Binomial Models of Reference Read Counts

Attila Gulyás-Kovács

February 19, 2016

1 Preliminaries

We have $i = 1, ..., I$ individuals and $g = 1, ..., G$ genes. I start with the simplest but unrealistic case that there is 0 or 1 SNP for each (i, g) pair for which i is heterozygous. Later I will generalize to allow more than 1 such SNPs. Let Y_{ig} be the read count for the reference allele and n_{ig} the total read count (by adding alternative alleles to Y_{ig}).

For each (i, g) we test $\mathcal{H}_0$ of biallelic expression against $\mathcal{H}_1$ of monoallelic expression. Let us denote $(i, g) \in \mathcal{H}_h$ that (i, g) conforms to $\mathcal{H}_h$ ($h = 0$ in the biallelic case and $h = 1$ in the monoallelic case).

2 Andy's model

In my understanding, in Andy's general model $\{Y_{ig}\}_{ig}$ are independent random variables and

$$Y_{ig} \sim \mathrm{Binom}(q_h \text{ or } 1 - q_h, n_{ig}) \text{ under } \mathcal{H}_h, \; h = 0, 1 \tag{1}$$

Let $p_h = \max(q_h, 1 - q_h)$. In Andy's specific model $p_0 = 1/2$ and $p_1 = 9/10$. To specify the model more completely, suppose $p_h = q_h$ with 1/2 probability *a priori*. Then for each (i, g) the probability mass function (p.m.f.) of Y_{ig}'s sampling distribution is

$$f(y|p_h, n_{ig}) = \frac{1}{2} \frac{n_{ig}!}{y!(n_{ig} - y)!} \left[p_h^y (1 - p)^{n_{ig} - y} + p^{n_{ig} - y} (1 - p)^y \right]. \tag{2}$$

Note that for homozygous (i, g) pairs $f(y = n_{ig}|p_h, n_{ig}) = 1$ for $h = 0, 1$ because all reads must surely come from a single variant regardless allelic exclusion.

For the observation $Y_{ig} = y_{ig}$ the p-value is

$$\sum_{y = y_{ig}}^{n_{ig}} f(y|p_0, n_{ig}). \tag{3}$$

Set classification threshold $n_{ig} t$ for any Y_{ig}. For instance, $t = 0.9$ means that we classify those pairs (i, g) for which at least 9/10 of the reads come from the reference allele. Let π_0 and π_1 be the fraction of (i, g) pairs when $(i, g) \in \mathcal{H}_0$ and when $(i, g) \in \mathcal{H}_1$, respectively. Note that $\pi_0 + \pi_1 = 1$.

The expected number of (i, g) pairs called monoallelic is then

$$\sum_{i,g} \pi_0 \overbrace{\sum_{y=t}^{n_{ig}} f(y|p_0, n_{ig})}^{\text{false positive rate}} + \pi_1 \overbrace{\sum_{y=t}^{n_{ig}} f(y|p_1, n_{ig})}^{\text{true positive rate}}. \tag{4}$$

So, given t, there are two ways to learn about the expected number of positives