

1. The problem

You want to analyse a large multilocus microsatellite dataset. Genotyping errors are inevitable, and if you fail to take them into account you are likely to reach erroneous conclusions^{1,7,8}. If you follow standard procedures, you will re-genotype a subset (e.g. 10%) of your samples and count the mismatches (e.g. Figure 1), allowing you to estimate the error rate across the whole data set².

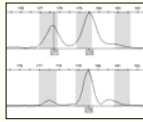


Figure 1 Allelic dropout or false allele?

A limitation of this approach is that it can't resolve two classes of microsatellite genotype errors, *allelic dropout* and *false alleles*, that affect data analyses in fundamentally different ways (Box A). Analysis methods that account for the rates of these two errors (ϵ_1 and ϵ_2) can make correct inferences even from data with high error rates³. Unfortunately, ϵ_1 and ϵ_2 can't be measured by simply comparing repeat genotypes, because you can't trust either genotype. Would you classify the repeat genotype shown in Figure 1 as allelic dropout in the second genotype or a false allele in the first?

Currently the only way to separate these two error classes is to compare the error-affected data with 'perfect' reference genotypes, obtained either from higher-quality sample tissue (possibly unavailable) or multiple repeats⁴ (costly and labour intensive).

We have developed a method that estimates the two error rates without the need for reference data. The method is implemented in a Windows program, PEDANT, which can be downloaded from www.stats.gla.ac.uk/~paulj/pedant.html.

Box A Why we need to estimate two error rates

Imagine you are doing a paternity analysis using microsatellites with high allelic dropout (ϵ_1) but very low false allele rate (ϵ_2) – a common scenario⁵. Allelic dropout affects homozygotes and heterozygotes in the following ways:

Homozygote: True AA \rightarrow Typed as AA ✓
Heterozygote: True AB \rightarrow Typed as AA✗ or BB✗

With frequent allelic dropouts and rare false alleles, heterozygous genotypes are unlikely to have been affected by error, and so are more reliable than homozygous genotypes:

Candidate father: AA AB
Offspring: CC CD
Exclusion confidence: Low High

But you need to estimate both ϵ_1 and ϵ_2 to make these judgements. Analysis methods that use a composite error rate (e.g. CERVENUS⁶) will have too much confidence in homozygotes and too little in heterozygotes.

Box B Estimating ϵ_1 and ϵ_2 using maximum likelihood (ML)

An allele drops out with probability ϵ_1 , and is read as a false allele with probability ϵ_2 . For each genotype, the probabilities of no dropouts is $p_0 = (1 - \epsilon_1)^2$ and the probability of one given allele dropping out is $p_1 = \epsilon_1(1 - \epsilon_1)$. Double dropouts are not counted as they are indistinguishable from other causes of genotype failure. Similarly, the probabilities of 0, 1 and 2 false alleles occurring are $f_0 = (1 - \epsilon_2)^2$, $f_1 = \epsilon_2(1 - \epsilon_2)$ and $f_2 = \epsilon_2^2$. Under simplifying assumptions, including HWE, the expected frequencies ($P_{1...7}$) of the seven repeat genotype classes are:

$$P_1 = P(AAAA|H, \epsilon_1, \epsilon_2) = (1 - H) [p_0^2 f_0^2 + 4p_0 p_1 f_1^2 + 4p_1 p_0 f_1 f_2 + 4p_1^2 f_2^2 + 8p_0^2 f_1 f_2 + 4p_1^2 f_2^2] + H [2p_1^2 f_0^2 + 4p_1^2 f_1 f_2 + 2p_1^2 f_2^2]$$

$$P_2 = P(ABAB|H, \epsilon_1, \epsilon_2) = H [p_0^2 f_0^2]$$

$$P_3 = P(AAAB|H, \epsilon_1, \epsilon_2) = (1 - H) [4p_0^2 p_1 f_1 f_2 + 8p_0 p_1 p_1 f_1^2 + 8p_0 p_1 p_1 f_2^2] + H [4p_0 p_1 p_1 f_0^2 + 8p_0 p_1 p_1 f_1 f_2 + 4p_0 p_1 p_1 f_2^2]$$

$$P_4 = P(AABB|H, \epsilon_1, \epsilon_2) = (1 - H) [4p_0^2 p_1 f_1 f_2 + 8p_0 p_1 p_1 f_1^2 + 8p_0 p_1 p_1 f_2^2] + H [2p_1^2 f_0^2 + 12p_1^2 f_1 f_2 + 14p_1^2 f_2^2 + 12p_1^2 f_1 f_2]$$

$$P_5 = P(ABAB|H, \epsilon_1, \epsilon_2) = (1 - H) [4p_0^2 p_1 f_1^2] + H [4p_0^2 p_1 f_1^2]$$

$$P_6 = P(ABAC|H, \epsilon_1, \epsilon_2) = (1 - H) [2p_0^2 f_1 f_2 + 8p_0 p_1 f_1^2 + 4p_0 p_1 f_1 f_2] + H [8p_0 p_1 f_1 f_2 + 12p_0 p_1 f_1^2 + 8p_0 p_1 f_1 f_2]$$

$$P_7 = P(ABCB|H, \epsilon_1, \epsilon_2) = H [2p_0^2 f_1 f_2 + 2p_0^2 f_1^2]$$

These frequencies must be adjusted to sum to one, giving the expected frequency of class i , $F_i = P_i / \sum P_i$. The likelihood of the data (the observed category counts $X_{1...7}$, which sum to n) is

$$L(X|H, \epsilon_1, \epsilon_2) = \frac{n!}{X_1! X_2! \dots X_7!} F_1^{X_1} F_2^{X_2} \dots F_7^{X_7}$$

allowing the ML estimates of ϵ_1 and ϵ_2 to be obtained (Figure 2).

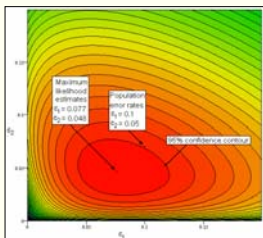


Figure 2 A log-likelihood contour plot showing the maximum-likelihood estimates of ϵ_1 and ϵ_2 for a simulated data set of 150 repeat genotypes. The population parameters in the simulated data were $\epsilon_1=0.1$ and $\epsilon_2=0.05$, and $H_i = 0.75$. The sample error rates were 0.070 and 0.048.

2. The solution

How the method works in theory...

If we can't trust either repeat genotype, how can we distinguish between allelic dropout and false alleles? The frequencies of classes of mismatch differ depending on the two error rates. For example, an excessive number of AB.AC-type mismatches intuitively suggests a high false allele rate. Using the same logic, we can calculate the expected frequencies for all seven possible classes of repeated genotypes:

1. AA.AA
2. AB.AB
3. AA.AB
4. AA.BB
5. AB.AC
6. AA.BC
7. AB.CD

The expected frequencies of each class is a function of ϵ_1 , ϵ_2 and H_i , the expected heterozygosity. The likelihood of each pair of errors is then a function of the expected and observed frequencies of the seven classes (Box B).

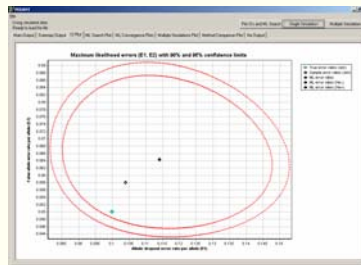


Figure 3 A plot showing error rates with 90% and 95% confidence limits estimated from a simulated data set (population parameters: $\epsilon_1 = 0.1$, $\epsilon_2 = 0.05$, $H_i = 0.75$, $n = 250$), produced by the program PEDANT. The population (light blue circle) and sample error rates (yellow diamond) are also shown.

...and in practice

PEDANT estimates ϵ_1 and ϵ_2 with confidence intervals from duplicated microsatellite genotypes. We have validated PEDANT using simulated and real data.

- In the simulations PEDANT estimates error rates accurately (e.g. Figure 3) and, remarkably, with greater precision than achievable using reference samples (Figure 4).
- When confronted with real data, PEDANT performs well overall, although there is evidence that our error model will fit some data better than others (Figure 5).

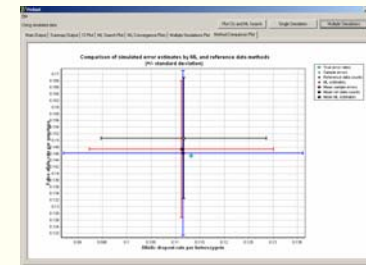


Figure 4 Comparison between ML (red) and reference data error estimates (blue). Mean error estimates (\pm SD) across 500 simulated data sets are shown ($\epsilon_1 = 0.06$, $\epsilon_2 = 0.08$, $H_i = 0.8$, $n = 200$), as are the population (light blue circle) and sample error rates (yellow diamond, black SD bars). Although the ref. data method has the advantage of knowledge of the true genotypes, it has higher variance because only half the number of genotyping errors are sampled. N.B. Here the error rates have been converted from per-allele to per-genotype.

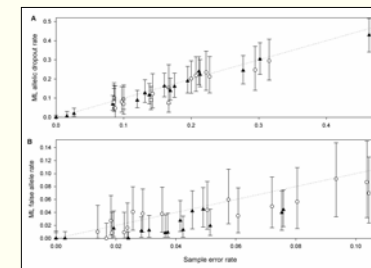


Figure 5 Performance of PEDANT in estimating allelic dropout rate per heterozygote (A) and false allele rate per genotype (B) from real data with known error rates. Two duplicated microsatellite data sets were analyzed: 149–182 red fox teeth genotyped at 16 loci (\blacktriangle), and 72–121 Ethiopian wolf faecal samples genotyped at 17 loci (\circ). The dotted line shows the equality between ML and sample error rates. Error bars show 95% confidence intervals for the ML estimates. Although PEDANT tends to underestimate the false allele rate in the fox genotypes (the error model will fit some data better than others), overall it performs well.

Additional features

- PEDANT's simulations help you decide how many samples to replicate to get a desired level of precision in your error rate estimates.
- PEDANT estimates H_i from the total data set and investigates the (usually negligible) effect of variance in the H_i estimate on the error rate estimates.