Lab 10

Instructions. Complete this lab during the lab session for HW10 credit.

Diploid organisms, such as humans and *C. elegans*, carry two copies of every chromosome, except the sex chromosome. For both humans and *C. elegans*, males carry only one copy of chromosome X. Thus, we have two copies of most sites in our genome.

An *allele* is a genetic variant at a nucleotide site or a gene variant at a gene, or a variant at any generic location of any breadth in a genome. For example, a special variant of the FOXP2 gene discovered in a British family has taken away the ability of affected individuals to adequately move their lips, tongues, and mouths to speak clearly. The SNP rs6025 in the human F5 gene causes a blood coagulation disease, where affected individuals suffer from an unusual frequency of blood clots. This SNP, and most SNPs, are biallelic, meaning there are only two possible nucleotides observed at the site.

The pair of alleles observed at a genomic site is the *genotype* at that site. At biallelic SNP sites, the individual is either homozygous normal (CC for rs6025), heterozygous (TC or CT, almost always the same thing), or homozygous mutant (TT). (Note, the coding normal allele at rs6025 is actually G, on the reverse complement strand.)

1. In class yesterday, we began talking about the methods to use NGS data to genotype or perform SNP/SNV calling (identification of sites that are polymorpic). For this question, suppose that you observe k C alleles at rs6025 and $n-k$ T alleles. The quality scores are, ordered with the k C alleles first, $q_1, q_2, \dots, q_k$, and the $n-k$ T alleles last, $q_{k+1}, q_{k+2}, \dots, q_n$. These quality scores can be converted as $e_i = 10^{-q_i/10}$ to error probabilities, $e_1, e_2, \dots, e_k, e_{k+1}, \dots, e_n$. In this question, assume reads are independent and the quality scores communicate the true error probabilities.

   (a) The C to T mutation at rs6025 causes a change in encoded amino acid from arginine to glutamine. Why would such a mutation be likely to affect the F5 gene? (Think about what you learned in the protein structure unit.)

   (b) If G is the unobserved genotype of the sequenced individual at rs6025 and N_C is the random variable representing the number of C alleles observed in the aligned NGS sample, derive formulae for the likelihoods $P(N_\mathrm{C} = k \mid G = \mathrm{CC})$ and $P(N_\mathrm{C} = k \mid G = \mathrm{TT})$.

(c) Show that

$$P(N_{\mathrm{C}} = k \mid G = \mathrm{TC}) = \frac{1}{2^n} \sum_{\substack{Z_1,\ldots,Z_n \\ \sum_i Z_i = k}} \prod_{i=1}^{k} \left[(1-e_i)^{Z_i} \left(\frac{e_i}{3}\right)^{1-Z_i} \right] \prod_{i=k+1}^{n} \left[\left(\frac{e_i}{3}\right)^{Z_i} (1-e_i)^{1-Z_i} \right],$$

where

$$Z_i = \begin{cases} 1 & \text{read } i \text{ is from chromosome bearing C} \\ 0 & \text{read } i \text{ is from chromosome bearing T.} \end{cases}$$

2. In this question, we will return the *C. elegans* alignments of the last lab. The files you will be working on are in `/ptmp/bcbio444/lab10`. The `vcf` files result from an incomplete run of `samtools mpileup` because the `slurm` job scheduler killed the job before it was complete, but about half of chromosome I was completed, and that is enough for our purposes.

   (a) Taking the data for the first position of *C. elegans* chromosome I, where the mechanics of extracting the necessary information is relatively easy, see if you can compute the PHRED-scaled genotype likelihoods, rounded to the nearest integer, reported in the file `SRR065390.vcf` created last week by `samtools pileup`. They are the last numbers on the this line from that `vcf` file:

```
I  1  .  G  <*>  0  .  DP=15;I16=13,0,0,0,438,14762,0,0,0,0,0,0,0,0,0,0;QS=1,0;\
MQ0F=1  PL  0,39,11
```

   Hint. `DP=15` indicates the read depth, but `I16=13,0,0,0` indicates that 2 reads were discarded, presumably for poor alignment quality.

(b) If you were going to estimate the genotype as that which maximized the likelihood, which genotype would you call at this site? What about site I:230580? Which genotype is most likely, and what is the likelihood (computed from the `vcf` file, not by hand)?
Hint. Genotypes are listed in the order: homozygote of most common allele first, then all heterozygotes involving most common allele, then homozygote of second most common allele, and so on.

(c) Assuming the likelihoods (roughly) reported for genotypes at site I:230580 compute the posterior probabilities $P(G = g \mid N_k = k)$ for $g \in \{\mathrm{AA}, \mathrm{AG}, \mathrm{GG}\}$ under two models. Both models assume $P(G = \mathrm{AG}) = r$ and $P(G = \mathrm{AA}) = P(G = \mathrm{GG}) = \frac{1-r}{2}$. The first, appropriate for SNP calling, assumes $r = 0.001$. The second, appropriate for genotype calling, assumes $r = 0.2$. If making your genotype call based on the maximum posterior probability, does your decision change under either of these models from that estimate obtained by maximizing the likelihood?

(d) Finally, use `samtools tview` to view the alignments at site I:230580. (Once in `tview`, type `gI:230550` and press enter to get a pretty good view.) Does looking at the data raise any concerns about the your conclusion that there is true variation (A) at this site?