Sequencing Errors and Molecular Evolutionary Analysis[1]

Andrew G. Clark and Thomas S. Whittam

Institute of Molecular Evolutionary Genetics and Department of Biology, Pennsylvania State University

Heuristic approaches were used to quantify the influence that sequencing errors have on estimates of nucleotide diversity, substitution rate, and the construction of genealogies. Error rates of <1 nucleotide/kb probably have little affect on conclusions about the evolutionary history of highly polymorphic organisms such as *Drosophila* and *Escherichia coli,* but organisms with very low nucleotide diversity, such as humans, require greater sequencing accuracy. A scan of GenBank for corrections of previous errors reveals that sequencing errors are highly nonrandom.

Introduction

Although many of the statistical properties of population samples of molecular sequences are well understood, remarkably little is known about the influence that errors in sequence determination have on the conclusions from comparative evolutionary analysis. It is clear that sequence data bases contain errors, because, nearly every time a listed gene is sequenced a second time, errors are reported. The incidence of corrections added to GenBank and other sequence data banks demonstrates that errors occur with regularity. Analyses of sequence variation in species with very low sequence diversity are particularly sensitive to sequencing errors, simply because the signal-to-noise ratio is lower than that for species with high levels of sequence diversity. Li and Sadler (1991) were acutely aware of this problem when they compiled data on human DNA sequence polymorphism, and they took extra care to select only the sequences that were most thoroughly verified. Their estimate of nucleotide diversity in humans—i.e., 0.0004—is probably close to the rate of sequencing error of the most careful laboratories. When the frequency of sequencing errors approaches that of polymorphic sites, there is clearly a serious problem.

Here we examine the effect that sequencing errors have on several aspects of molecular evolutionary analysis, including estimation of nucleotide diversity, synonymous and nonsynonymous rates of substitution, and construction of gene genealogies. By "sequencing error" we refer to the total number of erroneous nucleotides between an actual gene and the sequence as it appears in a data bank. The total error rate represents the accumulation of mistakes compounded through an experiment, including the misincorporation of bases in polymerase-chain-reaction amplification of templates, compressions in sequencing ladders, misreading of autoradiographs, mistyping of results, and miscommunication of the sequences to the data base. Although estimates of the error rate at several of these steps are available and are generally very small (e.g., the percentage of bases misincorporated by *Taq* polymerase during polymerase chain reaction is 10^{-3}; Kwiatowski et al. 1991), the order of magnitude

1. Key words: DNA sequencing error, gene genealogy.

Address for correspondence and reprints: Andrew G. Clark, Institute of Molecular Evolutionary Genetics and Department of Biology, 208 Mueller Laboratory, Pennsylvania State University, University Park, Pennsylvania 16802.

Mol. Biol. Evol. 9(4):744–752. 1992.

0737-4038/92/0904-0014$02.00

of the total error rate is unknown. Our purpose is to identify analyses that are particularly sensitive to sequence errors and to ascertain whether the current levels of sequencing precision are really adequate for the intended analyses. A simple but important observation is that effects of a given level of sequencing error depend on the true extant level of DNA polymorphism—less polymorphic data are more sensitive to errors.

Theory

First consider the effect that sequencing errors have on estimates of sequence divergence for a pair of sequences. Errors made on very similar sequences will almost always result in greater apparent divergence, whereas errors on more distantly related sequences will sometimes occur at sites that are already different. In the latter case, the errors may either make the sequences appear more similar or may result in no change in the estimate of sequence divergence. The situation is analogous to the divergence of nucleotide sequences by mutation (Jukes and Cantor 1969). Let p be the observed proportion of nucleotide sites that differ between a pair of sequences, and let d be the number of nucleotide substitutions per site that have occurred between the two sequences. For small degrees of divergence, $p \approx d$; but, as d increases, $p < d$, because of multiple hits and back mutations. In the absence of sequencing errors, the Jukes-Cantor formulation gives

$$p = (3/4)[1 - e^{-(4/3)d}] . \quad (1)$$

Suppose that sequencing errors occur uniformly along a sequence at a rate ε. When an error occurs, the base that is scored at a site is equally likely to be one of the three incorrect bases. If p_{err} is the observed proportion of nucleotide sites that differ as a result of sequencing errors, we get

$$p_{err} = (3/4)[1 - e^{-(4/3)(d+2\varepsilon)}] . \quad (2)$$

The increase of d by 2ε in the exponent results from errors in the two sequences, both contributing to the total number of differences between the sequences. The difference between the estimate of the apparent proportion of sites that differ (including errors) and the true proportion is

$$p_{err} - p = (3/4)[e^{-(4/3)d} - e^{-(4/3)(d+2\varepsilon)}] . \quad (3)$$

Expanding the Taylor series for equation (3) and ignoring terms in ε of second and higher order, we get

$$p_{err} - p \approx \varepsilon[2 - (8/3)d] . \quad (4)$$

This approximation is reasonably good for $d < 0.5$. When the true sequences are identical ($d = 0$), the deviation in the estimate of nucleotide diversity is twice the sequencing error rate. These equations also verify that more diverse sequences will exhibit a lower apparent error rate. For the range of nucleotide diversities and error rates that are reasonable, the error in estimates of diversity are very nearly linear functions of the error rate.

Errors Reported in GenBank Are Not Uniform

The uniformity of errors was tested by collecting a large set of sequence corrections in GenBank version 67. GenBank generally lists changes reported by the original author under the keyword "REVISION" and records changes found by other investigators as "CONFLICTS." Several of the comments indicate that previously reported conflicts were found to be caused by sequencing different strains, so there is an awareness of polymorphism. We identified 273 REVISIONS and 91 CONFLICTS, and a χ^2 test showed them to be homogeneous with respect to the base changes [$\chi^2 = 13.94$; 10 degrees of freedom (df); not significant]. After the REVISIONS and CONFLICTS are pooled, there were 364 corrections, including 201 single-base substitutions (e.g., G substituted by C) and 163 indels of one or more bases (e.g., GT changed to GCT). These corrections were tallied in the form of a transition matrix (table 1) indicating all possible single-base changes. The frequency of correction of bases departs significantly from uniformity ($\chi^2 = 109.5$; 12 df; $P < 0.001$), because the changes A↔C and G↔T are rarer than expected. This bias is not the same as the evolutionary transition bias, because the other two transversions (A↔T and G↔C) each occur as frequently as the transitions (A↔G and C↔T). The nonuniformity of errors can be incorporated into the model given in equations (1)–(4) in the same manner as models of evolutionary change in nucleotide sequences (e.g., see Kimura 1980; Gojobori et al. 1982; Tajima and Nei 1984), but even a strong transition bias in these models leads to little departure from the Jukes-Cantor formulation, unless $d > 0.5$. Because error rates will always (we hope) be <1%, there is little need to refine equation (3) to account for the error bias.

Erroneous assignment of a base may, of course, affect the inferred amino acid sequence of the gene product. When all codons are equally frequent and all single-base changes occur with equal frequency, then 76.0% of the nucleotide changes result in amino acid replacements. When we weighted the frequency of all possible single-base changes by the frequency of their occurrence (table 1), 74.8% of the changes were nonsynonymous. Although these percentages would change somewhat depending on the actual codon usage, nonuniformity of errors does not generally affect the change in the overall proportion of synonymous and nonsynonymous polymorphisms.

What Fraction of Nucleotides in the Data Bases Are Erroneous?

The above analysis does not allow an estimation of the total error rate, and no systematic study has been reported that allows the error rate to be estimated. A number

Table 1
Corrected Errors Reported in the GenBank Data Base Version 67

	Corrected Nucleotide				
Original Nucleotide	A	T	C	G	X
A		12 (0.26)	11 (0.24)	23 (0.50)	11
T	11 (0.24)		25 (0.56)	9 (0.20)	16
C	8 (0.14)	22 (0.38)		27 (0.47)	13
G	26 (0.49)	10 (0.19)	17 (0.32)		10
X	27	32	37	17	

Note.—Data are number of base changes. Numbers in parentheses are the fraction of single-base replacements (ignoring indels) that are of each type. The entire data base was scanned for the keywords "REVISION" and "CONFLICT," and reports of base changes were tallied. X = absence of a base at a site.

of genes have been sequenced more than once, and the concordance between the repeated sequencing efforts gives an indication of the error rate. For example, Laurie et al. (1991) report sequences of alleles KA12 and RI32 of *Adh* from *Drosophila melanogaster,* and they found two differences in KA12 and one in RI32, compared with the sequences originally reported by Kreitman (1983). This represents three differences in a total of 3,840 nucleotides. This is fairly close to the error rate, for manual sequencing, claimed in the advertisements for the Applied Biosystems automated sequencer (1 error/1,000 nucleotides).

Illustrations with *gnd, gapA,* and *Adh*

The effect that sequencing errors have on different types of molecular evolutionary analyses was explored by simulating errors in the eight *Escherichia coli gnd* sequences studied by Sawyer et al. (1987), the 14 *Salmonella gapA* sequences studied by Nelson et al. (1991), and the 11 *Drosophila Adh* sequences studied by Kreitman (1983). At a given error rate, single-base changes were placed with equal probability on the aligned sequences, and the pairwise distance matrix of nucleotide divergence (d_{ij}) and average divergence (π) were calculated from the simulated data. One hundred replicates were generated for each sequencing error rate.

The results exhibit a nearly linear relationship between the sequencing error rate and all distance measures (fig. 1). As predicted, the slope of the relationship is lower for the more diverse sequences. For highly diverse sequences, such as *gapA,* the 95% confidence intervals of these statistics do not overlap even with a very high error rate of 4%. Similarly, for long sequences such as *Adh,* the confidence intervals are sufficiently small that they do not overlap, even with high error rates. In contrast, the *gnd* data exhibit overlapping confidence intervals among distance measures, even with an error rate <1%. One might expect that the topology of the genealogy of *gnd* alleles would be sensitive to infrequent sequencing errors, and this was tested below.

For most coding regions there is an excess of synonymous polymorphism, because of the past action of natural selection constraining amino acid replacements. Inferences about the nature of selective constraints are frequently made by comparing the rate of synonymous substitution per site (d_s) and the rate of nonsynonymous substitution per site (d_n). These statistics were estimated by the method of Nei and Gojobori (1986) and in figure 1 are plotted for several error rates. The *gnd* and *gapA* data show a dramatic excess of synonymous substitution compared with the nonsynonymous substitution rate, and this pattern holds even for high error rates. Because most of the nucleotide sites are nonsynonymous, sequencing errors have a larger effect on d_n than on d_s. High rates of sequencing error lead to an overlap in the confidence intervals of d_s and d_n in the *Adh* data. Sequencing errors make the data appear closer to the expectation of the neutral theory for unconstrained coding regions, by increasing the ratio of d_n/d_s (fig. 2).

Sensitivity of Gene Genealogy Reconstruction to Sequencing Errors

Because the effect of sequencing errors is greater on pairs of sequences that are more similar, it is to be expected that the pairwise distances among a set of alleles will be distorted nonuniformly. Such distortion will have an effect on the inferred topology of a tree. To explore just how sensitive the tree construction is to sequencing errors, we introduced errors into published data sets, and for each error rate we calculated a distance matrix (with Jukes-Cantor correction) and generated a neighbor-joining tree (Saitou and Nei 1987). We then scored whether each node of the original tree was present in each of the error-laden trees. Figure 3 shows the original trees and plots,

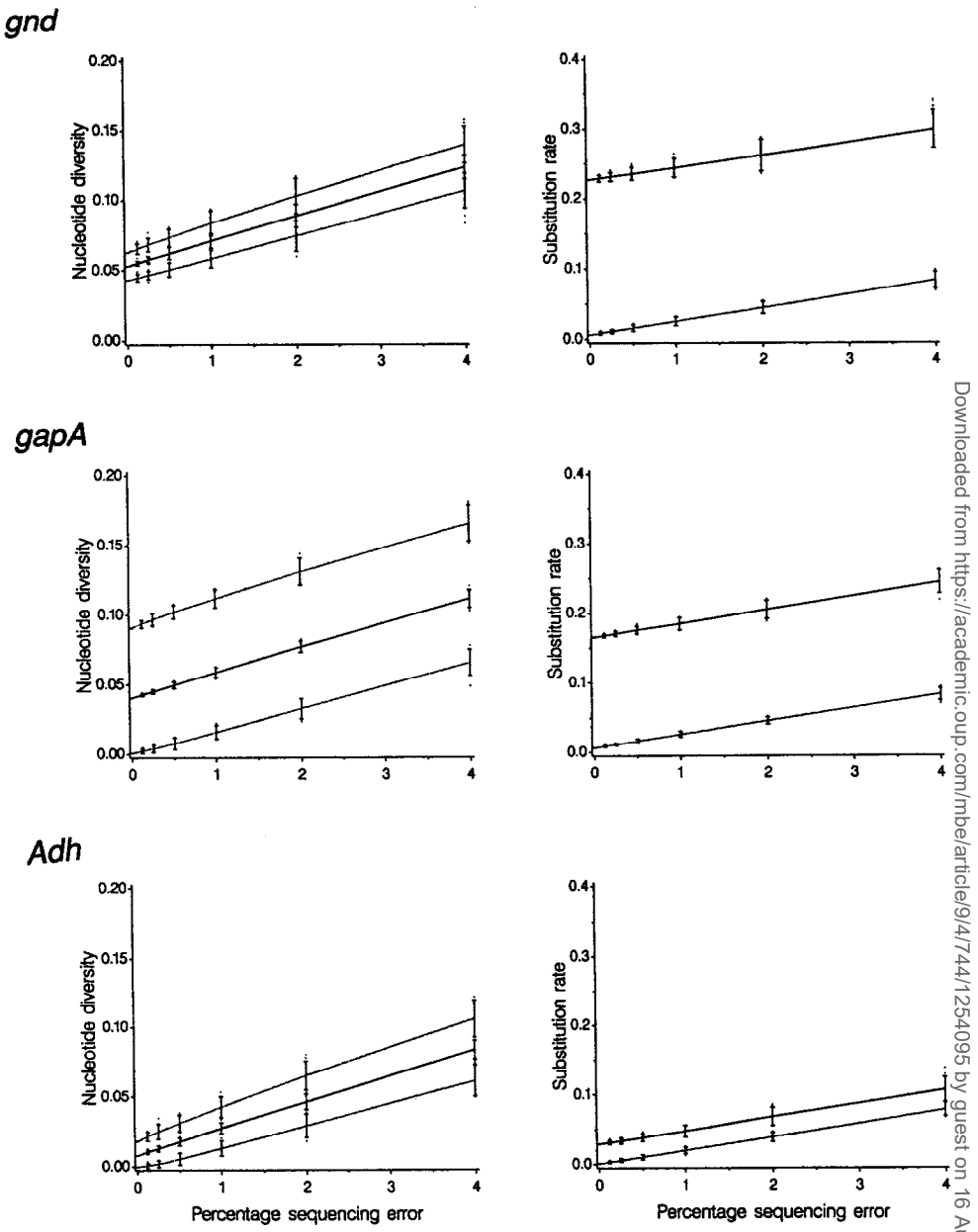

FIG. 1.—Effect that sequencing errors have on π (middle line in graphs in left col.), minimum divergence (lower line in graphs in left col.), and maximum divergence (upper line in graphs in left col.) and on d_s (upper lines in graphs in right col.) and d_n (lower lines in graphs in right col.). Errors were placed on the sequences with equal likelihood of all base changes. Plotted are the 95% confidence intervals for each statistic. Dots represent observations falling outside the confidence interval. When the errors were introduced according to the frequencies of table 1, the results were virtually indistinguishable (data not shown).

for each error rate, the percentage of trees that possess each node. In the case of *gnd,* three of the six nodes were quite sensitive to sequencing errors, and an error rate of just 0.1% resulted in a different tree more than half the time. The *gapA* sequences yielded the most stable tree, in part because the sequences were very divergent. In the

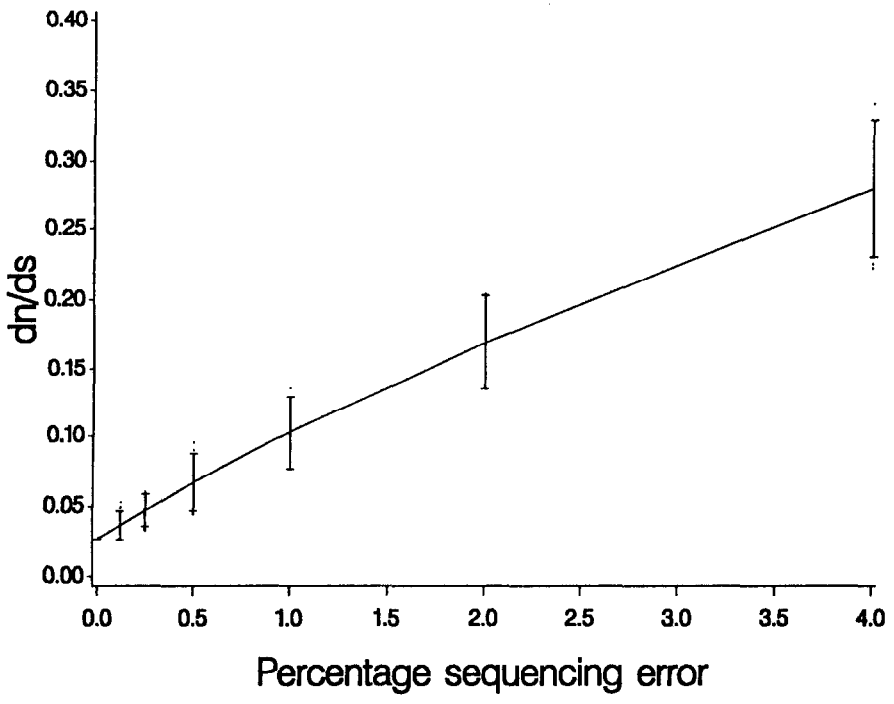

FIG. 2.—d_n/d_s for *gnd* data, over a range of sequencing error rates. Uniform errors were placed on the data of Sawyer et al. (1987), with 100 replicates for each error rate, and, for each altered data set, the method of Nei and Gojobori (1986) was used to estimate d_s and d_n. Because there are initially very few nonsynonymous differences, errors cause a predominance of nonsynonymous changes, making d_n/d_s increase rapidly with increasing error rate.

case of *gapA,* one node was considerably more sensitive to sequencing errors than were the other nodes, and 6 of the 12 nodes were correctly recovered with an error rate as high as 4%. The genealogy of *Adh* alleles of *Drosophila* was intermediate in sensitivity to sequencing errors.

Discussion and Conclusions

DNA sequence data bases clearly contain errors, and although some analyses may be robust to the probable error level (States and Botstein 1991), evolutionary analyses that focus on the rare variant sites may be greatly perturbed by even the lowest error rates. We do not know what the error rate in molecular sequence data bases is. The fact that repeated sequencing of genes, typically 1–2 kb in length, usually uncovered some errors suggests that the error rate is probably in the range of one error every 500–1,000 bp. This is a crude estimate at best and may be too high by an order of magnitude. For evolutionary analysis, investigators sequence the same gene multiple times, and this probably results in improved accuracy, because polymorphic sites are repeatedly checked. Also, repeated sequencing makes the investigator familiar with troublesome regions which require extra attention. The question that is addressed in this paper is whether this error rate has a serious effect on the inferences we wish to draw from the data. A general answer is that, for organisms with a high nucleotide diversity, such as bacteria and flies, it is probably good enough for all the analyses that have been done. But for humans and other mammals, this error rate is close to the level of polymorphism, and investigators need to be extra cautious to verify putative

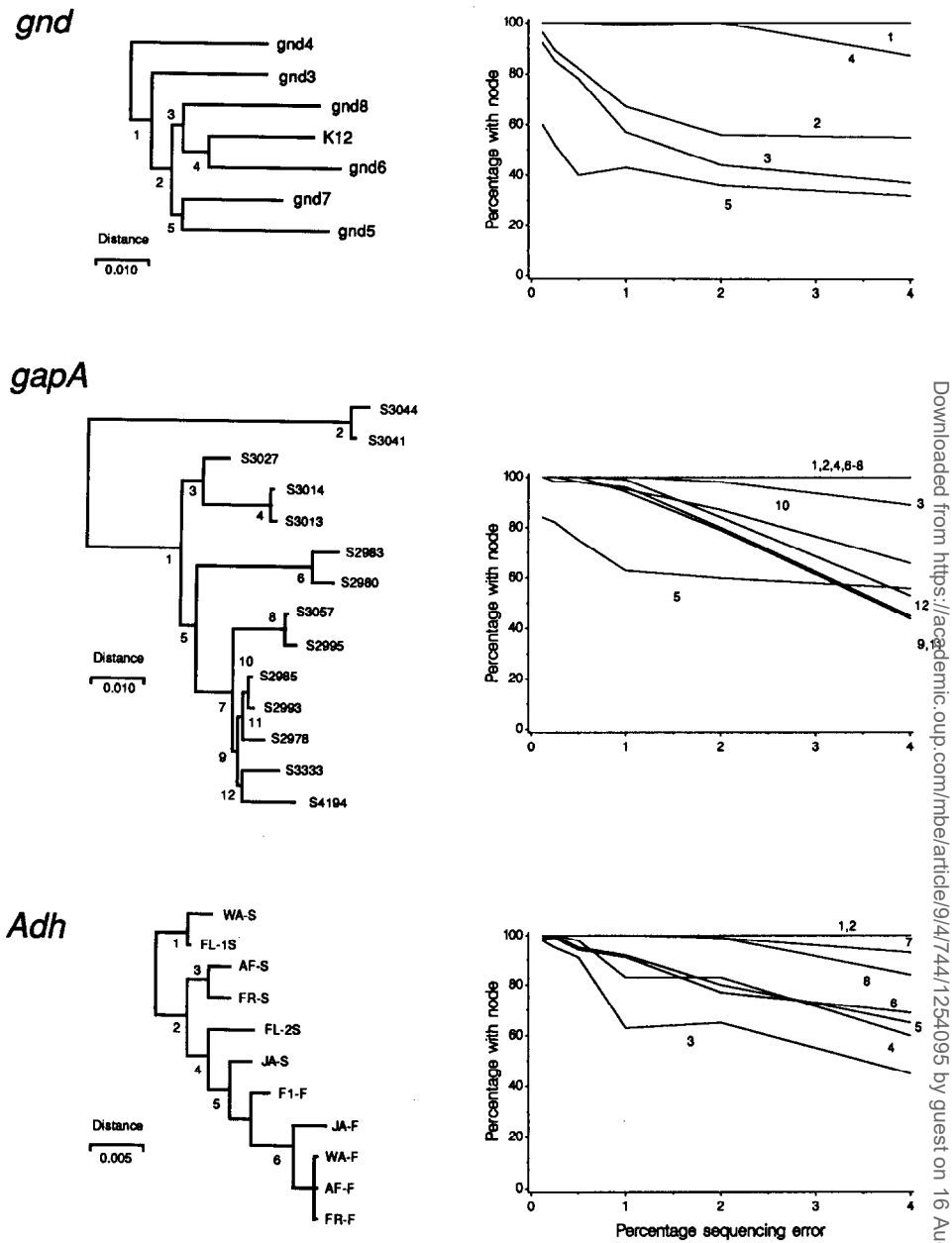

FIG. 3.—Effect that sequencing errors have on tree topology. The left column gives the neighbor-joining tree for the *gnd, gapA,* and *Adh* data. Graphs to the right of each tree indicate the stability to errors that is shown by each labeled node in the respective tree. One hundred trees were constructed for each error rate, with each tree representing a different set of introduced errors. As the sequencing error rate increases, more nodes differ from those in the original tree, and a smaller fraction of error-laden trees support each node. When errors were introduced with the same frequencies as those of the observed revisions to GenBank (table 1), results were virtually the same.

polymorphic sites. Fortunately there is already motivation for high levels of accuracy in human gene sequencing, because of the medical consequences of many of the polymorphisms. We have assumed throughout that errors are made uniformly

throughout the data set—and, if this is not the case, so that, say, one sequence has most of the errors, the effect on the tree topology could be much worse than our analysis indicates. The reason is that a uniform distribution of errors will tend to increase all pairwise differences fairly evenly, whereas, if a single allele has many errors, its position in a genealogy could be moved from within a tight cluster of alleles to a distant outlier. This is an important issue if one is using data generated from two or more different laboratories, a practice that is likely to occur with increasing frequency.

The four major conclusions of our analyses are as follows:

1. Apparent sequence divergence increases linearly with sequencing error rate (for error rates <1%), and the rate of increase in divergence depends on the nucleotide diversity of the sample. More-diverse sequences show a slower rate of apparent increase in diversity with increasing error rate.

2. Absolute estimates of parameters such as nucleotide diversity (π) are very sensitive to sequencing errors, especially for genes with low nucleotide diversity. In the case of *Adh* in *Drosophila,* the observed sequencing error rate of 3/3,840 nucleotides will not result in a π estimate that lies outside the 95% confidence interval based on the current data. More generally, for the magnitude of variation seen among *Escherichia coli* and *Drosophila,* the sequencing error rate is probably sufficiently low that spurious conclusions caused by errors are unlikely. However, nobody includes sequencing error in statements of confidence intervals on population genetic parameters—and, if they did, the confidence intervals could be considerably larger.

3. If most nucleotide polymorphisms are silent, sequencing errors will occur disproportionately at nonsynonymous sites and thus, by increasing the ratio of d_n/d_s, will weaken the evidence for purifying selection. If the data are being collected for studies of polymorphism, investigators probably pay more attention to nonsynonymous differences and carefully verify them.

4. Gene genealogies are particularly sensitive to sequencing errors, but the nodes that are least stable to error are generally either toward the tips of branches (e.g., node 3 of the *Adh* tree) or near very short branches (e.g., node 5 of the *gapA* tree). Minor changes in tree topology can be expected even with error rates <1/1,000. Investigators are beginning to perform bootstrap techniques to make statements about the confidence in trees, and, while nodes that are stable to resampling are generally the same nodes that are most stable to sequencing errors, this is not always the case. Exceptions can be seen whenever the lines on figure 3 cross, which indicate reversals in the ordering of the stability of nodes with changing error rates. Again, molecular evolutionists need to be aware of the fact that confidence in a tree depends on the sequencing error rate.

We have only considered single-base changes in our analyses, because indels (which account for about half of the revisions posted in GenBank) can result in frameshifts in coding sequences and are thus more easily detected in multiple allelic sequencing. True indel variation does occur with fairly high frequency in some species (e.g., *Drosophila*) and requires extra care in data collection.

Acknowledgments

We thank Masatoshi Nei for comments on an earlier draft of this paper. This work was supported by NIH grants AI00964 and HD00743.

LITERATURE CITED

GOJOBORI, T., K. ISHII, and M. NEI. 1982. Estimation of average number of nucleotide differences when the rate of substitution varies with nucleotide. J. Mol. Evol. **18**:414–423.

JUKES, T. H., and C. R. CANTOR. 1969. Evolution of protein molecules. Pp. 21–132 *in* H. N. MUNRO, ed. Mammalian protein metabolism III. Academic Press, New York.

KIMURA, M. 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

KREITMAN, M. 1983. Nucleotide polymorphism at the *alcohol dehydrogenase* locus of *Drosophila melanogaster*. Nature **304**:412–417.

KWIATOWSKI, J., D. SKARECKY, S. HERNANDEZ, D. PHAM, F. QUIJAS, and F. J. AYALA. 1991. High fidelity of the polymerase chain reaction. Mol. Biol. Evol. **8**:884–887.

LAURIE, C. C., J. T. BRIDGHAM, and M. CHOUDHARY. 1991. Association between DNA sequence variation and variation in expression of the *Adh* gene in natural populations of *Drosophila melanogaster*. Genetics **129**:489–499.

LI, W.-H., and L. A. SADLER. 1991. Low nucleotide diversity in man. Genetics **129**:513–523.

NEI, M., and T. GOJOBORI. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3**:418–426.

NELSON, K., T. S. WHITTAM, and R. K. SELANDER. 1991. Nucleotide polymorphism and evolution in the glyceraldehyde-3-phosphate dehydrogenase gene (*gapA*) in natural populations of *Salmonella* and *Escherichia coli*. Proc. Natl. Acad. Sci. USA **88**:6667–6671.

SAITOU, N., and M. NEI. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

SAWYER, S. A., D. E. DYKHUIZEN, and D. L. HARTL. 1987. Confidence interval for the number of selectively neutral amino acid polymorphisms. Proc. Natl. Acad. Sci. USA **84**:6225–6228.

STATES, D. J., and D. BOTSTEIN. 1991. Molecular sequence accuracy and the analysis of protein coding regions. Proc. Natl. Acad. Sci. USA **88**:5518–5522.

TAJIMA, F., and M. NEI. 1984. Estimation of evolutionary distance between nucleotide sequences. Mol. Biol. Evol. **1**:269–285.

WALTER M. FITCH, reviewing editor

Received August 2, 1991; revision received December 18, 1991

Accepted December 18, 1991