Journal of Advanced Bioinformatics Applications and Research
ISSN 0976-2604.Online ISSN 2278–6007
Vol 5, Issue2, 2014, pp56-71
http://www.bipublication.com

ON THE ESTIMATION OF MUTATION RATE BASED ON COALESCENCE GENEALOGY

Ao Yuan and Guanjie Chen

National Human Genome Center, Howard University, Washington DC, USA.
Center for Research on Genomics and Global Health, NHGRI, NIH, Bethesda, Maryland, USA.
Email:chengu@mail.nih.gov

[Received-03/01/2012, Accepted-01/04/2014]

ABSTRACT:

Estimating gene mutation rates using DNA data from a population since its coalescence is an important part in the study of evolutionary genetic history, and is of its own interest also. There are a number of existing methods for this goal and are successful in application, but formal asymptotic study is absent except for the Watterson estimator, such as consistency of the estimator and its rate. Since the observed data are genetically dependent, the problem is non-standard. Here we investigate the asymptotic property for the estimation of mutation rate in a given population under three common cases in practice, estimations based on coalescent tree, rooted tree and the total number of observed segregation sites only. We show, in each case, the strong consistency of the estimate, and its asymptotic normality with a very slow rate of $(\log n)^{1/2}$, in contrast to the standard rate of $n^{1/2}$for independently identically distributed data. Although this result is known several decades ago, we rediscovered this fact without knowing the previous results and the settings and consitions we used are not all the same as those previous studies. We also propose a simple simulation based method for the estimation of the mutation rate using DNA data. The result is illustrated by a segment of the mitochondrial data from an Amerindian tribal population, compared to some of the commonly used existing methods, and found to be consistent with these methods.

Key words: Coalescence, Convergence rate, DNA data, Genealogy, Mutation rate.

1. INTRODUCTION

In the past decades, considerable progress has been made in the field of population genetics. One of the main goals is to study the evolutionary history of a population under study since the time of their most recent common ancestor (MRCA), in which estimating the gene mutation rate in this population since coalescence is a precursory step and plays a fundamental role in the study. Also, mutation rate estimation has its Owen interest.

The coalescent theory is a retrospective of population genetics that traces all genes in a sample from a population to a single ancestral copy shared by all the members of the population. The coalescent time of a population is the time of their MRCA at

which their ancestral history converge. The gene mutation rate since that time in the population directly affects the results in the study of the population evolutionary history. As an indispensable element in such study, the mutation rate is to be estimated using the observed DNA sequence diversity from a sample of n individuals independently drawn from the population of known size N . The problem is not as seemingly simple as the gene frequency estimation, since often mutations are rare events, the generations spans a long history, and the data are a highly dependent sample with varying dependence structure over generations. Also, the data should be from a relatively closed population, so its expensive and time consuming to collect.

For this problem, there are many methods. Watterson [24] derived moment estimator using the number of observed segregating sites alone, Felsenstein [7] and Kuhner et al. [18] proposed maximum likelihood estimate (MLE) of related parameters. Fu and Li [8] and Fu [9] derived lower bound for the variance of the MLE and the best linear unbiased estimate, under the assumption that the number of mutations on each branch of the coalescent tree are known. Recently some studies [12,22,20,4,14,2,11] investigated mutation rate on on X or Y-chromosome microsatellite and mtDNA data without using the genealogy tree. However, asymptotical properties of these methods, such as consistency of the estimates and the asymptotic normality are not seen in the literature. Here the data is genealogically dependent, and evolve over time, the scenario is very different from the common case of independent and identically distributed (i.i.d.) data. Some simulation studies did not show apparent concentration of the estimator to the true parameter value as the sample size n increases, although the accuracy improves a little bit. A natural question is: whether the estimator will be consistent as the number of samples n tends to infinity? If it does, at what rate? Here we investigate this problem under three common cases in practice, estimations based coalescent tree, rooted tree and the total number of segregating sites only (the Watterson estimator). We show the answer to the above problem is affirmative for each case, with common convergence rate of $(\log n)^{1/2}$ to normality.

In contrast to the convergence rate of $n^{1/2}$ in the case of i.i.d. data, the above rate is very slow. This is why sometimes apparent improvement in estimation of mutation rate is not observed with increased sample size. After our work is done, some readers point out that the asymptotic normality result has been known since around 1970, and that for the Watterson estimator has already been done by Klein et al [17]. However, our proof method is very different from theirs, and the settings and conditions used here are not all the same as in those previous studies.

Also, often the DNA data only provides the number of segregating sites, not information in the form of a coalescent tree, nor the number of mutations in each branch of the genealogy tree. So methods based on likelihood models using coalescent tree or the number of mutations in each branch of the genealogy tree are not directly applicable. Also, the method using the number of segregating sites only does not use the prior information in the genealogy at all. Since the convergence rate of mutation rate estimation is very slow, estimation with finite sample size is more efficient if we implement such prior information. Here we propose a very simple simulation method for this problem, with the genealogy information - the prior coalescent times distribution implemented, without using neither the coalescent tree nor the number of mutations in each branch of the genealogy tree, as such information is not directly available for most DNA data. Also, the MLE of the mutation rate under the full data information of the form of a coalescent tree, the estimation turns out only depend on the total number of observed mutation sites in the data and the prior tree length. The latter is often unknown in practice. We implement such information by simulation from the prior coalescent distribution. The result is illustrated by a real example, compared to some of

the commonly used existing methods, and found to be consistent with these results.

Mitochondrial DNA data are commonly used in gene mutation studies. It is one of the few genes exist outside the cell nucleus, and for mammalian it is only maternally inherited. Human mtDNA is a double-stranded molecule sequence about 16,500 base pairs in length. It is known that the mutation rate in mtDNA is about 10 times that of the nuclear genes, and that on one part of the mitochondria, its control region, is even one order higher. The simple inheritance pattern and high variability make mtDNA an important source to study human evolutionary history. In Section 2, we give a brief review of the background of the problem and three commonly used methods. Section 3 studies the asymptotic behavior of these methods, with relevant proofs in the Appendix. Our study shows that for the estimation of the mutation rate θ , all these methods are asymptotically equivalent with convergence rate $(\log n)^{1/2}$, a much slower rate than the typical rate of $n^{1/2}$ for independent and identically distributed data. The reason is that the observed sequences are highly dependent. Thus, as the validity of asymptotic result requires impractically large sample size, in practice we should take as much genealogy information as possible. In section 4 we present our proposed method, and Section 5 illustrates its use on a segment of the mitochondrial data from an Amerindian tribal population, and compare the results with those of others, with brief concluding remarks in Section 6.

2. Brief Review of Background and Related Methods

The gene mutation rate is the probability that a mutant type occurs in a given reference time period. For example in many studies such time period is taken as per generation. In population history studies, the time period is since the coalescence of the population to present. The coalescent is a model for the genealogical tree of a random sample of ^{n}DNA sequences from a large population. An example of such coalescent tree of sample size n=7 is given in Figure 1.

{1cm} \centerline{Figure 1. about here} {1cm} \vspace*{1cm} %\centerline{\psfig {figure=mutfig1.ps,height=6cm,width=7cm,angle=0 }} %\begin{figure}[htb] %\caption{Coalescent tree for a sample of seven individuals.} %\end{figure}

In this Figure, w_2 is the time, in unit of 2N generations, between the points when the sample of seven people from a given population has 2 and a single common ancestor, w_3 is the time between the points when the sample has 3 and 2 common ancestors, etc. For more detailed reviews materials of this topic see Hudson [13], Donnelly and Tavaré S. [3].

In coalescence inference there are some explicite and implicite assumptions in the literatures, which we list below.

Basic assumptions:

The population size *N* is large, remain unchanged for many generations into the past, and is known, or can be estimated from other sources; the data is a random sample from the population; the number of births in each generation follows the Wright-Fisher model (since the population is of constant size, the number of deaths also follow the similar model); mutation (substitution) at any nucleotide site can occur only once in the ancestry and is irreversible; mutations occur in different time intervals are independent; all loci have the same mutation rate; the time point at which mutation occurs follow a Poisson distribution with rate $\theta/2$ to be defined latter, independently in each branch of the genealogy tree.

Here estimating θ based on the observed DNA data and investigating its asymptotic behavior are the goals of our study. The inference of θ and that of the coalescence time t_n of a sample population of size n has close relationship. The latter has two steps. The first step is modeling the distribution of without any data, the pre-data distribution; then in the second step, update the pre-data distribution, using the

observed data, to the post-data distribution, based on which the formal inference is conducted. The pre-data distribution is pioneered by Kingman [15,16], who showed that in time units of **N** generations,

$$t_n = \sum_{j=2}^{n} w_j, \quad (1)$$

where the w_j's are independent waiting times, w_j is the duration between the sample had j and j-1 common ancestors (see, for example, Tavaré S. , [25]). Here w_j is distributed as the exponential model $Exponential(j(j-1)/2)$ with mean $E(w_j) = 2/(j(j-1))$. The w_j's can be represented graphically as a coalescent tree as in Figure 1, then is the height of the tree. Define the tree's total branch length as

$$l_n = \sum_{j=2}^{n} j w_j,$$

then (Kingman)

$$E(t_n) = 2\left(1 - \frac{1}{n}\right), \quad Var(t_n) = 8\sum_{j=2}^{n} \frac{1}{j^2} - 4\left(1 - \frac{1}{n}\right)^2;$$

$$E(l_n) = 2\sum_{j=1}^{n-1} \frac{1}{j}, \quad Var(l_n) = 4\sum_{j=1}^{n-1} \frac{1}{j^2}. \quad (2)$$

Thus on average the coalescence time of **N** people of the given population is about **2N** generations back into the past. The time unit is transformed to years by the relationship $t_n NY$, where ***Y*** is the average years of each generation, which usually taken as 20-25.

Here we see that, as an initial analysis without the observed data, the coalescent time of a random sample of size ***n*** from a population of size ***N*** is roughly ***2N*** generations, as long as $n(\leq N)$ is moderately large. Thus the coalescent time of a sub-sample from a population is roughly the same as that of the total population (as long as the sample size is moderately large). Here the sample must be a random draw from the population, otherwise the result may not be reliable.

For mutation, the common assumption is that the times at which mutation occur follow a Poisson process with constant rate **θ/2**, so that in any branch of length l from the tree, the number of mutations on that branch has a Poisson distribution with mean $l\theta/2$, independently of the mutations on the other branches. For the time scale mentioned before, usually $\theta = 2N\mu$, where μ is the probability of a mutation occurs per sequence per generation. For DNA sequences, μ is the sequence length (number of bases) times the mutation rate per site per generation. and often available from existing studies. Since the coalescent time of a sample with moderate size is approximately $2N$ generations, θ can be approximately interpreted as the commulative (since the time of MRCA) mutation rate (number of mutations) per sequence. Also, since the population size is N, $\theta/2$ can also be interpreted as the mutation rate of the whole population per generation.

Thus given the mutation rate θ and the tree length l_n, the number of mutations s_n in a sample of n individuals from the given population follow the Poisson distribution $\mathrm{Po}(\theta l_n/2))$ [25].

$$P(s_n = k | l_n = l) = e^{-\theta l/2}\frac{(\theta l/2)^k}{k!} := Po(k, \theta l/2) \quad k = 0,1,2,\ldots (3)$$

Note this probability does not depend on n , but on k , l and θ.

Now we come to the problem of estimating the mutation rate θ . Given the sampled sequence data, we don't know which sites are mutant, and the data have inhomogeneous dependence among them. So the estimation may not be straight forward. The commonly used methods, depend on data type and how much information is used from the observed data, varies from very simple to very complicated. The method of moment (MM) estimate is to solve the equation

$$E(s) = \theta h_n, \quad h_n = 1 + 1/2 + \cdots + 1/(n-1)$$

by replacing $E(s)$ with the observed number of segregating sites, which is the same as the total number of mutations in the observed sequences, under the assumption that mutation can only occur at most once at each site and is irreversible. This method is extremely simple, but does not use the data structural information at all. Also, it does not provide estimated standard deviation. For more detail of this method see Watterson [24]. The method based on coalescent tree and that based on number of mutations on each tree segments are also simple and more informative than the moment method. But in practice, the observed data are often not in the form of a coalescent tree, nor the number of mutations on each tree segments are directly available. Thus to use these methods, one needs to first infer the genealogy information to construct the coalescent tree or the number of mutations on each tree segments, by some other methods. Generally, the observed data is only in the form of some DNA sequences, with no additional information to construct a coalescent tree, nor the number of mutations on each tree segments. For this type of data, often we only know the number of segregating sites. It can also be represented as an unrooted tree or a certain number of rooted trees, the well known method in Griffiths and Tavaré S. [10], hereafter GT, is based on the full data information represented by a set of rooted trees. Detailed description on coalescent tree, rooted tree and unrooted tree can be found in Tavaré S. [25]. This method is one of the basic tools for this problem using full data information, but is computationally complicated. GT used the probabilities recursion formula, derived in Ethier and Griffiths [5], which is not easy to use for many geneticists. The methods of Lundstrom et al. [19] are a type of least squares (LS) method and likelihood one, in which the likelihood is not the true data likelihood, so they call the estimate from this likelihood independent-sites (IS) estimator. This latter method is also not simple computationally.

3. Asymptotic Results of Some Existing Methods

In order to answer the basic questions of consistency and the convergence rate raised in the Introduction, here we first investigate the asymptotic properties of the mutation rate estimations under three commonly used data forms for this problem: coalescent tree, segments of mutations, and number of mutations only, although the first two data forms are generally not directly available from the observed DNA sequence data.

3.1 Estimation Based on Coalescent Tree.

We first consider the case the data is in the form of a given coalescent tree of n sequences, as in Figure 1. This type of data assumes the coalescent times w_i's and the number of mutations k_{ij}'s on each branch are known. There are $n-1$ nodes (splitting points) in the tree numbered 2 to n in their time order. Recall the definition of the ith coalescent time w_i. Between the $(i-1)$th and ith node there are exactly i segments, denote them as $w_{i1},\ldots,w_{ii}$ from left to right, each has length w_i. Then k_{ij} is the number of mutations on segment w_{ij}. Denote $\mathbf{k}=\{k_{ij}: i=2,\ldots,n; j=1,\ldots,i.\}$ and $\mathbf{w}=\{w_{ij}: i=2,\ldots,n; j=1,\ldots,i.\}$. Since the numbers of mutations on different segments are independent, the probability of $\mathbf{k}$ given $\mathbf{w}$ and the mutation rate θ is

$$P(\mathbf{k}|\mathbf{w},\theta)=\prod_{i=2}^{n}\prod_{j=1}^{i} Po(k_{ij}, w_{ij}\theta/2). \quad (4)$$

To compute the MLE $\theta_{1,n}$ of θ under the above model, let $L_{1,n}(\theta)$ be the log-likelihood of θ under the above probability, and $L'_{1,n}(\theta)$ be its first derivative with respect to θ. Set $L'_{1,n}(\theta)=0$, we get $\theta_{1,n}=\frac{2\sum_{i=2}^{n}\sum_{j=1}^{i}k_{ij}}{\sum_{i=2}^{n} i w_i}=\frac{2k}{l_n}$,

where $k = \sum_{i=2}^{n} \sum_{j=1}^{i} k_{ij}$ is the total number of segregating sites in the observed data, it is also the total number of mutations in the observed sequences under the common assumption mentioned before. For i.i.d. observations x_i's with common density function $f(\cdot|\theta)$, the Fisher information plays an important role in the asymptotic distribution of estimators. It is defined as $I(\theta) = -E_\theta[\partial^2 \log f(x|\theta)/\partial\theta^2]$, and is equal to

$$I(\theta) = \lim_n -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \log f(x_i|\theta)}{\partial\theta^2}.$$

In our case, there is no common density function, we define the asymptotic Fisher information as

$$I(\theta) = \lim_n -\frac{1}{h_n}\frac{\partial^2 \log P(\mathbf{k}|\mathbf{w},\theta)}{\partial^2\theta}.$$

The reason we use h_n here instead of the common n for the i.i.d. case will be clear in the proof. Denote $\xrightarrow{a.s.}$ for convergence almost surely, and $\xrightarrow{D}$ for convergence in distribution, and θ_0 be the true parameter value generates the data.

Proposition 1. As $n \to \infty$, we have i) $\theta_{1,n} \xrightarrow{a.s.} \theta_0$.

ii) If $\theta_0 > 0$, then $h_n^{1/2}(\theta_{1,n} - \theta_0) \xrightarrow{D} N(0, I^{-1}(\theta_0))$

where $I(\theta_0) = 1/\theta_0$.

Remark 1. Since $h_n/\log(n) \to 1$, the convergence rate of $\theta_{1,n}$ to θ_0 is $(\log n)^{1/2}$. In contrast to the standard rate of $n^{1/2}$ for the independent and identically distributed (i.i.d.) observation case, this rate is much slower. However this result tells us that as the number of observations increase without bound, the inference of θ can achieve arbitrary accuracy.

3.2 Estimation Based on Segments Mutations. The coalescent tree provides the full information for estimating the mutation rate, but generally the observed data are not in the form of a coalescent tree, i.e. often we do not have the w_{ij}'s nor the k_{ij}'s. Suppose we have the number of mutations k_i's on segments of the coalescent tree, but not the coalescent times w_{ij}'s. Let $f(w_i)$ be the density function of $w_i = \sum_{j=1}^{i} w_{ij}$. In this case the likelihood is [8].

$$P(\mathbf{k}|\theta) = \prod_{i=2}^{n} \int \prod_{j=1}^{i} [Po(k_{ij}, w_i\theta/2) f(w_i)] dw_i = \prod_{i=2}^{n} \frac{(1+i-1)\theta^{k_i}}{(\theta+i-1)^{k_i+1}},$$

and the corresponding log-likelihood is, up to a constant,

$$L_{2,n}(\theta) = \sum_{i=2}^{n} [k_i \log(\theta) - (k_i+1)\log(\theta+i-1)].$$

Set $L'_{2,n}(\theta) = 0$, the MLE $\theta_{2,n}$ of θ is the solution of the equation $\sum_{i=2}^{n} \frac{k_i+1}{\theta+i-1} = \frac{k}{\theta}$, which has no closed form. We still have

Proposition 2. Suppose $c \le \theta_{2,n} \le C$, all n, for some $0 < c \le C < \infty$, then as $n \to \infty$, we have i) $\theta_{2,n} \xrightarrow{a.s.} \theta_0$.

ii) If $\theta_0 > 0$, then $h_n^{1/2}(\theta_{2,n} - \theta_0) \xrightarrow{D} N(0, I^{-1}(\theta_0))$.

3.3. Estimation Based on Number of observed segregating sites only.

In this case, let be the total number of observed of segregating sites, it is also the number of mutations in the collected sequences as explained in Section 3.1. The estimation of mutation rate is given by the moment estimator [24],

as $\theta_{3,n} = \frac{k}{h_n}$. We have

Proposition 3.

As $n \to \infty$, we have i) $\theta_{3,n} \xrightarrow{a.s.} \theta_0$.

ii) If $\theta_0 > 0$, then $h_n^{1/2}(\theta_{3,n} - \theta_0) \xrightarrow{D} N(0, I^{-1}(\theta_0))$.

Remark 2:

a) Klein et al. [17] obtained the same result above. Their proof is based on Laplace transformation, while we used the Lindeberg condition.

b) Compare ii) of Propositions 1, 2 and 3, we see that the estimators $\theta_{1,n}$, $\theta_{2,n}$ and $\theta_{3,n}$ are asymptotically equivalent in the sense of asymptotical normality. This is a striking phenomenon, since for finite sample size, the lower bounds b_1, b_2 and b_3 for the variances of the three estimators satisfy [8] $b_1 \le b_2 \le b_3$.

Our results indicate that the estimation of mutation rate is eventually determined by the number of observed mutations, the effect from the genealogy information will be dominated as the sample size increases without bound (similar phenomenon was also found in Fu and Li, [8]).

c) On the other hand, for i.i.d. data, the convergence rate is $n^{1/2}$, the asymptotic normality is valid roughly for sample size $n \ge 20$. In our case since the convergence rate $(\log n)^{1/2}$ is much slower, the sample size for validity of normal approximation will be much larger. If we require $\log(n) \ge 20$, or $n \ge 485165195$ for this problem, this will be an impractical number. So, for moderate sample size, the estimation from the three methods can be significantly different as shown in Griffiths and Tavaré S. [10].

Thus in practice, estimation of the mutation rate should take into account of as much information in the genealogy structure in the data as possible.

4. The Proposed Method

In practice, often the observed data of n sequences are not in the form of a coalescent tree, since the w_j's are unknown; nor does it provide the number of mutations k_i's on the tree segments, but we have the number of segregating sites from the data.

For example the mitochondrial data used in GT is given in Table 1 below. In this case, the method for $\theta_{1,n}$ and $\theta_{2,n}$ can not be directly used. A classical tool is the method of GT.

The distinct sequences in the data are called lineages . Site at which not all the observed sequences have the same base is a segregating site, as shown in the following Table.

{1cm} \centerline{Table 1. about here} {1cm}

Each row of the table represents a DNA sequence lineage. In this data, there are 18 segregating sites, there are transitions but no transversion observed.

Suppose we have the w_{ij}'s and k_{ij}'s, then the model is given by (4) and the estimate of θ is given by the MLE $\theta_{1,n} = 2k/\sum_{i=2}^{n} i w_i$, here k is just the number of segregating sites. We see that given the full information of w_{ij}'s and k_{ij}'s, only $k = \sum_{i=2}^{n}\sum_{j=1}^{i} k_{ij}$ and $\sum_{i=2}^{n} i w_i$ are relevant in the estimation of θ. This suggest a very simple algorithm for estimating θ. As the w_i's are unknown, we generate them by simulation. To be specific, for $m = 1, \ldots, M$, let $w_i^{(m)}$ be the simulated value of w_i at iteration m, we set our estimate of as

$$\theta_n = \frac{1}{M}\sum_{m=1}^{M} \frac{2k}{\sum_{i=2}^{n} i w_i^{(m)}}. \quad (5)$$

The simulation method is described below. Specify a simulation size M (typically $M \ge 10{,}000$), for $m = 1, \ldots, M$, do the following steps

i) Sample $\mathbf{w}^{(m)} = (w_2^{(m)}, \ldots, w_n^{(m)})$ from the coalescent distribution as in (1), i.e., the $w_i^{(m)}$'s are

independent, with $w_i^{(m)} \sim Exponential(i(i-1)/2)$. Or equivalently, sample $u \sim U(0,1)$ and set $w_i^{(m)} = -2/(i(i-1))\ln(1-u)$.

ii) Compute θ_n as in (5).

iii) Compute the estimated variance σ_n^2 of θ_n by Bootstrap, as: set $N \approx M/1000$, for $j = 1, \ldots, N$ compute

$$\theta_{nj} = \frac{1}{N} \sum_{m=(j-1)N+1}^{jN} \frac{2k}{\sum_{i=1}^{n} i w_i^{(m)}}, \quad \sigma_n^2 = \frac{1}{N} \sum_{j=1}^{N} \theta_{nj}^2 - \theta_n^2.$$

Then the estimated standard error of θ_n is σ_n.

Similarly as in the proof of Proposition 2, we have the following

Corollary. First let $M \to \infty$, then $n \to \infty$, we have i) $\theta_n \xrightarrow{a.s.} \theta_0$.

ii) If $\theta_0 > 0$, then $h_n^{1/2}(\theta_n - \theta_0) \xrightarrow{D} N(0, I^{-1}(\theta_0))$.

Note $I^{-1}(\theta) = \theta$, from the above result, the standard deviation (SD) of θ_n is approximated as $SD(\theta_n) \approx h_n^{-1/2}\theta_n^{1/2}$.

5. Mutation Inference

Now we use our method to analyze the data in Table 1, which was taken from part of the data in Ward et al. [23]. They are from a segment of the control region, with 352 base pairs (sites), out of which 159 of them are purine sites and 193 are pyrimidine sites. This data contains 63 sequences sampled from a North American Indian tribe, the Nuu-Chah-Nulth from Vancouver Island. After eliminating sequences with multiple mutations on some single sites, so that the assumption of at most one mutation each site is met. The remaining data has 55 sequences, with 14 distinct lineages. The whole sequences are long, but only the segregating sites are informative for the analysis, the other sites are ignored.

The mentioned full data has k = 18 segregating sites, with n = 55 and is presented in Table 1. We estimate the mutation rate by using all the data, the Purine data (k= 5, n=5) and the Pyrimidine (k = 13, n=5) respectively, and compare the results with those from some other methods.

The results from the proposed method and those from other methods are shown as in Table 2. Numbers are estimated mutation rate **θ**, in parenthesis are estimated SDs when available. {1cm} \centerline{Table 2. about here} {1cm}

We see that the results from ours and those from GT and MM are basically consistent.

6. Concluding Remarks.

We investigated the asymptotic properties of the mutation rate estimation under some common settings, we show the commonly used estimators are consistent and asymptotic normal, but with a very slow rate of $(\log n)^{1/2}$. We proposed a simulation based method of mutation rate estimation. This method implements the prior knowledge of the population genealogy, is simple to use, and yields comparable results with other methods.

Appendix.

Proof of Proposition 1.

i) Recall the w_i 's are independent with $w_i \sim Exponential(i(i-1)/2)$, $E(w_i) = 2/(i(i-1))$, and the k_{ij} 's are independent with $k_{ij}|w_i \sim Po(\cdot, w_i\theta_0/2)$. Let $y_i = i(i-1)w_i/2$, then the y_i 's are i.i.d. with $y_i \sim \exp(1)$ and $E(y_i) = 1$. Let $k_i = \sum_{j=1}^{i} k_{ij}$, then the k_i's are independent with $k_i|w_i \sim Po(\cdot, iw_i\theta_0/2)$ and $E(k_i|w_i) = iw_i\theta/2$. Similarly, let $m_i = (i-1)k_i$, then the m_i 's are independent and identically distributed with

$E(m_i) = (i-1)E[E(k_i|w_i)] = (i-1)E(iw_i\theta_0/2) = \theta_0.$

We have

$$\hat{\theta}_{1,n} = \frac{2\sum_{i=2}^{n}\sum_{j=1}^{i} k_{ij}}{\sum_{i=2}^{n} iw_i} = \frac{\sum_{i=2}^{n}(i-1)^{-1}m_i}{\sum_{i=2}^{n}(i-1)^{-1}y_i} = \frac{\sum_{i=2}^{n} a_{n,i}m_i}{\sum_{i=2}^{n} a_{n,i}y_i},$$

where $a_{n,i} = (i-1)^{-1}/h_n$. Note $\lim_n a_{n,i} = 0$ for each fixed i, and $\sum_{i=2}^{n}|a_{n,i}| = 1 < \infty$, i.e. $\{a_{n,i}\}$ is a Toeplitz sequence. By the strong law of large numbers for weighted sum of i.i.d. random variables (see Bingham 1986, for a review of such results), we have

$$\sum_{i=2}^{n} a_{n,i}(y_i - E(y_i)) \xrightarrow{a.s.} 0, \quad \sum_{i=2}^{n} a_{n,i}(m_i - E(m_i)) \xrightarrow{a.s.} 0.$$

Since $\sum_{i=2}^{n} a_{n,i}E(y_i) = 1$, we have

$$\hat{\theta}_{1,n} = \frac{\sum_{i=2}^{n} a_{n,i}E(m_i) + \sum_{i=2}^{n} a_{n,i}(m_i - E(m_i))}{\sum_{i=2}^{n} a_{n,i}E(y_i) + \sum_{i=2}^{n} a_{n,i}(y_i - E(y_i))}$$

$$\xrightarrow{a.s.} \lim_n \frac{\sum_{i=2}^{n} a_{n,i}E(m_i)}{\sum_{i=2}^{n} a_{n,i}E(y_i)} = \lim_n \frac{\sum_{i=2}^{n} a_{n,i}\theta_0}{\sum_{i=2}^{n} a_{n,i}} = \theta_0.$$

ii) Since $\hat{\theta}_{1,n} = \sum_{i=2}^{n} k_i/(\sum_{i=2}^{n} iw_i/2)$, we have

$$h_n^{1/2}(\hat{\theta}_{1,n} - \theta_0) = \frac{h_n}{\sum_{i=2}^{n} iw_i/2} \frac{\sum_{i=2}^{n}(k_i - iw_i\theta_0/2)}{h_n^{1/2}}.$$

In the proof of i) we have that

$$\frac{h_n}{\sum_{i=2}^{n} iw_i/2} = \frac{1}{\sum_{i=2}^{n} a_{n,i}y_i} \xrightarrow{a.s.} 1,$$

so we only need to show

$$\frac{\sum_{i=2}^{n}(k_i - iw_i\theta_0/2)}{h_n^{1/2}} \xrightarrow{D} N(0, I^{-1}(\theta_0)). \quad (A.1)$$

For this, we only need to check the Lindeberg condition (Feller [6]) for the sum of independent variables

$$\sum_{i=2}^{n}(k_i - iw_i\theta_0/2).$$

Let $x_i = k_i - iw_i\theta_0/2$, $S_n = \sum_{i=2}^{n} x_i$ and $\sigma^2(S_n) = \sum_{i=2}^{n} Var(x_i)$. Recall $k_i|w_i \sim Po(\cdot, iw_i\theta_0/2)$, so

$$E(k_i|w_i) = Var(k_i|w_i) = iw_i\theta_0/2 \text{ and } E(k_i^2|w_i) = Var(k_i|w_i) + E^2(k_i|w_i) = iw_i\theta_0/2 + i^2w_i^2\theta_0^2/4;$$

$w_i \sim \exp(i(i-1)/2)$, so $E(w_i) = 2/(i(i-1))$and $Var(w_i) = 4/(i^2(i-1)^2)$. We have

$$E(x_i) = E[E(x_i|w_i)] = E[iw_i\theta_0/2 - iw_i\theta_0/2] = 0,$$

$$Var(x_i) = E[Var(x_i|w_i)] + Var[E(x_i|w_i)] = E[Var(x_i|w_i)] = E[E(x_i^2|w_i)]$$

$$= E[E((k_i^2 - ik_iw_i\theta_0 + i^2w_i^2\theta^2/4)|w_i)] = E[iw_i\theta_0^2/2] = \theta_0/(i-1).$$

This gives $\sigma^2(S_n) = h_n\theta_0$.

Let $F_i(\cdot)$be the distribution function of x_i, the Lindeberg condition in our case is, for all $\epsilon > 0$,

$$\frac{1}{\sigma^2(S_n)}\sum_{i=2}^{n}\int_{|x|\geq\epsilon\sigma(S_n)} x^2dF_i(x) \to 0.$$

Let $\chi(\cdot)$be the indicator function, then for each fixed i,

$$\int_{|x|\geq\epsilon\sigma(S_n)} x^2dF_i(x) = E(x_i^2\chi(|x_i|\geq \epsilon\theta_0^{1/2}h_n^{1/2}))$$

$$= E[E(x_i^2\chi(|x_i|\geq \epsilon\theta_0^{1/2}h_n^{1/2})|w_i)] := \int Q_{n,i}(w)f_i(w)dw,$$

where $f_i(w)$is the density function of w_i, and $Q_{n,i}(w) = E[x_i^2\chi^2(|x_i|\geq \epsilon\theta_0^{1/2}h_n^{1/2})|w_i = w]$.

Note $\int Q_n(w_i)f(w_i)dw_i$is decreasing in $n$for each fixed i, and in $i$for fixed n, denote $c$as a generic constant $0 < c < \infty$, then for all $n$and i, we have

$$\int Q_{n,i}(w)f_i(w)dw \leq cQ_{n,i}(E(w_i))$$

$$= c\sum_{r\geq iE(w_i)\theta_0/2+\epsilon\theta_0^{1/2}h_n^{1/2}} (r - iE(w_i)\theta_0/2)^2e^{-iE(w_i)\theta_0/2}\frac{(iE(w_i)\theta_0/2)^r}{r!}$$

$$= c\sum_{r\geq\theta_0/(i-1)+\epsilon\theta_0^{1/2}h_n^{1/2}} (r - \theta_0/(i-1))^2e^{-\theta_0/(i-1)}\frac{(\theta_0/(i-1))^r}{r!}$$

$$= c \sum_{r \geq \theta_0/(i-1)+\epsilon\theta_0^{1/2}h_n^{1/2}} \frac{(r-\theta_0/(i-1))^2}{r(r-1)} e^{-\theta_0/(i-1)} \left(\frac{\theta_0}{i-1}\right)^2 \frac{(\theta_0/(i-1))^{r-2}}{(r-2)!}$$

$$\leq c \sum_{r \geq \theta_0/(i-1)+\epsilon\theta_0^{1/2}h_n^{1/2}} \frac{(\theta_0/(i-1))^{r-2}}{(r-2)!} \leq c \sum_{r \geq \epsilon\theta_0^{1/2}h_n^{1/2}} \frac{(\theta_0/(i-1))^{(r-2)}}{(r-2)!}$$

$$\leq c \frac{\theta_0^{(r_n-2)}}{(r_n-2)!} \frac{1}{(i-1)^{(r_n-2)}},$$

where $r_n = [\epsilon\theta_0^{1/2}h_n^{1/2}]$, the integer part of $\epsilon\theta_0^{1/2}h_n^{1/2}$. The above inequality holds for all $i$with some common $0 < c < \infty$. Also, $r_n \to \infty$as $h_n \to \infty$. Since $h_n/\log(n) \to 1$, so for large n, $r_n - 2 > 2$. Also, $\theta_0^{(r_n-2)}/(r_n-2)! \to 0$, thus

$$\frac{1}{\sigma^2(S_n)} \sum_{i=2}^{n} \int_{|x| \geq \epsilon\sigma(S_n)} x^2 dF_i(x) \leq c \frac{1}{\theta_0 h_n} \frac{\theta_0^{(r_n-2)}}{(r_n-2)!} \sum_{i=2}^{n} \frac{1}{(i-1)^{(r_n-2)}}$$

$$\leq c \frac{1}{\theta_0 h_n} \frac{\theta_0^{(r_n-2)}}{(r_n-2)!} \sum_{i=2}^{n} \frac{1}{(i-1)^2} \leq c \frac{1}{\theta_0 h_n} \frac{\theta_0^{(r_n-2)}}{(r_n-2)!} \to 0.$$

Thus by the classical central limit theorem,

$$\frac{\sum_{i=2}^{n} x_i}{\sigma(S_n)} \xrightarrow{D} N(0,1)$$

which is the same as (A.1), as long as we show

$$I(\theta_0) = 1/\theta_0.$$

In fact, it is easy to see

$$L''_n(\theta_0) = -\sum_{i=2}^{n} \sum_{j=1}^{i} \frac{k_{ij}}{\theta_0^2} = -\sum_{i=2}^{n} \frac{k_i}{\theta_0^2}$$

so, by the result in the proof of i), $\lim_n \sum_{i=2}^{n} a_{n,i} m_i \to \theta_0$, hence

$$I(\theta_0) = -\lim_n \frac{L''_n(\theta_0)}{h_n} = \lim_n \frac{\sum_{i=2}^{n} k_i}{h_n \theta_0^2} = \lim_n \frac{\sum_{i=2}^{n} a_{n,i} m_i}{\theta_0^2} = 1/\theta_0, \quad (a.s.)$$

This completes the proof of ii).

Proof of Proposition 2. As $\theta_{2,n}$is not in closed form, the proof here is different from those in Propositions 1 and 3. Note

$$L'_{2,n}(\theta) = \sum_{i=2}^{n} \frac{k_i(i-1)-\theta}{\theta(\theta+i-1)},$$

$$L''_{2,n}(\theta) = \sum_{i=2}^{n} -\frac{k_i}{\theta^2} + \frac{k_i+1}{(\theta+i-1)^2}.$$

Since $L'_{2n}(\theta_{2,n}) = 0$, so $-L'_{2,n}(\theta_0) = L'_{2,n}(\theta_{2,n}) - L'_{2,n}(\theta_0) = L''_{2,n}(\tilde{\theta}_n)(\theta_{2,n}-\theta_0)$, where $\tilde{\theta}_n$is an

immediate value between $\theta_0$and $\theta_{2,n}$. We have

$$\theta_{2,n} - \theta_0 = [-L''_{2,n}(\tilde{\theta}_n)/h_n]^{-1} L'_{2,n}(\theta_0)/h_n.$$

i) We only need to show

$$\lim_n -L''_{2,n}(\tilde{\theta}_n)/h_n > 0, \qquad \lim_n L'_{2,n}(\theta_0)/h_n = 0. (A.2)$$

In fact,

$$-\frac{L''_{2,n}(\tilde{\theta}_n)}{h_n} = \frac{1}{\tilde{\theta}_n^2}\frac{\sum_{i=2}^{n} k_i}{h_n} - \frac{1}{h_n}\sum_{i=2}^{n}\frac{k_i+1}{(\tilde{\theta}_n+i-1)^2}.$$

As in the proof of Proposition 3, i) below, $\sum_{i=2}^{n} k_i/h_n \to \theta_0$(a.s.). By the similar way as in the proof of

Proposition 1, i) $\sum_{i=2}^{n} \frac{k_i+1}{(i-1)^2}$ can be written as a Toeplitz summation times a bounded normalizing constant, and we have

$$\sum_{i=2}^{n}\frac{k_i+1}{(\tilde{\theta}_n+i-1)^2} \le \sum_{i=2}^{n}\frac{k_i+1}{(i-1)^2} \xrightarrow{a.s.} \lim_n \sum_{i=2}^{n}\frac{E(k_i)+1}{(i-1)^2} = \lim_n \sum_{i=2}^{n}\frac{\theta_0/(i-1)+1}{(i-1)^2} < \infty.$$

Since $h_n \to \infty$, so $h_n^{-1}\sum_{i=2}^{n}\frac{k_i+1}{(\tilde{\theta}_i+i-1)^2} \to 0$ by the assumption on $\theta_{2,n}$, we get

$$\lim_n -\frac{L''_{2,n}(\tilde{\theta}_n)}{h_n} \ge \frac{1}{C}\lim_n \frac{\sum_{i=2}^{n} k_i}{h_n} - \lim_n \frac{1}{h_n}\sum_{i=2}^{n}\frac{k_i+1}{(i-1)^2} = \frac{\theta_0}{C} > 0.$$

Also, formulate $L'_{2,n}(\theta_0)$as a Toeplitz summation times a bounded normalizing constant, by the result for weighted summation we have

$$\lim_n \frac{L'_{2,n}(\theta_0)}{h_n} = \lim_n h_n^{-1} \sum_{i=2}^{n} \frac{(i-1)E(k_i) - \theta_0}{\theta_0(\theta_0 + i - 1)} = 0.$$

Thus (A.2) is true.

ii). By i), we have $\tilde{\theta}_n \to \theta_0$(a.s.), and $-L''_{2,n}(\tilde{\theta}_n)/h_n \to 1/\theta_0$(a.s.). So we only need to show

$$\frac{L'_{2,n}(\theta_0)}{h_n^{1/2}} \xrightarrow{D} N(0, \theta_0). \quad (A.3)$$

For this we only need to check Lindeberg condition for

$$S_{2,n} = \sum_{i=2}^{n} \frac{(i-1)k_i - \theta_0}{\theta_0(\theta_0 + i - 1)} := \sum_{i=2}^{n} x_{2,i}.$$

Note $E(x_{2,i}) = 0$all i, and $Var(x_{2,i}) = E(x_{2,i}^2) = 1/(\theta_0(\theta_0 + i - 1))$, we have

$$\sigma^2(S_{2,n}) := \sum_{i=2}^{n} Var(x_{2,i}) = \sum_{i=2}^{n} \frac{1}{\theta_0(\theta_0 + i - 1)} = \frac{1}{\theta_0}\left(\sum_{i=2}^{n} \frac{1}{i-1} - \sum_{i=2}^{n} \frac{\theta_0}{(i-1)(\theta_0 + i - 1)}\right),$$

and $\sigma^2(S_{2,n})/h_n \to 1/\theta_0$. Lindeberg condition is checked the same way as in the proof of Proposition 1, ii), so

$$\frac{S_{2,n}}{\sigma(S_{2,n})} \xrightarrow{D} N(0, 1),$$

which is the same as (A.3).

Proof of Proposition 3. i) As in the proof of Proposition 1, i), recall the notation m_i 's and $a_{n,i}$'s there, we rewrite $\theta_{3,n}$ as

$$\theta_{3,n} = \sum_{i=2}^{n-1} a_{n,i} m_i \xrightarrow{a.s.} \lim_n \sum_{i=2}^{n-1} a_{n,i} E(m_i) = \theta_0.$$

ii) Note $E(k_i) = E[E(k_i|w_i)] = E[iw_i\theta_0/2] = (i-1)^{-1}\theta_0$, so we can rewrite $\theta_{3,n}$as

$$\theta_{3,n} = \frac{\sum_{i=2}^{n} k_i}{h_n} = \frac{\sum_{i=2}^{n} k_i}{\sum_{i=2}^{n} E(k_i)/\theta_0}.$$

Denote $h_{3,n} = h_{3,n}(\theta_0) = h_n + b_n\theta_0$, with $b_n = \sum_{i=2}^{n-1} i^{-2}$. We have

$$h_{3,n}^{1/2}(\theta_{3,n} - \theta_0) = \frac{h_{3,n}^{1/2}}{h_n^{1/2}} \frac{\sum_{i=2}^{n}(k_i - E(k_i))}{h_{3,n}^{1/2}}.$$

$$\frac{\sum_{i=2}^{n}(k_i - E(k_i))}{h_{3,n}^{1/2}} \xrightarrow{D} N(0, \Gamma^{-1}(\theta_0)). \quad (A.3)$$

Since $h_{3,n}/h_n \to 1$, we only need to show

Equivalently, we only need to check the Lindeberg condition for the sum of independent random variables

$$S_{3,n} = \sum_{i=2}^{n}(k_i - E(k_i)) = \sum_{i=2}^{n}(k_i - \theta_0/(i-1)) := \sum_{i=2}^{n} x_{3,i}.$$

Let $\sigma^2(S_{3,n}) = \sum_{i=2}^{n} Var(x_{3,i})$. Note

$$E(x_{3,i}|w_i) = iw_i\theta_0/2 - \theta_0/(i-1), \quad E(x_{3,i}^2|w_i) = iw_i\theta_0/2 + i^2w_i^2\theta_0^2/4 - iw_i\theta_0^2 + \theta_0^2/(i-1)^2$$

$$Var(x_{3,i}|w_i) = E(x_{3,i}^2|w_i) - E^2(x_{3,i}|w_i) = iw_i\theta_0/2.$$

$$E[Var(x_{3,i}|w_i)] = i(\theta_0/2)E(w_i) = \theta_0/(i-1),$$

$$Var(E(x_{3,i}|w_i)) = i^2(\theta_0^2/4)Var(w_i) = \theta_0^2/(i-1)^2,$$

$$\sigma^2(S_{3,n}) = \sum_{i=2}^{n} Var(x_{3,i}) = \sum_{i=2}^{n} E[Var(x_{3,i}|w_i)] + Var[E(x_{3,i}|w_i)] = h_{3,n}\theta_0.$$

The Lindeberg condition is checked the same way as in the proof of Proposition 1, ii), and (A.3) is proved.

ACKNOWLEDGEMENT:

This work is supported in part by the National Center for Research Resources at NIH grant 2G12RR003048, and by the Center for Research on Genomics and Global Health (CRGGH) at NHGRI/NIH.

REFERENCES

1. Bingham, N.H. (1986). Extensions of the strong law. Advances in Applied Probability, special supplement 1986, 27-36. \item Booth, K. and Lueker, G. (1976). Testing for the consecutive ones property, interval graphs, and graph planarity using PQ-tree algorithms. {\it Journal of Computer and System Sciences}, {\bf 13} 335-379. \vspace{-0.3cm} %\item Camin, J. and Sokal, R. (1965). A method for deducing branching sequences in phylogeny. {\it Evolution}, {\bf 19} 311-326. \vspace{-0.3cm}
2. Chandler, J. (2006). Estimating per-locus mutation rates. Journal of Genetic Genealogy, 2, 27-33.
3. Donnelly, P. and Tavaré S. (1995). Coalescents and genealogical structure under neutrality. Annual Review of Genetics, 29, 401-421.
4. Dupuy, B.M. et al (2004). Y-chromosomal microsatellite mutation rates: differences in mutation rate between and within loci. Human Mutation, 23, 117-124.
5. Ethier, S.N. and Griffiths, R.C. (1987). The infinitely-many-sites model as a measure valued diffusion. Annals of Probability, 15 515-545. \item Farris, J.S. (1967). Inferring phylogenetic trees from chromosome inversion data. {\it Systematic Zoology}, {\bf 27} 275-284. \vspace{-0.3cm}
6. Feller, W. (1968). An Introduction to Probability Theory and Its Applications, 3rd ed. New York: Wiley. \item Felsenstein, J. (1982). Numerical methods for inferring evolutionary trees. {\it The Quarterly Review of Biology}, {\bf 57} 379-404. \vspace{-0.3cm}

7. Felsenstein, J. (1992). Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. Genetical Research, 59 139-147.
8. Fu, Y. and Li, W. (1993). Maximum likelihood estimation of population parameters. Genetics, 134 1261-1270.
9. Fu, Y. (1994). A phylogenetic estimator of effective population size or mutation rate. Genetics, 136 685-692. \item Griffiths, R.C. (1987). An algorithm for constructing genealogical trees. Statistics Research Report 163, Dept. Mathematics, Monash Univ., Australia. \vspace{-0.3cm}
10. Griffiths, R.C. and Tavaré S. (1994). Ancestral inference in population genetics. Statistical Science, 9 307-319. \item Griffiths, R.C. and Tavaré S.$\acute{e}$, S. (1995). Unrooted genealogical tree probabilities in the infinitely-many-sites model. {\it Mathematical Biosciences}, {\bf 127} 77-89. \vspace{-0.3cm} %\item Gusfield, D. (1991). Efficient algorithms for inferring evolutionary trees. {\it Networks} {\bf 21} 19-28. \vspace{-0.3cm}
11. Haag-Liautard, C., Coffey, N. Houe, D., Charlesworth, B. et al (2008). Direct estimation of the mitochondrial DNA mutation rates in Drosophila melanogaster. PLoS Biology. 6(8): e204. doi:10.1371/journal.pbio.0060204.
12. Heyer, E. et al (1997). Estimating Y-chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. Human Molecular Genetics, 6, 799-803. \item Hastings, W. (1970). Monte Carlo sampling methods using Markov %chains and their applications. {\it Biometrika}, {\bf 57} 97-109. %\vspace{-0.3cm}
13. Hudson, R.R. (1991). Gene genealogies and the coalescent process, pp. 1-44 in Oxford Surveys in Evolutionary Biology, Eds. D. Futuyma and J. Antonovics. Oxford University Press.
14. Hutchison, L. et al (2004). Direct determination of mutation characteristics of Y chromosome STR loci. Poster presented at the American Society of Human Genetics 2004 Annual Meeting, October 2004, Toronto.
15. Kingman, J.F.C. (1982a). On the genealogy of large populations. Journal of Applied Probability, 19A 27-43.
16. Kingman, J.F.C. (1982b). Exchangeability and the evolution of large populations, pp. 97-112 in Exchangeability in Probability and Statistics, Eds. G. Koch and F. Spizzchino. North-Holland Publishing Company, Amsterdam.
17. Klein, E.K., Austerliz, F., Laredo, C. (1999). Some statistical improvements for estimating population size and mutation rate from segregating sites in DNA sequences, Theoretical Population Biology, 55, 235-247.
18. Kuhner, M.K., Yamato, J., Felsenstein, J. (1998). Maximum likelihood estimation of population growth rates based on the coalescent. Genetics, 149 429-434.
19. Lundstrom, R., Tavaré S., Ward, R.H. (1992). Estimating substitution rates from molecular data using the coalescent. Proceedings of the National Academic of Since, USA, 89 5961-5965.
20. Natchman, M., Crowell, S. (2000). Estimate the mutation rate per nucleotide in humans. Genetics, 156, 297-304. \item Meligkotsidou, L. and Fearnhead, P. (2005). Maximum-likelihood %estimation of coalescent times in genealogical trees. {\it Genetics}, %{\bf 171} 2073-2084. %\vspace{-0.3cm} %\item
21. Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. Teller, %E. (1953). Equations of state calculations by fast computing machines. %{\it Journal of Chemical Physics}, {\bf 21} 1087-1092. %\vspace{-0.3cm} %\item Neumann, J. (1951). Various techniques used in connection with %random digits in Monte Carlo methods. {\it National Bureau Standards}, %{\bf 12} 36-38. %\vspace{-0.3cm} %\item Robert, C.P. and Casella, G. (2004). {\it Monte Carlo Statistical %Methods} (2nd Edition). New York: Springer-Verlag.
22. Stephens, M. and Donnelly, P. (2000). Inference in Molecular Population Genetics. Journal of the Royal Statistical Society, Series B, 62, 605-655 Tavaré S. (1997). Ancestral inference from DNA sequence data. Chapter 5 in Case Studies in Mathematical Modeling: Ecology, Pysiology and Cell Biology, Eds. Othmer HG, Adler FR, Lewis MA and Dallon J, pp 81-96, Prentice-Hall. \item Tavaré S.$\acute{e}$, S., Balding, D.J., Griffiths, R.C., Donnelly, P. (1997). Inferring coalescence times from DNA sequence {\it Genetics}, {\bf 145} 505-518.\vspace{-0.3cm}
23. Ward, R.H., Frazier, B.L., Dew, K., Pääbo, S. (1991). Extensive mitochondrial diversity within a single Amerindian tribe. Proc. Nat. Acad. Sci. U.S.A., 88 8720-8724.
24. Watterson, G.A. (1975). On the number of segregating sites in genetical models without recombination. Theory of Population Biology, 7 256-276. \item Watterson, G.A. (1982). Mutant substitutions at linked nucleotide %sites. {\it Advances in Applied Probability}, {\bf 14} 206-224.

Tables and Figures

Table 1. Nucleotide position in control region

Site	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	Linkage
	Purines					Pyrimidines													freqs
Linkage																			
a	A	G	G	A	A	T	C	C	T	C	T	T	C	T	C	T	T	C	2
b	A	G	G	A	A	T	C	C	T	T	T	T	C	T	C	T	T	C	2
c	G	A	G	G	A	C	C	C	T	C	T	T	C	C	C	T	T	T	1
d	G	G	A	G	A	C	C	C	C	C	T	T	C	C	C	T	T	C	3
e	G	G	G	A	A	T	C	C	T	C	T	T	C	T	C	T	T	C	19
f	G	G	G	A	G	T	C	C	T	C	T	T	C	T	C	T	T	C	1
g	G	G	G	G	A	C	C	C	T	C	C	C	C	C	C	T	T	T	1
h	G	G	G	G	A	C	C	C	T	C	C	C	T	C	C	T	T	T	1
i	G	G	G	G	A	C	C	C	T	C	T	T	C	C	C	C	C	T	4
j	G	G	G	G	A	C	C	C	T	C	T	T	C	C	C	C	T	T	8
k	G	G	G	G	A	C	C	C	T	C	T	T	C	C	C	T	T	C	5
l	G	G	G	G	A	C	C	C	T	C	T	T	C	C	C	T	T	T	4
m	G	G	G	G	A	C	C	T	T	C	T	T	C	C	C	T	T	C	3
n	G	G	G	G	A	C	T	C	T	C	T	T	C	C	T	T	T	C	1

Table 2. Comparison of Mutation Estimates

Data	GT	MM	IS	Ours
Purine & Pyrimidine	4.80(1.48)	3.93		4.196(0.351)
Pyrimidine	3.31(1.14)	2.84	4.63	3.041(0.254)
Purine	1.22(0.61)	1.09	1.27	1.170(0.103)

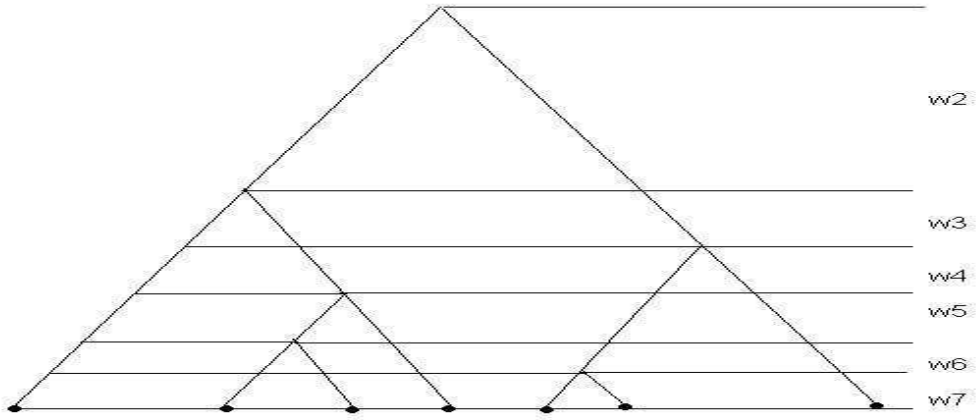

Figure 1: Coalescent tree for a sample of seven individuals.