# A non-zero variance of Tajima's estimator for two sequences even for infinitely many unlinked loci

Léandra King[1], John Wakeley[1], and Shai Carmi*[2]

[1]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA
[2]Braun School of Public Health and Community Medicine, The Hebrew University of Jerusalem, Israel

## Abstract

The population-scaled mutation rate, $\theta$, is informative on the effective population size and is thus widely used in population genetics. We show that for two sequences and $n$ unlinked loci, Tajima's estimator ($\hat{\theta}$), which is the average number of pairwise differences, is not consistent and therefore its variance does not vanish even as $n \to \infty$. The non-zero variance of $\hat{\theta}$ results from a (weak) correlation between coalescence times even at unlinked loci, which, in turn, is due to the underlying fixed pedigree shared by all genealogies. We derive the correlation coefficient under a diploid, discrete-time, Wright-Fisher model, and we also derive a simple, closed-form lower bound. We also obtain empirical estimates of the correlation of coalescence times under demographic models inspired by large-scale human genealogies. While the effect we describe is small ($\mathrm{Var}\left[\hat{\theta}\right]/\theta^2 \approx \mathcal{O}\left(N_e^{-1}\right)$), it is important to recognize this feature of statistical population genetics, which runs counter to commonly held notions about unlinked loci.

**Keywords.** Coalescent Theory; Recombination; Heterozygosity; Effective Population Size; Pedigrees; Genealogies; Markov Chains

## 1 Introduction

The population mutation rate, $\theta$, is defined as $4N_e\mu$, where $N_e$ is the effective population size and $\mu$ is the mutation rate per locus per generation. Two classic estimators were developed for $\theta$, Watterson's (based on the number of segregating sites (Watterson, 1975)) and Tajima's (based on the average number of pairwise differences (Tajima, 1983, 1989)). For a single pair of sequences, both estimators are identical (denoted here as $\hat{\theta}$) and equal to the number of differences between the sequences.

Increasing the number of sampled individuals has limited ability to improve these estimates of $\theta$, because shared ancestry reduces the number of independent branches on which mutations can arise (Rosenberg and Nordborg, 2002). Felsenstein (2006) showed that the variance of maximum likelihood estimates of $\theta$ decreases approximately logarithmically with the number of individuals sampled. In contrast, the variance decreases inversely with the number of independent loci. Thus, to increase the accuracy of estimates of $\theta$, it is generally more effective to increase the number of independent loci than the sample size at each locus (see also e.g., (Pluzhnikov and Donnelly, 1996) and references within).

Consider a set of $n$ unlinked loci located on different (non-homologous) chromosomes. We show here that even as $n \to \infty$, the variance of the resulting estimate of $\theta$ does not converge to zero, in contrast to what we may have naïvely assumed. This behavior results from the fact that coalescence times, even at unlinked loci, are in fact weakly correlated, due to the sharing the same fixed underlying pedigree across all genealogies at all loci (Wakeley et al., 2012). By conditioning on the number of shared genealogical common ancestors, we derive a simple lower bound, as a function of $N_e$, on the variance of $\hat{\theta}$.

Unlinked loci may also be sampled from the same chromosome, separated by an infinitely high recombination rate. The correlation of coalescence times in such a case is higher, as the two loci may travel together for the first few generations. Therefore, the extent of the correlation, and thereby, the variance of $\hat{\theta}$, also depend on the *sampling configuration*. We derive the correlation coefficient analytically, as a function of the configuration and the effective population size, using a diploid discrete time Wright-Fisher model (DDTWF). This model is an extension of the haploid DTWF model, previously advocated by Bhaskar et al. (2014) for the study of large samples from finite populations.

Our results for the variance of $\hat{\theta}$ were obtained under the Wright-Fisher demographic model. To shed light on the variance of $\hat{\theta}$ under more realistic demographic models, we run simulations based on real, large-scale human genealogical data (Erlich, 2016). The pedigrees inspired by different human populations differ from each other and from the Wright Fisher pedigrees in a number of ways, for example in the variance of the relatedness of any two randomly chosen individuals. These differences lead to differences in the variance of $\hat{\theta}$ for each population, even

*shai.carmi@huji.ac.il

if they have the same effective population size.

## 2 The relation of the variance of $\hat{\theta}$ to the correlation of the coalescence times

For a sample of size two at $n$ loci, the estimator of $\theta$ can be expressed as

$$\hat{\theta}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i, \quad (1)$$

where $\hat{\theta}_i$ is the number of differences at locus $i$. If we assume the loci are exchangeable, we have:

$$\mathrm{Var}\left[\hat{\theta}_{(n)}\right] = \frac{\mathrm{Var}\left[\hat{\theta}_i\right]}{n} + \frac{n-1}{n}\mathrm{Cov}\left[\hat{\theta}_i, \hat{\theta}_j\right] \;;\; i \neq j. \quad (2)$$

Under the standard coalescent model (Kingman, 1982), $\hat{\theta}_i$ is Poisson distributed with mean $2\mu T_i$, where $T_i$ is the time until coalescence at locus $i$ in generations and $\mu$ is the mutation rate per locus per generation. Using the law of total covariance,

$$\begin{aligned}\mathrm{Cov}\left[\hat{\theta}_i, \hat{\theta}_j\right] &= \mathrm{E}\left[\mathrm{Cov}\left[\hat{\theta}_i, \hat{\theta}_j | T_i, T_j\right]\right] \\ &+ \mathrm{Cov}\left[\mathrm{E}\left[\hat{\theta}_i | T_i, T_j\right], \mathrm{E}\left[\hat{\theta}_j | T_i, T_j\right]\right] \\ &= 4\mu^2\mathrm{Cov}\left[T_i, T_j\right], \end{aligned} \quad (3)$$

since conditional on $T_i$ and $T_j$, $\hat{\theta}_i$ and $\hat{\theta}_j$ are independent. Thus, for infinitely many sites,

$$\mathrm{Var}\left[\hat{\theta}\right] = \lim_{n\to\infty}\mathrm{Var}\left[\hat{\theta}_{(n)}\right] = 4\mu^2\mathrm{Cov}\left[T_i, T_j\right]. \quad (4)$$

Because $T_i$ is distributed exponentially with rate $1/(2N_e)$ under the standard coalescent model (Kingman, 1982; Tajima, 1983), $\mathrm{Var}\left[T_i\right] = 4N_e^2$. Since $\mathrm{Cov}\left[T_i, T_j\right] = \mathrm{Corr}\left[T_i, T_j\right] \times \mathrm{Var}\left[T_i\right]$, we can write:

$$\begin{aligned}\mathrm{Var}\left[\hat{\theta}\right] &= (4\mu N_e)^2\mathrm{Corr}\left[T_i, T_j\right] \\ &= \theta^2\mathrm{Corr}\left[T_i, T_j\right], \end{aligned} \quad (5)$$

and we focus henceforth on the correlation of $T_i$ and $T_j$. Studying the correlation instead of the covariance allows us, later on, to visually compare the results across different effective population sizes.

We note that the variance of $\hat{\theta}$ is calculated over independent repeats over the entire evolutionary process, including the generation of the population pedigree (family relationships between all individuals), as well as the gene genealogies. We elaborate below on this important point (sections 3, 5, and the Discussion).

## 3 Modeling the effect of the shared pedigree

In this section, we provide an intuitive derivation of the role of the shared underlying pedigree in generating a non-zero variance of $\hat{\theta}$.

### 3.1 Inconsistency of $\hat{\theta}$ due to the underlying pedigree

We begin with a general analysis of the inconsistency of the estimator of $\theta$. The value of $\hat{\theta}$ is a function of the pedigree that connects the two individuals in our sample, where the pedigree itself is randomly drawn from a demographic model (e.g., the Wright-Fisher model) with parameter $\theta$. If the sampled individuals happen to be more closely related than average, then $\hat{\theta}$ will tend to underestimate the true value of $\theta$. The opposite is true if the sampled individuals are less closely related than average.

Let $\delta$ be the probability that a randomly sampled pair of individuals is very closely related, for example as full siblings. Let $\epsilon$ be some arbitrary value smaller than the difference between $\theta$ and $\hat{\theta}^*$, where $\hat{\theta}^*$ is estimated from a sample of full siblings. By sampling sufficiently many loci (or gene genealogies), we could theoretically infer the common ancestry of the sampled pair to any desired accuracy. However, this would not give information about the pedigree beyond the ancestry of the sampled pair, and as the sampled pair is related more closely than average, $\hat{\theta}^*$ would underestimate $\theta$. For this fixed $\epsilon$ and $\delta$, we therefore cannot find $n$ large enough such that $\mathrm{Prob}(|\hat{\theta}_{(n)} - \theta| > \epsilon) < \delta$. This implies that there is no convergence in probability, which means that this estimate of $\theta$ is not consistent. In turn, this inconsistency implies that the variance of $\hat{\theta}_{(n)}$ does not tend to 0 as $n$ increases.

### 3.2 A lower bound on the limiting variance

Next, we derive an intuitive lower bound on the limiting variance of $\hat{\theta}$ for a sample of two loci on non-homologous chromosomes, where according to Eq. (4), we only need the covariances of $T_i$ and $T_j$. To compute these covariances, we condition on a vector of variables $\{x\} = x_1, x_2, ..., x_G$, where $x_g$ is the number of shared ancestors $g$ generations ago. The vector $\{x\}$ is, in a sense, a lower dimensional representation of the shared pedigree, and can be used to compute the probability of coalescence at each generation. For example, if $x_1 = 2$ (full siblings), then all loci have the same 25% probability of coalescing within a single generation. We only consider the first $G = \log_2 N_e$ generations, where $N_e$ is the (constant) effective population size, as it was shown that the effect of the shared pedigree is important only up to $\approx \log_2 N_e$ generations (Wakeley et al., 2012; Derrida et al., 2000; Chang, 1999). Beyond that time, almost all ancestors are shared,

and the distribution of the contribution of each ancestor to the present day sample is approximately stationary.

By the law of total covariance, we have:

$$\begin{aligned}\mathrm{Cov}\left[T_i, T_j\right] &= \mathrm{E}_{\{x\}}\left[\mathrm{Cov}\left[T_i, T_j|\{x\}\right]\right] \\ &\quad + \mathrm{Cov}_{\{x\}}\left[\mathrm{E}\left[T_i|\{x\}\right], \mathrm{E}\left[T_j|\{x\}\right]\right]. \end{aligned} \quad (6)$$

$\mathrm{E}_{\{x\}}\left[\mathrm{Cov}\left[T_i, T_j|\{x\}\right]\right] \approx 0$, because conditioning on the pedigree, the loci are independently segregating. Therefore:

$$\begin{aligned}\mathrm{Cov}\left[T_i, T_j\right] &= \mathrm{Cov}_{\{x\}}\left[\mathrm{E}\left[T_i|\{x\}\right], \mathrm{E}\left[T_j|\{x\}\right]\right] \\ &= \mathrm{Var}_{\{x\}}\left[\mathrm{E}\left[T_i|\{x\}\right]\right]. \end{aligned} \quad (7)$$

To compute $\mathrm{E}\left[T_i|\{x\}\right]$, we condition on whether coalescence has occurred in the first $G$ generations. If it has not occurred, we assume that the process then behaves just as the standard coalescent, or $\mathrm{E}\left[T_i|\text{no coal}\right] = 2N_e + G$. We can write:

$$\begin{aligned}\mathrm{E}\left[T_i|\{x\}\right] &= (2N_e + G)P\left(\text{no coal by } G|\{x\}\right) \\ &\quad + \sum_{g=1}^{G} gP\left(\text{coal at } g|\{x\}\right). \end{aligned} \quad (8)$$

As computed in Wakeley et al. (2012), the coalescence probability is roughly given by $P\left(\text{coal at } g|\{x\}\right) = \alpha(g)\prod_{g'=1}^{g-1}\left[1-\alpha(g')\right]$, where $\alpha(g) = x_g/2^{2g+1}$ and $\mathrm{Prob}\left\{\text{no coal by } G|\{x\}\right\} = \prod_{g'=1}^{G}\left[1-\alpha(g')\right]$. Since $\alpha(g) \ll 1$ (see below), we approximate $P\left(\text{coal at } g|\{x\}\right) \approx \alpha(g)$ and $P\left(\text{no coal by } G|\{x\}\right) \approx 1 - \sum_{g=1}^{G}\alpha(g)$. Thus,

$$\mathrm{E}\left[T_i|\{x\}\right] \approx (2N_e + G) - \sum_{g=1}^{G}(2N_e + G - g)\,\alpha(g) \quad (9)$$

and

$$\begin{aligned}\mathrm{Var}_{\{x\}}\left[\mathrm{E}\left[T_i|\{x\}\right]\right] &\approx \mathrm{Var}\left[\sum_{g=1}^{G}(2N_e + G - g)\,\alpha(g)\right] \\ &\approx 4N_e^2\mathrm{Var}\left[\sum_{g=1}^{G}\frac{x_g}{2^{2g+1}}\right], \end{aligned} \quad (10)$$

since $G \ll N_e$.

In Supplementary Material Section S1, we provide a numerical method to calculate the exact covariances of the $x_g$'s under a diploid, discrete-time Wright-Fisher model (see the next section for definitions). To proceed here, we assume that the $x_g$'s are independent. While the $x_g$'s are clearly positively correlated, the independence assumption allows us to derive a lower bound on $\mathrm{Cov}\left[T_i, T_j\right]$, and thereby, the variance of $\hat{\theta}$. Under that assumption, Eq. (10) becomes

$$\mathrm{Var}_{\{x\}}\left[\mathrm{E}\left[T_i|\{x\}\right]\right] \gtrsim N_e^2\sum_{g=1}^{G}\frac{\mathrm{Var}\left[x_g\right]}{2^{4g}}. \quad (11)$$

To compute the variance of $x_g$, we note that the distribution of $x_g$ is roughly hypergeometric with parameters $2^g$ potential successes (the number of ancestors of one individual), $N_e - 2^g$ potential failures (all individuals in the population who are not ancestors of that individual), and $2^g$ draws (the number of ancestors of the other individual), giving $\mathrm{Var}\left[x_g\right] \approx 2^{2g}(N_e - 2^g)^2/N_e^3$. We provide the exact distribution of the variance of $x_g$ in Supplementary Material Section S1. Substituting the hypergeometric variance in Eq. (11),

$$\mathrm{Var}_{\{x\}}\left[\mathrm{E}\left[T_i|\{x\}\right]\right] \gtrsim \frac{1}{N_e}\sum_{g=1}^{G}\frac{(N_e - 2^g)^2}{2^{2g}}. \quad (12)$$

Using $G = \log_2 N_e$, we have $\sum_{g=1}^{G}\frac{(N_e-2^g)^2}{2^{2g}} = \left(\frac{N_e^2}{3} - 2N_e + \frac{3\log N_e}{\log 8} + \frac{5}{3}\right) \approx \frac{N_e^2}{3}$ for large $N_e$, and hence, using Eq. (7),

$$\mathrm{Cov}\left[T_i, T_j\right] \gtrsim \frac{N_e}{3}. \quad (13)$$

Using Eq. (4) and $\theta = 4\mu N_e$, we finally obtain

$$\mathrm{Var}\left[\hat{\theta}\right] \gtrsim \frac{\theta^2}{12N_e}. \quad (14)$$

In summary, the variance due to the shared pedigree is of order $\theta^2/N_e$, independently of the number of regions $n$. Thus, as argued above, even for a large number of chromosomes, the variance of $\hat{\theta}$ does not decay to zero, but rather to a constant that depends on the effective population size.

To intuitively explain the non-zero variance, we note that the pedigree itself is the product of a stochastic model (Wright-Fisher or another). Thus, even a fully specified pedigree, as obtained by sampling infinitely many loci, leaves uncertainty regarding the value of $\theta$. In other words, the uncertainty in the estimate of $\theta$ results from having at hand only a single instance of a pedigree generated from the stochastic model governed by that parameter (see also Ralph (2015)).

# 4 Exact results for the correlation of the coalescence times at unlinked loci

In this section, we provide an exact derivation of the correlation of coalescence times at unlinked loci under a diploid, discrete-time, Wright-Fisher model. Further, we consider multiple sampling configurations for those loci, as explained below.

## 4.1 The sampling configurations

To compute the correlation of coalescence times at a pair of unlinked loci, we first note that there are multiple ways

by which two such loci can be sampled in two individuals (or sequences). The six *sampling configurations* are shown in Figure 1. Four of these configurations involve a sample of two individuals, and we start by describing these.

In the first configuration, the loci are located effectively infinitely far apart on the same chromosome in both individuals. This means that these loci will be coupled for the first few generations, until separated by a recombination event. Once separated, they may later back-coalesce onto the same chromosome, and again resume percolating together through the pedigree for a period of time that is expected to be short. (In the event of back-coalescence, two ancestral loci not sharing genetic material come to be located on the same chromosome, which essentially undoes the effect of recombination.) In the second configuration, the loci are on different homologous chromosomes, meaning they will necessarily be present in different parents in the immediately preceding generation, as each chromosome was inherited from a different parent. It is then also possible for them to back-coalesce in later generations. The third configuration is a mixture of the first two: the loci are located on the same chromosome in one individual, and on homologous chromosomes in the other. In the fourth configuration, the loci are sampled from non-homologous chromosomes in both individuals. This configuration is different from the previous three in that back-coalescence is not possible.

In the fifth and sixth sampling configurations, all sequences are sampled from a single individual. This is common in practice, as measuring the heterozygosity in a single individual does not require haplotype phasing. In configuration 5, we sample two loci from the same chromosome (and their pairs from the homologous chromosome). Given that each homologous chromosome must originate from a different parent, in one generation the sampled loci will transition to configuration 1 with probability 0.25, to configuration 2 with probability 0.25, and to sampling configuration 3 with probability 0.5. In sampling configuration 6, the sampled loci are on different (non-homologous) chromosomes. This configuration is reduced in one generation to sampling configuration 4, and therefore has the same correlation properties as that configuration.

## 4.2 The DDTWF model

To study the correlation of coalescence times under the different sampling configurations, we use a discrete-time Wright-Fisher (DTWF) model. This class of models has been advocated as an alternative to the coalescent when the sample size is large relative to the population size, as it can accommodate multiple and simultaneous mergers (Bhaskar et al., 2014).

In our case, we assume non-overlapping generations, a constant population size of $N_e$ *diploid* individuals, half of which are males and half of which are females, random mating between the sexes, no selection, and no migration. There are three possible events: recombination, coalescence, and back-coalescence. Because the population size is finite, combinations of these events can occur in a single generation. We also keep track of whether lineages are in the same individual or not, as this determines their trajectory in the immediately preceding generation. We refer to this model as the 2-sex DDTWF. (Later, we also consider a simplified (1-sex) DDTWF). The dynamics of this 2-sex DDTWF model can be summarized by a Markov transition matrix (Supplementary Material Section S2) with 17 states, where the initial state is one of the sampling configurations 1, 2, 3, or 5.

The model described above represents pairs of loci sampled from either the same chromosome or homologous chromosomes, as the notion of back-coalescence and recombination only applies for these cases. Nevertheless, we found that the same transition matrix applies to sampling configurations 4 and 6 (non-homologous chromosomes), albeit with a different interpretation of the states (not shown).

Given the transition matrix, we can write a system of equations using a first step analysis for all states $x$ such that $E[T_iT_j|x] > 0$:

$$\begin{aligned} E[T_iT_j|x] &= \sum_k p_{xk} E[(T_i+1)(T_j+1)|k] \\ &= 1 + \sum_k p_{xk} E[T_i|k] + \sum_k p_{xk} E[T_j|k] + \sum_k p_{xk} E[T_iT_j|k] \\ &= E[T_i|x] + E[T_j|x] + \sum_k p_{xk} E[T_iT_j|k] - 1, \end{aligned} \tag{15}$$

where $p_{xk}$ is the transition probability between states $x$ and $k$.

Solving this system of equations allows us to obtain exact results for $\mathrm{Cov}\,[T_i, T_j|x]$. As a note, $E[T_i|x]$ can be different from $E[T_j|x]$ depending on the state $x$. For example, if the pair of lineages at locus $i$ is located on two different chromosomes in the same individual, whereas the pair of lineages at locus $j$ is located in two different individuals, then $E[T_i|x] = E[T_j|x] + 1$. See more details in Supplementary Material Section S2. To obtain the correlation coefficient, we then normalize the covariance by the variance of the coalescence time at a locus, which is the same regardless of whether the lineages were sampled from the same or from different individuals. The variance can be calculated using the aforementioned system of equations with $i = j$.

Figure 2 shows the correlation coefficient of the coalescence times for each sampling configuration. The highest correlation is found for configuration 1. As the two loci are located on the same chromosome in both sampled individuals, they must have originated from the same parent in the previous generation. Therefore, both loci either both coalesce to the same parent or both do not, introducing correlation between the coalescence times. The effect of this sampling configuration then persists, as long as

there is no recombination. As $N_e$ increases, the correlation decreases, as it is much more likely for a recombination event to occur before a coalescence event. Sampling configuration 3 (two loci located far apart on the same chromosome in one individual, and on different chromosomes in the second individual) shows the lowest correlation. In fact, it is slightly negative for very small values of $N_e$, for if one of the loci coalesces in the first generation, then it is impossible for the other locus to coalesce. The correlation in other configurations is intermediate between those of configurations 1 and 3.

Figure 2 also shows results for a simplified DDTWF model, which is similar to the 2-sex DDTWF, except that individuals are monoecious and we do not keep track of whether lineages are in the same individual or not. There are fewer states in this model than in the 2-sex DDTWF, and it is therefore significantly easier to analyze. The simplified model displays a slightly higher correlation compared to the 2-sex model for $N_e \lesssim 40$, but is a good approximation otherwise (as we also show in Section 6). More details on both models are given in Supplementary Material Section S2.

# 5 Simulations

## 5.1 Wright-Fisher simulations

In this section, we use simulation of the 2-sex diploid, discrete-time Wright-Fisher model to support our analytical results from Section 3.2. To estimate the correlation coefficient of the coalescence times at two loci, we first simulate many Wright-Fisher pedigrees and sample, for each pedigree, two individuals from the current generation. We set the population size $N_e$ to be the same in every generation, with equal numbers of males and females. We then consider two loci on non-homologous chromosomes and simulate the path through the pedigree that connects the two lineages at each locus to their most recent common ancestor. In each generation and for each locus, lineages that are found in the same individual coalesce with probability 1/2, in which case the coalescence time is recorded. Loci on different chromosomes in the same individual coalesce neither in that generation nor in the previous generation.

We repeat this process multiple times for each pedigree to obtain an estimate of $\mathrm{E}[T|\mathrm{ped}]$. We then compute its variance over many simulated pedigrees to obtain $\mathrm{Var}_{\mathrm{ped}}[\mathrm{E}[T|\mathrm{ped}]]$. By the same logic as Eq. (7) above, $\mathrm{Var}_{\mathrm{ped}}[\mathrm{E}[T|\mathrm{ped}]]$ is equal to $\mathrm{Cov}[T_i, T_j]$. To obtain the correlation coefficient, we divide $\mathrm{Cov}[T_i, T_j]$ by $\mathrm{Var}[T] = \mathrm{Var}_{\mathrm{ped}}[\mathrm{E}[T|\mathrm{ped}]] + \mathrm{E}_{\mathrm{ped}}[\mathrm{Var}[T|\mathrm{ped}]]$. The simulation results are shown in Figure 3. Our analytical lower bound, which, based on Eqs. (14) and (5), can be written as $\mathrm{Corr}[T_i, T_j] \gtrsim 1/(12N)$, is well supported by the simulations, and is in fact relatively tight.

## 5.2 Simulations based on real human pedigrees

The Wright-Fisher model is only one way to generate pedigrees having a given effective population size. In real human populations, pedigrees have complex structures that depend on their geographical region. For example, there are different rates of consanguineous marriages in different countries (Bittles and Black, 2015), different distributions of the number of children per family, and different mating structures, leading to differences in the number of full-siblings and half-siblings. To gain insight on the effect of these differences on the ability to estimate $\theta$, we constructed a Wright-Fisher-like model, but which is constrained by patterns of real human pedigrees. Specifically, we used the FAMILINX database, compiled by Erlich (2016), which carries information on about 44 million individuals from different countries.

We extracted genealogical data for three countries (Kenya, Sweden, USA) from FAMILINX; these countries were arbitrarily selected among those with sufficient data. We then used these genealogies to simulate pedigrees by breaking down and reassembling small family units, as previously described for a different dataset (Wakeley et al., 2012). Specifically, we first split the genealogies into two-generational family units of children and their parents. To belong to a unit, a child must share at least one parent with at least one other child in the family unit. Because FAMILINX contains data on more than the three countries we chose, then in order not to create a bias in favor of smaller, simpler family units, we only require that the first sampled child be in the corresponding country data set. These family units then serve as building blocks to generate pedigrees with the same mating patterns and distribution of the number of children as in the reference population. Under these models, the effective population size $N_e$ is not guaranteed to equal the census population size. Therefore, to determine the effective population size for each model, we estimated $N_e$ as half the empirical average time until coalescence across randomly sampled pairs and random pedigrees. We could then fine-tune the census size, for each country, until reaching a pre-specified $N_e$. Once the pedigrees were generated, we simulated genealogies through those pedigrees as described in Section 5.1. Additional details are provided in Supplementary Material Section S3.

For each country and for a range of $N_e$'s, we then used the simulated data to compute the correlation coefficient of the coalescence times, as in Section 5.1 (i.e., $\mathrm{Var}_{\mathrm{ped}}[\mathrm{E}[T|\mathrm{ped}]]$ divided by $\mathrm{Var}[T]$). The results, shown in Figure 4, demonstrate that $\mathrm{Corr}[T_i, T_j]$, and consequently, $\mathrm{Var}\left[\hat{\theta}\right]$, vary across populations, and are higher in the FAMILINX-inspired models compared to the Wright-Fisher model. A plausible explanation is that in the Wright-Fisher model, the ratio of half siblings to full-siblings is much higher than in the human pedigrees; this implies higher variance in the degree of relatedness in

many real-world pedigrees relative to Wright-Fisher pedigrees. Therefore, it would be more difficult to estimate $\theta$ (i.e., the variance of $\hat{\theta}$ will be higher) in real-world populations than based on the expectation from the Wright-Fisher model. Further deviations are expected if we were to impose realistic first-cousin mating rates (Bittles and Black, 2015).

# 6 Linked sites and model comparisons

We have so far only studied unlinked sites; however, our analytical results for the DDTWF models can be relatively easily extended to the case of linked loci. Such an extension is important, since, for example, the covariance of coalescence times at two loci is directly related to the $r^2$ measure of linkage disequilibrium (McVean, 2002). Quantifying the behavior of different models in terms of the covariance of coalescence times can thus provide insight into the importance of certain modeling assumptions.

In the DDTWF model with linked sites, the transition probabilities are expressed in terms of the per generation recombination probability, $r$, which has been so far set to 0.5. The transition matrix of Supplementary Material Section S2 is straightforward to adapt for any $r < 0.5$, and the covariance or correlation coefficient of the coalescence times can be computed. The correlation coefficient under the 2-sex DDTWF model is plotted in Figure 5 vs the scaled recombination rate $\rho = 4N_e r$, showing perfect agreement with simulations.

These results now enable us to compare the exact 2-sex DDTWF model to the simplified DDTWF model, as well as to the coalescent with recombination and its Markovian approximations. Let $\rho = 4Nr$. Under the ancestral recombination graph (ARG) (Griffiths and Marjoram, 1997), which is the standard model for the coalescent with recombination, the covariance of coalescence times at two loci satisfies (e.g., Simonsen and Churchill (1997)),

$$\mathrm{Cov}_{\mathrm{ARG}}\left[T_i, T_j\right] = \frac{18+\rho}{18+13\rho+\rho^2}. \tag{16}$$

Under the Sequentially Markov Coalescent (SMC) (McVean and Cardin, 2005), each new genealogy (following recombination) depends only on the previous genealogy (as opposed to the ARG (Wiuf and Hein, 1999)), and the new coalescence time must differ from the previous time (no back-coalescence allowed). In this case, we have,

$$\mathrm{Cov}_{\mathrm{SMC}}\left[T_i, T_j\right] = \frac{1}{1+\rho}. \tag{17}$$

The SMC' model (Marjoram and Wall, 2006) is a variant of SMC where back-coalescence is allowed. Under SMC' (Eriksson et al., 2009; Wilton et al., 2015),

$$\mathrm{Cov}_{\mathrm{SMC'}}\left[T_i, T_j\right] = 2^{\rho/2} e^{-\rho/4} (-\rho)^{-1/2-\rho/4} \times \left[\Gamma\left(\frac{2+\rho}{4}\right) + \Gamma\left(\frac{2+\rho}{4}, -\frac{\rho}{4}\right)\right]. \tag{18}$$

(The covariances of Eq. (16)-(18) are also equal to their respective correlation coefficients, since $\mathrm{Var}\left[T\right] = 1$ under either the ARG, SMC, and SMC'). In Figure 6, we compare the correlation of $T_i$ and $T_j$ across the different models as a function of $\rho$ for $N_e = 100$ and different values of $r$. The ARG provides a very good approximation under these conditions. In turn, the SMC' model shows very slight deviations compared to the ARG, while, as previously shown, the SMC model deviates more substantially (Wilton et al., 2015).

The 2-sex DDTWF model is compared to the simplified DDTWF model in Figure 7. Compared to the full 2-sex model, the simplified model is an extremely good approximation even for $N_e$ as small as 100: the maximum difference in the correlation coefficient (across different values of $r$) between these two models was less than .005 (see also Figure 2). Therefore, the simplified model should be preferred due its much reduced complexity. For $N_e = 10$, we observe a more noticeable difference between the 2-sex and the simplified DDTWF models, with a maximal difference around .025.

# 7 Discussion

Increasing the size of the sample is known to have limited ability to improve estimates of $\theta$, as the individuals in the sample share most of their genealogy (Rosenberg and Nordborg, 2002). For this reason, it was recommended to use data from many unlinked gene loci from a small number of individuals (Felsenstein, 2006). While this intuition still holds, we have shown that the estimator of $\theta$ based on the average number of pairwise differences at many loci is not consistent and has non-zero variance, even when sampling infinitely many loci. We have provided an approximate lower bound for the variance for loci on non-homologous chromosomes, as well as exact results for diploid, discrete time Wright-Fisher models under various configurations of two sampled loci.

Fundamentally, the non-zero variance of $\hat{\theta}$ is a result the underlying pedigree shared between all loci. The shared pedigree itself is assumed to be a single draw from a random demographic process (Wright-Fisher or another), with a characteristic effective population size. Thus, even if we were able to perfectly characterize the single pedigree at hand, we cannot hope to infer with complete certainty the parameters of the demographic model. It is worth noting that one can adopt a different (philosophical) view, under which the pedigree itself is the subject of inference, and is not a product of a random demographic process (Ralph, 2015). Under such a view, there is no such thing as an estimator of the effective population size.

The analytical results in this paper are based on the Wright-Fisher model. To gain insight on the behavior of more realistic demographic models, we adapted the Wright-Fisher model according to the family structure of real human populations. The results demonstrated that the correlation of coalescence times is higher in the human-inspired models than in the WF model; therefore, $\theta$ should be more difficult to estimate than expected under the pure WF model.

When using a demographic model, it is not always clear which features of the real population are crucial (e.g., two sexes, diploidy, etc.), or whether simplified models could display similar characteristics. We used our analytical framework to study the correlation of coalescence times as a function of the scaled recombination rate, $\rho$, for the 2-sex and the simplified DDTWF models, and compared the results to the coalescent with recombination and its Markovian approximations. We found that, as expected, for sufficiently large effective population size ($N \gtrsim 100$), the results for the coalescent (as well as for its SMC' approximation, but not for SMC) were extremely close to those of the DDTWF models. In contrast, differences were observed for $N = 10$, even between the 2-sex and the simplified DDTWF.

We have focused here on a sample of two individuals at two loci. For unlinked loci, we showed that the variance of $\hat{\theta}$ for any number of loci is reduced to the two-loci problem. Extending the sample size to more than two individuals is expected to be significantly more complicated. Deviations between the coalescent and the discrete time haploid Wright-Fisher model for increasing sample sizes were recently studied and shown to be important for realistic human demographic histories (Bhaskar et al., 2014). We similarly speculate the presence of a shared pedigree to have an increasingly significant effect on the variance of Tajima's estimator as the sample size grows, but this analysis is left for future studies.

# References

A. Bhaskar, A. G. Clark, and Y. S. Song. Distortion of genealogical properties when the sample is very large. *Proc. Natl. Acad. Sci. U. S. A.*, 111:2385–2390, 2014.

A. H. Bittles and M. L. Black. Global patterns and tables of consanguinity, 2015. URL `http://consang.net`.

J. T. Chang. Recent common ancestors of all present-day individuals. *Adv. Appl. Probab.*, 31:1002–1026, 1999.

B. Derrida, S. C. Manrubia, and D. H. Zanette. On the genealogy of a population of biparental individuals. *J. Theor. Biol.*, 203:303–315, 2000.

A. Eriksson, B. Mahjani, and B. Mehlig. Sequential Markov coalescent algorithms for population models with demographic structure. *Theor. Popul. Biol.*, 76: 84–91, 2009.

Y. Erlich. Crowd-sourced genealogy for human genetics, 2016. URL `http://erlichlab.wi.mit.edu/familinx/`.

J. Felsenstein. Accuracy of coalescent likelihood estimates: Do we need more sites, more sequences, or more loci? *Mol. Biol. Evol.*, 23:691–700, 2006.

R. Griffiths and P. Marjoram. *Progress in Population Genetics and Human Evolution*, chapter An ancestral recombination graph, pages 257–270. Springer Verlag, 1997.

J. F. C. Kingman. The coalescent. *Stoch. Proc. Appl.*, 13:235–248, 1982.

P. Marjoram and J. D. Wall. Fast coalescent simulation. *BMC Genet.*, 7:16, 2006.

G. A. T. McVean. A genealogical interpretation of linkage disequilibrium. *Genetics*, 162:987–991, 2002.

G. A. T. McVean and N. J. Cardin. Approximating the coalescent with recombination. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 360:1387–1393, 2005.

A. Pluzhnikov and P. Donnelly. Optimal sequencing strategies for surveying molecular genetic diversity. *Genetics*, 144:1247–62, 1996.

P. L. Ralph. An empirical approach to demographic inference. arXiv:1505.05816, 2015.

N. A. Rosenberg and M. Nordborg. Genealogical trees, coalescent theory, and the analysis of genetic polymorphisms. *Nat. Rev. Genet.*, 3:380–390, 2002.

K. L. Simonsen and G. A. Churchill. A Markov chain model of coalescence with recombination. *Theor. Popul. Biol.*, 52:43–59, 1997.

F. Tajima. Evolutionary relationship of DNA sequences in finite populations. *Genetics*, 105:437–460, 1983.

F. Tajima. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123:585–595, 1989.

J. Wakeley, L. King, B. S. Low, and S. Ramachandran. Gene genealogies within a fixed pedigree, and the robustness of kingman's coalescent. *Genetics*, 190:1433–1435, 2012.

G. A. Watterson. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.*, 7:256–276, 1975.

P. R. Wilton, S. Carmi, and A. Hobolth. The SMC' is a highly accurate approximation to the Ancestral Recombination Graph. *Genetics*, 200:343–355, 2015.

C. Wiuf and J. Hein. Recombination as a point process along sequences. *Theor. Popul. Biol.*, 55:248–259, 1999.

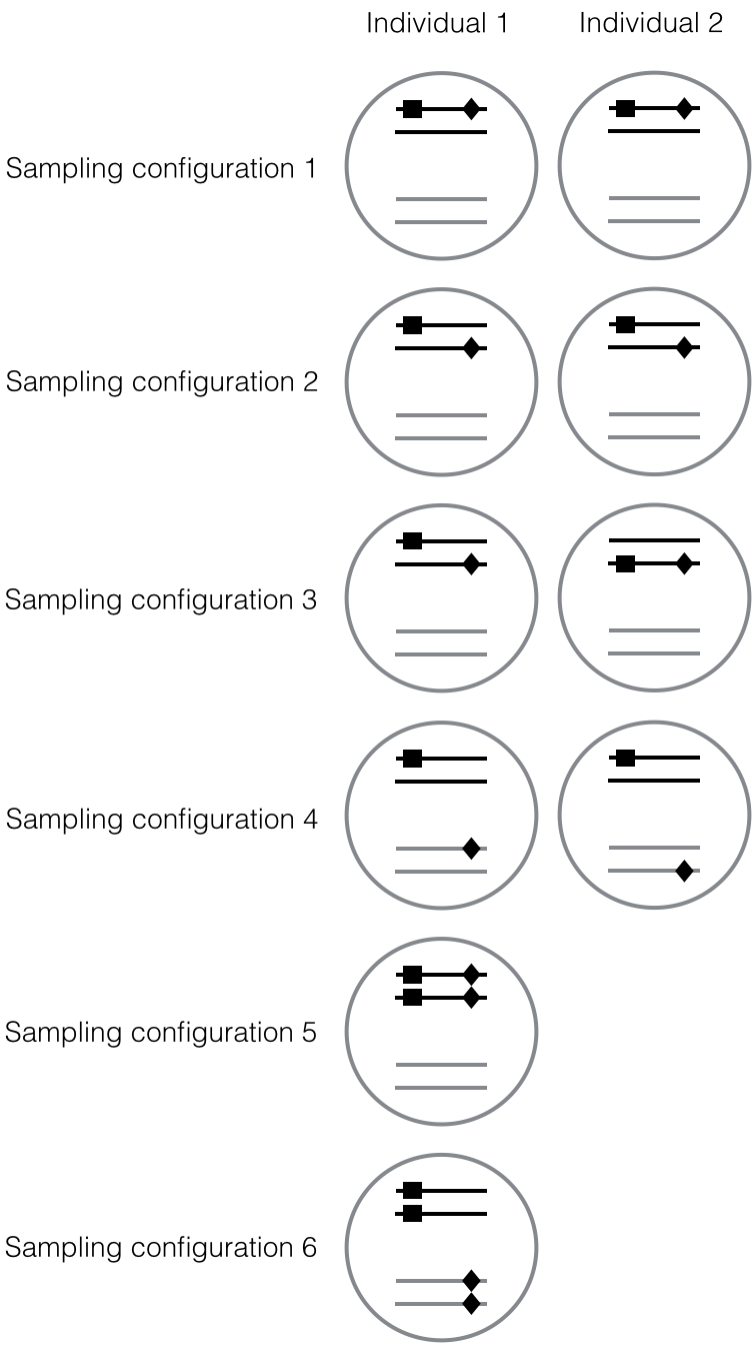



**Figure** 1: **The sampling configurations.** Sampling configurations 1 to 4 involve a sample of two individuals, depicted by two circles. Sampling configurations 5 and 6 involve a single individual, depicted by a single circle. The lines within each circle correspond to two pairs of homologous chromosomes. The two loci are indicated by squares and diamonds.

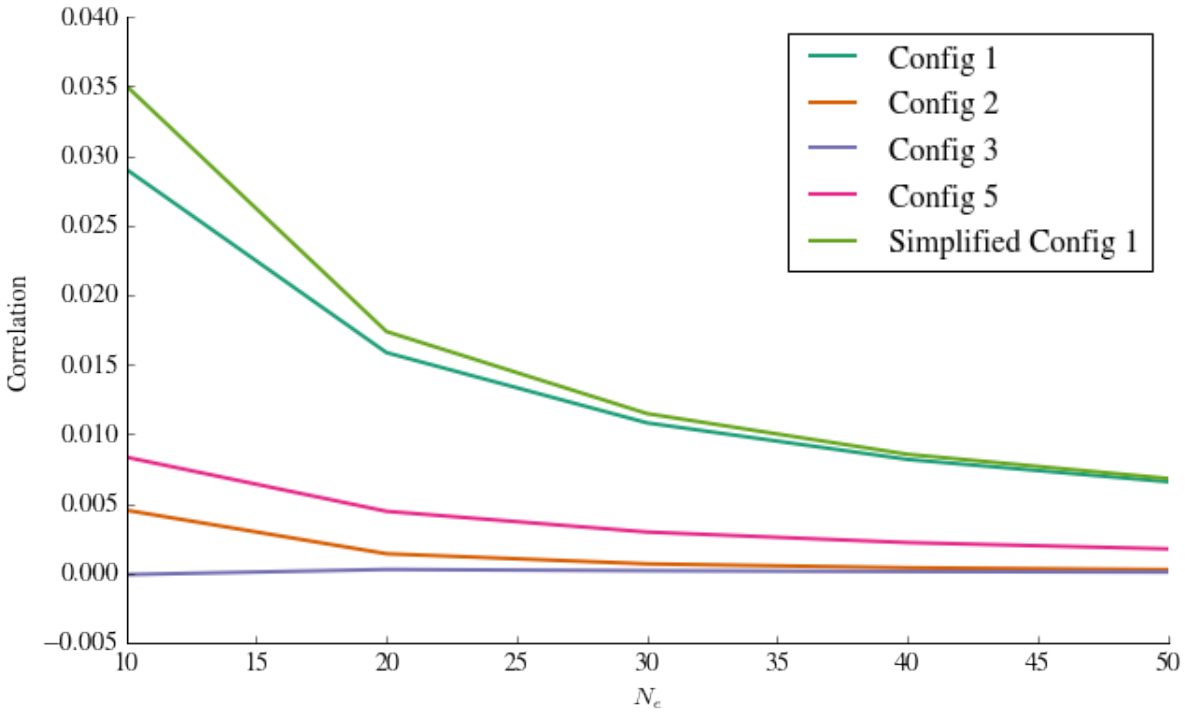



**Figure** 2: **Correlation of coalescence times for a sample of size 2.** We plot the correlation coefficients for the different sampling configurations under the 2-sex DDTWF and the simplified DDTWF vs the effective population size $N_e$. The calculations are described in detail in Supplementary Section S2.

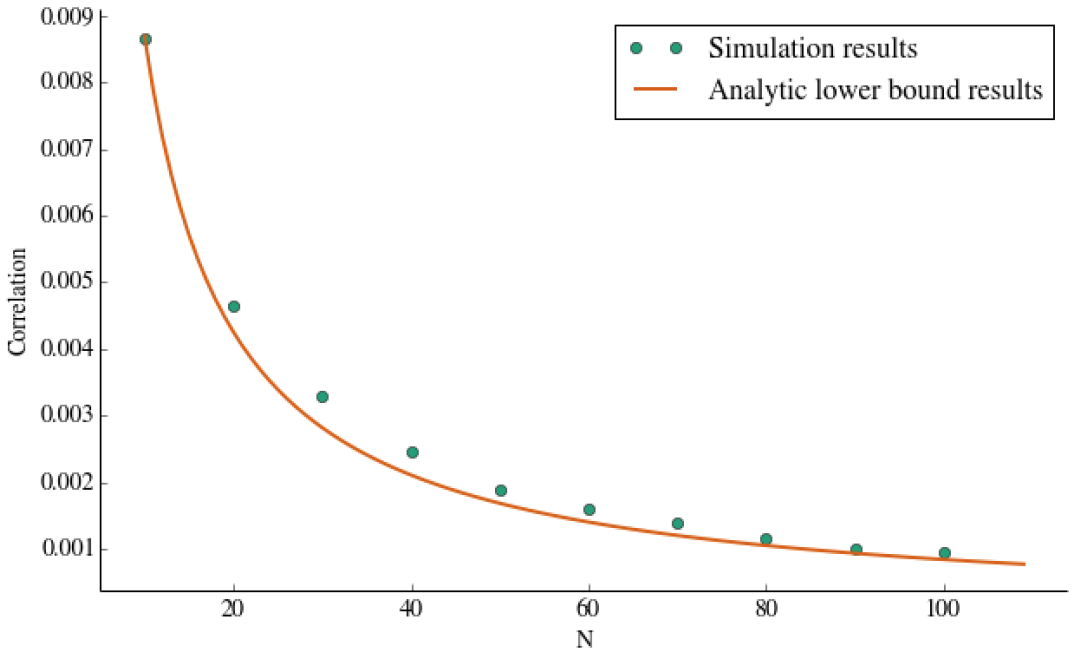



**Figure** 3: **Analytical lower bound for the correlation of coalescence times at unlinked loci.** We plot the correlation coefficient of the coalescence times at unlinked loci sampled from non-homologous chromosomes under the 2-sex, diploid, discrete-time Wright-Fisher model (green circles) as a function of the effective population size $N_e$. The analytical lower bound ($\mathrm{Corr}\,[T_i, T_j] \gtrsim 1/(12N)$) is plotted as a solid line.

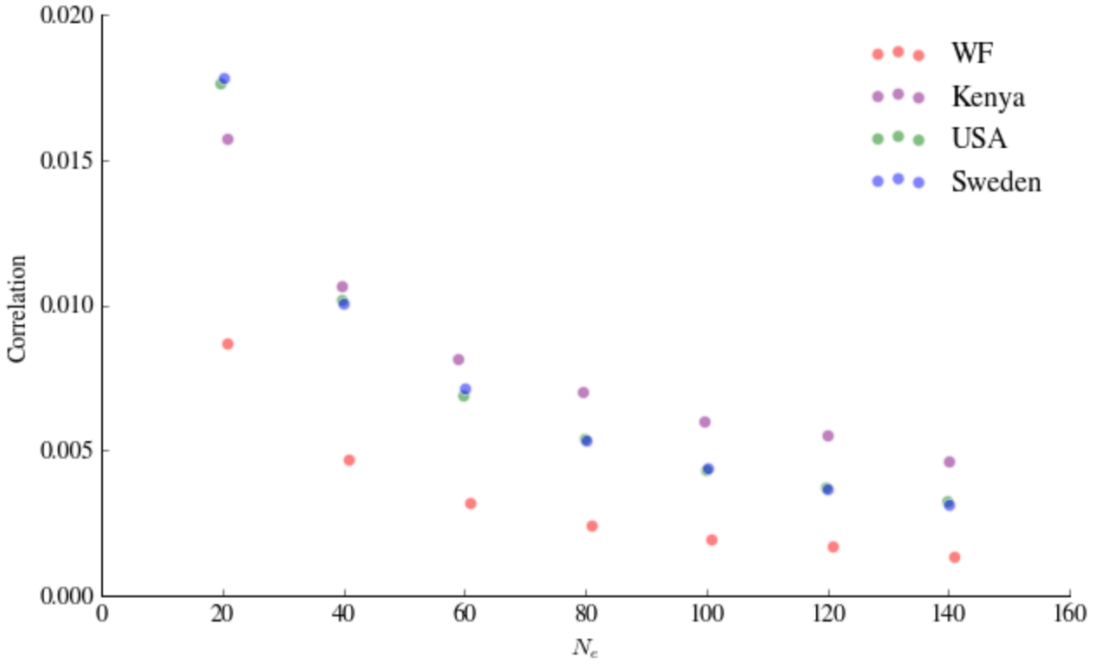



**Figure** 4: **The correlation coefficient of coalescence times at two unlinked loci under synthetic pedigrees constructed using the Familinx dataset.** Results are shown for three countries, as well as for the 2-sex DDTWF model. The correlation coefficient is plotted vs the effective population size $N_e$ (see the main text on how $N_e$ was set for the FAMILINX pedigrees). The two loci were sampled from non-homologous chromosomes. It can be seen that the correlation depends on the structure of the pedigree in ways that cannot be summarized by $N_e$.

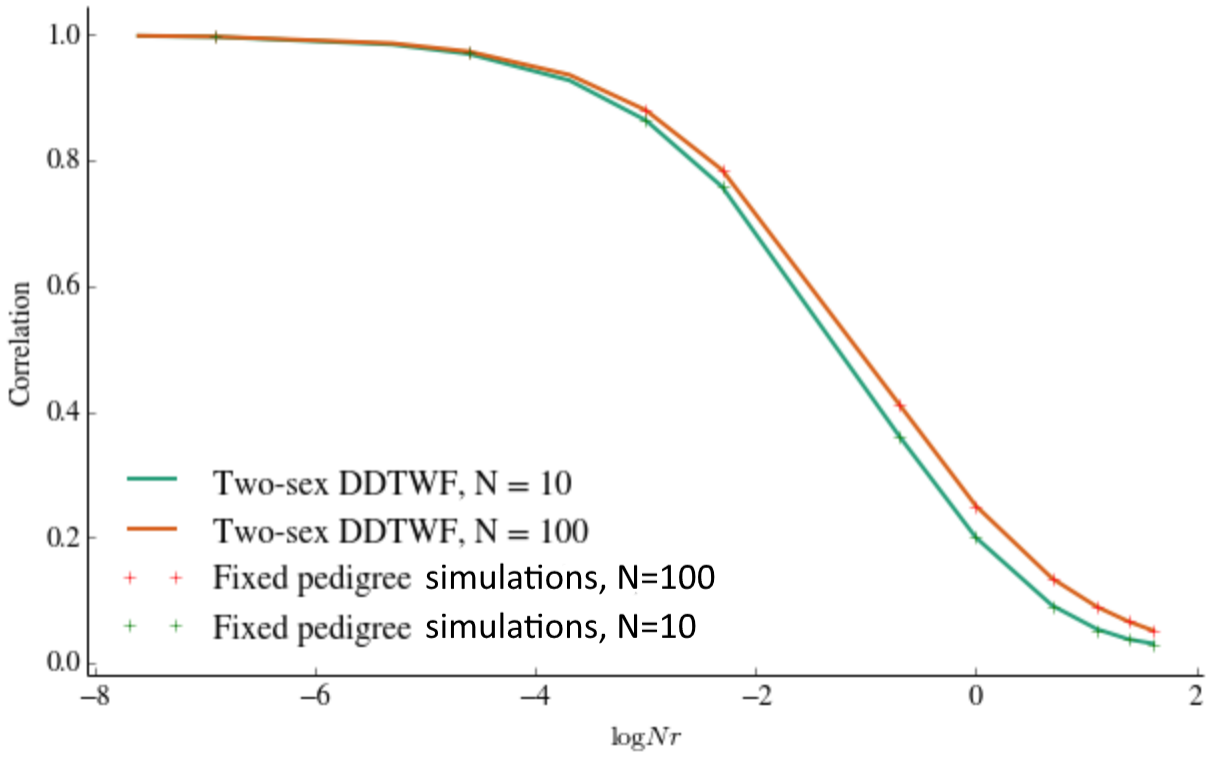



**Figure 5: The correlation coefficient of coalescence times at two linked loci under the 2-sex DDTWF model.** The correlation coefficients are plotted as lines for two values of $N_e$ vs the scaled recombination rate $\rho = 4Nr$. Simulation results are shown as + symbols. The two loci were sampled in configuration 1.

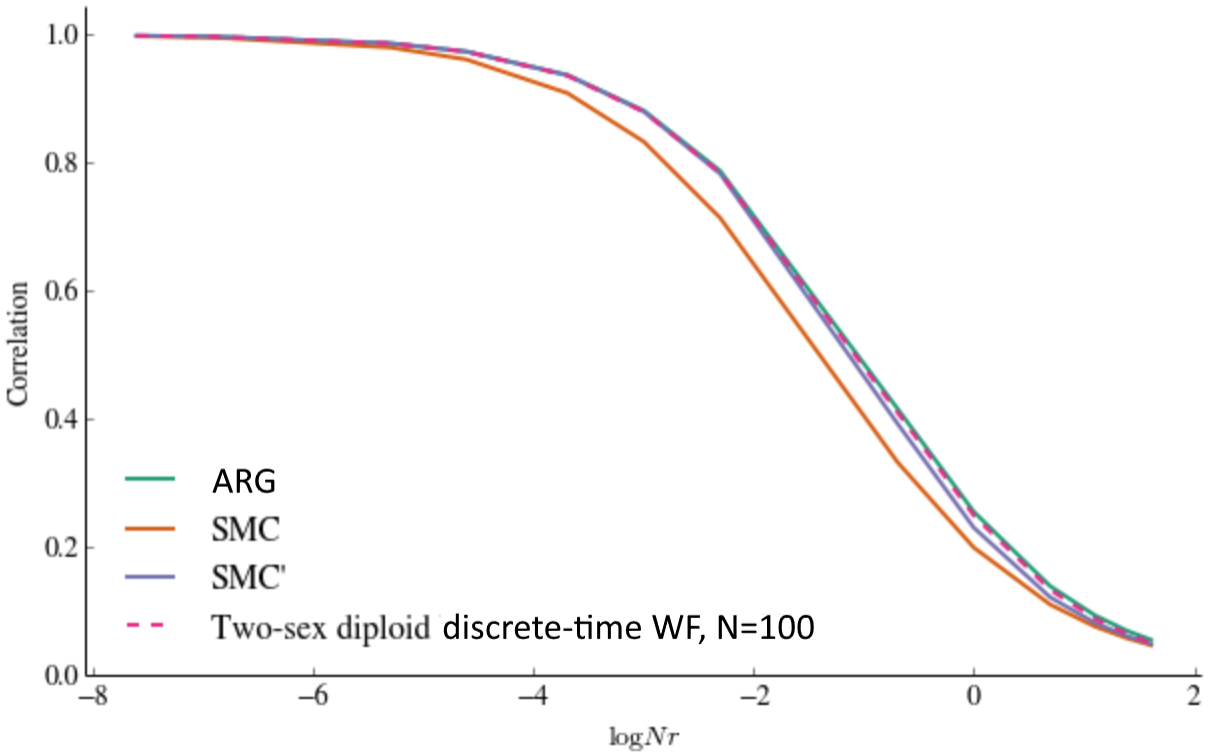



**Figure** 6: **A comparison of the correlation coefficient of the coalescence times at two linked loci under models of increasing complexity.** We compare the ARG, SMC, SMC', and the 2-sex DDTWF with $N_e = 100$, across different values of $\rho = 4N_e r$. The predictions of the ARG and SMC' are very good approximations for those of the 2-sex diploid Markov Chain model (for the value of $N_e$ shown here).

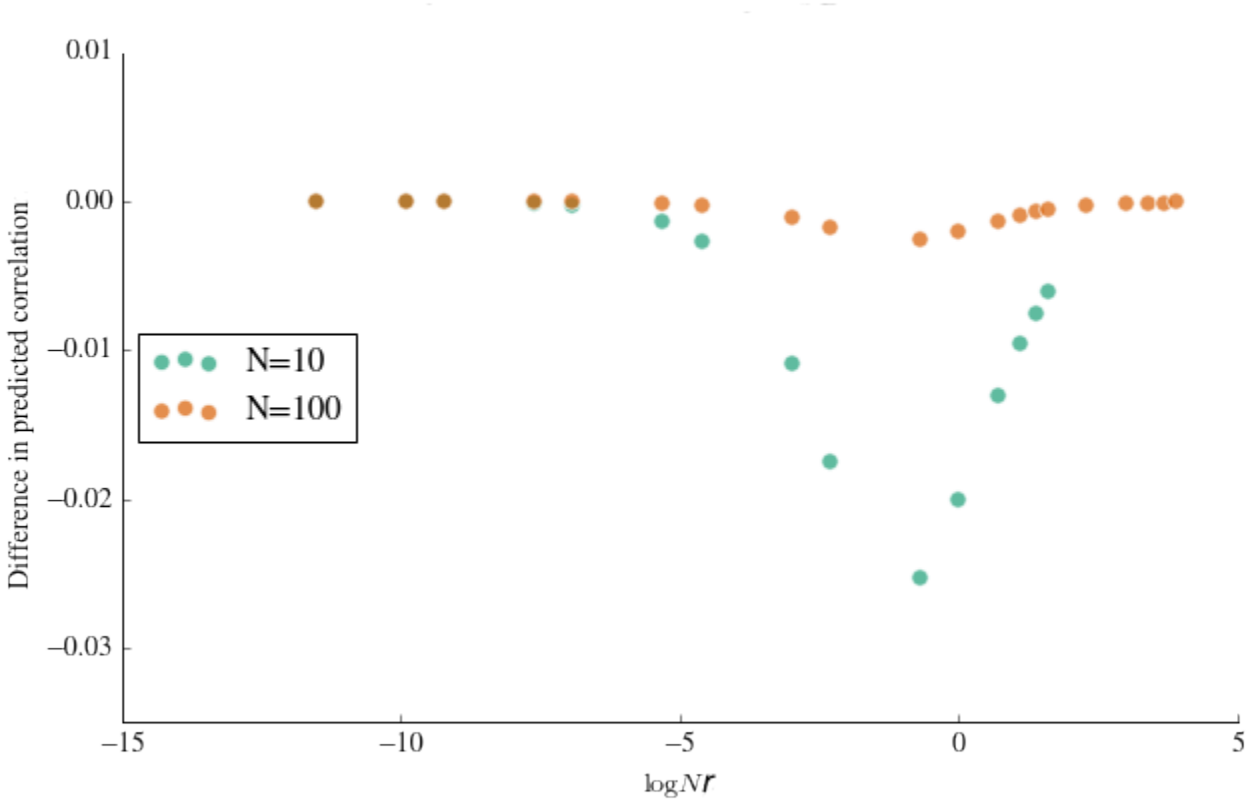



**Figure 7**: **A comparison of the correlation coefficient of the coalescence times at two linked loci between the 2-sex and the simplified DDTWF models.** We plot the difference between the correlation coefficients of the two models for $N_e = 10$ and $N_e = 100$ and for different values of $r$. The predictions of the two models slightly diverge at $N_e = 10$.

# Supplementary Material

## Extended methods and analytical results

## S1 The number of shared ancestors

In this section, we derive the covariance of the number of shared ancestors at each generation for the 2-sex diploid discrete-time Wright-Fisher model (DDTWF), denoted $x_g$ in section 3.2 of the main text. The model is defined in Section 4.2 of the main text and Section S2 below. We proceed in three steps.

### S1.1 Distribution of the number of ancestors from one generation to the next

Consider a single individual in a population with non-overlapping generations in the 2-sex model. Each generation $g$, there are $N_f$ males and $N_m$ females, where $N_f + N_m = N_e$, and typically $N_f = N_m = N_e/2$. Let $y_g$ be the number of ancestors of a particular individual at generation $g$ in the past. During the first few generations, the number of ancestors grows very fast, and we expect $y_g \approx 2^g$. As the number of ancestors in a given generation starts to approach the size of the population, the ancestors overlap with one another, and the growth of ancestors slows down until an equilibrium distribution is reached. We are interested in modeling the exact distribution of the number of ancestors in generation $g+1$, $y_{g+1}$, given the number of ancestors in generation $g$, $y_g$.

We can first divide the number of ancestors in generation $g+1$ into males and females:

$$y_{g+1} = F + M, \tag{1}$$

where $F$ is the number of fathers of individuals in $y_g$, and $M$ is the number of mothers of individuals in $y_g$. We have

$$P(F = f|y_g) = \frac{\binom{N_f}{f} f! S_2(y_g, f)}{(N_f)^{y_g}}, \tag{2}$$

where $S_2$ is the Stirling number of the second kind. The intuition behind this formula is that there are $\binom{N_f}{f}$ possible ways of choosing $f$ fathers among the $N_f$ available. There are then $f!$ possible orderings of these chosen males. The Stirling number of the second kind is the number of ways we can partition a set of $y_g$ individuals into $f$ categories. We divide all this by the total number of ways of making $y_g$ choices of fathers among the $N_f$ available, or $(N_f)^{y_g}$. Likewise,

$$P(M = m|y_g) = \frac{\binom{N_m}{m} m! S_2(y_g, m)}{(N_m)^{y_g}}. \tag{3}$$

We then obtain the following convolution for the number of ancestors $a$ in generation $g+1$,

$$P(y_{g+1} = a|y_g) = \sum_{f=1}^{a-1} P(F = f|y_g) P(M = a - f|y_g). \tag{4}$$

The numbers $y_1, ..., y_G$ form a Markov Chain. The preceding formula defines the transition matrix of $y_{g+1}$ given $y_g$.

If we did not have a 2-sex model, but instead a bi-parental monoecious model, the formula for the number of ancestors in generation $g+1$ would be the following simpler expression,

$$P(y_{g+1} = a|y_g) = \frac{\binom{N_e}{a} a! S_2(2y_g, a)}{N_e^{2y_g}}. \tag{5}$$

## S1.2 Overlap in the number of ancestors each generation

In the previous subsection, we described the distribution of the number of ancestors at each generation. Here, we start with a sample of two individuals, A and B, and are interested in the distribution of the number of shared ancestors each generation. If this sample consists of a pair of full siblings, then the number of shared ancestors grows according to the formula provided in the previous section, as full siblings share all of their ancestors.

Let $X_g$ be the set of common ancestors in generation $g$, $A_g$ be the ancestors of A that are not in $X_g$, and $B_g$ the set of ancestors of B that are not in $X_g$. Let $|A_g|$, $|B_g|$ and $|X_g|$ ($= x_g$ in the notation of the main text) be the cardinality of these three disjoint sets. Let $F_A$ be the set of fathers of individuals in $A_g$, and let $|F_A|$ be the cardinality of $F_A$. Likewise, we define $F_X$, $F_B$, $|F_X|$, and $|F_B|$. Given $|A_g|$, $|B_g|$, and $|X_g|$, the distribution of $|F_A|$, $|F_B|$, and $|F_X|$ is as described in the previous subsection,

$$P\left(|F_A| = f \middle| A_g\right) = \frac{\binom{N_f}{f} f! S_2(|A_g|, f)}{N_f^{|A_g|}}, \tag{6}$$

$$P\left(|F_B| = f \middle| B_g\right) = \frac{\binom{N_f}{f} f! S_2(|B_g|, f)}{N_f^{|B_g|}}, \tag{7}$$

$$P\left(|F_X| = f \middle| X_g\right) = \frac{\binom{N_f}{f} f! S_2(|X_g|, f)}{N_f^{|X_g|}}. \tag{8}$$

The number of fathers in common between individuals in $A_g$ and $X_g$, $x_a$, follows a hypergeometric distribution with $F_X$ success states, $N_{f(g+1)} - |F_X|$ failure states, and $|F_A|$ draws,

$$P\left(|F_A \cap F_X| = x_a\right) = \frac{\binom{|F_X|}{x_a}\binom{N_f - |F_X|}{|F_A| - x_a}}{\binom{N_f}{|F_A|}}. \tag{9}$$

The probability that individuals in $B_g$ have $x_b$ fathers in common with individuals in $X_g$, and $a_b$ fathers in common with individuals in $A_g$ (but not with individuals already in $X_g$), given that $|F_A \cap F_X| = x_a$, is defined by a trivariate hypergeometric distribution,

$$P\left(|F_B \cap F_X| = x_b \text{ and } |F_B \cap F_A| = a_b \middle| |F_A \cap F_X| = x_a\right) = \frac{\binom{|F_X|}{x_b}\binom{|(F_A - F_X \cap F_A|}{a_b}\binom{N_f - |(F_X \cup F_A)|}{|F_B| - x_b - a_b}}{\binom{N_f}{|F_B|}}. \tag{10}$$

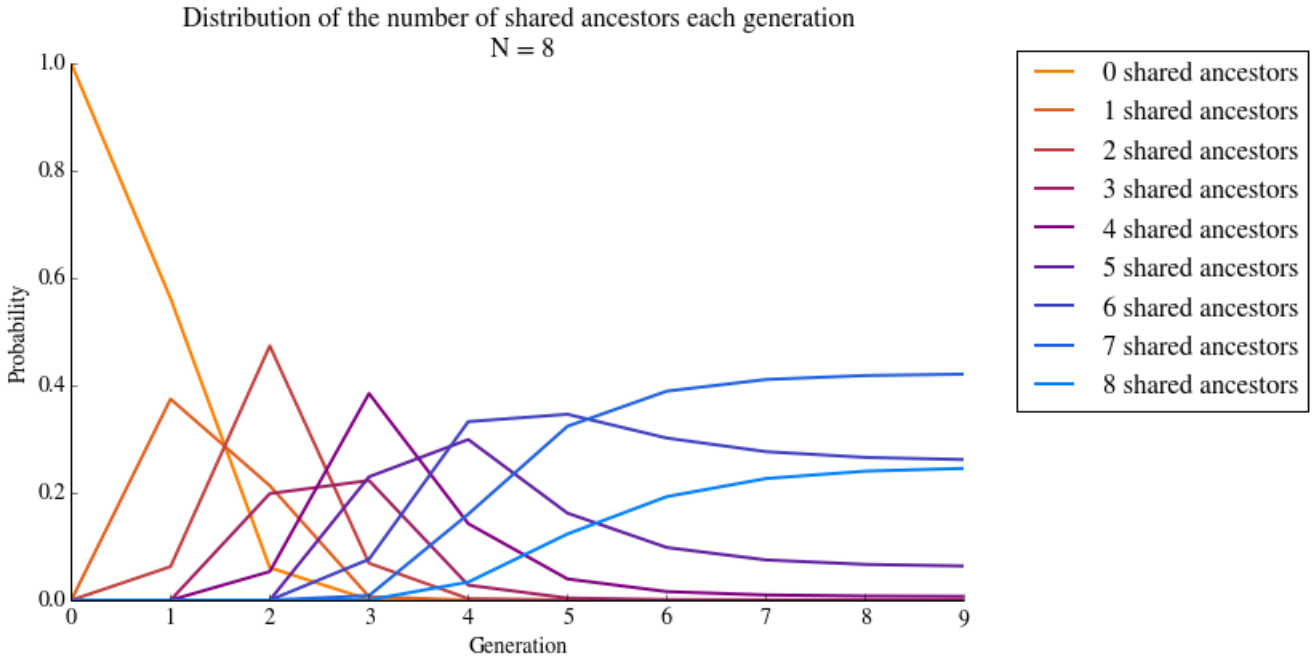
The number of shared male ancestors in generation $g+1$ is $|X_{f(g+1)}| = |F_X| + a_b$, the number of male ancestors exclusive to A is $|A_{f(g+1)}| = |F_A| - a_b - x_a$, and the number of male ancestors exclusive to B is $|B_{f(g+1)}| = |F_B| - a_b - x_b$. To obtain the number of shared female ancestors, $|X_{m(g+1)}|$, we use the same protocol, except replacing $N_f$ by $N_m$. Finally, to derive the joint distribution of $X_{g+1}$, $A_{g+1}$ and $B_{g+1}$, we take the convolution over the number of male and female ancestors.

In this way, we can derive a transition matrix $T$. The entries $T_{ij}$ of the transition matrix give the probability of entering state $j = (|A_{g+1}|, |B_{g+1}|, |X_{g+1}|)$ given state $i = (|A_g|, |B_g|, |X_g|)$.

We plot the dynamics of the number of shared ancestors along the generations in Supplementary Figure 1. The distribution of the number of shared ancestors in generation $g$ is obtained by considering the $g$-th power of $T$, assuming a sampling configuration of $(1, 1, 0)$ and then summing over the probabilities of all configurations with same $|X_g|$.

Finally, consider the simpler bi-parental monoecious model, and let $K_A$, $K_B$, and $K_X$ be the parents of individuals in $A_g$, $B_g$, and $X_g$, respectively. As in the previous subsection,

$$P\left(|K_A| = k \middle| A_g\right) = \frac{\binom{N_e}{k} k! S_2(2|A_g|, k)}{N_e^{2|A_g|}}, \tag{11}$$




**Supplementary Figure** 1: **The distribution of the number of shared ancestors in each generation for the 2-sex DDTWF model.** We used $N_e = 8$. The process reaches an equilibrium distribution after about 7 generations.

and similarly for $K_B$ and $K_X$. As above, the number of parents in common between individuals in $A_g$ and $X_g$, $x_a$, follows a hypergeometric distribution,

$$P\left(|K_A \cap K_X| = x_a\right) = \frac{\binom{|K_X|}{x_a}\binom{N_e - |K_X|}{|K_A| - x_a}}{\binom{N_e}{|K_A|}}. \tag{12}$$

The number of parents common to $B_g$ and $A_g$ or $X_g$ is similarly given by

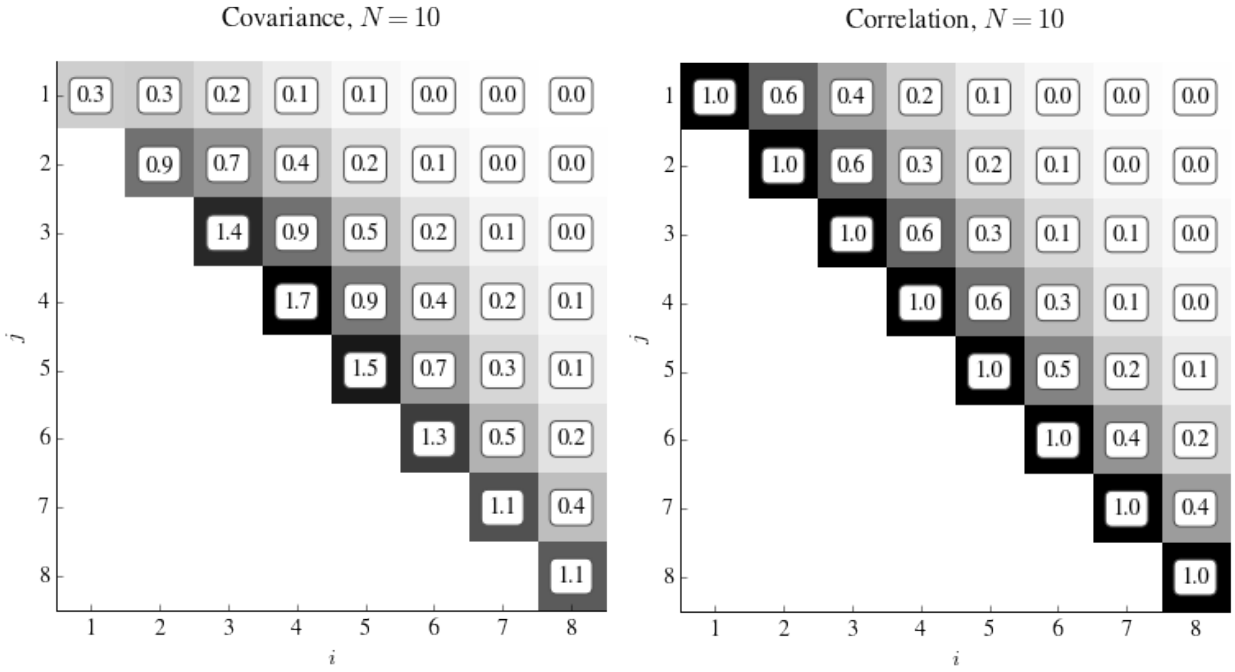
$$P(|K_B \cap K_X| = x_b \text{ and } |K_B \cap K_A| = a_b \mid |K_A \cap K_X| = x_a) = \frac{\binom{|K_X|}{x_b}\binom{|(K_A - K_X \cap K_A)|}{a_b}\binom{N_e - |(K_X \cup K_A)|}{|K_B| - xb - a_b}}{\binom{N_e}{|K_B|}}. \tag{13}$$

As above, we have $|X_{g+1}| = |K_X| + a_b$, $|A_{g+1}| = |K_A| - a_b - x_a$, and $|B_{g+1}| = |K_B| - a_b - x_b$. This fully specifies the distribution of the configuration in generation $g+1$, given the configuration in generation $g$.

## S1.3 Variance and covariances of the number of ancestors each generation

Finally, we calculate the covariances between the number of shared ancestors in generations $i$ and $j$, $\mathrm{Cov}(x_i, x_j)$, using the transition matrix $T$ derived as described in the previous section. Let state 0 be the index of the sampling configuration, $(1, 1, 0)$. We have, for $i \leq j$,

$$\begin{aligned}\mathrm{Cov}[x_i, x_j] &= \mathrm{E}[x_i x_j] - \mathrm{E}[x_i]\mathrm{E}[x_j] = \mathrm{E}[x_i \mathrm{E}[x_j | x_i]] - \mathrm{E}[x_i]\mathrm{E}[x_j] \\ &= \sum_{z=0}^{N_e} \left( z P(x_i = z) \sum_{k=0}^{N_e} k P(x_j = k | x_i = z) \right) - \sum_{z=0}^{N_e} z P(x_i = z) \sum_{z=0}^{N_e} z P(x_j = z).\end{aligned} \tag{14}$$




**Supplementary Figure** 2: **The covariance and correlation of the number of shared ancestors across the generations.** In the left panel, we show the covariance of the number of shared ancestors, $x_g$, for each generation $g$, and for $N_e = 10$. The diagonal represents the variance of the number of shared ancestors, and is highest in generations 3-5. In the right panel, we show the correlation coefficients. The correlation between $x_g$ and $x_{g+1}$ decreases with $g$.

Each value of the number of shared ancestors, $z$ is represented by multiple states of the transition matrix. We refer to the set of these states as "Conf z", or

$$P(x_i = z) = \sum_{\zeta \in \text{Conf z}} T^i[0][\zeta]. \tag{15}$$

Thus,

$$\sum_{z=0}^{N_e} \left( zP(x_i = z) \sum_{k=0}^{N_e} kP(x_j = k | x_i = z) \right) = \sum_{z=0}^{N_e} z \sum_{\zeta \in \text{Conf z}} T^i[0][\zeta] \sum_{k=0}^{N_e} k \sum_{\kappa \in \text{Conf k}} T^{j-i}[\zeta][\kappa]. \tag{16}$$

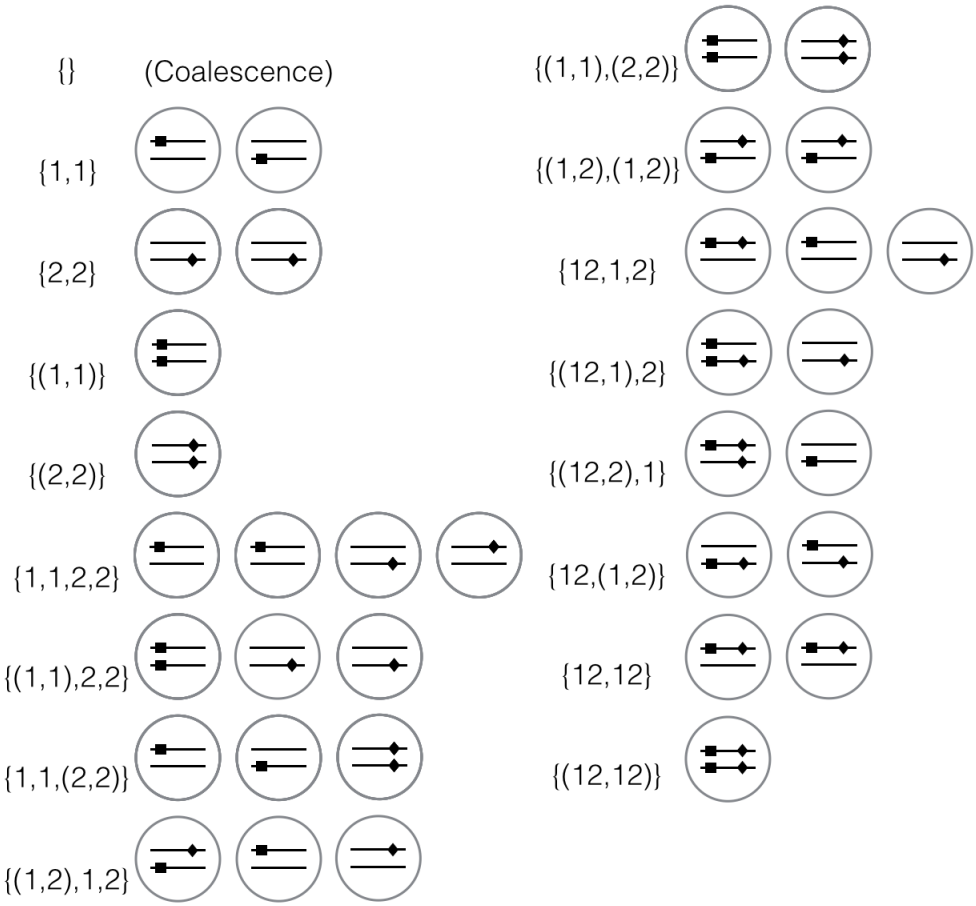
We plot the covariances and correlations for the 2-sex DDTWF model and for $N_e = 10$ in Supplementary Figure 2 .

We note that the entire derivation of this section can be generalized to the case when the number of males and females is allowed to differ as well as change along the generations.

# S2 The DDTWF models

## S2.1 The 2-sex DDTWF model and transition matrix

The notation we use to label the states in this transition matrix is derived from the notation of Wakeley and Lessard (2003), who used a similar transition matrix to analyze patterns of linkage disequilibrium in a 2-locus multi-deme model. The notation is explained in Supplementary Figure 3. For example, state {12,12}




**Supplementary Figure** 3: **The states of the 2-sex DDTWF model.** Circles represent individuals; the two lines within each individual represents a pair of homologous chromosomes; the square represent the first locus and the diamond represents the second locus. For example, {12,12} corresponds to the sampling configuration 1 in main text Figure 1.

represents the case where two copies of the first locus are located in two different individuals, and on the same chromosome as this first locus is the second locus. The comma separates the different chromosomes on which genetic material os tracked, and the numbers 1 and 2 represent the loci on each chromosome.

In state {(12,12)}, the parentheses indicate that the tracked pairs of loci are present on two different chromosomes in the same individual. If the tracked lineages are on different chromosomes of the same individual, then they must be located in different individuals in the previous generation. So, for example, state {(1,1)} transitions to state {1,1} in one generation with probability 1.

The set of all possible states in our model is : {}, {1,1}, {2,2}, {(1,1)}, {(2,2)}, {1,1,2,2}, {(1,1),2,2}, {1,1,(2,2)}, {1,2,(1,2)}, {(1,1),(2,2)}, {(1,2),(1,2)}, {12,1,2}, {(12,1),2}, {(12,2),1}, {12,(1,2)}, {12,12} and {(12,12)}. We show the communicating states in this transition matrix in table S2.1.

2-sex diploid DTWF model

| State | coal | 1,1 | 2,2 | (1,1) | (2,2) | 1,1,2,2 | (1,1), 2, 2 | 1,1,(2,2) | (1,2),1, 2 | (1, 1), (2, 2) | (1,2),(1,2) | 12,1,2 | (12,1),2 | (12,2),1 | 12, (1,2) | 12, 12 | (12, 12) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| coal | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1,1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2,2 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| (1,1) | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| (2,2) | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1,1,2,2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| (1,1), 2,2 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 0 |
| 1,1,(2,2) | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 |
| (1,2), 1,2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| (1,1),(2,2) | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| (1,2),(1,2) | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| 12,1,2 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| (12,1),2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 |
| (12,2),1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 |
| 12, (1,2) | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| 12,12 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 |
| (12,12) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 |

The cell at coordinates (i,j) is 1 if the probability of transitioning to state j starting from state i in one generation is non-zero.

## S2.2 The simplified DDTWF model

We also consider a simplified version of this model, a monoecious bi-parental DDTWF model. In this model, we do not keep track of whether lineages are in the same individual or not. The diploidy only comes into play in that recombination is impossible in a haploid context. The complete list of states in this model is: {}, {1,1}, {2,2}, {1,1,2,2}, {12,1,2}, and {12,12}, far fewer than in the 2-sex DDTWF model. For this reason, this model can only be used to show the effect of a limited number of sampling configurations. For example, it is not possible to model sampling configuration 2, where loci are sampled from different homologous chromosomes in the same individual. We show a matrix of communicating states in table S2.2.

| Simplified diploid DTWF model | | | | | | |
|---|---|---|---|---|---|---|
| State | coal | 1,1 | 2,2 | 1,1,2,2 | 12,1,2 | 12, 12 |
| coal | 1 | 0 | 0 | 0 | 0 | 0 |
| 1,1 | 1 | 1 | 0 | 0 | 0 | 0 |
| 2,2 | 1 | 0 | 1 | 0 | 0 | 0 |
| 1,1,2,2 | 1 | 1 | 1 | 1 | 1 | 1 |
| 12,1,2 | 1 | 1 | 1 | 1 | 1 | 1 |
| 12,12 | 1 | 1 | 1 | 1 | 1 | 1 |

The cell at coordinates (i,j) is 1 if the probability of transitioning to state j starting from state i in one generation is non-zero.

## S2.3 The expected generation time in both models

If two lineages are located in two different individuals, then the probability they coalesce in a single generation is just $1/(2N_e)$. However, if they are present in different chromosomes of the same individual, they must have originated from two different individuals in the previous generation. Because of this, the expected time until coalescence will be different than $2N_e$ in the 2-sex DDTWF, as opposed to in the simplified DDTWF where it is just equal to $2N_e$.

The process retains some memory of the fact that lineages were initially sampled in two different individuals. Indeed, the time until coalescence at generation $g$, given no coalescence in any previous generation, will be different than the expected time until coalescence at generation $g+1$, given no coalescence in any previous generation. As $g$ increases, this difference in coalescence times decreases from one generation to the next, and the process converges to an average generation time.

Consider a pair of lineages in two individuals. In generation $g+1$, given that no coalescence events have occurred in any of the previous $g$ generations, the probability of the two lineages to coalesce is

$$C(g+1) = \frac{1}{4}P(F(g)|\text{No Coal at 1,...,}g) + \frac{1}{8}P(H(g)|\text{No Coal at 1,...,}g), \tag{17}$$

where $P(F(g)|\text{No Coal at 1,...,}g)$ and $P(H(g)|\text{No Coal at 1,...,}g)$ are the probabilities that the two lineages are located in full siblings and half siblings, respectively, in generation $g$, given no coalescence in that generation or any of the previous generations. Next, we write

$$P(F(g)|\text{No Coal at 1,...,}g) = \frac{P(F(g), \text{No coal at } g|\text{No coal at 1,...,}g-1)}{P(\text{No coal at } g|\text{No coal at 1,...,}g-1)}. \tag{18}$$

The denominator is simply given by $1-C(g)$. For the numerator, we note that for the two lineages to arrive at full siblings in generation $g$, then first, we must exclude the possibility that the lineages are at the same individual in generation $g$ (given no previous coalescence), which happens with probability $2C(g)$ (since the probability of coalescence is half the probability to arrive at the same individual). Second, the probability that these two individuals share both parents is $1/(N_e/2)^2$. Therefore,

$$P(F(g)|\text{No Coal at 1,...,}g) = \frac{1-2C(g)}{1-C(g)}\frac{1}{(N_e/2)^2}. \tag{19}$$

In the same way, we have

$$P(H(g)|\text{No Coal at } 1,...,g) = \frac{1-2C(g)}{1-C(g)}\frac{2}{N/2}\frac{N_e/2-1}{N_e/2}, \tag{20}$$

By solving $C(g+1) = C(g)$, we obtain the limiting coalescence probability as a function of $N_e$. As the equilibrium distribution of the time until MRCA is geometric, the equilibrium generation time is the inverse of the probability of coalescence each generation, or

$$E[T_i] = \frac{2N_e}{1+N_e-\sqrt{1+N_e^2}}. \tag{21}$$

This generation time is always slightly greater than $2N_e$.$E[T_i]/2N_e$ converges to 1 as $N_e$ becomes large.

## S3 Building pedigrees with Familinx

We simulated our Familinx-based pedigrees over `GEN` $= 100$ non-overlapping generations. For each generation, we selected family units at random from the data until the total number of children across all family units was greater than some pre-determined $N_c$ (the population census size). In addition, we required the total number of parents among the selected family units to be less than or equal to $N_c$. Then, we connected the `GEN` generations together by randomly assigning each parent in generation $g$ to be one of the children in generation $g+1$, disallowing sibling mating. Finally, we connected the first and last generation so that the pedigree is cyclical, with a period of `GEN` generations.

As a note, this procedure will not be appropriate for datasets where a substantial number of family units contain only one child, because the algorithm requires the number of children to be greater than or equal to the number of parents. When many families have only one child, families with more children will be over-sampled, and the family structure of our constructed pedigrees will be very different from the family structure we are attempting to replicate.

The value of $N_c$ was chosen to generate pedigrees with a target effective population size, $N_e$. For each $N_c$, we estimated the effective population size of our pedigree by calculating the average time until coalescence over 50 sampled pairs, and setting $N_e$ as half of that time. We then discarded the pedigree unless this value is within $\sigma_{N_e}$ of the target $N_e$, where $\sigma_{N_e}$ is the standard deviation of the observed coalescent effective sizes for a population of size $N_c = N_e$ in a Wright-Fisher model. We constrain our pedigrees to be close to the target effective population size because we want to make sure that the higher covariance we observe in the Familinx pedigree simulations relative to the WF model is not only due to potentially higher variance of $N_e$.

We note that under our algorithm, some information on the country-specific pedigree structure is lost by breaking large genealogies into family units (e.g., inter-generational correlations in family size, or the rate of first and second cousin matings). Nevertheless, sufficient information is retained so that pedigrees with the same $N_e$ generated based on data from different countries are distinguished by their correlation of coalescence times (main text Figure 4).

## References

J. Wakeley and S. Lessard. Theory of the effects of population structure and sampling on patterns of linkage disequilibrium applied to genomic data from humans. *Genetics*, 164:1043–1053, 2003.