The Interdiciplinary Center, Herzliya
Efi Arazi School of Computer Science
M.Sc. Program - Research Track

A New Bayesian Method for Comparing Demographic Models

by
Ron Visbord

M.Sc. dissertation, submitted in partial fulfillment of the requirements
for the M.Sc. degree, research track, School of Computer Science
The Interdisciplinary Center, Herzliya

May, 2018

This work was carried out under the supervision of Dr. Ilan Gronau from the Efi Arazi School of Computer Science, The Interdiciplinary Center, Herzliya.

Abstract

The advent of high throughput sequencing has greatly improved our ability to investigate the evolutionary history of species using detailed demographic models. A popular approach for inferring parameters in these demographic models is to sample genealogical histories at many short unlinked loci using a Markov Chain Monte Carlo algorithm. The use of explicit coalescent models by these methods makes them powerful for inferring demographic parameters, but they are limited in their ability to assess the fit of the inferred model to data. The purpose of this research is to examine a new approach, based on Relative Bayes Factors, for using genealogy samples to compare different evolutionary hypotheses.

In this work we review Bayesian inference of parameterized demographic models and formalize the model selection problem. We then define Relative Bayes Factors (RBFs), which represent demographic model fit relative to some reference demographic model. We further derive RBFs for two types of reference models - Clade models and Comb models. The two types are useful for different model-selection problem instances. Having reached tractable formulae for relative model fit, we describe in detail how they are calculated in an efficient manner, without incurring significant computational overhead during MCMC sampling. Finally, we test these model-fit assessments using a series of model-selection experiments based on simulated sequence data. Our results show RBFs significantly improving on the base-line harmonic mean model fit estimator in the model selection task.

Contents

1 Introduction

In recent years, advances in high throughput DNA sequencing have made it easy to sequence many genomes of individuals from closely related species. This allows evolutionary biologists to examine the evolution of recently diverged species by employing sophisticated computational methods and statistical models. Typically, an evolutionary biologist, having obtained and aligned genome sequences of individuals from closely related species or populations, would like to reconstruct the evolutionary history of these populations. This evolutionary history includes a series of population splits, population size changes and post-divergence gene flow.

Evolutionary history is often modeled using a parameterized probabilistic demographic model $\mathcal{M}$, which induces a probability distribution over observed genomic data $\mathbf{X}$. The structural components of $\mathcal{M}$ consist of a population phylogeny $\mathcal{T}$ and a collection of migration bands B that indicate ordered pairs of populations between which gene flow is allowed. The free parameters of $\mathcal{M}$, such as population divergence times, population sizes and migration rates, are denoted by $\mathbf{\Theta}$. The model $\mathcal{M}$ is thus defined by specifying the structural components $(\mathcal{T}, B)$ and a prior distribution over the free parameters of the model $P(\mathbf{\Theta}|\mathcal{M})$. The conditional probability distribution for the observed genomic data $P(\mathbf{X}|\mathcal{M}, \mathbf{\Theta})$ is defined by standard models for molecular evolution and population genetics (e.g., Jukes and Cantor (1969); Kingman (1982)).

A common approach to inferring parameters of a demographic model $\mathcal{M}$ is to assume the model structure $(\mathcal{T}, B)$, and to explicitly represent the genealogy of the sequenced individuals at short unlinked loci. These genealogies are used along-side the target model parameters as hidden variables in a Markov chain Monte Carlo (MCMC) sampling algorithm. The algorithm effectively integrates out the genealogical relationships between individuals and produces Bayesian estimates of target parameters. These methods have two key advantages. 1) The full probabilistic generative model of the data at their core allows modeling of more complex evolutionary history, with more free parameters; 2) The parameter values sampled by the MCMC provide means to assess the uncertainty in the resulting estimates. However, because these methods condition on a given model structure, they provide no straightforward way to compare demographic model hypotheses.

In principle, measuring of model fit $P(X|\mathcal{M})$ can be approximated by using importance sampling on the approximated posterior distribution (Newton and Raftery (1994)), and this could be used to compare models. However, it was shown that estimates tend to be biased upward, and they are more biased the more parameter-rich the model is (Xie et al., 2011). There have been several methods suggested to improve the accuracy of importance sampling estimation by sampling from "hybrid" models (Lartillot and Philippe, 2006; Xie et al., 2011). These methods are very effective, but they require an order of magnitude more sampling iterations (10x) compared to the number of iterations required for the MCMC of parameter inference. So they are not very practical in our setting.

The goal of our research is thus to improve on existing importance-sampling approaches to model selection, without incurring significant additional computational cost. We accomplish this by estimating model fit relative to some reference model $\mathcal{M}_{ref}$. Reference models are base-line phylogenetic structures used to asses model fit within a specific context, allowing us to better select between competing model candidates. We implement the model-selection algorithm based on the parameter-inference framework G-PhoCS , but our theory and approach can be applied to all bayesian demography inference methods.

We will start in subsection 1.1 by overviewing relevant work in the field. Subsections 1.2-1.3 present background on the demography inference problem and state the model selection problem. Section 2 formally introduces the concept of reference models and explains how they relate to phylogenetic population models. It then derives the theory behind our relative Bayes factors (RBFs), and explains how they are used as model selection criteria. Section 3 explains in depth our implementation of McRef - our model selection algorithm which uses the G-PhoCS parameter-inference framework. McRef earned it's nickname due to it's employment of reference models in the MCMC process. Finally, in section 4 we share empirical results from our model-selection experiments on simulated sequence data, showcasing the advantages and limitations of our method.

1.1 Related work

There are several common approaches for demography inference; Likelihood-based models associate each model $\mathcal{M}$ with the most likely parameter values $\mathbf{\Theta}$. The joint likelihood $P(\mathbf{X}|\mathcal{M}, \mathbf{\Theta})$ is then approximated by making additional simplifying assumptions on the population genetic model, or the data. There are methods which assume that all sites are independent (i.e. allow free recombination between sites) and use a combination of analytic calculations and simulations to estimate $P(\mathbf{X}|\mathcal{M}, \mathbf{\Theta})$ (Gutenkunst et al., 2009; Kamm, Terhorst and Song, 2017; Kamm et al., 2018). Other methods use summary statistics extracted from the data, such as the lengths of shared haplotypes (Harris and Nielsen, 2013; Browning and Browning, 2015). The key disadvantages of these methods is that 1) they make many simplifying assumptions, and 2) they associate a model with its most likely parameter values. This means they give an advantage to models which imply high confidence in the parameter values compared to models where the likelihood is more spread out across the parameter space.

Bayesian model-based methods, such as IM (Nielsen and Wakeley, 2001) (most updated version IMa2p (Hey and Nielsen, 2007; Sethuraman and Hey, 2016)), MCMCcoal Rannala and Yang (2003) (most updated version BPP (Yang, 2015)), and G-PhoCS (Gronau et al., 2011) all explicitly model genealogies coalescing in a population phylogeny, and differ mostly in additional modeling assumptions and software design. BPP does not model gene flow between populations and is thus mostly used for relatively diverged species. IM was originally developed for analyzing data from models with only two leaf populations. It has since been extended for larger population phylogenies, but its design limits its use to relatively small data sets (few populations and up to 1,000 loci). Importantly, all methods use MCMC to generate posterior samples of the model parameters, and the model selection methods we develop here can be applied to all of them.

Regarding estimation of Bayes factors, the basic idea to use importance sampling (IS) to estimate $P(\mathbf{X}|\mathcal{M})$ in a Bayesian setting was suggested by Newton and Raftery (1994). This idea has since become the standard way to estimate model fit in a Bayesian setting, but experience has shown it to be very noisy and biased toward more complex models (Xie et al., 2011). In particular, it was shown that estimates tend to be biased upward, and they are more biased the more parameter-rich the model is. Several methods have suggested ways to improve the accuracy of IS estimation by sampling from "hybrid" models, which combine the prior $P(\mathbf{\Theta}|\mathcal{M})$ times some power of the conditional $P(\mathbf{X}, \mathbf{G}|\mathcal{M}, \mathbf{\Theta})$ (Lartillot and Philippe, 2006; Xie et al., 2011). Unfortunately, though these methods are effective, they require an order of magnitude more sampling iterations compared to the number of iterations required for the MCMC of parameter estimation.

1.2 Bayesian inference and G-PhoCS

The objective of demography inference methods is to infer values for $\mathbf{\Theta}$ that have high joint probability with the data: $P(\mathbf{X}, \mathbf{\Theta}|\mathcal{M}) = P(\mathbf{\Theta}|\mathcal{M})P(\mathbf{X}|\mathcal{M}, \mathbf{\Theta})$, where $\mathbf{\Theta}$ consist of divergence times, $\boldsymbol{\tau} = \{\tau_p : p \text{ is an ancestral population in } \mathcal{T}\}$, effective population sizes, $\boldsymbol{\theta} = \{\theta_p : p \text{ is a population in } \mathcal{T}\}$, and migration rates, $\mathbf{m} = \{m_b : b \in B\}$. Values of parameters in $\mathbf{\Theta}$ are scaled by mutation rate.

Because the conditional probability $P(\mathbf{X}|\mathcal{M}, \mathbf{\Theta})$ does not typically have a closed-form expression, an increasingly popular approach for inference is to introduce additional hidden variables $\mathbf{G}$, which represent genealogical relationships between the sampled individuals. The benefit of this is that given the genealogical information, the data $\mathbf{X}$ becomes independent of the model $\mathcal{M}$ and parameters $\mathbf{\Theta}$, and the likelihood can be expressed as a product of three tractable terms:

$$P(\mathbf{X}, \mathbf{G}, \mathbf{\Theta}|\mathcal{M}) = P(\mathbf{\Theta}|\mathcal{M})P(\mathbf{G}|\mathcal{M}, \mathbf{\Theta})P(\mathbf{X}|\mathbf{G}) . \tag{1}$$

This joint probability function may be used by a Markov chain Monte Carlo (MCMC) algorithm to generate a sample of model parameters together with genealogies according to a probability distribution approximating the posterior, $P(\mathbf{G}, \mathbf{\Theta}|\mathcal{M}, \mathbf{X})$. Consequently, the sampled parameter values have high joint probability with the data. A major advantage of this approach to parameter inference is that it is extremely flexible and can be applied to a wide range of demographic models and different types of genomic data.

G-PhoCS is one such Bayesian demography inference method. G-PhoCS considers a model of sequence data at short unlinked loci, where $\mathbf{G}$ contains the information on the local tree in each locus, and loci are assumed to be independent (Figure 1) (e.g., Nielsen and Wakeley (2001); Rannala and Yang (2003); Gronau et al. (2011)). Equation 2 shows the probability distribution approximated by G-PhoCS.

$$P(\mathbf{X}, \mathbf{G}, \mathbf{\Theta}|\mathcal{M}) \;=\; P(\mathbf{\Theta}|\mathcal{M})P(\mathbf{G}|\mathcal{M}, \mathbf{\Theta})P(\mathbf{X}|\mathbf{G}) \;=\; P(\mathbf{\Theta}|\mathcal{M})\prod_{l} P(\mathbf{G}_l|\mathcal{M}, \mathbf{\Theta})P(\mathbf{X}_l|\mathbf{G}_l). \qquad (2)$$

In the above Equation 2 $P(\mathbf{\Theta}|\mathcal{M})$ is the prior probability of model parameters. $P(\mathbf{G}_l|\mathcal{M}, \mathbf{\Theta})$ is the probability of local genealogy G_l at locus l given the model parameters. This is calculated under the Kingman Coalescent model, with special regard to migration events. $P(\mathbf{X}_l|\mathbf{G}_l)$ is the local data likelihood given local genealogy G_l, which is computed using standard DNA substitution models (Jukes and Cantor (1969)). In each MCMC update step G-PhoCS proposes a new instace of $\mathbf{G}\&\mathbf{\Theta}$. It then decides whether to accept the proposal based on the ratio between complete likelihoods of the current instance and proposed instance. Each G-PhoCS update step is divided into a series of Metropolis-Hastings updates of subsets of variables. The update steps are:

1. Update coalescent times: For each individual coalescent event in each population, perturb the time of the event without changing the topology of the genealogy or any other coalescent time.

2. Update genealogy structure: For each subtree of each genealogy, alter the subtree using a subtree prune-and-regraft operation.

3. Update θ_p: For each population p, perturb θ_p.

4. Update τ_p: For each population p, perturb τ_p. If nescessary, also “stretch” or "squeeze" each genealogy G_i as needed to accommodate the proposed change in τ_p.

5. Rescale all parameters: Slightly perturb all model parameters θ_p, τ_p, m_b and all coalescent times across all genealogies by a multiplicative factor sampled close to 1.

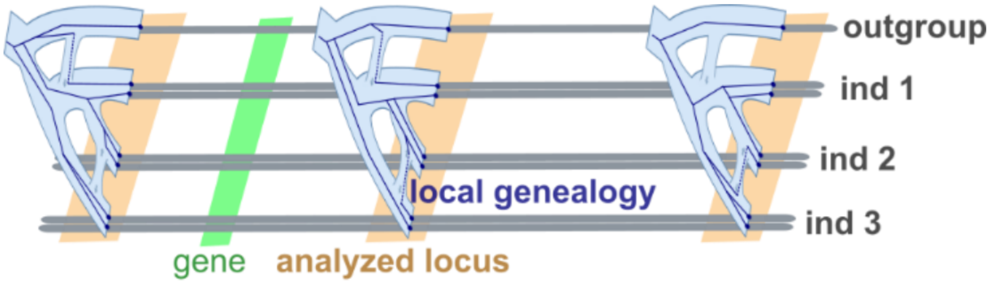


Figure 1. G-PhoCS uses independant loci chosen to be far away from genes and from each other to reduce the influence of selection and accomodate the assumption of independence. A local genealogy is represented over each locus and embedded in the population phylogeny.

1.3 The model selection problem

The model selection problem takes sequence data and a series of demography models $\mathcal{M}_1, ...\mathcal{M}_n$, which differ in their structural components, and aims to find the one which best fits the data set, i.e. select the model $\mathcal{M}_i$ with maximal $P(X|\mathcal{M}_i)$. Typically only the structural assumptions of the models are compared ($\mathcal{T}$ and B), and not specific parameter values ($\mathbf{\Theta}$). Figure 2 is an example instance of the model selection problem, in which we need to choose the best fitting model amongst three structural hypotheses.

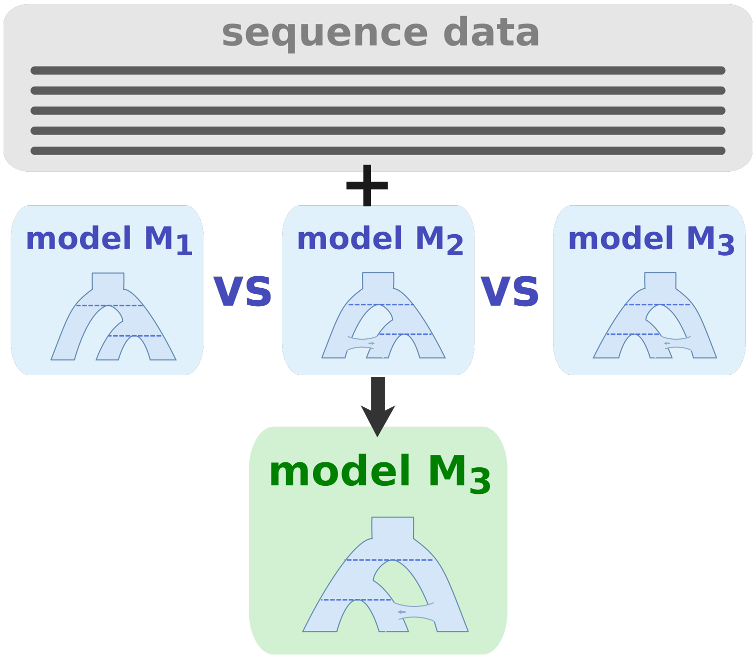


Figure 2. An example problem of selecting between three models with different topological structures. Model $\mathcal{M}_1$ has three leaf populations and no migration. Model $\mathcal{M}_2$ has the same phylogeny as $\mathcal{M}_1$ but with an additional migration band. In model $\mathcal{M}_3$ the relationship between leaves is different. Given aligned sequence data and a set of structural hypotheses, we wish to choose which structural model best fits the data.

In this study, building upon the G-PhoCS demography inference method and MCMC sampler, we develop the theoretical framework for a robust model-selection scheme, and implement a method for comparing models and their fit to data, this without analytically calculating $P(X|\mathcal{M}_i)$.

2 Methods

2.1 Estimating data likelihood via importance sampling

Model fit is best captured by the marginal data likelihood, $P(\mathbf{X}|\mathcal{M})$, whose computation involves integration over the space of unknown parameter values and genealogical relationships, denoted jointly by $\mathbf{G\Theta}$. This high-dimensional integral may be approximated via importance sampling using a collection of instances $\{\mathbf{G\Theta}^{(i)}\}$ sampled via MCMC conditioned on $\mathbf{X}$ and $\mathcal{M}$. The approximation is established by expressing the inverse of the likelihood as an expected value under the posterior distribution of $\mathbf{G\Theta}$ given $\mathcal{M}$ and $\mathbf{X}$:

$$\begin{aligned}
\frac{1}{P(\mathbf{X}|\mathcal{M})} &= \frac{\int P(\mathbf{G\Theta}|\mathcal{M}) d\mathbf{G\Theta}}{P(\mathbf{X}|\mathcal{M})} \\
&= \int \frac{P(\mathbf{G\Theta}|\mathcal{M})}{P(\mathbf{X}|\mathcal{M})} \frac{P(\mathbf{X},\mathbf{G\Theta}|\mathcal{M})}{P(\mathbf{X},\mathbf{G\Theta}|\mathcal{M})} d\mathbf{G\Theta} \\
&= \int \frac{P(\mathbf{G\Theta},\mathbf{X}|\mathcal{M})}{P(\mathbf{X}|\mathcal{M})} \Bigg/ \frac{P(\mathbf{X},\mathbf{G\Theta}|\mathcal{M})}{P(\mathbf{G\Theta}|\mathcal{M})} d\mathbf{G\Theta} \\
&= \int \frac{P(\mathbf{G\Theta}|\mathcal{M},\mathbf{X})}{P(\mathbf{X}|\mathcal{M},\mathbf{G\Theta})} d\mathbf{G\Theta} \\
&= \int \frac{1}{P(\mathbf{X}|\mathbf{G})} P(\mathbf{G\Theta}|\mathcal{M},\mathbf{X}) d\mathbf{G\Theta} \\
&= \mathbb{E}_{\mathbf{G\Theta}|\mathcal{M},\mathbf{X}} \left[\frac{1}{P(\mathbf{X}|\mathbf{G})} \right] \\
&\approx \frac{1}{N} \sum_{i=1}^{N} \frac{1}{P(\mathbf{X}|\mathbf{G}^{(i)})} \,. \qquad (3)
\end{aligned}$$

This *harmonic mean estimator* is straightforward and can be applied in a very general setting, but its practical use is often limited due to very high variance of the inverse likelihood, $1/P(\mathbf{X}|\mathbf{G})$. This high variance means that only models with very different levels of fit may be compared reliably via harmonic mean estimators of $P(\mathbf{X}|\mathcal{M})$. The main objective of the approach we propose next is to correlate the sensitivity of model comparison with the level of similarity between the models being compared.

2.2 Relative Bayes factors

We propose here an alternative way to evaluate the fit of model $\mathcal{M}$ by estimating its likelihood relative to some reference model $\mathcal{M}_{ref}$. As before, assume a collection $\{\mathbf{G\Theta}^{(i)}\}$ sampled via MCMC according to an approximate posterior probability distribution $P(\mathbf{G\Theta}|\mathcal{M},\mathbf{X})$. We wish to use these MCMC samples to estimate the *Bayes factor of $\mathcal{M}$ relative to $\mathcal{M}_{ref}$*, defined as the ratio $P(\mathbf{X}|\mathcal{M})/P(\mathbf{X}|\mathcal{M}_{ref})$. The Bayes factor can be estimated by running an additional MCMC for $\mathcal{M}_{ref}$ and taking the ratio of the two harmonic-mean estimates for $P(\mathbf{X}|\mathcal{M})$ and $P(\mathbf{X}|\mathcal{M}_{ref})$. However, in some cases the relative Bayes factor may be estimated directly from $\{\mathbf{G\Theta}^{(i)}\}$ without the need of an additional MCMC for $\mathcal{M}_{ref}$. This is done by connecting the models $\mathcal{M}$ and $\mathcal{M}_{ref}$ via a conditional distribution over the the hidden variables of $\mathcal{M}$, $\widetilde{P}(\mathbf{G\Theta}|\mathcal{M}_{ref})$, which satisfies the following two requirements:

$$P(\mathbf{X}|\mathcal{M}_{ref}) \;=\; \int \widetilde{P}(\mathbf{G\Theta}|\mathcal{M}_{ref})\, P(\mathbf{X}|\mathbf{G})\, d\mathbf{G\Theta} \qquad (4)$$

$$P(\mathbf{G\Theta}|\mathcal{M},\mathbf{X}) = 0 \;\Rightarrow\; \widetilde{P}(\mathbf{G\Theta}|\mathcal{M}_{ref}) = 0 \qquad (5)$$

The *model pairing conditional distribution*, $\widetilde{P}(\mathbf{G\Theta}|\mathcal{M}_{ref})$, plays a key role in our estimator of the relative Bayes factor. The special notation $\widetilde{P}$ indicates that this probability function is not naturally defined by either $\mathcal{M}$ or $\mathcal{M}_{ref}$, and there will typically be some degree of freedom associated with its specification. Given a model-pairing conditional distribution, the relative Bayes factor may be expressed as an expected value under the posterior distribution of $\mathbf{G}\boldsymbol{\Theta}$ given $\mathcal{M}$ and $\mathbf{X}$, implying the following approximation:

$$\frac{1}{\text{BF}(\mathcal{M}:\mathcal{M}_{ref}|\mathbf{X})} \triangleq \frac{P(\mathbf{X}|\mathcal{M}_{ref})}{P(\mathbf{X}|\mathcal{M})} = \frac{\int \widetilde{P}(\mathbf{G}\boldsymbol{\Theta}|\mathcal{M}_{ref})\ P(\mathbf{X}|\mathbf{G})\ d\mathbf{G}\boldsymbol{\Theta}}{P(\mathbf{X}|\mathcal{M})} \tag{6}$$

$$= \int \frac{\widetilde{P}(\mathbf{G}\boldsymbol{\Theta}|\mathcal{M}_{ref})\ P(\mathbf{X}|\mathbf{G})}{P(\mathbf{X}|\mathcal{M})}\ \frac{P(\mathbf{G}\boldsymbol{\Theta}|\mathcal{M},\mathbf{X})}{P(\mathbf{G}\boldsymbol{\Theta}|\mathcal{M},\mathbf{X})} d\mathbf{G}\boldsymbol{\Theta} \tag{7}$$

$$= \int \frac{\widetilde{P}(\mathbf{G}\boldsymbol{\Theta}|\mathcal{M}_{ref})\ P(\mathbf{X}|\mathbf{G})}{P(\mathbf{X},\mathbf{G}\boldsymbol{\Theta}|\mathcal{M})} P(\mathbf{G}\boldsymbol{\Theta}|\mathcal{M},\mathbf{X}) d\mathbf{G}\boldsymbol{\Theta}$$

$$= \int \frac{\widetilde{P}(\mathbf{G}\boldsymbol{\Theta}|\mathcal{M}_{ref})}{P(\mathbf{G}\boldsymbol{\Theta}|\mathcal{M})} P(\mathbf{G}\boldsymbol{\Theta}|\mathcal{M},\mathbf{X}) d\mathbf{G}\boldsymbol{\Theta} \tag{8}$$

$$= \mathbb{E}_{\mathbf{G}\boldsymbol{\Theta}|\mathcal{M},X}\left[\frac{\widetilde{P}(\mathbf{G}\boldsymbol{\Theta}|\mathcal{M}_{ref})}{P(\mathbf{G}\boldsymbol{\Theta}|\mathcal{M})}\right].$$

$$\approx \frac{1}{N}\sum_{i=1}^{N} \frac{\widetilde{P}(\mathbf{G}\boldsymbol{\Theta}^{(i)}|\mathcal{M}_{ref})}{P(\mathbf{G}\boldsymbol{\Theta}^{(i)}|\mathcal{M})}. \tag{9}$$

Note that the condition of Equation 4 implies the equality in Equation 6, and the condition of Equation 5 guarantees no division-by-zero in Equation 7. Interestingly, the contribution of the data to the likelihood cancels out in Equation 8 (because it is equal in both models). Thus the ratio used for estimation, $\widetilde{P}(\mathbf{G}\boldsymbol{\Theta}|\mathcal{M}_{ref})/P(\mathbf{G}\boldsymbol{\Theta}|\mathcal{M})$, is not a direct function of the data ($\mathbf{X}$), and the data affects the estimate only through its influence on the sampled instances $\{\mathbf{G}\boldsymbol{\Theta}^{(i)}\}$. We refer to the ratio in Equation 6 as the *relative Bayes factor (RBF) ratio*, and employ it as a model selection criteria by comparing RBFs of competing hypothesis models, calculated using the same reference model -

$$\frac{1}{\text{BF}(\mathcal{M}_i:\mathcal{M}_{ref}|\mathbf{X})} > \frac{1}{\text{BF}(\mathcal{M}_j:\mathcal{M}_{ref}|\mathbf{X})} \Rightarrow P(\mathbf{X}|\mathcal{M}_j) > P(\mathbf{X}|\mathcal{M}_i)$$

Importantly, the variance of the RBF depends on the definition of the model-pairing conditional, $\widetilde{P}$, and it will typically decrease as $\mathcal{M}$ and $\mathcal{M}_{ref}$ become more similar. For instance, in the trivial case where $\mathcal{M}_{ref} = M$, we can define $\widetilde{P}(\mathbf{G}\boldsymbol{\Theta}|\mathcal{M}_{ref}) = P(\mathbf{G}\boldsymbol{\Theta}|\mathcal{M})$ and the RBF ratio becomes 1 for all instances $\{\mathbf{G}\boldsymbol{\Theta}^{(i)}\}$. This is the key advantage of direct estimation of the Bayes factor, when compared to estimation via harmonic mean. Realizing this advantage requires construction of an effective model-pairing conditional distribution for $\mathcal{M}$ and $\mathcal{M}_{ref}$. The following sections present specific constructions for $\widetilde{P}$ in a series of cases.

2.3 The null reference model $\mathcal{M}_0$

We start by considering a simple case where $\mathcal{M}$ is a demographic model with no migration bands and $\mathcal{M}_{ref}$ is the simplest possible model with a single population p_0 of constant size θ_0. We refer to this simple one-parameter model as the *null reference model* $\mathcal{M}_0$ (Figure 3). The first step of constructing a model-pairing conditional for the two models is to identify a mapping F from the space of hidden variables in $\mathcal{M}$ to the space of hidden variables in $\mathcal{M}_0$. In our case, denote by $\widetilde{\mathbf{G}}$ and $\widetilde{\boldsymbol{\Theta}}$ the hidden variables of $\mathcal{M}_0$. Since both $\mathcal{M}$ and $\mathcal{M}_0$ have no migration bands, we may assume that the genealogical information used by both models is the same, implying a natural one-to-one mapping between $\mathbf{G}$ and $\widetilde{\mathbf{G}}$ (the implications of migration are discussed in the next subsection). A mapping between $\boldsymbol{\Theta} = (\boldsymbol{\tau}, \boldsymbol{\theta})$ and $\widetilde{\boldsymbol{\Theta}} = (\theta_0)$ can be defined by selecting one of the population size parameters in $\boldsymbol{\Theta}$ to be associated with θ_0. This can be the size of the root population, θ_{root}, or any other population that we expect to best represent the single population in $\mathcal{M}_0$. The model pairing conditional is obtained by applying this mapping and extending it to the unmapped hidden variables, $\mathbf{Z} = (\boldsymbol{\tau}, \boldsymbol{\theta}\backslash\{\theta_{root}\})$, with the use of a conditional distribution, $\widetilde{P}(\mathbf{Z}|\mathbf{G}\boldsymbol{\Theta}\backslash\mathbf{Z})$:

$$\widetilde{P}(\mathbf{G}\boldsymbol{\Theta}|\mathcal{M}_0) = P(\theta_0 = \theta_{root}|\mathcal{M}_0)\ P(\widetilde{\mathbf{G}} = \mathbf{G}|\mathcal{M}_0, \theta_0 = \theta_{root})\ \widetilde{P}(\mathbf{Z}|\mathbf{G}, \theta_{root}). \tag{10}$$

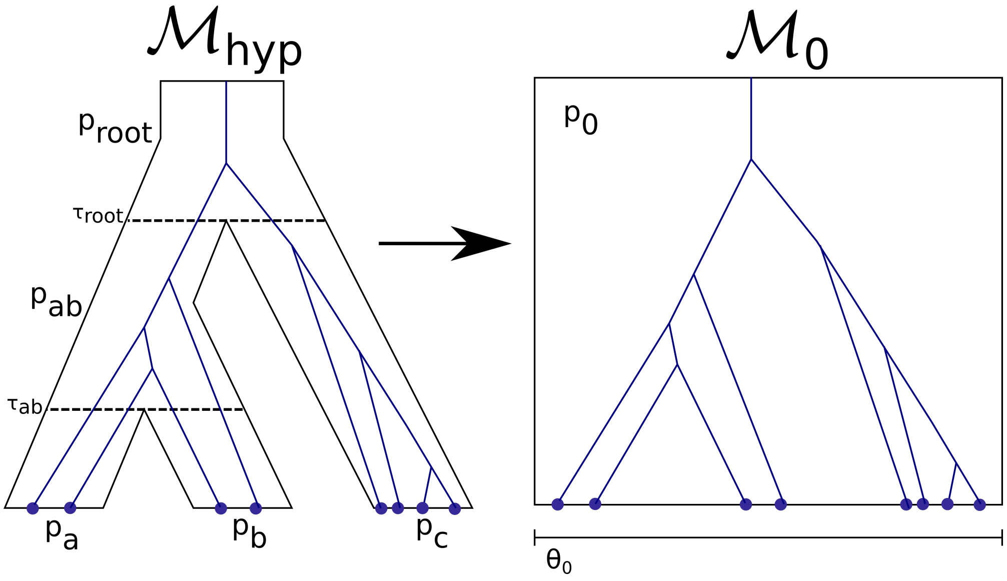


Figure 3. Mapping the hypothesis model $\mathcal{M}_{hyp}$ onto the null reference model $\mathcal{M}_0$. Genealogies are mapped as-is. The null population size θ_0 is defined by associating it with the size of the population in $\mathcal{M}$ which we expect to best represent p_0 (usually θ_{root}). The remaining model parameters are mapped so as to satisfy the model-pairing requirements (4 & 5): Population sizes $\{\theta_{p_a}, \theta_{p_b}, \theta_{p_c}, \theta_{p_{ab}}\}$ are mapped according to their prior probability in $\mathcal{M}_{hyp}$. These have no effect on the reference model structure. Divergence times $\{\tau_{ab}, \tau_{root}\}$ are mapped onto a uniform distribution with upper bound calculated (see Appendix A).

The model-pairing condition of Equation 4 is thus established, regardless of how $\widetilde{P}(\mathbf{Z}|\mathbf{G}, \theta_{root})$ is defined:

$$\begin{aligned}
P(\mathbf{X}|\mathcal{M}_0) = & \int P(\widetilde{\mathbf{\Theta}}|\mathcal{M}_0)\ P(\widetilde{\mathbf{G}}|\mathcal{M}_0, \widetilde{\mathbf{\Theta}})\ P(\mathbf{X}|\widetilde{\mathbf{G}})\ d\widetilde{\mathbf{G}} d\widetilde{\mathbf{\Theta}} \\
= & \int P(\theta_0 = \theta_{root}|\mathcal{M}_0)\ P(\widetilde{\mathbf{G}} = \mathbf{G}|\mathcal{M}_0, \theta_0 = \theta_{root})\ P(\mathbf{X}|\mathbf{G})\ d\mathbf{G} d\theta_{root} \\
= & \int P(\theta_0 = \theta_{root}|\mathcal{M}_0)\ P(\widetilde{\mathbf{G}} = \mathbf{G}|\mathcal{M}_0, \theta_0 = \theta_{root})\ P(\mathbf{X}|\mathbf{G})\ \left(\int \widetilde{P}(\mathbf{Z}|\mathbf{G}, \theta_{root})\ d\mathbf{Z}\right) d\mathbf{G} d\theta_{root} \\
= & \int P(\theta_0 = \theta_{root}|\mathcal{M}_0)\ P`(\widetilde{\mathbf{G}} = \mathbf{G}|\mathcal{M}_0, \theta_0 = \theta_{root})\ \widetilde{P}(\mathbf{Z}|\mathbf{G}, \theta_{root})\ P(\mathbf{X}|\mathbf{G})\ d\mathbf{G}\mathbf{\Theta} \\
= & \int \widetilde{P}(\mathbf{G}\mathbf{\Theta}|\mathcal{M}_0)\ P(\mathbf{X}|\mathbf{G})\ d\mathbf{G}\mathbf{\Theta}\ . \qquad (11)
\end{aligned}$$

We are left to construct $\widetilde{P}(\mathbf{Z}|\mathbf{G}, \theta_{root})$ so that it ensures the model-pairing condition of Equation 5, and we wish to use the remaining degree of freedom to minimize the variance of the RBF ratio. Equation 5 is guaranteed by constricting $\widetilde{P}(\mathbf{Z}|\mathbf{G}, \theta_{root})$ to have zero values whenever $P(\mathbf{G}, \theta_{root}, \mathbf{Z}|\mathcal{M}, \mathbf{X}) = 0$. Among the unmapped variables $\mathbf{Z} = (\boldsymbol{\tau}, \boldsymbol{\theta}\backslash\{\theta_{root}\})$, the population size parameters $\boldsymbol{\theta}\backslash\{\theta_{root}\}$ do not pose any restrictions on the mapped variables $\mathbf{G}, \theta_{root}$. This means that Equations 5 is guaranteed regardless of how their marginal distribution is defined. We thus define their conditional probability distribution according to their prior probability in $\mathcal{M}$, to cancel out terms in the RBF ratio and reduce its variance.

$$\begin{aligned}\frac{\widetilde{P}(\mathbf{G\Theta}|\mathcal{M}_0)}{P(\mathbf{G\Theta}|\mathcal{M})} &= \frac{P(\theta_0=\theta_{root}|\mathcal{M}_0)\,P(\mathbf{G}|\mathcal{M}_0,\theta_0=\theta_{root})\,\widetilde{P}(\mathbf{Z}|\mathbf{G},\theta_{root})}{P(\mathbf{G\Theta}|\mathcal{M})}\\ &= \frac{P(\mathbf{G}|\mathcal{M}_0,\theta_0=\theta_{root})}{P(\mathbf{G}|\mathcal{M},\mathbf{\Theta})}\;\frac{P(\theta_0=\theta_{root}|\mathcal{M}_0)\prod_{p\neq\theta_{root}}\widetilde{P}(\theta_p|\mathbf{G},\theta_{root})}{P(\theta_{root}|\mathcal{M})\prod_{p\neq\theta_{root}}P(\theta_p|\mathcal{M})}\;\frac{\widetilde{P}(\boldsymbol{\tau}|\mathbf{G},\boldsymbol{\theta})}{P(\boldsymbol{\tau}|\mathcal{M})}\\ &= \frac{P(\mathbf{G}|\mathcal{M}_0,\theta_0=\theta_{root})}{P(\mathbf{G}|\mathcal{M},\mathbf{\Theta})}\;\frac{P(\theta_0=\theta_{root}|\mathcal{M}_0)}{P(\theta_{root}|\mathcal{M})}\;\frac{\widetilde{P}(\boldsymbol{\tau}|\mathbf{G},\boldsymbol{\theta})}{P(\boldsymbol{\tau}|\mathcal{M})}\,. \end{aligned} \tag{12}$$

Note that if we assume that $\mathcal{M}$ and $\mathcal{M}_0$ use the same prior distribution over θ_{root} and θ_0 (resp.), then the middle term in Equation 12 also cancels out. We cannot similarly define $\widetilde{P}(\boldsymbol{\tau}|\mathbf{G},\boldsymbol{\theta}) = P(\boldsymbol{\tau}|\mathcal{M})$, because this may lead to conflicts between divergence times and coalescence times in $\mathbf{G}$, which result in violation of the model-pairing condition of Equation 5. Such conflicts occur when a divergence time τ_p is deeper than the most recent common ancestor in $\mathbf{G}$ of two individuals that are each a descendant of a different daughter population of population p. Thus, the final step of constructing $\widetilde{P}(\mathbf{G\Theta}|\mathcal{M}_{ref})$ is to construct $\widetilde{P}(\boldsymbol{\tau}|\mathbf{G},\boldsymbol{\theta}) = \widetilde{P}(\boldsymbol{\tau}|\mathbf{G})$ to have zero values whenever $P(\mathbf{G}|\mathcal{M},\boldsymbol{\tau},\boldsymbol{\theta}) = 0$. This guarantee is achieved by computing for each τ_p an upper bound based on the coalescent events in $\mathbf{G}$ and defining $\widetilde{P}(\boldsymbol{\tau}|\mathbf{G})$ as a product of uniform distributions in the feasible ranges of $\boldsymbol{\tau}$ (see Appendix A for complete derivation and proof).

2.4 Models with gene flow

Assume now that the reference model is still the null model, $\mathcal{M}_0$, but the model of interest, $\mathcal{M}$, has a non-empty set of migration bands, B, associated with migration rates, $\mathbf{m} = \{m_b : b \in B\}$. Migrations complicate the mapping between $\mathcal{M}$ and $\mathcal{M}_0$ because the genealogies in $\mathcal{M}$ hold information about migration events, but the genealogies in $\mathcal{M}_0$ do not (Figure 4).

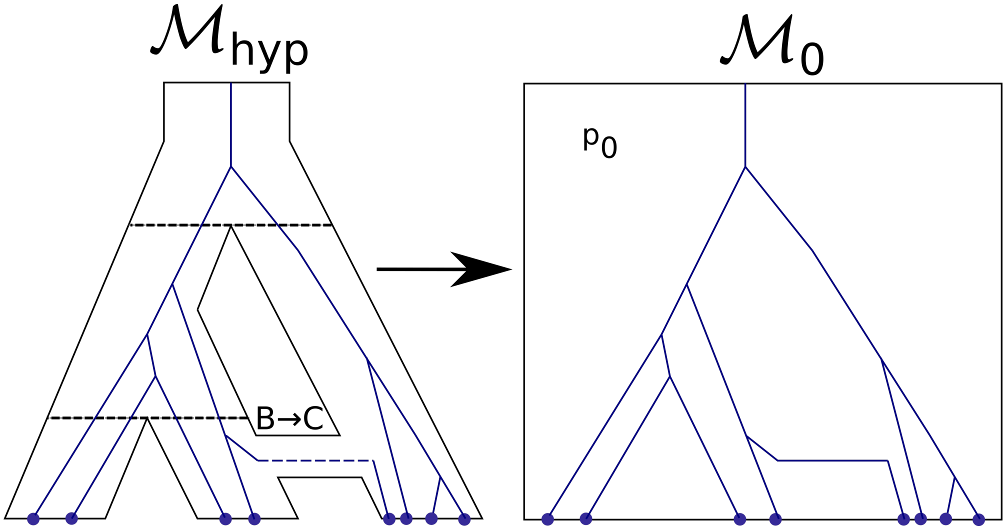


Figure 4. Mapping a model with migration onto the null reference model. Genealogies in p_0 do not hold information about migration events. A complex interplay between migration events and coalescence times makes defining the conditional probability distribution $\widetilde{P}$ challenging. Appendix B specifies the generative process used to address this.

For a sequence of local genealogies $\mathbf{G}$ in $\mathcal{M}$, denote by $\mathbf{G}_c$ the coalescent trees implied by $\mathbf{G}$ and denote by $\mathbf{G}_m$ the information on migration events in $\mathbf{G}$ (locus, timing of event, branch in $\mathbf{G}_c$, source and target populations). Thus, a mapping between the hidden variables of $\mathcal{M}$ ($\mathbf{G}_c, \mathbf{G}_m, \mathbf{\Theta}$) and the hidden variables of $\mathcal{M}_0$ ($\widetilde{\mathbf{G}}, \theta_0$) can be defined by mapping $\mathbf{G}_c$ to $\widetilde{\mathbf{G}}$ and mapping some $\theta_{root} \in \mathbf{\Theta}$ to θ_0. Consequently, the set of unmapped hidden variables is $\mathbf{Z} = (\mathbf{G}_m, \boldsymbol{\tau}, \mathbf{m}, \boldsymbol{\theta} \backslash \{\theta_{root}\})$. This implies a slight modification of the model-pairing conditional specified in Equation 10:

$$\widetilde{P}(\mathbf{G\Theta}|\mathcal{M}_0) = \; P(\theta_0 = \theta_{root}|\mathcal{M}_0)\; P(\widetilde{\mathbf{G}} = \mathbf{G}_c|\mathcal{M}_0, \theta_0 = \theta_{root})\; \widetilde{P}(\mathbf{Z}|\mathbf{G}_c, \theta_{root})\;. \tag{13}$$

The model-pairing condition of Equation 4 can be confirmed by following a sequence of equalities similar to the ones we derived for the scenario without migration (see Equation 11). We are thus left to specify the conditional distribution $\widetilde{P}(\mathbf{Z}|\mathbf{G}_c, \theta_{root})$ to ensure that all $\mathbf{G\Theta}$ for which $P(\mathbf{G}_c, \theta_{root}, \mathbf{Z}|\mathcal{M}, \mathbf{X}) = 0$ also satisfy $\widetilde{P}(\mathbf{Z}|\mathbf{G}_c, \theta_{root}) = 0$. Since the genealogy trees $\mathbf{G}_c$ do not restrict the population size and migration rate parameters, we may define the conditional probability for these parameters based on their prior probability under $\mathcal{M}$, so that their terms cancel out in the RBF ratio:

$$\begin{aligned}
\frac{\widetilde{P}(\mathbf{G\Theta}|\mathcal{M}_0)}{P(\mathbf{G\Theta}|\mathcal{M})} &= \frac{P(\theta_0 = \theta_{root}|\mathcal{M}_0)\; P(\widetilde{\mathbf{G}} = \mathbf{G}_c|\mathcal{M}_0, \theta_0 = \theta_{root})\; \widetilde{P}(\mathbf{Z}|\mathbf{G}_c, \theta_{root})}{P(\mathbf{G\Theta}|\mathcal{M})} \\
&= \frac{P(\mathbf{G}_c|\mathcal{M}_0, \theta_0 = \theta_{root})}{P(\mathbf{G}_c, \mathbf{G}_m|\mathcal{M}, \mathbf{\Theta})}\; \frac{P(\theta_0 = \theta_{root}|\mathcal{M}_0) \prod_{p \neq root} \widetilde{P}(\theta_p|\mathbf{G}_c, \theta_{root})\; \prod_b \widetilde{P}(m_b|\mathbf{G}_c, \theta_{root})}{P(\theta_{root}|\mathcal{M}) \prod_{p \neq \theta_{root}} P(\theta_p|\mathcal{M})\; \prod_b P(m_b|\mathcal{M})}\; \frac{\widetilde{P}(\boldsymbol{\tau}, \mathbf{G}_m|\mathbf{G}_c)}{P(\boldsymbol{\tau}|\mathcal{M})} \\
&= \frac{P(\mathbf{G}_c|\mathcal{M}_0, \theta_0 = \theta_{root})}{P(\mathbf{G}_c, \mathbf{G}_m|\mathcal{M}, \mathbf{\Theta})}\; \frac{P(\theta_0 = \theta_{root}|\mathcal{M}_0)}{P(\theta_{root}|\mathcal{M})}\; \frac{\widetilde{P}(\boldsymbol{\tau}, \mathbf{G}_m|\mathbf{G}_c)}{P(\boldsymbol{\tau}|\mathcal{M})}\;.
\end{aligned} \tag{14}$$

As in the case without migration, we are left to define the conditional probability distribution over the restricting hidden variables, which are in this case the divergence times $\boldsymbol{\tau}$ and the migration events $\mathbf{G}_m$. The complex dependence between divergence times and migration events makes this particularly challenging. For instance, a migration event between populations p_1 and p_2 at time t implies that the divergence times of all populations ancestral to p_1 and p_2 is at least t, but at the same time this migration event may also relax the upper bound of these divergence times. Thus, bounds on divergence times cannot be determined solely based on $\mathbf{G}_c$, and the conditional $\widetilde{P}(\boldsymbol{\tau}, \mathbf{G}_m|\mathbf{G}_c)$ cannot be factored into a product of two separate probability distributions for $\boldsymbol{\tau}$ and $\mathbf{G}_m$. In Appendix B we present a specification for the joint conditional distribution $\widetilde{P}(\boldsymbol{\tau}, \mathbf{G}_m|\mathbf{G}_c)$, which addresses this complex dependence and ensures that $\widetilde{P}(\boldsymbol{\tau}, \mathbf{G}_m|\mathbf{G}_c) = 0$ whenever $P(\boldsymbol{\tau}, \mathbf{G}_m, \mathbf{G}_c|\mathcal{M}) = 0$. This construction results in additional terms canceling out with terms in the genealogy likelihood $P(\mathbf{G}_c, \mathbf{G}_m|\mathcal{M}, \mathbf{\Theta})$, to further reduce the variance of the RBF ratio.

2.5 The comb reference model

The null model has the unique advantage of being a valid reference for the comparison of any two models. This advantage, however, comes at the cost of collapsing all population structure. In many cases we know the population designation of the sampled individuals, and model uncertainty is restricted to the relationships between the sampled populations. To capture this simple structure we use a population phylogeny with a single ancestral population splitting simultaneously into all sampled populations. We refer to such reference models as *comb* models and denote them by $\mathcal{M}_{\sqcap\!\sqcap}$, due to the comb-like structure of the population phylogeny (Figure 5). A comb model is defined by: (1) a set of sampled (leaf) populations, L; (2) an ancestral population, $comb$; and (3) a set of migration bands B_L between populations in L. The resulting demographic model, $\mathcal{M}_{\sqcap\!\sqcap}(L, B_L)$, has $|B_L| + |L| + 2$ parameters: $\widetilde{\mathbf{\Theta}} = (\tau_{comb}, \widetilde{\boldsymbol{\theta}}, \widetilde{\mathbf{m}})$, where $\widetilde{\boldsymbol{\theta}} = \{\theta_p : p \in L \cup \{comb\}\}$ and $\widetilde{\mathbf{m}} = \{m_b : b \in B_L\}$.

Consider a demographic model, $\mathcal{M}(\mathcal{T}, B)$, and its corresponding comb model, $\mathcal{M}_{\sqcap\!\sqcap}(L, B_L)$, defined by $L = leaves(\mathcal{T})$ and $B_L = B \cap (L \times L)$. For brevity, we refer to $\mathcal{M}_{\sqcap\!\sqcap}(L, B_L)$ simply as $\mathcal{M}_{\sqcap\!\sqcap}$. The model-pairing conditional distribution for $\mathcal{M}$ and $\mathcal{M}_{\sqcap\!\sqcap}$ is constructed by first defining a mapping between

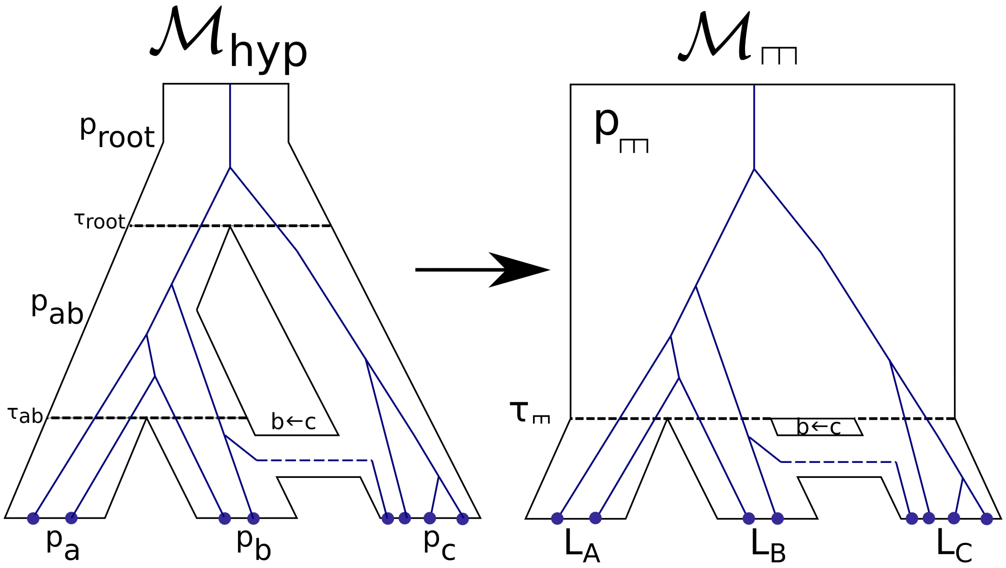


Figure 5. A mapping from hypothesis $\mathcal{M}_{hyp}$ onto the reference model $\mathcal{M}_{comb}$. Genealogies and model parameters above $min(\tau)$ are mapped according to the mapping into the null model (Subsection 2.3) and those below $min(\tau)$ are mapped as is. The remaining parameter p_{comb} can be freely mapped in order to improve RBF estimation.

the hidden variables of $\mathcal{M}$ ($\mathbf{G\Theta}$) and the hidden variables of $\mathcal{M}_{\sqcap\sqcap}$ ($\widetilde{\mathbf{G}}\widetilde{\mathbf{\Theta}}$). This mapping is derived from the requirement that below the comb divergence time (τ_{comb}) the comb model is identical to $\mathcal{M}$ and above it $\mathcal{M}_{\sqcap\sqcap}$ is identical to the null model $\mathcal{M}_0$. We thus set $\tau_{comb} = \tau_{\text{min}} \overset{\Delta}{=} \min(\boldsymbol{\tau})$, to guarantee that all population divergence events in $\mathcal{M}$ map to the comb population in $\mathcal{M}_{\sqcap\sqcap}$. The migration rates of bands in $\mathbf{B} \cap (L \times L)$ and effective sizes of populations in L are mapped into their counterparts in $\widetilde{\mathbf{\Theta}}$, and following the mapping for the null model, a single ancestral population size parameter (θ_{root}) is chosen to be mapped into θ_{comb}. We denote the set of mapped migration rate and population size parameters of $\mathcal{M}$ collectively as $\mathbf{\Theta}_{\sqcap\sqcap}$. Mapping between genealogies is obtained by removing from $\mathbf{G}$ all migration events above time τ_{min}. The resulting collection of local genealogies are denoted by $\mathbf{G}_{\sqcap\sqcap}$ and are directly mapped to $\widetilde{\mathbf{G}}$. The remaining unmapped hidden variables ($\mathbf{Z}$) of $\mathcal{M}$ consist of the following components:

1. Unmapped population size parameters: $\{\theta_p : p \notin L \cup \{root\}\ \}$.

2. Unmapped migration rate parameters: $\{m_b : b \notin L \times L\}$.

3. The identity of the ancestral population in $\mathcal{T}$ with minimum divergence time: $minAncPop = \text{argmin}(\boldsymbol{\tau})$. Note that this population may be *any ancestral population with two leaf daughters*, and its identity is lost when mapping $\boldsymbol{\tau}$ into τ_{comb}.

4. The divergence times of all other populations: $\{\tau_p : p \neq minAncPop\}$.

5. Information on all migration events in $\mathbf{G}$ above time τ_{comb}, which we denote by $\mathbf{G}_{m|>\tau_{\text{min}}}$.

A model-pairing conditional distribution for $\mathcal{M}$ and $\mathcal{M}_{\sqcap\sqcap}$ is thus established by applying the mapping described above and specifying a conditional distribution over the unmapped parameters, $\widetilde{P}(\mathbf{Z}|\mathbf{G}_{\sqcap\sqcap}, \mathbf{\Theta}_{\sqcap\sqcap}, \tau_{\text{min}})$.

The proof of the condition in Equation 4 is given below:

$$\widetilde{P}(\mathbf{G}\mathbf{\Theta}|\mathcal{M}_{\sqcap\!\sqcap}) = P(\widetilde{\mathbf{\Theta}} = (\mathbf{\Theta}_{\sqcap\!\sqcap}, \tau_{\text{min}})|\mathcal{M}_{\sqcap\!\sqcap})\ P(\widetilde{\mathbf{G}} = \mathbf{G}_{\sqcap\!\sqcap}|\mathcal{M}_{\sqcap\!\sqcap}, \mathbf{\Theta}_{\sqcap\!\sqcap}, \tau_{\text{min}})\ \widetilde{P}(\mathbf{Z}|\mathbf{G}_{\sqcap\!\sqcap}, \mathbf{\Theta}_{\sqcap\!\sqcap}, \tau_{\text{min}})\ . \tag{15}$$

$$\begin{aligned}
P(\mathbf{X}|\mathcal{M}_{\sqcap\!\sqcap}) &= \int P(\widetilde{\mathbf{\Theta}}|\mathcal{M}_{\sqcap\!\sqcap})\ P(\widetilde{\mathbf{G}}|\mathcal{M}_{\sqcap\!\sqcap}, \widetilde{\mathbf{\Theta}})\ P(\mathbf{X}|\widetilde{\mathbf{G}})\ d\widetilde{\mathbf{G}}\widetilde{\mathbf{\Theta}} \\
&= \int P(\widetilde{\mathbf{\Theta}} = (\mathbf{\Theta}_{\sqcap\!\sqcap}, \tau_{\text{min}})|\mathcal{M}_{\sqcap\!\sqcap})\ P(\widetilde{\mathbf{G}} = \mathbf{G}_{\sqcap\!\sqcap}|\mathcal{M}_{\sqcap\!\sqcap}, \mathbf{\Theta}_{\sqcap\!\sqcap}, \tau_{\text{min}})\ P(\mathbf{X}|\mathbf{G}_{\sqcap\!\sqcap})\ \ d\mathbf{G}_{\sqcap\!\sqcap}\mathbf{\Theta}_{\sqcap\!\sqcap}\tau_{\text{min}} \\
&= \int P(\widetilde{\mathbf{\Theta}} = (\mathbf{\Theta}_{\sqcap\!\sqcap}, \tau_{\text{min}})|\mathcal{M}_{\sqcap\!\sqcap})\ P(\widetilde{\mathbf{G}} = \mathbf{G}_{\sqcap\!\sqcap}|\mathcal{M}_{\sqcap\!\sqcap}, \mathbf{\Theta}_{\sqcap\!\sqcap}, \tau_{\text{min}})\ P(\mathbf{X}|\mathbf{G}_{\sqcap\!\sqcap})\left(\int \widetilde{P}(\mathbf{Z}|\mathbf{G}_{\sqcap\!\sqcap}, \mathbf{\Theta}_{\sqcap\!\sqcap}, \tau_{\text{min}})d\mathbf{Z}\right) d\mathbf{G}_{\sqcap\!\sqcap}\mathbf{\Theta}_{\sqcap\!\sqcap}\tau_{\text{min}} \\
&= \int P(\widetilde{\mathbf{\Theta}} = (\mathbf{\Theta}_{\sqcap\!\sqcap}, \tau_{\text{min}})|\mathcal{M}_{\sqcap\!\sqcap})\ P(\widetilde{\mathbf{G}} = \mathbf{G}_{\sqcap\!\sqcap}|\mathcal{M}_{\sqcap\!\sqcap}, \mathbf{\Theta}_{\sqcap\!\sqcap}, \tau_{\text{min}})\ \widetilde{P}(\mathbf{Z}|\mathbf{G}_{\sqcap\!\sqcap}, \mathbf{\Theta}_{\sqcap\!\sqcap}, \tau_{\text{min}})\ P(\mathbf{X}|\mathbf{G})\ d\mathbf{G}\mathbf{\Theta} \\
&= \int \widetilde{P}(\mathbf{G}\mathbf{\Theta}|\mathcal{M}_0)\ P(\mathbf{X}|\mathbf{G})\ d\mathbf{G}\mathbf{\Theta}\ .
\end{aligned} \tag{16}$$

The conditional distribution $\widetilde{P}(\mathbf{Z}|\mathbf{G}_{\sqcap\!\sqcap}, \mathbf{\Theta}_{\sqcap\!\sqcap}, \tau_{\text{min}})$ is defined similar to its specification in the null model. The unmapped population size and migration rate parameters are distributed according to their prior probability under $\mathcal{M}$ to eliminate terms in the RBF ratio. The identity of the minimal ancestral population, $minAncPop$, is distributed uniformly among all ancestral populations in $\mathcal{T}$ with two leaf daughters. We denote the number of such populations in $\mathcal{T}$ by $\kappa(\mathcal{T})$. The only unmapped variables restricted by $\mathbf{G}_{\sqcap\!\sqcap}$ and τ_{min} are the unmapped divergence times and migration events above time τ_{min}. Their conditional distribution, $\widetilde{P}(\boldsymbol{\tau}\backslash\{\tau_{\text{min}}\}, \mathbf{G}_{m|>\tau_{\text{min}}}|\mathbf{G}_c)$, is defined using the process described for the null model (see Appendices A and B). This specification thus guarantees the condition of Equation 5, as in the case of the null reference model. The resulting RBF ratio is expressed as follows:

$$\begin{aligned}
\frac{\widetilde{P}(\mathbf{G}\mathbf{\Theta}|\mathcal{M}_{\sqcap\!\sqcap})}{P(\mathbf{G}\mathbf{\Theta}|\mathcal{M})} &= \frac{P(\widetilde{\mathbf{\Theta}} = (\mathbf{\Theta}_{\sqcap\!\sqcap}, \tau_{\text{min}})|\mathcal{M}_{\sqcap\!\sqcap})\ P(\widetilde{\mathbf{G}} = \mathbf{G}_{\sqcap\!\sqcap}|\mathcal{M}_{\sqcap\!\sqcap}, \mathbf{\Theta}_{\sqcap\!\sqcap}, \tau_{\text{min}})\ \widetilde{P}(\mathbf{Z}|\mathbf{G}_{\sqcap\!\sqcap}, \mathbf{\Theta}_{\sqcap\!\sqcap}, \tau_{\text{min}})}{P(\mathbf{G}\mathbf{\Theta}|\mathcal{M})} \\
&= \frac{P(\widetilde{\mathbf{G}} = \mathbf{G}_{\sqcap\!\sqcap}|\mathcal{M}_{\sqcap\!\sqcap}, \mathbf{\Theta}_{\sqcap\!\sqcap}, \tau_{\text{min}})}{P(\mathbf{G}|\mathcal{M}, \mathbf{\Theta})}\ \frac{P(\widetilde{\mathbf{\Theta}} = (\mathbf{\Theta}_{\sqcap\!\sqcap}, \tau_{\text{min}})|\mathcal{M}_{\sqcap\!\sqcap})}{P(\mathbf{\Theta}_{\sqcap\!\sqcap}|\mathcal{M})}\ \frac{\frac{1}{\kappa(\mathcal{T})}\widetilde{P}(\boldsymbol{\tau}\backslash\{\tau_{\text{min}}\}, \mathbf{G}_{m|>\tau_{\text{min}}}|\mathbf{G}_c)}{P(\boldsymbol{\tau}|\mathcal{M})}\ .
\end{aligned} \tag{17}$$

As in the case of the null reference model, the above RBF ratio has several terms canceling out. First, the conditional probabilities of the unmapped population size and migration rate parameters cancel out with their priors under $\mathcal{M}$. Second, if we assume identical priors in both models for the mapped parameters, then these cancel out as well in the second term of Equation 17. Terms in the genealogy likelihood contributed by migration events above time τ_{min} also cancel out in the ratio (see Appendix B). Finally, the contribution of all events below time τ_{min} (coalescence and migration) also cancel out. If we denote the portion of $\mathbf{G}$ below time τ_{min} by $\mathbf{G}_{<\tau_{\text{min}}}$, and the portion above it by $\mathbf{G}_{>\tau_{\text{min}}}$, then the contribution of $\mathbf{G}_{<\tau_{\text{min}}}$ to the first term of the RBF ratio cancels out as follows:

$$\begin{aligned}
\frac{P(\mathbf{G}_{\sqcap\!\sqcap}|\mathcal{M}_{\sqcap\!\sqcap}, \mathbf{\Theta}_{\sqcap\!\sqcap}, \tau_{\text{min}})}{P(\mathbf{G}|\mathcal{M}, \mathbf{\Theta})} &= \frac{P(\mathbf{G}_{{\sqcap\!\sqcap}_{<\tau_{\text{min}}}}|\mathcal{M}_{\sqcap\!\sqcap}, \mathbf{\Theta}_{\sqcap\!\sqcap}, \tau_{\text{min}})P(\mathbf{G}_{{\sqcap\!\sqcap}_{>\tau_{\text{min}}}}|\mathcal{M}_{\sqcap\!\sqcap}, \mathbf{\Theta}_{\sqcap\!\sqcap}, \tau_{\text{min}})}{P(\mathbf{G}_{<\tau_{\text{min}}}|\mathcal{M}, \mathbf{\Theta})P(\mathbf{G}_{>\tau_{\text{min}}}|\mathcal{M}, \mathbf{\Theta})} \\
&= \frac{P(\mathbf{G}_{<\tau_{\text{min}}}|\mathcal{M}_{\sqcap\!\sqcap}, \mathbf{\Theta}_{\sqcap\!\sqcap},\ \tau_{comb} = \tau_{\text{min}})}{P(\mathbf{G}_{<\tau_{\text{min}}}|\mathcal{M}, \mathbf{\Theta}_{\sqcap\!\sqcap},\ \min(\boldsymbol{\tau}) = \tau_{\text{min}})}\ \frac{P(\mathbf{G}_{c|>\tau_{\text{min}}}|\mathcal{M}_{\sqcap\!\sqcap}, \theta_{comb} = \theta_{root})}{P(\mathbf{G}_{>\tau_{\text{min}}}|\mathcal{M}, \mathbf{\Theta})} \\
&= \frac{P(\mathbf{G}_{c|>\tau_{\text{min}}}|\mathcal{M}_0, \theta_0 = \theta_{root})}{P(\mathbf{G}_{>\tau_{\text{min}}}|\mathcal{M}, \mathbf{\Theta})}\ .
\end{aligned} \tag{18}$$

The RBF may thus be re-expressed as follows:

$$\frac{\widetilde{P}(\mathbf{G}\mathbf{\Theta}|\mathcal{M}_{\sqcap\!\sqcap})}{P(\mathbf{G}\mathbf{\Theta}|\mathcal{M})} = \frac{1}{\kappa(\mathcal{T})}\ \frac{\widetilde{P}(\mathbf{G}_{>\tau_{\text{min}}},\ \mathbf{\Theta}\backslash\{\tau_{\text{min}}\}\ |\ \mathcal{M}_0)}{P(\mathbf{G}_{>\tau_{\text{min}}},\ \mathbf{\Theta}\backslash\{\tau_{\text{min}}\}\ |\ \mathcal{M})}\ \frac{P(\widetilde{\mathbf{\Theta}}\backslash\{\theta_{comb}\} = (\mathbf{\Theta}_{\sqcap\!\sqcap}\backslash\{\theta_{root}\}, \tau_{\text{min}})|\mathcal{M}_{\sqcap\!\sqcap})}{P(\mathbf{\Theta}_{\sqcap\!\sqcap}\backslash\{\theta_{root}\}|\mathcal{M})}\ . \tag{19}$$

2.6 Constructing a reference model

Subsections 2.3-2.5 described two examples of reference models - the null reference model and the comb reference model. During construction of both these reference models, the structure of the entire phylogeny (sans a portion of the leaves in case of a comb mapping) is collapsed into a single population, and a mapping is derived from this. However, in many cases of interest the modeling uncertainty is restricted to a certain subtree in the population phylogeny. In such cases, we wish to consider a reference model where only a subset of the sampled populations is collapsed into a clade or a comb submodel.

In general, a reference model $\mathcal{M}_{ref}$ for hypothesis model $\mathcal{M}$ may be obtained by applying the following three-step process:

1. First, **choose a subtree** of the population phylogeny of $\mathcal{M}$. The subtree is associated with the population p at its root.
2. Then **collapse the subtree structure** into either a clade structure, i.e. a single population p_{clade}, or a comb structure, i.e. an ancestral population p_{comb} and a set of leaf populations and migration bands L, B_L.
3. Finally, **map the hidden parameters** of $\mathcal{M}$ onto parameters of $\mathcal{M}_{ref}$, defining the model-pairing conditional distribution $\widetilde{P}$ such that the necessary conditions (4 & 5) are met. This mapping should cancel-out as many terms of the RBF ratio as possible (equations (14) & (19)).

Identically mapping all structure and parameters outside the subtree of p during step 3 leads to canceling-out of all corresponding terms in the RBF of $\mathcal{M}$ relative to $\mathcal{M}_{ref}$.

3 RBF Computational Scheme

Having defined the concept of reference models and formulated their relative Bayes factors, we now describe the computational scheme we use to estimate RBFs as derived in subsection 2.2:

$$\frac{1}{\text{BF}(\mathcal{M} : \mathcal{M}_{ref}|\mathbf{X})} \approx \frac{1}{N}\sum_{i=1}^{N}\frac{\widetilde{P}(\mathbf{G\Theta}^{(i)}|\mathcal{M}_{ref})}{P(\mathbf{G\Theta}^{(i)}|\mathcal{M})} \tag{20}$$

This RBF is further derived for clade and comb reference models (Equations 14 & 19). We now focus our attention on the components making up the model pairing conditional. Consider for example the RBF derivation for a null reference model in equation 14 -

$$\frac{\widetilde{P}(\mathbf{G\Theta}|\mathcal{M}_0)}{P(\mathbf{G\Theta}|\mathcal{M})} \approx \frac{P(\mathbf{G}_c|\mathcal{M}_0, \theta_0 = \theta_{root})}{P(\mathbf{G}_c, \mathbf{G}_m|\mathcal{M}, \mathbf{\Theta})}\frac{P(\theta_0 = \theta_{root}|\mathcal{M}_0)}{P(\theta_{root}|\mathcal{M})}\frac{\widetilde{P}(\boldsymbol{\tau}, \mathbf{G}_m|\mathbf{G}_c)}{P(\boldsymbol{\tau}|\mathcal{M})}$$

The two denominators $P(\mathbf{G}_c, \mathbf{G}_m|\mathcal{M}, \mathbf{\Theta})$ and $P(\boldsymbol{\tau}|\mathcal{M})$ are calculated as part of the G-PhoCS MCMC flow. During RBF estimation these values are taken as-is from G-PhoCS and utilized as explained in section 3.5. In the derivation we suggest that the parameter priors $P(\theta_{root}|\mathcal{M})$ and $P(\theta_0 = \theta_{root}|\mathcal{M}_0)$ may share the same distribution and thus cancel out. However, in theory and practice, any parameter prior or constant value can be sused as $P(\theta_0|\mathcal{M}_0)$. In such a case, $P(\theta_{root}|\mathcal{M})$ is taken from G-PhoCS as-is and $P(\theta_0|\mathcal{M}_0)$ is plugged into the calculation in section 3.2. The condtional distribution $\widetilde{P}(\boldsymbol{\tau}, \mathbf{G}_m|\mathbf{G}_c)$ is calculated as described in appendices A and B and utilized as described in section 3.5. Lastly, the genealogy likelihood in the reference model $P(\mathbf{G}_c|\mathcal{M}_0, \theta_0 = \theta_{root})$ is calculated from scratch under Kingman's coalescent. We consider this the main component of the model pairing conditional, as it represents the bulk of our computational challenge. The rest of section 3 details its calculation in an efficient manner.

3.1 Maximizing algorithm flexibility

A main objective of our computational scheme is allowing maximal flexibility in choice of reference model, while attaining reasonable algorithm run time and space usage. Since the most time consuming step is the MCMC sampling algorithm, we assume only a single MCMC chain per hypothesis. With this in mind, we note that there exists a clear trade-off between flexibility in choice of reference model and amount of data the MCMC process is required to emit. For example, if the reference model is predetermined before MCMC execution (i.e. no flexibility is required), the RBF ratio can be calculated during MCMC iteration and only the final RBF estimation need be emitted. Unfortunately, this approach would require another full MCMC execution in order to estimate RBF of any other reference model. On the other hand, the RBF for every reference model could be computed in post-processing if the MCMC would print out the full hidden state $\mathbf{G}\mathbf{\Theta}$ in each iteration. This, however, would yield an unreasonable amount of traced information - in proportion to the size of the model and to the number of loci.

Our computational scheme aims to find a reasonable middle ground between these two extremes. Our objective is to maximize the number of reference models we can consider using a single MCMC sampling chain without blowing up the output trace. This is accomplished by identifying a collection of sufficient statistics for $\mathbf{G}$ that satisfy three conditions:

1. The sufficient statistics allow calculation of $P(\mathbf{G}|\mathbf{\Theta}, \mathcal{M}_{ref})$ for a wide variety of reference model structures, i.e. for any model structure obtained by applying a comb or clade collapse operation on an ancestral population.

2. Given a reference model structure, the sufficient statistics allow calculation of $P(\mathbf{G}|\mathbf{\Theta}, \mathcal{M}_{ref})$ for any value of the freely parameter θ_{root}.

3. The number of sufficient statistics depends on the complexity of the hypothesis model $\mathcal{M}_{hyp}$, but not on the size of the data (i.e. the number of individuals and of loci).

We then perform the RBF calculation in two phases. Phase 1, which is performed jointly with the MCMC sampling process, emits intermediate summary statistics which meet the above three conditions. Phase 2 is then given a definition of specific reference model structure and mapping of free reference parameters. This phase assembles the relevant statistics, plugs in the appropriate parameter priors and emits the final estimate of $\frac{1}{\mathrm{BF}(\mathcal{M}:\mathcal{M}_{ref}|\mathbf{X})}$. Phase 2 can be repeatedly rerun with different reference models, utilizing the same sufficient statistics emitted by phase 1, thus calculating RBFs of different reference models.

Subsection 3.2 explains how to calculate sufficient statistics which meet conditions 2 & 3 for a single model structure. Subsections 3.3 and 3.4 attain condition 1 by efficiently extending these statistics to all comb and clade reference models. Later, section 3.5 explains how the intermediate sufficient statistics are combined with other statistics into an RBF estimate for a specific reference model.

3.2 Efficient sufficient statistics for reference model genealogy likelihood

Sufficient statistics that satisfy conditions 2 & 3 are derived from the expression for the genealogy likelihood $P(\mathbf{G}|\mathbf{\Theta}, \mathcal{M}_{ref})$ under Kingman's coalescent, which we briefly recall here. First, because the loci are assumed to be freely recombining, then the local genealogies $\mathbf{G} = (G_1, ...G_L)$ are conditionally independent given the model parameters and the likelihood may be expressed as a product of locus-specific likelihoods, $P(G_l|\mathbf{\Theta}, \mathcal{M}_{ref})$. Each locus-specific likelihood is a product of exponentially distributed waiting times for coalescent and migration events. The rates of these exponential distributions depend on the model parameters (population sizes and migration rates) as well as the number of lineages considered for coalescence and migration. We thus identify for each population the set of coalescent and migration events that change the number of lineages modeled in that population in G_l. Each time interval I between two consecutive events is associated with the following properties:

- $t(I)$ – the elapsed time of the interval.
- $n(I)$ – the number of lineages of $\mathbf{G}_l$ alive during that time in the target population.
- $isCoal(I)$, $isInMig(I)$ – binary values that indicate whether the event above the interval is a coalescent event or incoming migration event (respectively).

The contribution of population p to $P(G_l|\mathbf{\Theta}, \mathcal{M}_{ref})$ can then be expressed as a product over the set of relevant time intervals $\mathcal{I}(p, l)$:

$$f_{coal}(\mathbf{G}_l, p|\mathbf{\Theta}, \mathcal{M}_{ref}) \triangleq \prod_{I \in \mathcal{I}(p,l)} \left(\frac{2}{\theta_p}\right)^{isCoal(I)} \exp\left(-\frac{2}{\theta_p}\binom{n(I)}{2} t(I)\right) . \tag{21}$$

Similarly, the contribution of migration band b to $P(G_l|\mathbf{\Theta}, \mathcal{M}_{ref})$ can be expressed as a product over the set of time intervals $\mathcal{I}(b, l)$ defined by events in the target population of the migration band:

$$f_{mig}(\mathbf{G}_l, b|\mathbf{\Theta}, \mathcal{M}_{ref}) \triangleq \prod_{I \in \mathcal{I}(b,l)} m_b^{isInMig(I)} \exp\left(-m_b\, n(I)\, t(I)\right) . \tag{22}$$

Using these notations, the genealogy log likelihood can be expressed as follows:

$$\begin{aligned}
\ln\left(P(\mathbf{G}|\mathbf{\Theta}, \mathcal{M}_{ref})\right) &= \ln\left(\prod_l P(G_l|\mathbf{\Theta}, \mathcal{M}_{ref})\right) \\
&= \ln\left(\prod_l \left(\prod_p f_{coal}(\mathbf{G}_l, p|\mathbf{\Theta}, \mathcal{M}_{ref}) \prod_b f_{mig}(\mathbf{G}_l, b|\mathbf{\Theta}, \mathcal{M}_{ref})\right)\right) \\
&= \sum_p \sum_l \ln\left(f_{coal}(\mathbf{G}_l, p|\mathbf{\Theta}, \mathcal{M}_{ref})\right) + \sum_b \sum_l \ln\left(f_{mig}(\mathbf{G}_l, b|\mathbf{\Theta}, \mathcal{M}_{ref})\right) . \qquad (23)
\end{aligned}$$

The key to likelihood calculation is to sum over the log-likelihood contributions across time intervals and across loci (Figure 6):

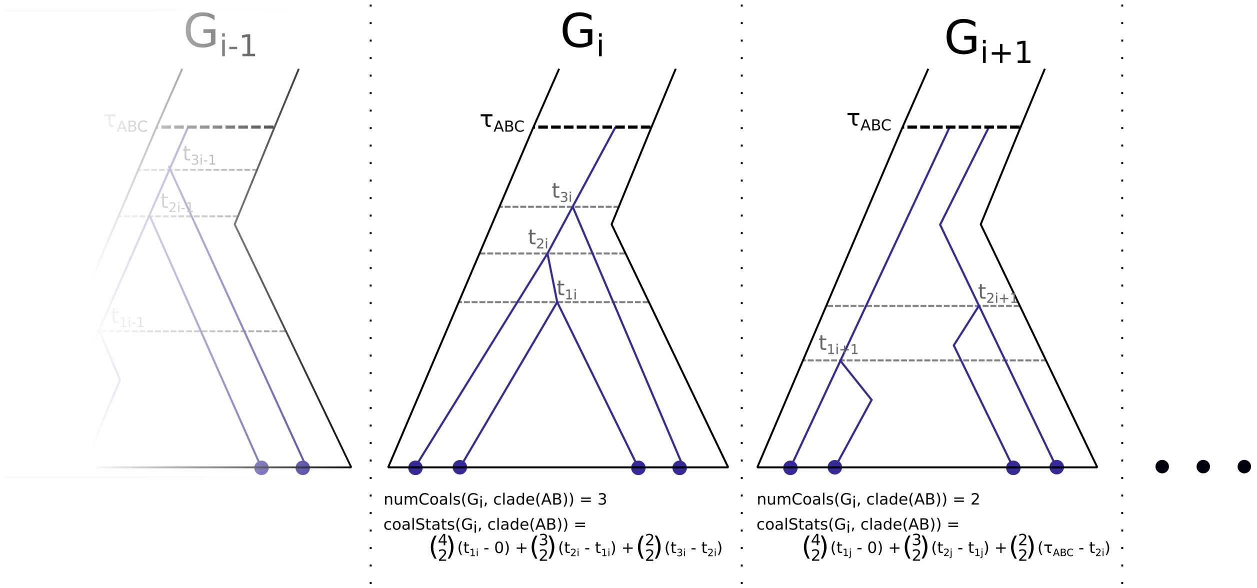
$$\sum_l \ln\left(f_{coal}(\mathbf{G}_l, p|\mathbf{\Theta}, \mathcal{M}_{ref})\right) = \ln\left(\frac{2}{\theta_p}\right) \sum_l \sum_{I \in \mathcal{I}(p,l)} isCoal(I) - \frac{2}{\theta_p} \sum_l \sum_{I \in \mathcal{I}(p,l)} \binom{n(I)}{2} t(I) . \tag{24}$$

$$\sum_l \ln\left(f_{mig}(\mathbf{G}_l, b|\mathbf{\Theta}, \mathcal{M}_{ref})\right) = \ln(m_b) \sum_l \sum_{I \in \mathcal{I}(p,l)} isInMig(I) - m_b \sum_l \sum_{I \in \mathcal{I}(p,l)} n(I) t(I) . \tag{25}$$

Note that the four double sums in these expressions depend on the local genealogies $\mathbf{G}$ and the divergence times $\{\tau_p\}$, but they do not depend on the population size and migration rate parameters. We thus denote these sums respectively as $numCoals(\mathbf{G}, p)$, $coalStats(\mathbf{G}, p)$, $numMigs(\mathbf{G}, b)$, and $migStats(\mathbf{G}, b)$, and the log-likelihood can be expressed as follows:

$$\begin{aligned}
\ln\left(P(\mathbf{G}|\mathbf{\Theta}, \mathcal{M}_{ref})\right) &= \sum_p \ln\left(\frac{2}{\theta_p}\right) \cdot numCoals(\mathbf{G}, p) - \frac{1}{\theta_p} \cdot coalStats(\mathbf{G}, p) \qquad (26) \\
&\quad + \sum_b \ln(m_b) \cdot numMigs(\mathbf{G}, b) - m_b \cdot migStats(\mathbf{G}, b) . \qquad (27)
\end{aligned}$$

The summary statistics $numCoals(\mathbf{G}, p)$, $coalStats(\mathbf{G}, p), numMigs(\mathbf{G}, b) \& migStats(\mathbf{G}, b)$ aggregate all genealogy state information, and postpone the plugging in of parameters θ_p and m_b. This enables computation of $P(\mathbf{G}|\mathbf{\Theta}, \mathcal{M}_{ref})$ for different parameters in a later stage, when settling on a specific free parameter mapping, as specified in our 2nd requirement from the sufficient statistics.



G_{i-1}
G_i
G_{i+1}
τ_{ABC}
t_{3i-1}
t_{2i-1}
t_{1i-1}
τ_{ABC}
t_{3i}
t_{2i}
t_{1i}
τ_{ABC}
t_{2i+1}
t_{1i+1}
numCoals(G_i, clade(AB)) = 3
coalStats(G_i, clade(AB)) =
$\binom{4}{2}(t_{1i} - 0) + \binom{3}{2}(t_{2i} - t_{1i}) + \binom{2}{2}(t_{3i} - t_{2i})$
numCoals(G_i, clade(AB)) = 2
coalStats(G_i, clade(AB)) =
$\binom{4}{2}(t_{1j} - 0) + \binom{3}{2}(t_{2j} - t_{1j}) + \binom{2}{2}(\tau_{ABC} - t_{2i})$

Figure 6. The sufficient statistic $coalStat(G, clade(AB))$ is calculated by accumulating the Kingman Coalescent genealogy log-likelihod across loci. The contribution of each locus is calculated via the set of intervals $\mathcal{I}(clade(AB), l_i)$. The sufficient statistic $numCoals(G, clade(AB))$ is simply the sum across loci of the amount of coalescence events inside $clade(AB)$.

3.3 Sufficient statistics for all clade models

We now consider an example hypothesis and reference setting in order to describe all statistics required in computing a single RBF estimation. The hypothesis model $\mathcal{M}$ has a set of leaf populations A, B, C and ancestral populations AB and ABC, as well as possibly other irrelevant populations. To create the reference model $\mathcal{M}_{C(AB)}$ we collapse the clade under population AB and associate θ_0 with θ_{AB}. A snippet of the clade population is seen in Figure 6. The hypothesis and reference model are identical everywhere outside the AB clade, so to compute the RBF we need only calculate terms inside the clade -

$$\frac{\widetilde{P}(\mathbf{G\Theta}|\mathcal{M}_{C(AB)})}{P(\mathbf{G\Theta}|\mathcal{M})} \approx \frac{P(\mathbf{G}_{<\tau_{ABC}}|\mathcal{M}_{C(AB)}, \theta_0 = \theta_{AB})}{P(\mathbf{G}_{\in A}|\theta_A)P(\mathbf{G}_{\in B}|\theta_B)P(\mathbf{G}_{\in AB}|\theta_{AB})} \frac{P(\theta_0 = \theta_{AB}|\mathcal{M}_0)}{P(\theta_{AB}|\mathcal{M})} \frac{\widetilde{P}(\tau_{AB}|\mathbf{G}_{<\tau_{ABC}})}{P(\tau_{AB}|\mathcal{M})}$$

A similiar derivation can be done for all reference models generated by the reference construction process (subsection 2.6). To support calculating RBF of all these models we must calculate all relevant terms for each reference model. Fortunately, statistics heavily reappear in different RBFs; To fulfill all hypothesis genealogy likelihoods we emit per iteration the genealogy likelihood of each population. These are already calculated during MCMC. This fulfils terms $P(\mathbf{G}_{\in A}|\theta_A), P(\mathbf{G}_{\in B}|\theta_B)$ and $P(\mathbf{G}_{\in AB}|\theta_{AB})$ in the above example. All theta values and theta and tau priors are also emitted in each iteration. Reference tau priors are calculated as described in Appendix A and the rest are readily available from the MCMC process. This fulfils terms $P(\theta_{AB}|\mathcal{M}), P(\theta_0 = \theta_{AB}|\mathcal{M}_0), \widetilde{P}(\tau_{AB}|\mathbf{G}_{<\tau_{ABC}})$ and $P(\tau_{AB}|\mathcal{M})$ in the above example. Finally, sufficient statistics for all possible collapsed clades are emitted - $\{\ numCoals(\mathbf{G}, clade(p)),\ \ coalStats(\mathbf{G}, clade(p))\ \}_p$

To efficiently calculate sufficient statistics for all clades, calculation of $numCoals$ and $coalStats$ is done recursively down the population phylogeny of M as implemented in the pseudo-python code below. This implementation uses a function for computing $coalStats$ given a sorted list of intervals (function calculate_coal_stats), as well as accessors to data from G-PhoCS (functions num_coals_from_gphocs and sorted_intervals_from_gphocs):

```
def recursive_num_coals(pop):
    """recursively calculate and store num of coalescence
    events in clade(pop) as well as all descendant clades"""

    pop_num_coals = num_coals_from_gphocs(pop)

    if is_leaf(pop):
        return pop_num_coals

    left_num_coals = recursive_num_coals(pop.left)
    right_num_coals = recursive_num_coals(pop.right)

    current_num_coals = pop_num_coals + left_num_coals + right_num_coals
    store(current_num_coals)

    return current_num_coals


def recursive_coal_stats(pop):
    """recursively calculate and store coalescence stats
    of clade(pop) as well as all descendant clades"""

    pop_intervals = sorted_intervals_from_gphocs(pop)

    if is_leaf(pop):
        return pop_intervals

    left_intervals = recursive_coal_stats(pop.left)
    right_intervals = recursive_coal_stats(pop.right)
    merged_intervals = merge_sort(left_intervals, right_intervals)

    clade_intervals = merged_intervals.append(pop_intervals)

    clade_coal_stats = calculate_coal_stats(clade_intervals)
    store(clade_coal_stats)

    return clade_intervals
```

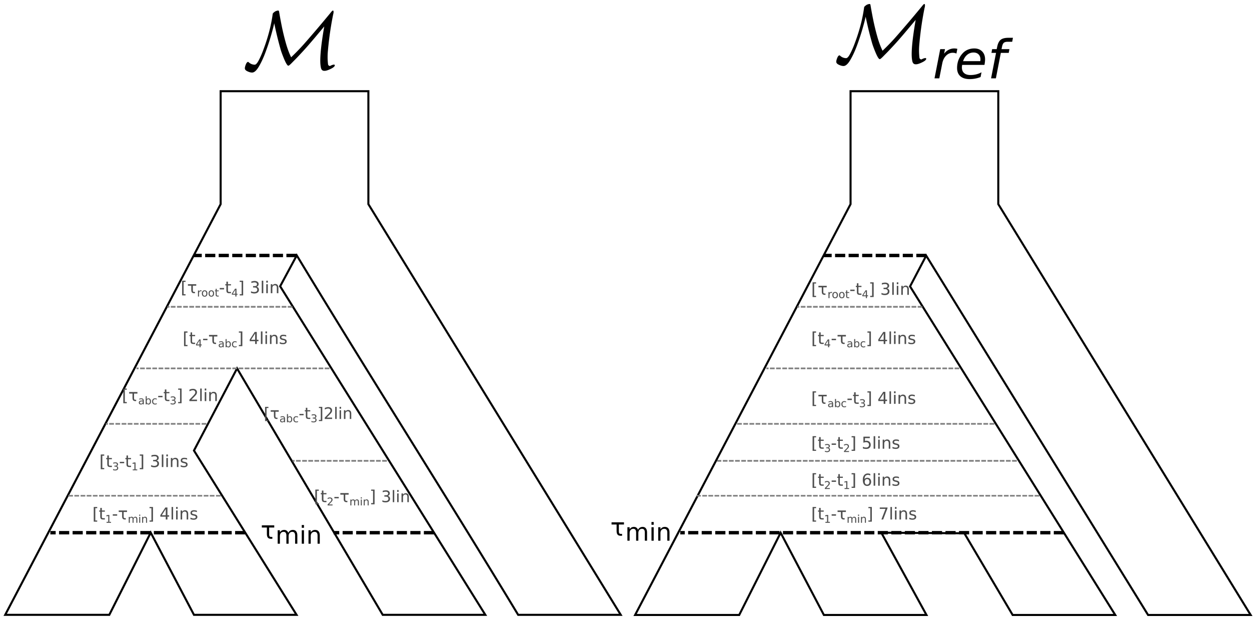
3.4 Recursive Sufficient Statistics for All Comb Models

Equation 18 shows how for a reference model created by comb-collapsing the root population, contribution of the genealogy-likelihood to the model-pairing conditional is reduced to contribution of the portion of genealogies above τ_{min} -

$$\frac{P(\mathbf{G}_{c|>\tau_{\text{min}}}|\mathcal{M}_0, \theta_0 = \theta_{root})}{P(\mathbf{G}_{>\tau_{\text{min}}}|\mathcal{M}, \mathbf{\Theta})}$$

When comb-collapse is applied to a subtree, we apply the same idea to the portion of the genealogy contained in that subtree. Figure 7 illustrates the intervals relevant for genealogy-likelihood calculation in the hypothesis and reference models.

As in the case for clade reference models, we wish to calculate statistics for all viable comb reference models after only one MCMC chain. We do this by storing for every ancestral population p the log of the denominator $ln(P(\mathbf{G}_{>\tau_{\text{min}}}|\mathcal{M}, \mathbf{\Theta}))$ and the two sufficient statistics involved in the calculation of the enumerator - $(\{\ numCoals(\mathbf{G}, comb(p)),\ \ coalStats(\mathbf{G}, comb(p))\ \}_p)$. This is again calculated recursively down the population phylogeny of $\mathcal{M}$, but the function `calculate_coal_stats` now takes into account only intervals inside the subtree of p and above τ_{min}.



$\mathcal{M}$
[τ_{root}-t_4] 3lin
[t_4-τ_{abc}] 4lins
[τ_{abc}-t_3] 2lin
[τ_{abc}-t_3]2lin
[t_3-t_1] 3lins
[t_2-τ_{min}] 3lin
[t_1-τ_{min}] 4lins
τ_{min}
$\mathcal{M}_{ref}$
[τ_{root}-t_4] 3lin
[t_4-τ_{abc}] 4lins
[τ_{abc}-t_3] 4lins
[t_3-t_2] 5lins
[t_2-t_1] 6lins
[t_1-τ_{min}] 7lins
τ_{min}

Figure 7. In comb reference models, genealogy-likelihood need only be calculated strictly within the bounds of the comb population $comb(p)$. Outside this area of the topology, genealogy likelihoods of the two models cancel out in the RBF.

3.5 Finalizing the RBF calculation using McRef

After the MCMC process is completed, we are left with sufficient statistics and parameter priors per iteration for each clade and comb reference model. These are stored in multiple trace files (see example trace snippet in appendix C). The remaining step is to calculate the estimated relative model fit $P(X|\mathcal{M}_{ref})/P(X|\mathcal{M}_{hyp})$ for a chosen reference model (or several). For this purpose we developed the McRef utility.

When setting up McRef, several parameters are configured. The main configurations is the chosen reference model. This is specified by simply stating on what hypothesis population to perform a comb/clade collapse operation. The remaining configuration options pertain to standard I/O (e.g. where the trace data files are stored and where to store output), to G-PhoCS configuration (e.g. what alpha & beta to use for gamma prior, what print multipliers were applied to trace data when emitted by gphocs etc.), to internal statistical calculations (e.g. number of bootstrap iterations for confidence calculation and burn-in and sample-dilution to apply on MCMC traces) and to debugging (e.g. what debug calculations to run and visualizations to emit). See appendix C for an example configuration file.

McRef finishes calculating Equation 27 by plugging in the chosen parameters for each configured reference population. In our current implementation, θ_{root} of the comb/clade population in the reference model is set to the value of theta of the population at the root of the comb/clade in the hypothesis, but this can be easily adjusted if ever we decide to consider other theta mappings. In addition to evaluating RBF, McRef also roughly measures the accuracy of estimation using bootstrapping of traced samples. The bootstrapping algorithm used is a simple monte-carlo case resampling. It's results appear in the bar charts of section 4 as the error interval.

With the goal of optimizing the practical run-time and usability of McRef, several techniques were employed; trace data files, which are repeatedly read and used, are lazily loaded and cached in each mcref process. Multiple McRef processes are launched using a single command-line tool and are cocurrently run on different processors, eventually aggregating summary results to a single log file. See appendix C for an example output.

To clarify results and to help understand and debug McRef runs, several visualizations are emitted. Each McRef run emits plots of the genealogy-log-likelihood for the reference and hypothesis models, as well as a plot of the RBF and harmonic mean estimations across G-PhoCS iterations. Multiple debug plots are also emitted by mcref. Their goal is to help the researcher assert the experiment executed as planned. These plots contain the kingman coalescence and migration likelihoods of every population and migration band in the hypothesis and reference models. They also contain the aggregate coalescence stats of the hypothesis and reference model. See appendix C for example debug graphs. The McRef code resides on Github, along with an installation guide and examples - https://github.com/selotape/McRef .

4 Results

In order to evaluate our new method we designed a series of experiments which measure it's strengths and weaknesses. The experiments were set up to try to distinguish between different hypothesis structures; Experiment 1 tries to decide whether a divergence between populations did or did not occur. Experiment 2 tries to decide the ancestral relationship between three leaf populations and an out-group, and experiment 3 tries to identify the true migration pattern between leaf populations. A secondary goal of the experiment design is to learn how best to employ RBFs and how to choose a reference model for a given set of hypotheses.

4.1 General setup

We generated data sets under different demographic scenarios, using the following constant setup. In experiments I and III, the generative population models (the "true" models) have 3 leaf populations A, B and C, an ancestral population AB and a root ancestral population ABC. In experiment II the true model has another ancestor population $ROOT$ which splits to ABC and an outgroup leaf population O. We use the coalescent software *ms* (Hudson, 2002) to generate four haploid sequences per leaf population. Sequence data contains 5000 loci of length 1000 bases. See appendix C for a sample *ms* sequence generation script. For each demographic scenario we generated two independent data sets (using the same generative hypothesis) to examine replication of results. To further assess replication we ran 2 independent MCMC runs in each G-PhoCS setting. See appendix C for an example G-PhoCS MCMC configuration file.

In each comparison instance we compared two hypothesis models $\mathcal{M}_1$ and $\mathcal{M}_2$ on a given data set. Depending on the specific test, comparison was done using relative Bayes factors with a differing reference model $\mathcal{M}_{ref}$ and using the harmonic mean as a benchmark comparison. On each data set we ran G-PhoCS twice with $\mathcal{M}_1$ and twice with $\mathcal{M}_2$, yielding four potential differences between the relevant stats (e.g., $HM(\mathcal{M}_\mathbf{1}, data) - HM(\mathcal{M}_\mathbf{2}, data)$, $RBF(\mathcal{M}_\mathbf{1}, \mathcal{M}_{null}, data) - RBF(\mathcal{M}_\mathbf{2}, \mathcal{M}_{null}, data)$ etc). The differences represent the algorithms final "choice" between $\mathcal{M}_1$ and $\mathcal{M}_2$, i.e. which model has higher estimated data likelihood relative to the reference model.

For each comparison we recorded the maximum and minimum of the 4 differences with their standard error margins (which McRef computes via bootstrap). For the max value we recorded $max + ste$ and for the min value we recorded $min - ste$. Since these errors correspond to the difference between two values, we took the square root of the sum of the two appropriate errors. As a result we attained 4 values for each comparison of M_1 and M_2 on a given data set and each method of comparison (e.g. HM, RBF with null model, etc). We used these values to plot the confidence intervals seen in the figures of each experiment.

4.2 Experiment I - Identifying population separation

In this experiment we generated data sets in which population ABCs divergence time is fixed to 0.00300, and perturbed ABs divergence time from 0 up to 0.00050. No migration was allowed between any population. We considered 2 hypotheses:

1. $\mathcal{M}_{3pops}$ - A model with 3 leaf populations A, B and C. This is the true model used to generate the sequence data
2. $\mathcal{M}_{2pops}$ - A model with 2 leaf populations AB and C, where the sequenced individuals of the original A and B populations are grouped into a single leaf population AB. This model coincides with the true model in the data set with $\tau_{AB} := 0$

and compared these two models using each of three techniques:

1. The harmonic mean
2. Relative Bayes factors with a reference model of $\mathcal{M}_{null}$
3. Relative Bayes factors with a reference model of $\mathcal{M}_{clade(AB)}$ (the original model with a clade rooted at population AB). Note that when $\mathcal{M}_{hyp} := \mathcal{M}_{2pops}$ we get $\mathcal{M}_{ref} = \mathcal{M}_{hyp}$

We used these three techniques to compare models $\mathcal{M}_{3pops}$ and $\mathcal{M}_{2pops}$ (i.e., log $\frac{P(X|\mathcal{M}_{3pops})}{P(X|\mathcal{M}_{2pops})}$)) with each variant of the data set.

We observe that when computing $RBF(\mathcal{M}_{hyp} = \mathcal{M}_{2pops}, \mathcal{M}_{ref} = \mathcal{M}_{clade(AB)})$, we get values near 0 ($< 1e^{-6}$). This is because the reference and hypothesis models converge to the same model. We consider this a simple validation of our RBF calculation. We also notice that both HM and the two RBFs are able to determine the correct model ($\mathcal{M}_{3pops}$) for $div50$, and they don't reject $\mathcal{M}_{2pops}$ for $div00$ (although $null\_RBF$ does give positive values). In $div20$ we see that both RBFs determine the correct model, while the harmonic mean does not significantly reject $\mathcal{M}_{2pops}$. We see that when we use $\mathcal{M}_{clade(AB)}$, the estimates of RBF are less noisy than when using $\mathcal{M}_{null}$. Lastly, $\mathcal{M}_{null}$ appears to bias upward the RBF estimates, resulting in false-positives for low divergence data sets ($div00$ and $div10$).

4.3 Experiment II - Determining model topology

In this experiment the true model contained an additional outgroup leaf population O. The divergence time of population $ROOT$ to populations O & ABC was set to 0.01000. The divergence time of ABC was again fixed to 0.00300 and the divergence time of AB was perturbed between 0.00300 and 0.00180. Again, no migration was allowed between any population. We considered 3 hypotheses:

1. $\mathcal{M}_{AB\_C\_O}$ - A model with four leaf populations A, B and C and O where A and B are siblings. This is the true model used to generate the sequence data

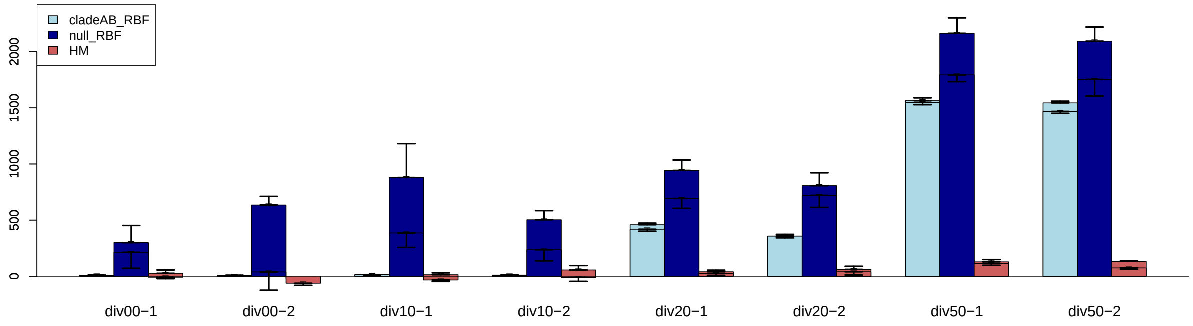


Figure 8. Results of experiments selecting between $\mathcal{M}_{3pops}$ and $\mathcal{M}_{2pops}$ using three techniques - 1) RBF of the null reference model, 2) RBF of a $clade(AB)$ reference model and 3) HM. Data sets are marked on the X-axis by *divXX-a*, where the value of *XX* stands for $\tau_{AB} \times 10,000$ and *a* indicates the data replicate (1 or 2). The divergence time of AB used in the generation of the data set increases between comparisons (left to right). The bars heights are the values of the comparison metric $\log \frac{P(X|\mathcal{M}_{3pops})}{P(X|\mathcal{M}_{2pops})}$. Each experiment was repeated twice to assess reproducability. We see in the graph that for $\tau_{AB} \leqslant 0.00020$ the harmonic mean does not confidently prefer the true hypothesis $\mathcal{M}_{3pops}$ over the competing hypothesis $\mathcal{M}_{2pops}$. RBFs, however, prefer $\mathcal{M}_{3pops}$ starting from $\tau_{AB} \geqslant 0.00010$, regardless of the chosen reference model.

2. $\mathcal{M}_{A\_BC\_O}$ - A similiar model but in which B and C are siblings

3. $\mathcal{M}_{AC\_B\_O}$ - A similiar model but in which A and C are siblings

Note that when $\tau(AB) = 0.00300 = \tau(ABC)$, the simulated model is one in which the three populations instantaneously diverge, so we expect the three hypotheses to have similiar fit to data. We compared these models using each of five techniques:

1. The harmonic mean

2. Relative Bayes factors with a reference model of $\mathcal{M}_{Clade(ROOT)}$ ($\mathcal{M}_{null}$)

3. Relative Bayes factors with a reference model of $\mathcal{M}_{Clade(ABC)}$

4. Relative Bayes factors with a reference model of $\mathcal{M}_{Comb(ROOT)}$

5. Relative Bayes factors with a reference model of $\mathcal{M}_{Comb(ABC)}$

Similiarly to experiment I, we used these techniques to compare the true model $\mathcal{M}_{AB\_C\_O}$ against the alternatives $\mathcal{M}_{A\_BC\_O}$ and $\mathcal{M}_{AC\_B\_O}$. Figure 9 shows the results of the comparisons.

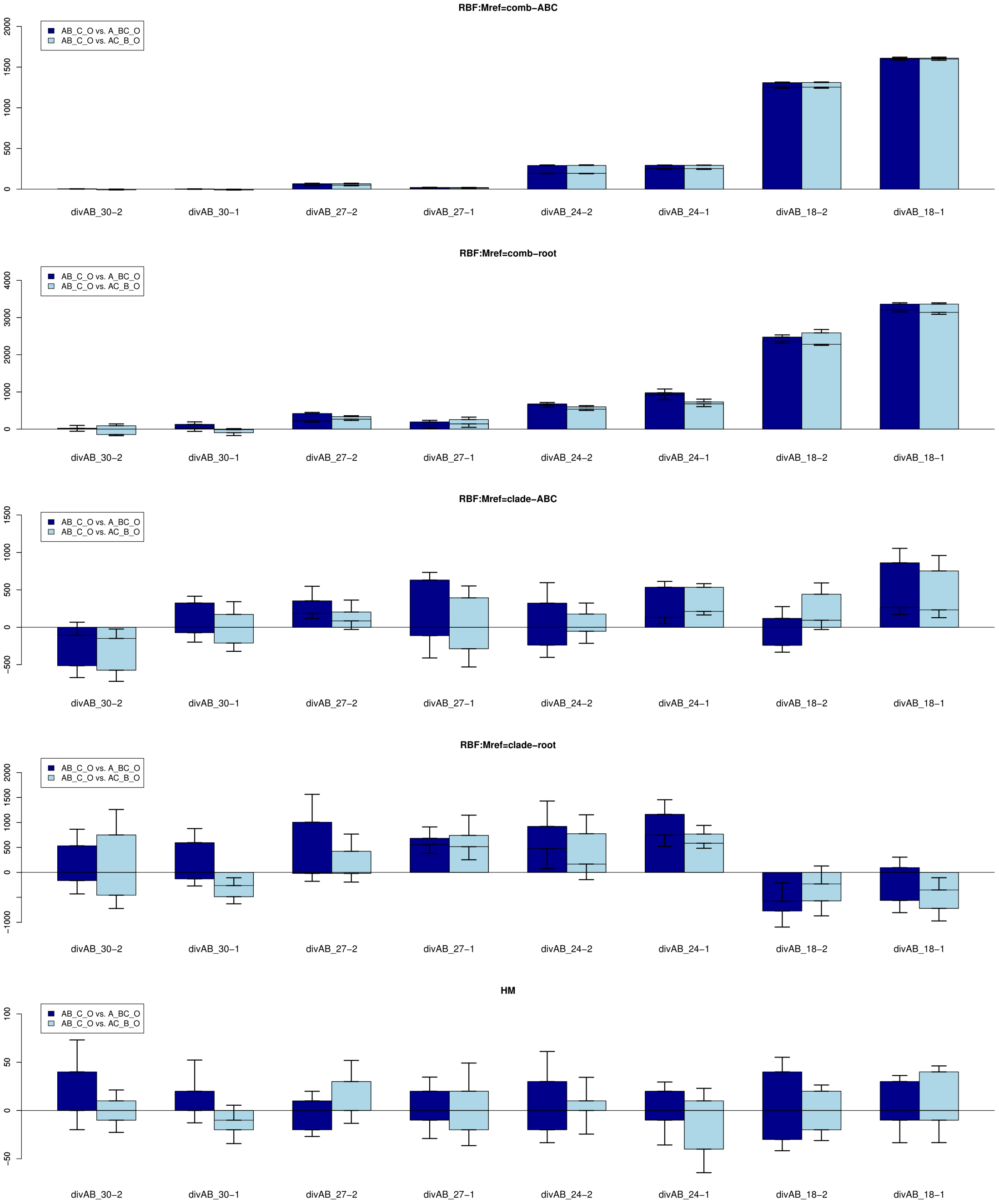


Figure 9. Results of experiments selecting between three model structures using each of multiple RBFs and HM. Data sets are marked on the X-axis by *divAB_XX-a*, where the value of *XX* stands for $\tau(AB) \times 10,000$ and *a* indicates the data replicate (1 or 2). The true gap between divergence times $\tau(ABC)$ and $\tau(AB)$ starts from zero on the left -most bar (where $\tau_{ABC} = 0.00300 = \tau_{AB}$) and increases between comparisons (left to right). We see that the more informative reference models ($Comb(ABC)$ followed by $Comb(ROOT)$) successfully select the true model, whereas the more general methods are very noisy and uncertain, even when the hypotheses should be indistinguishable.

We see that the two comb reference methods (first two panels of Figure 9) clearly and confidently choose the true hypothesis model, $\mathcal{M}_{AB\_C\_O}$. The comb reference methods also correctly show no preference to any model when the hypotheses are eqivalent. This is not true for the other methods. Amongst the two comb reference methods, the more localized $comb(ABC)$ provides a stronger and more confident signal. However, when using the $clade(ABC)$ reference model (3rd panel) we see at most a gentle upward trend in results, but no reproducable clear selection. In the remaining two experiments ,$clade(ROOT)$ and HM (seen in 4th and 5th panels) we see no selection and a high degree of uncertainty.

4.4 Experiment III - Determining direction of gene flow

In this experiment we generated data sets where the divergence times are fixed to $\tau_{ABC} = 0.00300$, and $\tau_{AB} = 0.00150$ and simulated different migration rates from population C to population B. We considered four hypotheses:

1. $\mathcal{M}_{migCB}$ - A model with a migration band from C to B (the true model)
2. $\mathcal{M}_{nomig}$ - A model with no migration bands
3. $\mathcal{M}_{migALL}$ - A model with migration bands between all pairs of sampled populations (6 migration bands total)
4. $\mathcal{M}_{migBC}$ - A model with migration band from B to C

and examined two ways to compare these 4 models:

1. Using the harmonic mean estimator (HM)
2. Using RBF where $\mathcal{M}_{ref} = \mathcal{M}_{null}$

To present the results, we conducted a comparison between each of the three models with migration against $\mathcal{M}_{nomig}$ as a base model (e.g. $\log \frac{P(X|\mathcal{M}_{migBC})}{P(X|\mathcal{M}_{nomig})}$). We ploted for each hypothesis and each data set the results when comparing using an RBF with the null modeland when using the harmonic mean (Figure 10). Because the conditional distribution for migration events is not fully implemented, we applied a small shortcut and assumed migration priors of the hypothesis and reference models cancel out. We estimate that this results in a relatively small correction, and believe it does not affect any trend in results.

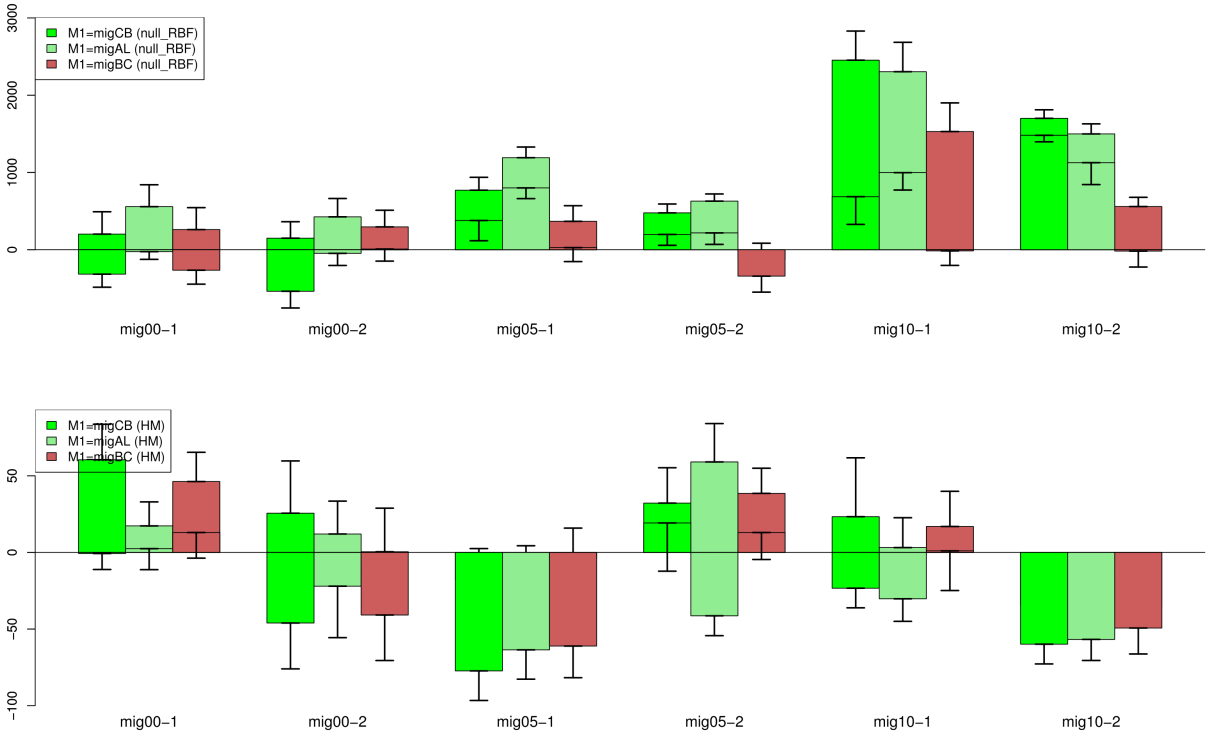


Figure 10. Results of experiments selecting between multiple migration patterns using the null reference model (panel 1) and using HM (panel 2). Data sets are marked by *migXX-a*, where *XX* stands for the migration rate from population C to B and a indicates the data replicate (1 or 2). The migration rate from C to B used in the generation of the data set increases between comparisons (left to right). The bars heights are the values of the comparison metric against $\mathcal{M}_{nomig}$, $\log \frac{P(X|\mathcal{M}_{migBC})}{P(X|\mathcal{M}_{nomig})}$. Each experiment was repeated twice to assess reproducability. We see that the harmonic mean (2nd panel) does not consistently prefer any model over another, whereas the null RBF (1st panel) prefers models with migration to the migration-less base model.

We see that the harmonic mean scores the three models with migrations similarly and it never significantly prefers models with migration to $\mathcal{M}_{nomig}$ (Figure 10). RBFs however consistently score $\mathcal{M}_{migCB}$ and $\mathcal{M}_{migAL}$ higher than $\mathcal{M}_{nomig}$ in the 4 data sets with migration. The preference is correlated to the simulated migration rate. RBFs also score $\mathcal{M}_{migCB}$ and $\mathcal{M}_{migALL}$ higher than $\mathcal{M}_{migBC}$. This shows that they are able to identify the direction of migration ($C \rightarrow B$ instead of $B \rightarrow C$). There doesn't seem to be a significant difference between the scores of $\mathcal{M}_{migCB}$ and $\mathcal{M}_{migALL}$. In principle, we would've liked to give a higher score to the most "compact" model, but this is not attained.

4.5 Summary

We've utilized RBFs in answering three model selection questions; 1) whether a divergence event occured, 2) what is the true migration pattern and 3) what is the relationship between leaf populations. In all three scenarios, model selection using relative Bayes factors significantly outperformed the harmonic mean estimator. We saw in experiments 1 and 2 that the choice of reference model has a great impact on algorithm performance. Generaly speaking, the best performing reference model is the most informative reference model that can be used, i.e. the one closest to all models being compared. We also see that, as expected, the success of the algorithm is correlated with the true distance between models, but it's estimations are not of high certainty. Finally we note that in experiment 3 RBFs did not succeed in choosing the most parsimonious hypothesis.

5 References

Browning SR, Browning BL. 2015. Accurate Non-parametric Estimation of Recent Effective Population Size from Segments of Identity by Descent. Am. J. Hum. Genet. 97:404–418.

Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A. 2011. Bayesian inference of ancient human demography from individual genome sequences. Nat. Genet. 43:1031–1034.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 5:e1000695.

Harris K, Nielsen R. 2013. Inferring demographic history from a spectrum of shared haplotype lengths. PLoS Genet. 9:e1003521.

Hey J, 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. 104:2785–2790.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics. 18:337–338.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro H, editor, Mammalian Protein Metabolism, New York: Academic Press, pp. 21–132.

Kamm JA, Terhorst J, Durbin R, Song YS. 2018. Efficiently inferring the demographic history of many populations with allele count data. bioRxiv preprint. Https://doi.org/10.1101/287268.

Kamm JA, Terhorst J, Song YS. 2017. Efficient computation of the joint sample frequency spectra for multiple populations. J Comput Graph Stat. 26:182–194.

Kingman J. 1982. The coalescent. Stoch. Process. Appl. 13:235–248.

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

Newton MA, Raftery AE. 1994. Approximate bayesian inference with the weighted likelihood bootstrap. Journal of the Royal Statistical Society. Series B (Methodological). pp. 3–48.

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

Rannala B, Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics. 164:1645–1656.

Sethuraman A, Hey J. 2016. IMa2p–parallel MCMC and inference of ancient demography under the Isolation with migration (IM) model. Mol Ecol Resour. 16:206–215.

Xie W, Lewis PO, Fan Y, Kuo L, Chen MH. 2011. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. Syst. Biol. 60:150–160.

Yang Z. 2015. The BPP program for species tree estimation and species delimitation. Curr Zool. 61:854–865.

A The conditional distribution $\widetilde{P}(\boldsymbol{\tau}|\mathrm{G})$ for models without migration

When the hypothesis model $\mathcal{M}$ has no migration, its model-pairing conditional distribution with the null model $\mathcal{M}_0$ is determined by specifying a conditional distribution for the divergence times, $\widetilde{P}(\boldsymbol{\tau}|\mathbf{G})$, such that $\widetilde{P}(\boldsymbol{\tau}|\mathbf{G}) > 0$ if and only if $P(\mathbf{G}|\boldsymbol{\tau}, \mathcal{M}) > 0$ (see Equations 5 and 12). Let $(\mathcal{T}, \boldsymbol{\tau})$ be a timed population phylogeny and let $\mathbf{G}$ be a collection of coalescent trees in which every leaf is mapped to a leaf population in $\mathcal{T}$ and every internal vertex v corresponds to a coalescent event at time $t(v)$. Then $P(\mathbf{G}|\boldsymbol{\tau}, \mathcal{M}) > 0$ if and only if the trees in $\mathbf{G}$ can be *embedded* in $(\mathcal{T}, \boldsymbol{\tau})$, as defined below.

Definition 1. *An embedding of a collection of local genealogies* $\mathbf{G}$ *in a timed population phylogeny* $(\mathcal{T}, \boldsymbol{\tau})$ *is a mapping, pop* : $\mathbf{G} \to \mathcal{T}$*, which satisfies the following conditions for every coalescence event* $v \in \mathbf{G}$*:*

1. *$pop(v)$ is alive at time $t(v)$:* $\tau(pop(v)) \;<\; t(v) \;\leqslant\; \tau(parent(pop(v)))$.
   (if p is a leaf population then $\tau(p) = 0$ and if p is the root population then $\tau(parent(p)) = \infty$.)

2. *$pop(parent(v))$ is ancestral (or equal) to $pop(v)$:* $pop(parent(v)) \geqslant_{\mathcal{T}} pop(v)$.

Note that if $\mathbf{G}$ is embeddable in $(\mathcal{T}, \boldsymbol{\tau})$, then this embedding is unique, because given a coalescent event v with daughter u, there is only one population that is alive at time $t(v)$ (condition 1) and ancestral or equal to $pop(u)$ (condition 2). A similar argument is used to establish a sufficient and necessary condition for embeddability below.

Definition 2 ($mrcaPop$)**.** *Given a coalescence event v in a local genealogy whose leaves are assigned to the leaves of a population phylogeny $\mathcal{T}$, let $mrcaPop(v)$ denote the most recent common ancestor (MRCA) in $\mathcal{T}$ of all populations to which leaves in the subtree rooted at v are mapped.*

Lemma 1. *A collection of local genealogies* $\mathbf{G}$ *has an embedding in a timed population phylogeny* $(\mathcal{T}, \boldsymbol{\tau})$ *iff for every $v \in \mathbf{G}$ we have $t(v) > \tau(mrcaPop(v))$.*

Proof.
$\Rightarrow$: Consider an embedding $pop : \mathbf{G} \to \mathcal{T}$, and let v be an arbitrary coalescence event in $\mathbf{G}$. Condition 2 implies that $pop(v) \geqslant_{\mathcal{T}} pop(l)$ for all leaves in the subtree rooted at v. We thus get $pop(v) \geqslant_{\mathcal{T}} mrcaPop(v)$, and by condition 1: $t(v) > \tau(pop(v)) \geqslant \tau(mrcaPop(v))$.
$\Leftarrow$: Let v be an arbitrary coalescence event in $\mathbf{G}$, and assume that $t(v) > \tau(mrcaPop(v))$. This means that there is a (unique) population, p^*, ancestral to $mrcaPop(v)$ that is also alive at time $t(v)$ (i.e., $\tau(p^*) \;<\; t(v) \;\leqslant\; \tau(parent(p^*))$). Define the embedding by mapping v to population p^*. Condition 1 is guaranteed by construction. Condition 2 is proved by considering an arbitrary coalescence event v and its parent $u = parent(v)$. Both $pop(u)$ and $pop(v)$ are ancestral (or equal) to $mrcaPop(v)$, because $mrcaPop(u) \geqslant_{\mathcal{T}} mrcaPop(v)$. Thus either $pop(v) \geqslant_{\mathcal{T}} pop(u)$ or $pop(u) \geqslant_{\mathcal{T}} pop(v)$. Condition 1 implies that $pop(v)$ cannot be strictly ancestral to $pop(u)$ via the following sequence of inequalities:

$$\tau(parent(pop(u)) \;\geqslant\; t(u) \;>\; t(v) \;>\; \tau(pop(v)) \;.$$

Hence, $pop(u) \geqslant_{\mathcal{T}} pop(v)$, establishing condition 2. □

Lemma 1 directly implies a feasible range of every divergence time τ_p:

Claim 1. *Let* $\mathbf{G}$ *be a collection of local genealogies whose leaves are mapped to leaves of a population phylogeny $\mathcal{T}$. Then for every ancestral population p, $P(\mathbf{G}|\tau_p = \tau, \mathcal{M}) > 0$ iff $\tau \in [0, ubound(p|\mathbf{G}))$, where the upper bound of the feasible range for τ_p is given by:*

$$ubound(p|\mathbf{G}) = \quad \min\{t(v) : mrcaPop(v) \geqslant_{\mathcal{T}} p\} \tag{28}$$

We thus define $\widetilde{P}(\boldsymbol{\tau}|\mathbf{G})$ as a product of uniform distributions for $\boldsymbol{\tau}$ in their feasible ranges, as defined by Claim 1.

B The conditional distribution $\widetilde{P}(\boldsymbol{\tau}, \mathbf{G}_m|\mathbf{G}_c, \mathbf{m})$ for models with migration

As with the case without migration, the conditional distribution $\widetilde{P}(\boldsymbol{\tau}, \mathbf{G}_m|\mathbf{G}_c, \mathbf{m})$ is constructed by first specifying the necessary and sufficient conditions under which a genealogy with migration events $\mathbf{G} = (\mathbf{G}_c, \mathbf{G}_m)$ is embeddable in a timed population phylogeny $(\mathcal{T}, \boldsymbol{\tau})$. Migration complicates these conditions because of two main reasons: (1) migration breaks the fundamental assumption that genealogy branches move from a population to its parent in the phylogeny, and (2) unlike coalescent events, migration events are mapped to specific populations and thus pose strict constraints on the embedding. The first issue is addressed by examining *migration-free* trees, obtained by cutting branches of the local genealogies in $\mathbf{G}$ at migration events. We associate each migration event $w \in \mathbf{G}_m$ with the branch in $\mathbf{G}_c$ on which it is placed, a specific time along that branch, a source population for migration, and a target population for migration. Thus, each migration event, $w \in \mathbf{G}_m$, is a root of one migration-free tree mapped to population $target(w)$ and a leaf of another tree mapped to population $source(w)$. In each migration-free tree, leaves are mapped to populations in $\mathcal{T}$ and branches move from a population to its parent, as assumed in condition 2 of Definition 1. Hence, we can extend the operator $mrcaPop(v)$ of Definition 2 as the MRCA of all populations to which the leaves of the migration-free subtree rooted at v are mapped. The following lemma specifies embeddability conditions based on this extended $mrcaPop$ operator and on the restriction that at the time of each migration event, the source and target populations must be alive.

Lemma 2. *A collection of local genealogies* $\mathbf{G}$ *consisting of coalescent trees* $\mathbf{G}_c$ *and migration events* $\mathbf{G}_m$ *has an embedding in a timed population phylogeny* $(\mathcal{T}, \boldsymbol{\tau})$ *iff the following four conditions are satisfied:*

1. $\forall v \in \mathbf{G}_c : t(v) > \tau(mrcaPop(v))$
2. $\forall w \in \mathbf{G}_m : target(w) \geqslant_{\mathcal{T}} mrcaPop(w)$
3. $\forall w \in \mathbf{G}_m : t(w) > \max(\ \tau(source(w))\ ,\ \tau(target(w))\ )$
4. $\forall w \in \mathbf{G}_m : t(w) \leqslant \min(\ \tau(parent(source(w)))\ ,\ \tau(parent(target(w)))\ )$

Proof.
$\Rightarrow$: Assume a collection of local genealogies $\mathbf{G}$ embedded in a timed population phylogeny $(\mathcal{T}, \boldsymbol{\tau})$. For every coalescent event $v \in \mathbf{G}_c$, we know that $t(v) \geqslant \tau(pop(v))$, and $pop(v) \geqslant_{\mathcal{T}} mrcaPop(v)$ (considering the migration-free tree that v belongs to), implying condition 1. Now consider an arbitrary migration event $w \in \mathbf{G}_m$, which is a root of some migration-free tree in $\mathbf{G}$. Because this root is mapped to population $target(w)$, we get that $target(w) \geqslant_{\mathcal{T}} mrcaPop(w)$ (condition 2). Finally, conditions 3 and 4 are implied by the fact that w is mapped to populations $target(w)$ (as the root of a migration-free tree) and $source(w)$ (as a leaf of a migration-free tree).
$\Leftarrow$: Assume a collection of local genealogies $\mathbf{G}$ and a timed population phylogeny $(\mathcal{T}, \boldsymbol{\tau})$ satisfying the four conditions of the lemma. We embed $\mathbf{G}$ in $(\mathcal{T}, \boldsymbol{\tau})$ by mapping every coalescent event to the population ancestral to $mrcaPop(v)$ that is also alive at time $t(v)$. This is the same mapping used in the proof of Lemma 1, when no migration was assumed, and as in that case, we can show that such a population exists (through condition 1) and that for each coalescent event v we have $pop(parent(v)) \geqslant_{\mathcal{T}} pop(v)$. Hence, the two conditions of Definition 1 are satisfied for all coalescent events. The same holds for all migration events, because conditions 3 and 4 imply that each migration event w is mapped to source and target populations that are both alive at time $t(w)$, and condition 2 implies that $target(w)$ is ancestral to the population to which the event at the bottom of the branch below w is mapped. Thus the mapping satisfies the two conditions of Definition 1 with respect to all migration-free trees in $\mathbf{G}$, implying that $\mathbf{G}$ is embeddable in $(\mathcal{T}, \boldsymbol{\tau})$. □

Note that condition 2 of the lemma specifies constraints on migration events in $\mathbf{G}_m$ and conditions 1, 3, and 4 define the feasible range for divergence times, as defined below.

Claim 2. *Let* $\mathbf{G}$ *be a collection of local genealogies with migration events. Then for every ancestral population* p*,* $P(\mathbf{G}|\tau_p = \tau, \mathcal{M}) > 0$ *iff for every* $w \in \mathbf{G}_m$ *we have* $target(w) \geqslant_{\mathcal{T}} mrcaPop(w)$ *and* $\tau \in [lbound(p|\mathbf{G}), ubound(p|\mathbf{G}))$*, where the bounds of the feasible range for* τ_p *are given by:*

$$lbound(p|\mathbf{G}) = \max\{t(w)|w \in \mathbf{G}_m \wedge (p \geqslant_{\mathcal{T}} parent(source(w)) \vee p \geqslant_{\mathcal{T}} parent(target(w)))\} \tag{29}$$

$$ubound(p|\mathbf{G}) = \min(ubound_1(p|\mathbf{G}), ubound_2(p|\mathbf{G})) \tag{30}$$

$$ubound_1(p|\mathbf{G}) = \min\{t(v)|v \in \mathbf{G}_c \wedge mrcaPop(v) \geqslant_{\mathcal{T}} p\} \tag{31}$$

$$ubound_2(p|\mathbf{G}) = \min\{t(w)|w \in \mathbf{G}_m \wedge (source(w) \geqslant_{\mathcal{T}} p \vee target(w) \geqslant_{\mathcal{T}} p)\} \tag{32}$$

We thus define the conditional distribution $\widetilde{P}(\mathbf{G}_m, \boldsymbol{\tau}|\mathbf{G}_c, \mathbf{m}) = \widetilde{P}(\boldsymbol{\tau}|\mathbf{G})\widetilde{P}(\mathbf{G}_m|\mathbf{G}_c, \mathbf{m})$, where $\widetilde{P}(\boldsymbol{\tau}|\mathbf{G})$ is the product of uniform distributions for $\boldsymbol{\tau}$ in their feasible ranges, as defined by Claim 2, and $\widetilde{P}(\mathbf{G}_m|\mathbf{G}_c, \mathbf{m})$ is defined using a probabilistic protocol for sampling migration events. This protocol mimics the true migration model of $\mathcal{M}$ as much as possible without knowing the divergence times. Migration events are sampled backward in time by holding for each branch $(u, v) \in \mathbf{G}_c$ the set of populations it may be embedded in (those ancestral to $mrcaPop(v)$), and allowing the branch to migrate back along any migration band whose target population is one of those populations. The protocol starts by enabling migration in all bands, and it removes a migration band b from consideration when the protocol reaches time $t = \min(ubound(parent(source(b))|\mathbf{G}), ubound(parent(target(b))|\mathbf{G}))$, as defined by Equations 30-32. By doing this, the protocol ensures that the resulting $\mathbf{G}$ will be embeddable in some timed version of the population phylogeny (see Claim 3 below).

Sampling protocol for $\widetilde{P}(\mathbf{G}_m|\mathbf{G}_c, \mathbf{m})$:

1. **Initialization:**

   (a) Initialize set of living branches: $E_{live} \leftarrow \{(u, v) \in E(\mathbf{G}_c)|v \text{ is a leaf}\}$. Map each $(u, v) \in E_{live}$ to the sampling population of the leaf v and all populations ancestral to it: $pops((u, v)) \leftarrow \{p|p \geqslant_{\mathcal{T}} pop(v)\}$.

   (b) Initialize living migration bands: $B_{live} \leftarrow B$.

   (c) Initialize time: $t \leftarrow 0$.

2. **Determine current migration rates:** Determine the number of branches currently mapped to each population, $n[p] = |\{e \in E_{live} : p \in pops(e)\}|$, and compute the *effective rate* of each living migration band: $\lambda[b] = m_b \times n[target(b)]$ (the migration rate scaled by the number of potentially migrating branches).

3. **Sample time of next migration:** Sample a waiting time Δt for the next migration event according to an exponential distribution with rate $\lambda = \sum_{b \in B_{live}} \lambda[b]$. If there are no live migration bands with positive rates, then $\lambda = 0$ and the scan terminates (no more migration events to sample). Otherwise, set $t \leftarrow t + \Delta t$ and compare t to the time of the next coalescent event back in time, v.

4. If $t < t(v)$, then **sample migration event:**

   (a) Sample a migration band $b \in B_{live}$ using a categorical distribution with $p_b = \frac{\lambda[b]}{\lambda}$.

   (b) Select a branch for migration $e \in E_{live}$ uniformly at random among the $n[target(b)]$ branches mapped to the target population of the selected migration band.

   (c) Add a new migration event w to $\mathbf{G}_m$ on branch e from population $source(b)$ to population $target(b)$ at time t.

   (d) Update the population mapping of edge e: $pops(e) \leftarrow \{p : p \geqslant_{\mathcal{T}} source(b)\}$.

   (e) Remove from B_{live} all migration bands whose source or target population is a strict descendant of either $source(b)$ or $target(p)$. Formally, remove band b' iff there is $p' \in \{source(b'), target(b')\}$ and $p \in \{source(b), target(b)\}$ s.t. $p \geqslant_{\mathcal{T}} parent(p')$.

(f) Go to Step 2.

5. If $t \geqslant t(v)$, then **encounter coalescence event:**

   (a) Let e_1 and e_2 be the two branches coalescing in v, and let e be the branch above v.

   (b) Update current branches: $\mathbb{E}_{live} \leftarrow E_{live} \backslash \{e_1, e_2\} \cup \{e\}$.

   (c) Map the new branch: $pops(e) = pops(e_1) \cap pops(e_2)$.

   (d) Remove from B_{live} all migration bands whose source or target is a strict descendant of the most recent population in $pops(e)$. Formally, if p_0 is the most recent population in $pops(e)$, then remove band b iff $p_0 \geqslant_{\mathcal{T}} parent(source(b))$ or $p_0 \geqslant_{\mathcal{T}} parent(target(b))$.

   (e) Set $t \leftarrow t(v)$ and go to Step 2.

The following claim establishes the validity and completeness of the above protocol for $\widetilde{P}(\mathbf{G}_m|\mathbf{G}_c, \mathbf{m})$:

Claim 3. $\widetilde{P}(\mathbf{G}_m|\mathbf{G}_c, \mathbf{m}) > 0$ *iff there exist* $\boldsymbol{\tau}$ *s.t.* $P(\mathbf{G}_c, \mathbf{G}_m|\boldsymbol{\tau}, \mathbf{m}, \mathcal{M}) > 0$.

Proof. First, note that the protocol maps each branch (u, v) to the set of populations ancestral to $mrcaPop(v)$: $pops((u, v)) = \{p : p \geqslant_{\mathcal{T}} mrcaPop(v)\}$. This is done by the appropriate initialization of the mapping in leaf branches in step 1a and branches above migration events in step 4d, and by the appropriate intersection update in branches above coalescent events in step 5c. Both directions of the claim are now proved using this observation and the conditions of Claim 2

$\Rightarrow$

Let $\mathbf{G}_m$ be the set of migration events sampled by the protocol given $\mathbf{G}_c$ and $\mathbf{m}$. To establish that there exist $\boldsymbol{\tau}$ s.t. $P(\mathbf{G}_c, \mathbf{G}_m|\boldsymbol{\tau}, \mathbf{m}, \mathcal{M}) > 0$ using Claim 2, we need to show that: (1) every sampled migration event in $\mathbf{G}_m$ satisfies $target(w) \geqslant_{\mathcal{T}} mrcaPop(w)$, and (2) the resulting $\mathbf{G}$ satisfies $lbound(p|\mathbf{G}) < ubound(p|\mathbf{G})$ for every ancestral population p. Let $w \in \mathbf{G}_m$ be an arbitrary migration event and denote by $e(w)$ the branch in $\mathbf{G}_c$ on which w is sampled. Then, $target(w) \subseteq pops(e(w))$ (step 4b), implying that $target(w) \geqslant_{\mathcal{T}} mrcaPop(w)$, as required by Claim 2. Now, consider an arbitrary ancestral population p, and denote for brevity $lb = lbound(p|\mathbf{G})$, $ub_1 = ubound_1(p|\mathbf{G})$, and $ub_2 = ubound_2(p|\mathbf{G})$ (Equations 29-32). We will show that $lb < \min(ub_1, ub_2) = ubound(p|\mathbf{G})$.

Let v be the coalescent event realizing ub_1 and let w and w' be the migration events realizing ub_2 and lb, respectively. Note that if one of these events does not exist, then the appropriate bound is set to its extreme value (0 for lb and ∞ for ub_1 and ub_2), and the inequality above holds. Otherwise, the definition of w' and lb implies that either $p \geqslant_{\mathcal{T}} parent(source(w'))$ or $p \geqslant_{\mathcal{T}} parent(target(w'))$, and the definition of w and ub_2 implies that either $source(w) \geqslant_{\mathcal{T}} p$ or $target(w) \geqslant_{\mathcal{T}} p$. Hence, the condition of step 4e of the protocol is satisfied for the migration band of event w' (b') when the protocol samples event w. This means that migration band b' is not alive after sampling w and $lb = t(w') < t(w) = ub_2$. Similarly, the condition of step 5d of the protocol is satisfied for migration band b' when the protocol encounters coalescent event v ($p_0 = mrcaPop(v)$). Hence, migration band b' is not alive after encountering v and $lb = t(w') < t(v) = ub_1$, completing the requirements of Claim 2.

$\Leftarrow$

Let $(\mathbf{G}, \boldsymbol{\tau})$ be a collection of local genealogies and divergence times s.t. $P(\mathbf{G}_c, \mathbf{G}_m|\boldsymbol{\tau}, \mathbf{m}, \mathcal{M}) > 0$. We will show that the migration events in $\mathbf{G}_m$ can be sampled by the protocol (with some positive probability). Consider an arbitrary migration event $w \in \mathbf{G}_m$ and assume that the protocol reached time $t(w)$ in $\mathbf{G}_c$ after having correctly sampled all events $w' \in \mathbf{G}_m$ s.t. $t(w') < t(w)$. To prove that event w can be sampled with positive probability we need to establish that: (1) its migration band $(p_s, p_t) = (source(w), target(w))$ is alive at time $t(w)$, and (2) its branch, e, is mapped to the target population p_t. The second requirement follows from Claim 2, which implies that $p_t \geqslant_{\mathcal{T}} mrcaPop(w)$, and our observation on the mapping that states that each branch is mapped to the set of populations ancestral to its $mrcaPop$.

To establish the first requirement we need to prove that migration band $b = (p_s, p_t)$ was not removed from B_{live} before time $t(w)$. The protocol removes migration bands from B_{live} either after sampling migration events (step 4e) or after encountering a coalescent events (step 5d). Let $w' \in \mathbf{G}_m$ be an arbitrary migration event sampled before w s.t. $t(w') < t(w)$. Claim 2 implies that for $p' \in \{source(w'), target(w')\}$ we have $\tau(p') < ubound_2(p'|\mathbf{G}) \leqslant t(w') < t(w)$, and for $p \in \{p_s, p_t\}$ we have $t(w) \leqslant lbound(parent(p)|\mathbf{G}) \leqslant \tau(parent(p))$. Hence, $\tau(p') < \tau(parent(p))$, implying that populations $source(w')$ and $target(w')$ are not strictly ancestral to populations p_s and p_t, and so the migration band (p_s, p_t) is not removed from B_{live} after sampling event w' (see step 4e).

Now, let $v \in \mathbf{G}_c$ be an arbitrary coalescent event encountered before sampling w s.t. $t(v) < t(w)$. Claim 2 implies that for $p' = mrcaPop(v)$ we have $\tau(p') < ubound_1(p'|\mathbf{G}) \leqslant t(v) < t(w)$ and for $p \in \{p_s, p_t\}$ we have $t(w) \leqslant lbound(parent(p)|\mathbf{G}) \leqslant \tau(parent(p))$. This means that $\tau(p') < \tau(parent(p))$, implying that population $mrcaPop(v)$ is not strictly ancestral to populations p_s and p_t, and so the migration band (p_s, p_t) is not removed from B_{live} after encountering event v (see step 5d). Thus, migration band (p_s, p_t) is alive at time $t = t(w)$, and the branch e is mapped to p_t, allowing the protocol to sample w at time $t(w)$ with positive probability. □

Computing the conditional probability

Now that we have fully defined the conditional probability distribution $\widetilde{P}(\mathbf{G}_m, \boldsymbol{\tau}|\mathbf{G}, \mathbf{m})$, we turn to describe how to compute it for given values of $(\mathbf{G}, \boldsymbol{\tau}, \mathbf{m})$. The divergence time conditionals, $\widetilde{P}(\boldsymbol{\tau}|\mathbf{G})$, are defined as a product of uniform distributions in the feasible space of every parameter, as defined by Claim 2. The bounds $lbound$ and $ubound_2$ are easy to compute by traversing all migration events in $\mathbf{G}_m$, and the bound $ubound_1$ can be computed by recursively computing $mrcaPop$ for all coalescent events in $\mathbf{G}_c$, as described in the previous section. This is done by considering the migration-free trees defined by $\mathbf{G}$. The conditional probability for the migration events, $\widetilde{P}(\mathbf{G}_m|\mathbf{G}_c, \mathbf{m})$ is computed according to the sampling protocol described above. As in a standard model of migration at constant rate, this probability can be expressed as a product of contributions across migration bands:

$$\ln\left(\widetilde{P}(\mathbf{G}_m|\mathbf{G}_c, \mathbf{m})\right) \;=\; \sum_b \left(\ln(m_b) \cdot numMigs(\mathbf{G}_m, b)^{m_b} - m_b \cdot \widetilde{migStats}(\mathbf{G}, b)\right) . \tag{33}$$

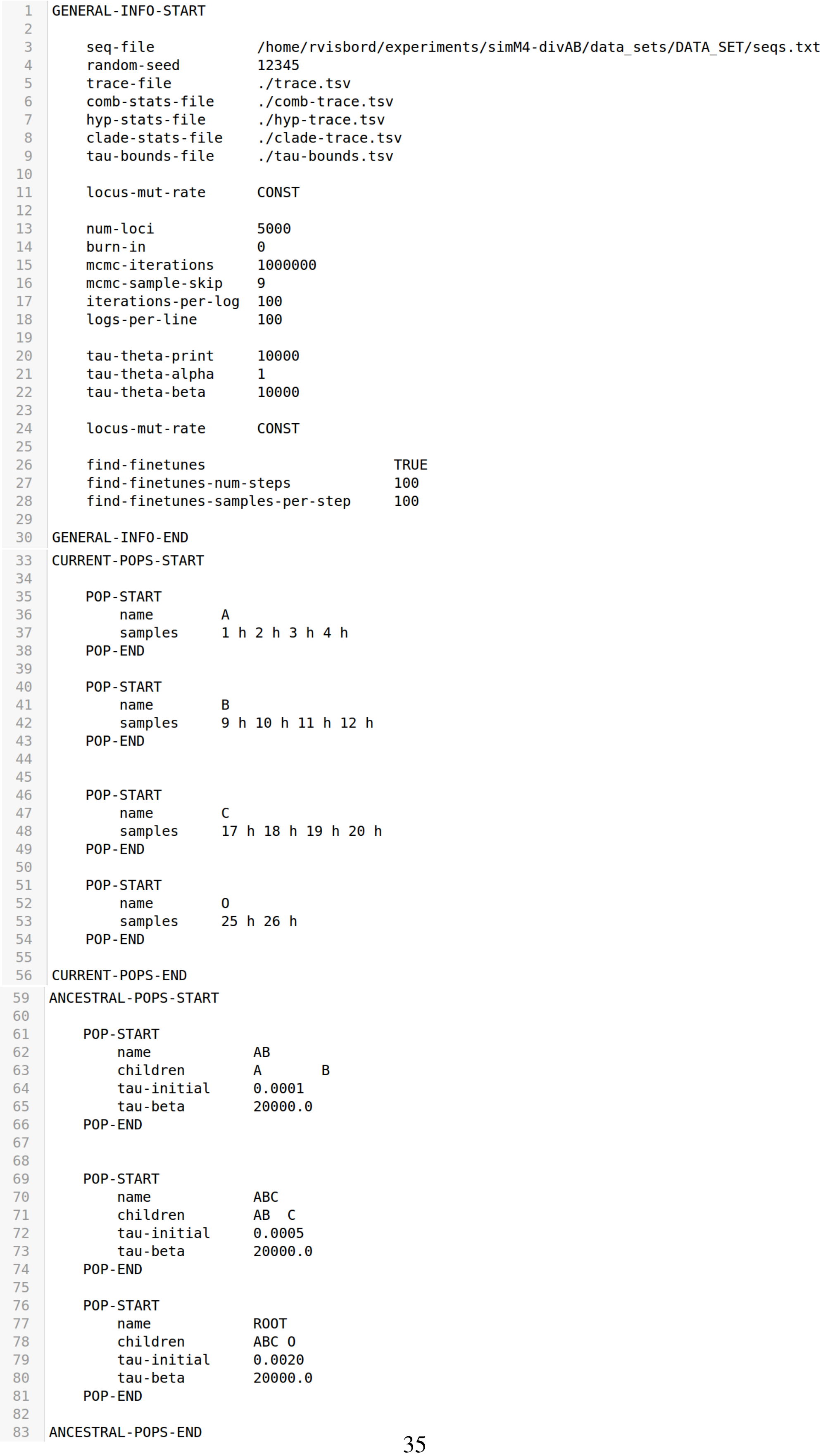
Consequently, the contribution of migration band b to $\widetilde{P}(\mathbf{G}_m|\mathbf{G}_c, \mathbf{m})$ is very similar to its contribution to $P(\mathbf{G}|\boldsymbol{\Theta}, \mathcal{M})$, and the ratio between these contributions is defined by the difference between $migStats(\mathbf{G}, b)$ and $\widetilde{migStats}(\mathbf{G}, b)$. Both migration statistics are defined as sum across time intervals in population $target(b)$ across the life span of the migration band. In model $\mathcal{M}$, the life span starts at $t = \max(\tau(source(b)), \tau(target(b)))$ and ends at
$t = \min(\tau(parent(source(b))), \tau(parent(target(b))))$. In the sampling protocol the life span starts at time $t = 0$ and ends at $t = \min(ubound(parent(source(b))|\mathbf{G}), ubound(parent(target(b))|\mathbf{G}))$. Note that the life span in $\mathcal{M}$ is contained in the protocol life span, and in this time the lineages mapped to population $target(b)$ are the same in both cases. Thus the residual difference, $migStats(\mathbf{G}, b) - \widetilde{migStats}(\mathbf{G}, b)$, is computed by considering intervals mapped to $target(b)$ in the protocol and not in $\mathcal{M}$. For instance, if b is a migration band between two sampled populations, then its life span in $\mathcal{M}$ and in the protocol starts at $t = 0$, and the residual is computed by determining which branches of $\mathbf{G}$ are mapped to population $target(p)$ in the time interval between $t = \min(\tau(parent(source(b))), \tau(parent(target(b))))$ and $t = \min(ubound(parent(source(b))|\mathbf{G}), ubound(parent(target(b))|\mathbf{G}))$.

C Pipeline examples

Listing 1. *ms* script used to generate data-sets in experiment III

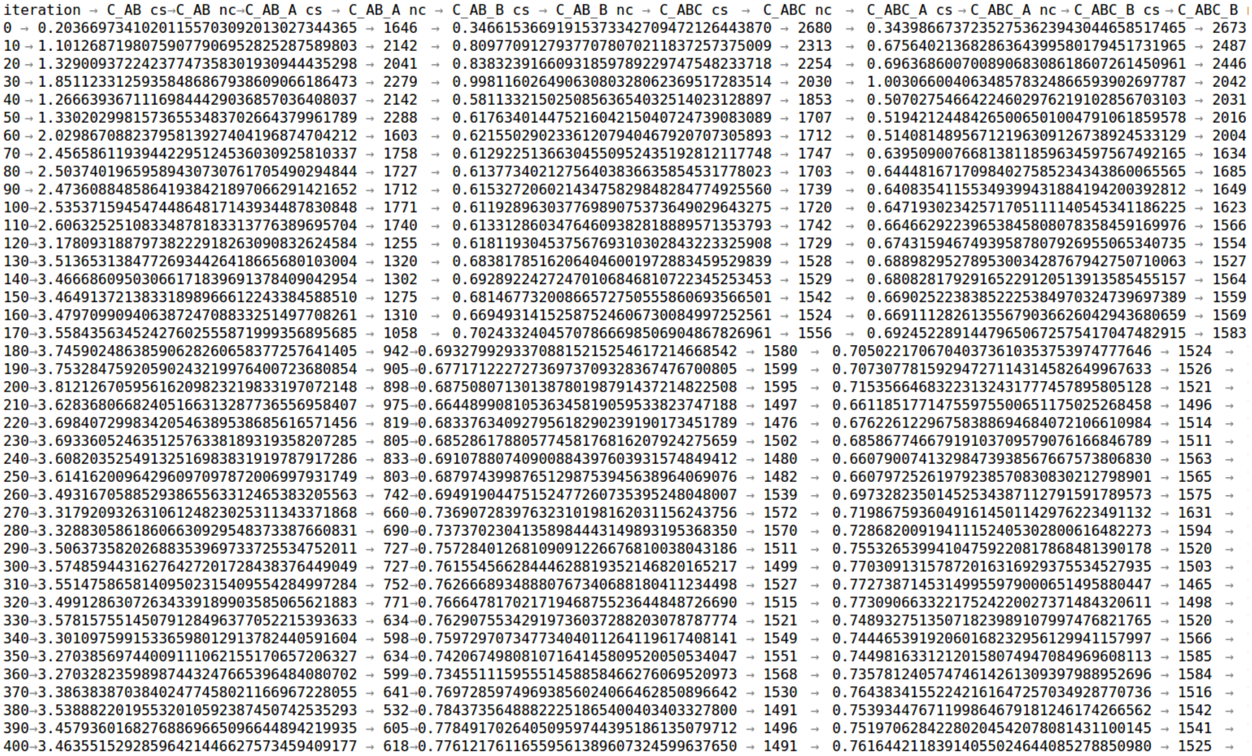
```
#!/bin/bash

# M4 - four population model with 26 individuals (8 per pop + 2 in outgroup),
# theta = 0.001, tau_AB = TAU, tau_ABC=0.0003, tau_ABCD=0.001, mig C-->A (M_CA)
# and mig A-->C (M_AC)
ms      26          5000    -T    -r 0.000001    1000            -I 4 8 8 8 2
     -n 1 100  -n 2 100  -n 3 100 -n 4 100
     -m 1 3 M_CA -m 3 1 M_AC
     -ej TAU     2 1  -en TAU 1 100 -em TAU 1 3 0.0 -em TAU 3 1 0.0
     -ej 30      3 1  -en 30  1 100
     -ej 100     4 1  -en 100 1 100
```

Sample G-PhoCS MCMC configuration



```
GENERAL-INFO-START

    seq-file            /home/rvisbord/experiments/simM4-divAB/data_sets/DATA_SET/seqs.txt
    random-seed         12345
    trace-file          ./trace.tsv
    comb-stats-file     ./comb-trace.tsv
    hyp-stats-file      ./hyp-trace.tsv
    clade-stats-file    ./clade-trace.tsv
    tau-bounds-file     ./tau-bounds.tsv

    locus-mut-rate      CONST

    num-loci            5000
    burn-in             0
    mcmc-iterations     1000000
    mcmc-sample-skip    9
    iterations-per-log  100
    logs-per-line       100

    tau-theta-print     10000
    tau-theta-alpha     1
    tau-theta-beta      10000

    locus-mut-rate      CONST

    find-finetunes                      TRUE
    find-finetunes-num-steps            100
    find-finetunes-samples-per-step     100

GENERAL-INFO-END
CURRENT-POPS-START

    POP-START
        name        A
        samples     1 h 2 h 3 h 4 h
    POP-END

    POP-START
        name        B
        samples     9 h 10 h 11 h 12 h
    POP-END


    POP-START
        name        C
        samples     17 h 18 h 19 h 20 h
    POP-END

    POP-START
        name        O
        samples     25 h 26 h
    POP-END

CURRENT-POPS-END
ANCESTRAL-POPS-START

    POP-START
        name            AB
        children        A       B
        tau-initial     0.0001
        tau-beta        20000.0
    POP-END


    POP-START
        name            ABC
        children        AB  C
        tau-initial     0.0005
        tau-beta        20000.0
    POP-END

    POP-START
        name            ROOT
        children        ABC O
        tau-initial     0.0020
        tau-beta        20000.0
    POP-END

ANCESTRAL-POPS-END
```

Sample G-PhoCS traces of sufficient stats



```
iteration → C_AB cs→C_AB nc→C_AB_A cs → C_AB_A nc → C_AB_B cs → C_AB_B nc → C_ABC cs  →  C_ABC nc  →  C_ABC_A cs → C_ABC_A nc → C_ABC_B cs → C_ABC_B
0 → 0.20366973410201155703092013027344365 → 1646  →  0.34661536691915373342709472126443870 → 2680  →  0.34398667372352753623943044658517465 → 2673
10 → 1.10126871980759077906952825287589803 → 2142  →  0.80977091279377078070211837257375009 → 2313  →  0.67564021368286364399580179451731965 → 2487
20 → 1.32900937224237747358301930944435298 → 2041  →  0.83832391660931859789229747548233718 → 2254  →  0.69636860070089068308618607261450961 → 2446
30 → 1.85112331259358486867938609066186473 → 2279  →  0.99811602649063080328062369517283514 → 2030  →  1.00306600406348578324866593902697787 → 2042
40 → 1.26663936711169844429036857036408037 → 2142  →  0.58113321502508563654032514023128897 → 1853  →  0.50702754664224602976219102856703103 → 2031
50 → 1.33020299815736553483702664379961789 → 2288  →  0.61763401447521604215040724739083089 → 1707  →  0.51942124484265006501004791061859578 → 2016
60 → 2.02986708823795813927404196874704212 → 1603  →  0.62155029023361207940467920707305893 → 1712  →  0.51408148956712196309126738924533129 → 2004
70 → 2.45658611939442295124536030925810337 → 1758  →  0.61292251366304550952435192812117748 → 1747  →  0.63950900766813811859634597567492165 → 1634
80 → 2.50374019659589430730761705490294844 → 1727  →  0.61377340212756403836635854531778023 → 1703  →  0.64448167170984027585234343860065565 → 1685
90 → 2.47360884858641938421897066291421652 → 1712  →  0.61532720602143475829848284774925560 → 1739  →  0.64083541155349399431884194200392812 → 1649
100→2.53537159454744864817143934487830848 → 1771  →  0.61192896303776989075373649029643275 → 1720  →  0.64719302342571705111140545341186225 → 1623
110→2.60632525108334878183313776389695704 → 1740  →  0.61331286034764609382818889571353793 → 1742  →  0.66466292239653845808078358459169976 → 1566
120→3.17809318879738222918263090832624584 → 1255  →  0.61811930453756769310302843223325908 → 1729  →  0.67431594674939587807926955065340735 → 1554
130→3.51365313847726934426418665680103004 → 1320  →  0.68381785162064046001972883459529839 → 1528  →  0.68898295278953003428767942750710063 → 1527
140→3.46668609503066171839691378409042954 → 1302  →  0.69289224272470106846810722345253453 → 1529  →  0.68082817929165229120513913585455157 → 1564
150→3.46491372138331898966612243384588510 → 1275  →  0.68146773200866572750555860693566501 → 1542  →  0.66902522383852225384970324739697389 → 1559
160→3.47970990940638724708833251497708261 → 1310  →  0.66949314152587524606730084997252561 → 1524  →  0.66911128261355679036626042943680659 → 1569
170→3.55843563452427602555871999356895685 → 1058  →  0.70243324045707866698506904867826961 → 1556  →  0.69245228914479650672575417047482915 → 1583
180→3.74590248638590628260658377257641405 → 942→0.69327992933708815215254617214668542 → 1580  →  0.70502217067040373610353753974777646 → 1524  →
190→3.75328475920590243219976400723680854 → 905→0.67717122272736973709328367476700805 → 1599  →  0.70730778159294727114314582649967633 → 1526  →
200→3.81212670595616209823219833197072148 → 898→0.68750807130138780198791437214822508 → 1595  →  0.71535664683223132431777457895805128 → 1521  →
210→3.62836806682405166313287736556958407 → 975→0.66448990810536345819059533823747188 → 1497  →  0.66118517714755975500651175025268458 → 1496  →
220→3.69840729983420546389538685616571456 → 819→0.68337634092795618290239190173451789 → 1476  →  0.67622612296758388694684072106610984 → 1514  →
230→3.69336052463512576338189319358207285 → 805→0.68528617880577458176816207924275659 → 1502  →  0.68586774667919103709579076166846789 → 1511  →
240→3.60820352549132516983831919787917286 → 833→0.69107880740900884397603931574849412 → 1480  →  0.66079007413298473938567667573806830 → 1563  →
250→3.61416200964296097097872006997931749 → 803→0.68797439987651298753945638964069076 → 1482  →  0.66079725261979238570830830212798901 → 1565  →
260→3.49316705885293865563312465383205563 → 742→0.69491904475152477260735395248048007 → 1539  →  0.69732823501452534387112791591789573 → 1575  →
270→3.31792093263106124823025311343371868 → 660→0.73690728397632310198162031156243756 → 1572  →  0.71986759360491614501142976223491132 → 1631  →
280→3.32883058618606630929548373387660831 → 690→0.73737023041358984443149893195368350 → 1570  →  0.72868200919411152405302800616482273 → 1594  →
290→3.50637358202688353969733725534752011 → 727→0.75728401268109091226676810038043186 → 1511  →  0.75532653994104759220817868481390178 → 1520  →
300→3.57485944316276427201728438376449049 → 727→0.76155456628444628819352146820165217 → 1499  →  0.77030913157872016316929375534527935 → 1503  →
310→3.55147586581409502315409554284997284 → 752→0.76266689348880767340688180411234498 → 1527  →  0.77273871453149955979000651495880447 → 1465  →
320→3.49912863072634339189903585065621883 → 771→0.76664781702171946875523644848726690 → 1515  →  0.77309066332217524220027371484320611 → 1498  →
330→3.57815755145079128496377052215393633 → 634→0.76290755342919736037288203078787774 → 1521  →  0.74893275135071823989107997476821765 → 1520  →
340→3.30109759915336598012913782440591604 → 598→0.75972970734773404011264119617408141 → 1549  →  0.74446539192060168232956129941157997 → 1566  →
350→3.27038569744009111062155170657206327 → 634→0.74206749808107164145809520050534047 → 1551  →  0.74498163312120158074947084969608113 → 1585  →
360→3.27032823598987443247665396484080702 → 599→0.73455111595551458858466276069520973 → 1568  →  0.73578124057474614261309397988952696 → 1584  →
370→3.38638387038402477458021166967228055 → 641→0.76972859749693856024066462850896642 → 1530  →  0.76438341552242161647257034928770736 → 1516  →
380→3.53888220195532010592387450742535293 → 532→0.78437356488822251865400403403327800 → 1491  →  0.75393447671199864679181246174266562 → 1542  →
390→3.45793601682768869665096644894219935 → 605→0.77849170264050959744395186135079712 → 1496  →  0.75197062842280204542078081431100145 → 1541  →
400→3.46355152928596421446627573459409177 → 618→0.77612176116559561389607324599637650 → 1491  →  0.76164421183914055024644085278850980 → 1525  →
```

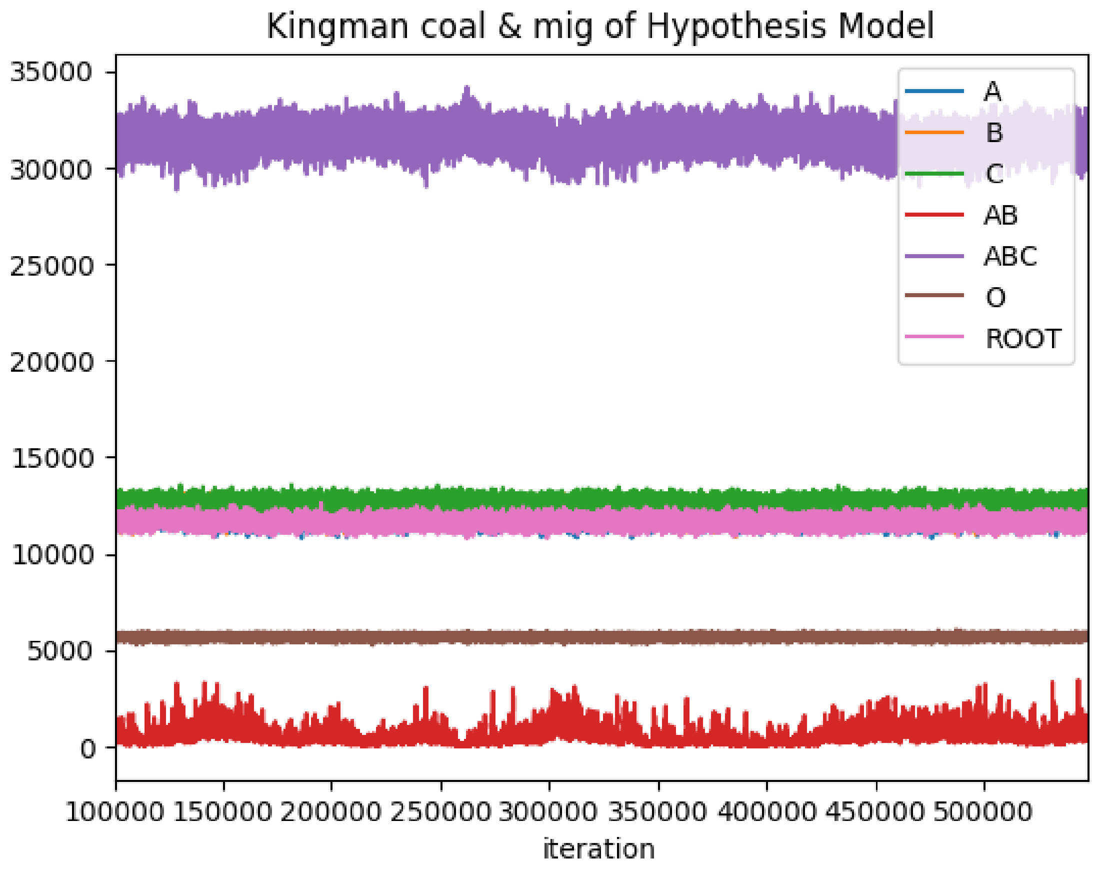
Sample McRef *config.ini* for a comb reference model from experiment II

```
[ReferenceModel]
comb = ABC
hyp_pops = 0,ROOT
comb_leaves = A,B,C
hyp_mig_bands =

[Input]
trace_file = ./trace.tsv
comb_stats_file = ./comb-trace.tsv
clade_stats_file = ./clade-trace.tsv
hyp_stats_file = ./hyp-trace.tsv
tau_bounds_file = ./tau-bounds.tsv
tau-theta-print = 10000.0
tau-theta-alpha = 1.0
tau-theta-beta = 10000.0
mig-rate-print = 0.001

[Output]

[Data]
skip_rows = 100000
number_of_rows = 400000

[Debug]
enabled = true
hypothesis_pops = A,B,C,AB,ABC,0,ROOT
hypothesis_migbands =
```

Sample McRef Output from experiment II

```
=== Running McRef with ABC_COMB reference model ===

simulation                                                              rbf_mean    rbf_bootstrap      hm_mean    hm_bootstrap
-------------------------------------------------------------------  -----------  ---------------  -----------  --------------
/home/ron/Desktop/data_sets/M4.divAB_18-1/results/AB_C_O/seed_12345  -1600.53          12.7698     1.45392e+06        23.2929
/home/ron/Desktop/data_sets/M4.divAB_18-1/results/AB_C_O/seed_54321  -1590.44          15.6897     1.45388e+06         2.09028
/home/ron/Desktop/data_sets/M4.divAB_18-1/results/A_BC_O/seed_12345      9.32425        0.0690867  1.45391e+06         6.10843
/home/ron/Desktop/data_sets/M4.divAB_18-1/results/A_BC_O/seed_54321      9.55824        0.15798    1.45391e+06         5.02275
/home/ron/Desktop/data_sets/M4.divAB_18-1/results/AC_B_O/seed_12345      9.74785        0.0713541  1.45391e+06         5.4703
/home/ron/Desktop/data_sets/M4.divAB_18-1/results/AC_B_O/seed_54321      9.64641        0.0842066  1.45392e+06         5.1779
/home/ron/Desktop/data_sets/M4.divAB_18-2/results/AB_C_O/seed_12345  -1299.88           6.9838     1.45394e+06         6.78321
/home/ron/Desktop/data_sets/M4.divAB_18-2/results/AB_C_O/seed_54321  -1244.18          12.8009     1.45397e+06        11.09
/home/ron/Desktop/data_sets/M4.divAB_18-2/results/A_BC_O/seed_12345      9.01693        0.0534758  1.45394e+06         4.4059
/home/ron/Desktop/data_sets/M4.divAB_18-2/results/A_BC_O/seed_54321      9.17147        0.0697426  1.45398e+06        13.7233
/home/ron/Desktop/data_sets/M4.divAB_18-2/results/AC_B_O/seed_12345     10.6226         0.765409   1.45395e+06         2.07356
/home/ron/Desktop/data_sets/M4.divAB_18-2/results/AC_B_O/seed_54321      9.21732        0.172327   1.45396e+06         2.51385
/home/ron/Desktop/data_sets/M4.divAB_24-1/results/AB_C_O/seed_12345   -242.607          9.60566    1.45472e+06         4.47794
/home/ron/Desktop/data_sets/M4.divAB_24-1/results/AB_C_O/seed_54321   -283.387          3.4061     1.45475e+06        23.7023
/home/ron/Desktop/data_sets/M4.divAB_24-1/results/A_BC_O/seed_12345      9.22721        0.168712   1.45474e+06         7.01179
/home/ron/Desktop/data_sets/M4.divAB_24-1/results/A_BC_O/seed_54321      9.84558        0.0597326  1.45474e+06         8.57476
/home/ron/Desktop/data_sets/M4.divAB_24-1/results/AC_B_O/seed_12345      9.15577        0.111286   1.45471e+06         2.13128
/home/ron/Desktop/data_sets/M4.divAB_24-1/results/AC_B_O/seed_54321      9.15118        0.100971   1.45473e+06        12.0132
/home/ron/Desktop/data_sets/M4.divAB_24-2/results/AB_C_O/seed_12345   -281.177          6.77911    1.45599e+06        12.6686
/home/ron/Desktop/data_sets/M4.divAB_24-2/results/AB_C_O/seed_54321   -185.257          5.12774    1.456e+06          13.2996
/home/ron/Desktop/data_sets/M4.divAB_24-2/results/A_BC_O/seed_12345      9.30068        0.132353   1.45602e+06        27.6855
/home/ron/Desktop/data_sets/M4.divAB_24-2/results/A_BC_O/seed_54321      9.21249        0.0485166  1.45598e+06         4.66902
/home/ron/Desktop/data_sets/M4.divAB_24-2/results/AC_B_O/seed_12345     10.1056         0.882446   1.456e+06          20.3955
/home/ron/Desktop/data_sets/M4.divAB_24-2/results/AC_B_O/seed_54321      8.54255        0.0600942  1.456e+06          13.4949
/home/ron/Desktop/data_sets/M4.divAB_30-1/results/AB_C_O/seed_12345      7.14187        0.194358   1.45669e+06        11.2392
/home/ron/Desktop/data_sets/M4.divAB_30-1/results/AB_C_O/seed_54321      7.11989        0.840854   1.45669e+06        11.7376
/home/ron/Desktop/data_sets/M4.divAB_30-1/results/A_BC_O/seed_12345      5.58964        0.162472   1.45669e+06         5.6343
/home/ron/Desktop/data_sets/M4.divAB_30-1/results/A_BC_O/seed_54321      9.80763        0.601917   1.45671e+06        28.8325
/home/ron/Desktop/data_sets/M4.divAB_30-1/results/AC_B_O/seed_12345      5.22395        0.406371   1.45667e+06         8.07406
/home/ron/Desktop/data_sets/M4.divAB_30-1/results/AC_B_O/seed_54321     -4.03428        0.420154   1.45668e+06         9.85529
/home/ron/Desktop/data_sets/M4.divAB_30-2/results/AB_C_O/seed_12345      4.62991        0.277489   1.45634e+06         7.22576
/home/ron/Desktop/data_sets/M4.divAB_30-2/results/AB_C_O/seed_54321      8.1412         0.16902    1.45636e+06         9.2261
/home/ron/Desktop/data_sets/M4.divAB_30-2/results/A_BC_O/seed_12345      8.427          0.359837   1.45638e+06        31.6903
/home/ron/Desktop/data_sets/M4.divAB_30-2/results/A_BC_O/seed_54321      8.32354        0.19312    1.45636e+06        17.9924
/home/ron/Desktop/data_sets/M4.divAB_30-2/results/AC_B_O/seed_12345      4.52845        0.618151   1.45635e+06         5.19023
/home/ron/Desktop/data_sets/M4.divAB_30-2/results/AC_B_O/seed_54321     -1.38435        0.835659   1.45635e+06         8.56084
```

Sample McRef debug plot of reference population genealogy likelihoods



תקציר

בזכות פריצות הדרך בריצוף גנטי בקצב גבוה השתפרה משמעותית יכולתנו לחקור את היסטוריית האבולוציה של מינים באמצעות מודלים דמוגרפיים מפורטים. גישה פופולרית להסקת פרמטרים של מודלים דמוגרפיים אלו היא לדגום גנאולוגיות מעל לוקוסים קצרים ובלתי תלויים, באמצעות אלגוריתמי שרשראות-מרקוב מונטה-קרלו (MCMC). השימוש של אלגוריתמים אלו במודלי התמזגות גנאולוגיות מפורשים מקנה להם כוח רב בתהליך הסקת פרמטרים דמוגרפיים, אך יכולתם לשערך את התאימות בין המודל לנתונים הגנטיים מוגבלת. מטרת מחקרנו היא לבחון גישה חדשה, המבוססת על גורמים בייסיאנים יחסיים, לניצול תהליכי דגימת הגנאולוגיות הללו לטובת השוואה, בחינה ובחירה בין מודלים אבולוציונים שונים.

בעבודה זו נסקור שיטות בייסיאניות להסקת פרמטרים דמוגרפיים ונתאר את בעיית בחירת המודל. לאחר מכן נגדיר גורמי בייס יחסיים (RBFs), המייצגים התאמה של מודל דמוגרפי לנתונים גנטיים, יחסית למודל השוואה. נפתח RBFs עבור שני טיפוסי מודלי השוואה - מודל ענף ומודל מסרק. טיפוסים אלו שימושיים עבור מופעים שונים של בעיית בחירת המודל. לאחר שנציג נוסחאות סגורות לחישוב RBFs, נתאר בפירוט איך אילו מחושבות באופן יעיל, תוך מזעור התקורה החישובית על תהליך הMCMC. לבסוף, נבחן את הRBFs בסדרת השוואות מודלים בעזרת דנא מסומלץ. בתוצאות שנציג ניכר כי ביצועי הRBFs טובים משמעותית מאילו של התוחלת ההרמונית במבחן השוואת מודלים דמוגרפיים.

עבודה זו בוצעה בהדרכתו של דר' אילן גרונאו מבי"ס אפי ארזי למדעי המחשב, המרכז
הבינתחומי, הרצליה.

הבינתחומי הרצליה
בית ספר אפי ארזי
למדעי המחשב

המרכז הבינתחומי בהרצליה
בית-ספר אפי ארזי למדעי המחשב
התכנית לתואר שני (M.Sc.) - מסלול מחקרי

שיטה בייסיאנית חדשה להשוואת מודלים פילוגנטיים

מאת
רון ויסבורד

עבודת תזה המוגשת כחלק מהדרישות לשם קבלת תואר מוסמך M.Sc.
במסלול המחקרי בבית ספר אפי ארזי למדעי המחשב, המרכז הבינתחומי הרצליה

אפריל 2018