Theor. Appl. Genet. 51, 217-222 (1978)

Sampling Strategies for Using Female Gametophytes to Estimate Heterozygosity in Conifers

R.W. Morris and P.T. Spieth
Department of Genetics, University of California, Berkeley, California (USA)

Summary. Unbiased estimators of genotype and allele frequencies and their respective variances are obtained for loci identified by mendelian segregation in haploid female gametophytes from individual trees. By a minimum sampling variance criterion, the allocation of experimental effort between the number of female gametophytes analysed per tree and the number of trees sampled per population is examined for a fixed total amount of experimental effort. For estimating heterozygosity, the optimum sampling design for many (generally most) cases is three female gametophytes per tree, but may be more than three depending upon the true genotype frequencies in the population. For estimating allele frequencies, the optimum sampling design is one female gametophyte per tree except in cases where a strong negative correlation exists between alleles within genotpyes. Guidelines are discussed for determining a suitable number of female gametophytes to be analysed per tree in order to estimate heterozygosity.

Key words: Allozyme Data - Genetic Variation - Sampling Efficiency

Introduction

Techniques of gel electrophoresis are currently being extended to an ever-increasing array of organisms for the purpose of identifying and quantifying allelic variation at structural gene loci within natural populations (Lewontin 1974; Powell 1975). Coniferous trees are no exception; however, certain features of conifer biology give allozyme studies of coniferous trees a unique character that leads to certain methodological questions. Properly done, allozyme studies of natural populations have two phases. First, allelic relationships among protein phenotypes must be determined by genetic analyses. Second, the genetic organization of natural populations must be characterized by surveys that estimate allele frequencies and genotype frequencies within a number of populations. As a rule, the precision of such surveys depends upon analyzing as many individuals and populations as time and resources permit.

For conifers, the use of diploid sporophyte tissue to perform genetic analyses and test hypotheses concerning allelism among protein mobility variants is impeded by long generation time and difficulties inherent in the technology required for controlled matings. Investigators have circumvented these obstacles by utilizing haploid female gametophyte tissue ("endosperm") in seeds for genetic analysis of electrophoretically detectable protein variation. Female gametophyte tissue develops from the functional megaspore, a product of meiosis, and thereby provides the basis for a direct analysis of gene segregation and determination of the genetic basis of the variants. A number of enzyme loci in several species of coniferous trees have been identified with this method of genetic analysis (Bartels 1971; Bergmann 1973a, 1973b; Feret 1974; Simonsen and Wellendorf 1975).

In principle, estimates of allele and genotype frequencies can be obtained directly from electrophoretic analyses of diploid sporophyte tissue in a manner analogous to the methods used for studies of animal populations. However, tissue-specific differences in gene expression and enzyme extractability make it unwise to survey populations using tissue different from the tissue used for genetic analyses. Consequently, the method adopted for allozyme surveys of conifer populations is based upon inferring the diploid genotype of mature trees by analyzing the haploid genotypes of a number of female gametophytes from each tree. This method avoids the problem of tissue specificity and has several additional benefits such as ease of storage, ease of enzyme extraction, and control of

0040-5752/78/0051/0217/$ 1.20

developmental stage. On the other hand, this approach introduces a sampling dilemma.

The accuracy with which genotype frequencies can be estimated when using haplophase segregation analysis depends not only upon the number of trees included in the sample, but also upon the probability of correctly identifying each tree's genotype. As the number of female gametophytes analysed per tree is increased, the likelihood of a correct classification of genotypes is increased, but so is the amount of experimental effort required to classify each genotype. As a consequence of finite experimental resources, increasing the number of female gametophytes per tree generally results in a reduction in the number of trees included in the sample. Different workers have responded to this dilemma in different ways. For example, Bergmann (1973c) analysed 9 female gametophytes per tree and 15 to 20 trees per population; Tigerstedt (1973) analysed 6 female gametophytes for each of 50 trees per population. The purpose of this paper is to analytically determine an appropriate allocation of experimental effort between the number of female gametophytes analysed per tree and the number of trees sampled per population.

To state the problem more precisely, let k denote the number of female gametophytes analysed per tree and let n denote the number of different trees sampled. The total number of analyses to be performed is $N = k \cdot n$. For a given amount of experimental effort, N, we shall determine the value of k that yields the most efficient estimate of the true genotype frequencies in a population. Our criterion for an optimum (i.e. most efficient) value of k is that value which minimizes the sampling vairance of an unbiased estimator of genotype frequency. By this criterion we minimize the expected squared error of estimation. In this sense we shall claim to be finding an optimum value for k. Generally both allele and genotype frequencies are of interest in population surveys; we, therefore, include analyses of the sampling variances associated with estimators for each.

Analysis

The typical sampling situation can be represented as follows. Suppose m alleles, $A_1, A_2, \ldots, A_m$, are segregating at a particular locus within a diploid population of a coniferous tree species. Let P_{ii} denote the true genotype frequency of the homozygote A_iA_i and P_{ij} $(i \neq j)$ that of individuals heterozygous for alleles A_i and A_j; $\sum_{i=1}^{m} \sum_{j=1}^{i} P_{ij} = 1$. Seed samples are collected from each of n different trees. For each tree, k female gametophytes are analysed. If all k female gametophytes exhibit the same allele, the tree is classified as a homozygote; otherwise, it is classified as a heterozygote. Let n_{ii} denote the number of trees classified as having the homozygous genotype A_iA_i. Let n_{ij} $(i \neq j)$ denote the number of trees classified as heterozygous for alleles A_i and A_j. The total number of trees samples is $n = \sum_{i=1}^{m} \sum_{j=1}^{i} n_{ij}$. For simplicity, we assume that genotype sampling occurs with replacement; in practice, this usually implies the actual numer of trees in the population is much greater than n. We also assume that those trees which produce seeds constitute a random sample of genotypes within the population.

Heterozygous trees willl be mistakenly classified as homozygotes when, as a chance result of mendelian segregation, all k female gametophytes exhibit the same allele. For any heterozygous genotype the probability of this event occurring is $\lambda = (1/2)^{k-1}$. On the average, half of the misclassified individuals of genotype A_iA_j will be assigned to each of the two homozygous classes A_iA_i and A_jA_j. As a consequence of the non-zero probability of misclassifying heterozygous genotypes, the random variables, n_{ij} are multinomially distributed with parameters n and

$$\begin{array}{ll} P_{ii} + \frac{1}{2}\lambda \sum_{j \neq i} P_{ij} & \text{for } A_iA_i \\ \text{and} & \\ (1 - \lambda)P_{ij} & \text{for } A_iA_j. \end{array} \qquad (1)$$

Estimation of Genotype Frequencies

Estimators for the various genotype frequencies are obtained by direct compensation for the inflation of

the homozygous classes - and the corresponding reduction of the heterozygous classes - that occurs from misclassification of heterozygotes. In particular, unbiased estimators for the individual genotype frequencies are

$$\tilde{P}_{ii} = \frac{n_{ii}}{n} - \frac{1}{2}\lambda \sum_{j \neq i} \frac{n_{ij}}{(1-\lambda)n} \quad \text{for } A_iA_i$$

and (2)

$$\tilde{P}_{ij} = \frac{n_{ij}}{(1-\lambda)n} \quad \text{for } A_iA_j.$$

Rather than obtain a variance for the estimator of each genotype, consider an arbitrary grouping, denoted by S, of one or more of the $\frac{m(m-1)}{2}$ different heterozygotes. S may represent any particular collection of heterozygotes, ranging from a single heterozygous class (e.g. A_1A_2) to the collection of all heterozygotes in the population. Let $H_S = \Sigma_S P_{ij}$ denote the true frequency of the combined group of heterozygotes. The estimator of H_S is

$$\tilde{H}_S = \Sigma_S \tilde{P}_{ij} = \frac{\Sigma_S n_{ij}}{(1-\lambda)n}.$$

As the sum of unbiased estimators, it too is unbiased. When the combined group of heterozygotes, S, is distinguished from all the other genotypes, it follows from the parameters of the multinomial distribution given by (1) that $\Sigma_S n_{ij}$ is binomially distributed with parameters n and $(1-\lambda)H_S$. Therefore,

$$\mathrm{Var}(\tilde{H}_S) = \frac{H_S[1-(1-\lambda)H_S]}{(1-\lambda)n}.$$

Rearranging, substituting $\lambda = (1/2)^{k-1}$ and imposing the constraint $n = N/k$ gives,

$$\mathrm{Var}(\tilde{H}_S) = \frac{kH_S}{N}\left[\frac{1}{1-(1/2)^{k-1}} - H_S\right].$$

The problem now is to find the value of k that minimizes $\mathrm{Var}(\tilde{H}_S)$ for any given N and H_S. Since k takes on only integral values, the effect on $\mathrm{Var}(\tilde{H}_S)$ caused by increasing the number of female gametophytes analysed per tree from k to $k+1$, while keeping the total amount of effort N constant, can be examined by a difference equation. Let $\Delta V = \mathrm{Var}(H_S : k+1) - \mathrm{Var}(H_S : k)$ denote the change in variance that results from a unit increase in k. Using the above expression for the $\mathrm{Var}(\tilde{H}_S)$ gives

$$\Delta V = H_S[C_k - H_S]/N,$$

where

$$C_k = \frac{k+1}{1-(1/2)^k} - \frac{k}{1-(1/2)^{k-1}}.$$

Consequently, increasing the number of gametophytes per tree from k to $k+1$ will reduce the variance of $\tilde{H}_S$ (that is, ΔV will be negative) if, and only if, the true value H_S is greater than C_k, which is a function of k alone. For each integral value of k we refer to C_k as the critical value of heterozygosity. Table 1 lists the values of C_k for a number of values of k.

Because $\Delta V < 0$ if, and only if, $H_S > C_k$, it follows that a particular k will minimize the variance of the estimate $\tilde{H}_S$ when the true heterozygosity is such that $C_{k-1} < H_S < C_k$. For example, using three female gametophytes per tree minimizes the variance when $H_S < .57$; using four does so when $.57 < H_S < .76$; using five does so when $.76 < H_S < .86$; and so forth. The key point from this result is that the optimum value of k, by a minimum variance criterion, depends upon the true frequency of the class of heterozygotes whose frequency is to be estimated.

Table 1. Critical Values of True Heterozygosity

k	C_k	n_e
2	0	1
3	0.5714	2.33
4	0.7619	4.20
5	0.8602	7.15
6	0.9176	12.14
7	0.9519	20.79
8	0.9723	36.10
9	0.9843	63.69
10	0.9912	113.64

Increasing the number of macrogametophytes per tree from k to $k+1$ will give a more efficient estimate of heterozygosity if, and only if, the true heterozygosity is greater than the critical value, $[C_k/n_e]$ is the effective number of alleles corresponding to the critical level of heterozygosity

Although S denotes an arbitrary class of heterozygotes, in practice, two cases are of most interest: (i) the frequency of each particular genotype, for which the estimators are given by (2); and (ii) the total frequency of heterozygotes in the population for which the estimator is $\sum_{i=1}^{m}\sum_{j=i+1}^{m} n_{ij}/\{n(1-\lambda)\}$. In the first case, significant perturbation from Hardy-Weinberg proportions in the direction of heterozygote excess is necessary for the true frequency of any one heterozygote to exceed .57. Consequently, three female gametophytes per tree is generally the optimum strategy for estimating individual genotype frequencies. In the second case, with a large number of alleles, the total frequency of heterozygotes may exceed some of the critical values in Table 1, thereby making a strategy of more than three female gametophytes per tree optimum.

If some preliminary work has been done, an indication of the optimum k can be obtained from the number of alleles known to be at the locus in question. Table 1 lists the effective number of alleles, n_e, that correspond to the critical values of heterozygosity. Equating the actual number of alleles to the effective number of alleles provides a crude indication of the true level of heterozygosity in the population: n_e is based upon equally frequent alleles and random mating; in a real population, with unequal allele frequencies and a certain amount of positive assortative mating, the true level of heterozygosity will be less than that given by equating the actual to the effective number of alleles. Consequently, the values of n_e in Table 1 indicate that using more than four female gametophytes per tree is appropriate only for loci with a large actual number of alleles.

As a rule, therefore, using three or, in some cases, four female gametophytes per tree provides the optimum sampling strategy for the majority (but not all) of situations that might be encountered.

Estimation of Allele Frequencies

The true frequency of allele A_i in the population is $p_i = P_{ii} + 1/2 \sum_{j\neq i} P_{ij}$. The obvious method of estimating p_i is to simply count the A_i genes, so that the estimator is

$$\tilde{p}_i = \frac{n_{ii} + \frac{1}{2}\sum_{j\neq i} n_{ij}}{n}.$$

From (1), the expected value of $\tilde{p}_i$ is p_i, hence the estimator is unbiased. Misclassification of heterozygotes does not affect the expected value of $\tilde{p}_i$ because a given heterozygote is equally likely to be classified as homozygous for either of the two segregating alleles. An expression for the variance of $\tilde{p}_i$ can be obtained through straight-forward application of statistical theory, but the derivation is tedious and we give only the result:

$$\mathrm{Var}(\tilde{p}_i) = \frac{2p_i(1-p_i) - (1-\lambda)(p_i - P_{ii})}{2n}.$$

This expression can be put in a form more amenable to interpretation by making some changes of variables. Substituting $\lambda = (1/2)^{k-1}$ and $P_{ii} = p_i^2 + f_i p_i(1-p_i)$ and imposing the constraint $n = N/k$ yields

$$\mathrm{Var}(\tilde{p}_i) = \frac{p_i(1-p_i)}{n}\cdot\frac{k}{2}\left[2-(1-f_i)\left(1-\left(\tfrac{1}{2}\right)^{k-1}\right)\right].$$

The quantity f_i $(-1 \leq f_i \leq 1)$ is introduced simply as a change of variable for the homozygote frequency, P_{ii}; however, it can be interpreted as the correlation between A_i and non-A_i allelic states within individuals and reflects any perturbation from Hardy-Weinberg proportions that may exist at the time of sampling.

For $k = 1$, the variance of $\tilde{p}_i$ reduces to the binomial variance, $p_i(1-p_i)/N$. In general, any increase in k will increase the variance of $\tilde{p}_i$. This generality breaks down only if there exists, for whatever reason, a sufficiently strong negative correlation between allelic states within individuals in the population. In particular, when $-1 \leq f_i < -.707$ some values of k other $k = 1$ will give a lower sampling variance. However, such strongly negative values of f_i can occur in natural populations only in rather unusual situations approaching that of balanced lethals. Consequently, it is safe to assert that analysing one female gametophyte per tree is the optimum sampling design for estimating allele frequencies.

Discussion

The preceding analyses show that no single sampling strategy is universally optimum for estimating both allele frequencies and genotype frequencies. For allele frequencies the optimum procedure, except in rare cases, is to use one female gametophyte per tree, while the optimum for genotype frequencies is three or more female gametophytes per tree depending upon the true value of heterozygosity. Choosing a sampling strategy, therefore, requires subjectively weighing the objectives of each particular study. Using only a single female gametophyte from each tree, while maximizing the efficiency of estimating allele frequencies, excludes all information about genotype frequencies. Using two or more female gametophytes per tree provides information on both kinds of frequencies; however, there is a reduction in efficiency in the estimation of allele frequencies because the number of independently sampled alleles is reduced.

When both kinds of frequencies are of interest, a compromise strategy of two female gametophytes per tree may be considered. The consequence of such a compromise is an increase in the standard deviations of both estimates relative to their respective optimum strategies. We can place rough bounds upon the resulting increases. For $H_S < .5$, using $\underline{k} = 2$ instead of $\underline{k} = 3$ will increase the standard deviations of the estimates for genotype frequencies by less than ten percent. Concommitantly, the standard deviations of the estimates for allele frequencies will be increased by ten to fifteen percent above their values for $\underline{k} = 1$.

The strategy of using only three or four female gametophytes to determine the genotype of a parental tree is somewhat surprising. It is worthwile to attempt a heuristic explanation. Classification based upon so few female gametophytes per tree results in substantial misclassification: with $\underline{k} = 3$, one of every four heterozygotes will be erroneously classified as a homozygote. The estimators given by (2) correct for misclassification at the population level; in terms of individual trees, the mistakes in classification remain. The increased accuracy of estimation is, therefore, solely at the population level, where the ability to correct for errors of classification allows an increased number of different parental trees to be included in the study without changing the total number of assays $\underline{N}$.

Several points are worth specific mention. First, when working with the array of "observed" genotype frequencies, it is necessary to keep in mind that there is a high level of misclassification. For example, if the array is to be tested against Hardy-Weinberg proportions, the expected Hardy-Weinberg frequencies need to be corrected. In particular, expected genotype frequencies should be calculated as $n[\tilde{p}_i^2 + \lambda\tilde{p}_i(1 - \tilde{p}_i)]$ for genotype A_iA_i and as $n(1 - \lambda)2\tilde{p}_i\tilde{p}_j$ for A_iA_j and compared with the "observed" frequencies as classified by segregation.

Second, if for any reason it should be necessary to determine the genotype of an individual tree with a high degree of accuracy, then more female gametophytes must be examined than are needed for estimating population parameters. So doing will, of course, necessitate either doing more electrophoretic analyses or reducing the accuracy with which population parameters are estimated.

Finally, two assumptions about the experimental system are central to the preceding mathematical analysis and need specific mention to allow proper application of the recommended sampling strategy. The analyses in this paper are based upon exclusive use of female gametophytes for estimating maternal allele and genotype frequencies and, therefore, are peculiar to conifers. Other methods often employed for estimating these parameters in plant populations depend on some form of progeny testing. There are strong similarities between the theoretical considerations associated with sampling strategies for progeny tests and those discussed here; however, the strategies differ slightly, and for progeny testing the papers by Brown should be consulted (Brown and Allard 1970; Brown, Weir, and Marshall 1970; Brown 1975).

The principal assumption in the treatment of this paper is that the total number of electrophoretic assays to be done is a given constant, so that using fewer female gametophytes per tree allows a greater number of trees to be sampled. That is, $\underline{k}$ and $\underline{n}$ are related by the constant constraint $\underline{nk} = \underline{N}$. As a rule, an upper bound on $\underline{N}$ is imposed by laboratory considerations; however, it is possible that field conditions might limit the number of trees available for

sampling to a degree that makes the suggested strategy inappropriate. Suppose, for example, that laboratory resources allow 150 assays for each locus. The optimum strategy for estimating genotype frequencies would, usually, be to assay three female gametophytes from each of 50 trees. If the maximum number of assays should subsequently be raised to 240, the best response would be to increase the number of trees sampled to 80, while continuing to use only three gametophytes per tree. However, if seed samples can only be obtained from a maximum of 30 trees, doing 150 assays using five gametophytes from each of the 30 trees will give better data than would staying with three gametophytes and doing only 90 assays.

The point is that the strategy advocated by this paper is applicable to situations in which the limiting factor is the total number of electrophoretic assays rather than the availability of trees for sampling. As long as such is the case, the sampling strategies given above will, for the amount of work done, yield the most accurate estimates of allele frequencies and genotype frequencies.

Literature

Bartels, H.: Genetic control of multiple esterases from needles and macrogametophytes of *Picea abies*. Planta 99, 283-289 (1971)

Bergmann, F.: Genetic studies in *Picea abies* with the aid of isoenzyme identification. II. Genetic control of esterase and leucine aminopeptidase isoenzymes in haploid macrogametophytes of dormant seeds. (In German with English summary). Theor. Appl. Genet. 43, 222-225 (1973a)

Bergmann, F.: Analysis of genetic variation in a Douglas-Fir provenance by means of isozyme polymorphisms. IUFRO working party on Douglas-Fir provenances, Göttingen, West Germany. S2.02-05, 207-215 (1973b)

Bergmann, F.: Geographic pattern of genetic variation at four isozyme loci in the Finnish spruce population (*Picea abies*). IUFRO Joint Workshop and Symp. on "Norway Spruce Provenances", Biri/Norway 2.02.11, 6 pp. (1973c)

Brown, A.H.D.: Efficient experimental designs for the estimation of genetic parameters in plant populations. Biometrics 31, 145-160 (1975)

Brown, A.H.D.; Allard, R.W.: Estimation of the mating system in open-pollinated maize populations using isozyme polymorphisms. Genetics 66, 133-145 (1970)

Brown, A.H.D.; Weir, B.S.; Marshall, D.R.: Optimum family size for the estimation of heterozygosity in plant populations. Heredity 25, 233-239 (1970)

Feret, P.P.: Genetic differences among three small stands of *Pinus pungens*. Theor. Appl. Genet. 44, 173-177 (1974)

Lewontin, R.C.: The genetic basis of evolutionary change. New York: Columbia University Press 1974

Powell, J.R.: Protein variation in natural populations of animals. Evol. Biol. 8, 79-119 (1975)

Simonsen, V.; Wellendorf, H.: Some polymorphic isoenzymes in the seed endosperm of Sitka Spruce (*Picea sitchensis* (Bong.) Carr.), 20 pp. Kobenhaven: Forest Tree Improvement 9, Adademisk Forlag 1975

Tigerstedt, P.M.A.: Studies on isozyme variation in marginal and central populations of *Picea abies*. Hereditas 75, 47-60 (1973)

Received November 5, 1976/July 5, 1977
Communicated by W.J. Libby

Prof. P.T. Spieth
Department of Genetics
University of California
Berkely, California 94720
(USA)
(to whom reprint request should be directed)