Inferring Population Mutation Rate and Sequencing Error Rate Using the SNP Frequency Spectrum in a Sample of DNA Sequences

Xiaoming Liu, Taylor J. Maxwell, Eric Boerwinkle, and Yun-Xin Fu

Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston

One challenge of analyzing samples of DNA sequences is to account for the nonnegligible polymorphisms produced by error when the sequencing error rate is high or the sample size is large. Specifically, those artificial sequence variations will bias the observed single nucleotide polymorphism (SNP) frequency spectrum, which in turn may further bias the estimators of the population mutation rate $\theta = 4N\mu$ for diploids. In this paper, we propose a new approach based on the generalized least squares (GLS) method to estimate θ, given a SNP frequency spectrum in a random sample of DNA sequences from a population. With this approach, error rate ε can be either known or unknown. In the latter case, ε can be estimated given an estimation of θ. Using coalescent simulation, we compared our estimators with other estimators of θ. The results showed that the GLS estimators are more efficient than other θ estimators with error, and the estimation of ε is usable in practice when the θ per bp is small. We demonstrate the application of the estimators with 10-kb noncoding region sequence sampled from a human population and provide suggestions for choosing θ estimators with error.

1. Introduction

Sequencing errors can produce many problems for evolutionary or population genetical analysis of samples of DNA sequences (Clark and Whittam 1992; Johnson and Slatkin 2008). Given a random sample of sequence from a population, artificial polymorphisms caused by sequencing error will skew both the number and the frequency spectrum of the observed single nucleotide polymorphisms (SNPs). This will further skew any estimations or tests based on the number and/or frequency spectrum of the SNPs if errors are not properly taken into account. The bias will be more prominent with increased sample size because sequencing error accumulates linearly with sample size while θ increases slower, as implied by coalescent theory. Johnson and Slatkin (2008) suggested a rule of thumb that an uncorrected estimate will be biased significantly if $n\varepsilon \geqslant \theta/L$, where n is the sample size (i.e., number of sequences), ε is the average error rate per site, L is the sequence length of the given locus, and θ is the population mutation rate of the locus. θ is equal to $4N\mu$ for diploids and $2N\mu$ for haploids, where N is the effective population size and μ is the mutation rate for the given locus. θ/L typically ranges from 10^{-4} for species with extremely low diversity, such as the bog turtle (Rosenbaum et al. 2007), to 10^{-2} for species with extremely high diversity, such as human immunodeficiency virus (Achaz et al. 2004). For human populations, θ/L is approximately 10^{-3} (Crawford et al. 2005). Therefore, with a typical sequencing error rate of 10^{-5} on the Sanger sequencing platform (Zwick 2005), uncorrected estimates will have problem with larger than 50 individuals (100 chromosomes) in human populations. If next-generation sequencing technologies are used, their higher sequencing error (10^{-4}) (Mackay et al. 2008; Shendure and Ji 2008) may further reduce the upper limit of sample size to only 10 chromosomes or less.

Key words: coalescent theory, sequencing error, mutation rate, SNP frequency spectrum, generalized least squares.

E-mail: yunxin.fu@uth.tmc.edu.

Mol. Biol. Evol. 26(7):1479–1490. 2009
doi:10.1093/molbev/msp059
Advance Access publication March 24, 2009

© The Author 2009. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved.
For permissions, please e-mail: journals.permissions@oxfordjournals.org

In response to the problem, researchers are developing new unbiased estimators and tests for population samples with sequencing error incorporated into the analysis. Johnson and Slatkin (2006) proposed maximum likelihood estimators for θ and the scaled exponential growth rate using the SNP frequency spectrum while accounting for sequencing error via Phred quality scores. They also studied the effects of sequencing error on two uncorrected estimators of θ, Tajima's $\hat{\theta}_\pi$ (Tajima 1983) and Watterson's $\hat{\theta}_K$ (Watterson 1975), and proposed two unbiased estimators of θ assuming known ε (see discussions below) (Johnson and Slatkin 2008). Hellmann et al. (2008) proposed a method to correct $\hat{\theta}_K$, assuming known ε. It is similar to Johnson and Slatkin (2008)'s θ estimator based on the total number of polymorphic sites but takes account of the uncertainty of chromosome sampling in a shot-gun sequencing of mixed diploid individuals. Lynch (2008) proposed several methods to correct $\hat{\theta}_\pi$ for high-coverage shot-gun genomic sequences from a single diploid individual. Knudsen and Miyamoto (2007) incorporated missing data, sequencing error, and multiple reads for diploid individuals into a full-likelihood coalescent model of a sample of DNA sequences. Their model can be used to estimate θ and ε jointly. However, the computational intensity of calculating the likelihood limits the application of the model to small sample sizes (i.e., less than 20 sequences). Assuming a low but unknown ε, Achaz (2008) proposed two new moment estimators of θ based on the SNP number and frequency spectrum while ignoring singletons, on which sequencing error skews the most, and constructed two new test statistics for neutrality based on the new estimators. Jiang et al. (2009) developed a method to estimate θ and recombination rate for shot-gun resequencing data and proposed some refinements to increase its robustness to sequencing errors.

Here, we propose a new approach based on the generalized least squares (GLS) method to estimate θ using the SNP frequency spectrum of a random sample of DNA sequences from a population. It can be considered as a modification and extension of Fu (1994b)'s best linear unbiased estimator (BLUE) estimator under the assumption of a moderate to low ε, either known or unknown. When ε is unknown, it can be estimated sequentially after the estimation of θ. The rational here is to incorporate the variance–covariance structure of the SNP frequency spectrum, with

either known or unknown ancestral states of the SNPs, and reduce the variation of the estimator. Its computational intensity is higher than Achaz (2008)'s moment estimators but much lower than Knudsen and Miyamoto (2007)'s full-likelihood method, which makes it applicable to samples with thousands of sequences. In the following sections, we first introduce our GLS estimators along with some corrected or uncorrected estimators of θ; then, we compare these estimators using coalescent simulation assuming different population parameters. We demonstrate their application on a sample of DNA sequences of a 10-kilobase noncoding region on human chromosome 22. Finally, we discuss the advantages, limitations, and some technical issues of the GLS estimators.

2. Methods

2.1 *Assumptions*

Sequencing error at each site is usually modeled as a binomial or Poisson variable with parameter $p = n\varepsilon$. As defined in the Introduction, n is the sample size (i.e., number of sequences) and ε is the average error rate per site. Here, we simplify the model by assuming that each site has at most one sequencing error, which occurs with probability P. As a rule of thumb, this assumption is approximately valid with $P < 0.01$. We further assume that when a sequencing error occurs on an ancestral (or mutant) allele, the probability that it leads to an existing allele (mutant or ancestral) is u and the probability that it yields a new allele is $1-u$. For example, if only point mutation is considered and assuming that a nucleotide has equal probability to be read incorrectly as one of the other three nucleotides, $u = 1/3$. For a sample of n sequences with length L randomly sampled from a population, we designate ξ_i as the (unknown) number of true polymorphic sites with mutation size i (or SNP class i, i.e., there are i copies of the mutant allele in the sample) and $\xi_{i,k}$ as the observed number of sites with $n-i-k$ copies of the ancestral allele, i copies of the mutant alleles and k copy of the new allele (other than mutant and ancestral alleles) produced by sequencing error ($k = 0, 1$).

2.2 *Approximate GLS Estimators of θ When ε is Unknown*

If ε or p is unknown but assumed to be relatively small, it is easy to see that most errors are singletons ($\xi_{0,1}$ or $\xi_{n-1,1}$) because $\xi_0 \gg \sum_{i=1}^{n-1} \xi_i$ with assumption of infinite site model and a reasonable θ (Achaz 2008). In another words, $\xi_i \approx \xi_{i,0}$ for $2 \leqslant i \leqslant n-1$. By ignoring singletons, we removed most of the errors so that $\varepsilon \approx 0$ in the remaining SNP classes based on which new estimators can be developed. For example, Achaz (2008) proposed $\hat{\theta}_{K-\xi_1}$ and $\hat{\theta}_{K-\eta_1}$ estimators, which are modified Watterson (1975)'s $\hat{\theta}_K$ estimator and $\hat{\theta}_{\pi-\xi_1}$ and $\hat{\theta}_{\pi-\eta_1}$ estimators, which are modified Tajima (1983)'s $\hat{\theta}_\pi$ estimator (see details below). Fu (1994b) proposed a θ estimator based on the GLS method ($\hat{\theta}_{BLUEu}$ and $\hat{\theta}_{BLUEf}$, called BLUE estimators in Fu 1994b), which makes full use of the observed numbers of SNP frequency classes, so that it has much less variability than either Watterson (1975)'s or Tajima (1983)'s estimator.

Following Achaz (2008)'s idea that because singletons are much more likely to be errors comparing to other SNP class, an estimator based on the SNP spectrum excluding singletons will be robust to errors. Here, we briefly describe the modification of Fu (1994b)'s GLS method by ignoring all singletons and triple allele. See Fu (1994a, 1994b) for more technique details of the method. We first rewrite the expectations, variances, and covariances of the observed number of SNP classes as:

$$E(\xi_i) = \rho_i \theta \tag{1}$$

$$Var(\xi_i) = Cov(\xi_i, \xi_i) = \rho_i \theta + \sigma_{ii} \theta^2 \tag{2}$$

$$Cov(\xi_i, \xi_j) = \sigma_{ij} \theta^2, \tag{3}$$

where $1 \leqslant i \leqslant n-1$. The values of ρ_i and σ_{ij} are dependent on the population model, which can be obtained theoretically or numerically using simulation. Assuming a Wright–Fisher model with constant population size, no selection, and recombination, their explicit formulas are known (Fu 1995). Writing in a matrix form, we have

$$E(Y) = X\theta$$

$$Var(Y) = \hat{V},$$

where

$$Y = \begin{pmatrix} \xi_2 \\ \vdots \\ \xi_{n-1} \end{pmatrix}, \quad X = \begin{pmatrix} \rho_2 \\ \vdots \\ \rho_{n-1} \end{pmatrix}$$

and $\hat{V}$ is calculated with equations (2) and (3). Then, the corresponding GLS estimator of θ is

$$\hat{\theta}_{BLUE-\xi_1} = \left[X' \hat{V}^{-1} X\right]^{-1} X' \hat{V}^{-1} Y.$$

Because the calculation of $\hat{V}$ needs prior estimation of θ, the $\hat{\theta}_{BLUE-\xi_1}$ is calculated iteratively (Fu 1994a, 1994b): 1) give an initial value of θ, say $\theta_0 = \hat{\theta}_\pi$, 2) calculate $\hat{V}$ with θ_0; 3) calculate $\hat{\theta}_{BLUE-\xi_1}$ with $\hat{V}$ and let $\theta_1 = \hat{\theta}_{BLUE-\xi_1}$; 4) update $\hat{V}$ with θ_1; and 5) repeat step 3 and 4 until $\hat{\theta}_{BLUE-\xi_1}$ converges.

If the ancestral state of each SNP is unknown (i.e., we do not know which allele is ancestral and which is mutant) then we need to fold the counting of ξ_i and ξ_{n-i} to include all SNPs with a minor allele frequency i and major allele frequency $n-i$ or vice versa. In the folded case,

$$Y = \begin{pmatrix} \xi_2 + \xi_{n-2} \\ \vdots \end{pmatrix}, X = \begin{pmatrix} \rho_2 + \rho_{n-2} \\ \vdots \end{pmatrix}$$

and $\hat{V}$ is changed accordingly. The corresponding estimator, designated as $\hat{\theta}_{BLUE-\eta_1}$, can be calculated with the same iterative process as described above.

After the θ is estimated, p can be approximated as

$$\hat{p} = \frac{\xi_{1,0} + \xi_{0,1} - E\left(\xi_1 | \hat{\theta}\right)}{\xi_{1,0} + \xi_{0,1} - E\left(\xi_1 | \hat{\theta}\right) + \xi_{0,0}} \tag{4}$$

for unfolded counting or

$$\hat{p}=\frac{\begin{array}{c}\xi_{1,0}+\xi_{0,1}+\xi_{n-1,0}+\xi_{n-1,1}\\ -E\left(\xi_1|\hat{\theta}\right)-E\left(\xi_{n-1}|\hat{\theta}\right)\end{array}}{\begin{array}{c}\xi_{1,0}+\xi_{0,1}+\xi_{n-1,0}+\xi_{n-1,1}-E\left(\xi_1|\hat{\theta}\right)\\ -E\left(\xi_{n-1}|\hat{\theta}\right)+\xi_{n,0}+\xi_{0,0}\end{array}}, \quad (5)$$

for folded counting, where $E\left(\xi_1|\hat{\theta}\right)$ and $E\left(\xi_{n-1}|\hat{\theta}\right)$ are calculated using equation (1) with θ replaced by $\hat{\theta}$. Then, $\hat{\varepsilon}$ is simply $\hat{p}/n$. For convenience, we use the subscript of the θ estimator used in $\hat{\varepsilon}$ to designate different $\hat{\varepsilon}$s. For example, $\hat{\varepsilon}_{BLUE-\xi_1}$ means the $\hat{\varepsilon}$ calculated using $\hat{\theta}_{BLUE-\xi_1}$.

2.3 *GLS Estimators of θ When ε is Known*

When ε is known, we can develop a more precise GLS estimator of θ by making full use of this information. First, we need to derive expectations, variance, and covariances of $\xi_{i,k}$. For simplicity, letting $\xi_i=0$ if $i<0$ or $i\geqslant n$ then

$$E\left(\xi_{i,0}|p,\theta\right)=E\left(\xi_i\right)+p\sum_{k=i-1}^{i+1}b_{i-k,0}^{k}E\left(\xi_k\right)\times(i=0,\cdots,n) \quad (6)$$

$$E\left(\xi_{i,1}|p,\theta\right)=p\sum_{k=i}^{i+1}b_{i-k,1}^{k}E\left(\xi_k\right)\times(i=0,\cdots,n-1) \quad (7)$$

$$\begin{aligned}Var\left(\xi_{i,0}|p,\theta\right)=&\,p\sum_{k=i-1}^{i+1}c_{i-k,0,i-k,0}^{k}E\left(\xi_k\right)\\&+p^2\sum_{k=i-1}^{i+1}d_{i-k,0,i-k,0}^{k}E\left(\xi_k\right)\\&+\sum_{k=i-1}^{i+1}\sum_{l=j-1}^{j+1}\left(a_{i-k,0}^{k}+b_{i-k,0}^{k}p\right)\\&\times\left(a_{j-l,0}^{l}+b_{j-l,0}^{l}p\right)Cov\left(\xi_k,\xi_l\right)\\&(i=0,\cdots,n)\end{aligned} \quad (8)$$

$$\begin{aligned}Var\left(\xi_{i,1}|p,\theta\right)=&\,p\sum_{k=i}^{i+1}c_{i-k,1,i-k,1}^{k}E\left(\xi_k\right)\\&+p^2\sum_{k=i}^{i+1}d_{i-k,1,i-k,1}^{k}E\left(\xi_k\right)\\&+p^2\sum_{k=i}^{i+1}\sum_{l=j}^{j+1}\left(b_{i-k,1}^{k}\right)\\&\times\left(b_{j-l,1}^{l}\right)Cov\left(\xi_k,\xi_l\right)\\&(i=0,\cdots,n-1)\end{aligned} \quad (9)$$

$$\begin{aligned}&Cov\left(\xi_{i,0},\xi_{j,0}|p,\theta\right)\\&=p\sum_{k=i-1}^{i+1}\sum_{l=j-1}^{j+1}\delta_{k=l}c_{i-k,0,j-k,0}^{k}E\left(\xi_k\right)\\&\quad+p^2\sum_{k=i-1}^{i+1}\sum_{l=j-1}^{j+1}\delta_{k=l}d_{i-k,0,j-k,0}^{k}E\left(\xi_k\right)\\&\quad+\sum_{k=i-1}^{i+1}\sum_{l=j-1}^{j+1}\left(a_{i-k,0}^{k}+b_{i-k,0}^{k}p\right)\\&\quad\times\left(a_{j-l,0}^{l}+b_{j-l,0}^{l}p\right)Cov\left(\xi_k,\xi_l\right)\\&\quad(i=0,\cdots,n;j=0,\cdots,n;i\neq j)\end{aligned} \quad (10)$$

$$\begin{aligned}Cov\left(\xi_{i,1},\xi_{j,1}|p,\theta\right)=&\,p\sum_{k=i}^{i+1}\sum_{l=j}^{j+1}\delta_{k=l}c_{i-k,1,j-k,1}^{k}E\left(\xi_k\right)\\&+p^2\sum_{k=i}^{i+1}\sum_{l=j}^{j+1}\delta_{k=l}d_{i-k,1,j-k,1}^{k}E\left(\xi_k\right)\\&+p^2\sum_{k=i}^{i+1}\sum_{l=j}^{j+1}\left(b_{i-k,1}^{k}\right)\\&\times\left(b_{j-l,1}^{l}\right)Cov\left(\xi_k,\xi_l\right)\\&(i=0,\cdots,n-1;j=0,\cdots,n-1;i\neq j)\end{aligned} \quad (11)$$

$$\begin{aligned}Cov\left(\xi_{i,0},\xi_{j,1}|p,\theta\right)=&\,p\sum_{k=i-1}^{i+1}\sum_{l=j}^{j+1}\delta_{k=l}c_{i-k,0,j-k,1}^{k}E\left(\xi_k\right)\\&+p^2\sum_{k=i-1}^{i+1}\sum_{l=j}^{j+1}\delta_{k=l}d_{i-k,0,j-k,1}^{k}E\left(\xi_k\right)\\&+\sum_{k=i-1}^{i+1}\sum_{l=j}^{j+1}\left(a_{i-k,0}^{k}+b_{i-k,0}^{k}p\right)\\&\times\left(b_{j-l,1}^{l}p\right)Cov\left(\xi_k,\xi_l\right)\\&(i=0,\cdots,n;j=0,\cdots,n-1),\end{aligned} \quad (12)$$

where $\delta_{k=l}$ is an index function, which equals to 1 if $k=l$ or 0 otherwise; $a_{m,j}^{i}$, $b_{m,j}^{i}$, c_{m,j_1,k,j_2}^{i}, and d_{m,j_1,k,j_2}^{i} are functions of i, j, j_1, j_2, m, k, n, and u. Their detailed expression and the derivation of the above formulas can be found in the Appendix. $E\left(\xi_i\right)$, $Var\left(\xi_i\right)$, and $Cov\left(\xi_i,\xi_j\right)$ $(1\leqslant i,j\leqslant n-1)$ can be calculated using equations (1)–(3).

In practice, $\xi_{1,0}$ and $\xi_{0,1}$ are indistinguishable even when the ancestral states are known. To avoid the above difficulties, we combine observations of $\xi_{1,0}$ and $\xi_{0,1}$. The general formulas for calculating the expectation of the combined observations and the variance–covariance of the combined observation with other (combined)

observations are

$$E\left(\sum_i \xi_{x_i,y_i}|p,\theta\right)=\sum_i E\left(\xi_{x_i,y_i}|p,\theta\right) \tag{13}$$

$$Var\left(\sum_i \xi_{x_i,y_i}|p,\theta\right)=\sum_i Var\left(\xi_{x_i,y_i}|p,\theta\right)+\sum_{j=1}^{i}\sum_{k=1}^{i} Cov\left(\xi_{x_j,y_j},\xi_{x_k,y_k}|p,\theta\right) \tag{14}$$

$$Cov\left(\sum_i \xi_{x_i,y_i},\xi_{w,z}|p,\theta\right)=\sum_i Cov\left(\xi_{x_i,y_i},\xi_{w,z}|p,\theta\right) \tag{15}$$

$$Cov\left(\sum_i \xi_{x_i,y_i},\sum_j \xi_{w_j,z_j}|p,\theta\right)=\sum_i\sum_j Cov\left(\xi_{x_i,y_i},\xi_{w_j,z_j}|p,\theta\right), \tag{16}$$

where ξ_{x_i,y_i}, $\xi_{w,z}$, and ξ_{w_j,z_j} can be any given $\xi_{k,0}$ or $\xi_{k,1}$. For example, to apply equations (13)–(14) to $\xi_{1,0}$ and $\xi_{0,1}$, we simply let $x_1=1$, $y_1=0$, $x_2=0$, and $y_2=1$. By incorporating $E(\xi_0)=L-\sum_{i=1}^{n-1}E(\xi_i)$, we have

$$E(\xi_{1,0}+\xi_{0,1}|p,\theta)=\rho_1\theta+Lp+\left(-\sum_{i=1}^{n-1}\rho_i+\frac{2u}{n}\rho_2+\frac{1-u-n}{n}\rho_1\right)\theta p.$$

A GLS estimator can be calculated as follows. If the ancestral state of each site is known or can be inferred from outgroups, let

$$Y=\begin{pmatrix}\xi_{1,0}+\xi_{0,1}\\ \xi_{2,0}\\ \vdots\\ \xi_{n-1,0}\\ \xi_{1,1}\\ \vdots\\ \xi_{n-1,1}\end{pmatrix}, X_{10}=\begin{pmatrix}\rho_1\\ \rho_2\\ \vdots\\ \rho_{n-1}\\ 0\\ \vdots\\ 0\end{pmatrix},$$

$$X_{01}=\begin{pmatrix}L\\ 0\\ \vdots\\ 0\\ 0\\ \vdots\\ 0\end{pmatrix}, X_{11}=\begin{pmatrix}-\sum_{i=1}^{n-1}\rho_i+\frac{2u}{n}\rho_2+\frac{1-u-n}{n}\rho_1\\ \sum_{k=2-1}^{2+1}b_{2-k,0}^{k}\rho_k\\ \vdots\\ \sum_{k=n-1-1}^{n-1+1}b_{n-2-k,0}^{k}\rho_k\\ \sum_{k=1}^{1+1}b_{1-k,1}^{k}\rho_k\\ \vdots\\ \sum_{k=n-1}^{n-1+1}b_{n-2-k,1}^{k}\rho_k\end{pmatrix}$$

then

$$E(Y)=X_{10}\theta+X_{01}p+X_{11}p\theta. \tag{17}$$

The GLS estimator of θ is

$$\hat{\theta}_{GLS}=\left[(X_{10}+X_{11}p)'\hat{V}^{-1}(X_{10}+X_{11}p)\right]^{-1}\times(X_{10}+X_{11}p)'\hat{V}^{-1}(Y-X_{01}p), \tag{18}$$

again $\hat{V}$ is an estimation of the variance–covariance matrix of Y , which can be calculated using equations (8)–(16). The same iterative process described above is then used to calculate $\hat{\theta}_{GLS}$. In the case that ancestral states are unknown, we need to further fold $\xi_{i,0}$ with $\xi_{n-i,0}$ $\left(2\leqslant i<\frac{n}{2}\right)$, and $\xi_{i,1}$ with $\xi_{n-i-1,1}$ $\left(2\leqslant i<\frac{n-1}{2}\right)$ because they are indistinguishable. Let

$$Y=\begin{pmatrix}\xi_{1,0}+\xi_{0,1}+\xi_{n-1,0}+\xi_{n-1,1}\\ \xi_{2,0}+\xi_{n-2,0}\\ \vdots\\ \xi_{1,1}+\xi_{n-2,1}\\ \vdots\end{pmatrix},$$

$$X_{10}=\begin{pmatrix}\rho_1+\rho_{n-1}\\ \rho_2+\rho_{n-2}\\ \vdots\\ 0\\ \vdots\end{pmatrix},$$

$$X_{01}=\begin{pmatrix}L\\ 0\\ \vdots\\ 0\\ \vdots\end{pmatrix},$$

$$X_{11}=\begin{pmatrix}-\sum_{i=1}^{n-1}\rho_i+\frac{2u}{n}\rho_2+\frac{1-u-n}{n}\rho_1+\frac{2u}{n}\rho_{n-2}\\ +\frac{1-u-n}{n}\rho_{n-1}\sum_{k=2-1}^{2+1}b_{2-k,0}^{k}\rho_k\\ +\sum_{k=n-2-1}^{n-2+1}b_{n-2-k,0}^{k}\rho_k\\ \vdots\\ \sum_{k=1}^{1+1}b_{1-k,1}^{k}\rho_k+\sum_{k=n-2}^{n-2+1}b_{n-2-k,1}^{k}\rho_k\\ \vdots\end{pmatrix}$$

and follow the same steps as described for $\hat{\theta}_{GLS}$. We can calculate the GLS estimator with folded observations. Later in this paper, we designate $\hat{\theta}_{GLSf}$ to be the GLS estimator using folded observations and $\hat{\theta}_{GLSu}$ to be that using unfolded observations.

2.4 *Comparing to Other Estimators of θ Using Simulation*

Three types of corrected and uncorrected estimators of θ were compared in this study, which will be briefly introduced below. Johnson and Slatkin (2006)'s method was not compared because here, we assume the detailed Phred quality scores is unknown. Neither did we include

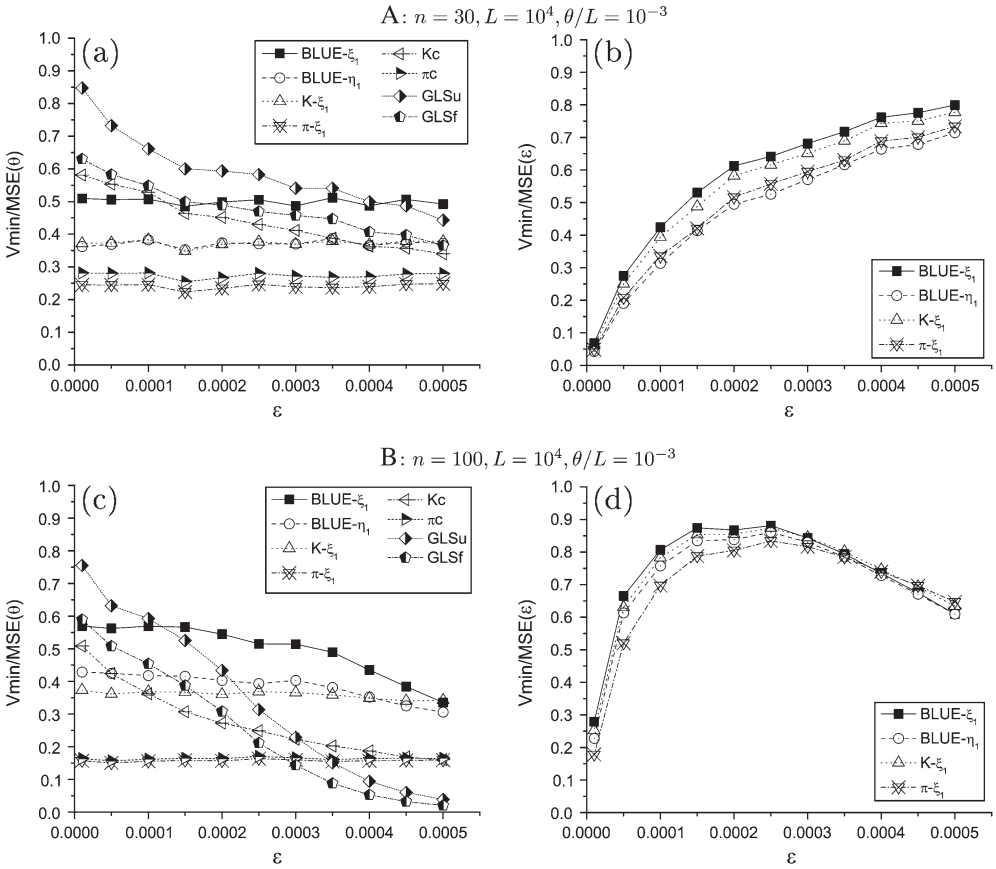

FIG. 1.—Efficiency of θ and ε estimators with increase of ε. Only the subscripts of the estimators were shown.

Knudsen and Miyamoto (2007)'s method because of its computational intensity and limitation to small sample sizes. Hellmann et al. (2008) and Johnson and Slatkin (2008)'s estimators based on the total number polymorphic sites should have similar performances, so only the latter was included in our comparison.

The first type of estimator (type I) is based on the total number of polymorphic sites. An uncorrected estimator of this type is the widely used Watterson's estimator $\hat{\theta}_K = K/a_n$ (Watterson 1975), where $a_n = \sum_{i=1}^{n-1} 1/i$ and K is the total number of polymorphic sites in the sample. Assuming a known ε, Johnson and Slatkin (2008) proposed a corrected estimator (with modification):

$$\hat{\theta}_{Kc} = \frac{K - E[q_2]L}{a_n(1 - E[q_1] - E[q_2])},$$

where

$$E[q_1] = \frac{1}{a_n}\sum_{i=1}^{n-1}\frac{1}{i}\left[\left(\frac{\varepsilon}{m-1}\right)^i (1-\varepsilon)^{n-i} + (1-\varepsilon)^i\left(\frac{\varepsilon}{m-1}\right)^{n-i} + \varepsilon\frac{m-2}{m-1}\left(\frac{\varepsilon}{m-1}\right)^{n-1}\right],$$

$$E[q_2] = 1 - (1-\varepsilon)^n - \varepsilon\left(\frac{\varepsilon}{m-1}\right)^{n-1},$$

and m is the total number of possible alleles in a site. Assuming an unknown but small ε, Achaz (2008) proposed two corrected estimators of this type. They are $\hat{\theta}_{K-\xi_1} = \left(\sum_{i=2}^{n-1}\xi_i\right)/(a_n - 1)$ for unfolded observations and $\hat{\theta}_{K-\eta_1} = \left(\sum_{i=2}^{n-2}\xi_i\right)/[a_n - n/(n-1)]$ for folded observations.

The second type of estimator (type II) is based on the average difference between two sequences. An uncorrected estimator of this type is Tajima's $\hat{\theta}_\pi = \binom{n}{2}^{-1}\sum_{i<j}^{n}\pi_{ij}$ (Tajima 1983), where π_{ij} is the number of difference between sequences i and j. Johnson and Slatkin (2008)'s corrected estimator of this type (with modification) is

$$\hat{\theta}_{\pi c} = \frac{\hat{\theta}_\pi - E[p_2]L}{1 - E[p_1] - E[p_2]},$$

where

$$E[p_1] = \frac{2\varepsilon(1-\varepsilon)}{m-1} + \frac{(m-2)\varepsilon^2}{(m-1)^2}$$

$$E[p_2] = 2\varepsilon(1-\varepsilon) + \frac{m-2}{m-1}\varepsilon^2.$$

Achaz (2008)'s corrected estimators of this type are

$$\hat{\theta}_{\pi-\xi_1} = \frac{2}{(n-1)(n-2)}\sum_{i=2}^{n-1} i(n-i)\xi_i$$

for unfolded observations and

$$\hat{\theta}_{\pi-\eta_1} = \frac{2}{n(n-3)}\sum_{i=2}^{n-2} i(n-i)\xi_i$$

for folded observations.

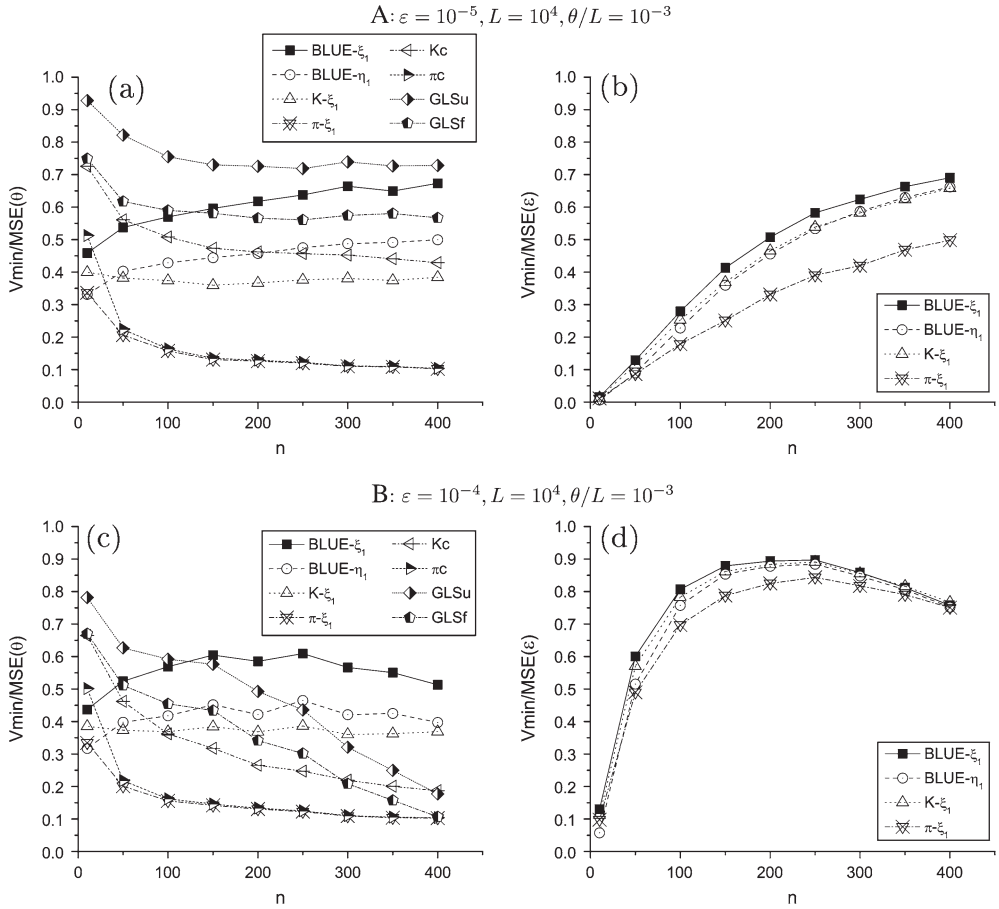

FIG. 2.—Efficiency of θ and ε estimators with increase of n. Only the subscripts of the estimators were shown.

The third type of estimator (type III) compared is the GLS estimator. Our estimators are corrected GLS estimators, which can be considered an extension to Fu (1994b)'s uncorrected GLS estimators $\hat{\theta}_{BLUEu}$ and $\hat{\theta}_{BLUEf}$ for unfolded and folded counting, respectively.

We compared the performances of the estimators using coalescent simulations (e.g., Hudson 2002). We assumed a Wright–Fisher model with constant population size, no selection, no recombination and a simple infinite sites model for mutation. For each combination of parameters, 10,000 samples were simulated. Only point mutations ($u = 1/3$) were simulated. The GLS estimation iteration was stopped either when the absolute value change between the update of $\hat{\theta}_{GLS}$ is smaller than 10^{-3} or when 200 updates of equation (18) was conducted (see Methods for details). Sequencing error at each site was simulated to follow a binomial distribution with parameter p. Because $\hat{\theta}_{BLUE-\xi_1}$, $\hat{\theta}_{BLUE-\eta_1}$, $\hat{\theta}_{K-\xi_1}$, $\hat{\theta}_{K-\eta_1}$, $\hat{\theta}_{\pi-\xi_1}$, $\hat{\theta}_{\pi-\eta_1}$, $\hat{\theta}_{BLUEu}$, and $\hat{\theta}_{BLUEf}$ assume at most two alleles at each site, for those estimators if a site has more than one type of non-ancestral allele, the one with the largest count is regarded as the mutant allele and all other non-ancestral alleles are regarded as errors from ancestral alleles. For example, if a site has two non-ancestral alleles with counts 1 and 2 then it will be added to ξ_2 instead of ξ_1. $\hat{p}$ was calculated with equation (4) using $\hat{\theta}_{BLUE-\xi_1}$, $\hat{\theta}_{K-\xi_1}$, and $\hat{\theta}_{\pi-\xi_1}$ or with equation (5) using $\hat{\theta}_{BLUE-\eta_1}$, $\hat{\theta}_{K-\eta_1}$, and $\hat{\theta}_{\pi-\eta_1}$.

3. Results

3.1 *Simulation Comparison*

For each set of simulated samples with a given combination of parameters, we compared the mean and variance of each estimator. We also calculated the mean squared error (MSE) of each estimator and the variance of an optimal estimator, V_{min}. For θ, the large-sample approximation of V_{min} is $V_{min}(\theta) = \theta\left[\sum_{i=1}^{n-1}(\theta+i)^{-1}\right]^{-1}$ (Fu and Li 1993). Because we know the actual number of errors introduced in each simulated sample, say n_{err}, we simply use n_{err}/nL as the optimal estimator of ε and calculated its variance $V_{min}(\varepsilon)$ in simulated samples. The ratio V_{min}/MSE indicates the relative efficiency of each estimator to the optimal estimator.

As expected, with increasing p, via either an increasing error rate ε or an increasing sample size n, the uncorrected θ estimators perform poorly (data not shown). The four uncorrected θ estimators ($\hat{\theta}_K$, $\hat{\theta}_\pi$, $\hat{\theta}_{BLUEu}$, and $\hat{\theta}_{BLUEf}$) are upper biased significantly with increasing ε because they assume every polymorphism is a result of mutation. As a result, the uncorrected θ estimators are the least efficient estimators comparing to the corrected ones

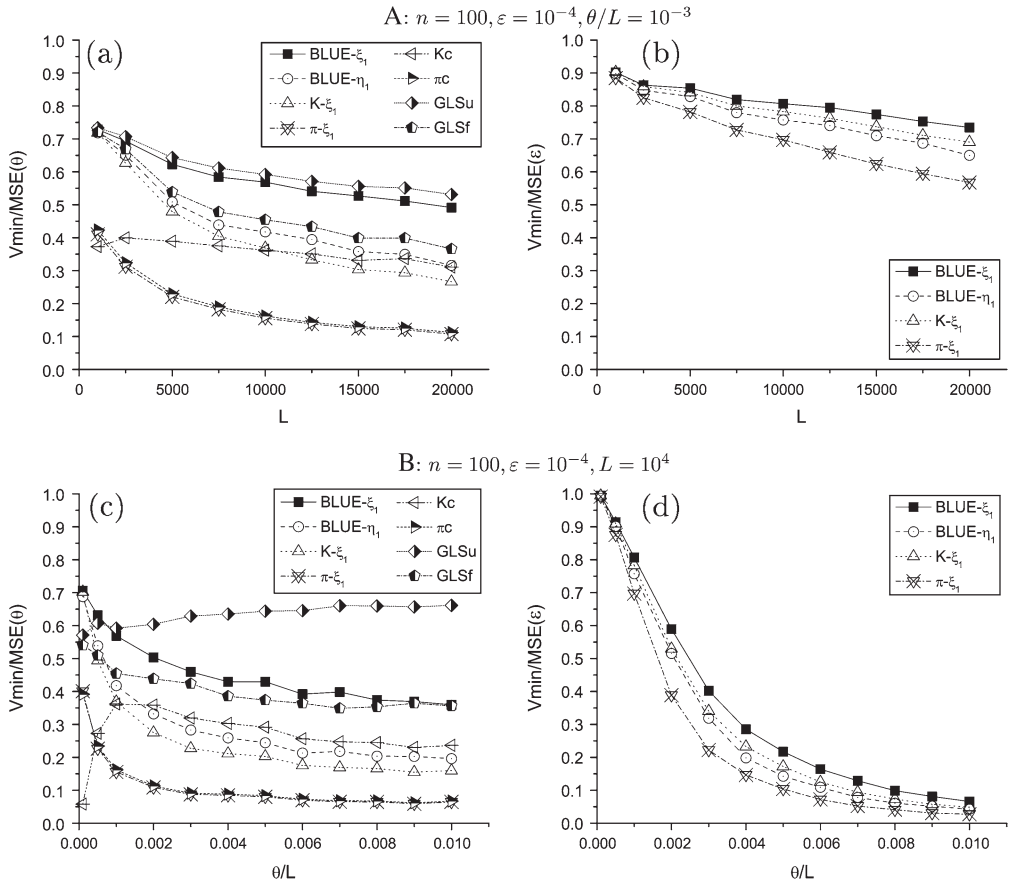

FIG. 3.—Efficiency of θ and ε estimators with increase of L and θ/L. Only the subscripts of the estimators were shown.

(data not shown). Among them, $\hat{\theta}_{BLUEu}$ and $\hat{\theta}_{BLUEf}$ are the most sensitive to sequencing error, whereas $\hat{\theta}_\pi$ is the least sensitive. Looking at this in another way, although under the null hypothesis of no sequencing error, the order of their relative efficiency is $\hat{\theta}_{BLUEu} > \hat{\theta}_{BLUEf} > \hat{\theta}_K > \hat{\theta}_\pi$, the order is totally reversed when p is larger. The above observations can be partially explained by the weight each estimator puts on ξ_i. Although $\hat{\theta}_K$ puts equal weight on every ξ_i, $\hat{\theta}_\pi$ puts less weight on ξ_i than on ξ_j if $i < j$. On the contrary, $\hat{\theta}_{BLUEu}$ puts more weight on ξ_i because under the assumption of no error, ξ_i is a more reliable observation than ξ_j. However, the opposite is true with errors.

On the other hand, the efficiency of corrected θ estimators is less affected by increasing p. As to those with unknown ε, all are approximately unbiased. Their variances are relatively unchanged with the increase of ε or even decrease slightly with an increase of n (data not shown). The order of their efficiency is $\hat{\theta}_{BLUE-\xi_1} > \hat{\theta}_{BLUE-\eta_1} \gtrsim \hat{\theta}_{K-\xi_1} \approx \hat{\theta}_{K-\eta_1} > \hat{\theta}_{\pi-\xi_1} \approx \hat{\theta}_{\pi-\eta_1}$ (figs. 1 and 2, $\hat{\theta}_{K-\eta_1}$ and $\hat{\theta}_{\pi-\eta_1}$ are not shown because they have a very similar efficiency as $\hat{\theta}_{K-\xi_1}$ and $\hat{\theta}_{\pi-\xi_1}$, respectively). For the estimators with known ε, all of them are approximately unbiased except that $\hat{\theta}_{GLSu}$ and $\hat{\theta}_{GLSf}$ are upper biased when $p > 0.01$. The variances of $\hat{\theta}_{Kc}$, $\hat{\theta}_{GLSu}$ and $\hat{\theta}_{GLSf}$ increase with the increase of ε or n except when n is small (say < 100). As a result, their efficiency decreases with increasing p (figs. 1 and 2). With $p < 0.01$, the order of their efficiency is $\hat{\theta}_{GLSu} > \hat{\theta}_{GLSf} > \hat{\theta}_{Kc} > \hat{\theta}_{\pi c}$ (figs. 1 and 2). In summary, with a small to moderate ε (when $P < 0.01$ and known), $\hat{\theta}_{GLSu}$ and $\hat{\theta}_{GLSf}$ are the most efficient θ estimators; otherwise, $\hat{\theta}_{BLUE-\xi_1}$ and $\hat{\theta}_{BLUE-\eta_1}$ are the most efficient estimators.

Increasing θ, either by increasing θ/L or by L, affects the variances but not the means of the θ estimators. The variances of all estimators increase with the increase of θ. As a result, the efficiency of corrected estimators decrease with an increase of θ, with the exception of $\hat{\theta}_{Kc}$, which actually increases when θ/L is small (fig. 3). Among the corrected estimators, the GLS estimators perform better than others, whereas those with unfolded counting perform better than those with folded counting (fig. 3).

The differences between the estimators of p are smaller compared with their corresponding θ estimators. Because the estimation of p is based on prior estimation of θ, it is easy to understand that a more accurate estimation of θ will help with a more accurate estimation of p. In general, using a θ estimator with unfolded counting will provide a more efficient estimation of p than using an estimator with folded counting. Among those θ estimators with unfolded (or folded) counting, the GLS estimator usually has the highest efficiency, whereas the type II estimator usually has the lowest efficiency. With an increase of p, the efficiency of the estimators of p increase to a certain level

and then decreases (figs. 1 and 2). Even though their variances converge toward $V_{min}(\varepsilon)$ with an increase of p, they tend to underestimate ε at the same time; and at a certain point, the efficiency gain from the variance is not sufficient to compensate for the efficiency loss from bias. With an increase of L, the biases of the estimators do not change but their variances decrease at the same pace as the optimal estimator. With the increase of θ/L, both the biases and the variances of the estimators increase. As a result, their efficiency always decreases with an increase of θ, which is faster with an increase of θ/L than with an increase of L (fig. 3).

3.2 *Application Example*

Zhao et al. (2000) sequenced a 10-kilobase noncoding region on human chromosome 22 from 64 individuals collected worldwide. Homologous sequences were obtained from a chimpanzee and an orangutan as outgroups. Among the 78 variant sites originally found from the alignment, 43 (including all 24 singletons) were verified by restriction fragment length polymorphism, resequencing, or subcloning. Four errors were found from the singletons, including three originated from nonvariant sites and one originated from a doubleton. The authors also described an error that changed a site from the 10 mutant class to the 9 mutant class. So it is possible for us to recover an uncleaned data set with five errors.

Here, we demonstrate the application of the estimators using both the cleaned data set and the uncleaned data set. The cleaned sequences were retrieved from GenBank and aligned using MUSCLE (Edgar 2004). Ancestral states of alleles were determined using the outgroup sequences. After trimming and removing insertions and deletions, the data set includes $n = 128$ and $L = 9413$ and 69 polymorphic sites (table 1). $\hat{\theta}_{BLUE-\xi_1} = 14.964$ and $\hat{\varepsilon}_{BLUE-\xi_1} = 2.5 \times 10^{-6}$ were estimated using the cleaned data (table 2). If we assume there is no error ($\varepsilon = 0$), $\hat{\theta}_{GLSu} = 16.300$ and $\hat{\theta}_{BLUEu} = 16.238$ (table 2) can be regarded as an upper bound of the true θ. We then restored the five errors back to the data set according to Zhao et al. (2000) (table 1). Results show that $\hat{\theta}_K$, $\hat{\theta}_\pi$, and $\hat{\theta}_{BLUEu}$ are inflated by the erroneous singletons, whereas $\hat{\theta}_{BLUE-\xi_1}$ is not (table 2). Actually, $\hat{\theta}_{BLUE-\xi_1}$ decreases slightly to 14.499 mostly due to a decrease in the number of doubletons. Because we already know there are at least five errors, we know a lower bound of $\hat{\varepsilon} = 5/nL = 4.15 \times 10^{-6}$. Using this $\hat{\varepsilon}$, $\hat{\theta}_{GLSu} = 15.452$, which is our best approximation of an upper bound of the true θ with the uncleaned data.

Table 1
Counting of Sites with Different Number of Mutant Alleles in the Example Data Set

No. of Mutant	Site Counts (Cleaned)	Site Counts (Uncleaned)
1	18	22
2	20	19
3	3	3
4	3	3
5	1	1
7	1	1
8	3	3
9	0	1
10	4	3
12	1	1
21	1	1
27	1	1
32	1	1
45	1	1
46	1	1
54	1	1
55	1	1
56	1	1
59	1	1
69	1	1
77	1	1
79	1	1
89	1	1
114	1	1
118	1	1
Total	69	72

4. Discussion

The results reported here show that the GLS estimators of θ perform well with a small to moderate sequencing error rate ε. When $P < 0.01$, the GLS estimators with a known ε ($\hat{\theta}_{GLSu}$ and $\hat{\theta}_{GLSf}$) are the most efficient estimators. For the θ estimators with unknown ε, the GLS estimators ($\hat{\theta}_{BLUE-\xi_1}$ and $\hat{\theta}_{BLUE-\eta_1}$) are relatively more efficient than other corrected estimators. In general, the GLS estimators using unfolded observation are superior to those using folded observations. In addition, because the GLS estimators are based on summary statistics of the sample, the computation is much faster than Knudsen and Miyamoto (2007)'s method based on full likelihood. From this, we conclude that the GLS estimators are a good balance between estimation efficiency and computation efficiency.

The GLS estimators also have obvious limitations. First, it assumes a maximum of one sequencing error for each site, which may not be true when n and ε are large. When $P > 0.01$, $\hat{\theta}_{GLSu}$ and $\hat{\theta}_{GLSf}$ tend to overestimate θ. We also observe that the efficiency of $\hat{\theta}_{GLSu}$ and $\hat{\theta}_{GLSf}$ worsens faster than that of other estimators with increase of p. The reason is similar to that of the efficiency reverse of $\hat{\theta}_{BLUEu}$, $\hat{\theta}_{BLUEf}$, $\hat{\theta}_K$, and $\hat{\theta}_\pi$ with and without error, as we briefly discussed in the Results. That is, when there are more errors than assumed (or can be corrected), the counting of rare variants is further skewed. Comparing to other estimators, $\hat{\theta}_{GLSu}$ and $\hat{\theta}_{GLSf}$ put much more weight on rare variants than on common variants, which causes them to be biased more than other estimators. Although $\hat{\theta}_{BLUE-\xi_1}$ and $\hat{\theta}_{BLUE-\eta_1}$ are more robust to p, their efficiency decreases when p is large. Second, they cannot easily handle missing data. However, with our limited simulations, we observed that if missing data is simply imputed using the observed allele frequencies, a random missing rate of up to 10% seems to only have a small effect on the efficiencies of $\hat{\theta}_{GLSu}$, $\hat{\theta}_{GLSf}$, $\hat{\theta}_{BLUE-\xi_1}$, and $\hat{\theta}_{BLUE-\eta_1}$ (data not shown). Third, the computational complexity is higher than those of the type I and type II estimators.

Choosing the most efficient θ estimator with error depends on our knowledge of the error rate. If ε is known and $P < 0.01$, $\hat{\theta}_{GLSu}$ or $\hat{\theta}_{GLSf}$ is probably a good choice. If ε is unknown, we could first use $\hat{\theta}_{BLUE-\xi_1}$ or $\hat{\theta}_{BLUE-\eta_1}$ to

Table 2
Estimations of θ and ε Using the Example Data Set

	$\hat{\theta}_\pi$	$\hat{\theta}_K$	$\hat{\theta}_{BLUEu}$	$\hat{\theta}_{BLUE-\xi_1}$	$\hat{\varepsilon}_{BLUE-\xi_1}$	$\hat{\theta}_{GLSu}$ (Assumed ε)
Cleaned	8.634	12.718	16.238	14.964	2.5×10^{-6}	16.300 (0)
Uncleaned	8.652	13.271	17.420	14.499	6.3×10^{-6}	15.452 (4.15×10^{-6})

estimate θ and then estimate p. When a sequence is very long or the sample size is very large, computational intensity may be a limitation. In this case, $\hat{\theta}_{K-\xi_1}$ or $\hat{\theta}_{K-\eta_1}$ is a good alternative.

In this paper, ε is estimated sequentially after an estimation of θ. Although this estimation is biased downward, it is reasonable when θ/L is small. It is possible to estimate ε and θ jointly using a similar iterative process as that used in the GLS estimators. That is, before we update $\hat{\theta}_{GLS}$, we update p with

$$\hat{p}_{GLS} = \left[\left(X_{01}+X_{11}\hat{\theta}\right)' \hat{V}^{-1}\left(X_{01}+X_{11}\hat{\theta}\right)\right]^{-1} \times \left(X_{01}+X_{11}\hat{\theta}\right)' \hat{V}^{-1}\left(Y-X_{10}\hat{\theta}\right), \qquad (19)$$

where $\hat{\theta}$ is the estimation of θ in the previous step. However, our simulation shows that the performance is not as good as a simple estimate of θ ignoring singletons and then estimating ε using singletons and the estimate of θ.

There are some technical issues associated with the computational complexity of the GLS estimators. One is the calculation of $\hat{V}^{-1}$ in equation (18). Depending on the numerical methods used, sometimes a true inverse of $\hat{V}$ may be hard to calculate. In our experience, when that happens, a generalized inverse can be used without noticeable problems. Another issue is the convergence of $\hat{\theta}_{GLS}$. With our limited simulations, the convergence rate is typically larger than 99.9% for most combination of parameters and never less than 99.7%.

Several java programs for calculating the GLS estimators are available upon request or can be downloaded from http://sites.google.com/site/jpopgen/.

5. Acknowledgments

This research was supported by National Institutes of Health grant 5P50GM065509-07. We thank the two anonymous reviewers for their comments and suggestions.

Appendix

Assume that originally, there are ξ_i sites of mutation size i. Under the sequencing error model, it can produce sites with configuration (mutant, ancestral, and error) $(i,n-i,0)$, $(i-1,n-i+1,0)$, $(i-1,n-i,1)$, $(i+1,n-i-1,0)$, or $(i,n-i-1,1)$. We denote the configuration as $(i+m,n-i-m-j,j)$ and its number as $X^i_{i+m,n-i-m-j,j}$, with $\sum X^i_{i+m,n-i-m-j,j} = \xi_i$. Then,

$$E\left(X^i_{i,n-i,0}|\xi,p\right) = \xi_i(1-p)$$

$$E\left(X^i_{i-1,n-i+1,0}|\xi,p\right) = \xi_i\frac{i}{n}up$$

$$E\left(X^i_{i-1,n-i,1}|\xi,p\right) = \xi_i\frac{i}{n}(1-u)p$$

$$E\left(X^i_{i+1,n-i-1,0}|\xi,p\right) = \xi_i\frac{n-i}{n}up$$

$$E\left(X^i_{i,n-i-1,1}|\xi,p\right) = \xi_i\frac{n-i}{n}(1-u)p$$

$$Var\left(X^i_{i,n-i,0}|\xi,p\right) = \xi_i(1-p)p$$

$$Var\left(X^i_{i-1,n-i+1,0}|\xi,p\right) = \xi_i\frac{i}{n}up\left(1-\frac{i}{n}up\right)$$

$$Var\left(X^i_{i-1,n-i,1}|\xi,p\right) = \xi_i\frac{i}{n}(1-u)p\left(1-\frac{i}{n}(1-u)p\right)$$

$$Var\left(X^i_{i+1,n-i-1,0}|\xi,p\right) = \xi_i\frac{n-i}{n}up\left(1-\frac{n-i}{n}up\right)$$

$$Var\left(X^i_{i,n-i-1,1}|\xi,p\right) = \xi_i\frac{n-i}{n}(1-u) \times p\left(1-\frac{n-i}{n}(1-u)p\right)$$

$$\begin{aligned} &Cov\left(X^i_{i,n-i,0},X^i_{i+1,n-i-1,0}|\xi,p\right)\\ &= E\left(X^i_{i,n-i,0}X^i_{i+1,n-i-1,0}|\xi,p\right)\\ &\quad - E\left(X^i_{i,n-i,0}|\xi,p\right)E\left(X^i_{i+1,n-i-1,0}|\xi,p\right)\\ &= \sum_{a=0}^{\xi_i} p_{\xi_i,a}\sum_{b=0}^{a}p_{a,b}\sum_{c=0}^{b}p_{b,c}\sum_{d=0}^{a-b}p_{a-b,d}(\xi_i-a)d\\ &\quad - \xi_i(1-p)\xi_i\frac{n-i}{n}up\\ &= u\xi_i(\xi_i-1)p(1-p)\left(1-\frac{i}{n}\right) - \xi_i(1-p)\xi_i\frac{n-i}{n}up\\ &= -u\xi_i p(1-p)\left(1-\frac{i}{n}\right) \end{aligned}$$

where

$$p_{\xi_i,a} = \binom{\xi_i}{a}p^a(1-p)^{\xi_i-a}$$

$$p_{a,b} = \binom{a}{b}\left(\frac{i}{n}\right)^b\left(1-\frac{i}{n}\right)^{a-b}$$

$$p_{b,c} = \binom{b}{c}(u)^c(1-u)^{b-c}$$
$$p_{a-b,d} = \binom{a-b}{d}(u)^d(1-u)^{a-b-d}.$$

Similarly, we have

$$\begin{aligned}
&Cov\left(X^i_{i,n-i,0}, X^i_{i-1,n-i+1,0}|\xi,p\right)\\
&\quad = -u\xi_i p(1-p)\frac{i}{n}\\
&Cov\left(X^i_{i,n-i,0}, X^i_{i-1,n-i,1}|\xi,p\right)\\
&\quad = -(1-u)\xi_i p(1-p)\frac{i}{n}\\
&Cov\left(X^i_{i,n-i,0}, X^i_{i,n-i-1,1}|\xi,p\right)\\
&\quad = -(1-u)\xi_i p(1-p)\left(1-\frac{i}{n}\right)\\
&Cov\left(X^i_{i-1,n-i+1,0}, X^i_{i-1,n-i,1}|\xi,p\right)\\
&\quad = -u(1-u)\xi_i p^2\left(\frac{i}{n}\right)^2\\
&Cov\left(X^i_{i-1,n-i+1,0}, X^i_{i+1,n-i-1,0}|\xi,p\right)\\
&\quad = -u^2\xi_i p^2\left(\frac{i}{n}\right)\left(1-\frac{i}{n}\right)\\
&Cov\left(X^i_{i-1,n-i+1,0}, X^i_{i,n-i-1,1}|\xi,p\right)\\
&\quad = -u(1-u)\xi_i p^2\left(\frac{i}{n}\right)\left(1-\frac{i}{n}\right)\\
&Cov\left(X^i_{i-1,n-i,1}, X^i_{i+1,n-i-1,0}|\xi,p\right)\\
&\quad = -u(1-u)\xi_i p^2\left(\frac{i}{n}\right)\left(1-\frac{i}{n}\right)\\
&Cov\left(X^i_{i-1,n-i,1}, X^i_{i,n-i-1,1}|\xi,p\right)\\
&\quad = -(1-u)^2\xi_i p^2\left(\frac{i}{n}\right)\left(1-\frac{i}{n}\right)\\
&Cov\left(X^i_{i+1,n-i-1,0}, X^i_{i,n-i-1,1}|\xi,p\right)\\
&\quad = -u(1-u)\xi_i p^2\left(1-\frac{i}{n}\right)^2.
\end{aligned}$$

Assume all ξ_is (represented by ξ) are known. Then, X^i_* and X^j_* are independent, where $*$ represents any valid configuration and $i \neq j$. So that $Cov(X^i_*, X^j_*) = 0$ when $i \neq j$. Let $\xi_{-1} = 0$ and $\xi_{n+1} = 0$, then

$$\begin{aligned}
&E\left(X^i_{i+m,n-i-m-j,j}|\xi,p\right) = a^i_{m,j}\xi_i + b^i_{m,j}\xi_i p\\
&Var\left(X^i_{i+m,n-i-m-j,j}|\xi,p\right) = c^i_{m,0,m,0}\xi_i p + d^i_{m,0,m,0}\xi_i p^2\\
&Cov\left(X^i_{i+m,n-i-m,0}, X^i_{i+k,n-i-k,0}|\xi,p\right) = c^i_{m,0,k,0}\xi_i p\\
&\qquad + d^i_{m,0,k,0}\xi_i p^2\\
&Cov\left(X^i_{i+m,n-i-m-1,1}, X^i_{i+k,n-i-k-1,1}|\xi,p\right) = c^i_{m,1,k,1}\xi_i p\\
&\qquad + d^i_{m,1,k,1}\xi_i p^2\\
&Cov\left(X^i_{i+m,n-i-m,0}, X^i_{i+k,n-i-k-1,1}|\xi,p\right) = c^i_{m,0,k,1}\xi_i p\\
&\qquad + d^i_{m,0,k,1}\xi_i p^2
\end{aligned}$$

where

$$a^i_{m,j} = \begin{cases} 1 & (m=0, j=0)\\ 0 & \text{otherwise}\end{cases}$$

$$b^i_{m,j} = \begin{cases} -1 & (m=0, j=0)\\ \frac{i}{n}u & (m=-1, j=0)\\ \frac{i}{n}(1-u) & (m=-1, j=1)\\ \frac{n-i}{n}u & (m=1, j=0)\\ \frac{n-i}{n}(1-u) & (m=0, j=1)\\ 0 & \text{otherwise}\end{cases}$$

$$c^i_{m,0,k,1} = \begin{cases} -(1-u)\frac{i}{n} & (m=0, k=-1)\\ -(1-u)\left(1-\frac{i}{n}\right) & (m=0, k=0)\\ 0 & \text{otherwise}\end{cases}$$

$$d^i_{m,0,k,1} = \begin{cases} (1-u)\frac{i}{n} & (m=0, k=-1)\\ (1-u)\left(1-\frac{i}{n}\right) & (m=0, k=0)\\ -u(1-u)\left(\frac{i}{n}\right)^2 & (m=-1, k=-1)\\ -u(1-u)\frac{i}{n}\left(1-\frac{i}{n}\right) & (m=-1, k=0)\\ -u(1-u)\frac{i}{n}\left(1-\frac{i}{n}\right) & (m=1, k=-1)\\ -u(1-u)\left(1-\frac{i}{n}\right)^2 & (m=1, k=0)\\ 0 & \text{otherwise}\end{cases}$$

$$c^i_{m,0,k,0} = \begin{cases} 1 & (m=0, k=0)\\ \frac{i}{n}u & (m=-1, k=-1)\\ \frac{n-i}{n}u & (m=1, k=1)\\ -u\left(1-\frac{i}{n}\right) & (m=0, k=1\, or\, k=0, m=1)\\ -u\frac{i}{n} & (m=0, k=-1\, or\, k=0,\\ & \quad m=-1)\\ 0 & \text{otherwise}\end{cases}$$

$$d^i_{m,0,k,0} = \begin{cases} -1 & (m=0, k=0)\\ -\left(\frac{i}{n}u\right)^2 & (m=-1, k=-1)\\ -\left(\frac{n-i}{n}u\right)^2 & (m=1, k=1)\\ u\left(1-\frac{i}{n}\right) & (m=0, k=1\, or\, k=0,\\ & \quad m=1)\\ u\frac{i}{n} & (m=0, k=-1\, or\, k=0,\\ & \quad m=-1)\\ -u^2\frac{i}{n}\left(1-\frac{i}{n}\right) & (m=-1, k=1\, or\, k=-1,\\ & \quad m=1)\\ 0 & \text{otherwise}\end{cases}$$

$$c^i_{m,1,k,1} = \begin{cases} \frac{i}{n}(1-u) & (m=-1, k=-1)\\ \frac{n-i}{n}(1-u) & (m=0, k=0)\\ 0 & \text{otherwise}\end{cases}$$

$$d^i_{m,1,k,1}=\begin{cases}-\left(\frac{i}{n}(1-u)\right)^2 & (m=-1,k=-1)\\ -\left(\frac{n-i}{n}(1-u)\right)^2 & (m=0,k=0)\\ -(1-u)^2\frac{i}{n}\left(1-\frac{i}{n}\right) & (m=-1,k=0\,or\\ & \quad k=-1,m=0).\\ 0 & \text{otherwise}\end{cases}$$

Then,

$$\begin{aligned}E(\xi_{i,0}|\xi,p)&=E\left(X^{i-1}_{i,n-i,0}|\xi,p\right)+E\left(X^{i}_{i,n-i,0}|\xi,p\right)\\&\quad+E\left(X^{i+1}_{i,n-i,0}|\xi,p\right)\\&=\xi_i+p\sum_{k=i-1}^{i+1}b^k_{i-k,0}\xi_k\,(i=0,\cdots,n)\end{aligned}$$

$$\begin{aligned}E(\xi_{i,1}|\xi,p)&=E\left(X^{i}_{i,n-i-1,1}|\xi,p\right)\\&\quad+E\left(X^{i+1}_{i,n-i-1,1}|\xi,p\right)\\&=p\sum_{k=i}^{i+1}b^k_{i-k,1}\xi_k\,(i=0,\cdots,n-1)\end{aligned}$$

$$\begin{aligned}Var(\xi_{i,0}|\xi,p)&=Var\left(X^{i-1}_{i,n-i,0}|\xi,p\right)+Var\left(X^{i}_{i,n-i,0}|\xi,p\right)\\&\quad+Var\left(X^{i+1}_{i,n-i,0}|\xi,p\right)\\&=p\sum_{k=i-1}^{i+1}c^k_{i-k,0,i-k,0}\xi_k\\&\quad+p^2\sum_{k=i-1}^{i+1}d^k_{i-k,0,i-k,0}\xi_k\,(i=0,\cdots,n)\end{aligned}$$

$$\begin{aligned}Var(\xi_{i,1}|\xi,p)&=Var\left(X^{i}_{i,n-i-1,1}|\xi,p\right)\\&\quad+Var\left(X^{i+1}_{i,n-i-1,1}|\xi,p\right)\\&=p\sum_{k=i}^{i+1}c^k_{i-k,1,i-k,1}\xi_k\\&\quad+p^2\sum_{k=i}^{i+1}d^k_{i-k,1,i-k,1}\xi_k\,(i=0,\cdots,n-1)\end{aligned}$$

$$\begin{aligned}Cov\left(\xi_{i,0},\xi_{j,1}|\xi,p\right)&=\sum_{k=i-1}^{i+1}\sum_{l=j}^{j+1}Cov(X^k_{i,n-i,0},X^l_{j,n-j-1,1}|\xi,p)\\&=p\sum_{k=i-1}^{i+1}\sum_{l=j}^{j+1}\delta_{k=l}c^k_{i-k,0,j-k,1}\xi_k\\&\quad+p^2\sum_{k=i-1}^{i+1}\sum_{l=j}^{j+1}\delta_{k=l}d^k_{i-k,0,j-k,1}\xi_k\end{aligned}$$

$$\begin{aligned}Cov\left(\xi_{i,0},\xi_{j,0}|\xi,p\right)&=\sum_{k=i-1}^{i+1}\sum_{l=j-1}^{j+1}Cov(X^k_{i,n-i,0},X^l_{j,n-j,0}|\xi,p)\\&=p\sum_{k=i-1}^{i+1}\sum_{l=j-1}^{j+1}\delta_{k=l}c^k_{i-k,0,j-k,0}\xi_k\\&\quad+p^2\sum_{k=i-1}^{i+1}\sum_{l=j-1}^{j+1}\delta_{k=l}d^k_{i-k,0,j-k,0}\xi_k\end{aligned}$$

$$\begin{aligned}Cov\left(\xi_{i,1},\xi_{j,1}|\xi,p\right)&=\sum_{k=i}^{i+1}\sum_{l=j}^{j+1}Cov(X^k_{i,n-i-1,1},X^l_{j,n-j-1,1}|\xi,p)\\&=p\sum_{k=i}^{i+1}\sum_{l=j}^{j+1}\delta_{k=l}c^k_{i-k,1,j-k,1}\xi_k\\&\quad+p^2\sum_{k=i}^{i+1}\sum_{l=j}^{j+1}\delta_{k=l}d^k_{i-k,1,j-k,1}\xi_k.\end{aligned}$$

Since ξ_is are unknown,

$$\begin{aligned}E(\xi_{i,0}|p,\theta)&=E\left[E(\xi_{i,0}|\xi,p)\right]\\&=E(\xi_i)+p\sum_{k=i-1}^{i+1}b^k_{i-k,0}E(\xi_k)\quad(i=0,\cdots,n)\end{aligned}$$

$$\begin{aligned}E(\xi_{i,1}|p,\theta)&=E\left[E(\xi_{i,1}|\xi,p)\right]\\&=p\sum_{k=i}^{i+1}b^k_{i-k,1}E(\xi_k)\quad(i=0,\cdots,n-1)\end{aligned}$$

$$\begin{aligned}Var(\xi_{i,0}|p,\theta)&=E\left[Var(\xi_{i,0}|\xi,p)\right]+Var\left[E(\xi_{i,0}|\xi,p)\right]\\&=p\sum_{k=i-1}^{i+1}c^k_{i-k,0,i-k,0}E(\xi_k)\\&\quad+p^2\sum_{k=i-1}^{i+1}d^k_{i-k,0,i-k,0}E(\xi_k)\\&\quad+\sum_{k=i-1}^{i+1}\sum_{l=j-1}^{j+1}\left(a^k_{i-k,0}+b^k_{i-k,0}p\right)\\&\quad\times\left(a^l_{j-l,0}+b^l_{j-l,0}p\right)Cov(\xi_k,\xi_l)\end{aligned}$$

$$\begin{aligned}Var(\xi_{i,1}|p,\theta)&=E\left[Var(\xi_{i,1}|\xi,p)\right]+Var\left[E(\xi_{i,1}|\xi,p)\right]\\&=p\sum_{k=i}^{i+1}c^k_{i-k,1,i-k,1}E(\xi_k)\\&\quad+p^2\sum_{k=i}^{i+1}d^k_{i-k,1,i-k,1}E(\xi_k)\\&\quad+p^2\sum_{k=i}^{i+1}\sum_{l=j}^{j+1}\left(b^k_{i-k,1}\right)\left(b^l_{j-l,1}\right)Cov(\xi_k,\xi_l)\end{aligned}$$

$$\begin{aligned}Cov\left(\xi_{i,0},\xi_{j,0}|p,\theta\right)&=E\left[Cov\left(\xi_{i,0},\xi_{j,0}|\xi,p\right)\right]\\&\quad+Cov\left[E(\xi_{i,0}|\xi,p),E\left(\xi_{j,0}|\xi,p\right)\right]\\&=p\sum_{k=i-1}^{i+1}\sum_{l=j-1}^{j+1}\delta_{k=l}c^k_{i-k,0,j-k,0}E(\xi_k)\end{aligned}$$

$$+p^2\sum_{k=i-1}^{i+1}\sum_{l=j-1}^{j+1}\delta_{k=l}d^k_{i-k,0,j-k,0}E(\xi_k)$$

$$+\sum_{k=i-1}^{i+1}\sum_{l=j-1}^{j+1}\left(a^k_{i-k,0}+b^k_{i-k,0}p\right)$$

$$\times\left(a^l_{j-l,0}+b^l_{j-l,0}p\right)Cov(\xi_k,\xi_l)$$

$$Cov\left(\xi_{i,1},\xi_{j,1}|p,\theta\right)=E\left[Cov\left(\xi_{i,1},\xi_{j,1}|\xi,p\right)\right]$$

$$+Cov\left[E\left(\xi_{i,1}|\xi,p\right),E\left(\xi_{j,1}|\xi,p\right)\right]$$

$$=p\sum_{k=i}^{i+1}\sum_{l=j}^{j+1}\delta_{k=l}c^k_{i-k,1,j-k,1}E(\xi_k)$$

$$+p^2\sum_{k=i}^{i+1}\sum_{l=j}^{j+1}\delta_{k=l}d^k_{i-k,1,j-k,1}E(\xi_k)$$

$$+p^2\sum_{k=i}^{i+1}\sum_{l=j}^{j+1}\left(b^k_{i-k,1}\right)\left(b^l_{j-l,1}\right)$$

$$\times Cov(\xi_k,\xi_l)$$

$$Cov\left(\xi_{i,0},\xi_{j,1}|p,\theta\right)=E\left[Cov\left(\xi_{i,0},\xi_{j,1}|\xi,p\right)\right]$$

$$+Cov\left[E\left(\xi_{i,0}|\xi,p\right),E\left(\xi_{j,1}|\xi,p\right)\right]$$

$$=p\sum_{k=i-1}^{i+1}\sum_{l=j}^{j+1}\delta_{k=l}c^k_{i-k,0,j-k,1}E(\xi_k)$$

$$+p^2\sum_{k=i-1}^{i+1}\sum_{l=j}^{j+1}\delta_{k=l}d^k_{i-k,0,j-k,1}E(\xi_k)$$

$$+\sum_{k=i-1}^{i+1}\sum_{l=j}^{j+1}\left(a^k_{i-k,0}+b^k_{i-k,0}p\right)$$

$$\times\left(b^l_{j-l,1}p\right)Cov(\xi_k,\xi_l).$$

Literature Cited

Achaz G. 2008. Testing for neutrality in samples with sequencing errors. Genetics. 179:1409–1424.

Achaz G, Palmer S, Kearney M, Maldarelli F, Mellors JW, Coffin JM, Wakeley J. 2004. A robust measure of HIV-1 population turnover within chronically infected individuals. Mol. Biol. Evol. 21:1902–1912.

Clark AG, Whittam TS. 1992. Sequencing errors and molecular evolutionary analysis. Mol. Biol. Evol. 9:744–752.

Crawford DC, Akey DT, Nickerson DA. 2005. The patterns of natural variation in human genes. Annu. Rev. Genomics. Hum. Genet. 6:287–312.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic. Acids. Res. 32:1792–1797.

Fu YX. 1994a. A phylogenetic estimator of effective population-size or mutation-rate. Genetics. 136:685–692.

Fu YX. 1994b. Estimating effective population size or mutation rate using the frequencies of mutations of various classes in a sample of DNA sequences. Genetics. 138:1375–1386.

Fu YX. 1995. Statistical properties of segregating sites. Theor. Popul. Biol. 48:172–197.

Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. Genetics. 133:693–709.

Hellmann I, Mang Y, Gu Z, Li P, de la Vega FM, Clark AG, Nielsen R. 2008. Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals. Genome. Res. 18:1020–1029.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics. 18: 337–338.

Jiang R, Tavaré S, Marjoram P. 2009. Population genetic inference from resequencing data. Genetics. 181:187–197.

Johnson PLF, Slatkin M. 2006. Inference of population genetic parameters in metagenomics: a clean look at messy data. Genome. Res. 16:1320–1327.

Johnson PLF, Slatkin M. 2008. Accounting for bias from sequencing error in population genetic estimates. Mol. Biol. Evol. 25:199–206.

Knudsen B, Miyamoto MM. 2007. Incorporating experimental design and error into coalescent/mutation models of population history. Genetics. 176:2335–2342.

Lynch M. 2008. Estimation of nucleotide diversity, disequilibrium coefficients, and mutation rates from high-coverage genome-sequencing projects. Mol. Biol. Evol. 25: 2409–2419.

Mackay T, Richards S, Gibbs R. 2008. Proposal to sequence a Drosophila genetic reference panel: a community resource for the study of genotypic and phenotypic variation [white paper]. FlyBase. org. Available from http://flybase.org/static_pages/news/wpapers.html. Last accessed on April 14, 2009.

Rosenbaum P, Robertson J, Zamudio K. 2007. Unexpectedly low genetic divergences among populations of the threatened bog turtle (Glyptemys muhlenbergii). Conserv. Genet. 8: 331–342.

Shendure J, Ji H. 2008. Next-generation DNA sequencing. Nat. Biotechnol. 26:1135–1145.

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

Watterson GA. 1975. Number of segregating sites in genetic models without recombination. Theor. Popul. Biol. 7:256–276.

Zhao Z, Jin L, Fu YX et al. 2000. Worldwide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22. Proc. Natl. Acad. Sci. U.S.A. 97:11354–11358.

Zwick ME. 2005. A genome sequencing center in every lab. Eur. J. Hum. Genet. 13:1167–1168.

Hideki Innan, Associate Editor

Accepted March 18, 2009