Published in final edited form as:

Theor Popul Biol. 2008 August ; 74(1): 56–67. doi:10.1016/j.tpb.2008.04.006.

Summary Statistics of Neutral Mutations in Longitudinal DNA Samples

Xiaoming Liu and Yun-Xin Fu^{*}

Human Genetics Center, School of Public Health, University of Texas at Houston, 1200 Herman Pressler, Houston, Texas 77030, USA

Abstract

Longitudinal samples of DNA sequences are the DNA sequences sampled from the same population at different time points. For fast evolving organisms, e.g. RNA virus, these kind of samples have increasingly been used to study the evolutionary process in action. Longitudinal samples provide some interesting new summary statistics of genetic variation, such as the frequency of mutation of size i in one sample and size j in another, the average number of mutations accumulated since the common ancestor of two sequences each from a different sample, and number of private, shared and fixed mutations within samples. To make the results more applicable, we used in this study a general two-sample model, which assumes two longitudinal samples were taken from the same measurably evolving population. Inspired by the HIV study, we also studied a two-sample-two-stage model, which is a special case of two-sample model and assumes a treatment after the first sampling instantaneously changes the population size. We derived the formulas for calculating statistical properties, e.g. expectations, variances and covariances, of these new summary statistics under the two models. Potential applications of these results were discussed.

Keywords

longitudinal sample; coalescent theory; summary statistics; measurably evolving population

1 Introduction

One fundamental goal of population genetics is to understand the change of genetic variation of a population in the process of evolution. Longitudinal samples, or serial samples, are samples taken at a series of time points from the same population. For fast evolving organisms (e.g. RNA virus), longitudinal DNA/RNA samples are often used to study the change of the population. Other applications of longitudinal DNA/RNA samples include ancient DNA studies (Willerslev and Cooper, 2005) and artificial evolution experiments on RNA enzymes (Johns and Joyce, 2005). Several new statistical methods have been developed for analyzing longitudinal DNA samples (reviewed by Drummond et al., 2003). Most of the methods are based on likelihood frameworks (e.g. Rodrigo and Felsenstein, 1999; Rambaut, 2000; Drummond et al., 2001; Seo et al., 2002; Drummond et al., 2002; Rodrigo et al., 2003), that try to capture all information from the data and thus to be more

^{*}Corresponding author. Tel.: (713)-500-9813, Fax: (713)-500-0900, Yunxin.Fu@uth.tmc.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

powerful. However, there are some difficulties faced by those methods, such as high computational intensity and high model dependency. Especially, calculating likelihood of haplotypes with recombination is computationally so intensive that unrealistic assumptions such as no recombination or free recombination between markers are often made. Such assumptions may cause large bias on parameter estimations. On the other hand, using summary statistics may suffer from some information loss. But with carefully chosen summary statistics, it is possible to find an effective way to capture the essential information in a sample, while avoiding the bias introduced and some difficulties faced by likelihood-base methods.

In this paper we studied some new summary statistics of mutation number provided by longitudinal samples, namely the frequency of mutation of size i in one sample and size j in another, the average number of mutations accumulated since the common ancestor of two sequences each from a different sample, and number of private, shared and fixed mutations within samples (see definitions below). Assuming no selection and no recombination, we derived the formulas for calculating the expectations, variances and covariances of these summary statistics. The statistical properties of these summary statistics are independent to mutation models. Under the infinite site model, the number of mutations and the number of segregating sites is equal, so that the results of this paper can be applied directly to the sequence data. Under a finite site model, the number of mutations and the number of segregating sites is not equal. But as long as the mutation model is given, the number of mutations can be recovered adequately in most cases and these summary statistics of mutations can be widely applicable also. While the expectations of these summary statistics are not affected by recombination events, their variances will be smaller when there is recombination. So that, with applying some estimators or test statistics built on the variances of these summary statistics, cautions are needed when interpreting the results, if recombination can not be ignored.

In the remaining of this paper we first introduced the definitions of different summary statistics with our two-sample model and its special case, the two-sample-two-stage model, which is inspired by HIV treatment. Then we showed the formulas for calculating the expectations, variances and covariances of these summary statistics. Finally, we discussed their potential applications. Most technical details can be found in Appendix.

2 Summary statistics for longitudinal samples

Among the most well-known summary statistics for DNA polymorphism (segregating sites) are the number of mutations of various sizes (ξ_i), number of segregating sites (K) and average difference between two sequences (Π) (see definitions below). The properties of these statistics have been well studied for the case of single sample assuming infinite site model (Watterson, 1975; Tajima, 1983; Tavaré, 1984; Fu, 1995). Since in infinite site model each new mutation produce a new segregating site, segregating site and mutation are mutually exchangeable. In this paper, we borrowed the notations of these classic summary statistics for our new statistics studied, but keep in mind our new summary statistics are for mutations. They can apply to sequence analysis directly with assumption of the infinite site model. But with finite site model, a transformation to statistics of segregating is needed.

We begin with the definition of these summary statistics under the infinite site model. A mutation on the genealogy of a DNA sample divides the sample into two groups. One consists of sequences that inherit the mutation and the other does not. If there are i descendants of the mutation in the sample, the mutation is said to be size i . Let ξ_i be the total number (or frequency) of mutations of size i in that sample, the range of possible i is 1 to $n - 1$, where n is the sample size. The total number of segregating sites, K , is just the total

number of mutations on the genealogy of the sample after their most recent common ancestor (MRCA). It can also be regarded as the sum of all ξ_i , i.e.

$$K = \sum_{i=1}^{n-1} \xi_i. \quad (1)$$

The average number of difference between two sequences, Π , is defined as

$$\Pi = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^n d_{ij} \quad (2)$$

where d_{ij} is the number of differences between the i -th and j -th sequence of the sample of the total n sequences. Again Π can be written as a linear combination of ξ_i , i.e.

$$\Pi = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} i(n-i) \xi_i. \quad (3)$$

An example of a genealogy of five sequences was presented in Figure 1. The sample sequences and their MRCA sequence were shown at the bottom and top of the tree. There were four mutations (shown as gray dots) on the genealogy with mutant ancestor sequences shown on the left-hand side. The t_i ($i \geq 2$) shown on the right-hand side of the tree was the i -th coalescent time. If we define state i as the state when there are exact i lines on the genealogy (or i ancestral sequences of the sample), the i -th coalescent time is the time duration in state i . Given these sequences, it is easy to calculate that $\xi_1 = 2$, $\xi_2 = \xi_3 = 1$, $\xi_4 = 0$, $K = 4$, $\Pi = 2$.

The above summary statistics describe the mutations in a single sample. Because longitudinal samples consist of more than one correlated samples, there are some interesting summary statistics describing the variation between samples. The purpose of this study is to investigate the statistical properties of these summary statistics, based on which methods for analyzing longitudinal samples can be developed.

The type of summary statistics we were interested in this paper was best illustrated in Figure 2. It showed a genealogy of three longitudinal samples s_1 , s_2 and s_3 sampled at three different time points, and each sample had three sequences. There were five mutations numbered 1 to 5 on the genealogy, which were shown as gray dots. Consider a pair of longitudinal samples first. Define ξ_{ij} be the number of mutations which is size i in the earlier sample and size j in the later sample. If sample s_1 is assumed to be the earlier sample and sample s_2 as the later sample, we can see: mutation 1 is size 1 in both samples; mutation 2 is size 1 in sample s_1 and size 0 in sample s_2 ; mutation 3 is size 0 in sample s_1 and size 1 in sample s_2 . So that, for sample s_1 and s_2 , $\xi_{11} = \xi_{10} = \xi_{01} = 1$ and all other ξ_{ij} s equal 0. Now look at sample s_2 and sample s_3 . We can see: mutation 3 is size 1 in sample s_2 and size 3 in sample s_3 ; mutation 4 is size 0 in sample s_2 and size 3 in sample s_3 ; mutation 5 is size 0 in sample s_2 and size 2 in sample s_3 . Assume sample s_2 as the earlier sample and sample s_3 as the later sample, then $\xi_{13} = \xi_{03} = \xi_{02} = 1$ and all other ξ_{ij} s equal 0.

We can divide ξ_{ij} s into different groups to form new summary statistics. Define the number of private mutations, K_p , as the number of mutations which are polymorphic only in one sample but not the other. K_p can be further divided to K_{p1} and K_{p2} , the number of mutations

which is polymorphic only in the earlier sample or the later sample, respectively. Define the number of shared mutations, K_s , as the number of mutations which are polymorphic in both samples. Define the number of fixed mutations, K_f , as the number of mutations which is monomorphic in both samples but the monomorphic type of the earlier sample is different to that of the later sample. Write them in the form of functions of ξ_{ij} s:

$$K_{p1} = \sum_{i=1}^{n_1-1} (\xi_{i0} + \xi_{in_2}) \quad (4)$$

$$K_{p2} = \sum_{i=1}^{n_2-1} (\xi_{0i} + \xi_{n_1i}) \quad (5)$$

$$K_s = \sum_{i=1}^{n_1-1} \sum_{j=1}^{n_2-1} \xi_{ij} \quad (6)$$

$$K_f = \xi_{0n_2} + \xi_{n_10} \quad (7)$$

where n_1 and n_2 are sizes of the first and the second sample, respectively. Look at sample s_1 and s_2 again, we can see mutation 1 is a shared mutation of sample s_1 and s_2 , mutation 2 is a private mutation of sample s_1 and mutation 3 is a private mutation of sample s_2 . If we look at sample s_2 and s_3 , then mutation 3 is a private mutation of sample s_2 , mutation 4 is a fixed mutation and mutation 5 is a private mutation of sample s_3 . So that, for s_1 and s_2 , $K_{p1} = K_{p2} = K_s = 1$ and $K_f = 0$. For s_2 and s_3 , $K_{p1} = K_{p2} = K_f = 1$ and $K_s = 0$. Actually, if we only consider two longitudinal samples, shared mutations and fixed mutations are mutually exclusive, that is, either K_s or K_f must equal 0 (Wakeley and Hey, 1997). This fact is easy to understand by considering the possible position of mutations on the genealogy.

As to the average number of mutations accumulated since the common ancestor of two sequences (i.e. the average differences between two sequences under the infinite site model), other than the cases both sequences are from the same sample, now we can define Π_{12} as the average number of mutations accumulated since the common ancestor of two sequences, with one sequence from the earlier sample and one sequence from the later sample:

$$\Pi_{12} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} d_{ij}^* \quad (8)$$

where d_{ij}^* is the number of mutation accumulated on the i -th sequence of the earlier sample and the j -th sequence of the later sample after their MRCA, and again n_1 and n_2 are sizes of the two samples. Write it as a function of ξ_{ij} s:

$$\Pi_{12} = \frac{1}{n_1 n_2} \sum_{i=0}^{n_1} \sum_{j=0}^{n_2} [(n_1 - i)j + (n_2 - j)i] \xi_{ij}. \quad (9)$$

3 Two-sample model and two-sample-two-stage model

One of the simple models for longitudinal samples is the two-sample model. Under it only two longitudinal samples are considered. So that ξ_{ij} , K_p , K_s , K_f and Π_{12} are all summary statistics for two longitudinal samples. In this model, we assumed the population followed the commonly used Wright-Fisher model for haploids (Fisher, 1930; Wright, 1931) and the population size change follows a determinate function

$$N(t) = N(0) \omega(t) \quad (10)$$

where $\omega(t)$ is a function of time t ; $N(0)$ and $N(t)$ are the population sizes at time 0 and time t , respectively. The second sample (sample 2, size n_2) is taken at time 0 and the first sample (sample 1, size n_1) is taken at g generations before. (We followed the retrospective view of coalescent theory and defined time 0 as the time point when sample 2 was taken. However, the “before” and “after” mentioned in this paper still referred to the normal prospective view.) Unless otherwise mentioned, time in the paper was scaled in $N(0)$ generations and let $T = g/N(0)$. Let N_1 and N_2 be the population size at time points of the first sampling and the second sampling, respectively. That is, $N_1 = N(T)$ and $N_2 = N(0)$. We further assumed that all mutations were neutral; the mutation rate was μ per sequence per generation and remained constant over time; and the number of mutations accumulated in a given period of time τ follow the Poisson distribution with parameter $\mu\tau$. Some other parameters are defined as following: $\lambda = N_1/N_2$, $\theta_1 = 2N_1\mu$, $\theta_2 = 2N_2\mu$.

In HIV study, people can use longitudinal samples to study the effect of a treatment, for example, giving a treatment after the first sampling and taking a second sample after a period. In this case we may assume a two-sample-two-stage model, which is a special case of the general two-sample model. Figure 3 illustrated the model with two longitudinal samples each with three sequences. In this model, we assume that before the earlier sample (sample 1) was taken, the effective population size, N_1 , remained constant. Then right after sample 1 was taken, a treatment is given and the effective population size experienced an instantaneously change to N_2 , then it remained constantly thereafter. g generations after sample 1 was taken, the later sample, sample 2, was taken. This two-sample-two-stage model is closely related to Wakeley and Hey (1997)’s isolation model and size-change model. Actually, some of the expectations of the new summary statistics in this study were derived from their results. However, they provided no formulas for calculating the variance and covariances of the statistics, besides they missed some other important expectations. That largely limited the application of these statistics and we hope this study can fill these gaps.

The basis of two-sample model is coalescent times. Let $t_m(T)$ and $t_m(0)$ be the m -th coalescent times of samples taken at time T and time 0, respectively. Let t_{ma} and t_{mb} be the part of $t_m(0)$ after and before T , respectively. Suppose at T , there are k ancestral sequences of sample 2. Then given k , t_{ka} and t_{kb} are independent; if $m < k$, $t_{ma} = 0$ and $t_{mb} = t_m(0)$; if $m > k$, $t_{ma} = t_m(0)$ and $t_{mb} = 0$. If $k = 1$, as shown in Figure 3, define $t_1(0) = t_{1a}$ as the time interval between T and the MRCA of sample 2, that is,

$$t_1(0) = \max \left(0, T - \sum_{m=2}^{n_2} t_m(0) \right) \quad (11)$$

where $\max(\cdot)$ stands for maximum.

With the probability densities of coalescent times under the two-sample model or two-sample-two-stage model (see Appendix A and B) we can calculate $E(t_m(T))$, $E(t_{ma})$ and $E(t_{mb})$. Then all the means of the summary statistics studied in this paper can be calculated directly (see details in following sections). Variances and covariances of the summary statistics can also be derived under the two-sample model, which need

$E(t_m(T) t_{m'}(T))$, $E(t_m^2(T))$, $E(t_{ma}^2)$ and $E(t_{ma} t_{m'a})$, where $m' \neq m$. The mathematics becomes more complicated. For a simpler model, such as the two-sample-two-stage model, it is possible to calculate them directly (see Appendix B). For more complicated demographic change, simulation may be used to obtain those quantities. Since only coalescent times need to be simulated, the simulation process can be very fast and accurate estimations of these quantities can be obtained practically.

4 Number of mutations of various sizes in two samples

4.1 Calculate $E(\xi_{ij})$

As defined in Section 2, ξ_{ij} is the number of mutations which are size i in sample 1 and size j in sample 2. Further define ξ_{ijb} and ξ_{ija} as the parts of ξ_{ij} that occurred before and after T , respectively. Assume there are n' ($n' \leq n_2$) ancestral sequences of sample 2 at time T , then

$$E(\xi_{ijb}|n') = \begin{cases} \sum_{k'=1}^{\min(n'-1, j)} P(k' \rightarrow j|n_2, n') \frac{\binom{n_1}{i} \binom{n'}{k'}}{\binom{n_1+n'}{i+k'}} E(\xi_{i+k'|n_1+n'}(T)) & \text{if } 0 < j < n_2 \\ \frac{\binom{n_1}{i}}{\binom{n_1+n'}{i}} E(\xi_{i|n_1+n'}(T)) & \text{if } j=0 \\ \frac{\binom{n_1}{i}}{\binom{n_1+n'}{i+n'}} E(\xi_{i+n'|n_1+n'}(T)) & \text{if } j=n_2 \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

$$E(\xi_{ija}|n') = \begin{cases} \frac{\theta_2}{2} \sum_{k=\max(2, n')}^{n_2} k P(k, j|n_2) E(t_{ka}|n') & \text{if } i=0, 1 < j < n_2 \\ \frac{\theta_2}{2} E(t_{1a}|1) & \text{if } i=0, j=n_2, n'=1 \\ 0 & \text{otherwise} \end{cases} \quad (13)$$

where $E(\cdot|n')$ means the expectation of some statistic, specified by $\cdot$, given there are n' ancestral sequences of sample 2 at time T ; $\min(\cdot)$ stands for the minimum; $\xi_{i|n}(T)$ is the number of mutations of size i in a sample of size n and taken at time T ; $P(k' \rightarrow j|n, k)$ is a function of k', j, n, k , whose meaning and formula can be found in Appendix C.

The properties of $\xi_{i|n}(T)$ have been studied by Fu (1995) and for $1 \leq j < n$

$$E(\xi_j(T)) = \sum_{k=2}^n k P(k, j|n) E(\zeta_k(T)) \quad (14)$$

where $P(k, i|n)$ is a function of k, i and n , whose meaning and formula can be found in Appendix C; $\zeta_k(T)$ is the number of mutations accumulated on a line of state k .

$$E(\zeta_k(T)) = \frac{\theta_2}{2} E(t_k(T)) \quad (15)$$

Under two-sample-two-stage model, it can be simplified (Fu, 1995) as

$$E_{TT}(\zeta_k(T)) = \frac{\theta_1}{k(k-1)} \quad (16)$$

where “ E_{TT} ” means the expectation are only applicable to the two-sample-two-stage model. Substitute (16) into (14) and (12), we have

$$E_{TT}(\xi_j(T)) = \frac{\theta_1}{k} \quad (17)$$

(Fu, 1995) and

$$E_{TT}(\xi_{ijb}|n') = \begin{cases} \theta_1 \sum_{k'=1}^{\min(n'-1, j)} P(k' \rightarrow j|n_2, n') \frac{\binom{n_1}{i} \binom{n'}{k'}}{\binom{n_1+n'}{i+k'}} \frac{1}{(i+k')} & \text{if } 0 < j < n_2 \\ \theta_1 \frac{\binom{n_1}{i}}{\binom{n_1+n'}{i}} \frac{1}{i} & \text{if } j=0 \\ \theta_1 \frac{\binom{n_1}{i}}{\binom{n_1+n'}{i+n'}} \frac{1}{i+n'} & \text{if } j=n_2 \\ 0 & \text{otherwise} \end{cases} \quad (18)$$

For the shared mutations, the expectation of ξ_{ij} can be derived from the results of Wakeley and Hey (1997):

$$E(\xi_{ij}) = E(\xi_{ijb}) \quad (19)$$

$$= \sum_{k=2}^{n_2} P_{n_2, k}(T) \left[\sum_{k'=1}^{k-1} P(k' \rightarrow j|n_2, k) \frac{\binom{n_1}{i} \binom{k}{k'}}{\binom{n_1+k}{i+k'}} E(\xi_{i+k'}(T)) \right] \quad (20)$$

$$E_{TT}(\xi_{ij}) = \theta_1 \sum_{k=2}^{n_2} P_{n_2,k}(T) \left[\sum_{k'=1}^{k-1} P(k' \rightarrow j | n_2, k) \frac{\binom{n_1}{i} \binom{k}{k'}}{\binom{n_1+k}{i+k'}} \frac{1}{(i+k')} \right] \quad (21)$$

with $0 < i \leq n_1$, $0 < j < n_2$, where $P_{n,k}(T)$ is a function of n , k and T , whose meaning and formula can be found in Appendix C.

Similarly, we can compute the expectations for the fixed or private mutations:

$$E(\xi_{i0}) = E(\xi_{i0b}) \quad (22)$$

$$= \sum_{k=1}^{n_2} P_{n_2,k}(T) \frac{\binom{n_1}{i}}{\binom{n_1+k}{i}} E(\xi_i(T)) \quad (23)$$

$$E_{TT}(\xi_{i0}) = \theta_1 \sum_{k=1}^{n_2} P_{n_2,k}(T) \frac{\binom{n_1}{i}}{\binom{n_1+k}{i}} \frac{1}{i} \quad (24)$$

$$E(\xi_{in_2}) = E(\xi_{in_2b}) \quad (25)$$

$$= \sum_{k=1}^{n_2} P_{n_2,k}(T) \frac{\binom{n_1}{i}}{\binom{n_1+k}{i+k}} E(\xi_{i+k}(T)) \quad (26)$$

$$E_{TT}(\xi_{in_2}) = \theta_1 \sum_{k=1}^{n_2} P_{n_2,k}(T) \frac{\binom{n_1}{i}}{\binom{n_1+k}{i+k}} \frac{1}{(i+k)} \quad (27)$$

$$E(\xi_{0n_2b}) = \sum_{k=1}^{n_2} P_{n_2,k}(T) \frac{1}{\binom{n_1+k}{k}} E(\xi_k(T)) \quad (28)$$

$$E_{TT}(\xi_{0n_2b}) = \theta_1 \sum_{k=1}^{n_2} P_{n_2,k}(T) \frac{1}{\binom{n_1+k}{k}} \frac{1}{k} \quad (29)$$

$$E(\xi_{0n_2a}) = \frac{\theta_2}{2} E(t_{1a}) \quad (30)$$

$$E(\xi_{0n_2}) = E(\xi_{0n_2b}) + E(\xi_{0n_2a}) \quad (31)$$

$$E(\xi_{0jb}) = \sum_{k=2}^{n_2} P_{n_2,k}(T) \left[\sum_{k'=1}^{\min(k-1,j)} P(k' \rightarrow j|n_2, k) \frac{\binom{k}{k'}}{\binom{n_1+k}{k'}} E(\xi_{k'}(T)) \right] \quad (32)$$

$$E_{TT}(\xi_{0jb}) = \theta_1 \sum_{k=2}^{n_2} P_{n_2,k}(T) \left[\sum_{k'=1}^{\min(k-1,j)} P(k' \rightarrow j|n_2, k) \frac{\binom{k}{k'}}{\binom{n_1+k}{k'}} \frac{1}{k'} \right] \quad (33)$$

$$E(\xi_{0ja}) = \frac{\theta_2}{2} \sum_{k'=2}^{n_2} k' P(k', j|n_2) E(t_{k'a}) \quad (34)$$

$$= \frac{\theta_2}{2} \sum_{k=1}^{n_2} P_{n_2,k}(T) \left[\sum_{k'=\max(k,2)}^{n_2} k' P(k', j|n_2) E(t_{k'a}|k) \right] \quad (35)$$

$$E(\xi_{0j}) = E(\xi_{0jb}) + E(\xi_{0ja}) \quad (36)$$

where $0 < i \leq n_1$, $0 < j < n_2$, except $0 < i < n_1$ in (25).

4.2 Calculate $E(\xi_{ij}^2)$ and $E(\xi_{ij}\xi_{lm})$

Again, assume there are n' ($n' \leq n_2$) ancestral sequences of sample 2 at time T . $E(\xi_{ijb}\xi_{lmb}|n')$ can be computed as

$$E(\xi_{ijb}\xi_{lmb}|n')$$

$$= \delta_{(i=l, j=m)} \sum_{k=2}^{n_1+n'} \sum_{j'=0}^{n'} k P(k, i+j'|n_1+n') \frac{\binom{n_1}{i} \binom{n'}{j'}}{\binom{n_1+n'}{i+j'}} P(j' \rightarrow j|n_2, n') E(\xi_k^2(T)) + \delta_{(i+l \leq n_1, j+m \leq n_2)} \sum_{k=2}^{n_1+n'} \sum_{j'=0}^{n'} \sum_{m'=0}^{n'-j'} k(k-1) P(k, i+j$$

(37)

and

$$\begin{aligned} E_{TT}(\xi_{ijb}\xi_{lmb}|n') = & \delta_{(i=l, j=m)} \sum_{j'=0}^{\min(n', j)} \frac{\binom{n_1}{i} \binom{n'}{j'}}{\binom{n_1+n'}{i+j'}} P(j' \rightarrow j|n_2, n') \left(\frac{\theta_1}{i+j'} + \beta_{n_1+n'}(i+j') \theta_1^2 \right) + \delta_{(i+l=n_1, j+m=n_2)} \theta_1^2 \sum_{j'=0}^{\min(n', j)} \frac{\binom{n_1}{i} \binom{n'}{j'}}{\binom{n_1+n'}{i+j'}} P(j' \rightarrow j|n_2, n') \\ & \left(\frac{a_{n_1+n'}-a_{i+j'}}{n_1+n'-i-j'} + \frac{a_{n_1+n'}-a_{i+n'-j'}}{n_1+j'-l} - \frac{\beta_{n_1+n'}(i+j') + \beta_{n_1+n'}(l+n'-j')}{2} \right) + \delta_{(i+l < n_1, j+m=n_2)} \theta_1^2 \sum_{j'=0}^{\min(n', j)} \frac{\binom{i+j'}{i} \binom{l+n'-j'}{l}}{\binom{n_1+n'}{n_1}} P(j' \rightarrow j|n_2, n') \\ & \left(\frac{1}{(i+j')(l+n'-j')} - \frac{\gamma_{n_1+n'}(i+j') + \gamma_{n_1+n'}(l+n'-j')}{2} \right) + \delta_{(i+l \leq n_1, j+m < n_2)} \theta_1^2 \sum_{j'=0}^{\min(n'-1, j)} \sum_{m'=0}^{\min(n'-1-j', m)} \frac{\binom{i+j'}{i} \binom{l+m'}{l} \binom{n_1+n'-i-j'-l-m'}{n_1-i-l}}{\binom{n_1+n'}{n_1}} \\ & P(j' \rightarrow j, m' \rightarrow m|n_2, n') \left(\frac{1}{(i+j')(l+m')} - \frac{\gamma_{n_1+n'}(i+j') + \gamma_{n_1+n'}(l+m')}{2} \right) + \delta_{(i > l, j=m)} \theta_1^2 \sum_{j'=0}^{\min(n', j)} \frac{\binom{l+j'}{l} \binom{n_1+n'-i-j'}{n_1-i}}{\binom{n_1+n'}{n_1}} P(j' \rightarrow j|n_2, n') \frac{\gamma_{n_1+n'}(l+j')}{2} + \delta_{(i \geq l, j > m)} \theta_1^2 \sum_{j'=1}^{\min(n', j)} \sum_{m'=0}^{\min(m, j'-1)} \frac{n_1+n'-n_1-l-m'}{n_1-l} \\ & P(j'-m' \rightarrow j-m, m' \rightarrow m|n_2, n') \frac{\gamma_{n_1+n'}(l+m')}{2} + \delta_{(l > i, m=j)} \theta_1^2 \sum_{j'=0}^{\min(n', j)} \frac{\binom{i+j'}{i} \binom{n_1+n'-l-j'}{n_1-l}}{\binom{n_1+n'}{n_1}} P(j' \rightarrow j|n_2, n') \frac{\gamma_{n_1+n'}(l+j')}{2} + \delta_{(l \geq i, m > j)} \theta_1^2 \sum_{m'=1}^{\min(n', m)} \sum_{j'=0}^{\min(m'-1, j)} \frac{n_1+n'-l-m'}{n_1-l} \frac{\binom{i+j'}{i}}{\binom{n_1+n'}{n_1}} \\ & P(m'-j' \rightarrow m-j, j' \rightarrow j|n_2, n') \frac{\gamma_{n_1+n'}(i+j')}{2} \end{aligned}$$

(38)

where

$$E(\xi_k^2(T)) = \frac{\theta_2}{2} E(t_k(T)) + \left(\frac{\theta_2}{2} \right)^2 E(t_k^2(T)) \quad (39)$$

$$E_{TT}(\xi_k^2(T)) = \frac{1}{k(k-1)} \theta_1 + \frac{2}{k^2(k-1)^2} \theta_1^2; \quad (40)$$

$$E(\zeta'_k(T) \zeta_k(T)) = \left(\frac{\theta_2}{2}\right)^2 E(t_k^2(T)) \quad (41)$$

$$E_{TT}(\zeta'_k(T) \zeta_k(T)) = \frac{2}{k^2(k-1)^2} \theta_1^2; \quad (42)$$

$$E(\zeta_k(T) \zeta_{k'}(T)) = \left(\frac{\theta_2}{2}\right)^2 E(t_k(T) t_{k'}(T)) \quad (43)$$

$$E_{TT}(\zeta_k(T) \zeta_{k'}(T)) = \frac{1}{k(k-1)k'(k'-1)} \theta_1^2; \quad (44)$$

$\delta_{(r)}$ is an index function that takes the value 1 if the relationship r is true and 0 otherwise; a_n is a function of n ; $\beta_n(i)$ and $\gamma_n(i)$ are functions of n and i ; $P(j' \rightarrow j, m' \rightarrow m | n, n')$ is a function of j', j, m', m, n' and n ; $P(k, i; k, j | n)$ is a function of i, j, k and n ; $P_a(k, i; k', j | n)$ and $P_b(k, i; k', j | n)$ are functions of i, j, k, k' and n ; the meanings and formulas of these functions can be found in Appendix C.

Then

$$E(\xi_{ijb} \xi_{lmb}) = \sum_{n'=1}^{n_2} P_{n_2, n'}(T) E(\xi_{ijb} \xi_{lmb} | n') \quad (45)$$

For $E(\xi_{ija} \xi_{lma})$, consider the following conditions:

-
1	$i = l = 0, j < n_2, m < n_2$
2	$i = l = 0, j < n_2, m = n_2$
3	$i = l = 0, j = m = n_2$
4	otherwise
-

Condition 1: $i = l = 0, j < n_2, m < n_2$

$$\begin{aligned}
& E(\xi_{0ja}\xi_{0ma}) \\
&= \delta_{(j=m)} \sum_{k=2}^{n_2} k P(k, j|n_2) E(\zeta_{ka}^2) \\
&+ \delta_{(j+m \leq n_2)} \sum_{k=2}^{n_2} k(k-1) P(k, j; k, m|n_2) E(\zeta_{ka}\zeta'_{ka}) \\
&+ \sum_{k=2}^{n_2-1} \sum_{k'=k+1}^{n_2} kk' [P(k, j; k', m|n_2) + P(k, m; k', j|n_2)] E(\zeta_{ka}\zeta'_{k'a}).
\end{aligned} \tag{46}$$

Condition 2: $i = l = 0, j < n_2, m = n_2$

$$E(\xi_{0ja}\xi_{0n_2a}) = \sum_{k=2}^{n_2} k P(k, j|n_2) E(\zeta_{ka}\zeta_{1a}) \tag{47}$$

Condition 3: $i = l = 0, j = m = n_2$

$$E(\xi_{0n_2a}^2) = E(\zeta_{1a}^2) \tag{48}$$

Condition 4: otherwise

$$E(\xi_{ija}\xi_{lma}) = 0 \tag{49}$$

where ζ_{ka} and ζ'_{ka} are the numbers of mutations accumulated on two different lines of state k after T ;

$$E(\zeta_{ka}^2) = \frac{\theta_2}{2} E(t_{ka}) + \left(\frac{\theta_2}{2}\right)^2 E(t_{ka}^2) \tag{50}$$

$$E(\zeta_{ka}\zeta'_{ka}) = \left(\frac{\theta_2}{2}\right)^2 E(t_{ka}^2) \tag{51}$$

$$E(\zeta_{ka}\zeta'_{k'a}) = \left(\frac{\theta_2}{2}\right)^2 E(t_{ka}t_{k'a}). \tag{52}$$

Since the numbers of mutations before and after T are independent to each other given the state at T ,

$$E(\xi_{ij}\xi_{lm}|n') = \begin{cases} E(\xi_{ijb}\xi_{lmb}|n') & \text{if } i>0, l>0 \\ E(\xi_{ijb}\xi_{0mb}|n') + E(\xi_{ijb}|n') E(\xi_{0ma}|n') & \text{if } i>0, l=0 \\ E(\xi_{0jb}\xi_{lmb}|n') + E(\xi_{lmb}|n') E(\xi_{0ja}|n') & \text{if } i=0, l>0 \\ E(\xi_{0jb}\xi_{0mb}|n') + E(\xi_{0ja}\xi_{0ma}|n') & \\ + E(\xi_{0jb}|n') E(\xi_{0ma}|n') & \text{if } i=l=0 \\ + E(\xi_{0mb}|n') E(\xi_{0ja}|n') & \end{cases} \quad (53)$$

$$E(\xi_{ij}\xi_{lm}) = \sum_{n'=1}^{n_2} P_{n_2, n'}(T) E(\xi_{ij}\xi_{lm}|n'). \quad (54)$$

Finally,

$$Var(\xi_{ij}) = E(\xi_{ij}\xi_{ij}) - E^2(\xi_{ij}) \quad (55)$$

$$Cov(\xi_{ij}, \xi_{lm}) = E(\xi_{ij}\xi_{lm}) - E(\xi_{ij}) E(\xi_{lm}). \quad (56)$$

5 Number of private, shared and fixed mutations

The expectations of $E(K_f)$, $E(K_s)$, $E(K_{p1})$ and $E(K_{p2})$ can be derived from the results of Wakeley and Hey (1997). That is,

$$E(K_f \text{ before } T) = \sum_{k=1}^{n_2} P_{n_2, k}(T) \frac{E(\xi_{n_1}(T)) + E(\xi_k(T))}{\binom{n_1+k}{n_1}} \quad (57)$$

$$E_{TT}(K_f \text{ before } T) = \theta_1 \sum_{k=1}^{n_2} P_{n_2, k}(T) \frac{\left(\frac{1}{n_1} + \frac{1}{k}\right)}{\binom{n_1+k}{n_1}} \quad (58)$$

$$E(K_f \text{ after } T) = \frac{\theta_2}{2} E(t_{1a}) \quad (59)$$

$$E_{TT}(K_f \text{ after } T) = \frac{\theta_2 T}{2} - \frac{\theta_2}{2} \sum_{i=2}^{n_2} \frac{1 - \exp\left(-\binom{i}{2} T\right)}{\binom{i}{2} \prod_{j=2, j \neq i}^{n_2} \left[1 - \frac{i(i-1)}{j(j-1)}\right]} \quad (60)$$

$$E(K_f) = E(K_f \text{ before } T) + E(K_f \text{ after } T) \quad (61)$$

$$E(K_s) = \sum_{k=2}^{n_2} P_{n_2,k}(T) \left[\frac{E(\xi_{n_1}(T)) + E(\xi_k(T))}{\binom{n_1+k}{n_1}} - \sum_{k'=k}^{n_1+k-1} E(\xi_{k'}(T)) \right] + (1 - P_{n_2,1}(T)) \sum_{k'=1}^{n_1-1} E(\xi_{k'}(T)) \quad (62)$$

$$E_{TT}(K_s) = \theta_1 \sum_{k=2}^{n_2} P_{n_2,k}(T) \left[a_{n_1+k} - a_{n_1+k-1} - \frac{\left(\frac{1}{n_1} + \frac{1}{k}\right)}{\binom{n_1+k}{n_1}} \right] \quad (63)$$

$$E(K_{p_1}) = \sum_{k=1}^{n_2} P_{n_2,k}(T) \left[\sum_{k'=k}^{n_1+k-1} E(\xi_{k'}(T)) - \frac{E(\xi_{n_1}(T)) + E(\xi_k(T))}{\binom{n_1+k}{n_1}} \right] \quad (64)$$

$$E_{TT}(K_{p_1}) = \theta_1 \sum_{k=1}^{n_2} P_{n_2,k}(T) \left[a_{n_1+k} - a_k - \frac{\left(\frac{1}{n_1} + \frac{1}{k}\right)}{\binom{n_1+k}{n_1}} \right] \quad (65)$$

$$E(K_{p_2} \text{ before } T) = \sum_{k=2}^{n_2} P_{n_2,k}(T) \left[\sum_{k'=n_1}^{n_1+k-1} E(\xi_{k'}(T)) - \frac{E(\xi_{n_1}(T)) + E(\xi_k(T))}{\binom{n_1+k}{n_1}} \right] \quad (66)$$

$$E_{TT}(K_{p_2} \text{ before } T) = \theta_1 \sum_{k=2}^{n_2} P_{n_2,k}(T) \left[a_{n_1+k} - a_{n_1} - \frac{\left(\frac{1}{n_1} + \frac{1}{k}\right)}{\binom{n_1+k}{n_1}} \right] \quad (67)$$

$$E(K_{p_2} \text{ after } T) = \sum_{k'=1}^{n_2-1} E(\xi_{k'}(0)) - \sum_{k=2}^{n_2} P_{n_2,k}(T) \left[\sum_{k'=1}^{k-1} E(\xi_{k'}(T)) \right] \quad (68)$$

$$E_{TT}(K_{p_2} \text{ after } T) = \theta_2 \left[a_{n_2} - \sum_{k=2}^{n_2} P_{n_2 k}(T) a_k \right] \quad (69)$$

$$E(K_{p_2}) = E(K_{p_2} \text{ before } T) + E(K_{p_2} \text{ after } T) \quad (70)$$

In Section 2 we also showed that K_f , K_s , K_{p_1} and K_{p_2} can also be expressed as the sum of ξ_{ij} using (4–7). Obviously, their expectations can also be calculated using $E(\xi_{ij})$ s, i.e.,

$$E(K_{p_1}) = \sum_{i=1}^{n_1-1} [E(\xi_{i0}) + E(\xi_{in_2})] \quad (71)$$

$$E(K_{p_2}) = \sum_{i=1}^{n_2-1} [E(\xi_{0i}) + E(\xi_{n_1i})] \quad (72)$$

$$E(K_s) = \sum_{i=1}^{n_1-1} \sum_{j=1}^{n_2-1} E(\xi_{ij}) \quad (73)$$

$$E(K_f) = E(\xi_{0n_2}) + E(\xi_{n_10}) \quad (74)$$

We can also use the summation of $E(\xi_{ij}\xi_{lm})$ to calculate the expectations of the products of K_f , K_s , K_{p_1} and K_{p_2} :

$$E(K_f^2) = E(\xi_{0n_2}^2) + E(\xi_{n_10}^2) + 2E(\xi_{0n_2}\xi_{n_10}) \quad (75)$$

$$E(K_s^2) = \sum_{i=1}^{n_1-1} \sum_{j=1}^{n_2-1} \sum_{l=1}^{n_1-1} \sum_{m=1}^{n_2-1} E(\xi_{ij}\xi_{lm}) \quad (76)$$

$$E(K_{p_1}^2) = \sum_{i=1}^{n_1-1} \sum_{j=1}^{n_1-1} E(\xi_{i0}\xi_{j0}) + E(\xi_{in_2}\xi_{jn_2}) + 2E(\xi_{i0}\xi_{jn_2}) \quad (77)$$

$$E(K_{p_2}^2) = \sum_{i=1}^{n_2-1} \sum_{j=1}^{n_2-1} E(\xi_{0i}\xi_{0j}) + E(\xi_{n_1i}\xi_{n_1j}) + 2E(\xi_{0i}\xi_{n_1j}) \quad (78)$$

$$E(K_f K_s) = 0 \quad (79)$$

$$E(K_f K_{p_1}) = \sum_{i=1}^{n_1-1} [E(\xi_{i0}\xi_{0n_2}) + E(\xi_{i0}\xi_{n_1 0}) + E(\xi_{in_2}\xi_{0n_2}) + E(\xi_{in_2}\xi_{n_1 0})] \quad (80)$$

$$E(K_f K_{p_2}) = \sum_{i=1}^{n_2-1} [E(\xi_{0i}\xi_{0n_2}) + E(\xi_{0i}\xi_{n_1 0}) + E(\xi_{n_1 i}\xi_{0n_2}) + E(\xi_{n_1 i}\xi_{n_1 0})] \quad (81)$$

$$E(K_s K_{p_1}) = \sum_{i=1}^{n_1-1} \sum_{j=1}^{n_2-1} \sum_{k=1}^{n_1-1} E(\xi_{ij}\xi_{k0}) + E(\xi_{ij}\xi_{kn_2}) \quad (82)$$

$$E(K_s K_{p_2}) = \sum_{i=1}^{n_1-1} \sum_{j=1}^{n_2-1} \sum_{k=1}^{n_2-1} E(\xi_{ij}\xi_{0k}) + E(\xi_{ij}\xi_{n_1 k}) \quad (83)$$

$$E(K_{p_1} K_{p_2}) = \sum_{i=1}^{n_1-1} \sum_{j=1}^{n_2-1} [E(\xi_{i0}\xi_{0j}) + E(\xi_{i0}\xi_{n_1 j}) + E(\xi_{in_2}\xi_{0j}) + E(\xi_{in_2}\xi_{n_1 j})], \quad (84)$$

with which variances and covariances of K_f , K_s , K_{p_1} and K_{p_2} can be calculated easily.

6 Average number of mutations since the MRCA of two sequences between samples

It is easy to see that

$$E(\Pi_{12}) = 2E(\zeta_2(T)) + \mu g \quad (85)$$

$$E_{TT}(\Pi_{12}) = \theta_1 + \frac{\theta_2 T}{2}. \quad (86)$$

and

$$E(\Pi_{12}^2 | n_1, n_2) = \frac{1}{n_1^2 n_2^2} E \left[\left(\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} d_{ij}^* \right)^2 \right] \quad (87)$$

$$= \frac{1}{n_1 n_2} E(d_{ij}^{*2}) + \frac{n_2 - 1}{n_1 n_2} E(d_{ij}^* d_{ik}^*) + \frac{n_1 - 1}{n_1 n_2} E(d_{ij}^* d_{lj}^*) + \frac{(n_1 - 1)(n_2 - 1)}{n_1 n_2} E(d_{ij}^* d_{lk}^*), \quad (88)$$

where $i \neq l$ and $j \neq k$. Since $E(\Pi_{12}^2 | n_1, n_2)$ can also be expressed and calculated as

$$E(\Pi_{12}^2 | n_1, n_2) = \frac{1}{n_1^2 n_2^2} \sum_{i=0}^{n_1} \sum_{j=0}^{n_2} \sum_{l=0}^{n_1} \sum_{m=0}^{n_2} (n_1 j + n_2 i - 2ij)(n_1 m + n_2 l - 2lm) E(\xi_{ij} \xi_{lm}), \quad (89)$$

$E(d_{ij}^{*2})$, $E(d_{ij}^* d_{kj}^*)$, $E(d_{ij}^* d_{lj}^*)$ and $E(d_{ij}^* d_{lk}^*)$ can be obtained by solving equations for $E(\Pi_{12}^2 | 1, 1)$, $E(\Pi_{12}^2 | 1, 2)$, $E(\Pi_{12}^2 | 2, 1)$ and $E(\Pi_{12}^2 | 2, 2)$. That is

$$\begin{cases} E(d_{ij}^{*2}) = E(\Pi_{12}^2 | 1, 1) \\ E(d_{ij}^* d_{ik}^*) = 2E(\Pi_{12}^2 | 1, 2) - E(\Pi_{12}^2 | 1, 1) \\ E(d_{ij}^* d_{lj}^*) = 2E(\Pi_{12}^2 | 2, 1) - E(\Pi_{12}^2 | 1, 1) \\ E(d_{ij}^* d_{lk}^*) = 4E(\Pi_{12}^2 | 2, 2) + E(\Pi_{12}^2 | 1, 1) - 2E(\Pi_{12}^2 | 1, 2) - 2E(\Pi_{12}^2 | 2, 1) \end{cases}. \quad (90)$$

With $E(\Pi_{12}^2 | n_1, n_2)$ and $E(\Pi_{12})$,

$$\text{Var}(\Pi_{12} | n_1, n_2) = E(\Pi_{12}^2 | n_1, n_2) - E^2(\Pi_{12}) \quad (91)$$

7 Discussion

Longitudinal samples of DNA sequences are the DNA sequences sampled from the same population at different time points. For fast evolving organisms, e.g. RNA virus, these kind of samples have increasingly been used to study the evolutionary process in action.

Longitudinal samples provide some interesting new summary statistics of genetic variation, such as the frequency of mutation of size i in one sample and size j in another (ξ_{ij}), the average number of mutations accumulated since the MRCA of two sequences each from a different sample (Π_{12}), and the number of private, shared and fixed mutations (K_{p1} , K_{p2} , K_s and K_f). All these statistics were studied under the two-sample model, which assumes two longitudinal samples. Because in the application of longitudinal samples of RNA viruses, a sampling is often followed by medical treatments, we also studied a two-sample-two-stage model, which further assumes an instantaneous population size change right after the first sampling and no population size change before or after that event. We derived the formulas for calculating statistical properties, e.g. expectations, variances and covariances, of these new summary statistics that can be used to develop methods for better analysis of longitudinal samples.

The results developed in this paper can be utilized in several ways. One is to estimate various parameters. For example, θ is a very important parameter in population genetics. Fu and Li (1993) showed that a better estimator than the widely used Watterson (1975)'s can be found by partitioning the total number of mutations K into some non-overlapping categories and building an estimator on the more detailed information. The best linear unbiased estimators (BLUEs) were developed on that rational (Fu, 1994a,b). One natural partition for longitudinal samples is ξ_{ij} , which can provide much more detailed partition than that of ξ in a single sample ($n_1 n_2 + n_1 + n_2 - 4$ possible partitions vs. $n - 1$ possible partitions). So

based on ξ_{ij} it is possible to build a better estimator of θ . The construction of BLUE needs to know the estimation, variances and covariance of these summary statistics, which can be computed using the formulas derived in this paper. One advantage of BLUE based on the ξ s over the estimators based on the likelihood of genealogy is its robustness when the sequences have undergone recombination. Recombination does not affect the expectations of ξ s, but it will reduce the variances of them, which actually will improve the accuracy of the estimations. So that, for studying the organisms with relative high recombination rate, such as HIV, the BLUE is preferred. For the parameters whose linear or quadratic relationship with summary statistics does not exist, e.g. growth rate, the BLUE can not apply. One alternative is to use χ^2 estimator developed by Fu and Chakraborty (1998). Another alternative is to use moment based estimators by solving a group of equations numerically, since the estimations of summary statistics can be expressed as functions of these parameters. For instance, Wakeley and Hey (1997) showed an example of using K_{p1} , K_{p2} , K_s and K_f to estimate θ_1 , θ_2 and T under the infinite site model.

Hypothesis tests can also be constructed with the summary statistics studied in this paper. One of the most interested tests in empirical population genetics studies is testing the hypothesis of neutral mutations. Significant progress has been made over the last decade in developing such statistical tests (see reviews by Kreitman, 2000; Ford, 2002). One type of tests uses the difference between two canonical summary statistics of genetic variation within a single population (Tajima, 1989; Fu and Li, 1993; Fu, 1997; Fay and Wu, 2000). Most of the tests of this type are of the form

$$\frac{L_1 - L_2}{\sqrt{\text{Var}(L_1 - L_2)}} \quad (92)$$

where L_1 and L_2 are summary statistics with the expectation of $L_1 - L_2$ equals zero under the null hypothesis that all mutations are selective neutral. Since longitudinal samples provide more summary statistics than a single sample, there exists a potential to find more powerful tests of form (92). Other hypotheses tests useful for longitudinal samples include testing the significance of genetic difference between two longitudinal samples and testing the population size change or mutation rate change. One example was given by Liu and Fu (2007), which proposed several methods for testing genetical isochronism or detecting significant genetical heterochronism between two longitudinal samples. Two best tests named T_{c2} and T_{c3} achieved 75% power with 5% significance level if $\mu g = 1$, $\theta_1 = 10$ and $n_1 = n_2 = 20$. With the same $\mu g = 1$ but much larger $\theta_1 = 40$, they can still achieve about 65% power with 5% significance level if $n_1 = n_2 = 40$. These methods can be used to determine the necessary sample size and sampling interval in experimental design or to combine genetically isochronic samples for better data analysis.

Acknowledgments

This research was supported by National Institutes of Health grants GM60777 and GM50428. We thank Sara Barton for preparing the manuscript. We thank three anonymous reviewers for their comments and suggestions. Part of the work was done during an academic visit to the Institute of Biostatistics, Fudan University, Shanghai, China. We thank Dr. Zewei Luo for his invitation.

References

- Drummond A, Forsberg R, Rodrigo AG. The inference of stepwise changes in substitution rates using serial sequence samples. *Mol Biol Evol*. 2001; 18(7):1365–1371. [PubMed: 11420374]

- Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*. 2002; 161(3):1307–1320. [PubMed: 12136032]
- Drummond AJ, Pybus OG, Rambaut A, Forsberg R, Rodrigo AG. Measurably evolving populations. *Trends Ecol Evol*. 2003; 18(9):481–488.
- Fay JC, Wu CI. Hitchhiking under positive Darwinian selection. *Genetics*. 2000; 155(3):1405–1413. [PubMed: 10880498]
- Fisher, RA. *The Genetical Theory of Natural Selection*. Oxford: Clarendon; 1930.
- Ford MJ. Applications of selective neutrality tests to molecular ecology. *Mol Ecol*. 2002; 11(8):1245–1262. [PubMed: 12144648]
- Fu YX. Estimating effective population size or mutation rate using the frequencies of mutations of various classes in a sample of DNA sequences. *Genetics*. 1994a Dec; 138(4):1375–1386. [PubMed: 7896116]
- Fu YX. A phylogenetic estimator of effective population-size or mutation-rate. *Genetics*. 1994b; 136(2):685–692. [PubMed: 8150291]
- Fu YX. Statistical properties of segregating sites. *Theor Popul Biol*. 1995; 48(2):172–197. [PubMed: 7482370]
- Fu YX. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics*. 1997 Oct; 147(2):915–925. [PubMed: 9335623]
- Fu YX, Chakraborty R. Simultaneous estimation of all the parameters of a stepwise mutation model. *Genetics*. 1998; 150(1):487–497. [PubMed: 9725863]
- Fu YX, Li WH. Statistical tests of neutrality of mutations. *Genetics*. 1993; 133(3):693–709. [PubMed: 8454210]
- Griffiths RC, Tavaré S. Sampling theory for neutral alleles in a varying environment. *Philos Trans R Soc Lond B Biol Sci*. 1994; 344(1310):403–410. [PubMed: 7800710]
- Hey J. The structure of genealogies and the distribution of fixed differences between DNA sequence samples from natural populations. *Genetics*. 1991; 128(4):831–840. [PubMed: 1916247]
- Johns G, Joyce G. The promise and peril of continuous in vitro evolution. *J Mol Evol*. 2005; 61:253–263. [PubMed: 15999246]
- Johnson, NL.; Kotz, S. *Urn Models and Their Application*. New York: Wiley; 1977.
- Kingman JFC. The coalescent. *Stoch Process Appl*. 1982a; 13:235–248.
- Kingman JFC. On the genealogy of large populations. *J Appl Prob*. 1982b; 19A:27–43.
- Kreitman M. Methods to detect selection in populations with applications to the human. *Annu Rev Genomics Hum Genet*. 2000; 1:539–559. [PubMed: 11701640]
- Liu X, Fu Y-X. Test of genetical isochronism for longitudinal samples of DNA sequences. *Genetics*. 2007; 176:327–342. [PubMed: 17277372]
- Polanski A, Bobrowski A, Kimmel M. A note on distributions of times to coalescence, under time-dependent population size. *Theor Popul Biol*. 2003; 63(1):33–40. [PubMed: 12464493]
- Rambaut A. Estimating the rate of molecular evolution: Incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics*. 2000; 16(4):395–399. [PubMed: 10869038]
- Rodrigo, AG.; Felsenstein, J. Coalescent approaches to HIV-1 population genetics. Ch. 8. In: Crandall, KA., editor. *The Evolution of HIV*. Baltimore, Maryland: Johns Hopkins University Press; 1999. p. 233-272.
- Rodrigo AG, Goode M, Forsberg R, Ross HA, Drummond A. Inferring evolutionary rates using serially sampled sequences from several populations. *Mol Biol Evol*. 2003; 20(12):2010–2018. [PubMed: 12949147]
- Seo T-K, Thorne JL, Hasegawa M, Kishino H. Estimation of effective population size of HIV-1 within a host: A pseudomaximum-likelihood approach. *Genetics*. 2002; 160(4):1283–1293. [PubMed: 11973287]
- Tajima F. Evolutionary relationship of DNA-sequences in finite populations. *Genetics*. 1983; 105(2):437–460. [PubMed: 6628982]

- Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*. 1989; 123(3):585–595. [PubMed: 2513255]
- Tavaré S. Line-of-descent and genealogical processes, and their applications in population-genetics models. *Theor Popul Biol*. 1984; 26(2):119–164. [PubMed: 6505980]
- Wakeley J, Hey J. Estimating ancestral population parameters. *Genetics*. 1997; 145(3):847–855. [PubMed: 9055093]
- Watterson GA. Number of segregating sites in genetic models without recombination. *Theor Popul Biol*. 1975; 7(2):256–276. [PubMed: 1145509]
- Willerslev E, Cooper A. Ancient DNA. *Proc Biol Sci*. 2005; 272:3–16. [PubMed: 15875564]
- Wright S. Evolution in mendelian populations. *Genetics*. 1931; 16:97–159. [PubMed: 17246615]

Appendix

A Probability density of coalescent time under two-sample model

Assuming the population size change follows a determinate function

$$N(t) = N(0) \omega(t)$$

where $N(0)$ and $N(t)$ are the population sizes at time 0 and time t respectively. Here time is scaled in $N(0)$ generations.

For sample 1,

$$f(t_m(T)=t) = \begin{cases} f_m^*(t|T) & \text{if } m=n_1 \\ \int_0^\infty f_{n_1,m}^*(s|T) f_m^*(t|s) ds & \text{if } m < n_1 \end{cases} \quad (\text{A.1})$$

where

$$f_m^*(t|s) = \frac{\binom{m}{2}}{\omega(s+t)} \exp\left(-\binom{m}{2} \int_s^{s+t} \frac{d\sigma}{\omega(\sigma)}\right) \quad (\text{A.2})$$

is the conditional probability density function of the m -th coalescent time given at time s there are m ancestral sequences (Griffiths and Tavaré, 1994).

$$f_{n,k}^*(t|s) = \phi(n, k) \sum_{i=k+1}^n \frac{\exp\left(-\binom{i}{2} \int_s^{s+t} \frac{d\sigma}{\omega(\sigma)}\right)}{\omega(s+t) \psi(n, k, i)}$$

is the conditional probability density of convolution of $f_n^*, f_{n-1}^*, \dots, f_{k+1}^*$ given at time s there are n ancestral sequences (Polanski et al., 2003).

For sample 2,

$$f(t_m(0)=t) = \begin{cases} f_m^*(t|0) & \text{if } m=n_1 \\ \int_0^\infty f_{n_2,m}^*(s|0) f_m^*(t|s) ds & \text{if } m < n_1 \end{cases} \quad (\text{A.3})$$

$$f^*(t_{ma}=t|k)=\begin{cases} \frac{f_{n_2,k}^*(T-t|0) P_{k,k}^*(t|T-t)}{P_{n_2,k}^*(T)} & \text{if } 1 \neq m=k \neq n_2 \\ \delta_{(t=T)} & \text{if } m=k=n_2 \\ \frac{f_{n_2,1}^*(T-t|0)}{P_{n_2,1}^*(T)} & \text{if } m=k=1 \\ 0 & \text{if } m < k \\ \frac{\int_0^{T-t} f_{n_2,m}^*(x|0) f_m^*(t|x) P_{m,1}^*(T-t-x|x+t) dx}{P_{n_2,k}^*(T)} & \text{if } m > k \end{cases} \quad (\text{A.4})$$

$$f^*(t_{mb}=t|k)=\begin{cases} f^*(t_m(T)=t) & \text{if } m \leq k \\ 0 & \text{if } m > k \end{cases}, \quad (\text{A.5})$$

where

$$P_{n,k}^*(t|s)=P_{n,k}\left(\int_s^{s+t} \frac{d\sigma}{\lambda(\sigma)}\right) \quad (\text{A.6})$$

is the conditional probability of n sequences coalesce to k sequences in t given there are n sequences at time s (Griffiths and Tavaré, 1994).

B Coalescent time under two-sample-two-stage model

The properties of coalescent time of sample 1 are the same as a single sample, which have been well known assuming the population size is large and a continuous approximation of time (Kingman, 1982a,b). Further, if $N_1 = N_2$ or $\lambda = 1$, the properties of $t_m(0)$ are the same as a single sample. Here we only consider $t_m(0)$ under the situation when $\lambda \neq 1$. Since this section is for the two-sample-two-stage model, we dropped the “ TT ” in the “ E_{TT} ” temporarily.

B.1 Calculate $E(t_m(0))$ and $E(t_m^2(0))$

Given there are k sequences at T , for all $m \leq k$, t_{mb} has the same properties as m -th coalescent time of a single sample. Obviously, we have

$$E(t_{mb}|k)=\begin{cases} \frac{2\lambda}{m(m-1)} & \text{if } m \leq k \\ 0 & \text{if } m > k \end{cases} \quad (\text{B.1})$$

$$E(t_{mb}^2|k)=\begin{cases} 2E^2(t_{mb}|k) & \text{if } m \leq k \\ 0 & \text{if } m > k \end{cases}. \quad (\text{B.2})$$

Considering the following conditions:

$$1 \quad k=1, m=k$$

$$\begin{array}{ll}
2 & k = 1, m \neq k \\
3 & k \neq 1, m = k \\
4 & k \neq 1, m \neq k
\end{array}$$

Condition 1: $k = m = 1$

As shown in Hey (1991),

$$\begin{aligned}
f(t_{ma}=t|k) &= \frac{f_{n_2,1}(T-t)}{P_{n_2,1}(T)} \\
&= \frac{\phi(n_2,1) \sum_{i=2}^{n_2} \frac{\exp\left(-\binom{i}{2}(T-t)\right)}{\psi(n_2,1,i)}}{P_{n_2,1}(T)}
\end{aligned} \tag{B.3}$$

$$E(t_m(0)|k) = E(t_{ma}|k) \tag{B.4}$$

$$\begin{aligned}
&T + \sum_{i=2}^{n_2} \frac{\phi(n_2,1) \exp\left(-\binom{i}{2}T\right)}{\binom{i}{2}^2 \psi(n_2,1,i)} - \frac{1}{\binom{i}{2}} \\
&= \frac{P_{n_2,1}(T)}{P_{n_2,1}(T)}
\end{aligned} \tag{B.5}$$

$$E(t_m^2(0)|k) = E(t_{ma}^2|k) \tag{B.6}$$

$$\begin{aligned}
&\left(T - \sum_{i=2}^{n_2} \frac{1}{\binom{i}{2}}\right)^2 + \sum_{i=2}^{n_2} \frac{1}{\binom{i}{2}^2} - \frac{2\phi(n_2,1) \exp\left(-\binom{i}{2}T\right)}{\binom{i}{2}^3 \psi(n_2,1,i)} \\
&= \frac{P_{n_2,1}(T)}{P_{n_2,1}(T)}.
\end{aligned} \tag{B.7}$$

Condition 2: $k = 1, m > 1$

$$\begin{aligned}
f(t_{ma}=t|k) &= \frac{f_m(t) P_{n_2,1,-m}(T-t)}{P_{n_2,1}(T)} \\
&= \frac{\binom{m}{2} e^{-\binom{m}{2}t} \left(1 - \phi_{-m}(n_2,1) \sum_{i=2, i \neq m}^{n_2} \frac{\exp\left(-\binom{i}{2}(T-t)\right)}{\psi_{-m}(n_2,1,i) \binom{i}{2}}\right)}{P_{n_2,1}(T)}
\end{aligned} \tag{B.8}$$

$$\begin{aligned}
 E(t_m(0)|k) &= E(t_{ma}|k) \\
 &= \frac{g_1(m, T) - \phi(n_2, 1) \sum_{i=2, i \neq m}^{n_2} \frac{h_1(m, i, T)}{\binom{i}{2}^{\psi-m(n_2, 1, i)}}}{P_{n_2, 1}(T)}
 \end{aligned} \tag{B.9}$$

$$\begin{aligned}
 E(t_m^2(0)|k) &= E(t_{ma}^2|k) \\
 &= \frac{g_2(n_2, T) - \phi(n_2, 1) \sum_{i=2, i \neq m}^{n_2} \frac{h_2(m, i, T)}{\binom{i}{2}^{\psi-m(n_2, 1, i)}}}{P_{n_2, 1}(T)}.
 \end{aligned} \tag{B.10}$$

Condition 3: $k \neq 1, m = k$

if $k = m = n_2$

$$f(t_{ma}=t|k) = \delta_{(t=T)} \tag{B.11}$$

$$E(t_{ma}|k) = T \tag{B.12}$$

$$E(t_{ma}^2|k) = T^2 \tag{B.13}$$

otherwise

$$f(t_{ma}=t|k) = \frac{f_{n_2, k}(T-t) P_{k, k}(t)}{P_{n_2, k}(T)} \tag{B.14}$$

$$\begin{aligned}
 &= \frac{\phi(n_2, k) \sum_{i=k+1}^{n_2} \frac{\exp\left(-\left(\binom{i}{2}\right)T - \left(\binom{k}{2} - \binom{i}{2}\right)t\right)}{\psi(n_2, k, i)}}{P_{n_2, k}(T)}
 \end{aligned} \tag{B.15}$$

$$E(t_{ma}|k) = \frac{\phi(n_2, k) \sum_{i=k+1}^{n_2} \frac{h_1(k, i, T)}{\psi(n_2, k, i)}}{P_{n_2, k}(T)} \tag{B.16}$$

$$E(t_{ma}^2|k) = \frac{\phi(n_2, k) \sum_{i=k+1}^{n_2} \frac{h_2(k, i, T)}{\psi(n_2, k, i)}}{P_{n_2, k}(T)}. \tag{B.17}$$

Then

$$E(t_m(0)|k) = E(t_{mb}|k) + E(t_{ma}|k) \quad (\text{B.18})$$

$$E(t_m^2(0)|k) = E(t_{mb}^2|k) + E(t_{ma}^2|k) + 2E(t_{mb}|k)E(t_{ma}|k). \quad (\text{B.19})$$

Condition 4: $k \neq 1, m \neq k$

if $m < k$

$$E(t_m(0)|k) = E(t_{mb}|k) \quad (\text{B.20})$$

$$= \frac{2\lambda}{m(m-1)} \quad (\text{B.21})$$

$$E(t_m^2(0)|k) = E(t_{mb}^2|k) \quad (\text{B.22})$$

$$= 2E^2(t_{mb}|k) \quad (\text{B.23})$$

if $m > k$

$$(t_{ma}=t|k) = \frac{f_m(t)P_{n_2,k,-m}(T-t)}{P_{n_2,k}(T)} \quad (\text{B.24})$$

$$= \frac{\phi(n_2, k) \sum_{i=k, i \neq m}^{n_2} \frac{\exp\left(-\binom{i}{2}T - \left(\binom{m}{2} - \binom{i}{2}\right)t\right)}{\psi_{-m}(m, k-1, i)}}{P_{n_2,k}(T)} \quad (\text{B.25})$$

$$E(t_m(0)|k) = E(t_{ma}|k) \quad (\text{B.26})$$

$$= \frac{\phi(n_2, k) \sum_{i=k, i \neq m}^{n_2} \frac{h_1(n_2, i, T)}{\psi_{-m}(m, k-1, i)}}{P_{n_2,k}(T)} \quad (\text{B.27})$$

$$E(t_m^2(0)|k) = E(t_{ma}^2|k) \quad (\text{B.28})$$

$$= \frac{\phi(n_2, k) \sum_{i=k, i \neq m}^{n_2} \frac{h_2(n_2, i, T)}{\psi_{-m}(m, k-1, i)}}{P_{n_2, k}(T)} \quad (\text{B.29})$$

where

$$f_m(t) = \binom{m}{2} e^{-\binom{m}{2} t} \quad (\text{B.30})$$

is the probability density of coalescent time m ; the details of functions $f_{n,k}(t)$, $f_{n,k,-m}(t)$, $P_{n,k}(T)$, $P_{n,k,-m}(T)$, $\phi(n, k)$, $\psi(n, k, r)$, $\phi_{-m}(n, k)$, $\psi_{-m}(n, k, r)$, $g_1(n, T)$, $g_2(n, T)$, $h_1(n, i, T)$ and $h_2(n, i, T)$ can be found in Appendix C.

All together,

$$E(t_m(0)) = \sum_{k=1}^{n_2} P_{n_2, k}(T) E(t_m(0)|k) \quad (\text{B.31})$$

$$E(t_m^2(0)) = \sum_{k=1}^{n_2} P_{n_2, k}(T) E(t_m^2(0)|k). \quad (\text{B.32})$$

B.2 Calculate $E(t_m(0) t_{m'}(0))$

If $\lambda \neq 1$, the coalescent time may not be independent. Given there are k sequences at T , $t_m(0)$ and $t_{m'}(0)$ are dependent only if $m, m' \geq k$. Here we show how to compute $E(t_{ma} t_{m'a}|k)$ assuming $m > m' \geq k$ without loss of generality.

We consider the following conditions:

-
1	k = 1, k ≠ n ₂ - 1, m' = k
2	k = 1, k ≠ n ₂ - 1, m' > k
3	k = n ₂ - 1, m' = k, m = n ₂
4	1 < k < n ₂ - 1, m' = k
5	1 < k < n ₂ - 1, m' > k
-

.

Condition 1: $k = 1, k \neq n_2 - 1, m' = k$

$$f(t_{ma}=t, t_{m'a}=t'|k) = \frac{f_m(t) f_{n_2,1,-m}(T-t-t')}{P_{n_2,1}(T)} \quad (\text{B.33})$$

$$E(t_{ma}t_{m'a}|k) = \frac{\phi(n_2, 1) \sum_{i=2, i \neq m}^{n_2} \frac{w_1(m, i, T)}{\psi_{-m}(n_2, 1, i)}}{P_{n_2,1}(T)}. \quad (\text{B.34})$$

Condition 2: $k = 1, k \neq n_2 - 1, m' > k$

$$f(t_{ma}=t, t_{m'a}=t'|k) = \frac{f_m(t) f_{m'}(t') P_{n_2,1,-m,-m'}(T-t-t')}{P_{n_2,1}(T)} \quad (\text{B.35})$$

$$E(t_{ma}t_{m'a}|k) = \frac{w_1(m, m', T)}{P_{n_2,1}(T)} - \frac{\phi(n_2, 1) \sum_{i=2, i \neq m, i \neq m'}^{n_2} \frac{w_2(m, m', i, T)}{\psi_{-m,-m'}(n_2, 1, i) \binom{i}{2}}}{P_{n_2,1}(T)}. \quad (\text{B.36})$$

Condition 3: $k = n_2 - 1, m' = n_2 - 1, m = n_2$

$$\begin{aligned} E(t_{ma}t_{m'a}|k) &= E(t_{n_2}(0)(T-t_{n_2}(0))) \\ &= T \cdot E(t_{n_2}(0)|n_2-1) - E(t_{n_2}^2(0)|n_2-1). \end{aligned} \quad (\text{B.37})$$

Condition 4: $n_2 - 1 > k > 1, m' = k$

$$f(t_{ma}=t, t_{m'a}=t'|k) = \frac{f_m(t) f_{n_2,k,-m}(T-t-t') P_{k,k}(t')}{P_{n_2,k}(T)} \quad (\text{B.38})$$

$$E(t_{ma}t_{m'a}|k) = \frac{\phi(n_2, k) \sum_{i=k+1, i \neq m}^{n_2} \frac{w_2(m, m', i, T)}{\psi_{-m}(n_2, k, i)}}{P_{n_2,k}(T)}. \quad (\text{B.39})$$

Condition 5: $n_2 - 1 > k > 1, m' \neq k$

$$f(t_{ma}=t, t_{m'a}=t'|k) = \frac{f_m(t) f_{m'}(t') P_{n_2,k,-m,-m'}(T-t-t')}{P_{n_2,k}(T)} \quad (\text{B.40})$$

$$E(t_{ma}t_{m'a}|k) = \frac{\phi(n_2, k) \sum_{i=k, i \neq m, i \neq m'}^{n_2} \frac{w_2(m, m', i, T)}{\psi_{-m,-m'}(n_2, k-1, i)}}{P_{n_2,k}(T)} \quad (\text{B.41})$$

where details of functions $f_{n,k,-m,-m'}(T)$, $P_{n,k,-m,-m'}(T)$, $\phi_{-m,-m'}(n, k)$, $\psi_{-m,-m'}(n, k, r)$, $w_1(m, i, T)$ and $w_2(m, m', i, T)$ can be found in Appendix C.

In general, given $m > m'$

$$E(t_m(0) t_{m'}(0)|k) = \begin{cases} E(t_{ma} t_{m'a}|k) + E(t_{m'b}|k) E(t_{ma}|k) & \text{if } m'=k \\ [E(t_{ma}|k) + E(t_{mb}|k)] E(t_{m'b}|k) & \text{if } m=k \\ E(t_{ma} t_{m'a}|k) & \text{if } m' > k \\ E(t_{mb}|k) E(t_{m'b}|k) & \text{if } m < k \\ E(t_{ma}|k) E(t_{m'b}|k) & \text{otherwise} \end{cases} \quad (\text{B.42})$$

All together,

$$E(t_m(0) t_{m'}(0)) = \sum_{k=1}^{n_2} P_{n_2,k}(T) E(t_m(0) t_{m'}(0)|k). \quad (\text{B.43})$$

C Functions

$$P(k, i|n) = \begin{cases} \frac{\binom{n-i-1}{k-2}}{\binom{n-1}{k-1}} & \text{if } k \geq 2, 1 \leq i < n \\ 0 & \text{otherwise} \end{cases} \quad (\text{C.1})$$

is the probability that a randomly chosen line of state k is of size i at state n ;

$$P(k' \rightarrow j|n, k) = \frac{\binom{j-1}{k'-1} \binom{n-j-1}{k-k'-1}}{\binom{n-1}{k-1}} \quad (\text{C.2})$$

is the probability that k' lines in state k grow to j lines in state n ;

$$P_{n,k}(T) = \begin{cases} \phi(n, 1) \sum_{i=2}^n \frac{1 - \exp\left(-\binom{i}{2} T\right)}{\binom{i}{2} \psi(n, 1, i)} & \text{if } k=1 \\ \phi(n, k) \sum_{i=k}^n \frac{\exp\left(-\binom{i}{2} T\right)}{\psi(n, k-1, i)} & \text{if } 2 \leq k < n \\ \exp\left(-\binom{n}{2} T\right) & \text{if } k=n \end{cases} \quad (\text{C.3})$$

is the probability of n sequences coalesce to k sequences in T ;

$$\phi(n, k) = \begin{cases} 1 & \text{if } n=k \\ \prod_{s=k+1}^n \binom{s}{2} & \text{otherwise} \end{cases}; \quad (\text{C.4})$$

$$\psi(n, k, r) = \begin{cases} 1 & \text{if } r=n, k=n-1 \\ \prod_{s=k+1, s \neq r}^n \left(\binom{s}{2} - \binom{r}{2} \right) & \text{otherwise} \end{cases}; \quad (\text{C.5})$$

$$a_n = \begin{cases} 1 + \frac{1}{2} + \cdots + \frac{1}{n-1} & \text{if } n > 1 \\ 0 & \text{otherwise} \end{cases}; \quad (\text{C.6})$$

$$\beta_n(i) = \begin{cases} \frac{2n}{(n-i+1)(n-i)}(a_{n+1} - a_i) - \frac{2}{n-i} & \text{if } i < n \\ 0 & \text{otherwise} \end{cases}; \quad (\text{C.7})$$

$$\gamma_n(i) = \beta_n(i) - \beta_n(i+1); \quad (\text{C.8})$$

$$P(j' \rightarrow j, m' \rightarrow m | n, n') = \begin{cases} 1 & \text{if } j'=j=m'=m=0 \\ P(m' \rightarrow m | n, n') & \text{if } j'=j=0, 0 < m' \leq m \\ P(j' \rightarrow j | n, n') & \text{if } m'=m=0, 0 < j' \leq j \\ P(j' \rightarrow j | n, n') & \text{if } j'+m'=n', j+m=n, \\ & 0 < j' \leq j, 0 < m' \leq m \\ \frac{\binom{j-1}{j'-1} \binom{m-1}{m'-1} \binom{n-j-m-1}{n'-j'-m'-1}}{\binom{n-1}{n'-1}} & \text{if } j'+m' < n', j+m < n, \\ & 0 < j' \leq j, 0 < m' \leq m, \\ & n'-j'-m' \leq n-j-m \\ 0 & \text{otherwise} \end{cases} \quad (\text{C.9})$$

is the joint probability of mutations of size j' at state n' which grow to size j at state n and mutations of size m' at state n' which grow to size m at state n (Johnson and Kotz, 1977);

$$P(k, i; k, j | n) = \begin{cases} \frac{\binom{n-i-j-1}{k-3}}{\binom{n-1}{k-1}} & \text{if } i+j < n, k > 2, i \geq 1, j \geq 1 \\ \frac{1}{n-1} & \text{if } i+j=n, k=2, i \geq 1, j \geq 1 \\ 0 & \text{otherwise} \end{cases} \quad (\text{C.10})$$

is the probability that two randomly chosen lines of state k are of size i and j at state n ;

$$P(k, i; k', j|n) = \delta_{(i \geq j)} P_a(k, i; k', j|n) + \delta_{(i+j \leq n)} P_b(k, i; k', j|n)$$

is the probability that a line of state k and a line of state k' are of size i and j at state n respectively (Fu, 1995), where

$$P_a(k, i; k', j|n) = \begin{cases} \frac{\sum_{t=2}^{\min(k'-k+1, i-j+1)} \frac{\binom{k'-k}{t-1}}{\binom{k'-1}{t}} \frac{k-1}{k'} \frac{\binom{i-j-1}{t-2} \binom{n-i-1}{k'-t-1}}{\binom{n-1}{k'-1}}}{\binom{n-i-1}{k'-2}} & \text{if } i > j \\ \frac{k-1}{(k'-1)k'} \frac{\binom{n-i-1}{k'-2}}{\binom{n-1}{k'-1}} & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \quad (\text{C.11})$$

is the probability for the case that the line of state k' is a descendant of the line of state k and

$$P_b(k, i; k', j|n) = \begin{cases} \frac{\sum_{t=1}^{\min(k'-2, k'-k+1, i)} \frac{\binom{k'-k}{t-1}}{\binom{k'-1}{t}} \frac{(k-1)(k'-t)}{tk'} \frac{\binom{i-1}{t-1} \binom{n-i-j-1}{k'-t-2}}{\binom{n-1}{k'-1}}}{\frac{1}{jk'} \frac{\binom{n-k'}{j-1}}{\binom{n-1}{j}}} & \text{if } i+j < n, k \geq 2 \\ \frac{1}{jk'} \frac{\binom{n-k'}{j-1}}{\binom{n-1}{j}} & \text{if } i+j = n, k = 2 \\ 0 & \text{otherwise} \end{cases} \quad (\text{C.12})$$

is the probability for the case that the line of state k' is not a descendant of the line of state k ;

$$P_{n,k,-m}(T) = \begin{cases} \phi_{-m}(n, k) \sum_{i=k, i \neq m}^n \frac{\exp\left(-\binom{i}{2} T\right)}{\psi_{-m}(n, k-1, i)} & \text{if } 2 \leq k < n \\ \phi_{-m}(n, 1) \sum_{i=2, i \neq m}^n \frac{1 - \exp\left(-\binom{i}{2} T\right)}{\binom{i}{2} \psi_{-m}(n, 1, i)} & \text{if } k = 1 \end{cases} ; \quad (\text{C.13})$$

$$\phi_{-m}(n, k) = \begin{cases} 1 & \text{if } n = k \\ \prod_{s=k+1, s \neq m}^n \binom{s}{2} & \text{otherwise} \end{cases} ; \quad (\text{C.14})$$

$$\psi_{-m}(n, k, r) = \begin{cases} 1 & \text{if } r=n, k=n-1 \\ \prod_{s=k+1, s \neq r, s \neq m}^n \left(\binom{s}{2} - \binom{r}{2} \right) & \text{otherwise} \end{cases}; \quad (\text{C.15})$$

$$g_1(n, T) = -Te^{-\binom{n}{2}T} - \frac{e^{-\binom{n}{2}T} - 1}{\binom{n}{2}}; \quad (\text{C.16})$$

$$g_2(n, T) = -T^2e^{-\binom{n}{2}T} + \frac{2}{\binom{n}{2}} \left[-Te^{-\binom{n}{2}T} - \frac{e^{-\binom{n}{2}T} - 1}{\binom{n}{2}} \right]; \quad (\text{C.17})$$

$$h_1(n, i, T) = -\frac{Te^{-\binom{n}{2}T}}{\binom{n}{2} - \binom{i}{2}} - \frac{e^{-\binom{n}{2}T} - e^{-\binom{i}{2}T}}{\left(\binom{n}{2} - \binom{i}{2} \right)^2}; \quad (\text{C.18})$$

$$h_2(n, i, T) = -\frac{T^2e^{-\binom{n}{2}T}}{\binom{n}{2} - \binom{i}{2}} - \frac{2Te^{-\binom{n}{2}T}}{\left(\binom{n}{2} - \binom{i}{2} \right)^2} - \frac{2 \left(e^{-\binom{n}{2}T} - e^{-\binom{i}{2}T} \right)}{\left(\binom{n}{2} - \binom{i}{2} \right)^3}; \quad (\text{C.19})$$

$$f_{n,k}(t) = \phi(n, k) \sum_{i=k+1}^n \frac{\exp\left(-\binom{i}{2}t\right)}{\psi(n, k, i)} \quad (\text{C.20})$$

is the probability density of the convolution of $f_n, f_{n-1}, \dots, f_{k+1}$;

$$f_{n,k,-m}(t) = \phi_{-m}(n, k) \sum_{i=k+1, i \neq m}^n \frac{\exp\left(-\binom{i}{2}t\right)}{\psi_{-m}(n, k, i)} \quad (\text{C.21})$$

is the probability density of the convolution of $f_n, f_{n-1}, \dots, f_{m+1}, f_{m-1}, \dots, f_{k+1}$;

$$f_{n,k,-m,-m'}(T) = \phi_{-m,-m'}(n,k) \sum_{i=k+1, i \neq m, i \neq m'}^n \frac{\exp\left(-\binom{i}{2}T\right)}{\psi_{-m,-m'}(n,k,i)} \quad (\text{C.22})$$

is the probability density of the convolution of $f_n, f_{n-1}, \dots, f_{m+1}, f_{m-1}, \dots, f_{m'+1}, f_{m'-1}, \dots, f_{k+1}$;

$$P_{n,k,-m,-m'}(T) = \begin{cases} \phi_{-m,-m'}(n,k) \sum_{i=k, i \neq m, i \neq m'}^n \frac{\exp\left(-\binom{i}{2}T\right)}{\psi_{-m,-m'}(n,k-1,i)} & \text{if } 2 \leq k \\ \phi_{-m,-m'}(n,1) \sum_{i=2, i \neq m, i \neq m'}^n \frac{1 - \exp\left(-\binom{i}{2}T\right)}{\binom{i}{2} \psi_{-m,-m'}(n,1,i)} & \text{if } k=1 \end{cases} ; \quad (\text{C.23})$$

$$\phi_{-m,-m'}(n,k) = \begin{cases} 1 & \text{if } n=k \\ \prod_{s=k+1, s \neq m, s \neq m'}^n \binom{s}{2} & \text{otherwise} \end{cases} ; \quad (\text{C.24})$$

$$\psi_{-m,-m'}(n,k,r) = \begin{cases} 1 & \text{if } r=n, k=n-1 \\ \prod_{s=k+1, s \neq r, s \neq m, s \neq m'}^n \left(\binom{s}{2} - \binom{r}{2} \right) & \text{otherwise} \end{cases} ; \quad (\text{C.25})$$

$$w_1(m,i,T) = \left(T - \frac{1}{\binom{i}{2}} \right) \frac{g_1(m,T)}{\binom{i}{2} \binom{m}{2}} - \frac{g_2(m,T)}{\binom{i}{2} \binom{m}{2}} + \frac{h_1(m,i,T)}{\binom{i}{2}^2}; \quad (\text{C.26})$$

$$w_2(m,m',i,T) = - \left(T + \frac{1}{\left[\binom{m'}{2} - \binom{i}{2} \right]} \right) \frac{h_1(m,m',T)}{\left[\binom{m'}{2} - \binom{i}{2} \right]} + \frac{h_2(m,m',T)}{\left[\binom{m'}{2} - \binom{i}{2} \right]} + \frac{h_1(m,i,T)}{\left[\binom{m'}{2} - \binom{i}{2} \right]^2}. \quad (\text{C.27})$$

Figure 1.
A genealogy of five sequences.

Figure 3.
two-sample-two-stage model.