Following Li [1], we will study the problem of SNP and genotype calling using Illumina-style next generation sequencing data. We assume whole genome or targeted resequencing has been undertaken of a random sample of individuals from a well-defined population. We assume the genome of the organism has been previously sequenced, and that the reads have been successfully aligned to a good reference genome. We assume sequencing errors at sites within the reads are independent, and we assume reads are independent conditional on the true source sequence. Following Li's notation, we assume a sample of size n individuals, and we focus on a particular site in the genome where there is read coverage from the sequencing data. For the ith individual, the ploidy level is m_i, the (unknown) genotype $G_i \in \{0, 1, 2, \ldots, m_i\}$ is the number of reference nucleotides, the sequence data are $\boldsymbol{D}_i = (D_{i1}, D_{i2}, \ldots, D_{il_i})$, and the quality scores are $\boldsymbol{Q}_i = (Q_{i1}, Q_{i2}, \ldots, Q_{il_i})$, where l_i is the coverage (number of sequences aligned) to the chosen genomic position. Throughout this problem, we will condition on the quality scores. When deriving equations, show your work and justify every step.

1. Read through the newly updated SNP/Genotyping handout. Various errors have been corrected. In this question, we focus on the ith individual and drop the subscript i in the notation.

   (a) The notes state that the likelihood of the data $\boldsymbol{D}$ as given in Eq. (2) of Li [1] is not correct because there is a missing factor of $\frac{1}{3}$. One might take issue with the presentation in the notes because Li [1] actually claims that Eq. (2) is the likelihood of observing k reference alleles and $l-k$ alleles not equal to the reference (here and in the notes the role of k and l are swapped relative to Li [1]). Let $Z_j = \mathbb{1}\{D_j = b_r\}$, where b_r is the reference allele, indicate if the jth observed read matches the reference allele. So, Li [1] is claiming that Eq. (2) is the likelihood of the data $Z_1, Z_2, \ldots, Z_l$ given the genotype g, when there are $\sum_{j=1}^{l} Z_j = k$ reference alleles. WLOG, reorder the data so the first k reads match the reference allele. Show that in fact the likelihood of data $Z_1, Z_2, \ldots, Z_l$ is

$$\begin{aligned} L(g \mid Z_1, Z_2, \ldots, Z_l) &= \Pr(Z_1 = z_1, Z_2 = z_2, \ldots, Z_l = z_l \mid G = g) \\ &= \prod_{j=1}^{k} \left[(1 - e_j)\frac{g}{m} + \frac{e_j(m-g)}{3m} \right] \\ &\quad \times \prod_{j=k+1}^{l} \left[\frac{e_j g}{m} + \frac{(1-e_j)(m-g)}{m} + \frac{2e_j(m-g)}{3m} \right] \end{aligned}$$

   (b) Use the data from three diploids at `http://dorman.stat.iastate.edu/files/genotyping.txt` to compute the likelihood of the data in individual 0 at sites 962 and 964 assuming the reference base in both cases is $n_b = \text{A}$. What are your maximum likelihood estimates $\hat{G}_{MLE,962}$ and $\hat{G}_{MLE,964}$ of the genotypes? (These data are unforgivingly huge, so you may want to do selective reading of the data in some smart way.)

   (c) Now find the maximum a posterior estimates $\hat{G}_{MAP,962}$ and $\hat{G}_{MAP,964}$ and explain any additional assumptions you make.

   (d) What is your confidence in the MAP estimates? Provide a numeric measure of that confidence.

   (e) You should always look at your data and numeric summaries of it to make sure the data are consistent with assumptions your model is making. Do you see anything unusual in the data from these two sites? Are your conclusions and confidence affected?

2. The full dataset contains reads from three individuals. Test all sites for evidence of variation (a possible SNP). While it is possible to combine the data into a population model, where you assume HWE and estimate the reference allele relative frequency ψ at each site, there are only three individuals and thus very little information about ψ. Instead, use the mutant calling method to test the hypothesis that the number of reference alleles in each of these three individuals at each of these sites is 2, i.e.

$$H_0 : G_{is} = 2$$

   against $H_1 : G_{is} < 2$ for all individuals $i \in \{0, 1, 2\}$ and sites $s \in \{764, 765, \ldots, 1199\}$.

(a) Produce a space-separated text file with the individual in set $\{0, 1, 2\}$, the site in set $\{764, 765, \ldots, 1199\}$, and the p-value for rejecting H_0. For computing the p-value, use Monte Carlo sampling to estimate

$$\Pr\left[T(\boldsymbol{D}_{is}) \leq T(\boldsymbol{d}_{is})\right] \approx \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\left\{\sum_{j=1}^{l_{ij}} z_{isjb} \leq T(\boldsymbol{d}_{is})\right\},$$

where $z_{isjb} \overset{\text{ind.}}{\sim} \text{Bernoulli}(1 - e_{isj})$ and e_{isj} is the probability of error of the sth nucleotide in the jth read of the ith individual.

(b) Do any of the four approximations described at `https://stats.stackexchange.com/questions/177199/success-of-bernoulli-trials-with-different-probabilities` work well? Add four columns to the file in Part a.

(c) Interpret your results. Do they make sense? What might have gone wrong?

3. We have made several asssumptions about the meaning of the NGS quality scores. The latest version of the SNP genotyping notes assume that (1) the probability of error is independent of the true base, (2) each possible error is equally likely when an error has occurred, and (3) the quality score can be taken literally. Under these conditions,

$$\begin{aligned}\Pr(R = b \mid B = b, Q = q) &= 1 - 10^{-q/10} \\ \Pr(R = b' \mid B = b, Q = q) &= \frac{1}{3}10^{-q/10}.\end{aligned}$$

for all $b \in \{\text{A}, \text{C}, \text{G}, \text{T}\}$ and $b' \neq b$. In this question, you will test the validity of these three assumptions. Using shorthand notation

$$p_{qbb'} := \Pr(R = b' \mid B = b, Q = q).$$

you will test various assumptions about these quantities. Please note that maximum likelihood estimation is analytically possible for all parts of this question, though of course you will have to use a computer to compute the value of the formula since the data are so large.

(a) Generally, we cannot observe the true base B that produced read nucleotide R, but for those sites where you are confident the individual is homozygous, we can be very confident that we know the true base. (Notice for heterozygous positions, we cannot.) Identify all the sites whose support for homozygosity is high (see Question 1, Part d). Use data from these sites to test the null hypothesis that the true nucleotide does not affect the error probability for each q,

$$H_0 : \sum_{\substack{b' \in \{\text{A},\text{C},\text{G},\text{T}\} \\ b' \neq b}} p_{qbb'} = p_q \;\forall\; b.$$

(b) Keep the model complexity supported by your conclusions from Part a. Test whether all substitutions are equally likely when an error has occurred,

$$H_0 : p_{qbb'} = p_{qb} \;\forall\; b' \neq b,$$

for each $b \in \{\text{A}, \text{C}, \text{G}, \text{T}\}$ if the H_0 in Part a is rejected (otherwise, drop b from test).

(c) Test support for

$$H_0 : p_{qbb'} = \frac{1}{3}10^{-q/10} \;\forall\; b, b' \in \{\text{A}, \text{C}, \text{G}, \text{T}\}.$$

against the alternative model H_1 you found supported by Parts a and b,

References

[1] Heng Li. "A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data." In: *Bioinformatics* 27.21 (2011), pp. 2987–2993.