Title
Statistical Inference under the Multispecies Coalescent: Methods and Theory

Author
Guerra, Geno A

Publication Date
2019

Peer reviewed|Thesis/dissertation

Statistical Inference under the Multispecies Coalescent: Methods and Theory

by

Geno A. Guerra

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Rasmus Nielsen, Chair
Professor Yun S. Song
Professor John P. Huelsenbeck

Fall 2019

Statistical Inference under the Multispecies Coalescent: Methods and Theory

Abstract

Statistical Inference under the Multispecies Coalescent: Methods and Theory

by

Geno A. Guerra

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Rasmus Nielsen, Chair

The rising availability of genome-scale data for a large number of species has allowed for more in-depth studies of the genetics between species using increasingly sophisticated methods. The accumulation of pairwise differences between individuals are indicative of how diverged they are in time. The multi-species coalescent (MSC) has been the most popular framework with which to model the dynamics of the coalescent process in the presence of species barriers, such as a tree structure. Modelling using the MSC in the presence of increasing amounts of data (loci and species) while maintaining feasible computational times is the main focus of many emerging methods.

In this dissertation, I explore the use of the MSC in 3 different ways, using classical and novel statistical analysis to provide insight into species divergence parameters. I begin by constructing a novel statistical method for inferring species tree divergence times and population size parameters for any given tree topology from sequence data. The program COAL-PHYRE, presented here, makes use of the MSC marginally between individuals, as I demonstrate that pairwise information within the MSC is sufficient to learn times and population sizes on a tree. My focus then shifts to the derivation of the covariance between pairs of coalescence times and its application to studying average pairwise differences and the commonly used statistic, F_{ST}. I confirm that estimates of F_{ST} are biased, and quantify the effect of not accounting for this bias in different applications. I conclude by continuing to study the covariance between coalescence times and its use in inferring species tree topologies. I define a metric based on these statistics which, when paired with the minimum spanning tree algorithm, provides estimates of species tree topologies. I provide partial proofs of statistical consistency of the approach.

To mom and dad.

Contents

List of Figures

List of Tables

Acknowledgments

I am delighted to acknowledge the many friends, loved ones, and mentors who have made my time at Berkeley a happy and rewarding one. Firstly, I must thank my advisor Rasmus Nielsen. I didn't join his lab until my fifth year at Berkeley, but I am fortunate to get to work under his guidance for as many years as I did. Rasmus is the definition of a great advisor. His expertise in statistics and biology never failed to impress me, and his personability and patience made him a wonderful mentor.

As I didn't join Rasmus until later on in my graduate career, I also want to acknowledge and thank Yun Song, my advisor for my first years at Berkeley. He welcomed me into his group with open arms and started me down my path into computational biology. I was able to learn so much from him. His intellectual curiosity with the work ethic to match makes for a truly remarkable researcher and advisor.

While I was a part of the labs of Rasmus and Yun, my fellow lab members helped to teach me, support me, and even humor me. A few I would especially like to acknowledge are Jonathan Fischer, Matthias Steinruecken, Jack Kamm, Sara Mathieson, Jonathan Terhorst, Jeffrey Spence, Jeffrey Chan, April Wei, Jasmine Nirody, Shishi Luo, Neil Thomas, Ethan Jewett, Aaron Stern, Jane Yu, and the rest of the Song and Nielsen labs.

The best part of graduate school has been the many brilliant friends I have made over the years. Jonathan Fischer and Andre Waschka, thank you guys for being my closest friends, and the groomsmen in my wedding. Partow Imani, Matthew Harrison-Trainor, Joe Borja, Kelly Street, Gwen Tindula, Suzanne Dufault, and Yannik Pitcan, thank you for the wonderful memories in our time at Berkeley. Go Outliers!

The highest acknowledgements must got to my mom and to my dad. From such a young age, the amount of love, support and absolute prioritization of my (and my brother's) education and well-being has been unfathomable. I wouldn't be where I am today without the confidence my parents gave me throughout my life. I owe them everything. Thank you to my brother, Andre. I am lucky to have a lifelong friend like you in my life. To my grandpa Monte, the world's biggest Geno fan. Thank you so much for your infectious positivity, your Doctor buddy is finally legit.

Lastly, none of this would have been possible without the love and support of my dear wife and love of my life, Megan. Thank you for keeping me sane during the long hours/days/weeks/months/years it took to get here. Looking forward to what comes next for us!

Chapter 1

Introduction

The diversity of all walks of life we see today on earth is due to the process of evolution over time [5]. This was likely begun with a single common ancestor where new species were generated through a branching process across millions and billions of years. The genetic relationship between species is therefore often modelled using phylogenetic trees. A phylogenetic tree (or phylogeny) traces the evolutionary history of species through time, with each node on the tree representing a distinct species. Internal nodes are viewed as species which are the ancestors of all nodes subtending it. While there may be some semantic difference, we use the terms phylogeny and species tree interchangeably throughout. Species evolve and split into new species for many reasons, including adaptation to new environments, and geographic isolation. At a genetic level, species diverge by evolution through the continual acquisition of new mutations across the genome.

1.1 Background

Due to the random accumulation of new mutations, comparing the differences in the DNA sequence data of different species can be used to make estimates of the divergence patterns in the past. Most notably during meiosis, the process of recombination, where aligned homologous chromosomes exchange tracts of their DNA, allows a single DNA sequence to be a mosaic of genetic histories of one's ancestors. Studying different parts of the genome can result in differing estimates of the genetic history of a set of individuals/species. This evolutionary history of a set of individuals at a particular segment of the genome is known as a *gene tree*. Gene trees can differ in topology (ordering of events) from the overall species tree for many reasons, such as migration between species, gene duplication/loss, and most commonly

through the random population genetic processes.

This genetic process is commonly modelled by the coalescent [20], which models the likelihood of genealogies within a population. The process models the backwards in time probability of finding time to a common ancestor when the parent of each individual is chosen at random from the previous generation. A multi-species extension, known as the multispecies coalescent (MSC) allows for a local genealogy of individuals from separate populations to be jointly modelled assuming a species tree structure. Barring the presence of migration, individuals find a common ancestor with an individual from another species at some time more ancient than the species time of divergence.

The random process of finding a common ancestor, or *coalescing*, can potentially result in a very recent, or very ancient time to a common ancestor (TMRCA). Under the MSC, lineages in a single population can fail to coalesce, an event known as incomplete lineage sorting (ILS). These failures to coalesce in the given time span can result in gene trees which are incongruent with the species tree topology. Gene tree discordance due to ILS is a common problem in phylogenetics, and is the main focus of many researchers (see chapter 2 for references).

Another complicating factor in the estimation of local gene trees is the random process of mutation. As sequence data is finite, and thus the amount of differing mutations is also finite, estimating a time to coalescence between lineages comes with some error. Disentangling gene tree discordance due to mutation from that due to the MSC is a complicated problem. In chapters 2 and 3, we explore the effects of estimation error from mutation.

1.2 Outline

As DNA sequencing costs have dropped dramatically in recent years, the availability of large multi-locus sequences across many species is becoming increasingly more common. The ability to model these large data sets in efficient ways to estimate aspects of the unknown evolutionary history of the set of species. In this thesis, I develop new statistical methods to make estimates of the evolutionary history of species, and study the effects of estimation error on a set of existing theory, as well as provide new theory to help quantify the error. While each chapter focuses on a different problem in phylogenetics, they have the common theme of utilizing the multispecies coalescent to model gene tree discordance due to ILS. The following chapters can be succinctly summarized as follows:

- Chapter 2: COAL-PHYRE: A novel, scalable, method which models both ILS

and mutational variance in sequence data to estimate species divergence times and population sizes.

- Chapter 3: An exploration of the distribution of pairwise differences between individuals using the MSC, and its applications to the error in estimates of F_{ST}. This includes new theory on the covariance in coalescence times and in the covariance of pairwise differences.

- Chapter 4: A fast summary method to estimate species tree topologies using the covariance between pairwise coalescence times and the minimum spanning tree algorithm.

Any software developed to accompany each chapter can be found on github, with links in the respective chapters.

Chapter 2

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

This is joint work with Rasmus Nielsen.

Genome-scale data are increasingly being used for inferences of phylogenetic trees. When using genomic data from multiple species it is common that different regions of the genome have local topologies that differ from the species tree. One major source of this discrepancy is incomplete lineage sorting (ILS) which is well-modeled using the multi-species coalescent (MSC). 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 within a locus using a normal approximation. To evaluate the accuracy and speed of the method, we compare the method against BPP, a powerful MCMC full-likelihood method, as well as ASTRAL, a fast approximate method. We

show that COAL_PHYRE's divergence time and population size estimates are much 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 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.

2.1 Introduction

With the continued rise of modern day sequencing technology, inferences 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 in the past. 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) [30]. The most ubiquitous of these, ILS, is of particular focus in the field [8], and can be well-modeled using the multi-species coalescent (MSC) (see e.g., [40]). 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 [23, 32, 29]. 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 [14, 15, 21]. A class of Bayesian hierarchical methods exist, which jointly model gene and species tree topologies in a full likelihood framework (e.g. [26, 7, 28, 10, 13, 57]), and account for both coalescent and mutational variance, but these approaches have been shown to be computationally intensive and not able to scale to large amounts of genes or species [27, 31, 47].

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. [50, 25, 12] discuss results showing that concatenation-based approaches are not always outperformed by more ILS-sensitive methods. In short, concatenation methods seem to be predictably worse 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 [43]. 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 [4, 41, 38]. 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 [57] provides highly accurate results under the assumption of a molecular clock and the Jukes and Cantor model of sequence evolution [18]. 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) blocks of loci 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 [57], with at least an order of magnitude decrease in run time. We also compare to the popular approximate likelihood method ASTRAL-III [58], 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 [44], and find highly similar estimated parameters.

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 2.1(a) for an illustration. Given a recombination-free region of the genome, l, 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

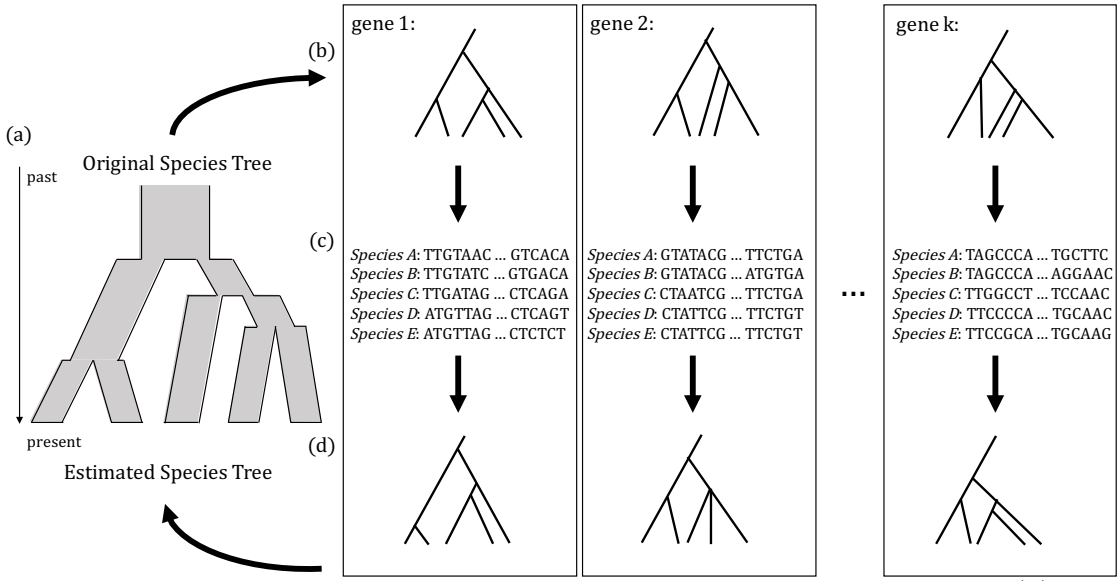


Figure 2.1: **Contribution of coalescent and mutational variance.** (a) Original bifurcating species tree. (b) K gene trees, each a different realization of a stochastic lineage sorting process on the original species tree. (c) Sequences created from the mutational process on each gene tree. (d) Gene trees estimated from the sequence data, which can differ in topology and branch length from the true gene trees due to mutational variance.

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 [23]. 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 $\mathcal{S}$.

2.2 Methods

Jointly modeling mutation and coalescent variation in a full likelihood framework has been studied, but is a challenging problem. Unlike the coalescent process, incorporating a full model of the mutational process requires simulation that is computationally intensive [14]. Existing approximation methods are unable to separate the effects of the two sources of variance in their inference procedure, and simulation studies have been required to test the methods' accuracy under varying levels of mutational variance [15]. The authors of ASTRAL [32] studied the effect of mutational variance, and proposed a data pre-processing weighted statistical binning approach [33, 1] where

loci with a high "combinability" are used to estimate a single gene tree, and each gene tree is weighted by the corresponding bin-size used.

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

An advantage of this model 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. A more detailed introduction to the model is left for later.

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.

Mutational variance

As is common in most species tree inference methods ([57, 32, 23] for example), we assume that genomic data can be divided recombination-free regions, with free

recombination between them. 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 2.1). 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. 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, in any case, evaluate all estimators presented in this paper using simulations under finite sites models. The component of the variance in the estimate of the coalescent time contributed by mutational noise is then

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

where $\omega = \frac{1}{\theta\mathcal{L}}$.

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 A.1 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 is given below.

The composite likelihood

The input for the algorithm is K sets of aligned sequences $(\vec{h_1}, ..., \vec{h_K})$, 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 is estimated from haplotypes $\vec{h_j}$ at locus j from the pairwise number of differences between the sequences.

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 indiviudla likelihood functions. For M individuals in the tree ($M \geq N$), we decompose the likelihood into Q univariate quantities:

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

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 the pair indexed by 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 zero coalescence times needs to be explicitly modeled, as the MSC-normal distribution does not adequately account for the point mass of zeros.

MSC-Normal distribution

For two individuals, a, b (indexed by i), the divergence time for the species A, B ($a \in A$, $b \in B$) is denoted by τ_{AB}. For a given locus, we observe some estimated coalescence time $\bar{g}_j(i)$ between the pair, based on the reconstructed gene tree at the

locus. 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 $\{(\eta_0,\tau_0),\ldots,(\eta_{R-1},\tau_{R-1})\}$, where $\eta_0=\eta_{AB}$ and $\tau_0=\tau_{AB}$. At each branch on the tree, we can calculate the likelihood of $\mathcal{G}_j(i)$ given the coalescence event occurs within the branch $\Big(\mathcal{G}_j(i)\in(\tau_r,\tau_{r+1})\Big)$. 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

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

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},∞)

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

Where

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

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

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

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

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.

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

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

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

Likelihood weighting

In the calculation of the composite likelihood, the same information is 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. 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.

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 [16] 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 [18] through Seq-Gen [39] to produce haplotypes under various mutation rates to introduce varying levels of mutational variance. See Appendix A.2 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.

2.3 Simulation Results

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 A.2 for exact simulation details. We compare our method, COAL_PHYRE, to BPP [57] and ASTRAL-III [58]. COAL_PHYRE and BPP provides seperate estimates of colaescence times and population sizes, while ASTRAL-III provides estimates of the coalescence rate of each branch (recall coalescence rate = branch length/ population size), but is unable to separate the two parameters. 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, 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 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 $\theta = 0.01$ because of 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 2.2.

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 as to the true 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 ($\eta_1 \dots \eta_5$) and COAL_PHYRE has more uncertainty than BPP in estimation of internal population sizes, reflecting the well-known challenge of disentangling internal branch length from population sizes.

When comparing run times, ASTRAL completed on average in about 1 second per

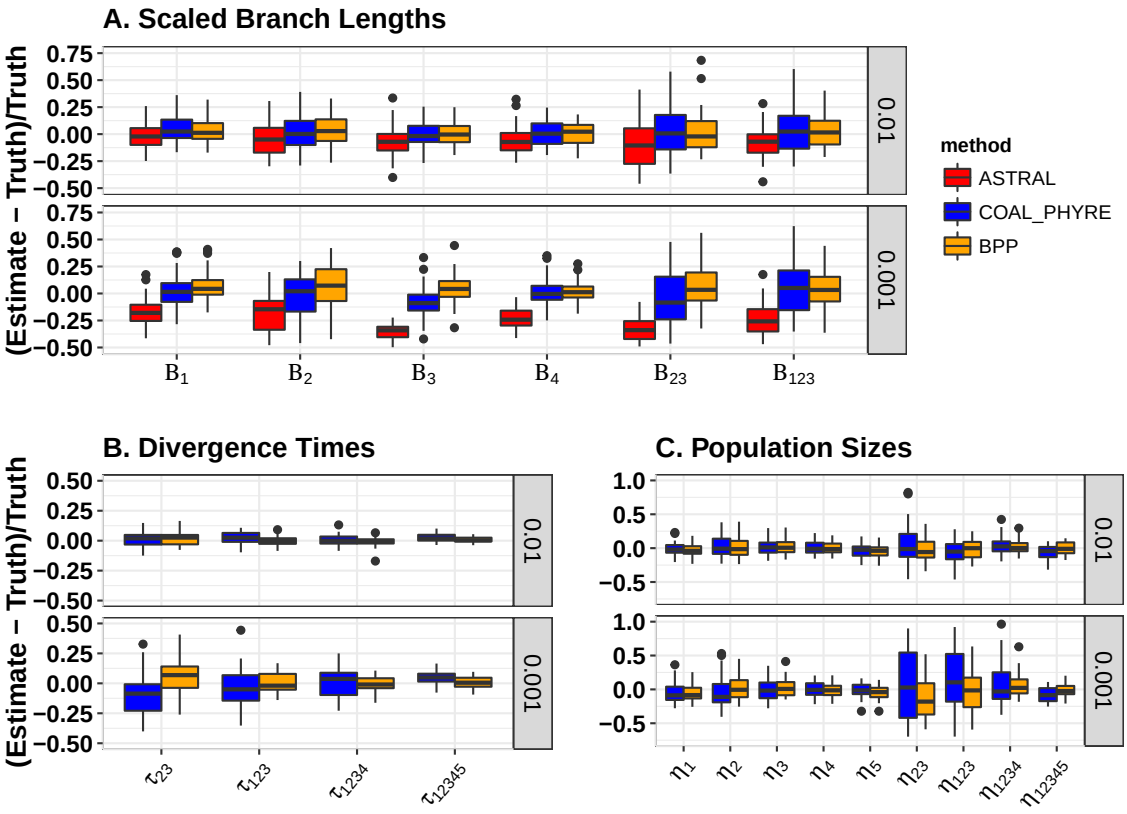


Figure 2.2: **Full parameter estimation for the fixed species tree topology (5,(4,(1,(2,3)))).** Comparison of parameter estimates between COAL_PHYRE, ASTRAL and BPP by branch over 40 iterations, using 100 sampled loci each iteration. The y-axis gives the standardized deviation from the true parameter value. In each panel, the top plot represents a high mutation rate setting, where mutational variance is low, and the bottom represents a $\times 10$ 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.

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.

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 among branches [RN: insert reference to where full details can be found]. Again, we compare COAL_PHYRE to BPP [57], and ASTRAL-III [58]. 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 2.3). Similarly to the 5 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.

2.4 Analysis of Gibbon Data

Here we analyze two full-genome data sets from [2] and [51] of four gibbon species: (*Hylobates moloch* (Hm), *Hylobates pileatus* (Hp)), *Nomascus leucogenys* (N), *Symphalangus syndactylus*(S), and *Hoolock leuconedys* (B). Gibbons (Hylobatidae),

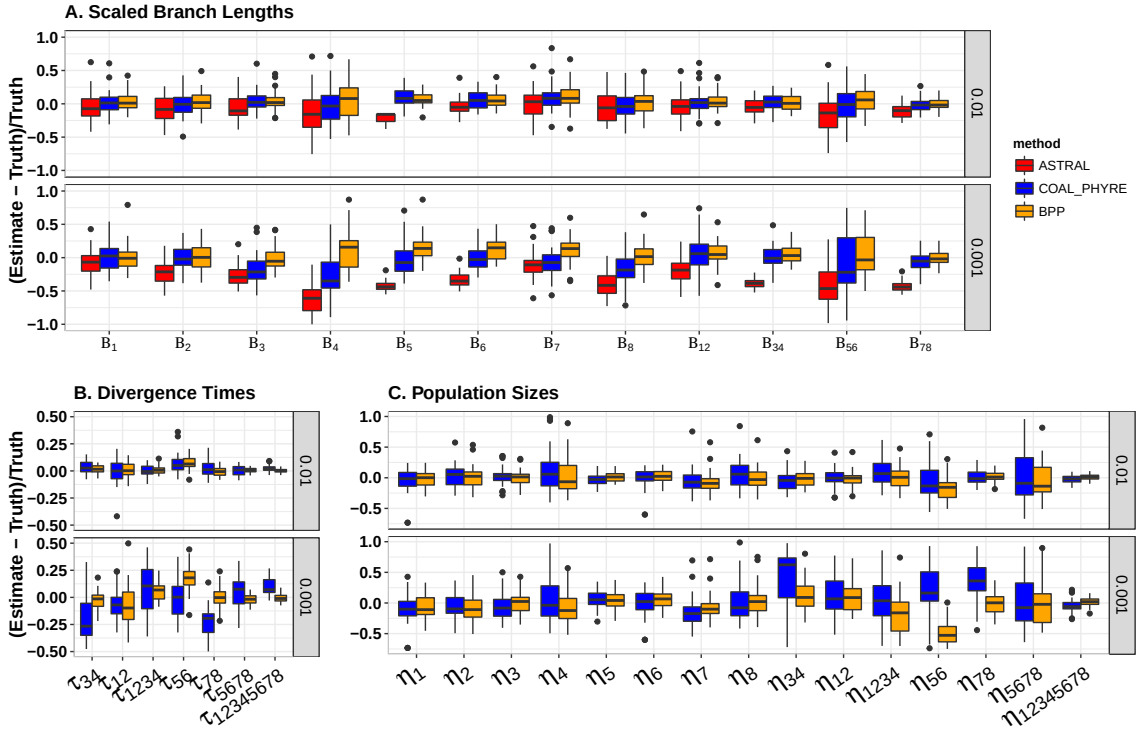


Figure 2.3: **Full parameter estimation for the fixed species tree topology (((1,2),(3,4)),((5,6),(7,8))).** Comparison of the parameter estimation accuracy between COAL_PHYRE (blue) and BPP (orange), and ASTRAL-III (red) using 100 independent genes, across 40 independent replicates. **A)** A comparison 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.

close relatives to humans and great apes, are found throughout Southeast Asia's tropical forests. A recent study, Shi and Yang, 2017[44] (hereby referred to as SY17) used the MCMC program, BPP ([57]), 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).

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.

Results

We use COAL_PHYRE to analyse 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 [18] model of mutation. As well, we use mutation rate parameters consistent with the means of the Gamma priors used in SY17.

Divergence time and population sizes estimates

The parameter values estimated using COAL_PHYRE, along with those estimated using BPP in are presented in Tables 2.1, 2.2, 2.3, 2.4. In each scenario, we found that COAL_PHYRE assigned the highest likelihood (between Tree 1 and Tree 2) to 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 there is no information 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 estimates $\tau_{\text{COAL\_PHYRE}} = 0.81\tau_{\text{BPP}}$. 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 branch.

Each of the four tables demonstrates one run of COAL_PHYRE, which on a single core is able to run on average in 10($\pm$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.

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 2.1 and 2.2 (see Figures 2.5 and 2.4, 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 2.4 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 2.5 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.

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($\pm$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 2.1: Table of Gibbon results: (H, (N, (B, S))) coding

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

Table 2.2: Table of Gibbon results: (H, (N, (B, S))) noncoding

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

Table 2.3: Table of Gibbon results: (N, (H, (B, S))) coding

Method	η_B	η_S	η_{Hm}	η_{Hp}	η_N
COAL_PHYRE	0.91	1.13	1.27	0.73	1.61
BPP	0.6	0.8	0.8	0.4	1.2
	η_{BS}	η_H	η_{HBS}	η_{HNBS}	η_{OHNBS}
COAL_PHYRE	3.87	1.43	24.87	3.22	8.43
BPP	22.3	2.0	2.6	1.9	7.8
	τ_{BS}	τ_H	τ_{HBS}	τ_{HNBS}	τ_{OHNBS}
COAL_PHYRE	1.66	1.04	1.82	2.14	10.85
BPP	1.9	1.0	3.0	3.05	11.5

Table 2.4: Table of Gibbon results: (N, (H, (B, S))) noncoding

Method	η_B	η_S	η_{Hm}	η_{Hp}	η_N
COAL_PHYRE	0.9	1.18	1.22	0.62	1.82
BPP	0.9	1.3	1.3	0.6	2.0
	η_{BS}	η_H	η_{HBS}	η_{HNBS}	η_{OHNBS}
COAL_PHYRE	3.73	112.32	inf	2.00	5.10
BPP	12.5	2.3	14.4	2.5	5.5
	τ_{BS}	τ_H	τ_{HBS}	τ_{HNBS}	τ_{OHNBS}
COAL_PHYRE	2.29	1.12	4.16	4.45	15.17
BPP	3.75	1.6	4.3	4.8	15.25

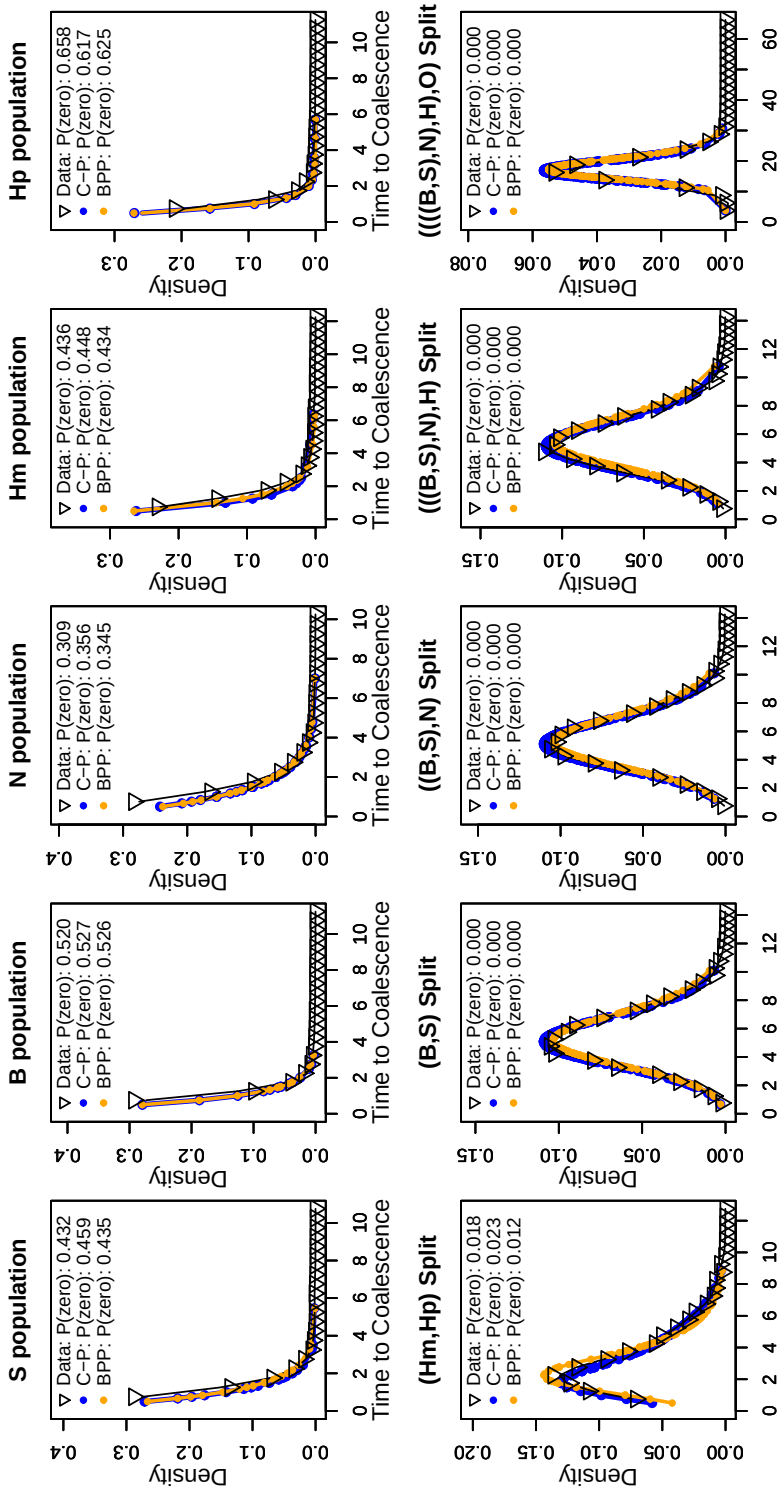


Figure 2.4: **Distribution of Estimated Coalescent Time: Tree 1 of Noncoding Gibbon data.** Comparison of zero-inflated MSC Normal distributions of estimated coalescence times using parameters inferred by COAL_PHYRE (C-P) (Blue) and BPP (Orange) along with the distribution of estimated coalescence times 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.

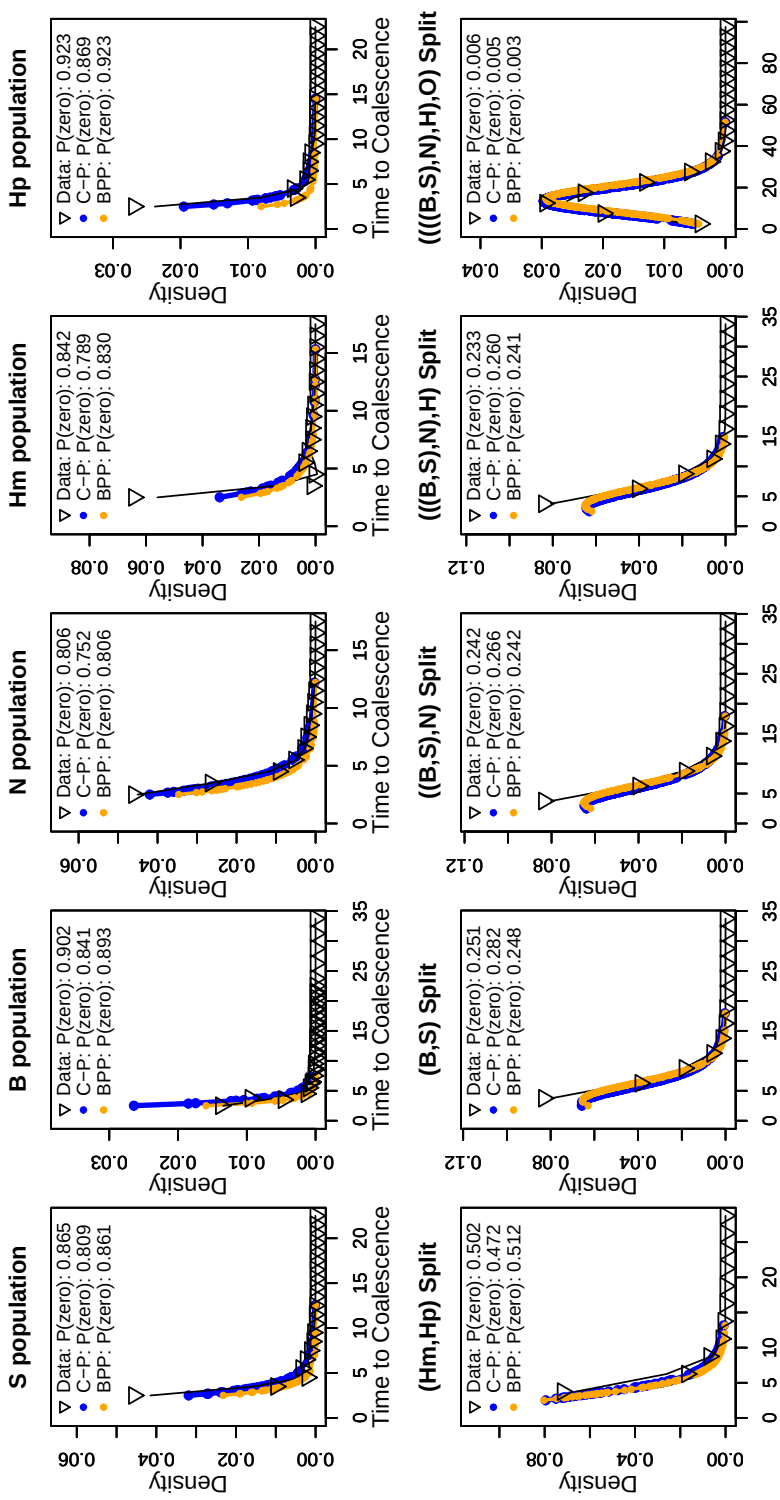


Figure 2.5: **Distribution of Estimated Coalescent Time: Tree 1 of Coding Gibbon data.** Comparison of zero-inflated MSC Normal distributions of estimated coalescence times using parameters inferred by COAL_PHYRE (Blue) and BPP (Orange) along with the distribution of estimated coalescence times from 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.

2.5 Discussion

Our simulations suggest 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. We acknowledge work done in [1, 33] 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.

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 times. 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 [44], 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 2.2, and the resulting effect in Figure 2.4.

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 the better method 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. For example, it might be possible to use ASTRAL to estimate the most likely topology (or set of topologies), and then using our method to estimate parameters of the topologies of interest.

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 [11]. 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.

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.

Chapter 3

Covariances of Pairwise Differences on a Multi-Species Coalescent Tree with Implications for the Statistical Properties of Sequenced-based F_{ST} Estimators

This is joint work with Rasmus Nielsen.

We here derive the variances and covariances of pairwise coalescence times in a general phylogenetic model with piecewise constant changes in population size. We use these expressions to derive the variance in average pairwise differences within and between groups and to derive approximate expressions for the expectation and bias of a sequence-based estimator of F_{ST}. We show that the commonly used estimator of F_{ST} is generally biased and will consequently lead to biases in standard applications such as the estimation of effective rates of migration. We also explore the accuracy of the common log transformation and ratio transformation for linearizing F_{ST} and show that the latter performs better. A freely available software package is provided, STCov, to calculate mean, variances, and covariances in coalescence times and pairwise differences, under arbitrary piecewise constant species phylogenies.

3.1 Introduction

Takahata and Nei [49] derived expressions for the variance in average pairwise nucleotide differences and Nei and Li's 'net number of differences'[35] (d). They assumed a Kingman coalescent model [20] of two diverging populations, and an infinite sites model of mutation [19, 55]. These classical results provided insights into when the net number of differences can be used as a reliable estimator for species divergence, and the appropriate sampling schemes to combat increased variance. However, the results relied on the assumption of constant and equal population sizes among populations and through time. Using the multispecies coalescent (MSC) we extend these results to arbitrary piecewise constant population size histories along a phylogeny. To do so, we present general equations for calculating the covariance of pairwise coalescence times, for any 2,3 or 4 individuals, arbitrarily chosen within the phylogeny. We also derive expressions for the expected shared branch length between sets of lineages. We provide a software package, STCov, for calculating these quantities. We also use the results to study the sampling distribution of a statistic measuring F_{ST}, [45], and the effects of sampling variance and demographic changes on various F_{ST}-based measurements, and demonstrate potential large bias when using F_{ST} estimated from a small number of segregating sites.

3.2 Average Pairwise Differences

Borrowing notation from Takahata and Nei [49], let d_X and d_Y be the mean number of nucleotide differences between two (haploid) individuals sampled from within population X or Y, respectively. Similarly, let d_{XY} be the average number of nucleotide differences between two individuals randomly sampled from populations X and Y. We can calculate d_X, d_Y, and d_{XY} based on sample sizes of n_X and n_Y from populations X and Y, respectively, as follows:

$$d_X = \frac{2}{n_X(n_X-1)} \sum_{i=1}^{n_X-1} \sum_{i'=i+1}^{n_X} k_{i,i'}$$

$$d_Y = \frac{2}{n_Y(n_Y-1)} \sum_{i=1}^{n_Y-1} \sum_{i'=i+1}^{n_Y} k_{j,j'}$$

$$d_{XY} = \frac{1}{n_X n_Y} \sum_{i=1}^{n_X} \sum_{j=1}^{n_Y} k_{i,j}$$

Where $k_{i,i'}$ is the number of pairwise nucleotide differences between individuals i and i'.

To measure the net number of nucleotide differences between two populations, Nei and Li's [35] d is defined as

$$d = d_{XY} - \frac{1}{2}(d_X + d_Y).$$

The relationship between differences within and between populations gives an indication of the degree of population subdivision. d specifically measures the excess number of substitutions between populations, which quantifies the extent of divergence between populations. These measures of species divergence form the basis for many evolutionary analyses and are among the most basic and commonly used inferential tools in modern population genetics.

Understanding the mean, variance, and covariance of these statistics (d_X, d_Y, d_{XY}, d) under arbitrary genetic and species tree models is essential for their biological interpretability, and considerable previous work has been devoted to understanding their properties. Tajima [48], and Takahata and Nei [49] studied the variance of average pairwise differences in a panmictic population, and in a split model with constant population size. In a series of papers, Wakeley studied the variance in pairwise differences in a general model of population sub-division [52], the average pairwise differences in a model with migration [53]. He demonstrated the usefulness of the variance as an estimator of recombination rates [54].

Here we extend this work to the general case of a multispecies split model with no migration, but arbitrary piecewise constant population size along the phylogeny. We derive exact expressions for the means, variances and, in particular, covariances of coalescence times and of average pairwise differences as functions of the mutation rate, sample size, divergence times, and effective population sizes. In order to do so, we first derive the covariance of pairwise coalescence times and expected shared branch length between pairs of lineages, under arbitrary piecewise-constant species tree demographic models. We then use these results to demonstrate the effects of various demographic, mutational, and sampling size changes on the distribution of d, and extend the discussion to the investigate the statistical properties of Slatkin's F_{ST} estimator [45], and some of its various applications [37, 3, 17].

3.3 Mean, Variance, and Covariance of Average Pairwise Differences

In this section, we review previous results for the mean, variance and covariance of average pairwise nucleotide differences for individuals sampled from two populations, X and Y, as functions of the individual pairwise difference terms ($k_{i,i'}$, $k_{i,j}$...). Suppose i, i', i'', i''' are individuals from population X, and j, j', j'', j''' are individuals from population Y. By definition we have:

$$\mathbb{E}(d_X) = \mathbb{E}(k_{i,i'})$$

and likewise for population Y. Suppose i, j are individuals from X, Y respectively, then:

$$\mathbb{E}(d_{XY}) = \mathbb{E}(k_{i,j}).$$

Following the derivations in Tajima [48], Takahata and Nei [49], and Wakeley [53], the variance and covariance of d_X, d_Y, d_{XY}, and d can be written as follows:

$$\begin{aligned}
\mathrm{Var}(d_X) &= \frac{1}{n_X(n_X-1)}\Big[2\mathbb{E}(k_{i,i'}^2) + 4(n_X-2)\mathbb{E}(k_{i,i'}k_{i,i''}) + (n_X-2)(n_X-3) \\
&\quad \times \mathbb{E}(k_{i,i'}k_{i'',i'''})\Big] - \mathbb{E}(k_{i,i'})^2 \\
\mathrm{Var}(d_Y) &= \frac{1}{n_Y(n_Y-1)}\Big[2\mathbb{E}(k_{j,j'}^2) + 4(n_Y-2)\mathbb{E}(k_{j,j'}k_{j,j''}) + (n_Y-2)(n_Y-3) \\
&\quad \times \mathbb{E}(k_{j,j'}k_{j'',j'''})\Big] - \mathbb{E}(k_{j,j'})^2 \\
\mathrm{Var}(d_{XY}) &= \frac{1}{n_X n_Y}\Big[\mathbb{E}(k_{i,j}^2) + (n_Y-1)\mathbb{E}(k_{i,j}k_{i',j}) + (n_X-1)\mathbb{E}(k_{i,j}k_{i,j'}) + (n_X-1) \\
&\quad \times (n_Y-1)\mathbb{E}(k_{i,j}k_{i',j'})\Big] - \mathbb{E}(k_{ij})^2 \\
\mathrm{Var}(d) &= \mathrm{Var}(d_{XY}) + \frac{1}{4}\Big[\mathrm{Var}(d_X) + \mathrm{Var}(d_Y) + 2\mathrm{Cov}(d_X, d_Y)\Big] - \mathrm{Cov}(d_{XY}, d_X) \\
&\quad - \mathrm{Cov}(d_{XY}, d_Y)
\end{aligned}$$

Lastly, formulas for the covariance of average pairwise difference terms:

$$\mathrm{Cov}(d_X, d_Y) = \mathrm{Cov}(k_{i,i'}, k_{j,j'})$$

with the result due to the fact that the covariance of sums can be decomposed into the sums of covariances.

Derived in [49], covariance equations involving the cross population:

$$\mathrm{Cov}(d_{XY}, d_X) = \frac{2}{n_X}\mathbb{E}(k_{i,i'}k_{i,j}) + \frac{n_X - 2}{n_X}\mathbb{E}(k_{i,i'}k_{i'',j}) - \mathbb{E}(k_{i,i'})\mathbb{E}(k_{j,j'})$$
$$\mathrm{Cov}(d_{XY}, d_Y) = \frac{2}{n_Y}\mathbb{E}(k_{j,j'}k_{i,j}) + \frac{n_Y - 2}{n_Y}\mathbb{E}(k_{j,j'}k_{i,j''}) - \mathbb{E}(k_{i,i'})\mathbb{E}(k_{j,j'})$$

These expressions are all functions of the individual pairwise differences. In what proceeds we demonstrate that these expressions can be generalized as functions of pairwise coalescence times.

3.4 Pairwise Mutational Differences

In this section, we generalize previous work, [49, 48], by deriving expressions for the covariance of pairwise differences under arbitrary demographic settings using the coalescent. Throughout we will assume an infinite sites model [19, 55]. We first review results on the mean and variance from previous work (e.g., [48], [49],[53]).

Mean and Variance

Recall that for a given coalescence time $t_{i,j}$ between two lineages, i and j, the expected number of nucleotide differences between the pair is equal to $2\mu t_{i,j}$, i.e.

$$\mathbb{E}(k_{i,j}) = 2\mu\mathbb{E}(t_{i,j}).$$

Under the assumption that mutations can be modelled by a Poisson distribution, it follows that:

$$\mathrm{Var}(k_{i,j}|t_{i,j}) = \mathbb{E}(k_{i,j}|t_{i,j})$$

Applying the law of total variance, we see:

$$\begin{aligned}\sigma^2_{k_{i,j}} = \mathrm{Var}(k_{i,j}) &= \mathbb{E}\big(\mathrm{Var}(k_{i,j}|t_{i,j})\big) + \mathrm{Var}\big(\mathbb{E}(k_{i,j}|t_{i,j})\big)\\ &= \mathbb{E}(2\mu t_{i,j}) + \mathrm{Var}(2\mu t_{i,j})\\ &= 2\mu\mathbb{E}(t_{i,j}) + 4\mu^2\mathrm{Var}(t_{i,j})\end{aligned}$$

and we can then get the second moment of the distribution of pairwise nucleotide differences, $\mathbb{E}(k^2_{i,j})$, from the definition of variance:

$$\mathbb{E}(k^2_{i,j}) = \sigma^2_{k_{i,j}} + \mathbb{E}(k_{i,j})^2 = 2\mu\mathbb{E}(t_{i,j}) + 8\mu^2\mathbb{E}(t_{i,j})^2$$

Covariance

Let i, i', j, j' be four individuals from arbitrary populations. Let T be a local coalescent tree relating the four individuals at a non-recombining region of the genome. Here we show that:

$$\text{Cov}(k_{i,i'}, k_{j,j'}|T) = \mu t_{i,i'\cap j,j'} \tag{3.1}$$

and consequently, the unconditional quantity,

$$\text{Cov}(k_{i,i'}, k_{j,j'}) = \mu\mathbb{E}(t_{i,i'\cap j,j'}) + 4\mu^2\text{Cov}(t_{i,i'}, t_{j,j'}) \tag{3.2}$$

where $t_{i,i'\cap j,j'}$ denotes the amount of branch length on T shared between the branch connecting pair i, i' and the branch connecting pair j, j'. Figure 3.1 provides an illustrative example of this quantity, and the appendix section B.5 provides a more technical treatment. To prove these results, we start by revisiting the idea that the mutational process given a branch length follows a Poisson distribution. Given T, with coalescence times $t_{i,i'}$ and $t_{j,j'}$ from T, we know that

$$k_{i,i'}|t_{i,i'} \sim \text{Poisson}(2\mu t_{i,i'}) \quad \text{and} \quad k_{j,j'}|t_{j,j'} \sim \text{Poisson}(2\mu t_{j,j'})$$

where $2t_{i,i'}$ is the amount of total branch length between the two individuals. A key feature of the Poisson distribution is that the sum of Poisson random variables also follows a Poisson distribution. To exploit this, let $t_{i,i'\cap j,j'}$ denote the amount of branch length on T shared by pairs i, i' and j, j'. The branch length between i, i' not shared with pair j, j' is denoted by $t_{i,i'\setminus j,j'}$, with similar notation for pair j, j' by swapping labels. We can decompose the branch lengths into the shared and non-shared segments as:

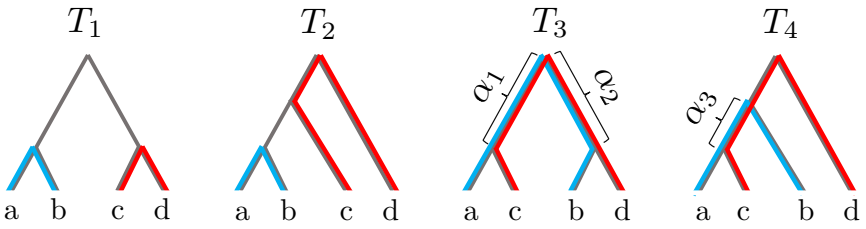
$$2t_{i,i'} = t_{i,i'\cap j,j'} + t_{i,i'\setminus j,j'} \quad \text{and} \quad 2t_{j,j'} = t_{i,i'\cap j,j'} + t_{j,j'\setminus i,i'}$$

Notice that $k_{i,i'\cap j,j'}|T$, $k_{i,i'\setminus j,j'}|T$, and $k_{j,j'\setminus i,i'}|T$ are independent Poisson random variables. Similarly, $k_{i,i'} = k_{i,i'\cap j,j'} + k_{i,i'\setminus j,j'}$ and $k_{j,j'} = k_{i,i'\cap j,j'} + k_{j,j'\setminus i,i'}$, where $k_{i,i'\cap j,j'}$, $k_{j,j'\setminus i,i'}$ and $k_{i,i'\setminus j,j'}$ are independent of each other conditionally on T.

We can expand $\text{Cov}(k_{i,i'}, k_{j,j'}|T)$ as follows:

$$\begin{aligned}
\text{Cov}(k_{i,i'}, k_{j,j'}|T) &= \text{Cov}(k_{i,i'\cap j,j'} + k_{i,i'\setminus j,j'}, k_{i,i'\cap j,j'} + k_{j,j'\setminus i,i'}|T) \\
&= \text{Var}(k_{i,,i'\cap j,j'}|T) + \text{Cov}(k_{i,i'\cap j,j'}, k_{i,i'\setminus j,j'}|T) \\
&\quad + \text{Cov}(k_{i,i'\cap j,j'}, k_{j,j'\setminus i,i'}|T) + \text{Cov}(k_{i,i'\setminus j,j'}, k_{j,j'\setminus i,i'}|T) \\
&= \text{Var}(k_{i,i'\cap j,j'}|T) \\
&= \mu t_{i,i'\cap j,j'}
\end{aligned}$$

A Possible topologies



B Shared branch length

$t_{a,b\cap c,d}$	$=$	0	0	$\alpha_1 + \alpha_2$	α_3

C Expected shared branch length

$$\mathbb{E}(t_{a,b\cap c,d}) = \mathbb{E}(\alpha_1 + \alpha_2|T_3)P(T_3) + \mathbb{E}(\alpha_3|T_4)P(T_4)$$

Figure 3.1: (A-C) Explanation of expected shared branch length for 4 unique individuals. Blue lines indicate the branch length between individuals a and b. Red lines indicate branch length between c and d. Overlapping blue and red lines (along with α terms) indicate shared branch length. The 4 tree topologies are representative of the possible gene tree orderings, but it should be noted that these representative trees assume a and b are exchangeable, as well as c and d. The expected shared branch length is a weighted sum of the shared branch lengths across all possible topology orderings.

The overall result is that the covariance of pairwise differences given the coalescent tree T is equal to the mutation rate times the shared branch length.

To get the unconditional quantity, $\text{Cov}(k_{i,i'}, k_{j,j'})$, we apply the law of total covariance:

$$\begin{aligned}
\text{Cov}(k_{i,i'}, k_{j,j'}) &= \mathbb{E}\Big(\text{Cov}(k_{i,i'}, k_{j,j'}|T)\Big) + \text{Cov}\Big(\mathbb{E}(k_{i,i'}|T), \mathbb{E}(k_{j,j'}|T)\Big) \\
&= \mathbb{E}(\mu t_{i,i'\cap j,j'}) + \text{Cov}(2\mu t_{i,i'}, 2\mu t_{j,j'}) \\
&= \mu\mathbb{E}(t_{i,i'\cap j,j'}) + 4\mu^2\text{Cov}(t_{i,i'}, t_{j,j'})
\end{aligned}$$

The case when we have only three unique individuals $(k_{i,i'}, k_{i,j})$ has the same form, by replacing j' with i in the equations above.

Takahata and Nei [49] have previously derived formulas for the covariance under constant population size, see Appendix section B.3 which presents their results and

a comparison to the generalized results presented here.

3.5 Mean, Variance and Covariance in Pairwise Coalescence Times

We assume species evolution follows a bifurcating species tree $\mathcal{S} = (S, \tau, \eta)$, with no migration. Each branch, i, of $\mathcal{S}$ is a parameterized by constant diploid population size η_i, start time, τ_i, and end time $\tau_{p(i)}$, where $p(i)$ is the parent branch of i. Let μ be the mutation rate (constant across the genome/species) per sequence per generation. Time is measured in units of generations in the past. We implicitly assume that all coalescent calculations here are conditioned on a fixed species tree $\mathcal{S}$, although the tree is not always indicated in the notation for the sake of simplicity.

Mean and Variance in Coalescence Times

Let $t_{i,j}$ be the coalescence time of two individuals, i and j sampled from species X and Y, respectively respectively. For species tree $\mathcal{S}$, denote the marginal tree $\mathcal{S}_{XY} = (\tau_{XY}, \eta_{XY})$ of two species, where τ_{XY} represents the set of divergence times of species ancestral to both X and Y, indexed by $(\tau_1, \tau_2, ...)$, where $\tau_1 := D_{XY}$, the divergence time for species X and Y. Similarly, η_{XY} represents the corresponding population sizes. Suppose there are $V \geq 1$ intervals in $\mathcal{S}_{XY}$.

Under this marginal tree, we can analytically calculate the first two moments of the distribution of $t_{i,j}$ as:

$$\begin{aligned}
\mathbb{E}(t_{i,j}|\mathcal{S}) &= \sum_{k=1}^{V} P_{22}(\tau_1,\tau_k)\int_{\tau_k}^{\tau_{k+1}} t_{i,j}P(t_{i,j}|\mathcal{S},\tau_k)dt_{i,j} \\
&= \sum_{k=1}^{V} P_{22}(\tau_1,\tau_k)\int_{\tau_k}^{\tau_{k+1}} \frac{t_{i,j}}{2\eta_k}e^{-\frac{(t_{i,j}-\tau_k)}{2\eta_k}}dt_{i,j} \\
&= \sum_{k=1}^{V} P_{22}(\tau_1,\tau_k)\big[-(\tau_{k+1}+2\eta_k)e^{-\frac{\tau_{k+1}-\tau_k}{2\eta_k}}+\tau_k+2\eta_k\big]
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}(t_{i,j}^2|\mathcal{S}) &= \sum_{k=1}^{V} P_{22}(\tau_1,\tau_k)\int_{\tau_k}^{\tau_{k+1}} t_{i,j}^2P(t_{i,j}|\mathcal{S},\tau_k)dt_{i,j} \\
&= \sum_{k=1}^{V} P_{22}(\tau_1,\tau_k)\int_{\tau_k}^{\tau_{k+1}} \frac{t_{i,j}^2}{2\eta_k}e^{-\frac{(t_{i,j}-\tau_k)}{2\eta_i}}dt_{i,j} \\
&= \sum_{k=1}^{V} P_{22}(\tau_1,\tau_k)\big[-(\tau_{k+1}^2+4\tau_{k+1}\eta_k+8\eta_k^2)e^{-\frac{(\tau_{k+1}-\tau_k)}{2\eta_k}}+\tau_k^2+4\tau_k\eta_k+8\eta_k^2\big]
\end{aligned}$$

Where $P_{22}(\tau_1,\tau_k)$ represents the probability that lineages i and j fail to coalesce in the time interval (τ_1,τ_k). $P_{22}(\tau_1,\tau_k)$ is the probability that 2 lineages which enter the time interval at τ_1 (backwards in time) have not coalesced by time τ_k:

$$P_{22}(\tau_1,\tau_k) = \prod_{\tau_1\le\tau_l<\tau_k} e^{-\frac{(\tau_{l+1}-T_l)}{2\eta_l}}$$

Note that the mean $\mathbb{E}(t_{i,j}|\mathcal{S})$, and variance $\mathrm{Var}(t_{i,j}|\mathcal{S}) = \mathbb{E}(t_{i,j}^2|\mathcal{S}) - \mathbb{E}(t_{i,j}|\mathcal{S})^2$ of coalescence times under the standard piecewise constant coalescent process are just a simply weighted sums over coalescence intervals.

Covariance in Pairwise Coalescence Times

The challenge in calculating the covariance terms from a species tree, $\mathcal{S}$, comes from the combinatorial problem of integrating over all of the possible times and orderings of the coalescent events along the multi-species tree. The general formula for covariance in this case is

$$\mathrm{Cov}(t_{i,i'},t_{j,j'}|\mathcal{S}) = \mathbb{E}(t_{i,i'}t_{j,j'}|\mathcal{S}) - \mathbb{E}(t_{i,i'}|\mathcal{S})\mathbb{E}(t_{j,j'}|\mathcal{S})$$

where the last term is simply a product of independent expectations. The first term on the right hand side of the equation is what we will focus on, in particular we write:

$$\mathbb{E}(t_{i,i'}t_{j,j'}|\mathcal{S}) = \int_{D_{j,j'}}^{\infty} t_{j,j'}P(t_{j,j'}|\mathcal{S}) \int_{D_{i,i'}}^{\infty} t_{i,i'}P(t_{i,i'}|t_{j,j'},\mathcal{S})dt_{i,i'}dt_{j,j'}$$

where $D_{i,i'}$ is the species divergence time between individuals i, i' from $\mathcal{S}$, where $D_{i,i'} = 0$ if i, i' are of the same species (similarly for $D_{j,j'}$). We assume all coalescence events must be at least as ancient as the species divergence time, (e.g. $t_{j,j'} \geq D_{j,j'}$), i.e. we assume no introgression or admixture.

To evaluate this quantity, $\mathbb{E}(t_{i,i'}t_{j,j'}|\mathcal{S})$, we consider 6 separate conditional cases. Recall for a bifurcating tree of 4 individuals, there are 3 unique coalescence events. The 6 cases correspond to the possible orderings of coalescence events for this local tree of 4 individuals, given that we structure the joint likelihood as $P(t_{i,i'}|t_{j,j'},\mathcal{S})$ $\times P(t_{j,j'}|\mathcal{S})$:

C1. $t_{i,i'}$ is the first coalescent event.
C2. $t_{i,i'}$ is the second event, $t_{j,j'}$ is the third.
C3. $t_{i,i'} = t_{j,j'}$ as the third coalescent event.
C4. $t_{j,j'}$ is the second event, $t_{i,i'}$ is the third.
C5. $t_{j,j'}$ is the first event, $t_{i,i'}$ is the second.
C6. $t_{j,j'}$ is the first event, $t_{i,i'}$ is the third.

Here, "first event" implies most recent, and "third" implies most ancient. Conditioning on each of these 6 events, and evaluating each expectation separately, the expression for the joint expectation becomes:

$$\mathbb{E}(t_{i,i'}t_{j,j'}|\mathcal{S}) = \sum_{k=1}^{6} \mathbb{E}(t_{i,i'}t_{j,j'}|\mathcal{S},\mathrm{C}_k)P(\mathrm{C}_k|\mathcal{S})$$

In the presence of no population isolation (all individuals from the same species), but piecewise constant population size history, the set of recursions and integrals is presented in its entirety in the appendix. This calculation is useful in the instance that all 4 lineages survive to a common population without having coalesced with one another, which occurs with some probability in each case.

Introducing a species tree structure on top of the 6 cases increases the number of cases to consider. There are 5 general possible species tree configurations that can arise, see figure B.11, located in the appendix. We have derived exact equations and recursions to evaluate all 6 cases across the 5 general possible tree configurations, and have implemented them in C++ code which is freely available to use (more information in the code availability section). From this implementation we are able to

calculate exact theoretical quantities for these statistics under any piecewise constant scenario.

3.6 Accuracy of Coalescent Calculations

To demonstrate the accuracy of the coalescent equations above, as implemented in our software, STCov, we compare the theoretical results against empirical estimates from gene trees using ms, [16]. We test 2 demographic scenarios for a tree of species X and Y: $\eta_Y = \eta_X$, and $\eta_Y = 2\eta_X$, where η represents scaled effective population size. We assume $\eta_{XY} = \eta_X$ in both scenarios. Let lineages i_1, i_2, i_3 originate in population X, and lineages j_1, j_2, j_3 originate in Y. We generate 500 independent gene trees from ms for each demographic scenario (population sizes and divergence time), and calculate sample mean, variance, and covariance terms. Overall we see that the theoretical calculations from STCov match simulations (dots) well, while variation in the empirical estimates can be attributed to a finite sample size.

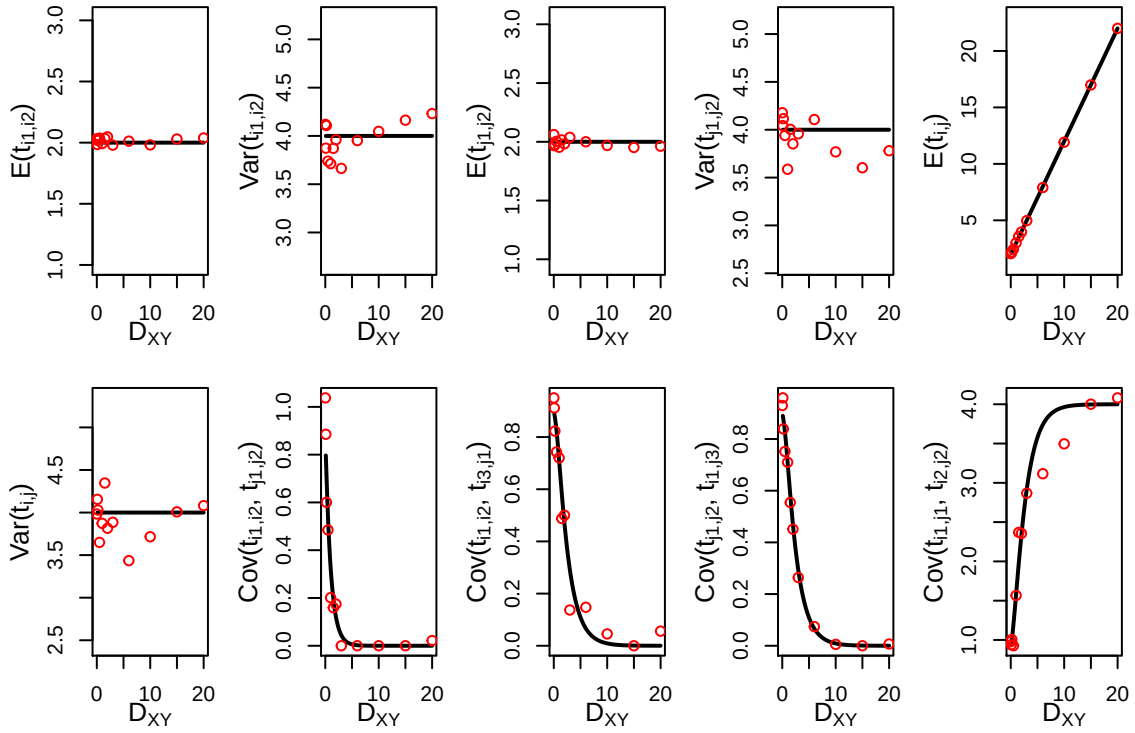


Figure 3.2: $\eta_Y = \eta_X$. Theoretical results from STCov are plotted in black, with dots representing empirical estimates from 500 independent local trees.

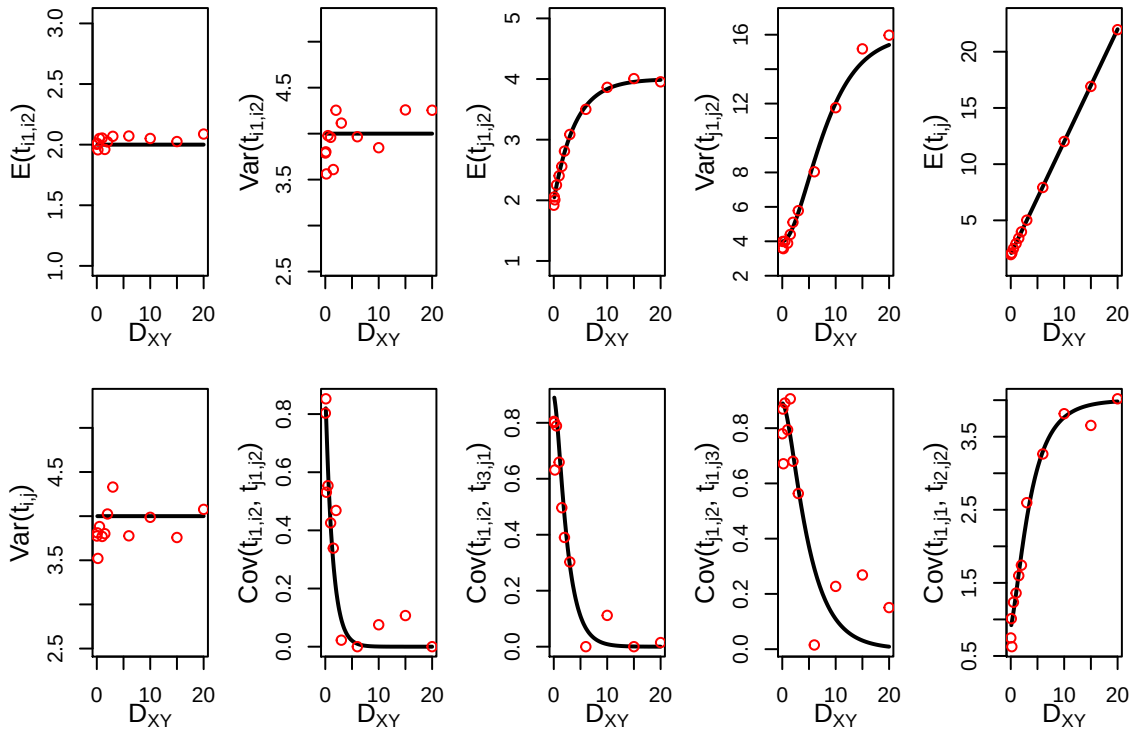


Figure 3.3: $\eta_Y = 2\eta_X$. Theoretical results from STCov are plotted in black, with dots representing empirical estimates from 500 independent local trees.

3.7 Accuracy of Pairwise Difference Calculations

In this section we evaluate the accuracy of our results under varying mutation rates, divergence times and population sizes. We compare our results to simulated data sets.

We compare 3 population size change models, denoted by $\eta_Y = 1\eta_X$, $\eta_Y = 2\eta_X$ and $\eta_Y = 10\eta_X$, and 3 mutation rates $2\mu\eta_X = 10, 1, 0.1$, for a total of 9 simulation scenarios. We present 1 of those scenarios here (figure 3.4), and leave the full set of results to the appendix. While allowing for variance in the empirical estimates from sample size, coalescent and mutational variation, there is strong agreement between the theoretical and simulated results. Note that the theoretical quantities assume an infinite-sites model of mutation, whereas our simulations are performed assuming a realistic, finite-sites model. We choose to compare this finite-sites model over simulations using a model of infinite sites to demonstrate the applicability of the results to the types of data that will be used in practice.

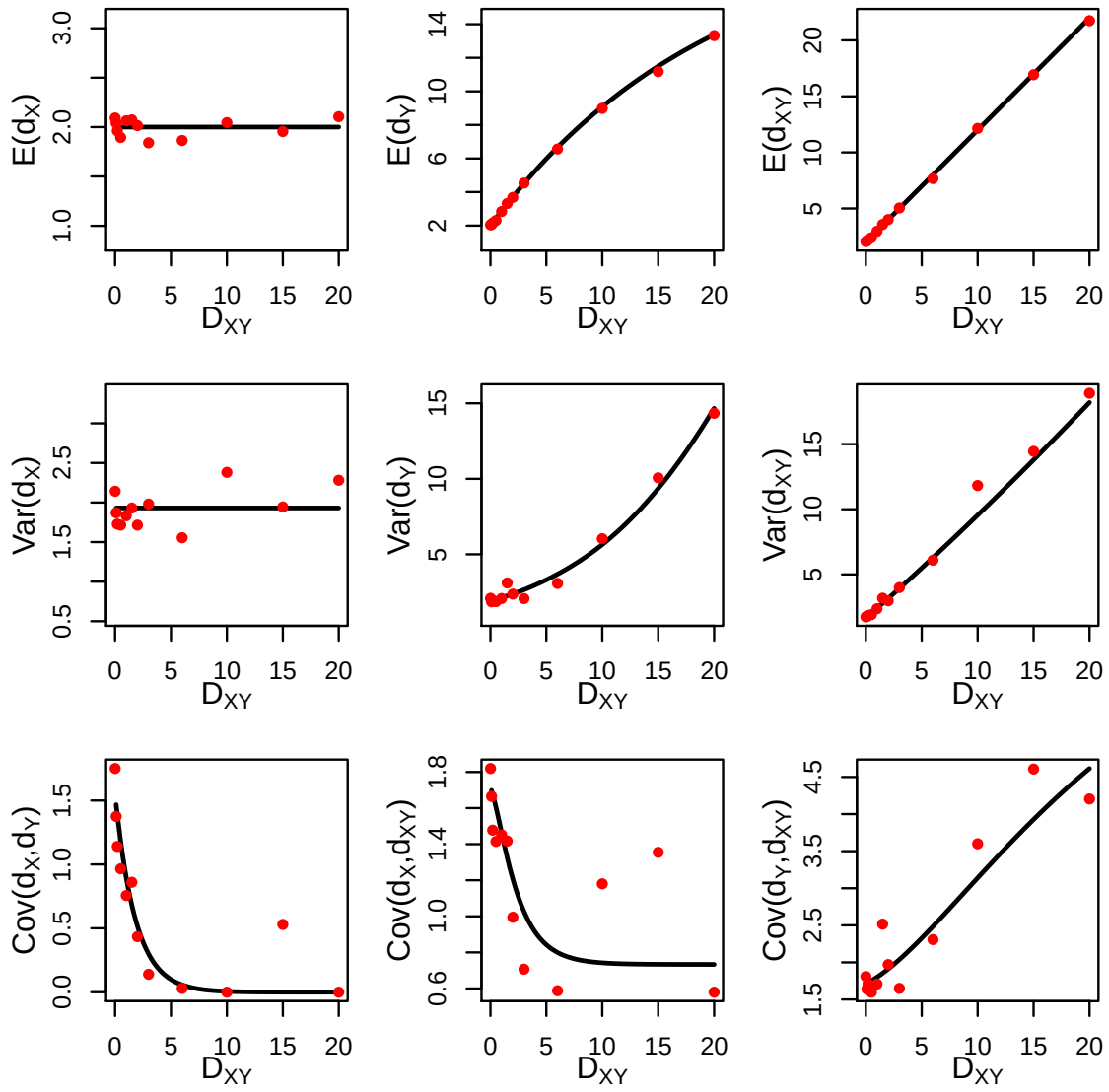


Figure 3.4: **Average Pairwise Coalescent results,** $2\mu\eta_X = 1$, $\eta_Y = 10\eta_X$**.** We compare our theoretical results (black line) with simulated estimated values from 250 independent genes (red dots), $n_X = n_Y = 10$ sampled individuals.

3.8 Accuracy in Approximating F_{ST}

A direct extension of our discussion on the mean and variance of average pairwise nucleotide differences is to the measurement, F_{ST}, for a given species tree, mutation

rate, and sample size. Slatkin 1991 [45] presented a coalescent-based definition of $\mathrm{F_{ST}}$ as a function of the difference in expected time to coalescence:

$$\mathrm{F_{ST}} = \frac{\mathbb{E}(t_{i,j}) - \frac{1}{2}\big(\mathbb{E}(t_{i,i'}) + \mathbb{E}(t_{j,j'})\big)}{\mathbb{E}(t_{i,j})} \tag{3.3}$$

Where i, i' are from population X, and j, j' are individuals sampled from population Y. This definition of $\mathrm{F_{ST}}$ relies on estimates of average coalescence times, where average pairwise differences in DNA sequence data are used as the proxy for the unknown coalescence times. Discussed in [45, 17], for two populations X and Y, $\mathrm{F_{ST}}$ can be estimated from sequence data using:

$$\mathrm{F_{ST}} \approx \frac{d_{XY} - \frac{1}{2}(d_X + d_Y)}{d_{XY}} \overset{define}{=} \mathrm{F^G_{ST}} \tag{3.4}$$

As we have shown above, we can get exact expressions for the expectation, variance and covariance of these sample average pairwise differences from theory, for a given mutation parameter μ and sample size. We can use these to study the accuracy of the $\mathrm{F^G_{ST}}$ estimator to Slatkin's $\mathrm{F_{ST}}$ under an arbitrary species tree, $\mathcal{S}$.

To begin, it is important to note that the mean of a ratio is not the ratio of means, specifically it is the case that:

$$\mathbb{E}(\mathrm{F^G_{ST}}) \neq \frac{\mathbb{E}(d_{XY}) - \frac{1}{2}\big(\mathbb{E}(d_X) + \mathbb{E}(d_Y)\big)}{\mathbb{E}(d_{XY})} = \frac{2\mu\mathbb{E}(t_{i,j}) + \mu\big(\mathbb{E}(t_{i,i'}) + \mathbb{E}(t_{j,j'})\big)}{2\mu\mathbb{E}(t_{i,j})} = \mathrm{F_{ST}}$$

This implies that the estimator $\mathrm{F^G_{ST}}$ is potentially a biased estimator of $\mathrm{F_{ST}}$, such that $\mathrm{F_{ST}} - \mathbb{E}(\mathrm{F^G_{ST}}) \neq 0$. To study this bias, we need an expression for the mean of $\mathrm{F^G_{ST}}$. In general, there is no closed form for the mean of a ratio of dependent random variables, so we will first simplify our terms, and then approximate the mean and variance using a Taylor expansion. We can first simplify the expressions for $\mathbb{E}(\mathrm{F^G_{ST}})$:

$$\mathbb{E}(\mathrm{F^G_{ST}}) = \mathbb{E}\Big(\frac{d_{XY} - \frac{1}{2}(d_X + d_Y)}{d_{XY}}\Big) = 1 - \frac{1}{2}\mathbb{E}\Big(\frac{d_X + d_Y}{d_{XY}}\Big)$$

$$\mathrm{Var}(\mathrm{F^G_{ST}}) = \mathrm{Var}\Big(1 - \frac{1}{2}\frac{d_X + d_Y}{d_{XY}}\Big) = \frac{1}{4}\mathrm{Var}\Big(\frac{d_X + d_Y}{d_{XY}}\Big)$$

We are now interested in the mean and variance of the ratio $(d_X + d_Y)/d_{XY}$. We can use a second order Taylor expansion of $f(A, B) = \frac{A}{B}$ around the mean values, $(\mathbb{E}(d_X) + \mathbb{E}(d_Y), \mathbb{E}(d_{XY}))$, to get an approximation to the mean, and a first order expansion around the means to get an approximation of the variance of the ratio term:

$$\begin{aligned}\mathbb{E}\Big(\frac{d_X + d_Y}{d_{XY}}\Big) &\approx \frac{\mathbb{E}(d_X) + \mathbb{E}(d_Y)}{\mathbb{E}(d_{XY})} - \frac{\mathrm{Cov}(d_X + d_Y, d_{XY})}{\mathbb{E}(d_{XY})^2} + \frac{\mathbb{E}(d_X) + \mathbb{E}(d_Y)}{\mathbb{E}(d_{XY})^3}\mathrm{Var}(d_{XY}) \\ &= \frac{\mathbb{E}(d_X) + \mathbb{E}(d_Y)}{\mathbb{E}(d_{XY})} - \frac{1}{\mathbb{E}(d_{XY})^2}\Big[\mathrm{Cov}(d_X, d_{XY}) + \mathrm{Cov}(d_Y, d_{XY})\Big] \\ &\quad + \frac{\mathbb{E}(d_X) + \mathbb{E}(d_Y)}{\mathbb{E}(d_{XY})^3}\mathrm{Var}(d_{XY})\end{aligned}$$

By rearranging terms, observe that $\mathbb{E}(\mathrm{F}_{\mathrm{ST}}^G)$ is a function of F_{ST}, along with other mean, variance, and covariance terms.

$$\begin{aligned}\mathbb{E}(\mathrm{F}_{\mathrm{ST}}^G) &= 1 - \frac{1}{2}\mathbb{E}\Big(\frac{d_X + d_Y}{d_{XY}}\Big) \\ &\approx \frac{1}{2\mathbb{E}(d_{XY})^2}\left(\mathrm{Cov}(d_X, d_{XY}) + \mathrm{Cov}(d_Y, d_{XY}) - \frac{\mathbb{E}(d_X) + \mathbb{E}(d_Y)}{\mathbb{E}(d_{XY})}\mathrm{Var}(d_{XY})\right) \\ &\quad + \mathrm{F}_{\mathrm{ST}}^E\end{aligned} \tag{3.5}$$

Using this, we can get an expression for the bias of $\mathbb{E}(\mathrm{F}_{\mathrm{ST}}^G)$:

$$\begin{aligned}&\mathbb{E}(\mathrm{F}_{\mathrm{ST}}^G) - \mathrm{F}_{\mathrm{ST}} \\ &\quad\approx \frac{1}{2\mathbb{E}(d_{XY})^2}\left(\mathrm{Cov}(d_X, d_{XY}) + \mathrm{Cov}(d_Y, d_{XY}) - \frac{\mathbb{E}(d_X) + \mathbb{E}(d_Y)}{\mathbb{E}(d_{XY})}\mathrm{Var}(d_{XY})\right)\end{aligned} \tag{3.6}$$

Similarly, we can get an first-order approximation for the variance of $\mathrm{F}_{\mathrm{ST}}^{G}$:

$$\begin{aligned}
\mathrm{Var}(\mathrm{F}_{\mathrm{ST}}^{G}) &= \frac{1}{4}\mathrm{Var}\Big(\frac{d_X + d_Y}{d_{XY}}\Big) \\
&\approx \frac{1}{4}\Big(\frac{\mathrm{Var}(d_X + d_Y)}{\big(\mathbb{E}(d_X) + \mathbb{E}(d_Y)\big)^2} + \frac{\big(\mathbb{E}(d_X) + \mathbb{E}(d_Y)\big)^2}{\mathbb{E}(d_{XY})^4}\mathrm{Var}(d_{XY}) - 2\frac{\mathbb{E}(d_X) + \mathbb{E}(d_Y)}{\mathbb{E}(d_{XY})^3} \\
&\quad \times \mathrm{Cov}(d_X + d_Y, d_{XY})\Big) \\
&= \frac{1}{4}\Bigg(\frac{\mathrm{Var}(d_X) + \mathrm{Var}(d_Y) + 2\mathrm{Cov}(d_X, d_Y)}{\big(\mathbb{E}(d_X) + \mathbb{E}(d_Y)\big)^2} + \frac{\big(\mathbb{E}(d_X) + \mathbb{E}(d_Y)\big)^2}{\mathbb{E}(d_{XY})^4}\mathrm{Var}(d_{XY}) \\
&\quad - 2\frac{\mathbb{E}(d_X) + \mathbb{E}(d_Y)}{\mathbb{E}(d_{XY})^3}\big(\mathrm{Cov}(d_X, d_{XY}) + \mathrm{Cov}(d_Y, d_{XY})\big)\Bigg)
\end{aligned}$$

Figure 3.5 shows the accuracy of the two Taylor approximations under a constant population size model. The approximation for the mean is a good one, however the first-order approximation to the variance is insufficient for low divergence times, as it can be seen there are higher order terms involved. From this we decide that we cannot approximate the variance in $\mathrm{F}_{\mathrm{ST}}^{G}$ well with this method, and do not pursue this aspect further.

In what follows we will evaluate the bias in the $\mathrm{F}_{\mathrm{ST}}^{G}$ estimator under different demographic and genetic parameters, using results for the mean of $\mathrm{F}_{\mathrm{ST}}^{G}$.

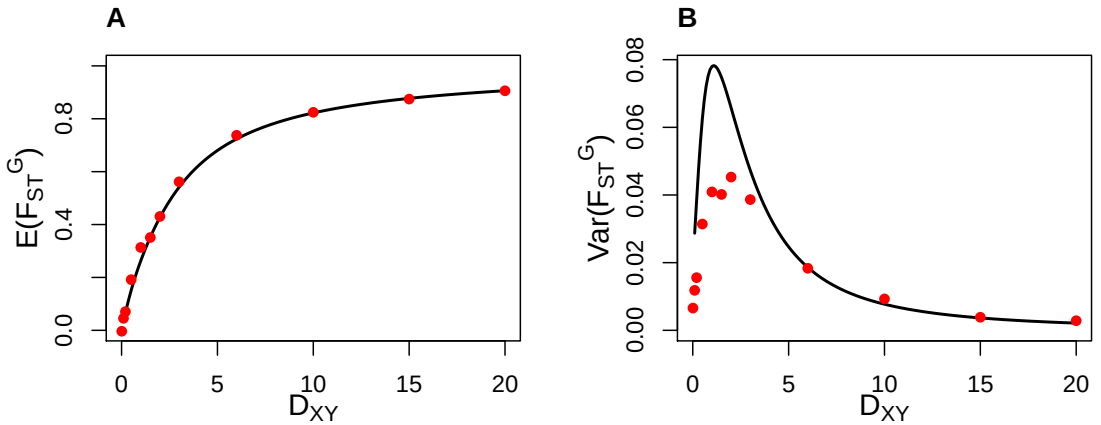


Figure 3.5: **$\mathbf{F_{ST}}$ mean and variance approximation accuracy,** $2\mu\eta_X = 1, \eta_Y = 1\eta_X$**.** (A) The approximated value to $\mathbb{E}(\mathrm{F}_{\mathrm{ST}}^{G})$ is shown as a black curve as a function of the divergence time D_{XY} for equal sample sizes n_x, n_y. (B) The first-order approximation for the variance $\mathrm{Var}(\mathrm{F}_{\mathrm{ST}}^{G})$ as a function of the divergence time.

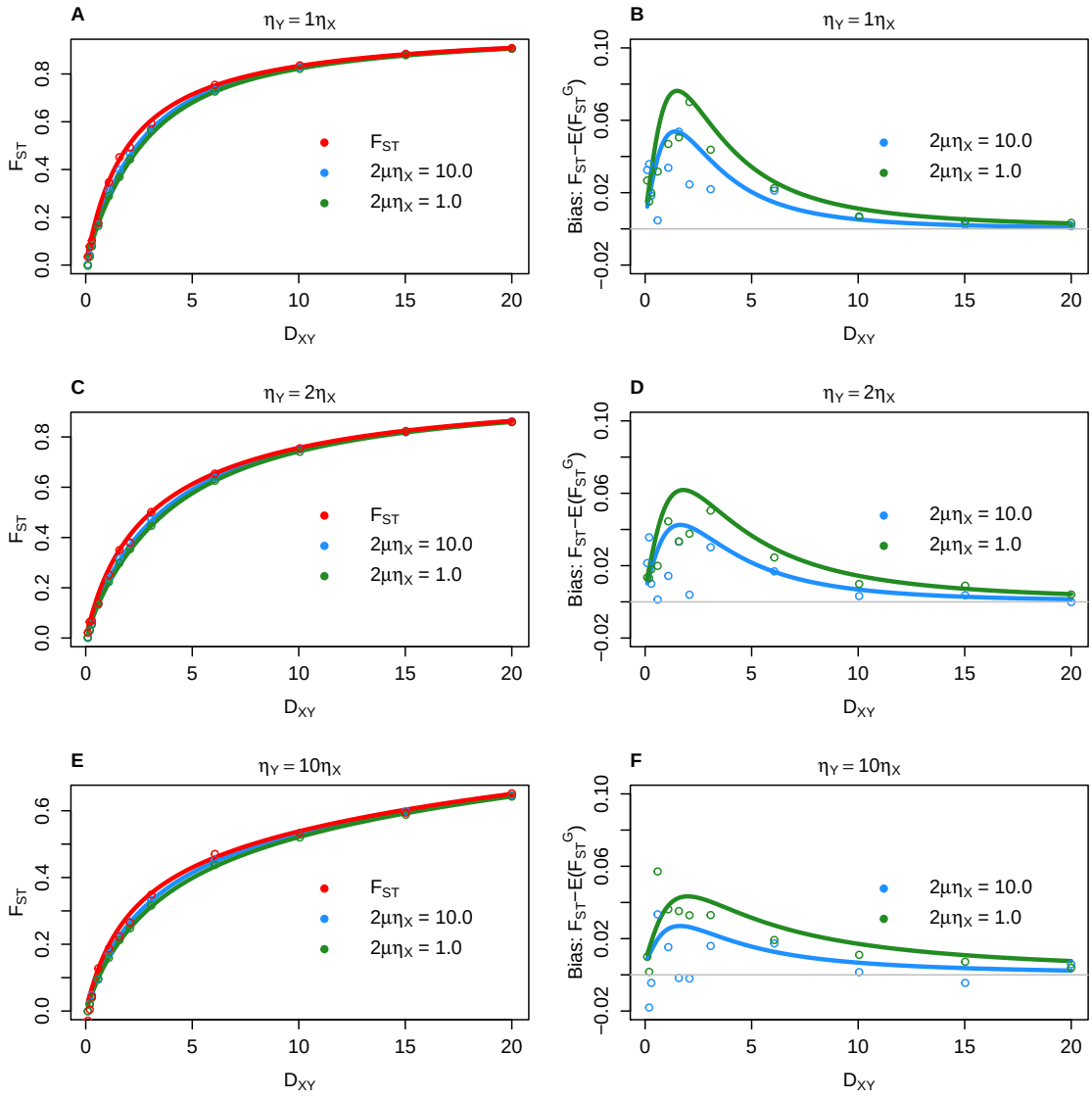


Figure 3.6: **F_{ST} approximation bias across divergence times.** (A,C,E) On the y axis are values $\mathbb{E}(F_{ST}^G)$ and F_{ST} as functions of divergence time D_{XY}. We plot the true value of F_{ST} in red, and approximations $\mathbb{E}(F_{ST}^G)$, for 2 mutation rates. (B,D,E) The difference between the true F_{ST} (red line in adjacent plot) and the expected sample quantity, to represent the bias in estimation. We simulated assuming equal sample sizes $n_X = n_Y = 10$. Each row of plots corresponds to different population size history as indicated at the top of each. In all figures, dots represent simulated estimates from 250 independent genes.

Results for the Mean and Bias of $\mathbf{F}_{\mathbf{ST}}^{G}$

In this section, we study the effects of varying demographic, and genetic parameters on the expectation of $\mathrm{F}_{\mathrm{ST}}^{G}$ and consequently its bias as an estimator of F_{ST}. First we start with a discussion on the differences between $\mathbb{E}(\mathrm{F}_{\mathrm{ST}}^{G})$ and F_{ST}, both as described above. Suppose we had access to the true values, we calculate F_{ST} only using the individual expectations of d_X, d_Y, and d_{XY}. We can write:

$$\mathrm{F}_{\mathrm{ST}} = \frac{\mathbb{E}(d_{XY}) - \frac{1}{2}\big(\mathbb{E}(d_X) + \mathbb{E}(d_Y)\big)}{\mathbb{E}(d_{XY})} = 1 - \frac{1}{2}\frac{\mathbb{E}(d_X) + \mathbb{E}(d_Y)}{\mathbb{E}(d_{XY})} = 1 - \frac{1}{2}\frac{\mathbb{E}(t_{i,i'}) + \mathbb{E}(t_{j,j'})}{\mathbb{E}(t_{i,j})}$$

Immediately we can note that F_{ST} is not dependent on n_X, n_Y or the mutation rate, μ. Instead, it is solely a function of mean coalescence times, and is only variable in the demographic parameter space. Also, notice the fundamental difference between $\mathbb{E}(\mathrm{F}_{\mathrm{ST}}^{G})$ and F_{ST} is the term

$$\mathbb{E}\Big(\frac{d_X + d_Y}{d_{XY}}\Big) \text{ vs. } \frac{\mathbb{E}(d_X) + \mathbb{E}(d_Y)}{\mathbb{E}(d_{XY})}$$

It is known that ratio estimators are in general biased. Jensen's inequality tells us that for two random variables, A, B ,

$$\mathbb{E}\Big(\frac{A}{B}\Big) \geq \frac{\mathbb{E}(A)}{\mathbb{E}(B)}$$

with equality holding when A and B are independent. While $d_X + d_Y$ and d_{XY} are not independent due to a shared ancestry, this gives a sense of direction of the bias. As the divergence time between X and Y becomes deeper (more ancient), we expect $d_X + d_Y$ to become increasingly independent from d_{XY} and $\mathbb{E}(\mathrm{F}_{\mathrm{ST}}^{G})$ to become increasingly closer to F_{ST}. Figure 3.6 demonstrates the relationship between $\mathbb{E}(\mathrm{F}_{\mathrm{ST}}^{G})$ and F_{ST} under varying divergence times D_{XY}, population sizes, and mutation rates μ. As discussed above, the relative bias of $\mathrm{F}_{\mathrm{ST}}^{G}$ is much less under a deep divergence model ($D_{XY} = 20.0$, in units of $2\eta_X$ generations) as d_X, d_Y and d_{XY} are more independent, compared to a more shallow divergence ($D_{XY} = 1.0$), where we see in our example F_{ST} is 3 times as large as $\mathbb{E}(\mathrm{F}_{\mathrm{ST}}^{G} | 2\mu\eta_X = 0.1)$. It is clear that $\mathrm{F}_{\mathrm{ST}}^{G}$ is a good estimator of F_{ST} under very high mutation rates, however, it is biased downwards for small values of μ, although the bias is reduced for deep divergence models.

Bias in the F_{ST} estimator for gene flow

The value of F_{ST} is often used to estimate levels of gene flow between populations. Wright [56] first derived the relationship between F_{ST} to estimate Nm in an Island models, where N is the number of individuals in each deme (sub-population), and m is the fraction of migrants into the deme in each generation. Hudson, Slatkin, and Maddison [17] use this relationship to estimate Nm using the following expression:

$$< Nm >_F = \frac{1}{2}\Big(\frac{1}{F_{ST}} - 1\Big) \tag{3.7}$$

where F_{ST} is an estimate from sequence data, i.e., F_{ST}^G in our notation. The results of the simulations done in the paper show estimates using $< Nm >_F$ are upward-biased. There are two potential sources of this bias, the estimator function, $< Nm >_F$, and the estimate, F_{ST}^G. The scope of this paper concerns the role of estimator F_{ST}^G, and we can study the effect of this estimator compared to using the true value, F_{ST}. We note that we do not intend to estimate or study gene flow in this manuscript, but simply evaluate the accuracy of the function $< Nm >_F$ when an estimate of F_{ST} is used.

To start, we can once again use a Taylor expansion to get an approximation for the mean of $< Nm >_F$, when using F_{ST}^G:

$$\begin{aligned}
\mathbb{E}(< Nm >_F) &= \frac{1}{4}\mathbb{E}\Big(\frac{d_X + d_Y}{d_{XY} - \frac{1}{2}(d_X + d_Y)}\Big) \\
&= \frac{1}{4}\frac{\mathbb{E}(d_X) + \mathbb{E}(d_Y)}{E(d_{XY}) - \frac{1}{2}\big(\mathbb{E}(d_X) + \mathbb{E}(d_Y)\big)} \\
&\times \Bigg[1 - \frac{\mathrm{Cov}(d_X + d_Y, d_{XY}) - \frac{1}{2}\big(\mathrm{Var}(d_X) + \mathrm{Var}(d_Y)\big) - \mathrm{Cov}(d_X, d_Y)}{\big(\mathbb{E}(d_X) + \mathbb{E}(d_Y)\big)\big(\mathbb{E}(d_{XY}) - \frac{1}{2}\big(\mathbb{E}(d_X) - \mathbb{E}(d_Y)\big)\big)} \\
&+ \frac{\mathrm{Var}(d_{XY}) + \frac{1}{4}\big(\mathrm{Var}(d_X) + \mathrm{Var}(d_Y) + 2\mathrm{Cov}(d_X, d_Y)\big)}{\Big(\mathbb{E}(d_{XY}) - \frac{1}{2}\big(\mathbb{E}(d_X) - \mathbb{E}(d_Y)\big)\Big)^2} \\
&- \frac{\mathrm{Cov}(d_X, d_{XY}) + \mathrm{Cov}(d_Y, d_{XY})}{\Big(\mathbb{E}(d_{XY}) - \frac{1}{2}\big(\mathbb{E}(d_X) - \mathbb{E}(d_Y)\big)\Big)^2}\Bigg]
\end{aligned} \tag{3.8}$$

We can use this expression to study the difference between using the estimator F_{ST}^G and the (unknown) true value, F_{ST} in the expression for $< Nm >_F$. Figure 3.7 shows the difference between using F_{ST} and F_{ST}^G in $< Nm >_F$ under different mutation rates, population sizes, and species divergence times. From the figure we see that

the expectations are, in fact, overestimates. In our figure, 10 individuals are sampled from each population. We see that when the divergence time D_{XY} is low, the bias relative to the true value is substantial, resulting in an estimate twice as large as that would have been obtained using an accurate estimate of F_{ST}. For high values of the mutation rate, μ, this bias decreases rapidly as D_{XY} increases. For a low mutation rate, $2\mu\eta_X$, a bias of greater than 50% overestimation persists. Even at high mutation rates, an upwards bias of about approximately 5% exists even at large divergence time values. Note, however, that we do not see a large difference in the bias across different population size models. The results here can explain (at least a portion of) the bias seen in [17], that using an estimate of F_{ST} can result in an artificial increase in the function $< Nm >_F$.

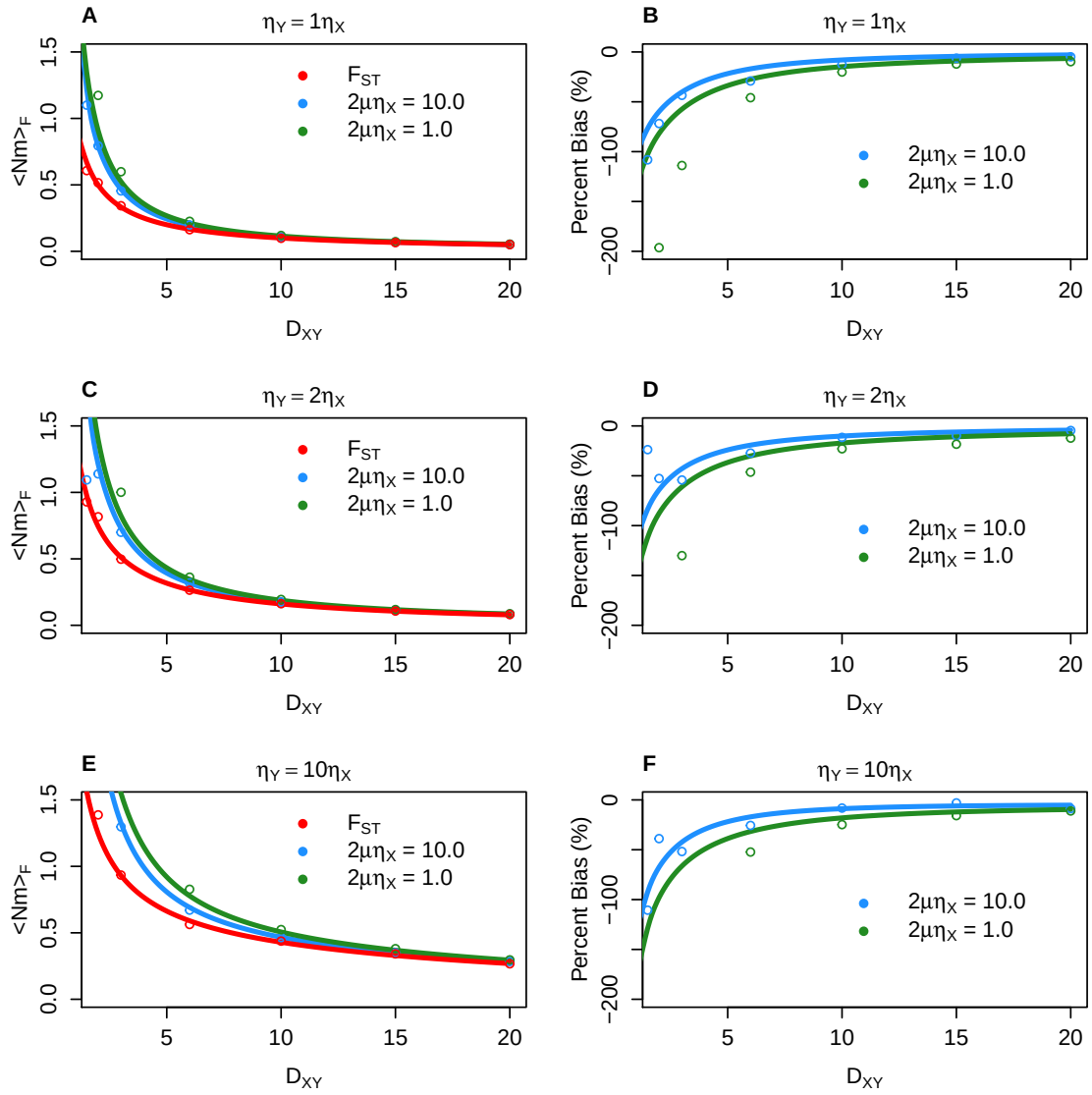


Figure 3.7: $< Nm >_F$ **approximation bias across divergence times.** (A, C, E) On the y axis are values $< Nm >_F$ as functions of divergence time D_{XY}. We plot the value of using the true F_{ST} in red, and approximations $\mathbb{E}(< Nm >_F |\theta)$, for mutation rates $\theta = 10.0$ and 1.0, in blue and green, respectively. (B, D, F) The percent difference between $< Nm >_F$ using F_{ST} (red line in A) and the expected sample quantity to represent the bias in estimation. We simulated assuming equal sample sizes $n_X = n_Y = 10$, and population size structure as indicated at the top of each plot. For a fixed sample size, the expected sample quantity tends to overestimate the 'true' value, with the amount of overestimation a function of μ and D_{XY}.

Accuracy of log transform for linearizing F_{ST}

Under a neutral divergence model, F_{ST} has also commonly been transformed as a linear approximation to the population divergence time, D_{XY}. Discussed in [3], and later in [37], is that given an estimate of F_{ST}, D_{XY} can be estimated by the transformation:

$$\hat{D}_{XY} \propto -log(1 - F_{ST}^{G}) \tag{3.9}$$

Another commonly used transformation, presented in [46] relates the time of divergence to a ratio of F_{ST} values:

$$\hat{D}_{XY} \propto \frac{F_{ST}^{G}}{1 - F_{ST}^{G}} \tag{3.10}$$

Here we evaluate the accuracy of these transformations by approximating the expected value of each using similar Taylor expansions as earlier. Without having an accurate approximation of $\mathrm{Var}(F_{ST}^{G})$, we can only make a first order approximation of equation 3.9 such that:

$$\mathbb{E}(-log(1 - F_{ST}^{G})) \approx -log(1 - \mathbb{E}(F_{ST}^{G})) \tag{3.11}$$

For equation 3.10, by plugging in the estimator for F_{ST} from equation 3.4, we find

$$\frac{F_{ST}^{G}}{1 - F_{ST}^{G}} = 2\frac{d_{XY}}{d_X + d_Y} - 1$$

Taking the expectation of this quantity,

$$\mathbb{E}\Big(\frac{F_{ST}^{G}}{1 - F_{ST}^{G}}\Big) = 2\mathbb{E}\Big(\frac{d_{XY}}{d_X + d_Y}\Big) - 1$$

By deriving a similar second-order Taylor approximation for the expectation on the RHS, as we did earlier with $\mathbb{E}\Big(\frac{d_X + d_Y}{d_{XY}}\Big)$, we get:

$$\begin{aligned}\mathbb{E}\Big(\frac{d_{XY}}{d_X + d_Y}\Big) \approx \frac{\mathbb{E}(d_{XY})}{\mathbb{E}(d_X) + \mathbb{E}(d_Y)} - \frac{\mathrm{Cov}(d_X, d_{XY}) + \mathrm{Cov}(d_Y, d_{XY})}{\big(\mathbb{E}(d_X) + \mathbb{E}(d_Y)\big)^2} \\ + \frac{\mathrm{Var}(d_X) + \mathrm{Var}(d_Y) + 2\mathrm{Cov}(d_X, d_Y)}{\big(\mathbb{E}(d_X) + \mathbb{E}(d_Y)\big)^3}\mathbb{E}(d_{XY})\end{aligned} \tag{3.12}$$

and we have a second-order Taylor approximation of the expectation of equation 3.10.

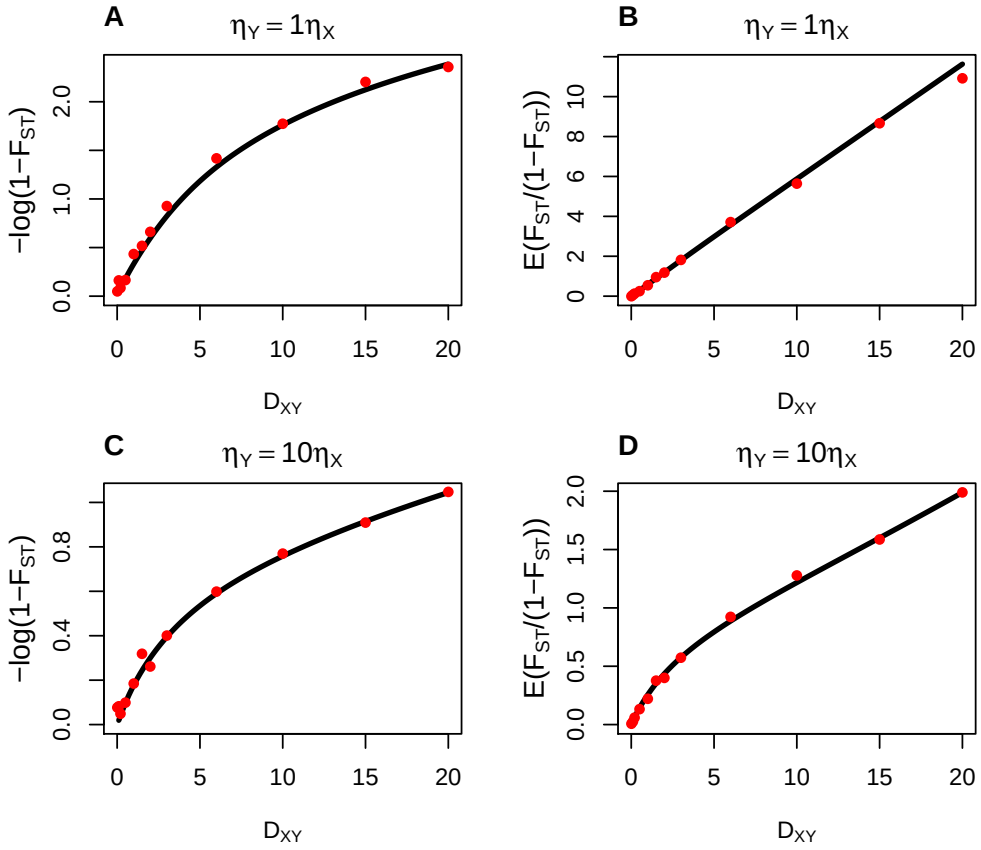


Figure 3.8: **Linearized F_{ST} estimates.** Testing the linearity of two F_{ST} transformations plotted against species divergence time. On the left is the approximate mean log transformed value. On the right is the approximated mean fraction transformed value. Both using F_{ST}^G as a proxy for the unknown F_{ST}. Plotted on the x-axis of both is the simulated divergence time. The red circles correspond to empirical values of $\mathbb{E}(-log(1-F_{ST}))$ and $\mathbb{E}(F_{ST}/(1-F_{ST}))$ to verify the accuracy of the approximation (line in black). (A,B) correspond to the approximations under a constant population size model. (C,D) correspond to the $\eta_Y = 10\eta_X$ imbalanced population size model.

In figure 3.8 we evaluate the linearity between these expressions and divergence time (x versus y axis), and the accuracy of our approximations against simulated data (dots versus line), under two different population size models. It is clear that Slatkin's [46] linear F_{ST} is a linear predictor of divergence time under the constant population size model assumed in its derivation. However, under a model where the population size of species Y is 10 times higher than X, the linearity disappears. The log transformation of [3] and [37] performs worse and can only be used as a

local-linear approximation. Across large values of D_{XY}, it demonstrates clear non-linear behavior and Slatkin's [46] transformation is preferable under all conditions investigated here.

3.9 Discussion/Conclusion

In this paper we have discussed the equations needed to calculate exact values for the covariance between pairs of coalescence times in a species tree model, under piecewise constant population sizes. Using these expressions, we are able to get exact values for the mean, variance, and covariance of average pairwise differences for a given mutation rate and sample size. We have demonstrated that in the constant population size scenario, we can exactly recreate the results of Takahata and Nei [49]. Using our results, we have further explored properties of $\mathrm{F_{ST}}$ and its approximation $\mathrm{F_{ST}^{G}}$ under a divergence model. In particular we demonstrate the downward bias in $\mathrm{F_{ST}}$ estimation using sequence data, and show it is non-vanishing for low mutation rates. As well, the results of the transformation used for gene flow estimation can be biased upwards when using an empirical estimate of $\mathrm{F_{ST}}$. Finally we study the accuracy of a couple of commonly-used linear transformations of $\mathrm{F_{ST}}$ as approximate measures of population divergence times, and find for equal population sizes, the estimator proposed in [46] has the best performance, but when population sizes are no longer equal, even this transformation shows deviations from linearity.

There are many interesting properties to study with the covariance in pairwise coalescent times. In this manuscript we presented one such application, the distribution of average pairwise differences. We hope that the software provided, STCov, will allow for greater investigation into the properties and usefulness of these quantities for estimating various species tree properties, such as topology reconstruction, divergence time and population size estimation, gene flow and admixture detection.

3.10 Software Availability

Along with this manuscript, we provide code (implemented in C++) available for download which calculates the various coalescent quantities presented here (means, variances, covariances, and shared branch length). We have designed the code to be very flexible to user inputted species trees. The program outputs exact quantities for any user-defined rooted, bifurcating, piecewise-constant population size species tree. The code is implemented in C++ and freely available at https://github.com/gaguerra/STCov.

Chapter 4

Statistically Consistent Species Topology Inference using Coalescent Covariance and Minimum Spanning Trees

This is joint work with Rasmus Nielsen.

Methods for estimating species tree topologies have been on the rise with the increase in the availability of large scale sequence data for many species. These methods mainly focus on variation between local gene trees due to the process of incomplete lineage sorting (ILS), with some also modelling the variance in constructing each gene tree estimate from a finite amount of data, although the latter can be computationally intensive. Here we define a new distance metric which can be used to infer species tree topologies based on the estimated covariance between coalescence events within local gene trees. We first demonstrate that this is in fact a metric that uniquely defines a tree shape, and provide (partial) proofs that, in the limit of infinite gene trees, the graph-theoretic minimum spanning tree (MST) algorithm recovers this shape (statistical consistency). This approach is an extremely fast, summary statistic based method that averages information across input sequences to return an estimated species topology. We compare against the quartet-based method ASTRAL on sets of simulated data under various sampling efforts.

4.1 Introduction

It has become clear in recent years that trees relating a set of species at a local region of the genome can differ from one another, and from the evolutionary history of the species as a whole [30, 8, 36]. A highly common source of gene tree discrepancy due to coalescent variation, incomplete lineage sorting (ILS), is an inherent result of the stochastic coalescent process. The level of ILS is a function of the demographic parameters surrounding the evolutionary history. Short branch lengths and large ancestral population sizes on a species tree are where ILS is most present across local trees. In the presence of high ILS, traditional methods that attempt to concatenate multiple sequences and estimate a single gene tree, or estimate a most common gene tree, have been shown to be potentially statistically inconsistent in the presence of ILS [24, 9] and the anomaly zone [6].

The multispecies coalescent (MSC) framework has become the common way to model deviations of gene trees from the underlying species tree. From a set of estimated gene trees, a class of computationally efficient summary statistic methods have been developed to estimate species trees using large amounts of genes, and large amounts of species, for example [23, 58, 29]. Here we present a new summary statistic method based on the previously unstudied covariance in pairwise coalescence times, and show that a statistically consistent topology estimator exists using this information.

In the previous chapter, we introduce theory and a package to calculate this covariance in coalescence times. Here we expand the theory to demonstrate that a distance metric based on this covariance, specifically correlation, can be used to construct a fully connected graph between all pairs of individuals, and a minimum spanning tree through this graph is guaranteed to reconstruct the species tree topology accurately, under the assumption that the covariance terms are estimated without error. As the covariance can never be known without variance in estimation, we present empirical simulations to demonstrate the effect of estimation error from too little genes, and its diminishing effect as the number of genes increases.

4.2 Tree Estimation Method

Constructing the fully connected graph $\mathcal{G}_{\mathcal{S}}$

Let $\mathcal{S}$ represent a bifurcating, rooted, species tree of n species, labeled $(1, ..., n)$. Viewing the topology of $\mathcal{S}$ as a graph, let each leaf node and branching point in the tree represent a node on the graph. This is a graph consisting of $2n-1$ nodes (labeled

$N_1, ..., N_{2n-1}$), with $2n-2$ edges connecting the nodes in a tree pattern. See figure 4.1A for an example visualization of $n = 4$ species. Take node N_6 in the figure as an example, this point on the tree represents the speciation event of species 3 from the species ancestral to (1,2). We associate a random variable with node N_6 which has distribution equal to the distribution of time to coalescence between two lineages, where one is a modern-day individual sampled from species 3, and one is from species 1 or 2. For a leaf node, e.g. node N_2, we associate the random variable of the time to coalescence for any two individuals sampled from species 2. A fully-connected graph of these nodes can be denoted by $\mathcal{G}_\mathcal{S}$ and can be seen in figure 4.1B.

We can define a distance between nodes on $\mathcal{G}_\mathcal{S}$ as a function of their associated random variables, as such:

Definition 4.2.1. *For nodes N_X, N_Y on species graph $\mathcal{G}_\mathcal{S}$, define the distance between the nodes (edge weight) to be*

$$d(N_X, N_Y) = 1 - \frac{Cov(T_{x1,x2}, T_{y1,y2})}{\sqrt{Var(T_{x1,x2})\, Var(T_{y1,y2})}}$$

where x_1 and x_2 are individuals sampled from each side of the divergence event N_X on $\mathcal{S}$, and similarly for y_1, y_2. We assume that all 4 individuals are unique.

The distance metric defined here is non-additive, the distance between two nodes is not equal to the sum of weights on any non-direct path between the two.

To construct $\mathcal{G}_\mathcal{S}$, the topology of $\mathcal{S}$ must already be known. This is rarely known exactly, and must be estimated. To be able to do the estimation, we must generalize our fully connected graph $\mathcal{G}_\mathcal{S}$ to something that can be constructed without knowledge of $\mathcal{S}$.

Constructing the fully connected graph $\mathcal{G}$

In reality, we do not have the information necessary to determine which species diverged at which node on the species tree. As such, the distance in definition 4.2.1 cannot be calculated without first knowing the tree, $\mathcal{S}$. However, what can be calculated is the distance between pairs of species.

Let $\mathcal{G}$ also be a fully connected graph, which requires no knowledge of $\mathcal{S}$ to construct. For a set of n species, let $\mathcal{G}$ have $M = \binom{n}{2} + n$ nodes. Each node represents one of the M pairs of taxa, allowing pairs of the same species. Here $\binom{n}{2}$ represents the number of nodes where the pair of taxa are unique, i.e. species i and species j. The next n nodes represent pairs of taxa where both individuals are from the same species, i.e. i, i. We refer to the nodes on $\mathcal{G}$ as "pseudo-nodes" in this paper as we reserve the proper term "node" to be a vertex on the species graph, $\mathcal{G}_\mathcal{S}$.

Definition 4.2.2. *For any two species, with species labels i, j, denote the pseudo-node $Z_{i,j} \in \mathcal{G}$, to be the speciation event of species i and j. $Z_{i,j} \in \mathcal{G}$.*

Definition 4.2.3. *For pseudo-nodes $Z_{i,j}$, $Z_{l,k}$ on graph $\mathcal{G}$, define their edge weight to be:*

$$d(Z_{i,j}, Z_{k,l}) = 1 - \frac{Cov(T_{i,j}, T_{k,l})}{\sqrt{Var(T_{i,j})\, Var(T_{k,l})}}$$

Where $T_{i,j}$ is the time to coalescence for a lineage sampled from species i, and one from j. Again, note that this distance metric in non-additive.

Figure 4.1C gives a visualization of $\mathcal{G}$ from a set of $n = 4$ species, whose unknown species topology is shown in panel A.

To clarify the use of the term 'pseudo-node', we use the term pseudo-node and node as distinct terms to refer to the nodes of $\mathcal{G}$ and $\mathcal{G}_{\mathcal{S}}$, respectively. All nodes, N_X, are unique, but all pseudo-nodes are not necessarily unique. Each node (on $\mathcal{G}_{\mathcal{S}}$) represents a unique speciation event in the evolutionary history of the set of species. Each pseudo-node, however, only represents the speciation event for a single pair of taxa, and this can be redundant for speciation events deep in a tree. Observe in figure 4.1C that the pseudo-nodes grouped by a grey dotted circle are all represented by the same, single, speciation event. This leads us to the following definition:

Definition 4.2.4. *A pseudo-node $Z_{i,j}$ maps to a node N_X if the speciation event of i and j occurs at node N_X on species tree $\mathcal{S}$.*

This will be useful later on when we begin proofs of statistical consistency.

Estimating a tree topology

From a set of K gene trees of n species, we can estimate the edge weights of the graph, $\mathcal{G}$ from sample variances and covariances. The goal is to estimate the species tree topology, $\mathcal{S}$. We do this using the minimum spanning tree algorithm over $\mathcal{G}$.

Minimum spanning tree

From the fully connected graph, $\mathcal{G}$, define MST($\mathcal{G}$) to be the minimum spanning tree (MST). The MST is a graph-theoretic procedure which determines the dominant tree-like pattern of the entire set of edge-weighted nodes by outlining the shortest path of nearest-neighbor connections. In general, a spanning tree is an acyclic subgraph which passes through all nodes contained in $\mathcal{G}$. The MST is the spanning tree whose total edge weight is minimized. Using Kruskal's algorithm [22], the MST can be found in $M\log(M)$ time.

Tree topology from the MST

Given the constructed MST($\mathcal{G}$) from the set of distances, the estimated unrooted species tree topology $\hat{S}$ can be discerned by pruning the MST. We use two steps sequentially to prune the tree:

- Step 1: Prune MST($\mathcal{G}$)
  - For any node $Z_{i,j}$ of degree 1, in MST($\mathcal{G}$), if $i \neq j$, remove node and its edge from MST($\mathcal{G}$).
  - Recalculate node degrees
  - Repeat until all $Z_{i,j}, i \neq j$ are of degree 2 or higher.
- Step 2: Contract MST($\mathcal{G}$)
  - For any node $Z_{i,j}$ of degree 2, if $i \neq j$, contract node.
  - Recalculate node degrees.
  - Repeat until all $Z_{i,j}, i \neq j$ are of degree 3 or higher.

Here, *contract* consists of two steps. First, remove the node and its edges. Second, add an edge directly between the two nodes formerly connected by the removed node. Note that Step 2 is performed on the pruned version of MST($\mathcal{G}$) from Step 1. Due to Step 2, any root node (for example node N_7 in figure 4.1) will be removed. This means the algorithm will return an unrooted species tree, and it is the user's responsibility to provide a rooting location/outgroup.

In what immediately follows, we will show that (conditional on a conjecture) the MST algorithm is guaranteed to return the true topology when the covariance/variance between coalescence times are known without error.

4.3 Statistical Consistency

In this section, we prove (contingent on conjecture 1) that when the covariance between pairs of coalescence times is known without error, that the estimated tree via the MST algorithm is guaranteed to recover the true species tree topology.

Species tree $\mathcal{S}$, known

First, we start with the proof of consistency when the tree is known. This is a sanity check that the distance metric and minimum spanning tree approach is guaranteed

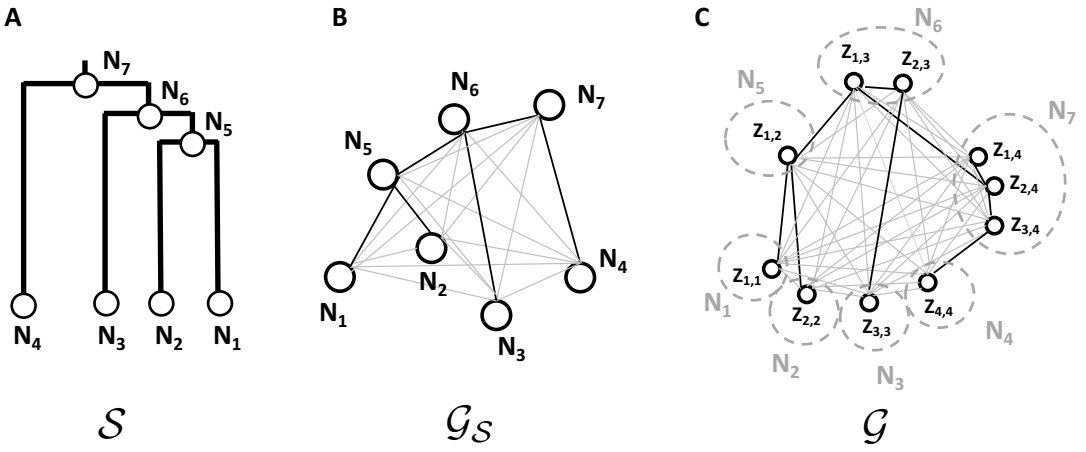


Figure 4.1: **Graph Theoretic View of Species Trees:** (A) A standard bifurcating, rooted species tree, $\mathcal{S}$ of 4 species. (B) A fully connected graph, $\mathcal{G}_\mathcal{S}$ between the leaf and internal nodes of $\mathcal{S}$. (C) Assuming the species tree $\mathcal{S}$ is unknown, and therefore the nodes are unknown, a fully connected graph, $\mathcal{G}$ between what we call "pseudo-nodes", indicating the speciation event between a pair of species. A graph $\mathcal{G}$ can be contracted/pruned into $\mathcal{G}_\mathcal{S}$ once the species tree is known. In both (B) and (C) an example spanning tree which matches the topology of $\mathcal{S}$ is highlighted in black.

to return the correct tree. Given the species tree, we can reduce the fully connected graph, $\mathcal{G}$ of M nodes to the much smaller complete sub-graph $\mathcal{G}_\mathcal{S}$ of $2n-1$ nodes. The set of nodes, $\mathcal{N}$, in this subgraph can be chosen randomly from the pseudo-nodes that map to each node. By definition, $N_1, ..., N_n$ are the pseudo-nodes $Z_{1,1}, ..., Z_{n,n}$. For an internal node N_x, define X_1 and X_2 to be the sets of species on each side of the node in $\mathcal{S}$. We can arbitrarily choose any pseudo-node Z_{x_1,x_2} ($x_1 \in X_1, x_2 \in X_2$) to represent N_X.

Definition 4.3.1. *For two nodes, N_I and N_J, and species tree $\mathcal{S}$, define $\mathcal{P}(N_I, N_J)$ to be the set of nodes in the path between the two nodes, obeying the edges of $\mathcal{S}$.*

Lemma 1. *For any two nodes (leaf or internal), denoted by N_I, N_J, and topology $\mathcal{S}$:*

$$d(N_I, N_J) \geq d(N_{p1}, N_{p2})$$

for all N_{p1} and $N_{p2} \in \mathcal{P}(N_I, N_J)$.

Proof. Let N_I and N_J be nodes on species tree $\mathcal{S}$. As well, let N_X be a node in the path from N_I to N_J along $\mathcal{S}$. We will first show that $d(N_I, N_J) \geq d(N_X, N_J)$ and iteratively apply this rule to show that for any two nodes N_X, N_Y in the path that $d(N_I, N_J) \geq d(N_X, N_Y)$.

Decompose the distance matrix to note that:

$$d(N_I, N_J) \geq d(N_X, N_J)$$

is equivalent to

$$\frac{\text{Cov}(T_{i1,i2}, T_{j1,j2})}{\sigma_i \sigma_j} \leq \frac{\text{Cov}(T_{x1,x2}, T_{j1,j2})}{\sigma_x \sigma_j}$$

where notation σ_a is the standard deviation in time to coalescence for a pair of individuals originating at node N_a.

Without loss of generality (WLOG), assume N_X is more ancient than node N_I. (This must be true of either N_I or N_J) Observe that we can write the distribution of coalescence times for pairs of individuals whose species diverge at node N_X as a conditional function of the distribution of times at node N_I. Let τ_X and τ_I represent the timing of the nodes on the species tree $\mathcal{S}$, (divergence times). Then we have:

$$T_{x1,x2} \stackrel{d}{=} T_{i1,i2} | T_{i1,i2} \geq \tau_X$$

where "$\stackrel{d}{=}$" indicates equality in distribution. We can use the law of total variance to get an expression of σ_i^2 as a function of σ_X^2. Define random indicator variable

$$\Phi = \begin{cases} 1, & \text{w.p. } P(T_{i1,i2} > \tau_X | \mathcal{S}) \\ 0, & \text{w.p. } 1 - P(T_{i1,i2} > \tau_X | \mathcal{S}) \end{cases}$$

Note that $T_{j1,j2}$ is independent of Φ, as the pair of individuals (j_1, j_2) are distinct from (i_1, i_2), and marginally, $T_{j1,j2}$ does not depend on these individuals. As well, observe $T_{i1,i2}$ is independent of $T_{j1,j2}$ conditional on the event $(T_{i1,i2} < \tau_X)$, as the event implies i_1, i_2 have coalesced before the lineages reach τ_j (as $\tau_X < \tau_j$).

The law of total variance of $T_{i1,i2}$ gives us:

$$\begin{aligned}\sigma_i^2 &= \mathbb{E}\big(\mathrm{Var}(T_{i1,i2}|\Phi)\big) + \mathrm{Var}\big(\mathbb{E}(T_{i1,i2}|\Phi)\big) \\ &= P(T_{i1,i2} > \tau_X|\mathcal{S})\mathrm{Var}\big(T_{i1,i2}|T_{i1,i2} \geq \tau_X\big) + \Big(1 - P(T_{i1,i2} > \tau_X|\mathcal{S})\Big) \\ &\times \mathrm{Var}\big(T_{i1,i2}|T_{i1,i2} < \tau_X\big) + P(T_{i1,i2} \geq \tau_X|\mathcal{S})\Big(\mathbb{E}(T_{i1,i2}|T_{i1,i2} \geq \tau_X) - \mathbb{E}(T_{i1,i2})\Big)^2 \\ &+ (1 - P(T_{i1,i2} \geq \tau_X|\mathcal{S}))\Big(\mathbb{E}(T_{i1,i2}|T_{i1,i2} < \tau_X) - \mathbb{E}(T_{i1,i2})\Big)^2\end{aligned}$$

Where observe $\mathrm{Var}(T_{i1,i2}|T_{i1,i2} \geq \tau_X) = \mathrm{Var}(T_{x1,x2}) = \sigma_X^2$. We can then write:

$$\sigma_i^2 = P(T_{i1,i2} \geq \tau_X|\mathcal{S})\sigma_X^2 + \zeta$$

where ζ is a sum of non-negative terms, therefore is greater than or equal to 0.

Similarly, we can use the law of total covariance to get an expression for the term $\mathrm{Cov}(T_{i1,i2}, T_{j1,j2})$:

$$\begin{aligned}\mathrm{Cov}(T_{i1,i2}, T_{j1,j2}) &= \mathbb{E}\big(\mathrm{Cov}(T_{i1,i2}, T_{j1,j2}|\Phi)\big) + \mathrm{Cov}\big(\mathbb{E}(T_{i1,i2}|\Phi), \mathbb{E}(T_{j1,j2}|\Phi)\big) \\ &= P(T_{i1,i2} \geq \tau_X, \mathcal{S})\mathrm{Cov}(T_{i1,i2}, T_{j1,j2}|T_{i1,i2} \geq \tau_X) \\ &+ \big(1 - P(T_{i1,i2} \geq \tau_X|\mathcal{S})\big)\mathrm{Cov}(T_{i1,i2}, T_{j1,j2}|T_{i1,i2} < \tau_X) \\ &+ P(T_{i1,i2} \geq \tau_X|\mathcal{S})\big(\mathbb{E}(T_{i1,i2}|T_{i1,i2} \geq \tau_X) - \mathbb{E}(T_{i1,i2})\big) \\ &\times \big[\mathbb{E}(T_{j1,j2}|T_{i1,i2} \geq \tau_Z) - \mathbb{E}(T_{j1,j2})\big] \\ &+ \Big(1 - P(T_{i1,i2} \geq \tau_X|\mathcal{S})\Big)\Big(\mathbb{E}(T_{i1,i2}|T_{i1,i2} < \tau_X) - \mathbb{E}(T_{i1,i2})\Big) \\ &\times \Big(\mathbb{E}(T_{j1,j2}|T_{i1,i2} \geq \tau_X) - \mathbb{E}(T_{j1,j2})\Big) \\ &= P(T_{i1,i2} \geq \tau_X|\mathcal{S})\mathrm{Cov}(T_{x1,x2}, T_{j1,j2})\end{aligned}$$

As we have mentioned, marginally, $T_{j1,j2} \perp\!\!\!\perp \Phi$ and so the last two terms equal 0, as well, $T_{j1,j2} \perp\!\!\!\perp T_{i1,i2}|\big(T_{i1,i2} < \tau_Z\big)$ making the second covariance term 0. We then have the result:

$$\mathrm{Cov}(T_{i1,i2}, T_{j1,j2}) = P(T_{i1,i2} \geq \tau_X|\mathcal{S})\mathrm{Cov}(T_{x1,x2}, T_{j1,j2})$$

With these two results, we observe:

$$\begin{aligned}\frac{\text{Cov}(T_{i1,i2}, T_{j1,j2})}{\sigma_I \sigma_J} &= \frac{P(T_{i1,i2} \geq \tau_Z | \mathcal{S}) \text{Cov}(T_{x1,x2}, T_{j1,j2})}{\sigma_j \sqrt{P(T_{i1,i2} \geq \tau_X | \mathcal{S}) \sigma_X^2 + \zeta}} \\ &\leq \frac{P(T_{i1,i2} \geq \tau_Z | \mathcal{S}) \text{Cov}(T_{x1,x2}, T_{j1,j2})}{\sigma_j \sqrt{P(T_{i1,i2} \geq \tau_X | \mathcal{S}) \sigma_X^2}} \\ &= \frac{\sqrt{P(T_{i1,i2} \geq \tau_X | \mathcal{S})} \text{Cov}(T_{x1,x2}, T_{j1,j2})}{\sigma_J \sigma_X} \\ &\leq \frac{\text{Cov}(T_{x1,x2}, T_{j1,j2})}{\sigma_J \sigma_X}\end{aligned}$$

where the first inequality comes from $\zeta \geq 0$, and the second comes from $0 \leq \sqrt{P(T_{i1,i2} \geq \tau_X | \mathcal{S})} \leq 1$.

So we have now shown for any three nodes, in order (N_I, N_X, N_J), that:

$$d(N_I, N_J) \geq d(N_X, N_J)$$

To show the general case, let (N_I, N_X, N_Y, N_J) be an ordering of nodes on tree $\mathcal{S}$, meaning $N_X, N_Y \in \mathcal{P}(N_I, N_J)$. Again, without loss of generality assume node N_X is more ancient on the tree than N_I, and N_Y is more ancient on the tree than N_J. We will show that $d(N_I, N_J) \geq d(N_X, N_Y)$. To do so, first consider the triplet of nodes (N_I, N_X, N_J), from above we know

$$d(N_I, N_J) \geq d(N_X, N_J)$$

Next, consider the second triplet of nodes (N_X, N_Y, N_J), as we have already removed N_J from the expression. Applying the result above again, we see:

$$d(N_X, N_J) \geq d(N_X, N_Y)$$

which gives us our result:

$$d(N_I, N_J) \geq d(N_X, N_Y)$$

for any nodes in the path from N_I to N_J on species tree $\mathcal{S}$. □

We have defined the metric $d(\cdot, \cdot)$, with the property that the direct distance between any two nodes is greater than or equal to any individual distance on the path connecting them via species tree $\mathcal{S}$. Denote the MST of graph $\mathcal{G}_\mathcal{S}$ to be $\text{MST}(\mathcal{G}_\mathcal{S})$. Assuming each edge weight is unique, the MST is unique.

Theorem 1. *For the fully connected graph, $\mathcal{G}_{\mathcal{S}}$, induced by metric $d(\cdot,\cdot)$ on tree $\mathcal{S}$, known, the minimum spanning tree, MST($\mathcal{G}_{\mathcal{S}}$), is the same unrooted topology of leaf nodes as the species tree $\mathcal{S}$, assuming the distances $d(\cdot,\cdot)$ are known without error and no two distances are exactly the same.*

Proof. We will show this result by showing that all edges that are not in species tree $\mathcal{S}$ cannot be in any minimum spanning tree of $\mathcal{G}_{\mathcal{S}}$, and therefore only edges matching $\mathcal{S}$ remain. To do this, we exploit the cycle property of MSTs.

The cycle property states that for any cycle $\mathbb{C}$ in the graph: The edge f in cycle $\mathbb{C}$ whose edge weight is larger than every other edge in the cycle, cannot be an edge in a MST. This is easily proven using a contradiction argument.

Let N_I and N_J be any two non-adjacent nodes in species tree $\mathcal{S}$ (meaning there exists some other node N_k such that to traverse from N_I to N_J via $\mathcal{S}$ you must pass through N_k). Define cycle $\mathbb{C}_{i,j}$ to be the set of edges which connect N_I and N_J on tree $\mathcal{S}$, as well as the direct edge connecting N_I and N_J. Combining the result of Lemma 1, and the assumption that no two distances are exactly the same value, it is true that $d(N_I, N_J)$ is the maximum edge length in $\mathcal{C}_{i,j}$. Therefore, by the cycle property, the edge directly connecting N_I and N_J cannot be in any MST.

By iteratively applying this argument to all pairs of nodes which are not directly adjacent on tree $\mathcal{S}$, we see that no edge directly connecting the pair can be in a MST. Therefore, the only path which remains to connect all nodes is the same path as $\mathcal{S}$. □

Species tree, $\mathcal{S}$, unknown

In practice, we do not directly know the internal nodes. Instead, we have distances between all M pseudo-nodes, pairs of lineages. To show statistical consistency still holds in the presence of pseudo-nodes, it suffices to prove that the distance between two pseudo-nodes which map to the same node, N_I, is smaller than the distance from either of the two pseudo-nodes to a pseudo-node which does not map to N_I. If this is true, than the minimum spanning tree result above still holds, as all pseudo-nodes which map to the same node will cluster with one another, forming a node on the tree. This section relies on a (currently) unproven conjecture:

Conjecture 1. *For any two pseudo-nodes $Z_{i,j}, Z_{k,l}$ which both map to a node N_X, and any pseudo-node $Z_{w,y}$ which maps to a node N_A more ancient on $\mathcal{S}$ than N_X:*

$$d(Z_{i,j}, Z_{k,l}) \leq min\Big(d(Z_{i,j}, Z_{w,y}), d(Z_{k,l}, Z_{w,y})\Big)$$

In the proof of the following lemma, we consider a pseudo-node $Z_{w,y}$ which does not map to a node N_X. There are then three possibilities to where this pseudo-node could map to: a node more recent on $\mathcal{S}$, a node more ancient on $\mathcal{S}$, or a sister-node on $\mathcal{S}$ (meaning it is not an ancestral node to N_X, nor N_X to it). Conjecture 1 accounts for the case when the node is more ancient than N_X.

Lemma 2. *For pseudo-nodes $Z_{i,j}$, $Z_{k,l}$ which both map to node N_X, and any node $Z_{w,y}$ which does not map to N_X:*

$$d(Z_{i,j}, Z_{k,l}) \leq min\Big(d(Z_{i,j}, Z_{w,y}), d(Z_{k,l}, Z_{w,y})\Big)$$

Proof. WLOG assume that $d(Z_{i,j}, Z_{w,y}) \leq d(Z_{k,l}, Z_{w,y)}$, then it suffices to show :

$$d(Z_{i,j}, Z_{k,l}) \leq d(Z_{i,j}, Z_{w,y})$$

There are three cases to consider here: 1. $Z_{w,y}$ occurs more recently on $\mathcal{S}$ than N_X, 2. $Z_{w,y}$ occurs more anciently than N_X, and 3. $Z_{w,y}$ maps to a sister node of N_X.

1: Let $Z_{w,y}$ occur more recently than N_X. Then, similar to the proof of Lemma 1, we can write $T_{k,l}$ as a conditional expression of $T_{w,y}$:

$$T_{k,l} \overset{d}{=} T_{w,y} | T_{w,y} \geq \tau_{k,l}$$

Applying the laws of total variance and covariance, we see:

$$\mathrm{Var}(T_{w,y}) = \mathrm{Var}(T_{k,l}) P(T_{w,y} \geq \tau_{k,l} | \mathcal{S}) + \zeta$$

$$\mathrm{Cov}(T_{i,j}, T_{w,y}) = \mathrm{Cov}(T_{i,j}, T_{k,l}) P(T_{w,y} \geq \tau_{k,l} | \mathcal{S})$$

where ζ is a sum of non-negative terms. Applying these results we get the following:

$$\begin{aligned}
\frac{\mathrm{Cov}(T_{i,j}, T_{w,y})}{\sigma_{i,j}\sigma_{w,y}} &= \frac{\mathrm{Cov}(T_{i,j}, T_{k,l}) P(T_{w,y} \geq \tau_{k,l} | \mathcal{S})}{\sigma_{i,j}\sqrt{\mathrm{Var}(T_{k,l}) P(T_{w,y} \geq \tau_{k,l} | \mathcal{S}) + \zeta}} \\
&\leq \frac{\mathrm{Cov}(T_{i,j}, T_{k,l}) \sqrt{P(T_{w,y} \geq \tau_{k,l} | \mathcal{S})}}{\sigma_{i,j}\sigma_{k,l}} \\
&\leq \frac{\mathrm{Cov}(T_{i,j}, T_{k,l})}{\sigma_{i,j}\sigma_{k,l}}
\end{aligned}$$

which implies $d(Z_{i,j}, Z_{k,l}) \leq d(Z_{i,j}, Z_{w,y})$ for case 1.

2: Let $Z_{w,y}$ occur more anciently than N_X. We must currently conjecture this is true (see conjecture 1).

3: Let $Z_{w,y} \to N_R$, a sister node to N_X. This case is straightforward. As N_R and N_X are sister nodes, there exists a common node ancient to both, denote this N_A, $N_A \in \mathcal{P}(N_X, N_R)$. Lemma 1 tells us that $d(N_X, N_R) \geq d(N_X, N_A)$. By applying case 2 to any pseudo-node which maps to N_A, we have our result. □

We have now shown that the distance between pseudo-nodes which map to the same node is smaller than the distance to any other pseudo-node, conditional on conjecture 1 being true. Let $\mathcal{G}$ be the fully connected graph over the M pseudo-nodes with our distance metric. Denote the MST of graph $\mathcal{G}$ to be MST($\mathcal{G}$). Assuming each edge weight is unique, the MST is unique.

Theorem 2. *For the fully connected graph, $\mathcal{G}$, induced by metric $d(\cdot,\cdot)$ on tree $\mathcal{S}$, the minimum spanning tree, MST($\mathcal{G}$), is guaranteed to return the same topology of leaf nodes as the species tree $\mathcal{S}$ when pruned, assuming the distances $d(\cdot,\cdot)$ are known without error and no two distances are exactly the same. Conjecture 1 must be assumed to be true.*

Proof. Here it suffices to show that the path between any two pseudo-nodes $Z_{i,j}$, $Z_{k,l}$ which map to the same node N_X only contains other pseudo-nodes which map to N_X in their path. Given that this is true, we can effectively collapse the set of pseudo-nodes into N_X, and apply Theorem 1.

We will show this by contradiction.

Suppose there exists a pseudo-node $Z_{s,t}$ which does not map to N_X such that $Z_{s,t} \in \mathcal{P}(Z_{i,j}, Z_{k,l})$ on MST($\mathcal{G}$). Let D be the total weight of this path $\mathcal{P}(Z_{i,j}, Z_{k,l})$. Insert the edge that directly connects $Z_{i,j}$ and $Z_{k,l}$, to form a cycle, $\mathbb{C}$, in the minimum spanning tree, $\mathbb{C} = (Z_{i,j}, Z_{k,l}) + \mathcal{P}(Z_{i,j}, Z_{k,l})$. WLOG assume $Z_{i,j}$ and $Z_{s,t}$ are directly connected on MST($\mathcal{G}$). If we apply the result of lemma 2, we know:

$$d(Z_{i,j}, Z_{k,l}) < d(Z_{i,j}, Z_{s,t})$$

By removing the edge $(Z_{i,j}, Z_{s,t})$ from C, the total weight connecting all nodes in $\mathbb{C}$ is now:

$$\Big(D + d(Z_{i,j}, Z_{k,l}) - d(Z_{i,j}, Z_{s,t})\Big) < D$$

and so the spanning tree which replaces edge $(Z_{i,j}, Z_{s,t}) \in$ MST($\mathcal{G}$) with edge $(Z_{i,j}, Z_{k,l})$ results in a spanning tree of less total weight than MST($\mathcal{G}$), therefore MST($\mathcal{G}$) is not the minimum spanning tree, a contradiction.

From this, we know that all pseudo-nodes of a node N_X form a self-contained spanning tree in the MST($\mathcal{G}$). Denote this self contained spanning tree as N_X^T. To form a spanning tree, any two self contained spanning trees, N_X^T, N_Y^T can only be connected by one single edge, otherwise a cycle will form in the graph. Therefore we can treat these self contained spanning trees, N_1^T,..,N_{2n-1}^T as we would the nodes $N_1, ..., N_{2n-1}$. From this we apply theorem 1. By collapsing all self contained spanning trees into single nodes, we exactly recover the species tree topology $\mathcal{S}$. □

We have conditionally proven that if the covariance between all pairs of individuals is known without error, constructing a minimum spanning tree using the correlation-based distance will exactly recover the underlying species tree, $\mathcal{S}$.

4.4 Consistency Simulation

As the covariances, and therefore distances, cannot be estimated without error, we present simulations to show as the amount of information increases, the error in estimation decreases, and the MST approach is asymptotically consistent. We note that there are two sources of variance in our simulation, a finite number of genes, and a finite amount of pairwise differences within a gene. We simulate a species tree of 8 individuals, and test the claim of consistency by evaluating the accuracy under different levels of sequencing efforts. Specifically, we test the accuracy in reconstruction for 100, 500, 1000, 2500, 5000, 10000, and 20000 sampled genes. For each, we produce 100 independent replicates, and assess the average performance. We measure two metrics: First, the percentage of replicates in which the MST returns the true tree, and second, the average Robinson-Foulds (RF) [42] distance between the estimated and true species tree topology. As expected, as the proportion of correct trees fixes to 1.0 with a sufficient number of sequences sampled, as the covariance terms can be estimated with a decreasing amount of sampling variance. See figure 4.2 for the results of this test.

Tree parameters

We simulate an 8 species topology, (H,((G,(F,E)), (D,(C,(B,A))))), with 4 individuals per species, a requirement for our approach. We generate K gene trees using ms [16] and gene sequences of length $\mathcal{L} = 1000$bp conditional on each tree using Seq-Gen [39] with population scaled mutation rate $\theta = 0.001$. The species tree is parameterized as follows, $\tau_{A,B} = 2.5$, $\tau_{E,F} = 2.0$, $\tau_{(A,B),C} = 3.0$, $\tau_{((A,B),C),D} = 3.5$, $\tau_{(E,F),G} = 4.0$, $\tau_{(((A,B),C),D),((E,F),G)} = 5.0$, $\tau_{H,((G,(F,E)),(D,(C,(B,A))))} = 8.0$, with all population sizes $= \eta_0 = 1.0$.

4.5 Performance against ASTRAL

5 species simulation

We test the performance of our method against ASTRAL-III for various number of genes sampled for a tree of 5 species (4 individuals per species), under a fixed

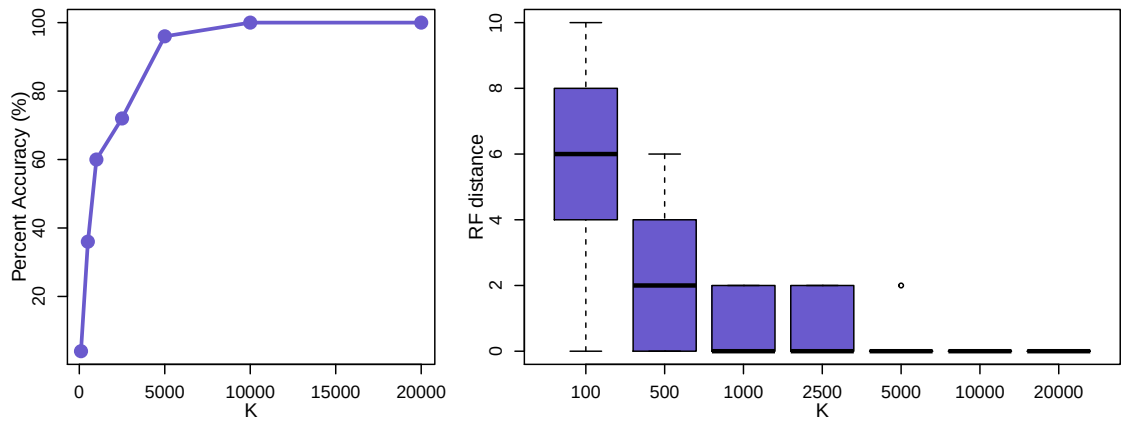


Figure 4.2: **Consistency Plot:** For a fixed species tree, gene length, and mutation rate, we simulate K independent loci and estimate the species tree topology. (A) The y-axis represents the percentage of 100 independent replicates in which the true tree is correctly estimated. (B) The distribution of RF-distances between the estimated and true topologies, with 25 independent replicates per K.

species tree and mutation rate. We simulate a tree topology ((A,B),(C,(D,E))), where in particular the branch length connecting node (D,E) to C is short. We expect there to be ILS between the lineages C,D,E, making tree inference difficult between topological orderings (C,(D,E)), (D,(C,E)), and (E,(C,D)). We assume a constant population size across the phylogeny (scaled to equal 1.0), and the following species divergence times: $\tau_{D,E} = 2.0$, $\tau_{C,D,E} = 2.2$, $\tau_{A,B} = 1.5$, $\tau_{ABCDE} = 5.0$. For each replicate we simulate K gene trees using ms [16], and generate sequences of length $\mathcal{L} = 1000$bp for each tree using Seq-Gen [39], with mutation parameter θ. We simulate a number of independent genes, K in $\{100, 500, 1000, 2500, 5000, 10000\}$, with $\theta = 0.001$. This gives a total of 6 simulation scenarios. For each scenario we simulate 100 independent replicates, and count the number of replicates in which each method produced the correct topology. Figure 4.3 shows the results of this simulation for our MST method against ASTRAL.

We see for low sample size, K, our method does not have sufficient information to accurately estimate the variance covariance structure between individuals, and so cannot discern the correct branching pattern of C,D,E. As the number of genes increases, our method has an increasingly better approximation to the variance/covariance, and can accurately estimate the topology. ASTRAL is much more accurate than our method for low numbers of genes, as the local topology at every gene is informative for ASTRAL, whereas we rely on pooling information across genes. Note, for any pair of individuals (i, j), the expected number of pairwise differences

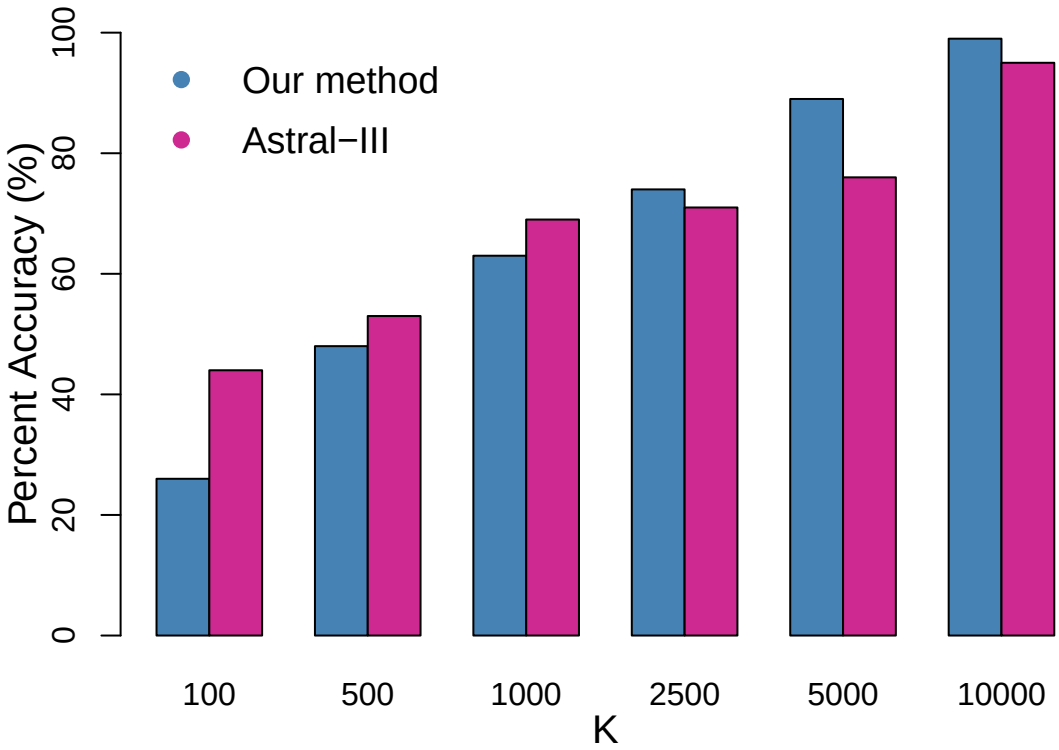


Figure 4.3: **Performance, 5 species:** The number of independent genes is indicated on the x-axis. Accuracy is measured as the percentage of correctly estimated topologies across 100 replicates for each method.

is $\theta\mathcal{L}\mathbb{E}(t_{i,j}|\mathcal{S})$.

4.6 Discussion

We have defined a new metric between nodes on a species tree using the correlation between coalescence times. We have conditionally proven that constructing a minimum spanning tree over a fully connected graph of what we call "pseudonodes", where weights on the graph are given by the distance presented here, is guaranteed to return the underlying unrooted true species tree topology, $\mathcal{S}$. Our proof, as presented here, relies on conjecture 1 that has yet to be shown.

We have shown through 2 simulation scenarios the performance of this approach on the problem of species tree inference. The results of our simulation against ASTRAL are promising for large numbers of genes. Further simulation studies and

larger numbers of replicates are needed to make any claims about our performance, in general.

The proofs of consistency here only rely on proving inequalities of correlations/covariances between pairs of coalescence times. The distance we define in this manuscript is therefore not required for our statistical guarantees. In fact there are likely other distance metrics which make use of this correlation/covariance that have better performance and properties. A problem with our distance metric as it is defined here is that the distance is highly non-additive. A transformation of the correlation/covariance that maintains some additive property will probably have better performance in the presence of estimation error of the covariances. Defining a distance metric using the only the covariances would be most preferred as the error in variance estimation can be ignored. Note that the proofs are all still correct for any distance metric which only uses the relative difference between covariances, and not necessarily normalized by their variances.

4.7 Notation reference

- n: Number of modern-day species, here labeled $\{1, ..., n\}$.
- $\mathcal{S}$: Species tree relating the n species.
- $\mathcal{N}$: The set of $2n-1$ nodes on $\mathcal{S}$ connected by $2n-2$ edges. Labeled $N_1, N_2, ..., N_{2n-1}$. Nodes $1, .., n$ represent leaves, $n+1, .., 2n-1$ internal nodes.
- τ_X: The speciation time of N_X according to $\mathcal{S}$.
- $\mathcal{G}_{\mathcal{S}}$: A fully connected graph between all nodes on $\mathcal{S}$.
- $\tau_{i,j}$: The split time between species i, j on $\mathcal{S}$.
- $Z_{i,j}$: A "pseudo-node" indicating the speciation event of species i and j.
- $T_{i,j}$: The random variable of time to coalescence between an individual sampled from species i and one from j.
- M: The set of $n(n+1)/2$ total pseudo-nodes across n species.
- $\mathcal{G}$: A fully connected graph between all pseudo-nodes.
- $d(N_X, N_Y)$: The edge weight between two nodes on $\mathcal{G}_{\mathcal{S}}$.
- $d(Z_{i,j}, Z_{k,l})$: The edge weight between two pseudo-nodes on $\mathcal{G}$.

- $T_{i,j} \stackrel{d}{=} T_{k,l}$: Indicates that $T_{i,j}$ is equal in distribution to $T_{k,l}$

Bibliography

[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] Luigi L Cavalli-Sforza. "Human diversity". In: *Proc. 12th Int. Congr. Genet.* Vol. 2. 1969, pp. 405–416.

[4] 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.

[5] Charles Darwin. *On the Origin of Species by Means of Natural Selection.* or the Preservation of Favored Races in the Struggle for Life. London: Murray, 1859.

[6] James H Degnan and Noah A Rosenberg. "Discordance of species trees with their most likely gene trees". In: *PLoS genetics* 2.5 (2006), e68.

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

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

[9] Scott V Edwards et al. "Implementing and testing the multispecies coalescent model: a valuable paradigm for phylogenomics". In: *Molecular phylogenetics and evolution* 94 (2016), pp. 447–462.

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

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

[17] Richard R Hudson, Montgomery Slatkin, and Wayne P Maddison. "Estimation of levels of gene flow from DNA sequence data." In: *Genetics* 132.2 (1992), pp. 583–589.

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

[19] Motoo Kimura. "Theoretical foundation of population genetics at the molecular level". In: *Theoretical population biology* 2.2 (1971), pp. 174–208.

[20] John Frank Charles Kingman. "The coalescent". In: *Stochastic processes and their applications* 13.3 (1982), pp. 235–248.

[21] 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.

[22] Joseph B Kruskal. "On the shortest spanning subtree of a graph and the traveling salesman problem". In: *Proceedings of the American Mathematical society* 7.1 (1956), pp. 48–50.

[23] 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.

[24] Laura Salter Kubatko and James H Degnan. "Inconsistency of phylogenetic estimates from concatenated data under coalescence". In: *Systematic Biology* 56.1 (2007), pp. 17–24.

[25] 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.

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

[27] 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.

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

[29] 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.

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

[31] 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.

[32] 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.

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

[34] Masatoshi Nei and Li Jin. "Variances of the average numbers of nucleotide substitutions within and between populations." In: *Molecular Biology and Evolution* 6.3 (1989), pp. 290–300.

[35] Masatoshi Nei and Wen-Hsiung Li. "Mathematical model for studying genetic variation in terms of restriction endonucleases". In: *Proceedings of the National Academy of Sciences* 76.10 (1979), pp. 5269–5273.

[36] Richard Nichols. "Gene trees and species trees are not the same". In: *Trends in Ecology & Evolution* 16.7 (2001), pp. 358–364.

[37] Rasmus Nielsen et al. "Maximum-likelihood estimation of population divergence times and population phylogeny in models without mutation". In: *Evolution* 52.3 (1998), pp. 669–677.

[38] 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.

[39] 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.

[40] 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.

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

[42] David F Robinson and Leslie R Foulds. "Comparison of phylogenetic trees". In: *Mathematical biosciences* 53.1-2 (1981), pp. 131–147.

[43] 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.

[44] 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.

[45] Montgomery Slatkin. "Inbreeding coefficients and coalescence times". In: *Genetics Research* 58.2 (1991), pp. 167–175.

[46] Montgomery Slatkin. "Isolation by distance in equilibrium and non-equilibrium populations". In: *Evolution* 47.1 (1993), pp. 264–279.

[47] 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.

[48] Fumio Tajima. "Evolutionary relationship of DNA sequences in finite populations". In: *Genetics* 105.2 (1983), pp. 437–460.

[49] Naoyuki Takahata and Masatoshi Nei. "Gene genealogy and variance of interpopulational nucleotide differences". In: *Genetics* 110.2 (1985), pp. 325–344.

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

[51] 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.

[52] John Wakeley. "Pairwise differences under a general model of population subdivision". In: *Journal of Genetics* 75.1 (1996), pp. 81–89.

[53] John Wakeley. "The variance of pairwise nucleotide differences in two populations with migration". In: *Theoretical population biology* 49.1 (1996), pp. 39–57.

[54] John Wakeley. "Using the variance of pairwise differences to estimate the recombination rate". In: *Genetics Research* 69.1 (1997), pp. 45–48.

[55] GA Watterson. "On the number of segregating sites in genetical models without recombination". In: *Theoretical population biology* 7.2 (1975), pp. 256–276.

[56] Sewall Wright. "The genetical structure of populations". In: *Annals of eugenics* 15.1 (1949), pp. 323–354.

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

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

Appendix A

Supplement for Chapter 2

A.1 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.

A.2 Further Simulation Details

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)))).

For a single replicate, we use ms to generate K independent gene trees of 10 individuals, 2 from each species, and Seq-Gen [39] 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 [16], 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 into COAL_PHYRE.

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 into COAL_PHYRE.

A.3 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 A.1 legend), to generate an empirical distribution of estimated coalescence times from the number of pairwise differences. Figure A.1 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.

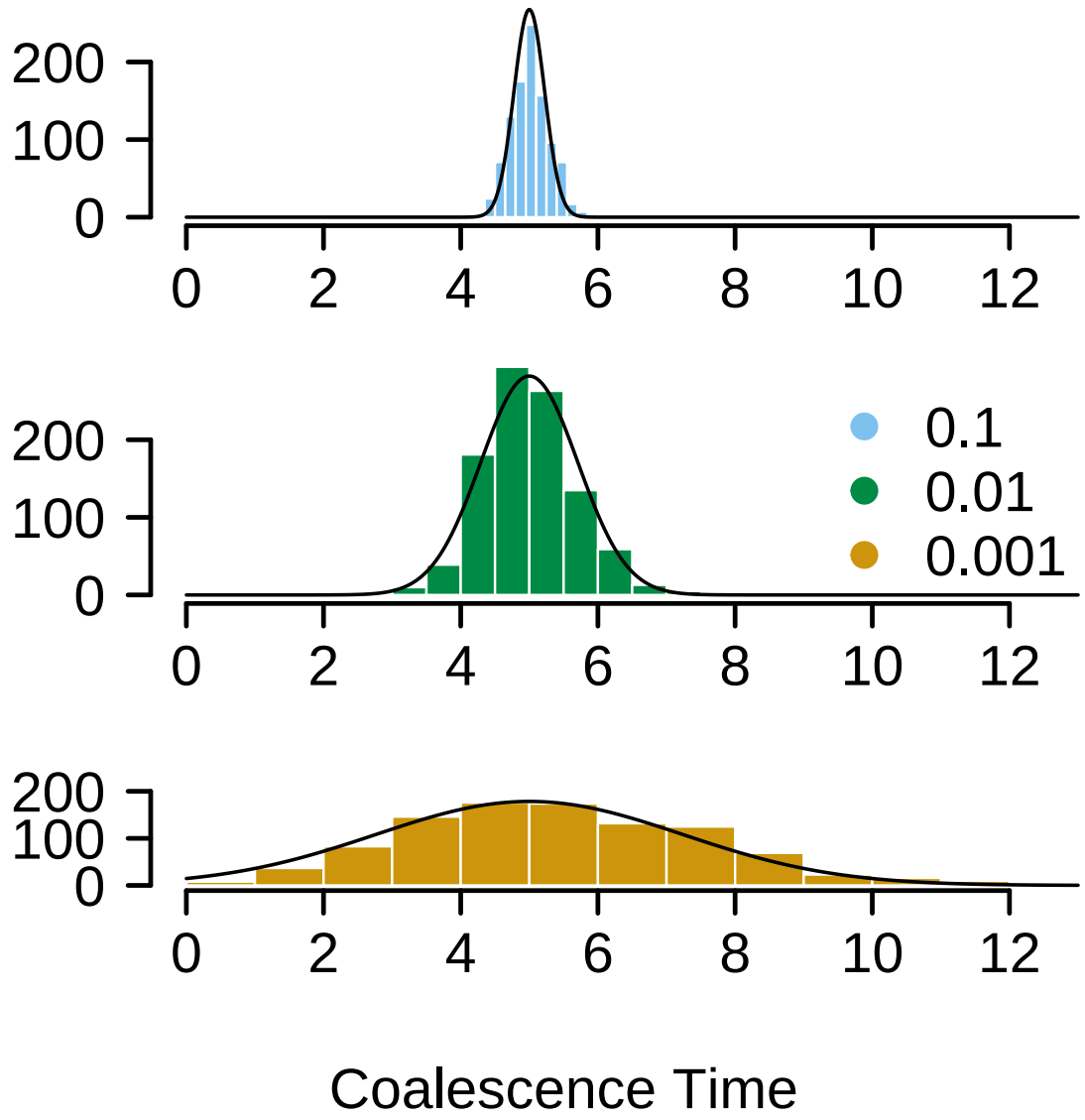


Figure A.1: **Modeling Mutational Variance:** The Normal approximation to Possion variance in coalescent time estimation error due to the mutational process for varying mutation rates.

Appendix B

Supplement for Chapter 3

B.1 Simulation Details

For a given divergence time, we simulated (n_X, n_Y) haploid individuals from each population under a species tree with split time parameterized by the value seen on the x-axis on the figures. In every simulation, we use the population size of species X, η_X, to be the reference population size. As well, we assume $\eta_X = \eta_{XY}$, but we vary population size η_Y to be 1, 2, and 10 times the value of η_X. Gene trees are generated using ms, and empirical estimates indicated by open circles in the figures are calculated using independent trees. With the simulated gene trees, to evaluate empirical estimates sequence data (10,000 bp per gene) is generated under a finite-sites model with the Jukes-Cantor model of evolution, using Seq-Gen, and a mutation parameter, which we vary $2\mu\eta_X = 0.1, 1.0, 10.0$. From the pairwise differences in each sequence, we generate empirical estimates using the sample means, variances and covariances from the independent replicates, to simulate estimates from the type of data used in practice (finite sites).

B.2 Mean, Variance and Covariance of Average Pairwise Differences

Using results presented earlier, we can write exact expressions for the variance and covariance of average pairwise differences as functions of the mutation rate (μ), sample sizes (n_X, n_Y), and coalescence times:

$$\begin{aligned}\mathrm{Var}(d_X) = \frac{\mu}{n_X(n_X-1)}\Big[&4\mathbb{E}(t_{i,i'}) + 8\mu\mathrm{Var}(t_{i,i'}) + 4(n_X-2)\big(\mathbb{E}(t_{i,i'}\cap t_{i,i''})\\ &+ 4\mu\mathrm{Cov}(t_{i,i'},t_{i,i''})\big) + (n_X-2)(n_X-3)\big(\mathbb{E}(t_{i,i'}\cap t_{i'',i'''})\\ &+ 4\mu\mathrm{Cov}(t_{i,i'},t_{i'',i'''})\big)\Big] - 4\mu^2\mathbb{E}(t_{i,i'})^2\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(d_Y) = \frac{\mu}{n_Y(n_Y-1)}\Big[&4\mathbb{E}(t_{j,j'}) + 8\mu\mathrm{Var}(t_{j,j'}) + 4(n_Y-2)\big(\mathbb{E}(t_{j,j'}\cap t_{j,j''})\\ &+ 4\mu\mathrm{Cov}(t_{j,j'},t_{j,j''})\big) + (n_Y-2)(n_Y-3)\big(\mathbb{E}(t_{j,j'}\cap t_{j'',j'''})\\ &+ 4\mu\mathrm{Cov}(t_{j,j'},t_{j'',j'''})\big)\Big] - 4\mu^2\mathbb{E}(t_{j,j'})^2\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(d_{XY}) = \frac{\mu}{n_X n_Y}\Big[&2\mathbb{E}(t_{i,j}) + 4\mu\mathrm{Var}(t_{i,j})\\ &+ (n_Y-1)\big(E(t_{i,j}\cap t_{i',j}) + 4\mu\mathrm{Cov}(t_{i,j},t_{i',j})\big)\\ &+ (n_X-1)\big(\mathbb{E}(t_{i,j}\cap t_{i,j'}) + 4\mu\mathrm{Cov}(t_{i,j},t_{i,j'})\big)\\ &+ (n_X-1)(n_Y-1)\big(\mathbb{E}(t_{i,j}\cap t_{i',j'}) + 4\mu\mathrm{Cov}(t_{i,j},t_{i',j'})\big)\Big]\\ - 4\mu^2\mathbb{E}(t_{i,j})&\end{aligned}$$

Similar to the the variance equations, the covariance terms can be expressed as functions of coalescence times by plugging in the expressions presented earlier:

$$\mathrm{Cov}(d_X, d_Y) = \mu\mathbb{E}(t_{i,i'\cap j,j'}) + 4\mu^2\mathrm{Cov}(t_{i,i'}, t_{j,j'})$$

$$\begin{aligned}\mathrm{Cov}(d_{XY}, d_X) = \frac{2}{n_X}\Big(&\mu\mathbb{E}(t_{i,i'\cap i,j}) + 4\mu^2\mathrm{Cov}(t_{i,i'}, t_{i,j})\Big)\\ &+ \frac{n_X-2}{n_X}\Big(\mu\mathbb{E}(t_{i,i'\cap i'',j}) + 4\mu^2\mathrm{Cov}(t_{i,i'}, t_{i'',j})\Big)\end{aligned}$$

$$\begin{aligned}\mathrm{Cov}(d_{XY}, d_Y) = \frac{2}{n_Y}\Big(&\mu\mathbb{E}(t_{j,j'\cap i,j}) + 4\mu^2\mathrm{Cov}(t_{j,j'}, t_{i,j})\Big)\\ &+ \frac{n_Y-2}{n_Y}\Big(\mu\mathbb{E}(t_{j,j'\cap i,j''}) + 4\mu^2\mathrm{Cov}(t_{j,j'}, t_{i,j''})\Big)\end{aligned}$$

B.3 Comparing against Takahata and Nei's Results

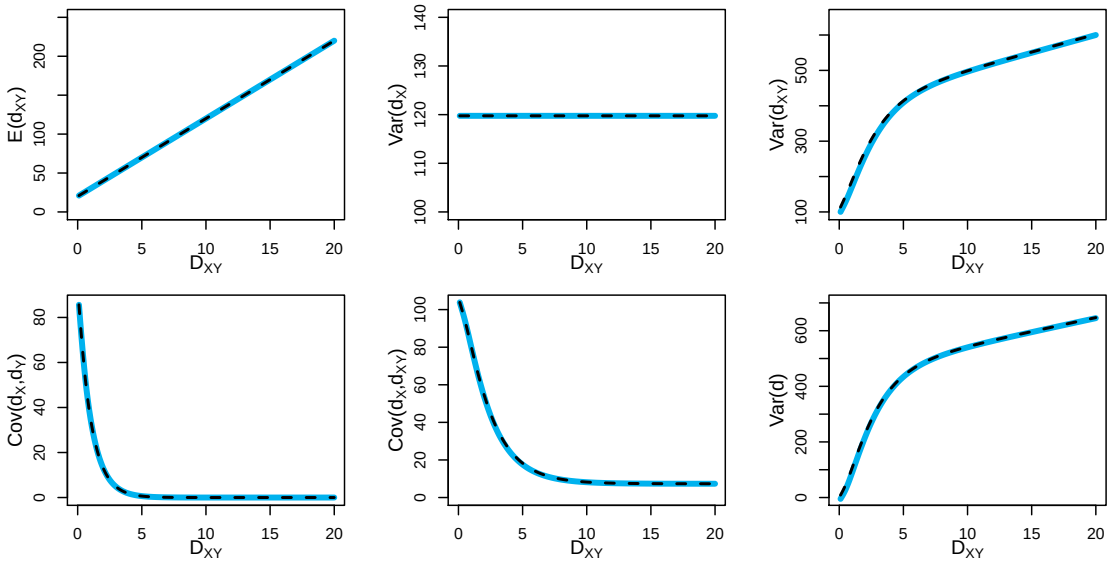


Figure B.1: **Comparing Results to Takahata and Nei, constant population size.** We compare the theoretical results derived by Takahata and Nei (blue solid lines) with those derived here (black dashed lines), for a sample size of $n_x = 10$, and variable divergence time D_{XY}. Note that $\text{Var}(d_Y)$ and $\text{Cov}(d_Y, d_{XY})$ are not pictured as they are identical to their d_X counterparts, by swapping n_X for n_Y in the equations.

Here we compare against Takahata and Nei's [49] results for the variance and covariance in pairwise differences under a constant population size model, and sample sizes, n_X, n_Y. Under this scenario, Takahata and Nei (with partial results also derived by

Tajima [48]) find:

$$
\begin{aligned}
\mathbb{E}(d_{XY}) &= \mathbb{E}(d_X)\big(1+\frac{1}{2}D_{XY}\big)\\
\mathrm{Var}(d_X) &= \frac{n_X+1}{3(n_X-1)}\mathbb{E}(d_X) + \frac{2(n_X^2+n_X+3)}{9n_X(n_X-1)}\mathbb{E}(d_X)^2\\
\mathrm{Var}(d_{XY}) &= (1-e^{-D_{XY}/2})^2\Big[(D_{XY}/2+1-2F)\mathbb{E}(d_X)+\mathbb{E}(d_X)^2\Big]\\
&\quad + 2e^{-D_{XY}/2}(1-e^{-D_{XY}/2})\Big[\big(\frac{D_{XY}}{4}+\frac{1}{2}-F\big)\mathbb{E}(d_X)+\frac{1}{3}\mathbb{E}(d_X)^2\Big]\\
&\quad + e^{-D_{XY}}\Big[\frac{1}{3}\mathbb{E}(d_X)+\frac{2}{9}\mathbb{E}(d_X)^2\Big]\\
\mathrm{Cov}(d_X,d_Y) &= e^{-D_{XY}}\Big(\frac{1}{3}\mathbb{E}(d_X)+\frac{2}{9}\mathbb{E}(d_X)^2\Big)\\
\mathrm{Cov}(d_{XY},d_X) &= \frac{2}{n_X}\Big[(1-e^{-D_{XY}/2})\mathbb{E}(d_X)F + e^{-D_{XY}/2}\big\{\frac{1}{2}(\frac{D_{XY}}{2}+1)\mathbb{E}(d_X)\\
&\quad + \frac{1}{3}\mathbb{E}(d_X)^2\big\}\Big] + \frac{n_X-2}{n_x}\Bigg[\frac{1}{3}\Big(1-\frac{3}{2}e^{-D_{XY}/2}+\frac{1}{2}e^{-3D_{XY}/2}\Big)\mathbb{E}(d_X)Z_1\\
&\quad + \frac{3}{2}(e^{-D_{XY}/2}-e^{-3D_{XY}/2})\Big\{\frac{1}{3}(\frac{D_{XY}}{2}+1-Z_2)\mathbb{E}(d_X)+\frac{2}{9}\mathbb{E}(d_X)^2\Big\}\\
&\quad + e^{-3D_{XY}/2}\Big(\frac{1}{3}\mathbb{E}(d_X)+\frac{2}{9}\mathbb{E}(d_X)^2\Big)\Bigg]
\end{aligned}
$$

Where $\mathbb{E}(d_X)=\mathbb{E}(d_Y)$, $F=\frac{1}{2}\big(1-(D_{XY}/2+1)e^{-D_{XY}/2}\big)$,

$Z_1 = 1-\frac{3}{2}(1+D_{XY}/2)e^{-D_{XY}/2}+\frac{1}{2}(1+3D_{XY}/2)e^{-3D_{XY}/2}$, and $Z_2=\frac{1}{3}\big\{1-(1+3D_{XY}/2)e^{-3D_{XY}/2}\big\}$. Results for $\mathrm{Var}(d_Y)$, $\mathrm{Cov}(d_{XY},d_Y)$ can be obtained by replacing n_x with n_Y.

Figure B.1 demonstrates that the results from Takahata and Nei match those derived in this manuscript with calculations done using STCov. The results presented in this manuscript can therefore be viewed as generalizations to those above.

B.4 Average Pairwise Difference Accuracy Plots

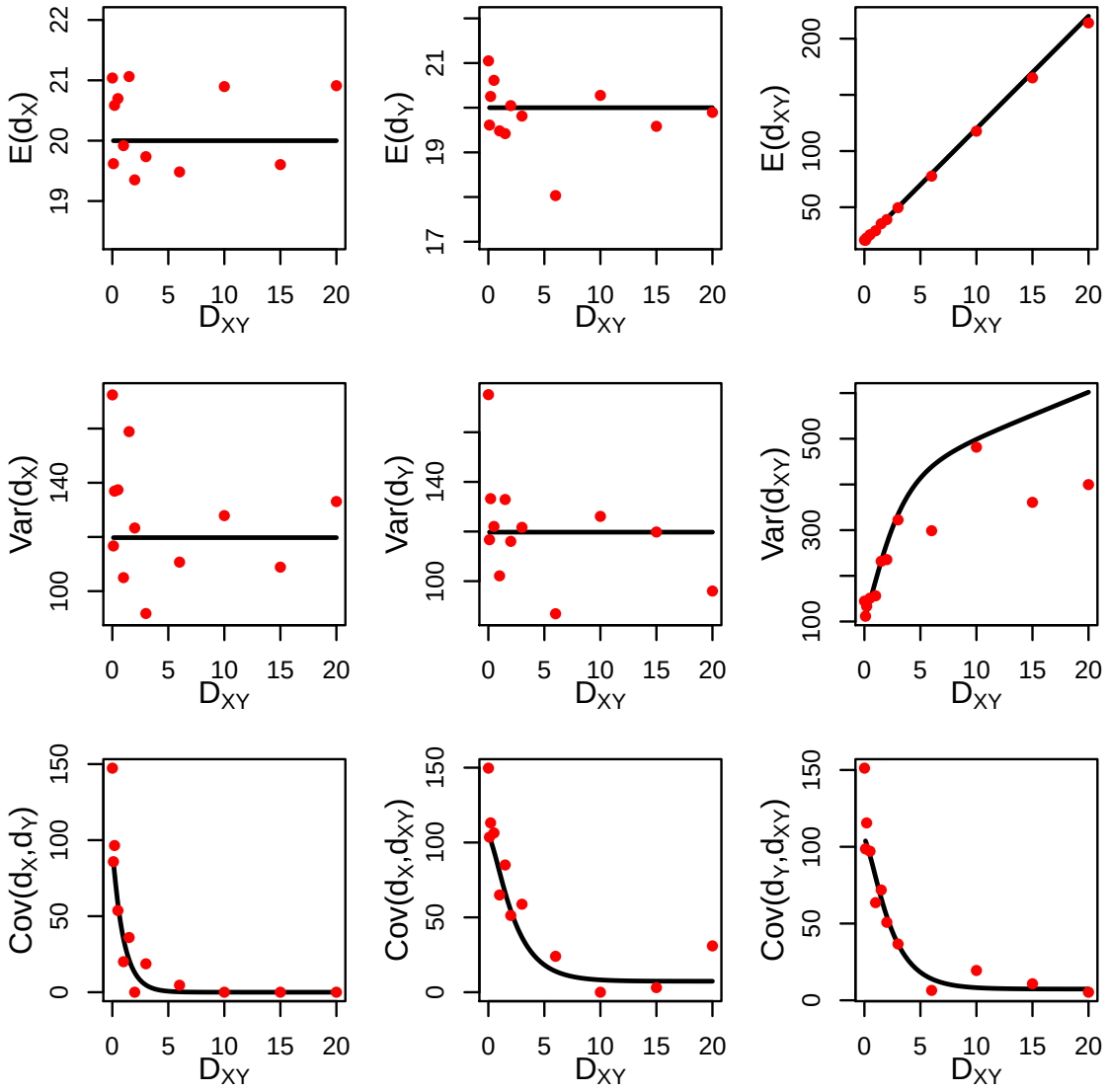


Figure B.2: **Average pairwise difference results,** $2\mu\eta_X = 10$, $\eta_Y = 1\eta_X$.

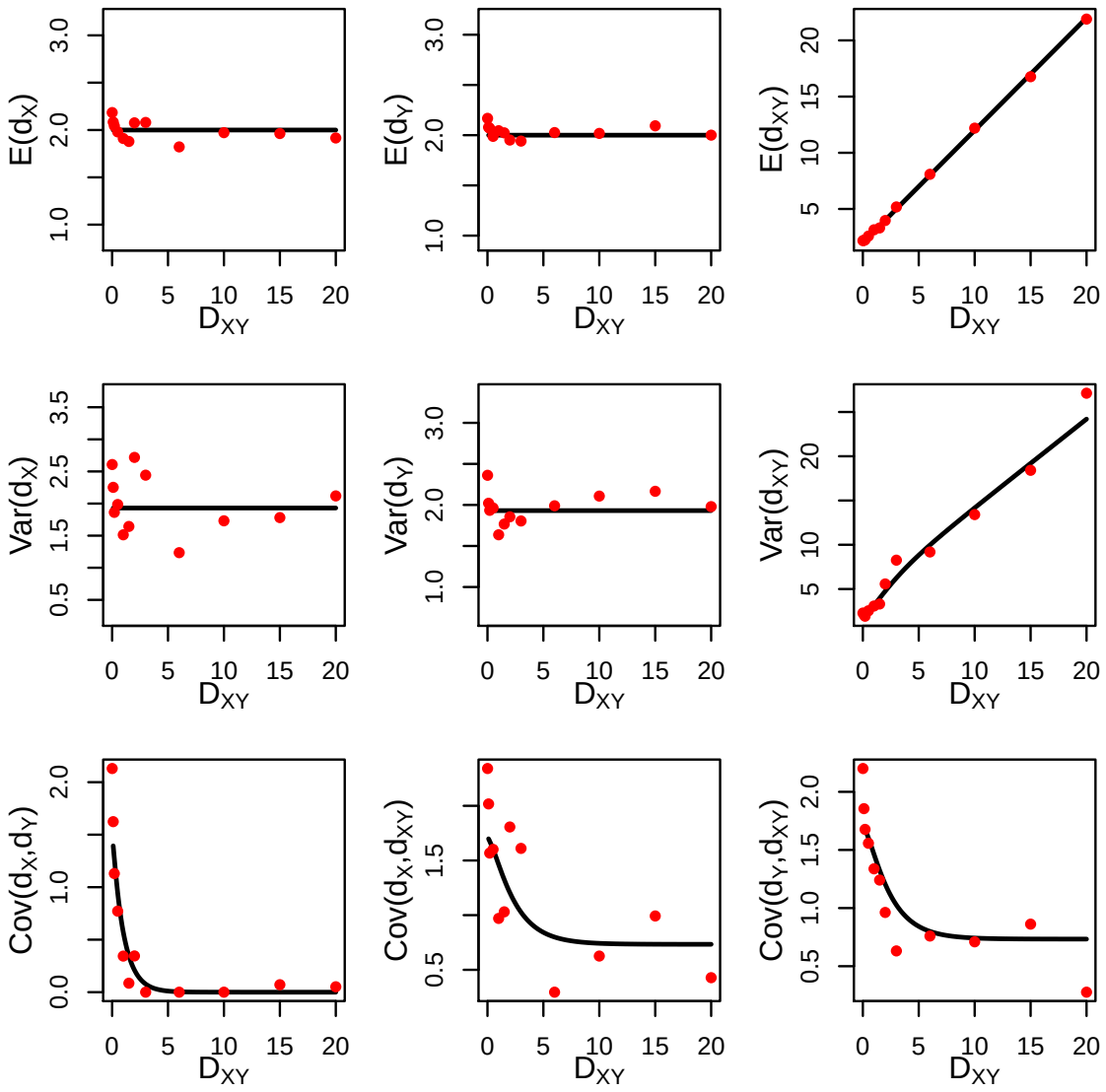


Figure B.3: **Average pairwise difference results,** $2\mu\eta_X = 1$, $\eta_Y = 1\eta_X$.

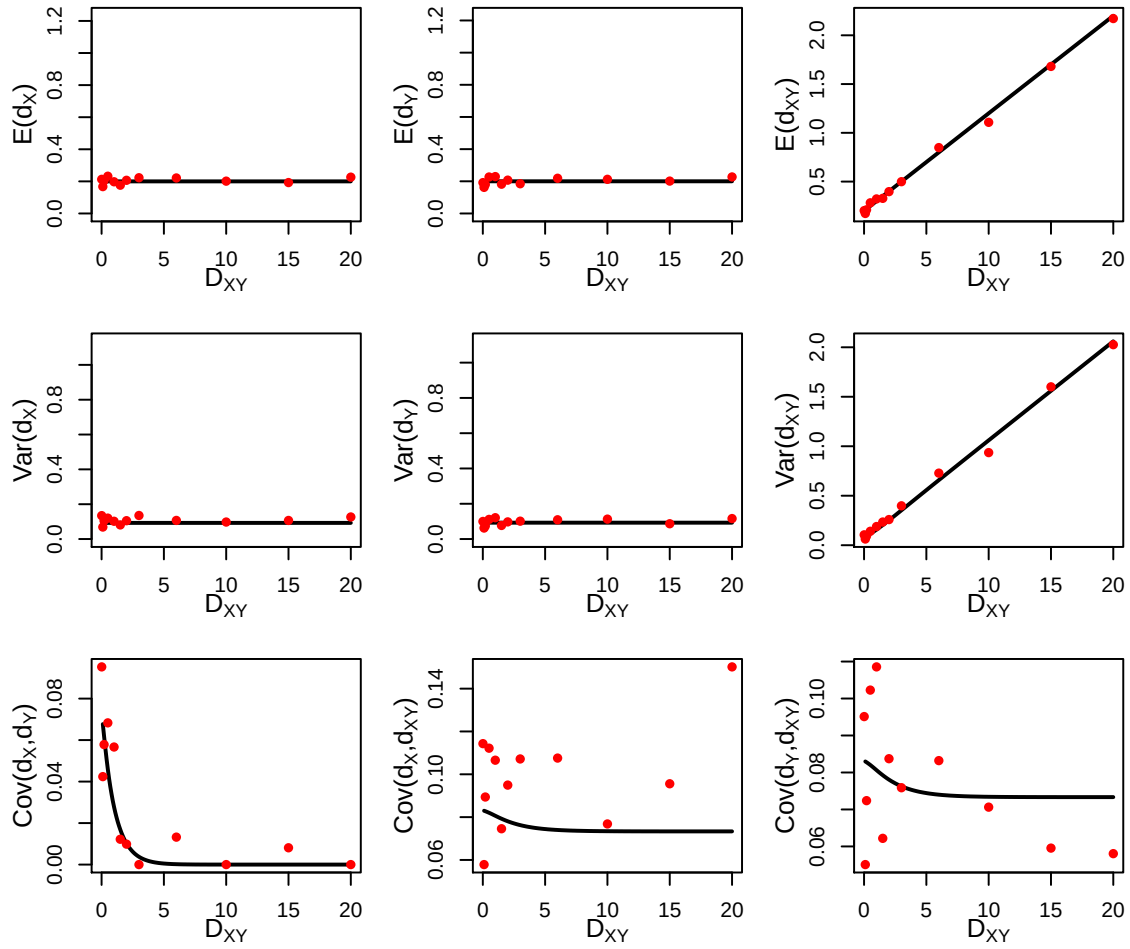


Figure B.4: **Average pairwise difference results,** $2\mu\eta_X = 0.1$, $\eta_Y = 1\eta_X$**.**

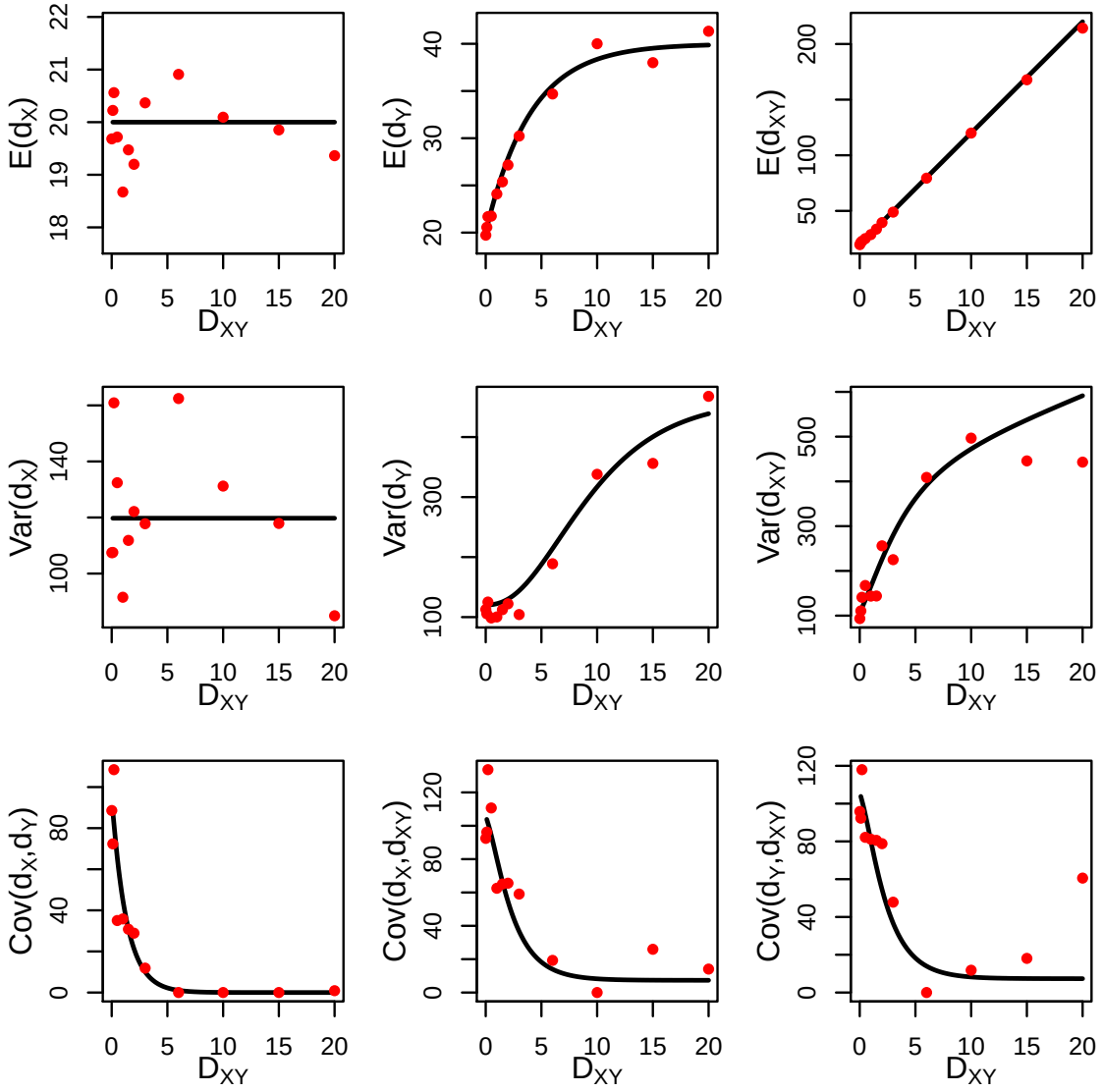


Figure B.5: **Average pairwise difference results,** $2\mu\eta_X = 10$, $\eta_Y = 2\eta_X$.

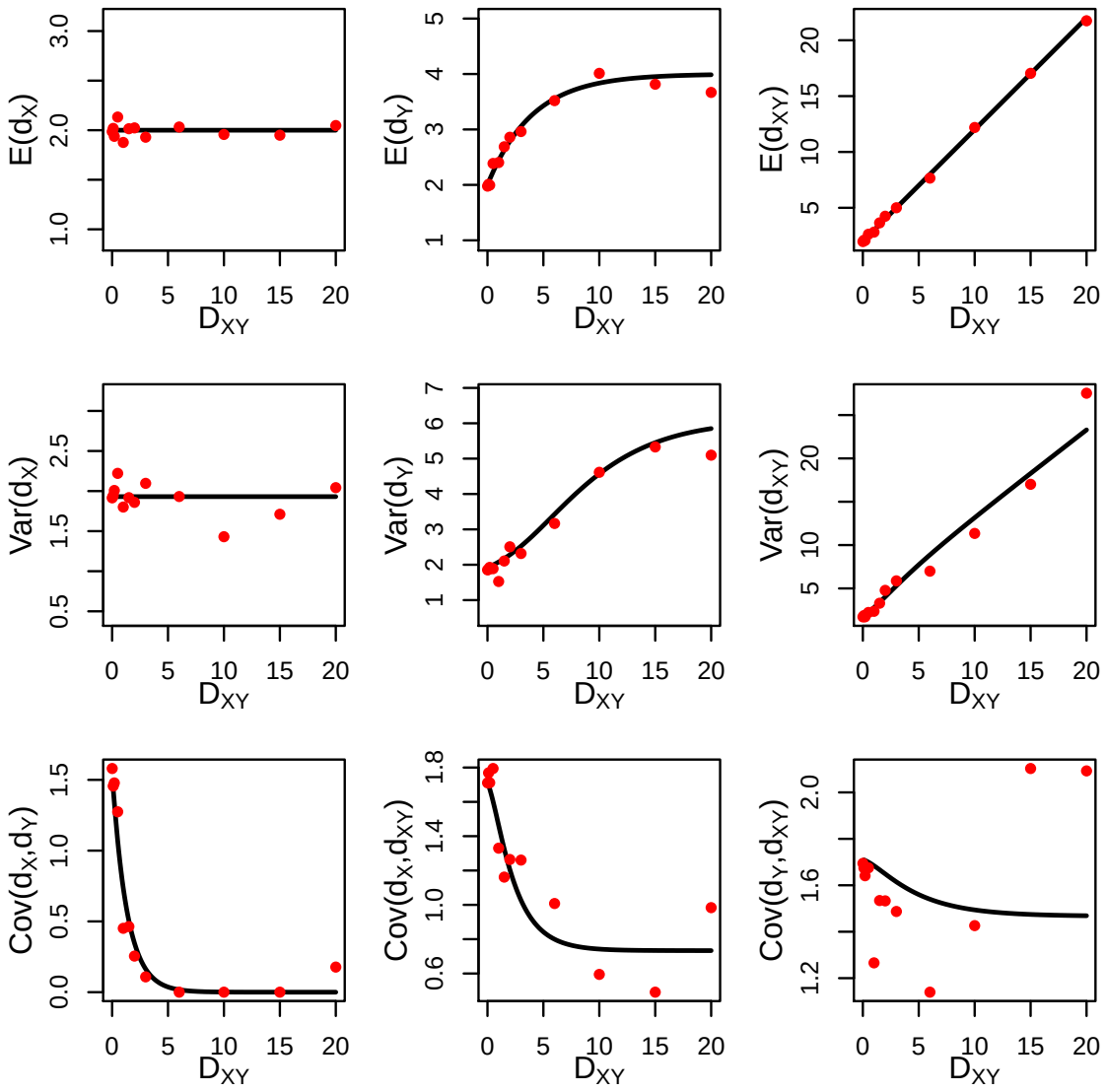


Figure B.6: **Average pairwise difference results,** $2\mu\eta_X = 1$**,** $\eta_Y = 2\eta_X$**.**

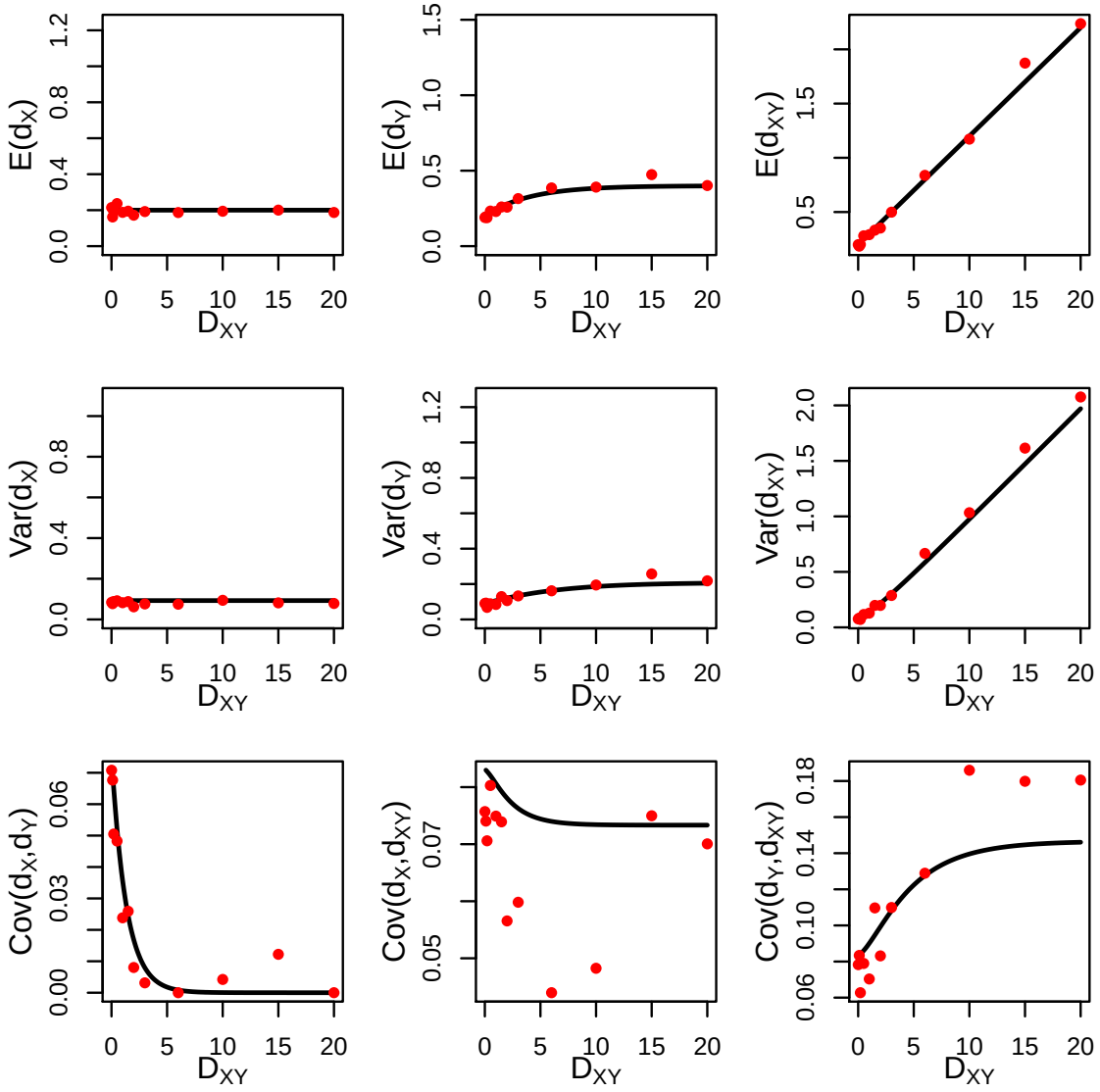


Figure B.7: **Average pairwise difference results,** $2\mu\eta_X = 0.1$**,** $\eta_Y = 2\eta_X$**.**

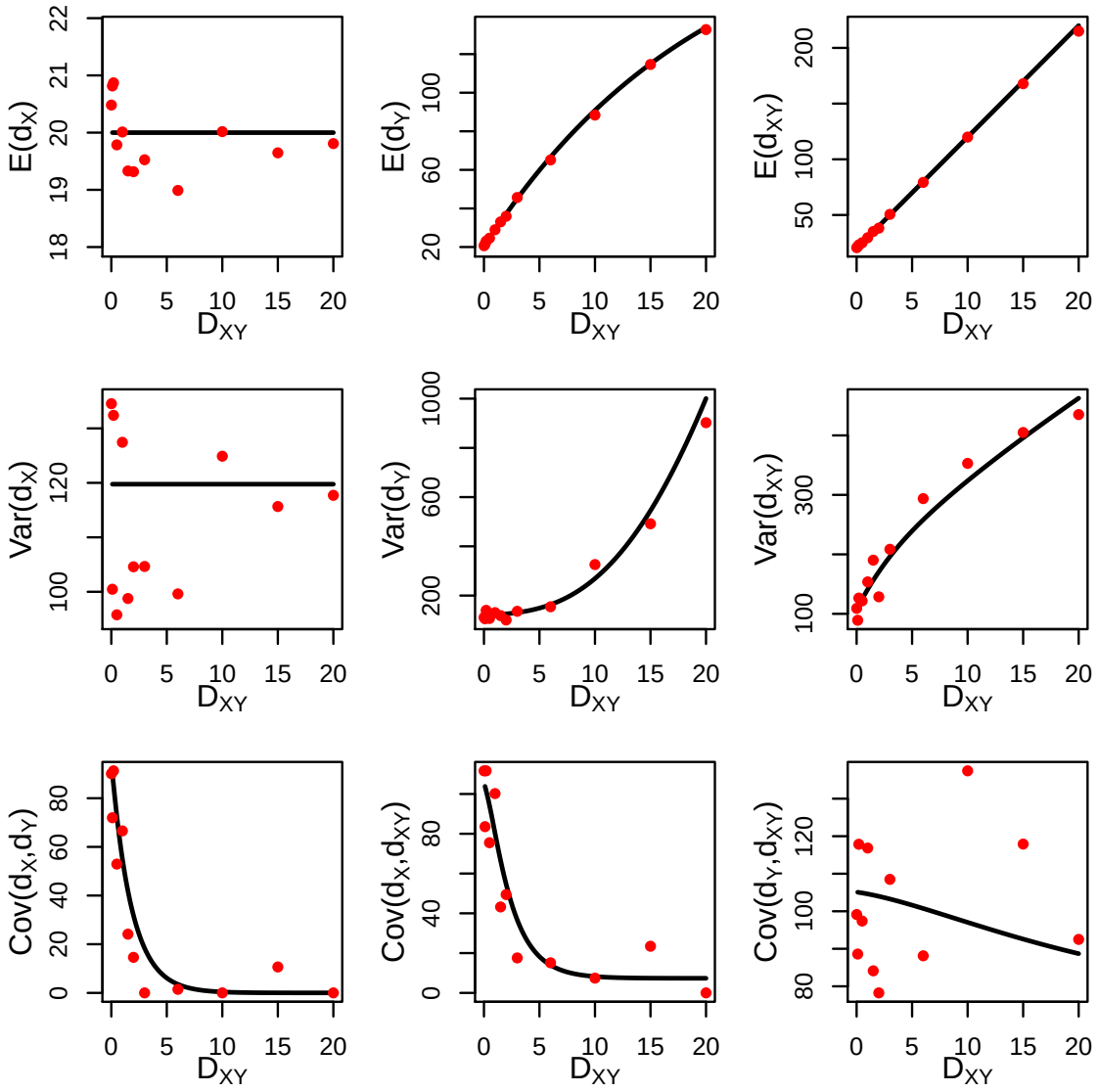


Figure B.8: **Average pairwise difference results,** $2\mu\eta_X = 10$, $\eta_Y = 10\eta_X$**.**

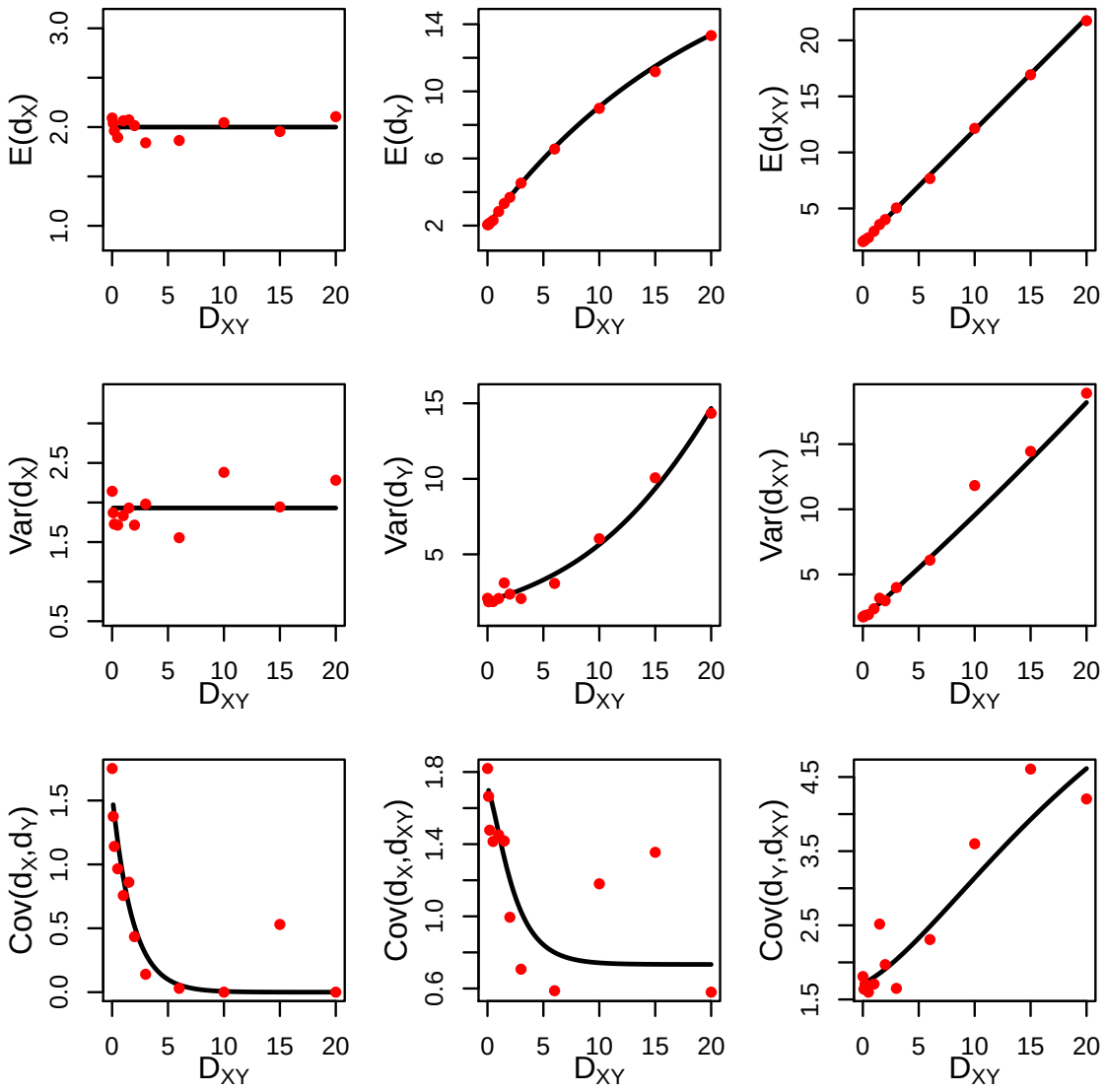


Figure B.9: **Average pairwise difference results,** $2\mu\eta_X = 1$, $\eta_Y = 10\eta_X$**.**

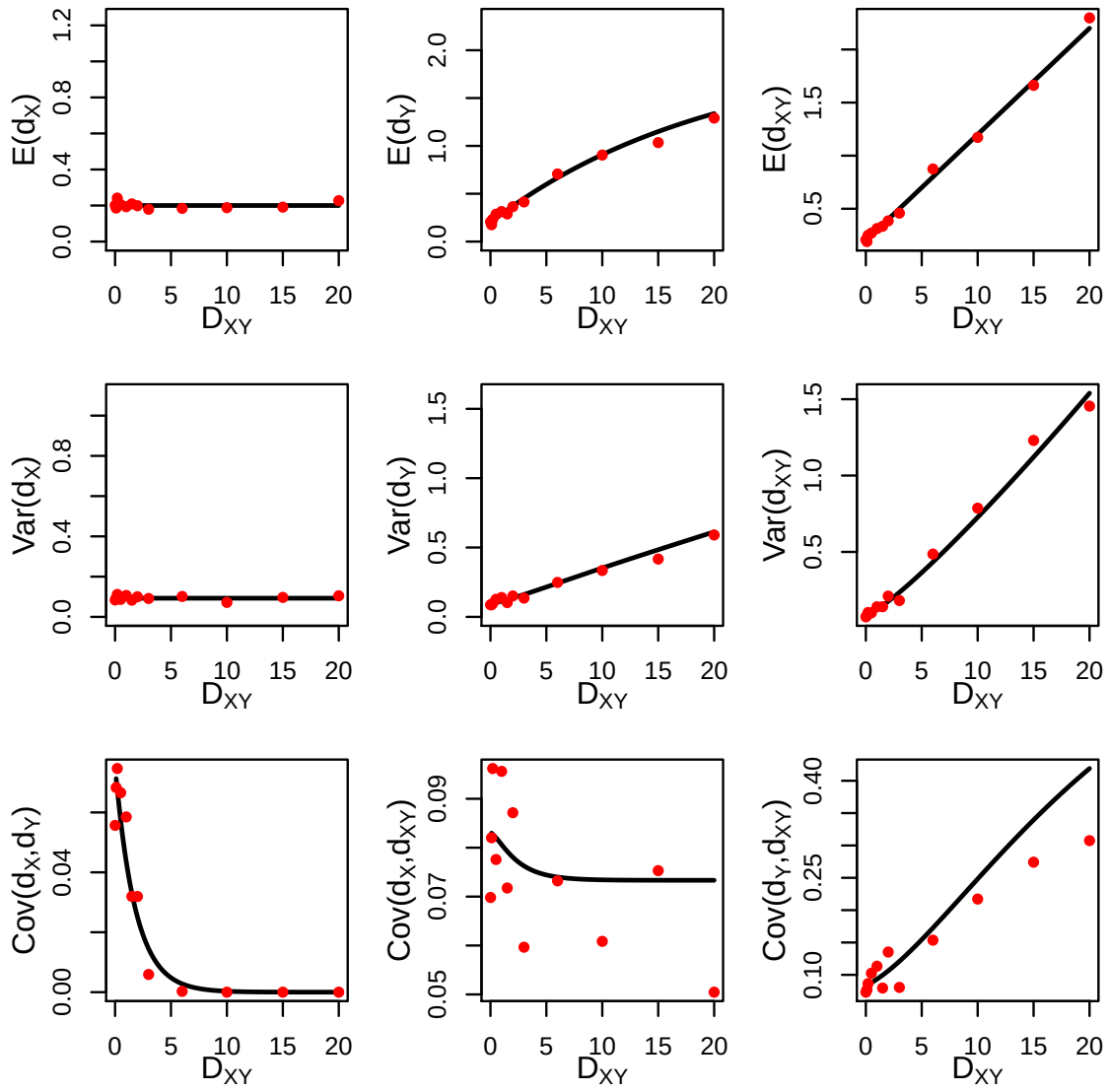


Figure B.10: **Average pairwise difference results,** $2\mu\eta_X = 0.1$, $\eta_Y = 10\eta_X$.

B.5 Covariance and Shared Branch Length

In this section we provide further details on the calculations of covariance and expected shared branch length for a pair of coalescence events. As mentioned in the main text, these calculations depend on the species tree topology. For 4 unique individuals a, b, c, d, there are 5 canonical topologies which need to be considered. Figure

B.11 illustrates these possibilities. Note that for events in which multiple individuals are in the same population, we simply assume divergence times are 0 on the tree.

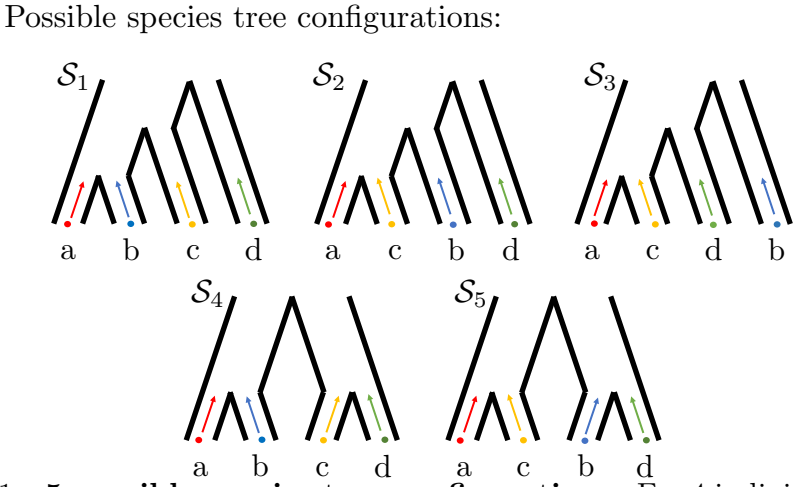


Figure B.11: **5 possible species tree configurations.** For 4 individuals,(a, b, c, d) there are 5 canonical species tree configurations to consider when calculating quantities $\text{Cov}(t_{a,b}, t_{c,d})$ and $\mathbb{E}(t_{a,b\cap c,d})$. Here we assume some exchangeability between lineages a,b and between lineages c,d.

In the following sections we discuss the general equations and tools needed to calculate the expected shared branch length and covariance as presented in the main text.

Expected shared branch length, 4 unique individuals

Here we present some details into the equations/recursions for the expected shared branch length of two coalescence events t_{ab} and t_{cd} where a, b, c, d are all unique individuals, who can originate from different modern day species. There is some underlying species tree $\mathcal{S}$, which we condition all of our calculations on, implicitly. Let D_{ab} and D_{cd} represent the divergence times on $\mathcal{S}$ of these pairs of individuals.

Defining shared branch length

To understand the expected shared branch length between two coalescence branches $2t_{ab}$ and $2t_{cd}$, we need an explicit formula to calculate shared branch length. Following the logic of [34], we take a two step approach:
Step 1:

Compute

$$\beta_1 = t_{ab} + t_{cd}$$
$$\beta_2 = t_{ac} + t_{bd}$$
$$\beta_3 = t_{ad} + t_{bc}$$

Step 2:

Let $\beta_m = \min(\beta_1, \beta_2, \beta_3)$. Then compute

$$t_{a,b\cap c,d} = (\beta_1 - \beta_m)$$

where $t_{a,b\cap c,d}$ represents the total length of overlap shared by the branch between a, b the branch connecting c, d. Note when there is no shared branch length, $t_{a,b\cap c,d} = 0$, which occurs when t_{ab} or t_{cd} is the minimum coalescence time on the tree of these 4 individuals. See figure 3.1 for an illustrative definition of expected shared branch length.

We are interested in

$$\mathbb{E}[t_{a,b\cap c,d}] = \mathbb{E}[\beta_1 - \beta_m] = \mathbb{E}[t_{ab} + t_{cd} - \beta_m] = \mathbb{E}[t_{ab}] + \mathbb{E}[t_{cd}] - \mathbb{E}[\beta_m]$$

We already know $\mathbb{E}[t_{ab}]$ and $\mathbb{E}[t_{cd}]$ from earlier mean calculations. What is left to do is derive the equations for $\mathbb{E}[\beta_m]$.

Expanding $\mathbb{E}[\beta_m]$

Using the definitions above, we can expand β_m to the following, by conditioning on the probability that quantity β_1, β_2 or β_3 is the minimum value,

$$\begin{aligned}
\mathbb{E}[\beta_m] =& \mathbb{E}[\min(\beta_1, \beta_2, \beta_3)] \\
=& \mathbb{E}[\min(t_{ab} + t_{cd}, t_{ac} + t_{bd}, t_{ad} + t_{bc})] \\
=& \mathbb{E}[t_{ab} + t_{cd} | t_{ab} + t_{cd} = \beta_m] P(t_{ab} + t_{cd} = \beta_m) \\
&+ \mathbb{E}[t_{ac} + t_{bd} | t_{ac} + t_{bd} = \beta_m] P(t_{ac} + t_{bd} = \beta_m) \\
&+ \mathbb{E}[t_{ad} + t_{bc} | t_{ad} + t_{bc} = \beta_m] P(t_{ad} + t_{bc} = \beta_m)
\end{aligned}$$

Next, it is important to observe that $t_{ab} + t_{cd}$ achieves the minimum if and only if t_{ab} or t_{cd} is the first (most recent) coalescence event on the tree of 4 individuals a, b, c, d.

For the sake of generality, we will focus on one of the three terms, although all results hold for the other two, by just appropriately re-labeling terms.

Focusing on $\mathbb{E}[t_{ab} + t_{cd} | t_{ab} + t_{cd} = b_m] P(t_{ab} + t_{cd} = \beta_m)$

Using the observation that the minimum is achieved if and only if one of the two times is the minimum coalescence time, we can further expand this quantity. For notation, let us denote C_m to be the first coalescence time on the tree.

$$\begin{aligned}
&\mathbb{E}[t_{ab} + t_{cd} | t_{ab} + t_{cd} = \beta_m] P(t_{ab} + t_{cd} = \beta_m) \\
&= \mathbb{E}\big[t_{ab} + t_{cd} | (t_{ab} = C_m) \text{ OR } (t_{cd} = C_m)\big] P\big((t_{ab} = C_m) \text{ OR } (t_{cd} = C_m)\big)
\end{aligned}$$

If we write out the integrals required to compute this quantity, and employ the use of Bayes' Theorem, we can form this into more recognizable terms,

$$\begin{aligned}
&\mathbb{E}\big[t_{ab} + t_{cd} | (t_{ab} = C_m) \text{ OR } (t_{cd} = C_m)\big] P\big((t_{ab} = C_m) \text{ OR } (t_{cd} = C_m)\big) \\
&\quad = \int_{D_{ab}}^{\infty} \int_{D_{cd}}^{\infty} (t_{ab} + t_{cd}) P\big(t_{ab}, t_{cd} | (t_{ab} = C_m) \text{ OR } (t_{cd} = C_m)\big) dt_{cd} t_{ab} \\
&\qquad \times P\big((t_{ab} = C_m) \text{ OR } (t_{cd} = C_m)\big) \\
&\quad = \int_{D_{ab}}^{\infty} \int_{D_{cd}}^{\infty} (t_{ab} + t_{cd}) \frac{P\big((t_{ab} = C_m) \text{ OR } (t_{cd} = C_m) | t_{ab}, t_{cd}\big) P(t_{ab}, t_{cd})}{P\big((t_{ab} = C_m) \text{ OR } (t_{cd} = C_m)\big)} dt_{cd} dt_{ab} \\
&\qquad \times P\big((t_{ab} = C_m) \text{ OR } (t_{cd} = C_m)\big) \\
&\quad = \int_{D_{ab}}^{\infty} \int_{D_{cd}}^{\infty} (t_{ab} + t_{cd}) P\big((t_{ab} = C_m) \text{ OR } (t_{cd} = C_m) | t_{ab}, t_{cd}\big) P(t_{ab}, t_{cd}) dt_{cd} dt_{ab}
\end{aligned}$$

Now, let's expand $P\big((t_{ab} = C_m) \text{ OR } (t_{cd} = C_m) | t_{ab}, t_{cd}\big)$. Observe, by the inclusion-exclusion principle

$$\begin{aligned}
&P\big((t_{ab} = C_m) \text{ OR } (t_{cd} = C_m) | t_{ab}, t_{cd}\big) \\
&\qquad = P(t_{ab} = C_m | t_{ab}, t_{cd}) + P(t_{cd} = C_m | t_{ab}, t_{cd}) \\
&\qquad\quad - P\big((t_{ab} = C_m) \text{ AND } (t_{cd} = C_m) | t_{ab}, t_{cd}\big) \\
&\qquad = P\big(t_{ab} = C_m | t_{ab}, t_{cd}\big) + P(t_{cd} = C_m | t_{ab}, t_{cd})
\end{aligned}$$

since $P\big(t_{ab} = C_m \text{ AND } t_{cd} = C_m | t_{ab}, t_{cd}\big) = 0$ as we assume all coalescence events are bifurcating between only two individuals at a time.

Applying Bayes' Theorem once more to each of these terms we see

$$\begin{aligned}
&P\big(t_{ab} = C_m | t_{ab}, t_{cd}) + P(t_{cd} = C_m | t_{ab}, t_{cd}) \\
&\quad = \frac{P(t_{ab}, t_{cd} | t_{ab} = C_m) P(t_{ab} = C_m)}{P(t_{ab}, t_{cd})} + \frac{P(t_{ab}, t_{cd} | t_{cd} = C_m) P(t_{cd} = C_m)}{P(t_{ab}, t_{cd})}
\end{aligned}$$

By plugging in this expression to our density in the double integral above, we can get a much more clear expression for the expectation

$$\begin{aligned}
&\int_{D_{ab}}^{\infty} \int_{D_{cd}}^{\infty} (t_{ab} + t_{cd}) P\big((t_{ab} = C_m) \text{ OR } (t_{cd} = C_m) | t_{ab}, t_{cd}\big) P(t_{ab}, t_{cd}) dt_{cd} dt_{ab} \\
&= \int_{D_{ab}}^{\infty} \int_{D_{cd}}^{\infty} (t_{ab} + t_{cd}) \Big[\frac{P(t_{ab}, t_{cd} | t_{ab} = C_m) P(t_{ab} = C_m)}{P(t_{ab}, t_{cd})} \\
&\quad + \frac{P(t_{ab}, t_{cd} | t_{cd} = C_m) P(t_{cd} = C_m)}{P(t_{ab}, t_{cd})} \Big] P(t_{ab}, t_{cd}) dt_{cd} dt_{ab} \\
&= \int_{D_{ab}}^{\infty} \int_{D_{cd}}^{\infty} (t_{ab} + t_{cd}) \Big[P(t_{ab}, t_{cd} | t_{ab} = C_m) P(t_{ab} = C_m) \\
&\quad + P(t_{ab}, t_{cd} | t_{cd} = C_m) P(t_{cd} = C_m) \Big] dt_{cd} dt_{ab} \\
&= \int_{D_{ab}}^{\infty} \int_{D_{cd}}^{\infty} (t_{ab} + t_{cd}) P(t_{ab}, t_{cd} | t_{ab} = C_m) P(t_{ab} = C_m) dt_{cd} dt_{ab} \\
&\qquad + \int_{D_{ab}}^{\infty} \int_{D_{cd}}^{\infty} (t_{ab} + t_{cd}) P(t_{ab}, t_{cd} | t_{cd} = C_m) P(t_{cd} = C_m) dt_{cd} dt_{ab} \\
&= \mathbb{E}[t_{ab} + t_{cd} | t_{ab} = C_m] P(t_{ab} = C_m) + \mathbb{E}[t_{ab} + t_{cd} | t_{cd} = C_m] P(t_{cd} = C_m)
\end{aligned}$$

To summarize this subsection, we have shown that

$$\begin{aligned}
&\mathbb{E}\big[t_{ab} + t_{cd} | (t_{ab} = C_m) \text{ OR } (t_{cd} = C_m)\big] P\big((t_{ab} = C_m) \text{ OR } (t_{cd} = C_m)\big) \\
&\qquad = \mathbb{E}[t_{ab} + t_{cd} | t_{ab} = C_m] P(t_{ab} = C_m) + \mathbb{E}[t_{ab} + t_{cd} | t_{cd} = C_m] P(t_{cd} = C_m)
\end{aligned}$$

Next we will focus on one of these two expectation terms, $\mathbb{E}[t_{ab} + t_{cd} | t_{ab} = C_m] P(t_{ab} = C_m)$, as it can be generalized to all terms in this calculation.

Immediately note that we can expand each expectation using the linearity of expectations

$$\begin{aligned}\mathbb{E}[t_{ab} + t_{cd}|t_{ab} = C_m]P(t_{ab} = C_m) \\ = \underbrace{\mathbb{E}[t_{ab}|t_{ab} = C_m]P(t_{ab} = C_m)}_{\text{Term 1}} + \underbrace{\mathbb{E}[t_{cd}|t_{ab} = C_m]P(t_{ab} = C_m)}_{\text{Term 2}}\end{aligned}$$

Recall that the expectation is still over the joint conditional distribution $P(t_{ab}, t_{cd}|t_{ab} = C_m)$ (and similarly with t_{cd} in the conditional for the second term).

In the next two subsections, we present a derivations of terms 1 and 2 in case when a, b, c, d are all in the same (potentially ancient) population. The rest of the cases are implemented in our C++ code.

Calculating Term 1: $\mathbb{E}[t_{ab}|t_{ab} = C_m]P(t_{ab} = C_m)$

Under the condition that all 4 individuals having survived to the same species without coalescing, the probability that any one of the possible 6 pairs is the first to coalesce is $\frac{1}{6}$. So we know immediately $P(t_{ab} = C_m) = \frac{1}{6}$. Let D_{abcd} be the time at which all 4 individuals have a common species ancestor. Let H_{abcd} represent the population size change history back in time starting at time D_{abcd}. H_{abcd} can be viewed as a list of time/size pairs (τ_i, η_i) where $\tau_1 = D_{abcd}$.

$$\begin{aligned}\text{Term 1} &= \mathbb{E}[t_{ab}|t_{ab} = C_m]P(t_{ab} = C_m) \\ &= \int_{D_{abcd}}^{\infty} t_{ab}P(t_{ab}|t_{ab} = C_m)P(t_{ab} = C_m)dt_{ab} \\ &= \sum_{i \in H_{abcd}} P(t_{ab} > t_i) \int_{\tau_i}^{\tau_{i+1}} t_{ab}P(t_{ab}|t_{ab} = C_m)P(t_{ab} = C_m)dt_{ab} \\ &= \sum_{i \in H_{abcd}} P(t_{ab} > t_i)\frac{1}{6} \int_{\tau_i}^{\tau_{i+1}} t_{ab}P(t_{ab}|t_{ab} = C_m)dt_{ab} \\ &= \sum_{i \in H_{abcd}} P(t_{ab} > \tau_i)\frac{1}{6} \int_{\tau_i}^{\tau_{i+1}} t_{ab}P(\text{6p Coal} = t_{ab}|\tau_i)dt_{ab} \\ &= \sum_{i \in H_{abcd}} P(t_{ab} > \tau_i)\frac{1}{6} \int_{\tau_i}^{\tau_{i+1}} t_{ab}\frac{6}{2\eta_i}e^{\frac{-6(t_{ab}-\tau_i)}{2\eta_i}}dt_{ab}\end{aligned}$$

Notation: $\Big(\text{6p Coal} = t|\tau_i\Big)$ is the event that 6 possible pairs can coalesce starting at time τ_i, and the time of the first event is $t > \tau_i$. These 6 pairs represent the

6 possible pairings: $\{a|b, a|c, a|d, b|c, b|d, c|d\}$. We formally present this and more notation at the end of this manuscript.

Calculating Term 2: $\mathbb{E}[t_{cd}|t_{ab} = C_m]P(t_{ab} = C_m)$

$$\begin{aligned}
\text{Term 2} &= \mathbb{E}[t_{cd}|t_{ab} = C_m]P(t_{ab} = C_m) \\
&= \int_{D_{abcd}}^{\infty}\int_{D_{abcd}}^{\infty} t_{cd}P(t_{ab}, t_{cd}|t_{ab} = C_m)P(t_{ab} = C_m)dt_{ab}dt_{cd} \\
&= \sum_{i\in H_{abcd}} P(T_{cd} > \tau_i)\int_{\tau_i}^{\tau_{i+1}} t_{cd}\int_{D_{abcd}}^{t_{cd}} P(t_{ab}|t_{cd}, t_{ab} = C_m)P(t_{ab} = C_m|t_{cd}) \\
&\quad \times P(t_{cd})dt_{cd}dt_{ab} \\
&= \sum_{i\in H_{abcd}} P(t_{cd} > \tau_i) \\
&\quad \times \underbrace{\int_{\tau_i}^{\tau_{i+1}} t_{cd}P(t_{cd})\int_{D_{abcd}}^{t_{cd}} P(t_{ab}|t_{cd}, t_{ab} = C_m)P(t_{ab} = C_m|t_{cd})dt_{cd}dt_{ab}}_{\text{Term 2}(i)}
\end{aligned}$$

$$\begin{aligned}
\text{Term 2(i)} &= \int_{\tau_i}^{\tau_{i+1}} t_{cd}P(t_{cd})\int_{D_{abcd}}^{t_{cd}} P(t_{ab}|t_{cd}, t_{ab} = C_m)P(t_{ab} = C_m|t_{cd})dt_{cd}dt_{ab} \\
&= \int_{\tau_i}^{\tau_{i+1}} t_{cd}P(t_{cd})\Bigg[\sum_{k\in H_{abcd}, k<i}\Big(P(\text{5p No Coal } \in (0,\tau_k)) \\
&\quad \times \int_{\tau_k}^{\tau_{k+1}} P(t_{ab}|t_{cd}, t_{ab} = C_m)dt_{ab}\Big) \\
&\quad + P(\text{5p No Coal } \in (0,\tau_i))\int_{\tau_i}^{t_{cd}} P(t_{ab}|t_{cd}, t_{ab} = C_m)dt_{ab}\Bigg]dt_{cd} \\
&= \int_{\tau_i}^{\tau_{i+1}} t_{cd}P(t_{cd})\Bigg[\sum_{k\in H_{abcd}, k<i}\Big(P(\text{5p No Coal } \in (0,\tau_k)) \\
&\quad \times \int_{\tau_k}^{\tau_{k+1}} \frac{1}{5}P(\text{5p Coal} = t_{ab}|\tau_k)dt_{ab}\Big) \\
&\quad + P(\text{5p No Coal } \in (0,\tau_i))\int_{\tau_i}^{t_{cd}} \frac{1}{5}P(\text{5p Coal} = t_{ab}|\tau_i, t_{cd})dt_{ab}\Bigg]dt_{cd}
\end{aligned}$$

$$
\begin{aligned}
&= \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}) \Bigg[\sum_{k \in H_{abcd}, k<i} \Big(P(\text{5p No Coal } \in (0, \tau_k)) \\
&\quad \times \frac{1}{5} \Big[1 - P(\text{5p No Coal } \in (\tau_k, \tau_{k+1})) \Big] \Big) \\
&\quad + P(\text{5p No Coal } \in (0, \tau_i)) \frac{1}{5} \Big[1 - P(\text{5p No Coal } \in (\tau_i, t_{cd})) \Big] \Bigg] dt_{cd} \\
&= \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}) dt_{cd} \Bigg[\sum_{k \in H_{abcd}, k<i} \Big(P(\text{5p No Coal } \in (0, \tau_k)) \\
&\quad \times \frac{1}{5} \Big[1 - P(\text{5p No Coal } \in (\tau_k, \tau_{k+1})) \Big] \Big) \Bigg] \\
&\quad + P(\text{5p No Coal } \in (0, \tau_i)) \frac{1}{5} \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}) \Big[1 - P(\text{5p No Coal } \in (\tau_i, t_{cd})) \Big] dt_{cd} \\
&= \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}) dt_{cd} \Bigg[\sum_{k \in H_{abcd}, k<i} \Big(P(\text{5p No Coal } \in (0, \tau_k)) \\
&\quad \times \frac{1}{5} \Big[1 - P(\text{5p No Coal } \in (\tau_k, \tau_{k+1})) \Big] \Big) \Bigg] \\
&\quad + P(\text{5p No Coal } \in (0, \tau_i)) \frac{1}{5} \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}) dt_{cd} \\
&\quad - P(\text{5p No Coal } \in (0, \tau_i)) \frac{1}{5} \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}) P(\text{5p No Coal } \in (\tau_i, t_{cd})) dt_{cd}
\end{aligned}
$$

Notation: $\Big(\text{5p No Coal } \in (\tau_i, t)\Big)$ is the event that none of the 5 possible pairs coalesce in the time interval (τ_i, t). We formally present this and more notation at the end of this manuscript.

Expected Shared Branch Length, 3 unique individuals

Here I am deriving the equations/recursions for the expected shared branch length of two coalescence events t_{ab}, t_{bc}, where a, b, c are three unique individuals, not necessarily of the same species. There is some underlying species tree $\mathcal{S}$, which we condition all of our calculations on, implicitly.

Defining shared branch length

To understand the expected shared branch length between two coalescent branches t_{ab} and t_{bc}, we need to take a different approach than in the 4 individual case. Note that there are three possible scenarios for the ordering of our pairs: $(t_{ab} = t_{bc})$, $(t_{ab} < t_{bc})$, $(t_{ab} > t_{bc})$. Let $t_{a,b\cap b,c}$ denote the shared branch length. We can expand the expectation into these three cases, and calculate their values separately.

$$\begin{aligned}
\mathbb{E}(t_{a,b\cap b,c}) &= \mathbb{E}(t_{a,b\cap b,c}|t_{ab} = t_{bc})P(t_{ab} = t_{bc}) + \mathbb{E}(t_{a,b\cap b,c}|t_{ab} < t_{bc})P(t_{ab} < t_{bc}) \\
&\quad + \mathbb{E}(t_{a,b\cap b,c}|t_{ab} > t_{bc})P(t_{ab} > t_{bc}) \\
&= \Big[2\mathbb{E}(t_{ab}|t_{ab} = t_{bc}) - \mathbb{E}(t_{ac}|t_{ab} = t_{bc})\Big]P(t_{ab} = t_{bc}) \\
&\quad + \mathbb{E}(t_{ab}|t_{ab} < t_{bc})P(t_{ab} < t_{bc}) \\
&\quad + \mathbb{E}(t_{bc}|t_{ab} > t_{bc})P(t_{ab} > t_{bc}) \\
&= 2\underbrace{\mathbb{E}(t_{ab}|t_{ab} = t_{bc})P(t_{ab} = t_{bc})}_{\text{Term 3}} - \underbrace{\mathbb{E}(t_{ac}|t_{ab} = t_{bc})P(t_{ab} = t_{bc})}_{\text{Term 4}} \\
&\quad + \underbrace{\mathbb{E}(t_{ab}|t_{ab} < t_{bc})P(t_{ab} < t_{bc})}_{\text{Term 5}} + \underbrace{\mathbb{E}(t_{bc}|t_{ab} > t_{bc})P(t_{ab} > t_{bc})}_{\text{Term 5}}
\end{aligned}$$

Here let us observe that, in a tree of only three individuals, the event $\Big(t_{ab} = t_{bc}\Big)$ is equivalent to events $\Big(t_{ac} < t_{ab}\Big)$, $\Big(t_{ac} < t_{bc}\Big)$ and $\Big(t_{ac} = C_m\Big)$ where C_m denotes the first coalescence event. Using this, we can see that Terms 4 and 5 are symbolically equivalent to calculating the expected coalescence time conditional on being the first event. From this, we will derive the forms for Term 3 and Term 4, which is sufficient to symbolically represent all terms in the expression.

Calculating Term 3: $\mathbb{E}(t_{ab}|t_{ab} = t_{bc})P(t_{ab} = t_{bc})$

Denote C_m to be the first coalescence event among the three pairs $\{a|b, a|c, b|c\}$. Note that when $t_{ab} = t_{bc}$, it must be that $t_{ac} = C_m$.

$$
\begin{aligned}
\text{Term 3} &= \mathbb{E}(t_{ab}|t_{ab}=t_{bc})P(t_{ab}=t_{bc}) \\
&= \int_{D_{abc}}^{\infty} t_{ab}P(t_{ab}|t_{ab}=t_{bc})P(t_{ab}=t_{bc})dt_{ab} \\
&= \sum_{i\in H_{abc}} P(t_{ab}>\tau_i)\int_{\tau_i}^{\tau_{i+1}}\int_{D_{abc}}^{t_{ab}} t_{ab}P(t_{ab}|t_{ac}=C_m)P(t_{ac}=C_m)dt_{ac}dt_{ab} \\
&= \sum_{i\in H_{abc}} P(t_{ab}>\tau_i)\int_{\tau_i}^{\tau_{i+1}} t_{ab}P(t_{ab})\int_{D_{abc}}^{t_{ab}} t_{ab}P(t_{ac}=C_m|t_{ab})dt_{ac}dt_{ab} \\
&= \sum_{i\in H_{abc}} P(t_{ab}>\tau_i)\int_{\tau_i}^{\tau_{i+1}} t_{ab}P(t_{ab})\int_{D_{abc}}^{t_{ab}} \frac{1}{2}P(\text{2p Coal } = t_{ac})dt_{ac}dt_{ab} \\
&= \sum_{i\in H_{abc}} P(t_{ab}>\tau_i)\int_{\tau_i}^{\tau_{i+1}} t_{ab}P(t_{ab})\Bigg[\sum_{k\in H_{abc},k<i} P(\text{2p No Coal } \in (D_{abc},\tau_k)) \\
&\quad \times \int_{\tau_k}^{\tau_{k+1}} \frac{1}{2}P(\text{2p Coal } = t_{ac})dt_{ac} \\
&\quad + P(\text{2p No Coal } \in (D_{abc},\tau_i))\int_{t_i}^{t_{ab}} \frac{1}{2}P(\text{2p Coal } = t_{ac})dt_{ac}\Bigg]dt_{ab} \\
&= \sum_{i\in H_{abc}} P(t_{ab}>\tau_i)\Bigg(\int_{\tau_i}^{\tau_{i+1}} t_{ab}P(t_{ab})dt_{ab}\Bigg[\sum_{k\in H_{abc},k<i} P(\text{2p No Coal } \in (D_{abc},\tau_k)) \\
&\quad \times \int_{\tau_k}^{\tau_{k+1}} \frac{1}{2}P(\text{2p Coal } = t_{ac})dt_{ac}\Bigg] \\
&\quad + P(\text{2p No Coal } \in (D_{abc},\tau_i))\int_{\tau_i}^{\tau_{i+1}} t_{ab}P(t_{ab})\int_{\tau_i}^{t_{ab}} \frac{1}{2}P(\text{2p Coal } = t_{ac})dt_{ac}dt_{ab}\Bigg) \\
&= \sum_{i\in H_{abcd}} P(t_{ab}>\tau_i)\Bigg(\int_{\tau_i}^{\tau_{i+1}} t_{ab}P(t_{ab})dt_{ab}\Bigg[\sum_{k\in H_{abc},k<i} P(\text{2p No Coal } \in (D_{abc},\tau_k)) \\
&\quad \times \frac{1}{2}\int_{\tau_k}^{\tau_{k+1}} P(\text{2p Coal } = t_{ac})dt_{ac}\Bigg] + P(\text{2p No Coal } \in (D_{abc},\tau_i)) \\
&\quad \times \int_{\tau_i}^{\tau_{i+1}} t_{ab}P(t_{ab})\frac{1}{2}\Big[1-P(\text{2p No Coal } \in (\tau_i,t_{ab}))\Big]dt_{ab}\Bigg)
\end{aligned}
$$

$$= \sum_{i \in H_{abcd}} P(\tau_{ab} > \tau_i) \Bigg(\int_{\tau_i}^{\tau_{i+1}} t_{ab} P(t_{ab}) dt_{ab} \Bigg[\sum_{k \in H_{abc}, k<i} P(\text{2p No Coal } \in (D_{abc}, \tau_k)) \times \frac{1}{2} \int_{\tau_k}^{\tau_{k+1}} P(\text{2p Coal } = t_{ac}) dt_{ac} \Bigg] + P(\text{2p No Coal } \in (0, \tau_i)) \Big[\frac{1}{2} \int_{\tau_i}^{\tau_{i+1}} t_{ab} P(t_{ab}) dt_{ab} - \frac{1}{2} \int_{\tau_i}^{\tau_{i+1}} t_{ab} P(t_{ab}) P(\text{2p No Coal } \in (\tau_i, t_{ab})) dt_{ab} \Big] \Bigg)$$

Term 4: $\mathbb{E}(t_{ac}|t_{ab} = t_{bc})P(t_{ab} = t_{bc})$

$$\begin{aligned}
\text{Term 4} &= \mathbb{E}(t_{ac}|t_{ab} = t_{bc})P(t_{ab} = t_{bc}) \\
&= \int_{D_{abc}}^{\infty} t_{ac} P(t_{ac}|t_{ab} = t_{bc}) P(t_{ab} = t_{bc}) dt_{ac} \\
&= \int_{D_{abc}}^{\infty} t_{ac} P(t_{ac} = C_m|t_{ac}) P(t_{ac}) dt_{ac} \\
&= \sum_{i \in H_{abc}} P(t_{ac} > \tau_i) \int_{\tau_i}^{\tau_{i+1}} P(t_{ac} = C_m|t_{ac}) P(t_{ac}) dt_{ac} \\
&= \sum_{i \in H_{abc}} P(t_{ac} > \tau_i) \int_{\tau_i}^{\tau_{i+1}} P(\text{2p No Coal} \in (D_{abc}, t_{ac})) P(t_{ac}) dt_{ac} \\
&= \sum_{i \in H_{abc}} P(t_{ac} > \tau_i) P(\text{2p No Coal } \in (D_{abc}, \tau_i)) \\
&\quad \times \int_{t_i}^{\tau_{i+1}} P(t_{ac}) P(\text{2p No Coal} \in (\tau_i, t_{ac})) dt_{ac} \\
&= \sum_{i \in H_{abc}} P(t_{ac} > \tau_i) P(\text{2p No Coal } \in (D_{abc}, \tau_i)) \int_{t_i}^{\tau_{i+1}} \frac{1}{2\eta_i} e^{\frac{-3(t_{ac} - \tau_i)}{2\eta_i}} dt_{ac}
\end{aligned}$$

Covariance calculation, 4 individuals (same species)

In this section, we will describe the equations needed to derive a key quantity in the covariance between pairs of coalescence events, where we explicitly assume all

individuals have failed to coalesce with one another until that point, and there are 4 unique individuals.

Let us suppose we are interested in calculating $\text{Cov}(T_{ab}, T_{cd}|\mathcal{S})$ for some individuals labeled a, b, c, d in our set of species. Note we may have piecewise constant population sizes within the ancestral species, which we get from species tree $\mathcal{S}$. The formula for the covariance of our two random variables is as follows:

$$\text{Cov}(T_{ab}, T_{cd}|\mathcal{S}) = \mathbb{E}(T_{ab}T_{cd}|\mathcal{S}) - \mathbb{E}(T_{ab}|S)\mathbb{E}(T_{cd}|\mathcal{S})$$

The second part of the right hand side is a simple exercise in the multispecies coalescent framework, which has been presented in the main text. The challenge comes from calculating $\mathbb{E}(T_{ab}T_{cd}|\mathcal{S})$, which is what we will focus on in this section. Specifically, we focus on the component when all lineages have already survived to a common ancestral population. Denote the time of this event to be D_{abcd} the divergence time of all 4 individuals/species. Further, let us use the symbol $> D_{abcd}$ to indicate all lineages being more ancient than this time.

To begin, note that for 4 individuals there will be 3 distinct coalescence events, with the last (most ancient) being the TMRCA of the set. Using this, we are interested in evaluating

$$\mathbb{E}(T_{ab}T_{cd}|S, > D_{abcd}) = \int_{D_{abcd}} t_{cd}P(t_{cd}|S, > D_{abcd}) \int_{D_{abcd}} t_{ab}P(t_{ab}|t_{cd}, \mathcal{S}, > D_{abcd})dt_{ab}dt_{cd}$$

by conditioning on all of the possible orderings of coalescence events.

Here is what we define as the canonical 6 orderings of events:

C1. T_{ab} is the first coalescent event.
C2. T_{ab} is the second event, T_{cd} is the third.
C3. $T_{ab} = T_{cd}$ as the third coalescent event.
C4. T_{cd} is the second event, T_{ab} is the third.
C5. T_{cd} is the first event, T_{ab} is the second.
C6. T_{cd} is the first event, T_{ab} is the third.

From here on out, we will use C_i for $i \in \{1, ..., 6\}$ to denote each case.

For any split on the species tree, here the split denoted by time D_{abcd}, let H_{abcd} be the sequence constant population size intervals back in time that trace the single ancestry from time D_{abcd} back. For instance $H_{abcd}[1] = (\tau_1, \eta_1)$, where $\tau_1 = D_{abcd}$ and η_1 represents the population size of this ancient population. So we will denote

$i \in H_{abcd}$ to be the i^{th} branch segment in the 'history' of this ancient species.

So we can begin our equation. For notation's sake, we drop the $> D_{abcd}$ notation from the following calculations as a space saving measure, but implicitly assume it is present. It will be made clear when this is no longer the case.

$$
\begin{aligned}
\mathbb{E}(T_{ab}T_{cd}|S) &= \int_{D_{cd}} t_{cd}P(t_{cd}|S)\int_{D_{abcd}} t_{ab}P(t_{ab}|t_{cd},S)dt_{ab}dt_{cd} \\
&= \sum_{i\in H_{abcd}} P(T_{cd}>\tau_i)\int_{\tau_i}^{\tau_{i+1}} t_{cd}P(t_{cd}|S)\int_{D_{abcd}}^{t_{cd}} t_{ab}P(t_{ab}|t_{cd},S)dt_{ab}dt_{cd} \\
&= \sum_{i\in H_{abcd}} P(T_{cd}>\tau_i)\Big[\int_{\tau_i}^{\tau_{i+1}} t_{cd}P(t_{cd}|S) \\
&\quad \times \int_{D_{abcd}}^{\infty} t_{ab}P(t_{ab}|t_{cd},S)\mathbb{1}(t_{ab}<t_{cd})dt_{ab}t_{cd} \\
&\quad + \mathbb{1}(T_{ab}=T_{cd})\int_{\tau_i}^{\tau_{i+1}} t_{cd}^2P(t_{cd}|S)P(t_{ab}=t_{cd}|t_{cd},S)dt_{cd} \\
&\quad + \int_{\tau_i}^{\tau_{i+1}} t_{cd}P(t_{cd}|S)\int_{t_{cd}}^{\infty} t_{ab}P(t_{ab}|t_{cd},S)\mathbb{1}(t_{ab}>t_{cd})dt_{ab}dt_{cd}\Big] \\
&= \sum_{i\in H_{abcd}} P(T_{cd}>\tau_i) \\
&\quad \times \Big[\int_{\tau_i}^{\tau_{i+1}} t_{cd}P(t_{cd}|S)\int_{D_{abcd}}^{t_{cd}} t_{ab}P(t_{ab}|t_{cd},[T_{ab}\ 1^{st}\text{ event}],S) \\
&\quad \times P([T_{ab}\ 1^{st}\text{ event}]|t_{cd},S)dt_{ab}dt_{cd} \\
&\quad + \int_{\tau_i}^{\tau_{i+1}} t_{cd}P(t_{cd}|S) \\
&\quad \times \int_{D_{abcd}}^{t_{cd}} t_{ab}P(t_{ab}|t_{cd},[T_{ac,ad,bc,bd}\ 1^{st}\text{ event}],[T_{ab}\ 2^{nd}\text{ event}],S) \\
&\quad \times P([T_{ab}\ 2^{nd}\text{ event}]|t_{cd},[T_{ac,ad,bc,bd}\ 1^{st}\text{ event}],S) \\
&\quad \times P([T_{ac,ad,bc,bd}\ 1^{st}\text{ event}]|t_{cd},S)dt_{ab}dt_{cd} \\
&\quad + \int_{\tau_i}^{\tau_{i+1}} {t_{cd}}^2P(t_{cd}|S)\int_{t_0}^{t_{cd}} \frac{4}{5}P(t_{ac,ad,bc,bd}=\zeta_1|t_{cd},S) \\
&\quad \times \int_{\zeta_1}^{t_{cd}} \frac{1}{2}P(\text{Coal event not } T_{ab} \text{ or } T_{cd}=\zeta_2|\zeta_1,t_{cd},S)d\zeta_2 d\zeta_1 dt_{cd}
\end{aligned}
$$

$$
\begin{aligned}
&+ \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) \int_{t_0}^{t_{cd}} \frac{4}{5} P(T_{ac,ad,bc,bd} \ 1^{st} \text{ event} = \zeta_1 | t_{cd}, S) \\
&\times P(\text{2p No Coal in } (\zeta_1, t_{cd}) | t_{cd}, S) d\zeta_1 \int_{t_{cd}}^{\infty} t_{ab} P(t_{ab}|t_{cd}, S) dt_{ab} dt_{cd} \\
&+ \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) P(\text{5p No Coal in } (t_0, t_{cd}) | t_{cd}, S) \\
&\times \int_{t_{cd}}^{\infty} \frac{1}{3} t_{ab} P(\text{3p coal} = t_{ab} | t_{cd}, S) dt_{ab} dt_{cd} \\
&+ \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) P(\text{5p No Coal in } (t_0, t_{cd}) | t_{cd}, S) \int_{t_{cd}}^{\infty} \frac{2}{3} P(\text{3p Coal} = \zeta | t_{cd}, S) \\
&\times \int_{\zeta}^{\infty} t_{ab} P(t_{ab}|\zeta, S) dt_{ab} d\zeta dt_{cd} \Bigg]
\end{aligned}
$$

Where inside the bracket of the last equation is a sum of 6 quantities which correspond directly to each of the 6 cases presented above, in order. Also, note we use notation, ζ, to represent coalescence events not equal to t_{ab} or t_{cd}.

We will go through each one of these 6 equations, and evaluate the integrals, noting that these are all conditional on the event that T_{cd} occurs in the interval (τ_i, τ_{i+1}).

Case 1: T_{ab} is the first coalescent event.

$$
\begin{aligned}
C1(i) &= \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) \int_{D_{abcd}}^{t_{cd}} t_{ab} P(t_{ab}|t_{cd}, [T_{ab}\ 1^{st} \text{ event}], S) \\
&\quad \times P([T_{ab}\ 1^{st} \text{ event}]|t_{cd}, S) dt_{ab} dt_{cd} \\
&= \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) \Big[\sum_{k<i\in H_{abcd}} \Big(P(5\text{p No Coal in } (t_0, \tau_k)) \\
&\quad \times \int_{\tau_k}^{\tau_{k+1}} t_{ab} P(t_{ab}|t_{cd}, [T_{ab}\ 1^{st} \text{ event}], S) dt_{ab} \Big) \\
&\quad + P(5\text{p No Coal in } (t_0, \tau_i)) \int_{\tau_i}^{t_{cd}} t_{ab} P(t_{ab}|t_{cd}, [T_{ab}\ 1^{st} \text{ event}], S) dt_{ab} \Big] dt_{cd} \\
&= \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) \Big[\sum_{k<i\in H_{abcd}} \Big(P(5\text{p No Coal in } (t_0, \tau_k)) \\
&\quad \times \int_{\tau_k}^{\tau_{k+1}} t_{ab} \frac{1}{5} \frac{5}{2\eta_k} e^{-\frac{5(t_{ab}-\tau_k)}{2\eta_k}} dt_{ab} \Big) \\
&\quad + P(5\text{p No Coal in } (t_0, \tau_i)) \int_{\tau_i}^{t_{cd}} t_{ab} \frac{1}{5} \frac{5}{2\eta_i} e^{-\frac{5(t_{ab}-\tau_i)}{2\eta_i}} dt_{ab} \Big] dt_{cd} \\
&= \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) dt_{cd} \\
&\quad \times \Big[\sum_{k<i\in H_{abcd}} \Big(P(5\text{p No Coal in } (t_0, \tau_k)) \int_{\tau_k}^{\tau_{k+1}} t_{ab} \frac{1}{5} \frac{5}{2\eta_k} e^{-\frac{5(t_{ab}-\tau_k)}{2\eta_k}} dt_{ab} \Big) \Big] \\
&\quad + P(5\text{p No Coal in } (t_0, \tau_i)) \\
&\quad \times \underbrace{\int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) \int_{\tau_i}^{t_{cd}} t_{ab} \frac{1}{5} \frac{5}{2\eta_i} e^{-\frac{5(t_{ab}-\tau_i)}{2\eta_i}} dt_{ab} dt_{cd}}_{C1A(i)}
\end{aligned}
$$

The first quantity is separated into easy to evaluate integrals for every interval in the species tree. Now we look specifically at the remaining double integral denoted C1A(i).

$$
\begin{aligned}
C1A(i) &= \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) \frac{1}{5}\Big[-(\frac{2}{5}\eta_i + t_{cd}) e^{-\frac{5(t_{cd}-\tau_i)}{2\eta_i}} + \frac{2}{5}\eta_i + \tau_i \Big] dt_{cd} \\
&= -\frac{2}{25}\eta_i \int_{\tau_i}^{\tau_{i+1}} t_{cd} \frac{1}{2\eta_i} e^{-\frac{6(t_{cd}-\tau_i)}{2\eta_i}} dt_{cd} - \frac{1}{5}\int \tau_i{}^{\tau_{i+1}} t_{cd}^2 \frac{1}{2\eta_i} e^{-\frac{6(t_{cd}-\tau_i)}{2\eta_i}} dt_{cd} \\
&\quad (\frac{2}{5}\eta_i + \tau_i) \int_{\tau_i}^{\tau_{i+1}} t_{cd} \frac{1}{2\eta_i} e^{-\frac{t_{cd}-\tau_i}{2\eta_i}} dt_{cd}
\end{aligned}
$$

Which is a sum of easily evaluatable integrals.

Case 2: T_{ab} is the second event, T_{cd} is the third.

$$
\begin{aligned}
C2(i) &= \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) \int_{D_{abcd}}^{t_{cd}} t_{ab} P(t_{ab}|t_{cd}, [T_{ac,ad,bc,bd}\ 1^{st} \text{ event}], [T_{ab}\ 2^{nd} \text{ event}], S) \\
&\quad \times P([T_{ab}\ 2^{nd} \text{ event}]|t_{cd}, [T_{ac,ad,bc,bd}\ 1^{st} \text{ event}], S) \\
&\quad \times P([T_{ac,ad,bc,bd}\ 1^{st} \text{ event}]|t_{cd}, S) dt_{ab} dt_{cd} \\
&= \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) \Bigg[\sum_{k=0}^{i-1} \Big(P(\text{5p No Coal in } (t_0, \tau_k)) \\
&\quad \times \int_{\tau_k}^{\tau_{k+1}} \frac{4}{5} P(\text{5p Coal } = \zeta|t_{cd}, S) \\
&\quad \times \int_{\tau}^{t_{cd}} t_{ab} \frac{1}{2} P(\text{2p coal } = t_{ab}|t_{cd}, \tau, S) dt_{ab} d\zeta \Big) \Bigg] dt_{cd} \\
&\quad + P(\text{5p No Coal in } (t_0, \tau_i)) \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) \\
&\quad \times \Bigg[\int_{\tau_i}^{t_{cd}} \frac{4}{5} P(\text{5p Coal } = \zeta|t_{cd}, S) \\
&\quad \times \int_{\tau}^{t_{cd}} t_{ab} \frac{1}{2} P(\text{2p Coal } = t_{ab}|\zeta, t_{cd}, S) dt_{ab} d\tau \Bigg] dt_{cd} \\
&= C2A(i) + P(\text{5p No Coal in } (t_0, \tau_i)) C2B(i)
\end{aligned}
$$

So let's look at each of these two triple integrals, starting with C2A(i), which is the event that T_{ab} occurs more recently than the interval that contains T_{cd} (before τ_i).

$$
\begin{aligned}
C2A(i) = &\int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) \Bigg[\sum_{k=0}^{i-1} \Big(P(\text{5p No Coal in } (t_0, \tau_k)) \\
&\quad \times \int_{\tau_k}^{\tau_{k+1}} \frac{4}{5} P(\text{5p Coal } = \zeta | t_{cd}, S) \\
&\quad \times \int_{\zeta}^{t_{cd}} t_{ab} \frac{1}{2} P(\text{2p coal } = t_{ab} | t_{cd}, \tau, S) dt_{ab} d\zeta \Big) \Bigg] dt_{cd} \\
= &\int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) \Bigg[\sum_{k=0}^{i-1} \Big(P(\text{5p No Coal in } (t_0, \tau_k)) \\
&\quad \times \int_{\tau_k}^{\tau_{k+1}} \frac{4}{5} P(\text{5p Coal } = \zeta | t_{cd}, S) \\
&\quad \times \Big(\int_{\zeta}^{\tau_{k+1}} t_{ab} \frac{1}{2} P(\text{2p Coal } = t_{ab} | \zeta, t_{cd}, S) dt_{ab} d\zeta \\
&\quad + P(\text{2p No Coal in}(\zeta, \tau_{k+1})) \int_{\tau_{k+1}}^{t_{cd}} t_{ab} \frac{1}{2} P(\text{2p Coal } = t_{ab} | \tau_{k+1}, t_{cd}, S) \Big) \Bigg] dt_{cd} \\
= &\int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) dt_{cd} \Bigg[\sum_{k=0}^{i-1} P(\text{5p No Coal in } (t_0, \tau_k)) \\
&\quad \times \underbrace{\int_{\tau_k}^{\tau_{k+1}} \frac{4}{5} P(\text{5p Coal } = \tau | t_{cd}, S)}_{C2A(i)_1(k)} \\
&\quad \underbrace{\times \int_{\zeta}^{\tau_{k+1}} t_{ab} \frac{1}{2} P(\text{2p Coal } = t_{ab} | \zeta, t_{cd}, S) dt_{ab} d\zeta \Bigg]}_{C2A(i)_1(k) \text{ con't}} \\
&+ \sum_{k=0}^{i-1} P(\text{5p No Coal in } (t_0, \tau_k) \int_{\tau_k}^{\tau_{k+1}} \frac{4}{5} P(\text{5p Coal } = \zeta | t_{cd}, S) \\
&\quad \times P(\text{2p No Coal in } (\zeta, \tau_{k+1})) d\zeta \\
&\quad \times \underbrace{\int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) \int_{\tau_{k+1}}^{t_{cd}} P(\text{2p Coal } = t_{ab} | \tau_{k+1}, t_{cd}, S) dt_{ab} dt_{cd}}_{C2A(i)_2(k)}
\end{aligned}
$$

$$
\begin{aligned}
&= \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) dt_{cd} \Big[\sum_{k=0}^{i-1} P(\text{5p No Coal in } (t_0, \tau_k)) \times C2A(i)_1(k) \Big] \\
&+ \sum_{k=0}^{i-1} P(\text{5p No Coal in } (t_0, \tau_k) \int_{\tau_k}^{\tau_{k+1}} \frac{4}{5} P(\text{5p Coal } = \tau | t_{cd}, S) t_{ab} \\
&\quad \times \frac{1}{2} P(\text{2p No Coal in } (\zeta, \tau_{k+1})) d\zeta \times C2A(i)_2(k)
\end{aligned}
$$

So let's evaluate the double integrals $C2A(i)_1(k)$ and $C2A(i)_2(k)$ next.

$$
\begin{aligned}
C2A(i)_1(k) &= \int_{\tau_k}^{\tau_{k+1}} \frac{4}{5} P(\text{5p Coal } = \zeta | t_{cd}, S) \\
&\quad \times \int_{\zeta}^{\tau_{k+1}} t_{ab} \frac{1}{2} P(\text{2p Coal } = t_{ab} | \tau, t_{cd}, S) dt_{ab} d\zeta \Big] \\
&= \int_{\tau_k}^{\tau_{k+1}} \frac{4}{5} P(\text{5p Coal } = \zeta | t_{cd}, S) \Big[\frac{1}{4} \big(-(\eta_k + \tau_{k+1}) e^{-\frac{2(\tau_{k+1} - \zeta)}{2\eta_k}} \\
&\quad + \eta_k + \zeta \big) \Big] d\zeta \\
&= -\frac{1}{5} (\eta_i + \tau_{k+1}) \int_{\tau_k}^{\tau_{k+1}} P(\text{5p Coal } = \zeta | t_{cd}, S) e^{-\frac{2(\tau_{k+1} - \zeta)}{2\eta_k}} d\zeta \\
&\quad + \frac{1}{5} \eta_k \int_{\tau_k}^{\tau_{k+1}} P(\text{5p Coal } = \zeta | t_{cd}, S) d\zeta \\
&\quad + \frac{1}{5} \int_{\tau_k}^{\tau_{k+1}} \zeta P(\text{5p Coal } = \zeta | t_{cd}, S) d\zeta
\end{aligned}
$$

which are simple to evaluate integrals for each interval.

$$
\begin{aligned}
C2A(i)_2(k) &= \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) \int_{\tau_{k+1}}^{t_{cd}} t_{ab} \frac{1}{2} P(\text{2p Coal } = t_{ab}|\tau_{k+1}, t_{cd}, S) dt_{ab} dt_{cd} \\
&= \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) \Bigg[\int_{\tau_{k+1}}^{\tau_i} t_{ab} \frac{1}{2} P(\text{2p Coal } = t_{ab}|\tau_{k+1}, \tau_i, S) dt_{ab} \\
&\quad + P(\text{2p No Coal in } (\tau_{k+1}, \tau_i)) \\
&\quad \times \int_{\tau_i}^{t_{cd}} \frac{t_{ab}}{2} P(\text{2p Coal } = t_{ab}|\tau_i, t_{cd}, S) dt_{ab} \Bigg] dt_{cd} \\
&= \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) dt_{cd} \Bigg(\sum_{j=k+1}^{i-1} P(\text{2p No Coal in } (\tau_{k+1}, \tau_j)) \\
&\quad \times \int_{\tau_j}^{\tau_{j+1}} t_{ab} \frac{1}{2} P(\text{2p Coal } = t_{ab}|\tau_j, \tau_{j+1}, S) dt_{ab} \Bigg) \\
&\quad + P(\text{2p No Coal in } (\tau_{k+1}, \tau_i)) \\
&\quad \times \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) \frac{1}{2} \big(-(\eta_i + t_{cd}) e^{-\frac{2(t_{cd} - \tau_i)}{2\eta_i}} + \eta_i + \tau_i \big) dt_{cd} \\
&= \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) dt_{cd} \\
&\quad \times \Bigg(\sum_{j=k+1}^{i-1} P(\text{2p No Coal in } (\tau_{k+1}, \tau_j)) \\
&\quad \times \int_{\tau_j}^{\tau_{j+1}} t_{ab} \frac{1}{2} P(\text{2p Coal } = t_{ab}|\tau_j, \tau_{j+1}, S) dt_{ab} \Bigg) \\
&\quad + \frac{1}{2} P(\text{2p No Coal in } (\tau_{k+1}, \tau_i)) \Bigg[-\eta_i \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) e^{-\frac{2(t_{cd} - \tau_i)}{2\eta_i}} dt_{cd} \\
&\quad - \int_{\tau_i}^{\tau_{i+1}} t_{cd}^2 P(t_{cd}|S) e^{-\frac{2(t_{cd} - \tau_i)}{2\eta_i}} dt_{cd} + (\eta_i + \tau_i) \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) dt_{cd} \Bigg]
\end{aligned}
$$

So this finishes Case 2A, I currently leave the single integrals up to the reader. See the notation section for assistance.

Let's look at $C2B(i)$ now. Note this is the event where all three coalescent events occur in (τ_i, τ_{i+1}) with T_{ab} as the second event, and T_{cd} is the last.

$$
\begin{aligned}
C2B(i) &= \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) \Bigg[\int_{\tau_i}^{t_{cd}} \frac{4}{5} P(\text{5p Coal } = \zeta | t_{cd}, S) \\
&\quad \times \int_{\zeta}^{t_{cd}} t_{ab} \frac{1}{2} P(\text{2p Coal } = t_{ab} | \zeta, t_{cd}, S) dt_{ab} d\zeta \Bigg] dt_{cd} \\
&= \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) \Bigg[\int_{\tau_i}^{t_{cd}} \frac{4}{5} P(\text{5p Coal } = \zeta | t_{cd}, S) \\
&\quad \times \frac{1}{2} \Big(-(\eta_i + t_{cd}) e^{-2\frac{(t_{cd}-\zeta)}{2\eta_i}} + \eta_i + \tau \Big) \Bigg] d\zeta dt_{cd} \\
&= \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) \Bigg[-\frac{2}{5}(\eta_i + t_{cd}) \frac{5}{3} \Big(e^{-2\frac{(t_{cd}-\tau_i)}{2\eta_i}} - e^{-5\frac{(t_{cd}-\tau_i)}{2\eta_i}} \Big) \\
&\quad + \frac{2}{5}\eta_i \big(1 - e^{-5\frac{(t_{cd}-\tau_i)}{2\eta_i}}\big) + \frac{2}{5} \Big(-(\frac{2}{5}\eta_i + t_{cd}) e^{-5\frac{(t_{cd}-\tau_i)}{2\eta_i}} + \frac{2}{5}\eta_i + \tau_i \Big) \Bigg] dt_{cd} \\
&= -\frac{2}{3}\eta_i \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) e^{-2\frac{(t_{cd}-\tau_i)}{2\eta_i}} dt_{cd} - \frac{2}{3} \int_{\tau_i}^{\tau_{i+1}} t_{cd}^2 P(t_{cd}|S) e^{-2\frac{(t_{cd}-\tau_i)}{2\eta_i}} dt_{cd} \\
&\quad + \frac{8}{75}\eta_i \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) e^{-5\frac{(t_{cd}-\tau_i)}{2\eta_i}} dt_{cd} \\
&\quad + \frac{4}{15} \int_{\tau_i}^{\tau_{i+1}} t_{cd}^2 P(t_{cd}|S) e^{-2\frac{(t_{cd}-\tau_i)}{2\eta_i}} dt_{cd} \\
&\quad + \frac{2}{5} \Big(\frac{2}{5}\eta_i + \tau_i \Big) \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) dt_{cd}
\end{aligned}
$$

Case 3: $T_{ab} = T_{cd}$ as the last coalescent event.

$$
\begin{aligned}
C3(i) = &\int_{\tau_i}^{\tau_{i+1}} {t_{cd}}^2 P(t_{cd}|S) \int_{t_0}^{t_{cd}} \frac{4}{5} P(t_{ac,ad,bc,bd} = \zeta_1 | t_{cd}, S) \\
&\times \int_{\zeta_1}^{t_{cd}} \frac{1}{2} P(\text{Coal event not } T_{ab} \text{ or } T_{cd} = \zeta_2 | \zeta_1, t_{cd}, S) d\zeta_2 d\zeta_1 dt_{cd} \\
= &\int_{\tau_i}^{\tau_{i+1}} {t_{cd}}^2 P(t_{cd}|S) \Bigg[\sum_{k=0}^{i-1} P(\text{5p No Coal in } (t_0, \tau_k)) \\
&\times \int_{\tau_k}^{\tau_{k+1}} \frac{4}{5} P(\text{5p Coal } = \zeta_1 | t_{cd}, S) \int_{\zeta_1}^{t_{cd}} \frac{1}{2} P(\text{2p Coal } = \zeta_2 | t_{cd}, \zeta_1, S) d\zeta_2 d\zeta_1 \\
&+ P(\text{5p No Coal in } (t_0, \tau_i)) \int_{\tau_i}^{t_{cd}} \frac{4}{5} P(\text{5p Coal } = \zeta_1 | t_{cd}, S) \\
&\times \int_{\zeta_1}^{t_{cd}} \frac{1}{2} P(\text{2p Coal } = \zeta_2 | t_{cd}, \zeta_1, S) d\zeta_2 d\zeta_1 \Bigg] dt_{cd} \\
= &\int_{\tau_i}^{\tau_{i+1}} {t_{cd}}^2 P(t_{cd}|S) \Bigg[\sum_{k=0}^{i-1} P(\text{5p No Coal in } (t_0, \tau_k)) \\
&\times \int_{\tau_k}^{\tau_{k+1}} \frac{4}{5} P(\text{5p Coal } = \zeta_1 | t_{cd}, S) \\
&\times \Bigg(\int_{\zeta_1}^{\tau_{k+1}} \frac{1}{2} P(\text{2p Coal } = \zeta_2 | t_{cd}, \zeta_1, S) d\zeta_2 \\
&+ P(\text{2p No Coal in } (\zeta_1, \tau_{k+1})) \int_{\tau_{k+1}}^{t_{cd}} \frac{1}{2} P(\text{2p Coal } = \zeta_2 | t_{cd}, \tau_{k+1}, S) d\zeta_2 \Bigg) d\zeta_1 \Bigg] dt_{cd} \\
&+ P(\text{5p No Coal in } (t_0, \tau_i)) \\
&\times \underbrace{\int_{\tau_i}^{\tau_{i+1}} {t_{cd}}^2 P(t_{cd}|S) \int_{\tau_i}^{t_{cd}} \frac{4}{5} P(\text{5p Coal } = \zeta_1 | t_{cd}, S)}_{C3A(i)} \\
&\times \underbrace{\int_{\zeta_1}^{t_{cd}} \frac{1}{2} P(\text{2p Coal } = \zeta_2 | t_{cd}, \zeta_1, S) d\zeta_2 d\zeta_1 dt_{cd}}_{C3A(i) \text{ con't}}
\end{aligned}
$$

$$
\begin{aligned}
&= P(\text{5p No Coal in } (t_0, \tau_i)) \times C3A(i) \\
&\quad + \int_{\tau_i}^{\tau_{i+1}} {t_{cd}}^2 P(t_{cd}|S) dt_{cd} \Bigg[\sum_{k=0}^{i-1} P(\text{5p No Coal in } (t_0, \tau_k)) \\
&\quad \times \underbrace{\int_{\tau_k}^{\tau_{k+1}} \frac{4}{5} P(\text{5p Coal } = \zeta_1 | t_{cd}, S) \int_{\zeta_1}^{\tau_{k+1}} \frac{1}{2} P(\text{2p Coal } = \zeta_2 | t_{cd}, \zeta_1, S) d\zeta_2 d\zeta_1}_{C3B(k)} \Bigg] \\
&\quad + \Bigg[\sum_{k=0}^{i-1} P(\text{5p No Coal in } (t_0, \tau_k)) \\
&\quad \times \int_{\tau_k}^{\tau_{k+1}} \frac{4}{5} P(\text{5p Coal } = \zeta_1 | t_{cd}, S) P(\text{2p No Coal in } (\zeta_1, \tau_{k+1})) d\zeta_1 \\
&\quad \times \underbrace{\int_{\tau_i}^{\tau_{i+1}} {t_{cd}}^2 P(t_{cd}|S) \int_{\tau_{k+1}}^{t_{cd}} \frac{1}{2} P(\text{2p Coal } = \zeta_2 | t_{cd}, \tau_{k+1}, S) d\zeta_2 dt_{cd}}_{C3C(i,k)} \Bigg] \\
&= P(\text{5p No Coal in } (t_0, \tau_i)) \times C3A(i) \\
&+ \int_{\tau_i}^{\tau_{i+1}} {t_{cd}}^2 P(t_{cd}|S) dt_{cd} \Bigg[\sum_{k=0}^{i-1} P(\text{5p No Coal in } (t_0, \tau_k)) \times C3B(k) \Bigg] \\
&+ \Bigg[\sum_{k=0}^{i-1} P(\text{5p No Coal in } (t_0, \tau_k)) \int_{\tau_k}^{\tau_{k+1}} \frac{4}{5} P(\text{5p Coal } = \zeta_1 | t_{cd}, S) \\
&\quad \times P(\text{2p No Coal in } (\zeta_1, \tau_{k+1})) d\zeta_1 \times C3C(i,k) \Bigg]
\end{aligned}
$$

So let's evaluate $C3A(i)$, $C3B(k)$ and $C3C(i,k)$. The rest of the equation are easy to evaluate single integrals (with a constant population size).

$$
\begin{aligned}
C3A(i) &= \int_{\tau_i}^{\tau_{i+1}} {t_{cd}}^2 P(t_{cd}|S) \int_{\tau_i}^{t_{cd}} \frac{4}{5} P(\text{5p Coal } = \zeta_1 | t_{cd}, S) \\
&\quad \times \int_{\zeta_1}^{t_{cd}} \frac{1}{2} P(\text{2p Coal } = \zeta_2 | t_{cd}, \zeta_1, S) d\zeta_2 d\zeta_1 dt_{cd} \\
&= \int_{\tau_i}^{\tau_{i+1}} {t_{cd}}^2 P(t_{cd}|S) \int_{\tau_i}^{t_{cd}} \frac{4}{5} P(\text{5p Coal } = \zeta_1 | t_{cd}, S) \left[\frac{1}{2} \left(1 - e^{-\frac{2(t_{cd}-\zeta_1)}{2\eta_i}} \right) \right] d\zeta_1 dt_{cd} \\
&= \int_{\tau_i}^{\tau_{i+1}} {t_{cd}}^2 P(t_{cd}|S) \Bigg[\int_{\tau_i}^{t_{cd}} \frac{2}{5} P(\text{5p Coal } = \zeta_1 | t_{cd}, S) d\zeta_1 \\
&\quad - \int_{\tau_i}^{t_{cd}} \frac{2}{5} P(\text{5p Coal } = \zeta_1 | t_{cd}, S) e^{-\frac{2(t_{cd}-\zeta_1)}{2\eta_i}} d\zeta_1 \Bigg] dt_{cd} \\
&= \int_{\tau_i}^{\tau_{i+1}} {t_{cd}}^2 P(t_{cd}|S) \left[\frac{2}{5} \left(1 - e^{-\frac{5(t_{cd}-\tau_i)}{2\eta_i}} \right) - \frac{2}{3} \left(e^{-\frac{2(t_{cd}-\tau_i)}{2\eta_i}} - e^{-\frac{5(t_{cd}-\tau_i)}{2\eta_i}} \right) \right] dt_{cd} \\
&= \frac{2}{5} \int_{\tau_i}^{\tau_{i+1}} {t_{cd}}^2 P(t_{cd}|S) dt_{cd} - \frac{2}{3} \int_{\tau_i}^{\tau_{i+1}} {t_{cd}}^2 P(t_{cd}|S) e^{-\frac{2(t_{cd}-\tau_i)}{2\eta_i}} dt_{cd} \\
&\quad + \frac{4}{15} \int_{\tau_i}^{\tau_{i+1}} {t_{cd}}^2 P(t_{cd}|S) e^{-\frac{5(t_{cd}-\tau_i)}{2\eta_i}} dt_{cd}
\end{aligned}
$$

Now for $C3B(k)$:

$$
\begin{aligned}
C3B(k) &= \int_{\tau_k}^{\tau_{k+1}} \frac{4}{5} P(\text{5p Coal } = \zeta_1 | t_{cd}, S) \int_{\zeta_1}^{\tau_{k+1}} \frac{1}{2} P(\text{2p Coal } = \zeta_2 | t_{cd}, \zeta_1, S) d\zeta_2 d\zeta_1 \\
&= \frac{2}{5} \int_{\tau_k}^{\tau_{k+1}} P(\text{5p Coal } = \zeta_1 | t_{cd}, S) \left[1 - e^{-\frac{5(\tau_{k+1}-\zeta_1)}{2\eta_k}} \right] d\zeta_1 \\
&= \frac{2}{5} \int_{\tau_k}^{\tau_{k+1}} P(\text{5p Coal } = \zeta_1 | t_{cd}, S) d\zeta_1 \\
&\quad - \frac{2}{5} \int_{\tau_k}^{\tau_{k+1}} P(\text{5p Coal } = \zeta_1 | t_{cd}, S) e^{-\frac{5(\tau_{k+1}-\zeta_1)}{2\eta_k}} d\zeta_1
\end{aligned}
$$

Now for $C3C(i,k)$:

$$
\begin{aligned}
C3C(i,k) &= \int_{\tau_i}^{\tau_{i+1}} {t_{cd}}^2 P(t_{cd}|S) \int_{\tau_{k+1}}^{t_{cd}} \frac{1}{2} P(\text{2p Coal } = \zeta_2 | t_{cd}, \tau_{k+1}, S) d\zeta_2 dt_{cd} \\
&= \int_{\tau_i}^{\tau_{i+1}} {t_{cd}}^2 P(t_{cd}|S) dt_{cd} \int_{\tau_{k+1}}^{\tau_i} \frac{1}{2} P(\text{2p Coal } = \zeta_2 | \tau_{k+1}, \tau_i, S) d\zeta_2 \\
&\quad + P(\text{2p No Coal in } (\tau_{k+1}, \tau_i)) \\
&\quad \times \int_{\tau_i}^{\tau_{i+1}} {t_{cd}}^2 P(t_{cd}|S) \int_{\tau_i}^{t_{cd}} \frac{1}{2} P(\text{2p Coal } = \zeta_2 | \tau_i, t_{cd}, S) d\zeta_2 dt_{cd} \\
&= \int_{\tau_i}^{\tau_{i+1}} {t_{cd}}^2 P(t_{cd}|S) dt_{cd} \Bigg[\sum_{j=k+1}^{i-1} P(\text{2p No Coal in } (\tau_{k+1}, \tau_j)) \\
&\quad \times \int_{\tau_j}^{\tau_{j+1}} \frac{1}{2} P(\text{2p Coal } = \zeta_2 | \tau_j, \tau_{j+1}, S) d\zeta_2 \Bigg] \\
&\quad + P(\text{2p No Coal in } (\tau_{k+1}, \tau_i)) \Bigg[\int_{\tau_i}^{\tau_{i+1}} {t_{cd}}^2 P(t_{cd}|S) dt_{cd} \\
&\quad - \int_{\tau_i}^{\tau_{i+1}} {t_{cd}}^2 P(t_{cd}|S) e^{-\frac{2(t_{cd}-\tau_i)}{2\eta_i}} dt_{cd} \Bigg]
\end{aligned}
$$

Case 4: T_{cd} is the second event, T_{ab} is the third.

$$
\begin{aligned}
C4(i) &= \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) \int_{t_0}^{t_{cd}} \frac{4}{5} P(T_{ac,ad,bc,bd} \ 1^{st} \text{ event} = \zeta_1 | t_{cd}, S) \\
&\quad \times P(\text{2p No Coal in } (\zeta_1, t_{cd}) | t_{cd}, S) d\zeta_1 \int_{t_{cd}}^{\infty} t_{ab} P(t_{ab}|t_{cd}, S) dt_{ab} dt_{cd} \\
&= \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) \int_{t_0}^{t_{cd}} \frac{4}{5} P(\text{5p Coal } = \zeta_1 | t_{cd}, S) \\
&\quad \times P(\text{2p No Coal in } (\zeta_1, t_{cd}) | t_{cd}, S) d\zeta_1 \int_{t_{cd}}^{\infty} t_{ab} P(t_{ab}|t_{cd}, S) dt_{ab} dt_{cd} \\
&= \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) \Bigg[\sum_{k=0}^{i-1} \Big(P(\text{5p No Coal in } (t_0, \tau_k)) \\
&\quad \times \int_{\tau_k}^{\tau_{k+1}} \frac{4}{5} P(\text{5p Coal } = \zeta_1 | t_{cd}, S) P(\text{2p No Coal in } (\zeta_1, t_{cd}) | t_{cd}, S) d\zeta_1
\end{aligned}
$$

$$
\begin{aligned}
&\times P(\text{2p No Coal in } (\tau_k, t_{cd})|t_{cd}, S)\Big) \\
&+ \int_{\tau_i}^{t_{cd}} \frac{4}{5} P(\text{5p Coal } = \zeta_1|t_{cd}, S) P(\text{2p No Coal in } (\zeta_1, t_{cd})|t_{cd}, S) d\zeta_1 \Bigg] \\
&\times \Bigg[\int_{t_{cd}}^{\tau_{i+1}} t_{ab} P(t_{ab}|t_{cd}, S) dt_{ab} dt_{cd} + P(\text{1p No Coal in } (t_{cd}, \tau_{i+1})|t_{cd}, S) \\
&\times \int_{\tau_{i+1}}^{\infty} t_{ab} P(t_{ab}|\tau_{i+1}, S) dt_{ab} \Bigg] \\
= &\Bigg[\sum_{k=0}^{i-1} P(\text{5p No Coal in } (t_0, \tau_k)) \int_{\tau_k}^{\tau_{k+1}} \frac{4}{5} P(\text{5p Coal } = \zeta_1|t_{cd}, S) \\
&\times P(\text{2p No Coal in } (\zeta_1, \tau_{k+1})|t_{cd}, S) d\zeta_1 \\
&\times P(\text{2p No Coal in } (\tau_{k+1}, \tau_i)) \Bigg] \\
&\times \Bigg[\underbrace{\int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) P(\text{2p No Coal in } (t_i, t_{cd})) \int_{t_{cd}}^{\tau_{i+1}} t_{ab} P(t_{ab}|t_{cd}, S) dt_{ab} dt_{cd}}_{C4A(i)} \\
&+ \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) P(\text{2p No Coal in } (t_i, t_{cd})) P(\text{1p No Coal in } (t_{cd}, \tau_{i+1}) dt_{cd} \\
&\times \int_{\tau_{i+1}}^{\infty} t_{ab} P(t_{ab}|\tau_{i+1}, S) dt_{ab} \Bigg] + P(\text{5p No Coal in } (t_0, \tau_i)) \\
&\times \Bigg[\underbrace{\int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) \int_{\tau_i}^{t_{cd}} \frac{4}{5} P(\text{5p Coal } = \zeta_1|t_{cd}, S)}_{C4B(i)} \\
&\times \underbrace{P(\text{2p No Coal in } (\zeta_1, t_{cd})|t_{cd}, S) d\zeta_1 \int_{t_{cd}}^{\tau_{i+1}} t_{ab} P(t_{ab}|t_{cd}, S) dt_{ab} dt_{cd}}_{C4B(i) \text{ con't}} \\
&+ \underbrace{\int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) \int_{\tau_i}^{t_{cd}} \frac{4}{5} P(\text{5p Coal } = \zeta_1|t_{cd}, S)}_{C4C(i)} \\
&\times \underbrace{P(\text{2p No Coal in } (\zeta_1, t_{cd})|t_{cd}, S) d\zeta_1 P(\text{1p No Coal in } (t_{cd}, \tau_{i+1})) dt_{cd}}_{C4C(i) \text{ con't}} \\
&\times \int_{\tau_{i+1}}^{\infty} t_{ab} P(t_{ab}|\tau_{i+1}, S) dt_{ab}
\end{aligned}
$$

So let's evaluate $C4A(i)$, $C4B(i)$, and $C4C(i)$.

$$\begin{aligned}
C4A(i) &= \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) P(\text{2p No Coal in } (t_i, t_{cd})) \int_{t_{cd}}^{\tau_{i+1}} t_{ab} P(t_{ab}|t_{cd}, S) dt_{ab} dt_{cd} \\
&= \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) e^{-\frac{2(t_{cd}-\tau_i)}{2\eta_i}} \Big(-(2\eta_i + \tau_{i+1}) e^{-\frac{(\tau_{i+1}-t_{cd})}{2\eta_i}} + 2\eta_i + t_{cd} \Big) dt_{cd} \\
&= -(2\eta_i + \tau_{i+1}) \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) e^{-\frac{2(t_{cd}-\tau_i)}{2\eta_i}} e^{-\frac{(\tau_{i+1}-t_{cd})}{2\eta_i}} dt_{cd} \\
&\quad + 2\eta_i \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) e^{-\frac{2(t_{cd}-\tau_i)}{2\eta_i}} dt_{cd} + \int_{\tau_i}^{\tau_{i+1}} t_{cd}^2 P(t_{cd}|S) e^{-\frac{2(t_{cd}-\tau_i)}{2\eta_i}} dt_{cd}
\end{aligned}$$

Now on to $C4B(i)$:

$$\begin{aligned}
C4B(i) &= \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) \int_{\tau_i}^{t_{cd}} \frac{4}{5} P(\text{5p Coal } = \zeta_1 | t_{cd}, S) \\
&\quad \times P(\text{2p No Coal in } (\zeta_1, t_{cd}) | t_{cd}, S) d\zeta_1 \\
&\quad \times \int_{t_{cd}}^{\tau_{i+1}} t_{ab} P(t_{ab}|t_{cd}, S) dt_{ab} dt_{cd} \\
&= \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) \left[\frac{4}{3} \Big(e^{-\frac{2(t_{cd}-\tau_i)}{2\eta_i}} - e^{-\frac{5(t_{cd}-\tau_i)}{2\eta_i}} \Big) \right] \\
&\quad \times \left[-(2\eta_i + \tau_{i+1}) e^{-\frac{(\tau_{i+1}-t_{cd})}{2\eta_i}} + 2\eta_i + t_{cd} \right] dt_{cd} \\
&= -\frac{4}{3}(2\eta_i + \tau_{i+1}) \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) e^{-\frac{2(t_{cd}-\tau_i)}{2\eta_i}} e^{-\frac{(\tau_{i+1}-t_{cd})}{2\eta_i}} dt_{cd} \\
&\quad + \frac{4}{3}(2\eta_i + \tau_{i+1}) \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) e^{-\frac{5(t_{cd}-\tau_i)}{2\eta_i}} e^{-\frac{(\tau_{i+1}-t_{cd})}{2\eta_i}} dt_{cd} \\
&\quad + \frac{8}{3}\eta_i \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) e^{-\frac{2(t_{cd}-\tau_i)}{2\eta_i}} dt_{cd} - \frac{8}{3}\eta_i \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) e^{-\frac{5(t_{cd}-\tau_i)}{2\eta_i}} dt_{cd} \\
&\quad + \frac{4}{3} \int_{\tau_i}^{\tau_{i+1}} t_{cd}^2 P(t_{cd}|S) e^{-\frac{2(t_{cd}-\tau_i)}{2\eta_i}} dt_{cd} - \frac{4}{3} \int_{\tau_i}^{\tau_{i+1}} t_{cd}^2 P(t_{cd}|S) e^{-\frac{5(t_{cd}-\tau_i)}{2\eta_i}} dt_{cd}
\end{aligned}$$

Next, $C4C(i)$:

$$
\begin{aligned}
C4C(i) = &\int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) \int_{\tau_i}^{t_{cd}} \frac{4}{5} P(\text{5p Coal } = \zeta_1 | t_{cd}, S) \\
&\times P(\text{2p No Coal in } (\zeta_1, t_{cd}) | t_{cd}, S) d\zeta_1 P(\text{1p No Coal in } (t_{cd}, \tau_{i+1})) dt_{cd} \\
= &\int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) P(\text{1p No Coal in } (t_{cd}, \tau_{i+1})) \\
&\times \left(\frac{4}{3} e^{-\frac{2(t_{cd}-\tau_i)}{2\eta_i}} - \frac{4}{3} e^{-\frac{5(t_{cd}-\tau_i)}{2\eta_i}} \right) dt_{cd} \\
= &\frac{4}{3} \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) e^{-\frac{2(t_{cd}-\tau_i)}{2\eta_i}} e^{-\frac{(\tau_{i+1}-t_{cd})}{2\eta_i}} dt_{cd} \\
&- \frac{4}{3} \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) e^{-\frac{5(t_{cd}-\tau_i)}{2\eta_i}} e^{-\frac{(\tau_{i+1}-t_{cd})}{2\eta_i}} dt_{cd}
\end{aligned}
$$

Case 5: T_{cd} is the first event, T_{ab} is the second.

$$
\begin{aligned}
C5(i) &= \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) P(\text{5p No Coal in } (t_0, t_{cd})|t_{cd}, S) \\
&\quad \times \int_{t_{cd}}^{\infty} \frac{1}{3} t_{ab} P(\text{3p coal} = t_{ab}|t_{cd}, S) dt_{ab} dt_{cd} \\
&= P(\text{5p No Coal in } (t_0, \tau_i|S) \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) P(\text{5p No Coal in } (\tau_i, t_{cd})|t_{cd}, S) \\
&\quad \times \Big[\int_{t_{cd}}^{\tau_{i+1}} t_{ab} \frac{1}{3} P(\text{3p Coal } = t_{ab}|t_{cd}, S) dt_{ab} \\
&\quad + P(\text{3p No Coal in } (t_{cd}, \tau_{i+1})) \\
&\quad \times \int_{\tau_{i+1}}^{\infty} t_{ab} \frac{1}{3} P(\text{3p Coal } = t_{ab}|\tau_{i+1}, S) dt_{ab} \Big] dt_{cd} \\
&= P(\text{5p No Coal in } (t_0, \tau_i|S) \\
&\quad \times \Big[\underbrace{\int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) P(\text{5p No Coal in } (\tau_i, t_{cd})|t_{cd}, S)}_{C5A(i)} \\
&\quad \underbrace{\times \int_{t_{cd}}^{\tau_{i+1}} t_{ab} \frac{1}{3} P(\text{3p Coal } = t_{ab}|t_{cd}, S) dt_{ab} dt_{cd}}_{C5A(i) \text{ con't}} \\
&\quad + \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) P(\text{5p No Coal in } (\tau_i, t_{cd})|t_{cd}, S) \\
&\quad \times P(\text{3p No Coal in } (t_{cd}, \tau_{i+1})) dt_{cd} \int_{\tau_{i+1}}^{\infty} t_{ab} \frac{1}{3} P(\text{3p Coal } = t_{ab}|\tau_{i+1}, S) dt_{ab} \Big]
\end{aligned}
$$

So let's evaluate $C5A(i)$:

$$
\begin{aligned}
C5A(i) &= \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) P(\text{5p No Coal in } (\tau_i, t_{cd}))|t_{cd}, S) \\
&\quad \times \int_{t_{cd}}^{\tau_{i+1}} t_{ab} \frac{1}{3} P(\text{3p Coal } = t_{ab}|t_{cd}, S) dt_{ab} dt_{cd} \\
&= \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) P(\text{5p No Coal in } (\tau_i, t_{cd}))|t_{cd}, S) \\
&\quad \times \frac{1}{3}\Big[-\big(\frac{2}{3}\eta_i + \tau_{i+1}\big) e^{-\frac{3(\tau_{i+1} - t_{cd})}{2\eta_i}} + \frac{2}{3}\eta_i + t_{cd} \Big] dt_{cd} \\
&= -\frac{1}{3}\big(\frac{2}{3}\eta_i + \tau_{i+1}\big) \\
&\quad \times \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) P(\text{5p No Coal in } (\tau_i, t_{cd}))|t_{cd}, S) e^{-\frac{3(\tau_{i+1} - t_{cd})}{2\eta_i}} dt_{cd} \\
&\quad + \frac{2}{9}\eta_i \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) P(\text{5p No Coal in } (\tau_i, t_{cd}))|t_{cd}, S) dt_{cd} \\
&\quad + \frac{1}{3} \int_{\tau_i}^{\tau_{i+1}} t_{cd}^2 P(t_{cd}|S) P(\text{5p No Coal in } (\tau_i, t_{cd}))|t_{cd}, S) dt_{cd}
\end{aligned}
$$

Case 6: T_{cd} is the first event, T_{ab} is the third.

$$
\begin{aligned}
C6(i) &= \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) P(\text{5p No Coal in } (t_0, t_{cd}))|t_{cd}, S) \int_{t_{cd}}^{\infty} \frac{2}{3} P(\text{3p Coal} = \tau|t_{cd}, S) \\
&\quad \times \int_{\tau}^{\infty} t_{ab} P(t_{ab}|\tau, S) dt_{ab} d\tau dt_{cd} \Big] \\
&= \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) P(\text{5p No Coal in } (t_0, t_{cd}))|t_{cd}, S) \\
&\quad \times \Big[\int_{t_{cd}}^{\tau_{i+1}} \frac{2}{3} P(\text{3p Coal} = \tau|t_{cd}, S) \int_{\tau}^{\infty} t_{ab} P(t_{ab}|\tau, S) dt_{ab} d\tau \\
&\quad + P(\text{3p No Coal in } (t_{cd}, \tau_{i+1}))|t_{cd}, S) \int_{\tau_{i+1}}^{\infty} \frac{2}{3} P(\text{3p Coal} = \tau|\tau_{i+1}, S) \\
&\quad \times \int_{\tau}^{\infty} P(t_{ab}|\tau, S) dt_{ab} d\tau \Big] dt_{cd}
\end{aligned}
$$

$$
\begin{aligned}
&= \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) P(\text{5p No Coal in } (t_0, t_{cd})|t_{cd}, S) \\
&\quad \times \Bigg[\int_{t_{cd}}^{\tau_{i+1}} \frac{2}{3} P(\text{3p Coal} = \tau|t_{cd}, S) \int_{\tau}^{\tau_{i+1}} t_{ab} P(t_{ab}|\tau, S) dt_{ab} d\tau \\
&\quad + \int_{t_{cd}}^{\tau_{i+1}} \frac{2}{3} P(\text{3p Coal} = \tau|t_{cd}, S) P(\text{1p No Coal in}(\tau, \tau_{i+1})|S) d\tau \\
&\quad \times \int_{\tau_{i+1}}^{\infty} t_{ab} P(t_{ab}|\tau_{i+1}, S) dt_{ab} \\
&\quad + P(\text{3p No Coal in } (t_{cd}, \tau_{i+1})|t_{cd}, S) \int_{\tau_{i+1}}^{\infty} \frac{2}{3} P(\text{3p Coal} = \tau|\tau_{i+1}, S) \\
&\quad \times \int_{\tau}^{\infty} t_{ab} P(t_{ab}|\tau, S) dt_{ab} d\tau \Bigg] dt_{cd} \\
&= \underbrace{\int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) P(\text{5p No Coal in } (t_0, t_{cd})|t_{cd}, S) \int_{t_{cd}}^{\tau_{i+1}} \frac{2}{3} P(\text{3p Coal} = \tau|t_{cd}, S)}_{C6A(i)} \\
&\quad \underbrace{\times \int_{\tau}^{\tau_{i+1}} t_{ab} P(t_{ab}|\tau, S) dt_{ab} d\tau dt_{cd}}_{C6A(i) \text{ con't}} \\
&+ \underbrace{\int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) P(\text{5p No Coal in } (t_0, t_{cd})|t_{cd}, S)}_{C6B(i)} \\
&\quad \underbrace{\times \int_{t_{cd}}^{\tau_{i+1}} \frac{2}{3} P(\text{3p Coal} = \tau|t_{cd}, S) P(\text{1p No Coal in}(\tau, \tau_{i+1})|S) d\tau dt_{cd}}_{C6B(i) \text{ con't}} \\
&\quad \times \int_{\tau_{i+1}}^{\infty} t_{ab} P(t_{ab}|\tau_{i+1}, S) dt_{ab} \\
&\quad + \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) P(\text{5p No Coal in } (t_0, t_{cd})|t_{cd}, S) P(\text{3p No Coal in}(t_{cd}, \tau_{i+1})|S) dt_{cd} \\
&\quad \underbrace{\times \int_{\tau_{i+1}}^{\infty} \frac{2}{3} P(\text{3p Coal} = \tau|\tau_{i+1}, S) \int_{\tau}^{\infty} t_{ab} P(t_{ab}|\tau, S) dt_{ab} d\tau}_{C6C(i)} \\
&= C6A(i) + C6B(i) \int_{\tau_{i+1}}^{\infty} t_{ab} P(t_{ab}|\tau_{i+1}, S) dt_{ab} + C6C(i) \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) \\
&\quad \times P(\text{5p No Coal in } (t_0, t_{cd})|t_{cd}, S) P(\text{3p No Coal in}(t_{cd}, \tau_{i+1})|S) dt_{cd}
\end{aligned}
$$

So let's evaluate each of these three components.

$$
\begin{aligned}
C6A(i) &= \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) P(\text{5p No Coal in } (t_0, t_{cd})|t_{cd}, S) \\
&\quad \times \int_{t_{cd}}^{\tau_{i+1}} \frac{2}{3} P(\text{3p Coal} = \tau|t_{cd}, S) \int_{\tau}^{\tau_{i+1}} t_{ab} P(t_{ab}|\tau, S) dt_{ab} d\tau dt_{cd} \\
&= \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) P(\text{5p No Coal in } (t_0, t_{cd})|t_{cd}, S) \\
&\quad \times \int_{t_{cd}}^{\tau_{i+1}} \frac{2}{3} P(\text{3p Coal} = \tau|t_{cd}, S) \\
&\quad \times \Big[- \Big(2\eta_i + \tau_{i+1}\Big) P(\text{1p No Coal in } (\tau, \tau_{i+1})|S) + 2\eta_i + \tau \Big] d\tau dt_{cd} \\
&= \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) P(\text{5p No Coal in } (t_0, t_{cd})|t_{cd}, S) \\
&\quad \times \Big[- \frac{2}{3}\Big(2\eta_i + \tau_{i+1}\Big) \int_{t_{cd}}^{\tau_{i+1}} P(\text{3p Coal} = \tau|t_{cd}, S) \\
&\quad \times P(\text{1p No Coal in } (\tau, \tau_{i+1})|S) d\tau + \frac{4}{3}\eta_i \int_{t_{cd}}^{\tau_{i+1}} P(\text{3p Coal} = \tau|t_{cd}, S) d\tau \\
&\quad + \frac{2}{3} \int_{t_{cd}}^{\tau_{i+1}} \tau P(\text{3p Coal} = \tau|t_{cd}, S) d\tau \Big] dt_{cd} \\
&= P(\text{5p No Coal in } (t_0, \tau_i)|S) \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) \\
&\quad \times P(\text{5p No Coal in } (\tau_i, t_{cd})|t_{cd}, S) \\
&\quad \times \Big[- \Big(2\eta_i + \tau_{i+1}\Big)\Big(e^{-\frac{(\tau_{i+1} - t_{cd})}{2\eta_i}} - e^{-\frac{3(\tau_{i+1} - t_{cd})}{2\eta_i}}\Big) + \frac{4}{3}\Big(1 - e^{-\frac{3(\tau_{i+1} - t_{cd})}{2\eta_i}}\Big) \\
&\quad + \frac{2}{3}\Big[- \Big(\frac{2}{3}\eta_i + \tau_{i+1}\Big) e^{-\frac{3(\tau_{i+1} - t_{cd})}{2\eta_i}} + \frac{2}{3}\eta_i + t_{cd} \Big] \Big] dt_{cd}
\end{aligned}
$$

$$
\begin{aligned}
&= P(\text{5p No Coal in } (t_0, \tau_i)|S) \\
&\quad \times \Bigg[\Big(2\eta_i + \tau_{i+1} - \frac{4}{3} - \frac{2}{3}\big(\frac{2}{3}\eta_i + \tau_{i+1}\big) \Big) \\
&\quad \times \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) P(\text{5p No Coal in } (\tau_i, t_{cd})|t_{cd}, S) e^{-\frac{3(\tau_{i+1}-t_{cd})}{2\eta_i}} dt_{cd} \\
&\quad - \Big(2\eta_i + \tau_{i+1} \Big) \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) P(\text{5p No Coal in } (\tau_i, t_{cd})|t_{cd}, S) e^{-\frac{(\tau_{i+1}-t_{cd})}{2\eta_i}} dt_{cd} \\
&\quad + \Big(\frac{4}{3} + \frac{4}{9}\eta_i \Big) \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) P(\text{5p No Coal in } (\tau_i, t_{cd})|t_{cd}, S) dt_{cd} \\
&\quad + \frac{2}{3} \int_{\tau_i}^{\tau_{i+1}} t_{cd}^2 P(t_{cd}|S) P(\text{5p No Coal in } (\tau_i, t_{cd})|t_{cd}, S) dt_{cd} \Bigg]
\end{aligned}
$$

So now lets look at $C6B(i)$:

$$
\begin{aligned}
C6B(i) &= \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) P(\text{5p No Coal in } (t_0, t_{cd})|t_{cd}, S) \\
&\quad \times \int_{t_{cd}}^{\tau_{i+1}} \frac{2}{3} P(\text{3p Coal} = \zeta|t_{cd}, S) P(\text{1p No Coal in}(\zeta, \tau_{i+1})|S) d\zeta dt_{cd} \\
&= P(\text{5p No Coal in } (t_0, \tau_i)|S) \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) P(\text{5p No Coal in } (\tau_i, t_{cd})|t_{cd}, S) \\
&\quad \times \frac{2}{3} \Bigg[\frac{3}{2} \Big(e^{-\frac{(\tau_{i+1}-t_{cd})}{2\eta_i}} - e^{-\frac{3(\tau_{i+1}-t_{cd})}{2\eta_i}} \Big) \Bigg] dt_{cd} \\
&= P(\text{5p No Coal in } (t_0, \tau_i)|S) \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) \\
&\quad \times P(\text{5p No Coal in } (\tau_i, t_{cd})|t_{cd}, S) e^{-\frac{(\tau_{i+1}-t_{cd})}{2\eta_i}} dt_{cd} \\
&\quad - P(\text{5p No Coal in } (t_0, \tau_i)|S) \int_{\tau_i}^{\tau_{i+1}} t_{cd} P(t_{cd}|S) \\
&\quad \times P(\text{5p No Coal in } (\tau_i, t_{cd})|t_{cd}, S) e^{-\frac{3(\tau_{i+1}-t_{cd})}{2\eta_i}} dt_{cd}
\end{aligned}
$$

Lastly, let's look at $C6C(i)$:

$$
\begin{aligned}
C6C(i) &= \int_{\tau_{i+1}}^{\infty} \frac{2}{3} P(\text{3p Coal} = \zeta | \tau_{i+1}, S) \int_{\zeta}^{\infty} t_{ab} P(t_{ab}|\zeta, S) dt_{ab} d\zeta \\
&= \sum_{j=i+1}^{n-1} (\text{3p No Coal in } (\tau_{i+1}, \tau_j)|S) \int_{\tau_j}^{\tau_{j+1}} \frac{2}{3} P(\text{3p Coal} = \zeta|\tau_j, S) \\
&\quad \times \Bigg[\int_{\zeta}^{\tau_{j+1}} t_{ab} P(t_{ab}|\zeta, S) dt_{ab} + P(\text{1p No Coal in } (\tau, \tau_{j+1})) \\
&\quad \times \int_{\tau_{j+1}}^{\infty} t_{ab} P(t_{ab}|\tau_{j+1}, S) dt_{ab} \Bigg] d\zeta \\
&= \sum_{j=i+1}^{n-1} (\text{3p No Coal in } (\tau_{i+1}, \tau_j)|S) \\
&\quad \times \Bigg[\int_{\tau_j}^{\tau_{j+1}} \frac{2}{3} P(\text{3p Coal} = \zeta|\tau_j, S) \Big(- \big(2\eta_j + \tau_{j+1}\big) e^{-\frac{(\tau_{j+1} - \zeta)}{2\eta_j}} + 2\eta_j + \zeta \Big) d\zeta \\
&\quad + \int_{\tau_{j+1}}^{\infty} t_{ab} P(t_{ab}|\tau_{j+1}, S) dt_{ab} \int_{\tau_j}^{\tau_{j+1}} \frac{2}{3} P(\text{3p Coal} = \zeta|\tau_j, S) e^{-\frac{(\tau_{j+1} - \zeta)}{2\eta_j}} d\zeta \Bigg] \\
&= \sum_{j=i+1}^{n-1} (\text{3p No Coal in } (\tau_{i+1}, \tau_j)|S) \\
&\quad \times \Bigg[\Big(\int_{\tau_{j+1}}^{\infty} t_{ab} P(t_{ab}|\tau_{j+1}, S) dt_{ab} - \big(2\eta_j + \tau_{j+1}\big) \Big) \\
&\quad \times \int_{\tau_j}^{\tau_{j+1}} \frac{2}{3} P(\text{3p Coal} = \zeta|\tau_j, S) e^{-\frac{(\tau_{j+1} - \zeta)}{2\eta_j}} d\zeta \\
&\quad + \frac{4}{3}\eta_j \int_{\tau_j}^{\tau_{j+1}} \frac{2}{3} P(\text{3p Coal} = \zeta|\tau_j, S) d\tau + \frac{2}{3} \int_{\tau_j}^{\tau_{j+1}} \frac{2}{3} \zeta P(\text{3p Coal} = \zeta|\tau_j, S) d\zeta \Bigg]
\end{aligned}
$$

By piecing together these quantities, which have all been presented as single integrals over a constant population window, we can get the exact value for $\mathbb{E}(T_{ab}, T_{cd}|\mathcal{S}, > D_{abcd})$ which is used in all subsequent covariance calculations.

We omit the presentation of similar calculations when only three individuals are present ($\mathbb{E}(t_{ab}, t_{bc}|S, > D_{abc})$) as the logic follows similar to the process presented here.

Notation for integrals

$$P(\text{K p No Coal } \in (\tau_i, \tau_{i+1})) = e^{-\frac{K(\tau_{i+1}-\tau_i)}{2\eta_i}}$$

indicates the probability that given K pairs of individuals entering branch i, none of them coalesce in the branch parameterized by times τ_i, τ_{i+1} and η_i.

$$P(\text{K p Coal } = \tau|\tau_i) = \frac{K}{2\eta_i} e^{-\frac{K(\tau-\tau_i)}{2\eta_i}}$$

is the probability that given K pairs of individuals have survived to branch i, that the first coalescence of the K pairs occurs at time τ.

$$T_{ac,ad,bc,bd}\ 1^{st} \text{ event} = \zeta_1$$

is the event that of the available 6 pairs of coalescence events, the first to occur happens at time ζ_1 and it is not pairs a, b or c, d.

In the equations presented here, we commonly see ambiguous notation like

$$\int_{\tau_i}^{\tau_{i+1}} P(t_{ab}) dt_{ab}$$

Here we are assuming t_{ab} has failed to coalesce before τ_i, and therefore the density $P(t_{ab})$ should also be conditioned on that value, such that

$$P(t_{ab}|\tau_i) = \frac{1}{2\eta_i} e^{-\frac{T_{ab}-\tau_i}{2\eta_i}}$$

we simply leave this off for sake of compactness in our equations.