Theoretical Population Biology ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

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

Léandra King [a], John Wakeley [a], Shai Carmi [b,*]

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

ARTICLE INFO

Article history:
Received 16 August 2016
Available online xxxx

Keywords:
Coalescent theory
Recombination
Heterozygosity
Effective population size
Pedigrees
Genealogies

ABSTRACT

The population-scaled mutation rate, θ, 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, the variance of Tajima's estimator ($\hat{\theta}$), which is the average number of pairwise differences, 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 gene genealogies at all loci. 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.

1. Introduction

The population-scaled mutation rate, θ, is defined as $4N_e\mu$, where N_e is the effective population size and μ is the mutation rate per locus per generation (Wakeley, 2009). Two classic estimators were developed for θ, 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 θ, 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 θ 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 θ, it is generally more effective to increase the number of independent loci than the sample size at each locus (see also e.g., Edwards and Beerli, 2000; 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 θ 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 of the same fixed underlying pedigree across all loci (Wakeley et al., 2012). By conditioning on the number of shared genealogical common ancestors, we derive a simple approximate lower bound, as a function of N_e, on the variance of $\hat{\theta}$ (Sections 2 and 3).

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*. In Section 4, 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, in Section 5 we run simulations based on real, large-scale human genealogical data (Kaplanis et al., 2017). The pedigrees inspired by different human populations differ from each other and from the

* Corresponding author.
E-mail address: shai.carmi@huji.ac.il (S. Carmi).

http://dx.doi.org/10.1016/j.tpb.2017.03.002
0040-5809/

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 if they have the same effective population size. Finally, we study some properties of linked sites in Section 6.

We note that the dependence of gene genealogies at unlinked loci has been previously recognized, most recently in the context of matching probabilities. Specifically, the probability of the genotypes of two individuals to match at two or more loci was computed under the Wright–Fisher and other models, and shown to differ from the product of the corresponding one-locus probabilities (Laurie and Weir, 2003; Song and Slatkin, 2007; Bhaskar and Song, 2009). In earlier literature, this effect was demonstrated in the context of identity-by-descent probabilities at unlinked loci (Weir and Cockerham, 1969) and implicitly in results on linkage disequilibrium (Ohta and Kimura, 1969). However, the treatment of this effect in the context of effective population size estimation is to our knowledge new.

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

For a sample of size two (haploids) at n loci, the estimator of θ can be expressed as

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

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

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

Under the standard coalescent model (Wakeley, 2009), $\hat{\theta}_i$ is Poisson distributed with mean $2\mu T_i$, where T_i is the time until coalescence at locus i in generations and μ is the mutation rate per locus per generation. Using the law of total covariance,

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

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

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

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

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

Studying the correlation instead of the covariance will allow 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 replicates over the entire evolutionary process, including both the population pedigree (family relationships between all individuals) and the gene genealogies. We elaborate below on this important point (Sections 3 and 5, and the Discussion).

3. Modeling the effect of the shared pedigree

In this section, we study the role of the shared underlying pedigree in the non-zero variance of $\hat{\theta}$. We first provide a formal derivation of the statistical inconsistency of $\hat{\theta}$, followed by an intuitive derivation of an approximate lower bound. Exact calculations appear in Section 4.

3.1. Statistical inconsistency of $\hat{\theta}$ due to the underlying pedigree

We begin with a general analysis of the inconsistency of the estimator of θ. An estimator is consistent if its sampling distribution converges in probability to the true parameter value (Wasserman, 2004). 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 θ. If the sampled individuals happen to be more closely related than average, then $\hat{\theta}$ will tend to underestimate the true value of θ. The opposite is true if the sampled individuals are less closely related than average.

To formally analyze the consistency of $\hat{\theta}$, recall that a sequence of random variables X_n converges in probability to C if, for every $\epsilon > 0$, $\lim_{n\to\infty}P(|X_n - C| \geq \epsilon) = 0$ (Wasserman, 2004). To show that $\hat{\theta}_{(n)}$ does not converge in probability to θ, it is sufficient to show that there exist $\epsilon > 0$ and $\delta > 0$, such that for each n, $P(|\hat{\theta}_{(n)} - \theta| \geq \epsilon) > \delta$. Let δ be the probability that a randomly sampled pair of individuals is as closely related as full siblings. Let ϵ be some arbitrary value smaller than the difference between θ 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 provide 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 θ. For those fixed ϵ and δ, we therefore cannot find n large enough such that $\mathrm{Prob}(|\hat{\theta}_{(n)} - \theta| \geq \epsilon) < \delta$. This implies that there is no convergence in probability to θ, and thus, this estimate of θ is not consistent. Since $\hat{\theta}_{(n)}$ is unbiased (from Eq. (1) and the discussion that followed, $\mathrm{E}\left[\hat{\theta}_{(n)}\right] = \mathrm{E}\left[\hat{\theta}_i\right] = \mathrm{E}\left[\mathrm{E}\left[\hat{\theta}_i | T_i\right]\right] = \mathrm{E}\left[2\mu T_i\right] = 4\mu N_e = \theta$), its inconsistency implies that its variance does not tend to 0 as n increases (Wasserman, 2004 Theorem 6.10).

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 two 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 $\{x\} = x_1, x_2, \ldots, x_G$, where x_g is the number of shared ancestors g generations ago. The vector $\{x\}$ is, in a sense, a low 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:

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

Please cite this article in press as: King, L., et al., A non-zero variance of Tajima's estimator for two sequences even for infinitely many unlinked loci. Theoretical Population Biology (2017), http://dx.doi.org/10.1016/j.tpb.2017.03.002.

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

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

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 $\mathrm{E}\left[T_i|\text{no coal}\right] \approx 2N_e + G$. We can write:

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

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 $\text{Prob}\left\{\text{no coal by } G|\{x\}\right\} = \prod_{g'=1}^{G}\left[1-\alpha(g')\right]$. 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,

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

and

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

since $G \ll N_e$.

In Supplementary material Section S1, we provide a numerical method to calculate the 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

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

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),

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

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),

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

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

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

In summary, the variance due to the shared pedigree is of order θ^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 as $\propto 1/N_e$. 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 (for the two individuals under consideration), as obtained by sampling infinitely many loci, leaves uncertainty regarding the value of θ. In other words, the uncertainty in the estimate of θ 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).

In yet another way, given that we are sampling infinitely many loci, and using Eq. (1), the assumption that $\hat{\theta}_i$ has mean $2\mu T_i$, and the law of large numbers, $\hat{\theta}$ converges to a fixed value, which is twice the mutation rate times the empirical mean TMRCA in the pedigree connecting the two individuals. This limit still has some degree of randomness about the model parameter θ, since different instances of the pedigree will have different empirical means. The variance of the estimator, $\mathrm{Var}\left[\hat{\theta}\right]$, is thus proportional to the variance of the realized mean TMRCA within the pedigree, as in our Eq. (7).

We could not find simple reasoning to the observed $1/N_e$ scaling. However, we can speculate that the dependence between the genealogies at the two loci results mostly from the rare events in which the two sampled individuals are closely related. Under the WF model, the probability of relatedness in the first $\mathcal{O}(1)$ generations is $\mathcal{O}(1/N_e)$, which may lead to the observed scaling. This result may also have some connections to the $\mathcal{O}(1/N_e)$ scaling of the σ_d^2 measure of linkage disequilibrium for unlinked loci (Ohta and Kimura, 1969).

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 from two sequences of present-day individuals. We focus on six particular *sampling configurations*, shown in Fig. 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 infinitely far apart on the same chromosome in both individuals. This means that these loci will be coupled for the first few generations, going backwards in time, 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. It is

Please cite this article in press as: King, L., et al., A non-zero variance of Tajima's estimator for two sequences even for infinitely many unlinked loci. Theoretical Population Biology (2017), http://dx.doi.org/10.1016/j.tpb.2017.03.002.

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 the two loci from the same 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 when going one generation backwards in time: 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 (diecious) 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 in 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 q such that $E[T_iT_j|q] > 0$:

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

where p_{qk} is the transition probability between states q and k. This system of equations is conceptually similar to that used by Laurie and Weir (2003), Song and Slatkin (2007), and Bhaskar and Song (2009) to calculate the probability of matching genotypes at two or more unlinked loci.

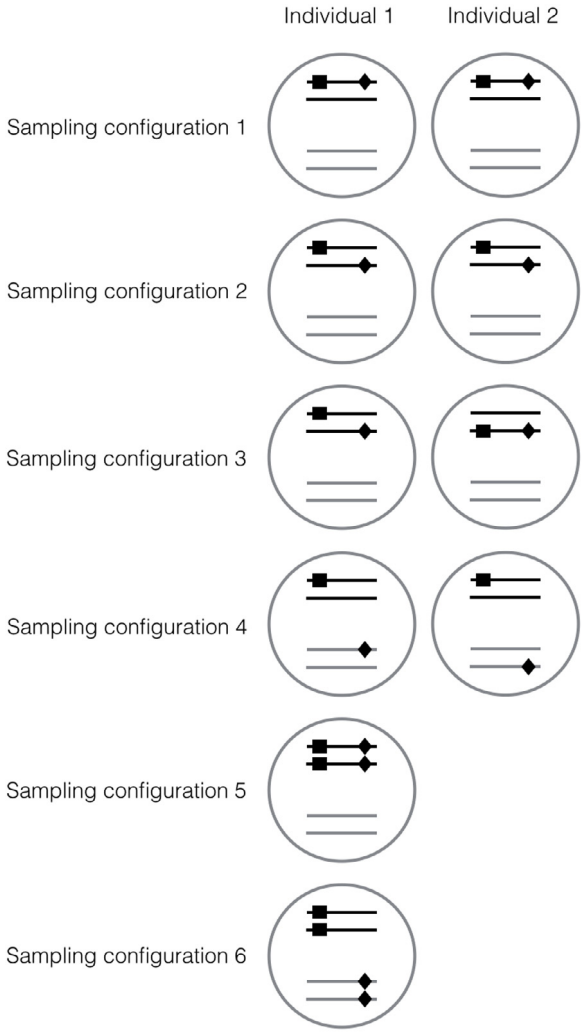

Fig. 1. The sampling configurations. Sampling configurations 1 to 4 involve a sample of two individuals, depicted by two circles. Sampling configurations 5 and 6 involve a single individual, depicted by a single circle. The lines within each circle correspond to two pairs of homologous chromosomes. The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

Solving this system of equations allowed us to obtain exact results for $\mathrm{Cov}\left[T_i, T_j|q\right]$. As a note, $E[T_i|q]$ can be different from $E[T_j|q]$, depending on the state q. For example, if the pair of lineages at locus i is located on two homologous chromosomes in the same individual, whereas the pair of lineages at locus j is located in two different individuals, then $E[T_i|q] = E[T_j|q] + 1$. 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 $i = j$.

Fig. 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, the two 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 the two loci remain on the same chromosome. As N_e increases, the correlation decreases, as it is much more likely for the two loci to split

Please cite this article in press as: King, L., et al., A non-zero variance of Tajima's estimator for two sequences even for infinitely many unlinked loci. Theoretical Population Biology (2017), http://dx.doi.org/10.1016/j.tpb.2017.03.002.

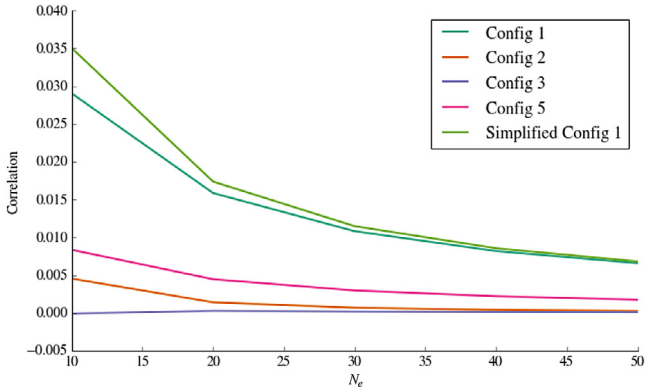

Fig. 2. Correlation of coalescence times for a sample of size 2. We plot the correlation coefficients for the different sampling configurations under the 2-sex DDTWF and the simplified DDTWF vs the effective population size N_e. The calculations are described in detail in Supplementary Section S2.

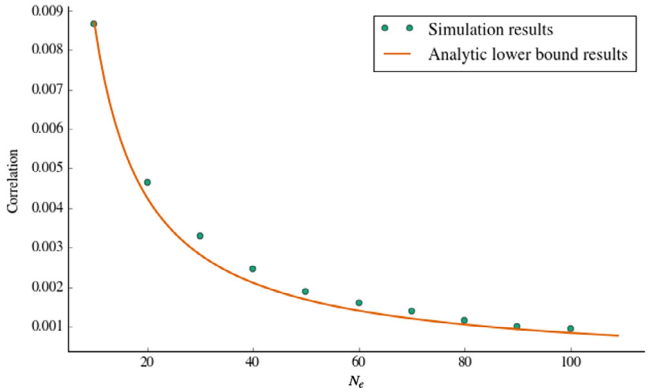

Fig. 3. Analytical lower bound for the correlation of coalescence times at unlinked loci. We plot the correlation coefficient of the coalescence times at unlinked loci sampled from non-homologous chromosomes under the 2-sex, diploid, discrete-time Wright–Fisher model (circles) as a function of the effective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

(probability 1/2 at each generation) before a coalescence event occurs. 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.

Fig. 2 also shows results for a simplified DDTWF model, which is similar to the 2-sex DDTWF, except that individuals are monecious and we do not keep track of whether any two 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. We then 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 other pedigrees and pairs of individuals to obtain an estimate of $\text{E}[T|\text{ped}]$. We then compute its variance over many simulated pedigrees to obtain $\text{Var}_{\text{ped}}[\text{E}[T|\text{ped}]]$. By the same logic as Eq. (7), $\text{Var}_{\text{ped}}[\text{E}[T|\text{ped}]]$ is equal to $\text{Cov}\left[T_i, T_j\right]$. To obtain the correlation coefficient, we divide $\text{Cov}\left[T_i, T_j\right]$ by $\text{Var}[T] = \text{Var}_{\text{ped}}[\text{E}[T|\text{ped}]] + \text{E}_{\text{ped}}[\text{Var}[T|\text{ped}]]$. The simulation results are shown in Fig. 3. Our analytical lower bound, which, based on Eqs. (14) and (5), can be written as $\text{Corr}\left[T_i, T_j\right] \gtrsim 1/(12N)$, 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 θ, 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 Kaplanis et al. (2017), 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. As FAMILINX contains data on more than the three countries we chose, and in order not to create a bias in favor of smaller, simpler family units, we only require that the first sampled child is in the corresponding country dataset. 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 the above-described scheme, the effective population size N_e is not guaranteed to equal the census population size (as is in the WF model). We thus defined N_e, for each country-inspired model, 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. We note that other definitions of N_e are possible, for example, based on the variance in the number of offspring (Wakeley, 2009). 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.

Please cite this article in press as: King, L., et al., A non-zero variance of Tajima's estimator for two sequences even for infinitely many unlinked loci. Theoretical Population Biology (2017), http://dx.doi.org/10.1016/j.tpb.2017.03.002.

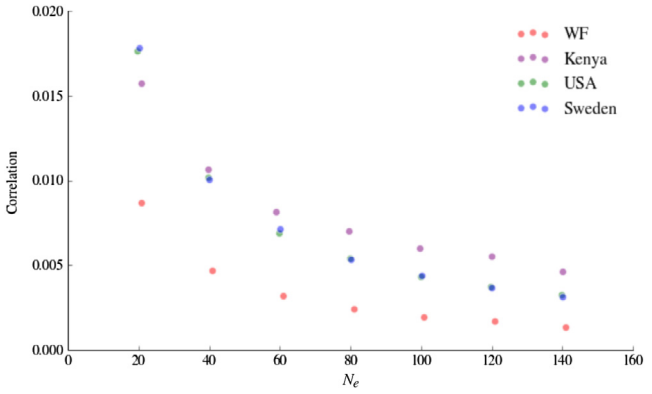

Fig. 4. The correlation coefficient of coalescence times at two unlinked loci in models inspired by the FAMILINX dataset. To generate the figure, synthetic pedigrees were constructed under Wright–Fisher-like models with single-generation genealogical patterns imitating those observed in FAMILINX. Results 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, defined here as half the mean coalescence time across randomly generated pedigrees (see the main text). 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, at least according to its definition used here.

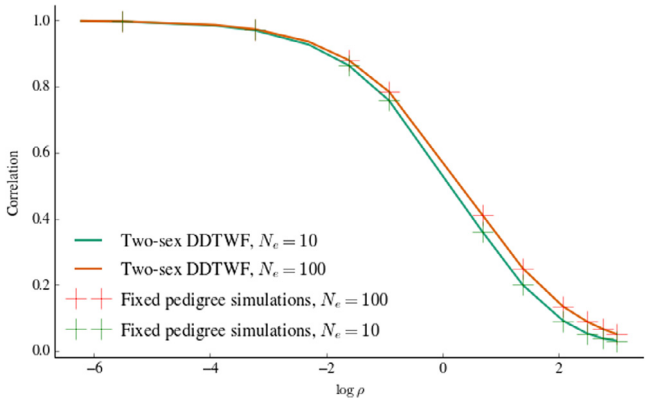

Fig. 5. The correlation coefficient of coalescence times at two linked loci under the 2-sex DDTWF model. The correlation coefficients are plotted as lines, for two values of N_e, vs the scaled recombination rate $\rho = 4N_e r$. Simulation results are shown as + symbols. The two loci were sampled in configuration 1. Note that since the x-axis corresponds to $\rho = 4N_e r$ (as opposed to just the recombination fraction r), the correlation for higher values of N_e need not be smaller.

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 Fig. 4, demonstrate that $\mathrm{Corr}\left[T_i, T_j\right]$, and consequently, $\mathrm{Var}\left[\hat{\theta}\right]$, vary across populations and between the FAMILINX-inspired models and the Wright–Fisher model. One plausible biological explanation for the differences is a different frequency of half siblings compared to full-siblings across the models. However, we note that a number of other factors could have affected the observed differences. For example, as mentioned above, N_e was defined as the mean coalescence time over pedigrees; using a different definition of N_e or a different method for fixing N_e in the model (see Supplementary material Section S3) could have led to different conclusions. Additionally, our FAMILINX-based models are not necessarily a truly faithful representation of actual human populations. For example, the real populations we considered are likely growing, while we have coerced the genealogies into a constant size population. Moreover, since FAMILINX is based on voluntarily donated data, different countries may differ in both the sub-populations represented, as well as in the genealogical error rate.

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 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 fraction, 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 Fig. 5 vs the scaled recombination rate $\rho = 4N_e r$, showing perfect agreement with simulations.

These results 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 $\rho = 4N_e r$. 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 Eqs. (16)–(18) are also equal to their respective correlation coefficients, since $\mathrm{Var}[T] = 1$ under either the ARG, SMC, and SMC'). In Fig. 6, we compare the correlation of T_i and T_j across the different models as a function of ρ 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 Fig. 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 0.005 (see also Fig. 2). Therefore, the simplified model should be preferred due its much reduced complexity. For $N_e = 10$, we observe a more pronounced difference between the 2-sex and the simplified DDTWF models, with a maximal difference around 0.025.

Please cite this article in press as: King, L., et al., A non-zero variance of Tajima's estimator for two sequences even for infinitely many unlinked loci. Theoretical Population Biology (2017), http://dx.doi.org/10.1016/j.tpb.2017.03.002.

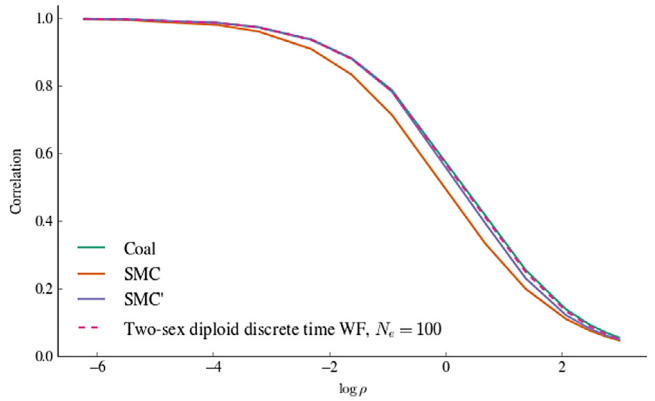

Fig. 6. A comparison of the correlation coefficient of the coalescence times at two linked loci under models of increasing complexity. We compare the ARG, SMC, SMC', and the 2-sex DDTWF with $N_e = 100$, across different values of $\rho = 4N_e r$. The predictions of the ARG and SMC' are very good approximations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

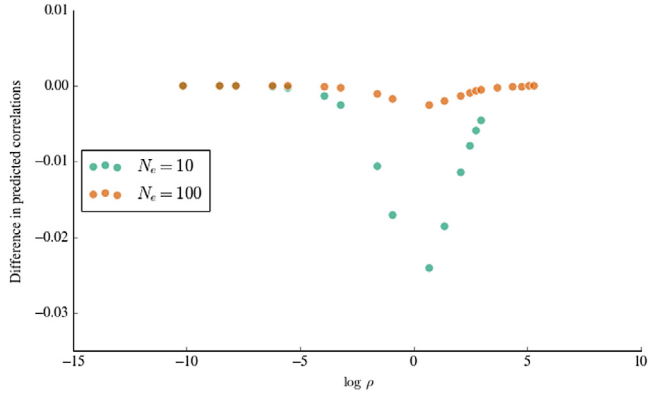

Fig. 7. A comparison of the correlation coefficient of the coalescence times at two linked loci between the 2-sex and the simplified DDTWF models. We plot the difference between the correlation coefficients of the two models vs ρ for $N_e = 10$ and $N_e = 100$. The predictions of the two models slightly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

7. Summary and discussion

Previous studies of estimators of θ using data from a single locus have revealed properties rather different from classical statistical results for independent samples, due to the non-independence of samples exerted by their shared gene genealogy. In particular, Tajima (1983) demonstrated that the average number of pairwise differences is an inconsistent estimator of θ as the sample size at the locus tends to infinity, and Joyce (1999) showed that there is no linear unbiased estimator using site-frequency information that has a uniformly lower variance than Watterson's estimator (Watterson, 1975). Our work adds a new dimension to such studies by considering the statistical properties of estimators as the number of independently segregating loci tends to infinity, but with non-independence exerted by the population pedigree that all loci share. Specifically, we have shown that even when sampling infinitely many loci, the estimator of θ based on the average number of pairwise differences at many loci is not consistent and has non-zero variance. We provided an approximate lower bound on 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.

As mentioned above, the non-zero variance of $\hat{\theta}$ is a result of the underlying pedigree shared between all loci (Laurie and Weir, 2003; Song and Slatkin, 2007; Bhaskar and Song, 2009). 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 different in the human-inspired models than in the WF model, which should lead to differences in the variance of $\hat{\theta}$.

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, ρ, 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 the 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-locus 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 underlying 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.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.tpb.2017.03.002.

References

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

Bhaskar, A., Song, Y.S., 2009. Multi-locus match probability in a finite population: a fundamental difference between the Moran and Wright-Fisher models. Bioinformatics 25, i187–i195.

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

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

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

Edwards, S.V., Beerli, P., 2000. Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. Evolution 54, 1839–1854.

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

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

Griffiths, R., Marjoram, P., 1997. An ancestral recombination graph. In: Tavaré, S., Donnelly, P. (Eds.), Progress in Population Genetics and Human Evolution. Springer Verlag, pp. 257–270.

Joyce, P., 1999. No BLUE among phylogenetic estimators. J. Math. Biol. 39, 421–438.

Kaplanis, J., Gordon, A., Wahl, M., Gershovits, M., Markus, B., Sheikh, M., Gymrek, M., Bhatia, G., MacArthur, D.G., Price, A., Erlich, Y., 2017. Quantitative analysis of population-scale family trees using millions of relatives. BioRxiv. http://dx.doi.org/10.1101/106427.

Kingman, J.F.C., 1982. The coalescent. Stochastic Process. Appl. 13, 235–248.

Laurie, C., Weir, B.S., 2003. Dependency effects in multi-locus match probabilities. Theor. Popul. Biol. 63, 207–219.

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

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

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

Ohta, T., Kimura, M., 1969. Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. Genetics 63, 229–238.

Pluzhnikov, A., Donnelly, P., 1996. Optimal sequencing strategies for surveying molecular genetic diversity. Genetics 144, 1247–1262.

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

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

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

Song, Y.S., Slatkin, M., 2007. A graphical approach to multi-locus match probability computation: revisiting the product rule. Theor. Popul. Biol. 72, 96–110.

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

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

Wakeley, J., 2009. Coalescent Theory: An Introduction. Roberts & Company Publishers, Greenwood Village, Colorado, USA.

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

Wasserman, L., 2004. All of Statistics: A Concise Course in Statistical Inference. Springer.

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

Weir, B.S., Cockerham, C.C., 1969. Group inbreeding with two linked loci. Genetics 63, 711–742.

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

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

Please cite this article in press as: King, L., et al., A non-zero variance of Tajima's estimator for two sequences even for infinitely many unlinked loci. Theoretical Population Biology (2017), http://dx.doi.org/10.1016/j.tpb.2017.03.002.