The Statistical Analysis of Relatedness

By Michael Yung

Mentor: Seth Temple

This quarter I started off by looking at the idea of kinship coefficient and how to calculate it with the method of path counting. The general idea of path counting is to count the number of segments between two individuals through their common ancestors in a pedigree and using that number to calculate a kinship coefficient. This formula started off simple with kindship coefficient being $(\frac{1}{2})^{(\#\ of\ segements + 1)}$, however, as the pedigree expanded and we started to add the variable of potential inbreeding as well, things got complicated very fast. The equation became $(\frac{1+f}{2})^{(\#\ of\ segements + 1)}$, where f is the inbreeding value, or the kindship coefficient between the parents of the individual. As the examples became more and more complicated, it was obvious that if one was computationally analyzing these values, it would be a lot quicker and accurate than by hand, where we could set each individual as a node, and create a graph relationship throughout the entire pedigree.

Then, I started to delve more into the idea of Identical by descent, which had values K_0, K_1, and K_2 defined, usually correlating them towards the kindship coefficient depending on the relationship. These variables are defined by if the two individuals we are looking at have either 0 matching alleles from their ancestors(K_0), one matching allele(K_1), or two matching alleles(K_2). These variables come in a proportion such that K_0 +K_1+K_2 = 1 since 1 represents the entire genome of the individuals that we are comparing. One example would a non-inbred relative, where K_2 = 0, K_1 = 4 * Kinship coefficient, and K_2 = 1-4* Kinship coefficient. This was very technical where each type of relationship had different IBD values, something very troublesome to memorize.

Lastly, after doing everything by hand and looking at how people used to determine relatedness by hand, we started to look more at concurrent research and programs that people use to determine relatedness instead. Since genome sequencing is possible in the modern age, we know we could get a very accurate determination of how much genetic materials two individuals shared, but it might take a long time since the entire genome is millions of nucleoids to look at. I was introduced to the format of these nucleoid data with a file system call “vcf”, which has to be condensed and compacted with other programs due to their sheer size. Furthermore, the nucleoid data coming in from the vcf files are usually jumbled as it does not contain data on if a nucleoid came from the maternal or paternal side. This required phasing, which is to unjumble these data and separate paternal and maternal nucleoids. Then we can use a matching program to infer a relationship between the two individuals. To get a better understanding of this process, we went through a lab of how people in the research field will be processing these vcf files, using phasing and matching programs like BEAGLE to analyze the data. We also delve into optimizing run time and how some of the newer programs uses the Hidden Markov Model to make the phasing of these nucleoids exponentially faster compared to the old methods.

Overall, I learnt a lot more on the background of biostatistics and how people determine the relatedness of individuals in the modern world. These ideas will help me substantially in my current work with professor Bruce Weir as I have a stronger foundation on the entire premise of relatedness and how to determine it.