

# Accounting for Bias from Sequencing Error in Population Genetic Estimates

Philip L. F. Johnson\* and Montgomery Slatkin†

\*Biophysics Graduate Group, University of California, Berkeley and †Department of Integrative Biology, University of California, Berkeley

Sequencing error presents a significant challenge to population genetic analyses using low-coverage sequence in general and single-pass reads in particular. Bias in parameter estimates becomes severe when the level of polymorphism (signal) is low relative to the amount of error (noise). Choosing an arbitrary quality score cutoff yields biased estimates, particularly with newer, non-Sanger sequencing technologies that have different quality score distributions. We propose a rule of thumb to judge when a given threshold will lead to significant bias and suggest alternative approaches that reduce bias.

## Introduction

Ever since the advent of high-throughput sequencing with the Human Genome Project, investigators have faced a trade-off between data quality and cost of acquisition (Bouck et al. 1998; Olson and Green 1998). Too much error will swamp the true signal of genetic diversity, but too little data will lead to increased variance in parameter estimates.

Fifteen years ago, Clark and Whittam (1992) recognized that sequencing error could be a problem for evolutionary analysis, and other groups tackled error in the context of alignments (States 1992) and genetic maps (Lincoln and Lander 1992); however, sequencing analysis software and technology have since changed in 2 key ways. First, standard algorithms now provide accurate estimates of the probability that a given base call is incorrect (Ewing and Green 1998). Second, newer, non-Sanger technologies such as pyrosequencing (Margulies et al. 2005) and oligonucleotide array-based resequencing (Frazer et al. 2004) have different error rate distributions than capillary-based Sanger sequencers.

One widely used technique for maintaining sequence quality involves picking an arbitrary quality cutoff, discarding all bases with quality below the threshold, and, in some studies, verifying the remaining data by manual inspection. Sometimes, this strategy is employed in tandem on forward and reverse reads that cover the same base twice (e.g., Brown et al. 2004; Rynänen and Primmer 2004). In this case, the effective threshold will range between the single-read threshold probability and the square of that probability, depending on the degree to which errors in the 2 reads are independent.

From a quantitative standpoint, eliminating bases below a threshold is equivalent to censoring a portion of the data (machine-called low quality plus human-called low quality) and then trusting what remains. If we assume that quality scores are independent of base identity, this censoring produces data with values missing at random and will not affect parameter estimates (Little and Rubin 2002). However, the second step of trusting what remains creates a more fundamental problem in that parameter estimates made on the basis of these data will be biased by any remaining error. In particular, error will often cause sites that are truly nonpolymorphic to appear polymorphic but

with only a single chromosome having a different allele. Estimators that are particularly sensitive to such singleton sites (e.g., Watterson's estimator of  $\theta$ ; 1975) thus will be particularly susceptible to bias.

Here, we apply simple analytic models to determine when this threshold strategy is acceptable versus when it leads to substantially biased results. In the latter situation, a modified estimation framework that incorporates quality scores may be able to recover an unbiased estimator (Johnson and Slatkin 2006). Although we refer to sequencing error in this report, the general principle holds for any type of error leading to base miscalls, including ancient DNA degradation (Pääbo 1989; Briggs et al. 2007) or phantom mutation hotspots (Brandstatter et al. 2005). We start by analyzing output from 3 sequencing technologies (traditional Sanger, pyrosequencing by 454 Life Sciences, and microarray data) to find their quality score distributions. Given these distributions, we quantify the bias in 4 estimators caused by using a threshold score with each of these technologies.

## Methods

First, we present theory that describes how a number of parameters are affected by error. Then, we show how the theory can be used to create improved estimators. Finally, we review data from 3 major sequencing technologies to determine their error distributions.

## Theory

The underlying concept behind our calculations is straightforward: error can cause truly fixed (nonpolymorphic) sites to appear segregating (polymorphic) and vice versa. However, error can also change one polymorphic site into a different polymorphic site by altering the frequency at which the 2 alleles appear in the sample. Estimators that depend on the allele frequency spectrum must take this last type of change into account along with the more obvious fixed-to-polymorphic and polymorphic-to-fixed changes.

We split our observed data between apparent fixed sites and apparent polymorphic sites, without subdividing the polymorphic sites on the basis of how many alleles are observed. Although apparent polymorphic sites with more than 2 alleles are likely the result of at least one error, they are also likely to be truly polymorphic. Thus, these polyallelic sites form a biased subsample of all sites, and a simple strategy of discarding such sites will lead to bias in parameter estimates. We avoid this problem and simplify

Key words: sequencing error, population genetics, bias, quality score.

E-mail: plfjohnson@berkeley.edu.

*Mol. Biol. Evol.* 25(1):199–206. 2008

doi:10.1093/molbev/msm239

Advance Access publication November 2, 2007

the theory by considering bi-, tri-, and quadallelic sites all to be segregating.

We highlight 4 estimators of 3 basic parameters: Watterson's estimator of  $\theta$  (1975), Tajima's estimator of  $\theta$  (1983), Tajima's  $D$  (1989), and Weir and Cockerham's estimator of  $F_{ST}$  (1984). The expected value of each estimator is calculated in turn, given the true value of the parameter and the distribution of quality scores. For some parameters, we use coalescent simulations instead of analytic theory. We assume that an error causes a switch to any of the other 3 nucleotides with equal probability.

First, we outline our notation. We define  $\theta = 4N_e\mu$ , where  $N_e$  is the effective population size and  $\mu$  is the mutation rate per nucleotide per generation. Note, this means  $\theta$  is a per-site measure rather than a per-locus measure. A phred style (Ewing and Green 1998) quality score,  $Q$ , for a single nucleotide maps directly to an error probability via the relation  $\Pr(\text{error}) = 10^{-Q/10}$ , so we will henceforth refer to error probabilities rather than quality scores. The distribution of error probabilities is represented by  $G$ , with the average per-site error probability  $\varepsilon = \mathbb{E}[G]$ . Observed values incorporating error are distinguished from true values without error by the subscripts  $o$  and  $t$  (i.e.,  $S_o$  and  $S_t$  represent observed and true numbers of segregating sites). Our sample consists of  $n$  chromosomes, each of which is  $L$  nucleotides long.

#### Tajima's $\hat{\theta}_\pi$

This estimator of  $\theta$  uses the average number of pairwise nucleotide differences (often denoted  $\pi$ ) between sequences in a sample of size  $n$  (Tajima 1983):

$$\hat{\theta}_\pi = \frac{2}{n(n-1)} \sum_{i < j}^n \pi_{ij},$$

where  $\pi_{ij}$  is the number of differences per nucleotide between sequences  $i$  and  $j$ . Without error, this is an unbiased estimator ( $\mathbb{E}[\hat{\theta}_\pi] = \theta$ ). With error:

$$\begin{aligned} \mathbb{E}[\hat{\theta}_\pi | \theta, G] &= \frac{2}{n(n-1)} \sum_{i < j}^n \mathbb{E}[\pi_{ij,o} | \theta, G] \\ &= \mathbb{E}[\pi_{ab,o} | \theta, G]. \end{aligned}$$

Because the  $\pi_{ij}$  are identically distributed for all  $i \neq j$ ,  $a$  and  $b$  represent arbitrary sequences where  $a \neq b$ . We write the observed differences ( $\pi_{ab,o}$ ) as a function of the true differences ( $\pi_{ab,t}$ ), those sites that were originally mismatched and error made matching ( $X_k$ ) and those sites that were originally matching and error made mismatch ( $Y_k$ ). To maintain  $\pi$  (and thereby  $\theta$ ) as a per-site measure, we divide by  $L$ :

$$\pi_{ab,o} = \pi_{ab,t} - \sum_{k=1}^m X_k/L + \sum_{k=m+1}^L Y_k/L,$$

where  $m$  is the number of true mismatches between 2 sequences,  $X_k \sim \text{Bernoulli}(p_1)$ ,  $Y_k \sim \text{Bernoulli}(p_2)$ ,  $p_1 = \Pr(\text{observed match} | \text{true mismatch}, \{g_1, g_2\})$ ,  $p_2 = \Pr(\text{observed mismatch} | \text{true match}, \{g_1, g_2\})$ , and  $g_1, g_2 \sim G$ . Note that the probabilities  $p_1$  and  $p_2$  are random variables in their own right that depend on the error probabilities,  $g_1$  and  $g_2$ ,

at a particular site. Continuing our derivation and using Wald's equation (Ross 1996),

$$\begin{aligned} \mathbb{E}[\hat{\theta}_\pi | \theta, G] &= \mathbb{E}[\pi_{ab,t} - \sum_{k=1}^m X_k/L + \sum_{k=m+1}^L Y_k/L | \theta, G] \\ &= \theta - (1/L)\mathbb{E}[m|\theta]\mathbb{E}[X_k|G] \\ &\quad + (1/L)\mathbb{E}[L - m|\theta]\mathbb{E}[Y_k|G] \\ &= \theta - \theta\mathbb{E}[X_k|G] + (1 - \theta)\mathbb{E}[Y_k|G] \\ &= \theta - \theta\mathbb{E}[\mathbb{E}[X_k|p_1]|G] + (1 - \theta)\mathbb{E}[\mathbb{E}[Y_k|p_2]|G] \\ &= \theta(1 - \mathbb{E}[p_1|G]) + (1 - \theta)\mathbb{E}[p_2|G]. \end{aligned} \tag{1}$$

The match-given-mismatch event of  $p_1$  can arise either when exactly one base switches to the other  $(1 - g_1)g_2(1/3) + (1 - g_2)g_1(1/3)$  or when both bases switch to the same new nucleotide  $g_1g_2(2/3)(1/3)$ :

$$\begin{aligned} p_1 &= (1 - g_1)g_2(1/3) + (1 - g_2)g_1(1/3) \\ &\quad + g_1g_2(2/3)(1/3). \end{aligned}$$

The mismatch-given-match event of  $p_2$  can arise either when exactly one base switches to another  $(1 - g_1)g_2(3/3) + (1 - g_2)g_1(3/3)$  or when both bases switch to different nucleotides  $g_1g_2(3/3)(2/3)$ :

$$p_2 = (1 - g_1)g_2 + (1 - g_2)g_1 + g_1g_2(2/3).$$

Going back to equation (1), we can now calculate the expectations of  $p_1$  and  $p_2$ . Because we assume that the error probabilities are independent, these expectations depend only on the average error probability,  $\varepsilon$ :

$$\begin{aligned} \mathbb{E}[p_1|G] &= (2/3)\varepsilon(1 - \varepsilon) + (2/9)\varepsilon^2, \\ \mathbb{E}[p_2|G] &= 2\varepsilon(1 - \varepsilon) + (2/3)\varepsilon^2. \end{aligned} \tag{2}$$

#### Watterson's $\hat{\theta}_S$

This estimator of  $\theta$  uses the number of segregating sites in a sample of size  $n$  (Watterson 1975):

$$\begin{aligned} \hat{\theta}_S &= S / (La_{1,n}), \\ a_{1,n} &= \sum_{i=1}^{n-1} 1/i, \end{aligned}$$

where  $S$  is the number of segregating sites. Without error, this is also an unbiased estimator. With error:

$$\begin{aligned} \mathbb{E}[\hat{\theta}_S | \theta, n, G] &= \mathbb{E}[S_o / (La_{1,n}) | \theta, G] \\ &= (1 / (La_{1,n})) \mathbb{E}[S_o | \theta, n, G] \\ &= (1 / a_{1,n}) \{ \theta(1 - \mathbb{E}[q_1 | n, G]) \\ &\quad + (1 - \theta) \mathbb{E}[q_2 | n, G] \}. \end{aligned} \tag{3}$$

The derivation for  $\mathbb{E}[S_o|\theta, n, G]$  is analogous to the derivation of equation (1) with 2 differences. First,  $S_o$  refers to the actual number of segregating sites, so its expected value does not have the factor  $(1/L)$ . Second, the match/mismatch probabilities (now denoted  $q_1$  and  $q_2$ ) are conditional on the sample size,  $n$ , and defined within the context of segregating sites such that  $q_1 = \Pr(\text{observed fixed}|\text{true segregating}, n, \{g_k\})$  and  $q_2 = \Pr(\text{observed segregating}|\text{true fixed}, n, \{g_k\})$ , where  $\{g_k\}$  is a set of  $n$  error probabilities each distributed according to  $G$ . For the reason noted earlier, we do not distinguish between sites observed to segregate with 2 alleles versus those observed with 3 or 4 alleles.

We start by calculating  $q_1$ :

$$q_1 = \sum_{i=1}^{n-1} \Pr(\text{fixed}|\text{segregating at frequency } i, n, \{g_k\})f_n(i),$$

where  $f_n(i)$  is the probability of a mutation segregating at frequency  $i$  in a sample of size  $n$  (i.e., the frequency spectrum):  $f_n(i) = (1/i)/a_{1,n}$ . Taking the first term in the above summation, we observe a fixed site given a particular frequency of true segregation if 1 of 3 events occurs: all of one type switch to the other  $\prod_{k=1}^i (g_k/3) \prod_{k=i+1}^n (1-g_k)$  or vice versa  $\prod_{k=1}^i (1-g_k) \prod_{k=i+1}^n (g_k/3)$  or all bases of both types switch to a new, third allele  $(2/3)(1/3)^{n-1} \prod_{k=1}^n g_k$ .

The  $q_2$  event (segregating-given-fixed) will arise if any error occurs  $1 - \prod_{k=1}^n (1-g_k)$ , except for the case when all bases switch to the same new allele  $(3/3)(1/3)^{n-1} \prod_{k=1}^n g_k$ .

As with equation (1), we assume the error probabilities are independent, so the expected values of  $q_1$  and  $q_2$  depend only on the average error probability,  $\varepsilon$ :

$$\begin{aligned} \mathbb{E}[q_1|n, G] &= (1/a_{1,n}) \sum_{i=1}^{n-1} (1/i) [(\varepsilon/3)^i (1-\varepsilon)^{n-i} \\ &\quad + (1-\varepsilon)^i (\varepsilon/3)^{n-i} + (2\varepsilon/3)(\varepsilon/3)^{n-1}], \\ \mathbb{E}[q_2|n, G] &= 1 - (1-\varepsilon)^n - \varepsilon(\varepsilon/3)^{n-1}. \end{aligned} \quad (4)$$

*Variance of Watterson's  $\hat{\theta}_S$*

Watterson (1975) derived the variance in his estimator of  $\theta$  without error, given a finite number of sites,  $L$ , and sample size of  $n$ :

$$\begin{aligned} \text{Var}[\hat{\theta}_S|L, n, \theta] &= \left(\frac{1}{La_{1,n}}\right)^2 \text{Var}[S|L, n, \theta] \\ &= (La_{1,n})^{-2} (\theta La_{1,n} + (\theta L)^2 a_{2,n}). \\ a_{2,n} &= \sum_{i=1}^{n-1} 1/i^2, \end{aligned}$$

We use the notation from the  $\hat{\theta}$  expectation calculations above. With error:

$$\text{Var}[\hat{\theta}_S|G, L, n, \theta] = (La_{1,n})^{-2} \text{Var}[S_o|G, L, n, \theta]. \quad (5)$$

Remember,

$$S_o = S_t - \sum_{i=1}^{S_t} X_i + \sum_{i=S_t+1}^L Y_i,$$

where  $X_i \sim \text{Bernoulli}(q_1)$  and  $Y_i \sim \text{Bernoulli}(q_2)$ . Now we condition on the true number of segregating sites,  $S_t$ , and, for brevity, no longer explicitly write the conditionals on  $G, L, n, \theta$ :

$$\text{Var}[S_o|G, L, n, \theta] = \text{Var}[\mathbb{E}[S_o|S_t]] + \mathbb{E}[\text{Var}[S_o|S_t]]. \quad (6)$$

Working with the first term:

$$\begin{aligned} \text{Var}[\mathbb{E}[S_o|S_t]] &= \text{Var}[S_t - S_t \mathbb{E}[\mathbb{E}[X_i|q_1]] + (L - S_t) \mathbb{E}[\mathbb{E}[Y_i|q_2]]] \\ &= \text{Var}[S_t(1 - \mathbb{E}[q_1]) + (L - S_t) \mathbb{E}[q_2]] \\ &= (1 - \mathbb{E}[q_1] - \mathbb{E}[q_2])^2 \text{Var}[S_t] \\ &= (1 - \mathbb{E}[q_1] - \mathbb{E}[q_2])^2 (\theta La_{1,n} + (\theta L)^2 a_{2,n}). \end{aligned} \quad (7)$$

Working with the second term from equation (6), we use the fact that errors are independent of each other and move the variance inside the summations:

$$\mathbb{E}[\text{Var}[S_o|S_t]] = \mathbb{E}\left[0 + \sum_{i=1}^{S_t} \text{Var}[X_i] + \sum_{i=S_t+1}^L \text{Var}[Y_i]\right]. \quad (8)$$

We calculate  $\text{Var}[X_i]$  by further conditioning on  $q_1$ :

$$\begin{aligned} \text{Var}[X_i] &= \text{Var}[\mathbb{E}[X_i|q_1]] + \mathbb{E}[\text{Var}[X_i|q_1]] \\ &= \text{Var}[q_1] + \mathbb{E}[q_1(1-q_1)] \\ &= \mathbb{E}[q_1^2] - \mathbb{E}[q_1]^2 + \mathbb{E}[q_1] - \mathbb{E}[q_1^2] \\ &= \mathbb{E}[q_1](1 - \mathbb{E}[q_1]). \end{aligned}$$

The derivation of the variance of  $Y_i$  follows identically to yield  $\mathbb{E}[q_2](1 - \mathbb{E}[q_2])$ . Continuing from equation (8) and by Wald's equation (Ross 1996):

$$\begin{aligned} \mathbb{E}[\text{Var}[S_o|S_t]] &= \mathbb{E}[S_t] \mathbb{E}[q_1](1 - \mathbb{E}[q_1]) \\ &\quad + \mathbb{E}[L - S_t] \mathbb{E}[q_2](1 - \mathbb{E}[q_2]) \\ &= \theta La_{1,n} \mathbb{E}[q_1](1 - \mathbb{E}[q_1]) \\ &\quad + (L - \theta La_{1,n}) \mathbb{E}[q_2](1 - \mathbb{E}[q_2]). \end{aligned} \quad (9)$$

We calculate the overall variance of  $\hat{\theta}_S$  by starting with equation (5) and substituting, in turn, equations (6, 7, 9, and 4).

*Tajima's  $D$*

Tajima's  $D$  statistic is defined as the normalized difference between the 2 estimators of  $\theta$  for a given length of sequence,  $L$ :

$$D = \frac{\hat{\theta}_\pi - \hat{\theta}_S}{C_{S,L,n}},$$

where  $C_{S,L,n}$  is defined such that, under neutrality,  $\mathbb{E}[D]=0$  and  $\text{Var}[D]=1$  (Tajima 1989). With error:

$$\mathbb{E}[D|\theta, G] = \mathbb{E}\left[\frac{\hat{\theta}_\pi - \hat{\theta}_S}{C_{S_o,L,n}}|\theta, n, G\right].$$

We approximate this expectation by taking the limit as the length of the sequence goes to infinity (which means  $S \rightarrow \infty$  as well, although  $S/L \rightarrow \text{constant}$ ) and using a first-order

Taylor series expansion:

$$\approx \frac{\mathbb{E}[\hat{\theta}_\pi|\theta, G] - \mathbb{E}[\hat{\theta}_S|\theta, n, G]}{\mathbb{E}[C_{S_o, \infty, n}|\theta, n, G]}. \quad (10)$$

Now we assume neutrality and replace the numerator with the expected values calculated earlier (eqs. 1 and 3). The denominator reduces to be proportional to  $\mathbb{E}[S_o|\theta, n, G]$ , which we also know from equation (3).

We evaluate our approximations via coalescent simulations (Hudson 2002) to which we add “sequencing errors” with fixed probability  $\varepsilon$ .

### Wright's $F_{ST}$

$F_{ST}$  is defined as the proportion of heterozygosity found between subpopulations:

$$F_{ST} = \frac{T - W}{T},$$

where  $W$  is the average within-subpopulation heterozygosity and  $T$  is the total population heterozygosity.

We use Weir and Cockerham's (1984) estimator for  $r$  subpopulations, each of size  $n$ :

$$\hat{F}_{ST} = \frac{s^2 - \frac{1}{n-1}(\bar{p}(1 - \bar{p}) - \frac{r-1}{r}s^2)}{\bar{p}(1 - \bar{p}) + s^2/r}, \quad (11)$$

where  $\bar{p}$  is the average allele frequency across the total sample and  $s^2$  is the estimated variance in allele frequency across the subpopulation samples:

$$s^2 = \sum_{i=1}^r (p_i - \bar{p})^2 / (r - 1).$$

We combine information from multiple sites by summing the numerator and denominator across sites, as suggested by Weir and Cockerham (1984).

The expected value of  $\hat{F}_{ST}$  with error is difficult to calculate because it is a ratio of random variables. We again turn to simulations (Hudson 2002) with  $n = 10$  samples from each of  $r = 10$  island subpopulations with symmetric migration rate  $M = 4N_e m$ . Given this idealized situation,  $F_{ST} \approx 1/(1 + M(r - 1)/r)$  (Cockerham and Weir 1987), so simulating for a range of values of  $M$  corresponds to a range of values for  $F_{ST}$ .

### Example Unbiased Estimators

Given the average amount of error remaining in an observed data set,  $\varepsilon$ , and the exact analytic derivations for  $\hat{\theta}_S$  and  $\hat{\theta}_\pi$ , we can construct unbiased estimators that take into account sequencing error by inverting these equations:

$$\hat{\theta}_\pi = \frac{\frac{2}{n(n-1)} \sum_{i < j} \pi_{ij_o} - \mathbb{E}[p_2]}{1 - \mathbb{E}[p_1] - \mathbb{E}[p_2]}, \quad (12)$$

substituting equation (2) for  $\mathbb{E}[p_1]$  and  $\mathbb{E}[p_2]$ .

$$\hat{\theta}_S = \left( \sum_{i=1}^{n-1} 1/i \right)^{-1} \frac{S_o/L - \mathbb{E}[q_2]}{1 - \mathbb{E}[q_1] - \mathbb{E}[q_2]}, \quad (13)$$

substituting equation (4) for  $\mathbb{E}[q_1]$  and  $\mathbb{E}[q_2]$ .

### Data

We calculated quality score distributions arising from 3 different sequencing technologies (traditional Sanger, pyrosequencing denoted “454,” microarray denoted “Chip”) by downloading publicly available data from the National Center for Biotechnology Information (NCBI) Trace Archive (<http://www.ncbi.nlm.nih.gov/Traces>) on 17 May 2007. To ensure a fair comparison, we took the Sanger and 454 data from the same sequencing center (Joint Genome Institute [JGI]) and the same species (monkey flower) using the queries “SPECIES\_CODE = ‘MIMULUS GUTTATUS’ AND CENTER\_NAME = ‘JGI’ AND TRACE\_TYPE\_CODE = ‘WGS’” and “SPECIES\_CODE = ‘MIMULUS GUTTATUS’ AND CENTER\_NAME = ‘JGI’ AND TRACE\_TYPE\_CODE = ‘454.’” We limited the Sanger query to the first 200,000 reads to reduce the load on the server. The only Chip project currently in the Trace Archive is from the Center for Rodent Genetics/Perlegen's resequencing of 15 mouse strains, and we downloaded the first 200,000 reads from that project with the query “TRACE\_TYPE\_CODE = ‘CHIP.’”

### Results

Figure 1 shows histograms of the quality scores from the 3 types of sequences (Sanger, 454, Chip) downloaded from the NCBI Trace Archive (see Methods). These distributions represent all scores found in the reads, without any trimming. Clearly these distributions have different shapes, which means that, given a fixed minimum quality score, the amount of error remaining in the data will depend on the sequencing technology. Because the 454 and Chip distributions are skewed toward lower quality, these 2 methods will have more error remaining than Sanger reads. For example, taking the typical threshold value of  $Q = 30$  ( $\equiv g = 1/1,000$  chance of an error), we find that the mean error remaining after discarding data with quality scores less than 30 is approximately  $\varepsilon = 1/10,000$  for Sanger,  $\varepsilon = 3/10,000$  for 454, and  $\varepsilon = 5/10,000$  for Chip. For the analyses below, we take each base in a given individual to be sequenced only once and apply these mean error rates directly. Note that, because the relationship between quality score,  $Q$ , and error probability,  $g$ , is nonlinear ( $g = 10^{-Q/10}$ ; Ewing and Green 1998), the mean error,  $\varepsilon = \mathbb{E}[g]$ , is not the same as the error corresponding to the mean quality score.

Given these remaining rates of error, we want to quantify the bias in estimates of population genetic parameters. First, we analyze 2 estimators of the scaled mutation rate,  $\theta$ , by comparing the expected value of each estimator with and without error for a sample size of  $n = 10$  (figs. 2A and B). In a neutrally evolving, random-mating population without error, both the estimator based on pairwise differences,  $\hat{\theta}_\pi$  (Tajima 1983), and the estimator based on segregating sites,

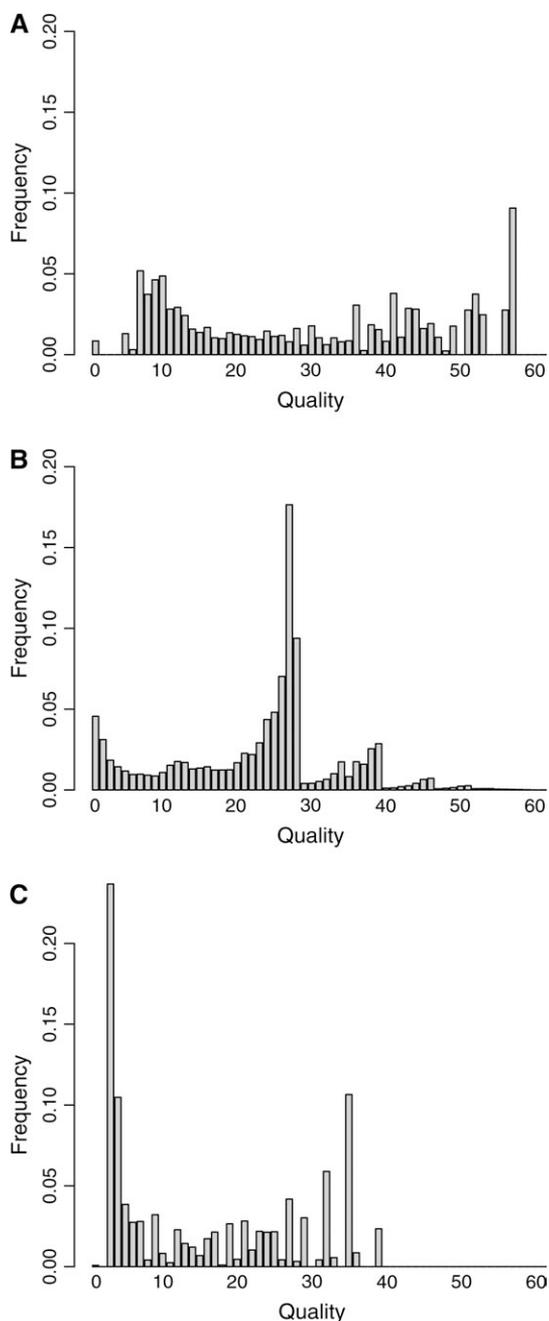


FIG. 1.—Quality score distributions for (A) Sanger, (B) 454 (pyrosequencing), and (C) Chip (Perlegen microarray). Qualities for 454 were truncated at 60 ( $\Pr(>60) = 1.2 \times 10^{-4}$ ) for ease of comparison. The multimodal nature of the 454 distribution is not exclusive to our data from JGI. A similar distribution arises from 454 sequencing of a mammoth bone conducted in a different laboratory (Poinar et al. [2006]; data not shown).

$\hat{\theta}_S$  (Watterson 1975), are unbiased, so this comparison is equivalent to comparing the estimator with error to the true value of  $\theta$ . Sequencing error has the potential to affect the latter estimator (fig. 2B) more than the former (fig. 2A) because the number of segregating sites does not distinguish between a site segregating because of a single individual in the sample (i.e., a singleton, likely caused by an error)

versus a site segregating at a higher frequency in the sample. As a result, the expected value of  $\hat{\theta}_S$  increases as the sample size,  $n$ , increases (see eq. 4 in Methods; note  $n = 10$  in fig. 2B).

Next, we combine these 2 estimators to calculate Tajima's  $D$ , a statistic that tests the null hypothesis of neutrality by taking the normalized difference between  $\hat{\theta}_\pi$  and  $\hat{\theta}_S$  (fig. 2C). Our analytic derivation (eq. 10; shown with lines in fig. 2C) makes several approximations (see Methods) but qualitatively matches simulation results (shown with points). The threshold for assessing a significant departure from neutrality is relatively strict (0.05 level of significance illustrated by the horizontal dotted line in fig. 2C), so error will not make  $D$  appear artificially significant unless the true  $\theta$  is extremely low.

Finally, we look at a measure of population subdivision,  $F_{ST}$  (Wright 1951), that compares heterozygosity within putative subpopulations versus the total population (fig. 2D). An  $F_{ST}$  value of 1 corresponds to fixed differences between subpopulations, whereas an  $F_{ST}$  value of 0 corresponds to the same allele frequencies in all subpopulations. In theory, error could obscure true subdivision (decreasing  $\hat{F}_{ST}$ ) by changing sites with fixed differences into sites with polymorphism within each subpopulation—the same outcome that would result if the subpopulations were truly interbreeding. However, error could also enhance true subdivision (increasing  $\hat{F}_{ST}$ ) by changing sites with the same allele frequencies across populations into sites with different allele frequencies across populations. The net effect of error will depend on the amount of true subdivision, the amount of mutation, and the amount of error. Our simulations of  $n = 10$  samples from each of  $r = 10$  subpopulations for migration rates ranging from 1 to 10 (corresponding to  $F_{ST}$  from ca. 0.5 to 0.1) resulted in error exclusively causing a decrease in  $\hat{F}_{ST}$ . For the parameters we explored, the relative amount of bias in  $\hat{F}_{ST}$  appears to depend only on  $\theta$  and not on the true  $F_{ST}$ .

## Discussion

The precise boundaries of the bias zone depend on the particular parameter being estimated, the form of the estimator, and the distribution of quality scores. As illustrated by the 2 estimators of  $\theta$ , a greater sensitivity toward singletons leads to increased bias in  $\hat{\theta}_S$  over  $\hat{\theta}_\pi$ . The estimator based on segregating sites,  $\hat{\theta}_S$ , generally has lower variance than the one based on pairwise differences,  $\hat{\theta}_\pi$  (Watterson 1975; Tajima 1983); however, the unequal bias introduced by sequencing error can lead to a higher mean squared error ( $= \text{bias}^2 + \text{var}$ ) for  $\hat{\theta}_S$ —unexpectedly making  $\hat{\theta}_\pi$  the preferred estimator (fig. 3A). As the sample size increases beyond the  $n = 10$  used here, the bias in  $\hat{\theta}_S$  increases as well, whereas the bias of  $\hat{\theta}_\pi$  remains constant. When we take the difference between the 2 estimators in the form of Tajima's  $D$ , it is less affected by error because both estimators are biased in the same direction. The relative bias in  $\hat{F}_{ST}$  depends primarily on the total population heterozygosity (i.e.,  $\theta$ ) rather than on the amount of subdivision, a fact that we discuss below. Although this relative bias in  $\hat{F}_{ST}$  can be substantial, its affect on the conclusion of subdivision

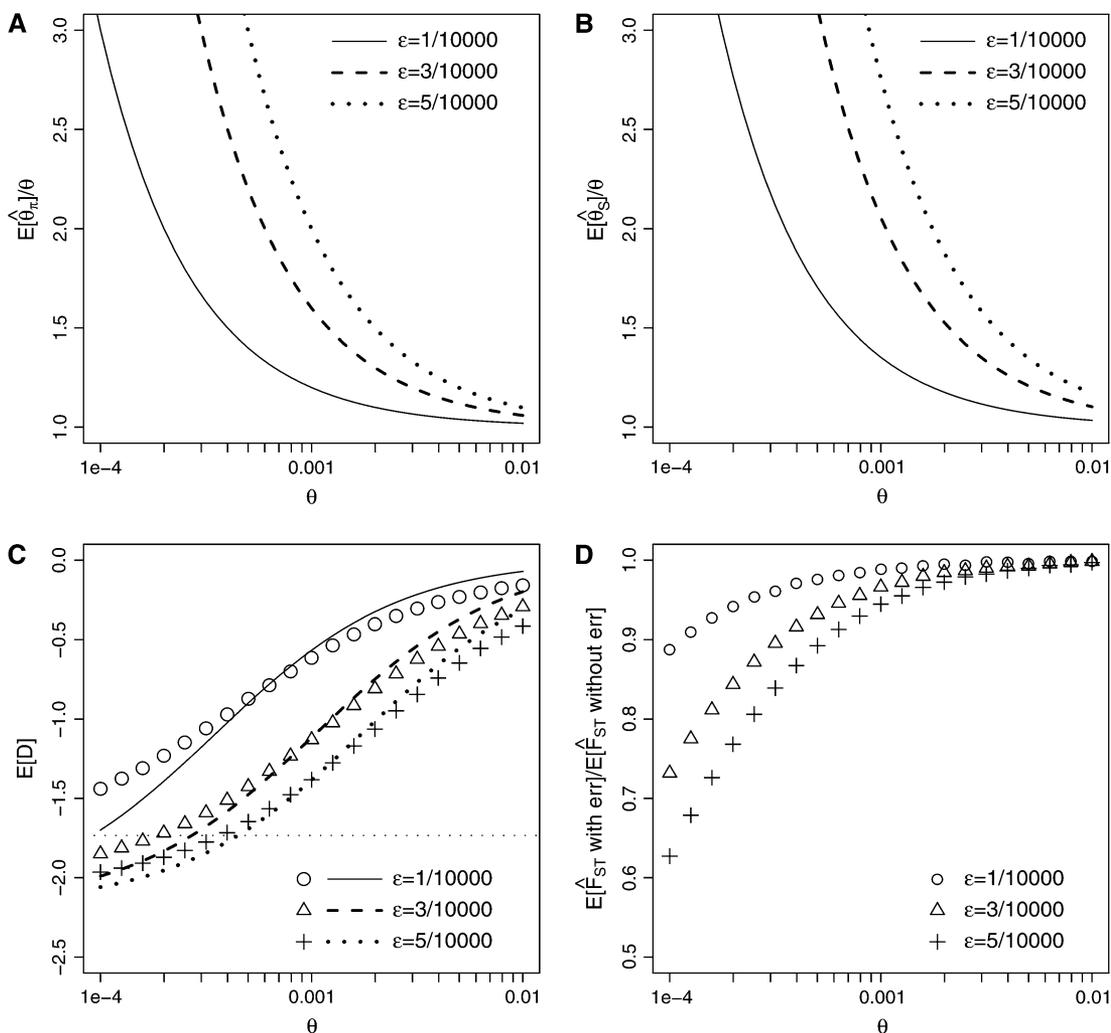


FIG. 2.—Expected value of parameter estimates with various levels of error (corresponding to Sanger, 454, Chip with quality threshold of 30) compared with expected value without error. (A)  $\hat{\theta}_\pi$  with error (analytic results) relative to without ( $=\theta$ ); expectation is independent of sample size,  $n$ , and sequence length,  $L$ . (B)  $\hat{\theta}_S$ , where  $n = 10$  (analytic results) with error relative to without ( $=\theta$ ); expectation is independent of  $L$ . (C) Tajima's  $D$  where true  $D = 0$ ,  $n = 10$ , and  $L = 100$  kb (lines show approximate analytic results; symbols show average of 100,000 replicate simulations). Dotted line represents 5% level of significance for test of neutrality from beta distribution of Tajima (1989). (D)  $F_{ST}$  with error relative to without as calculated from 10 samples from each of 10 subpopulations, where  $L = 10$  kb and migration  $M = 4N_e m = 1$  (average of 100,000 replicate simulations). Results shown for migration parameter  $M = 4N_e m = 1$ , although no qualitative difference was seen in the tested range ( $M$  from 1 to 10).

(or no subdivision) depends on the absolute value of  $F_{ST}$ , which depends, in turn, on the within-subpopulation heterozygosity.

The behavior of  $\hat{F}_{ST}$  with error warrants further discussion. Error tends to add minimally polymorphic sites where a single base in a single subpopulation has a different allele; when we apply equation (11) (see Methods) to such a site, we find 0 in the numerator and  $1/(nr)$  in the denominator. When  $m$  such sites are combined with more normal polymorphic sites, the true numerator remains unchanged ( $m \cdot 0 = 0$ ), whereas the true denominator increases by  $m/(nr)$ . Thus, the relative change in  $\hat{F}_{ST}$  depends only on the magnitude of  $m/(nr)$  versus the true denominator. Because the true denominator includes both the within- and between-subpopulation variances (Weir and Cockerham 1984), it is not directly affected by the true amount of subdivision, which explains why the relative amount of bias appears unaffected by the true  $F_{ST}$ .

Given a function that calculates the expected value of an estimator with error, inverting it should yield an unbiased estimator. When the function has a closed form, as in the case of  $\hat{\theta}_\pi$  and  $\hat{\theta}_S$ , this process for creating an unbiased estimator is straightforward (see eqs. 12 and 13). When the expected value with error cannot be calculated analytically, simulations can be used to approximate the expected value across a wide range of parameters, and then the resulting table of values can be inverted and interpolated to arrive at an approximate unbiased estimator for a given observation. However, all these unbiased estimators will still contain the variance from sequencing error in addition to the variance from the genealogical process. In the case of our  $\theta$  example, this additional variance means that the unbiased  $\hat{\theta}_\pi$  will still sometimes be preferred over the unbiased  $\hat{\theta}_S$  (fig. 3B).

The diversity of different populations spans orders of magnitude, but, in general, sequencing error will only be

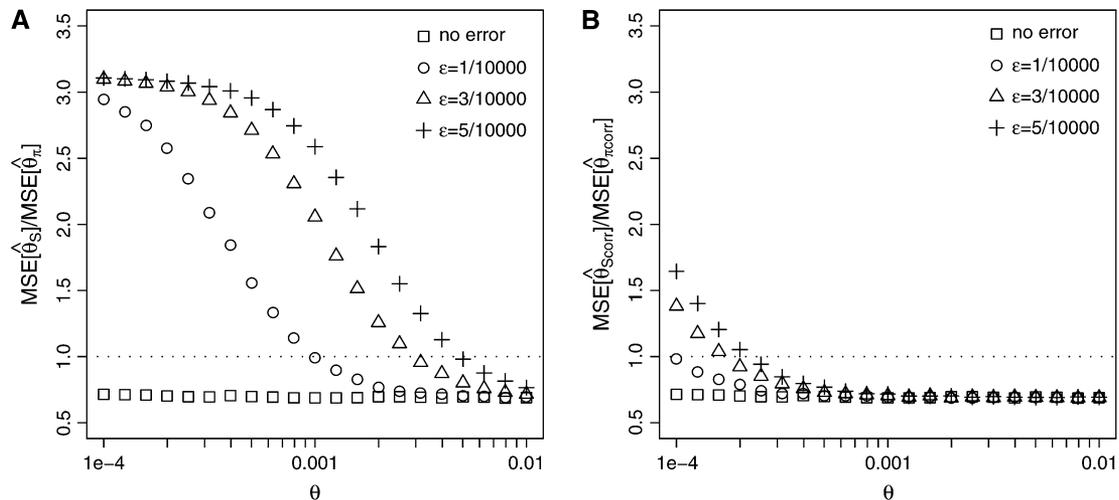


FIG. 3.—Mean squared error of  $\hat{\theta}_S$  relative to  $\hat{\theta}_\pi$  for  $n = 10$  sequences 100 kb in length calculated from 100,000 replicate simulations. Values above 1 (the dotted line) indicate  $\hat{\theta}_\pi$  is preferred, and values below indicate  $\hat{\theta}_S$  is preferred. (A) Comparison for standard estimators. (B) Comparison for new, unbiased estimators (eqs. 12 and 13).

a concern when looking at within-species diversity. Estimates of human diversity fall a little below  $\theta = 0.001$  (Crawford et al. 2005), but species with larger or smaller effective population sizes will have correspondingly larger or smaller  $\theta$ . For instance, endangered species can have extremely low diversity as a result of inbreeding (e.g., Atlantic salmon  $\theta \approx 5 \times 10^{-4}$  [Ryynänen and Primmer 2004] and bog turtle  $\theta \approx 10^{-4}$  [Rosenbaum et al. 2007]). Microbial populations, on the other hand, have enormous population sizes with the potential for high diversity (Allen et al. 2007), although near-clonal populations with very low diversity can also exist (Strous et al. 2006).

As a rule of thumb, an uncorrected estimate will be biased significantly if  $n\varepsilon \geq \theta$ , where  $n$  is the sample size,  $\varepsilon$  is the average error remaining in the data, and  $\theta$  is defined on a per-site basis (if  $\theta$  is defined per-locus, then  $n\varepsilon L \geq \theta$ ). However, estimators that focus on singletons as a means of detecting selection (e.g., Fu and Li's  $D$  [1993]) or population growth will encounter trouble much earlier. Given low diversity, a particular problem arises from “single-pass” data in which each nucleotide is sequenced at most once from a particular individual. Metagenomics and ancient DNA projects both fit into this category, along with any high-throughput sequencing where experiments cannot be repeated, either because of expense or insufficient quantity of biological sample.

The proper way to avoid bias is to explicitly incorporate quality scores into the parameter estimation framework, either by correcting the biased estimator as demonstrated above or by taking a likelihood-based approach as we did in an earlier study working with metagenomics data (Johnson and Slatkin 2006). An alternative method of reducing bias involves raising the threshold quality score and discarding even more data. However, not only does this latter strategy retain some bias (albeit a smaller amount) but it also leads to an increase in variance through the reduction in data. Thus, we urge empiricists to use threshold-based estimators with caution and theoreticians to develop estimators that avoid the problem altogether by accounting for data quality.

## Acknowledgments

Thanks to Weiwei Zhai for helpful discussions and to Michael Jordan for pointing out the broader connection to missing data problems. This research was supported by National Institutes of Health grant R01-GM40282 to M.S.

## Literature Cited

- Allen EE, Tyson GW, Whitaker RJ, Detter JC, Richardson PM, Banfield JF. 2007. Genome dynamics in a natural archaeal population. *Proc Natl Acad Sci USA*. 104:1883–1888.
- Bouck J, Miller W, Gorrell JH, Muzny D, Gibbs RA. 1998. Analysis of the quality and utility of random shotgun sequencing at low redundancies. *Genome Res*. 8:1074–1084.
- Brandstatter A, Sanger T, Lutz-Bonengel S, Parson W, Beraud-Colomb E, Wen B, Kong QP, Bravi CM, Bandelt HJ. 2005. Phantom mutation hotspots in human mitochondrial DNA. *Electrophoresis*. 26:3414–3429.
- Briggs AW, Stenzel U, Johnson PLF, et al. (11 co-authors). 2007. Patterns of damage in genomic DNA sequences from a Neandertal. *Proc Natl Acad Sci USA*. 104:14616–14621.
- Brown G, Gill G, Kuntz R, Langley C, Neale D. 2004. Nucleotide diversity and linkage disequilibrium in loblolly pine. *Proc Natl Acad Sci USA*. 101:15255–15260.
- Clark AG, Whittam TS. 1992. Sequencing errors and molecular evolutionary analysis. *Mol Biol Evol*. 9:744–752.
- Cockerham CC, Weir BS. 1987. Correlations, descent measures: drift with migration and mutation. *Proc Natl Acad Sci USA*. 84:8512–8514.
- Crawford DC, Akey DT, Nickerson DA. 2005. The patterns of natural variation in human genes. *Annu Rev Genomics Hum Genet*. 6:287–312.
- Ewing B, Green P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res*. 8:186–194.
- Frazer KA, Wade CM, Hinds DA, Patil N, Cox DR, Daly MJ. 2004. Segmental phylogenetic relationships of inbred mouse strains revealed by fine-scale analysis of sequence variation across 4.6 mb of mouse genome. *Genome Res*. 14:1493–1500.

- Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. *Genetics*. 133:693–709.
- Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*. 18: 337–338.
- Johnson PLF, Slatkin M. 2006. Inference of population genetic parameters in metagenomics: a clean look at messy data. *Genome Res*. 16:1320–1327.
- Lincoln SE, Lander ES. 1992. Systematic detection of errors in genetic linkage data. *Genomics*. 14:604–610.
- Little RJ, Rubin DB. 2002. *Statistical analysis with missing data*. 2nd ed. Hoboken (NJ): John Wiley & Sons.
- Margulies M, Egholm M, Altman WE, et al. (56 co-authors). 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*. 437:376–380.
- Olson M, Green P. 1998. A “quality-first” credo for the Human Genome Project. *Genome Res*. 8:414–415.
- Pääbo S. 1989. Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. *Proc Natl Acad Sci USA*. 86:1939–1943.
- Poinar HN, Schwarz C, Qi J, et al. (13 co-authors). 2006. Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *Science*. 311:392–394.
- Rosenbaum PA, Robertson JM, Zamudio KR. 2007. Unexpectedly low genetic divergences among populations of the threatened bog turtle (*Glyptemys muhlenbergii*). *Conserv Genet*. 8:331–342.
- Ross SM. 1996. *Stochastic Processes*. 2nd ed. New York: John Wiley & Sons, Inc.
- Ryynänen HJ, Primmer CR. 2004. Distribution of genetic variation in the growth hormone 1 gene in Atlantic salmon (*Salmo salar*) populations from Europe and North America. *Mol Ecol*. 13:3857–3869.
- States DJ. 1992. Molecular sequence accuracy: analysing imperfect data. *Trends Genet*. 8:52–55.
- Strous M, Pelletier E, Mangenot S, et al. (37 co-authors). 2006. Deciphering the evolution and metabolism of an anammox bacterium from a community genome. *Nature*. 440:790–794.
- Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics*. 105:437–460.
- Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*. 123:585–595.
- Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol*. 7:256–276.
- Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. *Evolution*. 38:1358–1370.
- Wright S. 1951. The genetical structure of populations. *Ann Eugen*. 15:323–354.

Yoko Satta, Associate Editor

Accepted October 27, 2007