Theor Appl Genet (1995) 90:462–470

O. Pons · R. J. Petit

Estimation, variance and optimal sampling of gene diversity
I. Haploid locus

Received: 25 July 1994 / Accepted: 30 September 1994

Abstract An extension of Nei's analysis of diversity in a subdivided population is proposed for a haploid locus. The differentiation G_{ST} becomes a natural extension of Wright's F_{ST} and generalizes Weir and Cockerham's parameter of co-ancestry by relaxing the assumption of identical correlation for all the alleles. Inter- and intra-population variances of the estimated diversities and differentiation are derived. Finally, the optimal sampling strategy for measuring G_{ST} when a fixed number of individuals can be analysed is considered. It is shown that, at a given locus, there is a unique sample size per population which yields the smallest variance of G_{ST}, regardless of the number of populations studied. These theoretical developments are illustrated with an analysis of chloroplast DNA diversity in a forest tree. The results emphasize the necessity of sampling many populations, rather than many individuals per population, for an accurate measurement of the subdivision of gene diversity at a single locus.

Key words Diversity · Differentiation · Variance Optimal design

Introduction

The measurement of genetic diversity in population surveys is part of most population genetic studies. Contrary to the situation in ecology, where a variety of indices are used to measure diversity, the studies of genetic diversity usually consider only a few indices, prominent among which is the gene diversity h of Nei (1973). This is simply defined as the probability that two sampled genes are different and is equally well adapted to haploid, diploid or polyploid genes. In a subdivided population, two parameters of diversity are defined, the average within-population diversity h_S and the total diversity h_T. The difference $h_T - h_S$ is a measure of the extent of the differentiation among the populations. Moreover, the ratio of this difference to the total diversity h_T, defined as G_{ST} by Nei (1973), measures the apportionment of diversity among the populations and is diversity independent when the number of populations studied is large. This property makes this index very useful when comparisons are needed among different organisms or among loci.

A method for estimating the parameters h_S, h_T and G_{ST} has been proposed by Nei and Chesser (1983) on the basis of a multinomial distribution of the alleles within the populations. However, the estimates only consider the sampling of individuals in fixed populations and not the sampling of populations. This restricts their usefulness since direct comparisons among species are not possible (Cockerham and Weir 1986).

Here, we extend Nei's approach to the total population, by considering that the sampled populations constitute a first level of sampling and the individuals within populations a second level of sampling. This point of view leads to new definitions of the total and average diversities and a new definition of the differentiation parameter G_{ST} follows which generalizes Wright's parameter F_{ST} (1943, 1951). Unbiased estimates are proposed in this setting, under assumptions similar to those of Nei. They are compared to other estimates in the literature. The two-stage sampling has already been considered by Weir and Cockerham using different methods and assumptions. These are discussed and contrasted with our estimates. Note that we consider here a haploid locus; the study of fixation indices (the so-called F-statistics) will be considered in a separate paper.

In order to study the accuracy of the estimates, we derive their analytical variances and we estimate them.

Communicated by P. M. A. Tigerstedt

O. Pons (✉)
Institut National de la Recherche Agronomique, Laboratoire de Biométrie, 78352 Jouy-en-Josas cedex, France

R. J. Petit
Institut National de la Recherche Agronomique, Laboratoire de Génétique et Amélioration des Arbres Forestiers, B.P. 45, 33611 Gazinet cedex, France

These results are expected to help investigators in the design of sampling schemes. More specifically, we consider the optimal strategy for the repartition of sampling effort within and among populations when the total number of individuals to be analysed is fixed and when an accurate measurement of the differentiation is of central interest. All these topics are illustrated using a study of chloroplast DNA diversity in European oak species (Petit et al. 1993).

Definition and estimation of diversity and differentiation indices

We consider a total population subdivided into a large number of populations in which I alleles are segregating. Let p_i be the frequency of the i-th allele in the general population and p_{ki} be the frequency of the i-th allele in the k-th population, conditionally on this population. The p_{ki}s are considered as random frequencies with expectation p_i and variance V_i. In addition, C_{ij} denotes the covariance between p_{ki} and p_{kj}, for $i \neq j$. Following Nei (1973), we define the diversity of the k-th population as

$$h_k = 1 - \sum_i p_{ki}^2. \tag{1}$$

We now define the average within-population diversity as the expectation Eh_k of h_k in the general population. From the definition of V_i as $Ep_{ki}^2 - p_i^2$, we get

$$h_S = 1 - \sum_i (p_i^2 + V_i) \tag{2}$$

and the total diversity is defined as

$$h_T = 1 - \sum_i p_i^2. \tag{3}$$

The parameters h_S and h_T do not depend on the number n of sampled populations as in the case of Nei who considered empirical means and defined h_S as $n^{-1} \sum_{k \leq n} h_k$ and h_T as $1 - \sum_i p_{\cdot i}^2$ where $p_{\cdot i} = n^{-1} \sum_{k \leq n} p_{ki}$ is the average frequency of the i-th allele for the n observed populations.

Nei's differentiation parameter G_{ST} is $(h_T - h_S)/h_T$ where $h_T - h_S = \sum_i n^{-1} \sum_k (p_{ki} - p_{\cdot i})^2$ would be the sums of empirical variances of the p_{ki}s if they were directly observed. With (2) and (3), it now becomes

$$G_{ST} = \frac{\sum_i V_i}{h_T} \tag{4}$$

where the empirical variances are replaced by the actual ones. It no longer depends on the number of observed populations and may be viewed as **a natural extension of Wright's parameter** F_{ST} (1943, 1951). This parameter satisfies $F_{ST} = \sigma_{\bar{p}}^2 / \bar{p}\bar{q}$ for two alleles with frequencies having means $\bar{p}$ and $\bar{q} = 1 - \bar{p}$ and a variance $\sigma_{\bar{p}}^2$ among the populations. If the Hardy-Weinberg equilibrium holds, $F_{ST} = (\sigma_{\bar{p}}^2 + \sigma_{\bar{q}}^2)(1 - \bar{p}^2 - \bar{q}^2)$ which generalizes as in (4).

A stratified random sampling is used to estimate these indices: n populations are drawn without replacement in the general population and a sample of n_k independent individuals is drawn from the k-th population. We assume that $n_k \geq 2$ for each k in order to observe some variability within the populations. Let n_{ki} be the number of individuals in the k-th population having the allele i and $x_{ki} = n_{ki}/n_k$ be the empirical frequency of the allele i in the population k, $i \leq I$, $k \leq n$. Conditionally on the populations sampled, the set $(n_{ki})_i$ follows a multinomial distribution with parameters n_k and the random frequencies $(p_{ki})_i$ (cf. Nei and Roychoudhury 1973). Let E be the general expectation, E_k be the expectation conditionally on the k-th population, and E^{pop} be the expectation conditionally on the n sampled populations. We have $Ex_{ki} = EE_k x_{ki} = Ep_{ki} = p_i$ thus $x_{\cdot i} = n^{-1} \sum_{k \leq n} x_{ki}$ is an unbiased estimate of p_i and $E^{pop} x_{\cdot i} = n^{-1} \sum_{k \leq n} p_{ki} = p_{\cdot i}$. Moreover, $E_k \sum_i x_{ki}^2 = (n_k - 1) n_k^{-1} \sum_i p_{ki}^2 + n_k^{-1}$.

Considering the k-th population as fixed, its diversity h_k (1) is then unbiasedly estimated by

$$\hat{h}_k = \frac{n_k}{n_k - 1} \left(1 - \sum_i x_{ki}^2 \right) \tag{5}$$

and an unbiased estimate of $h_S = Eh_k$ (2) may be obtained as an estimate of the empirical mean $\bar{h}_S = n^{-1} \sum_{k \leq n} h_k$ of the h_ks, namely

$$\hat{h}_S = \frac{1}{n} \sum_{k \leq n} \frac{n_k}{n_k - 1} \left(1 - \sum_i x_{ki}^2 \right). \tag{6}$$

It satisfies $E^{pop} h_S = \bar{h}_S$ and is therefore an unbiased estimate of Nei's within-population diversity in the setting of n fixed populations and a simple sampling procedure. The estimate

$$\hat{h}_S = n^{-1} \sum_k \tilde{h}_k \tag{7}$$

where $\tilde{h}_k = 1 - \sum_i x_{ki}^2$ could have been preferred since it is based on the efficient estimates x_{ki} of the p_{ki}s. However, it has the bias $-\tilde{n}^{-1} h_S$ where $\tilde{n} = n(\sum_k n_k^{-1})^{-1}$ is the harmonic mean of the n_ks and, after studying their variance in the two-stage sampling (cf. next section), we preferred to choose (6). Our estimate may be compared to that of Nei and Chesser (1983)

$$\hat{h}_{2S} = \frac{\tilde{n}}{\tilde{n} - 1} \left(1 - \frac{1}{n} \sum_{ki} x_{ki}^2 \right) \tag{8}$$

which is also an unbiased estimate of h_S in our framework. Equation (6) is just a different way of handling unequal sample sizes.

For the estimation of h_T, we consider

$$\hat{h}_T = 1 - \frac{1}{n(n-1)} \sum_{k \neq l} \sum_i x_{ki} x_{li}$$

$$= 1 - \sum_i x_{\cdot i}^2 + \frac{1}{n(n-1)} \sum_{ki} (x_{ki} - x_{\cdot i})^2. \tag{9}$$

If we assume that the populations are independent, $h_T = 1 - \sum_i E p_{ki} p_{li}$ for any $k \neq l$ and $\hat{h}_T$ is an unbiased estimate of h_T. Note that this assumption is both a genetical and a population sampling condition. It is currently used by most authors, at least implicitly (Cockerham and Weir 1986; Nei 1987), and it is a valuable approximation for weakly dependent populations.

Conditionally on the sampled populations, the mean of $\hat{h}_T$ (9) is $\bar{h}_T = 1 - \sum_i \bar{p}_{.i}^2 + n^{-1}(n-1)^{-1} \sum_{ki}(p_{ki} - \bar{p}_{.i})^2$. The expression (9) differs from the estimate of h_T in Nei (Nei and Chesser 1983; Nei 1987) where only $1 - \sum_i \bar{p}_{.i}^2$ is estimated by $1 - \sum_i x_{.i}^2 + (n\tilde{n})^{-1} \hat{h}_{2S}$, when each n_k is approximated by $\tilde{n}$, whereas (9) estimates $\bar{h}_T$.

An estimate of the differentiation parameter G_{ST} is deduced as

$$\hat{G}_{ST} = 1 - \hat{h}_S / \hat{h}_T. \tag{10}$$

This is a biased estimate because of the dependence between $\hat{h}_S$ and $\hat{h}_T$ and $\hat{h}_T^{-1}$ itself is not unbiased for h_T^{-1}. However the bias of $\hat{G}_{ST}$ is asymptotically negligible as we prove in the following under specified conditions.

Nei (1986) proposed a modified version of the differentiation parameter, F'_{ST} which satisfies $F'_{ST} = \sum_i (n-1)^{-1} \sum_k (p_{ki} - p_{.i})^2 / \bar{h}_T$ where $E \sum_i (n-1)^{-1} \sum_k (p_{ki} - p_{.i})^2 = \sum_i V_i$ and $E \bar{h}_T = h_T$ (3). If the p_{ki}s where directly observed, this new parameter F'_{ST} would then be an estimate of our differentiation parameter G_{ST} (4). Moreover, it is estimated by $\hat{F}'_{ST} = 1 - \hat{h}_{2S} / \hat{h}_T$ (Nei 1986) and is therefore equal to (10) if the populations have the same size. By analogy to Nei's indices, Lynch and Crease (1990) extended the F_{ST} to DNA sequences. It appears that their N_{ST} reduces to our $\hat{G}_{ST}$ when applied to gene diversity.

For populations of the same size, $\hat{G}_{ST}$ is also identical to the estimate of Weir and Cockerham's parameter θ (1984). In their approach, Wright's parameter F_{ST} is generalized into θ, the correlation of alleles present in different individuals in the same population, and the estimation relies on an analysis of variance of indicator functions. In our setting, the individuals within a fixed population are also considered as dependent in the general population, since they belong to the same random population. Weir and Cockerham assume an identical level of dependence for any allele as a property of the locus. This approach may also be viewed as a model $V_i = \theta p_i (1 - p_i)$ for the general variance of the populations mean frequencies. In this model, $\sum_i V_i = \theta h_T$ and our G_{ST} (4) is actually equivalent to θ. Thus our definition also generalizes Weir and Cockerham's point of view for a haploid locus. Because of the conditional independence of the inidividuals within a population, we follow Nei's multinomial distribution approach but with random populations. This multinomial model is the basis for the analyses of the statistical sampling effects, it allows for a study of the accuracy of the estimates which cannot be done in the analysis of variance framework as defined by Weir and Cockerham.

Variance of $\hat{h}_S$ and $\hat{h}_T$

The estimate $\hat{h}_S$ may be written in the form $n^{-1} \sum_k (\hat{h}_k - h_k) + n^{-1} \sum_k (h_k - h_S) + h_S$, where the h_ks (1) are independent and have the same distribution. If $E p_{ki}^4$ is finite, the variance of $\hat{h}_S$ follows as a sum of within-population and between-population variances

$$Var(\hat{h}_S) = n^{-2} \sum_k E Var_k(\hat{h}_k) + n^{-1} Var(h_k)$$

where Var_k denotes the variance conditionally on the k-th population and $Var_k(\hat{h}_k)$ is determined from the moments of the multinomial distribution, which yields

$$Var_k(\hat{h}_k) = \frac{2}{n_k(n_k-1)} \left\{ (3 - 2n_k) \left(\sum_i p_{ki}^2 \right)^2 + 2(n_k - 2) \left(\sum_i p_{ki}^3 \right) + \left(\sum_i p_{ki}^2 \right) \right\}.$$

If the populations have the same size, the different terms $Var_k(\hat{h}_k)$ have the same distribution and the first term of the development of $Var(\hat{h}_S)$ reduces to $n^{-1} E Var_k(\hat{h}_k)$. Furthermore, $Var(h_k) = E(h_k - h_S)^2 = E(\sum_i (p_{ki}^2 - p_i^2) - \sum_i V_i)^2$ may be written as $E(\sum_i p_{ki}^2)^2 - (1 - h_S)^2$.

The within-population variance, $Var_{intra}(\hat{h}_S) = n^{-2} \sum_k E Var_k(\hat{h}_k)$, depends on the populations sizes and on their number n, whereas the between-population variance, $Var_{inter}(\hat{h}_S) = n^{-1} Var(h_k)$, depends only on n. As Nei and Roychoudhury (1973) mentioned, n has to be large for reducing $Var_{inter}(\hat{h}_S)$. On the contrary, $Var_{intra}(\hat{h}_S)$ is small when the subsample sizes n_k are large or when n is large, these quantities being analogous weights in the expression of $Var_{intra}(\hat{h}_S)$ since $Var_k(\hat{h}_k) \simeq 4 n_k^{-1} \{ \sum_i p_{ki}^3 - (\sum_i p_{ki}^2)^2 \}$ for large n_ks and $Var_{intra}(\hat{h}_S)$ is of the order n^{-1} when n is large. That is why from now on we consider only a large number of populations with possibly small subsample sizes. None of the two variance terms is then predominant in the variance of $\hat{h}_S$,

$$Var(\hat{h}_S) = \frac{1}{n^2} \sum_k \frac{2}{n_k(n_k-1)} \left\{ (3 - 2n_k) E \left(\sum_i p_{ki}^2 \right)^2 + 2(n_k - 2) E \left(\sum_i p_{ki}^3 \right) + E \left(\sum_i p_{ki}^2 \right) \right\} + \frac{1}{n} \left\{ E \left(\sum_i p_{ki}^2 \right)^2 - (1 - h_S)^2 \right\} \tag{11}$$

which is of the order n^{-1}. Since the total number of individuals to be analysed is limited by practical constraints, the subsample sizes n_k may then be rather small if n is large and the bias of $\tilde{h}_S$ remains non-negligible ($-\tilde{n}^{-1} \tilde{h}_S$). The variances of $\hat{h}_S$ and $\tilde{h}_S$ being of the same order, we then opted for $\hat{h}_S$.

As in Nei and Roychoudhury (1973), the sampling variance $Var(\hat{h}_S)$ is estimated by the classical formula

$$Var(\hat{h}_S) = \frac{1}{n(n-1)} \sum_k (\hat{h}_k - \hat{h}_S)^2. \tag{12}$$

An estimation of the terms $Var(\hat{h}_k)$ calculated by replacing p_{ki} with x_{ki} provides a simple but biased estimate of $Var_{intra}(\hat{h}_S)$. In order to avoid a negative estimated value of this variance, an unbiased estimate is also defined using the estimates $(n_k - 1)^{-1} (n_k \sum_i x_{ki}^2 - 1)$ for $\sum_i p_{ki}^2$, $(n_k - 1)^{-1} (n_k - 2)^{-1} (n_k^2 \sum_i x_{ki}^3 - 3 n_k \sum_i x_{ki}^2 + 2)$ for $\sum_i p_{ki}^3$, and $(n_k - 1)^{-1} (n_k - 2)^{-1} (n_k - 3)^{-1} [n_k^3 (\sum_i x_{ki}^2)^2 - 4 n_k^2 \sum_i x_{ki}^3 - 2 n_k (n_k - 5) \sum_i x_{ki}^2 + (n_k - 6)]$ for $(\sum_i p_{ki}^2)^2$ if the n_ks are

all greater than 3. An estimate of $Var_{inter}(\hat{h}_S)$ follows by a difference.

We assumed that n is large, therefore $\hat{h}_S$ converges in probability to h_S as n tends to infinity, because $Var(\hat{h}_S)$ tends to zero. The distribution of $\hat{h}_S$ may be approximated by its limit: if the subsample sizes are bounded and if $n^{-1}\sum_k EVar_k(\hat{h}_k)$ has a limit as n tends to infinity, $\sqrt{n}(\hat{h}_S - h_S)$ converges in distribution to a Gaussian variable with zero mean and a variance σ_S^2 which is the limit of $nVar(\hat{h}_S)$ and which is unbiasedly estimated by $\hat{\sigma}_S^2 = n\hat{V}ar(\hat{h}_S)$.

The variance of $\hat{h}_T$ is given by

$$Var(\hat{h}_T) = \frac{1}{n^2(n-1)^2}\sum_{k\neq l}\sum_{u\neq v}\sum_{ij}\{Ex_{ki}x_{li}x_{uj}x_{vj} - p_i^2 p_j^2\}$$

and this sum develops according to the number of distinct indices for populations (among k, l, u, v). Since n is assumed to be large, the terms of order n^{-2} may be neglected and an approximation of $\hat{h}_T$ is obtained by deleting the terms of this sum having not at least three distinct summation indices. With the notation $C_{ii} = V_i$, we get

$$Var(\hat{h}_T) = \frac{4}{n^2}\sum_k\sum_{ij} p_i p_j\{Ex_{ki}x_{kj} - p_i p_j\} + 0(n^{-2})$$

$$= \frac{4}{n}\sum_{ij} p_i p_j C_{ij} + \frac{4}{n\tilde{n}}\left\{\sum_i p_i^3 - \sum_{ij} p_i p_j(p_i p_j + C_{ij})\right\} + 0(n^{-2}). \tag{13}$$

This splits into a within-population variance

$$Var_{intra}(\hat{h}_T) = \frac{1}{n^2(n-1)^2}\sum_{k\neq l}\sum_{u\neq v}\sum_{ij} E(x_{ki}x_{li}x_{uj}x_{vj} - p_{ki}p_{li}p_{uj}p_{vj})$$

$$= \frac{4}{n\tilde{n}}\left\{\sum_i p_i^3 - \sum_{ij} p_i p_j(p_i p_j + C_{ij})\right\} + 0(n^{-2}). \tag{14}$$

and a between-population variance

$$Var_{inter}(\hat{h}_T) = \frac{1}{n^2(n-1)^2}\sum_{k\neq l}\sum_{u\neq v}\sum_{ij} E(p_{ki}p_{li}p_{uj}p_{vj} - p_i^2 p_j^2)$$

$$= \frac{4}{n}\sum_{ij} p_i p_j C_{ij} + 0(n^{-2}). \tag{15}$$

Estimates of these variances are obtained by replacing C_{ij} by its unbiased estimate

$$\hat{C}_{ij} = \frac{1}{n-1}\sum_k (x_{ki} - x_{.i})(x_{kj} - x_{.j}) + \frac{1}{n}\sum_k \frac{x_{ki}x_{kj}}{n_k - 1}, \text{ if } i \neq j,$$

and $C_{ii} = V_i$ by

$$\hat{V}_i = \frac{1}{n-1}\sum_k (x_{ki} - x_{.i})^2 - \frac{1}{n}\sum_k \frac{x_{ki}(1 - x_{ki})}{n_k - 1}.$$

This provides the estimates

$$\hat{V}ar(\hat{h}_T) = \frac{4}{n(n-1)}\sum_{ij} x_{.i}x_{.j}\sum_k (x_{ki} - x_{.i})(x_{kj} - x_{.j}), \tag{16}$$

$$\hat{V}ar_{inter}(\hat{h}_T) = \frac{4}{n}\sum_{ij} x_{.i}x_{.j}\hat{C}_{ij} \tag{17}$$

and by difference,

$$\hat{V}ar_{intra}(\hat{h}_T) = \frac{4}{n^2}\left\{\sum_i x_{.i}^2\sum_k \frac{x_{ki}}{n_k - 1} - \sum_{ij} x_{.i}x_{.j}\sum_k \frac{x_{ki}x_{kj}}{n_k - 1}\right\}. \tag{18}$$

The covariance of $\hat{h}_S$ and $\hat{h}_T$ is approximated in a similar way. The between-population covariance is also of the order n^{-1} and the within-population covariance of the order $(n\tilde{n})^{-1}$, their expressions and estimates are detailed in Appendix 1.

Since $1/\tilde{n} \leq 1/2$ if $n_k \geq 2$ for each k, $\hat{V}ar(\hat{h}_T)$ tends to zero as $n \to \infty$ and $\hat{h}_T$ converges in probability to h_T (cf. Appendix 1). Furthermore $n^{1/2}(\hat{h}_S - h_S, \hat{h}_T - h_T)$ converges to a two-dimensional Gaussian variable, its components having respectively variances $\sigma_S^2 = \lim_n nVar(\hat{h}_S)$ and $\sigma_T^2 = \lim_n nVar(\hat{h}_T)$ and a covariance $\sigma_{ST}^2 = \lim_n nCov(\hat{h}_S, \hat{h}_T)$.

Approximation of the variance of $\hat{G}_{ST}$ and optimal sampling design

From the convergence results concerning $(\hat{h}_S - h_S, \hat{h}_T - h_T)$, the estimate $\hat{G}_{ST}$ (10) of the differentiation parameter converges to G_{ST}. Under the previous conditions, $n^{1/2}(G_{ST} - \hat{G}_{ST})$ has the same limit distribution as $n^{1/2}(\hat{h}_S - h_S)(h_T)^{-1} - n^{1/2}(\hat{h}_T - h_T)h_S(h_T)^{-2}$ which converges to a Gaussian variable having the mean zero and the variance $(h_T)^{-2}\sigma_S^2 - 2h_S(h_T)^{-3}\sigma_{ST}^2 + h_S^2(h_T)^{-4}\sigma_T^2$. If n is large, the variance of $\hat{G}_{ST}$ is then approximated by

$$(h_T)^{-2}Var(\hat{h}_S) - 2h_S(h_T)^{-3}Cov(\hat{h}_S, \hat{h}_T) + h_S^2(h_T)^{-4}Var(\hat{h}_T), \tag{19}$$

and an estimate of $Var(\hat{G}_{ST})$ follows as

$$\hat{V}ar(\hat{G}_{ST}) = \frac{1}{n(n-1)\hat{h}_T^2}\sum_k (\hat{h}_k - \hat{h}_S)^2 - \frac{4\hat{h}_S}{(n-2)\hat{h}_T^3}$$
$$\left\{\hat{h}_S(1 - \hat{h}_T) - \frac{1}{n-1}\sum_{ki}\left(x_{.i} - \frac{x_{ki}}{n}\right)\hat{h}_k x_{ki}\right\}$$
$$+ \frac{4\hat{h}_S^2}{n(n-1)\hat{h}_T^4}\sum_{kij} x_{.i}x_{.j}(x_{ki} - x_{.i})(x_{kj} - x_{.j})$$

from (12), (16) and Appendix 1.

The variance of $\hat{G}_{ST}$ may also be split into within- and between-population variances which satisfy a formula similar to (19) with the corresponding within and between terms. An approximation of the within-population variance of $\hat{G}_{ST}$ is for instance

$$(h_T)^{-2}Var_{intra}(\hat{h}_S) - 2h_S(h_T)^{-3}Cov_{intra}(\hat{h}_S, \hat{h}_T)$$
$$+ h_S^2(h_T)^{-4}Var_{intra}(\hat{h}_T),$$

an estimate of $Var_{intra}(\hat{G}_{ST})$ is deduced by replacing the different terms in this expression by their estimates and the same holds for the between-population terms.

We proved that $Var_{inter}(\hat{h}_S)$, $Var_{inter}(\hat{h}_T)$, $Cov_{inter}(\hat{h}_S, \hat{h}_T)$, and therefore $Var_{inter}(\hat{G}_{ST})$, do not depend on the n_ks whereas $Var_{intra}(\hat{h}_S)$, $Var_{intra}(\hat{h}_T)$, $Cov_{intra}(h_S, \hat{h}_T)$, and hence $Var_{intra}(\hat{G}_{ST})$, do. This decomposition of (19) into terms depending, or not depending, on the n_ks may be used to find an optimal value of n which minimizes $\hat{V}ar(\hat{G}_{ST})$ in a sampling design with populations of equal sizes, $\tilde{n}$, and when the total number $n\tilde{n}$ has a fixed value α. In that case, using (11), (13), (19) and Appendix 1, $Var(\hat{G}_{ST})$ is approximated in the form

$$f(n,\tilde{n}) = \frac{A}{n} + \frac{B}{n\tilde{n}} + \frac{1}{n\tilde{n}(\tilde{n}-1)}\{C(3-2\tilde{n}) + D(\tilde{n}-2) + E\},$$

where A, B, C, D and E are approximately constants under the conditions of the asymptotic normality of $\hat{G}_{ST}$, and are defined in Appendix 2. With $n\tilde{n} = \alpha$, $f(n,\tilde{n})$ becomes

$$f(n) \doteq \frac{A}{n} + \frac{B}{\alpha} + \frac{(3n-2\alpha)C + (\alpha-2n)D + E}{\alpha(\alpha-n)} \qquad (20)$$

and it is minimum at n_{opt} such as $f'(n_{opt}) = 0$, i.e., n_{opt} is a solution of $n^2(C-D+E-A) + 2nA\alpha - A\alpha^2 = 0$. If the discriminant $\Delta = A\alpha^2(C-D+E)$ of this equation is positive, the optimal values n_{opt} and $\tilde{n}_{opt}$ of n and $\tilde{n}$ follow,

$$n_{opt} = \alpha\frac{A - \sqrt{A(C-D+E)}}{A-C+D-E}, \qquad (21)$$

$$\tilde{n}_{opt} = \frac{A-C+D-E}{A-\sqrt{A(C-D+E)}}. \qquad (22)$$

It is noticeable that $\tilde{n}_{opt}$ does not depend on the total number α of individuals in the study, it is a constant which depends only on the distribution of the variables p_{ki}. When $\Delta < 0$, $f(n)$ is strictly decreasing to zero and $Var(\hat{G}_{ST})$ is minimum when n is maximum, i.e., $n_{opt} = \alpha/2$ and $\tilde{n}_{opt} = 2$ since there must be at least two individuals in each population to estimate h_S.

An estimate of $\tilde{n}_{opt}$ is obtained from (22), replacing the constants by estimates based on a preliminary sample with populations of possibly varying sizes n_k. They are given more precisely in Appendix 2.

Numerical example

The data set that was used originates from a survey of the chloroplast DNA diversity of oaks in Europe (Petit et al. 1993). A total of 90 populations sampled over the European range are included here. Four cytotypes (Table 1) were detected in the survey. The analyses were made both at the level of the chloroplast locus and at the level of the individual cytotypes (by considering each of the four cytotypes along with the pooled three remaining cytotypes are equivalent to four diallelic loci). The arithmetic and harmonic mean numbers of genotypes per population were respectively 8.02 and 6.73, for a total of 722 individuals analysed. Cytotype 4 was found only in two populations where it was fixed.

Table 1 Comparison of several methods of analysis of diversity of cpDNA in *Quercus sp.* The method used is given in brackets: (*bias*), uncorrected definitions of the parameter; (*Nei*), estimates of Nei and Chesser (1983); (*Neib*), alternative estimates of Nei (1986); (*WC*), estimates of Weir and Cockerham (1984); (*PP*), present study. *SD*, standard deviates of the estimates (*PP*)

Haplotypes	1	2	3	4	Total
Frequency	**0.2288**	**0.3906**	**0.3584**	**0.0222**	**1.0000**
$h_S(bias)$	0.0328	0.0461	0.0416	0.0000	0.0707
$h_S(Nei)$	0.0385	0.0541	0.0488	0.0000	0.0707
$h_S(WC)$	0.0389	0.0513	0.0525	0.0000	0.0714
$\boldsymbol{h_S}$**(PP)**	**0.0379**	**0.0540**	**0.0481**	**0.0000**	**0.0700**
SD	0.0132	0.0162	0.0166	0.0000	0.0184
$h_T(bias)$	0.3529	0.4761	0.4599	0.0435	0.6661
$h_T(Nei)$	0.3530	0.4762	0.4600	0.0435	0.6663
$h_T(WC)$	0.3372	0.4629	0.4880	0.0465	0.6673
$\boldsymbol{h_T(PP)}$	**0.3565**	**0.4809**	**0.4646**	**0.0440**	**0.6730**
SD	0.0460	0.0215	0.0275	0.0299	0.0161
$G_{ST}(bias)$	0.9071	0.9033	0.9096	1.0000	0.9096
$G_{ST}(Nei)$	0.8909	0.8864	0.8938	1.0000	0.8939
$G_{ST}(Neib)$	0.8920	0.8875	0.8949	1.0000	0.8949
$\theta(WC)$	0.8845	0.8892	0.8924	1.0000	0.8949
$\boldsymbol{G_{ST}(PP)}$	**0.8936**	**0.8876**	**0.8964**	**1.0000**	**0.8930**
SD	0.0381	0.0337	0.0333	–	0.0276

Estimates

First, our estimates were compared with the simple but biased estimates $\tilde{h}_S$ (7), $\tilde{h}_T$ given by (3) where $x_{.i}$ replaces p_i and the corresponding $\tilde{G}_{ST} = 1 - \tilde{h}_s/\tilde{h}_T$ (Table 1). Note that $\tilde{h}_S$ and $\tilde{G}_{ST}$ poorly estimate h_S and G_{ST} as judged from their differences with the unbiased estimates.

Second, our estimates were compared with those of Nei and Chesser (1983) and Weir and Cockerham (1984). All of them are quite close. Our estimate of h_S is closer to Nei and Chesser's estimate $\hat{h}_{2S}$, whereas our estimate of h_T is closer to that of Weir and Cockerham. This last result is expected since our estimates are valid for a larger set of populations than those sampled, like with Weir and Cockerham's method. For all three parameters, our estimates should give exactly the same results as those of Weir and Cockerham if all population sizes were identical. Here, this is not the case and the differences observed (especially for h_S) probably derive from the way populations are weighted in the alternative estimations procedures. We assume that differences in sample sizes are independent of the true effective population sizes and all populations receive similar weights in the computations, as proposed by Nei (1987). In Weir and Cockerham's method, on the other hand, weights are proportional to the sample sizes. Hence, our estimates combine properties of both former methods. If a

lower number of populations had been sampled, our estimates would have been much more different from those of Nei and Chesser, since their estimates are not independent of the number of populations sampled.

Variances

We computed total, intra- and inter-population variances for the estimates of h_S, h_T and G_{ST}. The results are given in Table 2. The estimates of h_S and h_T have similar variances but the estimation of G_{ST} (which directly derives from the other two parameters) is less precise. Moreover, the variance due to sampling within populations accounted for a small fraction of the total variance for h_S and G_{ST} and especially h_T. This was a first indication that the sampling of populations, rather than that of individuals within populations, was limiting the precision of our G_{ST} estimates in this example.

Table 2 Estimation of total, intra- and inter-population variance components of gene diversity and differentiation

Item	Variances × 10^4		
	h_S	h_T	G_{ST}
Variance (total)	3.38	2.60	7.64
Variance (intra)	0.15	0.00	0.31
Variance (inter)	3.23	2.60	7.33

Optimal sampling design

The proposed statistical methods make it possible to determine *a posteriori* what would have been an 'optimal' sampling design i.e., a sampling design which would have yielded the smallest variance for $\hat{G}_{ST}$ with the same total sample size. It was indeed shown analytically that there is a unique sample size per population which is universally optimal (i.e., regardless of the total number of analyses made) in terms of minimizing the differentiation sample variance at a given locus. By estimating $\tilde{n}_{opt}$ (22), we find this optimal sample size per population to be 2.5. More precisely, if we had analysed 250 populations instead of 90, with 2–3 individuals per population instead of the 6.7 actually studied (i.e., with the same total sample size of about 600 individuals), we would expect a variance of 4.3×10^{-4} instead of the variance of 7.6×10^{-4} that we found. In Fig. 1, the relation between the expected variance of $\hat{G}_{ST}$ and the number of sampled individuals per population, $\tilde{n}$, is illustrated in various situations where the total sample size $\sum_k n_k$ takes different values ranging from 100 to 5 000. It is apparent that the minimum variance is always obtained for $\tilde{n} = 2.5$, regardless of the total sampling effort. Moreover, it is also clear that, although the use of a slightly suboptimal sampling design may be tolerated for large samples, very significant losses in precision can be expected when the sample size uses either too many or too few individuals

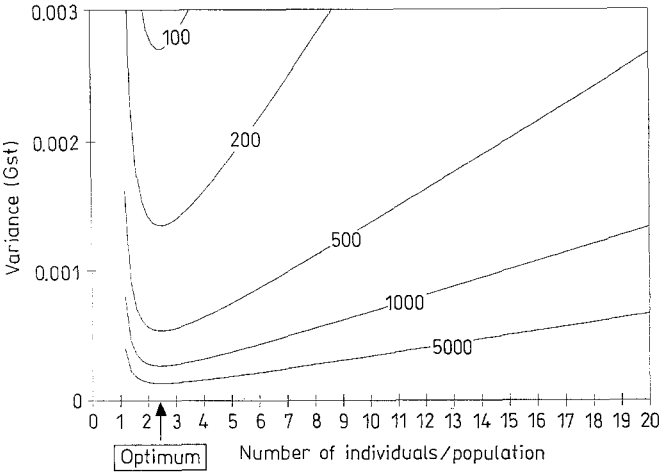
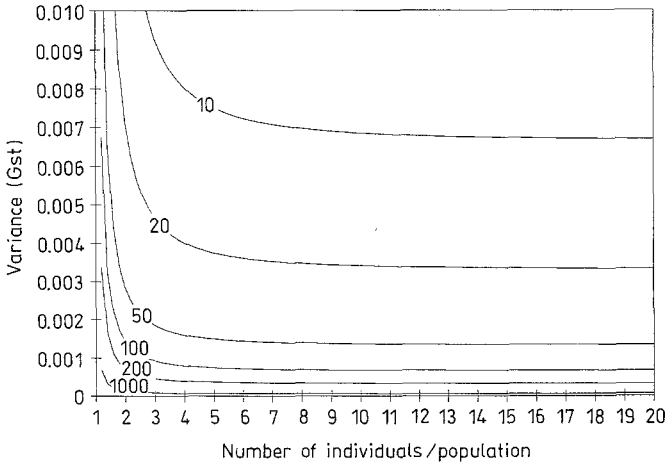


Fig. 1 Evolution of $Var(G_{ST})$ as a function of the sample size per population, for fixed total sample sizes. The optimum (2.5 individuals per population) corresponds to the smallest variance of $\hat{G}_{ST}$ for a fixed number of individuals analysed (ranging from 100 to 5 000)

per population, though especially so when the number of individuals per population, $\tilde{n}$, is smaller than the optimum. The very small value found for $\tilde{n}_{opt}$ is striking, and indicates that our initial sampling scheme was adequate since it emphasised the sampling of as many populations as was possible, at the expense of the number of individuals per population.

Since the choice of the sample size per population may be dictated by considerations other than maximizing the precision of $\hat{G}_{ST}$ (as when precise estimates of allele frequencies are required for each population, or when gametic disequilibria need to be studied, etc.), we provide in Figs. 2 and 3 the evolution of variance of $\hat{G}_{ST}$ for a fixed number of individuals per population or for a fixed number of populations. For the cpDNA locus studied here, this gives predictions for the precision of the

Fig. 2 Evolution of the variance of G_{ST} as a function of the sample size per population. The number of populations analysed are given for each curve. An increase in the number of individuals per population is soon inefficient in improving the precision of G_{ST} when the number of populations studied is limited



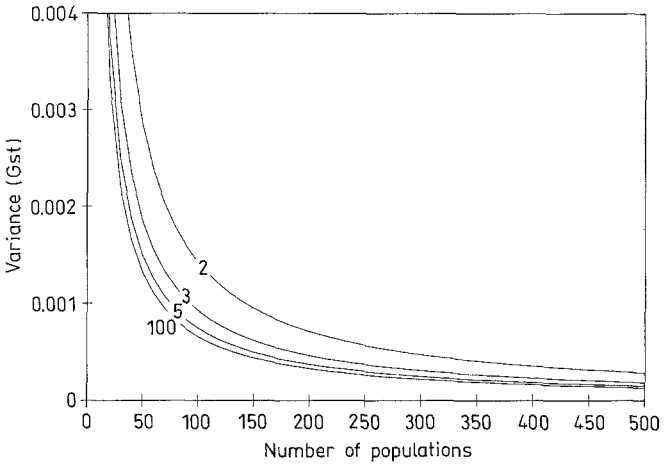


Fig. 3 Evolution of the variance of G_{ST} as a function of the number of populations analysed. The sample sizes per population are given for each curve. A continuous decrease in $var(G_{ST})$ is achieved by increasing the total number of populations sampled

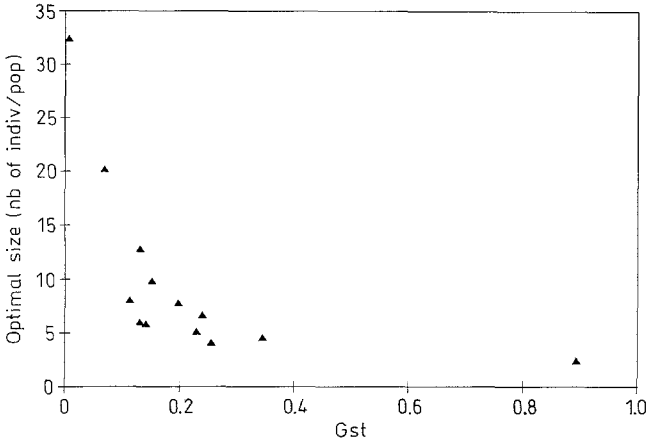


Fig. 4 Relationship between optimal sampling and level of differentiation. The optimal sample size per population is plotted against the level of differentiation for nuclear loci (values of G_{ST} ranging from 0 to 0.36) and for the chloroplast DNA locus (G_{ST} approximately 0.9). The higher the differentiation, the smaller the optimal sample size per population necessary to achieve a correct estimation of G_{ST}

measure of differentiation in a variety of sampling schemes. It is clear, from the comparison of both figures (note that very different scales are used), that increasing the number of populations is essential in order to reach reasonably precise estimates (Fig. 2), whereas, when the number of populations is limited, there is little point in increasing the sample size per population past values situated around five individuals per population, since the decay of $Var(\hat{G}_{ST})$ becomes very low after this limit (Fig. 3).

Because the example studied included a case of rather extreme differentiation, it is unclear from the data whether optimal sample size will always be as low as that found here. We therefore computed G_{ST} and optimal sample sizes for 12 additional loci to further examine the relationship between the level of differentiation and optimal sampling. The additional data (Petit and Bahrman, unpublished) were obtained in a study of six populations of maritime pine (*Pinus pinaster* Ait.) using isozymes and abundant proteins (see Bahrman and Damerval 1989, for a study of inheritance of these markers). Sample sizes per population range from 32 trees for protein loci to over120 for the isozyme data. The tissue analysed is the megagametophyte of the seeds, and the data obtained are therefore haploid for these nuclear genes. The values of differentiation obtained range from close to zero to over 0.3 depending on the locus. A relationship is apparent between optimal sample size and G_{ST} (Fig. 4). A higher within-population sampling is necessary for loci showing little differentiation. However, even in these cases, the optimal sample sizes remain relatively low by comparison to current sampling practices.

Discussion

A precise knowledge of the subdivision of genetic variation within and among populations is of particular importance for the management of genetical resources. Indeed, genetically heterogeneous species will require different strategies of conservation than comparatively uniform species. A popular parameter used for such purposes is F_{ST}, the genetic differentiation first defined by Wright (1951), and generalised to multiple alleles and any ploidy levels by Nei (1973). Two of its properties probably explain its popularity. First, it is diversity independent, being defined as a ratio of gene diversities. Hence, measurements of genetical differentiation obtained with different (neutral) genetic markers, such as protein polymorphisms or coding or non-coding DNA polymorphisms, are expected to give similar values for differentiation regardless of their level of variability. Second, if a set of assumptions are verified, a population genetic model predicts that this parameter F_{ST} (or G_{ST}) is related to the amount of gene flow among populations in an island model of population structure (Wright 1951). Therefore, a precise estimate of differentiation is necessary in order to obtain an estimate of gene flow.

Our results indicate that, for accurate measurements of genetic differentiation at level of the single locus or allele, more extensive surveys of genetic variation are necessary than what is usually encountered. Note, for instance, that the average number of populations analyzed in 655 studies of gene diversity in plants is only 12.3 (compiled by Hamrick et al. 1992).

Until now, a detailed study of the elementary components of gene differentiation at the single locus level has been lacking. Though tests of the null hypothesis of no differentiation have been proposed, either analytically (Long 1986) or using permutational procedures (Excoffer et al. 1992), more general methods were still needed for the study of the heterogeneity of gene differentiation. This led us to develop variance estimates for all parameters of gene diversity, with a particular emphasis on G_{ST}, at the level of the individual locus or haplotype.

We also considered the optimal sampling strategy for the study of gene differentiation. Attempts to define optimal sampling strategies have already been considered in the context of the collection of material for ex situ genetic conservation (Marshell and Brown 1975; Bogyo et al. 1980; Brown and Munday 1982; Yonezawa 1985). In these approaches, however, the goal is the collection of the maximal amount of 'genetically useful' variability, and not the study of the geographic distribution of diversity per se. The optimal strategy in then to collect a single individual from as many populations as is feasible. However, because the cost of collecting new populations is generally higher than the collection of additional individuals, the measure of the partitioning of variability within and among populations remains important in that context. If the estimation of the genetic differentiation is not the single goal, the definition of the optimal sampling strategy will have to be a compromise among conflicting sampling options. However, we reached the same main conclusion as did most previous published works, i.e., that an increase in the number of populations, rather than of individuals per population, is of primary importance in surveys of genetic variation (see for instance Marshall and Brown 1975; Yonezawa 1985; Lynch and Crease 1990, in the context of the study of DNA sequence variation). In some of these papers, considerations of optimal sampling strategies are derived from the examination of the relative importance of intra- and inter-population variance components. In our study, by contrast, a direct estimate of the optimal sample size per population is derived. Completely unexpectedly, it was shown to be independent of the total sampling effort. This is an interesting property in large-scale sequential studies.

We also found that with increasing levels of genetic differentiation, decreasing optimal sample sizes per population will be necessary. Since our optimum sampling strategy is valid for a single locus or allele, it is not obvious for selecting a strategy in multilocus surveys if the level of differentiation is heterogeneous among loci. In that respect, Marshall and Brown (1975) recommended the sampling of a small number of individuals per population, with the rationale that this strategy would still ensure the sampling of 'locally common' alleles in collections of germplasm. This assumes some heterogeneity of differentiation among alleles or loci and emphasizes the importance of the fraction of the genome showing the highest degree of population subdivision, because of its possible involvement in adaptation to local conditions.

Appendix 1: covariance of $\hat{h}_S$ and $\hat{h}_T$

As $\hat{h}_S = n^{-1}\sum_k \hat{h}_S$ and from (9)

$$E(\hat{h}_S \hat{h}_T) = h_S - \frac{1}{n^2(n-1)} \sum_{u\neq v,ki} E(x_{ui} x_{vi} \hat{h}_k)$$

and

$$Cov(\hat{h}_S, \hat{h}_T) = \frac{2}{n}\left\{h_S(1-h_T) - \frac{1}{n}\sum_{ki} p_i E(\hat{h}_k x_{ki})\right\} + 0(n^{-2})$$

then, using

$$E(\hat{h}_k x_{ki}) = p_i - \frac{n_k - 2}{n_k}\sum_j E(p_{ki} p_{kj}^2) - \frac{2}{n_k} E p_{ki}^2,$$

$$Cov(\hat{h}_S, \hat{h}_T) = \frac{2}{n}\left\{-(1-h_S)(1-h_T) + \left(1 - \frac{2}{\tilde{n}}\right)\sum_{ij} p_i E(p_{ki} p_{kj}^2) + \frac{2}{\tilde{n}}\sum_i p_i E p_{ki}^2\right\} + 0(n^{-2}),$$

$$Cov_{inter}(\hat{h}_S, \hat{h}_T) = \frac{2}{n}\left\{\sum_{ij} p_i E(p_{ki} p_{kj}^2) - (1-h_S)(1-h_T)\right\} + 0(n^{-2}),$$

$$Cov_{intra}(\hat{h}_S, \hat{h}_T) = \frac{4}{n\tilde{n}}\left\{\sum_i p_i E(p_{ki}^2) - \sum_{ij} p_i E(p_{ki} p_{kj}^2)\right\} + 0(n^{-2}).$$

They are unbiasedly estimated by

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

$$\hat{Cov}_{intra}(\hat{h}_S, \hat{h}_T) = \frac{4}{n(n-1)}\sum_{ki}\left(x_{.i} - \frac{x_{ki}}{n}\right)\frac{n_k}{(n_k-1)(n_k-2)} x_{ki}\left(x_{ki} - \sum_j x_{kj}^2\right)$$

and

$$\hat{Cov}_{inter}(\hat{h}_S, \hat{h}_T) = \hat{Cov}(\hat{h}_S, \hat{h}_T) - Cov_{intra}(\hat{h}_S, \hat{h}_T).$$

The asymptotic behaviour of $\hat{h}_T$ derives from studying $n^{1/2}\sum_i(p_i^2 - x_{.i}^2)$ or equivalently $n^{1/2}\sum_i 2p_i(p_i - x_{.i})$ which have the same limit distribution as $n^{1/2}(\hat{h}_T - h_T)$. Now $(x_{.i} - p_i)_i = n^{-1}\sum_k\{(x_{ki} - p_i)_i\}$ is a normalized sum of independent variables with zero mean and a variance matrix with diagonal terms $\sigma_{ki}^2 = n_k^{-1}U_i + V_i$ where $U_i = p_i(1-p_i) - V_i$ and other terms $\sigma_{kij}^2 = C_{ij} - n_k^{-1}(p_i p_j + C_{ij})$. If the n_ks are bounded, $n^{1/2}(p_i - x_{.i})_i$ converges to a Gaussian variable with zero mean and a finite variance and $n^{1/2}(\hat{h}_S - h_S, \hat{h}_T - h_T)$ converges to two-dimensional Gaussian variable.

Appendix 2: Estimation of $\tilde{n}_{opt}$

The terms A, B, C, D, and E appearing in (20) are defined by

$$A = \frac{n}{h_T^2} Var_{inter}(\hat{h}_S) + \frac{n h_S^2}{h_T^4} Var_{inter}(\hat{h}_T) - \frac{2n h_S}{h_T^3} Cov_{inter}(\hat{h}_S, \hat{h}_T),$$

$$B = \alpha\frac{h_S^2}{h_T^4} Var_{intra}(\hat{h}_T) - \frac{2\alpha h_S}{h_T^3} Cov_{intra}(\hat{h}_S, \hat{h}_T),$$

$$C = \frac{2}{h_T^2} E\left(\sum_i p_{ki}^2\right)^2,$$

$$D = \frac{2}{h_T^2} E\sum_i p_{ki}^3,$$

$$E = \frac{2}{h_T^2} E\left(\sum_i p_{ki}^2\right).$$

An estimate of $\tilde{n}_{opt}$ is defined in the same form, replacing each term of the above expressions by the estimate defined in the previous sections.

Acknowledgement The authors thank A. Kremer who made suggestions to improve the manuscript and provided great support during the project. A Turbo-Pascal program used to compute all diversity measures (estimates, variances, optimal sample sizes) is available upon request from R.J.P.

References

Bahrman N, Damerval C (1989) Linkage relationships of loci controlling protein amounts in maritime pine (*Pinus pinaster* Ait.). Heredity 63:267–274

Bogyo TP, Porceddu E, Perrino P (1980) Analysis of sampling stategies for collecting genetic material. Econ Bot 34:160–174

Brown AHD, Munday J (1982) Population-genetic structure and optimal sampling of land races of barley from Iran. Genetica 58:85–96

Cockerham CC, Weir BS (1986) Estimation of inbreeding parameters in stratified populations. Ann Hum Genet 50:271–281

Excoffer L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among haplotypes: application to human mitochondrial DNA restriction data. Genetics 131:479–491

Hamrick JL, Godt MJW, Sherman-Broyles SL (1992) Factors influencing levels of genetic diversity in woddy plant species. New Forests 6:95–124

Long JC (1986) The allelic correlation structure of Gainj- and Kalam-speaking people. I. The estimation and interpretation of Wright's *F*-statistics. Genetics 112:629–647

Lynch M, Crease TJ (1990) The analysis of population survey data on DNA sequence variation. Mol Biol Evol 7:377–394

Marshall DR, Brown AHD (1975) Optimal sampling strategies, genetic conservation. In: Frankel OH, Hawkes JG (eds) Crop genetic resources for today and tomaorrow. Cambridge University Press, pp 53–80

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

Nei M (1986) Definition and estimation of fixation indices. Evolution 40:643–645

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York

Nei M, Chesser RK (1983) Estimation of fixation indices and gene diversities. Ann Hum Genet 47:253–259

Nei M, Roychoudhury AK (1973) Sampling variances of heterozygosity and genetic distance. Genetics 76:379–390

Petit RJ, Kremer A, Wagner DB (1993) Geographic structure of chloroplast DNA polymorphisms in European oaks. Theor Appl Genet 87:122–128

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. Evolution 38:1358–1370

Wright S (1943) Isolation by distance. Genetics 28:114–138

Wright S (1951) The genetical structure of populations. Eugenics 15:323–354

Yonezawa K (1985) A definition of the optimal allocation of effort in conservation of plant genetic resources with application to sample size determination for field collection. Euphytica 34:345–354