workout_probabilities

Brian Mannakee

4/22/2019

Intro

Variant callers are statistical classifiers. Usefulness of any classifier is a combination of the ranking performance of the classifier, and the quality of decisions based on the classifier in the real world. Here we describe an empirical Bayes method which uses very high confidence variant calls from MuTect to create a biologically inspired prior probability of mutation based in the mutation spectrum of the tumor. We use realistic tumor simulations to show that this method is superior to MuTect as a classifier based on AUROC, while equal or inferior to MuTect based on AUPRC except in situations where the distribution of allele frequencies is extreme and unrealistic. We then implement a method of false discovery rate control that recovers the superiority of the classifier in the decision context.

Emprical Bayes estimate of the posterior probability of mutation

At every site in the genome with non-zero coverage, Next Generation Sequencing (NGS) produces a vector $\mathbf{x} = (\{b_i\}, \{q_i\}), i = 1 \ldots D$ of base calls and their associated quality scores, where D is total read depth. The goal is to use $\mathbf{x}$ to select between competing hypotheses;

$$\begin{array}{rl} \mathbf{H_0}: & \text{Alt allele} = m; \quad \nu = 0 \\ \mathbf{H_1}: & \text{Alt allele} = m; \quad \nu = \hat{f}, \end{array}$$

where ν is the variant allele frequency, $\hat{f}$ is the maximum likelihood estimate of ν given data $\mathbf{x}$, i.e. the ratio of the count of variant reads and total read depth, and m is any of the 3 possible alternative non-reference bases. For a given read with base b_i and q-score q_i, the density function under a hypothesis is $\mathrm{f}_{\nu,m}(b_i, q_i)$ is defined as

$$\mathrm{f}_{\nu,m}(b_i, q_i) = \begin{cases} \nu \frac{10^{-q_i/10}}{3} + (1-\nu)(1 - 10^{-q_i/10}) & b_i = \text{reference} \\ \nu(1 - 10^{-q_i/10}) + (1-\nu)\frac{10^{-q_i/10}}{3} & b_i = m \\ \frac{10^{-q_i/10}}{3} & \mathit{otherwise}. \end{cases}$$

The likelihood under the hypothesis is then $\mathcal{L}_{\nu,m}(\mathbf{x}) = \prod_{i=1}^{D} \mathrm{f}_{\nu,m}(x_i)$. MuTect reports the log likelihood ratio $log(\mathcal{L}_{\nu=\hat{f},m}(\mathbf{x})/\mathcal{L}_{\nu=0,m}(\mathbf{x}))$ as either TLOD or t_lod_fstar depending on the version. By fixing the threshold posterior odds at two, and the prior probability of mutation a constant $\mu = 1\text{e}{-6}$, they derive a TLOD threshold of 6.3. Here we examine the effect of the assumption of a constant prior probability of mutation.