Measuring and Testing Genetic Differentiation With Ordered *Versus* Unordered Alleles

O. Pons* and R. J. Petit†

*Institut National de la Recherche Agronomique, *Laboratoire de Biométrie, 78352 Jouy-en-Josas cedex, France and †Laboratoire de Génétique et Amélioration des Arbres Forestiers, 33611 Gazinet cedex, France*

Manuscript received December 28, 1995
Accepted for publication July 13, 1996

ABSTRACT

Estimates and variances of diversity and differentiation measures in subdivided populations are proposed that can be applied to haplotypes (ordered alleles such as DNA sequences, which may contain a record of their own histories). Hence, two measures of differentiation can be compared for a single data set: one (G_{ST}) that makes use only of the allelic frequencies and the other (N_{ST}) for which similarities between the haplotypes are taken into account in addition. Tests are proposed to compare N_{ST} and G_{ST} with zero and with each other. The difference between N_{ST} and G_{ST} can be caused by several factors, including sampling artefacts, unequal effect of mutation rates and phylogeographic structure. The method presented is applied to a published data set where a nuclear DNA sequence had been determined from individuals of a grasshopper distributed in 24 regions of Europe. Additional insights into the genetic subdivision of these populations are obtained by progressively combining related haplotypes and reanalyzing the data each time.

So far, in population genetics surveys, differentiation has been typically estimated on the basis of the frequency of unordered alleles. With the advent of more efficient molecular techniques, genetic distances between alleles (which are then called haplotypes, *i.e.*, a particular sequence or restriction map) are becoming available. This molecular information may be used to improve our knowledge of the genetic structure of populations. As early as 1979, NEI and LI proposed to input such information in measures of genetic diversity. NEI (1982) then introduced an index of differentiation at the nucleotide level for fixed populations. For population studies involving multiple samples, LYNCH and CREASE (1990) then proposed estimates of the diversity and differentiation indices for a general population subdivided into an unknown number of subpopulations, however they did not define the indices themselves. Later, HUDSON *et al.* (1992) proposed a method to test genetic differences between populations for ordered alleles.

Here we propose to extend the work of PONS and PETIT (1995) on the estimation and variance of gene diversity and differentiation to the case of ordered alleles. A distance between the haplotypes is now introduced as a weighting matrix. For constant and parametric weights, we analyze the variance of the estimates by an expansion according to nucleotide, within-population and between-population variations, together with their interactions. LYNCH and CREASE (1990) already studied such a partition of the variance but their results depend on the data, which should not be the case for variances, and some of them may have negative values. More precise statements were therefore necessary.

We further show that G_{ST}, as defined in PONS and PETIT (1995), can be considered as a particular case of N_{ST}, as defined here. The existence of a significant genetic structure can be tested by comparing either G_{ST} or N_{ST} with zero, but the latter test will generally be more powerful. Furthermore, the comparison of N_{ST} with G_{ST} may also give some useful insights into the pattern of subdivision of the species studied, and we therefore show here how to test whether these two measures actually differ.

The theory is applied to a recently published data set where 561 noncoding nuclear sequences (haplotypes) from 24 geographic regions were obtained in a grasshopper (COOPER *et al.* 1995).

DEFINITION AND ESTIMATION OF THE HAPLOTYPE DIVERSITIES AND DIFFERENTIATION

We consider I distinct haplotypes at a locus in a general haploid population subdivided into a large number of populations. The relative sizes of the populations are generally unknown and we assume that they are similar so that we can estimate the haplotype frequencies in the total population by the average of the within-population frequencies. Moreover, the population sizes are assumed to be much larger than the sample sizes. As usual, at least implicitly in estimating the average within-population diversity, we make the approximation that the populations are independent.

Corresponding author: R. J. Petit, Institut National de la Recherche Agronomique, Laboratoire de Génétique et Amélioration des Arbres Forestiers, B.P.45, 33611 Gazinet cedex, France.
E-mail: remy@pierroton.inra.fr

Let p_i be the frequency of the ith haplotype in the general population and p_{ki} be the frequency of the ith haplotype within the kth population. The p_{ki}'s are considered as random frequencies with mean p_i, variance v_i and c_{ij} denotes the covariance between p_{ki} and p_{kj} in the general population. Weights π_{ij} are defined as distances between the haplotypes i and j. They may be the proportion of different nucleotides or restriction sites between the haplotypes, the proportion of nucleotide substitutions per site between the haplotypes or any other relevant measure, with $\pi_{ii} = 0$ and $\pi_{ij} = \pi_{ji}$. They also adapt to multiloci studies, as the proportion loci that have different alleles between two loci sets. The diversity of the kth population is

$$v_k = \sum_{ij} \pi_{ij} p_{ki} p_{kj}. \quad (1)$$

We define the average within-population diversity as the expectation of v_k in the general population, namely

$$v_S = \sum_{ij} \pi_{ij}(p_i p_j + c_{ij}), \quad (2)$$

and the total diversity is defined as

$$v_T = \sum_{ij} \pi_{ij} p_i p_j. \quad (3)$$

Introducing (2) and (3) instead of NEI's diversities in the differentiation parameter G_{ST}, we obtain $N_{ST} = (v_T - v_S)/v_T$, *i.e.*,

$$N_{ST} = \frac{\sum_{ij} \pi_{ij} c_{ij}}{v_T}, \quad (4)$$

which does not depend any more on the observations but only on the model parameters.

The weights π_{ij} (*i.e.*, the distances between the haplotypes i and j) can be known constants or parameters that need to be estimated. For instance, if complete gene sequences are available, π_{ij} can be chosen as the proportion of different nucleotides between two sequences. This purely descriptive distance is a constant known without error but the results cannot be extrapolated to the whole genome or even to a longer sequence. If the available sequence [or restriction fragment length polymorphism (RFLP) data] is considered as a sample representative of a longer sequence that is of interest, π_{ij} should be considered as a parameter that needs to be estimated. The diversities can additionally be related to the evolutionary process that has generated the sequence of interest. Then, whether or not an extrapolation to a longer sequence is to be made, π_{ij} can be defined as a parametric phylogenetic distance, such as the percentage of nucleotide substitution between haplotypes or the coalescing time. They have to be estimated in an evolutionary model that will not be considered here.

For a single locus, let $\pi_{ij} = 1$, if $i \neq j$, and 0 otherwise. In that case, v_S and v_T are equal to $h_S = 1 - \Sigma_i (p_i^2 + v_i)$ and $h_T = 1 - \Sigma_i p_i^2$, the diversity indices defined for a general population subdivided into an unknown number of subpopulations (PONS and PETIT 1995). When applied to multiloci analyses, the above diversities v_S and v_T are related to the classical gene diversities and, for independent loci, they may be compared to NEI's average diversities. For L independent loci having the average and total diversities h_{Sl} and h_{Tl}, $l \leq L$, if the distance π is the proportion of different alleles between two sets of alleles at the L loci, we get $v_S = \bar{h}_{S,L}$ defined as $\Sigma_{l \leq L} h_{Sl}/L$, $v_T = \bar{h}_{T,L} \equiv \Sigma_{l \leq L} h_{Tl}/L$ and the associated multiloci differentiation is $\bar{F}_{ST,L} = 1 - \bar{h}_{S,L}/\bar{h}_{T,L}$, which is similar to NEI's mean multiloci differentiation (1977).

The diversity and differentiation parameters are estimated using a two-level random sampling for the frequency parameters and an independent nucleotide sampling for the π_{ij} parameters when they are unknown parameters, such as a proportion of nucleotide substitutions between haplotypes. First, n populations are drawn independently and uniformly in the general population, then a sample of n_k independent individuals is drawn from the kth sampled population (where n is assumed to be large and $n_k > 3$ for each k). In the kth sampled population, we observe x_{ki}, the empirical frequency of the ith haplotype. We assume that estimates $\hat{\pi}_{ij}$ of the parameters π_{ij} are obtained from the nucleotide sampling when necessary and that the haplotype and nucleotide sampling are independent. In the following, we denote by E the expectation associated with the global sampling distribution, E_k is the expectation within the kth population, E_{pop} is the expectation with respect to the distribution of the populations. Similar notations are also used for the variances.

Extending the previous works about nucleotide diversity (NEI 1987; LYNCH and CREASE 1990) and about gene diversity (NEI 1987; PONS and PETIT 1995), unbiased estimates are based on multinomial distributions. If the weights π_{ij} are known constants, we get

$$\hat{v}_k = \frac{n_k}{n_k - 1} \sum_{ij} \pi_{ij} x_{ki} x_{kj}, \quad (5)$$

$$\hat{v}_S = \frac{1}{n} \sum_{k \leq n} \hat{v}_k = \frac{1}{n} \sum_k \frac{n_k}{n_k - 1} \sum_{ij} \pi_{ij} x_{ki} x_{kj}, \quad (6)$$

$$\hat{v}_T = \frac{1}{n(n-1)} \sum_{k \neq l} \sum_{ij} \pi_{ij} x_{ki} x_{lj} = \sum_{ij} \pi_{ij} x_{.i} x_{.j} - \frac{1}{n(n-1)} \sum_k \sum_{ij} \pi_{ij} (x_{ki} - x_{.i})(x_{kj} - x_{.j}) \quad (7)$$

and an estimate of the differentiation parameter N_{ST} is deduced as

$$\hat{N}_{ST} = 1 - \frac{\hat{v}_S}{\hat{v}_T}. \quad (8)$$

If the weights π_{ij} are parameters, they are replaced by their estimators in (5)–(7). The estimators (5)–(7) are

only denoted $\hat{v}$ (*i.e.*, respectively, $\hat{v}_k$, $\hat{v}_S$ and $\hat{v}_T$) when there is no ambiguity or $\hat{v}(\hat{\pi})$ for estimated π and $\hat{v}(\pi)$ for constant π when it must be specified. We also denote $v_k(\pi)$ for v_k and $v_k(\hat{\pi})$ for $\Sigma_{ij}\,\hat{\pi}_{ij}p_{ki}p_{kj}$. In the special case where π_{ij} is the proportion of nucleotide substitution per site between haplotypes i and j, (6) and (7) are equal to LYNCH and CREASE's estimators (1990) with $\hat{v}_b = \hat{v}_T - \hat{v}_S$ and $\hat{v}_w = \hat{v}_S$ in their notations, though they did not define the diversity indices.

VARIANCE OF THE DIVERSITIES WITH CONSTANT π

If the weights π_{ij} are known constants, the variance of $\hat{v}_S$ is the sum of the within-population and between-population variances

$$Var(\hat{v}_S) = n^{-2}\sum_k EVar_k(\hat{v}_k) + n^{-1}Var(v_k) \qquad (9)$$

because of the two-level random sampling. The within-population variance $Var_{intra}(\hat{v}_S) = n^{-2}\,\Sigma_k\, EVar_k(\hat{v}_k)$ is given from NEI (1987, ch. 10)

$$Var_k(\hat{v}_k) = \frac{2}{n_k(n_k-1)}\left\{(3-2n_k)v_k^2 + 2(n_k-2)\times\sum_{ija}\pi_{ij}\pi_{ia}p_{ki}p_{kj}p_{ka} + \sum_{ij}\pi_{ij}^2p_{ki}p_{kj}\right\}, \qquad (10)$$

and the between-population variance is

$$Var_{inter}(\hat{v}_S) = \frac{1}{n}\{E(v_k^2) - v_S^2\}. \qquad (11)$$

The sampling variance of $\hat{v}_S$ is unbiasedly estimated by

$$\hat{V}ar(\hat{v}_S) = \frac{1}{n(n-1)}\sum_k(\hat{v}_k - \hat{v}_S)^2, \qquad (12)$$

and an unbiased estimate of $Var_{inter}(\hat{v}_S)$ is obtained as

$$\hat{V}ar_{inter}(\hat{v}_S) = \frac{1}{n}\{\hat{E}(v_k^2) - \hat{v}_S^2 + \hat{V}ar(\hat{v}_S)\},$$

where $\hat{E}(v_k^2)$ is the following unbiased estimate of $E(v_k^2)$

$$\hat{E}(v_k^2) = \frac{1}{n}\sum_k\frac{1}{(n_k-1)(n_k-2)(n_k-3)}\times\left\{n_k^3\left(\sum_{ij}\pi_{ij}x_{ki}x_{kj}\right)^2 + 2n_k\sum_{ij}\pi_{ij}^2x_{ki}x_{kj} - 2n_k^2\times\left[2\sum_{ija\neq}\pi_{ij}\pi_{ia}x_{ki}x_{kj}x_{ka} + \sum_{ij}\pi_{ij}^2(x_{ki}^2x_{kj} + x_{ki}x_{kj}^2)\right]\right\},$$

then an unbiased estimate of $Var_{intra}(\hat{v}_S)$ follows from the difference $\hat{V}ar_{intra}(\hat{v}_S) = \hat{V}ar(\hat{v}_S) - \hat{V}ar_{inter}(\hat{v}_S)$.

The variance of $\hat{v}_T$ and the covariance of $\hat{v}_S$ and $\hat{v}_T$ can be approximated and estimated as in PONS and PETIT (1995) when n is assumed to be large. We get

$$Var(\hat{v}_T) = \frac{4}{n^2}\sum_k\sum_{ij}\sum_{ab}\pi_{ij}\pi_{ab}(Ex_{ki}x_{ka} - p_ip_a)p_jp_b + O(n^{-2}),$$

and it splits into a between-population variance

$$Var_{inter}(\hat{v}_T) = \frac{4}{n}\sum_{ij}\sum_{ab}\pi_{ij}\pi_{ab}c_{ia}p_jp_b + O(n^{-2}) \qquad (13)$$

and a within-population variance

$$Var_{intra}(\hat{v}_T) = \frac{4}{n\tilde{n}}\left\{\sum_{ija}\pi_{ij}\pi_{ia}p_ip_jp_a - \sum_{ij}\sum_{ab}\pi_{ij}\pi_{ab}(p_ip_a + c_{ia})p_jp_b\right\} + O(n^{-2}), \qquad (14)$$

where $\tilde{n}$ is the harmonic mean of the n_k's. Consistent estimates of the variances of $\hat{v}_T$ are given by

$$\hat{V}ar(\hat{v}_T) = \frac{4}{n(n-1)}\sum_{ij}\sum_{ab}\pi_{ij}\pi_{ab}x_{\cdot j}x_{\cdot b}\sum_k(x_{ki}-x_{\cdot i})(x_{ka}-x_{\cdot a}), \qquad (15)$$

$$\hat{V}ar_{intra}(\hat{v}_T) = \frac{4}{n^2}\sum_k\frac{1}{n_k-1}\sum_{ijb}\pi_{ij}x_{ki}x_{\cdot j}x_{\cdot b}\left\{\pi_{ib} - \sum_a\pi_{ab}x_{ka}\right\} \qquad (16)$$

then $\hat{V}ar_{inter}(\hat{v}_T) = \hat{V}ar(\hat{v}_T) - \hat{V}ar_{intra}(\hat{v}_T)$.

Moreover, the covariance of $\hat{v}_S$ and $\hat{v}_T$ is approximated as

$$Cov(\hat{v}_S, \hat{v}_T) = \frac{2}{n}\left\{\frac{1}{n}\sum_k\sum_{ij}\pi_{ij}p_iE(x_{kj}\hat{v}_k) - v_Sv_T\right\} + O(n^{-2}), \qquad (17)$$

it is the sum of the between-population covariance

$$Cov_{inter}(\hat{v}_S, \hat{v}_T) \simeq \frac{2}{n}\left\{\sum_{ij}\sum_{ab}\pi_{ij}\pi_{ab}p_iE(p_{kj}p_{ka}p_{kb}) - v_Sv_T\right\} \qquad (18)$$

and of the within-population covariance

$$Cov_{intra}(\hat{v}_S, \hat{v}_T) \simeq \frac{4}{n\tilde{n}}\left\{\sum_{ij}\pi_{ij}p_i\left\{\sum_b\pi_{jb}E(p_{kj}p_{kb})\right\} - \sum_{ab}\pi_{ab}E(p_{kj}p_{ka}p_{kb})\right\}. \qquad (19)$$

From the moments of the multinomial distribution, they are consistently estimated by

$$\hat{C}ov(\hat{v}_S, \hat{v}_T) = \frac{2}{n-2}\left\{\frac{1}{n-1}\sum_{kij}\pi_{ij}\left(x_{\cdot i} - \frac{x_{ki}}{n}\right)\hat{v}_kx_{kj} - \hat{v}_S\hat{v}_T\right\}, \qquad (20)$$

$$\hat{Cov}_{intra}(\hat{v}_S, \hat{v}_T) = \frac{4}{n(n-1)} \sum_{kij} \pi_{ij}\left(x_{.i} - \frac{x_{ki}}{n}\right) \times \frac{n_k}{(n_k-1)(n_k-2)} \times \left\{\sum_b \pi_{jb} x_{kj} x_{kb} - \sum_{ab} \pi_{ab} x_{kj} x_{ka} x_{kb}\right\},$$

and $\hat{Cov}_{inter}(\hat{v}_S, \hat{v}_T)$ is obtained by difference.

If n is large, the variance of $\hat{N}_{ST}$ is approximated by

$$(v_T)^{-2} Var(\hat{v}_S) - 2v_S(v_T)^{-3} Cov(\hat{v}_S, \hat{v}_T) + v_S^2(v_T)^{-4} Var(\hat{v}_T), \quad (21)$$

and it splits into within- and between-population variances of the same form but with the corresponding decompositions of the variances and covariance of $\hat{v}_S$ and $\hat{v}_T$. When a locus with unordered alleles was considered, an optimal sampling design was defined to yield the optimal sample size per population to have the smallest variance of $\hat{G}_{ST}$ (PONS and PETIT 1995). The same result applies for haplotypes and we obtain the optimal population sampling size

$$\tilde{n}_{opt} = \frac{A - C + D - E}{A - \sqrt{A(C - D + E)}},$$

where $A = nVar_{inter}(\hat{N}_{ST})$ is approximately a constant as n increases and the other terms are the coefficients of $(3 - 2\tilde{n})/\tilde{n}(\tilde{n} - 1)$, $(\tilde{n} - 2)/\tilde{n}(\tilde{n} - 1)$ and $1/\tilde{n}(\tilde{n} - 1)$ that appear in $2v_T^{-2} Var_{intra}(\hat{v}_S)$. Thus we have

$$A = \frac{n}{v_T^2} Var_{inter}(\hat{v}_S) + \frac{nv_S^2}{v_T^4} Var_{inter}(\hat{v}_T) - \frac{2nv_S}{v_T^3} Cov_{inter}(\hat{v}_S, \hat{v}_T),$$

$$C = \frac{2}{v_T^2} E(v_k^2),$$

$$D = \frac{4}{v_T^2} E\left(\sum_{ija} \pi_{ij}\pi_{ia} p_{ki} p_{kj} p_{ka}\right),$$

$$E = \frac{2}{v_T^2} E\left(\sum_{ij} \pi_{ij}^2 p_{ki} p_{kj}\right).$$

These quantities are estimated by plugging estimates of each parameter in these expressions, with $n^{-1} \Sigma_k \hat{E}(v_k^2)$ for $E(v_k^2)$, $n^{-1} \Sigma_k n_k^2[(n_k - 1)(n_k - 2)]^{-1}\{\Sigma_{ija} \pi_{ij}\pi_{ia} x_{ki} x_{kj} x_{ka} - n_k^{-1} \Sigma_{ij} \pi_{ij}^2 x_{ki} x_{kj}\}$ for $E(\Sigma_{ija} \pi_{ij}\pi_{ia} p_{ki} p_{kj} p_{ka})$, and $n^{-1}\Sigma_k n_k(n_k - 1)^{-1} \Sigma_{ij} \pi_{ij}^2 x_{ki} x_{kj}$ for $E(\Sigma_{ij} \pi_{ij}^2 p_{ki} p_{kj})$.

VARIANCE OF THE DIVERSITIES WITH PARAMETER π

When $\hat{v}_S$ depends on estimates of the parameters π_{ij}, the variance of $\hat{v}_S(\hat{\pi})$ develops as a sum of variances due to the nucleotides sampling, populations sampling, individuals within populations sampling, and nucleotides-populations and nucleotides-individuals interaction terms. We denote E_{pop} the mean under the populations sampling distribution, E_n, Var_n and Cov_n the mean, variance and covariance under the nucleotides sampling distribution and $Var_{n,inter}$ (or $Var_{n,pop}$) and $Var_{n,intra}$ (or $Var_{n,k}$) the interaction variances due to nucleotide sampling and population or, respectively, individuals sampling. We have

$$Var(\hat{v}_S(\hat{\pi})) = \frac{1}{n^2}\left\{\sum_k Var(\hat{v}_k(\hat{\pi})) + \sum_{k \neq l} Cov(\hat{v}_k(\hat{\pi}), \hat{v}_l(\hat{\pi}))\right\}.$$

By the assumed independence of the populations, the covariance of $\hat{v}_k(\hat{\pi})$ and $\hat{v}_l(\hat{\pi})$ is only due to the nucleotide sampling if $k \neq l$ (*cf.* formula 12 in LYNCH and CREASE 1990), therefore

$$Cov(\hat{v}_k(\hat{\pi}), \hat{v}_l(\hat{\pi})) = E_{pop} Cov_n(v_k(\hat{\pi}), v_l(\hat{\pi}))$$

and this is also equal to the nucleotide sampling variance of $\hat{v}_S(\hat{\pi})$

$$Var_n(v_S(\hat{\pi})) = \sum_{ijab} Cov_n(\hat{\pi}_{ij}, \hat{\pi}_{ab})(p_i p_j + c_{ij})(p_a p_b + c_{ab}).$$

The variance of $\hat{v}_k(\hat{\pi})$ has a decomposition of the same type as (5) in LYNCH and CREASE (1990), but with means under the general distribution of individuals, populations and nucleotides. More precisely,

$$Var(\hat{v}_k(\hat{\pi})) = E_{pop}E_n Var_k(\hat{v}_k(\hat{\pi})) + E_{pop} Var_n(v_k(\hat{\pi})) + Var(v_k(\pi))$$

and LYNCH and CREASE'S residual term $\Delta_r(\hat{v}_k)$ equals $Var(v_k(\pi))$, it is therefore always positive though they claim the opposite. The results obtained for constant weights apply, yielding

$$E_n Var_k(\hat{v}_k(\hat{\pi})) = Var_k(\hat{v}_k(\pi)) + Var_{n,k}(\hat{v}_k(\hat{\pi})),$$

where the interaction term between the nucleotides and individuals sampled is

$$Var_{n,k}(\hat{v}_k(\hat{\pi})) = \frac{2}{n_k(n_k - 1)}\left\{(3 - 2n_k) Var_n(v_k(\hat{\pi})) + 2(n_k - 2)\left[\sum_{ija} Cov_n(\hat{\pi}_{ij}, \hat{\pi}_{ia}) p_{ki} p_{kj} p_{ka}\right] + \sum_{ij} Var_n(\hat{\pi}_{ij}) p_{ki} p_{kj}\right\}$$

and

$$Var_n(v_k(\hat{\pi})) = \sum_{ijab} Cov_n(\hat{\pi}_{ij}, \hat{\pi}_{ab}) p_{ki} p_{kj} p_{ka} p_{kb}.$$

Then $E_{pop} Var_n(v_k(\hat{\pi}))$ is the sum of the nucleotide sampling variance of $\hat{v}_S(\hat{\pi})$ and of the nucleotide-population interaction variance of $\hat{v}_k(\hat{\pi})$,

$$Var_{n,pop}\hat{v}_k(\hat{\pi}) = \sum_{ijab} Cov_n(\hat{\pi}_{ij}, \hat{\pi}_{ab}) Cov(p_{ki} p_{kj}, p_{ka} p_{kb}).$$

Finally, the total variance of $\hat{v}_S(\hat{\pi})$ is the sum of the nucleo-

tide sampling variance $Var_n(\hat{v}_S(\hat{\pi}))$, the between-population variance $Var_{inter}(\hat{v}_S(\hat{\pi})) = n^{-1}Var(v_k(\pi))$, the within-population variance $Var_{intra}(\hat{v}_S(\hat{\pi})) = n^{-2}\Sigma_k E_{pop}Var_k(\hat{v}_k(\pi))$, the interaction variance due to the nucleotide and individual sampling, $Var_{n,intra}(\hat{v}_S(\hat{\pi})) = n^{-2}\Sigma_k E_{pop}Var_{n,k}(\hat{v}_k(\hat{\pi}))$ and the interaction variance due to the nucleotide and population sampling, $Var_{n,inter}(\hat{v}_S(\hat{\pi})) = Var_{n,pop}\hat{v}_k(\hat{\pi})$.

Alternatively, if we first consider the variance of $\hat{v}_S(\hat{\pi})$ conditionally on the nucleotide sampling and denote $Var(\hat{v}_S)(\hat{\pi})$ the value at $\hat{\pi}$ of $Var(\hat{v}_S(\pi))$ given by (9), (10) and (11),

$$Var(\hat{v}_S(\hat{\pi})) = E_nVar(\hat{v}_S)(\hat{\pi}) + Var_n(v_S(\hat{\pi})),$$

where $Var(\hat{v}_S)(\hat{\pi})$ develops as above. Using estimates of the covariances between the $\hat{\pi}_{ij}$'s that are independent of the x_{ki}'s, we derive the next estimates

$$\hat{V}ar(\hat{v}_S(\hat{\pi})) = \hat{V}ar(\hat{v}_S)(\hat{\pi}) + \hat{V}ar_n(v_S(\hat{\pi}))$$

$$\text{where } \hat{V}ar(\hat{v}_S)(\hat{\pi}) = \frac{1}{n(n-1)}\sum_k (\hat{v}_k(\hat{\pi}) - \hat{v}_S(\hat{\pi}))^2,$$

$$\hat{V}ar_n(v_S(\hat{\pi})) = \frac{1}{n(n-1)}\sum_{k\neq l}\sum_{ijab}\hat{C}ov_n(\hat{\pi}_{ij}, \hat{\pi}_{ab})\left(\frac{n_k}{n_k-1}\right)^2 x_{ki}x_{kj}x_{la}x_{lb}.$$

A similar decomposition holds for $\hat{v}_T(\hat{\pi})$,

$$Var(\hat{v}_T(\hat{\pi})) = E_nVar(\hat{v}_T)(\hat{\pi}) + Var_n(v_T(\hat{\pi})),$$

where the second term is the part of the variance due to nucleotide sampling,

$$Var_n(v_T(\hat{\pi})) = \sum_{ijab} Cov_n(\hat{\pi}_{ij}, \hat{\pi}_{ab})p_ip_jp_ap_b$$

and $Var(\hat{v}_T)(\hat{\pi})$ is given by the sum of (13) and (14) at $\hat{\pi}$ instead of π. This provides

$$E_nVar(\hat{v}_T)(\hat{\pi}) = \frac{4}{n^2}\sum_k\sum_{ijab}E_n\hat{\pi}_{ij}\hat{\pi}_{ab}E_{pop}E_k(x_{ki}x_{ka} - p_ip_a)p_jp_b$$

$$+ O(n^{-2}) = Var(\hat{v}_T(\pi)) + \frac{4}{n}\sum_{ijab}Cov_n(\hat{\pi}_{ij}, \hat{\pi}_{ab})c_{ia}p_jp_b$$

$$+ \frac{4}{n\tilde{n}}\left\{\sum_{ija}Cov_n(\hat{\pi}_{ij}, \hat{\pi}_{ia})p_ip_jp_a - \sum_{ijab}Cov_n(\hat{\pi}_{ij}, \hat{\pi}_{ab})(p_ip_a + c_{ia})p_jp_b\right\} + O(n^{-2})$$

and a decomposition of $E_nVar(\hat{v}_T)(\hat{\pi})$ follows as a sum of within- and between-populations variances $Var_{intra}(\hat{v}_T(\pi))$ and $Var_{inter}(\hat{v}_T(\pi))$, just as in $Var(\hat{v}_T(\pi))$, and two variation terms due to the interaction of the nucleotide sampling and these terms

$$Var_{n,intra}(\hat{v}_T(\hat{\pi})) = 4/n\tilde{n}\{\Sigma_{ija}Cov_n(\hat{\pi}_{ij},\hat{\pi}_{ia})p_ip_jp_a - \Sigma_{ijab}Cov_n(\hat{\pi}_{ij}, \hat{\pi}_{ab})(p_ip_a + c_{ia})p_jp_b\} + O(n^{-2})$$

and

$$Var_{n,inter}(\hat{v}_T(\hat{\pi})) = 4/n\,\Sigma_{ijab}Cov_n(\hat{\pi}_{ij}, \hat{\pi}_{ab})c_{ia}p_jp_b + O(n^{-2}).$$

An estimate of $Var(\hat{v}_T(\hat{\pi}))$ is given by the sum of estimates of $E_nVar(\hat{v}_T)(\hat{\pi})$ and $Var_n(v_T(\hat{\pi}))$, namely

$$\hat{V}ar(\hat{v}_T)(\hat{\pi}) = \frac{4}{n(n-1)}\sum_{ijab}\hat{\pi}_{ij}\hat{\pi}_{ab}x_{.j}x_{.b}\sum_k (x_{ki} - x_{.i})(x_{ka} - x_{.a})$$

and

$$\hat{V}ar_n(\hat{v}_T(\hat{\pi})) = \sum_{ijab}\hat{C}ov(\hat{\pi}_{ij}, \hat{\pi}_{ab})x_{.i}x_{.j}x_{.a}x_{.b}.$$

In the same way, a decomposition of the covariance of $\hat{v}_S(\hat{\pi})$ and $\hat{v}_T(\hat{\pi})$ derives from the results with constant π and the partition

$$Cov(\hat{v}_S(\hat{\pi}), \hat{v}_T(\hat{\pi})) = E_nCov(\hat{v}_S, \hat{v}_T)(\hat{\pi}) + Cov_n(v_S(\hat{\pi}), v_T(\hat{\pi})).$$

The nucleotide sampling variance is

$$Cov_n(v_S(\hat{\pi}), v_T(\hat{\pi})) = \sum_{ijab}Cov_n(\hat{\pi}_{ij}, \hat{\pi}_{ab})p_ip_j(p_ap_b + c_{ab}),$$

$$E_nCov(\hat{v}_S, \hat{v}_T)(\hat{\pi}) = \frac{2}{n}\left[\sum_{ij}p_i\left\{\left(1 - \frac{2}{\hat{n}}\right)\sum_{ab}Cov_n(\hat{\pi}_{ij}, \hat{\pi}_{ab})E(p_{kj}p_{ka}p_{kb}) + \frac{2}{\hat{n}}\sum_b Cov_n(\hat{\pi}_{ij}, \hat{\pi}_{jb})E(p_{kj}p_{kb})\right\}\right] + Cov(\hat{v}_S(\pi), \hat{v}_T(\pi)) + O(n^{-2})$$

and a decomposition into within and between populations terms and interactions terms with the nucleotide level is similar to that which is described for constant π.

A consistent estimate of $E_nCov(\hat{v}_S, \hat{v}_T)(\hat{\pi})$ is given by

$$\hat{C}ov(\hat{v}_S, \hat{v}_T)(\hat{\pi}) = \frac{2}{n(n-2)}\sum_k\sum_{ij}x_{.i}\hat{\pi}_{ij}\hat{v}_k(x_{kj} - x_{.j}),$$

whereas the covariance at the nucleotide level is estimated by

$$\hat{C}ov_n(\hat{v}_S(\hat{\pi}), \hat{v}_T(\hat{\pi})) = \sum_{ijab}x_{.i}x_{.j}\hat{C}ov(\hat{\pi}_{ij}, \hat{\pi}_{ab})\frac{1}{n}\sum_k\frac{n_k}{n_k-1}x_{ka}x_{kb},$$

then $\hat{C}ov(\hat{v}_S(\hat{\pi}), \hat{v}_T(\hat{\pi}))$ is the sum of these two variance estimates. Finally, the variance of $\hat{N}_{ST}$ is estimated introducing these variances and covariances in (21) and it develops according to the same sampling variations.

TEST OF THE GENETIC STRUCTURE

Several tests can be proposed to examine the significance of the genetic structure. First, one may wish to test, for ordered and unordered alleles, whether there is indeed a geographic structure. For this purpose, a test of the hypothesis $N_{ST} = 0$ or $G_{ST} = 0$ may be used.

Since the population number n is assumed to be large, their estimates $\hat{N}_{ST}$ (8) and $\hat{G}_{ST}$ (PONS and PETIT 1995) are approximated by Gaussian variables having the means G_{ST} and N_{ST}, respectively. It follows that the statistics

$$U_G = \frac{\hat{G}_{ST}}{\sqrt{Var(\hat{G}_{ST})}} \quad \text{and} \quad U_N = \frac{\hat{N}_{ST}}{\sqrt{Var(\hat{N}_{ST})}}$$

are approximately standard Gaussian variables under the hypotheses $N_{ST} = 0$ and $G_{ST} = 0$, respectively. We will therefore use them as test statistics. Up to the expression of its denominator, U_N^2 is similar to LYNCH and CREASE's statistic D and both are approximately χ_1^2 distributed.

It may be also useful to test whether $N_{ST} = G_{ST}$. More specifically, since there are several causes that may lead to a larger value of N_{ST} relative to G_{ST}, an unilateral test of $N_{ST} \leq G_{ST}$ against $N_{ST} > G_{ST}$ may be preferred. Such a test can be based on the comparison of the estimates $\hat{N}_{ST}$ and $\hat{G}_{ST}$. Using again their Gaussian approximation, we consider the test statistic

$$U = \frac{\hat{N}_{ST} - \hat{G}_{ST}}{\{\hat{Var}(\hat{N}_{ST}) + \hat{Var}(\hat{G}_{ST}) - 2\hat{Cov}(\hat{N}_{ST}, \hat{G}_{ST})\}^{1/2}},$$

where the covariance between $\hat{N}_{ST}$ and $\hat{G}_{ST}$ is

$$Cov(\hat{N}_{ST}, \hat{G}_{ST}) \simeq \frac{Cov(\hat{h}_S, \hat{v}_S)}{h_T v_T} + \frac{h_S v_S}{h_T v_T^2} Cov(\hat{h}_T, \hat{v}_T)$$

$$+ \frac{h_S v_S}{h_T v_T^3} \hat{Var}(\hat{v}_T) + \frac{h_S v_S}{h_T^3 v_T} \hat{Var}(\hat{h}_T) - \frac{1}{h_T v_T}\left\{\frac{h_S}{v_T} Cov(\hat{v}_S, \hat{v}_T)\right.$$

$$\left. + \frac{h_S}{h_T} Cov(\hat{v}_S, \hat{h}_T) + \frac{v_S}{v_T} Cov(\hat{h}_S, \hat{v}_T) + \frac{v_S}{h_T} Cov(\hat{h}_S, \hat{h}_T)\right\}$$

and it is estimated by replacing h_S, h_T, v_S and v_T by their estimates and the various covariances terms by the estimated ones, with

$$\hat{Cov}(\hat{h}_S, \hat{v}_S) = \frac{1}{n(n-1)} \sum_k (\hat{v}_k - \hat{v}_S)(\hat{h}_k - \hat{h}_S),$$

$$\hat{Cov}(\hat{h}_S, \hat{v}_T) = \frac{2}{n-2}\left\{\frac{1}{n-1} \sum_{kij} \pi_{ij}\left(x_{.i} - \frac{x_{ki}}{n}\right)\hat{h}_k x_{kj} - \hat{h}_S \hat{v}_T\right\},$$

$$\hat{Cov}(\hat{h}_T, \hat{v}_S) = \frac{2}{n-2}\left\{\hat{v}_S(1 - \hat{h}_T) - \frac{1}{n-1} \sum_{ki}\left(x_{.i} - \frac{x_{ki}}{n}\right)\hat{v}_k x_{ki}\right\},$$

$$\hat{Cov}(\hat{h}_T, \hat{v}_T) = \frac{4}{n(n-1)} \sum_{ij} \pi_{ij} x_{.j} x_{.b} \sum_k (x_{ki} - x_{.i})(x_{ka} - x_{.a}).$$

For large n, U is approximated by a Gaussian variable having the mean $N_{ST} - G_{ST}$ and the variance 1. Under the hypothesis $N_{ST} - G_{ST} \leq 0$ the statistic U has a negative mean, and under the alternative U increases with n so that a one-sided test has to be used. The hypothesis is therefore rejected if the value of U is larger than the upper quantile of the standard Gaussian variable.

APPLICATION

We analyzed the data presented by COOPER *et al.* (1995) on a grasshopper sampled over its natural range in Europe (*Chorthippus parallelus*). The sample consists of 561 nuclear DNA sequences belonging to individuals from 24 geographic regions (mean sample size per geographic region: 23 haplotypes, ranging from six to 71). The sequence is an anonymous noncoding segment of 281–286 bp, determined by direct sequencing after PCR amplification. There are 73 polymorphic nucleotide sites and 71 haplotypes. We computed the percentage of nucleotide differences (including insertion-deletions, all very short and therefore never encompassing more than one polymorphic nucleotide site) between all pairs of haplotypes, relative to the number of polymorphic sites. Note that we are not interested here in measures of diversity but in the parameter of differentiation, which is not modified by the introduction of the monomorphic sites. In their paper, COOPER *et al.* (1995) mentioned but did not present a phylogenetic tree obtained using the maximum parsimony method. Although the phylogenetic distances obtained by such methods, if available, may be preferred to evaluate the distances between haplotypes, they can be difficult to obtain for potentially recombining nuclear DNA sequences. In addition, the simple metric distances used here are easy to obtain and sufficient to illustrate the advantages inherent in the use of ordered haplotypes. In the absence of informations on the flanking sequences of the DNA fragment studied, we do not wish here to make extrapolations to an unknown longer sequence and we therefore consider that the π_{ij} are known constants.

Estimates of N_{ST} and G_{ST} were therefore computed. Note that G_{ST} can be considered as a particular case of N_{ST} where all the distances between the haplotypes are equal to one, except for the diagonal that is zero. The results are given in Table 1, which shows that $\hat{N}_{ST}$ is larger than $\hat{G}_{ST}$ and that $\hat{G}_{ST}$ and $\hat{N}_{ST}$ are much larger than zero. The population number $n = 24$ is large enough to ensure the Gaussian approximation of $\hat{G}_{ST}$ and $\hat{N}_{ST}$. It appears that the species is clearly structured since $U_N = 7.91$ and $U_G = 5.62$. A comparison of N_{ST} and G_{ST} was then tested and the result is significant ($U = 2.32$, $P = 0.01$).

Following our previous study on the estimation and optimal sampling of gene diversity (PONS and PETIT 1995), the optimal sampling for N_{ST} was computed. It is equal to 6.6 individuals per population, close to the

TABLE 1

Parameters of diversity and differentiation for ordered and unordered haplotypes

$\hat{v}_S$	$\hat{v}_T$	$\hat{N}_{ST}$
0.044 (0.007)[a]	0.059 (0.008)	0.253 (0.032)
$\hat{h}_S$	$\hat{h}_T$	$\hat{G}_{ST}$
0.796 (0.022)	0.933 (0.013)	0.146 (0.026)

[a] SD calculated with constant distances π_{ij}.

value 5.7 obtained for G_{ST}. As outlined in previous papers, these results emphasize the necessity of sampling as many populations as possible, with homogeneous sample sizes, since the sampling of the populations (here the geographic regions) has a predominant influence on the accuracy of the differentiation estimate. Increasing the sample size within population will improve only to a small extent the precision of the estimates. Note also the relative importance of the inter- and intrapopulation components of the variance of diversity and differentiation estimates in Table 2.

It is of interest, now that we have identified the existence of a significant geographic structure, to analyze it further to examine whether all levels of genetic relatedness between haplotypes contribute to the geographic structure, or whether some levels are more important than others. We therefore made successive analyses by progressively pooling the haplotypes that were the most closely related. To do this, we gave a value of zero for all distances between haplotypes differing at less than two, three, four, . . . , nucleotide sites. A similar operation was made for G_{ST}, using the distance approach of N_{ST} but with a matrix of ones and zeros. Actually the measure obtained is not a true G_{ST} since the distances between haplotypes are used to identify the closest haplotypes. The results are given in Figure 1. When only differences between haplotypes separated by at least three nucleotide sites are considered, N_{ST} is no longer larger than G_{ST}. Moreover, if only differences between divergent haplotypes are considered it results in higher levels of subdivision (both G_{ST} and N_{ST} become higher, with values above 0.30 when haplotypes differing by three to seven nucleotides are pooled).

TABLE 2

Estimation of intra- and interpopulation and total variances components for ordered and unordered haplotypes

	Variances × 10^5			
	Var_{intra}	Var_{inter}	Var_{tot}	Var_{inter}/Var_{tot}
$Var(\hat{v}_S)$	0.42	4.14	4.56	90.7
$Var(\hat{v}_T)$	1.12	5.37	6.49	79.8
$Var(\hat{N}_{ST})$	10.24	94.02	104.27	90.2
$Var(\hat{h}_S)$	9.35	37.06	46.42	79.8
$Var(\hat{h}_T)$	6.83	10.28	17.11	60.1
$Var(\hat{G}_{ST})$	9.73	57.47	67.20	85.5

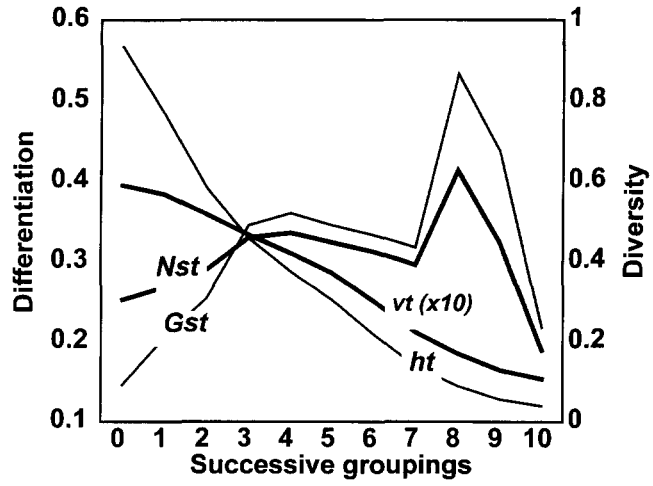

FIGURE 1.—Evolution of the genetic diversity and differentiation for ordered and unordered alleles as a function of the minimum divergence considered between haplotypes. In the successive groupings, haplotypes differing at an increasing number of nucleotide sites were pooled. At each step, all parameters were computed to study how they depend on the relatedness between haplotypes. When all nucleotides sites are considered, N_{ST} is significantly higher than G_{ST}, as expected for phylogeographically organized populations (case 1 in Figure 2). However, when the haplotypes differing by less than three sites are pooled, this is no longer true. As a consequence of pooling haplotypes, the diversity progressively decreases (the nucleotidic diversity is multiplied by a factor of 10 for illustration purpose). On the other hand, differentiation increases when only divergent haplotypes are considered in the analyses (groupings 8 and 9).

The values obtained by merging even more divergent haplotypes should be considered with caution since the level of diversity becomes very low and the precision is therefore limited.

DISCUSSION

LYNCH and CREASE (1990) mention that "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." This is illustrated in a particular simple example involving two populations and four haplotypes (Figure 2). In all cases, G_{ST} is the same since it depends only on the frequencies of the haplotypes and not on their similarities. On the other hand, N_{ST} varies according to the distances between haplotypes. In case 1, where the haplotypes found in the same population are closely related, N_{ST} is close to one and the distance between haplotypes indicates that the two populations are strongly differentiated genetically. If the distance between the four haplotypes are similar (case 2), N_{ST} is equal to G_{ST} since the haplotypes are phylogenetically equivalent. Finally, if haplotypes A and C are strongly related as are haplotypes B and D (case 3), then N_{ST} is much smaller than G_{ST} and very close to zero. The relative geographic distribution of haplotypes may have nothing to do with

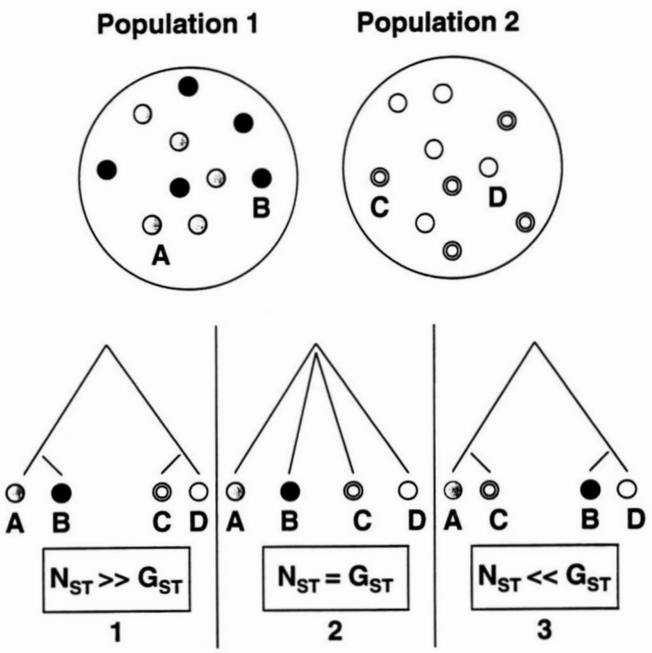

FIGURE 2.—Illustration of the correspondence between the phylogenies of the haplotypes and their geographic distribution. When there is correspondence (case 1), the differentiation measured by taking into account the similarities between haplotypes (N_{ST}) is greater than the differentiation based only on the frequency of the haplotypes (G_{ST}). When the haplotypes are equally related, $N_{ST} = G_{ST}$ (case 2). Finally, if the most strongly related haplotypes are never together but are always found in different populations, $N_{ST} \leq G_{ST}$ (case 3).

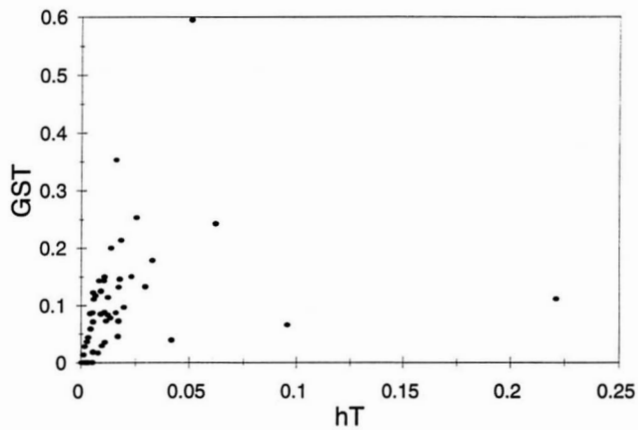

FIGURE 3.—Relation between differentiation (G_{ST}) and diversity (h_T) measured for each haplotype considered as a diallelic locus. The estimates of these two parameters were computed for the 71 haplotypes of *C. parallelus*. Note that differentiation is not independent of diversity, a situation that can bias the G_{ST} values at very polymorphic loci.

their genetic distances in the cases of old lineages that have had ample time to become geographically redistributed since they first appeared through mutations. But identifying an evolutionary process that would lead to situations where distantly related alleles are more often found within the same populations (like in case 3) seems difficult.

It is interesting to point out that COOPER *et al.* (1995) did not identify any clear pattern of phylogenetic subdivision in this grasshopper with the exception of a monophyletic group containing only Turkish and Greek haplotypes. On the other hand, they concluded that the important differences found between most populations could not have been produced by chance fluctuations due to sampling over a nondifferentiated species range. Hence, the expectation here is that subdivision should not be much higher when ordered haplotypes are considered (N_{ST}) rather than unordered ones (G_{ST}).

However, other factors than the divergence between haplotypes can bias G_{ST} downward relative to N_{ST}. In particular, G_{ST} may be differentially affected by large mutation rates at the whole haplotype level. Theoretical models (*e.g.*, TAKAHATA and NEI 1984) and simulations (*e.g.*, SLATKIN 1993) indeed indicate that for highly variable loci subdivision of gene diversity may be lower than for less variable ones submitted to similar migration rates. Although the differential effect of mutation rate on subdivision measured using unordered or ordered alleles does not seem to have been studied to date, it could be significant. Another problem for G_{ST} when dealing with very variable loci derives from the fact that in the extreme case when all alleles become extremely rare, estimates of G_{ST} will tend to zero since differentiation for each allele is negatively correlated with its diversity, regardless of the estimate of differentiation used (PETIT *et al.* 1995). This is illustrated in Figure 3 where the relationship between diversity and differentiation was studied for the 71 haplotypes (each being considered as a diallelic locus). All G_{ST} values lower than 0.05 involve rare haplotypes with a frequency lower than 0.01 (*i.e.*, h_T lower than 0.02). These rare haplotypes tend therefore to decrease the multiallelic G_{ST} value.

The availability of ordered alleles makes it possible to progressively pool some of the haplotypes according to the distance measured between them. This leads to results such as those of Figure 1. The fact that G_{ST} is rapidly no longer smaller than N_{ST} when similar haplotypes are pooled may indicate that only the very similar haplotypes are not randomly associated within populations. On the other hand, it could also indicate that the artefacts due to the increased sensibility of G_{ST} to mutation rates or to the presence of many low frequencies alleles are progressively eliminated. Nevertheless, the increase in both G_{ST} and N_{ST} when only the more remote haplotypes are distinguished from each other may indicate that the older lineages are less frequently present together in the same populations.

Since several reasons can account for a difference between N_{ST} and G_{ST}, most of them tending to minimize G_{ST} relative to N_{ST}, N_{ST} should in general be more useful than G_{ST}. Moreover, if N_{ST} is significantly higher than G_{ST}, this could be an indication that the species presents some degree of phylogenetic subdivision, *i.e.*, that on average strongly related haplotypes are more often

found together within the same populations than less related ones. In all cases, molecular informations are more efficiently used when ordered alleles are considered, leading to a more powerful test of the existence of a genetic structure, but also a better understanding of its nature.

There has been considerable debate about which method is most appropriate to analyze empirical data on the population structure of species. NEI (1986) and CHAKRABORTY and DANKER-HOPFE (1991) argue at length that NEI's (1973) gene diversity approach should be preferred for natural populations, in particular because it makes no assumption about the evolutionary process involved in shaping this structure. We follow them and consider here only the sampling variance rather than the variance caused by the evolutionary process. In analyzing natural populations likely to evolve in a very complex manner (see for instance the recent analysis of TEMPLETON *et al.* 1995), a first statistical description dealing only with sampling effects may be preferred. It may eventually be followed by more detailed and complex analyses aimed at testing specific evolutionary hypotheses, if a significant genetic structure was discovered in the initial descriptive study.

In this study, we assume that the observed populations are independent. This assumption is useful to develop a complete statistical analysis (*cf.* the discussion in LYNCH and CREASE 1990, p. 382) but it may seem unrealistic for natural populations. Note however that this assumption is actually made in all current methods for estimating the diversities and evaluating the variance of the estimated parameters, as also in resampling methods such as jackknife or bootstrap and in permutation tests. When all populations of a species are analyzed (see, for instance, the example discussed in CHAKRABORTY and DANKER-HOPFE 1991), a model with fixed populations is more appropriate than our model with sampled populations. In that case, the population sampling variations disappear from the variances but the assumption of independent populations is still used.

Our study of N_{ST} is closely related to the work by LYNCH and CREASE (1990). However, they only considered the nucleotide diversity whereas our results hold for any distance matrix as in EXCOFFIER *et al.* (1992) in the analysis of variance context. Moreover, some of their formulations were in terms of the data and not in terms of the underlying parameters. Here we provide a detailed variance partition for the diversity indices and the differentiation, then each term is consistently estimated. If the distance matrix is a parameter to be estimated, this also requires estimation of the covariance between its estimated elements using a model, as in LYNCH and CREASE (1990). Our tests for detecting genetic subdivision are only valid for a large population number. Other tests have been proposed in the literature and HUDSON *et al.* (1992) compared several of them through simulations, especially permutation tests that are quite relevant when the population number is small ($n = 2$ in their study). Their permutation tests are based on the average diversity and they are equivalent to permutation tests based on the differentiation. Our general average diversity v_S and differentiation N_{ST} can be used in the same way to perform permutation tests for small n.

Due to the absence of any genetic model, the analysis presented here may be of interest in other fields: for instance, in ecology, one may wish to compare species diversity of different ecosystems. Information about the taxonomic similarities between species may be introduced to improve the description and an analogue of the N_{ST} parameter described here may be estimated from such data.

The computations described here were made with a program written in Turbo-Pascal (to work with a PC or a Sun station). This program is available upon request to R.J.P. (please send a formatted IBM-compatible floppy disk).

LITERATURE CITED

COOPER, S. J. B., K. M. IBRAHIM and G. M. HEWITT, 1995 Postglacial expansion and genome subdivision in the European grasshopper *Chorthippus parallelus*. Mol. Ecol. **4:** 49–60.

EXCOFFIER, L., P. E. SMOUSE and J. M. QUATTRO, 1992 Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics **131:** 479–491.

HUDSON, R. R., D. D. BOOS and N. L. KAPLAN, 1992 A statistical test for detecting geographic subdivision. Mol. Biol. Evol. **9:** 138–151.

LYNCH, M., and T. J. CREASE, 1990 The analysis of population survey data on DNA sequence variation. Mol. Biol. Evol. **7:** 377–394.

NEI, M., 1977 *F*-statistics and analysis of gene diversity in subdivided populations. Ann. Hum. Genet. Lond. **41:** 225–233.

NEI, M., 1982 Evolution of human races at the gene level, pp. 167–181 in *Human Genetics, Part A: The Unfolding Genome,* edited by B. BONNE-TAMIR, T. COHEN and R. M. GOODMAN. Alan R. Liss, New York.

NEI, M., 1987 *Molecular Evolutionary Genetics.* Columbia University Press, New York.

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.

PETIT, R. J., N. BAHRMAN and PH. BARADAT, 1995 Comparison of genetic differentiation in maritime pine (*Pinus pinaster* Ait.) estimated using isozyme, total protein and terpenic loci. Heredity **75:** 382–389.

PONS, O., and R. J. PETIT, 1995 Estimation, variance and optimal sampling of gene diversity. I: haploid locus. Theor. Appl. Genet. **90:** 462–470.

SLATKIN, M., 1993 Isolation by distance in equilibrium and non-equilibrium populations. Evolution **47:** 264–279.

TAKAHATA, N., and M. NEI, 1984 F_{ST} and G_{ST} statistics in the finite island model. Genetics **107:** 501–504.

TEMPLETON, A. R., ROUTMAN, E. and PHILLIPS, C. A., 1995 Separating population structure from population history: a cladistic analysis of the geographical distribution of mitochondrial DNA haplotypes in the tiger salamander, *Ambystoma tigrinum.* Genetics **140:** 767–782.

Communicating editor: M. SLATKIN