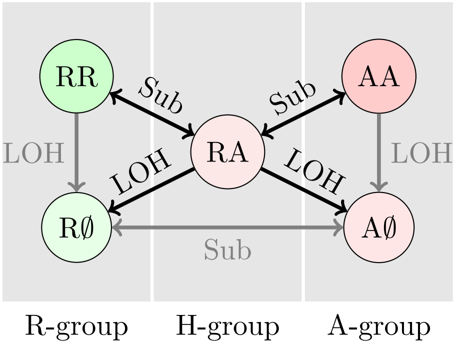
Lab rotation: theories (draft)

Last modified on June 17, 2022

Model

- There are only two types of mutation: point substitution and loss of heterozygosity (LOH).
- A cell has exactly one (in case a LOH event has occurred) or two copies (in case no LOH event has occurred) of any locus.
- A copy of a locus is either the reference (R) or the alternative (A) allele. The other two nucleotides are excluded from data.
- At most one mutation can occur at any locus.
- The genotype at the root of the tree is always one without LOH.



R: reference allele; A: alternative allele; ∅: lost allele;
Sub: substitution; LOH: loss of heterozygosity

Based on the model above, the genotype of a cell at any locus must be one of five: RR, R∅, AA, A∅ or RA. It is worth noting that a substituion that brings the genotype from R∅ to A∅ or vice versa will never have a chance to occur, because one mutation is required to reach the genotype R∅ or A∅ at first place, and a second mutation is not allowed by the model.

Mutation detection

For the purpose of mutation detection, each genotype is assigned to one of three groups: R (reference only), A (alternative only) and H (heterozygous). This is because, due to limitations of the sequencing method, copy number changes are much harder to detect compared to mutations that introduces a new allele or removes an allele completely. As a result, the genotypes RR and R$\emptyset$ are deemed indistinguishable and so are AA and A$\emptyset$.

The mutation state S of a locus is denoted by a two-letter code. The first letter is the genotype group of the root, while the second is the genotype group of cells affected by a mutation. If no detectable mutation occurs at this locus, the second letter is the same as the first. For example, if the root has genotype RR (group R) and a substitution converts some of the cells into genotype RA (group H), then the mutation state of this locus is represented by the code "RH". With three genotypes groups, there are nine possible mutation states. However, two of them, namely RA and AR, will never occur since conversions between these groups R and A require at least two mutations. Hence we are left with three states with a mutation (RR, AA and HH) and four states without a mutation (RH, HR, AH and HA).

Given a locus, a cell and the corresponding genotype group $G \in \{\text{R, A, H}\}$, the probability of an observation $D = (n_R, n_A)$ is assumed to follow a beta-binomial distribution. Hence, the likelihood of G given observation D is

$$\mathcal{L}(G \mid D) = P(D \mid G) = \text{BetaBin}(n_R \mid n_A + n_R, f_G, \omega_G)$$

where $f_G = \frac{\alpha}{\alpha+\beta}$ is the expected frequency of the reference read among all reads, and $\omega_G = \alpha + \beta$ controls the uncertainty of f. Suppose the sequencing method has been tested on known sequences of total length n and showed an an error rate of ε. Assuming that an error leads to the three wrong reads with equal chance, we get

$$f_\text{R} = \frac{1 - \varepsilon}{1 - \frac{2}{3}\varepsilon}$$

$$f_\text{A} = \frac{\frac{1}{3}\varepsilon}{1 - \frac{2}{3}\varepsilon} = 1 - f_\text{R}$$

and

$$\omega_\text{R} = \omega_\text{A} = n + 2$$

For the heterozygous case $G = \text{H}$, the situation is a bit more complicated. First, since the two alleles are present in equal amount, their corresponding read counts should also be equal, hence

$$f_\text{H} = \frac{1}{2}$$

Second, before being sequenced,the two alleles need to be amplified. This step introduces additional variance to the number of reads of the two alleles, so ω_H needs to be adjusted accordingly. For now, $\omega_H = \frac{1}{2}\omega_R$ is used to test the codes.

Let N be the total number of cells, $\mathbf{G} = (G_1, ..., G_N)$ be the genotypes and $\mathbf{D} = \{D_1, ..., D_N\}$ be the observations. Since the cells are sequenced separately, the distribution of observations should depend only on the corresponding genotype. Hence, when $\mathbf{G}$ is given, individual observations are treated as independent on each other:

$$\mathcal{L}(\mathbf{G} \mid \mathbf{D}) = P(\mathbf{D} \mid \mathbf{G}) = \prod_{i=1}^{N} P(D_i \mid G_i)$$

To detect loci that likely contain a mutation, we want to find the posterior probabilities for each mutation state. Let $S = XX'$, where $X, X' \in \{\text{R, H, A}\}$. When $X = X'$ (without mutation), we have

$$P(S = XX \mid \mathbf{D}) = \frac{\mathcal{L}(S = XX \mid \mathbf{D}) P(S = XX)}{P(\mathbf{D})}$$

Since $S = XX$ is equivalent to all genotypes being X, we have

$$\mathcal{L}(S = XX \mid \mathbf{D}) = \mathcal{L}(\forall i : G_i = X \mid \mathbf{D}) = \prod_{i=1}^{N} P(D_i \mid G_i = X)$$

When $X \neq X'$ (with mutation), the posterior can be obtained by marginalizing over the number of cells that are affected by the mutation. To do this, first define $\ell(n, k)$ as the likelihood that k of the first n cells are affected given the first n observations:

$$\ell(n, k) = P\left(\mathbf{D}^{(n)} \middle| \sum_{i=1}^{n} \delta_{G_i, X'} = k\right)$$

where $\mathbf{D}^{(n)} = \{D_1, .., D_n\}$ is the first n observations and δ is the Kronecker delta.

When $n = 1$, $\ell(1, 1) =$ and.

Then we have

$$\begin{aligned} P(S = XX' \mid \mathbf{D}) &= \sum_{k=1}^{N} P\left(\sum_{i=1}^{N} \delta_{G_i, X'} = k \middle| \mathbf{D}\right) \\ &= \frac{1}{P(\mathbf{D})} \sum_{k=1}^{N} \ell(N, k) P\left(\sum_{i=1}^{N} \delta_{G_i, X'} = k\right) \end{aligned}$$

$$
\begin{aligned}
\ell(n+1,k) &= P\left(\mathbf{D}^{(n+1)}\middle|\sum_{i=1}^{n+1}\delta_{G_i,X'}=k\right)\\
&= \sum_{\sum_{i=1}^{n+1}\delta_{G_i,X'}=k} P\left(\mathbf{D}^{(n+1)}\middle|\mathbf{G}^{(n+1)}\right)P\left(\mathbf{G}^{(n+1)}\middle|\sum_{i=1}^{n+1}\delta_{G_i,X'}=k\right)\\
&= \frac{1}{\binom{n+1}{k}}\sum_{\sum_{i=1}^{n+1}\delta_{G_i,X'}=k} P\left(\mathbf{D}^{(n+1)}\middle|\mathbf{G}^{(n+1)}\right)\\
&= \frac{1}{\binom{n+1}{k}}\left(\sum_{\substack{G_{n+1}=X\\ \sum_{i=1}^{n}\delta_{G_i,X'}=k}} P\left(\mathbf{D}^{(n+1)}\middle|\mathbf{G}^{(n+1)}\right)+\sum_{\substack{G_{n+1}=X'\\ \sum_{i=1}^{n}\delta_{G_i,X'}=k-1}} P\left(\mathbf{D}^{(n+1)}\middle|\mathbf{G}^{(n+1)}\right)\right)\\
&= \frac{1}{\binom{n+1}{k}}\Big(\ell(n,k)P(D_{n+1}|G_{n+1}=X)+\ell(n,k-1)P(D_{n+1}|G_{n+1}=X')\Big)
\end{aligned}
$$

Underflow problem

When a large number of probabilities are multiplied together, the result can be so small that the floating point number underflows to zero. This problem is bypassed by using the log-scale notation, i.e. denoting the numbers by their logarithms. Then the calculation of the product becomes addition of logarithms of individual probabilities.

When marginalizing over all possible genotype profiles, the joint probabilities of each genotype profile need to be added up. Let p_1 and p_2 be two very small probabilities. To avoid underflow, p_1 and p_2 are represented by their logarithms $q_1 = \log p_1$ and $q_2 = \log p_2$, respectively. Then the logarithm of the sum of p_1 and p_2 can be calculated as follows

$$
\begin{aligned}
\log(p_1+p_2) &= \log(e^{q_1}+e^{q_2})\\
&= \log(e^{q_1}(1+e^{q_2-q_1}))\\
&= q_1+\log(1+e^{q_2-q_1})
\end{aligned}
$$

This method avoids directly calculating p_1 and p_2 and reduces the risk of underflow. Since q_1 and q_2 tend to be similar, $e^{q_1-q_2}$ will usually not underflow. Even if $e^{q_1-q_2}$ underflows, it will have little effect on the final result because in this case p_2 is much smaller than p_1 and hence we have $p_1+p_2 \approx p_1$. The only problem comes when $e^{q_1-q_2}$ is too large that it overflows. It remains to be assessed how often and under which conditions this situation may occur.