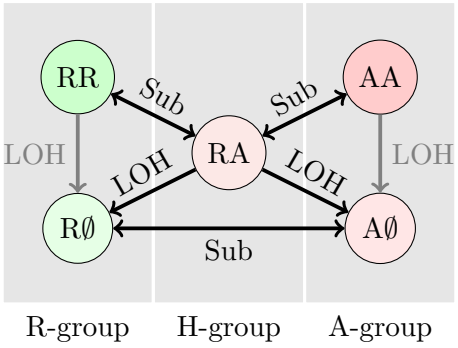
Lab rotation: theories (draft)

June 1, 2022

1 Mutation detection

The loci are treated as independent. The genotype of a cell at any locus is assigned to one of three groups: R (reference only), A (alternative only) and H (heterozygous). This grouping is dependent on which alleles are present in the cell and does not take copy numbers into account. More specifically, it is assumed that each locus has exactly one or two alleles in any cell. This means, while an H-group cell is guaranteed to have one reference and one alternative allele, an R- or A-group cell can possess either one or two of the same allele.



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

Given a locus, a cell and the corresponding observation, the likelihood of the number of reference reads given a genotype G follows a beta-binomial distribution. Let $D = \{n_R, n_A\}$ be the numbers of reference and alternative reads, respectively. Then the likelihood is

$$P(D \mid G) = \text{BetaBin}(n_R, n_A + n_R, f\omega, \omega - f\omega)$$

where $f = \frac{\alpha}{\alpha+\beta}$ and $\omega = \alpha + \beta$ are specific for each G. Let N be the number of cells, then the joint likelihood of all observations at this locus is

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

For any locus, we want to find the probability that a mutation exists. Specifically, we only consider mutations that move the cell from one genotype group to another, namely LOH in heterozygous cells and substitutions. Other mutations, especially copy number changes (e.g. LOH in homozygous cells), are harder to detect using the data available and are therefore ignored for now.

Given the observations $\mathbf{D}$ at a locus, the probability that a mutation eixsts here is equal to one minus the probability that no mutation exists, i.e. that all cells are of the same genotype group. Since we assumed that all cells are independent, we have

$$\begin{aligned} P(\text{mutation} \mid \mathbf{D}) &= 1 - P(\forall i : G_i = \text{R} \mid \mathbf{D}) - P(\forall i : G_i = \text{H} \mid \mathbf{D}) - P(\forall i : G_i = \text{A} \mid \mathbf{D}) \\ &= 1 - \sum_{G_0 \in \{\text{R, H, A}\}} \frac{P(\mathbf{D} \mid \forall i : G_i = G_0) P(\forall i : G_i = G_0)}{P(\mathbf{D})} \\ &= 1 - \sum_{G_0 \in \{\text{R, H, A}\}} \frac{P(\forall i : G_i = G_0)}{P(\mathbf{D})} \prod_{i=1}^{N} P(D_i \mid G_i = G_0) \end{aligned}$$

While we already have a model for $P(D_i \mid G_i = G_0)$, it remains a question how we can estimate the values of $P(\forall i : G_i = G_0)$ and $P(\mathbf{D})$.

If we assume that genotypes of the cells are independent given the observations, we get

$$\begin{aligned} P(\text{mutation} \mid \mathbf{D}) &= \sum_{G_0 \in \{\text{R, H, A}\}} \prod_{i=1}^{N} P(G_i = G_0 \mid D_i) \\ &= \sum_{G_0 \in \{\text{R, H, A}\}} \prod_{i=1}^{N} \frac{P(D_i \mid G_i = G_0) P(G_i = G_0)}{P(D_i)} \end{aligned}$$

In this case, $P(G_i = G_0)$ is easier to estimate than $P(\forall i : G_i = G_0)$, and once it is known for all G_0 values, the normalizing factor $P(D_i)$ is also easy to obtain.
However, the assumption that the genotypes are independent might not be plausible.