Conservation Genetics (2006)
DOI 10.1007/s10592-006-9134-9

Maximum likelihood estimation of the frequency of null alleles at microsatellite loci

Steven T. Kalinowski* & Mark L. Taper
*Department of Ecology, Montana State University, 310 Lewis Hall, Bozeman, MT, 59717, USA (*Corresponding author: Phone: +1-406-994-3232; Fax: +1-406-994-3190; E-mail: skalinowski@montana.edu)*

Received 27 May 2005; accepted 1 February 2006

Key words: allele, estimation, frequency, microsatellite, null

Abstract

We review three methods for estimating the frequency of null alleles at codominant loci (such as microsatellite loci) and present a new maximum likelihood approach. Computer simulations show that the maximum likelihood estimator has a smaller root mean squared error than previous estimators.

Introduction

Microsatellite loci are the markers of choice for estimating evolutionary relationships between populations and genealogical relationships between individuals. When using microsatellite loci, however, care must be taken that "null alleles" do not distort conclusions. Null alleles are alleles that consistently do not amplify during PCR, and thus are not detected when individuals are genotyped (see Dakin and Avis 2004 for a review). If, for example, A_n is a null allele, an individual with the genotype A_iA_n will be indistinguishable from a A_iA_i homozygote. If an individual is homozygous for a null allele, genotyping will fail.

Null alleles can distort several types of conservation genetic research. Null alleles decrease the apparent heterozygosity in a sample, thus interfering with efforts to measure genetic diversity in populations. They lead to over estimates of the frequencies of non-null alleles, thereby interfering with estimates of population structure. They tend to decrease estimates of relatedness. Last, and perhaps most important, null alleles can interfere with parentage identification (e.g. Dakin and Avis 2004). Consider a cross between a dam with genotype A_iA_i and a sire with genotype A_jA_n. With these parental genotypes, there is a 50% chance that an offspring will have the genotype A_iA_n, and thus appear to not be an offspring of its actual sire.

There are currently three methods for estimating the frequency of null alleles from co-dominant genotypes, such as those at microsatellite loci (Chakraborty et al. 1992; Brookfield 1996; Summers and Amos 1997). All three methods have been described as "likelihood" approaches, but they have not been discussed thoroughly enough for a critical reader to choose among them. In this note, we briefly review each method and present a new maximum likelihood estimator for the frequency of a null allele that uses more information than the three methods currently available.

Previous estimators

Chakraborty et al. (1992) provided the first estimator of the frequency of a null allele in microsatellite data. Chakraborty et al.'s estimate, $\hat{p}_{n(\text{Chakraborty})}$, is calculated from the difference between the heterozygosity observed in a sample,

H_{obs}, and the heterozygosity expected from the allele frequencies observed in the sample, H_{exp}

$$\hat{p}_{n(\text{Chakraborty})} = \frac{H_{exp} - H_{obs}}{H_{exp} + H_{obs}} \quad (1)$$

This estimate is reasonably accurate (see below), but its statistical basis is not clear. It appears to be a method-of-moments estimator, but this has not been established. Brookfield (1996) claimed Equation (1) is a maximum likelihood estimator, but we have been unable to confirm this without assuming that the apparent frequency of the *i*-th allele in a sample, $\tilde{p}_i$, is equal to the actual frequency of the *i*-th allele divided by $(1-p_n)$ – an assumption which precludes Equation (1) from being a maximum likelihood estimator.

In any case, Brookfield (1996) identified a drawback to Equation (1) that is probably more important. Equation (1) is calculated from the number of heterozygotes observed and expected in a sample. No reference is made to the number of individuals for which genotyping failed. Chakraborty et al. deliberately did not include missing data as an observation because genotyping may fail either because an individual is homozygous for a null allele or because there was an unrelated technical problem (degraded DNA, contamination etc). Brookfield (1996), however, pointed out that if there are no missing genotypes in a sample, this information should be used in the estimation of the frequency of a null allele, and derived the estimator, $\hat{p}_{n(\text{Brookfield})}$,

$$\hat{p}_{n(\text{Brookfield})} = \frac{H_{exp} - H_{obs}}{H_{exp} + 1}. \quad (2)$$

Brookfield framed the derivation of this formula in a maximum likelihood context, but explicitly assumed that $\tilde{p}_i$, is equal to $p_i/(1-p_n)$. Therefore, the theoretical basis of Equation (2) is uncertain. It does not seem to be a maximum likelihood estimator.

Summers and Amos (1997) provide the third estimator of the frequency of null alleles. They describe their method as a "likelihood approach," but do not define the likelihood or show how it is maximized. Their method works well with simulated data (see below). There are, however, two reasons to suspect that a more informative estimator can be derived. First, none of the Chakraborty, the Brookfield, or the Summers/Amos estimators are calculated from the actual genotype counts in the data. Each method summarizes the data before estimating the frequency of a null allele. For example, Equation (1) treats all homozygotes as equal, when in fact a homozygote for a rare allele is stronger evidence for a null allele than a homozygote for a common allele. Second, none of the three methods take full advantage of the number of individuals for which there is no data. If, for example, a large sample has only one individual with missing data, this observation should be used when estimating the frequency of a null allele because it helps set a bound on how high the frequency of the allele is likely to be.

A maximum likelihood estimator

We propose a maximum likelihood estimator of the frequency of null alleles in a sample that may or may not have missing data. The approach is a modest extension of the method used to estimate the frequency of the *O* allele at the *ABO* blood protein locus (Ceppellini et al. 1955; see Weir 1996, Chapter 2, for a review). As we mentioned above, missing data may be caused by a A_nA_n homozygote or because of some other reason (degraded DNA, PCR failure etc). Therefore, a likelihood model for estimating the frequency of null alleles needs a parameter to incorporate this type of non-null missing data. Let β represent the probability that genotyping fails at a locus for a reason other than the locus being homozygous for a null allele. Let p_n represent the frequency of the null allele and let p_i represent the frequency of the *i*-th visible allele. The probability that the genotype A_iA_i is observed in a sample is $(p_i^2 + 2p_ip_n)(1-\beta)$. The probability that the genotype A_iA_j is observed in a sample (where *j* is a visible allele distinct from *i*) is $2p_i\,p_j\,(1-\beta)$. And lastly, the probability that genotyping fails to produce a genotype is $\beta + {p_n}^2\,(1-\beta)$. Each of these three probabilities assumes that the genotypes in the population are in Hardy–Weinberg proportions, and that the probability of genotyping failure, β, is independent of genotype.

We calculate the likelihood from the genotype counts observed in a sample. Let n_{ii} represent the number of samples apparently homozygous for the *i*-th allele (out of *k* total visible alleles). Let n_{ij} represent he number of $A_i\,A_j$ heterozygotes, and let n_{mm} represent the number of individuals that were not genotyped successfully (i.e. n_{mm} is the number

of individuals lacking any visible alleles). The likelihood is calculated

$$L = \left\{ \prod_{i=1}^{k} \left[\left(p_i^2 + 2p_i p_n \right) (1-\beta) \right]^{n_{ii}} \right\} \times \left\{ \prod_{i \neq j}^{k} \left[\left(2p_i p_j \right) (1-\beta) \right]^{n_{ij}} \right\} \times \left\{ \left[\beta + p_n^2 (1-\beta) \right]^{n_{mm}} \right\}. \quad (3)$$

We have been unable to find expressions for p_1, p_2 ... p_k, p_n, and β that maximize Equation (3). Numerical optimization, however, is easily accomplished with any program with an optimization routine (e.g. Matlab, Mathcad, Microsoft Excel). The EM algorithm (Dempster 1977) is also convenient. The relevant iterative equations for EM optimization are

$$\hat{p}_i' = \frac{1}{2N} \left[2n_{ii} \left(\frac{\hat{p}_i^2}{\hat{p}_i^2 + 2\hat{p}_i \hat{p}_n} \right) + n_{ii} \left(\frac{2\hat{p}_i \hat{p}_n}{\hat{p}_i^2 + 2\hat{p}_i \hat{p}_n} \right) + \sum_{j \neq i}^{k} n_{ij} + 2n_{mm} \left(\frac{\hat{\beta}}{\hat{\beta} + \hat{p}_n^2 \left(1 - \hat{\beta} \right)} \right) \hat{p}_i \right]$$

$$\hat{p}_n' = \frac{1}{2N} \left[\sum_{i=1}^{k} n_{ii} \left(\frac{2\hat{p}_i \hat{p}_n}{\hat{p}_i^2 + 2\hat{p}_i \hat{p}_n} \right) + 2n_{mm} \left(\frac{\left(1 - \hat{\beta} \right) \hat{p}_n^2}{\hat{\beta} + \hat{p}_n^2 \left(1 - \hat{\beta} \right)} \right) \right] + 2n_{mm} \left(\frac{\hat{\beta}}{\hat{\beta} + \hat{p}_n^2 \left(1 - \hat{\beta} \right)} \right) \hat{p}_n \quad (4)$$

$$\hat{\beta}' = \frac{1}{N} \left[n_{mm} \left(\frac{\hat{\beta}}{\hat{\beta} + \hat{p}_n^2 \left(1 - \hat{\beta} \right)} \right) \right].$$

The equations for $\hat{p}_i'$ and $\hat{p}_n'$ can be simplified slightly

$$\hat{p}_i' = \frac{1}{2N} \left[\frac{2n_{ii}(\hat{p}_i + \hat{p}_n)}{\hat{p}_i + 2\hat{p}_n} + \sum_{j \neq i}^{k} n_{ij} + 2n_{mm} \left(\frac{\hat{\beta}}{\hat{\beta} + \hat{p}_n^2 - \hat{\beta}\hat{p}_n^2} \right) \hat{p}_i \right] \quad (5)$$

$$\hat{p}_n' = \frac{1}{2N} \left[\sum_{i=1}^{k} n_{ii} \left(\frac{2\hat{p}_n}{\hat{p}_i + 2\hat{p}_n} \right) + 2n_{mm} \left(\frac{\hat{p}_n^2 - \hat{\beta}\hat{p}_n^2 + \hat{\beta}\hat{p}_n}{\hat{\beta} + \hat{p}_n^2 - \hat{\beta}\hat{p}_n^2} \right) \right]$$

for easier computation. $\hat{p}_{n(\text{Chakraborty})}$ (Equation 1) provides a convenient starting point for iteration. The EM algorithm is guaranteed to climb the likelihood surface, but may get stuck on a sub-optimal peak. Therefore, several starting points should be tried. Broken stick random numbers make good starting points because they sum to one (as allele frequencies must) and are distributed uniformly in multi-dimensional space (Devroye 1986).

Methods

We used computer simulation to compare the accuracy of our maximum likelihood estimates with the estimates of Chakraborty et al. (1992) and Summers and Amos (1997). The method of Brookfield was not tested because it only can be used when there are no missing genotypes in a sample. We simulated genotypes by first simulating allele frequencies in a population and then drawing alleles from these frequencies. Allele frequencies in a population were simulated with broken stick random numbers; individual genotypes were simulated by drawing alleles with replacement from the population allele frequencies. One of the alleles in the population was chosen to be a null allele. Individuals with genotype A_iA_n were converted to A_iA_i. Individuals with genotype A_nA_n were considered missing data. In addition, a small proportion of the genotyping was selected to fail for other reasons at rate β. The frequency of the null allele was then estimated using the EM algorithm described above, the method of Chakraborty et al. (1992), and the method of Summers and Amos (1997). Fifty thousand simulated data sets were generated, and the root mean squared error, RMSE, was calculated from difference between the actual frequency of the null allele (p_n) and the estimated frequency ($\hat{p}_n$)

$$\text{RMSE} = \sqrt{\text{Avg}(p_n - \hat{p}_n)^2} \quad (6)$$

where Avg() indicates that the arithmetic mean was taken across 50,000 estimates. In addition, the root mean square error was calculated for visible alleles

$$\mathrm{RMSE} = \sqrt{\mathrm{Avg}\left(\frac{\sum_{i=1}^{k}(p_i - \hat{p}_i)^2}{k}\right)} \qquad (7)$$

where k is the number of visible alleles at a locus.

Three parameters were varied during these simulations: the number of visible alleles ($k = 2$, 4, and 8), the genotyping failure rate ($\beta = 0.0$, 0.2, and 0.05), and the total sample size ($N = 50$ and 100). Broken stick random numbers were used to simulate allele frequencies, so the parametric frequency of the null allele varied in each iteration of the simulation.

Results and discussion

The maximum likelihood estimators had the lowest root mean squared error in all scenarios tested (Table 1). This included maximum likelihood estimates of the frequency of null alleles as well as estimates of the frequency of visible alleles (Table 1). Maximum likelihood estimates outperformed the other methods most at loci with few alleles. For example, when there were two visible alleles, the maximum likelihood estimator of p_n had a RMSE that was approximately 40% less than the estimators of Chakraborty et al. and Summers and Amos (Table 1). Even larger improvements were obtained for estimates of the frequency of visible alleles. Maximum likelihood estimates of the frequencies of visible alleles had a RMSE that was up to 70% lower than estimator of Summers and Amos (1997) and up to 95% lower than the estimator of Chakraborty et al. (1992).

Maximum likelihood estimates of β were reasonably accurate (results not shown). They were most accurate (RMSE ≈ 0.005) when sample size was large and loci had many alleles. They were least accurate (RMSE ≈ 0.06) when sample size was small and loci had few alleles.

These results lead us to recommend our estimator in place of previous formulations. A Windows based computer program, *ML-NullFreq* is available to perform the necessary calculations. It

Table 1. Root mean squared error for estimates of the frequency of a null allele and visible alleles in simulated data having a total of N individuals, k visible alleles per locus, and a genotyping failure rate of β

β	k	N	Null allele			Visible alleles		
			ML	CHAC	SA	ML	CHAC	SA
0	2	50	0.117	0.202	0.218	0.062	0.663	0.162
0	2	200	0.067	0.112	0.173	0.036	0.908	0.124
0	4	50	0.058	0.078	0.080	0.020	0.276	0.033
0	4	200	0.029	0.038	0.040	0.010	0.252	0.016
0	8	50	0.035	0.044	0.045	0.008	0.067	0.011
0	8	200	0.018	0.022	0.023	0.004	0.059	0.006
0.02	2	50	0.117	0.199	0.219	0.062	0.662	0.164
0.02	2	200	0.072	0.112	0.174	0.039	0.908	0.123
0.02	4	50	0.061	0.079	0.081	0.021	0.283	0.035
0.02	4	200	0.032	0.039	0.041	0.011	0.256	0.015
0.02	8	50	0.038	0.044	0.045	0.009	0.066	0.011
0.02	8	200	0.020	0.023	0.023	0.004	0.060	0.006
0.05	2	50	0.124	0.205	0.221	0.065	0.644	0.164
0.05	2	200	0.073	0.115	0.175	0.039	0.906	0.124
0.05	4	50	0.063	0.079	0.080	0.022	0.279	0.032
0.05	4	200	0.035	0.039	0.041	0.012	0.244	0.015
0.05	8	50	0.041	0.045	0.046	0.009	0.068	0.011
0.05	8	200	0.022	0.023	0.023	0.005	0.061	0.006

Results are presented for three methods: maximum likelihood (ML), the heterozygote deficiency estimator of Chakraborty et al. (CHAC), and the estimator of Summers and Amos (SA).

is available at http://www.montana.edu/kalinowski, and uses GENEPOP data files (Raymond and Rousset 1995) for input.

An estimate of the frequency of a null allele will seldom be useful unless that estimate can be used during data analysis (e.g. while estimating F_{ST} or paternity). Accommodating null alleles in such calculations is straightforward. For example, Wagner et al. (2006) and Kalinowski et al. (2006) have developed a statistical method and software for accommodating null alleles while estimating genealogical relationship. Additional work will be necessary to provide similar accommodations for other analyses.

Acknowledgements

This research has been supported by NSF grant DEB-0415932 (MLT).

References

Brookfield JFY (1996) A simple new method for estimating null allele frequency from heterozygote deficiency. *Mol. Ecol.*, **5**, 4534–4555.

Ceppellini R, Siniscalco M, Smith CAB (1955) The estimation of gene frequencies in a randomly mating population. *Ann. Hum. Genet.*, **20**, 97–115.

Chakraborty R, De Andrade M, Daiger SP, Budowle B (1992) Apparent heterozygote deficiencies observed in DNA typing and their implications in forensic applications. *Ann. Hum. Genet.*, **56**, 455–457.

Dakin EE, Avis JC (2004) Microsatellite null alleles in parentage analysis. *Heredity*, **93**, 504–509.

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood estimation from incomplete data via the EM algorithm. *J. R. Stat. Soc., B*, **39**, 1–38.

Devroye L (1986) *Non-uniform Random Variate Generation*, Springer-Verlag, New York.

Kalinowski ST, Wagner AP, Taper ML (2006) ML-Relate: Software for estimating relatedness and relationship from multilocus genotypes. *Mol. Ecol. Notes* In press.

Raymond M, Rousset F (1995) GENEPOP (version 1.2): Population genetics software for exact tests and ecumenicism. *J. Hered.*, **86**, 248–249.

Summers K, Amos W (1997) Behavioral, ecological, and molecular genetic analyses of reproductive strategies in the Amazonian dart-poison frog, *Dendrobates ventrimaculatus*. *Behav. Ecol.*, **8**, 260–267.

Wagner AP, Creel S, Kalinowski ST (2006) Maximum likelihood estimation of relatedness and relationship using microsatellite loci with null alleles. *Heredity* Accepted pending minor revision.

Weir BS (1996) *Genetic Data Analysis II*, Sunderland, Massachusetts, Sinauer Associates Inc.