Week 5:
Genes mirror geography within Europe

Assignment 2

1. Improve on your data ingestion from Part 1
2. Extract SNPs from VCF and generate principal component plot
    a. Likely will have to explore different parameters / cutoffs
    b. Up to you to identify/remove outliers (if necessary)
    c. We'll test on part of chr22, but you'll likely want to use more than one chromosome
3. Ideally, will have pipeline code:
    a. Take a FASTQ file to a BAM
    b. Take BAM to VCF

Data

GitHub: https://github.com/ShanEllis/Genetic-Variation/tree/master/project

- Small test files

Data on Server: /datasets/dsc180a-wi20-public/Genome

- What would be helpful to have here?

Discussion Questions

- How did scientists recreate the map of Europe (what data? what techniques?)
- What is population structure (also sometimes referred to as: population substructure)?
- Why do geneticists care about population (sub)structure?

Student Questions

Theme 1: PCA

- I would like to discuss more on the PCA that was done just to make sure that I understood their methods correctly.
- Why do we only care about PC1 and PC2? Why is two dimensional enough for our study?
- How could we extrapolate the results from the paper to make genetic clusters of the whole world?

Discussion Questions

- How did scientists recreate the map of Europe (what data? what techniques?)
- What is population structure (also sometimes referred to as: population substructure)?
- Why do geneticists care about population (sub)structure?

Student Questions

<u>Theme 2: Progress</u>

- (From the paper): *In addition, the PCA-based methods used here are based on genotypic patterns of variation and do not take advantage of signatures of population structure that are contained in patterns of haplotype variation.* **What would be a "signature" of population structure.**
- This paper was written in 2008 and says "*Soon-to-be-available whole-genome re-sequencing will give us access to informative low-frequency alleles, and further statistical method development will allow us to leverage patterns of haplotype variation. The prospect of these developments suggests the geographic resolution presented here is only a lower bound on the performance possible in the near future.*" **What methods have been developed in the past 12 years?**

Next week's readings

1. Technology : Sequencing
    a. Don't worry about the details; understand the basics
    b. Think: how these technologies could be used to answer questions
2. Vox: GWAS
    a. A good introduction to GWAS and PRS
    b. Think: is there/are there questions you want to answer using these approaches

…*starting* to think about what you want to work on next quarter