An Application of Principal Components Analysis in Genetics

Samuel Morrissette

April 14, 2020

Section 1

Background

Genetic Association Studies

- As its name implies, genetic association studies test for an association between certain genetic variants and a particular disease or trait. For example, sickle cell anemia is caused by an abnormal *allele* in the HBB gene.
- Genetic association studies are frequently conducted through a case-control study.
- If there is an association between a disease and a certain allele, we would expect that this allele would occur more frequently in individuals suffering from the disease (the case group) than in those who do not suffer from the disease (the control group).

Population Stratification

- Case-control studies may be confounded by *population stratification*.
- Population stratification refers to the differences in allele frequencies arising from systematic ancestral differences. In other words, some alleles naturally occur more frequently in certain groups as a result of their ancestry.
- Suppose that this particular group is overrepresented in the case group of a case-control study. Some alleles may be falsely associated with occurrence of the disease when, in reality, they are simply a result of ancestry.

Correcting for Population Stratification

- To avoid spurious associations, we need to either entirely avoid or correct for population stratification.
- To avoid population stratification, cases and controls would have to be genetically homogenous, but this may be unrealistic for a variety of reasons (e.g. individuals may be unaware of their exact ancestry).
- A statistical method called EIGENSTRAT was proposed by Price et. al in 2006. EIGENSTRAT utilizes principal components analysis to correct for population stratification.
  - Other methods of correction include genomic control and structured association.
  - Genomic control corrects test statistics by dividing them by a uniform inflation factor. (ISSUES WITH THIS?)
  - At the time of publication, these genomic control and structure association were the main methods of correction. Since then, (in large part due to this paper) PCA has become widely used.

Section 2

Eigenstrat

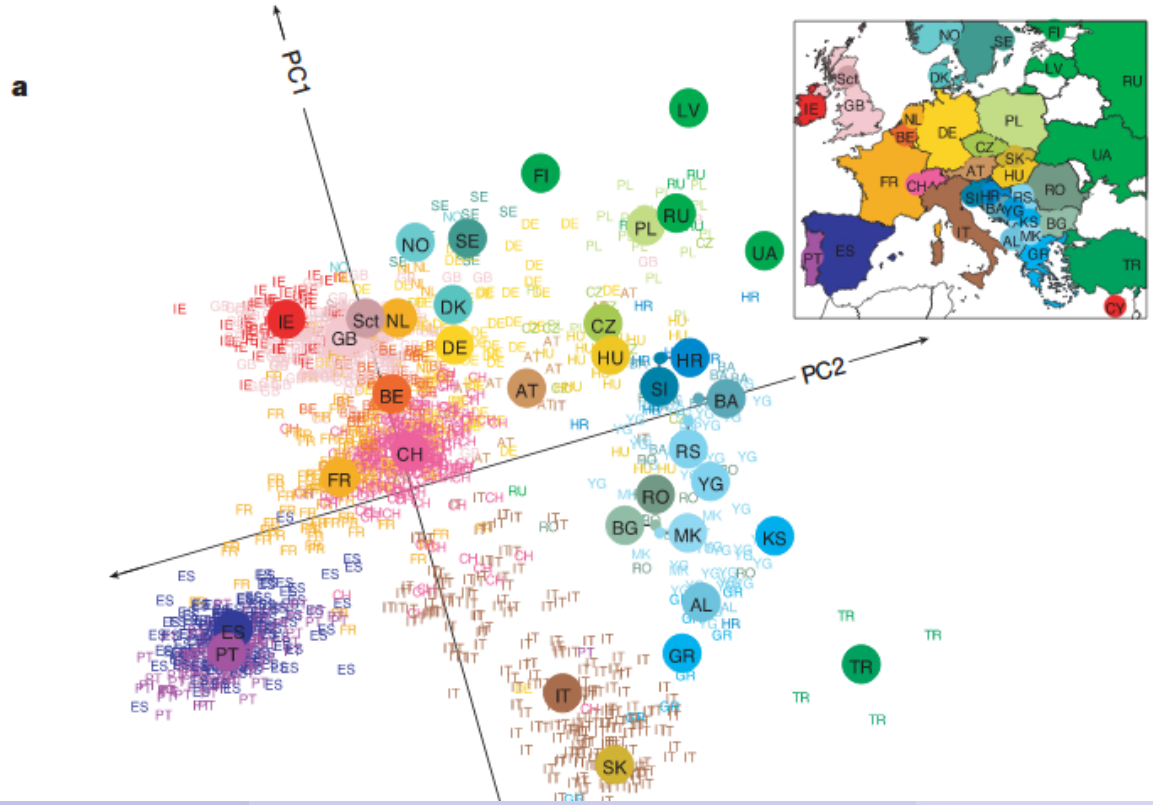
Eigenstrat Algorithm

- Eigenstrat is an algorithm proposed by Price, et al. (2006) in the paper "Principal components analysis corrects for stratification in genome-wide association studies"
- Eigenstrat consists of three main steps:

1. Apply PCA to random SNPs (preferably unrelated to the candidate SNPs of interest) to infer "axes of variation"
2. Adjust the candidate SNPs and phenotypes based on these axes
3. Compute a test statistic for association using the adjusted genotypes and phenotypes

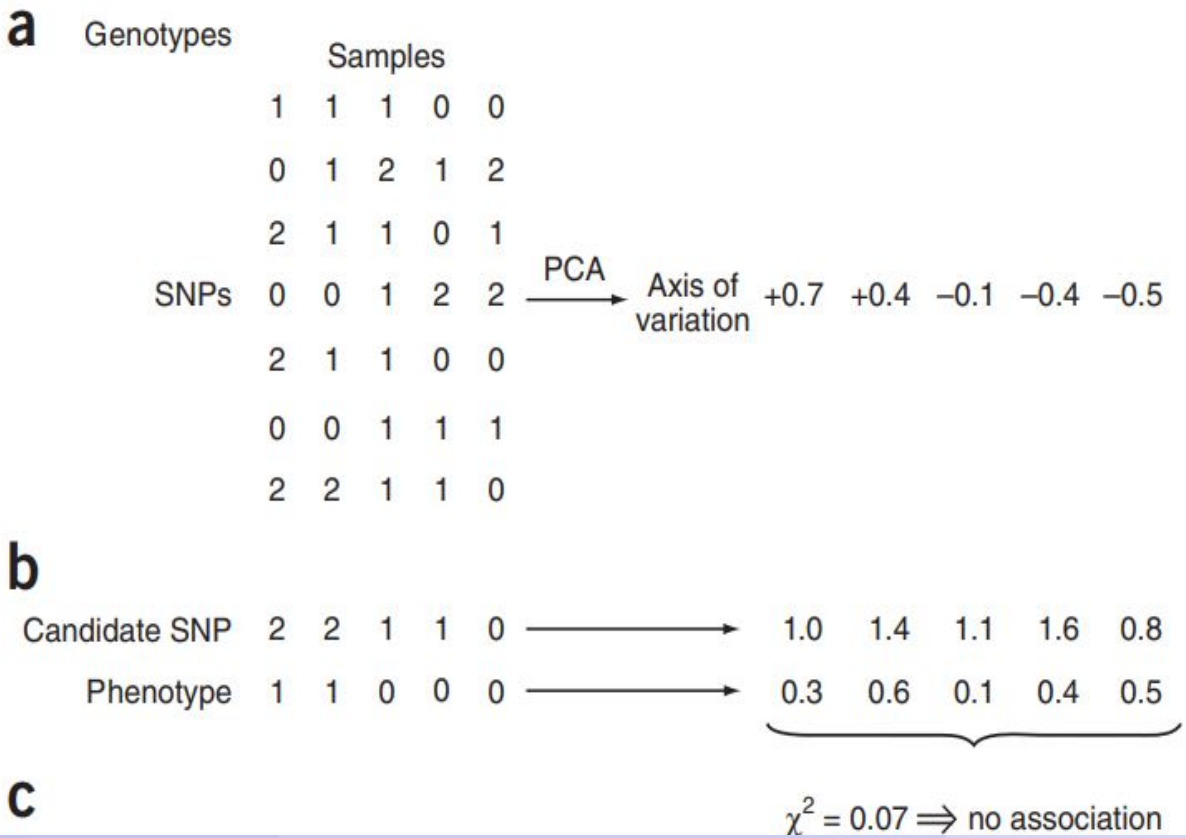
1. Axes of Variation

- Defined as "the top eigenvectors of a covariance matrix between samples" (i.e. the top principal components).
- These principal components can capture differences in genetic variation attributable to ancestry and can sometimes even have a geographical interpretation within continents



2. Adjustment and 3. Calculations

- Using values based on the axes of variation, the genotypes and phenotypes of the candidate SNPs are adjusted.
- The EIGENSTRAT test statistic is then calculated based on the adjusted genotypes and phenotypes.



Section 3

Results

asdf

- The authors tested the EIGENSTRAT algorithm on both simulated data and a real data set consisting of samples of European Americans. They compared the results with:
  - Armitage trend test statistic (uncorrected for population stratification)
  - Genomic control (a method that corrects for population stratification using an inflation factor)
- In several scenarios, EIGENSTRAT was able to detect and correct for population stratification better than the uncorrected and genomic control-corrected test statistics.

Section 4

Implementation in R

Bovine data

- We can see how PCA corrects for population stratification in bovines using data from the "adegenet"" package in R.
- In the following data set, we have a sample of 704 cattle sampled from Africa and France genotyped at 374 SNPs.

R Code

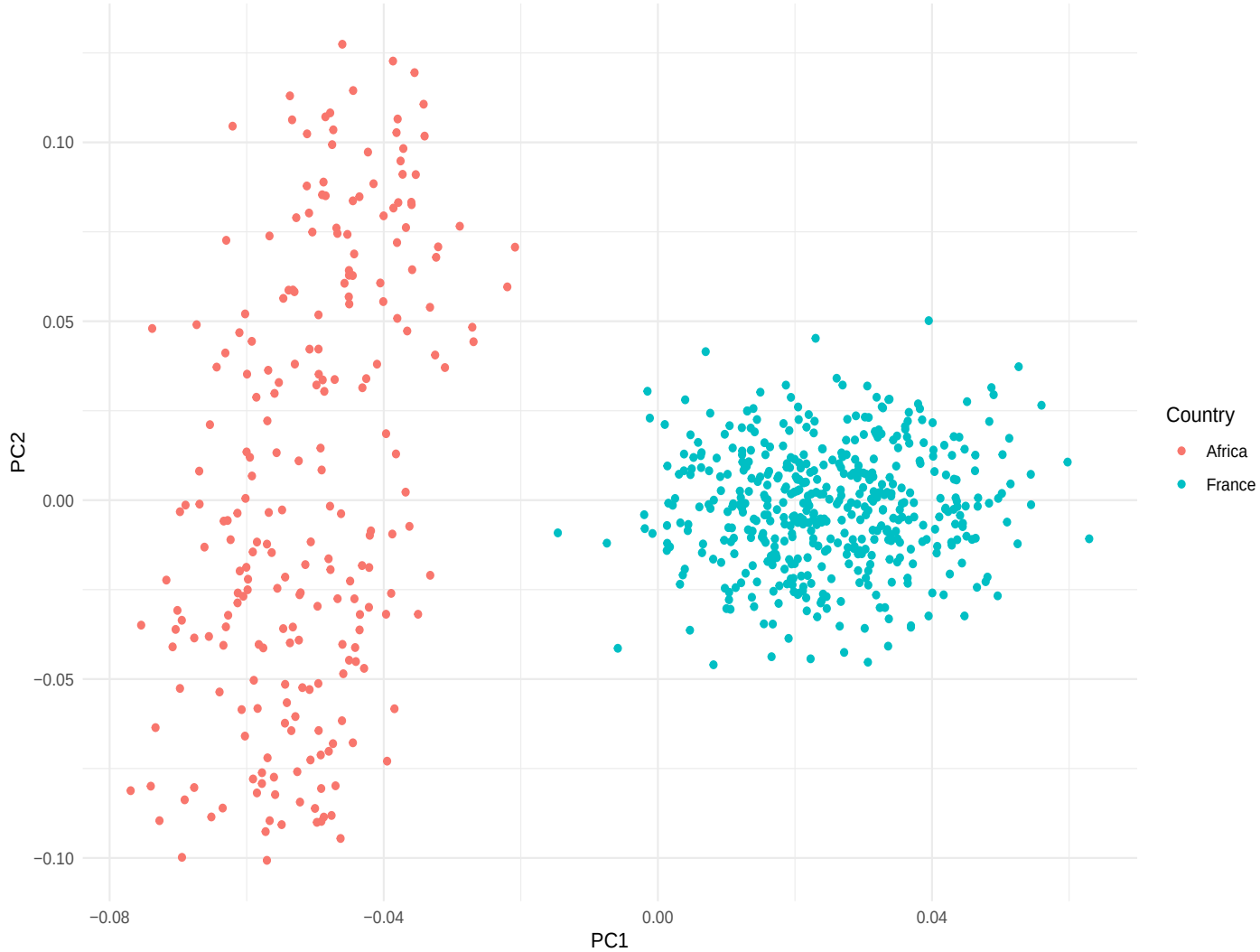
```
library(adegenet)


data(microbov)
data <- t(tab(microbov, freq=FALSE, NA.method="zero"))
dim(data)
```

```
## [1] 373 704
```

Bovines by country

- After applying PCA to the dataset we can see clear separation between the cattles' country of origin with only the first principal component.



Bovines by breed

- We also have each bovine's breed. We can see some separation with the first two principal components.