Fall 2025 Biostat 915/Stat 915/Compsci 762 Homework 1

Option B

Due: 9/30 (Tuesday) to the TA's email (zhengchi.ma@duke.edu)

For this data analysis assignment, we will be analyzing genetic data from worldwide goat populations! The raw data can be accessed via the original paper: https://doi.org/10.1186/s12711-018-0422-x

Q1: Exploratory data analysis

Explore the three files and describe the data. How many subjects/individuals are in this goat dataset? How many different populations of goats are in this dataset? Create at least two exploratory data analysis plots to help us visualize the data. Describe what you observe about missingness in each of the files, if any.

Q2: Genotype missingnes

Describe what you observe about missingness or unknown variables in each of the files, if any. How does the genotype (SNP) missingness distribution look for each individual from the bed file? According to the paper, this data has already been filtered for initial quality control with individual genotype call rate, SNP call rate, etc. and is ready for analysis.

Q3: Use PCA to visualize ancestry distribution

Studying worldwide goat population can help us understand molecular diversity across and within continents. Principal component analysis (PCA) is a standard method used for ancestry inference, such as adjusting for population structure and identifying ethnic origins. Failing to account for differences in genetic ancestry may lead to biases in various genetic discoveries. We want to produce a few PCA plots (similar to Figure 4 in the literature) to better understand how we can subset out data for Q4, association analysis. Moreover, determine a reasonable method to detach the components used in PCA.

Q4: Simulate Phenotypes and run a simple GWAS

Typically for Genome-wide Association Studies, we will have individuals labelled as case or control (in the fam file) for the purpose of identifying genes associated with a particular disease or trait. A standard regression model for GWAS assumes the Y variable as the vector of phenotype values and the X variable as the vector of genotype values for all individuals at each SNP. There are several sophisticated ways to simulate quantitative traits and to assign cases and controls to individuals. But for the purpose of this exercise, let's just randomly assign phenotypes (assign cases or control in the "pheno" column of the fam file) so that we can practice running a simple GWAS. You can use a subset of the data that you find interesting according to your Q3 PCA analysis. Feel free to use any GWAS algorithm/package, the output should give you summary statistics of SNPs/markers, which you can then try to visualize with a manhattan plot, or with p-value diagnostic plots.

Note:

Run time for these analyses will largely depend on your computational power and resources. Given the limited amount time/resources, please feel free to select a few ancestries from the worldwide population for this analysis. This could be a few breeds from different ancestries, or you can focus on specific continents or countries. Just be sure to document your data processing steps, how you subset the data, and interpret your results. You should turn in a short write up to answer these four questions. You can submit your code in a separate file or together with the write-up. There are many packages and libraries out there that can help you with this analysis, such as plink or SNPRelate (R-package). No need to reinvent the wheel, but it may take some time to familiarize yourself with different resources and various formats required to each tool.

Background:

Here are some background to help you get started:

The bim/bed/fam files are in plink "binary ped" format. Plink is an open-source whole genome data analysis toolset, typically used as a command-line tool for analyzing high dimensional data. The files can be read into R with the `read_plink()` function from the **genio** R-package. They can also be manipulated and analyzed with command-line tools (which is often times preferred due to large file size).

```
#install.packages("genio")
library(genio)
library(tidyverse)

# name of the files without the extension
name <- 'hw-optionB-files/ADAPTmap_genotypeTOP_20160222_full'
goat = read_plink(name)
```

The bim file

Info on SNP location.

```
goat$bim %>% head() %>% as.data.frame()
```

```
##   chr                          id posg pos alt ref
## 1   0  snp10134-scaffold1361-15149    0   0   A   G
## 2   0  snp10135-scaffold1361-44576    0   0   A   G
## 3   0  snp10136-scaffold1361-91495    0   0   G   A
## 4   0 snp10412-scaffold1372-579082    0   0   0   G
## 5   0 snp10413-scaffold1372-610565    0   0   G   A
## 6   0 snp10415-scaffold1372-688806    0   0   G   A
```

```
goat$bim %>% tail() %>% as.data.frame()
```

```
##   chr                          id posg       pos alt ref
## 1  30  snp8814-scaffold1316-122968    0 121695220   A   C
## 2  30   snp8813-scaffold1316-69235    0 121748953   G   A
## 3  30   snp8812-scaffold1316-38259    0 121779929   G   A
## 4  30 snp32736-scaffold3775-128450    0 121824403   G   A
## 5  30  snp32735-scaffold3775-77895    0 121874958   A   G
## 6  30  snp32734-scaffold3775-36517    0 121916336   G   A
```

The fam file

Info on each individual. Total number of rows = total number of subjects. The first column "fam" is typically a larger grouping of the individuals, the second column is the individual id. The "sex" column have values 0 (unknown), 1 (males), or 2 (females). The last "pheno" column typically indicates case or control, in this case -9 is just a default value.

```
goat$fam %>% head() %>% kable()
```

fam	id	pat	mat	sex	pheno
ABR	ET_ABR0001	0	0	0	-9
ABR	ET_ABR0002	0	0	0	-9
ABR	ET_ABR0003	0	0	0	-9
ABR	ET_ABR0004	0	0	0	-9
ABR	ET_ABR0005	0	0	0	-9
ABR	ET_ABR0006	0	0	0	-9

The bed file

Genotype information is stored in the **X** matrix. The row names are SNP id's and the column names are individual id's.

```
goat$X[1:5, 1:4]
```

```
##                               ET_ABR0001 ET_ABR0002 ET_ABR0003 ET_ABR0004
## snp10134-scaffold1361-15149           NA         NA         NA         NA
## snp10135-scaffold1361-44576           NA         NA         NA         NA
## snp10136-scaffold1361-91495            1          2          1          1
## snp10412-scaffold1372-579082          NA         NA         NA         NA
## snp10413-scaffold1372-610565           0          0          0          0
```