The Analysis of Population Survey Data on DNA Sequence Variation[1]

Michael Lynch and Teresa J. Crease†*

*Department of Biology, University of Oregon; and †Department of Biology, University of Windsor

A technique is presented for the partitioning of nucleotide diversity into within- and between-population components for the case in which multiple populations have been surveyed for restriction-site variation. This allows the estimation of an analogue of F_{ST} at the DNA level. Approximate expressions are given for the variance of these estimates resulting from nucleotide, individual, and population sampling. Application of the technique to existing studies on mitochondrial DNA in several animal species and on several nuclear genes in *Drosophila* indicates that the standard errors of genetic diversity estimates are usually quite large. Thus, comparative studies of nucleotide diversity need to be substantially larger than the current standards. Normally, only a very small fraction of the sampling variance is caused by sampling of individuals. Even when 20 or so restriction enzymes are employed, nucleotide sampling is a major source of error, and population sampling is often quite important. Generally, the degree of population subdivision at the nucleotide level is comparable with that at the haplotype level, but significant differences do arise as a result of inequalities in the genetic distances between haplotypes.

Introduction

For the past 2 decades enzyme electrophoresis has been a paradigm for the assessment of population genetic variability (Nei 1975). However, because of the redundancy of the genetic code and the inability of electrophoresis to detect all amino acid replacements, measures of variation derived from protein surveys are somewhat lacking in quantitative reliability. As the newly developed DNA technologies become more economical and more accessible to population biologists, restriction-site surveys are becoming popular, and with the advent of the polymerase chain reaction we can expect population surveys of sequence variation to become common in the near future. With these types of data, it is possible to define measures of genetic variation and their sampling variances without ambiguity (Weir 1983; Nei 1987).

Surveys of restriction-site variation are now being used to study a number of population genetic problems including the influence of selective constraint on locus-specific diversity; the relative rates of evolutionary divergence for loci associated with cytoplasmic organelles, autosomes, and sex chromosomes; and the influence of intragenic recombination on heterozygosity and linkage disequilibrium. The study of these and other issues requires that an estimate of genetic variation have a reasonably small standard error (SE).

[1] Key words: genetic variation, mitochondrial DNA, nucleotide sequence divergence, population subdivision, restriction analysis.

Address for correspondence and reprints: Michael Lynch, Department of Biology, University of Oregon, Eugene, Oregon 97403.

Mol. Biol. Evol. 7(4):377–394. 1990.

0737-4030/90/0704-0008$02.00

Aspects of the variance of nucleotide diversity measures associated with haplotype sampling have been studied by Nei and Tajima (1981, 1983), and the problem of nucleotide sampling has been investigated by Nei and Jin (1989). However, their work was concerned primarily with pairs of populations, whereas population studies usually involve multiple samples, often at several hierarchical levels. They did not consider the issue of population sampling, nor did they evaluate the relative magnitude of the error resulting from the three levels of sampling. Information on these matters should be of use in the design of future population surveys.

Our purpose is to add to the earlier work of Nei, Tajima, and Jin in developing a general method for the description of within- and between-population variation at the nucleotide level. In addition, we introduce a measure of population subdivision at the nucleotide level, N_{ST}. Large-sample variance expressions are presented for these estimates, taking into account the sampling of nucleotides, individuals, and populations. Our primary focus is on variation in restriction sites, as this is still the most common method of DNA analysis at the population level, but the derived methods also apply to sequence data.

Most empirical studies have relied entirely on the equations of Nei and Tajima (1981) for the estimation of SEs for genetic diversity measures. We show below that these SEs are biased substantially downward because of their failure to account for nucleotide and population sampling. This is of concern, since the SEs of genetic diversity estimates in the literature are already quite large.

The Evolutionary Distance between Haplotypes

We will refer to a particular variant for the stretch of DNA being probed as a *haplotype*. π_{xy} is the fraction of nucleotide sites that differ between two haplotypes (x and y). This quantity can be obtained without error from DNA sequence data, and there are several indirect methods for estimating it by restriction-site or fragment analysis (Nei and Li 1979; Engels 1981; Ewens et al. 1981; Kaplan 1983; Nei and Tajima 1983). These methods assume a random distribution of nucleotides in the DNA sequence, as well as an equal mutation rate at all sites, conditions which are rarely, if ever, met. However, for the small evolutionary distances that are usually the rule in population surveys, the results are quite robust to violations in assumptions (Tajima and Nei 1982; Golding 1983; Kaplan 1983). We recommend the maximum likelihood method of Nei and Tajima (1983) for obtaining the estimate $\hat{\pi}_{xy}$.

As a measure of variation we adopt the number of substitutions per nucleotide site. The observed differences can be converted to the average number of substitutions per nucleotide site ($\hat{\delta}_{xy}$) by the method of Jukes and Cantor (1969),

$$\hat{\delta}_{xy} = -\frac{3}{4}\left(1 - \frac{4\hat{\pi}_{xy}}{3}\right). \qquad (1)$$

For $\hat{\pi}_{xy} \leq 0.05$, which is usually the case for intraspecific variation, $\hat{\delta}_{xy} \simeq \hat{\pi}_{xy}$. The sampling variance of $\hat{\delta}_{xy}$, $\mathrm{Var}(\hat{\delta}_{xy})$, due to the analysis of a finite number of restriction sites can be estimated by use of equations (19) and (23) of Nei and Tajima (1983).

The method of Nei and Tajima (1983) yields an SE of zero for two sequences that have identical sets of observed restriction sites. This may seem undesirable, since the application of additional restriction enzymes often reveals differences where none appeared previously. One might prefer to be on the conservative side by assigning an approximate upper (rather than lower) limit to the sampling variance of the distance between apparently identical haplotypes. However, when this is done, the influence on the SEs of population parameter estimates is negligible (authors' unpublished data), so we will not pursue the matter any further.

Within-Population Variation

Nei and Tajima (1981) developed a simple method for estimating the average number of substitutions per nucleotide site for random pairs of sequences sampled from the same population,

$$\hat{v}_i = \frac{2}{n_i(n_i - 1)} \sum_{x<y} n_{ix} n_{iy} \hat{\delta}_{xy} , \tag{2}$$

where n_i is the total number of sequences assayed in population i and where n_{ix} is the observed number of haplotype x in the same population. Since it is common for surveys to involve several populations, it is useful to have a pooled estimate of the within-population differentiation. We define this to be simply

$$\hat{v}_w = \frac{\sum_i \hat{v}_i}{n_p} , \tag{3}$$

where n_p is the number of populations sampled.

Equation (3) yields an unbiased estimate of the within-population variation provided that the sample (and population) sizes of the different populations are uncorrelated with their genetic diversities. Sometimes a special effort is made to secure more samples in more diverse populations (or vice versa), in which case an alternative approach is necessary. The simplest solution would be to randomly reduce the sample sizes of all populations down to the same level.

When populations are stable in size for very long periods of time, a positive correlation is expected to develop between population size and v_i (Crow and Kimura 1970). However, natural populations are rarely stable for more than a few generations and frequently exhibit dramatic fluctuations in size, so this correlation is expected to be weak. In the absence of conflicting evidence, the use of equal weighting for different populations seems to be justified.

There are three sources of error in the estimation of the within-population variation: the sampling of haplotype frequencies (h), the sampling of nucleotide sites (n), and, in the case of the pooled estimate, the sampling of populations (p). For the ith population, the total sampling variance is obtained from the formula for the variance of a product under the assumption that haplotype and nucleotide sampling errors are independent,

$$\begin{aligned}\mathrm{Var}(\hat{v}_i) = \left(\frac{2n_i}{n_i - 1}\right)^2 \sum_{\substack{x<y \\ z<w}} & \hat{\delta}_{xy}\hat{\delta}_{zw}\mathrm{Cov}(p_{ix}p_{iy},\, p_{iz}p_{iw}) + p_{ix}p_{iy}p_{iz}p_{iw}\mathrm{Cov}(\hat{\delta}_{xy},\, \hat{\delta}_{zw}) \\ & + \mathrm{Cov}(p_{ix}p_{iy},\, p_{iz}p_{iw}) \cdot \mathrm{Cov}(\hat{\delta}_{xy},\, \hat{\delta}_{zw}) ,\end{aligned} \tag{4}$$

where $p_{ix} = n_{ix}/n_i$ is the estimated frequency of haplotype x in population i and where

Cov, the sampling covariance, should read "Var" when $x = z$ and $y = w$. For comparison with earlier work, it is useful to abbreviate equation (4) as

$$\mathrm{Var}(\hat{v}_i) = \mathrm{Var}_h(\hat{v}_i) + \mathrm{Var}_n(\hat{v}_i) + \Delta_r(\hat{v}_i)\,, \tag{5}$$

where $\mathrm{Var}_h(\hat{v}_i)$ and $\mathrm{Var}_n(\hat{v}_i)$ are the sampling variances that would be computed if haplotype (h) or nucleotide (n) sampling were solely considered and where $\Delta_r(\hat{v}_i)$ is the deviation of $\mathrm{Var}(\hat{v}_i)$ from the sum of haplotype and nucleotide sampling variances. The residual term is not necessarily positive.

Nei and Tajima [1981, eq. (11)] provide an expression for the first term which can be written as

$$\mathrm{Var}_h(\hat{v}_i) = \frac{4}{n_i(n_i-1)}\left\{\frac{(n_i-1)^2(3-2n_i)}{2n_i^2}\hat{v}_i^2 + 2(n_i-2)\sum_{\substack{x\neq y\\ y<z}} p_{ix}p_{iy}p_{iz}\hat{\delta}_{xy}\hat{\delta}_{xz} + \sum_{x<y}[1+(n_i-2)(p_{ix}+p_{iy})]p_{ix}p_{iy}\hat{\delta}_{xy}^2\right\}. \tag{6}$$

The sampling variance at the nucleotide level is

$$\mathrm{Var}_n(\hat{v}_i) = \left(\frac{2n_i}{n_i-1}\right)^2\left[\sum_{x<y} p_{ix}^2p_{iy}^2\mathrm{Var}(\hat{\delta}_{xy}) + \sum_{\substack{x<y\\ z<w\\ xy\neq zw}} p_{ix}p_{iy}p_{iz}p_{iw}\mathrm{Cov}(\hat{\delta}_{xy}, \hat{\delta}_{zw})\right], \tag{7}$$

which is in the general form of equation (3) of Nei and Jin (1989). The residual term in equation (5) is evaluated by letting

$$\mathrm{Cov}(p_{ix}p_{iy}, p_{iz}p_{iw}) = \frac{(n_i-1)p_{ix}p_{iy}}{n_i^3}[p_{iz}p_{iw}(6-4n_i)+\phi]\,, \tag{8}$$

where $\phi = (n_i - 2)(p_{ix}+p_{iy}) + 1$ when $x = z$ and $y = w$, $\phi = (n_i - 2)p_{iw}$ when $x = z$ and $y \neq w$, and $\phi = 0$ when $x \neq z$ and $y \neq w$.

The covariance terms involving haplotype distances in equations (4) and (7) account for the nonindependence results from phylogenetic relationships. If, for example, a is the common ancestor of haplotypes y and z and if x is a more remote relative, the evolutionary distances δ_{xy} and δ_{xz} share the evolutionary path between x and a. The sampling covariance of these two distances is then equivalent to the nucleotide sampling variance of the distance between x and a. Nei and Jin (1989) present a simple algorithm for estimating the shared evolutionary distance ($\hat{\delta}'$) for $\hat{\delta}_{xy}$ and $\hat{\delta}_{zw}$, where x may equal z. Once $\hat{\delta}'$ has been obtained, $\mathrm{Cov}(\hat{\delta}_{xy}, \hat{\delta}_{zw})$ is estimated as $\mathrm{Var}(\hat{\delta}')$ by use of equations (19) and (23) of Nei and Tajima (1983). As noted by Nei and Jin (1989), the procurement of estimates of δ' requires a phylogenetic tree of haplotypes. There are a number of ways to obtain this from the matrix of observed haplotype distances, and it should be kept in mind that all of these are subject to error (Saitou and Nei 1986). In the applications presented below, we relied on the unweighted-pair-group method (UPGMA) of Sneath and Sokal (1973).

The sampling variance of the pooled within-population variation is obtained by noting that

$$\hat{v}_i = v_i + e_i \,, \tag{9}$$

where e_i is the deviation between the estimate $\hat{v}_i$ and the parameter v_i caused by haplotype and nucleotide sampling. Under the assumption that the population-specific v_i are uncorrelated with their sampling errors,

$$\mathrm{Var}(\hat{v}_w) = \mathrm{Var}_p(\overline{v_i}) + \mathrm{Var}_e(\hat{v}_w) \,, \tag{10}$$

where $\mathrm{Var}_p(\overline{v_i}) = \mathrm{Var}(v_i)/n_p$ is the variance due to population sampling and where

$$\mathrm{Var}_e(\hat{v}_w) = \frac{\sum\limits_i \mathrm{Var}(\hat{v}_i) + 2 \sum\limits_{i<j} \mathrm{Cov}(\hat{v}_i, \hat{v}_j)}{n_p^2} \tag{11}$$

is the variance due to nucleotide and haplotype sampling. $\mathrm{Var}(\hat{v}_i)$ in equation (11) has been defined above, while $\mathrm{Cov}(\hat{v}_i, \hat{v}_j)$ is the sampling covariance between the estimates $\hat{v}_i$ and $\hat{v}_j$. Since haplotypes are sampled independently from different populations, such covariance is caused only by the nucleotide sampling error,

$$\begin{aligned} \mathrm{Cov}_n(\hat{v}_i, \hat{v}_j) = \left[\frac{4n_i n_j}{(n_i - 1)(n_j - 1)}\right] & \sum_{x<y} p_{ix}p_{iy}p_{jx}p_{jy}\mathrm{Var}(\hat{\delta}_{xy}) \\ & + \sum_{\substack{x<y \\ z<w \\ xy \neq zw}} p_{ix}p_{iy}p_{jz}p_{jw}\mathrm{Cov}(\hat{\delta}_{xy}, \hat{\delta}_{zw}) \,. \end{aligned} \tag{12}$$

If the assayed populations are the only ones of interest, then $\mathrm{Var}_p(\overline{v_i}) = 0$. However, the variance resulting from population sampling needs to be included whenever the assayed populations are being treated as a random sample of a larger group of populations. The variance of the parametric v_i can be estimated by computing the variance of the observed $\hat{v}_i$ and subtracting from this the inflation resulting from sampling error. We argue that, since all of the estimates of v_i are based on the same nucleotide sites, the variance among the $\hat{v}_i$ may be influenced by the particular sites sampled, but there is no reason to expect such sampling to cause an upward or downward bias relative to the variance of the v_i. (If a relatively invariant set of nucleotides has been sampled, then all of the $\hat{v}_i$ will be underestimated relative to the v_i.) On the other hand, haplotype sampling will cause the variance of the $\hat{v}_i$ to be overly dispersed with respect to the v_i. Thus, the variance of $\hat{v}_w$ due to population sampling is approximated by

$$\mathrm{Var}_p(\overline{v_i}) = \frac{\sum\limits_i \hat{v}_i^2 - n_p\hat{v}_w^2}{n_p(n_p - 1)} - \frac{\sum\limits_i \mathrm{Var}_h(\hat{v}_i)}{n_p^2} \,. \tag{13}$$

[We acknowledge that our recommendation of not subtracting any nucleotide sampling variance from equation (13) is controversial and that it may cause this expression to yield upwardly biased estimates. However, to the extent that such bias exists, it should be offset by a downward bias described below.]

The above approach to estimating $\mathrm{Var}_p(\overline{v_i})$ assumes implicitly that the assayed populations represent independent sampling units. This assumption is also implicit in the use of equation (3) to estimate v_w and, indeed, in all current methods for estimating average genetic diversity. Populations certainly are not evolutionarily independent, since they share haplotypes with common ancestry (Takahata and Nei 1985) and may be interconnected by migration or exposed to common selection pressures. Thus, the v_i may be historically correlated even if all populations are isolated simultaneously. In studies of natural populations, however, we generally have no information on these historical processes, and the focus of the present paper is on sampling variance rather than on variance caused by the evolutionary process. Although observable, similarity of haplotype distributions is not a reliable measure of population relationship at the level of sampling. Consider, for example, the extreme situation in which all assayed populations were isolated simultaneously by fragmentation of an ancestral base population; by chance, the similarity within some pairs of populations would be greater than that in others, but all populations would be independent. Thus, in the absence of information on the historical events leading to current population structure, the best we can do is to sample in such a way that the independence assumption is likely to be approximated; for example, multiple samples from the same woodlot, pond, or tide pool should be avoided.

Since nonindependence of populations usually will cause a positive covariance between the v_i, equation (13) provides a minimum estimate of the sampling variance of $\hat{v}_w$. A more general expression of $\mathrm{Var}_p(\overline{v_i})$ is $[\mathrm{Var}(v_i) + (n_p - 1)\mathrm{Cov}(v_i, v_j)]/n_p$, where $\mathrm{Cov}(v_i, v_j)$ represents the average covariance between parameters v_i and v_j, so, in principle, equation (13) can be improved if something is known about the latter term.

Between-Population Variation

As inferred from the computations of Nei and Li (1979), the average number of substitutions per nucleotide site between populations i and j is estimated by

$$\hat{v}_{ij} = \hat{v}'_{ij} - \frac{\hat{v}_i + \hat{v}_j}{2}\,, \tag{14}$$

where $\hat{v}'_{ij} = \sum_{x,y} p_{ix}p_{jy}\hat{\delta}_{xy}$. We define the pooled estimate to be

$$\hat{v}_b = \frac{2\sum\limits_{i<j} \hat{v}_{ij}}{n_p(n_p - 1)}\,. \tag{15}$$

The sampling variance of $\hat{v}_{ij}$ can be written as

$$\begin{aligned}\mathrm{Var}(\hat{v}_{ij}) = \mathrm{Var}(\hat{v}'_{ij}) + {}^1\!/\!_4[\mathrm{Var}(\hat{v}_i)+\mathrm{Var}(\hat{v}_j)] - \mathrm{Cov}(\hat{v}'_{ij}, \hat{v}_i)\\ - \mathrm{Cov}(\hat{v}'_{ij}, \hat{v}_j) + {}^1\!/\!_2\,\mathrm{Cov}(\hat{v}_i, \hat{v}_j)\,,\end{aligned} \tag{16}$$

(Nei and Jin 1989), where $\mathrm{Var}(\hat{v}_i)$, $\mathrm{Var}(\hat{v}_j)$, and $\mathrm{Cov}(\hat{v}_i, \hat{v}_j)$ are as they have been defined in the preceding section. The sampling variance of the uncorrected between-population variation, $\hat{v}'_{ij}$, is

$$\mathrm{Var}(\hat{v}'_{ij}) = \sum_{x,y,z,w} \hat{\delta}_{xy}\hat{\delta}_{zw}\mathrm{Cov}(p_{ix}p_{jy},\, p_{iz}p_{jw}) + p_{ix}p_{jy}p_{iz}p_{jw}\mathrm{Cov}(\hat{\delta}_{xy},\, \hat{\delta}_{zw})$$
$$+ \mathrm{Cov}(p_{ix}p_{jy},\, p_{iz}p_{jw})\cdot\mathrm{Cov}(\hat{\delta}_{xy},\, \hat{\delta}_{zw}) = \mathrm{Var}_h(\hat{v}'_{ij}) + \mathrm{Var}_n(\hat{v}'_{ij}) + \Delta_r(\hat{v}'_{ij})\,, \tag{17}$$

where x and y (and z and w) are pairs of haplotypes from different populations and where $\mathrm{Cov}(\hat{\delta}_{xy}, \hat{\delta}_{zw})$ is computed from the phylogenetic relationship of haplotypes across populations. The three terms in equation (17) are analogous to those in equation (5). Nei and Tajima (1981) have shown that

$$\mathrm{Var}_h(\hat{v}'_{ij}) = \frac{1}{n_i n_j}\{(1 - n_i - n_j)(\hat{v}'_{ij})^2 + \sum_{x,y} p_{ix}p_{jy}\hat{\delta}^2_{xy}$$
$$+ \sum_{x,y,z} [(n_i - 1)p_{iy}p_{iz}p_{jx} + (n_j - 1)p_{ix}p_{jy}p_{jz}]\hat{\delta}_{xy}\hat{\delta}_{xz}\}\,. \tag{18}$$

The variance due to nucleotide sampling alone is

$$\mathrm{Var}_n(\hat{v}'_{ij}) = \sum_{x<y} (p_{ix}p_{jy} + p_{iy}p_{jx})^2\mathrm{Var}(\hat{\delta}_{xy})$$
$$+ \sum_{\substack{x<y\\ z<w\\ xy\neq zw}} (p_{ix}p_{jy} + p_{iy}p_{jx})(p_{iz}p_{jw} + p_{iw}p_{jz})\mathrm{Cov}(\hat{\delta}_{xy},\, \hat{\delta}_{zw})\,, \tag{19}$$

and $\Delta_r(\hat{v}'_{ij})$ is evaluated by letting

$$\mathrm{Cov}(p_{ix}p_{jy},\, p_{iz}p_{jw}) = \frac{p_{ix}p_{jy}}{n_i n_j}[(1 - n_i - n_j)p_{ix}p_{jw} + \phi]\,, \tag{20}$$

where $\phi = 0$ when $x \neq z$ and $y \neq w$, $\phi = (n_j - 1)p_{jw}$ when $x = z$ and $y \neq w$, and $\phi = (n_i - 1)p_{ix} + (n_j - 1)p_{jy} + 1$ when $x = z$ and $y = w$.

The terms $\mathrm{Cov}(\hat{v}'_{ij}, \hat{v}_i)$ and $\mathrm{Cov}(\hat{v}'_{ij}, \hat{v}_j)$ also can be written as the sum of three components,

$$\mathrm{Cov}(\hat{v}'_{ij},\, \hat{v}_i) = \frac{2n_i}{n_i - 1}\sum_{\substack{x\neq y\\ z<w}} \hat{\delta}_{xy}\hat{\delta}_{zw}\mathrm{Cov}(p_{ix}p_{jy},\, p_{iz}p_{iw}) + p_{ix}p_{jy}p_{iz}p_{iw}\mathrm{Cov}(\hat{\delta}_{xy},\, \hat{\delta}_{zw})$$
$$+ \mathrm{Cov}(p_{ix}p_{jy},\, p_{iz}p_{iw})\cdot\mathrm{Cov}(\hat{\delta}_{xy},\, \hat{\delta}_{zw})$$
$$= \mathrm{Cov}_h(\hat{v}'_{ij},\, \hat{v}_i) + \mathrm{Cov}_n(\hat{v}'_{ij},\, \hat{v}_i) + \Delta_r(\hat{v}'_{ij},\, \hat{v}_i)\,, \tag{21}$$

with $\mathrm{Cov}(\hat{v}'_{ij}, \hat{v}_j)$ obtained by exchanging i and j. With a slight correction to the equation of Nei and Tajima (1981),

$$\mathrm{Cov}_h(\hat{v}'_{ij}, \hat{v}_i) = \frac{2}{n_i}\left(\frac{1 - n_i}{n_i}\,\hat{v}'_{ij}\hat{v}_i + \sum_{x,y,z} p_{ix}p_{iy}p_{jz}\hat{\delta}_{xy}\hat{\delta}_{xz}\right). \tag{22}$$

The nucleotide sampling component is

$$\begin{aligned}\mathrm{Cov}_n(\hat{v}'_{ij}, \hat{v}_i) = \frac{2n_i}{n_i - 1}\Big[&\sum_{x<y} p_{ix}p_{iy}(p_{ix}p_{jy} + p_{iy}p_{jx})\mathrm{Var}(\hat{\delta}_{xy}) \\ &+ \sum_{\substack{x<y \\ z<w \\ xy\neq zw}} p_{iz}p_{iw}(p_{ix}p_{jy} + p_{jx}p_{iy})\mathrm{Cov}(\hat{\delta}_{xy}, \hat{v}_{zw})\Big]\,,\end{aligned} \tag{23}$$

and $\Delta_r(\hat{v}'_{ij}, \hat{v}_i)$ is obtained by letting

$$\mathrm{Cov}(p_{ix}p_{jy}, p_{iz}p_{iw}) = \frac{1 - n_i}{n_i^2}\,p_{ix}p_{iw}p_{jy}(2p_{iz} + \phi)\,, \tag{24}$$

where $\phi = 0$ when $x \neq z \neq w$ and $\phi = -1$ when $x = z$ or $x = w$.

Finally, we come to the sampling variance of the pooled estimate of the between-population divergence. $\mathrm{Var}(\hat{v}_b)$ can be partitioned into the variance due to the sampling of the pairwise parameters v_{ij} and the variance due to the estimated $\hat{v}_{ij}$ caused by nucleotide and haplotype sampling,

$$\mathrm{Var}(\hat{v}_b) = \mathrm{Var}_p(\overline{v_i}) + \mathrm{Var}_e(\hat{v}_b)\,, \tag{25}$$

where

$$\mathrm{Var}_p(\overline{v_i}) = \frac{\mathrm{Var}(v_{ij}) + 2(n_p - 2)\mathrm{Cov}(v_{ij}, v_{ik})}{n_c} \tag{26}$$

and

$$\mathrm{Var}_e(\hat{v}_b) = \frac{\sum_{i<j}\mathrm{Var}(\hat{v}_{ij}) + 2\sum_{\substack{i<k \\ k<l \\ ij\neq kl}}\mathrm{Cov}(\hat{v}_{ij}, \hat{v}_{kl})}{n_c^2}\,, \tag{27}$$

with $n_c = n_p(n_p - 1)/2$.

Equation (26) assumes that all of the between-population diversities are independent except when they share a population. Violations of this assumption will cause our estimate of $\mathrm{Var}(\hat{v}_b)$ to be downwardly biased. According to the logic developed above for $\mathrm{Var}(v_i)$,

$$\mathrm{Var}(v_{ij}) = \frac{\sum_{i<j}\hat{v}_{ij}^2 - n_c\hat{v}_b^2}{n_c - 1} - \frac{\sum_{i<j}\mathrm{Var}_h(\hat{v}_{ij})}{n_c}\,, \tag{28}$$

and

$$\mathrm{Cov}(v_{ij}, v_{ik}) = \frac{\sum\limits_{i \neq j \neq k} \hat{v}_{ij}\hat{v}_{ik} - n'\hat{v}_b^2}{n' - 1} - \frac{\sum\limits_{i \neq j \neq k} \mathrm{Cov}_h(\hat{v}_{ij}, \hat{v}_{ik})}{n'}, \tag{29}$$

where $n' = 2n_c(n_p - 2)$.

Each of the sampling covariance terms in equation (27) expands to

$$\begin{aligned}\mathrm{Cov}(\hat{v}_{ij}, \hat{v}_{kl}) = {} & \mathrm{Cov}(\hat{v}'_{ij}, \hat{v}'_{kl}) \\ & + \tfrac{1}{4}[\mathrm{Cov}(\hat{v}_i, \hat{v}_k) + \mathrm{Cov}(\hat{v}_i, \hat{v}_l) + \mathrm{Cov}(\hat{v}_j, \hat{v}_k) + \mathrm{Cov}(\hat{v}_j, \hat{v}_l)] \\ & - \tfrac{1}{2}[\mathrm{Cov}(\hat{v}'_{ij}, \hat{v}_k) + \mathrm{Cov}(\hat{v}'_{ij}, \hat{v}_l) + \mathrm{Cov}(\hat{v}'_{kl}, \hat{v}_i) + \mathrm{Cov}(\hat{v}'_{kl}, \hat{v}_j)]\,.\end{aligned} \tag{30}$$

Of the nine components in equation (30), the four covariances between the within-population diversities have already been defined in equation (12) for $i \neq j$ and by the sum of equations (6) and (7) for $i = j$. Where $k = i$ or j, covariances of the form $\mathrm{Cov}(\hat{v}'_{ij}, \hat{v}_k)$ are defined by equation (21). For k unequal to either i or j, the covariance between $\hat{v}'_{ij}$ and $\hat{v}_k$ is caused only by nucleotide sampling,

$$\begin{aligned}\mathrm{Cov}_n(\hat{v}'_{ij}, \hat{v}_k) = \frac{2n_k}{n_k - 1}\Big[& \sum_{x<y} p_{kx}p_{ky}(p_{ix}p_{jy} + p_{iy}p_{jx})\mathrm{Var}(\hat{\delta}_{xy}) \\ & + \sum_{\substack{x<y \\ z<w \\ xy \neq zw}} p_{kx}p_{ky}(p_{iz}p_{jw} + p_{iw}p_{jz})\mathrm{Cov}(\hat{\delta}_{xy}, \hat{\delta}_{zw})\Big]\,.\end{aligned} \tag{31}$$

For the covariance between uncorrected measures of between-population diversity, the contribution from nucleotide sampling is

$$\begin{aligned}\mathrm{Cov}_n(\hat{v}'_{ij}, \hat{v}'_{kl}) = & \sum_{x<y} (p_{ix}p_{jy} + p_{iy}p_{jx})(p_{kx}p_{ly} + p_{ky}p_{lx})\mathrm{Var}(\hat{\delta}_{xy}) \\ & + \sum_{\substack{x<y \\ z<w \\ xy \neq zw}} (p_{ix}p_{jy} + p_{iy}p_{jx})(p_{kz}p_{lw} + p_{kw}p_{lz})\mathrm{Cov}(\hat{\delta}_{xy}, \hat{\delta}_{zw})\,.\end{aligned} \tag{32}$$

To the above quantity we must add contributions due to haplotype sampling when i or j is equal to k or l,

$$\mathrm{Cov}_h(\hat{v}'_{ij}, \hat{v}'_{ik}) = \frac{1}{n_i}\Big(\sum_{x,y,z} p_{ix}p_{jy}p_{kz}\hat{\delta}_{xy}\hat{\delta}_{xz} - \hat{v}'_{ij}\hat{v}'_{ik}\Big), \tag{33}$$

and

$$\Delta_r(\hat{v}'_{ij}, \hat{v}'_{ik}) = \sum_{x,y,z,w} \mathrm{Cov}(p_{ix}p_{jy}, p_{iz}p_{kw}) \cdot \mathrm{Cov}(\hat{\delta}_{xy}, \hat{\delta}_{zw}), \tag{34}$$

where

$$\mathrm{Cov}(p_{ix}p_{jy}, p_{iz}p_{kw}) = \frac{p_{ix}p_{jy}p_{kw}}{n_i}(\phi - p_{iz}), \tag{35}$$

with $\phi = 0$ when $x \neq z$ and $\phi = 1$ when $x = z$.

The Degree of Population Subdivision at the Nucleotide Level

Many attempts have been made to estimate the extent of population differentiation, through surveys of gene (haplotype) frequencies, by use of Weir and Cockerham's (1984) θ or Nei's (1973) G_{ST}, both of which are intended to measure Wright's (1951) F_{ST} (for a recent review, see Chakraborty and Danker-Hopfe, accepted).

As an analogue of these indices at the nucleotide level, we suggest

$$N_{ST} = \frac{\hat{v}_b}{\hat{v}_w + \hat{v}_b}, \tag{36}$$

which gives the ratio of the average genetic distance between genes from different populations relative to that among genes in the population at large. Extreme N_{ST} estimates of 0 and 1 indicate zero and complete population subdivision, respectively.

Nei (1982) introduced an index γ_{ST} which is similar to our N_{ST}. However, while our measures of the within-population diversity are the same, his measure of inter-population diversity includes comparisons of populations with themselves, making it approximately equivalent to $(n_p - 1)\hat{v}_b / n_p$. We suggest N_{ST} should be used whenever one is making inferences about a larger collection of populations from the restricted sample of n_p. γ_{ST} is more appropriate when the sampled populations are the only ones of interest, as in Nei's (1982) analysis of the three major races of man. In any event, γ_{ST} and N_{ST} are convergent as n_p becomes large, and they are not greatly different for $n_p > 5$.

Estimates ofpopulation subdivision that are based on nucleotide divergence need not be the same as those based on haplotype frequencies. The indices θ and G_{ST} treat the evolutionary distances between all pairs of haplotypes as being identical, while N_{ST} explicitly accounts for the variation in genetic identity. Both approaches will give identical results when there are only two alleles (haplotypes) per locus. With more than two alleles per locus, N_{ST} will be greater than or less than F_{ST}, depending on whether pairs of relatively divergent haplotypes tend to be distributed between or within populations.

The approximate sampling variance for N_{ST} is obtained by a first-order Taylor expansion (in the context of isozyme analysis, see Chakraborty 1974),

$$\mathrm{Var}(N_{ST}) = \left(\frac{N_{ST}}{\hat{v}_w + \hat{v}_b}\right)^2 \left[\left(\frac{\hat{v}_w}{\hat{v}_b}\right)^2 \mathrm{Var}(\hat{v}_b) - 2\left(\frac{\hat{v}_w}{\hat{v}_b}\right) \mathrm{Cov}(\hat{v}_w, \hat{v}_b) + \mathrm{Var}(\hat{v}_w)\right], \tag{37}$$

with

$$\mathrm{Cov}(\hat{v}_w, \hat{v}_b) = \frac{\sum_{i,j<k} \mathrm{Cov}(\hat{v}_{jk}, \hat{v}_i)}{n_p n_c},$$

and

$$\mathrm{Cov}(\hat{v}_{jk}, \hat{v}_i) = \mathrm{Cov}(\hat{v}'_{jk}, \hat{v}_i) - \tfrac{1}{2}[\mathrm{Cov}(\hat{v}_j, \hat{v}_i) + \mathrm{Cov}(\hat{v}_k, \hat{v}_i)] .$$

If we assume, as a first approximation, that N_{ST} is normally distributed, then, because $E(N_{ST}) = 0$ under the null hypothesis of no population subdivision, the test statistic

$$D = \frac{N_{ST}^2}{\mathrm{Var}(N_{ST})} \qquad (38)$$

should be χ^2 distributed with 1 degree of freedom under the null hypothesis. $D > 3.84$ and $D > 6.64$ then allow rejection of the null hypothesis at the 5% and 1% confidence levels, respectively. It seems likely that the distribution of N_{ST} will be positively skewed, because N_{ST} is constrained from taking on very negative values, so these are probably conservative criteria.

Application to Existing DNA Surveys

Among population studies at the DNA level, restriction-site surveys of the mitochondrial genome are by far the most common (Avise et al. 1987*a*; Moritz et al. 1987). For several such studies, we have procured the data necessary for the application of the preceding formulas. We now give a brief summary of the results, the primary emphasis being on the sources and magnitude of the sampling variance of the population parameter estimates.

Aside from the extensive human studies by Cann et al. (1984, 1987), our own study on the cyclically parthenogenetic microcrustacean *Daphnia pulex* (Crease et al. 1990) has a larger number of sampled genomes per population (~24) than does any other published study. The number of restriction sites sampled per genome (82) is also substantially greater than in most other studies. Nevertheless, the SEs of the measures of genetic variation within individual populations and between pairs of populations are quite large—in most cases, at least half as large as the parameter estimates (table 1). Nearly all of the sampling variance is attributable to nucleotide sampling. Thus, for this study, the exclusive reliance on Nei and Tajima's (1983) formula would have led to substantial underestimates of the SEs of evolutionary distances.

Table 2 summarizes the pooled estimates of the within- and between-population variation for mitochondrial surveys involving several species. Again, it can be seen that the SEs are usually quite large for both types of estimates and that haplotype sampling makes only a small contribution to the total sampling variance. The residual terms $\Delta_r(\hat{v}_w)$ and $\Delta_r(\hat{v}_b)$ are also usually of negligible importance. Population sampling is often of greater importance, sometimes substantially so (as in the case of our *Daphnia* study, where two populations contained a single haplotype but where the other two

Table 1
Summary Statistics for a Mitochondrial DNA Restriction-Site Survey of Four *Daphnia pulex* Populations

			Sampling Variance[b]			
Site[a]	$\hat{v}$	SE($\hat{v}$)	hap	nuc	res	n
Within populations						
BU	0.0000	0.0000	. . .	. . .	. . .	16
KA	0.0000	0.0000	. . .	. . .	. . .	21
PA	0.0017	0.0009	0.25	0.70	0.05	31
SA.	0.0052	0.0014	0.03	0.95	0.02	26
Between populations:						
BU-KA	0.0075	0.0029	0.00	1.00	0.00	
BU-PA	0.0047	0.0036	0.00	1.02	−0.02	
BU-SA	0.0025	0.0018	0.10	0.84	0.06	
KA-PA	0.0058	0.0035	0.00	1.02	−0.02	
KA-SA	0.0052	0.0020	0.01	0.90	0.08	
PA-SA	0.0013	0.0012	0.11	0.74	0.15	

[a] BU, KA, PA, and SA refer to different study populations (Crease et al. 1990).

[b] Columns "hap" and "nuc" give the fractions of the sampling variance that would be recognized, respectively, if only the sampling of indiviuals or if only the sampling of nucleotides were accounted for, whereas column "res" gives the fraction of the total sampling variance attributable to joint sampling of haplotypes and nucleotides.

contained two and six). The majority of the sampling variance is often a consequence of the limited number of restriction enzymes employed.

Table 3 summarizes data from a number of global surveys of restriction-site variation in nuclear genes of *Drosophila.* In all but one case, the SEs of $\hat{v}_b$ are substantially larger than the estimate. The SEs of $\hat{v}_w$ are less than the estimate $\hat{v}_w$ in all cases, but they are large enough that not much can be said about differences between loci or species.

Only seven of the 16 surveys in tables 2 and 3 reveal significant population subdivision at the nucleotide level, as is shown in table 4. Also included in table 4 are the estimates of population subdivision that are based on haplotype frequencies. In this case, we use the likelihood-ratio statistic G (Sokal and Rohlf 1981) as a criterion for population subdivision and find four significant cases. Note that there is significant subdivision at the level of haplotype frequencies but not at the level of nucleotide divergence for the *vermilion* locus in *Drosophila ananassae.* On the other hand, the fish *Opsanus beta* and *Stizostedion vitreum* exhibit significant population divergence at the nucleotide level but not at the level of haplotype frequencies. These types of discrepancies arise when the evolutionary distances between random pairs of *different* haplotypes are unequal at the within- and between-population levels (lower at the between-population level in the first case and lower at the within-population level in the latter). Such inequality may provide useful insight into the mechanisms responsible for the geographic structure of genetic differentiation.

Despite these interesting exceptions, the estimates of N_{ST} and G'_{ST} are highly correlated ($r = 0.92$; $P < 0.01$). The slope of the regression is not significantly different from one, and the intercept is not significantly different from zero. This is a useful result, since it supports the idea that, *on average,* N_{ST} estimates the same kind of population subdivision as do such haplotype analyses as the numerous isozyme surveys that have been performed in the past. There is still some need for caution, however,

Table 2
Summary of Data on Mitochondrial DNA Variation for Restriction-Site Surveys

Species and Component of Variation	$\hat{v}$	SE	Sampling Variance[a]				n_p[b]	$\bar{n}$[c]
			hap	nuc	pop	res		
Odocoileus virginianus (white-tailed deer):[d]								
Within population	0.0008	0.0007	0.04	0.26	0.67	0.03	4	11
Between population	0.0023	0.0028	0.01	1.05	0.02	−0.08		
Agelaius phoeniceus (red-winged blackbird):[e]								
Within population	0.0019	0.0012	0.02	0.96	0.00	0.02	14	7
Between population	0.0005	0.0013	0.02	0.94	0.00	0.04		
Arius felis (hardhead catfish):[c]								
Within population	0.0020	0.0011	0.10	0.80	0.06	0.04	9	6
Between population	−0.0002	0.0010	0.09	0.62	0.00	0.29		
Opsanus beta (Gulf toadfish):[f]								
Within population	0.0014	0.0006	0.34	0.19	0.24	0.23	3	6
Between population	0.0023	0.0019	0.03	0.94	0.00	0.03		
Stizostedion vitreum (walleye):[g]								
Within population	0.0013	0.0009	0.02	0.91	0.06	0.01	10	14
Between population	0.0005	0.0006	0.05	0.75	0.04	0.16		
Limulus polyphemus (horseshoe crab):[h]								
Within population	0.0005	0.0004	0.12	0.68	0.07	0.13	6	6
	0.0019	0.0011	0.09	0.30	0.57	0.04	6	6
Between population	−0.0000	0.0002	0.67	0.64	0.00	−0.31		
	0.0032	0.0026	0.03	1.01	0.00	−0.04		
Daphnia pulex:[i]								
Within population	0.0017	0.0013	0.01	0.12	0.87	0.00	4	24
Between population	0.0045	0.0025	0.00	0.87	0.12	0.01		
Drosophila melanogaster:[j]								
Within population	0.0051	0.0030	0.22	0.32	0.42	0.04	5	6
	0.0047	0.0026	0.04	0.80	0.14	0.02	9	6
Between population	0.0054	0.0050	0.14	0.90	0.00	−0.04		
	0.0033	0.0046	0.02	0.92	0.01	0.05		

[a] Column "pop" gives the population sampling variance for the population; other columns are as defined in table 1.
[b] Number of populations.
[c] Mean number of individuals sampled per population.
[d] Source: Carr et al. (1986).
[e] Source: Ball et al. (1988).
[f] Source: Avise et al. (1987*b*).
[g] Source: Billington and Hebert (1988).
[h] Source: Saunders et al. (1986) (separate analyses were performed for populations north of Florida and for Florida populations.
[i] Source: Crease et al. (1990).
[j] Source: Hale and Singh (1987) (separate analyses were performed for New World and Old World populations).

since it is well known that alleles revealed at the protein level can harbor many variants at the nucleotide level, and there is limited information on how such cryptic variants are distributed within and between populations (Aquadro et al. 1986). In principle, extensive population subdivision can exist at the nucleotide level even when none is revealed with isozymes.

Table 3
Summary of Data from Several Restriction-Site Surveys for Nuclear Gene Loci in *Drosophila*

			Sampling Variance[a]					
Species, Locus, and Component of Variation	$\hat{v}$	SE	hap	nuc	pop	res	n_p	$\bar{n}$
Drosophila ananassae:								
Vermilion:[b]								
Within population	0.0026	0.0019	0.04	0.37	0.54	0.04	3	20
Between population . . .	0.0021	0.0039	0.01	0.98	0.01	0.01		
Forked:[b]								
Within population	0.0057	0.0020	0.06	0.33	0.62	−0.01	3	20
Between population . . .	0.0026	0.0022	0.07	0.95	0.00	−0.02		
D. melanogaster:								
ADH:[c]								
Within population	0.0056	0.0028	0.03	0.89	0.06	0.02	4	12
Between population . . .	0.0013	0.0030	0.04	0.82	0.00	0.14		
Notch:[d]								
Within population	0.0045	0.0018	0.15	0.50	0.30	0.05	6	5
Between population . . .	0.0012	0.0024	0.11	0.81	0.00	0.08		
Zeste-tko:[e]								
Within population	0.0041	0.0018	0.03	0.88	0.09	0.00	3	21
Between population . . .	0.0004	0.0013	0.04	0.89	0.00	0.07		
Yellow-achaete-scute:[f]								
Within population	0.0003	0.0002	0.07	0.62	0.30	0.01	3	21
Between population . . .	0.0000	0.0002	0.01	0.79	0.00	0.20		

[a] Columns are as defined in table 2.
[b] Source: Stephan and Langley (1989).
[c] Source: Aquadro et al. (1986).
[d] Source: Schaeffer et al. (1988).
[e] Source: Aguadé et al. (1989b).
[f] Source: Aguadé et al. (1989*a*).

Discussion

Our main purpose has been to develop a general method for estimating the diversity at the DNA level within and between populations without making any assumptions about the evolutionary mechanisms that led to such variation. We did assume a uniform distribution of nucleotides and of their mutation rates. However, the results in tables 2 and 3 indicate that, for populations of the same species, the evolutionary distances between haplotypes are well below the level at which violations of these assumptions have quantitative significance (Tajima and Nei 1982; Golding 1983; Kaplan 1983). Our estimates for the variances of $\hat{v}_w$ and $\hat{v}_b$ due to population sampling are approximations, and at this point we do not have a good understanding of their degree of accuracy. This is a difficult area that merits further investigation.

With that caveat in mind, the vast majority (in most cases, >90%) of the sampling variance for estimates of nucleotide diversity appears to be attributable to nucleotide and population sampling, at least for surveys employing 10–20 restriction enzymes (i.e., almost all existing studies). These sources of variation have not been accounted for in previous population surveys. The fact that the variance due to haplotype sampling and to the residual terms defined above is relatively small is useful. If haplotype sampling is ignored completely (and hence all of the above terms involving Var_h, Cov_h, and Δ_r are dropped from consideration), a great deal of computational simplicity is

Table 4
Nucleotide (N_{ST}) and Haplotype (G'_{ST} and θ) Population Subdivision

	Nucleotides		Haplotypes	
	N_{ST}	SE[a]	G'_{ST}	θ[c]
Mitochondrial DNA:				
Odocoileus virginianus	0.75**	0.23	0.62**	0.69**
Agelaius phoeniceus	0.19	0.46	0.08	0.11
Arius felis	−0.08	0.59	0.00	0.08
Opsanus beta	0.62**	0.15	0.28	0.33
Stizostedion vitreum	0.26**	0.09	0.25	0.28
Limulus polyphemus	−0.06	0.52	−0.10	0.05
	0.64**	0.21	0.61*	0.60*
Daphnia pulex	0.72**	0.18	0.64**	0.67**
Drosophila melanogaster	0.52	0.25	0.38	0.41
	0.41	0.35	0.27	0.31
Nuclear genes:				
Drosophila ananassae:				
Vermilion	0.44	0.45	0.50**	0.51**
Forked	0.31**	0.08	0.13	0.14
D. melanogaster				
ADH	0.19	0.33	0.06	0.08
notch	0.21	0.34	0.11	0.15
zeste-tko	0.09	0.24	0.07	0.07
yellow-achaete-scute	0.03	0.45	0.10	0.11

[a] Square root of eq. (37).

[b] Computed by use of eqq. (3), (15), and (36) by setting all $\delta_{xy} = 1$; this is a slight deviation from Nei (1986) in the way populations are weighted but is otherwise identical in structure to N_{ST}.

[c] Calculated by the method of Weir and Cockerham (1984).

* $P < .05$.

** $P < .01$.

gained at the expense of only a slight (generally, <5%) downward bias to the SE. This is not of great concern, since the SEs are themselves estimates.

In the future, when population surveys involve direct sequencing rather than restriction-site analysis, the problem of nucleotide sampling will also be partially eliminated. However, unlike mitochondrial DNA restriction-site surveys, sequence analysis does not sample randomly over a whole genome, so the problem of nucleotide sampling is still of some concern if one wishes to make inferences about the entire genome.

The empirical information that we have provided on the relative magnitude of the sources of sampling variance should be of use in the future design of restriction-site surveys. If the populations are readily accessible, it will generally cost about the same to double the number of individuals sampled as to double the number of restriction enzymes. Thus, since most of the sampling variance in existing studies is caused by nucleotide sampling, future studies should concentrate more on the addition of restriction enzymes (or on the use of 4- as opposed to 6-bp cutters) or populations rather than on the enhancement of sample sizes within populations.

The procedures outlined above can be extended to a hierarchical analysis of population structure. For example, for surveys over broad geographic regions, it might be of interest to consider the variation within demes, between demes within sites, and between sites. The between-site component of variation is obtainable by pooling all of the data within sites, treating the latter as populations, and computing $\hat{v}_w$ and $\hat{v}_b$. In this case, v_w would contain both the within-deme and between-deme components of variation, leaving v_b as an estimate of the between-site variation.

As in all cases in which the form of the sampling distribution of parameter estimates is unknown, the SEs generated by our equations provide a rough guide as to the accuracy of the estimates. Regardless of the form of the distribution, by Chebyshev's inequality, the probability that the absolute difference between observed and actual distances exceeds x SEs is $<x^{-2}$.

Resampling procedures may be useful in the development of more explicit statistical tests. For example, one could test for significant evolutionary divergence between two populations by pooling the samples from both populations. An empirical distribution of the observed distance under the hypothesis of no evolutionary divergence could then be obtained by repeatedly and randomly drawing two sets of individuals and computing the evolutionary distance between them. One could then assess the probability of sampling a distance as great as (or as small as) that observed.

In closing, we emphasize that the sampling variance estimators that we have derived are purely a function of the limitations on the investigator. They do not include the variance among hypothetical (unobserved) replicate populations that is caused by the stochastic nature of the evolutionary process. In testing specific evolutionary hypotheses, this additional source of variation needs to be taken into account. Tajima (1983) and Takahata and Nei (1985) have analyzed the simplest situation in which mutation and random genetic drift are the only evolutionary processes. Their results indicate that the stochastic variance can often exceed the sampling variance, so this is not a trivial matter.

Application of the procedures discussed in the present paper requires a great deal of computation, even with quite simple data sets. We have therefore produced a computer program that we are willing to share with anyone who provides us with an IBM-compatible floppy disk.

Acknowledgments

This work has been supported by NSF grant BSR 86-00487 and by PHS grant R01 GM36827-01 to M.L. and by a Sloan Foundation Fellowship to T.J.C. We are very grateful to John Avise for providing unpublished data that allowed us to compute the red-winged blackbird statistics and to Neil Billington for access to the walleye data. R. Chakraborty, M. Nei, and K. Spitze provided very useful comments.

LITERATURE CITED

AGUADÉ, M., N. MIYASHITA, and C. H. LANGLEY. 1989*a*. Reduced variation in the yellow-achaete-scute region in natural populations of *Drosophila melanogaster*. Genetics **122**:607–615.

———. 1989*b*. Restriction-map variation at the *Zeste-tko* region in natural populations of *Drosophila melanogaster*. Mol. Biol. Evol. **6**:123–130.

AQUADRO, C. F., S. F. DESSE, M. M. BLAND, C. H. LANGLEY, and C. C. LAURIE-AHLBERG. 1986. Molecular population genetics of the alcohol dehydrogenase gene region of *Drosophila melanogaster*. Genetics **114**:1165–1190.

AVISE, J. C., J. ARNOLD, R. M. BALL, E. BERMINGHAM, T. LAMB, J. E. NEIGEL, C. A. REEB, and N. C. SAUNDERS. 1987*a*. Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. Annu. Rev. Ecol. Syst. **18**:489–522.

AVISE, J. C., C. A. REEB, and N. C. SAUNDERS. 1987*b*. Geographic population structure and species differences in mitochondrial DNA of mouthbrooding marine catfishes (Ariidae) and demersal spawning toadfishes (Batrachoididae). Evolution **41**:991–1002.

Ball, R. M. Jr., S. Freeman, F. C. James, E. Bermingham, and J. C. Avise. 1988. Phylogeographic population structure of red-winged blackbirds assessed by mitochondrial DNA. Proc. Natl. Acad. Sci. USA **85**:1558–1562.

Billington, N., and P. D. N. Hebert. 1988. Mitochondrial DNA variation in Great Lakes walleye (*Stizostedion vitreum*) populations. Can. J. Fisheries Aquatic Sci. **45**:643–654.

Cann, R. L., W. M. Brown, and A. C. Wilson. 1984. Polymorphic sites and the mechanism of evolution in human mitochondrial DNA. Genetics **106**:479–499.

Cann, R. L., M. Stoneking, and A. C. Wilson. 1987. Mitochondrial DNA and human evolution. Nature **325**:31–36.

Carr, S. M., S. W. Ballinger, J. N. Derr, L. H. Blankenship, and J. W. Bickham. 1986. Mitochondrial DNA analysis of hybridization between sympatric white-tailed deer and mule deer in west Texas. Proc. Natl. Acad. Sci. USA **83**:9576–9580.

Chakraborty, R. 1974. A note on Nei's measure of gene diversity in a substructured population. Humangenetik **21**:85–88.

Chakraborty, R., and H. Danker-Hopfe. A comparative study of different estimators of Wright's fixation indices for population structure analysis. Evol. Biol. (accepted).

Crease, T. J., M Lynch, and K. Spitze. 1990. Hierarchical analysis of population genetic variation in mitochondrial and nuclear genes of *Daphnia pulex*. Mol. Biol. Evol. **7** (accepted).

Crow, J. F., and M. Kimura. 1970. An introduction to population genetics theory. Harper & Row, New York.

Engels, W. R. 1981. Estimating genetic divergence and genetic variability with restriction endonucleases. Proc. Natl. Acad. Sci. USA **78**:6329–6333.

Ewens, W. J., R. S. Speilman, and H. Harris. 1981. Estimation of genetic variation at the DNA level from restriction endonuclease data. Proc. Natl. Acad. Sci. USA **78**:3748–3750.

Golding, G. B. 1983. Estimates of DNA and protein sequence divergence: an examination of some assumptions. Mol. Biol. Evol. **1**:125–142.

Hale, L. R., and R. S. Singh. 1987. Mitochondrial DNA variation and genetic structure in populations of *Drosophila melanogaster*. Mol. Biol. Evol. **4**:622–637.

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pp. 21–123 *in* H. N. Munro, ed. Mammalian protein metabolism. Academic Press, New York.

Kaplan, N. 1983. Statistical analysis of restriction enzyme map data and nucleotide sequence data. Pp. 75–106 *in* B. S. Weir, ed. Statistical analysis of DNA sequence data. Marcel Dekker, New York.

Moritz, C., T. E. Dowling, and W. M. Brown. 1987. Evolution of animal mitochondrial DNA: relevance for population biology and systematics. Annu. Rev. Ecol. Syst. **18**:269–292.

Nei, M. 1973. Analysis of gene diversity in subdivided populations. Proc. Natl. Acad. Sci. USA **70**:3321–3323.

———. 1975. Molecular population genetics and evolution. North-Holland, Amsterdam.

———. 1982. Evolution of human races at the gene level. Pp. 167–181 *in* B. Bonne-Tamir, P. Cohen, and R. N. Goodman, eds. Human genetics, part A: the unfolding genome. Alan R. Liss, New York.

———. 1986. Definition and estimation of fixation indices. Evolution **40**:643–645.

———. 1987. Molecular evolutionary genetics. Columbia University Press, New York.

Nei, M., and L. Jin. 1989. Variances of the average numbers of nucleotide substitutions within and between populations. Mol. Biol. Evol. **6**:290–300.

Nei, M., and W.-H. Li. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA **76**:5269–5273.

Nei, M., and F. Tajima. 1981. DNA polymorphism detectable by restriction endonucleases. Genetics **97**:145–163.

———. 1983. Maximum likelihood estimation of the number of nucleotide substitutions from restriction sites data. Genetics **105**:207–217.

Saitou, N., and M. Nei. 1986. The number of nucleotides required to determine the branching order of three species, with special reference to the human-chimpanzee-gorilla divergence. J. Mol. Evol. **24**:189–204.

SAUNDERS, N. C., L. G. KESSLER, and J. C. AVISE. 1986. Genetic variation and geographic differentiation in mitochondrial DNA of the horseshoe crab, *Limulus polyphemus*. Genetics **112**:613–627.

SCHAEFFER, S. W., C. F. AQUADRO, and C. H. LANGLEY. 1988. Restriction-map variation in the *Notch* region of *Drosophila melanogaster*. Mol. Biol. Evol. **5**:30–40.

SNEATH, P. H. A., and R. R. SOKAL. 1973. Numerical taxonomy. W. H. Freeman, San Francisco.

SOKAL, R. R., and F. J. ROHLF. 1981. Biometry, 2d ed. W. H. Freeman, San Francisco.

STEPHAN, W., and C. H. LANGLEY. 1989. Molecular genetic variation in the centromeric region of the X chromosome in three *Drosophila ananassae* populations. I. Contrasts between the vermilion and forked loci. Genetics **121**:89–99.

TAJIMA, F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics **105**:437–460.

TAJIMA, F., and M. NEI. 1982. Biases of the estimates of DNA divergence obtained by the restriction enzyme technique. J. Mol. Evol. **18**:115–120.

TAKAHATA, N., and M. NEI. 1985. Gene genealogy and variance of interpopulational nucleotide differences. Genetics **110**:325–344.

WEIR, B. S., ed. 1983. Statistical analysis of DNA sequence data. Marcel Dekker, New York.

WEIR, B. S., and C. C. COCKERHAM. 1984. Estimating F-statistics for the analysis of population structure. Evolution **38**:1358–1370.

WRIGHT, S. 1951. The genetical structure of populations. Ann. Eugenics **15**:323–354.

MASATOSHI NEI, reviewing editor

Received April 7, 1989; revision received March 21, 1990

Accepted March 27, 1990