Mathematical Notes on Mutect

David Benjamin* and Takuto Sato†
Broad Institute, 75 Ames Street, Cambridge, MA 02142
(Dated: July 12, 2017)

I. SOMATIC LIKELIHOODS MODEL

We have a set of potential somatic alleles and read-allele likelihoods $\ell_{ra} \equiv P(\text{read } r|\text{allele } a)$. We don't know which alleles are real somatic alleles and so we must compute, for each subset $\mathbb{A}$ of alleles, the likelihood that the reads come from $\mathbb{A}$. A simple model for this likelihood is as follows: each read r is associated with a latent indicator vector $\mathbf{z}_r$ with one-hot encoding $z_{ra} = 1$ iff read r came from allele $a \in \mathbb{A}$. The conditional probability of the reads $\mathbb{R}$ given their allele assignments is

$$P(\mathbb{R}|\mathbf{z}, \mathbb{A}) = \prod_{r \in \mathbb{R}} \prod_a \ell_{ra}^{z_{ra}}. \tag{1}$$

The alleles are not equally likely because there is a latent vector $\mathbf{f}$ of allele fractions – f_a is the allele fraction of allele a. Since the components of $\mathbf{f}$ sum to one it is a categorical distribution and can be given a Dirichlet prior,

$$P(\mathbf{f}) = \text{Dir}(\mathbf{f}|\boldsymbol{\alpha}). \tag{2}$$

Then f_a is the prior probability that a read comes from allele a and thus the conditional probability of the indicators $\mathbf{z}$ given the allele fractions $\mathbf{f}$ is

$$P(\mathbf{z}|\mathbf{f}) = \prod_r \prod_a f_a^{z_{ra}}. \tag{3}$$

The full-model likelihood is therefore

$$\mathbb{L}(\mathbb{A}) = P(\mathbb{R}, \mathbf{z}, \mathbf{f}|\mathbb{A}) = \text{Dir}(\mathbf{f}|\boldsymbol{\alpha}) \prod_a \prod_r (f_a \ell_{ra})^{z_{ra}}. \tag{4}$$

And the marginalized likelihood of $\mathbb{A}$, that is, the model evidence for allele subset $\mathbb{A}$, is

$$P(\mathbb{R}|\mathbb{A}) = \sum_{\mathbf{z}} \int d\mathbf{f}\, \text{Dir}(\mathbf{f}|\boldsymbol{\alpha}) \prod_a \prod_r (f_a \ell_{ra})^{z_{ra}}, \tag{5}$$

where the integral is over the probability simplex $\sum_a f_a = 1$.

The integral over $\mathbf{f}$ is the normalization constant of a Dirichlet distribution and as such we can simply look up its formula. However, the sum over all values of $\mathbf{z}$ for all reads has exponentially many terms. We will get around this difficulty by handling $\mathbf{z}$ with a mean-field approximation in which we factorize the likelihood as $\mathbb{L} \approx q(\mathbf{z})q(\mathbf{f})$. This approximation is exact in two limits: first, if there are many reads, each allele is associated with many reads and therefore the Law of Large Numbers causes $\mathbf{f}$ and $\mathbf{z}$ to become uncorrelated. Second, if the allele assignments of reads are obvious $\mathbf{z}_r$ is effectively not a random variable at all (there is no uncertainty as to which of component is non-zero) and also becomes uncorrelated with $\mathbf{f}$.

In the variational Bayesian mean-field formalism the value of $\mathbf{f}$ that $\mathbf{z}$ "sees" is the expectation of $\log \mathbb{L}$ with respect to $q(\mathbf{f})$ and vice versa. That is,

$$q(\mathbf{f}) \propto \text{Dir}(\mathbf{f}|\boldsymbol{\alpha}) \prod_a \prod_r f_a^{\bar{z}_{ra}} \propto \text{Dir}(\mathbf{f}|\boldsymbol{\alpha} + \sum_r \bar{\mathbf{z}}_r), \tag{6}$$

*Electronic address: davidben@broadinstitute.org
†Electronic address: tsato@broadinstitute.org

where $\bar{z}_{ra} \equiv E_q[z_{ra}]$, and

$$q(\mathbf{z}_r) = \prod_a \left(\tilde{f}_a \ell_{ra}\right)^{z_{ra}}, \tilde{f}_a = \exp E[\ln f_a] \tag{7}$$

Because $q(\mathbf{z})$ is categorical and $q(\mathbf{f})$ is Dirichlet[1] the necessary mean fields are easily obtained and we have

$$\bar{z}_{ra} = \frac{\tilde{f}_a \ell_{ra}}{\sum_{a'} \tilde{f}_{a'} \ell_{ra'}} \tag{8}$$

and

$$\ln \tilde{f}_a = \psi(\alpha_a + \sum_r \bar{z}_{ra}) - \psi(\sum_{a'} \alpha_{a'} + N) \tag{9}$$

where ψ is the digamma function and N is the number of reads. To obtain $q(\mathbf{z})$ and $q(\mathbf{f})$ we iterate Equations 8 and 9 until convergence. A very reasonable initialization is to set $\bar{z}_{ra} = 1$ if a is the most likely allele for read r, 0 otherwise. Having obtained the mean field of $\mathbf{z}$, we would like to plug it into Eq 5. We can't do this directly, of course, because Eq 5 says nothing about our mean field factorization. Rather, we need the variational approximation (Bishop's Eq 10.3) to the model evidence, which is

$$\ln P(\mathbb{R}|\mathbb{A}) \approx \sum_{\mathbf{z}} \int d\mathbf{f} q(\mathbf{z}) q(\mathbf{f}) \left[\ln P(\mathbb{R}, \mathbf{z}, \mathbf{f}|\mathbb{A}) - \ln q(\mathbf{z}) - \ln q(\mathbf{f})\right] \tag{10}$$

$$= E_q\left[\ln P(\mathbb{R}, \mathbf{z}, \mathbf{f}|\mathbb{A})\right] - E_q\left[\ln q(\mathbf{z})\right] - E_q\left[\ln q(\mathbf{f})\right]. \tag{11}$$

Before we proceed, let's introduce some notation. First, from Eq 6 the posterior $q(\mathbf{f})$ is

$$q(\mathbf{f}) = \mathrm{Dir}(\mathbf{f}|\boldsymbol{\beta}), \quad \boldsymbol{\beta} = \boldsymbol{\alpha} + \sum_r \bar{\mathbf{z}}_r. \tag{12}$$

Second, let's define the log normalization constant of a Dirichlet distribution as g so that

$$\ln \mathrm{Dir}(\mathbf{f}|\boldsymbol{\omega}) = g(\boldsymbol{\omega}) + \sum_a (\omega_a - 1) \ln f_a, \quad g(\boldsymbol{\omega}) = \ln \Gamma(\sum_a \omega_a) - \sum_a \ln \Gamma(\omega_a). \tag{13}$$

Finally, define the Dirichlet mean log (aka "that digamma stuff") as h:

$$E_{\mathrm{Dir}(\mathbf{f}|\boldsymbol{\omega})}\left[\ln f_a\right] = \psi(\omega_a) - \psi(\sum_{a'} \omega_{a'}) \equiv h_a(\boldsymbol{\omega}). \tag{14}$$

The log of Eq 4 is

$$\ln P(\mathbb{R}, \mathbf{z}, \mathbf{f}|\mathbb{A}) = g(\boldsymbol{\alpha}) + \sum_a (\alpha_a - 1) \ln f_a + \sum_{ra} z_{ra} (\ln f_a + \ln \ell_{ra}). \tag{15}$$

and thus the first term in Eq 11 is

$$E_q\left[\ln P(\mathbb{R}, \mathbf{z}, \mathbf{f}|\mathbb{A})\right] = g(\boldsymbol{\alpha}) + \sum_a (\alpha_a - 1) h_a(\boldsymbol{\beta}) + \sum_{ra} \bar{z}_{ra} \left(h_a(\boldsymbol{\beta}) + \ln \ell_{ra}\right) \tag{16}$$

$$= g(\boldsymbol{\alpha}) + \sum_a (\beta_a - 1) h_a(\boldsymbol{\beta}) + \sum_{ra} \bar{z}_{ra} \ln \ell_{ra}, \tag{17}$$

where we used the relationship $\boldsymbol{\beta} = \boldsymbol{\alpha} + \sum_r \bar{\mathbf{z}}_r$.

The second term in Eq 11 is

$$-E_q\left[\ln q(\mathbf{z})\right] = -\sum_{ra} \bar{z}_{ra} \ln \bar{z}_{ra}. \tag{18}$$

---

[1] Note that we didn't *impose* this in any way. It simply falls out of the mean field equations.

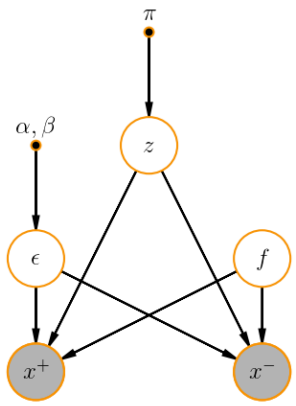

FIG. 1: The probabilistic graphical model

The third term in Eq 11 is

$$-E_q\left[\ln q(\mathbf{f})\right] = -g(\boldsymbol{\beta}) - \sum_a (\beta_a - 1) E_q[\ln f_a] = -g(\boldsymbol{\beta}) - \sum_a (\beta_a - 1) h_a(\boldsymbol{\beta}). \tag{19}$$

Adding Eqs 17, 18, and 19 and noting the cancellation between parts of Eqs 17 and 19 we obtain

$$\ln P(\mathbb{R}|\mathbb{A}) \approx g(\boldsymbol{\alpha}) - g(\boldsymbol{\beta}) + \sum_{ra} \bar{z}_{ra} \left(\ln \ell_{ra} - \ln \bar{z}_{ra}\right). \tag{20}$$

We now have the model evidence for allele subset $\mathbb{A}$. This lets us choose which alleles are true somatic variants. It also lets us make calls on somatic loss of heterozygosity events. Furthermore, instead of reporting max-likelihood allele fractions as before, we may emit the parameters of the Dirichlet posterior $q(\mathbf{f})$, which encode both the maximum likelihood allele fractions and their uncertainty.

II. STRAND ARTIFACT MODEL

- $\mathbf{z}$ is a latent random variable having a 1-of-K representation. For each variant locus, $\mathbf{z}$ encodes the presence of strand artifact in forward reads ($[1,0,0]$), artifact in reverse reads ($[0,1,0]$), or no artifact ($[0,0,1]$)
- $f \sim \text{Unif}(0,1)$ is a prior distribution over alt allele fraction f
- $\epsilon \sim \text{Beta}(\alpha, \beta)$ is a prior distribution over the error probability on a read on the artifact strand. For instance, if we have strand artifact on the reverse strand (i.e. $z = [0,1,0]$), ϵ is the probability that the sequencer reads a ref allele on a reverse read as alt
- $x^+|f, \epsilon, z$ is the number of forward reads with the alt allele. It's a mixture of binomials, defined as follows:

$$x^+|f,\epsilon,z \sim \begin{cases} \text{Bin}(n^+, f + \epsilon(1-f)) & z = \text{Art+} \\ \text{Bin}(n^+, f) & z = \text{Art}- \\ \text{Bin}(n^+, f) & z = \text{noArt} \end{cases} \tag{21}$$

We compute the conditional distributions of x^- analogously.

Having observed the read counts in the forward and reverse directions, we can compute the posterior probabilities of the latent variable z. Below we derive the unnormalized posterior probability of strand artifact in forward reads ($z = art+$), given that we observed x^+ forward alt reads and x^- reverse alt reads. We use a shorthand z_0 to denote $z = art+$ for conciseness.

First we will derive the likelihood $p(x^+, x^-, f, \epsilon|z_0)$

$$p(x^+, x^-|z_0) = \iint_{f,\epsilon} p(x^+, x^-, f, \epsilon|z_0)\, df\, d\epsilon \tag{22}$$

$$= \iint_{f,\epsilon} p(f)p(\epsilon)p(x^+, x^-|z_0, f, \epsilon)\, df\, d\epsilon \tag{23}$$

$$= \iint_{f,\epsilon} p(f)p(\epsilon)p(x^+|z_0, f, \epsilon)p(x^-|z_0, f, \epsilon)\, df\, d\epsilon \tag{24}$$

$$= \iint_{f,\epsilon} p(\epsilon)p(x^+|z_0, f, \epsilon)p(x^-|z_0, f, \epsilon)\, df\, d\epsilon \tag{25}$$

$$= \iint_{f,\epsilon} \text{Beta}(\epsilon|\alpha, \beta)\text{Bin}(x^+|f + \epsilon(1-f), n^+)\text{Bin}(x^-|f, n^-)\, df\, d\epsilon \tag{26}$$

The posterior probability of strand artifact in forward reads is therefore

$$p(z_0|x^+, x^-) \propto p(z_0)p(x^+, x^-|z_0) \tag{27}$$

$$= p(z_0) \iint_{f,\epsilon} \text{Beta}(\epsilon|\alpha, \beta)\text{Bin}(x^+|f + \epsilon(1-f), n^+)\text{Bin}(x^-|f, n^-)\, df\, d\epsilon \tag{28}$$

The derivation for the probability of strand artifact on reverse strand is analogous.

For the case of no strand artifact, the derivation of likelihoods is identical up to (25). Here we can simply the equation to a single integral over f because the conditional probabilities of x^+ and x^- do not depend on ϵ. We use a shorthand z_2 for $z = \text{noArt}$

$$p(x^+, x^-|z_2) = \iint_{f,\epsilon} p(\epsilon)p(x^+|z_2, f, \epsilon)p(x^-|z_2, f, \epsilon)\, df\, d\epsilon$$

$$= \int_f p(x^+|z_2, f)p(x^-|z_2, f)\, df \int_\epsilon p(\epsilon)d\epsilon \tag{29}$$

$$= \int_f \text{Bin}(x^+|f, n^+)\text{Bin}(x^-|f, n^-)\, df \tag{30}$$

And the posterior probability is

$$p(z_2|x^+, x^-) \propto p(z_2)p(x^+, x^-|z_2) \tag{31}$$

$$= p(z_2) \int_f \text{Bin}(x^+|f, n^+)\text{Bin}(x^-|f, n^-)\, df \tag{32}$$

III. GERMLINE FILTER

Suppose we have detected an allele such that its (somatic) likelihood in the tumor is ℓ_t and its (diploid) likelihood in the normal is ℓ_n. By convention, both of these are relative to a likelihood of 1 for the allele *not* to be found. If we have no matched normal, $\ell_n = 1$. Suppose we also have the population allele frequency f of this allele. Then the prior probability for the normal to be heterozygous or homozygous alt for the allele is $2f(1-f) + f^2$ and the prior probability for the normal genotype not to contain the allele is $(1-f)^2$. Finally, suppose that the prior for this allele to arise as a somatic variant is π.

We can determine the posterior probability that the variant exists in the normal genotype by calculating the unnormalized probabilities of four possibilities:

1. The variant exists is both the normal and the tumor samples. This has unnormalized probability $\left(2f(1-f) + f^2\right)\ell_n\ell_t(1-\pi)$.

2. The variant exists in the tumor but not the normal. This has unnormalized probability $(1-f)^2\ell_t\pi$.

3. The variant exists in neither the tumor nor the normal. This has unnormalized probability $(1-f)^2(1-\pi)$.

4. The variants exists in the normal but not the tumor. This is biologically very unlikely. Furthermore, if it *did* occur we wouldn't care about filtering the variant as a germline event because we wouldn't call it as a somatic event. Thus we neglect this possibility.

Normalizing, we obtain the following posterior probability that an allele is a germline variant:

$$P(\text{germline}) = \frac{(1)}{(1)+(2)+(3)} = \frac{\left(2f(1-f)+f^2\right)\ell_n\ell_t(1-\pi)}{\left(2f(1-f)+f^2\right)\ell_n\ell_t(1-\pi) + (1-f)^2\ell_t\pi + (1-f)^2(1-\pi)}. \tag{33}$$

To filter, we set a threshold on this posterior probability.

IV. CALCULATING CONTAMINATION

Below, we present the GATK's fast, simple, and accurate method for calculating the contamination of a sample without a matched normal and regardless of the number of contaminating samples that remains accurate even when the sample has a lot of copy number variation.

The inputs to our tool are a bam file and a list of common variants with their allele frequencies. The basic idea is simply to count ref reads at hom alt sites and subtract the number of ref reads expected from sequencing error to obtain the number of ref reads contaminating these hom alt sites. Finally, we use the allele frequencies to account for the fact that some contaminating reads have the alt allele. The only minor subtlety is in distinguishing hom alt sites from loss of heterozygosity events, which is not difficult to achieve as we describe below.

Suppose we have a set $\mathbb{H}$ of hom alt sites. Let N_{ref} be the total number of ref reads at these sites. We can decompose N_{ref} as follows:

$$N_{\text{ref}} = N_{\text{error}} + N_{\text{contamination}}, \tag{34}$$

where N_{error} and $N_{\text{contamination}}$ are as their names suggest. We can obtain N_{ref} by counting reads, and we estimate N_{error} as follows. Suppose, WLOG, that the ref allele is A and the alt is C. Then, assuming that all substitution errors are equally likely, N_{ref} is approximately half the number of Gs and Ts. This is, of course, not a perfect assumption for any one site, but on average over all the sites in $\mathbb{H}$ it is very good.

Next we take the expectation of both sides of Equation 34 to obtain

$$\langle N_{\text{ref}} - N_{\text{error}} \rangle = \sum_{s\in\mathbb{H}} \langle N^s_{\text{contamination}} \rangle \times \langle \text{fraction of contaminant reads that are ref} \rangle, \tag{35}$$

where $N^s_{\text{contamination}}$ is the number of contaminant reads at site s, the expected value of which is the depth d_s at site s times the contamination. The expected fraction of contaminant reads that are ref is one minus the alt allele frequency f_s. Crucially, this fact is independent of how many contaminating samples there are. Thus we have

$$\langle N_{\text{ref}} - N_{\text{error}} \rangle = \text{contamination} \sum_{s\in\mathbb{H}} d_s(1-f_s) \tag{36}$$

and obtain the estimate

$$\text{contamination} \approx \frac{\langle N_{\text{ref}} - N_{\text{error}} \rangle}{\sum_{s\in\mathbb{H}} d_s(1-f_s)} \tag{37}$$

Let us now roughly estimate the error bars on this result. SInce most alt alleles are not too common (by default, the tool looks only at sites with $f_s < 0.2$), we can calculate the error bars under the assumption that all contaminant reads are ref. Then the number of contaminating reads is a binomial distribution with the number of trials equal $\sum_{s\in\mathbb{H}} d_s$ and success probability equal to the contamination. Our contamination estimate is the number of proportion of successes drawn from this binomial distribution. Plugging in the variance of a binomial distribution, we find:

$$\text{var}(\text{contamination}) \approx \text{contamination} \times (1 - \text{contamination}) \frac{\sum_{s\in\mathbb{H}} d_s}{(\sum_{s\in\mathbb{H}} d_s)^2} \tag{38}$$

For the usual case of low contamination, we obtain a relative error

$$\frac{\text{std(contamination)}}{\text{contamination}} \approx \left(\sum_{s \in \mathbb{H}} d_s \right)^{-1/2} = (\text{average depth} \times \text{number of hom alt sites})^{-1/2} \tag{39}$$

The number of hom alt sites in an exome is easily several hundred, so the uncertainty in our estimate is quite small. In fact, for an average coverage of 30 reads and looking only at exonic hom alts in chromosome 1, we expect to find 50 or so hom alts, which yields a fairly small relative error.

It remains to describe how we determine which sites are hom alt. Our simple heuristic is to consider only sites with sufficient depth and fraction of alt reads, say 80 percent. (If this measure fails to find enough hom alt sites, we can conclude that contamination is greater than 20 percent, which is so large as to make the sample worthless for somatic variant calling). We must additionally reject apparent hom alt sites that are actually het sites with deletion of the ref allele. We do this by looking for copy number variation and loss of heterozygosity in the vicinity of this site. To find sites with anomalous copy number, we compute the average depth smoothed over a scale of one megabase and compare it to the average depth of the sample as a whole and reject sites that show evidence of a deletion. To find loss of heterozygosity, we obtain ref and alt read counts of common SNPs within a megabase and compare the number of hets found to that expected based on the known allele frequencies. If too few hets are found, we reject the site. These two heuristics suffice to screen out false hom alts.