Theor Appl Genet (1995) 91: 1284–1287

H. M. Alexander · N. A. Slade · R. Gomulkiewicz

A Bayesian approach to the inference of diploid genotypes using haploid genotypes

Received: 15 May 1995 / Accepted: 2 June 1995

Abstract Morris and Spieth (1978) described a method of calculating unbiased estimates of diploid genotype frequencies given information on the genotypes of haploid cells derived from diploid individuals. They concluded that three haploids per diploid would minimize sampling variance of genotype frequencies, given a fixed total number of haploids examined. If the identity of individual diploid genotypes is needed, Morris and Spieth (1978) stated that more haploids should be collected per diploid. We extend this work by showing from a Bayesian perspective that the probability of misclassification of individuals depends not only on the number of haploids sampled, but also on the genetic structure of the population since misclassification error will increase as the frequency of heterozygotes increases. Since information on the genetic structure (allele frequencies, inbreeding coefficient) of a population is rarely known prior to the initiation of an empirical study, the usefulness of our Bayesian approach is in experimental design, by revealing the magnitude of possible misclassification errors given a particular choice of number of haploids.

Key words Bayes's rule · Haploid · Experimental design · Genetic structure

Communicated by P. M. A. Tigerstedt

H. M. Alexander (✉)
Department of Botany and Department of Systematics and Ecology, University of Kansas, Lawrence, KS 66045-2106, USA

N. A. Slade
Department of Systematics and Ecology and Natural History Museum, University of Kansas, Lawrence, KS 66045-2106 and KS 66045-2454, respectively, USA

R. Gomulkiewicz
Department of Systematics and Ecology, University of Kansas, Lawrence, KS 66045-2106, USA

Introduction

A variety of technical and genetic factors associated with diploid tissue can make genetic analysis of haploid cells more tractable (Morris and Spieth 1978; Lewis and Snow 1992; Lynch and Milligan 1994). Since the derivation of haploid cells from diploid cells by meiosis is common to nearly all organisms, several workers have considered ways to infer diploid genotype frequencies from the derived haploid cells. This statistical problem was addressed by Tigerstedt (1973) and Bergmann (1973, as cited by Morris and Spieth 1978), who sought to infer the diploid genotype frequencies of conifers by genetic analysis of haploid female gametophytes. Morris and Spieth (1978) presented a formal discussion of the calculation of unbiased estimators of diploid genotype and allele frequencies based on data from haploid gametophytes. Additionally, they determined the optimum sampling design for a fixed experimental effort (i.e., given that a certain number of haploid genotypes can be determined, how many trees and gametophytes per tree should be sampled?). Morris and Spieth's 1978 paper has been heavily cited (32 citations through September 1994), primarily by conifer researchers who used their approach to design sampling strategies for analysis of genetic variation (e.g., Millar 1983; Cheliak et al. 1984; Strauss 1986; Yeh and Morgan 1987). The importance of the Morris and Spieth (1978) approach extends beyond conifers; their paper has been considered by workers concerned with experimental design and mating system theory (Shaw and Brown 1982; Cheliak et al. 1983; Ross 1986), and by researchers focusing on other types of organisms (Shoen 1979; Lobo and Krieger 1992).

The Morris and Spieth (1978) approach for estimating unbiased genotype frequencies focuses on the frequency of directly detected heterozygotes (i.e., cases where an array of haploids sampled from one diploid individual includes alleles of two types,=polytypic haploid array), with an adjustment for the frequency of heterozygotes that are not detected (i.e., by chance, a haploid array includes only alleles of one type,=monotypic haploid array). Specifi-

cally, the unbiased estimate of the frequency of the heterozygote genotype A_iA_j (or, equivalently, the estimated probability that a sampled individual has genotype A_iA_j) is:

$$P(A_iA_j) = \frac{n_{ij}}{\left(1-\frac{1}{2^{k-1}}\right)n}$$

and the unbiased estimate of the frequency of the homozygote genotype A_iA_i (or, equivalently, the estimated probability that a sampled individual has genotype A_iA_i) is:

$$P(A_iA_i) = \frac{n_{ii}}{n} - \frac{1}{2^k}\sum_{j\neq i}\frac{n_{ij}}{\left(1-\frac{1}{2^{k-1}}\right)n}$$

where k is the number of haploids sampled per diploid, n is the total number of diploids, and n_{ii} and n_{ij} are the number of monotypic and polytypic haploid arrays, respectively.

The probability of failing to directly detect a heterozygote is $1/2^{k-1}$. Morris and Spieth (1978) considered the allocation of a fixed number of haploids within and among diploids to minimize variation in estimated genotype frequencies. They recommended sampling three haploids per diploid, but also stated that more haploids per diploid should be used if the identification of individuals is the objective. Using Bayesian inference, we extend their work by considering how the probability of sampling a heterozygote, which itself depends on the genetic structure of the population, affects the probability of misclassification of individuals.

Bayesian approach

Using Bayesian inference (Schmitt 1969), we can calculate P(genotype | data), i.e., the probability of having sampled a particular diploid genotype *given* our data (data=observation of a specific haploid array). We, after all, know only the haploid genotypes and are unsure about the diploid genotype. Note that we use the term "probability" in the Bayesian sense of "our confidence or degree of belief in an individual having a particular genotype"; "probability" is incorrect in a strict sense because a given individual either has or doesn't have a specific genotype. Bayes's rule in this case is:

$$P(\text{genotype} \mid \text{data}) = \frac{P(\text{data} \mid \text{genotype})\ P(\text{genotype})}{P(\text{data})} \quad (1)$$

If one obtains a polytypic array of haploid genotypes, the probability of the diploid genotype being heterozygous is one. However, if one screens k haploid products and all have the A_i allele (i.e., data= k A_i's), there are two possible diploid genotypes, A_iA_i and A_iA_j. The conditional probability of the diploid genotype being a homozygote is:

$$P(A_iA_i \mid k\,A_i\text{'s}) = \frac{P(k\,A_i\text{'s} \mid A_iA_i)\,P(A_iA_i)}{P(k\,A_i\text{'s} \mid A_iA_i)\,P(A_iA_i) + \sum_j P(k\,A_i\text{'s} \mid A_iA_j)\,P(A_iA_j)}$$

while the conditional probability of it being a heterozygote is:

$$P(A_iA_j \mid k\,A_i\text{'s}) = \frac{P(k\,A_i\text{'s} \mid A_iA_j)\,P(A_iA_j)}{P(k\,A_i\text{'s} \mid A_iA_i)\,P(A_iA_i) + \sum_j P(k\,A_i\text{'s} \mid A_iA_j)\,P(A_iA_j)} \quad (2)$$

Consider now, for simplicity, a case with two alleles at a locus. If we were to classify the diploid individuals that produce monotypic arrays as homozygous, Eq. 2 describes the probability that we will misclassify (i.e., classify an individual as a homozygote when it is a heterozygote). It is convenient to reparameterize these equations by defining genotype frequencies in terms of the allele frequency (p) and the inbreeding coefficient (f) (Hartl and Clark 1989):

$$P(A_1A_1) = (1-f)\,p^2 + fp$$
$$P(A_1A_2) = (1-f)\,2p(1-p)$$
$$P(A_2A_2) = (1-f)\,(1-p)^2 + f(1-p)$$

Reparameterized in this way, Eq. 2 becomes:

$$P(A_1A_2 \mid k\,A_1\text{'s}) = \frac{\frac{1}{2^{k-1}}(1-f)(1-p)}{\frac{1}{2^{k-1}} + \left(1-\frac{1}{2^{k-1}}\right)(f+p-fp)} \quad (3)$$

Equation 3 shows that the probability of misclassification depends on p and f as well as k (see Fig. 1). There are two components to this misclassification error. First, the degree of misclassification will depend on the probability of sampling a heterozygote, which in turn depends on the genetic structure of the population as defined by p and f. For example, with high levels of inbreeding and a high fre-

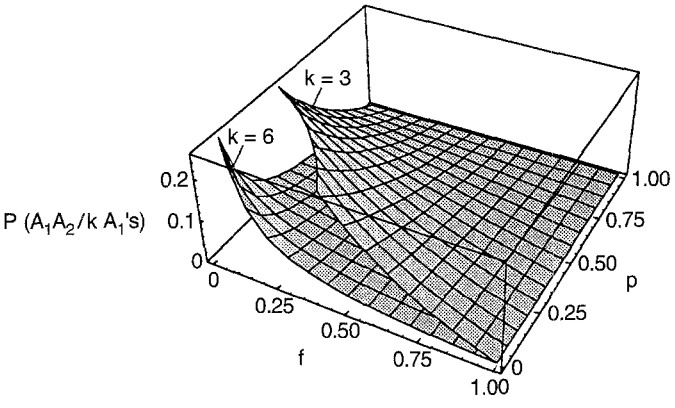

Fig. 1 The probability that an individual yielding a monotypic A_1 haploid array is heterozygous for different frequencies of the A_1 allele (*p*) and different values of the inbreeding coefficient (*f*). Surfaces are indicated for cases where k=3 and k=6; note that the vertical axis was truncated to allow comparison of the surfaces. As *p* and *f* approach 0, the expected frequency of heterozygotes approaches 1.0

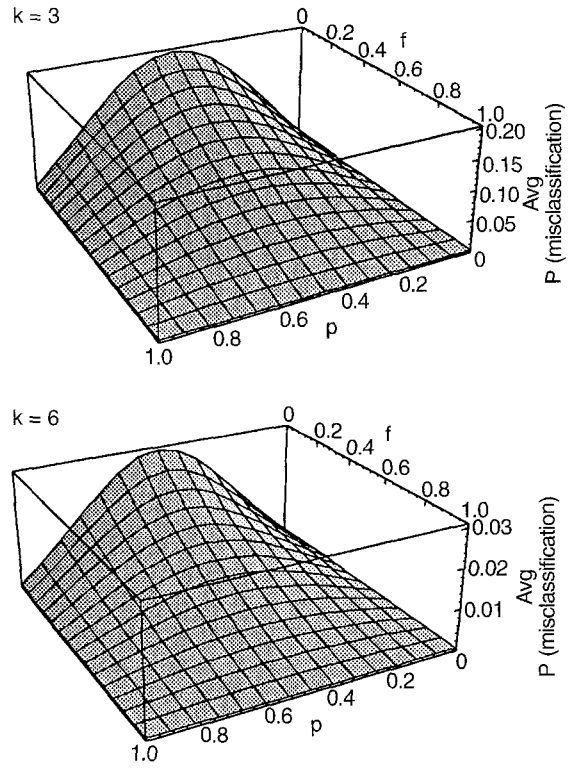

Fig. 2 The average probability of misclassification (=avg. expected frequencies of heterozygotes in a collection of A_1 and A_2 monotypic haploid arrays, see Eq. 4) for different frequencies of the A_1 allele (p) and different values of the inbreeding coefficient (f). Surfaces are shown for k=3 and k=6; note the differences in the vertical scale for the two graphs

quency of the A_1 allele, nearly all individuals in the population are A_1A_1 and thus the chance of misclassifying an individual when only A_1 types are recovered is low. Second, if a heterozygote is sampled, the probability of misclassification will depend on the number of haploids screened (k), as emphasized by Morris and Spieth (1978).

Taken at face value, Eq. 3 suggests that if A_1 is rare ($p \approx 0$) and the population is in Hardy-Weinberg equilibrium (f=0), the probability of misclassification of heterozygotes given monotypic A_1 arrays approaches one (Fig. 1). (Note that the magnitude of the misclassification is sample size-dependent). However, in a population with small f and p, monotypic A_1 arrays rarely will be encountered. A better representation of the possible impact of misclassification error is provided by Fig. 2 for the cases where k=3 and 6. The vertical axis is an average probability of misclassification error, which is the average of the probabilities of misclassifying A_1 and A_2 monotypic arrays, weighted by their expected frequencies:

$$\text{Avg P (misclassification)} = \frac{P(A_1A_2 \mid k\,A_1\text{'s})\left[\frac{pq(1-f)}{2^k} + p^2(1-f) + pf\right] + P(A_1A_2 \mid k\,A_2\text{'s})\left[\frac{pq(1-f)}{2^k} + q^2(1-f) + qf\right]}{\frac{pq(1-f)}{2^{k-1}} + p^2(1-f) + pf + q^2(1-f) + qf} \quad (4)$$

where q=1−p.

Figure 2 reveals that the highest average errors are obtained with Hardy-Weinberg equilibrium and equally frequent alleles. Note that the absolute magnitude of misclassification is greatly reduced with the increased number of haploids sampled.

Discussion

Identities of individual genotypes are necessary for studies of spatial substructuring of populations (Heywood 1991), paternity assignment (Meagher and Thompson 1987), and analysis of the relationship between fitness of an organism and its genotype (for example, examination of the relationship between fitness and heterozygosity: Ennos 1990). Although the actual genotype of individuals that produce monotypic arrays can never be known with complete certainty, one can reduce the chance of misclassification of individuals by increasing the number of haploids sampled per individual. The Bayesian approach quantifies the probabilities of correct identification and illustrates the importance of the genetic structure of the population to classifying individuals. However, one rarely knows the genetic structure of the sampled population prior to the study. The main usefulness of the Bayesian perspective is, therefore, not in actual calculation of misclassification error in an empirical study, but in experimental design. Consider, for example, the graphs for k=3 and k=6 in Fig. 2. For k=3, the actual range of misclassification errors (0–0.2) is an order of magnitude larger than the actual range of errors (0–0.03) for the k=6 case. Thus, since it is difficult to have a priori knowledge of the mating system (estimates of outcrossing rates for the same plant species can differ greatly: Waller 1986; Holtsford and Ellstrand 1989), it is useful to know the range of Bayesian misclassification errors associated with a particular value of k so that the maximum possible error in classification of individuals is understood. Such maximum errors will occur when p=0.5 and f=0 (Fig. 2); in this case, Eq. 4 can be simplified to:

$$\text{Avg P (misclassification)} = \frac{\frac{1}{2^{k-1}}}{1 + \frac{1}{2^{k-1}}} \quad (5)$$

Given a decision on the maximum possible misclassification error that can be tolerated (m), one can solve for the minimum number of haploids that must be sampled (k) by rearranging Eq. 5:

$$k \geq 1 + \frac{\log\left(\frac{1-m}{m}\right)}{\log 2}$$

In addition to its value in experimental design, our approach can be used in a retrospective way. For example, in work by Tigerstedt (1973), spatial maps of genotypes were created with individuals with monotypic arrays assigned as homozygotes and individuals with polytypic arrays considered as heterozygotes; a proportion of the individuals that were identified as homozygous will, however, actually be heterozygous. The Morris and Spieth (1978) approach can be used to calculate unbiased genotype frequencies for the population. Such estimates of p and f (the latter obtained using Wright's $F_{IS} = 1 - h_o/h_e$; where h_o and h_e refer, respectively, to the observed and Hardy-Weinberg expected frequency of heterozygotes), would allow one to estimate where a specific population is on the Bayesian surface (Fig. 2), allowing a more fine-tuned evaluation of the degree of misclassification in the analysis.

Acknowledgements This work was partially supported by NSF grant DEB-9119409 to H. M. A. and P. V. Oudemans. We appreciate discussion of these issues by J. Heywood, R. Shaw, and M. Whitlock. M. A. and R. G. acknowledge support from the University of Kansas General Research Fund.

References

Cheliak WM, Morgan K, Strobeck C, Yeh FCH, Dancik BP (1983) Estimation of mating system parameters in plant populations using the EM algorithm. Theor Appl Genet 65:157–161

Cheliak WM, Dancik BP, Yeh FCH, Strobeck C, Morgan K (1984) Segregation of allozymes in megagametophytes of viable seed from a natural population of jack pine, *Pinus banksiana* Lamb. Theor Appl Genet 69:145–151

Ennos RA (1990) Detection and measurement of selection: genetic and ecological approaches. In: Brown AHD, Clegg MT, Kahler AL, Weir BS (eds) plant population genetics, breeding, and genetic resources. Sinauer Associates, Sunderland, Mass, pp 200–214

Hartl DL, Clark AG (1989) Principles of population genetics. 2nd edn. Sinauer Associates, Sunderland, Mass.

Heywood JS (1991) Spatial analysis of genetic variation in plant populations. Annu Rev Ecol Syst 22:335–355

Holtsford TP, Ellstrand NC (1989) Variation in outcrossing rate and population genetic structure of *Clarkia tembloriensis* (Onagraceae). Theor Appl Genet 78:480–488

Lewis PO, Snow AA (1992) Deterministic paternity exclusion using RAPD markers. Mol Ecol 1:155–160

Lobo JA, Krieger H (1992) Maximum likelihood estimates of gene frequencies and racial admixture in *Apis mellifera* L. (Africanized honeybees). Heredity 68:441–448

Lynch M, Milligan BG (1994) Analysis of population genetic structure with RAPD markers. Mol Ecol 3:91–99

Meagher TR, Thompson E (1987) Analysis of parentage for naturally established seedlings of *Chamaelirium luteum* (Liliaceae). Ecology 68:803–812

Millar CI (1983) A step cline in *Pinus muricata*. Evolution 37: 311–319

Morris RW, Spieth PT (1978) Sampling strategies for using female gametophytes to estimate heterozygosity in conifers. Theor Appl Genet 51:217–222

Ross HA (1986) Estimation of mating system parameters in plant populations using marker loci with null alleles. Theor Appl Genet 72:322–327

Schmitt SA (1969) Measuring uncertainty: an elementary introduction to Bayesian statistics. Addison-Wesley, Reading , Mass.

Schoen DJ (1979) An angiosperm analogue to megagametophyte analysis. J Theor Biol 79: 543–546

Shaw DV, Brown AHD (1982) Optimum number of marker loci for estimating outcrossing rates in plant populations. Theor Appl Genet 61:321–325

Strauss SH (1986) Heterosis at allozyme loci under inbreeding and crossbreeding in *Pinus attenuata*. Genetics 113:115–134

Tigerstedt PMA (1973) Studies on isozyme variation in marginal and central populations of *Picea abies*. Hereditas 75:47–60

Waller DM (1986) Is there disruptive selection for self-fertilization? Am Nat 128:421–426

Yeh FC, Morgan K (1987) Mating system and multilocus associations in a natural population of *Pseudotsuga menziesii* (Mirb.) Franco. Theor Appl Genet 73:799–808