The Analysis of Population Survey Data on DNA Sequence Variation

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.
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. \( \pi_{xy} \) 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 (\( \delta_{xy} \)) by the method of Jukes and Cantor (1969),

\[
\delta_{xy} = -\frac{3}{4} \left( 1 - \frac{4\hat{\pi}_{xy}}{3} \right).
\]

For \( \hat{\pi}_{xy} \leq 0.05 \), which is usually the case for intraspecific variation, \( \delta_{xy} \approx \hat{\pi}_{xy} \). The sampling variance of \( \delta_{xy} \), \( \text{Var}(\delta_{xy}) \), 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,

\[
\hat{v}_i = \frac{2}{n_i(n_i - 1)} \sum_{x < y} n_{ix} n_{iy} \delta_{xy},
\]

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

\[
\hat{v}_w = \frac{1}{n_p} \sum_{i=1}^{n_p} \hat{v}_i,
\]

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 unrelated 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 \(i\)th 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,

\[
\text{Var}(\hat{v}_i) = \left(\frac{2n_i}{n_i - 1}\right)^2 \sum_{x < y} \delta_{xy} \delta_{zw} \text{Cov}(p_{ix}p_{iy}, p_{iz}p_{iw}) + p_{ix}p_{iy}p_{iz}p_{iw} \text{Cov}(\delta_{xy}, \delta_{zw})
\]

\[
+ \text{Cov}(p_{ix}p_{iy}, p_{iz}p_{iw}) \cdot \text{Cov}(\delta_{xy}, \delta_{zw})
\]

where \(p_{ix} = n_{ix}/n_i\) is the estimated frequency of haplotype \(x\) in population \(i\) and where
\[ \text{Cov, the sampling covariance, should read "Var" when } x = z \text{ and } y = w. \] 

For comparison with earlier work, it is useful to abbreviate equation (4) as

\[ \text{Var}(\hat{\theta}_i) = \text{Var}_h(\hat{\theta}_i) + \text{Var}_n(\hat{\theta}_i) + \Delta_i(\hat{\theta}_i), \] (5)

where \( \text{Var}_h(\hat{\theta}_i) \) and \( \text{Var}_n(\hat{\theta}_i) \) are the sampling variances that would be computed if haplotype \( (h) \) or nucleotide \( (n) \) sampling were solely considered and where \( \Delta_i(\hat{\theta}_i) \) is the deviation of \( \text{Var}(\hat{\theta}_i) \) 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

\[ \text{Var}_h(\hat{\theta}_i) = \frac{4}{n_i(n_i - 1)} \left\{ \frac{(n_i - 1)^2(3 - 2n_i)}{2n_i^2} \hat{\theta}_i^2 + 2(n_i - 2) \sum_{x \neq y, y < z} p_{lx}p_{ly}p_{lw} \delta_{xy}\delta_{xz} \right. \]

\[ \left. + \sum_{x < y} [1 + (n_i - 2)(p_{lx} + p_{ly})] p_{lx}p_{ly} \delta_{xy}^2 \right\}. \] (6)

The sampling variance at the nucleotide level is

\[ \text{Var}_n(\hat{\theta}_i) = \frac{2n_i}{n_i - 1} \left\{ \sum_{x < y} p_{lx}^2 p_{ly}^2 \text{Var}(\delta_{xy}) + \sum_{x < y, z < w, x < y} p_{lx}p_{ly}p_{lz}p_{lw} \text{Cov}(\delta_{xy}, \delta_{zw}) \right\}, \] (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

\[ \text{Cov}(p_{lx}p_{ly}, p_{iz}p_{iw}) = \frac{(n_i - 1)p_{lx}p_{ly}}{n_i^2} [p_{iz}p_{iw}(6 - 4n_i) + \phi], \] (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 \( \delta_{xy} \) and \( \delta_{xz} \) 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 \((\delta') \) for \( \delta_{xy} \) and \( \delta_{zw}, \) where \( x \) may equal \( z. \) Once \( \delta' \) has been obtained, \( \text{Cov}(\delta_{xy}, \delta_{zw}) \) is estimated as \( \text{Var}(\delta') \) by use of equations (19) and (23) of Nei and Tajima (1983). As noted by Nei and Jin (1989), the procurement of estimates of \( \delta' \) 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

$$\hat{\theta}_i = \theta_i + e_i, \quad (9)$$

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

$$\text{Var}(\hat{\theta}_w) = \text{Var}_p(\overline{\theta}_i) + \text{Var}_e(\hat{\theta}_w), \quad (10)$$

where $\text{Var}_p(\overline{\theta}_i) = \text{Var}(\theta_i)/n_p$ is the variance due to population sampling and where

$$\text{Var}_e(\hat{\theta}_w) = \frac{\sum_i \text{Var}(\hat{\theta}_i) + 2 \sum_{i<j} \text{Cov}(\hat{\theta}_i, \hat{\theta}_j)}{n_p^2}, \quad (11)$$

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

$$\text{Cov}_n(\hat{\theta}_i, \hat{\theta}_j) = \frac{4n_i n_j}{(n_i - 1)(n_j - 1)} \sum_{x<y} p_{ix}p_{iy}p_{jx}p_{jy} \text{Var}_h(\hat{\delta}_{xy})$$

$$+ \sum_{x<y, z<w} p_{ix}p_{iy}p_{jx}p_{jw} \text{Cov}(\hat{\delta}_{xy}, \hat{\delta}_{zw}). \quad (12)$$

If the assayed populations are the only ones of interest, then $\text{Var}_p(\overline{\theta}_i) = 0$. 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 $\theta_i$ can be estimated by computing the variance of the observed $\hat{\theta}_i$ and subtracting from this the inflation resulting from sampling error. We argue that, since all of the estimates of $\theta_i$ are based on the same nucleotide sites, the variance among the $\hat{\theta}_i$ 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 $\theta_i$. (If a relatively invariant set of nucleotides has been sampled, then all of the $\hat{\theta}_i$ will be underestimated relative to the $\theta_i$.) On the other hand, haplotype sampling will cause the variance of the $\hat{\theta}_i$ to be overly dispersed with respect to the $\theta_i$. Thus, the variance of $\hat{\theta}_w$ due to population sampling is approximated by

$$\text{Var}_p(\overline{\theta}_i) = \frac{\sum_i \hat{\theta}_i^2 - n_p \hat{\theta}_w^2}{n_p(n_p - 1)} - \frac{\sum_i \text{Var}_h(\hat{\theta}_i)}{n_p^2}. \quad (13)$$

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 $\text{Var}_p(\bar{v}_i)$ assumes implicitly that the assayed populations represent independent sampling units. This assumption is also implicit in the use of equation (3) to estimate $v$, 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 $\bar{v}_i$. A more general expression of $\text{Var}_p(\bar{v}_i)$ is $[\text{Var}(v_i) + (n_p - 1)\text{Cov}(v_i, v_j)]/n$, where $\text{Cov}(v_i, v_j)$ 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

$$\hat{\theta}_{ij} = \hat{\theta}_{ij} - \frac{\hat{\theta}_i + \hat{\theta}_j}{2},$$

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

$$\hat{\theta}_p = \frac{2 \sum_{i<j} \hat{\theta}_{ij}}{n_p(n_p - 1)}.$$

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

$$\text{Var}(\hat{\theta}_{ij}) = \text{Var}(\hat{\theta}_{ij}) + \frac{1}{4} [\text{Var}(\hat{\theta}_i) + \text{Var}(\hat{\theta}_j)] - \text{Cov}(\hat{\theta}_{ij}, \hat{\theta}_i)$$

$$- \text{Cov}(\hat{\theta}_{ij}, \hat{\theta}_j) + \frac{1}{2} \text{Cov}(\hat{\theta}_i, \hat{\theta}_j),$$

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

\[
\text{Var}(\hat{\theta}'_{ij}) = \sum_{x,y,z,w} \delta_{xy}\delta_{zw} \text{Cov}(p_{ix}p_{jy}, p_{iz}p_{jw}) + p_{ix}p_{jy}p_{iz}p_{jw} \text{Cov}(\delta_{xy}, \delta_{zw}) + \text{Cov}(p_{ix}p_{jy}, p_{iz}p_{jw}) \cdot \text{Cov}(\delta_{xy}, \delta_{zw}) = \text{Var}_h(\hat{\theta}'_{ij}) + \text{Var}_n(\hat{\theta}'_{ij}) + \Delta_r(\hat{\theta}'_{ij}),
\]

(17)

where \( x \) and \( y \) (and \( z \) and \( w \)) are pairs of haplotypes from different populations and where \( \text{Cov}(\delta_{xy}, \delta_{zw}) \) 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

\[
\text{Var}_h(\hat{\theta}'_{ij}) = \frac{1}{n_i n_j} \{(1 - n_i - n_j)(\hat{\theta}'_{ij})^2 + \sum_{x,y} p_{ix}p_{jy}\delta_{xy}^2
\]

\[
+ \sum_{x,y,z} [(n_i - 1)p_{ix}p_{jy}p_{ix}p_{jy} + (n_j - 1)p_{iz}p_{jw}p_{iz}p_{jw}]\delta_{xy}\delta_{zw}\}.
\]

(18)

The variance due to nucleotide sampling alone is

\[
\text{Var}_n(\hat{\theta}'_{ij}) = \sum_{x<y} (p_{ix}p_{jy} + p_{iy}p_{jx})^2 \text{Var}(\delta_{xy}),
\]

\[
+ \sum_{x<y, z<w} (p_{ix}p_{jy} + p_{iy}p_{jx})(p_{iz}p_{jw} + p_{iw}p_{jz})\text{Cov}(\delta_{xy}, \delta_{zw}),
\]

(19)

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

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

(20)

where \( \phi = 0 \) when \( x \neq z \) and \( y \neq w \), \( \phi = (n_j - 1)p_{iw} \) 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 \( \text{Cov}(\hat{\theta}'_{ij}, \hat{\theta}'_i) \) and \( \text{Cov}(\hat{\theta}'_{ij}, \hat{\theta}'_j) \) also can be written as the sum of three components,

\[
\text{Cov}(\hat{\theta}'_{ij}, \hat{\theta}'_i) = \frac{2n_i}{n_i - 1} \sum_{x \neq y} \delta_{xy}\delta_{zw} \text{Cov}(p_{ix}p_{jy}, p_{iz}p_{jw}) + p_{ix}p_{jy}p_{iz}p_{jw} \text{Cov}(\delta_{xy}, \delta_{zw})
\]

\[
+ \text{Cov}(p_{ix}p_{jy}, p_{iz}p_{jw}) \cdot \text{Cov}(\delta_{xy}, \delta_{zw})
\]

\[
= \text{Cov}_h(\hat{\theta}'_{ij}, \hat{\theta}'_i) + \text{Cov}_n(\hat{\theta}'_{ij}, \hat{\theta}'_i) + \Delta_r(\hat{\theta}'_{ij}, \hat{\theta}'_i),
\]

(21)

with \( \text{Cov}(\hat{\theta}'_{ij}, \hat{\theta}'_i) \) obtained by exchanging \( i \) and \( j \). With a slight correction to the equation of Nei and Tajima (1981),
The nucleotide sampling component is

\[
\text{COV}_n(\hat{\theta}_j, \hat{\theta}_i) = \frac{2n_i}{n_i - 1} \left( \frac{1 - n_i}{n_i} \hat{\theta}_j \hat{\theta}_i + \sum_{x,y,z} p_{ix}p_{iy}p_{jx}p_{jy} \text{COV}(8_{xy}, 8_{zw}) \right),
\]

and \( \Delta_r(\hat{\theta}_j, \hat{\theta}_i) \) is obtained by letting

\[
\text{COV}(P_{ix}P_{iy}, P_{iz}P_{iw}) = - \left( \frac{1}{n_i} \right) \sum_{x \neq y, z \neq w} \text{COV}(\hat{\theta}_j, \hat{\theta}_i).
\]

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. \( \text{Var}(\hat{\theta}_b) \) can be partitioned into the variance due to the sampling of the pairwise parameters \( \nu_{ij} \) and the variance due to the estimated \( \hat{\theta}_j \) caused by nucleotide and haplotype sampling,

\[
\text{Var}(\hat{\theta}_b) = \text{Var}_p(\bar{\nu}_i) + \text{Var}_e(\hat{\nu}_i).
\]

where

\[
\text{Var}_p(\bar{\nu}_i) = \frac{\text{Var}(\nu_{ij}) + 2(n_p - 2)\text{Cov}(\nu_{ij}, \nu_{ik})}{n_c}
\]

and

\[
\text{Var}_e(\hat{\nu}_i) = \frac{\sum_{i<j} \text{Var}(\hat{\nu}_{ij}) + 2 \sum_{i<k} \sum_{k<l} \text{Cov}(\hat{\nu}_{ij}, \hat{\nu}_{kl})}{n_c},
\]

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 \( \text{Var}(\hat{\theta}_b) \) to be downwardly biased. According to the logic developed above for \( \text{Var}(\nu_{ij}) \),

\[
\text{Var}(\nu_{ij}) = \frac{\sum_{i<j} \hat{\nu}_{ij}^2 - n_c \hat{\nu}_b^2}{n_c - 1} - \frac{\sum_{i<j} \text{Var}_h(\hat{\nu}_{ij})}{n_c},
\]
Each of the sampling covariance terms in equation (27) expands to

\[
\text{Cov}(\hat{\theta}_{ij}, \hat{\theta}_{kl}) = \text{Cov}(\hat{\theta}_{ij}, \hat{\theta}_{kl}) + \frac{1}{4} [\text{Cov}(\hat{\theta}_{i}, \hat{\theta}_{k}) + \text{Cov}(\hat{\theta}_{i}, \hat{\theta}_{l}) + \text{Cov}(\hat{\theta}_{j}, \hat{\theta}_{k}) + \text{Cov}(\hat{\theta}_{j}, \hat{\theta}_{l})] \\
- \frac{1}{2} [\text{Cov}(\hat{\theta}_{ij}', \hat{\theta}_{kl}) + \text{Cov}(\hat{\theta}_{ij}', \hat{\theta}_{il}) + \text{Cov}(\hat{\theta}_{kl}', \hat{\theta}_{il}) + \text{Cov}(\hat{\theta}_{kl}', \hat{\theta}_{ij})].
\]

(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 \( \text{Cov}(\hat{\theta}_{ij}', \hat{\theta}_{k}) \) are defined by equation (21). For \( k \) unequal to either \( i \) or \( j \), the covariance between \( \hat{\theta}_{ij}' \) and \( \hat{\theta}_{k} \) is caused only by nucleotide sampling,

\[
\text{Cov}_n(\hat{\theta}_{ij}, \hat{\theta}_{k}) = \frac{2n_k}{n_k - 1} \left[ \sum_{x < y} p_{x} p_{y} (p_{x} p_{y} + p_{y} p_{x}) \text{Var}(\hat{\theta}_{xy}) + \sum_{x < y} p_{x} p_{y} (p_{x} p_{y} + p_{y} p_{x}) \text{Cov}(\hat{\theta}_{xy}, \hat{\theta}_{zw}) \right].
\]

(31)

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

\[
\text{Cov}_n(\hat{\theta}_{ij}', \hat{\theta}_{kl}) = \sum_{x < y} (p_{x} p_{y} + p_{y} p_{x}) (p_{x} p_{y} + p_{y} p_{x}) \text{Var}(\hat{\theta}_{xy}) + \sum_{x < y} p_{x} p_{y} (p_{x} p_{y} + p_{y} p_{x}) (p_{x} p_{y} + p_{y} p_{x}) \text{Cov}(\hat{\theta}_{xy}, \hat{\theta}_{zw}).
\]

(32)

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

\[
\text{Cov}_h(\hat{\theta}_{ij}', \hat{\theta}_{kl}) = \frac{1}{n_i} \left( \sum_{x,y,z} p_{x} p_{y} p_{z} \text{Var}(\hat{\theta}_{xy}) \right).
\]

(33)

\[
\Delta_r(\hat{\theta}_{ij}', \hat{\theta}_{kl}) = \sum_{x,y,z,w} \text{Cov}(p_{x} p_{y} p_{z} p_{w}) \cdot \text{Cov}(\hat{\theta}_{xy}, \hat{\theta}_{zw}).
\]

(34)
where

\[
\text{Cov}(p_{ix}p_{jy}, p_{iz}p_{kw}) = \frac{p_{ix}p_{jy}p_{kw}}{n_i} (\phi - p_{iz}),
\]

(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) \(\Theta\) 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

\[
N_{ST} = \frac{\hat{d}_b}{\hat{d}_w + \hat{d}_b},
\]

(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 \(\gamma_{ST}\) 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 \((n_p - 1)\hat{d}_b/n_p\). We suggest \(N_{ST}\) should be used whenever one is making inferences about a larger collection of populations from the restricted sample of \(n_p\). \(\gamma_{ST}\) 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, \(\gamma_{ST}\) and \(N_{ST}\) are convergent as \(n_p\) becomes large, and they are not greatly different for \(n_p > 5\).

Estimates of population subdivision that are based on nucleotide divergence need not be the same as those based on haplotype frequencies. The indices \(\Theta\) 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),

\[
\text{Var}(N_{ST}) = \left(\frac{N_{ST}}{\hat{d}_w + \hat{d}_b}\right)^2 \left[ \left(\frac{\hat{d}_w}{\hat{d}_b}\right)^2 \text{Var}(\hat{d}_b) - 2 \left(\frac{\hat{d}_w}{\hat{d}_b}\right) \text{Cov}(\hat{d}_w, \hat{d}_b) + \text{Var}(\hat{d}_w) \right],
\]

(37)
with

\[ \text{Cov}(\hat{\theta}_{jk}, \hat{\theta}_i) = \frac{\sum_{i,j,k} \text{Cov}(\hat{\theta}_{jk}, \hat{\theta}_i)}{n_p n_c}, \]

and

\[ \text{Cov}(\hat{\theta}_{jk}, \hat{\theta}_i) = \text{Cov}(\hat{\theta}_{jk}, \hat{\theta}_i) - \frac{1}{2} \left[ \text{Cov}(\hat{\theta}_j, \hat{\theta}_i) + \text{Cov}(\hat{\theta}_k, \hat{\theta}_i) \right]. \]

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

\[ D = \frac{N_{ST}^2}{\text{Var}(N_{ST})} \]

should be \( \chi^2 \) 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. 1987a; 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 \( \Lambda_i(\hat{\theta}_w) \) and \( \Lambda_i(\hat{\theta}_h) \) 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

<table>
<thead>
<tr>
<th>Site</th>
<th>( \hat{\theta} )</th>
<th>SE(( \hat{\theta} ))</th>
<th>hap</th>
<th>nuc</th>
<th>res</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Within populations:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BU</td>
<td>0.0000</td>
<td>0.0000</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>16</td>
</tr>
<tr>
<td>KA</td>
<td>0.0000</td>
<td>0.0000</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>21</td>
</tr>
<tr>
<td>PA</td>
<td>0.0017</td>
<td>0.0009</td>
<td>0.25</td>
<td>0.70</td>
<td>0.05</td>
<td>31</td>
</tr>
<tr>
<td>SA</td>
<td>0.0052</td>
<td>0.0014</td>
<td>0.03</td>
<td>0.95</td>
<td>0.02</td>
<td>26</td>
</tr>
<tr>
<td>Between populations:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BU-KA</td>
<td>0.0075</td>
<td>0.0029</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>16</td>
</tr>
<tr>
<td>BU-PA</td>
<td>0.0047</td>
<td>0.0036</td>
<td>0.00</td>
<td>1.02</td>
<td>-0.02</td>
<td>16</td>
</tr>
<tr>
<td>BU-SA</td>
<td>0.0025</td>
<td>0.0018</td>
<td>0.10</td>
<td>0.84</td>
<td>0.06</td>
<td>16</td>
</tr>
<tr>
<td>KA-PA</td>
<td>0.0058</td>
<td>0.0035</td>
<td>0.00</td>
<td>1.02</td>
<td>-0.02</td>
<td>16</td>
</tr>
<tr>
<td>KA-SA</td>
<td>0.0052</td>
<td>0.0020</td>
<td>0.01</td>
<td>0.90</td>
<td>0.08</td>
<td>16</td>
</tr>
<tr>
<td>PA-SA</td>
<td>0.0013</td>
<td>0.0012</td>
<td>0.11</td>
<td>0.74</td>
<td>0.15</td>
<td>16</td>
</tr>
</tbody>
</table>

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

Columns "hap" and "nuc" give the fractions of the sampling variance that would be recognized, respectively, if only the sampling of individuals 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.

Table 1 contains 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 \( \hat{\theta}_w \) are substantially larger than the estimate. The SEs of \( \hat{\theta}_w \) are less than the estimate \( \hat{\theta}_w \) 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. 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

<table>
<thead>
<tr>
<th>SPECIES AND COMPONENT OF VARIATION</th>
<th>SAMPLING VARIANCE&lt;sup&gt;a&lt;/sup&gt;</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Odocoileus virginianus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(white-tailed deer):&lt;sup&gt;d&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within population</td>
<td>0.0008</td>
<td>0.0007</td>
<td>0.04</td>
<td>0.26</td>
<td>0.67</td>
<td>0.03</td>
</tr>
<tr>
<td>Between population</td>
<td>0.0023</td>
<td>0.0028</td>
<td>0.01</td>
<td>1.05</td>
<td>0.02</td>
<td>-0.08</td>
</tr>
<tr>
<td>Agelaius phoeniceus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(red-winged blackbird):&lt;sup&gt;e&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within population</td>
<td>0.0019</td>
<td>0.0012</td>
<td>0.02</td>
<td>0.96</td>
<td>0.00</td>
<td>0.02</td>
</tr>
<tr>
<td>Between population</td>
<td>0.0005</td>
<td>0.0013</td>
<td>0.02</td>
<td>0.94</td>
<td>0.00</td>
<td>0.04</td>
</tr>
<tr>
<td>Aris felis (hardhead catfish):&lt;sup&gt;e&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within population</td>
<td>0.0020</td>
<td>0.0011</td>
<td>0.10</td>
<td>0.80</td>
<td>0.06</td>
<td>0.04</td>
</tr>
<tr>
<td>Between population</td>
<td>-0.0002</td>
<td>0.0010</td>
<td>0.09</td>
<td>0.62</td>
<td>0.00</td>
<td>0.29</td>
</tr>
<tr>
<td>Opsanus beta (Gulf toadfish):&lt;sup&gt;f&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within population</td>
<td>0.0014</td>
<td>0.0006</td>
<td>0.34</td>
<td>0.19</td>
<td>0.24</td>
<td>0.23</td>
</tr>
<tr>
<td>Between population</td>
<td>0.0023</td>
<td>0.0019</td>
<td>0.03</td>
<td>0.94</td>
<td>0.00</td>
<td>0.03</td>
</tr>
<tr>
<td>Stizostedion vitreum (walleye):&lt;sup&gt;e&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within population</td>
<td>0.0013</td>
<td>0.0009</td>
<td>0.02</td>
<td>0.91</td>
<td>0.06</td>
<td>0.01</td>
</tr>
<tr>
<td>Between population</td>
<td>0.0005</td>
<td>0.0006</td>
<td>0.05</td>
<td>0.75</td>
<td>0.04</td>
<td>0.16</td>
</tr>
<tr>
<td>Limulus polyphemus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(horseshoe crab):&lt;sup&gt;h&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within population</td>
<td>0.0005</td>
<td>0.0004</td>
<td>0.12</td>
<td>0.68</td>
<td>0.07</td>
<td>0.13</td>
</tr>
<tr>
<td>Between population</td>
<td>0.0019</td>
<td>0.0011</td>
<td>0.09</td>
<td>0.30</td>
<td>0.57</td>
<td>0.04</td>
</tr>
<tr>
<td>Daphnia pulex:&lt;sup&gt;i&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within population</td>
<td>0.0017</td>
<td>0.0013</td>
<td>0.01</td>
<td>0.12</td>
<td>0.87</td>
<td>0.00</td>
</tr>
<tr>
<td>Between population</td>
<td>0.0045</td>
<td>0.0025</td>
<td>0.00</td>
<td>0.87</td>
<td>0.12</td>
<td>0.01</td>
</tr>
<tr>
<td>Drosophila melanogaster:&lt;sup&gt;j&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within population</td>
<td>0.0051</td>
<td>0.0030</td>
<td>0.22</td>
<td>0.32</td>
<td>0.42</td>
<td>0.04</td>
</tr>
<tr>
<td>Between population</td>
<td>0.0047</td>
<td>0.0026</td>
<td>0.04</td>
<td>0.80</td>
<td>0.14</td>
<td>0.02</td>
</tr>
<tr>
<td>Between population</td>
<td>0.0054</td>
<td>0.0050</td>
<td>0.14</td>
<td>0.90</td>
<td>0.00</td>
<td>-0.04</td>
</tr>
<tr>
<td></td>
<td>0.0033</td>
<td>0.0046</td>
<td>0.02</td>
<td>0.92</td>
<td>0.01</td>
<td>0.05</td>
</tr>
</tbody>
</table>

<sup>a</sup> Column "pop" gives the population sampling variance for the population; other columns are as defined in table 1.

<sup>b</sup> Number of populations.

<sup>c</sup> Mean number of individuals sampled per population.

<sup>d</sup> Source: Carr et al. (1986).

<sup>e</sup> Source: Ball et al. (1988).

<sup>f</sup> Source: Avise et al. (1987b).

<sup>g</sup> Source: Billington and Hebert (1988).

<sup>h</sup> Source: Saunders et al. (1986) (separate analyses were performed for populations north of Florida and for Florida populations).

<sup>i</sup> Source: Crease et al. (1990).

<sup>j</sup> 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

<table>
<thead>
<tr>
<th>SPECIES, LOCUS, AND COMPONENT OF VARIATION</th>
<th>( \bar{\theta} )</th>
<th>SE</th>
<th>hap</th>
<th>nuc</th>
<th>pop</th>
<th>res</th>
<th>( n_p )</th>
<th>( n_{\bar{\theta}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drosophila ananassae: Vermilion:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within population</td>
<td>0.0026</td>
<td>0.0019</td>
<td>0.04</td>
<td>0.37</td>
<td>0.54</td>
<td>0.04</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>Between population</td>
<td>0.0021</td>
<td>0.0039</td>
<td>0.01</td>
<td>0.98</td>
<td>0.01</td>
<td>0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forked:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within population</td>
<td>0.0057</td>
<td>0.0020</td>
<td>0.06</td>
<td>0.33</td>
<td>0.62</td>
<td>-0.01</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>Between population</td>
<td>0.0026</td>
<td>0.0022</td>
<td>0.07</td>
<td>0.95</td>
<td>0.00</td>
<td>-0.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D. melanogaster: ADH:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within population</td>
<td>0.0056</td>
<td>0.0028</td>
<td>0.03</td>
<td>0.89</td>
<td>0.06</td>
<td>0.02</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>Between population</td>
<td>0.0013</td>
<td>0.0030</td>
<td>0.04</td>
<td>0.82</td>
<td>0.00</td>
<td>0.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Notch:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within population</td>
<td>0.0045</td>
<td>0.0018</td>
<td>0.15</td>
<td>0.50</td>
<td>0.30</td>
<td>0.05</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>Between population</td>
<td>0.0012</td>
<td>0.0024</td>
<td>0.11</td>
<td>0.81</td>
<td>0.00</td>
<td>0.08</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zeste-tko:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within population</td>
<td>0.0041</td>
<td>0.0018</td>
<td>0.03</td>
<td>0.88</td>
<td>0.09</td>
<td>0.00</td>
<td>3</td>
<td>21</td>
</tr>
<tr>
<td>Between population</td>
<td>0.0004</td>
<td>0.0013</td>
<td>0.04</td>
<td>0.89</td>
<td>0.00</td>
<td>0.07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yellow-achaete-scute:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within population</td>
<td>0.0003</td>
<td>0.0002</td>
<td>0.07</td>
<td>0.62</td>
<td>0.30</td>
<td>0.01</td>
<td>3</td>
<td>21</td>
</tr>
<tr>
<td>Between population</td>
<td>0.0000</td>
<td>0.0002</td>
<td>0.01</td>
<td>0.79</td>
<td>0.00</td>
<td>0.20</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Columns are as defined in table 2.
* Source: Aquadro et al. (1986).
* Source: Schaeffer et al. (1988).
* Source: Aguadé et al. (1989b).
* Source: Aguadé et al. (1989a).

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 \( \bar{\theta}_w \) and \( \bar{\theta}_h \) 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 \( \text{Var}_h \), \( \text{Cov}_{h,h} \), and \( \Delta \), are dropped from consideration), a great deal of computational simplicity is
Table 4
Nucleotide ($N_{ST}$) and Haplotype ($G_{ST}$ and $\theta$) Population Subdivision

<table>
<thead>
<tr>
<th>NUCLEOTIDES</th>
<th>HAPLOTYPES</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_{ST}$</td>
<td>$SE^a$</td>
</tr>
</tbody>
</table>

Mitochondrial DNA:
- **Odocolleus virginianus**
  - $0.75^{**}$
  - $0.23$
  - $0.62^{**}$
  - $0.69^{**}$
- **Agelaius phoenicus**
  - $0.19$
  - $0.46$
  - $0.08$
  - $0.11$
- **Aris 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**
  - $-0.06$
  - $0.52$
  - $-0.10$
  - $0.05$
  - $0.62^{**}$
  - $0.21$
  - $0.61^*$
  - $0.66^*$
- **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**
    - $0.44$
    - $0.45$
    - $0.30^{**}$
    - $0.51^{**}$
  - **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-iko**
    - $0.09$
    - $0.24$
    - $0.07$
    - $0.07$
  - **yellow-achaete-scute**
    - $0.03$
    - $0.45$
    - $0.10$
    - $0.11$

* Square root of eq. (37).
* Computed by use of eqs. (3), (15), and (36) by setting all $\delta_{nm} = 1$; this is a slight deviation from Nei (1986) in the way populations are weighted but is otherwise identical in structure to $N_{ST}$.
  * $P < .05$.
  ** $P < .01$.

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 $\delta_w$ and $\delta_b$. 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 esti-
mates 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 sta-
tistical tests. For example, one could test for significant evolutionary divergence between
two populations by pooling the samples from both populations. An empirical distri-
bution 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 hy-
potheses, 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 com-
puter 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. 1989a. Reduced variation in the yellow-
acaete-scute region in natural populations of Drosophila melanogaster. Genetics 122:607-
615.
—. 1989b. Restriction-map variation at the Zeste-tko region in natural populations of
1986. Molecular population genetics of the alcohol dehydrogenase gene region of Drosophila
AVISE, J. C., J. ARNOLD, R. M. BALL, E. BERMINGHAM, T. LAMB, J. E. NEIGEL, C. A. REEB,
and N. C. SAUNDERS. 1987a. Intraspecific phylogeography: the mitochondrial DNA bridge
AVISE, J. C., C. A. REEB, and N. C. SAUNDERS. 1987b. Geographic population structure and
species differences in mitochondrial DNA of mouthbrooding marine catfishes (Ariidae) and demersal spawning toadfishes (Batrachoididae). Evolution 41:991-1002.


SAITOU, N., and M. NEI. 1986. The number of nucleotides required to determine the branching


MASATOSHI NEI, reviewing editor

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

Accepted March 27, 1990