Selma Chihab (Mentor: Seth Temple)

This quarter I studied a few concepts in statistical genetics as they concern relationship inference. We call two genes identical-by-descent (IBD) if they are inherited from a common ancestor. IBD states 0, 1, and 2 correspond to the probabilities that two individuals share 0, 1, and 2 of their genes IBD. Two measures of relatedness that we studied in relation to IBD are kinship and inbreeding. The kinship coefficient is the probability that different copies of a gene are IBD. The inbreeding coefficient denotes this value between the parents of an individual. Given a pedigree, we can compute the kinship coefficient using Wright's path counting formula, where A represents the common ancestor, f_A represents the inbreeding coefficient of A, and $n(P(A))$ represents the count of meioses in the equation shown below:

$$\psi = \sum_{A \in \mathcal{A}} \sum_{\mathcal{P}(A)} (1 + f_A)(1/2)^{n(\mathcal{P}(A))+1}$$

My project focused on inferring relationships from genetic data. I had a sample of 300 individuals from 4 different populations with various ancestries. All individuals in the sample were reported to be completely unrelated. Because I did not have pedigrees for the individuals in the sample, I had to use software to calculate the IBD states and kinship coefficient values desired.

To estimate the different IBD state probabilities and the kinship coefficients, we used two different software packages, PLINK and IBDkin, and juxtaposed their results. PLINK relies solely on the allele frequencies for single nucleotide polymorphisms and ran quickly and easily, unlike IBDkin that took much longer and required multiple steps. For IBDkin, I had to first phase our genotype data using the BEAGLE software, i.e. I had to identify at each marker whether the allele originated from the maternal or paternal haplotype. After phasing, I used hap-ibd to detect IBD segments and then ran IBDkin with the hap-ibd output. All the software was run using the command line, which is something I had previously not used before, thus making this project a great learning experience.

To conduct my analyses, I used R to combine, match, and clean the outputs of the two software into one table. After doing so, I created various plots. To compare the software used, I first plotted the calculated kinship values for each software against one another. I found that PLINK tended to overestimate its kinship values while IBDkin tended to underestimate. With my plots of the calculated kinship values against the IBD states, I found that there were in fact related individuals in the sample. In particular, I found five pairs of individuals to be parent/child relationships, one pair of individuals to be full siblings, and one pair of individuals to be first cousins. I made this assessment by comparing the calculated IBD state probabilities and kinship coefficients to the theoretical values we expect to see for the given relationships. Although I had previously worked with R before, this step of the project introduced me to many new functions, especially during the matching and cleaning phase, and I practiced interpreting plots.

Overall, through the DRP I gained better appreciation for how statistical models apply in the study of genetics.