Theoretical HET Sensitivity

Yossi Farjoun Jon Bloom

1 Introduction

Here we provide a theoretical estimate of the sensitivity to detect (snp) HET sites as a function of coverage distribution and base quality distribution. In particular, we will estimate the probability of calling a true HET site with confidence greater than a threshold T, e.g. 10^5 corresponding to Q50. The confidence is defined as the likelihood ratio between the HET model and the HOM_REF model for the data. The data consists of the allele (reference or alternate) and base quality of each base covering the HET site. Base quality is an integer q representing the phred-scaled probability of error, i.e. $-10 \log_{10} P(\text{error})$.

2 Simplifying Assumptions

To arrive at a simple model, we make the following assumptions:

1. The DNA is diploid (no cancer, no copy-number variation, no contamination).
2. At a HET site with genotype AB, the only possible calls are A and B (so an error switches between them).
3. There is no reference bias, e.g. variation does not affect mapping (essentially true for SNPs and possibly very short indels).
4. The coverage distribution $P(n)$ and base quality distribution $P(q)$ are known and statistically independent.

As noted throughout, if the joint distribution $P(n, q)$ is known then we can drop the independence assumption. This may be necessary if, for example, low coverage is significantly correlated with low base quality.

3 Generative model

The generative model of HET detection, a Bernoulli random variable, is as follows:

1. Draw depth n from the coverage distribution $P(n)$.

2. Draw the number of true alternate alleles $m \sim \text{binomial}(n, 0.5)$ covering the HET site.

3. Draw i.i.d. base qualities $q_1, \ldots q_n$ from the base quality distribution $P(q)$, where the first m are assigned to the alternates.

4. Compare the likelihood ratio of HET versus HOM_REF to the threshold T:

$$\frac{\binom{n}{m}\left(\frac{1}{2}\right)^n}{\binom{n}{m}\prod_{j=1}^{m} e_j \prod_{j=m+1}^{n}(1-e_j)} > T \tag{1}$$

   where $e_j = 10^{-q_j/10}$ is the probability of error.

So we can estimate HET sensitivity as the proportion of samples for which (1) is true (after cancelling the binomial coefficients). Note that one can remove the independence assumption by replacing $P(q)$ by $P(q \,|\, n)$ in Step 3.

Since we typically filter base quality at Q20, or $e_j < .01$, we can accelerate sampling with little loss in accuracy by approximating the event (1) by

$$\frac{1}{\prod_{j=1}^{m} e_j} > 2^n T. \tag{2}$$

Taking $\log_{10}$ of each side and rearranging terms yields

$$\sum_{j=1}^{m} q_j > C_n \tag{3}$$

where $C_n = 10\,(n \log_{10} 2 + \log_{10} T)$. Sampling now proceeds by drawing n, then m, then $q_1, \ldots, q_m$, and checking (3).

4 Analytic solution

For each $m \geq 0$, define the random variable Q_m to be the sum of m i.i.d. base qualities q_j drawn from $P(q)$. To estimate HET sensitivity without sampling, we write:

$$P\left(\sum_{j=1}^{m} q_j > C_n\right) = \sum_{n=0}^{\infty}\sum_{m=1}^{n} P\left(Q_m > C_n \,|\, n, m\right)\, P(m \,|\, n)\, P(n) \tag{4}$$

$$= \sum_{n=0}^{\infty} P(n) \sum_{m=1}^{n} P(m \,|\, n) \sum_{s=\lceil C_n \rceil}^{m q_{\max}} P(Q_m = s) \tag{5}$$

$$= \sum_{n=0}^{\infty} P(n) \sum_{m=1}^{n} P(m \,|\, n) \left(1 - \sum_{s=m q_{\min}}^{\lfloor C_n \rfloor} P(Q_m = s)\right) \tag{6}$$

where $q_{\min}$ (resp. $q_{\max}$) is the maximum (resp. minimum) possible base quality and $P(m \mid n) = \binom{n}{m} 2^{-n}$.

4.1 Computing the distribution of Q_m for small m

Let $B = q_{\max} - q_{\min} + 1$, the number of distinct base qualities. We can compute the distribution of Q_m for all $m \leq N$ in $\mathcal{O}(B^2N^2)$ operations by induction on m as follows. By definition Q_1 has the original base-quality distribution $P(q)$ and

$$P(Q_{m+1} = s) = \sum_{j=q_{\min}}^{q_{\max}} P(Q_1 = j)\, P(Q_m = s - j). \tag{7}$$

Here there are $(B-1)m + 1$ potential values of s, and B summands for each s, so the complexity of calculating Q_{m+1} from Q_m and Q_1 is bounded by $\mathcal{O}(B^2m)$. Inducting from m equal 1 to N yields the additional factor of N in $\mathcal{O}(B^2N^2)$.

Dropping independence, define the random variable $Q_{m,n}$ to be the sum of m i.i.d. base qualities q_j drawn from $P(q \mid n)$. We can compute the distribution of $Q_{m,n}$ for each $m \leq n \leq N$ by parallel $\mathcal{O}(B^2n^2)$ computations for each $n \leq N$ as above. The overall complexity is therefore $\mathcal{O}(B^2N^3)$.

4.2 Approximation for large m

In practice, $P(n)$ decays exponentially fast for large n. Furthermore, for large n, the distribution of $\frac{m}{n}$ with $m \sim \text{binomial}(n, 0.5)$ concentrates around $\frac{1}{2}$ and thus Q_m is well-approximated by the distribution of the sum of $\frac{n}{2}$ elements drawn from the distribution of base qualities. By the Central Limit Theorem, for large m this sum is approximately normally distributed with mean $m\mu$ and variance $m\sigma^2$, where μ and σ^2 are the mean and variance of the base quality distribution $P(q)$. Thus in (6) we can approximate the innermost sum by

$$\sum_{s=mq_{\min}}^{\lfloor C_n \rfloor} P(Q_m = s) \approx \sum_{s=mq_{\min}}^{\lfloor C_n \rfloor} \mathcal{N}(m\mu, m\sigma^2) \tag{8}$$

$$\approx \int_{s=mq_{\min}}^{C_n} \mathcal{N}(m\mu, m\sigma^2)\, ds \tag{9}$$

$$= \Phi\left(\frac{C_n - m\mu}{\sigma\sqrt{m}}\right) - \Phi\left(\frac{mq_{\min} - m\mu}{\sigma\sqrt{m}}\right). \tag{10}$$

Dropping independence, this approximation extends directly to $Q_{m,n}$ by using the mean and variance of $P(q \mid n)$.