Part2

September 16, 2020

1 Part 2

1.0.1 Question 2

Write the likelihood model that uses both the observed bases and the quality scores for a single site. The frequencies of the four bases are the parameters. Remember that the true bases are not observed. NB! you have to explain all of the notation i.e. variables/parameters you use in your model.

The following is the likelihood model:

$$\ell(f) = \sum_{i}^{N} \log \left(\sum_{j} p(b_i|G_j, f)(G_j|f) \right)$$

where

$$P(b_i|G_j) = \begin{cases} \frac{\epsilon_i}{3} & b \neq G_j \\ 1 - \epsilon_i & b = G_j \end{cases}$$

f is the base frequency to be estimated,
b_i is the base and N is the number of reads,
G_j is the haploid Genotype and $j \in \{A, C, G, T\}$,
ϵ is the probability of a base call error

Write the Q (estimation) and M step of the EM algorithm that you will need for the optimization. Use the same notation (variables/parameters) as you use to describe the likelihood.

Q Step:

$$Q_i(G_j) = p(G_j|b_i, f^{(n)}) = \frac{p(b_i|G_j, f^{(n)})p(G_j|f^{(n)})}{\sum_j p(b_i|G_j, f^{(n)})p(G_j|f^{(n)})}$$

M step:

$$f_j^{n+1} = \frac{\sum_i^N p(G_j|b_i, f^{(n)})}{\sum_i^N \sum_j p(G_j|b_i, f^{(n)})}$$

2 Question 3

- Estimate the allele frequencies for all sites
- How many sites are there where the allele frequency of the most common allele is less than 0.9?