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.

Somatic variant calling base statistical model

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;

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

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 particular hypothesis 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, the site-specific mutation probability a constant $\mathrm{p}(M) = \mu = 3\mathrm{e}{-6}$, and $\mathrm{p}(m \mid M)$ the prior probability of mutation to specific allele m constant $\mathrm{p}(m \mid M) = \mu/3 = 1\mathrm{e}{-6}$, they derive a TLOD threshold of 6.3 for classifying a site as a somatic variant. Here we examine the effect of the assumption of a constant prior probability of mutation.

Site-specific prior probability of mutation

While variant calling algorithms typically assume a constant probability of mutation at every site in the genome, work by Alexandrov and others show that the random mutation generating process actually varies from site to site in a nucleotide context specific manner. We develop a model of the prior probability of mutation to allele m conditional on the observed genomic context $\mathrm{p}(m, M \mid C)$, and demonstrate an empirical

Bayes method for computing this probability from MuTect output. The prior probability $\mathrm{p}(m, M \mid C)$ can be decomposed as

$$\mathrm{p}(m, M \mid C) = \mathrm{p}(m \mid C)\mathrm{p}(M \mid C) = \mathrm{p}(m \mid C)\mathrm{p}(C \mid M)\frac{p(M)}{p(C)}$$

since the probability of a mutation at a site and the probability that it is to allele m are independent conditional on the context. Here $\mathrm{p}(M) = \mu$ as above, and the empirical distribution of contexts $\mathrm{p}(C)$ is the fraction of the genome made up of each context. We model $\mathrm{p}(C \mid M)$ as a multinomial distribution with parameter $\boldsymbol{\pi} = \{\pi_i\}, i = 1 \ldots 96$. Mutations are drawn from this multinomial distribution such that $\mathrm{p}(C = i \mid M) = \pi_i$. The final quantity $\mathrm{p}(m \mid C)$, the probability of mutation to m given a particular three letter context, is a function of $\boldsymbol{\pi}$. We are left to estimate only the vector of probabilities $\boldsymbol{\pi}$.

$$C \mid M, \boldsymbol{\pi} \sim \text{Multinomial}(\boldsymbol{\pi})$$
$$\boldsymbol{\pi} \mid \boldsymbol{\alpha} \sim \text{Dirichlet}(\boldsymbol{\alpha}).$$

The posterior distribution of $\boldsymbol{\pi}$ is $\boldsymbol{\pi} \mid C, \boldsymbol{\alpha} \sim \text{Dirichlet}(\mathbf{C} + \boldsymbol{\alpha})$, where $\mathbf{C} = (C_1, \ldots, C_{96})$ are the counts of mutations present in the tumor for each of the 96 contexts. We compute an empirical bayes estimate of $\boldsymbol{\pi}$ by choosing $\mathbf{C}$ as the set of mutations assigned a TLOD by MuTect above some threshold, which we choose as 10. We show through extensive simulation that our estimate of $\boldsymbol{\pi}$ converges quickly its true simulated value.

Returning to the model above, we can calculate the log posterior odds in favor of $\mathbf{H_1}$ as

$$\log_{10}\left(\frac{(\mathcal{L}_{\nu=\hat{f},m}(\mathbf{x})\mathrm{p}(m, M \mid C)}{(\mathcal{L}_{\nu=0,m}(\mathbf{x})(1 - \mathrm{p}(m, M \mid C))}\right) = \text{TLOD} + \text{log prior odds},$$

and the posterior odds ratio in favor of $\mathbf{H_1}$ as

$$10^{(\text{TLOD+log prior odds})}.$$

We show via extensive simulation that under the assumption that the mutation signature describes the biological process generating mutations in a tumor, our posterior odds ratio is a better classifier than MuTect at any threshold. Unfortunately, this increased classification performance comes at a cost in terms of decision-making. As measured by AUPRC, performance is a complex function of sequencing depth, mutation signature concentration, and the allele frequency distribution. In our simulations there is always a point on the PRC curve where MuTect outperforms our method. This is a result of the fact that for a given there is an allele frequency at which the ratio of expect true positives to expected false positives from being greater than 1 to less than 1. There is also an asymmetry in that around any particular threshold, the gradient of the ratio is fairly steep (good lord, this is awful).

False positive rate control.

We develop a method, following Efron(2008), for controlling the false positive rate. Every site with sufficient coverage and at least 1 alternate read falls into one of two classes, they are either *null* (non-variant with $\nu = 0$) or *nonnull* (variant with $\nu = \hat{f}$) with prior probabilities p_0 and $p_1 = 1 - p_0$,

$$\begin{array}{lll} p_0 = \mathrm{P}\{\text{null}\} & \mathrm{f}_0(\mathbf{x}) & \text{density if null} \\ p_1 = \mathrm{P}\{\text{nonnull}\} & \mathrm{f}_1(\mathbf{x}) & \text{density if nonnull.} \end{array}$$

In this application we can a compute an estimate of p_0 as the ratio of the expected number of false positive sites and the number of expected true positive sites with variant allele frequency $\nu = \hat{f}$. At a given site, with read depth D, the probability of seeing n reads in error is binomial

$$p_e \sim \text{Bin}(n, D, p = \bar{q})$$

where $\bar{q}$ is the average quality score of the alternate reads. (This is an estimate, they are have different q-scores) The expected number of false positives with probability p_e is then $N * p_e$, where N is the number of sites sequenced. In order to compute the expected nubmer of true positives with a given vaf, we need to make an assumption about the allele frequency distribution. In a neutrally evolving tumor, the count of variants with a given allele frequency is(bozic et al)

$$N(f) = \frac{N\mu}{f}.$$