Final Project

STAT 499
Autumn 2021

Inferring relationships from genetic data is of popular intrigue. For instance, companies like 23andMe and AncestryDNA provide genealogical reports to consumers. One such report is to infer relationships between individuals in their large datasets, helping people to uncover long lost familial lineages. There are many algorithms to infer familial relationships. These split into (1) methods based on allele frequencies, (2) methods based on identity-by-descent segments, and (3) methods aware of admixed ancestry. Ramstetter et al. (2017) survey some of these methods applied to a sample of Hispanic families in the San Antonio area. Here we review the current white paper from AncestryDNA about discovering genetic relationships.

In this project, we will hone our skills in reading scientific articles and presenting statistical methods to a lay audience. Your audience, fellow undergraduate and graduate students, will know little about statistical genetics. The major challenge will be in communication. We will use the final two meetups to refine how we communicate statistical genetics and practice the presentation. Below are some questions to think about as you read the AncestryDNA white paper. (We may have to trim parts of the presentation to stay within 10 minutes.)

- Why do more distantly related people have shorter DNA segments identical-by-descent? Think about how to describe genetic recombination to a lay audience.
- Why do we have to use a statistical phasing program before inferring ibd segments? In what form does genotype data come in from SNP array sequencing?
- How do ibd segments imply certain familial relationships? Think about kinship and κ_0, κ_1, and κ_2 values.
- Give an overview of the hidden Markov model, with special consideration to BEAGLE as an example HMM.
- This paper describes a bioinformatics pipeline. Describe the following steps:
  - Phasing
  - ibd segment detection
  - Assessing informativeness
  - Estimating relationships