Germline Variant Calling, Revisited

BIOF2014

There are many types of variants that can occur in the DNA. We will focus on the simplest and most abundant type of variants known as **single-nucleotide variants** (SNV), which involves the change of a single nucleotide in the DNA (e.g. A>T, G>A).

In SNV calling, we are only interested in positions with a alternative alleles, the input to the mutation calling model would be the reads and their quality scores at selected positions.

The bioinformatic tool `bcftools` implements a simple statistical model for SNV calling, as detailed in Li 2011. Let's derive a Bayesian model for SNV calling.

Data

The data consists of the read $j \in \{1 \dots J\}$ that overlap with a specific genomic position l. Our model will analyze each genomic position separately, so we will omit index l from our model.

X_j represents an observed random variable indicator for whether read j contains the alternative allele at the base that aligns to genomic position l. If read j contains the alternative allele, $X_j = 1$; otherwise, $X_j = 0$. Let $\boldsymbol{X} = [X_j]$ be the vector of read indicators.

$e_j \in [0, 1]$ represents the probability of read error for read j at the base that aligns to position l. We treat e_j as an known fixed value.

Haploid model

Our goal is to infer the unknown genotype G at position l. Let's first consider a simple haploid model. Let m represent the ploidy (i.e. the total number of alleles).

We define the unknown random variable $G = 0$ if the genotype is reference, and $G = 1$ if the genotype is alternative.

The likelihood is

$$\begin{aligned}
p\left(x_j \mid G=0\right) &= \begin{cases} 1-e_j & \text{if } x_j = 0 \\ e_j & \text{if } x_j = 1 \end{cases} &&= (1-e_j)^{1-x_j} e_j^{x_j} \\
p\left(x_j \mid G=1\right) &= \begin{cases} e_j & \text{if } x_j = 0 \\ 1-e_j & \text{if } x_j = 1 \end{cases} &&= e_j^{1-x_j} (1-e_j)^{x_j}.
\end{aligned}$$

Since the reads are independent,

$$p\left(\boldsymbol{x} \mid g\right) = \prod_j p\left(x_j \mid g\right).$$

Let's start with an uniform prior:

$$p\left(g\right) = \frac{1}{2}.$$

Diploid model

Indeed, many organisms including humans are diploid. So, let's derive a SNV calling model for diploids ($m = 2$).

At position l, let $G \in \{0, 1, 2\}$ represent the diploid genotype;

- $G = 0$ for homozygous reference
- $G = 1$ for heterozygous
- $G = 2$ for homozygous alternative

The homozygous genotypes have the same likelihoods as in the corresponding genotypes under the haploid model:

$$\begin{aligned}
p\left(x_j \mid G=0\right) &= \begin{cases} 1-e_j & \text{if } x_j = 0 \\ e_j & \text{if } x_j = 1 \end{cases} \\
p\left(x_j \mid G=2\right) &= \begin{cases} e_j & \text{if } x_j = 0 \\ 1-e_j & \text{if } x_j = 1 \end{cases}.
\end{aligned}$$

As for the heterozygous genotype, observing the reference allele is equal probable to observing alternative allele. Therefore,

$$p\left(x_j \mid G=1\right) = \begin{cases} \frac{1}{2} & \text{if } x_j = 0 \\ \frac{1}{2} & \text{if } x_j = 1 \end{cases}.$$

A uniform prior on $G \in \{0, 1, 2\}$ would be

$$p\left(g\right) = \frac{1}{3}.$$

General model

Let's extend our model to support any ploidy $M = 1, 2, 3, 4, ...$

We now define $G \in \{0, ... M\}$ more generally as the number of copies of the alternative alleles at a specific position l in the genome.

To help us think through the problem, we introduce a **latent** (i.e. unknown and unobserved) random variable Z_j that represents whether read j truly correspond to the alternative allele. The **observed** random variable X_j represents whether read j is observed to correspond to the alternative allele, and Z_j represents the unknown underlying truth.

Similar to the haploid model, if we know Z_J, then,

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

Applying the rules of probability,

$$\begin{aligned} p\left(x_j \mid g, m\right) &= \sum_{z_j \in \{0,1\}} p\left(x_j, z_j, g, m\right) && \text{(justification?)} \\ &= \sum_{z_j \in \{0,1\}} p\left(x_j \mid z_j\right) p\left(z_j \mid g, m\right) && \text{(justification?)} \end{aligned}$$

Now, we need to derive $p\left(z_j \mid g, m\right)$.

Similar to the haploid model, $p\left(z_j \mid g, M = 1\right)$ is given by

$$\begin{aligned} p\left(z_j = 0 \mid G = 0, M = 1\right) &= 1 \qquad & p\left(z_j = 1 \mid G = 0, M = 1\right) &= 0 \\ p\left(z_j = 0 \mid G = 1, M = 1\right) &= 0 \qquad & p\left(z_j = 1 \mid G = 1, M = 1\right) &= 1. \end{aligned}$$

Similar to the diploid model, $p\left(z_j \mid G, M = 2\right)$ is given by

$$\begin{aligned} p\left(z_j = 0 \mid G = 0, M = 2\right) &= 1 \qquad & p\left(z_j = 1 \mid G = 0, M = 2\right) &= 0 \\ p\left(z_j = 0 \mid G = 1, M = 2\right) &= \frac{1}{2} \qquad & p\left(z_j = 1 \mid G = 1, M = 2\right) &= \frac{1}{2} \\ p\left(z_j = 0 \mid G = 2, M = 2\right) &= 0 \qquad & p\left(z_j = 1 \mid G = 2, M = 2\right) &= 1. \end{aligned}$$

Do you recognize the pattern?

$$\begin{aligned} p\left(z_j = 0 \mid G = 0, M = 3\right) &= & \qquad p\left(z_j = 1 \mid G = 0, M = 3\right) &= \\ p\left(z_j = 0 \mid G = 1, M = 3\right) &= & \qquad p\left(z_j = 1 \mid G = 1, M = 3\right) &= \\ p\left(z_j = 0 \mid G = 2, M = 3\right) &= & \qquad p\left(z_j = 1 \mid G = 2, M = 3\right) &= \\ p\left(z_j = 0 \mid G = 3, M = 3\right) &= & \qquad p\left(z_j = 1 \mid G = 3, M = 3\right) &= . \end{aligned}$$

After deriving $p(z_j \mid g, m)$ based on our understanding of genetics, we can now derive a closed-form expression for the likelihood $p\left(x_j \mid g, m\right)$.

Questions

Suppose that we are doing germline variant calling on human whole genomes.

1. Is a uniform prior on G appropriate? We know that a typical human genome differs from the reference genome at 4.1 to 5.0 million sites 2. We also know that the heterozygosity ratio is 1.36 - 1.73 3.
2. Derive a general expression for $p(z_j \mid g, m)$.
3. Derive $p(x_j \mid g, m)$.
4. Derive $p(g \mid x, m)$.
5. Derive an maximum likelihood estimator for g.
6. Derive an maximum *a posteriori* estimator for g.