# Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates

**JOSEPH FELSENSTEIN**

*Department of Genetics SK-50, University of Washington, Seattle, Washington 98195*

RUNNING TITLE: Effective population size from sequence samples

**Summary**

It is known that under neutral mutation at a known mutation rate a sample of nucleotide sequences, within which there is assumed to be no recombination, allows estimation of the effective size of an isolated population. This paper investigates the case of very long sequences, where each pair of sequences allows a precise estimate of the divergence time of those two gene copies. The average divergence time of all pairs of copies estimates twice the effective population number, and an estimate can also be derived from the number of segregating sites. One can alternatively estimate the genealogy of the copies. This paper shows how a maximum likelihood estimate of the effective population number can be derived from such a genealogical tree. The pairwise and the segregating sites estimates are shown to be much less efficient than this maximum likelihood estimate, and this is verified by computer simulation. The result implies that there is much to gain by explicitly taking the tree structure of these genealogies into account.

## 1. Introduction

The famous paper of Cann, Stoneking, & Wilson (1987) has focused attention on the potential of sequence samples from populations to illuminate population parameters such as effective population sizes and migration rates.

We can observe the numbers of substitutions by which sequences differ. Under the neutral mutation model these differences are expected to accumulate at a rate of $\mu$ per site per generation. We can estimate how long ago, in terms of mutational events, the sequences diverged. Under genetic drift, the actual divergence times of the sequences are related to the effective population size $N_e$. If $\mu$ is known, we can convert the mutational scale into a time scale and estimate $N_e$. If $\mu$ is not known, the best we can do is to estimate $N_e\mu$. In this paper I will discuss the problem in terms of the estimation of $4N_e\mu$. This is equivalent to estimation of $N_e$ if $\mu$ is known.

Nei and Tajima (1981) have suggested the use of the average number of differences per site between two sequences, which they call the *nucleotide diversity*, for estimation of $4N_e\mu$. Tajima (1983) and Nei (1987) give a formula for the variance of the estimate. A slightly different approach is used by Avise, Ball, and Arnold (1988; see also Avise 1989 and Ball, Neigel, and Avise 1990). They take pairs of sequences and make an estimate of divergence time from each. They avoid using all pairs of sequences in order to make the individual estimates more independent. In an isolated randomly mating population, we expect the divergence time for a randomly chosen pair of gene copies to be exponentially distributed with mean $2N_e$. They fit the observed distribution of pairwise estimates to an exponential distribution in order to estimate this quantity.

Watterson (1975) has presented results on the number of segregating sites at a locus under the neutral "infinite sites" model, which can also be used as the basis for an estimate the $N_e$ or $4N_e\mu$. It is important to realize when reading the literature on infinite sites models that the $\mu$ which is described there is the mutation rate *per locus*; throughout this paper it will be the mutation rate *per site*.

Neither of these estimates makes the most efficient use of such data. In this paper I will discuss maximum likelihood estimation, which I will show is con-

siderably more efficient. Its efficiency is demonstrated both theoretically and by computer simulation. The present paper discusses only the extreme case of an infinitely long nucleotide sequence; the more practical matter of dealing efficiently with sequences of finite length requires computationally intensive techniques that will be covered elsewhere. For the moment the objective is simply to show the weakness of the pairwise and segregating sites approaches.

## 2. A maximum likelihood method

In hopes that it will make efficient use of the data, let us make a maximum likelihood estimate of $4N_e\mu$ in the case of long sequences. We assume that the sequences allow us to estimate their genealogy without error, and that there is a single such genealogy, i.e., no recombination has occurred within the sequences during the relevant period of time. The genealogy is assumed to be produced by Kingman's "coalescent" process: we assume that to be a good enough approximation to the genealogy produced by random genetic drift in a finite population. This will be true if $N_e$ is not small.

The results of Kingman (1982a, b) on the coalescent and those of Harding (1971) on random trees establish that under the coalescent the prior distribution on the genealogy assigns equal probability to all possible bifurcating trees with interior nodes ordered in time. The $n$ tips are assumed to be contemporaneous. Each interior node has a time, and if two trees differ in the order of these times they are considered to be different. Kingman's coalescent also places a prior on the times. Starting from the $n$ tips, which occur at time 0 (the present) the time back to the most recent coalescent event is exponentially distributed with expectation $4N_e/[n(n-1)]$. This is an approximation, which is excellent for large $N_e$ when $n \ll N_e$, as is usually the case.

Tavaré (1984) has reviewed the logic of this approximation. Strictly speaking, it requires that as we take larger and larger values of $N_e$ we observe the process on a time scale whose units are $N_e$ generations. If we scale time in expected mutations per site (as we would if we did not know $\mu$), the mean of the

scaled time $u_n$ would be $4N_e\mu/[n(n-1)]$. We will in effect invoke the diffusion approximation, by assuming that $N_e \rightarrow \infty$ and $\mu \rightarrow 0$ in such a way that their product remains constant. Thus $4N_e\mu$ will equal some constant $\theta$, and we are approximating the genealogy of the actual population which has finite values of $N_e$ and $\mu$ by Kingman's coalescent process.

Prior to the most recent coalescence, uhere were $n-1$ tips, and the interval of scaled time $u_{n-1}$ back to the previous coalescent event is independently exponentially distributed with mean $4N_e\mu/[(n-1)(n-2)]$. In general, if $\theta = 4N_e\mu$ then, scaling time in expected mutations per site,

$$u_k \sim Exponential\ (\theta/[k(k-1)]) \tag{1}$$

with $k = n, n-1, ...2$.

The topology of the genealogical tree has no information about $\theta$. We have already seen that all topologies with time-ordered nodes are equiprobable, so that their distribution does not depend on $\theta$. All information about $\theta$ is contained in the scaled coalescence times and the intervals $u_k$ between them.

Assume that we have collected a sample of $n$ long sequences from a random-mating population whose sequences are diverging under neutral mutation. The sequences are sufficiently long that we can infer precisely the genealogical tree connecting those sequences, and from it the scaled time intervals $u_k$. Thus we will consider the $u_k$ to have been observed.

The $k$-th of these has the exponential density function

$$f_k(u) = \frac{k(k-1)}{\theta} \exp[-\frac{k(k-1)}{\theta}u] \tag{2}$$

so that the full set of $u_k$'s has a joint density function, the likelihood

$$\begin{aligned} L &= \prod_{k=2}^{n} f_k(u_k) \\ &= \prod_{k=2}^{n} \tfrac{k(k-1)}{\theta} \exp[-\tfrac{k(k-1)}{\theta}u_k] \end{aligned} \tag{3}$$

Taking logarithms, the log-likelihood is

$$\ln L = \sum_{k=2}^{n} \ln k + \sum_{k=2}^{n} \ln(k-1) - (n-1)\ln\theta - \frac{1}{\theta}\sum_{k=2}^{n} k(k-1)u_k. \qquad (4)$$

To find the maximum likelihood estimate for $\theta$ we differentiate with respect to it. The first two terms, which are logarithms of factorials, do not contain $\theta$ and disappear, so that

$$\frac{\partial \ln L}{\partial \theta} = -\frac{n-1}{\theta} + \frac{1}{\theta^2}\sum_{k=2}^{n} k(k-1)u_k. \qquad (5)$$

Equating this to zero and solving for $\theta$, we get as the maximum likelihood estimate

$$\hat{\theta} = \frac{\sum_{k=2}^{n} k(k-1)u_k}{n-1}. \qquad (6)$$

This is a simple average of the $k(k-1)u_k$, whose variance is easily obtained. Since the variance of the exponential variate $u_k$ is the square of its mean,

$$\begin{aligned} Var(\hat{\theta}) &= \tfrac{1}{(n-1)^2}\sum_{k=2}^{n} k^2(k-1)^2 Var(u_k) \\ &= \tfrac{1}{(n-1)^2}\sum_{k=2}^{n} k^2(k-1)^2(\theta/[k(k-1)])^2. \qquad (7) \\ &= \tfrac{\theta^2}{n-1} \end{aligned}$$

so that the squared coefficient of variation of $\theta$ is simply

$$\frac{Var(\hat{\theta})}{\theta^2} = \frac{1}{n-1}. \qquad (8)$$

Although the estimate in equation (6) is the maximum likelihood estimate, it could also be derived by many other methods. It is the minimum variance

unbiased estimate, the method of moments estimate, and the weighted least squares estimate as well.

Note that the quantities $k(k-1)u_k$ are independent and all exponentially distributed with the same expectation $\theta$. This suggests that it would be straightforward, given the $u_k$, to construct various goodness-of-fit tests that could detect whether there is a trend in the $u_k$, a trend that would indicate that the effective population sizes had changed through time.

### 3. The pairwise method

An attractive alternative to maximum likelihood would be to use pairs of sequences to estimate divergence time, and to average these estimates over all pairs of tips. Any two randomly chosen sequences have a time of divergence which is exponentially distributed with mean $2N_e$, so that if the divergence time is stated in mutations per site it has mean $2N_e\mu$ which is $\theta/2$. If we have long sequences, as we assume here, we can estimate $\theta$ by taking the mean of all these pairwise divergence times and then estimating $\theta$ by doubling that. Since the estimate is a mean of random variables, each of which has expectation $\theta$, it obviously makes an unbiased estimate of $\theta$. This method is analogous to the mean codon difference method of Nei and Tajima (1981) but is not identical to it: theirs is a pairwise method using mean codon difference, whereas the present method makes pairwise estimates of divergence time and then averages them. Pairwise methods are attractive because they do not involve estimating the tree topology and have an aura of robustness.

The aura is, I hope to show, misleading. To show this, we must compute the variance of the estimate. Each pair of sequences has a most recent common ancestor who occurred at the time of one of the coalescences. If $t_k$ is the time (scaled in mutations per site) from the present back until the coalescent event that reduced $k$ lineages to $k-1$, then by our earlier definition of the $u_k$,

$$t_k = u_n + u_{n-1} + u_{n-2} + ... + u_k = \sum_{i=k}^{n} u_i \tag{9}$$

Figure 1 shows the relationship between the $t_k$ and the $u_k$. Suppose that we define $m_k$ to be the number of pairs of sequences that have as their most recent common ancestor the coalescence that occurs when $k$ lineages are reduced to $k-1$. Since every one of the $n(n-1)/2$ pairs of sequences has one or another coalescence as their most recent common ancestor, it must be true that

$$\sum_{i=2}^{n} m_i = n(n-1)/2, \tag{10}$$

and we can express the pairwise estimate of $\theta$ in terms of the $m_k$ as

$$\hat{\theta}_P = \frac{4 \sum_{i=2}^{n} m_i t_i}{n(n-1)}. \tag{11}$$

The $t_i$ are random variables, but are not independent. We can use (9) to express them in terms of the $u_i$, which are independent, obtaining

$$\hat{\theta}_P = \frac{4 \sum_{i=2}^{n} m_i \sum_{j=i}^{n} u_j}{n(n-1)} \tag{12}$$

which on rearranging summations and changing their limits becomes

$$\hat{\theta}_P = \frac{4 \sum_{i=2}^{n} u_i \sum_{j=2}^{i} m_j}{n(n-1)}. \tag{13}$$

This is a weighted sum of the $u_i$ but not necessarily a weighted average. Let

$$C_i = \sum_{j=2}^{i} m_j \,. \tag{14}$$

Substituting this into (13),

$$\hat{\theta}_P = \frac{4 \sum_{i=2}^{n} u_i C_i}{n(n-1)}. \tag{15}$$

For each tree topology with ordered internal nodes, we can calculate the $m_i$, and from those using (14) the $C_i$. For that tree topology, we can use (15) to compute the expectation and variance of $\hat{\theta}_P$, using the fact that the $u_i$ are independently exponentially distributed according to (1). The expectation and variance given the $C_i$ are

$$E[\hat{\theta}_P \mid \mathbf{C}] = \frac{4}{n(n-1)} \sum_{k=2}^{n} \frac{\theta}{k(k-1)} C_k \tag{16}$$

and

$$Var[\hat{\theta}_P \mid \mathbf{C}] = \frac{16}{n^2(n-1)^2} \sum_{k=2}^{n} \frac{\theta^2}{k^2(k-1)^2} C_k^2. \tag{17}$$

The expectation (16) will not be the same from one ordered tree topology to another. The mean of these means will be $\theta$, but each ordered tree topology will have a slightly different mean. For each ordered tree topology the estimate is biased, but the mean bias is zero.

To complete the calculation from this formula of the variance of the pairwise estimate of $\theta$, we would need to sum over all ordered tree topologies, obtaining the $C_i$ for each, using formulas (16) and (17), and adding the mean of (17) to the variance of the means (16). Alternatively we would need a theory of the $C_i$ so that the summation over ordered tree topologies would not be necessary.

However, development of such a theory is not necessary, as Tajima (1983) has developed expressions for the mean and variance of the mean number of nucleotide differences between pairs of sequences in a sample from a single population without recombination. We use the modification of Tajima's expressions in equations (10.9) and (10.10) of Nei (1987) which takes the number of sites sampled into account. As the number of sites (Nei's $m_T$) becomes infinite we have, in our notation,

$$Var(\hat{\theta}_P) = \frac{2(n^2+n+3)\theta^2}{9n(n-1)}. \tag{18}$$

So that from (7) we can compute as the efficiency of the pairwise method

$$\frac{Var(\hat{\theta})}{Var(\hat{\theta}_P)} = \frac{9n}{2(n^2+n+3)}. \tag{19}$$

Variances of the maximum likelihood and pairwise estimators and the efficiency of the pairwise estimator, computed from (7), (18), and (19) are presented in Table 1. These are also shown as solid curves in Figures 2 and 3. It will immediately be apparent that the efficiency of the pairwise method rapidly becomes small, falling below 0.22 at 20 sequences, 0.11 at 40 sequences, and 0.045 at 100 sequences. The variances are computed for $\theta = 4$ but as they are proportional to $\theta^2$ they can be directly computed from this table for any value of $\theta$ by appropriately multiplying these values.

Note that the variance of the pairwise estimate does not fall to zero, approaching instead $2\theta^2/9$ as $n \to \infty$. The reason for this behavior is that the pairwise estimate takes most of its information from the times of the earliest few coalescences. It can be shown that of all pairs of species, a fraction $(n+1)/(3n-3)$ of them are expected to be separated by the bottom fork of the genealogical tree. This fraction is always greater than 1/3. This means that over 1/3 of all the information in the pairwise estimate comes from the time of this one fork!

That this is so is the consequence of a remarkable fact. If we consider the two lineages that result from this earliest fork, and wait until a total of $n$ lineages exist, the distribution of the number of descendants of the left lineage is uniform on $1, 2, ..., n-1$. This follows imediately from Theorem 1 of Harding (1971). It can also be obtained by realising that the process of splitting of the left and right lineages is modelled by Polya's Urn Model. Let us represent the two original lineages by balls of different colors. Splitting a random lineage corresponds to choosing a random ball, and adding to the urn another of that color. Expression (2.3) in Feller (1968) then establishes that when we reach $n$ balls the fraction that are of a given color is uniformly distributed. From this uniform

distribution the expected fraction of pairs of balls that are of different colors is easily calculated as being $(n+1)/(3n-3)$. Maddison and Slatkin (1991) and Slowinski and Guyer (1989) have also used this result in their work on random trees. A reviewer has pointed out to me that it also can be obtained directly from formula (2.3) of Saunders, Tavaré, and Watterson (1984) by considering the case $i = n, j = 2$, and $l_1 = l_2 = 2$. Their formula is then calculating the probability that among a population of $N$ organisms reproducing according to the asexual or haploid version of Moran (1958), when sample of size $n$ has been traced back to two ancestors, a subsample of a pair of organisms will still have two distinct parents. This result is interesting as in this case the formula holds even when the population size is finite.

In contrast to the pairwise estimate, the maximum likelihood estimate uses information from all coalescence events. As $n$ increases, it has more and more coalescences to work from and hence the estimate becomes more and more accurate.

Ball, Neigel, and Avise (1990) have presented simulation results for their technique of averaging divergence times for a subset of all pairs of sequences, chosen so that each sequence is only used once. Their results for an average divergence time of 50 pairs of sequences drawn from a simulated population of 100 sequences shows the expected lack of bias of the estimator, as well as substantial departures from independence of the 50 quantities, as expected from the argument given here.

### 4. Watterson's method

Watterson (1975) obtained the distributions of the number of segregating sites for a sample of $n$ sequences from a random-mating population under an infinite-sites model of mutation. The infinite-sites model is the limit of the present model as the mutation rate becomes small, and the number of sites large. Although Watterson does not discuss the estimation of $\theta$ directly, Nei (1987, p. 255) pointed out that an estimate can be based on the expectation

Watterson computed for the number of segregating sites in the sample, that is, the number of sites at which there is more than one base present. Watterson's derivation uses the assumption that there are finitely-many sites, but assumes that the mutation rate per site $\mu$ is allowed to get small and the population size large at the same rate, so that their product $N\mu$ remains constant. Watterson shows that if $K_n$ is the number of sites showing genetic variation among a sample of $n$ randomly chosen copies, that the expectation and variance of $K_n$ are to good approximation if $n \ll N$:

$$E(K_n) = \theta \sum_{i=1}^{n-1} \frac{1}{i} \tag{20}$$

and

$$Var(K_n) = E(K_n) + \theta^2 \sum_{i=1}^{n-1} \frac{1}{i^2} \tag{21}$$

We will in addition assume that there are a very large number of sites in the gene, so that although finite, $\theta$ is large. In this case, the term $E(K_n)$ makes an unimportant contribution to the right-hand-side of (21). An unbiased estimator of $\theta$ is (Nei, 1987, p.255), from (20),

$$\hat{\theta} = K_n \Big/ \sum_{i=1}^{n-1} \frac{1}{i} \tag{22}$$

and from (21) we can work out the variance of this estimate and compute its coefficient of variation to be

$$\frac{Var(\hat{\theta})}{\theta^2} = \frac{\sum_{i=1}^{n-1} 1/i^2}{\left(\sum_{i=1}^{n-1} 1/i\right)^2} \tag{23}$$

The estimate is unbiased, and as $n \to \infty$ its coefficient of variation decreases to zero, so that it makes a consistent estimate.

If the number of sites is not so large, then (23) is increased by the amount $1/\theta \sum_i (1/i)$, so that (23) is in effect a lower bound on the coefficient of variation of Watterson's estimate. We are interested in the cases where the sequences are very long and hence the number of segregating sites is large. Thus we are investigating Watterson's method in the cases most favorable to it.

The efficiency of Watterson's estimate must be, taking the ratio of (8) and (23), no more than

$$\frac{\left(\sum_{i=1}^{n-1} 1/i\right)^2}{(n-1)\sum_{i=1}^{n-1} 1/i^2} \tag{24}$$

The variance (from equation 21) and the efficiency (from 24) of Watterson's estimate are shown in the rightmost two columns of Table 1. Actually the variance shown is the lower bound, using only the second term of the right-hand side of (21). This is asymptotically valid, as mentioned above, for large values of $\theta$. The value shown is the bound for $\theta = 4$: as with the other variances in the table, the value will be proportional to $\theta^2$ and this can be used to compute this lower bound for all values of $\theta$, and hence compute the variance approximately for all large values of $\theta$.

The variance does decrease to zero with increasing $n$, but not as quickly as does the maximum likelihood estimator. Efficiency drops with larger $n$, and although not as low as that for the pairwise estimator, it is 0.42 for 20 sequences, 0.29 for 40, and 0.17 for 100 sequences.

### 5. Simulation results

One might well wonder whether the formula (18) can be applied to the pairwise estimation method as defined here. Tajima (1983) and Nei (1987) derived it as the variance of the average pairwise codon difference between sequences. This is not the same as the average scaled divergence time separating the sequences, which is the quantity of interest here. However for an infinitely long sequence, the Tajima-Nei variance formula is proportional to $\theta^2$. When sequences are infinitely long and $\theta$ is small the divergence time will be proportional to the codon difference between sequences. Equation (18) will be correct in that limit. For a set of sequences the joint distribution of estimated divergence times will be the same as it is when $\theta$ is small, but scaled proportional to $\theta$. Equation (18) should then continue to apply to the mean of the scaled divergence time estimates.

This argument is sufficiently indirect that it is helpful to check it by computer

simulation. A large computer simulation has been used both to compute the expected power of the pairwise method, and to check that the variances are correct. Two programs were written in MIPS Pascal on a Digital DECstation 3100 and a DECstation 5000. The first program is given the number of sequences to be sampled, and the number of replicates, as well as a random number seed. It simulates the coalescent, starting with a number of lineages equal to the desired number of sequences, drawing pairs of lineages and the sampling the times of their immediate common ancestor from the distribution (1). This technique of simulating the coalescent by working backwards was pioneered by Hudson (1983). The true genealogical tree is recorded, including the ordered tree topology as well as the actual times of the interior nodes.

From the time-ordered tree topology we can compute the quantities $m_i$ which are used in (14), (16), and (17). This gives us the expectation and variance of the estimate of $\theta$ for that ordered tree topology. The overall expectation of the estimate of $\theta$ will be the average of (16) over all tree topologies. The variance of the estimate will be the average of the within-topology variances (17), plus the variance of the expectations (16), averaging over all ordered tree topologies. Lacking a theory of the statistical behavior of the $C_i$ we cannot compute the expectation and variance of the estimate of $\theta$ in the pairwise method.

The approach taken here has been to compute these approximately by sampling a large number of ordered tree topologies. The second computer program takes each one and computes (16) and (17). These then can be used to compute the approximate expectation and variance of the estimate of $\theta$ under the pairwise method. Table 2 shows the results. Its penultimate column shows the variance of the pairwise estimate in the same case, as determined by this combination of simulation and theory. The computed expectations of the estimate of $\theta$ are not shown; they were always quite close to $\theta$ and support the conclusion that the pairwise method, like the maximum likelihood method, is unbiased. The variances of the pairwise estimate computed from the simulation are quite

close to the values obtained from the Tajima-Nei formula.

The computer programs that simulate the coalescent record not only the ordered tree topology, but the true ages of the interior nodes as well. This makes it possible in each case, using (6), (11), and (22), to compute what the maximum likelihood, the pairwise, and the segregating sites estimates of $\theta$ would be for that tree, given that long enough sequences were available to estimate the genealogical tree exactly. So for each simulated tree we get three estimates of $\theta$. The means of these estimates were in accord with the unbiasedness of the estimates and are not shown here. The first six columns of Table 2 show the empirical variances of the three estimates, and the ratios of variances, which are estimates of the efficiency of the pairwise and the segregating sites methods, computed directly from the simulations without using equations (8), (16), (17), or (24). Figures 2 and 3 show the variances and efficiencies from both tables. The lines connect the theoretical variances and efficiencies from Table 1, and the points are the empirical variances and efficiencies from the simulation results in Table 2. Again, the results show excellent agreement of the theory with the simulations.

### 6. Relation to infinite-sites models

The present approach may at first seem unrelated to the papers by Strobeck (1983), Ethier and Griffiths (1987), and Griffiths (1989) which calculate probabilities of different kinds of samples that might be taken from a population undergoing an infinite-sites model of mutation, in the absence of recombination. Strobeck (1983) used a diffusion approximation to derive recurrence relations among the probabilities of the different possible kinds of observed samples for two or three sequences. Ethier and Griffiths (1987) gave a general recursion formula for these probabilities for any number of sequnces. Strobeck (1983) shows how to use these formulas to make maximum likelihood estimates of $\theta$. Griffiths (1989) describes a computer program that can calculate these probabilities. The

probability of the observed sample is of course the likelihood, and by varying $\theta$ one can compute the likelihood curve and the maximum likelihood estimate.

The effect of having infinitely long sequences, as assumed here, is most easily seen by considering Strobeck's equations, for the case where two sequences are observed. As we use $\mu$ for the mutation rate at a single site, suppose that $U$ is the mutation rate for the whole locus, and that $\Omega = 4N_eU$. This is the quantity that these authors call $\theta$ in their equations. Strobeck's equation (2) for the case of two different sequences shows, for $n = 2$ and $m_1 = m_2 = 1$, that the distribution of the number of mutations by which two sequences differ is geometric, with mean $\Omega$. The variance of this distribution will be $\Omega^2 + \Omega$. The maximum likelihood estimate of $\Omega$ turns out to be simply the observed number of mutations by which the sequences differ, so that the mean and variance of the estimate are also $\Omega$ and $\Omega^2 + \Omega$. The squared coefficient of variation is the $1 + 1/\Omega$.

This exceeds the value calculated in this paper by $1/\Omega$. The extra variation is due to the inaccuracy of estimating the tree (which in this case is simply the divergence time of the two sequences). As the number of sites is taken larger and larger for a given value of $\theta$, $\Omega \to \infty$ so that $1 + 1/\Omega \to 1$. Thus the extra variability of the estimate due to the finiteness of $\Omega$ disappears, and we are left only with the variability due to the randomness of the true genealogy, the randomness accounted for in this paper. I expect that the same behavior will occur when more sequences are considered, and that this could be verified by a detailed consideration of the equations in Strobeck (1983) and Griffiths (1989). If so, there would be no conflict between their results and mine.

**7. Limitations**

The proofs above rely on a number of assumptions that are questionable:

*(i) No recombination.* It is assumed that the sequences have a genealogy which is a branching tree, and this can only happen when there is no recombination in the region in any of the lineages leading back to the common ancestor

sequence. Recombination would result in a single sequence (the recombinant) having contributions from two or more ancestors. With a single recombination, the front and rear ends of the sequences would have different, but similar, genealogical trees. Treating such cases is a major challenge for the future. For mitochondrial DNA sequences, strict maternal inheritance guarantees that this problem does not arise.

*(ii) Infinitely long sequences.* The analysis here was enormously facilitated by the assumption that we have infinitely long sequences so that they allow us to estimate the details of the genealogy without error. The statistical error in this study is thus only the error that comes from having a finite number of sequences. If instead we had sequences of finite length, as we always would, the maximum likelihood method becomes much more difficult computationally. I hope to present in a separate paper a computationally intensive procedure that can make a maximum likelihood estimate of effective population size from samples of finite-length sequences. In that case there is additional statistical error from the imprecision of estimation of both the topology of the genealogy and the divergence times. It would inflate the error of all of the estimates. It is not obvious which one would be affected most, but it is at least possible that, as both the numerator and the denominator of the efficiencies are affected, that the efficiency of pairwise and segregating sites methods would not be as low when the sequence lengths were small. However when sequences are long, the variances and efficiencies must approach the values given here.

*(iii) Lack of geographical subdivision.* It has been assumed that there is only one population, mating at random. If there are a number of local populations exchanging migrants, the notion of effective population size becomes complicated: we have both local effective sizes (Sewall Wright's (1940) concept of "neighborhood size") and an effective number for the whole species which can be considerably larger. Both contribute to the rate of coalescence of lineages. When two lineages are in the same local population, it will be possible for them

to coalesce in the previous generation, while when they are in different local populations they cannot. Slatkin (1987), Takahata (1988) and Takahata and Slatkin (1990) have investigated this for two populations exchanging migrants. The distribution of the time to coalescence of two lineages collected from the same local population is no longer exponential and is not easy to obtain. Slatkin and Maddison (1989) have proposed an estimate of migration rate between the populations for the case of infinitely long sequences. For the case where the sequences are not very divergent their estimate is probably close to being a maximum likelihood estimate.

It should be obvious that much remains to be done; it may be doubted whether the methods of analysis will ever catch up with the collection of data.

I am grateful to Monty Slatkin for frequent discussions of the coalescent, for pointing out the relevance of Polya's urn model, and for comments on the manuscript of this paper. I also thank Charles Geyer for comments on the lack of consistency of the pairwise method, David Aldous for comments on the logical status of the maximum likelihood estimator, and John Wakeley and Avigdor Beiles for finding typographical errors. I wish also to thank Richard Hudson, associate editor of *Genetics*, and his reviewers for helpful comments on an earlier version of this paper, in particular for pointing out the relevance of the work on infinite-sites models and suggesting notational reforms. This work was supported by National Science Foundation grants numbers BSR-8614807 and BSR-8918333, and by National Institutes of Health grant number 1 R01 GM 41716-01.

### References

Avise, J. C. (1989) Gene trees and organismal histories: a phylogenetic approach to population biology. *Evolution* **43**, 1192-1208.

Avise, J. C., Ball, R. M., Jr. & Arnold, J. (1988) Current versus historical population sizes in vertebrate species with high gene flow: a comparison based on

mitochondrial DNA polymorphism and inbreeding theory for neutral mutations. *Molecular Biology and Evolution* **5**, 331-344.

Ball, R. M., Jr., Neigel, J. E. & Avise, J. C. (1990) Gene genealogies within the organismal pedigrees of random-mating populations. *Evolution* **44**, 360-370.

Cann, R. L., Stoneking, M. & Wilson, A. C. (1987) Mitochondrial DNA and human evolution. *Nature* **325**, 31-36.

Ethier, S. N. & Griffiths, R. C. (1987) The infinitely-many-sites model as a measure-valued diffusion. *Annals of Probability* **15**, 515-545.

Feller, W. (1968) *An Introduction to Probability Theory and Its Applications*, 3rd edn. John Wiley, New York.

Griffiths, R. C. (1989) Genealogical tree probabilities in the infinitely-many-site model. *Journal of Mathematical Biology* **27**, 667-680.

Harding, E. F. (1971) The probabilities of rooted tree shapes generated by random bifurcation. *Advances in Applied Probability* **3**, 44-77.

Hudson, R. R. (1983) Testing the constant-rate neutral allele model with protein sequence data. *Evolution* **37**, 203-217.

Kingman, J. F. C. (1982a) The coalescent. *Stochastic Processes and Their Applications* **13**, 235-248.

Kingman, J. F. C. (1982b) On the genealogy of large populations. *Journal of Applied Probability* **19A**, 27-43.

Maddison, W. P. & Slatkin, M. (1991) Null models for the number of evolutionary steps in a character on a phylogenetic tree. *Evolution* **45**, 1184-1197.

Moran, P. A. P. (1958) Random processes in genetics. *Proc. Camb. Phil. Soc.* **54**, 60-71.

Nei, M. & Tajima, F. (1981) DNA polymorphism detectable by restriction endonucleases. *Genetics* 97, 145-163.

Nei, M. (1987) *Molecular Evolutionary Genetics.* Columbia University Press, New York.

Saunders, I. W., Tavaré, & Watterson, G. A. (1984) On the genealogy of nested subsamples from a haploid population. *Advances in Applied Probability* **16**, 471-491.

Slatkin, M. (1987) The average number of sites separating DNA sequences drawn from a subdivided population. *Theoretical Population Biology* **32**, 42-49.

Slatkin, M. (1989) Detecting small amounts of gene flow from phylogenies of alleles. *Genetics* **121**, 609-612.

Slatkin, M. & Maddison, W. P. (1989) Cladistic measure of gene flow inferred from the phylogenies of alleles. *Genetics* **123**, 603-613.

Slowinski, J. G. & Guyer, C. (1989) Testing the stochasticity of patterns of organismal diversity: an improved null model. American Naturalist **134**, 907-921.

Tajima, F. (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437-460.

Takahata, N. (1988) The coalescent in two partially isolated diffusion populations. *Genetical Research* **52**, 213-222.

Takahata, N. & Slatkin, M. (1990) Genealogy of neutral genes in two partially isolated populations. *Theoretical Population Biology* **38**, 331-350.

Tavaré, S. (1984) Line-of-descent and genealogical processes, and their applications in population genetics models. *Theoretical Population Biology* **26**, 119-164.

Watterson, G. A. (1975) On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* **7**, 256-276.

Wright, S. (1940) Breeding structure of populations in relation to speciation. *American Naturalist* **74**, 232-248.

Table 1. Theoretical variance of the maximum likelihood estimate of $\theta$, when $\theta = 4$, of the pairwise method, using equations (7) and (18), and of Watterson's method, from equation (23). The efficiencies of the pairwise method, from equation (19) and of Watterson's method, from equation (24), are also shown.

| $n$ | $Var(ML)$ | $Var(P)$ | $Efficiency(P)$ | $Var(W)$ | $Efficiency(W)$ |
|---|---|---|---|---|---|
| 2 | 16.0000 | 16.0000 | 1.00000 | 16.0000 | 1.00000 |
| 3 | 8.0000 | 8.8889 | 0.90000 | 8.8889 | 0.90000 |
| 4 | 5.3333 | 6.8148 | 0.78261 | 6.4793 | 0.82313 |
| 5 | 4.0000 | 5.8667 | 0.68182 | 5.2480 | 0.76220 |
| 6 | 3.2000 | 5.3333 | 0.60000 | 4.4917 | 0.71243 |
| 7 | 2.6667 | 4.9947 | 0.53390 | 3.9754 | 0.67080 |
| 8 | 2.2857 | 4.7619 | 0.48000 | 3.5980 | 0.63528 |
| 9 | 2.0000 | 4.5926 | 0.43548 | 3.3085 | 0.60451 |
| 10 | 1.7778 | 4.4642 | 0.39823 | 3.0784 | 0.57751 |
| 15 | 1.1429 | 4.1143 | 0.27778 | 2.3850 | 0.47918 |
| 20 | 0.8421 | 3.9579 | 0.21277 | 2.0259 | 0.41567 |
| 25 | 0.6667 | 3.8696 | 0.17228 | 1.8001 | 0.37034 |
| 30 | 0.5517 | 3.8130 | 0.14469 | 1.6424 | 0.33593 |
| 35 | 0.4706 | 3.7737 | 0.12470 | 1.5245 | 0.30868 |
| 40 | 0.4103 | 3.7447 | 0.10956 | 1.4323 | 0.28643 |
| 45 | 0.3636 | 3.7226 | 0.09768 | 1.3577 | 0.26784 |
| 50 | 0.3265 | 3.7050 | 0.08813 | 1.2957 | 0.25201 |
| 60 | 0.2712 | 3.6791 | 0.07371 | 1.1980 | 0.22637 |
| 70 | 0.2319 | 3.6608 | 0.06334 | 1.1236 | 0.20637 |
| 80 | 0.2025 | 3.6473 | 0.05553 | 1.0646 | 0.19024 |
| 90 | 0.1798 | 3.6368 | 0.04943 | 1.0163 | 0.17688 |
| 100 | 0.1616 | 3.6285 | 0.04454 | 0.9759 | 0.16561 |
| 150 | 0.1074 | 3.6038 | 0.02980 | 0.8405 | 0.12776 |
| 200 | 0.0804 | 3.5916 | 0.02239 | 0.7607 | 0.10569 |
| 300 | 0.0535 | 3.5795 | 0.01495 | 0.6661 | 0.08033 |
| 400 | 0.0401 | 3.5734 | 0.01122 | 0.6093 | 0.06582 |
| 500 | 0.0321 | 3.5698 | 0.00898 | 0.5700 | 0.05625 |

Table 2. Empirical variances of the coalescent, pairwise, and Watterson estimates of $\theta$ in the computer simulations, plus the empirical efficiency of the pairwise and Watterson method obtained by taking the ratio of the coalescent variance to those for these other two. The seventh column shows the value of the variance of the pairwise method obtained by simulation, computing within- and between tree topology variances by sampling ordered tree topologies for this value of $\theta$ and using equations (16) and (17), for different values of $n$. The eighth is the number of ordered tree topologies sampled in the simulation.

| $n$ | $Var(ML)$ | $Var(P)$ | $Eff(P)$ | $Var(W)$ | $Eff(W)$ | $Var'(P)$ | *Replicates* |
|---|---|---|---|---|---|---|---|
| 2 | 16.0510 | 16.0510 | 1.00000 | 16.0510 | 1.00000 | 16.0000 | 1,000,000 |
| 3 | 7.9562 | 8.8401 | 0.90001 | 8.8401 | 0.90001 | 8.9709 | 1,000,000 |
| 4 | 5.3259 | 6.8132 | 0.78170 | 6.4732 | 0.82277 | 6.7988 | 1,000,000 |
| 5 | 3.9900 | 5.8340 | 0.68391 | 5.2227 | 0.76397 | 5.8855 | 1,000,000 |
| 6 | 3.1971 | 5.3470 | 0.59793 | 4.5022 | 0.71012 | 5.3343 | 1,000,000 |
| 7 | 2.6553 | 4.9679 | 0.53449 | 3.9524 | 0.67181 | 4.9923 | 1,000,000 |
| 8 | 2.2895 | 4.7656 | 0.48042 | 3.6010 | 0.63579 | 4.7590 | 1,000,000 |
| 9 | 2.0020 | 4.5827 | 0.43686 | 3.2978 | 0.60707 | 4.5990 | 1,000,000 |
| 10 | 1.7770 | 4.4475 | 0.39954 | 3.0746 | 0.57796 | 4.4661 | 1,000,000 |
| 15 | 1.1433 | 4.1089 | 0.27824 | 2.3835 | 0.47966 | 4.1124 | 1,000,000 |
| 20 | 0.8446 | 3.9564 | 0.21346 | 2.0267 | 0.41671 | 3.9557 | 1,000,000 |
| 25 | 0.6648 | 3.8759 | 0.17153 | 1.8042 | 0.36848 | 3.8694 | 1,000,000 |
| 30 | 0.5524 | 3.8092 | 0.14501 | 1.6427 | 0.33625 | 3.8144 | 1,000,000 |
| 35 | 0.4704 | 3.7637 | 0.12499 | 1.5194 | 0.30963 | 3.7717 | 1,000,000 |
| 40 | 0.4098 | 3.7458 | 0.10940 | 1.4330 | 0.28596 | 3.7440 | 1,000,000 |
| 45 | 0.3636 | 3.7244 | 0.09762 | 1.3579 | 0.26775 | 3.7229 | 1,000,000 |
| 50 | 0.3257 | 3.6870 | 0.08833 | 1.2933 | 0.25183 | 3.7036 | 1,000,000 |
| 60 | 0.2707 | 3.6853 | 0.07345 | 1.2014 | 0.22530 | 3.6782 | 500,000 |
| 70 | 0.2314 | 3.6809 | 0.06287 | 1.1275 | 0.20526 | 3.6627 | 500,000 |
| 80 | 0.2019 | 3.6484 | 0.05534 | 1.0631 | 0.18993 | 3.6498 | 500,000 |
| 90 | 0.1794 | 3.6164 | 0.04960 | 1.0116 | 0.17731 | 3.6354 | 500,000 |
| 100 | 0.1614 | 3.6249 | 0.04451 | 0.9723 | 0.16594 | 3.6281 | 500,000 |
| 150 | 0.1073 | 3.5991 | 0.02983 | 0.8383 | 0.12805 | 3.6063 | 500,000 |
| 200 | 0.0803 | 3.6124 | 0.02222 | 0.7626 | 0.10526 | 3.5922 | 500,000 |
| 300 | 0.0536 | 3.5718 | 0.01500 | 0.6677 | 0.08022 | 3.5794 | 200,000 |
| 400 | 0.0400 | 3.5772 | 0.01119 | 0.6132 | 0.06527 | 3.5729 | 200,000 |
| 500 | 0.0325 | 3.8180 | 0.00852 | 0.5953 | 0.05462 | 3.5725 | 10,000 |

FIGURE CAPTIONS

Figure 1. A genealogical tree, showing the relationship between the $u_i$ and the $t_i$. Both are measured in generations back from the present.

Figure 2. The variances of the estimates of $\theta$ from the simulation from n=3 to n=500 when computed by the coalescent maximum likelihood method (circles) the pairwise method (squares), and Watterson's method (pluses). The continuous curves are the corresponding theoretical values from equations (7), (18), (23). The value for n=500 is based on many fewer simulations than the other values.

Figure 3. Theoretical and empirical values of the efficiency of the pairwise and segregating sites estimates of $\theta$ from n=3 to n=500. The lower curve shows the theoretical value, computed from equation (19), and the upper curve the theoretical efficiency of Watterson's method, from (24). Squares show the empirical values obtained by taking the ratio of the empirical variances among replicates of the coalescent and pairwise estimates, and pluses the empirical values for Watterson's method.