For all of these problems, if you need the prior probability of a genotype in the population, you can assume the following: $P(AA) = 0.95^2$, $P(BB) = 0.05^2$, and $P(AB)$ is the remainder.

Before generating random numbers, make sure to set a random seed so your results will be reproducible.

Problem 1

From biallelic models to triallelic models

In the "SNP or not SNP" lecture, we limited ourselves to biallelic variation. Fortunately, most single nucleotide variation is biallelic (estimates of multiallelic variation are 3% of SNPs). Let's assume that at this particular position in the genome, we can have 3 possibilities: $\{G, A, C\}$. This isn't crazy — sometimes for unstable mutations alleles can drift and then create more stable variation (e.g. a G can turn into a C in some finite number of generations). We are going to restrict ourselves to the case where only two chromosomes exist.

(a) How many possible genotypes are there? (i.e. GG, GA, ..., CC).

(b) If we assume an error can result in any base (e.g. a G can turn into a T), assume I observe a T in my data. How does that translate into the probabilistic model? Note: I'm not looking for a complicated answer. Simply describing how it might change the probability of observing in error is sufficient.

(c) Write the probability of observing read i for homozygous genotypes (i.e. GG, AA, CC). You can follow the example from class, but don't forget there are more possibilities than the example in class.

(d) Write the probability of observing read i given genotype GA. Remember that we have only two chromosomes but three possible alleles.

(e) Write down the probability of the remaining genotypes. If they reduce into functions of the others, feel free to be lazy and write them in terms of other probabilities.

Problem 2

Data analysis + the bootstrap

Consider the biallelic model again.

(a) Refer to the slides from `2_errors_and_snps.pdf`. If you assume $P(C_i = A) = P(C_i = B)$, does this reduce any of the likelihoods? For the remainder of this problem, please assume $P(C_i = A) = P(C_i = B)$ when there is than one allele in the truth.

(b) `reads.tsv` is some data in the following format:

observation	$P(E_i = 1)$	**indicator if actually an error**
A	0.02	FALSE
G	0.01	TRUE
...	...	...

This is a simulation and you don't need the final column; it is there for your personal enjoyment. You might be able to do something with it, but you don't need it to solve the problem. *Please don't use it to solve the problem.*

Write some code to estimate the posterior probability of the three possible genotypes given the data.

(c) Randomly sample 5 observations with replacement and re-estimate the posterior probability of each genotype. What are the results?

(d) Repeat (c) 1,000 times. That means you will have 1,000 estimates for each of your posterior probabilities, each using 5 observations. This procedure is a variation of the *bootstrap*. Make a histogram for each of the posterior probabilities. Please be mindful of the number of bins and the appearance of your histogram. No one likes an ugly histogram.

(e) Repeat (d), but this time instead of taking 5 observations, take 50. Again, make three histograms.

(f) How do the results from (d) and (e) compare? Feel free to take summary statistics like the mean and standard deviation from those resampled results.

(g) Implicitly, there are assumptions about the base caller, the prior probabilities, etc. What are these assumptions and how might they affect the results? An example, what if the base caller probability estimates were way off?