Lab rotation: theories

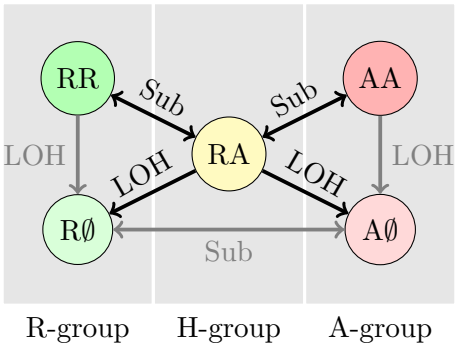
Xiaoyu Sun

Last modified on August 17, 2022

Model

The model is based on the following assumptions:

- 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. Reads of the other two nucleotides are excluded from data.
- At most one mutation can occur at any locus.
- The root of the tree contains no LOH.



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

With the assumptions above, the genotype of a cell at any locus must be one of five: RR, RØ, AA, AØ and RA. It is worth noting that not all possible mutations in this framework are taken into consideration. First, a substituion that brings the genotype from RØ to AØ or vice versa will never have a chance to occur, because one mutation is required to reach RØ or AØ at first place, and a second mutation is not allowed by the

model. Second, given the coverage, copy number changes has no effect on the distribution of the observation. Hence LOH events on homozygous loci are deemed undetectable.

The data include an observation $D = (n_R, n_A)$ for each cell and locus, where n_R and n_A are read counts of the reference and alternative allele, respectively. Given the coverage $n_R + n_A$, the observation is assumed to follow a beta-binomial distribution [1]:

$$P(D \mid G) = \text{BetaBin}(n_R \mid n_R + n_A, \alpha, \beta)$$

where α and β are defined by $f_G = \frac{\alpha}{\alpha+\beta}$, the expected frequency of reference reads, and $\omega_G = \alpha + \beta$, which controls the uncertainty of f.

The values of f_G and ω_G depend on the genotype G. For convenience, the genotypes are assigned to three groups: R (reference only), A (alternative only) and H (heterozygous). Suppose that (1) the sequencing method has been tested on known sequences of total length n and showed an error rate of ε and that (2) errors lead to the other three nucleotides with equal chance. When G is homozygous ($G \in \{\text{R}, \text{A}\}$), all reads except the erroneous ones should be of that allele. Hence

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

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

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 on average, 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.

Mutation detection

Given a locus, let N be the total number of cells, $\mathbf{G} = (G_1, ..., G_N)$ be their genotypes at this locus and $\mathbf{D} = \{D_1, ..., D_N\}$ be the observations. The distribution of an observation depends only on the genotype of the corresponding cell at that locus. Therefore, when $\mathbf{G}$ is given, individual observations are independent on each other:

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

To detect loci that likely contain a mutation, we want to find the posterior probabilities for each mutation state. A mutation state S can be denoted by a tuple XX',

where $X, X' \in \{\text{R, H, A}\}$ are genotypes before and after the mutation, respectively. If no mutation occurs ($X = X'$), we have

$$P(S = XX \mid \mathbf{D}) = \frac{1}{P(\mathbf{D})}\mathcal{L}(S = XX \mid \mathbf{D})P(S = XX)$$

where

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

In the presence of a mutation ($X \neq X'$), first define $\ell(n, k)$ as the likelihood that k of the first n cells are mutated:

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

where $\mathbf{D}^{(n)} = \{D_1, .., D_n\}$ is the first n observations and δ is the Kronecker delta. Then the posterior can be expressed as a function of $\ell(N, k)$ by marginalizing over the number of mutated cells:

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

To get the values of $\ell(N, k)$, note that when $n \geq 1$, $\ell(n, k)$ can be calculated recursively from $\ell(n-1, k)$ and $\ell(n-1, k-1)$:

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

In the case $n = 0$, since the number of mutated cells is always 0, we have $\ell(0, 0) = 1$ and $\ell(0, k) = 0$ for all $k \geq 1$.

Assuming that all trees have an equal chance to occur, the prior that k cells are affected by a mutation is given by [1]

$$P\left(\sum_{i=1}^{N} \delta_{G_i, X'} = k\right) = \frac{\binom{N}{k}^2}{(2k-1)\binom{2N}{2k}}$$

After the likelihoods and pirors are obtained, they can be multiplied and normalized to get the posteriors. The loci are selected according to the posteriors with the threshold $1 - \frac{1}{L}$, where L is the total number loci, so that the expected number of false positives is less than 1.

LOH detection

LOH differs from point substitution in that it simultaneously affects multiple consecutive loci. This means a cell must be either heterozygous or homozygous for all loci affected by the same LOH. Let $\mathbf{G}_1$ and $\mathbf{G}_2$ be genotypes of the cells at two mutated loci with states HX and HX', where $X, X' \in \{\text{R, A}\}$. If the two mutations are produced by a single LOH, then we have $G_{1,i} = X$ and $G_{2,i} = X'$ if and only if the i-th cell is affected by that LOH. In the remaining text, this situation is referred to as *correlated*.

Similar to mutation detection, the posterior that two mutated loci are correlated can be obtained by marginalizing over n_{homo}, the number of cells that are homozygous at both loci, which is equivalent to the number of affected cells in case of LOH. The only difference is that $n_{\text{homo}} = N$ doesn't need to be considered, because in that case the locus would have appeared as not mutated, and been filtered out during mutation detection. Let $\mathbf{D}_1$, $\mathbf{D}_2$ be the observations at the two loci, then

$$P(\text{corr.} \mid \mathbf{D}_1, \mathbf{D}_2) = \frac{1}{P(\mathbf{D}_1, \mathbf{D}_2)} \sum_{k=1}^{N-1} P(\mathbf{D}_1, \mathbf{D}_2 \mid n_{\text{homo}} = k) P(n_{\text{homo}} = k)$$

The posterior that two loci are independet is easy to obtain since the likelihood is given simply by the product of individual likelihoods:

$$P(\text{ind.} \mid \mathbf{D}_1, \mathbf{D}_2) = \frac{1}{P(\mathbf{D}_1, \mathbf{D}_2)} P(\mathbf{D}_1 \mid S_1) P(\mathbf{D}_2 \mid S_2) P(\text{ind.})$$

Once the posteriors are known, a LOH should appear as a section of loci in which all pairs have high posterior for correlation. But since correlation is transitive, the posterior needs only to be calculated for loci that neighbour each other.

It is worth mentioning that correlation does not imply the presence of a LOH, but can also be a result of mutations (be it point substitution or LOH) occurring on the same edge of the phylogenetic tree. Even if there is no LOH, a pair of loci should still have a chance of $1/(2N - 2)$ to be correlated since the tree has $2N - 2$ internal edges. Under the current model, there is no way to distinguish coincidental correlations from LOH. However, the chance of such coincidences shrinks exponentially as the section of correlated loci grows longer.

Tree inference

For tree T:

$$P(\mathbf{D} \mid T) = \prod_{i=1}^{L} \prod_{j=1}^{N} P(D_{ij} \mid G_{ij})$$

where $G_{ij} = S_{i2}$ when cell j is below mutation i in tree T, and $G_{ij} = S_{i1}$ otherwise. The likelihood ratio of mutation i is

$$\prod_{j=1}^{N} P(D_{ij} \mid G_{ij})$$

Likelihood ratio of mutation i above node v:

$$\begin{aligned}\lambda_{iv} &= \prod_{j:G_{ij}=S_{i1}} P(D_{ij} \mid S_{i1}) \prod_{j:G_{ij}=S_{i2}} P(D_{ij} \mid S_{i2}) \Big/ \prod_{j} P(D_{ij} \mid S_{i1}) \\ &= \prod_{j:G_{ij}=S_{i2}} \frac{P(D_{ij} \mid S_{i2})}{P(D_{ij} \mid S_{i1})}\end{aligned}$$

Let v_L, v_R be the two children of v, then we have

$$\lambda_{iv} = \lambda_{iv_L} \cdot \lambda_{iv_R}$$

This can be utilized to find out the maximum-likelihood location of any mutation.

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 denoting the numbers by their logarithms.

In the log space, multiplication becomes simple addition but addition requires a bit more calculation. Let p_1 and p_2 be two very small probabilities. W.l.o.g., let $p_1 \geq p_2$. To avoid underflow, they are represented by their logarithms, $q_1 = \log p_1$ and $q_2 = \log p_2$. The logarithm of the sum of p_1 and p_2 is then given by

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

which avoids directly calculating p_1 and p_2. Since $q_2 - q_1 \leq 0$, $e^{q_2 - q_1}$ will not overflow. When $e^{q_2 - q_1}$ underflows, it has little effect on the final result because in this case it must be that $p_2 \ll p_1$ and hence $p_1 + p_2 \approx p_1$.

References

[1] Singer, J., Kuipers, J., Jahn, K. *et al.* Single-cell mutation identification via phylogenetic inference. *Nat Commun* **9**, 5144 (2018). https://doi.org/10.1038/s41467-018-07627-7