Heredity **65** (1990) 429–433
© The Genetical Society of Great Britain
Received 17 April 1990

Allele frequency estimation at loci with incomplete co-dominant expression

Per Erik Jorde and
Nils Ryman

Department of Genetics, Stockholm University,
S–106 91 Stockholm, Sweden.

Although allelic variants at a locus usually are expressed either dominantly or co-dominantly, there are many cases when a gene is dominantly expressed in some individuals and co-dominantly in others. We present a maximum-likelihood procedure for allele frequency estimation in such situations of "incomplete" co-dominant gene expression at an autosomal locus that segregates for two alleles. Our proposed estimator generally is less biased and has a smaller sampling variance than those previously described.

INTRODUCTION

Estimation of allele frequencies presents a problem when all genotypes at a locus cannot be directly inferred from the phenotypes, as under dominant gene expression. Estimators for various specific cases, *e.g.*, di-allelic loci with complete dominance or the human ABO blood group system, are presented in many textbooks (*e.g.*, Cavalli-Sforza and Bodmer, 1971; Hedrick, 1983). However, it is not clear which estimator should be applied when a gene is dominantly expressed in some individuals of a sample and co-dominantly in others, a situation that is frequently encountered when dealing with, *e.g.*, isozyme or blood group data. In this paper we provide a maximum-likelihood estimator applicable to samples where this type of "mixed gene expression" occurs.

BACKGROUND

Although allelic variants detected by protein electrophoresis usually are co-dominantly expressed, and thus easily interpreted in terms of genotypes, there are several exceptions to this general rule. In addition to the presence of inactive ("null") alleles and co-migration of different gene products, these exceptions include variable activity of gene products and formation of secondary isozymes, phenomenons that are highly dependent on such variable factors as sample quality and technical procedures (reviewed by Utter *et al.*, 1987).

Gene expression may differ not only between samples but also among individuals of the same sample, and a locus may be treated as co-dominant in one subset of the sample and as dominant in another. Examples of such protein loci include *Cpk*-1 and *Ldh*-1 (coding for creatine kinase and lactate dehydrogenase) in brown trout (*Salmo trutta*; Taggart *et al.*, 1981; Ryman and Ståhl, 1981; Allendorf *et al.*, 1984) and *Sdh*-1 (coding for sorbitol dehydrogenase) in Atlantic salmon (*Salmo salar*; cf. Ståhl, 1983; Cross and King, 1983). Typical examples of mixed gene expression also occur for many blood group systems; using a single test serum all the individuals are classified as "dominant" and "recessive" phenotypes, whereas only a subset are separated into homozygotes and heterozygotes by use of additional test sera (*e.g.*, Mourant *et al.*, 1976).

An empirical example illustrating the problem of allele frequency estimation is presented in table 1. Tissue samples from a total of 95 brown trout were collected from a remotely located lake in central Sweden. Sample collection took several days. When scoring the autosomal *Cpk*-1 locus that segregates for two alleles (*1* and *2*), the samples collected during the first few days (subsample 1) showed lower activity than those collected subsequently (subsample 2). All the three genotypes could be distinguished in subsample 2.

Table 1 Distribution of phenotypes at the di-allelic *Cpk*-1 locus (alleles *1* and *2*) in two subsamples of brown trout. *11* and *12* phenotypes which cannot be unambiguously classified are designated "*1*-".

Sample	Number of individuals	Phenotype			
		11	*12*	*22*	*1-*
Subsample 1	41	—	—	2	39
Subsample 2	54	44	9	1	—
Total sample	95	44	9	3	39

In subsample 1 only the *22* homozygote could be unambiguously identified, whereas the *11* and *12* genotypes were indistinguishable and had to be lumped (*1*-).

Different standard procedures for estimating the population allele frequency are conceivable in this situation, none of which is entirely satisfactory. First, all the individuals may be lumped and the frequency of the *1*-allele estimated as (*e.g.*, Hedrick, 1983)

$$\hat{P}=1-\sqrt{c/n} \tag{1}$$

with the sampling variance

$$V(\hat{P})=\frac{1-(1-P)^2}{4n} \tag{2}$$

where c is the number of *22* homozygotes in a sample of n individuals. For the total data set ($n=95$) these estimators yield $\hat{P}=0{\cdot}822$ and $V(\hat{P})=0{\cdot}00255$. Although the information about P contained in the *11* and *12* phenotypes of subsample 2 is ignored, this approach may appear attractive because it makes use of all the individuals sampled. On the other hand, the maximum-likelihood estimator (1) is unbiased only for large sample sizes (Elandt-Johnson, 1971). The bias stems from the fact that the fraction of *22* genotypes (c/n) is an unbiased estimate of P^2, whereas the expectation (E) of a squared variable (x^2) is $E(x^2)=E^2(x)+V(x)$, (*e.g.*, Elandt-Johnson, 1971) so that $E(x^2)>E^2(x)$ when $V(x)>0$. Thus, formula (1) tends to overestimate P, and the bias may be considerable for population allele frequencies above, say 0·7 or 0·8, even for moderately large sample sizes (fig. 1).

Second, we may choose to ignore subsample 1 and estimate P from direct "counting" of the alleles in subsample 2 only, using the estimators (*e.g.*, Hedrick, 1983)

$$\hat{P}=\frac{2a+b}{2n} \tag{3}$$

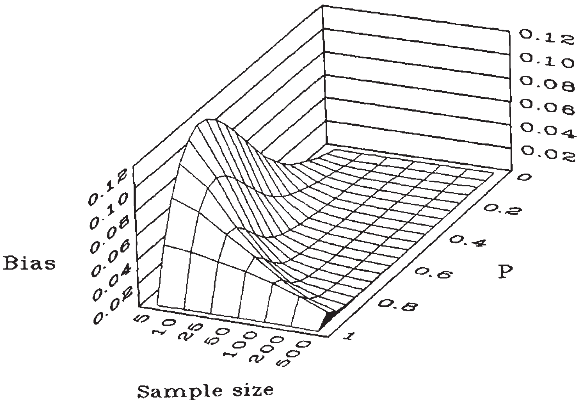

Figure 1 Bias ($E(\hat{P})-P$) of the estimator $\hat{P}=1-\sqrt{c/n}$ at a di-allelic autosomal locus (formula (1)). The expected values of $\hat{P}$ ($E(\hat{P})$) were calculated from the binomial distribution of dominant and recessive phenotypes for different sample sizes and population allele frequencies (P).

and

$$V(\hat{P})=\frac{P(1-P)}{2n} \tag{4}$$

where a and b are the number of *11* and *12* genotypes, respectively. For subsample 2 ($n=54$) we obtain $\hat{P}=0{\cdot}898$ and $V(\hat{P})=0{\cdot}00085$. This approach is attractive because (3) is an unbiased estimator of P, and in this particular example we also obtain a smaller variance than previously. However, the information about P contained in subsample 1 is disregarded.

Third, using (1) and (3) separately we may estimate P independently in each of the subsamples and produce a weighted average. Different weighing factors may be applied, but it is not clear which one is to be generally preferred. Weighing by the inverse of the sampling variance may produce problems if the estimated sampling variance is zero for either subsample. If weighing by sample size it is not clear whether the number of individuals (54) or the number of genes (108) is to be preferred for subsample 2.

The difficulty in obtaining an appropriate estimate is still more obvious if individuals cannot be identified as belonging to either of the subsamples 1 or 2, *i.e.*, if individuals exhibiting "dominant" and "co-dominant" expression are mixed in a single sample, as for the total of table 1.

DERIVATIONS

Consider an autosomal locus that segregates for two alleles (*1* and *2*) in a randomly mating population. The *22* genotype can always be identified. Among the individuals of the *11* and *12* genotypes a fraction D cannot be told apart (designated "*1-*"), whereas the genotype of the others (a fraction $1-D$) can be unambiguously identified (the probability D being the same for the *11* and *12* genotypes). The different phenotypes and their expected numbers in a sample of n individuals are given in table 2.

Estimation of allele frequency

Following the standard procedures for maximum-likelihood estimation (*e.g.*, Fisher, 1958; Elandt-Johnson, 1971) we choose as an estimate the value of P that maximizes the probability of the observed numbers. Given the expectations in table 2, this probability is

$$\begin{aligned} Prob = & \frac{n!}{a!b!c!d!}\{D[P^2+2P(1-P)]\}^d \\ & \times[(1-D)P^2]^a[(1-D)2P(1-P)]^b \\ & \times(1-P)^{2c}. \end{aligned}$$

The logarithm of this probability function attains its maximum for the same values of P and D as the probability function itself, so we simplify the computations by taking the logarithm (ln), obtaining

$$\begin{aligned} L &= \ln(Prob) \\ &= d\ln D+(a+b)\ln(1-D)+d\ln(2-P) \\ &\quad +(2a+b+d)\ln P+(b+2c)\ln(1-P)+K \end{aligned}$$

where K is a constant. The values of P and D that maximize L are the values for which the derivatives of L equal zero, such that

$$\frac{\delta L}{\delta P}=\frac{-d}{2-P}+\frac{2a+b+d}{P}-\frac{(b+2c)}{1-P}=0,$$

and

$$\frac{\delta L}{\delta D}=\frac{d}{D}-\frac{(a+b)}{1-D}=0.$$

Solving for P and D provides the estimators

$$\hat{P}=\frac{2n+a+b/2-\sqrt{(2n-a-b/2)^2-4nd}}{2n}, \tag{5}$$

and

$$\hat{D}=\frac{d}{n-c}. \tag{6}$$

Note that (5) simplifies to (1) and (3) when applied to cases of complete dominance ($a=b=0$) and co-dominance ($d=0$), respectively.

Sampling variance

The theoretical sampling variances are obtained through the second derivatives of the likelihood function L, which are

$$\frac{\delta^2 L}{\delta P^2}=-\frac{d}{(2-P)^2}-\frac{(2a+b+d)}{P^2}-\frac{(b+2c)}{(1-P)^2},$$

$$\frac{\delta^2 L}{\delta D^2}=-\frac{d}{D^2}-\frac{(a+b)}{(1-D)^2},$$

and

$$\frac{\delta^2 L}{\delta P \delta D}=0.$$

Substitution for the maximum-likelihood estimates in the variance-covariance matrix yields

$$V(\hat{P})=-\frac{1}{\delta^2 L/\delta P^2}=\frac{(2-P)P(1-P)}{4n-2n(1+D)P}, \tag{7}$$

$$V(\hat{D})=-\frac{1}{\delta^2 L/\delta D^2}=\frac{D(1-D)}{n-c}, \tag{8}$$

and

$$\operatorname{Cov}(\hat{D},\hat{P})=0. \tag{9}$$

Note that D and P are uncorrelated, and that (7) simplifies to (2) and (4) in the case of complete dominance ($D=1$) and co-dominance ($D=0$), respectively.

Table 2 Observed and expected number of phenotypes at an autosomal di-allelic locus (alleles *1* and *2*) in a sample of n individuals ($n=a+b+c+d$). *11* and *12* phenotypes which cannot be unambiguously classified are designated "*1-*". P is the frequency of the occasionally "dominant" *1*-allele, and D is the fraction of individuals expressing the dominant phenotype relative to all individuals carrying the *1-* allele.

Phenotype	*11*	*12*	*22*	*1-*
Observed	a	b	c	d
Expected	$n(1-D)P^2$	$n(1-D)2P(1-P)$	$n(1-P)^2$	$nD[P^2+2P(1-P)]$

Table 3 Mean ($\bar{p}$) and variance (s_p^2, multiplied by 10^5) of allele frequency estimates (p) obtained by computer simulation. P = population frequency of the occasionally dominant allele; n_1 and n_2 are the sizes of two independent subsamples 1 and 2 which are characterized by dominant and co-dominant gene expression, respectively ($n_1 + n_2 = 50$). Direct estimates were obtained by applying formulae (5) and (1) to the total sample ($n_1 + n_2$) and formula (3) to subsample 2 only (n_2); weighted estimates were obtained through weighing by the inverse of the variances, by n_1 and n_2, and by n_1 and $2n_2$. Each simulation was based on 10,000 runs.

			Direct estimates (sample size)						Weighted estimates (weighing factors)					
	Sample sizes		Formula (5) ($n_1 + n_2$)		Formula (1) ($n_1 + n_2$)		Formula (3) (n_2)		Inverse variances		Sample size (n_1 and n_2)		Sample size (n_1 and $2n_2$)	
P	n_1	n_2	$\bar{p}$	(s_p^2)	$\bar{p}$	(s_p^2)	$\bar{p}$	(s_p^2)	$\bar{p}$	(s_p^2)	$\bar{p}$	(s_p^2)	$\bar{p}$	(s_p^2)
0·95	25	25	0·951	(85)	0·983	(221)	0·950	(96)	0·953	(78)	0·969	(82)	0·963	(68)
0·95	40	10	0·954	(157)	0·983	(221)	0·950	(232)	0·960	(142)	0·977	(157)	0·973	(128)
0·95	10	40	0·950	(57)	0·982	(230)	0·950	(58)	0·951	(54)	0·958	(48)	0·954	(49)
0·90	25	25	0·901	(149)	0·938	(617)	0·900	(178)	0·908	(140)	0·927	(239)	0·918	(166)
0·90	40	10	0·906	(259)	0·939	(610)	0·901	(442)	0·918	(260)	0·935	(448)	0·929	(348)
0·90	10	40	0·901	(107)	0·940	(606)	0·900	(113)	0·903	(103)	0·914	(107)	0·908	(100)
0·70	25	25	0·702	(291)	0·709	(539)	0·700	(427)	0·706	(344)	0·713	(457)	0·708	(346)
0·70	40	10	0·704	(379)	0·709	(538)	0·699	(1050)	0·711	(458)	0·710	(505)	0·708	(438)
0·70	10	40	0·699	(236)	0·707	(529)	0·698	(262)	0·699	(272)	0·713	(320)	0·707	(255)

BIAS AND EFFICIENCY

An expression describing the efficiency of (5) relative to other estimators can be obtained from comparisons of the theoretical (large-sample) variances (*e.g.*, Elandt-Johnson, 1971). For instance, the efficiency of (5) relative to (1) (*i.e.*, estimating P from the frequency of recessive homozygotes) is given by the ratio of (2) and (7). This ratio is $\{(2-(1+D)P)/(2-2P)\}$, which is always larger than unity (*i.e.*, (5) is more efficient) except for $D = 1$, in which case (7) and (2) are identical (see above). Similarly, the efficiency of (5) relative to (3) applied to subsample 2 only (where the alleles are co-dominantly expressed) is $\{(n/n_2)(2-(1+D)P)/(2-P)\}$. This quantity is also always larger than unity except for the case of $n = n_2$ (implying $D = 1$), when (7) reduces to (4).

Strictly, the above efficiency comparisons are valid only for large samples, and bias is not taken into account. Of course, there is no point in striving at an estimator with a small sample variance if bias is unduly large; likewise, a minor bias can be accepted if a significant reduction of variance is gained.

We used computer simulations to evaluate bias and efficiency of our estimator (5) relative to a number of alternative ones, noting that only (1) and (5) can be applied in cases where individuals cannot be identified as belonging to a particular subsample (cf. table 1). Table 3 presents some simulation results for P values for which a large bias may be expected (fig. 1) and for the sample size $n = 50$. Each simulation comprised 10,000 runs. When weighing by the inverse of the sampling variances occasional runs gave $V(\hat{P}) = 0$; in those particular runs formula (1) was applied to avoid division by zero.

The simulation results indicate that our estimator (5) is to be preferred in most situations. First, it generally has a smaller sample variance than other estimators (table 3). Second, bias is always small or negligible, and the improvement is particularly conspicuous when comparing with (1) or with any of the weighted estimates (which all include the application of (1)). Of course, disregarding subsample 1 (dominant expression) and applying formula (3) to subsample 2 only (co-dominant expression) always yields an unbiased estimate; however, the smaller sample size results in an increase of the variance that generally cannot justify the reduction of bias.

Acknowledgments We thank Drs R. Chakraborty and F. Utter for valuable comments on an earlier version of the manuscript. Financial support was obtained from the Swedish Natural Science Research Council.

REFERENCES

ALLENDORF, F. W., STÅHL, G. AND RYMAN, N. 1984. Silencing of duplicate genes: a null allele polymorphism for lactate dehydrogenase in brown trout (*Salmo trutta*). *Mol. Biol. Evol.*, *1*, 238-248.

CAVALLI-SFORZA, L. L. AND BODMER, W. F. 1971. *The Genetics of Human Populations.* Freeman and Company, San Francisco.

CROSS, T. F. AND KING, J. 1983. Genetic effects of hatchery rearing in Atlantic salmon. *Aquaculture, 33,* 33-40.

ELANDT-JOHNSON, R. 1971. *Probability Models and Statistical Methods in Genetics.* John Wiley and Sons, New York.

FISHER, R. A. 1958. *Statistical Methods for Research Workers.* Oliver and Boyd, London.

HEDRICK, P. W. 1983. *Genetics of Populations.* Van Nostrand Reinhold, New York.

MOURANT, A. E., KOPEĆ, A. D. AND DOMANIEWSKA-SOBCZAK, K. 1976. *The Distribution of the Human Blood Groups and Other Polymorphisms.* Oxford University Press, London.

RYMAN, N. AND STÅHL, G. 1981. Genetic perspectives of the identification and conservation of Scandinavian stocks of fish. *Can. J. Fish. Aquat. Sci., 38,* 1562-1575.

STÅHL, G. 1983. Differences in the amount and distribution of genetic variation between natural populations and hatchery stocks of Atlantic salmon. *Aquaculture, 33,* 23-32.

TAGGART, J., FERGUSON, A. AND MASON, F. M. 1981. Genetic variation in Irish populations of brown trout (*Salmo trutta* L.): electrophoretic analysis of allozymes. *Comp. Biochem. Physiol., 69B,* 393-412.

UTTER, F., AEBERSOLD, P. AND WINANS, G. 1987. Interpreting genetic variation detected by electrophoresis. In Ryman, N. and Utter, F. (eds) *Population Genetics and Fishery Management,* University of Washington Press, Seattle.