Germline Variant Calling

BIOF2014

Short-read sequencing

High-throughput short-read sequencing data consists of millions of sequencing reads. Each read consists of a **base sequence** of a fixed length (50-150 bp), such as

```
TAGTAAAACTTGTTTTATAAGTTCCGTTAAGCACACCCAGTCAGAAAATG
```

(If the reads have been trimmed, their lengths may be variable.)

Each read is also associated with a **base quality score sequence**:

```
BBBFFFFFFFFFFFFFFFIIFIFIIIFIIFFIIIIIII<FFFBFIIFIIF
```

The base qualities are represented by ASCII characters. Different sequencing vendors have now standardized on the Phred+33 scale, which assigns a quality score of 0 to `!` (which has an integer value of 33 in the ASCII table). Additionally, quality scores of 32, 33, 34, … to `A`, `B`, `C`, … respectively (See FASTQ format for details). The quality score Q is defined on a Phred scale, which means that

$$Q = -10 \log_{10}(e),$$

where e is the **probability of base read error**.

Sequence alignment

In reference-based sequencing analysis, the short reads are aligned to the reference genome using a short-read aligner such as BWA-MEM. After alignment and pile-up, we can count the numbers of reference and alternative alleles at each genomic position. To simplify the task, only the reference allele and the most frequent non-reference allele would be considered, and we refer to the latter as the alternative allele.

```
reference
pos   12345678901234567890123456789012345678901234567890123456789012345678901234567890123
seq   TAGTAAAACTTGTTTTATAAGTTCCGTTAAGCACACCCAGTCAGAAAATGGCAATAAAGACCATATAGATGAG

alignment
reads TAGTAAAACTTGTTTTATAAGTTCCGTTAAGCACACCCAGTCAGAAAATG
        GTAAAACTTGTTTTATATGTTCCGTTAAGCACACCCAGTCAGAAAATGAC
          AAAACTTGTTTTATATGTTCCGTTAAGCACACCCAGTCAGAAAATGACAAT
            AACTTGTTTTATATGTTCCGTTAAGCACACCCAGTCAGAAAATGACAATAA
                 GTTTTATAAGTTCCGTTAAGCACACCCAGTCAGAAAATGGCAATAAAGACC
                      ATATGTTCCGTTAAGCACACCCAGTCAGAAAATGACAATAAAGACCATATA
                        AAGTTCCGTTAAGCACACCCAGTCAGAAAATGACAATAAAGACCATATAGA
                           TTCCGTTAAGCACACCCAGTCAGAAAATGGCAATAAAGACCATATAGATGA
                                  AAGCACACCCAGTCAGAAAATGACAATAAAGACCATATAGATGAG

allele counts
ref   11223344444555556673788888889999999999999999999999287776655555444443322 21
alt   00000000000000000004000000000000000000000000000000600000000000000000000 00
```

Variant calling

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.

Additionally, each aligned read is also given an **mapping quality score** by the aligner, which is a Phred scale score correpsonding to the probability that a read is mapped/aligned to the incorrect position.

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:

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

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,

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

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

1. Which distribution are the likelihoods in the models above based on?
2. Why is e_j always written in lower case?
3. Why don't we need to condition the likelihood on e_j?
4. Is a uniform prior on G appropriate? Can we do better?
5. Under the general model, which probability rule was applied at each step in the derivation of $p\left(x_j \mid g, m\right)$?
6. Derive a general expression for $p(z_j \mid g, m)$.
7. Derive $p(g \mid x, m)$.