Theor Appl Genet (1995) 91:122–130

O. Pons · K. Chaouche

Estimation, variance and optimal sampling of gene diversity
II. Diploid locus

Received: 15 November 1994 / Accepted: 15 December 1994

Abstract Nei's analysis of diversity at a diploid locus is extended to a population subdivided into a large number of subpopulations. The diversities and the heterozygotes frequency are defined with respect to the total population and unbiasedly estimated in a two-stage random cluster sampling. The fixation indices F_{IS}, F_{IT} and F_{ST} are derived, then inter- and intra-population variances of the estimated parameters are studied. We show that there is a unique sample size per population which yields the best accuracy in estimating F_{ST} and F_{IS}, respectively, at a given locus. These results are illustrated with an analysis of DNA diversity in a forest tree and compared to those obtained under the Hardy-Weinberg assumption.

Key words Diversity · Fixation indices · Diploid locus · Variance · Optimal design

Introduction

To describe the differentiation of populations, Wright (1943, 1951) introduced the fixations indices F_{IS}, F_{IT} and F_{ST} defined in terms of correlations between two uniting gametes within or between populations. Nei (1977) showed that these parameters are related to an analysis of the heterozygotes frequency based on the actual value H_0 of this frequency in the total population and on its expectation under Hardy-Weinberg equilibrium in the total population (H_T, the total diversity) and in the subpopulations (H_S, the average within-population diversity). This approach simply relies on the present state of the subpopulations without any evolutionary model. However, Nei's definitions depend on a fixed number of subpopulations, which is often regarded as restrictive for a population subdivided into a large number of subpopulations.

In the latter case, the variability of genotype frequencies among observed populations may be interpreted by considering that these frequencies are random variables and that the empirical frequencies are their realizations in random populations. Thus, the observed populations constitute a first level of sampling and the individuals within populations a second level of sampling. In this random setting, Pons and Petit (1994) generalized Nei's approach to the total population in the haploid case. New paramters h_S and h_T, corresponding to the previous H_S and H_T, were defined for the total population. Estimates of these parameters and of their sampling variance were proposed under assumptions similar to those used in Nei and Roychoudhury (1973) and Nei (1987). Then an optimal sampling size of the populations ensuring the most accurate estimate of the differentiation index G_{ST} followed.

In the present paper, we consider the case of diploid populations along the same lines. We extend Nei's notions of diversity and fixation indices to the total population and we adapt the estimates which were defined by Nei and Chesser (1983), for a fixed number of populations, to obtain unbiased estimates in the two-stage random sampling. The variances of all the estimates are developed in within-population variances and between-population variances due to the random sampling of the populations. Then the optimal sampling sizes of the populations to obtain minimal variance of the estimate of F_{ST} or F_{IS}, respectively, are deduced. A numerical example where h_0 is close to h_S is discussed.

Definition of diversity and differentiation indices

We consider a diploid population subdivided into independent subpopulations in which I alleles $A_1, A_2, \ldots, A_I$ are segregating at a diploid locus. We assume that the total number of populations and of individuals per population are very large with respect to the observed

Communicated by P. M. A. Tigerstedt

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

numbers, so that they are considered as infinity. We also assume that the populations have sizes of the same order and the same importance in the global population. Otherwise, weights could be added to take account of their respective sizes.

In the general population, the frequency of the genotype A_iA_j is denoted by P_{ij}, and $p_i = P_{ii} + \Sigma_{j \neq i} P_{ij}/2$ is the frequency of the allele A_i. In the k-th population, P_{kij} and p_{ki} are similarly defined as the frequencies of the genotype A_iA_j and of the allele A_i. For randomly selected populations, the P_{kij}s and p_{ki}s are considered as random frequencies with means $P_{ij} = EP_{kij}$ and $p_i = Ep_{ki}$ in the total population and their variances among the subpopulations are $V_{ij} = E(P_{kij}^2) - P_{ij}^2$ and $v_i = E(p_{ki}^2) - p_i^2$. Following Nei (1973) and Nei and Chesser (1983), if the k-th population is considered as fixed, we define its frequency of heterozygotes as

$$h_{0k} = 1 - \sum_i P_{kii} \tag{1}$$

and its diversity as

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

In the general population, these notion are generalized in the same way as for haploid individuals (Pons and Petit 1994), giving the mean frequency of heterozygotes as

$$h_0 = 1 - \sum_i P_{ii} \tag{3}$$

the mean within-population diversity as $h_S = Eh_k$, i.e.

$$h_S = 1 - \sum_i (p_i^2 + v_i), \tag{4}$$

and the total diversity as

$$h_T = 1 - \sum_i p_i^2. \tag{5}$$

The corresponding definitions in Nei and Chesser (1983) concerned n fixed populations, $H_0 = n^{-1}\Sigma_k h_{0k}$, $H_S = n^{-1}\Sigma_k h_k$, and $H_T = 1 - \Sigma_i p_{\cdot i}^2$ where $p_{\cdot i} = n^{-1}\Sigma_k p_{ki}$ is the average frequency of allele i over the n observed populations. If the P_{kij} were directly observed, they would be estimates of our parameters; however, the difference of view point leads to different unbiased estimations of the total diversities.

Nei's fixation indices are extended as functions of (3), (4) and (5) according to his definitions (Nei 1977):

$$F_{IS} = 1 - h_0/h_S$$

$$F_{IT} = 1 - h_0/h_T$$

$$F_{ST} = 1 - h_S/h_T.$$

Estimation

Since the actual frequencies and their variances among populations are not observed, we have to estimate the previous parameters from empirical frequencies. A two-stage random cluster sampling is used for that prupose: n independent populations are drawn with the same probability in the general population, then n_k individuals are drawn independently and uniformly from the k-th population. In the k-th population, n_{kij} denotes the number of individuals having the genotype A_iA_j, $X_{kij} = n_{kij}/n_k$ is the empirical frequency of this genotype and $x_{ki} = X_{kii} + \Sigma_{j \neq i} X_{kij}/2$ is the empirical frequency of allele i for $i, j \leq I$, $k \leq n$. As in Nei and Chesser (1983), we assume that the random vector $(n_{kij})_{ij}$ of length I^2 has a multinomial distribution $\mathcal{M}[n_k, (P_{kij})_{ij}]$ within the k-th population. The expectation with respect to this multinomial distribution, and conditionally on the k-the population, is denoted by E_k and E^{pop} is the expectation conditionally on the n sampled populations.

In the k-th population, the heterozygotes frequency (1) is unbiasedly estimated by

$$\hat{h}_{0k} = 1 - \sum_i X_{kii}.$$

Using an expansion of x_{ki}^2 in terms of the X_{kij}s, $j = 1, \ldots, I$, and the moments of these variables under the multinomial distribution (see Appendix), we get

$$1 - \sum_i E_k x_{ki}^2 = \frac{n_k - 1}{n_k} h_k + \frac{h_{0k}}{2n_k}$$

and Nei and Chesser's unbiased estimate of the diversity of the k-th population (2) is deduced as

$$\hat{h}_k = \frac{n_k}{n_k - 1}\left(1 - \sum_i x_{ki}^2 - \frac{\hat{h}_{0k}}{2n_k}\right). \tag{6}$$

In the global population, unbiased estimates of h_0 (3) and h_S (4) follow as

$$\hat{h}_0 = \frac{1}{n} \sum_{k \leq n} \hat{h}_{0k}, \tag{7}$$

$$\hat{h}_S = \frac{1}{n} \sum_{k \leq n} \hat{h}_k. \tag{8}$$

Note that if $\tilde{n} = n(\Sigma_k n_k^{-1})^{-1}$ is the harmonic mean of the n_ks, the estimate $\tilde{n}(\tilde{n} - 1)^{-1}(1 - n^{-1}\Sigma_{ki} x_{ki}^2 - \hat{h}_0/2\tilde{n})$ of h_S proposed by Nei and Chesser (1983) is unbiased for random populations but (8) will be prefered because of the homogeneity of the estimates.

The total diversity h_T (5) depends only on the mean allelic frequencies and not on the individuals within populations; the estimate that we previously defined for

haploid individuals still holds. If $x_{\cdot i} = n^{-1}\Sigma_k x_{ki}$, then

$$\hat{h}_T = 1 - \sum_i x_{\cdot i}^2 + \frac{1}{n(n-1)}\sum_{ki}(x_{ki} - x_{\cdot i})^2$$

$$= [n(n-1)]^{-1}\sum_{k\neq k'}\left(1 - \sum_i x_{ki}x_{k'i}\right) \quad (9)$$

is an unbiased estimate of h_T since the populations are assumed to be independent.

Estimates of the fixation indices are deduced by introducing the corresponding estimates of the diversities in their definitions, similarly to (12), (13) and (14) in Nei and Chesser (1983).

Variance of the estimates

For each parameter, the distance between the estimate and the parameter is the sum of within-population and between-population variations in the two-stage random sampling. If the populations were considered as fixed, the second term would be zero and the variance of the estimates would reduce to a within-population variance. Otherwise, the between-population variance of the estimates is of the order n^{-1} and the within-population variances are also of the order n^{-1} with respect to n, but it also depends on the n_k^{-1}s. We therefore consider that n is large in order to reduce the variance of the estimates.

For $\hat{h}_0$, this decomposition yields

$$Var(\hat{h}_0) = n^{-2}\sum_k EVar_k(\hat{h}_{0k}) + n^{-1}Var(h_{0k}).$$

For each k, $Var(h_{0k}) = \Sigma_i Var(P_{kii}) + \Sigma_{i\neq j} Cov(P_{kii}, P_{kjj})$, which is denoted as V_0; $Var_k(\hat{h}_{0k})$ is obtained from the multinomial distribution $Var_k(\hat{h}_{0k}) = n_k^{-1}h_{0k}(1-h_{0k})$. It follows that $Var(\hat{h}_0)$ is the sum of the within-population and between-population variances

$$Var_{intra}(\hat{h}_0) = \frac{1}{n\tilde{n}}[h_0(1-h_0) - V_0],$$

$$Var_{inter}(\hat{h}_0) = \frac{V_0}{n}.$$

As in the haploid case, the variance of $\hat{h}_S$ is the sum of within-population and between-population variances

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

where the h_ks are independent and identically distributed variables then

$$Var_{inter}(\hat{h}_S) = n^{-1}Var(h_k) = n^{-1}\left[E\left(\sum_i p_{ki}^2\right)^2 - (1-h_S)^2\right].$$

Using the moments of the multinomial distribution (see Appendix), the variance of $\hat{h}_k$ within the k-th population is

$$Var_k(\hat{h}_k) = \frac{1}{n_k(n_k-1)}\left[2(3-2n_k)\left(\sum_i p_{ki}^2\right)^2 + (n_k-2)\left(2\sum_i p_{ki}^3 + 2\sum_i p_{ki}^2 P_{kii} + \sum_{i\neq j} p_{ki}p_{kj}P_{kij}\right) + \frac{1}{2}\left(\sum_i p_{ki}^2 + \sum_i P_{kii}^2\right) + \sum_i p_{ki}P_{kii} + \frac{1}{8}\sum_{i\neq j} P_{kij}^2\right]$$

and the within-population variance of $\hat{h}_S$ follows as $Var_{intra}(\hat{h}_S) = n^{-2}EVar_k(\hat{h}_k)$.

The variance of h_T is defined from (5) and (9) as

$$Var(\hat{h}_T) = \frac{1}{n^2(n-1)^2}E\left[\sum_{k\neq l}\sum_i (p_i^2 - x_{ki}x_{li})\right]^2$$

$$= \frac{1}{n^2(n-1)^2}\left\{2\sum_{k\neq l}E\left[\sum_i (p_i^2 - x_{ki}x_{li})\right]^2 + 4\sum_{kll'\neq} E\sum_{ij}(p_i^2 - x_{ki}x_{li})(p_j^2 - x_{kj}x_{l'j})\right\}.$$

It splits into within- and between-population terms leading to the following approximations,

$$Var_{inter}(\hat{h}_T) = \frac{1}{n(n-1)}\left(2\sum_i\sum_j c_{ij}^2 + 4n\sum_i\sum_j c_{ij}p_ip_j\right)$$

$$= \frac{4}{n}\sum_i\sum_j c_{ij}p_ip_j + O(n^{-2}),$$

$$Var_{intra}(\hat{h}_T) = \frac{1}{n\tilde{n}}\left[\sum_{i\neq j} p_ip_jP_{ij} + \frac{1}{2}\sum_i p_i^2(p_i + P_{ii}) - 4\sum_{ij} p_ip_j(c_{ij} + p_ip_j)\right] + O(n^{-2}),$$

and $Var(\hat{h}_T)$ is approximated by their sum.

Variances of the fixation indices

The fixation indices are defined by two h parameters (h_0, h_S or h_T) on the form $1 - h_1/h_2$, where the estimates $\hat{h}_1$ and $\hat{h}_2$ of h_1 and h_2 are dependent. Their variance is then approximated on the form

$$\frac{Var(\hat{h}_1)}{h_2^2} - 2h_1\frac{Cov(\hat{h}_1, \hat{h}_2)}{h_2^3} + h_1^2\frac{Var(\hat{h}_2)}{h_2^4} \quad (10)$$

which requires an analytic expression of the covariance between the h parameters. Moreover, similar results hold for the within- and between-population variances, where the variances and covariance are simply replaced by the corresponding within- and between-population terms.

From (8), (9) and by independence of the populations,

$$Cov(\hat{h}_S, \hat{h}_T) = \frac{2}{n}\left[h_S(1-h_T) - \frac{1}{n}\sum_{ki} p_i E(\hat{h}_k x_{ki}) \right]; \qquad (11)$$

multinomial formulae then allow a closed form development of the last expectation and provides the expression of the covariance between $\hat{h}_S$ and $\hat{h}_T$ as the sum of the two following covariances:

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

$$Cov_{intra}(\hat{h}_S, \hat{h}_T) = \frac{2}{n\tilde{n}}\left\{ -2\sum_{ij} p_i E(p_{ki} p_{kj}^2) + \frac{1}{2}\sum_{ij} p_i E(p_{kj} P_{kij}) + \sum_i p_i E\left[p_{ki}\left(p_{ki} + \frac{P_{kii}}{2} \right) \right] \right\}.$$

Since $\hat{h}_0$ and $\hat{h}_S$ are respectively the means of independent variables $\hat{h}_{0k}$ and $\hat{h}_k$,

$$Cov(\hat{h}_0, \hat{h}_S) = n^{-2}\sum_k ECov_k(\hat{h}_{0k}, \hat{h}_k) + n^{-1} Cov(h_{0k}, h_k)$$

and the terms of the right member are equal to

$$Cov_{inter}(\hat{h}_0, \hat{h}_S) = \frac{1}{n}\left[\sum_{ij} E(p_{ki}^2 P_{kjj}) - (1-h_0)(1-h_S) \right], \text{ and}$$

$$Cov_{intra}(\hat{h}_0, \hat{h}_S) = \frac{2}{n\tilde{n}}\left[\sum_i E(p_{ki} P_{kii}) - \sum_{ij} E(p_{ki}^2 P_{kjj}) \right].$$

For the covariance between $\hat{h}_0$ and $\hat{h}_T$, note that a formula similar to (11) also holds

$$Cov(\hat{h}_0, \hat{h}_T) = \frac{2}{n}\left[h_0(1-h_T) - \frac{1}{n}\sum_{ki} p_i E(\hat{h}_{0k} x_{ki}) \right], \qquad (12)$$

which develops as the sum of

$$Cov_{inter}(\hat{h}_0, \hat{h}_T) = \frac{2}{n}\left[\sum_{ij} p_i E(p_{ki} P_{kjj}) - (1-h_0)(1-h_T) \right],$$

and

$$Cov_{intra}(\hat{h}_0, \hat{h}_T) = \frac{2}{n\tilde{n}}\left[\sum_i p_i P_i - \sum_{ij} p_i E(p_{ki} P_{kjj}) \right].$$

Estimation of the variances

Because $\hat{h}_0$ and $\hat{h}_S$ are empirical means, their total variances are unbiasedly estimated by the corresponding empirical variances,

$$\hat{Var}(\hat{h}_0) = \frac{1}{n(n-1)}\sum_k (\hat{h}_{0k} - \hat{h}_0)^2,$$

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

and the variance of $\hat{h}_T$ has the same consistent estimate as in the haploid case,

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

Estimation of the between-population variances are obtained by estimating the parameters which appear in their expressions, and the estimates of the within-population variances are deduced by difference from the estimated total variances. For $Var_{inter}(\hat{h}_0)$ we use the unbiased estimates of $C_{ij} = Cov(P_{kii}, P_{kjj})$ and $V_{ii} = Var(P_{kii})$ if $i \neq j$,

$$\hat{C}_{ij} = \frac{1}{n-1}\sum_k (X_{kii} - X_{.ii})(X_{kjj} - X_{.jj}) + \frac{1}{n}\sum_k \frac{X_{kij} X_{kjj}}{n_k - 1},$$

$$\hat{V}_{ii} = \frac{1}{n-1}\sum_k (X_{kii} - X_{.ii})^2 + \frac{1}{n}\sum_k \frac{X_{kii}(X_{kii} - 1)}{n_k - 1},$$

with the notation $X_{.ij} = n^{-1}\Sigma_k X_{kij}$ for $i, j = 1, \ldots, I$, which yields

$$\hat{Var}_{inter}(\hat{h}_0) = \frac{1}{n(n-1)}\sum_{kij} (X_{kii} - X_{.ii})(X_{kjj} - X_{.jj}) + \frac{1}{n^2}\sum_{kij} \frac{X_{kii} X_{kjj}}{n_k - 1} - \frac{1}{n^2}\sum_{ki} \frac{X_{kii}}{n_k - 1}$$

For $Var_{inter}(\hat{h}_S)$ we need an unbiased estimate of $S = E[(\Sigma_i p_{ki}^2)^2]$ and it can be defined as $n^{-1}\Sigma_k \hat{S}_k$ where $\hat{S}_k$ is an estimate of $(\Sigma_i p_{ki}^2)$ defined in the Appendix from the multinomial distribution. Then

$$\hat{Var}_{inter}(\hat{h}_S) = \frac{1}{n}\left[\frac{1}{n}\sum_k \hat{S}_k - (1 - \hat{h}_S)^2 + \hat{Var}(\hat{h}_S) \right].$$

Estimation of $Var_{inter}(\hat{h}_T)$ requires estimation of the parameters $c_{ij} = Cov(p_{ki}, p_{kj})$ for $i \neq j$ and $c_{ii} = v_i = Var(p_{ki})$. These estimates differ from those obtained in the haploid case, so we now have

$$\hat{c}_{ij} = \frac{1}{n-1}\sum_k (x_{ki} - x_{.i})(x_{kj} - x_{.j}) + \frac{1}{n}\sum_k \frac{1}{n_k - 1}\left(x_{ki}x_{kj} - \frac{X_{kij}}{4}\right),$$

$$\hat{v}_i = \frac{1}{n-1}\sum_k (x_{ki} - x_{.i})^2 + \frac{1}{n}\sum_k \frac{1}{n_k - 1} \times \left[x_{ki}\left(x_{ki} - \frac{1}{2}\right) - \frac{1}{2}X_{kii}\right],$$

and a consistent estimate of $Var_{inter}(\hat{h}_T)$ derives as

$$\hat{Var}_{inter}(\hat{h}_T) = \frac{4}{n}\sum_{ij} x_{.i}x_{.j}.\hat{c}_{ij}.$$

Estimation of the variances of the fixation indices

Estimates of the total and within- or between-population variances of the fixations indices are deduced from (10) by replacing each term by an unbiased estimate. Thus we also have to estimate the total and between-population covariances of the h parameters. From (11) and the similar (12)

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

$$\hat{Cov}(\hat{h}_0, \hat{h}_T) = \frac{2}{n-2}\left[\hat{h}_0(1 - \hat{h}_T) - \frac{1}{n-1}\sum_{ki}\left(x_{.i} - \frac{x_{ki}}{n}\right)\hat{h}_{0k} x_{ki}\right].$$

Moreover, the covariance between $\hat{h}_0$ and $\hat{h}_S$ is estimated by the empirical covariance based on the variable $\hat{h}_{0k}$ and $\hat{h}_k$,

$$\hat{Cov}(\hat{h}_0, \hat{h}_S) = \frac{1}{n(n-1)}\sum_k (\hat{h}_{0k} - \hat{h}_0)(\hat{h}_k - \hat{h}_S).$$

For the between-population covariances, we use the following unbiased estimates of $s_1 = \Sigma_{ij} p_i E(p_{ki} p_{kj}^2)$, $s_2 = \Sigma_{ij} E(p_{ki}^2 P_{kjj})$ and $s_3 = \Sigma_{ij} p_i E(p_{ki} P_{kjj})$, which are built from the multi nomial distribution,

$$\hat{s}_1 = \frac{1}{n-1}\sum_{ki}\frac{1}{(n_k-1)(n_k-2)}\left(x_{.i} - \frac{x_{ki}}{n}\right) \times \left\{n_k^2 x_{ki}\sum_j x_{kj}^2 - n_k[x_{ki}(x_{ki} + X_{kii}/2)] - \frac{n_k}{2}\sum_j [x_{ki}(x_{kj} + X_{kjj}) + x_{kj}X_{kij}] + (x_{ki} + X_{kii})\right\},$$

$$\hat{s}_2 = \frac{1}{n}\sum_{kij}\frac{1}{(n_k-1)(n_k-2)}\left\{n_k^2\sum_{ij} x_{ki}^2 X_{kjj} - \frac{n_k}{2} \times \left[\sum_i X_{kii} + \left(\sum_i X_{kii}\right)^2 + 4\sum_i x_{ki}X_{kii}\right] + 2\sum_i X_{kii}\right\},$$

$$\hat{s}_3 = \frac{1}{n-1}\sum_{ki}\frac{1}{(n_k-1)}\left(x_{.i} - \frac{x_{ki}}{n}\right)\left(n_k x_{ki}\sum_j X_{kjj} - X_{kii}\right).$$

Then the between-population covariances of the parameters are unbiasedly estimated by expressions of the same form,

$$\hat{Cov}_{inter}(\hat{h}_S, \hat{h}_T) = \frac{2}{n}[\hat{s}_1 - (1 - \hat{h}_S)(1 - \hat{h}_T) + \hat{Cov}(\hat{h}_S, \hat{h}_T)],$$

$$\hat{Cov}_{inter}(\hat{h}_0, \hat{h}_S) = \frac{1}{n}[\hat{s}_2 - (1 - \hat{h}_0)(1 - \hat{h}_S) + \hat{Cov}(\hat{h}_0, \hat{h}_S)],$$

$$\hat{Cov}_{inter}(\hat{h}_0, \hat{h}_T) = \frac{2}{n}[\hat{s}_3 - (1 - \hat{h}_0)(1 - \hat{h}_T) + \hat{Cov}(\hat{h}_0, \hat{h}_T)].$$

Optimal sampling designs

In the haploid case, Pons and Petit (1994) proposed an optimal sampling size of all the populations ($n_k = \tilde{n}$ for each k) such that the variance of a preliminary estimate $\hat{F}_{ST}$ is minimal for a fixed number of individuals to analyse, $n\tilde{n}$. This solution can be adapted here for the indices F_{ST} and F_{IS} according to the decomposition of their variance as the sum of expressions of the form (10) for within- and between-populations. For both $F = F_{ST}$ or F_{IS}, of form $1 - h_1/h_2$, under an equal size $\tilde{n}$ for all the populations, $Var(\hat{F})$ is approximated by

$$f(n, \tilde{n}) = \frac{A}{n} + \frac{B}{n\tilde{n}} + \frac{1}{n\tilde{n}(\tilde{n}-1)}[C(3 - 2\tilde{n}) + D(\tilde{n} - 2) + E] \quad (13)$$

and

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

where $A = nVar_{inter}(\hat{F})$, which is a constant for F_{IS} and tends to a constant as n increases for F_{ST} because of the

approximation in $Var_{inter}(\hat{h}_T)$; $B = n\tilde{n}[(h_1/h_2)^2 Var_{intra}(\hat{h}_2) - 2(h_1/h_2)Cov_{intra}(\hat{h}_1, \hat{h}_2)]/h_2^2$ is a constant for F_{ST} and F_{IS}; 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 $1/\tilde{n}(\tilde{n}-1)$ which appear in $Var_{intra}(\hat{h}_S)$. Denoting

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

$$d = E\left(2\sum_i p_{ki}^3 + 2\sum_i p_{ki}^2 P_{kii} + \sum_{i\neq j} p_{ki}p_{kj}P_{kij}\right),$$

$$e = E\left(\frac{1}{2}\sum_i p_{ki}^2 + 2\sum_i p_{ki}P_{kii} + \frac{1}{2}\sum_i P_{kii}^2 + \frac{1}{8}\sum_{i\neq j} P_{kij}^2\right),$$

for F_{ST} we have $C = c/h_T^2$, $D = d/h_T^2$ and $E = e/h_T^2$, and for F_{IS} we have $C = ch_0^2/h_S^4$, $D = dh_0^2/h_S^4$ and $E = eh_0^2/h_S^4$.

Thus, both criteria of minimal variance of $\hat{F}_{ST}$ and $\hat{F}_{IS}$ lead to solutions $\tilde{n}_{opt}(\hat{F}_{ST})$ and $\tilde{n}_{opt}(\hat{F}_{IS})$ which do not depend on the total number of individuals to analyse. This is not the same for $\hat{F}_{IT}$ since each intra-population variance or covariance appearing in the decomposition of $Var(\hat{F}_{IT})$ is of the order $(n\tilde{n})^{-1}$. In that case, the greatest number of populations provides the smallest variance of $\hat{F}_{IT}$ if $n\tilde{n}$ is fixed.

The optimal sizes $\tilde{n}_{opt}(\hat{F}_{ST})$ and $\tilde{n}_{opt}(F_{IS})$ may be estimated from a large enough preliminary sample: unbiased estimates of the constants d and e are

$$\hat{d} = \frac{1}{n}\sum_k \frac{1}{(n_k-1)(n_k-2)}\left[n_k^2\left(2\sum_i x_{ki}^3 + 2\sum_i x_{ki}^2 X_{kii} + \sum_{i\neq j} x_{ki}x_{kj}X_{kij}\right) - n_k\left(5\sum_i x_{ki}^2 + 6\sum_i x_{ki}X_{kii} + \sum_i X_{kii}^2 + \sum_{i\neq j} X_{kij}^2/4\right) + 3(2-n_k)\right],$$

$$\hat{e} = \frac{1}{n}\sum_k \frac{1}{n_k-1}\left[n_k\left(2\sum_i x_{ki}^2/2 + 2\sum_i x_{ki}X_{kii} + \sum_i X_{kii}^2/2 + \sum_{i\neq j} X_{kij}^2/8\right) - 3 + 5\hat{h}_{0k}/2\right],$$

and $\hat{c}$ derives from the estimate of $E(\Sigma_i p_{ki}^2)^2$ given in the Appendix. The corresponding $\hat{C}$, $\hat{D}$ and $\hat{E}$ may be defined as $\hat{C} = \hat{c}/\hat{h}_T^2$, $\hat{D} = \hat{d}/\hat{h}_T^2$ and $\hat{E} = \hat{e}/\hat{h}_T^2$ for F_{ST} and the analogous formulae with the coefficient $\hat{h}_0^2/\hat{h}_S^4$ for F_{IS}.

Estimation under the Hardy-Weinberg equilibrium

Although the diversities in the populations are defined as the Hardy-Weinberg expectation of heterozygosity, the above estimates do not take this assumption into account. Under the Hardy-Weinberg assumption, $h_{0k} = h_k$ for each k and the estimate of h_k simplifies as

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

which is similar to the estimate obtained for haploid individuals but with $2n_k$ observed alleles instead of n_k. This has to be related to the fact that under the Hardy-Weinberg assumption, and conditionally on the k-th population, the number of alleles of the different types $(n_{ki})_i = (n_k x_{ki})_i$ have the multinomial distribution $\mathcal{M}[2n_k, (p_{ki})_i]$. An estimate of h_S is deduced as

$$\hat{h}_S^{HW} = \frac{1}{n}\sum_k \frac{2n_k}{2n_k-1}\left(1 - \sum_i x_{ki}^2\right),$$

but the estimate of h_T remains unchanged under the Hardy-Weinberg assumption with random populations, contrary to the case of fixed populations (Nei and Chesser 1983).

Since the distribution of the x_{ki}s is now the same as that considered in a haploid population (Pons and Petit 1994), the variances of $\hat{h}_S$, $\hat{h}_T$ and $\hat{G}_{ST}$ are similar to those obtained in that case; but everywhere for the k-th population the number n_k of observed haploid individuals has to be replaced by the total number of observed alleles, $2n_k$.

Numerical example

To illustrate the results of this paper, we consider a data set from a large study of gene diversity in the sessile oak [*Quercus petraea* (Matt.) Liebl] in Europe using several isozyme markers. Sessile oak is a diploid, hermaphrodite and allogamous, wind-pollinated deciduous tree species. A preliminary study of 18 populations has already been published (Zanetto et al. 1993). Since then, the sample size has greatly increased and we selected a single locus (acid phosphatase, EC no. 3.1.3.2) and a total of 81 populations sampled over most of the European range of this species. A total of five alleles were detected at this locus; alleles 2 and 4 are largely predominant, and genotypes 2 2, 2 4 and 4 4 make up 98.3% of all the genotypes found. The arithmetic and harmonic mean number of genotypes per population were respectively 114.6 and 112.4.

We compare the estimates proposed in this paper for diploid individuals with the "haploid" estimates obtained under the Hardy-Weinberg equilibrium, as defined in Pons and Petit (1994). The estimates obtained by both methods are generally very close (Table 1) but with some differences in the decomposition of the vairances and covariances into within- and between-population terms (Tables 2 and 3). Owing to a large number of individuals per population, the variances and covariances of the estimates are very small, but for $\hat{F}_{IS}$ and $\hat{F}_{IT}$ they are larger than the others and quite similar.

Table 1 Comparison of the estimates

	h_0	h_S	h_T	F_{ST}	F_{IS}	F_{IT}
Haploid		0.494611	0.509333	0.028904		
Diploid	0.48506	0.494646	0.509333	0.028836	0.019407	0.047683

Table 2 Variances of the estimates $\times 10^6$

		h_0	h_S	h_T	F_{ST}	F_{IS}	F_{IT}
Total variance	Haploid		10.81	2.16	52.24		
	Diploid	56.42	10.84	2.16	52.53	220.89	224.23
Variance, inter	Haploid		8.91	1.73	47.60		
	Diploid	29.22	8.95	1.56	49.69	123.54	122.82
Variance, intra	Haploid		1.89	0.44	4.63		
	Diploid	27.21	1.89	0.60	5.97	97.36	101.41

Table 3 Covariances of the estimates $\times 10^6$

		$Cov(h_S, h_T)$	$Cov(h_0, h_S)$	$Cov(h_0, h_T)$
Total covariance	Haploid	−0.36		
	Diploid	−0.38	6.53	0.12
Covariance, inter	Haploid	−0.93		
	Diploid	−0.98	4.80	−0.36
Covariance, intra	Haploid	0.57		
	Diploid	0.59	1.73	0.47

The variance due to sampling within populations is lower than the variance due to the population sampling for $\hat{h}_S$ and $\hat{h}_T$, where the within-population variance accounts for about 20% of the total variance, and particularly for $\hat{F}_{ST}$, with a ratio of about 10%. By contrast, the partial variances of $\hat{h}_0$, $\hat{F}_{IS}$ and $\hat{F}_{IT}$ respectively, are of the same order.

The optimal sampling sizes for the minimal variance of $\hat{F}_{ST}$ are respectively estimated by 12.04 diploid individuals and 22.55 alleles under the Hardy-Weinberg assumption, corresponding approximately to 11 individuals. For $\hat{F}_{IS}$, the optimal size is estimated by 8.13 individuals, which differs from the value of 12.04 for $\hat{F}_{ST}$; this is not surprising since we now consider another criterion. All these values are very small when compared to the sampling used in most studies of gene diversity, as we already emphasized in Pons and Petit (1994) for a haploid locus. In this case, we related the small optimal size $\tilde{n}_{opt}$ obtained for estimating the differentiation G_{ST} to the low within-population variance of $\hat{G}_{ST}$. Moreover, it appeared through simulations that, with $\tilde{n}_{opt}$ sampled individuals per population, $Var_{intra}(\hat{G}_{ST}) \simeq Var_{inter}(\hat{G}_{ST})$ which corresponds to the optimal sampling strategy proposed by Nei and Roychoudhury (1973). The same conclusions still hold for $\hat{F}_{ST}$ but not for $\hat{F}_{IS}$: the optimal size $\tilde{n}_{opt}(\hat{F}_{IS})$ is much smaller than the observed $\tilde{n}$ but the estimated within- and between-population variances are almost equal, and clearly $Var_{intra}(\hat{G}_{ST})$ becomes larger than $Var_{inter}(\hat{G}_{ST})$ with $\tilde{n}_{opt}(F_{IS})$ sampled individuals per population. So the respective importance of the within- and between-population variances does not allow one to determine the sampling size which minimizes the variance of $\hat{F}_{IS}$.

As in Pons and Petit (1994), the expression $f(n, \tilde{n})$ (13) can be used to describe the evolution of the variance of $\hat{F}_{ST}$ and $\hat{F}_{IS}$, respectively, as a function of the number of sampled individuals per population or as a function of the number of sampled populations. The corresponding curves are similar to those obtained for G_{ST} in the haploid case. For each locus considered, a comparison of the curves relating to F_{ST} and F_{IS} allows us to ascertain how far the two sampling strategies agree or disagree, and to compound graphically when a precise estimation of both F_{ST} and F_{IS} is required.

Since $\hat{F}_{IS}$ is close to zero, it is interesting to test the hypothesis "$F_{IS} = 0$" or, equivalently, "$h_0 = h_S$". Several authors recommend a χ^2 test based on $\hat{F}_{IS}$; however, it is not clear how to perform such a test since they do not estimate the variance of $\hat{F}_{IS}$ as it is necessary to normalize such a statistic. A simpler test of the hypothesis "$h_0 = h_S$" can be based on the convergence of $\sqrt{n}(\hat{h}_0 - h_0, \hat{h}_S - h_S)$ to a bivariate Gaussian variable with zero means. Hence, the statistic

$$U(h_0, h_S) = \frac{\hat{h}_0 - \hat{h}_S}{\{\hat{Var}(\hat{h}_0) + \hat{Var}(\hat{h}_S) - 2\hat{Cov}(\hat{h}_0, \hat{h}_S)\}^{1/2}}$$

tends to a standard Gaussian variable under the null hypothesis and otherwise tends to infinity. Note that it is easily calculated from the data and the test can be performed before the estimation of the other variances. Here, we get $U(h_0, h_S) = -1.304$ and the Hardy-Weinberg assumption may be considered as a good approximation when estimating the diversities and the differentiation index F_{ST}. The variances of the estimates are then simpler to calculate though the results are very similar.

In the same way, a test of the hypothesis "$F_{ST} = 0$", or equivalently "$h_T = h_S$", is based on the statistic $U(h_T, h_S)$ where $\hat{h}_T$ replaces $\hat{h}_0$ in the expression of $U(h_0, h_S)$. Now $U(h_T, h_S) = 3.97$ which has to be compared to the Gaussian quantiles; h_T and h_S are therefore significantly different.

Acknowledgements We are very grateful to the Laboratoire de Génétique et Amélioration des Arbres Forestiers, Institut National de la Recherche Agronomique, B.P.45, 33611 Gazinet cedex, France, and especially to Anne Zanetto for providing partly unpublished material from her study on oak species and Rémy Petit for his suggestion to compare optimal sampling for F_{ST} and F_{IS} with the "haploid" results.

Appendix

Multinomial moments

Under the assumption of a multinomial distribution for the number of individuals of each genotype, the moments of the random vector $X_k=(X_{kij})_{ij}$ of length I^2 may be calculated conditionally on the k-th population. For distinct sets of indices (i, j), (u, v) and (a, b) in $(1,\ldots,I^2)$,

$$E_k(X_{kij}^2)=[(n_k-1)P_{kij}^2+P_{kij}]/n_k$$

$$E_k(X_{kij}^3)=[(n_k-1)(n_k-2)P_{kij}^3+3(n_k-1)P_{kij}^2+P_{kij}]/n_k^2$$

$$E_k(X_{kij}^4)=[(n_k-1)(n_k-2)(n_k-3)P_{kij}^4+6(n_k-1)(n_k-2)P_{kij}^3+7(n_k-1)P_{kij}^2+P_{kij}]/n_k^3$$

$$E_k(X_{kij}X_{kuv})=(n_k-1)P_{kij}P_{kuv}/n_k,$$

which generalizes, for the product of L distinct variables, into

$$E_k(\Pi_{l\le L}X_{ka_lb_l})=\Pi_{l\le L}(n_k-l)P_{ka_lb_l}/n_k^L$$

$$E_k(X_{kij}^2X_{kuv})=(n_k-1)P_{kij}P_{kuv}[(n_k-2)P_{kij}+1]/n_k^2$$

$$E_k(X_{kij}^2X_{kuv}^2)=(n_k-1)P_{kij}P_{kuv}[(n_k-2)(n_k-3)P_{kij}P_{kuv}+(n_k-2)(P_{kij}+P_{kuv})+1]/n_k^3$$

$$E_k(X_{kij}^3X_{kuv})=(n_k-1)P_{kij}P_{kuv}[(n_k-2)(n_k-3)P_{kij}^2+3(n_k-2)P_{kij}+1]/n_k^3$$

$$E_k(X_{kij}^2X_{kuv}X_{kab})=(n_k-1)(n_k-2)P_{kij}P_{kuv}P_{kab}[(n_k-3)P_{kij}+1]/n_k^3.$$

Moments of the empirical frequencies of the alleles

The moments of the random vector $x_k=(x_{ki})_{i\le I}$ are deduced by an expansion of the x_{ki} in terms of the X_{kij}s, $j=1,\ldots,I$. We also needed to use some moments of the x_ks and X_ks. For instance, if $i\neq j$ we get

$$E_kx_{ki}^2=\frac{n_k-1}{n_k}p_{ki}^2+\frac{1}{2n_k}(p_{ki}+P_{kii})$$

$$E_k(x_{ki}x_{kj})=\frac{n_k-1}{n_k}p_{ki}p_{kj}+\frac{P_{kij}}{4n_k}$$

$$E_k(x_{ki}X_{kii})=\frac{n_k-1}{n_k}p_{ki}P_{kii}+\frac{P_{kii}}{n_k}$$

$$E_k(x_{ki}X_{kjj})=\frac{n_k-1}{n_k}p_{ki}P_{kjj}$$

$$E_k(x_{ki}X_{kij})=\frac{n_k-1}{n_k}p_{ki}P_{kij}+\frac{P_{kij}}{2n_k}$$

$$E_k(x_{ki}x_{kj}^2)=\frac{(n_k-1)(n_k-2)}{n_k^2}p_{ki}p_{kj}^2+\frac{n_k-1}{2n_k^2}[p_{ki}(p_{kj}+P_{kjj})+p_{kj}P_{kij}]+\frac{P_{kij}}{8n_k^2}$$

$$E_k(x_{ki}^3)=\frac{(n_k-1)(n_k-2)}{n_k^2}p_{ki}^3+\frac{3(n_k-1)}{2n_k^2}p_{ki}(p_{ki}+P_{kii})+\frac{1}{4n_k^2}(p_{ki}+3P_{kii})$$

$$E_k(x_{ki}^4)=\frac{(n_k-1)(n_k-2)(n_k-3)}{n_k^3}p_{ki}^4+\frac{3(n_k-1)(n_k-2)}{n_k^3}(p_{ki}^3+p_{ki}P_{kii})+\frac{n_k-1}{n_k^3}\left(\frac{7}{4}p_{ki}^2+\frac{3}{4}P_{kii}^2+\frac{9}{2}P_{kii}p_{ki}\right)+\frac{1}{8n_k^3}(7P_{kii}+p_{ki})$$

$$E_k(x_{ki}^2x_{kj}^2)=\frac{(n_k-1)(n_k-2)(n_k-3)}{n_k^3}p_{ki}^2p_{kj}^2+\frac{(n_k-1)(n_k-2)}{n_k^3}\times\left\{\frac{1}{2}[p_{ki}^2(p_{kj}+P_{kjj})+p_{kj}^2(p_{ki}+P_{kii})]+p_{ki}p_{kj}P_{kij}\right\}+\frac{(n_k-1)}{n_k^3}\times\left[\left(P_{kii}+\frac{1}{4}\sum_{l\neq j}P_{kil}\right)\left(P_{kjj}+\frac{1}{4}\sum_{h\neq j}P_{kjh}\right)+\frac{1}{4}P_{kij}(p_{ki}+p_{kj})+\frac{1}{8}P_{kij}^2\right]+\frac{P_{kij}}{16n_k^3}$$

$$E_k(X_{kjj}x_{ki}^2)=\frac{(n_k-1)(n_k-2)}{n_k^2}P_{kjj}p_{ki}^2+\frac{n_k-1}{2n_k^2}P_{kjj}(p_{ki}+P_{kii})$$

$$E_k(X_{kii}x_{ki}^2)=\frac{(n_k-1)(n_k-2)}{n_k^2}P_{kii}p_{ki}^2+\frac{n_k-1}{n_k^2}P_{kii}\left(\frac{1}{2}P_{kii}+\frac{5}{2}p_{ki}\right)+\frac{1}{n_k^2}P_{kii}$$

$$E_k\left(\sum_{i\neq j}X_{kij}x_{ki}x_{kj}\right)=\frac{(n_k-1)(n_k-2)}{n_k^2}\sum_{i\neq j}P_{kij}p_{ki}p_{kj}+\frac{n_k-1}{n_k^2}\times\left[2\sum_i p_{ki}(p_{ki}-P_{kii})+\frac{1}{4}\sum_{i\neq j}P_{kij}^2\right]+\frac{1}{2n_k^2}h_{0k}.$$

Estimation of $E(\Sigma_i p_{ki}^2)^2$

From the expression of $E_kx_{ki}^4$ and $E_kx_{ki}^2x_{kj}^2$, $S_k=(\Sigma_i p_{ki}^2)^2$ is unbiasedly estimated by

$$\hat{S}_k=\frac{n_k^3(\Sigma_i x_{ki}^2)^2}{(n_k-1)(n_k-2)(n_k-3)}-\frac{1}{n_k-3}\times\left(\sum_{ij}\widehat{p_{ki}^2P_{kjj}}+2\sum_i\widehat{p_{ki}^3}+2\sum_i\widehat{p_{ki}^2P_{kii}}+\sum_{i\neq j}p_{ki}\widehat{p_{kj}P_{kij}}\right)-\frac{1}{(n_k-2)(n_k-3)}\left[\left(n_k+\frac{1}{2}\right)\sum_i\widehat{p_{ki}^2}+3\sum_i\widehat{p_{ki}P_{kii}}+\frac{1}{2}\sum_i\widehat{P_{kii}^2}+\frac{1}{4}\left(\sum_i\widehat{P_{kii}}\right)^2+\frac{1}{8}\sum_{i\neq j}\widehat{P_{kij}^2}\right]+\frac{1}{4(n_k-1)(n_k-2)(n_k-3)}[3n_k+1-\hat{h}_{0k}(2n_k+1)]$$

where the estimates in this formula are deduced from the previous moments. This entails

$$\hat{S}_k = \frac{1}{(n_k-1)(n_k-2)(n_k-3)}\left[n_k^3\left(\sum_i x_{ki}^2\right)^2 \right.$$

$$- n_k^2\left(\sum_{ij} x_{ki}^2 X_{kjj} + 2\sum_i x_{ki}^3 + 2\sum_i x_{ki}^2 X_{kii} + \sum_{i\neq j} x_{ki}x_{kj}X_{kij}\right)$$

$$+ n_k\left(5\sum_i x_{ki}X_{kii} + \frac{1}{2}\sum_i X_{kii}^2 + \frac{9-2n_k}{2}\sum_i x_{ki}^2\right.$$

$$\left.\left. - \frac{1}{8}\sum_{i\neq j} X_{kij}^2 + \frac{1}{4}\hat{h}_{0k}^2\right) + \frac{7-n_k}{2}\hat{h}_{0k} + n_k - 6\right]$$

and $E(\Sigma_i p_{ki}^2)^2$ is estimated by $\Sigma_k \hat{S}_k/n$.

References

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

Nei M (1977) F-statistics and analysis of gene diversity in subdivided populations. Ann Hum Genet 41:225–233

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

Pons O, Petit RJ (1995) Estimation, variance and optimal sampling of gene diversity. I. Haploid locus. Theor Appl Genet 90:462–470

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

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

Zanetto A, Kremer A, Labbe T (1993) Differences of genetic variation based on isozymes of primary and secondary metabolism in *Quercus petraea*. Ann Sci For 50, suppl 1:245s – 252s