GATK variant quality score model

David Benjamin[*]
Broad Institute, 75 Ames Street, Cambridge, MA 02142
(Dated: July 13, 2018)

I. INTRODUCTION

The variant quality model is used in both `HaplotypeCaller` and `GenotypeGVCFs`. In `HaplotypeCaller` the model is applied we have decided which alleles to consider and calculated the likelihoods of each possible genotype. In `GenotypeGVCFs` the model is applied after merging input GVCFs and harmonizing their allele representations, and the resulting score is used to decide which variants to emit. Let $\ell_{sg} \equiv P(\text{reads}_s|g)$ be the likelihood of genotype g in sample s. Then a simple model for the genotypes and observed reads is

$$P(\text{reads}, \mathbf{z}) = \prod_s P(\mathbf{z}_s) \prod_g \ell_{sg}^{z_{sg}}, \tag{1}$$

where z_{sg} is a binary indicator variable for sample s exhibiting genotype g. We can represent $P(\mathbf{z}_s)$ in terms of a latent vector π of population allele frequencies. Assuming independent alleles, the probability $P(\mathbf{z}_s|\pi)$ of a genotype g containing n_{ga} copies of allele a (i.e. for tetraploid genotype AABC, $n_A = 2$, $n_B = n_C = 1$) is

$$P(\mathbf{z}_s|\pi) = \prod_g \left[C_g \prod_a (\pi_a)^{n_{ga}} \right]^{z_{sg}}, \tag{2}$$

where $C_g = \text{ploidy}!/\left(\prod_a n_{ga}!\right)$ is the number of phased genotypes corresponding to unphased genotype g. For example, $C_{AAB} = 3$ because AAB, ABA, and BAA are distinct phased genotypes, while $C_{AAA} = 1$.

Because sites vary in their allele frequencies, π is not a constant. Rather, it is a random variable whose prior distribution should have the correct mean alt allele frequency (roughly 1 in 1000 for SNPs) and roughly the correct standard deviation. That is, the prior probabilities that a site has a rare alt with, say $\pi_B = 10^{-5}$ or a common alt with $\pi_B = 0.25$ should match the empirical distribution of allele frequencies. Note that the old model essentially assumed that every site had an alt allele with a constant allele frequency of 1/1000. We can achieve this by setting $\pi \sim \text{Dir}(\alpha)$, where α is a vector with one component per allele, and α_a is a prior pseudocount for allele a. We can set these pseudocounts separately for the ref allele, SNPs, and indels[1] to obtain the desired mean and standard deviation of allele frequencies. Unlike previous approaches which shoehorned multiallelic sites into biallelic model, this method extends naturally to multiallelic sites. Since α is not a random variable we can ignore the normalization constant of the Dirichlet, which depends only on α, to obtain $P(\pi|\alpha) \propto \prod_a \pi_a^{\alpha_a}$. This, combined with Equations 1 and 2, gives the joint posterior distribution

$$P(\pi, \mathbf{z}) \propto \left(\prod_a \pi_a^{\alpha_a} \right) \prod_{sg} \left[C_g \ell_{sg} \prod_a \pi_a^{n_{ga}} \right]^{z_{sg}} \tag{3}$$

We perform mean-field variational Bayesian inference on this posterior. In this framework we posit a factorized approximation: $P(\pi, \mathbf{z}) \approx q(\pi) q(\mathbf{z})$ and iterate, alternating between the following two updates:

$$q(\pi) \propto \exp E_{q(\mathbf{z})}[\ln P(\pi, \mathbf{z})] \propto \prod_a \pi_a^{\alpha_a + \sum_{sg} E[z_{sg}] n_{ga}} \tag{4}$$

$$q(\mathbf{z}) \propto \exp E_{q(\pi)}[\ln P(\pi, \mathbf{z})] \propto \prod_g \left[C_g \ell_{sg} \prod_a e^{E[\ln \pi_a] n_{ga}} \right]^{z_{sg}} \tag{5}$$

[*] Electronic address: `davidben@broadinstitute.org`

[1] In principle we could have a complicated model for prior pseudocounts depending on context. Furthermore, if we have a big call set like gnomAD, the best prior pseudocounts are simply the allele counts from that call set. `HaplotypeCaller` and `GenotypeGVCFs` do not currently exploit gnomAD for this.

The expectations required have analytic expressions. Using a well-known (i.e. see the Wikipedia entry for Dirichlet distribution) result:

$$\begin{aligned} E_{q(\pi)}[\ln \pi_a] =& \psi(\alpha_a + \sum_{sg} E[z_{sg}] n_{ga}) - \psi(\sum_a \alpha_a + \sum_{sg} E[z_{sg}] \sum_a n_{ga}) \qquad (6) \\ =& \psi(\alpha_a + \sum_{sg} E[z_{sg}] n_{ga}) - \psi(\sum_a \alpha_a + \text{total ploidy of all samples}), \qquad (7) \end{aligned}$$

where ψ is the digamma function, and

$$E_{q(\mathbf{z})}[z_{sg}] = \frac{C_g \ell_{sg} \prod_a e^{E[\ln \pi_a] n_{ga}}}{\text{normalizing constant}}, \qquad (8)$$

where the constant of normalization is chosen so that $\sum_g E_{q(\mathbf{z})}[z_{sg}] = 1$. Learning this model is just a matter of iterating Equations 4 and 5.

Once our iteration converges, we can extract several interesting things, some of which we already report and some of which we do not. First, we have the Dirichlet posterior $q(\pi)$ on allele frequencies. Letting $N_a \equiv \alpha_a + \sum_{sg} E[z_{sg}] n_{ga}$ be the prior plus observed pseudocounts for allele a, we have

$$\pi \sim \text{Dir}(N_0, N_1 \ldots) \qquad (9)$$

If we want a univariate allele frequency for a single allele (i.e. not the joint Dirichlet posterior on all allele frequencies) we simply marginalize, which is arithmetically easy:

$$\pi_a \sim \text{Beta}(N_a, \sum_{a' \neq a} N_{a'}) \qquad (10)$$

From this we can get the mean posterior allele frequency, which is $\bar{\pi}_a = N_a / \sum_{a'} N_a$, but we also have error bars on this estimate.

We can also obtain the variant quality score trivially. The probability that no variant exists among the samples is (we use the convention that $g = 0$ for the hom ref genotype)

$$P(\text{no variants}) = \prod_s P(z_{sg} = 0) = \prod_s E_{q(\mathbf{z})}[z_{s,g=0}] \qquad (11)$$

We can easily extend this to a per-allele variant quality score by considering not just the hom ref genotype, but all genotypes in which that allele is absent:

$$P(\text{no allele } a) = \prod_s \left(\sum_{g : n_{ga} = 0} E_{q(\mathbf{z})}[z_{s,g}] \right) \qquad (12)$$

Let's now consider the run time of this algorithm. Suppose there are S samples, G genotypes, and A alleles. The cost of calculating Equation 7 for all values of a is $O(SGA)$. The cost of calculating Equation 8 for all s and g is also $O(SGA)$. These things usually converge in not too many iterations so we have an $O(SGA)$ algorithm that yields genotype calls, genotype quals (these are the posteriors $E[z_{sg}]$), allele frequencies, per-site variant quality scores, and per-allele variant quality scores. The algorithm has two equations to learn the model and one equation to translate that into a variant quality score.