Homework2

Bowei Xiao

20/09/2019

Calculate Posterior Porbability

Define variables just as what we did in the lecture: Let $\mathbf{D} = \{\mathbf{D_1}, \mathbf{D_2}, \cdots \mathbf{D_n}\}$ be observed genotypes ($\mathbf{D_i} \perp\!\!\!\perp \mathbf{D_j}$, $\forall i \neq j$)on one specific genomic position from each reads passing that specific location. Let $\mathbf{R}$ be the reference genotype on this very same position, and $\mathbf{G}$ be underlined true genotype at this same location. Using the above settings, what we are interested in can be expressed as $P(\mathbf{G}|\mathbf{D}, \mathbf{R})$. Based on Baye's Rule, this can be rephrased as the following:

$$\begin{aligned} P(\mathbf{G}|\mathbf{D}, \mathbf{R}|) &= \frac{P(\mathbf{D}|\mathbf{G}, \mathbf{R}) * P(\mathbf{G}|\mathbf{R}))}{P(\mathbf{D}|\mathbf{R})} \\ &= \frac{P(\mathbf{D}|\mathbf{G}) * P(\mathbf{G}|\mathbf{R})}{P(\mathbf{D}|\mathbf{R})} \end{aligned}$$

given that $\mathbf{D}$ and $\mathbf{R}$ are conditional independent on $\mathbf{G}$.
Now, for each postion we want to calculate the probability of the true genotye ($\mathbf{G}$) being xx (0 copies of reference allele), xy (1 reference allele), and yy (2 reference allele) given the data.

1. To calculate the first case where $P(\mathbf{G} = xx|\mathbf{D}, \mathbf{R} = \mathbf{y}) = \frac{P(\mathbf{D}|\mathbf{G}=xx)*P(\mathbf{G}=xx|\mathbf{R}=\mathbf{y})}{P(\mathbf{D}|\mathbf{R}=\mathbf{y})}$. These three parts can be calculated separately as below:

   - $$\begin{aligned} P(\mathbf{D}|\mathbf{G} = xx) &= \prod_{i=1}^{n} P(D_i|\mathbf{G} = xx) \\ &= \prod_{\{i|D_i = x\}} P(D_i|\mathbf{G} = xx) * \prod_{\{i|D_i \neq x\}} P(D_i|\mathbf{G} = \mathbf{xx}) \end{aligned}$$

     The first probability is simply $(1 - \epsilon)$ where as the second probability is ϵ, assuming ϵ being the error made when genotyping on the read.

   - $P(\mathbf{G} = \mathbf{xx}|\mathbf{R} = \mathbf{y})$ is simply the frequency of non-reference allele squared ($(1 - \rho)^2$) based on Hardy-Weinberg Equilibirum (HWE).

   - $P(\mathbf{D}|\mathbf{R} = \mathbf{y})$: This is a scaling constant and does not depend on $\mathbf{G}$, we noted it as c for now. This can be solved (by computer of course) using the fact that $P(\mathbf{G} = xx|\mathbf{D}, \mathbf{R} = \mathbf{y}) + P(\mathbf{G} = xy|\mathbf{D}, \mathbf{R} = \mathbf{y}) + P(\mathbf{G} = yy|\mathbf{D}, \mathbf{R} = \mathbf{y}) = 1$

   To sum up, the probability that $P(\mathbf{G} = xx|\mathbf{D}, \mathbf{R} = \mathbf{y}) = \frac{(1-\rho)^2}{c} \prod_{\{i|D_i = x\}} (1 - \epsilon) \prod_{\{i|D_i \neq x\}} \epsilon$.

2. $P(\mathbf{G} = xy|\mathbf{D}, \mathbf{R} = \mathbf{y}) = \frac{P(\mathbf{D}|\mathbf{G}=xy)*P(\mathbf{G}=xy|\mathbf{R}=\mathbf{y})}{P(\mathbf{D}|\mathbf{R}=\mathbf{y})}$

   - $$\begin{aligned} P(\mathbf{D}|\mathbf{G} = xy) &= \prod_{i=1}^{n} P(D_i|\mathbf{G} = xy) \\ &= \prod_{i|D_i \in x,y} P(D_i|\mathbf{G} = xy) * \prod_{i|D_i \notin x,y} P(D_i|\mathbf{G} = xy) \end{aligned}$$

     The first probability is $\frac{1}{2}(1 - \epsilon) + \frac{\epsilon/3}{2}$ and the second probability is $\frac{\epsilon}{3}$.

- $P(\mathbf{G} = \mathbf{xy} | \mathbf{R} = \mathbf{y})$ is simply $2 * \rho * (1 - \rho)$ based on HWE.
- $P(\mathbf{D} | \mathbf{R} = \mathbf{y})$: as before, this is a constant c.

To sum up, the probability that $P(\mathbf{G} = xy | \mathbf{D}, \mathbf{R} = \mathbf{y}) = \frac{2 * \rho * (1-\rho)}{c} \prod_{\{i | D_i \in x, y\}} \frac{1 - 2/3\epsilon}{2} * \prod_{\{i | D_i \notin x, y\}} \frac{\epsilon}{3}$

3. The last probability $P(\mathbf{G} = yy | \mathbf{D}, \mathbf{R} = \mathbf{y})$ is very similar to the first case except that $P(\mathbf{G} = \mathbf{yy} | \mathbf{R} = \mathbf{y}) = \rho^2$, and thus the probability is $P(\mathbf{G} = yy | \mathbf{D}, \mathbf{R} = \mathbf{y}) = \frac{\rho^2}{c} \prod_{\{i | D_i = x\}} (1 - \epsilon) \prod_{\{i | D_i \neq x\}} \epsilon$.

For the sake of this homework, we can assume $\rho^2 = 0.999$ and thus $\rho = 0.999$. I think, we might have to estimate ϵ from the dataset?