Statistics in Molecular Biology
IMS Lecture Notes – Monograph Series (1999) Volume 33

LINEAR ESTIMATORS FOR THE EVOLUTION OF TRANSPOSABLE ELEMENTS

BY PAUL JOYCE[1], LINETTE FOX, N. CAROL CASAVANT, AND HOLLY A. WICHMAN

Department of Mathematics, Division of Statistics, University of Idaho, Department of Zoology and Genetics, Iowa State University and Department of Biological Sciences, University of Idaho

Pairwise differences and segregating sites are two measures of sequence divergence often used to estimate the rate of evolution. There are other measures of sequence divergence. Which method is most appropriate? Motivated by a study of the evolution of transposable elements, we develop a new and more precise method for estimating the rate of evolution. We apply our method to LINE-1 data from Casavant *et al.* (1996).

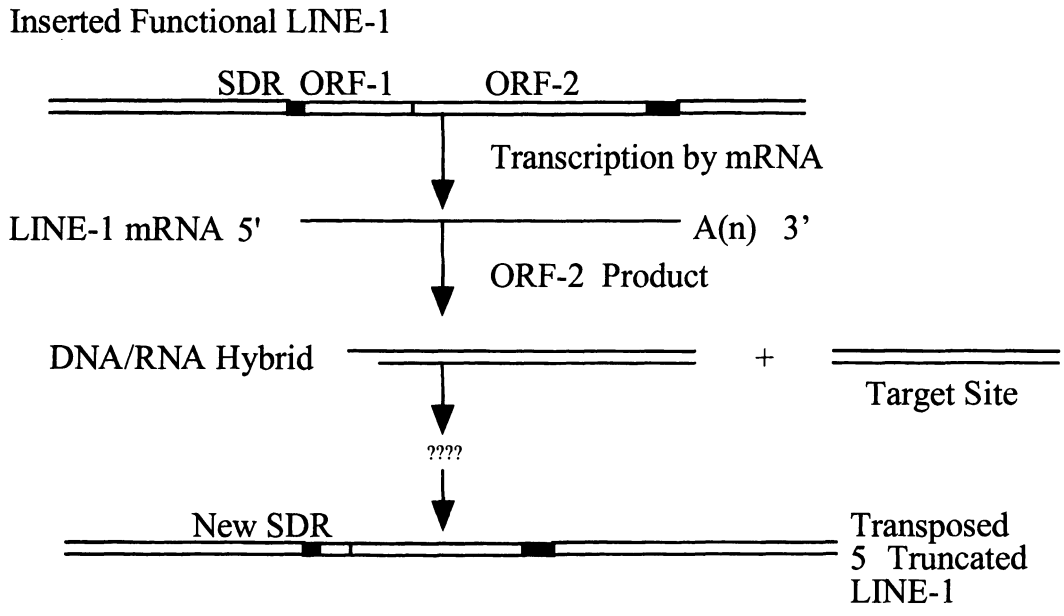
1. Introduction. The motivation for this paper grew out of a very curious discovery made while analyzing some DNA sequence data. See Fox (1997). Under certain model assumptions for the evolution of transposable elements (mobile repetitive DNA found dispersed throughout the genome), we considered two simple measures of sequence divergence to estimate the rate of evolution. One estimate was based on the number of pairwise differences and the other was based on the number of segregating sites. We showed that the estimator based on the number of pairwise differences was inconsistent (variance of the estimator does not go to zero), while the segregating sites method was consistent. This is not at all surprising. The same story is true for the well studied neutral coalescent model, see Watterson (1975) and Tajima (1983). However, not only was it demonstrated that the segregating sites estimator is consistent under our model assumptions, but the variance of the estimator goes to zero like $1/n$, where n is the sample size. For evolutionary models involving DNA data it is unusual for estimators to have such good asymptotic properties.

The most curious discovery came when we applied each method to data. We found that in most cases of biological relevance, the pairwise difference estimator actually outperformed the segregating sites estimator. In this paper we resolve this apparent paradox.

We begin with a brief description of the relevant biology followed by the assumptions of the single master model used to analyze the data. We then

[1]This research is supported by the National Science Foundation grant 96-26764.
AMS 1991 *subject classifications.* Primary 62F05; secondary 60G42, 60F15.
Key words and phrases. Coalescent, transposable elements, best linear unbiased estimator (BLUE).

FIG. 1. *LINE retrotransposition.*



discuss pairwise difference and segregating sites estimators in the context of our model. The properties of both estimators are best understood when considered as members of a class of estimators called linear unbiased. We thus develop the theory of linear unbiased estimation in this context. Finally, we apply our new method to data. Fu (1994) also considered unbiased estimation in the context of the neutral coalescent model.

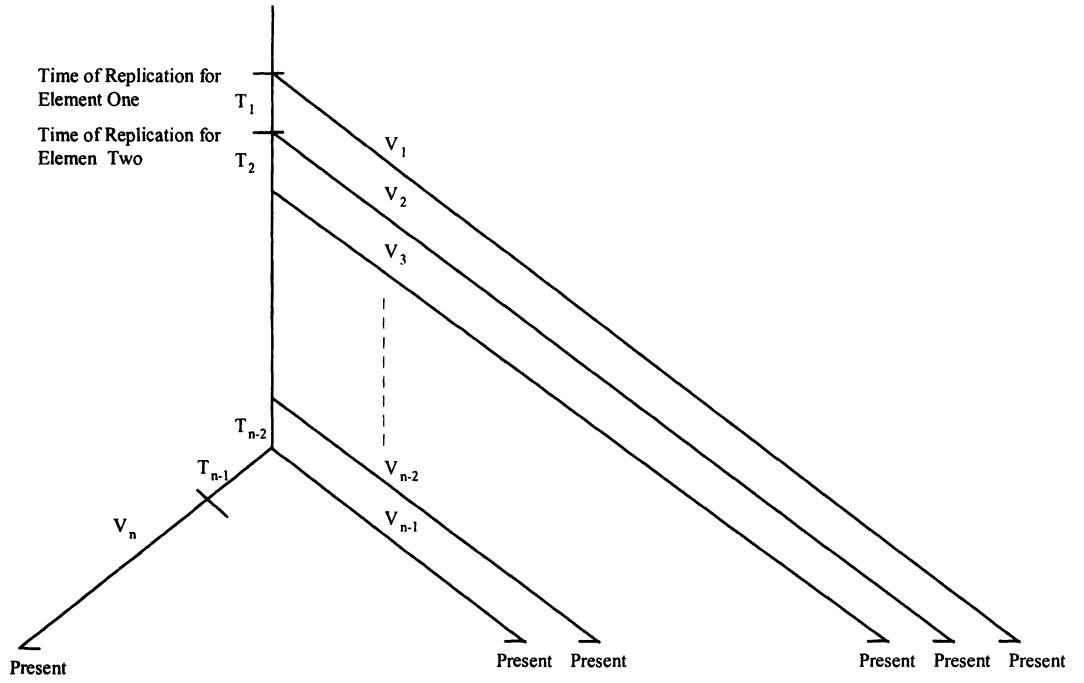
2. Mobile repetitive DNA. Mobile repetitive DNA sequences are found dispersed throughout the genome. LINEs (Long interspersed nuclear elements) have the capability of copying and inserting the copy into the genome at some other site by a process called retrotransposition. In this process DNA encodes RNA, RNA uses reverse transcription to code complementary DNA, and the DNA is integrated back into the chromosome at a different site in the genome. Figure 1 demonstrates one possible method of retrotransposition for LINEs. The functional LINE-1 with two open reading frames (ORF-1 and ORF-2) is transcribed by mRNA. The short direct repeats along the chromosomal DNA are shown by the filled in boxes and the open reading frames of the LINE-1 element by the open boxes. After transcription, RNA, which is shown as a single line, encodes a complementary DNA by reverse transcription. The reverse transcriptase protein in the ORF-2 region catalyzes the reverse transcription of

the RNA to form cDNA. The DNA/RNA hybrid integrates with the target site. The first open reading frame codes for a binding protein. This binding protein binds the DNA to the integration site. Retrotransposition usually produces a 5′ truncated LINE-1. At integration, the RNA is detached. When the RNA detaches, the DNA often folds back on itself and primes second strand synthesis. The loop formed by this synthesis is broken to allow cDNA to synthesize to the chromosomal DNA. The result is a 5′ truncation. Once integrated, the cDNA is ligated to the chromosomal DNA at the 3′ end. Repair synthesis builds the second strand of DNA on the homologous chromosome.

2.1. *Master copy model.* The Master Copy Model assumes that one or a few elements in the genome have the capacity to replicate and all other elements are pseudogenes. A version of this model dates back to Kaplan and Hudson (1989). They developed an equilibrium master copy model where the number of copies reaches equilibrium due to the balance between duplication and deletion. They showed their model was consistent with the *Alu* divergence data. Recent studies of *Alu* and other SINE (short interspersed nuclear elements) families is consistent with a transient master copy model (Tachida 1993, 1996) that considers successive waves of expansion. We are interested in master copy models that are consistent with data from LINE-1 families in mammalian genomes, in particular the deer mouse, *Peromyscus* (Casavant *et al.* 1996). Unlike the *Alu* SINE data, there is evidence of fairly young LINE-1. This indicates that the LINE-1 elements under study may be in the midst of an expansion period. The purpose is to develop statistical methodology that can be applied to master copy data.

A simple mathematical description of the master copy model is the following. Consider a population of elements that is generated by a single master element giving birth at a constant rate. After a fixed time t has evolved, sample n individuals at random. The rooted tree (rooted at the time of the first offspring) describing the relationship between individuals in the sample, will have one main branch with all of the offspring branches emanating from the main branch (see Figure 2) The expected age of a randomly chosen element is $t/2$. Conditional on the tree, place marks along the branches according to a Poisson process of rate θ_m. The marks represent observed mutations that have occurred over evolutionary time. The Poisson process is independent along each branch. Count the number of marks on each of the branches. Assume the tree topology together with the number of marks on each branch is observable. Based on this observation, the problem is to estimate the parameter $\theta = \theta_m t$, the mean number of marks accumulated by a randomly chosen element. We rescale time so that the (rescaled) age of the master element is 1.

FIG. 2. *Element replication.*



Let V_i be the age of the ith oldest replicate in a sample of size n. The difference between the age of element i, V_i, and the next youngest element, V_{i-1}, is denoted by T_i. The age of the youngest element in the sample is V_{n-1} and $V_{n-1} = T_{n-1}$. Let P_i be the number of private mutations accumulated by the ith element over the time period V_i. Let S_i be the number of shared variants between the ith and $i-1$th element accumualted over the time period T_i. Under the assumption of constant rate of element replication, the age of a randomly chosen element is uniformly distributed. So the ages of the elements are distributed according to the order statistics of the uniform. However, the master element will not appear in the sample. For this reason, we cannot determine which among the two youngest elements in the sample is indeed the youngest. This means that the coalescent time T_{n-1} is on average $2/(n+1)$, where as all other coalescent times are $E(T_i) = 1/(n+1)$ for $i < n-1$.

It will be convenient to use vector notation. Vectors will be column vectors unless superscripted by $'$ for transpose: thus we write

$$\mathbf{T}' = (T_1, T_2, \ldots, T_{n-1})$$

and similarly

$$\mathbf{B}' = (V_1, V_2, \cdots, V_{n-1}, T_1, T_2, \cdots, T_{n-1}).$$

It is convenient to record the private and shared mutations in the following order

$$\mathbf{S}' = (P_1, P_2, \cdots, P_{n-1}, S_1, S_2, \cdots, S_{n-2}, P_n).$$

Let $\boldsymbol{\mu} = E(\mathbf{T})$. We denote the variance of the $(n-1) \times 1$ vector $\mathbf{T}$ by the $(n-1) \times (n-1)$ matrix, $\boldsymbol{\Sigma} = \mathrm{Var}\mathbf{T}$, where $(\mathrm{Var}\mathbf{T})_{ij} = \mathrm{Cov}(T_i, T_j)$.

The branch lengths of a tree are related to the coalescent times, in that each branch is a sum of coalescent times. We may view the vector of branch lengths as a linear transformation of the coalescent times. If there are $n-1$ coalescent times, there will be $2(n-1)$ branches to the tree. As one traces the ancestry of the individuals, each coalescence introduces two new branches.

The linear transformation between coalescent times and branch lengths is given by an $(2n-2) \times (n-1)$ matrix $\mathbf{c}$, where the entries of $\mathbf{c}$ are $c_{ij} = 1$ or 0,

$$\mathbf{cT} = \mathbf{B}.$$

For the master locus model, relationship between branch lengths and coalescent times is given by

$$\mathbf{c} = \begin{pmatrix} 1\,1 \cdots 1 \\ 0\,1 \cdots 1 \\ \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot \\ 0\,0 \cdots 1 \\ 1\,0 \cdots 0 \\ 0\,1 \cdots 0 \\ \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot \\ 0\,0 \cdots 1 \end{pmatrix}$$

where $\mathbf{cT} = \mathbf{B}$,

$$\boldsymbol{\mu}' = \left(\tfrac{1}{n+1}, \tfrac{1}{n+1}, \cdot, \cdot \tfrac{1}{n+1}, \tfrac{2}{n+1}\right)$$

and

$$\Sigma = \begin{pmatrix} \frac{n}{(n+1)^2(n+2)} & \frac{-1}{(n+1)^2(n+2)} & \cdots & \frac{-2}{(n+1)^2(n+2)} \\ \frac{-1}{(n+1)^2(n+2)} & \frac{n}{(n+1)^2(n+2)} & \cdots & \frac{-2}{(n+1)^2(n+2)} \\ \vdots & \vdots & & \vdots \\ \frac{-2}{(n+1)^2(n+2)} & \frac{-2}{(n+1)^2(n+2)} & \cdots & \frac{2(n-1)}{(n+1)^2(n+2)} \end{pmatrix}$$

3. Pairwise differences versus segregating sites. Using the tree in Figure 2 one can formulate an unbiased estimate of the mean number of mutations on a randomly selected element using the segregating sites method. Private mutations P_i occur along the elements over the age of the element according to a Poisson process, and shared mutations S_i occur along the master between replication events. The total number of segregating sites S is

$$S = \sum_{i=1}^{n-2} S_i + \sum_{i=1}^{n} P_i.$$

If θ_m is the mutation rate, then an unbiased estimator for the parameter $\theta = \theta_m t$ can be calculated to be

$$\hat{\theta}_{ss} = \frac{2(n-1)}{n^2+3n-2} S.$$

We omit the details. The variance of the unbiased estimate using the segregating sites estimator is given as

$$\mathrm{Var}(\hat{\theta}_{ss}) = \frac{n^4+4n^3+5n^2+26n-24}{3(n^2+3n-2)^2(n+2)}\theta^2 + \frac{2(n+1)}{n^2+3n-2}\theta.$$

When time is distributed uniformly, the times between replication events are not independent, which complicates the variance. For a detailed derivation of this formula see Fox (1997). An asymptotic formula with only the leading terms is easier to absorb

$$\mathrm{Var}(\hat{\theta}_{ss}) \approx \frac{1}{3n}\theta^2 + \frac{2}{n}\theta.$$

For large values of n (relative to mean number of mutations accumulated by a randomly chosen element) the segregating sites method produces a small variance and is consistent.

The pairwise difference estimator sums the mutations for all pairs of elements. The sum of the pairwise differences, D, can be written as

$$D = (n-1)\sum_{i=1}^{n} P_i + \sum_{i=1}^{n-1} i(n-i)S_i$$

The pairwise difference estimator counts some mutations more often than others. An example illustrates the higher weights given to some mutations. If ten elements are sampled from the population and sorted by their times of replication, shared mutations that occur between the time of the replication of the 5th element and the replication of the 6th element are counted 25 times. Moreover, shared mutations between the replication of the ninth and tenth elements are counted nine times. Thus, the shared mutations between the 5th and 6th elements are counted more often and given a higher weight.

An unbiased estimator of θ based on the mean number of pairwise differences is given by

$$\hat{\theta}_{pd} = \frac{3}{4}\bar{D}.$$

The variance for the pairwise differences unbiased estimator is given by the following equation

$$\begin{aligned}\mathrm{Var}(\hat{\theta}_{pd}) &= \left(\frac{(2n-1)(2n^3+15n^2-41n+6)}{20n(n-1)(n+2)(n+1)^2} + \frac{9n^3}{4n^2(n-1)^2(n+1)^2(n+2)}\right)\theta^2 \\ &\quad + \left(\frac{9}{8n} + \frac{3(n^2+1)}{40n(n-1)}\right)\theta.\end{aligned}$$

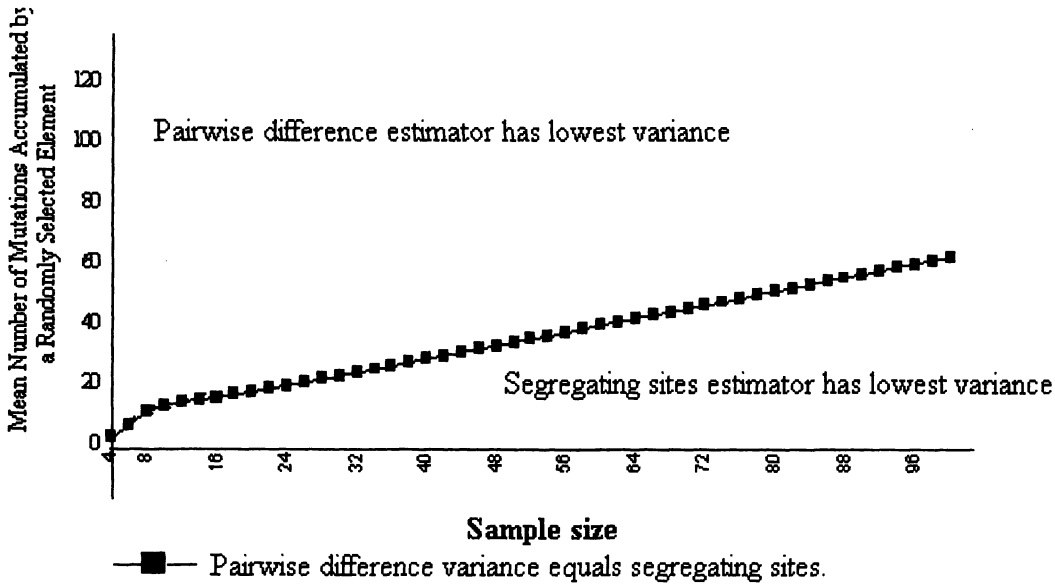
An asymptotic formula including only leading terms may be easier to absorb

$$\mathrm{Var}(\hat{\theta}_{pd}) \approx \frac{1}{5n}\theta^2 + \frac{3}{40}\theta.$$

Note that the variance of the pairwise difference estimator can never be smaller than $3\theta/40$, regardless of the sample size.

We introduced two familiar methods for estimating the parameter θ. One was based on the number of pairwise differences and the other on the number of segregating sites. It is not surprising to learn that the estimator based on

FIG. 3. *Pairwise differences vs. segregating sites.*



pairwise differences is inconsistent (that is, the error of the estimate does not go to zero as the sample size increases). One familiar with these types of problems may also expect that the estimate of θ based on the segregating sites is consistent. However, unlike the results of standard coalescent models,

1. the asymptotic properties of the segregating sites estimator are quite good

$$\mathrm{Var}(\hat{\theta}_{ss}) \to 0$$

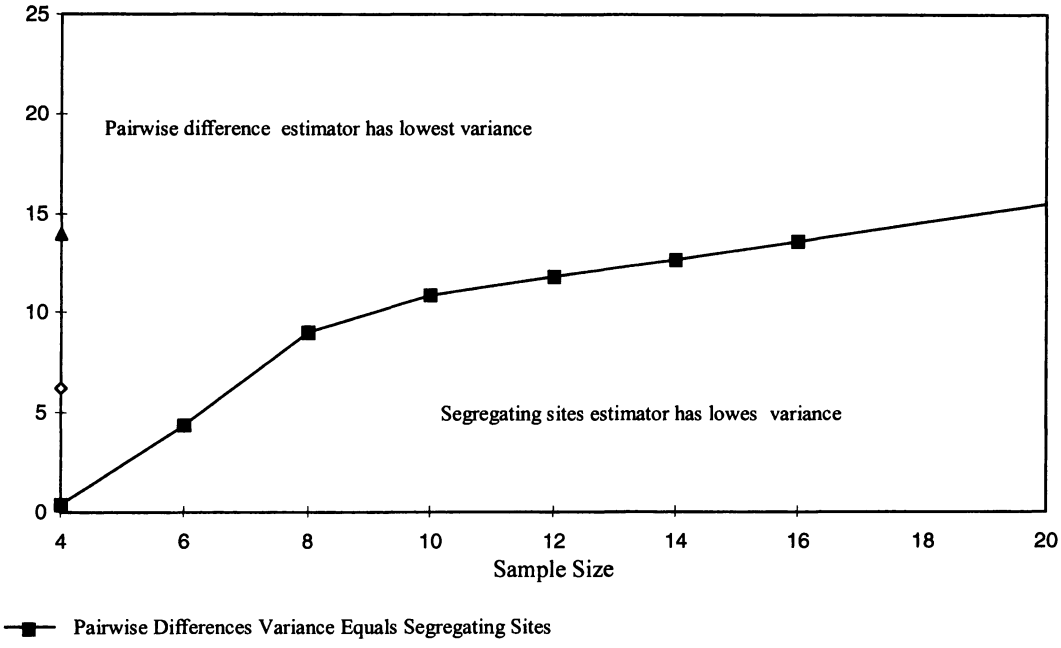
as $n \to \infty$ like $1/n$

2. for sample sizes and parameter values of practical interest, the inconsistent estimator based on pairwise differences actually outperforms the estimator based on segregating sites. All points above the line in the graphs below (Figures 3, 4) represent (n, θ) where the pairwise difference method outperforms segregating sites.

Resolving the apparent contradiction between statements (1) and (2) leads to the theory of linear unbiased estimators, and a new and better estimation procedure.

3.1. *Analyzing the sources of error.* If $\hat{\theta}$ is any estimate of the parameter θ, then we use the following conditioning argument to calculate the variance of $\hat{\theta}$

FIG. 4. *Pairwise differences vs. segregating sites (small sample size).*



given by

$$\mathrm{Var}(\hat{\theta}) = \mathrm{Var}(E(\hat{\theta}|\mathbf{B})) + E(\mathrm{Var}(\hat{\theta}|\mathbf{B})).$$

The above variance formula is central to understanding linear unbiased estimation. There are two sources of error in any estimate of θ. The first source is due to the stochastic nature of the coalescent process. That is, coalescent times are random. The second is due to the stochastic nature of the mutation process, that is, mutations accumulate over time on each branch according to a Poisson process.

The first source of error is measured by the quantity

$$\mathrm{Var}[E(\hat{\theta}|\mathbf{B})].$$

We refer to this as the **coalescent error**, because it is due to the stochastic nature of coalescent. The second source is measured by the quantity

$$E[\mathrm{Var}(\hat{\theta}|\mathbf{B})].$$

We refer to this as the **Poisson error**.

We will see that there is always a trade off between the two sources of error. If one chooses an estimate that has small coalescent error, it is likely to have larger Poisson error.

TABLE 1

Comparing the variance for segregating sites, pairwise differences, and CBLUE. The entries are the coalescent errors (CE), Poisson errors (PE) and the variances of the estimate (V).

		Segregating Sites			Pairwise Differences			CBLUE		
θ	Sample Size	CE	PE	V	CE	PE	V	CE	PE	V
9	9	2.20	1.70	3.90	2.03	1.89	3.93	0.82	2.62	3.43
17	6	11.00	4.57	15.58	10.29	4.76	15.05	6.02	6.04	12.06
18	14	6.14	2.28	8.43	5.26	2.91	8.17	1.45	4.23	5.68
15	7	7.52	3.53	11.04	7.06	3.75	10.81	3.57	4.93	8.50

It can be shown that the segregating sites estimator has the smallest Poisson error of all linear unbiased estimators. However, it has larger coalescent error than the estimator based on pairwise differences. Table 1 below shows that the coalescent error is often the dominant term for the variance of the estimate. Our new estimation procedure minimizes the coalescent error and thus outperforms pairwise differences and segregating sites in most cases. Our preliminary results show that this new method compares favorably to maximum likelihood.

4. Coalescent best linear unbiased estimator (CBLUE). The purpose of this section is to consider unbiased estimators of θ that can be described as linear combinations of the observed changes on the branches. Let $\mathbf{x}$ be an $(2n-2) \times 1$ vector of weights. We consider estimates of θ of the form

$$\hat{\theta} = \sum_{i=1}^{2n-2} x_i S_i = \mathbf{x}'\mathbf{S}.$$

If we assume that $\hat{\theta}$ is an unbiased estimator, then we have a linear constraint on the set of possible weight vectors $\mathbf{x}$. Note that

$$E(\hat{\theta}) = E(\mathbf{x}'\mathbf{S}) = \theta E(\mathbf{x}'\mathbf{B}) = \theta E(\mathbf{x}'\mathbf{c}\mathbf{T}) = \theta \mathbf{x}'\mathbf{c}\boldsymbol{\mu}.$$

Therefore, if $\hat{\theta}$ is unbiased, then $E(\hat{\theta}) = \theta$ implies $\mathbf{x}'\mathbf{c}\boldsymbol{\mu} = 1$. We now calculate

the variance of $\hat{\theta}$ as

$$\begin{aligned}
\mathrm{Var}(\hat{\theta}) &= E[\mathrm{Var}(\hat{\theta}|\mathbf{B})] + \mathrm{Var}[E(\hat{\theta}|\mathbf{B})] \\
&= E[\mathrm{Var}(\mathbf{x}'\mathbf{S}|\mathbf{B})] + \mathrm{Var}[E(\mathbf{x}'\mathbf{S}|\mathbf{B})] \\
&= E[\mathbf{x}'\mathrm{Var}(\mathbf{S}|\mathbf{B})\mathbf{x}] + \mathrm{Var}[\mathbf{x}'E(\mathbf{S}|\mathbf{B})] \\
&= \theta[\mathbf{x}'E(\mathrm{diag}(\mathbf{B}))\mathbf{x}] + \theta^2\mathrm{Var}(\mathbf{x}'\mathbf{B}) \\
&= \theta[\mathbf{x}'E(\mathrm{diag}(\mathbf{cT}))\mathbf{x}] + \theta^2\mathrm{Var}(\mathbf{x}'\mathbf{cT}) \\
&= \theta[\mathbf{x}'E(\mathrm{diag}(\mathbf{cT}))\mathbf{x}] + \theta^2[\mathbf{x}'\mathbf{c}\mathrm{Var}(\mathbf{T})\mathbf{c}'\mathbf{x}]
\end{aligned}$$

where $\mathrm{diag}(\mathbf{B})$ is a diagonal matrix with $(\mathrm{diag}(\mathbf{B}))_{ii} = B_i$ and $\mathrm{diag}(\mathbf{B})_{ij} = 0$ for $i \neq j$.

Let $\mathbf{y} = \mathbf{c}'\mathbf{x}$ and let $\mathbf{M} = E(\mathrm{diag}(\mathbf{cT}))$ then we can write

$$\mathrm{Var}(\hat{\theta}) = \theta[\mathbf{x}'\mathbf{M}\mathbf{x}] + \theta^2[\mathbf{y}'\boldsymbol{\Sigma}\mathbf{y}]. \tag{4.1}$$

Since the estimator must be unbiased we have the constraint $\boldsymbol{\mu}'\mathbf{y} = 1$. An estimator that minimizes the coalescent error is found by minimizing the quadratic form $\mathbf{y}'\boldsymbol{\Sigma}\mathbf{y}$, subject to the linear constraint $\boldsymbol{\mu}'\mathbf{y} = 1$. Thus there will be exactly one $\mathbf{y}$ that makes the coalescent error minimum. Notice that $\mathbf{y} = \mathbf{c}'\mathbf{x}$ is a system of $n-1$ linear equations with $2n-2$ unknowns. Because there are typically many solutions, there will be many choices of $\mathbf{x}$ that minimize the coalescent error. We then pick among these choices the one that minimizes the Poisson error. We call this the coalescent best estimators.

Definition. Let $\mathcal{C} = \{\hat{\theta} = \mathbf{x}'\mathbf{S} \mid \mathbf{x}'\mathbf{c}\boldsymbol{\mu} = 1, \ \mathrm{Var}(E(\hat{\theta}|\mathbf{B})) \leq \mathrm{Var}(E(\tilde{\theta}|\mathbf{B}))$ for all unbiased linear estimators $\tilde{\theta}\}$. $\hat{\theta}_c$ is the coalescent best linear unbiased estimator (CBLUE) if

1. $\hat{\theta}_c \in \mathcal{C}$
2. $E(\mathrm{Var}(\hat{\theta}_c|\mathbf{B})) \leq E(\mathrm{Var}(\tilde{\theta}|\mathbf{B}))$ for all $\tilde{\theta} \in \mathcal{C}$.

Lemma 1. *The collection of unbiased estimators that minimize the coalescent error $\mathcal{C}$ is a linear subspace of $\mathbb{R}^{2n-2}$ given by $\mathcal{C} = \{\hat{\theta} = \mathbf{x}'\mathbf{S} \mid \mathbf{x}$ is a solution to the following linear system $\dfrac{\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}{\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}} = \mathbf{c}'\mathbf{x}\}$, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean and variance of the coalescent times.*

Proof. It follows from equation (4.1) that if $\mathbf{y}$ minimizes the quadratic form $\mathbf{y}'\boldsymbol{\Sigma}\mathbf{y}$ subject to the linear constraint, $\boldsymbol{\mu}'\mathbf{y} = 1$, then any solution to the linear equations $\mathbf{y} = \mathbf{c}'\mathbf{x}$ will produce a $\hat{\theta}$ in $\mathcal{C}$. We need only to show that the solution to the minimization problem is $\mathbf{y} = \dfrac{\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}{\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}$. We use the method of Lagrange multipliers. Define

$$g(\mathbf{y}) = \mathbf{y}'\boldsymbol{\Sigma}\mathbf{y} - \lambda(\mathbf{y}'\boldsymbol{\mu} - 1).$$

Then the derivative of g is given by

$$\frac{dg}{d\mathbf{y}}(\mathbf{y}) = 2\boldsymbol{\Sigma}\mathbf{y} - \lambda\boldsymbol{\mu}.$$

Setting the derivative equal to zero and solving for the critical number gives

$$\mathbf{y} = \frac{\lambda}{2}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}.$$

Since $\boldsymbol{\mu}'\mathbf{y} = 1$ implies $\frac{\lambda}{2}\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = 1$. Therefore, $\lambda/2 = 1/(\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})$, implying

$$\mathbf{y} = \frac{\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}{\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}.$$

This completes the proof since $\mathbf{y} = \mathbf{c}'\mathbf{x}$ by definition. ∎

Theorem 1. *There exists a unique CBLUE, $\hat{\theta}_c$, for θ. If $\mathbf{T}$ is the vector of coalescent times, with $E(\mathbf{T}) = \boldsymbol{\mu}$, $\mathrm{Var}(\mathbf{T}) = \boldsymbol{\Sigma}$ and $E(\mathrm{diag}(\mathbf{cT})) = \mathbf{M}$ then the CBLUE estimate is given by $\hat{\theta}_c = \mathbf{S}'\mathbf{x}$, where*

$$\mathbf{x} = \mathbf{M}^{-1}\mathbf{c}\left(\mathbf{c}'\mathbf{M}^{-1}\mathbf{c}\right)^{-1}\frac{\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}{\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}. \tag{4.2}$$

Proof. It follows from the Lemma 1 and equation (4.1) that the CBLUE will be the $\mathbf{x}$ that minimizes the quadratic form $\mathbf{x}'\mathbf{M}\mathbf{x}$, subject to $n-1$ linear constraints induced by the linear equation $\mathbf{y} = \mathbf{c}'\mathbf{x}$, where $\mathbf{y} = \dfrac{\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}{\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}$. Again we use the method of Lagrange multipliers. This time the Lagrange multiplier is a $(n-1)\times 1$ vector denoted by $\boldsymbol{\lambda}$. Define

$$h(\mathbf{x}) = \mathbf{x}'\mathbf{M}\mathbf{x} - (\mathbf{x}'c - \mathbf{y}')\boldsymbol{\lambda}.$$

then the derivative of h is given by

$$\frac{dh}{d\mathbf{x}}(\mathbf{x}) = 2\mathbf{M}\mathbf{x} - \mathbf{c}\boldsymbol{\lambda}$$

TABLE 2
Estimates and standard deviations of Peromyscus data.

Estimation Method	*Peromyscus* Species	Sample Size	Mean Number of Mutations Per Element Estimate	Estimated Standard Deviation
Segregating sites	*P. californicus*	9	8.68	1.92
	P. maniculatus	6	16.42	3.83
Pairwise differences	*P. californicus*	9	8.63	1.92
	P. maniculatus	6	17.35	3.95
CBLUE	*P. californicus*	9	7.90	1.67
	P. maniculatus	6	18.08	2.89

Setting the derivative equal to zero and solving for $\mathbf{x}$ gives

$$\mathbf{x} = \frac{1}{2}\mathbf{M}^{-1}\mathbf{c}\boldsymbol{\lambda}.$$

Substituting the above $\mathbf{x}$ into the linear constraint $\mathbf{y} = \mathbf{c}'\mathbf{x}$ and solving for $\boldsymbol{\lambda}/2$ gives

$$\frac{1}{2}\boldsymbol{\lambda} = (\mathbf{c}'\mathbf{M}^{-1}\mathbf{c})^{-1}\mathbf{y}$$

which implies

$$\mathbf{x} = \mathbf{M}^{-1}\mathbf{c}(\mathbf{c}'\mathbf{M}^{-1}\mathbf{c})^{-1}\mathbf{y}. \blacksquare$$

We applied the CBLUE estimator given by $\hat{\theta}_c = \mathbf{S}'\mathbf{x}$ to two lineages of LINE-1 for two species of *Peromyscus* (deer mouse). The results are given in Table 2 below. Note that CBLUE method is a significant improvement over pairwise differences and segregating sites. A more complete analysis of the data can be found in Joyce *et al.*, in preparation.

5. Conclusion. The master locus model can be used to estimate rates of evolution and make comparisons for LINE-1 data. However, using traditional measures of sequence divergence to estimate the evolutionary parameters leads to the following puzzling conclusion. While the method of pairwise differences leads to an inconsistent estimator, it is in many cases more precise than the consistent estimator based on the number of segregating sites.

The puzzle is resolved when one realizes that the variance of a linear estimator is of the form

$$a_n\theta^2 + b_n\theta.$$

If θ is relatively large, then the $a_n\theta^2$ term will dominate the error in the estimate. This term is smaller for pairwise differences than for segregating sites.

By choosing an estimator that minimizes the dominant term of the error, one can often improve on both methods. This estimator is called the coalescent best linear unbiased estimator CBLUE.

While the CBLUE estimator was demonstrated for the master locus model, it applies in a more general setting. However, it can be shown that there does not exist a uniformly best linear unbiased estimator.

Acknowledgements. The authors would like to thank the American Mathematical Society for sponsoring the meeting 'Statistics in Molecular Biology,' at the University of Washington, Seattle in June 1997. A special thanks goes to Dr. Françoise Seillier-Moiseiwitsch for all her efforts in organizing the conference and editing this proceedings.

REFERENCES

CASAVANT, N. C., SHERMAN, A. N. and WICHMAN, H. A. (1996). Two Persistent LINE-1 Lineages in Peromyscus Have Unequal Rates of Evolution. *Genetics* **142** 1289–1298.

FU, Y.-X. (1994). A Phylogenetic Estimator of Effective Population Size or Mutation Rate. *Genetics* **136** 685–692.

FOX, L. (1997). Statistical Methods for Analyzing Transposable Elements. Unpublished Masters Thesis. University of Idaho.

JOYCE, P., FOX, L., CASAVANT, N. C., WICHMAN, H. A. and FOSTER, J. Statistical models and methods for LINE-1 evolution, in preparation.

KAPLAN, N. L. and HUDSON, R. R. (1989). An Evolutionary Model for Highly Repeated Interspersed DNA Sequences. In *Mathematical Evolutionary Theory*, Feldmen, M. W. (ed). Princeton University Press, New Jersey, 301–314.

TACHIDA, H. (1996). A Population Genetic Study of the Evolution of SINEs II Sequence Evolution Under the Master Copy Model. *Genetics* **143** 1033–1042.

WATTERSON, G. A. (1975). On the Number of Segregating Sites in Genetical Models Without Recombination. *Theoretical Population Biology* **7** 256–276.

DIVISION OF STATISTICS
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF IDAHO
MOSCOW, ID 83844-1103
JOYCE@UIDAHO.EDU
LFOX@PSG.HS.EDU

DEPARTMENT OF PLANT PATHOLOGY
IOWA STATE UNIVERSITY
AMES, IA 50011
CASAVANT@IASTATE.EDU

DEPARTMENT OF BIOLOGICAL SCIENCES
UNIVERSITY OF IDAHO
MOSCOW, ID 83844-3051
HWICHMAN@UIDAHO.EDU