Theor Appl Genet (1983) 65:197–200

Efficient estimation of allelic and genotypic proportions for electrophoretic loci

R. W. Marks
Department of Biology, Villanova University, Villanova, PA 19085, USA

C. A. Robertson
Department of Statistics, University of California, Riverside, Calif., USA

Received December 27, 1982
Communicated by R. W. Allard

Summary. The efficiency of estimating allelic and genotypic frequencies at allozyme loci can often be increased by running more than one individual simultaneously in each sample slot in a gel. Increased precision in estimating allelic frequencies is always possible if Hardy-Weinberg proportions can be assumed. However, if Hardy-Weinberg cannot be assumed, estimates of both allelic and genotypic frequencies can be improved for loci with hybrid dimers, but not for loci without them. The efficiency of all estimators depends jointly on allele frequencies and the number of individuals run simultaneously.

Key words: Allelic and genotypic frequencies – Gel electrophoresis

Introduction

Gel electrophoresis is widely used to determine allelic and/or genotypic proportions in population studies. When more individuals are available than can be run, the precision of estimation can be increased under many circumstances by running more than one individual per sample slot in the gel.

Two considerations are important. First, running gels is time consuming and can be expensive. The total number of gels which can be run is, therefore, limited. Second, in many organisms, a very large number of individuals can often be collected. Thus, in many circumstances far more individuals are available than can be run individually. The purpose of this note is to demonstrate how to use these individuals to get the best estimate of allelic and/or genotypic proportions.

Homogenates applied to the gel can easily contain tissue from more than one individual. Though combining individuals requires slightly more work, running these combined homogenates is no more expensive than running the same number of homogenates each with a single individual. Mixing causes some information to be lost because individuals cannot unambiguously be assigned to genotypes, but information is also gained from the increased number of individuals run. We will examine properties of estimators which assume more than one individual in each homogenate. We will examine the variances of these estimators and the dependence of these variances on allele frequency, presence or absence of hybrid dimers, deviations from Hardy-Weinberg proportions, and number of alleles. Our goal is to find the conditions under which it is advantageous to run more than one individual simultaneously.

Conventional estimates of allelic frequency

Electrophoretic genotypes are customarily assayed by homogenizing tissues from individuals and applying these homogenates, one at a time, to a gel. After running and staining the gel, individual genotypes may be read directly. The usual estimator of allele frequency, p^*, is given by $p^* = P^* + \frac{1}{2} H^*$, in which P^* and H^* are the estimates of P and H, the population frequencies of homozygote and heterozygote genotypes, respectively ($Q = 1 - P - H$). Let deviations Hardy-Weinberg proportions be measured as:

$$P = p^2 + F\,p\,q\,,$$

$$H = 2\,p\,q - 2F\,p\,q\,.$$

It can be shown that

$$\mathrm{Var}(p^*) = \frac{p\,q}{2N}(1 + F)\,. \qquad (1)$$

If the genotypes are in Hardy-Weinberg proportions, $F = 0$, and (1) is the familiar $\mathrm{Var}(p^*) = pq/2N$. We will assess the efficiency of other estimators by comparing their sampling variances with $\mathrm{Var}(p^*)$.

Estimates assuming more than one individual per homogenate

Let k represent the number of individuals per homogenate, and N the number of homogenates applied to the gel. In the usual case considered above, $k = 1$, and N is equal to the number of individuals. We will look at estimators for loci with two alleles in detail: one for loci which do not form a hybrid dimer, and two for loci which do. We will also look at estimators of genotypic proportions and briefly discuss the effects of additional alleles.

Two alleles, no hybrid dimer

Without a hybrid dimer three patterns on the gel are possible: two homomorphs and one heteromorph. This is still the case with k individuals in each homogenate. The homomorphs are obtained when the homogenate consists of k individuals all homozygous for the same allele; the heteromorph for all other cases. This is summarized in Fig. 1. The expected frequencies shown assume Hardy-Weinberg proportions (usually a good assumption: Cavalli-Sforza and Bodmer 1971, Lewontin 1974). Using standard maximum likelihood techniques (Cavalli-Sforza and Bodmer 1971) the estimator of p, $\hat{p}$, can be shown to be the solution of

$$0 = \frac{\hat{q}a - \hat{p}c}{\hat{p}\hat{q}} + \frac{b(\hat{q}^{2k-1} - \hat{p}^{2k-1})}{1 - \hat{p}^{2k} - \hat{q}^{2k}}.$$

Its variance is given approximately by

$$\mathrm{Var}(\hat{p}) \cong \frac{1}{4k^2N}\left[p^{2k-2} + q^{2k-2} + \frac{(q^{2k-1} - p^{2k-1})^2}{1 - p^{2k} - q^{2k}}\right].$$

We can assess the relative efficiencies of these estimators by examining $\mathrm{Var}(p^*)/\mathrm{Var}(\hat{p})$. If this ratio is greater than one, our new estimate, $\hat{p}$, is more efficient than p^*. Figure 2 plots this ratio, which is independent of N, as a function of allele frequency for several values of k.

For very small allele frequencies, increasingly better estimates will be obtained with larger values of k: we are likely to find a more representative number of rare types in collections of individuals larger than the collection on which p^* is based. For intermediate allele frequencies, high k's yield inefficient estimates: nearly every sample will show up as a heteromorph. However, note that for $k = 2$, $\hat{p}$ is better than p^* as an estimator of p at every allele frequency. We will therefore focus on $k = 2$ when looking at other cases.

Pattern on gel	—	— —	—
Expected frequencies assuming Hardy-Weinberg	p^{2k}	$1 - p^{2k} - q^{2k}$	q^{2k}
Observed numbers of homogenates	a	b ($a+b+c=N$)	c

Fig. 1. Electrophoretic banding patterns for a locus with two alleles, without a hybrid dimer, k individuals per homogenate. Allele frequencies are p and q ($p + q = 1$)

To obtain the estimator $\hat{p}$ we have assumed Hardy-Weinberg proportions for genotypes. It is also possible, of course, to estimate P, H, and Q, and use $\hat{P} + \frac{1}{2}\hat{H}$ as an estimator of allele frequency. However, for loci which do not form hybrid dimers, the variance of this estimator of p, for intermediate p is approximately as large, and for extremes of p is larger than the variance of p^*. On the other hand, in the next section we will show that the estimator based on genotypes, for loci which form hybrid dimers, is generally more efficient than p^*.

Two alleles, hybrid dimer, $k = 2$

For the case with hybrid dimers we will examine two estimators. The first, $\hat{p}_{HW}$, is obtained by assuming Hardy-Weinberg proportions and proceeding exactly as before. For the second, $\hat{p}_G$, we relax the assumption of Hardy-Weinberg, and find the maximum likelihood estimators for genotype frequencies, $\hat{P}$ and $\hat{H}$. The maximum likelihhod estimator of p is then obtained as $\hat{p}_G = \hat{P} + \frac{1}{2}\hat{H}$ (Breiman 1973).

We assume that if two individuals, each homozygous for a different allele, are homogenized together and run on a gel, they will form a hybrid dimer. (This is generally true – e.g., Hopkinson et al. 1976, but some

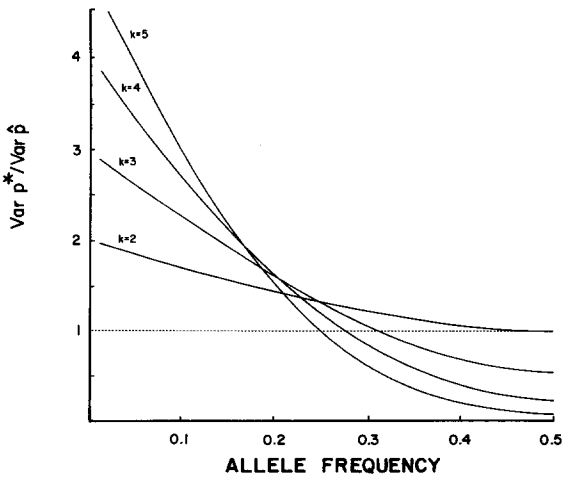

Fig. 2. Ratios of variances of the estimators of allele frequency at a two allele locus assuming Hardy-Weinberg, k individuals per homogenate. The curves are symmetrical for $0.5 < p < 1$

enzymes form hybrid dimers in vitro.) Thus there are four patterns possible on the gel. These patterns and their expected frequencies are shown in Fig. 3. Assuming Hardy-Weinberg proportions, the estimator $\hat{p}_{HW}$ is given by

$$0 = \frac{4a}{\hat{p}} - \frac{4d}{\hat{q}} + \frac{2b(\hat{q}-\hat{p})}{\hat{p}\hat{q}} + \frac{c(1-2\hat{p}-2\hat{p}\hat{q}(\hat{q}-\hat{p}))}{\hat{p}\hat{q}(1-\hat{p}\hat{q})},$$

in which the subscript HW has been dropped for readability. Its variance is given by

$$\mathrm{Var}(\hat{p}_{HW}) = \frac{1}{4N}(p^2+q^2+2pqA-B)^{-1}$$

in which

$$A = 2 + \frac{q-p}{pq}(1-2p),$$

and

$$B = \frac{(1-2p-2pq(q-p))^2}{pq(1-pq)} - 2(1-2pq+(q-p)^2).$$

If we do not assume Hardy-Weinberg, then $\hat{p}_G = \hat{P} + \frac{1}{2}\hat{H}$ in which the subscript G indicates estimators based on genotypic proportions. The estimates $\hat{P}$ and $\hat{H}$ are found by solving

$$\begin{pmatrix}0\\0\end{pmatrix} = \begin{pmatrix}\frac{2a+b}{\hat{P}} - \frac{2c(\hat{P}+\hat{Q})}{1-(\hat{P}+\hat{Q})^2}\\ \frac{2d+b}{\hat{Q}} - \frac{2c(\hat{P}+\hat{Q})}{1-(\hat{P}+\hat{Q})^2}\end{pmatrix}$$

using appropriate numerical methods (Cavalli-Sforza and Bodmer 1971).

The variance-covariance matrix for $\hat{P}$ and $\hat{Q}$ is obtained by the usual methods. From this, $\mathrm{Var}(\hat{p}_G)$ can be calculated, since $\mathrm{Var}(\hat{p}_G) = \frac{1}{4}(\mathrm{Var}\,P + \mathrm{Var}\,Q - 2\,\mathrm{Cov}(P,Q))$. We can look at the effect of deviations from Hardy-Weinberg on the efficiency of $\hat{P}_G$. As before, we use F to measure deviations from Hardy-Weinberg. Figure 4 shows the ratios of interest; the dashed curve for $\mathrm{Var}(p^*)/\mathrm{Var}(\hat{p}_{HW})$, and the solid curves for $\mathrm{Var}(p^*)/\mathrm{Var}(\hat{p}_G)$ for various values of F. Note that the estimate of p is improved only slightly by assuming that the genotypes are in Hardy-Weinberg proportions, in marked contrast to the case without a hybrid dimer.

Sampling variances for genotypic proportions, two alleles, k = 2

In the conventional method of estimation, the variance-covariance matrix for P* and Q*, call it **V***, is simply that of the multinomial distribution. We will compare **V*** with the variance-covariance (**V**) for $\hat{P}_G$ and $\hat{Q}_G$, by looking at the ratio of their determinants (generalized variance, Mood et al. 1974):

$$R = \frac{\mathbf{V}}{\mathbf{V}^*}.$$

With no hybrid dimer and k = 2, R is greater than 1 for all allele frequencies, and with departures from Hardy-Weinberg proportions. In addition, as we have already said, $\hat{p} = \hat{P} + \frac{1}{2}H$ in this case is generally no better an estimate than p*.

The situation is considerably different for loci with hybrid dimers. Figure 5 shows R as a function of allele frequencies for various values of F. Only in cases in which F < 0 would $\hat{P}$ and $\hat{Q}$ not always be better estimates than P* and Q*.

Pattern on gel	—	— —	— — —	—
Expected frequencies assuming Hardy-Weinberg	p^4	$2p^2q^2$	$1-(p^2+q^2)^2$	q^4
Expected frequencies not assuming Hardy-Weinberg	p^2	2PQ	$1-(P+Q)^2$	Q^2
Observed numbers of homogenates	a	b	c	d
			(a+b+c+d=N)	

Fig. 3. Electrophoretic banding patterns for a locus with two alleles, with a hybrid dimer, 2 individuals per homogenate (k = 2). P and Q are homozygote and heterozygote genotype frequencies, p and q the allele frequencies

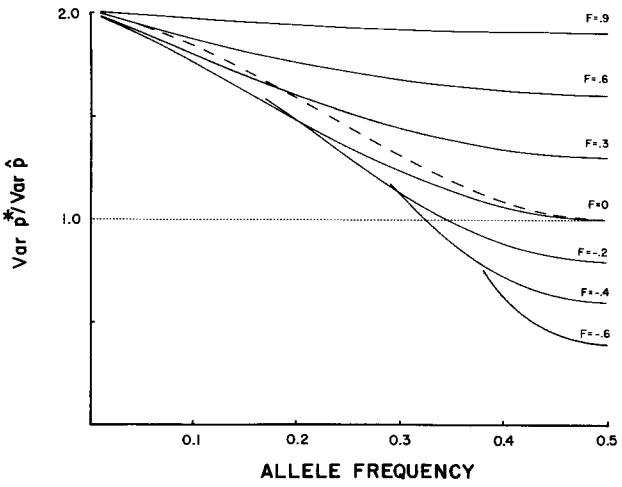

Fig. 4. Ratios of variances of estimators of allele frequency for a two allele locus, k = 2, with hybrid dimer. The dashed curve plots the ratio for the estimator assuming Hardy-Weinberg (with F = 0). The solid curves plot the ratio for estimators based on genotypic proportions. F measures the deviation from Hardy-Weinberg proportions. The curves for F < 0 do not extend to the ordinate because of the shape of the F – p state space

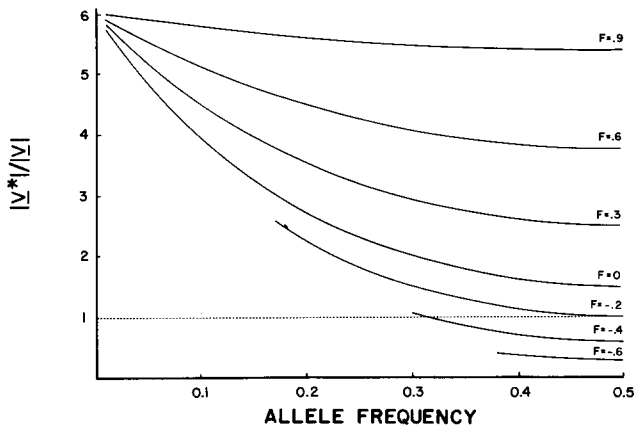

Fig. 5. Ratios of the generalized variances for estimators of genotypic proportions; two allele locus with hybrid dimer

Three or more alleles

The equations for the estimation of allele and genotype frequencies and their variances and covariances for loci with more than two alleles, with or without hybrid dimers, although cumbersome, are easily derived. We have looked explicitly at the properties of the estimators for three and four alleles, both with and without hybrid dimers ($k = 2$). In general, their behavior is much like the two allele case. With a hybrid dimer, both genotype and allele frequency estimates are improved by doubling up. The degree of overall improvement can be very large with more alleles, particularly when one genotype is in high frequency. (A more detailed discussion of how to derive estimators in general may be obtained by writing to R.W.M.)

Discussion

Running more than one individual per homogenate can be useful in a variety of electrophoretic studies. Spatial or temporal studies of electrophoretic variants are common. In studies such as these, and others in which the genetics and biochemistry of the loci of interest have been well characterized, the technique described here can be used very profitably. In fact, Slatkin and Charlesworth (1978) have demonstrated that patterns in the spatial distribution of rare alleles may be important in interpreting data from natural populations, and our technique is most useful for rare alleles.

An appropriate initial procedure for loci which form hybrid dimers would be to homogenize pairs of individuals ($k = 2$) because this improves estimates in virtually every case (provided hybrid dimers are not formed in vitro). Tests for fit to Hardy-Weinberg are possible since genotypic proportions are estimated. For loci without hybrid dimers, an initial check for fit to Hardy-Weinberg proportions is advisable because of the limitations of these particular estimators for $k > 1$. In either case, something is known about specific loci, k may be adjusted appropriately. Running more than one individual per homogenate is inadvisable only for loci without hybrid dimers which deviate from Hardy-Weinberg expectations.

There are, of course, practical operational limits on k. For example, if a large number of individuals is run simultaneously, staining for some bands may be so rapid or intense that the gel is impossible to read.

Some care must be taken in selecting k for a population in which the genotypes may not be in Hardy-Weinberg proportions. As mentioned above, for loci without hybrid dimers, $k = 1$ is appropriate. For loci with hybrid dimers, the direction of the deviation (sign of F) expected must be considered. One of the most common causes of such deviation is inbreeding, particularly in plants. Here $F > 0$ and the efficiency of the estimators for $k > 1$ is actually greater than those if $F = 0$ (Fig. 4). In addition to inbreeding, a deficiency of heterozygotes may be produced by Wahlund effect. If $F < 0$, $\mathrm{Var}(\hat{p}) < \mathrm{Var}(p^*)$ generally, and $k = 1$ is again appropriate. However, significant heterozygote excess is encountered far less often than heterozygote deficiency.

In a few other circumstances $k = 1$ would be appropriate. One obvious circumstance is when understanding of the formal genetics of the loci is incomplete. Another is when there is interest in correlating individual genotypes with some other attribute of the individual (for example, Tsakas and Krimbas 1970).

Using the guidelines suggested above it is possible to decide, in specific cases, whether running multiply is worthwhile. The increase in precision of estimates is worth the slight initial mathematical exercise required.

Acknowledgements. We would like to thank R. W. Allard, R. C. Lewontin, and T. Prout for helpful comments on the manuscript, and Mrs. M. Nordone for manuscript preparation.

References

Breiman L (1973) Statistics with a view toward applications. Houghton-Mifflin, Boston

Cavalli-Sforza LL, Bodmer WF (1971) The genetics of human populations. WH Freeman, San Francisco

Hopkinson DA, Edwards YH, Harris H (1976) The distributions of subunit numbers and subunit sizes of enzymes: a study of the products of 100 human gene loci. Ann Hum Genet 39: 383

Lewontin RC (1974) The genetic basis of evolutionary change. Columbia University Press, New York London

Mood AM, Graybill, FA, Boes, DC (1974) Introduction to the theory of statistics. McGraw Hill, New York

Slatkin M, Charlesworth D (1978) The spatial distribution of transient alleles in a subdivided population: a simulation study. Genetics 89:793–810

Tsakas S, Krimbas CB (1970) The genetics of *Dacus oleae*. 4. Relation between adult esterase genotypes and survival to organophosphate insecticides. Evolution 24:807–815