Segregation based metric for variant call QC

Richard Durbin

Version: June 30, 2014

Let us start by assuming we have N haploid samples, with equal average sequence coverage d, and that we have o_i observed copies of the non-reference variant in sample i, $(i = 1..N)$. Now let us consider two extreme hypotheses. In the first of these, H_1, the variant is a true variant represented in a subset M of the samples at some uniform mean depth (in principal d), and there is no noise. In the second, the null hypothesis H_0, the variant is false, and so all the non-reference observations are noise, and these are distributed randomly amongst the samples. In either case, the samples are a priori exchangeable, so the only information we can use from the observations is the distribution of in how many samples a variant was seen k times, $n_k = \sum_i \delta(o_i = k)$. We note that $\sum_k n_k = N$, and that under H_1 we must have $n_0 \geq N - M$.

Under H_0 the expected distribution of the o_i is multinomial with total counts $o = \sum_i o_i = \sum_k k\, n_k$ over N bins. This is very well approximated by independent identically distributed Poisson distributions $o_i \sim \mathrm{Poisson}(p)$ where $p = o/N$, so

$$P(o_i|H_0) = p^{o_i} e^{-p}/o_i! \ . \tag{1}$$

Under H_1, the o_i are spread across M bins with mean depth $q = o/M$. We have[1]

$$P(o_i|H_1, M) = \begin{cases} (M/N) q^{o_i} e^{-q}/o_i! & \text{if } o_i > 0 \\ (N-M)/N + (M/N)e^{-q} & \text{if } o_i = 0 \end{cases} \tag{2}$$

Let us define $Q(M)$ to be this value $P(0|H_1, M)$.

For fixed M we can calculate the log-likelihood ratio L_M of the distribu-

[1] $P(o_i|H_1, M) = P(i \text{ variant})P(o_i|i \text{ variant}) + P(i \text{ non-variant})P(o_i|i \text{ non-variant})$. We assume that non-carriers cannot show the variant (assumes no errors) and that the probability of getting zero occurrences for non-carriers is 1.

tion of the o_i under H_1 compared to under H_0,

$$L_M = \sum_i^N \log P(o_i|H_1) - \log P(o_i|H_0) \tag{3}$$

$$= \sum_i^N \begin{cases} \log(M/N)(q/p)^{o_i} e^{p-q} & \text{if } o_i > 0 \\ \log Q(M)/e^{-p} & \text{if } o_i = 0 \end{cases} \tag{4}$$

$$\tag{5}$$

$$= \sum_i^N \begin{cases} \log(N/M)^{o_i - 1} + (p-q) & \text{if } o_i > 0 \\ \log Q(M) + p & \text{if } o_i = 0 \end{cases} \tag{6}$$

$$\tag{7}$$

The only unknown variable in this equation is M which we approximate as $M = o/d$.

We need a diploid version of this. This will make little difference at low frequencies, but at higher frequencies it would incorporate Hardy-Weinberg: one would expect to get some double depths, some singles and some zero depth. Let M now be the number of variant alleles, and $f = M/2N$ be the allele frequency. The expected number of variant reads per-allele is $q = o/M$ in heterozygous individuals (RA) and $2q$ in homozygous non-reference individuals (AA) and assuming HWE we can write

$$P(o_i|H_1, M) = \begin{cases} \frac{1}{o_i!}\left[2f(1-f)q^{o_i}e^{-q} + f^2(2q)^{o_i}e^{-2q}\right] & \text{if } o_i > 0 \\ 2f(1-f)e^{-q} + f^2 e^{-2q} + (1-f)^2 & \text{if } o_i = 0 \end{cases} \tag{8}$$

We have very little power at very high frequencies near 1. We should ideally consider non-uniform depths di per sample. Without this, as long as the variation in depth is not extreme, I don't think we come to much harm; perhaps we lose a little power. For big variation in depth, there is a chance that H0 distributions will be misclassifed as H1 , because of errors clustering in particular samples at sufficient depth to appear like real calls.