Somatic Variant Calling

BIOF2014

Motivation

Building on our modeling experience in germline variant calling, we will now establish a model for somatic variant calling.

In cancer sequencing, we are interested in identifying variants that are in present in the cancer cells but not the normal cells. These are known as **somatic variants** (or mutations).

Let g_N represent the genotype in normal cells at a genomic position l, and let g_T represent the genotype in tumour cells. The joint genotypes are as follows.

Table 1: Joint genotype of a tumour-normal pair

$g_T =$	AA	AB	BB
$g_N =$ AA	wild type	somatic	somatic
AB	LOH	germline	LOH
BB	error	error	germline

LOH, loss of heterozygosity.

Since tumour cells come from normal cells, g_T is dependent on g_N, and some joint genotypes are likely to be due to sequencing errors. Therefore, calling the variants in a tumour sample jointly with its matched normal sample would yield higher accuracy than calling the genotype for normal and tumour samples separately.

Additionally, we will also enhance our previous variant calling model by accounting for whether a read is mapped correctly using its **mapping quality score**.

Somatic variant model for a single tumour

Our model will assess each genomic position l separately, so we will omit the l index.

The input consists of J reads with observed read status $\boldsymbol{X} = [X_j]$, base correctness probabilities $\boldsymbol{q} = [q_j]$, and mapping correctness probabilities $\boldsymbol{r} = [r_j]$. $X_j = 1$ if read j contains the alternative allele at genomic position l, and $X_j = 0$ otherwise. q_j represents probability that the base mapping to genomic position l is read correctly. In other words, $q_j = 1 - e_j$ where e_j is the probability of read error as defined in the previous tutorial. r_j represents the probability that read j is mapped correctly. We will treat x_j as observed random variables, while q_j and r_j are observed fixed values.

We define Z_j as the latent indicator random variable for X_j such that $Z_j = 1$ if read j truly contains the alternative allele and $Z_j = 0$ otherwise.

We further C_j as the latent indicator random variable such that $C_j = 1$ if read j is mapped correctly and $C_j = 0$ otherwise.

If read j is mapped correctly ($C_j = 1$), the likelihood is similar to the germline variant model:

$$p\left(x_j \mid Z_j = 1, C_j = 1\right) = \begin{cases} q_j & \text{if } x_j = 1 \\ 1 - q_j & \text{if } x_j = 0 \end{cases}$$
$$p\left(x_j \mid Z_j = 0, C_j = 1\right) = \begin{cases} 1 - q_j & \text{if } x_j = 1 \\ q_j & \text{if } x_j = 0 \end{cases}.$$

If read j is mapped incorrectly ($C_j = 0$), the likelihood is uniform over the domain of x_j:

$$p\left(x_j \mid Z_j = 1, C_j = 0\right) = \frac{1}{2}, \quad x_j \in \{0, 1\}.$$

The prior on the latent mapping indicator C_j is given by

$$p\left(c_j\right) = \begin{cases} r_j & \text{if } c_j = 1 \\ 1 - r_j & \text{if } c_j = 0 \end{cases}.$$

Now, the prior on the latent read indicator Z_j is a bit different from the germline variant model.

$$p\left(z_j \mid g, \mu_g\right) = \begin{cases} \mu_g & \text{if } z_j = 1 \\ 1 - \mu_g & \text{if } z_j = 0 \end{cases}.$$

In other words, μ_g is the probability that $z_j = 1$ given genotype $g \in \{0, 1, \ldots, M\}$, which represents the number of copies of the alternative allele.

Under the germline variant model, we derived a general expression for $p(z_j \mid g, m)$ based on genotype G and ploidy M. Recall that

$$p(z_j \mid g, m) = \begin{cases} \frac{g}{m} & \text{if } z_j = 1 \\ 1 - \frac{g}{m} & \text{if } z_j = 0 \end{cases}.$$

Therefore, $\mu_g = \frac{g}{m}$ for germline variants.

However, under a somatic variant model, μ_g is more difficult to define because it now depends on tumour ploidy, purity, and copy-number state at the genomic position. Rather than defining μ_g directly, we can define it probabilitistic by putting a prior on μ_g.

$$p\left(\mu_g\right) = \text{Beta}\left(\mu_g \mid a_g, b_g\right),$$

where a_g and b_g are model **hyperparameters** (i.e. tuning parameters that we can set to specific values). Hyperparameters are normally treated as fixed values.

Now, we are ready to define the prior on genotype G:

$$p\left(G = g \mid \boldsymbol{\theta}\right) = \text{Categorical}\left(g \mid \boldsymbol{\theta}\right), \quad g \in \{0, 1, \ldots, m\},$$

In turn, we put a prior on the parameter $\boldsymbol{\theta}$:

$$p\left(\boldsymbol{\theta}\right) = \text{Dirichlet}\left(\boldsymbol{\theta} \mid \boldsymbol{d}\right),$$

where $\boldsymbol{d} = [d_g]$ are hyperparameter weights.

Hence, we have extended our germline variant model to consider mapping errors and to incorporate uncertainty or variability in μ_g, as well as putting a flexible prior on the genotype G. The model that we described so far closely resembles the SNVMix2 model, though we have fixed a few errors in the latter model.

Somatic variant model for a tumour-normal sample pair

Previously, we defined the unknown genotype parameter G as the number of copies of the alternative allele in the genome of the sample at position l. Now, we have a tumour-normal pair, and we need to formally define G to correspond to the joint genotype of tumour-normal pair (Table 1).

Therefore, define joint genotype set as

$$\begin{aligned} \mathcal{S} = \{&(AA, AA), (AB, AA), (BB, AA), \\ &(AA, AB), (AB, AB), (BB, AB), \\ &(AA, BB), (AB, BB), (BB, BB)\}, \end{aligned}$$

where we enumerate the joint genotype table in column-major order. Then, random variable $G = k$ corresponds to the kth element of $\mathcal{S}$.

Finally, we can define $\boldsymbol{d}$ such that more probable joint genotypes are given higher weights, and d_k also corresponds to the kth element of $\mathcal{S}$.

Hence, we have a model similar to the JointSNVMix2 model.

Posterior

For variant calling, we would want to solve for the posterior distribution $p(g \mid \boldsymbol{x})$ for the random variable G with support $\mathcal{G}$, which can be derived using Bayes' theorem.

$$p(g \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid g)\ p(g)}{p(\boldsymbol{x})}.$$

The prior $p(g)$ is not immediately available from our model definition, so we need to derive it by

$$\begin{aligned} p(G) &= \int_{\boldsymbol{\Theta}} p(G \mid \boldsymbol{\theta})\ p(\boldsymbol{\theta})\ d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\Theta}} \boldsymbol{\theta}\, p(\boldsymbol{\theta})\ d\boldsymbol{\theta} \\ &= \mathrm{E}(\boldsymbol{\theta}) \quad \text{(definition of expectation)}. \end{aligned}$$

Therefore,

$$p(g) = \mathrm{E}(\theta_g)$$

Under the Dirichlet distribution, the expectation is given by

$$\mathrm{E}(\theta_g) = \frac{d_g}{\sum_k d_k}.$$

On the other hand, the model evidence $p(\boldsymbol{x})$ is given by

$$p(\boldsymbol{x}) = \sum_{g \in \mathcal{G}} p(\boldsymbol{x} \mid g)\ p(g).$$

Now, we need to derive the following quantities:

$$\begin{aligned} p(\boldsymbol{x} \mid g) &= \prod_j p(x_j \mid g) \\ p(x_j \mid g) &= \int_0^1 p(x_j \mid g, \mu_g)\, p(\mu_g)\ d\mu_g \\ p(x_j \mid g, \mu_g) &= \sum_{c=0}^{1} \sum_{z=0}^{1} p(x_j, c, z \mid g, \mu_g). \end{aligned}$$

Questions

1. Verify that each probability distribution that we define is proper (i.e. the distribution satisfies Kolmogorov's axiom 1 and 2).
2. Assign reasonable values to hyperparamter $\boldsymbol{d}$ in the somatic model for a single tumour.
3. Using the information in Table 1, assign appropriate values to the hyperparameter $\boldsymbol{d}$ in the somatic model for tumour-normal pair.
4. What does Table 1 assume about the ploidy of the tumour?
5. If you were to explicitly show the genomic position index l, which variables would be indexed by l?
6. If you were to explicity show the sample index i, which variables would be indexed by i?