COAL_PHYRE: A Composite Likelihood Method for Estimating Species Tree Parameters from Genomic Data Using Coalescent Theory

Geno Guerra[1,*], Rasmus Nielsen[1,2,3]

[1]Department of Statistics, University of California, Berkeley, California 94720, USA.

[2]Department of Integrative Biology, University of California, Berkeley, California 94720 USA.

[3]Lundbeck Foundations Centre for GeoGenetics, University of Copenhagen, Denmark.

[*] Corresponding author, email: gaguerra@berkeley.edu

1 Keywords

phylogenetics, species tree, maximum likelihood, composite likelihood, mutational variance, gene tree, estimation error, multispecies coalescent, incomplete lineage sorting, genomic data, sequence data

2 Abstract

Genome-scale data are increasingly being used to infer phylogenetic trees. A major challenge in such inferences is that different regions of the genome may have local topologies that differ from the species tree due to incomplete lineage sorting (ILS). Another source of gene tree discrepancies is estimation errors arising from the randomness of the mutational process during sequence evolution. There are two major groups of methods for estimating species tree from whole-genome data: a set of full likelihood methods, which model both sources of variance, but do not scale to large numbers of independent loci, and a class of faster approximation methods which do not model the mutational variance.

To bridge the gap between these two classes of methods, we present COAL_PHYRE (COmposite Approximate Likelihood for PHYlogenetic REconstruction), a composite likelihood based method for inferring population size and divergence time estimates of rooted species trees from aligned gene sequences. COAL_PHYRE jointly models coalescent variation across loci using the MSC and variation in local gene tree reconstruction using a normal approximation. To evaluate the accuracy and speed of the method, we compare against BPP, a powerful MCMC full-likelihood method, as well as ASTRAL-III, a fast approximate method. We show that COAL_PHYRE's divergence time and population size estimates are more accurate than ASTRAL, and comparable to those obtained using BPP, with an order of magnitude decrease in computational time. We also present results on previously published data from a set of Gibbon species to evaluate the accuracy in topology and parameter inference on real data, and to illustrate the method's ability to analyze data sets which are prohibitively large for MCMC methods.

3 Introduction

With the continued improvement of sequencing technologies, inferring evolutionary relationships between organisms using multi-gene sequences has become the standard in the field of phylogenetics. Bifurcating species trees are a common way to represent these relationships, with branching points representing speciation events. While a species tree represents the history of these species as a whole, trees in individual genome segments can have their own, potentially discordant, topology due to horizontal gene transfer, gene duplication/loss, and/or incomplete lineage sorting (ILS) [21]. The most ubiquitous of these, ILS, is of particular focus in the field [5], and can be well-modeled using the multi-species coalescent (MSC) (see e.g., [27]). Many methods exist to infer the species tree topology of a group of organisms using the MSC in the presence of ILS, and are shown to be statistically consistent assuming the gene tree topologies are known without error [15, 23, 20]. This assumption however is unrealistic, as gene trees typically are estimated from sequence data, with a finite amount of mutations present. The random process of mutation adds a second layer of variation among gene trees, and ignoring this can lead to poor method performance [10, 11, 14]. A class of Bayesian hierarchical methods exist, which jointly model gene and species tree topologies in a full likelihood framework (e.g. [17, 4, 19, 6, 9, 34]), and account for both coalescent and mutational variance, but these approaches have been shown to be computationally intensive (100s of hours) and not able to scale to large amounts of genes or species [18, 22, 31].

Although it has been known for decades that gene trees can differ in topology from an underlying species tree, a common approach to estimating trees and divergence times to avoid gene tree estimation error still relies on concatenated "super-matrices" of gene sequences (where multiple gene alignments are concatenated together to form one large "super gene"). Under high levels of mutational variation, this concatenation approach was justified as a way to pool information between highly noisy genes. [32, 16, 8] discuss results showing that concatenation-based approaches are not always outperformed by more ILS-sensitive methods. In short, concatenation methods seem to be predictably less accurate than coalescent based methods under high ILS (when there are short branches in the true species tree) and can even give high confidence to incorrect topologies [29]. Away from these scenarios, concatenation can empirically perform equal to or better than coalescent based methods. As such, concatenation is still widely used for inferring phylogenies in many empirical studies.

Divergence time estimates have become an essential addition in phylogenetic inference, as many studies utilize or require time-calibrated phylogenies, for example in biogeography, or in modeling of character evolution [3, 28, 25]. In particular, a challenging problem in phylogenetics is accurately inferring divergence times and population sizes in the presence of mutational variance. The Bayesian method, BPP [34] provides highly accurate results under the assumption of a molecular clock and the Jukes and Cantor model of sequence evolution [13]. However, this method, along with other Bayesian approaches, is unable to take advantage of the full information in genomic data sets, and must instead subdivide data into smaller (~ 100 segments) blocks of genes per run to perform inference in reasonable amounts of time.

In this paper, we present a coalescent based method to jointly infer species divergence times and ancient population sizes in the presence of mutational variance/gene tree estimation error. For a given topology, or set of k topologies, our method COAL-PHYRE (COmposite Approximate Likelihood for PHYlogenetic REconstruction) uses a composite likelihood approach to estimate tree parameters from DNA sequence data. COAL-PHYRE is able to analyze data with tens of thousands of genes/loci and multiple individuals in each sampled species. We show that the divergence time and population size estimates of COAL-PHYRE are comparable to the more time intensive estimates obtained using BPP [34], with at least an order of magnitude decrease in run time. We also compare to the popular approximate likelihood method ASTRAL-III [35], to compare the accuracy of our method against one that does not directly model mutational variance. Lastly, we analyze a data set of Gibbon species previously analyzed by BPP in [30], and find highly similar estimated parameters.

4 Methods

We consider a rooted bifurcating species tree $\mathcal{S} = (S, \tau, \eta)$ parameterized by topology S, divergence times τ, and population sizes η. See figure 1(a) for an illustration. Given a recombination-free region of the genome, l (interchangeably referred to as a gene or locus throughout), it is expected that that species tree topology S and the true local gene tree $\mathcal{G}_l$ will not always match due to incomplete lineage sorting (ILS), which is common when branch lengths are short relative to the effective population sizes. Let $\bar{g}_l$ represent an estimated rooted topology with branch lengths of the local ancestry from the region l. Note that $\bar{g}_l$ need not be bifurcating if the available genetic data is unable to resolve splits in the tree. This reconstructed gene tree is an estimate of the true local relationship between individuals, $\mathcal{G}_l$. For any finite amount of information, (number of pairwise mutational differences on l), there is estimation variance in $\bar{g}_l$. If $\bar{g}_l$ was known without error, meaning $\bar{g}_l = \mathcal{G}_l$, the MSC can be used to completely model the variation within and across gene trees, such as in STEM [15]. In reality, however, $\mathcal{G}_l$ cannot be reliably estimated without sampling variance, and accurate estimation of the species tree from a collection of estimated gene trees requires models accounting for both the distribution of $\bar{g}_l$ given $\mathcal{G}_l$, and $\mathcal{G}_l$ given $\mathcal{S}$.

Our goal is to incorporate the effect of mutational variance directly into the likelihood in an in-

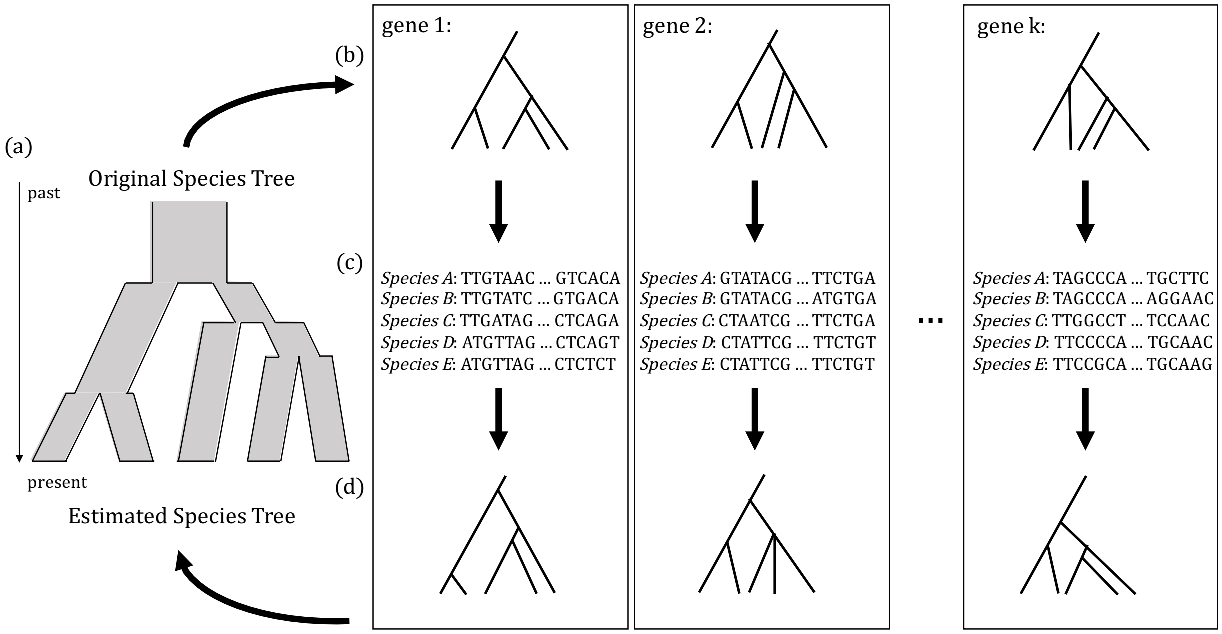
(a)
past
present
Original Species Tree
Estimated Species Tree
(b)
(c)
(d)
gene 1:
gene 2:
gene k:
Species A: TTGTAAC … GTCACA
Species B: TTGTATC … GTGACA
Species C: TTGATAG … CTCAGA
Species D: ATGTTAG … CTCAGT
Species E: ATGTTAG … CTCTCT
Species A: GTATACG … TTCTGA
Species B: GTATACG … ATGTGA
Species C: CTAATCG … TTCTGA
Species D: CTATTCG … TTCTGT
Species E: CTATTCG … TTCTGT
…
Species A: TAGCCCA … TGCTTC
Species B: TAGCCCA … AGGAAC
Species C: TTGGCCT … TCCAAC
Species D: TTCCCCA … TGCAAC
Species E: TTCCGCA … TGCAAG

Figure 1: **Contribution of coalescent and mutational variance.** (a) Original bifurcating species tree. (b) K gene trees, each a different realization of a stochastic lineage sorting process on the original species tree. (c) Sequences created from the mutational process on each gene tree. (d) Gene trees estimated from the sequence data, which can differ in topology and branch length from the true gene trees due to mutational variance.

terpretable way that is computationally tractable and scalable to many genes. We propose studying the observed distribution of individual coalescence times to do this.

We use the approximation that 'noisy' coalescence times (coalescence times estimated with mutational variation present) are well approximated by a hierarchical model of the MSC with an added normal distribution to capture both the coalescent and mutational variance, respectively. When coalescence times are estimated from sequence data, the layer of noise from gene tree reconstruction error (mutational variance) effectively smooths out the exponential-like distribution of the MSC, and the times fit closely to this hierarchical model.

Our method takes as input a set of aligned sequence data, and a rooted species tree topology (or set of topologies), and returns the inferred divergence times and population sizes which maximize the composite likelihood of pairwise coalescence times across the inputted loci, along with a likelihood, for each inputted topology. We assume there is no recombination within a locus, and allow free recombination between loci, and therefore assume loci are independent. To make use of the MSC, we assume the sequences have evolved on the gene tree under a molecular clock. Although not the goal of this paper, mutation rate variation between species can be incorporated into the gene tree estimation process if the computed gene trees have time measured in some real-time units as this satisfies the ultrametric property. We model each estimated pairwise coalescence time at a locus as an independent draw from a hierarchical MSC-normal distribution. The distribution of true coalescence times is modeled by the MSC, under a proposed species tree $\mathcal{S}$. Conditional on those times, the normal distribution is then parameterized using the approximated mutational variance, derived from properties of the Poisson distribution. Our goal is to infer a set of divergence times and population sizes that maximize the composite likelihood of the estimated gene trees.

4.1 Mutational variance

As is common in most species tree inference methods ([34, 23, 15] for example), we assume that genomic data can be divided into recombination-free regions, with free recombination between regions. At any given locus, l, the underlying true gene tree $\mathcal{G}_l$ (including branch lengths) is not known but can be estimated from aligned sequence data. This estimated gene tree $\bar{g}_l$ is a topology with estimated coalescence times.

For a specific time on the estimated tree, we can decompose the estimated time $\bar{g}_l(i)$ into a mixture of two components: the true coalescent time $\mathcal{G}_l(i)$, and then the estimation error resulting from having only a finite number mutations on each branch $\epsilon_l(i)$ (see figure 1). Mathematically, we can write this as:

$$\bar{g}_l(i) = \mathcal{G}_l(i) + \epsilon_l(i)$$

We approximate that error $\epsilon_l(i)$, the difference between the estimated and the (unknown) true coalescence time, as distributed with mean 0 and variance $\xi_l(i)$, i.e., we assume that an unbiased estimator has been used to estimate $\bar{g}_l(i)$. While $\mathcal{G}_l(i)$ can be modeled using the MSC, we use the Poisson distribution of mutations given a coalescence time to quantify the variance $\xi_l(i)$, meaning $\xi_l(i)$ is a function of the unknown true coalescence time $\mathcal{G}_l(i)$.

Under the infinite sites assumption, the number of mutations on a lineage is Poisson distributed and the variance in the estimate of the coalescence time will also follow that of a Poisson distribution. In real life applications, the divergence between sequences is often estimated using finite-sites models. However, even for these models the Poisson variance might be a reasonable approximation, and we will evaluate the performance of all estimators presented in this paper using simulations under finite sites models. The estimation variance from the mutation process is then

$$\xi_l(i) = \text{Var}(\bar{g}_l(i)|\mathcal{G}_l(i)) = \text{Var}\Big(\frac{k_l(i)}{\theta\mathcal{L}}|\mathcal{G}_l(i)\Big) = \frac{\text{Var}(k_l(i)|\mathcal{G}_l(i))}{\theta^2\mathcal{L}^2} = \frac{\theta\mathcal{L}\mathcal{G}_l(i)}{\theta^2\mathcal{L}^2} = \frac{\mathcal{G}_l(i)}{\theta\mathcal{L}} := \omega\mathcal{G}_l(i)$$

where $\omega = \frac{1}{\theta\mathcal{L}}$, $\mathcal{L}$ is the length of locus l, and $k_l(i)$ is the number of pairwise mutations for i on locus l .

While using the variance from the Poisson, we will approximate the sampling distribution of coalescence time estimates with a normal distribution for computational convenience. Figure 7 illustrates examples of distributions of estimated coalescence times produced under different mutation rates for a fixed locus and true coalescence time $\mathcal{G}_l(i)$, along with the variance approximated under a normal approximation. Further details for the normal approximation are given below.

4.2 The composite likelihood

The input for the algorithm are the haplotypes of M individuals, from K regions in the genome, $(\vec{h_1}, ..., \vec{h_K})$, where each $\vec{h_j}$ contains M haplotypes from locus j. We assume that the K genes are non-recombining blocks of the genome, and allow free recombination between genes. We allow for each locus to be of different length, and allow for missing characters in the sequences. The rooted gene tree topology, $\bar{g}_j$, of M individuals with branch lengths estimated from haplotypes $\vec{h_j}$ at locus j from the pairwise number of differences between the sequences. Each of the M individuals must also be assigned to be one of N present-day species.

We use a composite likelihood by maximizing the product of likelihoods of each independent gene tree:

$$L(\mathcal{S}|\{\bar{g}_1, ..., \bar{g}_K\}) = \prod_{j=1}^{K} f(\bar{g}_j|\mathcal{S})$$

To evaluate the likelihood of an estimated gene tree $\bar{g}_j$, $f(\bar{g}_j|\mathcal{S})$, we approximate it by the composite likelihood obtained as products of the individual likelihood functions. For M individuals in the tree $(M \geq N)$, we decompose the likelihood into Q univariate quantities:

$$f(\bar{g}_j|\mathcal{S}) = \prod_{i=1}^{Q} P_C(\bar{g}_j(i)|\mathcal{S})$$

where $Q = \binom{M}{2}$ is the number of pairs of individuals in the data set. We index each pair of individuals by a value i, $(i \in \{1, 2, ..., Q\})$, where $\bar{g}_j(i)$ is the estimated coalescence time of pair i on gene tree j. Note that these Q coalescence times are not all independent, as there are only $M - 1$ unique coalescence times on a tree of M individuals.

We model $P_C(\bar{g}_j(i)|\mathcal{S})$ with a zero-inflated MSC-normal hierarchical distribution. Due to the random process of mutation, the frequency of observing zero pairwise mutations at a locus needs to be explicitly modeled, as the MSC-normal distribution does not adequately account for the point mass at zero.

4.3 MSC-Normal distribution

For two individuals, a, b (indexed by i), the divergence time for the species A, B respectively ($a \in A$, $b \in B$) is denoted by τ_{AB}. For a given locus, we estimate a coalescence time $\bar{g}_j(i)$ for the pair, based on the estimated local gene tree. We know (assuming no recombination within the locus) that there is some underlying, but unknown, true coalescence time $\mathcal{G}_l(i)$.

We model the distribution of location-adjusted true coalescence times, $\mathcal{G}_l(i) - \tau_{AB}$, using the coalescent with piecewise constant population size history, with population sizes and times given by $\mathcal{S}_{AB}$. For notation's sake, we assume the history is a sequence of R population size- split time pairs $\mathcal{S}_{AB} = \{(\eta_0, \tau_0), \ldots, (\eta_{R-1}, \tau_{R-1})\}$, where $\eta_0 = \eta_{AB}$ and $\tau_0 = \tau_{AB}$. At each branch along the tree, we can calculate the likelihood of $\mathcal{G}_j(i)$ given the coalescence event occurs within the branch $\Big(\mathcal{G}_j(i) \in (\tau_r, \tau_{r+1})\Big)$, parameterized by the start and end times of the branch, along with the effective population size η_r. To get the overall likelihood of $\mathcal{G}_j(i)$, we sum over all the possible branches.

$$P(\mathcal{G}_j(i) = z, \mathcal{G}_j(i) \in (\tau_r, \tau_{r+1}) | \mathcal{S}) = P(\mathcal{G}_j(i) > \tau_r | \mathcal{S}) \frac{1}{2\eta_r} e^{-\frac{z-\tau_r}{2\eta_r}} \text{ for } z \in (\tau_r, \tau_{r+1})$$

$$P(\mathcal{G}_j(i) = z | \mathcal{S}) = \sum_{r=0}^{R-1} P(\mathcal{G}_j(i) = z, \mathcal{G}_j(i) \in (\tau_r, \tau_{r+1}) | \mathcal{S})$$

Assuming $\mathcal{G}_j(i) > \tau_{AB}$, and $\tau_0 = \tau_{AB}$.

Given $\mathcal{G}_j(i)$, we view the distribution of $\bar{g}_j(i)$ as normally distributed around mean $\mathcal{G}_j(i)$, with variance $\omega\mathcal{G}_j(i)$, as described earlier.

$$P(\bar{g}_j(i) = x | \mathcal{G}_j(i) = z, \omega) = \frac{1}{\sqrt{2\pi\omega z}} e^{-\frac{(x-z)^2}{2\omega z}}$$

Combining these distributions, we have

$$\begin{aligned} P(\bar{g}_j(i) = x, \mathcal{G}_j(i) = z | \mathcal{S}, \omega) &= \sum_{r=0}^{R-1} P(\bar{g}_j(i) = x | \mathcal{G}_j(i) = z, \omega) P(\mathcal{G}_j(i) = z, \mathcal{G}_j(i) \in (\tau_r, \tau_{r+1}) | \mathcal{S}) \\ &= \sum_{r=0}^{R-1} P(z > \tau_r | \mathcal{S}) \frac{1}{\sqrt{2\pi\omega z}} e^{-\frac{(x-z)^2}{2\omega z}} \frac{1}{2\eta_r} e^{-\frac{z-\tau_r}{2\eta_r}} \end{aligned}$$

To get the marginal distribution of estimated coalescence times, we need to integrate over the latent variable, $\mathcal{G}_j(i)$, the true coalescence time, which takes values in (τ_{AB}, ∞)

$$\begin{aligned} P(\bar{g}_j(i) = x | \mathcal{S}, \omega) &= \int_{\tau_{AB}}^{\infty} P(\bar{g}_j(i) = x, \mathcal{G}_j(i) = z | \mathcal{S}, \omega) dz \\ &= \int_{\tau_{AB}}^{\infty} \sum_{r=0}^{R-1} P(z > \tau_r | \mathcal{S}) \frac{1}{\sqrt{2\pi\omega z}} e^{-\frac{(x-z)^2}{2\omega z}} \frac{1}{2\eta_r} e^{-\frac{z-\tau_r}{2\eta_r}} dz \\ &= \sum_{r=0}^{R-1} P(z > \tau_r | \mathcal{S}) \int_{\tau_r}^{\tau_{r+1}} \frac{1}{\sqrt{2\pi\omega z}} e^{-\frac{(x-z)^2}{2\omega z}} \frac{1}{2\eta_r} e^{-\frac{z-\tau_r}{2\eta_r}} dz \\ &= \sum_{r=0}^{R-1} P(z > \tau_r | \mathcal{S}) \frac{\omega\Omega(r)}{4(\omega + \eta_r)} e^{\frac{\tau_r}{2\eta_r}} \left[e^{-x\Omega(r)} \big(\zeta(\tau_r) - \zeta(-\tau_{r+1}) \big) - e^{x\Omega(r)} \big(\zeta(\tau_r) - \zeta(\tau_{r+1}) \big) \right] \end{aligned}$$

Where

$$\Omega(r) = \sqrt{\frac{\omega + \eta_r}{\omega^2 \eta_r}}$$

$$\zeta(t) = \text{erf}\Big(\frac{t\ \omega\Omega(r) + x}{\sqrt{2}\sqrt{|t|}\sqrt{\omega}} \Big), \text{ with } \zeta(0) = 1$$

$$P(z > \tau_r | \mathcal{S}) = \sum_{l=0}^{r-1} e^{-\frac{\tau_{l+1} - \tau_l}{2\eta_l}}$$

$$\text{erf}(q) = \frac{2}{\sqrt{\pi}} \int_q^{\infty} e^{-y^2} dy$$

4.4 Accounting for no observed mutations

In studying sequence data it is common to encounter genes where two or more individuals have identical sequences, especially when genes are short, or the individuals are of the same species. In constructing a gene tree with no mutations between the two, this pair of individuals would have an estimated coalescence time of 0. For a given pair of individuals (indexed by i on the tree), we can calculate $P_0(\bar{g}_j(i) = 0|\mathcal{S}, \omega)$, using the MSC and a Poisson distribution of the mutation process. From the Poisson, for a given coalescence time, $\mathcal{G}_j(i)$, the probability of observing no mutations on the branch of length $2\mathcal{G}_j(i)$ is $p(\bar{g}_j(i) = 0|\mathcal{G}_j(i) = z, \omega) = e^{-z/\omega}$.

To obtain the unconditional probability of observing 0 mutations, we need to integrate over all of the possible values of the underlying (and unknown) true gene tree coalescence time, $\mathcal{G}_j(i) \in (0, \infty)$:

$$P_0(\bar{g}_j(i) = 0|\mathcal{S}, \omega) = \int_0^\infty p(\bar{g}_j(i) = 0|\mathcal{G}_j(i) = z, \omega)p(\mathcal{G}_j(i) = z|\mathcal{S})dz$$

We break the integral into regions of constant population size, indexed by $r \in \{0, ..., R-1\}$ and evaluate them separately.

$$\begin{aligned}
P_0(\bar{g}_j(i) = 0|\mathcal{S}, \omega) &= \sum_{r=0}^{R-1} P(\mathcal{G}_j(i) > \tau_r|\mathcal{S}) \int_{\tau_r}^{\tau_{r+1}} P(\bar{g}_j(i) = 0|\omega, \mathcal{G}_j(i) = z)P(\mathcal{G}_j(i) = z|\mathcal{S}, \tau_r)dz \\
&= \sum_{r=0}^{R-1} P(\mathcal{G}_j(i) > \tau_r|\mathcal{S}) \int_{\tau_r}^{\tau_{r+1}} \frac{1}{2\eta_r} e^{-z/\omega} e^{-\frac{z-\tau_r}{2\eta_r}} dz \\
&= \sum_{r=0}^{R-1} P(\mathcal{G}_j(i) > \tau_r|\mathcal{S}) \Big[\frac{1}{2\eta_r\omega + 1}\big(e^{-\tau_r/\omega} - e^{-\frac{(\tau_{r+1}-\tau_r)}{2\eta_r} - \tau_{r+1}/\omega}\big)\Big]
\end{aligned}$$

Where τ_0 is the species divergence time for the pair of individuals indexed by i. Calculating the quantity gives us the probability of encountering no mutations between pair i on gene j given species tree $\mathcal{S}$, gene length $\mathcal{L}$, and scaled mutation parameter θ. To distinguish this probability from the MSC-Normal distribution also presented above, we subscript the probability with a zero, $P_0(\bar{g}_j(i) = 0|\mathcal{S}, \theta, \mathcal{L})$, and write the complete likelihood as

$$P_C(\bar{g}_j(i) = x|\mathcal{S}, \omega) = \begin{cases} P_0(\bar{g}_j(i) = 0|\mathcal{S}, \omega) & \text{if } \; x = 0 \\ P(\bar{g}_j(i) = x|, \mathcal{S}, \omega) & \text{if } x > 0 \end{cases}$$

4.5 Likelihood weighting

In the composite likelihood, identical information is repeatedly used in multiple probability calculations. For a given node in a gene tree, let n_1 be the number of individuals on one side of the split, and n_2 be the number on the other. Being based on pairwise events, the composite likelihood would then use the information of that node split time $n_1 \times n_2$ times, which can become a large number for nodes deep in a gene tree. We apply a weight to the terms of the likelihood to down-weight this redundant use of information. As we do not observe the gene trees beforehand, we rely on the species tree topology to create the weight values. For a pair of individuals, $i = (i_1, i_2)$, $V(i)$ denotes the split on the tree such that i_1 is on one side of the split, and i_2 is on the other. Given $V(i)$, denote $n_1(i)$ and $n_2(i)$ to be the number of individuals on each side of the branch, such that $n_1(i) \times n_2(i)$ is the number of pairs of individuals who share the same split at $V(i)$. Define weight

$$w_V(i) = \frac{1}{n_1(i)n_2(i)}$$

such that, for a given split $V(i)$,

$$\sum_{j|V(j)=V(i)} w_V(j) = 1$$

where j indicates a pair of individuals (j_1, j_2) that share the same split event $V(i)$. We apply this weight to each term in the composite likelihood,

$$P_C(\bar{g}_j(i)|\mathcal{S}, \omega)^{w_V(i)}$$

so that the weight of information applied to each split on the species tree is equivalent.

It should be noted that these weights are only used in parameter inference, as using weights which depend on the topology can be problematic when comparing topologies. COAL_PHYRE is able to run with and without the weights applied.

4.6 Data simulation

To test the effectiveness of parameter inference of COAL-PHYRE, we conduct simulation studies under varying species tree topologies, divergence times, population sizes, mutation rates, and data set sizes. We simulate gene trees using ms [12] under a bifurcating species tree with piece-wise constant population size and no gene flow or migration after split. For consistency with the assumptions of BPP, we simulate the mutation process using the Jukes and Cantor mutation model [13] through Seq-Gen [26] to produce haplotypes under various mutation rates to introduce varying levels of mutational variance. See Appendix C for more details on the simulations. Although we use a simple model of evolution with a Jukes and Cantor model, performance using other models will likely be similar as long as gene tree estimation is done under the same model as used for simulation.

5 Simulation Results

5.1 5 species asymmetrical tree

We simulate a tree of 5 taxa, with asymmetric topology $(5, (4, (1, (3, 2))))$, where species 5 is the outgroup, and 2 individuals sampled per species. The population size within a branch is simulated to be constant, but different between branches, see Appendix section C for exact simulation details. We compare our method, COAL_PHYRE, to BPP [34] and ASTRAL-III [35]. COAL_PHYRE and BPP provide separate estimates of divergence times and population sizes, while ASTRAL-III provides estimates of the coalescence rate of each branch (coalescence rate = branch length/ population size), and does not attempt separate the parameters further. To accommodate the comparatively slow run time of the MCMC-based BPP, we simulate only 100 independent loci for each replicate. It should be noted that COAL_PHYRE can handle much larger sets of genes with only modest increases in run-time. For this data of 5 species, BPP and COAL_PHYRE provide estimates of all 4 split times, as well as the 9 separate population sizes (5 modern-day species and 4 ancestral populations). ASTRAL provides an estimate of 4 external branch lengths, and 2 internal. For each method, we provide as input the known species tree topology, and allow for parameter inference under the true topology. Note that BPP and COAL_PHYRE take as input the sequence data directly, but ASTRAL requires gene trees to be provided. As these simulations use the molecular clock, we use UPGMA (unweighted pair group method with arithmetic mean) to reconstruct gene trees as input to ASTRAL. We simulate under two different mutation rates, $\theta = 0.01$, and $\theta = 0.001$ (here $\theta = 4\eta_0\mu$ where μ is the per generation per base pair mutation rate), representing both high and low levels of mutation, with each locus chosen to be 1000 bp long. Under the $\theta = 0.01$ simulation, the the variance in the estimate of coalescence times is higher than for $\theta = 0.01$ due to the increased mutational noise.

We simulated 40 separate replicates under the two mutation rates, and used COAL_PHYRE, ASTRAL-III, and BPP to evaluate the accuracy of parameter reconstruction. The results of the estimation from all three methods can be seen in figure 2.

We can see that the performance of ASTRAL deteriorates under the low mutation rate model, as the method assumes gene trees are estimated without error, which is violated when the amount of phylogenetic signal in each gene is low. Divergence time estimates are nearly identical between COAL_PHYRE and BPP in the 0.01 mutation rate setting. Under the lower mutation rate, COAL_PHYRE tends to have higher variance and uncertainty in estimating divergence times than BPP. However, it is, similarly to BPP, approximately unbiased. Population size estimates are again nearly identical between COAL_PHYRE and BPP under the 0.01 mutation rate setting. For a lower mutation rate (0.001), the two methods are nearly identical in accuracy for the external population sizes ($\eta_1 \ldots \eta_5$) and COAL_PHYRE has more uncertainty than BPP in estimation of internal population sizes, reflecting the well-known challenge of disentangling internal branch lengths from population sizes.

When comparing run times, ASTRAL completed on average in about 1 second per replicate, much faster than either COAL_PHYRE or BPP, but requires pre-computed gene trees before running. COAL_PHYRE outputs results for each replicate in, on average, 1 minute whereas BPP required $\sim 10 - 20$ minutes to converge, both using a single-core on a standard laptop.

5.2 8 species symmetrical tree

Here we simulate a balanced tree topology of 8 species with 2 diploid individuals sampled per species. We simulate under the assumption of constant population size within each branch, but population sizes vary

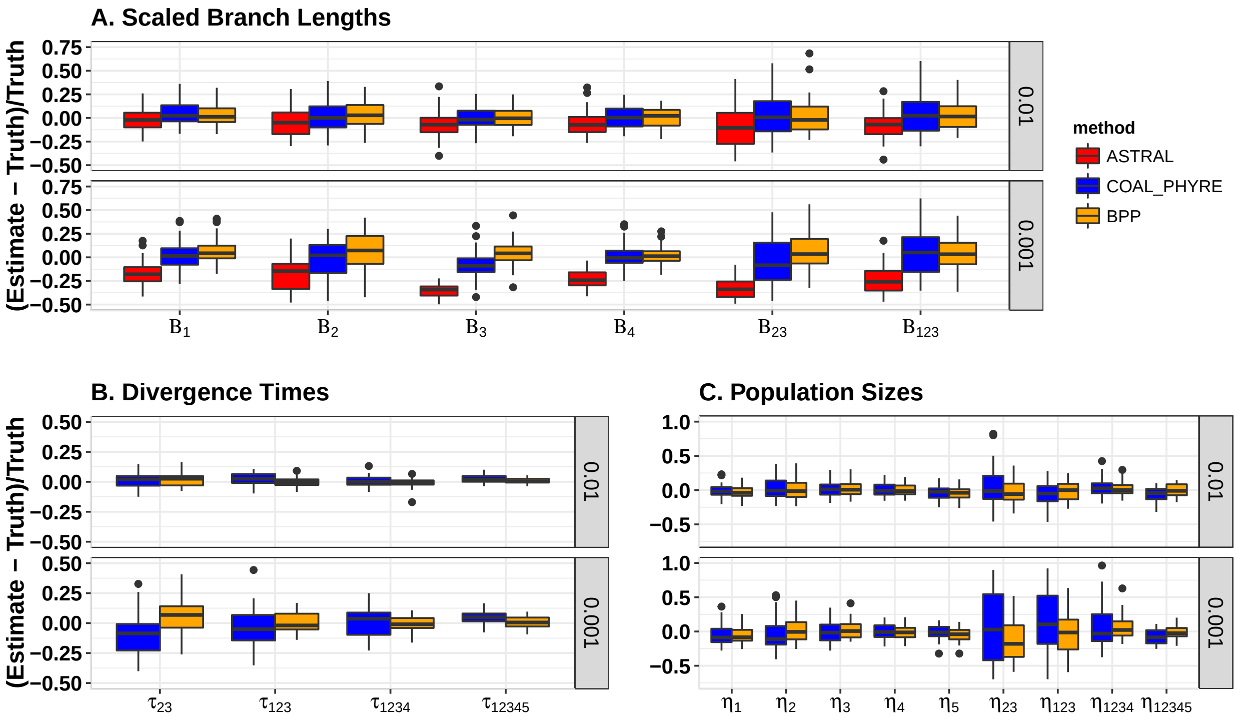

Figure 2: **Full parameter estimation for the fixed species tree topology (5,(4,(1,(2,3)))).** Comparison of parameter estimates between COAL_PHYRE, ASTRAL and BPP by branch over 40 iterations, using 100 independent loci each iteration. The y-axis gives the standardized deviation from the true parameter value. In each panel, the top plot represents a high mutation rate setting, where mutational variance is low, and the bottom represents a ×10 lower mutation rate, where mutational variance is larger. **A)** A comparison of estimated scaled branch lengths (branch length divided by population size) for the three methods. Only branches for which ASTRAL can provide an estimate are included. **B)** A comparison of divergence time estimates between COAL_PHYRE and BPP. **C)** A comparison of population size estimates between COAL_PHYRE and BPP.

among branches [RN: insert reference to where full details can be found]. Again, we compare COAL_PHYRE to BPP [34], and ASTRAL-III [35]. We simulate 100 independent sequences in each replicate, to compare against BPP at a reasonable run time. Both COAL_PHYRE and BPP can provide estimates of all 7 divergence times, and 15 population sizes (8 modern day, and 7 ancestral). ASTRAL only provides estimates for the leaf population branch lengths, and internal branches which are not directly adjacent to the ancestor of all species in the tree, (so not branch ”1234” or ”5678”). For BPP and COAL_PHYRE we provide as input the sequence data, the mutation rate, and the known species tree topology. To use ASTRAL, we provide a file of gene trees, pre-estimated using UPGMA, as well as the known species tree topology.

We simulate under two different mutation rates $\theta = 0.01$ and $\theta = 0.001$ (see above 5 species simulation for discussion on units), with each sequence simulated to be 1000 bp long (Figure 3). Similarly to the 5

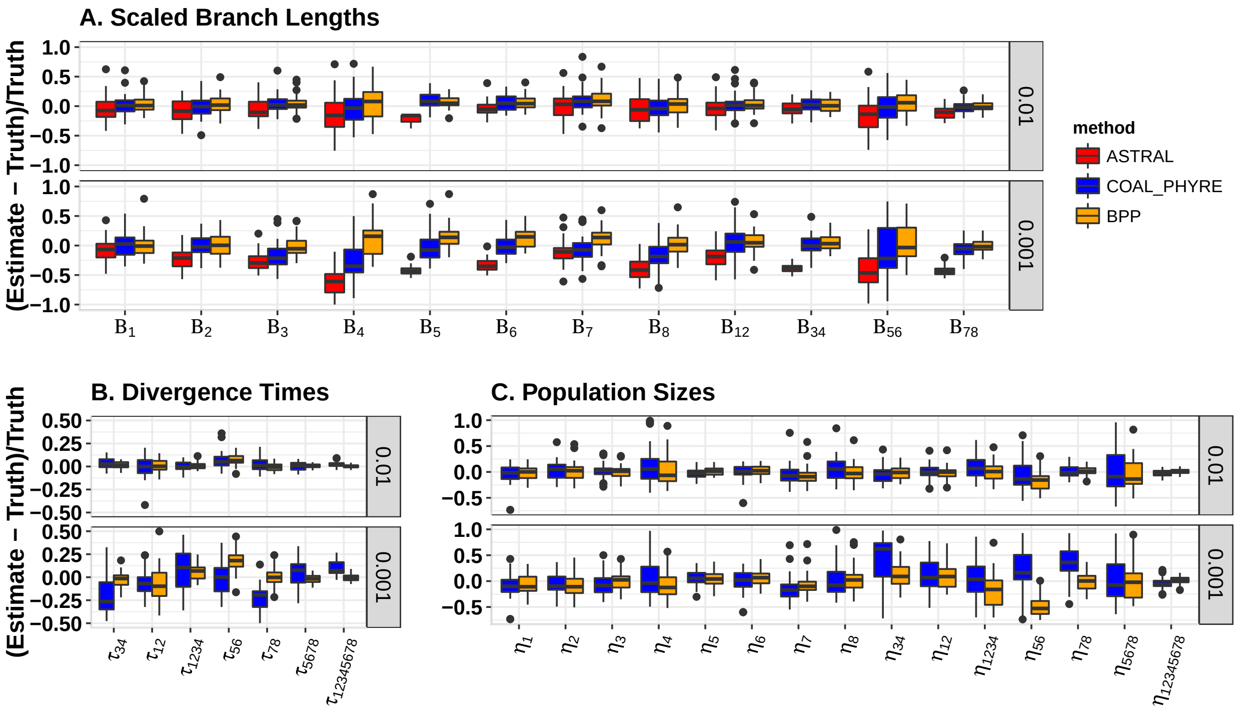

Figure 3: **Full parameter estimation for the fixed species tree topology (((1,2),(3,4)),((5,6),(7,8))).** Comparison of the parameter estimation accuracy between COAL_PHYRE (blue) and BPP (orange), and ASTRAL-III (red) using 100 independent genes, across 40 independent replicates. **A)** A comparison of estimated scaled branch lengths (branch length divided by population size) for the three methods. Only branches for which ASTRAL can provide an estimate are included. **B)** A comparison of divergence time estimates between COAL_PHYRE and BPP. **C)** A comparison of population size estimates between COAL_PHYRE and BPP.

species simulation, the branch length estimates of ASTRAL are biased downwards for the low mutation rate setting. As both COAL_PHYRE and BPP explicitly model the mutational noise, they do not experience the same bias. BPP and COAL_PHYRE demonstrate approximately the same level of performance at estimating divergence times and population sizes in the species tree. In particular, both methods provide highly accurate estimates of the leaf branch population sizes (η_1,...η_8). On a single-core laptop computer, COAL_PHYRE completed each of the replicates in 3-10 minutes. We were able to run BPP in approximately 30-60 minutes per replicate. We note that we allow BPP to complete under the recommended settings.

6 Analysis of Gibbon Data

Here we analyze two full-genome data sets from [2] and [33] of four gibbon species: (*Hylobates moloch* (Hm), *Hylobates pileatus* (Hp)), *Nomascus leucogenys* (N), *Symphalangus syndactylus*(S), and *Hoolock leuconedys*

(B). Gibbons (Hylobatidae), close relatives to humans and great apes, are found throughout Southeast Asia's tropical forests. A recent study, Shi and Yang, 2017[30] (hereby referred to as SY17) used the MCMC program, BPP ([34]), along with a suite of other methods, to attempt to resolve the phylogenetic relationship of these species. The results of the study show there are two most likely species tree topologies, (H, (N, (B, S))), which we will call Tree 1, and (N, (H, (B, S))), denoted by Tree 2. The authors also reported estimates for the population sizes and divergence times on the trees. (Note H= (Hm, Hp) indicating two subpopulations of the *Hylobates* species).

6.1 The data

The first data set (Noncoding) consists of 12,413 loci, each of 1,000 bp in length. The second data set (Coding) consists of 11,323 coding loci, each of 200bp in length. Within each data set one human haplotype (O) is used as an outgroup. There are a total of 17 haplotypes at each locus, with two diploid individuals from each Gibbon population, allowing for the estimation of leaf population sizes. See SY17 for a more detailed description of the data.

6.2 Results

We use COAL_PHYRE to analyze each of these data sets to provide a likelihood for each of the two topologies, and estimates of the divergence times and population sizes for each tree. To compare with the results of BPP we assume the JC69 [13] model of mutation. As well, we use mutation rate parameters consistent with the means of the Gamma priors used in SY17.

6.2.1 Divergence time and population sizes estimates

The parameter values estimated using COAL_PHYRE, along with those previously estimated in SY17 are presented in Tables 1, 2, 3, 4. In each scenario, we found that COAL_PHYRE assigned the highest likelihood to Tree 1, topology (H, (N, (B, S))), consistent with the findings in SY17. Also, note that population sizes are not reported for the human out group O, as only one haplotype was used, and so no information is available to estimate η_O.

Under the most likely topology (Tree 1) our estimates of the parameters are overall quite similar between coding and noncoding data sets, providing some evidence of internal consistency. To verify this, as suggested in SY17, we fit a regression line, $y = bx$ between the 5 parameter points (each point a pair of τ divergence time estimates, one from the noncoding dataset, the other from coding) to measure the internal consistency of the estimates from COAL_PHYRE. Our analysis under Tree 1 finds $\tau_{(C)} = 0.69\tau_{(NC)}$ with $r^2 = 0.988$. This demonstrates that our timing estimates are consistent between the two data sets, and that the mutation rate of the coding data is about 2/3 the rate of the non coding loci. SY17 found a rate of 0.73 with $r^2 = 0.985$, from their analysis. For the population size estimates (η's) of the leaf populations (B, S, N, Hm, Hp) we find $\eta_{(C)} = 0.95\eta_{(NC)}$ with a correlation of $r^2 = 0.995$ compared to $r^2 = 0.986$ from SY17.

We can also compare the correlation between our results and the results from BPP. Divergence time estimates for the (H,(N,(B,S))) coding data set show an $r^2 = 0.999$ between the divergence times estimated between the two methods, with $\tau_{\text{COAL\_PHYRE}} = 0.81\tau_{\text{BPP}}$. For the noncoding data set and tree (H,(N,(B,S)), we find an $r^2 = 0.9988$ with $\tau_{\text{COAL\_PHYRE}} = 0.94\tau_{\text{BPP}}$. When comparing the leaf population sizes we find for the coding data set, $\eta_{\text{COAL\_PHYRE}} = 1.43\eta_{\text{BPP}}$ with $r^2 = 0.995$. For the noncoding data set we find $\eta_{\text{COAL\_PHYRE}} = 0.97\eta_{\text{BPP}}$ with $r^2 = 0.998$.

We observe that our parameter estimates overall agree with the results of BPP, differing mainly in estimation of internal population sizes. The largest discrepancies occur on the (N,(H,(B,S)) tree (tree 2), which demonstrates how the two methods handle fitting parameters to a potentially incorrect topology. We acknowledge that SY17 observed BPP had mixing issues for such a large data set, and parameter estimation with short branch lengths can become highly variable. The extremely high population size estimate (which we write as "inf") of η_{HBS} in the noncoding tree 2 (N(H(B,S))) indicates that COAL_PHYRE attempts to model extremely high ILS in the HBS branch, attempting to fit a zero-probabilty of coalescence in that interval.

Each of the four tables demonstrates one run of COAL_PHYRE, which on a single core is able to run on average in 10(±5) hours. As reported in SY17, BPP took approximately 200 hrs for each analysis on a single core using the same data as COAL_PHYRE.

6.2.2 Predicted distribution of estimated coalescence times

Parameters on the species trees are estimated to best match the distribution of estimated coalescence times in the data, according to some likelihood function. In this section we assess the fit of the predicted distribution of estimated pairwise coalescence times of the Gibbon data when using the zero inflated MSC-Normal distribution implemented in COAL_PHYRE.

For a given set of tree parameters (topology, times and population sizes), we can study the resulting marginal distributions of estimated times. As we have two sets of tree parameters for each scenario, one from each method, we can compare the distributions predicted by each against the distribution of estimated times from data.

We specifically study the most likely tree topology, Tree 1 (H,(N,(B,S))), parameterized by the sets of divergence times and population sizes from Tables 1 and 2 (see Figures 5 and 4, respectively). Using the parameter values estimated by both methods, we can compare the predicted distribution under each set of parameters against the actual sampled distribution from the estimates across loci, and against one another to assess a level of 'best fit' to the data.

Figure 4 shows the distribution of binned estimated pairwise coalescence times from the data, along with the predicted distributions using the parameters of both COAL_PHYRE and BPP for the noncoding data set under Tree 1. From the plot, we can see that the predicted distributions between the two methods agree almost exactly in each panel. Figure 5 is the same approach, using the coding dataset.

Across all distributions of estimated coalescence times, it is expected that COAL_PHYRE should fit the data as well or better than the parameters from BPP, as the parameters inferred by COAL_PHYRE are estimated to fit specifically this likelihood.

Each plot also shows the predicted fraction of sequences that have no pairwise differences, as well as the observed frequency of zeros in the data. Comparing the parameters from COAL_PHYRE and BPP on the accuracy of predicting the fraction of zeros shows that BPP is slightly more accurate in this respect, on average.

Overall, the parameters inferred by each method fit the shape of the distribution of estimates well.

6.2.3 Run times

Each of the four tables demonstrates one run of COAL_PHYRE, which on a single core is able to run on average in 10(±5) hours. As reported in SY17, BPP took approximately 200 hrs for each analysis on a single core using the same data as COAL_PHYRE.

Table 1: Table of Results: (H, (N, (B, S))) Coding

Method	τ_{BS}	τ_{H}	τ_{NBS}	τ_{HNBS}	τ_{OHNBS}	η_{B}	η_{S}	η_{Hm}	η_{Hp}	η_{N}	η_{BS}	η_{H}	η_{NBS}	η_{HNBS}	η_{OHNBS}
COAL_PHYRE	1.65	0.96	2.11	2.12	10.87	0.91	1.12	1.22	0.70	1.61	11.84	1.97	0.18	3.27	8.41
BPP	2.13	0.8	2.7	2.75	11.9	0.6	0.8	0.9	0.4	1.2	26.7	2.1	10.4	1.9	7.8

Table 2: Table of Results: (H, (N, (B, S))) Noncoding

Method	τ_{BS}	τ_{H}	τ_{NBS}	τ_{HNBS}	τ_{OHNBS}	η_{B}	η_{S}	η_{Hm}	η_{Hp}	η_{N}	η_{BS}	η_{H}	η_{NBS}	η_{HNBS}	η_{OHNBS}
COAL_PHYRE	2.28	1.11	3.98	4.48	15.17	0.90	1.18	1.21	0.62	1.81	175.32	3.83	17.53	2.01	5.1
BPP	3.65	1.6	3.75	4.6	15.4	0.9	1.3	1.3	0.6	1.9	6.7	2.5	16.4	2.4	5.5

Table 3: Table of Results: (N, (H, (B, S))) Coding

Method	τ_{BS}	τ_{H}	τ_{HBS}	τ_{NHBS}	τ_{ONHBS}	η_{B}	η_{S}	η_{Hm}	η_{Hp}	η_{N}	η_{BS}	η_{H}	η_{HBS}	η_{NHBS}	η_{ONHBS}
COAL_PHYRE	1.66	1.04	1.82	2.14	10.85	0.91	1.13	1.27	0.73	1.61	3.87	1.43	24.87	3.22	8.43
BPP	1.9	1.0	3.0	3.05	11.5	0.6	0.8	0.8	0.4	1.2	22.3	2.0	2.6	1.9	7.8

Table 4: Table of Results: (N, (H, (B, S))) Noncoding

Method	τ_{BS}	τ_{H}	τ_{HBS}	τ_{NHBS}	τ_{ONHBS}	η_{B}	η_{S}	η_{Hm}	η_{Hp}	η_{N}	η_{BS}	η_{H}	η_{HBS}	η_{NHBS}	η_{ONHBS}
COAL_PHYRE	2.29	1.12	4.16	4.45	15.17	0.9	1.18	1.22	0.62	1.82	3.73	112.32	inf	2.00	5.10
BPP	3.75	1.6	4.3	4.8	15.25	0.9	1.3	1.3	0.6	2.0	12.5	2.3	14.4	2.5	5.5

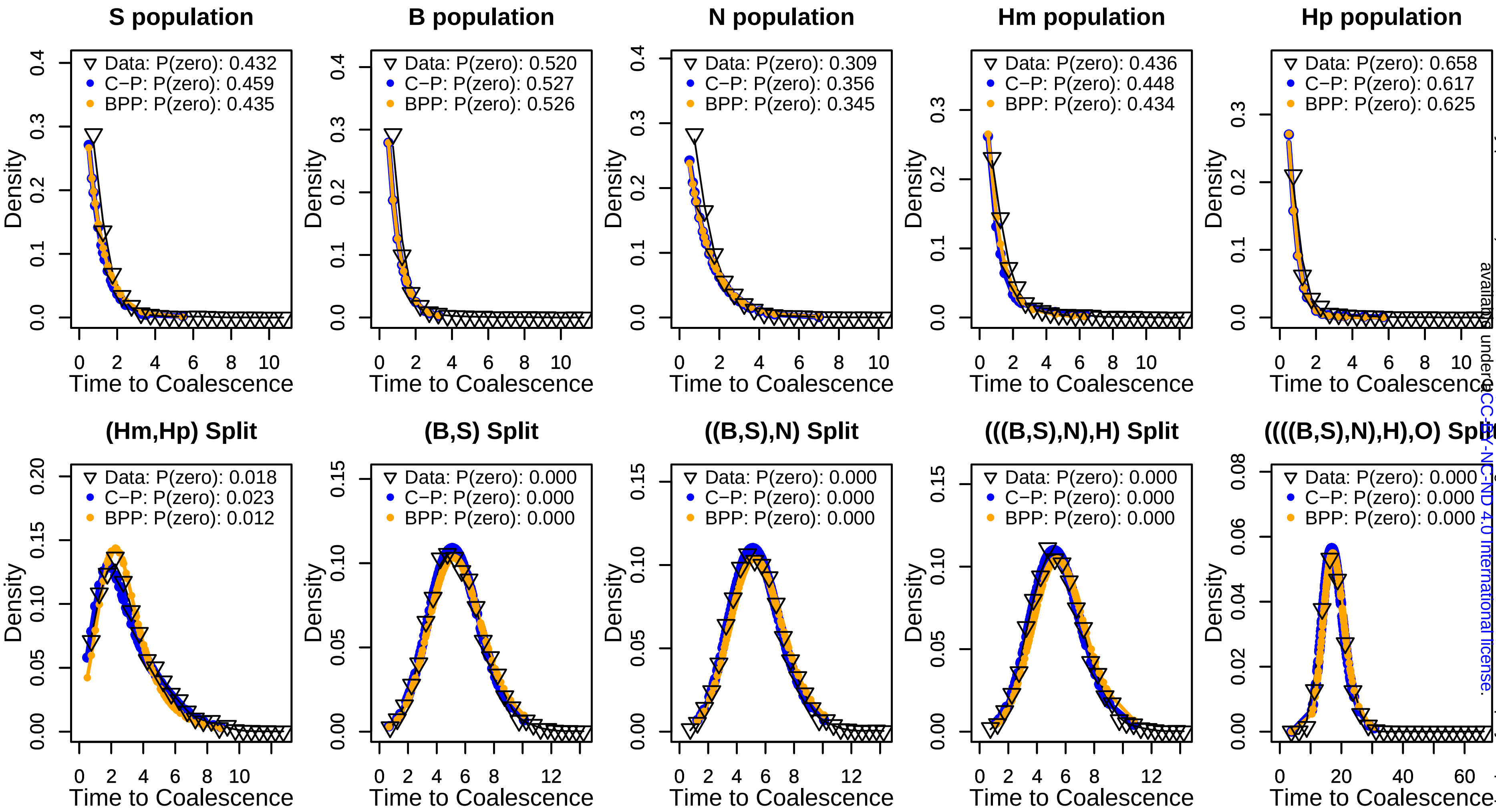

Figure 4: **Distribution of Estimated Coalescent Time: Tree 1 of Noncoding Gibbon data.** Comparison of zero-inflated MSC Normal distributions of estimated coalescence times using parameters inferred by COAL_PHYRE (Blue) and BPP (Orange) along with the distribution of estimated coalescence times from the data (black triangles) for each proposed split event. The top row: for two individuals sampled from the same population (as indicated by the plot header), the distribution of the estimated coalescence times. The bottom row: for a given split event, e.g. "((B,S),N) Split", the distribution plots the estimated time to coalescence for an individual sampled from the (B,S) subset, and one from the (N) subset. The figure legend also includes the observed (or predicted) fraction of zeros in the data.

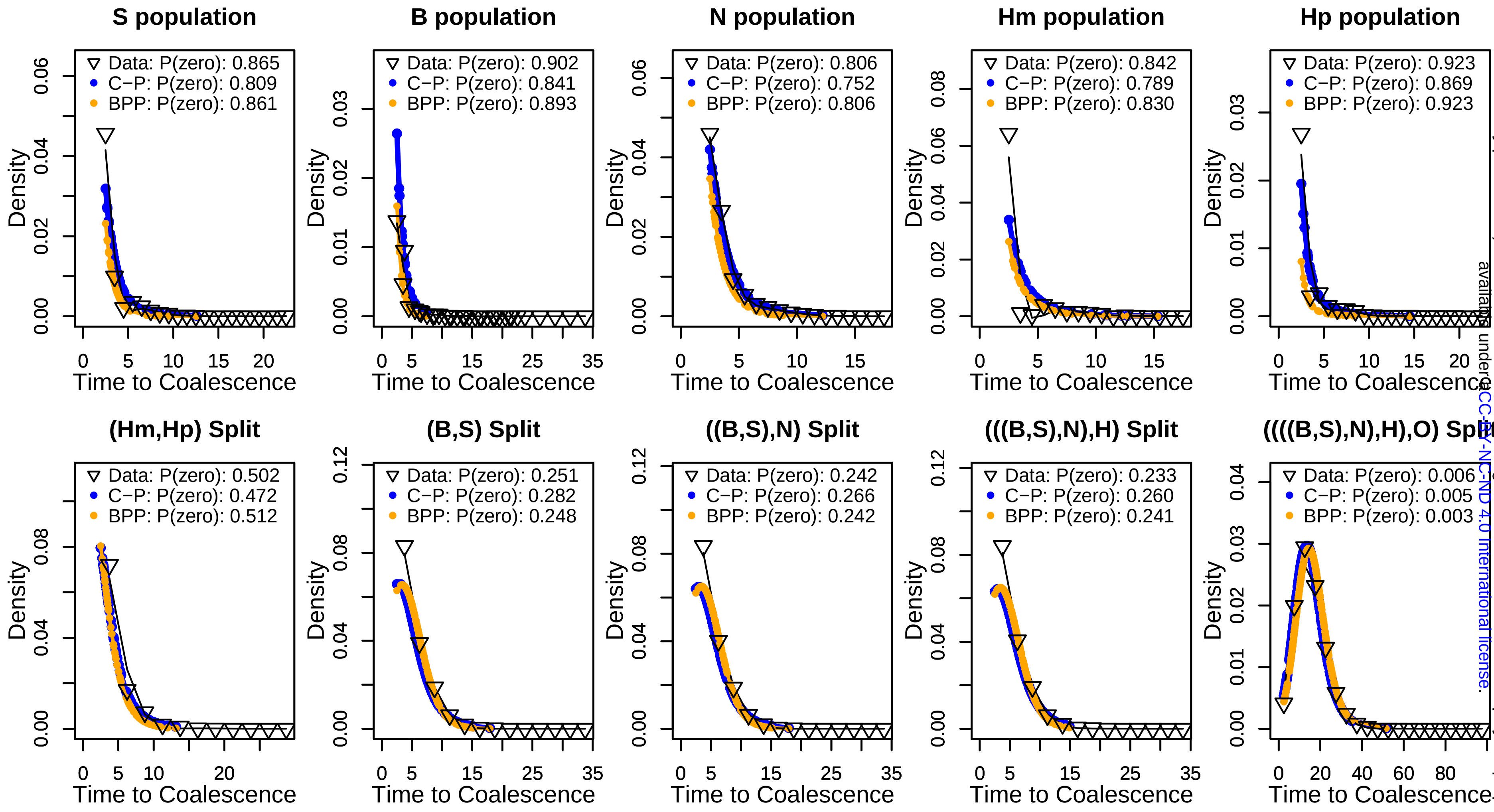

Figure 5: **Distribution of Estimated Coalescent Time: Tree 1 of Coding Gibbon data.** Comparison of zero-inflated MSC Normal distributions of estimated coalescence times using parameters inferred by COAL_PHYRE (Blue) and BPP (Orange) along with the distribution of estimated coalescence times from the data (black triangles) for each proposed split event. The top row: for two individuals sampled from the same population (as indicated by the plot header), the distribution of the estimated coalescence times. The bottom row: for a given split event, e.g. "((B,S),N) Split", the distribution plots the estimated time to coalescence for an individual sampled from the (B,S) subset, and one from the (N) subset. The figure legend also includes the observed (or predicted) fraction of zeros in the data.

7 Discussion

Our simulations show that COAL_PHYRE provides estimates that are comparable to BPP and much more accurate than estimates obtained using ASTRAL-III. We observe a strong effect of mutational variance on estimates obtained using ASTRAL in a low mutation rate setting. An advantage of COAL_PHYRE is that it is straightforward to separate the effects of the two genetic processes that generate the input data by studying the role of both the MSC and the normal distribution.

For the Gibbon data set, we showed that our method can analyze genomic-sized data sets with similar performance to BPP, with an order of magnitude decrease in run time. The composite likelihood approach of only using pairwise coalescence times implemented and presented here seems to sufficiently capture the relevant parts of the data needed to infer the tree parameters. COAL_PHYRE recovered the same most likely topology as presented in [30], for both the coding and non coding datasets. The largest discrepancies between our method and BPP in the analysis of the gibbon data was in fitting parameters to tree 2, which both methods infer to be an incorrect topology. We also see that large deviations in parameter estimates can have negligible effect on the estimated distribution of estimated coalescence times, for example η_{BS} in Table 2, and the resulting effect in Figure 4.

When studying species tree estimation, it is typical to also study topology reconstruction accuracy. We have found in our simulations that ASTRAL is superior in topology reconstruction, and with the speed of ASTRAL compared to COAL_PHYRE, we do not make claims that our method is superior for inferring topologies. The information extracted and used from the data by the two methods is largely orthogonal; ASTRAL uses purely the topological information from each estimated gene tree, and discards all information on coalescence times, whereas COAL_PHYRE only uses marginal coalescence times from each gene, and discards topology information. This lends itself to the idea that the information used in COAL_PHYRE and ASTRAL can be combined or that, at least, be employed in tandem. We also acknowledge work done in [1, 24] which presents a data pre-processing step to counter the effects of mutational variance for programs such as ASTRAL which do not directly model it.

Lastly, none of these methods account for migration/gene flow between species after divergence, something which is common in most real data sets. Failing to account for this potential gene flow can affect topology inference as well as drastically effect divergence time and population size estimation. Accounting for and modeling potential sources of admixture is a next step for these parameter inference methods. It is worth noting that a preprint for an extension of BPP implementing the full MSC with introgression (MSci) has recently been released [7]. Identifying locations of admixture and fitting admixture branches to a species tree are left to future work for COAL_PHYRE.

More studies are needed to understand the robustness of the different methods, for example with regards to substitution models or, and in particular, the effect of recombination within a block. Genomic data is not truly composed of free recombining segments with no internal recombination, which is effectively assumed by all methods analysed in this paper. To address the problem of recombination within blocks, a potential approach is to divide blocks into even smaller units, thereby increasing the amount of mutational variance within each unit, but decreasing the probability of recombination within the unit. As COAL_PHYRE is designed specifically to handle increased variance in estimation, this could be a potential work-around in cases where recombination might be a challenge.

7.1 Software Availability

Along with this manuscript, we provide code (implemented in C++) available for download which implements the likelihood presented here, named COAL-PHYRE. The code is implemented in C++ and freely available at `https://github.com/gaguerra/COAL_PHYRE`.

References

[1] Md Shamsuzzoha Bayzid et al. "Weighted statistical binning: enabling statistically consistent genome-scale phylogenetic analyses". In: *PLoS One* 10.6 (2015), e0129183.

[2] Lucia Carbone et al. "Gibbon genome and the fast karyotype evolution of small apes". In: *Nature* 513.7517 (2014), p. 195.

[3] Xin Chen et al. "Using phylogenomics to understand the link between biogeographic origins and regional diversification in ratsnakes". In: *Molecular phylogenetics and evolution* 111 (2017), pp. 206–218.

[4] Alexei J Drummond and Andrew Rambaut. "BEAST: Bayesian evolutionary analysis by sampling trees". In: *BMC evolutionary biology* 7.1 (2007), p. 214.

[5] Scott V Edwards. "Is a new and general theory of molecular systematics emerging?" In: *Evolution* 63.1 (2009), pp. 1–19.

[6] Thomas Flouri et al. "Species Tree Inference with bpp Using Genomic Sequences and the Multispecies Coalescent". In: *Molecular biology and evolution* (2018).

[7] Thomas Flouris et al. "A Bayesian implementation of the multispecies coalescent model with introgression for comparative genomic analysis". In: *bioRxiv* (2019), p. 766741.

[8] John Gatesy and Mark S Springer. "Phylogenetic analysis at deep timescales: unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum". In: *Molecular phylogenetics and evolution* 80 (2014), pp. 231–266.

[9] Sebastian Höhna et al. "RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language". In: *Systematic Biology* 65.4 (2016), pp. 726–736.

[10] Huateng Huang and L Lacey Knowles. "What is the danger of the anomaly zone for empirical phylogenetics?" In: *Systematic Biology* 58.5 (2009), pp. 527–536.

[11] Huateng Huang et al. “Sources of error inherent in species-tree estimation: impact of mutational and coalescent effects on accuracy and implications for choosing among different methods”. In: *Systematic Biology* 59.5 (2010), pp. 573–583.

[12] Richard R Hudson. “Generating samples under a Wright–Fisher neutral model of genetic variation”. In: *Bioinformatics* 18.2 (2002), pp. 337–338.

[13] Thomas H Jukes, Charles R Cantor, et al. “Evolution of protein molecules”. In: *Mammalian protein metabolism* 3.21 (1969), p. 132.

[14] L Lacey Knowles et al. “Full modeling versus summarizing gene-tree uncertainty: method choice and species-tree accuracy”. In: *Molecular phylogenetics and evolution* 65.2 (2012), pp. 501–509.

[15] Laura S Kubatko, Bryan C Carstens, and L Lacey Knowles. “STEM: species tree estimation using maximum likelihood for gene trees under coalescence”. In: *Bioinformatics* 25.7 (2009), pp. 971–973.

[16] Shea M Lambert, Tod W Reeder, and John J Wiens. “When do species-tree and concatenated estimates disagree? An empirical analysis with higher-level scincid lizard phylogeny”. In: *Molecular phylogenetics and evolution* 82 (2015), pp. 146–155.

[17] Bret R Larget et al. “BUCKy: gene tree/species tree reconciliation with Bayesian concordance analysis”. In: *Bioinformatics* 26.22 (2010), pp. 2910–2911.

[18] Adam D Leaché and Bruce Rannala. “The accuracy of species tree estimation under simulation: a comparison of methods”. In: *Systematic biology* 60.2 (2010), pp. 126–137.

[19] Liang Liu. “BEST: Bayesian estimation of species trees under the coalescent model”. In: *Bioinformatics* 24.21 (2008), pp. 2542–2543.

[20] Liang Liu, Lili Yu, and Dennis K Pearl. “Maximum tree: a consistent estimator of the species tree”. In: *Journal of mathematical biology* 60.1 (2010), pp. 95–106.

[21] Wayne P Maddison. “Gene trees in species trees”. In: *Systematic biology* 46.3 (1997), pp. 523–536.

[22] John E McCormack et al. “A phylogeny of birds based on over 1,500 loci collected by target enrichment and high-throughput sequencing”. In: *PLoS One* 8.1 (2013), e54848.

[23] Siavash Mirarab and Tandy Warnow. “ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes”. In: *Bioinformatics* 31.12 (2015), pp. i44–i52.

[24] Siavash Mirarab et al. “Statistical binning enables an accurate coalescent-based estimation of the avian tree”. In: *Science* 346.6215 (2014), p. 1250463.

[25] Ignacio Quintero and John J Wiens. “Rates of projected climate change dramatically exceed past rates of climatic niche evolution among vertebrate species”. In: *Ecology letters* 16.8 (2013), pp. 1095–1103.

[26] Andrew Rambaut and Nicholas C Grass. “Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees”. In: *Bioinformatics* 13.3 (1997), pp. 235–238.

[27] Bruce Rannala and Ziheng Yang. "Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci". In: *Genetics* 164.4 (2003), pp. 1645–1656.

[28] Robert E Ricklefs. "Estimating diversification rates from phylogenetic information". In: *Trends in Ecology & Evolution* 22.11 (2007), pp. 601–610.

[29] Sebastien Roch and Mike Steel. "Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent". In: *Theoretical population biology* 100 (2015), pp. 56–62.

[30] Cheng-Min Shi and Ziheng Yang. "Coalescent-based analyses of genomic sequence data provide a robust resolution of phylogenetic relationships among major groups of gibbons". In: *Molecular biology and evolution* 35.1 (2017), pp. 159–179.

[31] Brian Tilston Smith et al. "Target capture and massively parallel sequencing of ultraconserved elements for comparative studies at shallow evolutionary time scales". In: *Systematic biology* 63.1 (2013), pp. 83–95.

[32] João Tonini et al. "Concatenation and species tree methods exhibit statistically indistinguishable accuracy under a range of simulated conditions". In: *PLoS currents* 7 (2015).

[33] Krishna R Veeramah et al. "Examining phylogenetic relationships among gibbon genera using whole genome sequence data using an approximate Bayesian computation approach". In: *Genetics* 200.1 (2015), pp. 295–308.

[34] Ziheng Yang. "The BPP program for species tree estimation and species delimitation". In: *Current Zoology* 61.5 (2015), pp. 854–865.

[35] Chao Zhang et al. "ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees". In: *BMC bioinformatics* 19.6 (2018), p. 153.

Appendices

A Notation Reference

- N: Number of species considered.
- $\mathcal{S} = (S, \tau, \eta)$: A species tree parameterized by topology S, split times τ, and population sizes η.
- K: Number of independent loci/genes.
- M: Number of sampled individuals ($M \geq N$).
- $Q = \binom{M}{2}$: Number of pairs of individuals.

- $\vec{h_j}$: Set of M haplotypes at locus j, $j \in \{1, 2, ..., K\}$.
- $\bar{g}_j$: Estimated gene tree at locus j.
- $\mathcal{G}_j$: True gene tree at locus j.
- a, b: Individuals sampled from populations A, B, respectively.
- $\tau_{A,B}$: The split time of species A and B according to $\mathcal{S}$.
- Each pair of individuals are indexed by an integer i,in $(1, ..., Q)$.
- $\bar{g}_j(i)$: The estimated coalescence time of pair i at locus j.
- $\mathcal{G}_j(i)$: The true time to coalescence of pair i at locus j.
- μ: The per generation per base pair mutation rate.
- $\mathcal{L}$: The number of base pairs of gene.
- θ: The population scaled mutation rate, $\theta = 2\mu\eta_0$, for the reference population size, η_0.
- $\omega = \frac{1}{\theta \times \mathcal{L}}$
- $\omega\mathcal{G}_j(i)$: Mutational estimation variance of the true coalescence time.

B Fit of the MSC-Normal distribution

In this paper we have discussed that estimated coalescence times can be modelled with two sources of variance, one from the coalescent process, and the other from mutational process. In figure 6 we visualize an example of the approximation to the distribution of estimated coalescence times obtained using the MSC-Normal distribution. As well, we plot the true coalescence times when times are known without error. From the figure we notice that in the presence of estimation error, estimated coalescence times can be more recent than the species divergence time. The MSC-Normal distribution acts as an approximation to the distribution of estimated coalescence times, and captures this tail of recent times, as well as the overall distribution.

C Further Simulation Details

C.1 5-species simulation details

In this simulation study we analyzed a species tree of 5 species (labeled 1...5) with 10 individuals (labeled 1...10) where 2 individuals are from each species (i.e individuals 1 and 2 are from species 1). We simulate the rooted species topology (5,(4,(1,(2,3)))).

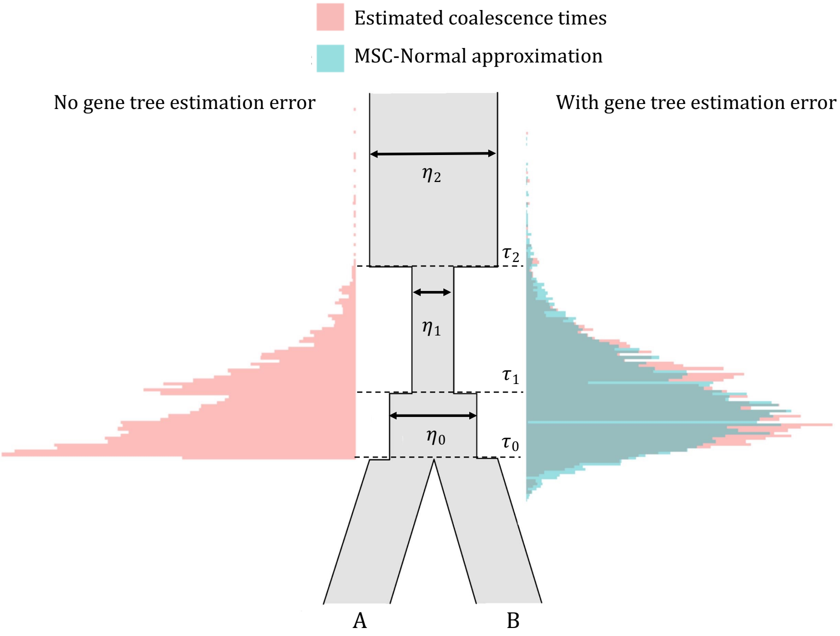

Figure 6: **MSC-Normal approximation of coalescence times with coalescence estimation error.** Center: Species tree of two individuals with a piece-wise constant population size history. On the left: Plotted in red is the distribution of coalescence times of these two individuals under simulation when the true gene trees are known, so no estimation variance is present. On the right: Plotted in red is the distribution of estimated coalescence times of the same two individuals from simulated sequence data, where variance is introduced due to the mutational process. In green is the MSC-normal approximation to the distribution of estimated coalescence times using the known tree parameters (τ's and η's) and mutation rate θ.

For a single replicate, we use ms to generate K independent gene trees of 10 individuals, 2 from each species, and Seq-Gen [26] to generate sequence data from the gene trees. To generate $K = 100$ gene trees of 10 individuals with species labeled as integers 1 through 5:

```
./ms 10 100 -T -I 5 2 2 2 2 2 -n 1 1.8 -n 2 2.4 -n 3 1.0 -n 4 2.0 -n 5 3.0 -ej 1.0 2 3
    -en 1.0 3 2.4 -ej 1.5 1 3 -en 1.5 3 3.0 -ej 2.2 3 4 -en 2.2 4 4.0 -ej 4.0 4 5
    -en 4.0 5 5.0 | tail +4 |grep -v // >gene.trees
```

In ms [12], time is measured in units of $4\eta_0$ generations, whereas COAL_PHYRE measures time in $2\eta_0$ generations, so that times from COAL_PHYRE must be halved to compare to the units of ms. As well, population sizes in ms are diploid, whereas in COAL_PHYRE we measure population sizes as haploid. To compare with ms, population sizes from COAL_PHYRE need to be doubled.

From the **gene.trees** file, and for a given mutation parameter θ (which we used either 0.01 or 0.001 in our simulation), and sequence length $\mathcal{L}$, we use Seq-Gen. For example, for $\theta = 0.001$ and $\mathcal{L} = 1000$:

```
./Seq-Gen -mHKY -l 1000 -s 0.001 <gene.trees >seqfile
```

We use this **seqfile** file as input to COAL_PHYRE.

C.2 8-species simulation details

For 8 species, 2 individuals sampled per species, we generated a single replicate of $K = 100$ independent gene trees using:

```
./ms 16 100 -T 8 2 2 2 2 2 2 2 2 -n 1 1.5 -n 2 2.5 -n 3 2.0 -n 4 6.0 -n 5 0.5 -n 6 1.0
-n 7 3.0 -n 8 4.0 -ej 0.5 2 1 -en 0.5 1 6.0 -ej 0.75 4 3 -en 0.75 3 1.0 -ej 0.8 8 7
-en 0.8 7 2.0 -ej 1.3 6 5 -en 1.3 5 4.0 -ej 1.5 3 1 -en 1.5 1 5.0 -ej 1.8 7 5 -en 1.8 1.5
-ej 2.0 5 1 -en 2.0 1 6.0 | tail +4 | grep -v // >gene.trees
```

For $\theta = 0.01$ and gene length $\mathcal{L} = 1000$, we generate sequence data with Seq-Gen:

```
./Seq-Gen -mHKY -l 1000 -s 0.01 <gene.trees >seqfile
```

We use this **seqfile** file as the input to COAL_PHYRE.

D Normal Approximation To Poisson, A Simulation

Throughout we discuss the distribution of estimated coalescence times. The estimation error from the mutation process, conditional on a branch length, follows a Poisson distribution. As our estimated coalescence times are not discrete, we use the Normal approximation to the Poisson. In this section we demonstrate in a simple simulation scenario, that this approximation is well fit to model the estimation error. For a given coalescence time (fixed here to be 5 in units of $2\eta_0$ generations), we simulate 1000 pairs of sequences, of length $\mathcal{L} = 1000$ base pairs, under varying scaled mutation rates, θ (indicated in figure 7 legend), to generate an empirical distribution of estimated coalescence times from the number of pairwise differences. Figure 7 shows these distributions versus the normal approximation presented earlier. This demonstrates the accuracy and suitability of the Normal approximation to mutational variance in time estimation.

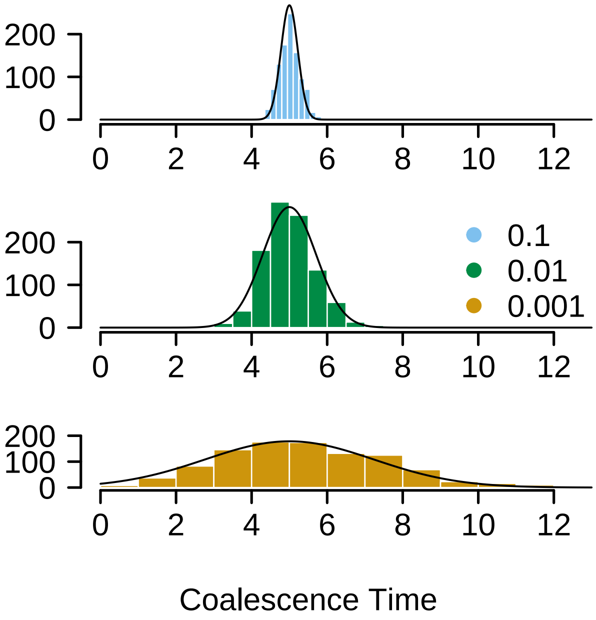

Figure 7: **Modeling Mutational Variance:** The Normal approximation to Possion variance in coalescent time estimation error due to the mutational process.