms of two-step approximate inference methods and full-likelihood methods based on the MSC.

3.1 Maximum likelihood method

Here we outline full-likelihood methods for estimating the species tree using multilocus sequence data under the MSC. Let the sequence alignment at locus i be X_i , with $i = 1, 2, \dots, L$. Let $\mathbf{X} = \{X_i\}$. Let θ_k be the MSC parameters (θ s and τ s) in species tree S_k . Let G_i be the gene tree at locus i . The main difference from traditional phylogenetic methods is that the gene trees are unobserved random variables, with distributions specified by the MSC model (Rannala and Yang, 2003). For example, the maximum likelihood method of species tree estimation maximizes the following likelihood function

$$f(\mathbf{X}|S_k, \theta_k) = \prod_{i=1}^L \left[\int f(G_i|S_k, \theta_k) f(X_i|G_i) dG_i \right], \quad (14)$$

where $f(G_i|S_k, \theta_k)$ is the MSC density for gene tree G_i at locus i discussed above (Rannala and Yang, 2003), and $f(X_i|G_i)$ is the probability of the sequence alignment at locus i or the phylogenetic likelihood (Felsenstein, 1981). The integral over gene tree G_i represents a summation over all possible gene tree topologies (labelled histories) for the locus and an integral over the coalescent times within each gene tree topology. In this formulation, the gene trees G_i are unobserved random variables (called latent variables), and the likelihood function for the species tree and MSC parameters has to average over all possible gene trees at each locus.

The S_k and θ_k that maximize the log-likelihood, $\ell = \log f(\mathbf{X}|S_k, \theta_k)$, will be the ML species tree and MLEs of parameters in that species tree. Note that both the MSC density $f(G_i|S_k, \theta_k)$ and the phylogenetic likelihood $f(X_i|G_i)$ are straightforward to calculate. The difficulty with the ML method lies in the averaging over the possible gene trees at each locus, because the number of possible gene trees is huge and the integral over coalescent times for each gene tree at a locus with n sequences is $(n - 1)$ -dimensional. The only ML implementation that has been achieved is the 3s program (Yang, 2002; Dalquen et al., 2017), which enumerates the gene tree topologies and uses numerical integration

3.3:10 Species Tree Inference

(Gaussian quadrature) to calculate the integrals. This is limited to three species and three sequences per locus, but can accommodate tens of thousands of loci.

3.2 Bayesian inference

In a Bayesian approach, we specify a prior probability distribution for all possible species trees, and for each species tree (which is an MSC model), we specify a prior for the parameters of the MSC model (θ s and τ s). Let $f(S_k)$ be the prior probability for species tree S_k . This can be a uniform distribution on the rooted species trees or on the labelled histories –ranked trees (Yang and Rannala, 2014). It is common to assign gamma or inverse priors on the MSC parameters given the species tree, $f(\theta_k|S_k)$. Inverse gamma priors are conjugate on θ , allowing θ s to be integrated out analytically (Hey and Nielsen, 2007). This may improve the Markov chain Monte Carlo (MCMC) mixing slightly during the tree search thanks to the reduced parameter space. Usually a gamma or inverse-gamma prior is assigned on the age of the root (τ_0), while the other node ages may be specified by a Dirichlet distribution (Yang and Rannala, 2010) or by a birth-death process model. Bayesian computation is achieved through MCMC algorithms, which generate a sample from the joint conditional distribution (joint posterior) of the species tree and the gene trees

$$f(S_k, \theta_k, \{G_i\}|\mathbf{X}) \propto f(S_k)f(\theta_k|S_k) \prod_{i=1}^L [f(G_i|S_k, \theta_k)f(X_i|G_i)]. \quad (15)$$

Compared with Equation 14, the integral over the gene trees disappears; instead integration occurs numerically through the MCMC. In other words, MCMC is used to traverse the joint space of gene trees as well as the species tree and the MSC parameters. The frequency at which the MCMC visits each species tree is the estimate of the posterior probability for that species tree. The first implementation of the Bayesian method is the Best program (Liu and Pearl, 2007), which used the posterior sample of gene trees from the MrBayes program (Ronquist and Huelsenbeck, 2003) and applied a correction for the gene tree density, because the gene trees from MrBayes are not generated with an MSC prior. Later implementations work directly on the MSC and sequence alignments, rather than processing MrBayes outputs (Liu, 2008), especially in the programs StarBeast (Heled and Drummond, 2010; Ogilvie et al., 2017) and bpp (Yang and Rannala, 2014; Rannala and Yang, 2017). Branch-swapping algorithms in phylogenetic tree search such as nearest-neighbor-interchange (NNI), subtree-pruning-and-regrafting (SPR), etc. have been adapted to become MCMC proposals, proposing changes from one species tree to another (Yang and Rannala, 2014; Rannala and Yang, 2017; Flouri et al., 2018), while other moves are used to change the gene trees.

The greatest challenge for MCMC algorithms for species tree inference appears to be the constraint between the species tree and the gene trees. If the species tree is changed when the gene trees at all loci are fixed, the current gene trees may place extremely stringent constraints on the species tree. Consider the simple move that changes the divergence time τ_{AB} between two species or clades A and B . In the coalescent model, the sequence divergence has to be older than species divergence, that is, $t_{ab} > \tau_{AB}$, and this constraint applies to every pair of sequences from A and B at every locus. In other words current gene trees provide a maximum bound for τ_{AB} , which is the minimum t_{ab} across all loci. If there are thousands of loci in the dataset and many sequences from A and B at each locus, the current value of τ_{AB} is often almost identical to this bound and τ_{AB} cannot possibly become even greater, and the algorithm is essentially stuck. An algorithm that appears to work well is the rubber-bound algorithm implemented in bpp (Rannala and Yang, 2003), which identifies the nodes on the gene trees that are affected by the change of τ_{AB} , and then modifies the ages of those gene-tree nodes at the same time that τ_{AB} is changed, in the same way that marked points on a rubber band move when with its two ends fixed, a rubber band is pulled from a point in the middle to one end. This move has been ported into StarBeast (Jones, 2017). Similar coordinated changes between the

species tree and the gene trees appear to improve MCMC mixing when the species tree topology is changed (Yang and Rannala, 2014; Rannala and Yang, 2017; Jones, 2017; Ogilvie et al., 2017). Those smart MCMC moves have made it possible to analyze large datasets with more than 10,000 loci (Rannala and Yang, 2017; Shi and Yang, 2018; Thawornwattana et al., 2018). Further improvements in both computational and mixing efficiency are clearly needed, as real datasets are often too large for Bayesian MCMC programs to handle.

3.3 Approximate species tree inference methods

Next-generation sequencing technologies are currently advancing at an astounding rate making dense genome sequences available for hundreds of individuals and species. This vast array of new data has driven demand for computational methods for inferring species trees that can be practically applied with thousands of loci and hundreds or thousands of sequences. Model-based methods for species tree estimation, such as the Bayesian inference procedure described above, are computationally intensive, and will lag behind the demands of many contemporary sequencing projects. As a result, many heuristic (or approximate) species tree inference methods have been proposed that use various shortcuts and heuristic approximations to improve computational efficiency for large datasets. Some of the heuristic methods discussed (MP-EST and ASTRAL) make an explicit attempt to accommodate ILS, while others (concatenation methods) do not.

One class of approximate species tree inference methods (super-tree methods such as MP-EST and ASTRAL) take the two-step approach of estimating the gene trees from phylogenetic analysis of sequence alignments at individual loci and then treating the gene trees as observed data. A second class of approximate species tree inference methods (sometimes referred to as supermatrix or super-gene methods) concatenate all the loci into a single sequence assuming that the gene tree of the supermatrix matches the species tree. This approach typically applies a standard maximum likelihood or Bayesian phylogenetic approach under the assumption of one gene tree that matches the species tree.

Two-step super-tree methods are much simpler to implement than full likelihood methods and are among the early approaches developed for inferring species trees under the MSC. Some of them use the estimated gene tree topologies with branch lengths (node ages), such as the Maximum Tree method of (Liu et al., 2010) implemented in the STEM program (Kubatko et al., 2009). A serious problem is that the method does not account for the sampling errors in the estimated tree topology and coalescent times. In particular, the coalescent times can have a major impact on species tree inference: for example, if two sequences from two species or clades A and B are identical at any locus, with $t_{ab} = 0$, then the species divergence time must be $\tau_{AB} = 0$. Such extreme estimates of species divergence times will influence the inference of the species tree topology. Other two-step methods use the estimated gene tree topologies as data, ignoring branch lengths or coalescent times. These methods use less information from the data but are also less affected by phylogenetic reconstruction errors. They often work on unrooted gene trees, which are estimated without the assumption of the molecular clock. These topology-only methods have been more successful than methods based on inferred gene trees with branch lengths. However, many of the two-step methods have poor statistical performance and the accuracy of some methods (such as the two-step likelihood method STEM [Kubatko et al. 2009]) even decreases with increasing numbers of loci (Leaché and Rannala, 2010; Mirarab et al., 2014). Concatenation-based super-gene methods rely on straightforward application of existing single-locus phylogenetic inference methods and are thus simple to apply. However, differences between gene trees and species trees (both in terms of branch lengths and topologies) resulting from the MSC and other processes can cause the methods to be statistically inconsistent.

3.3:12 Species Tree Inference

3.3.1 MP-EST: Maximum Pseudo-likelihood Estimation

The Maximum Pseudo-likelihood Estimation (MP-EST) method (Liu et al., 2010) is a two-step method based on species triplets. It extracts, for a tree with s species, all the $s(s-1)(s-2)/6$ rooted triplet “species subtrees” (each comprised of 3 species) to construct the likelihood function. The single internal branch length in each rooted species subtree is a sum of one or more internal branch lengths in the original s -species tree. The data input to the program are rooted gene tree topologies inferred using a maximum likelihood or Bayesian inference program. The number of each triplet gene tree topology given the species subtree follows a trinomial distribution with probabilities determined by the MSC (Equation 10)). The probabilities for gene tree topologies for triplets are multiplied across species subtrees and across loci. This is a pseudo-likelihood function as it ignores the fact that the triplet subtrees are not independent. The pseudo-likelihood is maximized using a heuristic search algorithm to infer the species tree. The MP-EST method may suffer from an information loss because it ignores branch lengths in the gene trees and because it ignores phylogenetic errors in the gene tree reconstruction; this is true of all super-tree methods.

Note that the theory underlying the MP-EST method is given in Equation 10. With the probabilities for the three gene trees given, one can find the most common gene tree topology, which is the species tree estimate, and estimate the internal branch length in the species tree (x). The method is clearly consistent, if the gene trees are known without error: when the number of loci or gene trees approaches infinity, the probability of recovering the correct species tree topology approaches one. Furthermore, in this case of three species and rooted gene trees, Yang (2002) showed that the most probable *estimated* gene tree topology is the one that matches the species tree, although phylogenetic reconstruction errors have the effect of inflating the gene tree-species tree mismatch probability. Thus the MP-EST method will be consistent when *estimated* gene tree topologies are used to estimate the species tree. The internal branch length in the species tree, however, is inconsistently estimated (and underestimated) because phylogenetic errors distort the gene tree probabilities and inflate the gene tree-species tree discordance.

3.3.2 ASTRAL: Accurate Species Tree Algorithm

ASTRAL (Mirarab et al., 2014; Mirarab and Warnow, 2015) is another two-step program that takes as input unrooted gene trees inferred using the maximum likelihood phylogenetic program RAxML (Stamatakis, 2006). The underlying method is based on quartets, with four species and four sequences, one sequence sampled from each species. The species tree is then chosen to be the one that agrees with the greatest number of quartet gene trees. If multiple sequences are available from one species, one sequence from each species is sampled to form the quartet. A motivation for using unrooted quartets for the optimization, rather than finding the species tree compatible with the largest number of complete gene trees (the “majority-vote tree”) is that there are no anomalous gene trees in the case of unrooted species trees for four species (Degnan, 2013).

The ASTRAL method essentially uses ML estimates of the gene tree topologies as summary statistics for inference. It does not use branch length information from the gene trees. Use of the gene tree topologies alone allows the identification of the species tree topology, as well as the internal branch lengths in coalescent units, but other parameters in the MSC model are not identifiable. Note that while the method is claimed to be consistent, the proof of consistency relies on the assumption that gene trees are known without error, and the impact of phylogenetic reconstruction errors is, in general, unknown although this is sometimes evaluated using computer simulation (Huang and Knowles, 2009).

3.3.3 Concatenation methods

A simple approach to inferring the species tree using multi-loci sequence data is to concatenate the sequences across loci and then infer a single tree using the “super-gene” sequence as the species tree estimate. This implicitly assumes that all gene loci share the same topology and branch lengths. Systematists have long struggled with the issue of whether to combine different genes into a single analysis (de Queiroz et al., 1995). From a statistical viewpoint, a standard approach for analyzing heterogeneous data is to do a combined analysis accommodating heterogeneity (Yang, 1996). However, until the development and implementation of the MSC model, no formal statistical method existed allowing multiple genes to be combined while respecting their different histories. When the species tree is easy, with long internal branches and small population sizes, one expects very little deep coalescence or incomplete lineage sorting. In such cases, concatenation and coalescent methods are expected to yield the same species topology (Edwards et al., 2007; Leaché and Rannala, 2010; Kubatko and Degnan, 2007). However, when the species tree is challenging, with short internal branches and large population sizes, concatenation may be inconsistent and may converge to an incorrect species tree topology (Roch and Steel, 2015).

Even if gene trees share topology, they may have different branches (coalescent times) due to coalescent fluctuations. In such cases, concatenation can lead to biases in estimation of major evolutionary parameters such as species divergence times, while coalescent methods (full-likelihood methods applied to sequence alignments) accommodate variable coalescence times providing reliable estimates (Ogilvie et al., 2017). A recent Bayesian analysis of diverse phylogenomic data sets (Jiang et al., 2019) suggests that (i) gene tree heterogeneity is real and abundant, even after accounting for gene tree errors; (ii) the concatenation assumption of topologically congruent gene trees can be rejected in almost all datasets; and (iii) the MSC model fits phylogenomic datasets better than the concatenation model. Concatenation continues to be a widely used approach (see Chapter 2.1 [Simion et al. 2020]), especially in comparative analyses of recently sequenced genomes, mainly because of its simplicity and lower computational burden. With the development of improved algorithms for MSC-based species tree inference and broader recognition of the importance of accommodating the coalescent process within species this situation may change.

3.4 Criticisms of MSC species tree inference methods

MSC-based methods of species tree inference make the assumption of no intra-locus recombination. Gatesy and Springer (2013) correctly noted that when multiple exons in transcriptome data are concatenated into one gene or locus, the exons may span large distances along the chromosome; this hybrid concatenation-coalescence approach may lead to violation of the MSC model. Based on empirical calculations, Springer and Gatesy (2016) predicted that the non-recombining unit in a typical species radiation is short enough to violate the MSC assumption of no recombination. However, their calculation does not account for the fact that recombination events during the time period when there is only one sequence in the sample are consistent with the MSC assumption (Edwards et al., 2016). Furthermore, simulation suggests that intra-locus recombination may be a problem for MSC methods under extreme levels of ILS only (Lanier and Knowles, 2012). The assumption of no recombination is more problematic for concatenation than for two-step coalescent methods because concatenation assumes the same genealogical history for all sites in all genes, which is almost certainly violated.

As noted above, two-step coalescent methods treat estimated gene trees as data and do not account for phylogenetic errors; this can cause two-step methods to underestimate internal branches in species trees and exaggerate the importance of ILS by inflating gene tree vs species tree discordance (Yang, 2002; Mirarab et al., 2016; Springer and Gatesy, 2014, 2016). This criticism applies to

3.3:14 Species Tree Inference

“two-step” coalescent methods specifically, because full likelihood methods accommodate gene tree errors correctly through the phylogenetic likelihood function (Equations 14 and 15). Recent efforts making use of the bootstrap and other measures of gene-tree uncertainty to correct for phylogenetic uncertainties in two-step methods may help reduce the impact of phylogenetic errors (Sayyari and Mirarab, 2016). The above discussion largely applies to shallow phylogenies for closely related species. For deep phylogenies involving distantly related species, a whole suite of complicating factors that affect phylogenetic analysis will affect species tree inference as well, including violation of the molecular clock, heterogeneity in the substitution process across genomic loci and across lineages (Yang, 2014). These factors operate in addition to deep coalescence, making inference of deep phylogenies for species that arose through ancient radiative speciation events a very challenging task (see Chapter 3.4 [Bryant and Hahn 2020]). We note that model violation is a common feature in phylogenetics, and whether a misspecified model is still useful may depend on a number of factors including the impact of the model on the analysis (see Chapter 2.1 [Simion et al. 2020]). More complex models, especially MSC models that account for migration or introgression, are likely to be even better than the basic MSC without gene flow and may lead to improved inference under complex scenarios where both deep coalescence and introgression exist (Bravo et al., 2019; Edwards et al., 2016; Nakhleh, 2013; Yu et al., 2013; Zhang et al., 2018).

4 Future Challenges

The development of the multispecies coalescent model is a major advance in molecular phylogenetics (Edwards, 2009). The model accommodates fluctuations in genealogical history across the genome and provides a natural framework for inference using genomic sequence data from closely related species, bridging the gap between phylogenetics and population genetics. The MSC forms the basis for addressing many exciting inference problems in phylogenomics and population genomics, including estimation of ancestral population sizes and inference of ancient hybridisation events – even those hybridization events involving species that have since gone extinct (Xu and Yang, 2016; Degnan, 2018).

Currently full-likelihood implementations of the MSC model, mostly in the form of MCMC algorithms, involve intensive computation. With the increase of data size (e.g., the number of species, the number of sites per sequence, the number of sequences per locus, and the number of loci), each MCMC iteration takes more computational effort. Furthermore there is a deterioration in MCMC mixing so that more MCMC iterations are necessary to generate an acceptable effective sample size. Most of the current MCMC implementations are not computationally feasible for genome-scale datasets with thousands of loci (see Chapter 1.4 [Lartillot 2020]), although implementations of smart MCMC moves in bpp that propose coordinated changes to both the gene trees and the species tree have made it possible to analyse datasets with over 10,000 loci (Rannala and Yang, 2017; Flouri et al., 2018). Further improvements in the computational and mixing efficiency of the algorithms are highly desirable.

The explosive growth of genomic sequence data means that approximate or heuristic methods will continue to play a major role in data analysis (see Chapter 1.2 [Stamatakis and Kozlov 2020]). Current two-step methods appear to make use of only a small portion of the information in genomic datasets, in particular in analysis of shallow phylogenies for closely related species, and as a result many parameters in the MSC model are unidentifiable by the two-step methods, even though the species tree topology is. Development of statistically more efficient heuristic methods should be a priority for future research.

For distantly related species, the molecular clock may be seriously violated. Even though one can adapt the relaxed-clock models developed in phylogenetics (dos Reis et al., 2016) to accommodate

the violation of the clock, the rate variation means that some of the temporal information in gene trees is eroded. It remains to be seen how full likelihood methods under the MSC with relaxed clock compare with heuristic methods using unrooted gene tree topologies and ignoring time information in gene-tree branch lengths.

We expect that accommodating cross-species gene flow in the MSC model will be a research hotspot in the next few years. Many recent empirical studies suggest that cross-species gene flow may be commonplace in animals as well as plants and indeed across the tree of life (Mallet et al., 2016; Folk et al., 2018; Degnan, 2018). The MSC model can be extended to accommodate cross-species gene flow. Two such models have been developed. The MSC-with-migration model, better known as the isolation-with-migration (IM) model (Hey and Nielsen, 2004), assumes continuous migration, with species exchanging migrants at certain rates every generation. This model is similar to population genetics models of population subdivision except that under the IM model the populations have a phylogenetic history with a branching order and divergence times. The probability density of the gene trees under the IM model is given by Hey and Nielsen (2004) and Hey (2010). The MSC with introgression (MSci) model (Flouri et al., 2020), also known as multispecies network coalescent (MSNC) model (Wen and Nakhleh, 2018), assumes episodic introgression/hybridization; in other words, introgression happened at a certain time point in the past. Important parameters in the model include the time of introgression and introgression probability. The gene tree density under the MSci model is given by Yu et al. (2014). Bayesian MCMC implementations include IMA3 (Hey, 2010; Hey et al., 2018) for the IM model, and StarBeast (Zhang et al., 2018; Jones, 2019) and PhyloNet (Wen and Nakhleh, 2018) for the MSci model. Those programs involve expensive computation and are not feasible for realistically sized datasets, with more than 200 loci, say. At the same time, the complexity of those models means that large datasets with thousands of loci may be necessary to obtain reliable parameter estimates. In the case of the IM model, the MCMC averages over a huge space of genealogical history at each locus, which includes the number and directions of migration events. At high migration rates, this space is in effect infinite and the likelihood surface is nearly flat over this space, because the sequence likelihood depends on the gene tree topology and divergence times but not on migration events. For the MSci model, a major stumbling block is the constraint between the species tree or network and the gene trees, as in the case of the simple MSC model. A recent effort to develop coordinated moves between the model parameters such as species divergence or hybridisation times and the gene trees in *bpp* has made it possible to analyse data of more than 10,000 loci (Flouri et al., 2020), but currently the model is fixed, with the number and directions of the introgression events specified by the user. MCMC proposals to allow moves between different MSci models are yet to be implemented. There is an urgent need to improve the computational efficiency of the full likelihood methods.

Several heuristic methods have been developed to detect cross-species gene flow and to estimate the introgression probability. Some take the two-step approach and use estimated gene tree topologies, such as SNaQ (Solis-Lemus et al., 2016, 2017). Others use other summaries of the multi-locus sequence data such as the counts of parsimony-informative site patterns for three or four species, including the popular ABBA-BABA test (Green et al., 2010; Durand et al., 2011) and the HyDe program (Blischak et al., 2018). Those methods do not use information in branch lengths on gene trees, although the recent heuristic method of Hibbins and Hahn (2019) does attempt to use branch length information. They can estimate the introgression probability and internal branch lengths in coalescent units on the species tree but other parameters in the model are unidentifiable. Moreover, many introgression scenarios are not identifiable and cannot be detected using those methods. In cases where the introgression parameter is identifiable the two-step methods appear to provide estimates with similar accuracy to full likelihood methods (Flouri et al., 2020). Developing statistically efficient heuristic methods should be a high priority in the next few years.

Another important avenue for future research concerning the MSci models is their identifiability (Degnan, 2018). The data may be either gene tree topologies (for the two-step heuristic methods) or multilocus sequence alignments (for full likelihood methods). Identifiability may concern either different introgression models (which assume different numbers of introgression events or assume introgressions involving different species) or parameters in a given introgression model (including the species divergence times, population sizes, and introgression probabilities). Some of the identifiability issues might be solved by using more informative summary statistics in two-step methods but that will likely make the derivation of a heuristic estimator more difficult.

Species tree inference is a difficult statistical problem, especially when factors such as introgression are incorporated. The MSC is a model that links population genetics with evolutionary history and it is for this reason central to the problem of species tree inference. We expect that the objective of efficiently and accurately inferring species trees will remain at the heart of the discipline of phylogenetic inference for the foreseeable future. Although much progress has been made during the last two decades many challenging problems remain.

References

- Avise, J. C. (1994). *Molecular Markers, Natural History and Evolution*. Chapman and Hall, New York.
- Avise, J. C., Arnold, J., Ball, R. M., Bermingham, E., Lamb, T., Neigel, J. E., Reeb, C. A., and Saunders, N. C. (1987). Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. *Annual Review of Ecology and Systematics*, pages 489–522.
- Avise, J. C., Lansman, R. A., and Shade, R. O. (1979). The use of restriction endonucleases to measure mitochondrial DNA sequence relatedness in natural populations. i. population structure and evolution in the genus *peromyscus*. *Genetics*, 92(1):279–95.
- Blischak, P. D., Chifman, J., Wolfe, A. D., and Kubatko, L. S. (2018). HyDe: A python package for genome-scale hybridization detection. *Syst. Biol.*, 67(5):821–829.
- Boussau, B., Szollosi, G. J., Duret, L., Gouy, M., Tannier, E., and Daubin, V. (2013). Genome-scale coestimation of species and gene trees. *Genome Res*, 23(2):323–30.
- Bravo, G. A., Antonelli, A., Bacon, C. D., Bartoszek, K., Blom, M. P. K., Huynh, S., Jones, G., Knowles, L. L., Lamichhaney, S., Marcussen, T., Morlon, H., Nakhleh, L. K., Oxelman, B., Pfeil, B., Schliep, A., Wahlberg, N., Werneck, F. P., Wiedenhoeft, J., Willows-Munro, S., and Edwards, S. V. (2019). Embracing heterogeneity: coalescing the tree of life and the future of phylogenomics. *PeerJ*, 7:e6399.
- Brown, W., Prager, E., and Wilson, A. (1982). Mitochondrial DNA sequences of primates: tempo and mode of evolution. *Journal of Molecular Evolution*, 18:225–39.
- Bryant, D. and Hahn, M. W. (2020). The concatenation question. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 3.4, pages 3.4:1–3.4:23. No commercial publisher | Authors open access book.
- Cloutier, A., Sackton, T. B., Grayson, P., Clamp, M., Baker, A. J., and Edwards, S. V. (2019). Whole-genome analyses resolve the phylogeny of flightless birds (palaeognathae) in the presence of an empirical anomaly zone. *Systematic Biology*, 68(6):937–955.
- Dalquen, D., Zhu, T., and Yang, Z. (2017). Maximum likelihood implementation of an isolation-with-migration model for three species. *Syst. Biol.*, 66:379–398.
- Darwin, C. (1859). *On the Origin of Species*. Harvard University Press, Cambridge.
- de Queiroz, A., Donoghue, M. J., and Kim, J. (1995). Separate versus combined analysis of phylogenetic evidence. *Annual Review of Ecology and Systematics*, 26:657–681.
- Degnan, J. H. (2013). Anomalous unrooted gene trees. *Systematic Biology*, 62(4):574–590.

- Degnan, J. H. (2018). Modeling hybridization under the network multispecies coalescent. *Syst. Biol.*, 67(5):786–799.
- Degnan, J. H. and Rosenberg, N. A. (2006). Discordance of species trees with their most likely gene trees. *PLoS Genetics*, 2(5):e68.
- Degnan, J. H. and Rosenberg, N. A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.*, 24:332–340.
- Degnan, J. H. and Salter, L. A. (2005). Gene tree distributions under the coalescent process. *Evolution*, 59:24–37.
- dos Reis, M., Donoghue, P. C. J., and Yang, Z. (2016). Bayesian molecular clock dating of species divergences in the genomics era. *Nat. Rev. Genet.*, 17:71–80.
- Durand, E. Y., Patterson, N., Reich, D., and Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.*, 28:2239–2252.
- Edwards, A. W. (1970). Estimation of the branch points of a branching diffusion process. *Journal of the Royal Statistical Society: Series B (Methodological)*, 32(2):155–164.
- Edwards, S. V. (2009). Is a new and general theory of molecular systematics emerging? *Evolution*, 63(1):1–19.
- Edwards, S. V. (2016). Inferring species trees. In Kliman, R., editor, *Encyclopedia of Evolutionary Biology*. Elsevier, New York.
- Edwards, S. V., Liu, L., and Pearl, D. K. (2007). High-resolution species trees without concatenation. *Proceedings of the National Academy of Sciences*, 104(14):5936–5941.
- Edwards, S. V., Xi, Z., Janke, A., Faircloth, B. C., McCormack, J. E., Glenn, T. C., Zhong, B., Wu, S., Lemmon, E. M., Lemmon, A. R., Leaché, A. D., Liu, L., and Davis, C. C. (2016). Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics. *Mol. Phylogenet. Evol.*, 94(Pt A):447–462.
- Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, 17:368–376.
- Flouri, T., Jiao, X., Rannala, B., and Yang, Z. (2018). Species tree inference with BPP using genomic sequences and the multispecies coalescent. *Mol. Biol. Evol.*, 35(10):2585–2593.
- Flouri, T., Jiao, X., Rannala, B., and Yang, Z. (2020). A bayesian implementation of the multispecies coalescent model with introgression for comparative genomic analysis. *Mol. Biol. Evol.*, page under review.
- Folk, R. A., Soltis, P. S., Soltis, D. E., and Guralnick, R. (2018). New prospects in the detection and comparative analysis of hybridization in the tree of life. *Am. J. Bot.*, 105(3):364–375.
- Gatesy, J. and Springer, M. S. (2013). Concatenation versus coalescence versus "concatalescence". *Proceedings of the National Academy of Sciences of the United States of America*, 110(13):E1179.
- Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H., Hansen, N. F., Durand, E. Y., Malaspina, A. S., Jensen, J. D., Marques-Bonet, T., Alkan, C., Prufer, K., Meyer, M., Burbano, H. A., Good, J. M., Schultz, R., Aximu-Petri, A., Butthof, A., Hober, B., Hoffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E. S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Z., Gusic, I., Doronichev, V. B., Golovanova, L. V., Lalueza-Fox, C., de la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R. W., Johnson, P. L., Eichler, E. E., Falush, D., Birney, E., Mullikin, J. C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D., and Paabo, S. (2010). A draft sequence of the neandertal genome. *Science*, 328:710–722.
- Griffiths, R. C. and Marjoram, P. (1996). Ancestral inference from samples of DNA sequences with recombination. *Journal of Computational Biology*, 3(4):479–502.
- Hare, M. (2001). Prospects for nuclear gene phylogeography. *Trends in Ecology and Evolution*, 16:700–706.

3.3:18 REFERENCES

- Heled, J. and Drummond, A. J. (2010). Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol.*, 27:570–580.
- Hey, J. (2010). Isolation with migration models for more than two populations. *Mol. Biol. Evol.*, 27:905–920.
- Hey, J., Chung, Y., Sethuraman, A., Lachance, J., Tishkoff, S., Sousa, V. C., and Wang, Y. (2018). Phylogeny estimation by integration over isolation with migration models. *Mol. Biol. Evol.*, 35(11):2805–2818.
- Hey, J. and Nielsen, R. (2004). Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics*, 167:747–760.
- Hey, J. and Nielsen, R. (2007). Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proc Natl Acad Sci U S A*, 104(8):2785–2790.
- Hibbins, M. S. and Hahn, M. W. (2019). The timing and direction of introgression under the multispecies network coalescent. *Genetics*, 211(3):1059–1073.
- Huang, H. and Knowles, L. L. (2009). What is the danger of the anomaly zone for empirical phylogenetics? *Syst. Biol.*, 58:527–536.
- Hudson, R. (1990). Gene genealogies and the coalescent process. In Futuyma, D. and Antonovics, J. D., editors, *Oxford Surveys in Evolutionary Biology*, pages 1–44. Oxford University Press, New York.
- Hudson, R. R. (1983). Testing the constant-rate neutral allele model with protein sequence data. *Evolution*, pages 203–217.
- Hudson, R. R. and Kaplan, N. L. (1985). Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, 111(1):147–164.
- Jiang, X., Edwards, S., and Liu, L. (2019). The multispecies coalescent model outperforms concatenation across diverse phylogenomic data sets. *bioRxiv*.
- Jones, G. (2017). Algorithmic improvements to species delimitation and phylogeny estimation under the multispecies coalescent. *J. Math. Biol.*, 74:447–467.
- Jones, G. R. (2019). Divergence estimation in the presence of incomplete lineage sorting and migration. *Syst. Biol.*, 68(1):19–31.
- Kingman, J. F. (1982). The coalescent. *Stochastic Processes and Their Applications*, 13(3):235–248.
- Kocher, T. D., Thomas, W. K., Meyer, A., Edwards, S. V., Pääbo, S., Villablanca, F. X., and Wilson, A. C. (1989). Dynamics of mitochondrial DNA evolution in animals: amplification and sequencing with conserved primers. *Proceedings of the National Academy of Sciences (USA)*, 86:6196–6200.
- Krone, S. M. and Neuhauser, C. (1997). Ancestral processes with selection. *Theoretical Population Biology*, 51(3):210–237.
- Kubatko, L. S., Carstens, B. C., and Knowles, L. L. (2009). STEM: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics*, 25(7):971–973.
- Kubatko, L. S. and Degnan, J. H. (2007). Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology*, 56(1):17–24.
- Lanier, H. C. and Knowles, L. L. (2012). Is recombination a problem for species-tree analyses? *Systematic Biology*, 61(4):691–701.
- Lartillot, N. (2020). The bayesian approach to molecular phylogeny. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.4, pages 1.4:1–1.4:17. No commercial publisher | Authors open access book.
- Leaché, A. D. and Rannala, B. (2010). The accuracy of species tree estimation under simulation: a comparison of methods. *Systematic Biology*, 60(2):126–137.

- Linkem, C. W., Minin, V. N., and Leaché, A. D. (2016). Detecting the anomaly zone in species trees and evidence for a misleading signal in higher-level skink phylogeny (squamata: Scincidae). *Systematic Biology*, 65(3):465–477.
- Liu, L. (2008). BEST: Bayesian estimation of species trees under the coalescent model. *Bioinformatics*, 24(21):2542–2543.
- Liu, L. and Pearl, D. K. (2007). Species trees from gene trees: reconstructing bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Syst. Biol.*, 56(3):504–514.
- Liu, L., Xi, Z., Wu, S., Davis, C. C., and Edwards, S. V. (2015). Estimating phylogenetic trees from genome-scale data. *Annals of the New York Academy of Sciences*, 1360:36–53.
- Liu, L., Yu, L., Kubatko, L., Pearl, D. K., and Edwards, S. V. (2009). Coalescent methods for estimating phylogenetic trees. *Mol Phylogenet Evol*, 53(1):320–8.
- Liu, L., Yu, L., and Pearl, D. K. (2010). Maximum tree: a consistent estimator of the species tree. *J. Math. Biol.*, 60:95–106.
- Maddison, W. P. (1997). Gene trees in species trees. *Systematic Biology*, 46(3):523–536.
- Mallet, J., Besansky, N., and Hahn, M. W. (2016). How reticulated are species? *BioEssays*, 38(2):140–149.
- Mirarab, S., Bayzid, M., and Warnow, T. (2016). Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. *Systematic Biology*, 65:366–380.
- Mirarab, S., Reaz, R., Bayzid, M. S., Zimmermann, T., Swenson, M. S., and Warnow, T. (2014). ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*, 30(17):i541–i548.
- Mirarab, S. and Warnow, T. (2015). ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*, 31(12):i44–i52.
- Nakhleh, L. (2013). Computational approaches to species phylogeny inference and gene tree reconciliation. *Trends in Ecology and Evolution*, 28(12):719–728.
- Nichols, R. (2001). Gene trees and species trees are not the same. *Trends Ecol. Evol.*, 16:358–364.
- Nordborg, M. (2007). Coalescent theory. In Balding, D., Bishop, M., and Cannings, C., editors, *Handbook of Statistical Genetics*, pages 843–877. Wiley, San Francisco.
- Ogilvie, H. A., Bouckaert, R. R., and Drummond, A. J. (2017). StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. *Molecular Biology and Evolution*, 34(8):2101–2114.
- Pamilo, P. and Nei, M. (1988). Relationships between gene trees and species trees. *Molecular Biology and Evolution*, 5(5):568–583.
- Rannala, B. and Yang, Z. (2003). Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, 164(4):1645–1656.
- Rannala, B. and Yang, Z. (2008). Phylogenetic inference using whole genomes. *Annual Review of Genomics and Human Genetics*, 9:217–231.
- Rannala, B. and Yang, Z. (2017). Efficient Bayesian species tree inference under the multispecies coalescent. *Syst. Biol.*, 66:823–842.
- Roch, S. and Steel, M. (2015). Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. *Theor. Popul. Biol.*, 100:56–62.
- Ronquist, F. and Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19:1572–1574.
- Rosenberg, N. A. (2002). The probability of topological concordance of gene trees and species trees. *Theoretical Population Biology*, 61(2):225–247.
- Sayyari, E. and Mirarab, S. (2016). Fast coalescent-based computation of local branch support from quartet frequencies. *Mol. Biol. Evol.*, 33(7):1654–1668.

3.3:20 REFERENCES

- Shi, C. and Yang, Z. (2018). Coalescent-based analyses of genomic sequence data provide a robust resolution of phylogenetic relationships among major groups of gibbons. *Mol. Biol. Evol.*, 35:159–179.
- Simion, P., Delsuc, F., and Philippe, H. (2020). To what extent current limits of phylogenomics can be overcome? In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 2.1, pages 2.1:1–2.1:34. No commercial publisher | Authors open access book.
- Solis-Lemus, C., Bastide, P., and Ane, C. (2017). PhyloNetworks: A package for phylogenetic networks. *Mol. Biol. Evol.*, 34(12):3292–3298.
- Solis-Lemus, C., Yang, M., and Ane, C. (2016). Inconsistency of species tree methods under gene flow. *Syst. Biol.*, 65(5):843–851.
- Springer, M. S. and Gatesy, J. (2014). Land plant origins and coalescence confusion. *Trends in Plant Science*, 19(5):267–9.
- Springer, M. S. and Gatesy, J. (2016). The gene tree delusion. *Molecular Phylogenetics and Evolution*, 94:1–33.
- Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690.
- Stamatakis, A. and Kozlov, A. M. (2020). Efficient maximum likelihood tree building methods. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.2, pages 1.2:1–1.2:18. No commercial publisher | Authors open access book.
- Szollósi, G. J., Tannier, E., Daubin, V., and Boussau, B. (2015). The inference of gene trees with species trees. *Syst. Biol.*, 64(1):e42–62.
- Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics*, 105(2):437–460.
- Takahata, N., Satta, Y., and Klein, J. (1995). Divergence time and population size in the lineage leading to modern humans. *Theor. Popul. Biol.*, 48:198–221.
- Thawornwattana, Y., Dalquen, D., and Yang, Z. (2018). Coalescent analysis of phylogenomic data confidently resolves the species relationships in the *Anopheles gambiae* species complex. *Mol. Biol. Evol.*, 35(10):2512–2527.
- Wakeley, J. (2009). *Coalescent Theory: An Introduction*. Roberts & Company, Greenwood Village, Colorado.
- Wen, D. and Nakhleh, L. (2018). Coestimating reticulate phylogenies and gene trees from multilocus sequence data. *Syst. Biol.*, 67(3):439–457.
- Wilson, A., Cann, R. L., Carr, S. M., George, M., Gyllensten, U. B., Helm-Bychowski, K. M., Higuchi, R. G., Palumbi, S. R., Prager, E. M., Sage, R. D., and Stoneking, M. (1985). Mitochondrial DNA and two perspectives on evolutionary genetics. *Biological Journal of the Linnaean Society*, 26:375–400.
- Wu, Y. (2012). Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. *Evolution: International Journal of Organic Evolution*, 66(3):763–775.
- Wu, Y. (2016). An algorithm for computing the gene tree probability under the multispecies coalescent and its application in the inference of population tree. *Bioinformatics*, 32(12):i225–i233.
- Xu, B. and Yang, Z. (2016). Challenges in species tree estimation under the multispecies coalescent model. *Genetics*, 204(4):1353–1368.
- Yang, Z. (1996). Maximum-likelihood models for combined analyses of multiple sequence data. *J. Mol. Evol.*, 42:587–596.
- Yang, Z. (2002). Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. *Genetics*, 162(4):1811–1823.

- Yang, Z. (2014). *Molecular Evolution A Statistical Approach*. Oxford University Press, Oxford, England.
- Yang, Z. and Rannala, B. (2010). Bayesian species delimitation using multilocus sequence data. *Proc. Natl. Acad. Sci. USA*, 107:9264–9269.
- Yang, Z. and Rannala, B. (2014). Unguided species delimitation using DNA sequence data from multiple loci. *Mol. Biol. Evol.*, 31(12):3125–3135.
- Yu, Y., Dong, J., Liu, K. J., and Nakhleh, L. (2014). Maximum likelihood inference of reticulate evolutionary histories. *Proc. Natl. Acad. Sci. U.S.A.*, 111(46):16448–16453.
- Yu, Y., Ristic, N., and Nakhleh, L. (2013). Fast algorithms and heuristics for phylogenomics under ILS and hybridization. *BMC Bioinformatics*, 14 Suppl 15:S6.
- Zhang, C., Ogilvie, H. A., Drummond, A. J., and Stadler, T. (2018). Bayesian inference of species networks from multilocus sequence data. *Mol. Biol. Evol.*, 35:504–517.