Esimation of Allele Frequencies from// Small Pool PCR (SPPCR)

Barry W. Brown

Michael J. Siciliano

1 Why Small Pool PCR?

Small pool PCR is used to quantitate the frequency of microsatellite alleles. In traditional large pool PCR, samples of thousands of alleles are amplified and their sizes measured. Because of the stutter bands near large peaks, it is difficult to detect a small frequency allele near a large peak and estimation of the frequency of the rare allele is almost impossible. The top panel of Figure 1 illustrates the problem; by viewing this panel only, it is not clear whether the peak at a size of 19 is real or not and it is certainly not clear how to estimate the frequency of allele 19.

In small pool PCR, small amounts, typically two alleles, of DNA are amplified. The small number of alleles distinguishes peaks, as shown by the middle and bottom panels of Figure 1. Many replicate examinations are performed to assure detection of rare alleles, and the proportion of replicates in which a particular allele is seen allows estimation of the frequency of the allele. This estimation is the subject of the current work.

2 Terminology

We consider a single locus. Progenitor alleles are identified by genotyping. Our primary interest is in the estimation of the frequency of mutant alleles.

An examination of a sample consists of the amplification of one or more amounts of DNA; the results from each amount amplified is termed a **run**. Replicate samples of each amplification are conducted, we term the replicate a **well**. The information obtained from a well consists of the identity of the alleles seen in it, for example, well 3 contained alleles 102 and 104.

A considerable reduction of the data is possible with no loss of information relevent to the estimation of allele frequencies. Only the number of wells in which an allele was seen in each run need be recorded; the combinations of alleles in wells contributes no additional information about allele frequencies.

The operational unit of the amount of amplified DNA is the allele equivalent (AE): one AE is that amount of DNA that, when amplified, produces *on average* one identifiable allele. Estimating the AE in one experimenter unit of initial DNA, e.g., picograms, is an important part of the analysis. c, denotes this number.

Alleles are labelled arbitrarily, usually by either the number of short DNA sequence repeats or the number of base pairs in the allele. The subscript i will be used to denote a particular allele.

To summarize, the information used in an analysis is:

- The experimental design: The number of runs, the amount of DNA amplified, and the number of wells in each run.
- The progenitor alleles.
- For each run, the number of wells in which each allele was seen.

Results of the analysis include:

- The calibration quantity, c – the AE in one experimenter DNA unit.

  Frequently, the amount of DNA that the experimenter amplifies at several loci is determined by the results at a different locus. Amplification may differ from locus to locus, so it is important to calibrate each separately.
- The frequency of each allele in the sample. The frequency of allele i is denoted f_i.
- The total mutation frequency.
- The variablitity of the estimates. This is needed, for example, to compare the mutation frequencies of two samples.

3 Overview of Statistical Methods

The analysis of SPPCR data involves maximum likelihood estimation. Descriptions of likelihood methods and the reasons for their use are found in standard texts on statistical theory, for example, Chapter 18 of Stuart (1991). Here, we attempt to emphasize the intuitive nature of the methods used.

The steps in the development of methods for analysing SPPCR data are:

1. **Develop a statistical model.** The model provides the probability of any outcome given the values of the parameters of the model, c and the f_i. The probability of the outcome of an experiment considered as a function of these parameters is termed the likelihood; the logarithm of this probability is the log-likelihood.

2. **Choose c and the f_i to maximize the log-likelihood.**

3. **Compute variances of the estimates.**

4 Statistical Model

The number of alleles in a particular well is distributed Poisson. We denote the DNA amount in run r in experimenter units by D_r. The mean number of alleles per well, AE, of this run is cD_r.

If there are N alleles in a well, the number of occurrences of each allele type is distributed according to the multinomial distribution with the probability that an allele is of type i being f_i.

The appendix shows that this distribution of number of alleles of a particular type, i, is Poisson; the mean of this Poisson distribution is $cD_r f_i$. The probability of n_i alleles of type i in a well is the same regardless of the number of alleles of a different type in the well – i.e., the distribution of different alleles are independent.

This result implies that knowledge of combination of alleles in a well provides no addition information over the number of wells in which each allele is seen. The probability of a combination is the product of the probabilities of the alleles in the combination; the particular combinations arising are purely due to chance.

The result also implies that that the mean number of alleles of different sizes can be estimated separately, ignoring those of other sizes. One-at-a-time estimation is a great computational simplification over simultaneous estimation.

4.1 The Likelihood

Let $\mu_i = cf_i$. The mean number of alleles labelled i in a well in run r is $D_r\mu_i$. The probability of *not* seeing allele i in a particular well in run r is the probability of no events in a Poisson process with this mean,

$$p_{ur} = \exp(-D_r\mu_i), \tag{1}$$

The probability of seeing allele i in a well is $p_{sr} = 1 - p_{ur}$. (u and s are mnemonic for 'unseen' and 'seen'.)

The probability of seening allele i in n_{sr} wells and of not seeing it in n_{ur} wells in a run is given by the binomial formula:

$$P_{ir} = \left(\begin{array}{c} n_{sr} + n_{ur} \\ n_{sr} \end{array} \right) p_{sr}^{n_{sr}} p_{ur}^{n_{ur}}. \tag{2}$$

This is the likelihood for allele i in this run.

It is traditional to work with the logarithm of the likelihood instead of the likelihood itself. Taking the logarithm frequently simplifies the calculation and the statistical theory tends to deal more naturally with the logarithmic values. The only operation which we perform on the log-likelihood is to find the value of μ_i that maximizes it. Consequently, we can omit the logarithm of the binomial coefficient from the log-likelihood, since it depends only on the data obtained and not on μ_i.

With this simplification, we can write the likelihood of seeing the i'th allele size in n_{sr} wells and not seeing it in n_{ur} wells for run r as

$$ll_{ir} = n_{sr}\log(p_{sr}) + n_{ur}\log(p_{ur}) \quad (3)$$

$$= n_{sr}\log(1-\exp(-D_r\mu_i)) - n_{ur}D_r\mu_i \quad (4)$$

where the last line fillows by replacing p_{sr} and p_{ur} from 1.

The total likelihood in i is the probability of P_ir over all runs r. Logarithms add over products so the total log-likelihood in i is

$$ll_i = \sum_r ll_{ir}.$$

For any one run, r, the estimation of μ_{ir} is straightforward. The maximum likelihood estimate of a binomial proportion of events is the observed proportion. Hence, the natural (and maximum likelihood) estimate of μ_i is obtained by solving

$$\hat{p_{ur}} = \frac{n_{ur}}{n_{sr}+n_{ur}} = \exp(-D_r\hat{\mu_i})$$

This yields the following estimate, $\hat{\mu_{ir}}$:

$$\hat{\mu_{ir}} = -\frac{\log(\hat{p_{ur}})}{D_r} \quad (5)$$

This solution can be verified by differentiating the log-likelihood and setting the derivative to zero.

If there are several runs, the likelihood must be maximized numerically. A good starting value for the maximization is the average over the runs of the $\hat{\mu_{ir}}$. We denote The maximum likelihood estimate of μ_i over the (possibly several) runs by $\hat{\mu_i}$.

5 A Problem

Suppose that there is only one run in an examination of a specimen and allele i was seen in every well. Then according to equation (5), the estimate of μ_i is infinite. A

value of infinity is absurd, because it implies that regardless of the dilution of the sample amplified, allele i will always be seen.

This problem persists even if the amount of DNA is small enough that there are wells in which i is not seen. The expectation (mean) of the estimate is

$$E(\hat{\mu_i}) = \sum_{u=0}^{N} \hat{\mu_i}(u) b(u, N, p_{ur})$$

where N is the number of wells in the run, u is the number of wells in which i is not seen. $b(u, N, p_{ur})$ is the binomial probability of not seeing the allele labelled i in precisely u of N wells when the probability of not seeing i in any one well is p_{ur}. Because $\hat{\mu_i}(0)$ is infinite and $b(0, N, p_{ur}) > 0$, the average is infinite. This in turn implies an infinite bias for the estimate since μ_i is finite. (The bias of an estimator is its expected value minus the true value.) The variance of the estimate will similarly be infinite. This problem is not resolved by having more than one run.

Theory provides no solution to this problem; any solution used will be *ad hoc*. Our solution: modify $\hat{\mu_i}(0)$ by increasing n_{ur} from 0 to 1/2 and correspondingly decrease n_{sr} by 1/2. If there are several runs and n_{ur} is 0 in all of them, only the value in the run with the largest amount of DNA amplified is modified.

6 Estimation of c. the f_i, and the total mutant frequency

The estimate of c is

$$\hat{c} = \sum_i \hat{\mu_i} \tag{6}$$

since $\sum_i f_i = 1$.

The estimate of f_i is thus

$$\hat{f_i} = \frac{\hat{\mu_i}}{\hat{c}} \tag{7}$$

and the estimate of the fraction of mutants is:

$$\hat{m} = \frac{\sum_k \hat{\mu_k}}{\sum_j \hat{\mu_j}}$$

The the last formula, the index j ranges over all alleles and k ranges over all except the progenitor alleles.

7 Estimates of the Variances

There are two methods for computing the variance of the estimates:

1. Asymptotic approximations. These approximations arise from the theory. Their accuracy improves with increases in the total number of wells. This method has two disadvantages. (1) It requires a bit of mathematical sophistication to derive the estimates. (2) The theory does not provide methods for determining when the number of wells is sufficiently large for these approximations to be useful.

2. **Simulation or bootstrap estimates.** New random data is generated from the original data and fit to obtain estimates of c and the f_i. The process is repeated a large number (e.g., 1000) times and the variance of the estimate is obtained from these replicates.

If the two methods disagree, we prefer the simulation method because it does not require a large number of wells for accuracy.

The simulation method does require more computation than does the asymptotic method. However, with a modern computer the generation and analysis of 1000 random replicates of the experiment requires only a fraction of a second.

The generation of new random data sets proceeds as follows: For each run, we know the number of wells and p_{sr} for allele i. The simulated value of n_{sr} for allele i is a random number from a binomial distribution. The number of trials in the binomial is the number of wells, and the probability of an event is p_{sr}.

8 Transformation of Data

One of the primary uses for SPPCR results is the comparison of mutation frequencies between specimens (normal versus tumor) or populations (genetic abnormality or not). The normal approximation to the binomial is frequently used to compare proportions. As the number of wells gets larger and larger, this approximation gets better and better. However, for any real experiment, the approximation can be poor.

Figure 2 (left panel) shows the distribution of 1000 random replicate estimates of a mutant frequency of 5%; the distribution is scaled from 0 to 1 to make it comparable to the rightmost panel. The distribution is notably skewed to the right; there are more values further from the mean on the right of the distribution than on the left.

The right panel shows the distribution when the arcsin transform is applied to each estimate. The arcsin transfromation of a proportion, m, is

$$t(m) = 2\arcsin(\sqrt{(m)})$$

and this transformation is frequently used to better approximate the normal distribution. The skew is noticibly less in the right panel than in the left; by the usual statistical measure of skewness, the left panel has a skewnes of 0.61, the rightmost panel of -0.18. The skewness of a symmetric distribution would be zero, so the transformation slightly over corrects in this case..

9 Example

Figure 1 shows some of the results of chromatograms of separation of the alleles of the trinucleotide (CTG) repeat, a microsatellite at the myotonic dystrophy locus (DMPK). The alleles were visualized by conducting PCR using primers that flanked the repeat. One of the primers was labeled with a fluor enabling the detection of

the amplified fragments with an ABI gel separation apparatus. The human tissue used was suspected of having some level of microsatellite instability. Twenty small pool PCR wells at an investigator estimated 1 g.e. (2 a.e.) were run. Of these, 16 contained the 5 repeat fragment, 14 contained the 20 repeat fragment, and 3 contained the mutant 19 repeat fragment.

The allele of size 5 was not seen in 4 of 20 wells, or a proportion of 0.2. From 5, we have

$$\hat{\mu_5} = -\frac{\log(0.2)}{2} = 0.8047 \tag{8}$$

The estimates of the other mu's are:

$$\hat{\mu_{19}} = 0.0812 \tag{9}$$

$$\hat{\mu_{20}} = 0.6019 \tag{10}$$

$$\tag{11}$$

From 6,

$$\hat{c} = 0.8047 + 0.0812 + 0.6019 = 1.4878$$

and finally from 7

$$\hat{f_5} = 0.8047/1.4878 = 0.5409 \tag{12}$$

$$\hat{f_{19}} = 0.0812/1.4878 = 0.0545 \tag{13}$$

$$\hat{f_{20}} = 0.6019/1.4878 = 0.4045 \tag{14}$$

$$\tag{15}$$

The 95% confidence limits on the mutant frequency (f_{19}) are $(0.0078, 0.1397)$. The mutant frequency is not determined with precision with only 20 wells.

10 Statistical Testing

This section discusses the most common statistical tests associated with SPPCR.

10.1 Comparing Mutant Frequencies Between Two Specimens

The two specimens on which mutant frequencies are compared might be normal and tumor tissue in the same individual; the usual comparison would be that of total mutants, however, the frequency of any particular mutant could be compared across the samples.

Let the estimates of the mean of the transformed frequencies of interest in the two specimens be F_1 and F_2 and let the corresponding estimated variances be V_1 and be V_2. Then, since $t(F_1)$ and $t(F_2)$ are approximately normal an appropriate statistic for assessing the significance of the difference between the two frequencies is:

$$Z = \frac{t(F_1) - t(F_2)}{\sqrt{(V_1 + V_2)}}$$

If the two frequencies are the same, then Z should be distributed as a unit normal, so a difference in absolute value of at least 1.96 is significant at the 0.5% level for a two-sided test.

10.2 Comparing Two Mutant Frequencies in a Single Specimen

The obvious method would be to use the normal approximation for the frequencies, f_i, taking as mean and standard deviation the mean and standard error of the estimates and the bootstrap standard error. This is incorrect since the f's are correlated due to their common dependence on the estimate of c.

The correct proceedure is to compare the μ_i using a normal approximation. The μ's are uncorrelated.

10.3 Comparing Mutant Frequencies Between Two Categories of Specimens

Categories are groups of samples identifiable by some criterion; for example, samples from individuals with cancer and others without cancer or those with some genetic abnormality and those without the same abnormality.

Our transformation, t, that makes the data more nearly normal in one sample, is of no help here because the transformed mean of individual frequencies is not necessarily near the mean of the transformed frequencies.

If we believed that the true frequency of mutation were the same in each individual in each category and that deviations were due to chance alone, then a Z statistic could be developed in analogy with the above one. This reasoning does not account for the individual to individual variation in frequency which is likely to be much greater than the chance variation of the experiment. Consequently, we recommend standard procedures for the comparison of two groups on the raw (not transformed) estimates of the mutant frequencies. In particular, we would run the t-test and the signed-rank tests and examine the data carefully if these two methods disagreed appreciably.

11 References

Monckton, D G, Wong,L-J C, Ashizawa T, Caskey C T, (1995). Somatic mosaicism, germline expansions, germline reversions and intergenerational reductions in myotonic dystrophy males: small pool PCR analyses. Hum. Mol. Genet. 4:1-8.

Stuart, A. and Ord, J. K. (1991) Kendall's Advanced Theory of Statistics. Oxford University Press, N.Y.

Appendix A: Demonstration that Alleles in a Well are Distributed as Independent Poisson Variates

An amount D of DNA is amplified in a well. The number of alleles in the well is distributed Poisson with mean cD, where c is the calibration constant. The probability of n alleles in a well is

$$\frac{(cD)^n}{n!}e^{-cD} \tag{16}$$

Suppose that there are three alleles labelled 1,2, and 3. The frequencies of the alleles are f_1, f_2, f_3, where the f's are positive and add to one. . Let n_1, n_2, n_3 be thre non-negative integers adding to n. Then given that there are n alleles in a well, the probability of n_1 of size 1, n_2 of size 2, and n_3 of size 3 is given by the multinomial distribution,

$$\frac{n!}{n_1!n_2!n_3!}f_1^{n_1}f_2^{n_2}f_3^{n_3} \tag{17}$$

We wish to show that the product of the two probabilities (16) and (17) is the same as the probability of (n_1, n_2, n_3) events from independent Poisson distributions with means (cDf_1, cDf_2, cDf_3). The latter probability is

$$\frac{(cDf_1)^{n_1}}{n_1!}e^{-(cDf_1)}\frac{(cDf_2)^{n_2}}{n_2!}e^{-(cDf_2)}\frac{(cDf_3)^{n_3}}{n_3!}e^{-(cDf_3)} \tag{18}$$

The factorial terms are obviously the same in the two expressions as are the powers of f_i. Because $f_1 + f_2 + f_3 = 1$, it follows that

$$e^{-(cDf_1)} + e^{-(cDf_2)} + e^{-(cDf_3)} = e^{-(cD)(f_1+f_2+f_3)} = e^{-cD}$$

finishing the demonstration.

The proof is the same for more than three alleles.

Appendix B: Asymptotic Variances

The asymptotic variance of $\hat{mui}$ is, from likelihood theory,

$$\mathrm{Var}(\mu_i) = -\left\{ \frac{\partial^2 ll}{\partial \mu_i}^2 \right\}^{-1}.$$

This, according to (4) and (1) is

$$\mathrm{Var}(\mu_i) = \sum_r \left(\frac{p_{sr}^2}{n_{sr} {D_r}^2 p_u} \right).$$

Since the μ_i are independent, their variances add, so

$$\mathrm{Var}(c) = \sum_i Var\mu_i.$$

To compute the variances of the f_i, we use the delta method also called the propogation of error For a description of this method see section 10.5 of Stuart and Ord, Vol. I (1993). Under certain conditions that are met in the cases to be discussed, the variance of a function, f, of two random variables is asymptotically approximated by:

$$\mathrm{Var}(f(X,Y)) \approx (\frac{\partial f}{\partial X})^2 \mathrm{Var}(X) + (\frac{\partial f}{\partial Y})^2 \mathrm{Var}(Y) + \frac{\partial^2 f}{\partial X \partial Y} \mathrm{Cov}(X,Y)$$

We need only the case in which

$$f(X,Y) = \frac{X}{X+Y}$$

and X and Y are independent so that $\mathrm{Cov}(X,Y) = 0$. The method yields the approximation,

$$\mathrm{Var}\left(\frac{X}{X+Y} \right) \approx \frac{Y^2 \mathrm{Var}(X) + X^2 \mathrm{Var}(Y)}{(X+Y)^4}$$

In calculating f_i, X is μ_i and Y is $\sum_j \mu_j$ over $j \neq i$.

In calculating the total proportion of mutants, X is $\sum_j \mu_j$ where j ranges over the labels of the mutants; Y is $\sum_p \mu_k$ where k ranges over the label(s) of the progenitor allele(s).

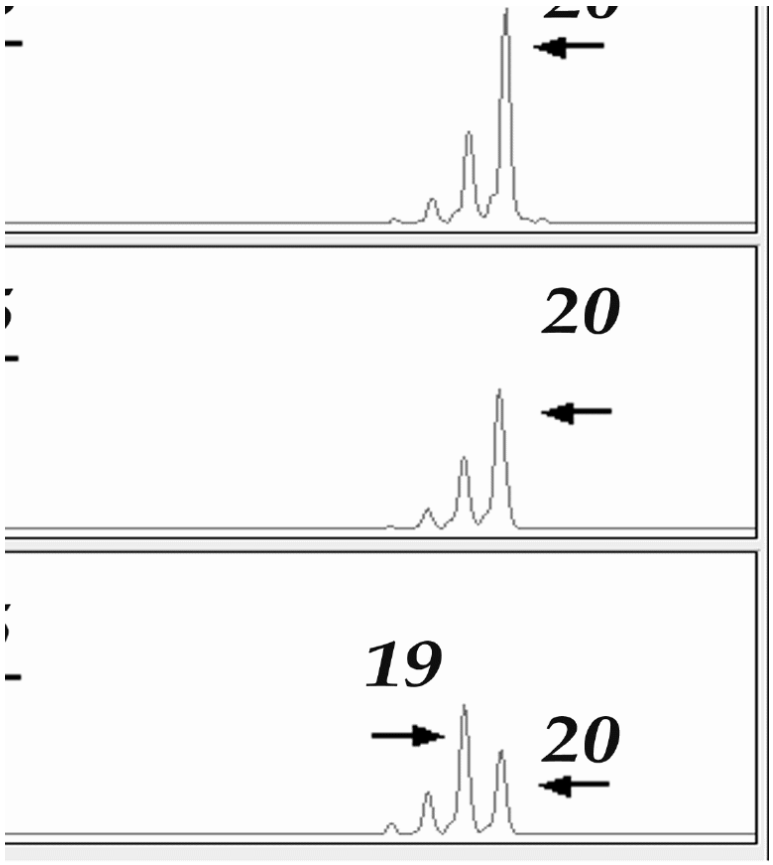

Fig. 1. The top panel shows a typical result from a chromatogram of a well containing at least 100 genome equivalents of DNA. The two "progenitor" fragments of this 5 repeat/20 repeat heterozygote are clearly visible with the attending smaller "satellite" bands leading single repeat units ahead. Assessing whether the satellite bands are real or only noise is difficult from an examination of this panel. The bottom two panels are two of 20 small pools in which the DNA had been diluted to contain only a single g.e. The middle panel contains only the expected 5 repeat and 20 repeat fragments. The bottom panel is one in which a 19 repeat fragment is clearly visible. The 19 repeat fragment is never unequivocably visualized in the "large" pool PCR reactions because it is present in low frequency and is consequently lost in the stutter band.

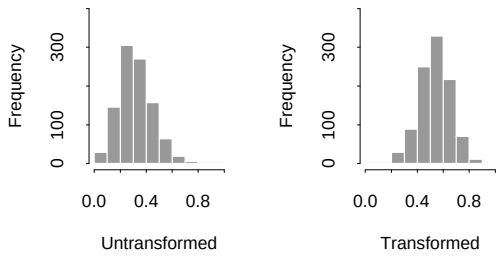

Fig. 2. Distribution of estimates of the mutation frequency in 1000 random replicates (left panel). Distribution after applying the arcsin transformation (right panel).