Probability Model

Brian Mannakee

6/25/2019

Introduction

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 on 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. The model above is poorly calibrated, so we use estimates of the mutation rate and the distribution of allele frequencies under the null hypothesis. These estimates rely only the well-supported assumption that after initiation the tumor is evolving essentially neutrally. We show that the model is well-calibrated with this full estimate of the posterior distribution.

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;

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

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 particular hypothesis is defined as

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

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 $\boldsymbol{\pi}$.

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

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 calibration, The probabilities from this model are substantially worse than those from Mutect, such that threshold selection is essentially impossible and precision/recall is substantially worse for this model. There are several distributional assumptions in this model that effect model calibration, and we address all of them below.

Estimating the mutation rate from the data

As discussed above, MuTect fixes the site-specific mutation probability at $\mathrm{p}(M) = \mu = 3\mathrm{e}{-6}$. All variant callers we are aware of either fix this parameter μ or allow the user to input the value, but there is no way to really know this value until the variant allele frequencies have been observed. As with estimation of the mutation signature, we can use high confidence mutations and a model of tumor evolution to compute the tumor-specific mutation rate. Bozic, Gerold, and Nowak (2016) show that for any variant allele frequency α, the total number of mutations with frequency greater than α and less than 0.25 is

$$N(\alpha) = N\mu\left(\frac{1}{\alpha} - \frac{1}{0.25}\right)$$

Where N is the total number of sites sequenced and μ is the per-site mutation probability. By selecting α such that we are highly confident in all calls at frequencies greater than α, we can compute μ and recompute the odds in favor of $\mathbf{H_1}$. This improves the calibration of the model, but it is still worse than MuTect. The final assumption embedded in the model is that all variants, regardless of allele frequency, have the same prior probability of being true variants. Below we address this assumption.

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 $p_1 = 1 - p_0$,

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

The local, or site-specific, true positive probability p_1 can be estimated as the fraction of all sequenced sites that are expected to be positive. In a neutrally evolving tumor, the number of cells is growing exponentially, and the count of variants with an allele frequency greater than a given allele frequency f is (Bozic, Gerold, and Nowak 2016,Williams et al. (2016))

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

Where N is the total number of sites sequenced and μ is the per-site mutation probability. The estimated fraction of all of the sites in the genome that will have a mutation with frequency f is

$$\hat{p}_1 = \frac{\int_{f_-}^{f_+} N(f)}{N} = \frac{\mu}{f - .1f} - \frac{\mu}{f + .1f}$$

and the estimated null probability $\hat{p}_0 = 1 - \hat{p}_1$. Williams et al. (2016) provides a full derivation for this, we are essentially computing the integral here of $N(f)$ in a small area around f.

Results

Figure 1 shows that the calibration is better than MuTect or the model without false positive rate control. In addition estimation of mu from the data results in better calibration than assuming a mis-specified rate.

Figures

References

Bozic, Ivana, Jeffrey M. Gerold, and Martin A. Nowak. 2016. "Quantifying Clonal and Subclonal Passenger Mutations in Cancer Evolution." *PLoS Computational Biology* 12 (2): e1004731. doi:10.1371/journal.pcbi.1004731.

Efron, Bradley. 2008. "Microarrays, Empirical Bayes and the Two-Groups Model." *Statistical Science* 23 (1). Institute of Mathematical Statistics: 45–47. doi:10.1214/08-sts236rej.

Williams, Marc J, Benjamin Werner, Chris P Barnes, Trevor A Graham, and Andrea Sottoriva. 2016. "Identification of neutral tumor evolution across cancer types." *Nature Genetics* 48 (3). Nature Publishing Group: 238–44. doi:10.1038/ng.3489.

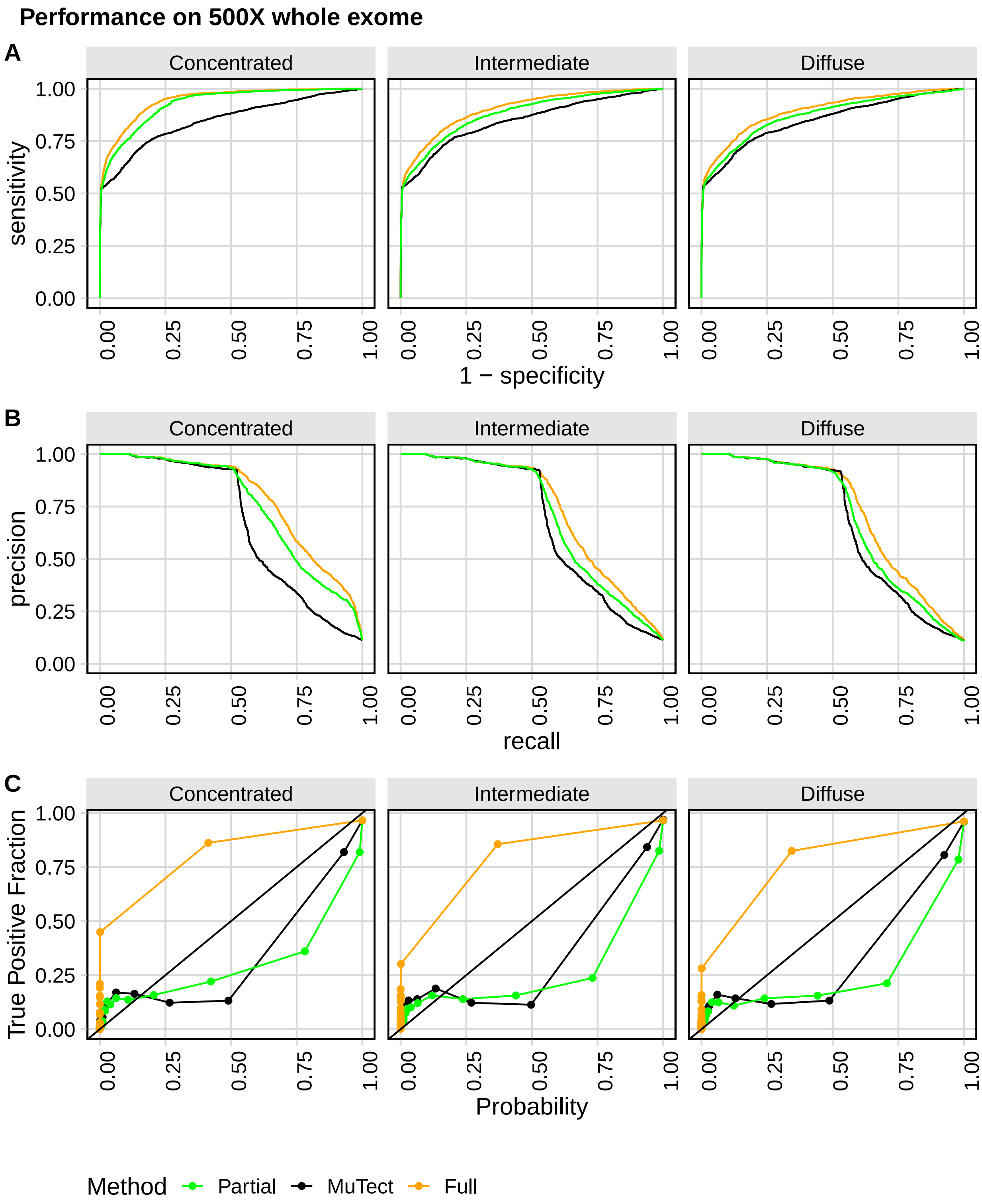

Figure 1: Model performance 500X whole exome for 3 classes of mutation spectrum

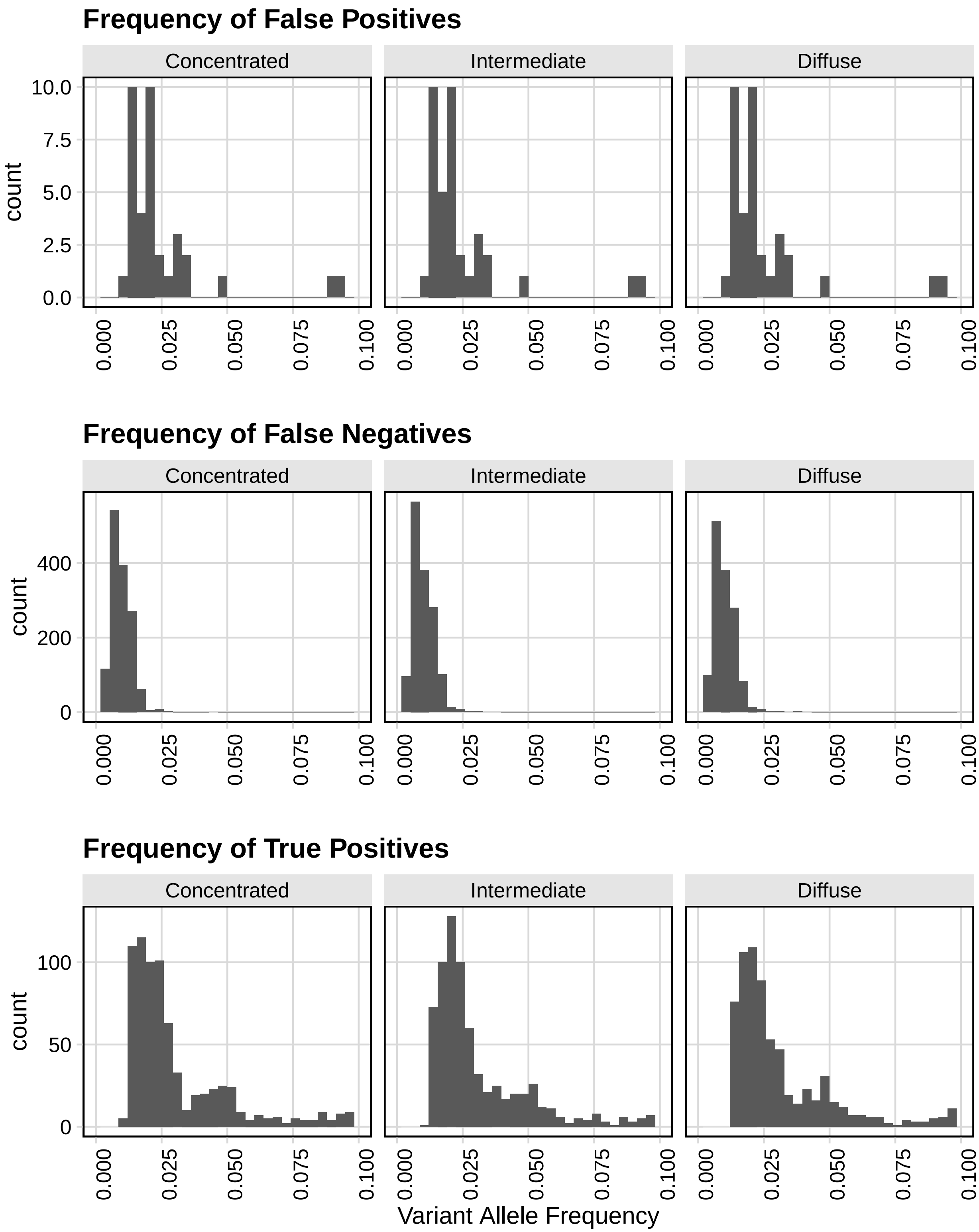

Figure 2: Allele frequency profile 500X whole exome for 3 classes of mutation spectrum. Frequencies at a threshold p(H1) = 0.8

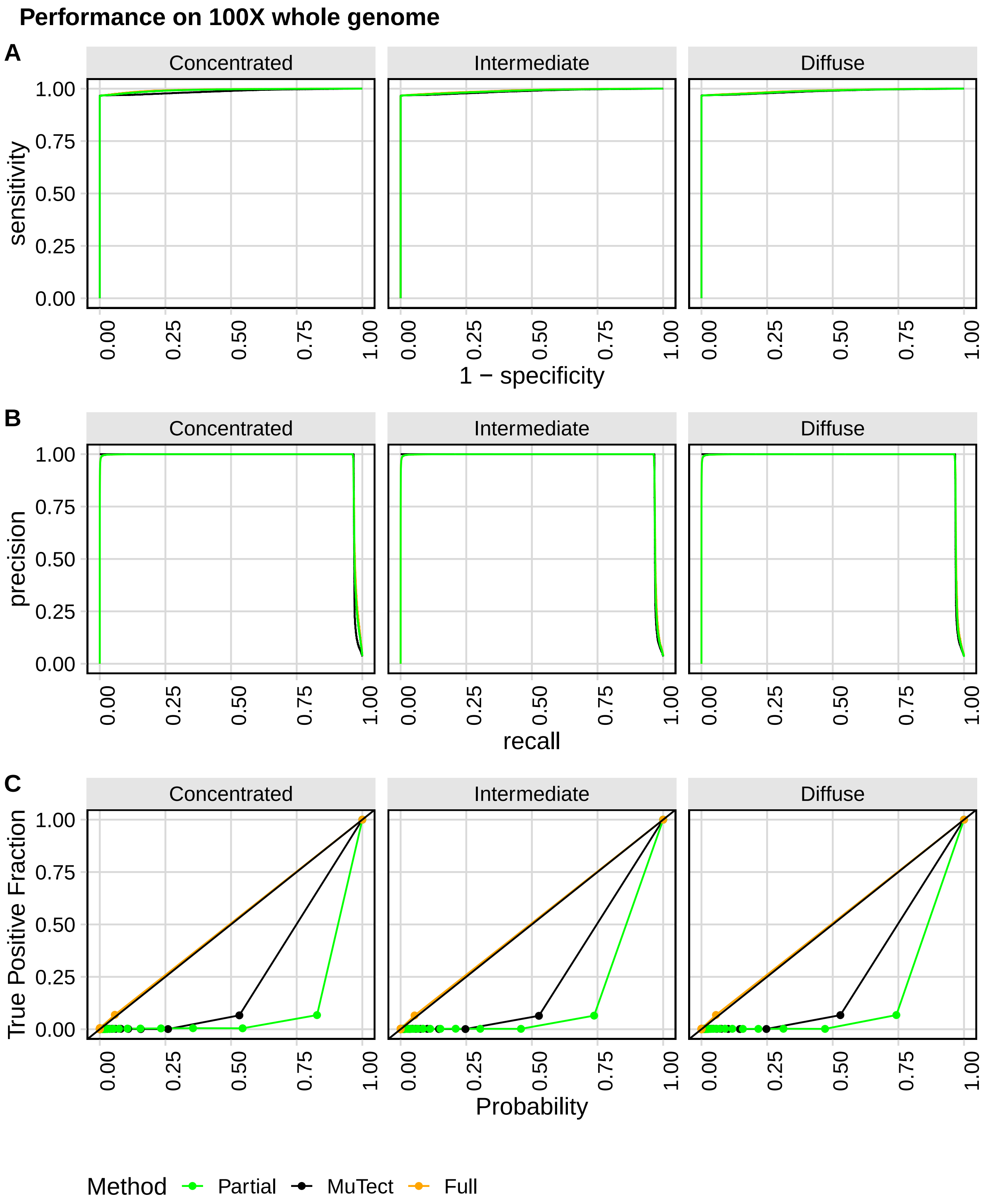

Figure 3: Model performance 100X whole genome for 3 classes of mutation spectrum

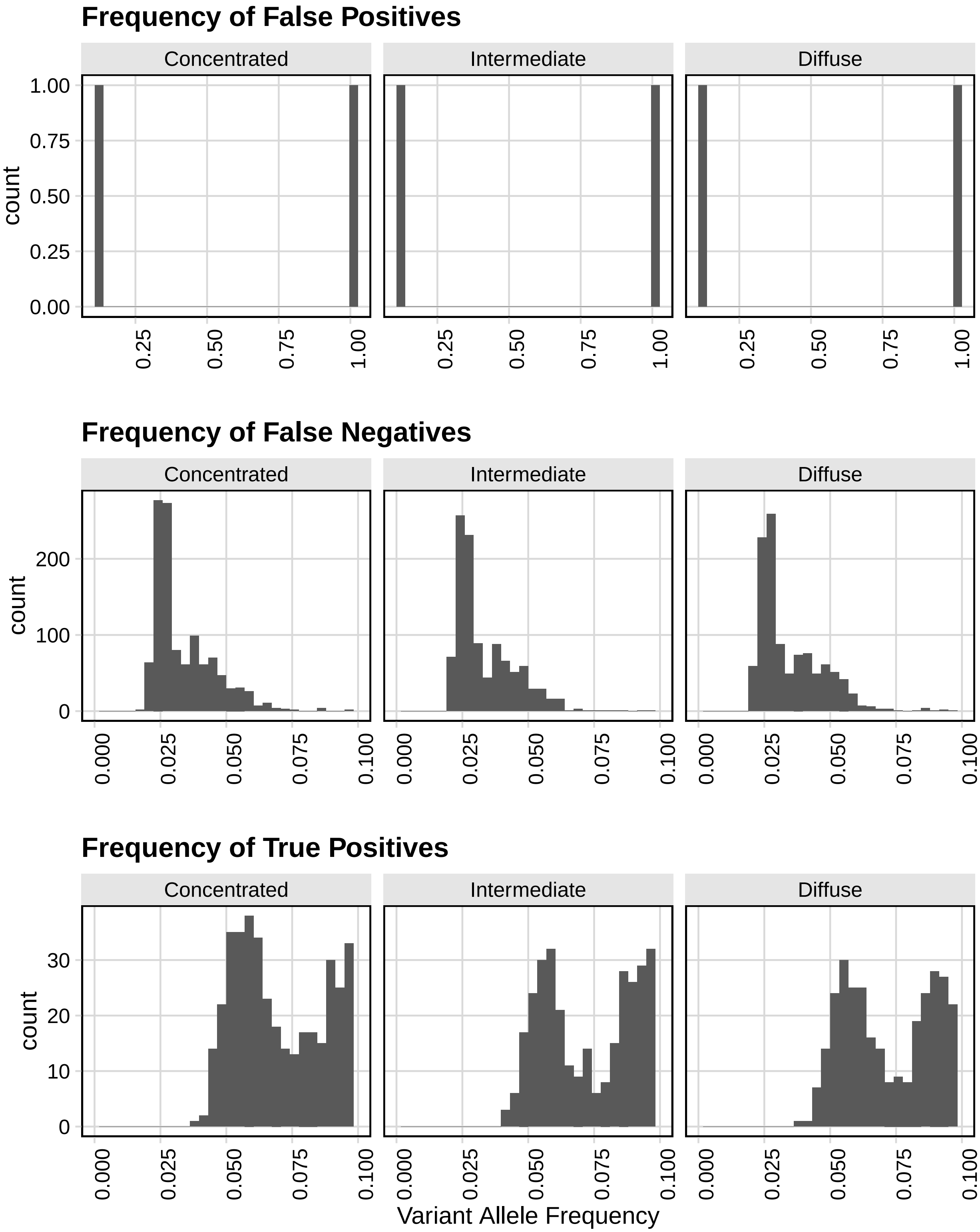

Figure 4: Allele frequency profile 100X whole genome for 3 classes of mutation spectrum. Frequencies at a threshold p(H1) = 0.8

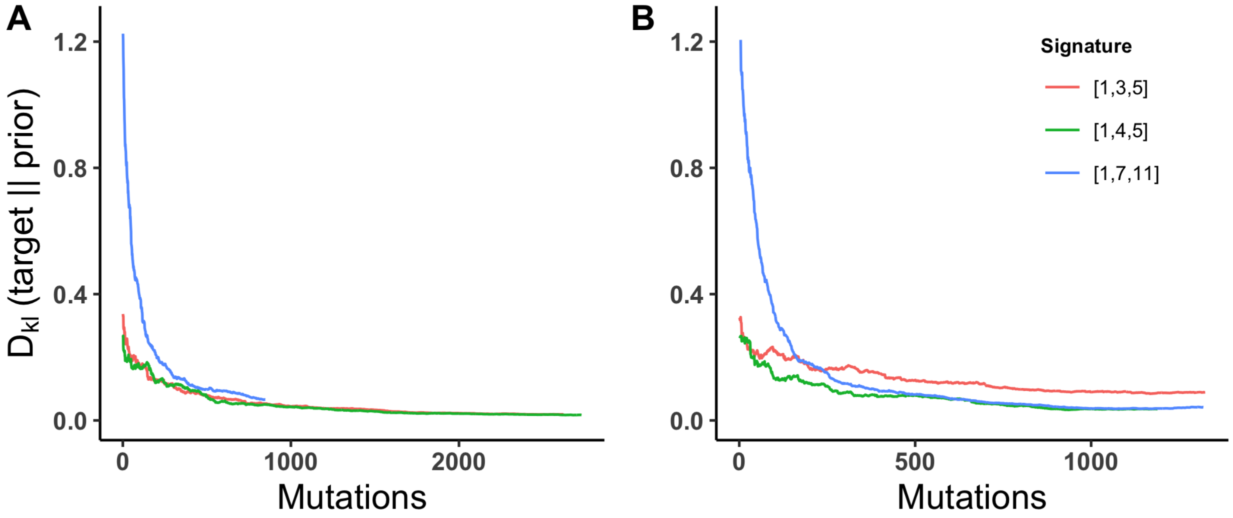

Figure 5: Convergence of empirical prior

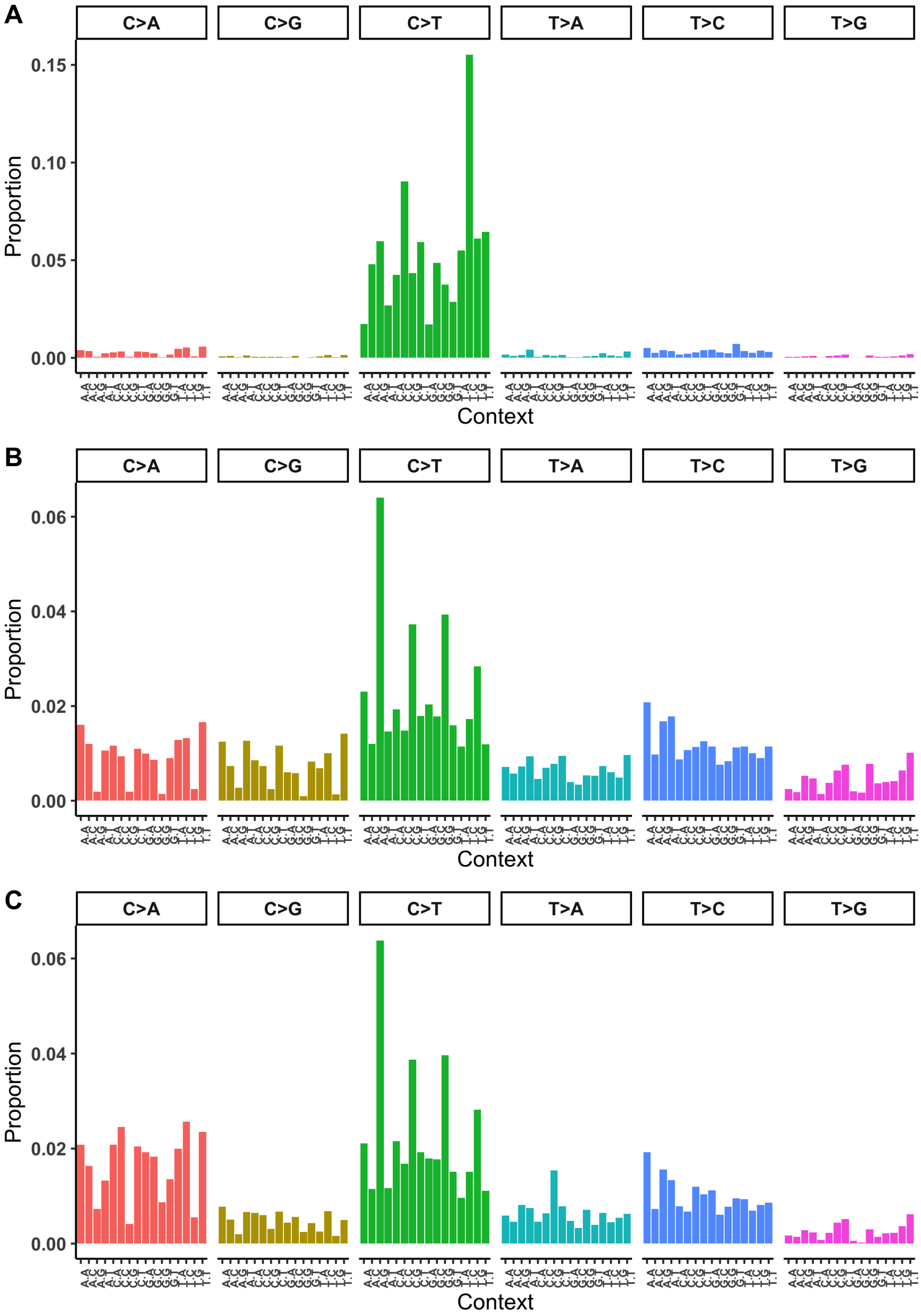

Figure 6: Referene input signatures