Title

Covariance of pairwise differences on a multi-species coalescent tree and implications for FST

Journal

Philosophical Transactions of the Royal Society B Biological Sciences, 377(1852)

ISSN

0962-8436

Authors

Guerra, Geno

Nielsen, Rasmus

Publication Date

2022-06-06

DOI

10.1098/rstb.2020.0415

PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Research

Cite this article: Guerra G, Nielsen R. 2022 Covariance of pairwise differences on a multi-species coalescent tree and implications for F_{ST}. *Phil. Trans. R. Soc. B* **377**: 20200415. https://doi.org/10.1098/rstb.2020.0415

Received: 30 June 2021
Accepted: 4 January 2022

One contribution of 15 to a theme issue 'Celebrating 50 years since Lewontin's apportionment of human diversity'.

Subject Areas:
computational biology, evolution, genetics, theoretical biology, genomics

Keywords:
multi-species coalescent, covariance, F_{ST}, population differentiation, pairwise differences

Author for correspondence:
Geno Guerra
e-mail: geno.guerra@ucsf.edu

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.5901223.

THE ROYAL SOCIETY PUBLISHING

Covariance of pairwise differences on a multi-species coalescent tree and implications for F_{ST}

Geno Guerra[1,3] and Rasmus Nielsen[1,2,4]

[1]Department of Statistics, and [2]Department of Integrative Biology, University of California, Berkeley, CA 94720, USA
[3]Department of Neurological Surgery, University of California, San Francisco, CA 94158, USA
[4]Lundbeck Foundations Centre for GeoGenetics, University of Copenhagen, Kobenhavn, Denmark

GG, 0000-0001-9870-9998; RN, 0000-0003-0513-6591

The multi-species coalescent (MSC) provides a theoretical foundation for modern phylogenetics and comparative population genetics. Its theoretical properties have been heavily studied but there are still aspects of the MSC that are largely unknown, including the covariances in pairwise coalescence times, which are fundamental for understanding the properties of statistics that combine data from multiple species, such as the fixation index (F_{ST}). The major contribution of this study is the derivation and implementation of exact expressions for the covariances of pairwise coalescence times under phylogenetic models with piecewise constant changes in population size, assuming no gene flow after species divergence. We use these expressions to derive the variance in average pairwise differences within and between populations. We then derive approximations for the expectation and bias of a sequence-based estimator of F_{ST}, a commonly used genetic measurement of population differentiation, when it is applied to a non-recombining region of the genome. We show that the estimator of F_{ST} is generally biased downward. A freely available software package is provided, STCov, to calculate the mean, variances and covariances in coalescence times presented here under user-defined piecewise-constant species trees.

This article is part of the theme issue 'Celebrating 50 years since Lewontin's apportionment of human diversity'.

1. Introduction

The multi-species coalescent (MSC) is a generalization of Kingman's coalescent [1] that describes the joint coalescence process in multiple species, or populations, as they diverge from each other. The MSC provides a theoretical foundation for phylogenetic analyses as it fully describes and characterizes the process of incomplete lineage sorting [2–5]. It is, therefore, central in the unification of the fields of population genetics and phylogenetics. It is also central for understanding divergence between populations and allows the theoretical prediction of the amount of variance within and between populations. In this sense, it provides a theoretical framework for relating apportionment of genetic variance within and between populations, as proposed by Lewontin [6], to specific models of population divergence.

One of the important utilities of theoretical models, such as the MSC, is to provide predictions regarding observed statistics, eventually leading to the development of estimators of population-level parameters. In this regard, an important use of the MSC has been to understand the properties of pairwise nucleotide differences within and between species, which is one of the most commonly used statistics to analyse population genetic data. Takahata & Nei [7] derived expressions for the variance in average pairwise nucleotide differences and Nei and Li's 'net number of differences' [8], (d). They assumed a

Kingman's coalescent model [1] of two diverging populations, and an infinite sites model of mutation [9,10]. 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 reduce the variance. It is also one of the first uses of the MSC.

Takahata & Nei [7] defined d_X and d_Y to be the mean number of nucleotide differences between two (haploid) individuals sampled from within population X or Y, respectively. Similarly, d_{XY} is the average number of nucleotide differences between two individuals randomly sampled from populations X and Y. The statistics d_X, d_Y and d_{XY} are then calculated based on sample sizes of n_X and n_Y from populations X and Y, respectively, as follows:

$$d_X = \frac{2}{n_X(n_X-1)} \sum_{i=1}^{n_X-1} \sum_{i'=i+1}^{n_X} k_{i,i'} \tag{1.1}$$

$$d_Y = \frac{2}{n_Y(n_Y-1)} \sum_{i=1}^{n_Y-1} \sum_{i'=i+1}^{n_Y} k_{j,j'} \tag{1.2}$$

$$\text{and} \quad d_{XY} = \frac{1}{n_X n_Y} \sum_{i=1}^{n_X} \sum_{j=1}^{n_Y} k_{i,j}, \tag{1.3}$$

where $k_{i,i'}$ is the number of pairwise nucleotide differences between individuals (haplotype genomic sequences) i and i'. Henceforth, in this study, an 'individual' is a non-recombining haploid genomic sequence.

To measure the net number of nucleotide differences between two populations, Nei & Li's [8] d is defined as

$$d = d_{XY} - \frac{1}{2}(d_X + d_Y). \tag{1.4}$$

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. 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.

The pairwise differences d_{XY}, d_X and d_Y provide measures of genetic variability within and between species/populations that are applicable to DNA sequencing data and have been fundamental in analyses of such data since the 1980s. However, since their invention, the question quickly arose of how they relate to older measures of genetic divergence and variability originally derived for independent loci such as allozymes, in particular, how are they related to Wright's F_{ST}? Furthermore, how should F_{ST} appropriately be calculated for DNA sequencing data? These questions were answered by Slatkin [11], who argued that F_{ST} is equivalent to a ratio of average coalescence times of different pairs of genes. Assuming an infinite sites model, he then showed that Wright's F_{ST} in the context of DNA sequencing data could be expressed in terms of d_{XY}, d_X, and d_Y (see equation (7.2) below).

The statistics d_{XY}, d_X and d_Y have been, and continue to be, a cornerstone of the analysis of DNA sequence data. Understanding their mean, variances and covariances 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 [12] and Takahata & Nei [7] 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 [13] and the average pairwise differences in a model with migration [14], and later demonstrated the impact of recombination on the numerical stability of such estimates [15]. Tang *et al*. [16] derived an estimator for the time to most recent common ancestor (TMRCA) of a sample of DNA sequences along with quantification of sampling error by leveraging pairwise differences, free of population structure assumptions.

The multi-species coalescent has received renewed attention in the age of genomics because of its applicability in phylogenetic analyses using multiple loci. Efromovich & Kubatko [17] presented a method to calculate the distribution of coalescent times at the root of a species tree with an arbitrary number of populations. In a pair of papers, Wilkinson-Herbots provided unified analytic results for both the distribution of coalescence times and pairwise differences under models of isolation with migration [18,19] under assumptions of constant population size. Heled [20] helped to further marry previously pairwise difference quantification and the multispecies coalescent by deriving closed-form exact results for the 'average sequence dissimilarity' between pairs of sequences drawn at random under a simple two-species coalescent process with constant population size. Many methods have also been developed to use pairwise differences under the MSC while leveraging large genomics datasets to infer species tree topologies and divergence times (e.g. [21–23]).

Takahata & Nei's [7] original results on d_{XY}, d_X and d_Y relied on the assumption of constant and equal population sizes among populations and through time. Using the MSC, we here extend these results to arbitrary piecewise constant population size histories along a phylogeny. To do so, we derive and present general equations for calculating the covariance of pairwise coalescence times, for any two, three or four haploid 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 theoretical MSC quantities. 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 specifically investigate the statistical properties of Slatkin's F_{ST} estimator [11], and some of its various applications [24–26], as it is the most commonly used measure of F_{ST} using sequence data. We investigate the effects of bottlenecks, sampling variance and demographic changes on various F_{ST}-based measurements, and present the magnitude of downward bias when using F_{ST} estimated from a 'ratio of averages' approach to Slatkin's estimator, as is typical in single gene analyses.

2. Mean, variance and covariance of average pairwise differences

We first 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 ($k_{i,i'}$, $k_{i,j}\cdots$). Suppose i, i', i'', i''' are individuals from population

X, and j, j', j'', j''' are individuals from population Y. By definition we have,

$$\mathbb{E}(d_X) = \mathbb{E}(k_{i,i'}), \tag{2.1}$$

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

$$\mathbb{E}(d_{XY}) = \mathbb{E}(k_{i,j}). \tag{2.2}$$

Following the derivations in Tajima [12], Takahata & Nei [7] and Wakeley [14], under an infinite-site model of mutation, the variance and covariance of d_X, d_Y, d_{XY} and d can be written as follows:

$$\begin{aligned}\mathrm{Var}(d_X) = \frac{1}{n_X(n_X-1)}\big[&2\mathbb{E}(k_{i,i'}^2) + 4(n_X-2)\mathbb{E}(k_{i,i'}k_{i,i''}) \\ &+(n_X-2)(n_X-3)\mathbb{E}(k_{i,i'}k_{i'',i'''})\big] - \mathbb{E}(k_{i,i'})^2,\end{aligned} \tag{2.3}$$

$$\begin{aligned}\mathrm{Var}(d_Y) = \frac{1}{n_Y(n_Y-1)}\big[&2\mathbb{E}(k_{j,j'}^2) + 4(n_Y-2)\mathbb{E}(k_{j,j'}k_{j,j''}) \\ &+(n_Y-2)(n_Y-3)\mathbb{E}(k_{j,j'}k_{j'',j'''})\big] - \mathbb{E}(k_{j,j'})^2,\end{aligned} \tag{2.4}$$

$$\begin{aligned}\mathrm{Var}(d_{XY}) = \frac{1}{n_X n_Y}\big[&\mathbb{E}(k_{i,j}^2) + (n_Y-1)\mathbb{E}(k_{i,j}k_{i',j}) + (n_X-1)\mathbb{E}(k_{i,j}k_{i,j'}) \\ &+(n_X-1)(n_Y-1)\mathbb{E}(k_{i,j}k_{i',j'})\big] - \mathbb{E}(k_{ij})^2\end{aligned} \tag{2.5}$$

$$\begin{aligned}\text{and}\qquad \mathrm{Var}(d) = \mathrm{Var}(d_{XY}) + \frac{1}{4}&[\mathrm{Var}(d_X) + \mathrm{Var}(d_Y)] \\ &+2\mathrm{Cov}(d_X, d_Y) - \mathrm{Cov}(d_{XY}, d_X)] \\ &- \mathrm{Cov}(d_{XY}, d_Y).\end{aligned} \tag{2.6}$$

Further, formulae for the covariance of average pairwise difference terms can also be reduced to functions of individual pairwise terms

$$\mathrm{Cov}(d_X, d_Y) = \mathrm{Cov}(k_{i,i'}, k_{j,j'}). \tag{2.7}$$

This simple result is due to the fact that the covariance of sums can be decomposed into the sums of covariances.

As presented in Takahata & Nei (equations 18a–d) [7], covariance equations involving the cross population can be expressed as follows:

$$\begin{aligned}\mathrm{Cov}(d_{XY}, d_X) = &\frac{2}{n_X}\mathbb{E}(k_{i,i'}k_{i,j}) + \frac{n_X-2}{n_X}\mathbb{E}(k_{i,i'}k_{i'',j}) \\ &- \mathbb{E}(k_{i,i'})\mathbb{E}(k_{j,j'})\end{aligned} \tag{2.8}$$

and

$$\begin{aligned}\mathrm{Cov}(d_{XY}, d_Y) = &\frac{2}{n_Y}\mathbb{E}(k_{j,j'}k_{i,j}) + \frac{n_Y-2}{n_Y}\mathbb{E}(k_{j,j'}k_{i,j''}) \\ &- \mathbb{E}(k_{i,i'})\mathbb{E}(k_{j,j'}).\end{aligned} \tag{2.9}$$

These expressions are all functions of the individual pairwise differences, e.g. $k_{i,i'}$. In what proceeds we demonstrate that these expressions can be further generalized as functions of pairwise coalescence times, e.g. $t_{i,i'}$.

3. Pairwise mutational differences

In this section, we generalize previous work [7,12] by deriving expressions for the covariance of pairwise differences under arbitrary piecewise-constant demographic settings using the MSC. Throughout this section, we will assume an infinite sites model [9,10], with no recombination. We first review results on the mean and variance from previous work (e.g. [7,12,14]), and then extend results to the covariance.

(a) Mean and variance

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

$$\mathbb{E}(k_{i,j}) = 2\mu\mathbb{E}(t_{i,j}). \tag{3.1}$$

Under the assumption that the number of mutations conditional on a genealogy is Poisson, the conditional expectation and variance of pairwise differences are equal.

$$\mathrm{Var}(k_{i,j}|t_{i,j}) = \mathbb{E}(k_{i,j}|t_{i,j}). \tag{3.2}$$

By applying the law of total variance, we can decompose the unconditional variance of pairwise differences as

$$\begin{aligned}\sigma^2_{k_{i,j}} = \mathrm{Var}(k_{i,j}) &= \mathbb{E}\big(\mathrm{Var}(k_{i,j}|t_{i,j})\big) + \mathrm{Var}\big(\mathbb{E}(k_{i,j}|t_{i,j})\big) \\ &= \mathbb{E}(2\mu t_{i,j}) + \mathrm{Var}(2\mu t_{i,j}) \\ &= 2\mu\mathbb{E}(t_{i,j}) + 4\mu^2\mathrm{Var}(t_{i,j}).\end{aligned} \tag{3.3}$$

We can obtain the second moment of the distribution of pairwise nucleotide differences, $\mathbb{E}(k_{i,j}^2)$, from the definition of variance,

$$\mathbb{E}(k_{i,j}^2) = \sigma^2_{k_{i,j}} + \mathbb{E}(k_{i,j})^2 = 2\mu\mathbb{E}(t_{i,j}) + 8\mu^2\mathbb{E}(t_{i,j})^2. \tag{3.4}$$

(b) Covariance

Let i, i', j, j' be four individuals sampled from arbitrary populations. Let T be a local coalescent tree relating the four individuals restricted to a non-recombining region. Here, we show that

$$\mathrm{Cov}(k_{i,i'}, k_{j,j'}|T) = \mu t_{i,i'\cap j,j'}. \tag{3.5}$$

Consequently, we further derive the unconditional quantity

$$\mathrm{Cov}(k_{i,i'}, k_{j,j'}) = \mu\mathbb{E}(t_{i,i'\cap j,j'}) + 4\mu^2\mathrm{Cov}(t_{i,i'}, t_{j,j'}), \tag{3.6}$$

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 1 provides an illustrative example of this quantity, and electronic supplementary material, §F, provides a more technical treatment.

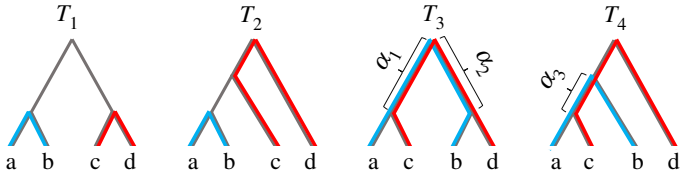
To prove these results, we start by revisiting the idea that under the infinite-site model, the mutational process given a branch length is Poisson. Given local tree, T, with coalescence times $t_{i,i'}$ and $t_{j,j'}$ from T, conditional pairwise differences follow a Poisson distribution, written as

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

where $2t_{i,i'}$ is the amount of total branch length locally between the two individuals. A key feature of the Poisson distribution is that the sum of Poisson random variables is also Poisson. 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' (figure 1). 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

$$2t_{i,i'} = t_{i,i'\cap j,j'} + t_{i,i'\setminus j,j'} \quad \text{and} \quad 2t_{j,j'} = t_{i,i'\cap j,j'} + t_{j,j'\setminus i,i'}. \tag{3.7}$$

(*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

$$\mathbb{E}(t_{a,b\cap c,d}) = \mathbb{E}(\alpha_1 + \alpha_2 | \mathrm{T}_3)\mathrm{P}(\mathrm{T}_3) + \mathbb{E}(\alpha_3 | \mathrm{T}_4)\mathrm{P}(\mathrm{T}_4)$$

Figure 1. (*a–c*) Explanation of expected shared branch length for four unique individuals. Bolded blue lines indicate the branch length between individuals a and b. Bolded red lines indicate branch length between c and d. Overlapping blue and red lines (along with α terms) indicate shared branch length. The four 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. (Online version in colour.)

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 therefore 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 $\mathrm{Cov}(k_{i,i'}, k_{j,j'} | T)$, (equation 3.5) as follows:

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

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, $\mathrm{Cov}(k_{i,i'}, k_{j,j'})$ (equation 3.6), we apply the law of total covariance:

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

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

Takahata & Nei [7] have previously derived formulas for the covariance under constant population size; see electronic supplementary material, §C, which presents a visualization of their results as a comparison to the generalized results presented here.

4. Mean, variance and covariance in pairwise coalescence times

We assume species evolution follows a bifurcating species tree $\mathcal{S} = (S, \vec{\tau}, \vec{\eta})$, with no migration (see figure 2*a*). Each branch, i, of $\mathcal{S}$ is 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 and compactness.

(a) 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, in a non-recombining region of the genome. For species tree $\mathcal{S}$, denote the marginal tree $\mathcal{S}_{XY} = (S_{XY}, \vec{\tau}_{XY}, \vec{\eta}_{XY})$ of two species (see figure 2*b*). Here, $\vec{\tau}_{XY}$ represents the set of divergence times of species ancestral to both X and Y, indexed by $(\tau_1, \tau_2, \ldots)$, where $\tau_1 := \tau_{XY}$, the divergence time for species X and Y. Similarly, $\vec{\eta}_{XY}$ 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

$$\begin{aligned}\mathbb{E}(t_{i,j}|\mathcal{S}) &= \sum_{k=1}^{V} P_{22}(\tau_1, \tau_k)\int_{\tau_k}^{\tau_{k+1}} t_{i,j}P(t_{i,j}|\mathcal{S}, \tau_k)\,\mathrm{d}t_{i,j}\\ &= \sum_{k=1}^{V} P_{22}(\tau_1, \tau_k)\int_{\tau_k}^{\tau_{k+1}} \frac{t_{i,j}}{2\eta_k}\,\mathrm{e}^{-((t_{i,j}-\tau_k)/2\eta_k)}\,\mathrm{d}t_{i,j}\\ &= \sum_{k=1}^{V} P_{22}(\tau_1, \tau_k)\Big[-(\tau_{k+1} + 2\eta_k)\,\mathrm{e}^{-((\tau_{k+1}-\tau_k)/2\eta_k)} + \tau_k + 2\eta_k\Big]\end{aligned} \tag{4.1}$$

and

$$\begin{aligned}\mathbb{E}(t_{i,j}^2|\mathcal{S}) &= \sum_{k=1}^{V} P_{22}(\tau_1, \tau_k)\int_{\tau_k}^{\tau_{k+1}} t_{i,j}^2P(t_{i,j}|\mathcal{S}, \tau_k)\,\mathrm{d}t_{i,j}\\ &= \sum_{k=1}^{V} P_{22}(\tau_1, \tau_k)\int_{\tau_k}^{\tau_{k+1}} \frac{t_{i,j}^2}{2\eta_k}\,\mathrm{e}^{-((t_{i,j}-\tau_k)/2\eta_i)}\,\mathrm{d}t_{i,j}\\ &= \sum_{k=1}^{V} P_{22}(\tau_1, \tau_k)\Big[-(\tau_{k+1}^2 + 4\tau_{k+1}\eta_k + 8\eta_k^2)\,\mathrm{e}^{-((\tau_{k+1}-\tau_k)/2\eta_k)}\\ &\quad + \tau_k^2 + 4\tau_k\eta_k + 8\eta_k^2\Big].\end{aligned} \tag{4.2}$$

$P_{22}(\tau_1, \tau_k)$ represents the probability that lineages i and j fail to coalesce in the time interval (τ_1, τ_k), (two lineages in, two

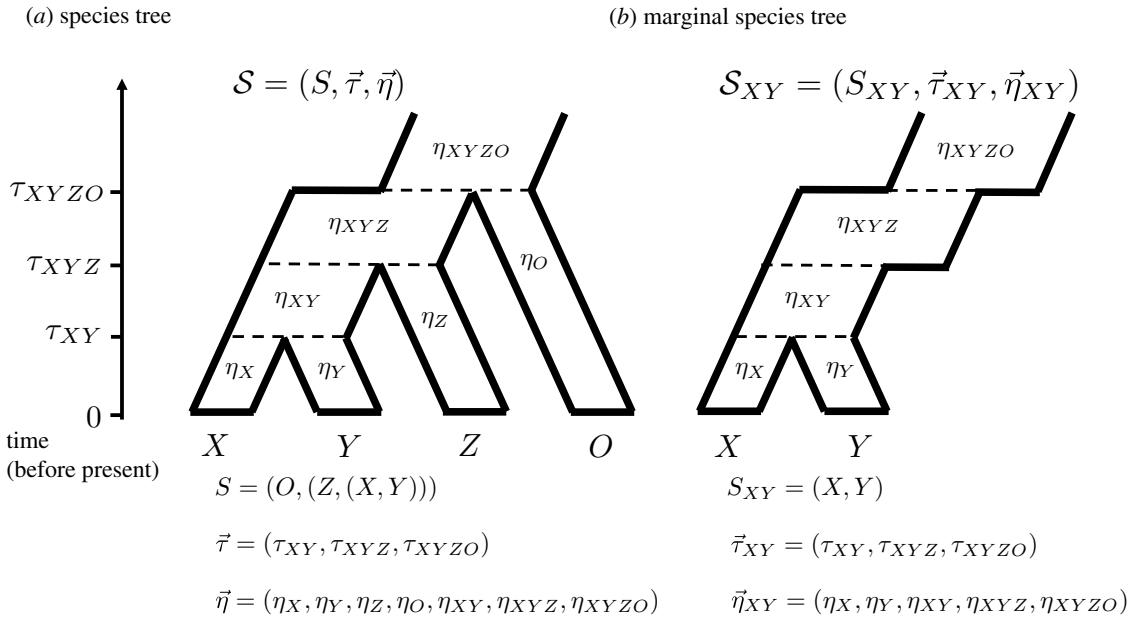


Figure 2. Species tree notation. (*a*) Example of notation used for a four-species tree with topology, divergence times and constant population sizes within each population which can vary between species. (*b*) Example of a marginal species tree, the result of subsetting a larger species tree. As a consequence, the population size histories are no longer constant within each species, but instead are piecewise constant.

lineages out). Formally, this is the probability that two lineages which exist in the same population at time interval τ_1 have not coalesced by time τ_k (backwards in time)

$$P_{22}(\tau_1, \tau_k) = \prod_{\tau_1 \le \tau_l < \tau_k} e^{-((\tau_{l+1} - T_l)/2\eta_l)}. \tag{4.3}$$

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 pairwise coalescence times under the standard piecewise constant coalescent process are just simply weighted sums over coalescence intervals.

(b) 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 given by

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

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

$$\mathbb{E}(t_{i,i'}t_{j,j'}|\mathcal{S}) = \int_{D_{j,j'}}^{\infty} t_{j,j'} P(t_{j,j'}|\mathcal{S}) \int_{D_{i,i'}}^{\infty} t_{i,i'} P(t_{i,i'}|t_{j,j'}, \mathcal{S})\, \mathrm{d}t_{i,i'}\, \mathrm{d}t_{j,j'}. \tag{4.4}$$

$D_{i,i'}$ 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. $t_{j,j'} \ge D_{j,j'}$), i.e. we assume no introgression, migration or admixture, etc.

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

C_1. $t_{i,i'}$ is the first coalescent event.
C_2. $t_{i,i'}$ is the second event, $t_{j,j'}$ is the third.
C_3. $t_{i,i'} = t_{j,j'}$ as the third coalescent event.
C_4. $t_{j,j'}$ is the second event, $t_{i,i'}$ is the third.
C_5. $t_{j,j'}$ is the first event, $t_{i,i'}$ is the second.
C_6. $t_{j,j'}$ is the first event, $t_{i,i'}$ is the third.

Here, 'first event' implies most recent, and 'third' implies most ancient. These events are further illustrated in detail in figure 3. Conditioning on each of these six events, and evaluating each expectation separately, the expression for the joint expectation becomes

$$\mathbb{E}(t_{i,i'}t_{j,j'}|\mathcal{S}) = \sum_{k=1}^{6} \mathbb{E}(t_{i,i'}t_{j,j'}|\mathcal{S}, C_k)P(C_k|\mathcal{S}). \tag{4.5}$$

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 electronic supplementary material, §G. This calculation is useful in the instance that all four 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 six cases multiplies the number of cases to consider. There are five general possible species tree configurations that can arise (see electronic supplementary material, figure S13). We have derived exact equations and recursions to evaluate all six cases ($C_1, \ldots, C_6$) across the five general possible tree configurations, and have implemented them in C++ code (STCov) 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.

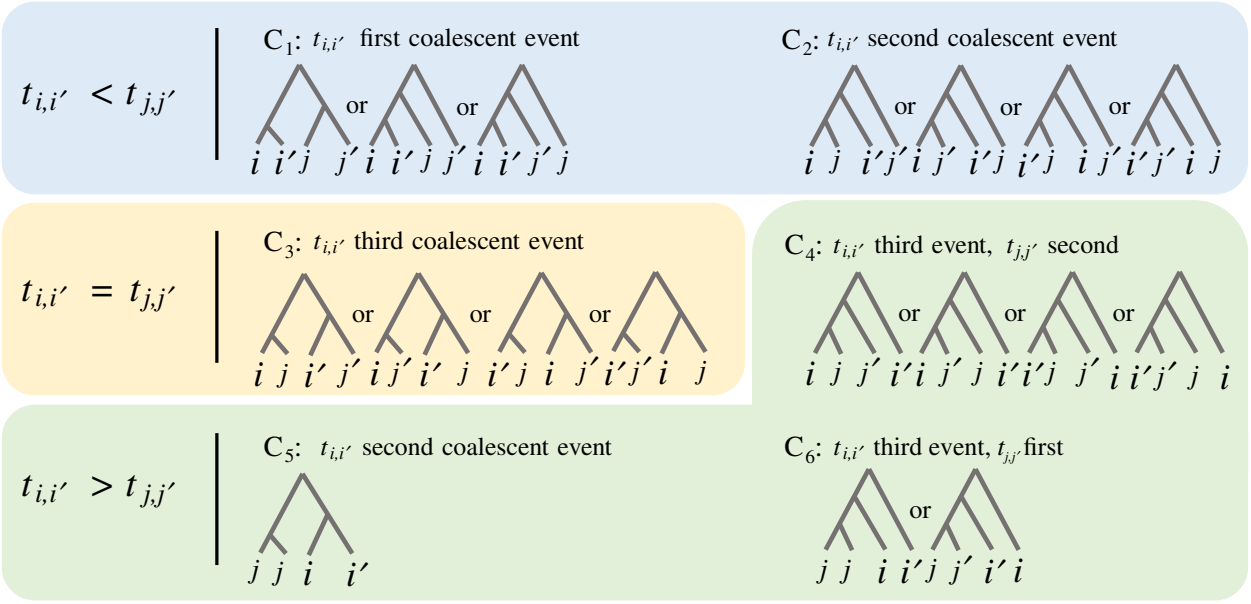


Figure 3. Ordered topologies to consider when calculating $t_{i,i'}|t_{j,j'}$. Given four individuals, i, i', j, j', the six cases presented outline the necessary labelled/ordered local trees essential for the conditional calculation of $P(t_{i,i'}|t_{j,j'})$. The cases can be grouped into three general scenarios based on the timing of $t_{i,i'}$ in relation to the conditional $t_{j,j'}$. All 18 possible ordered tree topologies are considered. (Online version in colour.)

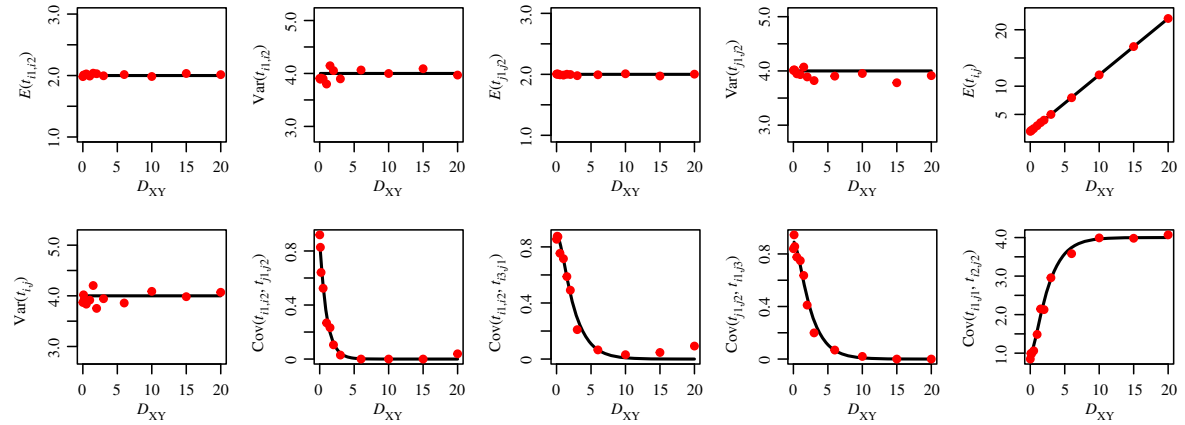


Figure 4. Assessing the accuracy of theoretical pairwise coalescent time calculations against simulated values, for population sizes: $\eta_Y = \eta_X$. Theoretical results from STCov are plotted as black curves, with dots representing empirical estimates of the quantity on the y-axis using 4500 independently simulated local trees.

5. Accuracy of coalescent calculations

To demonstrate the accuracy of the coalescent equations above, as implemented in our software STCov, we compare the theoretical results (assuming infinite-sites) against empirical estimates from gene trees under a finite-sites model using ms [27]. We first test two simple demographic scenarios for a tree of two species, X and Y: $\eta_Y = \eta_X$, and $\eta_Y = 2\eta_X$ (figures 4 and 5), 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 1500 independent gene trees from ms for each demographic scenario (with specified population sizes and single divergence time which we vary from 0–20 in units of $2\eta_X$ generations), and calculate sample mean, variance and covariance terms. The figures demonstrate that the theoretical calculations from STCov match simulations (dots) well, while variation in the empirical estimates can be attributed to a finite sample size.

6. 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 datasets.

We compare three population size change models, denoted by $\eta_Y = 1\eta_X$, $\eta_Y = 2\eta_X$ and $\eta_Y = 10\eta_X$, along with three mutation rates $2\mu\eta_X = 10, 1, 0.1$, for a total of nine simulation scenarios. We present one of those scenarios here (figure 6), and leave the full set of results to the electronic supplementary material, figures S2–S10. 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 (1500 independent genes of 10 000 bp each; see electronic supplementary material for full simulation details). We choose to compare this finite-

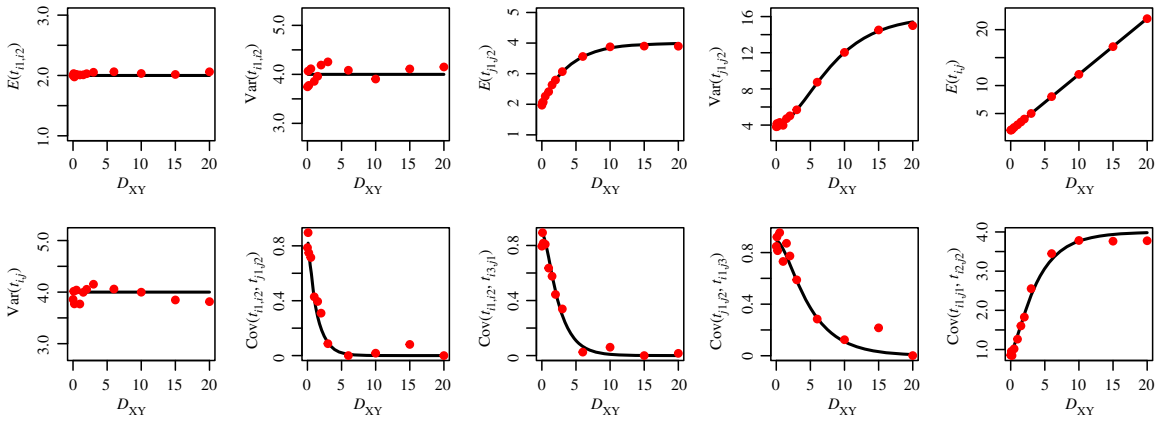


Figure 5. Assessing the accuracy of theoretical pairwise coalescent time calculations against simulated values, for population sizes: $\eta_Y = 2\eta_X$. Theoretical results from STCov are plotted as black curves, with dots representing empirical estimates of the quantity on the *y*-axis using 4500 independently simulated local trees. (Online version in colour.)

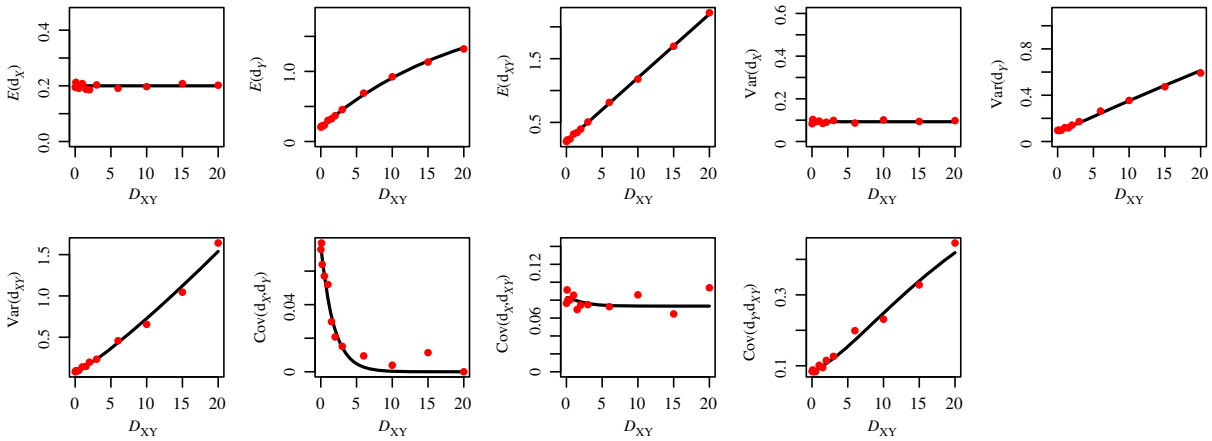


Figure 6. Assessing the accuracy of average pairwise difference results, $2\mu\eta_X = 1$, $\eta_Y = 10\eta_X$. We compare our theoretical results based on coalescence theory using equations presented here (black line) with empirical estimates using 1500 independently simulated gene sequences (red dots), $n_X = n_Y = 10$ sampled individuals. (Online version in colour.)

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, and to demonstrate when there are limitations. We leave a demonstration of the accuracy of our variance/covariance calculations in relation to the previous results derived for constant population size in Takahata & Nei [7] to the electronic supplementary material, §C.

7. Accuracy in estimating 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, equation 8) [11] presented a coalescent-based definition of F_{ST} as a function of the difference in expected time to coalescence for a collection of subpopulations. Specializing to two sub populations of interest, X and Y, Slatkin's F_{ST} can be expressed as

$$F_{ST} = \frac{\mathbb{E}(t_{i,j}) - (1/2)\big(\mathbb{E}(t_{i,i'}) + \mathbb{E}(t_{j,j'})\big)}{\mathbb{E}(t_{i,j})}, \tag{7.1}$$

where i, i' are from population X, and j, j' are individuals sampled from population Y. This definition of F_{ST} relies on a ratio of estimates of average coalescence times, where average pairwise differences in DNA sequence data are used as the proxy to estimate the unknown coalescence times. Discussed in Slatkin and Hudson *et al.* [11,26], for two populations X and Y, F_{ST} can be estimated from a non-recombining portion of the genome using

$$F_{ST} \approx \frac{d_{XY} - (1/2)(d_X + d_Y)}{d_{XY}} \overset{\text{define}}{=} F_{ST}^{G}. \tag{7.2}$$

For the sake of this paper, we differentiate F_{ST} and F_{ST}^{G} as the exact measurement from unobservable coalescence times and the estimate from pairwise differences across multiple sequences, respectively. As we have shown above, the expectation, variance and covariance of these sample average pairwise differences contained in equation (7.2) can be derived using coalescent theory, for a given mutation parameter μ and sample sizes. We can use these to study the accuracy of the F_{ST}^{G} estimator to Slatkin's 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

$$\begin{aligned}\mathbb{E}(F_{ST}^{G}) &\neq \frac{\mathbb{E}(d_{XY}) - (1/2)(\mathbb{E}(d_X) + \mathbb{E}(d_Y))}{\mathbb{E}(d_{XY})} \\ &= \frac{2\mu\mathbb{E}(t_{i,j}) + \mu\big(\mathbb{E}(t_{i,i'}) + \mathbb{E}(t_{j,j'})\big)}{2\mu\mathbb{E}(t_{i,j})} = F_{ST}.\end{aligned} \tag{7.3}$$

This implies that the estimator F_{ST}^G is potentially a biased estimator of F_{ST}, such that $F_{ST} - \mathbb{E}(F_{ST}^G) \neq 0$. To study this bias, we need an expression for the mean of F_{ST}^G. 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 $\mathbb{E}(F_{ST}^G)$

$$\begin{aligned}\mathbb{E}(F_{ST}^G) &= \mathbb{E}\left(\frac{d_{XY} - (1/2)(d_X + d_Y)}{d_{XY}}\right) \\ &= 1 - \frac{1}{2}\mathbb{E}\left(\frac{d_X + d_Y}{d_{XY}}\right)\end{aligned} \tag{7.4}$$

and

$$\begin{aligned}\mathrm{Var}(F_{ST}^G) &= \mathrm{Var}\left(\frac{d_{XY} - \frac{1}{2}(d_X + d_Y)}{d_{XY}}\right) \\ &= \frac{1}{4}\mathrm{Var}\left(\frac{d_X + d_Y}{d_{XY}}\right).\end{aligned} \tag{7.5}$$

We are now interested in the mean and variance of the ratio $(d_X + d_Y)/d_{XY}$. As generally discussed in Stuart & Kendall [28], we can use a second-order Taylor expansion of $f(A, B) = A/B$ 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. We can approximate the mean as

$$\begin{aligned}\mathbb{E}\left(\frac{d_X + d_Y}{d_{XY}}\right) \approx{}& \frac{\mathbb{E}(d_X) + \mathbb{E}(d_Y)}{\mathbb{E}(d_{XY})} + \frac{\mathbb{E}(d_X) + \mathbb{E}(d_Y)}{\mathbb{E}(d_{XY})^3}\mathrm{Var}(d_{XY}) \\ &- \frac{1}{\mathbb{E}(d_{XY})^2}[\mathrm{Cov}(d_X, d_{XY}) + \mathrm{Cov}(d_Y, d_{XY})].\end{aligned} \tag{7.6}$$

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

$$\begin{aligned}\mathbb{E}(F_{ST}^G) &= 1 - \frac{1}{2}\mathbb{E}\left(\frac{d_X + d_Y}{d_{XY}}\right) \\ &\approx F_{ST} + \frac{1}{2\mathbb{E}(d_{XY})^2}(\mathrm{Cov}(d_X, d_{XY}) + \mathrm{Cov}(d_Y, d_{XY}) \\ &\quad - \frac{\mathbb{E}(d_X) + \mathbb{E}(d_Y)}{\mathbb{E}(d_{XY})}\mathrm{Var}(d_{XY})\Big).\end{aligned} \tag{7.7}$$

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

$$\begin{aligned}\mathbb{E}(F_{ST}^G) - F_{ST} \approx{}& \frac{1}{2\mathbb{E}(d_{XY})^2}(\mathrm{Cov}(d_X, d_{XY}) + \mathrm{Cov}(d_Y, d_{XY}) \\ &- \frac{\mathbb{E}(d_X) + \mathbb{E}(d_Y)}{\mathbb{E}(d_{XY})}\mathrm{Var}(d_{XY})\Big).\end{aligned} \tag{7.8}$$

Similarly, we can get a first-order approximation for the variance of F_{ST}^G:

$$\begin{aligned}\mathrm{Var}(F_{ST}^G) &= \frac{1}{4}\mathrm{Var}\left(\frac{d_X + d_Y}{d_{XY}}\right) \\ &\approx \frac{1}{4}\Bigg(\frac{\mathrm{Var}(d_X) + \mathrm{Var}(d_Y) + 2\mathrm{Cov}(d_X, d_Y)}{(\mathbb{E}(d_X) + \mathbb{E}(d_Y))^2} \\ &\quad + \frac{(\mathbb{E}(d_X) + \mathbb{E}(d_Y))^2}{\mathbb{E}(d_{XY})^4}\mathrm{Var}(d_{XY}) \\ &\quad - 2\frac{\mathbb{E}(d_X) + \mathbb{E}(d_Y)}{\mathbb{E}(d_{XY})^3}(\mathrm{Cov}(d_X, d_{XY}) + \mathrm{Cov}(d_Y, d_{XY}))\Bigg).\end{aligned} \tag{7.9}$$

Figure 7 shows the accuracy of the two Taylor approximations under a constant population size model for mutation rate $2\mu\eta_X = 1$. 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 F_{ST}^G well with this method, and do not pursue this aspect further. Electronic supplementary material, figures S11 and S12, demonstrate the accuracy of the Taylor approximations under alternate mutation rates, and it can be seen that the approximation to $\mathbb{E}(F_{ST}^G)$ breaks down under a 10× reduction in the mutation rate ($2\mu\eta_X = 0.1$) due to the high variance in estimating variance/covariance terms of the d statistics.

In what follows, we will evaluate the bias in the F_{ST}^G estimator of F_{ST} under different demographic and genetic parameters, using the approximation given in equation (7.7).

(a) Results for the mean and bias of F_{ST}^G

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

$$\begin{aligned}F_{ST} &= \frac{\mathbb{E}(d_{XY}) - (1/2)(\mathbb{E}(d_X) + \mathbb{E}(d_Y))}{\mathbb{E}(d_{XY})} = 1 - \frac{1}{2}\frac{\mathbb{E}(d_X) + \mathbb{E}(d_Y)}{\mathbb{E}(d_{XY})} \\ &= 1 - \frac{1}{2}\frac{\mathbb{E}(t_{i,i'}) + \mathbb{E}(t_{j,j'})}{\mathbb{E}(t_{i,j})}.\end{aligned} \tag{7.10}$$

Immediately we can note that F_{ST} is not dependent on sample sizes 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, note the fundamental difference between $\mathbb{E}(F_{ST}^G)$ and F_{ST} is the term

$$\mathbb{E}\left(\frac{d_X + d_Y}{d_{XY}}\right) \text{ versus } \frac{\mathbb{E}(d_X) + \mathbb{E}(d_Y)}{\mathbb{E}(d_{XY})}. \tag{7.11}$$

It is known that ratio estimators are in general biased [29]. Jensen's inequality [30] tells us, for a convex function $f(t)$, that

$$\mathbb{E}(f(t)) \geq f(\mathbb{E}(t)). \tag{7.12}$$

Letting $f(t) = (d_X + d_Y)/d_{XY}$ and observing that $d_{XY} \geq 1/2(d_X + d_Y)$, the inequality implies

$$\mathbb{E}\left(\frac{d_X + d_Y}{d_{XY}}\right) \geq \frac{\mathbb{E}(d_X) + \mathbb{E}(d_Y)}{\mathbb{E}(d_{XY})}. \tag{7.13}$$

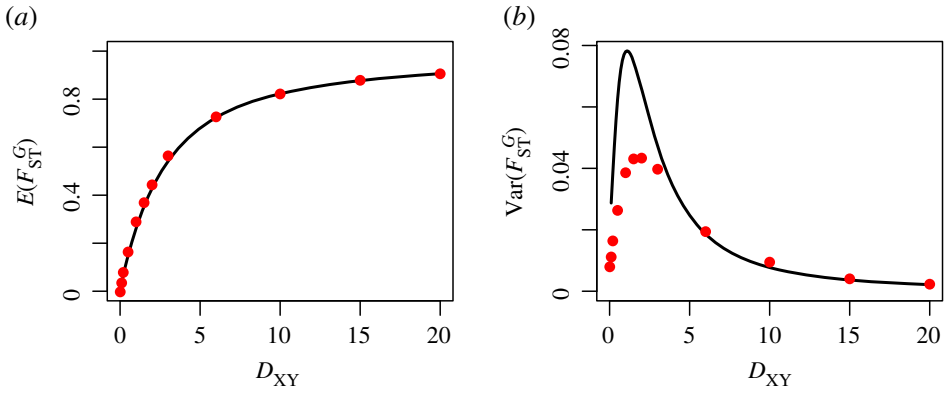


Figure 7. Accuracy of approximations to the mean and variance of F_{ST}^G, $2\mu\eta_X = 1$, $\eta_Y = \eta_X$. A comparison of the approximations in equations (7.7) and (7.9) (black curves) to values estimated from empirical simulations (red dots). (*a*) The approximated value to $\mathbb{E}(F_{ST}^G)$ as a function of divergence time, D_{XY}, for equal sample sizes n_X, n_Y accurately approximates simulated estimates. (*b*) The first-order approximation for the variance $\mathrm{Var}(F_{ST}^G)$ as a function of D_{XY} is a poor approximation for more recent divergence time models. (Online version in colour.)

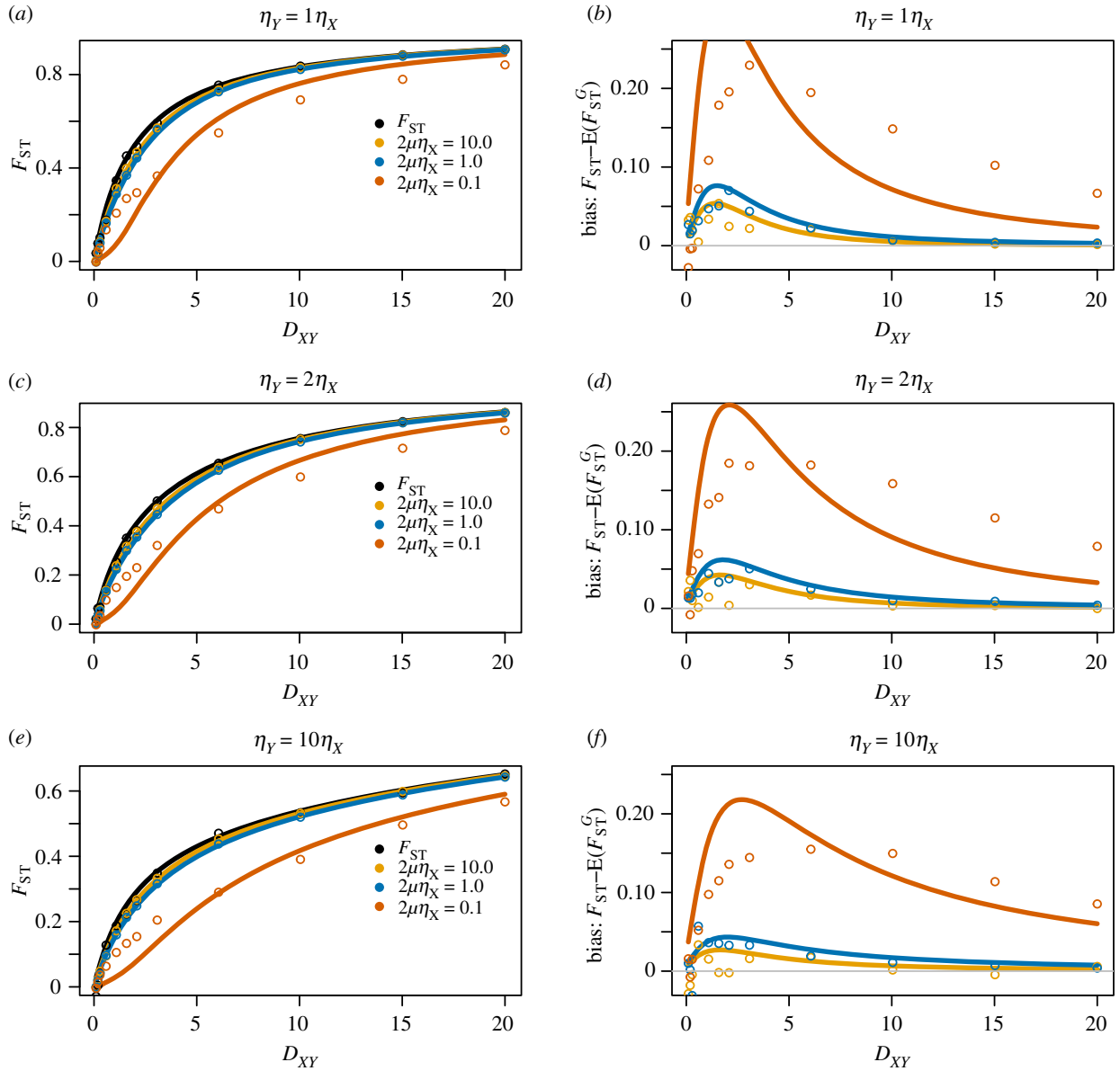


Figure 8. F_{ST} approximation bias using $\mathbb{E}(F_{ST}^G)$ across divergence times. Under varying population size scenarios (rows), we study the difference between theoretical F_{ST} and the expected estimate calculated from pairwise differences, $\mathbb{E}(F_{ST}^G)$, to highlight the potential biases in doing so. (*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 black, and approximations $\mathbb{E}(F_{ST}^G)$ using equation (7.7) under three mutation rates. (*b*,*d*,*e*) The difference between the true F_{ST} (black line in adjacent plot) and the expected sample quantity, to represent the bias in estimation. We simulated assuming equal sample sizes $n_X = n_Y = 10$. In all figures, dots represent simulated estimates from 1500 independent genes. (Online version in colour.)

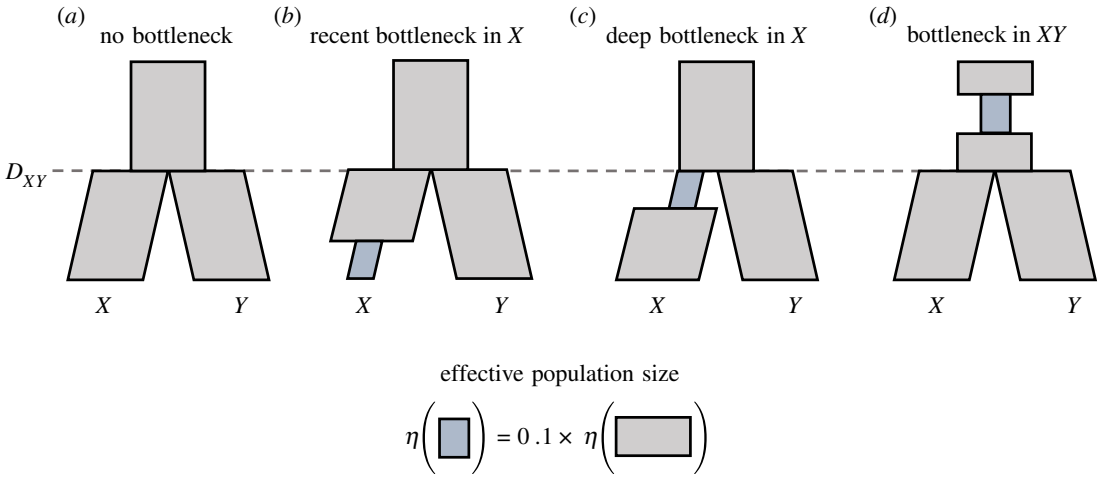


Figure 9. Bottlenecks considered in a tree of two species. (*a*) Constant population size tree with no bottleneck. (*b*) A bottleneck occurring in the recent history of species *X*. (*c*) A bottleneck which occurred directly after speciation in population *X*. (*d*) A bottleneck which occurred in the population ancestral to both *X* and *Y*. In all trees, the non-bottleneck population sizes are a fixed constant. (Online version in colour.)

Thus we expect $\mathbb{E}(F_{ST}^G)$ to be a negatively biased estimate of F_{ST}. 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 $\mathbb{E}(F_{ST}^G)$ to become increasingly closer to F_{ST}. Also, letting the number of mutations increase in an infinite-sites model, the estimates of d_X, d_Y and d_{XY} become closer to their expectations, bringing equation (7.13) closer to equality. Figure 8 demonstrates the relationship between $\mathbb{E}(F_{ST}^G)$ and F_{ST} under varying divergence times D_{XY}, population sizes and mutation rates μ. As discussed above, the relative bias of F_{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 three times as large as $\mathbb{E}(F_{ST}^G | 2\mu\eta_X = 0.1)$. It is clear that F_{ST}^G is a faithful estimator of F_{ST} under very high mutation rates, however, it is biased downward for small values of μ, although the bias is reduced for deep divergence models. When estimating F_{ST} from multiple genes across the genome, one approach used to reduce the estimation bias is to estimate each term in equation (7.10) individually and apply a 'ratio of averages' approach [31], as further highlighted in the discussion.

(b) Effect of bottleneck timing on F_{ST}

Population bottlenecks can drastically affect the genetic diversity of populations over evolutionarily short periods of time. In the context of F_{ST}, the question of when a bottleneck occurred in a history of evolution is key in understanding its impact on population differentiation. In this section, we use the flexibility of STCov to explore the effect of a population bottleneck placed at various times in the history of two theoretical species, X and Y, on F_{ST}. Here, we model a population bottleneck as a $10\times$ reduction in the population size η_0 for a fixed length of time (1.0 in units of $2\eta_0$ generations). We study four scenarios as described in figure 9. For varying divergence times D_{XY}, we use STCov to calculate F_{ST} under each scenario, and use empirical simulations via ms and SeqGen to validate our results. We find that a recent bottleneck has the largest impact on F_{ST} at every divergence time tested (figure 10), demonstrating an increased level of differentiation as compared to the scenario with no bottleneck. Both scenarios of deeper bottlenecks have much less effect on overall F_{ST} despite their bottlenecks being identical in size and length. This illustrates that the timing of variation-reducing events such as a bottleneck plays a large role in the impact to measured genetic differentiation using F_{ST}, where the impact can be effectively lost given sufficient time post-bottleneck.

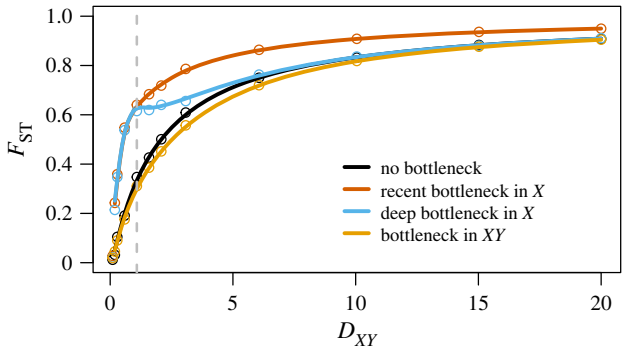


Figure 10. The effect of different population bottlenecks on F_{ST}. Four different bottleneck scenarios were considered in the genetic history of two species *X* and *Y*, as described in figure 9. Curves represent theoretical results from STCov, open circles are empirical estimates from 1500 independently simulated sequences under $2\mu\eta_X = 1.0$ mutation rate. Note that as bottleneck lengths in *X* were fixed to be min (1.0, D_{XY}), for $D_{XY} \leq 1$, the population histories of recent and deep bottlenecks in *X* are identical. The vertical dashed line at $D_{XY} = 1$ indicates this boundary. (Online version in colour.)

(c) 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 [32] first derived the relationship between F_{ST} to estimate Nm in an Island model, 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 *et al.* [26] used this relationship to estimate Nm using the following expression:

$$\langle Nm \rangle_F = \frac{1}{2}\left(\frac{1}{F_{ST}} - 1\right), \tag{7.14}$$

where F_{ST} is an estimate from sequence data, i.e. F_{ST}^G in our notation. The results of the simulations presented there

show estimates using $\langle Nm\rangle_F$ are upward-biased using an estimate of F_{ST} from sequence data in place of the unknown F_{ST} based on coalescence times. There are two potential sources of this bias, the estimator function, $\langle Nm\rangle_F$, and the estimate, F^G_{ST}. The scope of this study concerns the role of estimator F^G_{ST}, and we can investigate 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 $\langle Nm\rangle_F$ when an estimate of F_{ST} is used.

To start, we can once again use a Taylor expansion to get an approximation for the expected value of $\langle Nm\rangle_F$, when using F^G_{ST}

$$\begin{aligned}\mathbb{E}(\langle Nm\rangle_F) = \frac{1}{4}\mathbb{E}\left(\frac{d_X + d_Y}{d_{XY} - (1/2)(d_X + d_Y)}\right) &\approx \frac{1}{4}\frac{\mathbb{E}(d_X) + \mathbb{E}(d_Y)}{\mathbb{E}(d_{XY}) - (1/2)(\mathbb{E}(d_X) + \mathbb{E}(d_Y))} \\ &\times \left[1 - \frac{\mathrm{Cov}(d_X, d_{XY}) + \mathrm{Cov}(d_Y, d_{XY}) - (1/2)(\mathrm{Var}(d_X) + \mathrm{Var}(d_Y)) - \mathrm{Cov}(d_X, d_Y)}{(\mathbb{E}(d_X) + \mathbb{E}(d_Y))(\mathbb{E}(d_{XY}) - (1/2)(\mathbb{E}(d_X) - \mathbb{E}(d_Y)))}\right. \\ &\left. + \frac{\mathrm{Var}(d_{XY}) + (1/4)(\mathrm{Var}(d_X) + \mathrm{Var}(d_Y) + 2\mathrm{Cov}(d_X, d_Y)) - \mathrm{Cov}(d_X, d_{XY}) - \mathrm{Cov}(d_Y, d_{XY})}{(\mathbb{E}(d_{XY}) - (1/2)(\mathbb{E}(d_X) - \mathbb{E}(d_Y)))^2}\right].\end{aligned} \tag{7.15}$$

We can use this expression to study the difference between using the estimator, F^G_{ST}, and the (unknown) true value, F_{ST}, in the expression for $\langle Nm\rangle_F$. Figure 11 shows the difference between using F_{ST} and F^G_{ST} in $\langle Nm\rangle_F$ under different mutation rates, population sizes and species divergence times. From the figure, we see that the expectations are, in fact, overestimates. In this figure, 10 individuals are sampled from each population. 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 which would have been obtained using an accurate estimate of F_{ST}. For high mutation rates, this bias decreases rapidly as D_{XY} increases. For a low mutation rate, $2\mu\eta_X = 0.01$, 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 Hudson *et al.* [26], that using an estimate of F_{ST} can result in an artificial increase in the function $\langle Nm\rangle_F$.

(d) 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 Cavalli–Sforza [25], and later Nielsen *et al.* [24], is that given an estimate of F_{ST}, D_{XY} can be estimated by the transformation

$$\hat{D}_{XY} \propto -log(1 - F^G_{ST}). \tag{7.16}$$

Another commonly used transformation, presented in Slatkin [33], relates the time of divergence to a ratio of F_{ST} values

$$\hat{D}_{XY} \propto \frac{F^G_{ST}}{1 - F^G_{ST}}. \tag{7.17}$$

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 $\mathrm{Var}(F^G_{ST})$, we can only make a first-order approximation of equation (7.16) such that

$$\mathbb{E}(-\log(1 - F^G_{ST})) \approx -\log(1 - \mathbb{E}(F^G_{ST})). \tag{7.18}$$

For equation (7.17), by plugging in the estimator for F_{ST} from equation (7.2), we find

$$\frac{F^G_{ST}}{1 - F^G_{ST}} = 2\frac{d_{XY}}{d_X + d_Y} - 1.$$

Taking the expectation of this quantity

$$\mathbb{E}\left(\frac{F^G_{ST}}{1 - F^G_{ST}}\right) = 2\mathbb{E}\left(\frac{d_{XY}}{d_X + d_Y}\right) - 1. \tag{7.19}$$

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

$$\begin{aligned}\mathbb{E}\left(\frac{d_{XY}}{d_X + d_Y}\right) \approx \frac{\mathbb{E}(d_{XY})}{\mathbb{E}(d_X) + \mathbb{E}(d_Y)} &- \frac{\mathrm{Cov}(d_X, d_{XY}) + \mathrm{Cov}(d_Y, d_{XY})}{(\mathbb{E}(d_X) + \mathbb{E}(d_Y))^2} \\ &+ \frac{\mathrm{Var}(d_X) + \mathrm{Var}(d_Y) + 2\mathrm{Cov}(d_X, d_Y)}{(\mathbb{E}(d_X) + \mathbb{E}(d_Y))^3}\mathbb{E}(d_{XY}),\end{aligned} \tag{7.20}$$

and we have a second-order Taylor approximation of the expectation of equation (7.17).

In figure 12, we evaluate the linearity between these expressions and divergence time, and the accuracy of our approximations against simulated data (line versus dots), under two different population size models. It is clear that Slatkin's [33] 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 expectedly disappears. The log transformation of Nielsen *et al.* and Cavalli-Sforza [24,25] performs worse and can only be used as a local-linear approximation. Across large values of D_{XY}, it demonstrates clear nonlinear behaviour and Slatkin's [33] transformation is preferable under the conditions investigated here.

8. Discussion

In this study, we have derived the equations and recursions needed to calculate exact values for the covariance between pairs of coalescence times in a species tree model, allowing for piecewise constant changes in population sizes throughout the tree. Using these expressions, we are able to build on previous theory to get exact values for the mean, variance

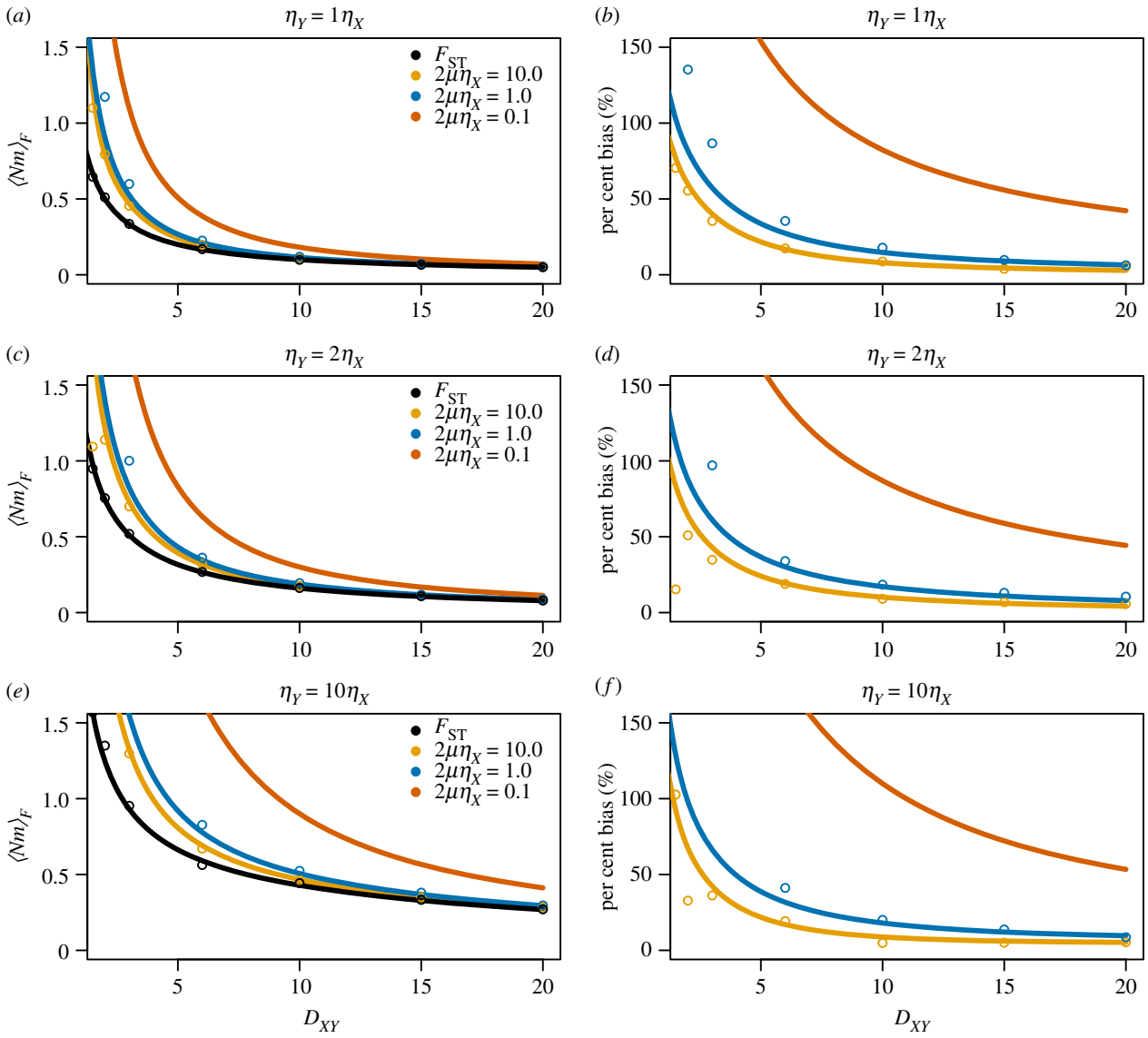


Figure 11. $\langle Nm \rangle_F$ approximation bias across divergence times and mutation rates. Under varying population size scenarios (rows), we demonstrate the difference between theoretical $\langle Nm \rangle_F$ and the expected estimate when calculating from pairwise differences using equation 7.15. (*a*,*c*,*e*) On the *y*-axis are values $\langle Nm \rangle_F$ as functions of divergence time D_{XY}. We plot the value when using the true F_{ST}, and approximations $\mathbb{E}(\langle Nm \rangle_F | 2\mu\eta_X)$, for mutation rates $2\mu\eta_X = 10.0, 1.0$ and 0.1. (*b*,*d*,*f*) The per cent difference between $\langle Nm \rangle_F$ using F_{ST} (black line in *a*,*c*,*e*) 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}.

and covariance of the average number of 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 covariance results of Takahata & Nei [7]. The equations and recursions derived here are implemented in a freely available software package, STCov, which allows for exact calculations under any piecewise constant model of divergence for arbitrary numbers of species/populations. While the covariance results presented here are interesting on their own, we imagine there are many further applications of the summary statistics presented here.

One such application we explored is the properties of Slatkin's F_{ST} and its approximation using sequence data, F_{ST}^G, under a divergence model. Under the infinite-sites model with no recombination, we demonstrate the known negative bias in estimating F_{ST} using sequence data and the 'average of ratios' approach. We show that the magnitude of the bias is a function of both mutation rate and population divergence time, with the amount of bias decreasing as both mutation rates and divergence times increase. The bias, however, is non-vanishing for low mutation rates, even as simulated divergence time increases, and is further exaggerated for imbalanced population sizes. As such, the results of the transformation for F_{ST} used for gene-flow estimation can be biased upwards when using empirical estimates, which reaffirms discussion in Hudson *et al.* [26] and provides further insight to the source of the bias. We therefore advocate that when looking at F_{ST} in a gene-by-gene fashion, such as when performing local F_{ST} scans, to consider that empirical estimates of Slatkin's F_{ST} are generally accurate for high values of mutation and deep divergence, but warn against its over-interpretation in low mutation or recent divergence scenarios, where the F_{ST} estimate can be uninformative. We recommend using equation (7.8) to estimate the expected level of bias upon application.

Throughout the theoretical equations presented here, we assumed an infinite-sites model of mutation with no recombination between sites. However, allowing for recombination

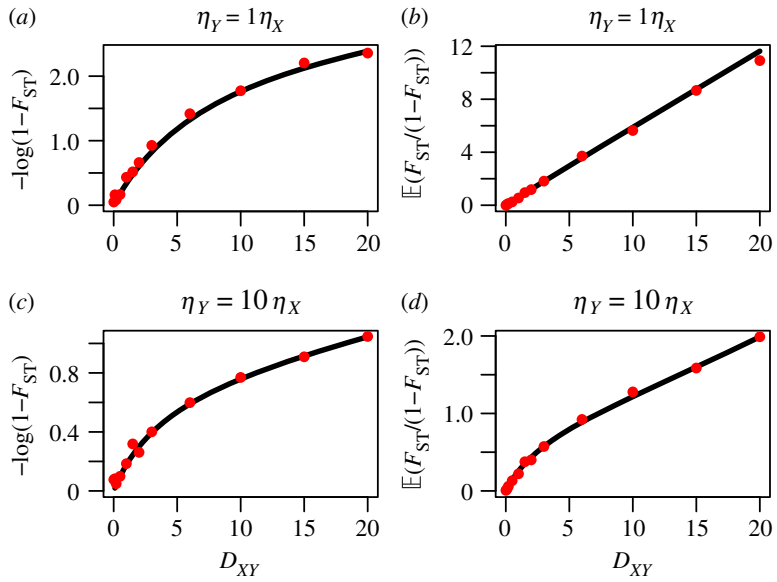


Figure 12. Linearized F_{ST} estimates. Testing the linearity of two F_{ST} transformations plotted against species divergence time, D_{XY}. On the left (*a*,*c*) is the approximate mean log transformed value. On the right (*b*,*d*) is the approximated mean fraction transformed value. Both use F_{ST}^G as a proxy for the unknown F_{ST}. Plotted on the *x*-axis of all 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. (Online version in colour.)

between sites provides more stable estimates of the expectations of pairwise differences. As discussed in Wakeley [15], allowing for an increasing amount of recombination between loci decreases the error in estimates of expectations of d_X, d_Y and d_{XY}. At the limit of infinitely free recombination between loci, estimates of equation (7.13) tend towards equality and thus the estimator $\mathbb{E}(F_{ST}^G)$ would converge to the value of F_{ST}, mitigating the negative bias seen here. Therefore, aligning with conclusions drawn in Bhatia *et al.* [31], in the age of whole-genome estimates of F_{ST}, taking a 'ratio of averages' across independent loci rather than the 'average of ratios' approach to F_{ST} can sidestep the bias we have presented when estimating F_{ST} from loci across an entire genome; the former also having the advantage of being a more numerically stable estimator.

Independent of bias, our equations demonstrate that the timing of a bottleneck can drastically impact measured levels of F_{ST}. Specifically, that the impact of population variation can vanish given enough time. Finally, we study the accuracy of a couple of commonly used linear transformations of F_{ST} as approximate measures of population divergence times, and find, for equal population sizes, the estimator proposed in Slatkin [33] has the best performance, but when population sizes are no longer equal, expectedly, even this transformation shows deviations from linearity.

There are many interesting properties to study with the covariance of pairwise coalescent times and pairwise differences. We hope that the software provided, STCov, will allow for further investigation into the properties and usefulness of these quantities for estimating various aspects of species trees, such as topology reconstruction, divergence time and population size estimation, gene flow and admixture detection.

9. Software availability

Along with this manuscript, we provide software (implemented in C++) freely 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. Download the code at https://github.com/gaguerra/STCov.

Data accessibility. All scripts used in this study are openly accessible through https://github.com/StochasticBiology/boolean-efflux.git. The data are provided in electronic supplementary material [34].

Authors' contributions. G.G.: conceptualization, data curation, formal analysis, investigation, methodology; R.N.: conceptualization, formal analysis, funding acquisition, investigation, methodology and project administration.

Both authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Competing interests. We declare we have no competing interests.

Funding. This research was supported by NIH grant no. R01GM138634 to R.N.

Acknowledgements. The authors thank Montgomery Slatkin for helpful discussions and comments.

References

1. Kingman JFC. 1982 The coalescent. *Stoch. Proc. Appl.* **13**, 235–248. (doi:10.1016/0304-4149(82)90011-4)
2. Maddison WP. 1997 Gene trees in species trees. *Syst. Biol.* **46**, 523–536. (doi:10.1093/sysbio/46.3.523)
3. Rosenberg NA. 2002 The probability of topological concordance of gene trees and species trees. *Theor. Popul. Biol.* **61**, 225–247. (doi:10.1006/tpbi.2001.1568)

4. Degnan JH, Rosenberg NA. 2006 Discordance of species trees with their most likely gene trees. *PLoS Genet.* **2**, e68. (doi:10.1371/journal.pgen.0020068)
5. Rosenberg NA, Tao R. 2008 Discordance of species trees with their most likely gene trees: the case of five taxa. *Syst. Biol.* **57**, 131–140. (doi:10.1080/10635150801905535)
6. Lewontin RC. 1972 The apportionment of human diversity. In *Evolutionary biology* (eds T Dobzhansky, MK Hecht, WC Steere), pp. 381–398. Berlin, Germany: Springer.
7. Takahata N, Nei M. 1985 Gene genealogy and variance of interpopulational nucleotide differences. *Genetics* **110**, 325–344. (doi:10.1093/genetics/110.2.325)
8. Nei M, Li WH. 1979 Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl Acad. Sci. USA* **76**, 5269–5273. (doi:10.1073/pnas.76.10.5269)
9. Kimura M. 1971 Theoretical foundation of population genetics at the molecular level. *Theor. Popul. Biol.* **2**, 174–208. (doi:10.1016/0040-5809(71)90014-1)
10. Watterson G. 1975 On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–276. (doi:10.1016/0040-5809(75)90020-9)
11. Slatkin M. 1991 Inbreeding coefficients and coalescence times. *Genet. Res.* **58**, 167–175. (doi:10.1017/S0016672300029827)
12. Tajima F. 1983 Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460. (doi:10.1093/genetics/105.2.437)
13. Wakeley J. 1996 Pairwise differences under a general model of population subdivision. *J. Genet.* **75**, 81–89. (doi:10.1007/BF02931753)
14. Wakeley J. 1996 The variance of pairwise nucleotide differences in two populations with migration. *Theor. Popul. Biol.* **49**, 39–57. (doi:10.1006/tpbi.1996.0002)
15. Wakeley J. 1997 Using the variance of pairwise differences to estimate the recombination rate. *Genet. Res.* **69**, 45–48. (doi:10.1017/S0016672396002571)
16. Tang H, Siegmund DO, Shen P, Oefner PJ, Feldman MW. 2002 Frequentist estimation of coalescence times from nucleotide sequence data using a tree-based partition. *Genetics* **161**, 447–459. (doi:10.1093/genetics/161.1.447)
17. Efromovich S, Kubatko LS. 2008 Coalescent time distributions in trees of arbitrary size. *Stat. Appl. Genet. Mol. Biol.* **7**, Article 2. (doi:10.2202/1544-6115.1319)
18. Wilkinson-Herbots HM. 2008 The distribution of the coalescence time and the number of pairwise nucleotide differences in the 'isolation with migration' model. *Theor. Popul. Biol.* **73**, 277–288. (doi:10.1016/j.tpb.2007.11.001)
19. Wilkinson-Herbots HM. 2012 The distribution of the coalescence time and the number of pairwise nucleotide differences in a model of population divergence or speciation with an initial period of gene flow. *Theor. Popul. Biol.* **82**, 92–108. (doi:10.1016/j.tpb.2012.05.003)
20. Heled J. 2012 Sequence diversity under the multispecies coalescent with Yule process and constant population size. *Theor. Popul. Biol.* **81**, 97–101. (doi:10.1016/j.tpb.2011.12.007)
21. Liu L, Yu L, Pearl DK, Edwards SV. 2009 Estimating species phylogenies using coalescence times among sequences. *Syst. Biol.* **58**, 468–477. (doi:10.1093/sysbio/syp031)
22. Mossel E, Roch S. 2008 Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **7**, 166–171. (doi:10.1109/TCBB.2008.66)
23. Jewett EM, Rosenberg NA. 2012 iGLASS: an improvement to the GLASS method for estimating species trees from gene trees. *J. Comput. Biol.* **19**, 293–315. (doi:10.1089/cmb.2011.0231)
24. Nielsen R, Mountain JL, Huelsenbeck JP, Slatkin M. 1998 Maximum-likelihood estimation of population divergence times and population phylogeny in models without mutation. *Evolution* **52**, 669–677. (doi:10.1111/evo.1998.52.issue-3)
25. Cavalli-Sforza LL. 1969 Human diversity. In *Proc. 12th Int. Congr. Genet., Tokyo*, vol. 3, pp. 405–416.
26. Hudson RR, Slatkin M, Maddison WP. 1992 Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**, 583–589. (doi:10.1093/genetics/132.2.583)
27. Hudson RR. 2002 Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–338. (doi:10.1093/bioinformatics/18.2.337)
28. Stuart A, Kendall MG. 1963 *The advanced theory of statistics*. London, UK: Griffin.
29. David IP, Sukhatme B. 1974 On the bias and mean square error of the ratio estimator. *J. Am. Stat. Assoc.* **69**, 464–466. (doi:10.1080/01621459.1974.10482975)
30. Jensen JLWV. 1906 Sur les fonctions convexes et les inegalites entre les valeurs moyennes. *Acta Math.* **30**, 175–193. (doi:10.1007/BF02418571)
31. Bhatia G, Patterson N, Sankararaman S, Price AL. 2013 Estimating and interpreting F_{ST}: the impact of rare variants. *Genome Res.* **23**, 1514–1521. (doi:10.1101/gr.154831.113)
32. Wright S. 1949 The genetical structure of populations. *Ann. Eugen.* **15**, 323–354. (doi:10.1111/j.1469-1809.1949.tb02451.x)
33. Slatkin M. 1993 Isolation by distance in equilibrium and non-equilibrium populations. *Evolution* **47**, 264–279. (doi:10.1111/evo.1993.47.issue-1)
34. Guerra G, Nielsen R. 2022 Covariance of pairwise differences on a multi-species coalescent tree and implications for F_{ST}. Figshare.