An Application of Principal Components Analysis in Genetics

Samuel Morrissette

April 14, 2020

1. Background and Terminology
2. Eigenstrat Algorithm and Definitions
3. Results
4. Example - Bovine Data
5. Conclusion

Section 1

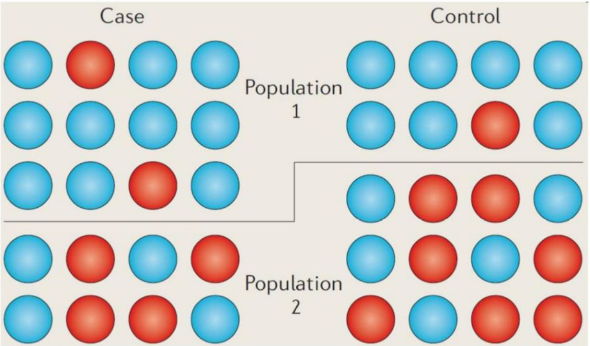
Background and Terminology

Genetic Association Studies

- Genetic association studies:
  - Test for an association between certain genetic variants (alleles) and a particular disease or trait.
  - Are frequently conducted through a case-control study.
- Expected occurrence of alleles in case group vs. control group

Population Stratification

- **Population stratification** refers to the differences in allele frequencies arising from systematic ancestral differences.
- Case-control studies may be confounded by population stratification.
  - Overrepresentation of a population in the case or control group can result in spurious associations.



Correcting for Population Stratification

- Avoiding population stratification is difficult and likely unrealistic
- Correcting for population stratification is more realistic
    - Genomic control and structured association were two of the most common methods
    - Eigenstrat, proposed by Price et al. in 2006, has since become the prevailing approach

Section 2

Eigenstrat Algorithm and Definitions

Eigenstrat Algorithm

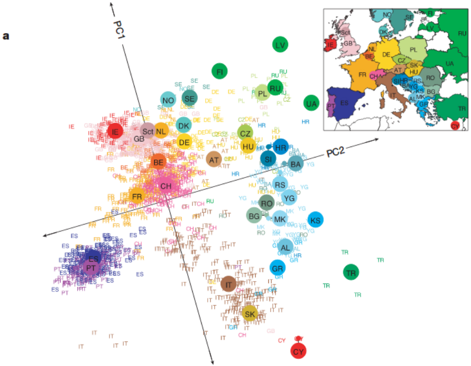
Eigenstrat consists of three main steps:

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 of the samples based on these axes
3. Compute a test statistic using adjusted values

Axes of Variation

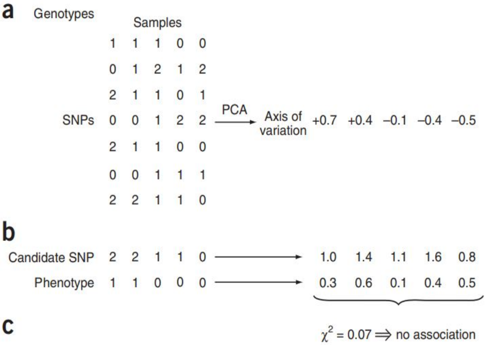
The axes of variation:

- Defined as the top principal components
- Can capture differences in genetic variation attributable to ancestry.
  - May have a geographical interpretation within continents (figure below)



Adjustment and Test Statistic Calculation

- The genotypes of the candidate SNPs and phenotypes of the samples are adjusted
  - Adjustment corrects for population stratification
- The Eigenstrat test statistic is then calculated based on these adjusted genotypes and phenotypes



1	1	1	0	0
0	1	2	1	2
2	1	1	0	1
0	0	1	2	2
2	1	1	0	0
0	0	1	1	1
2	2	1	1	0

Axis of variation: +0.7 +0.4 –0.1 –0.4 –0.5

Candidate SNP	2	2	1	1	0	1.0	1.4	1.1	1.6	0.8
Phenotype	1	1	0	0	0	0.3	0.6	0.1	0.4	0.5

$\chi^2 = 0.07 \Rightarrow$ no association

Section 3

Results

Testing Scenarios

- Price et al., tested the Eigenstrat algorithm on simulated data:
- Simulated candidate SNPs in three different categories:
    1. Random SNPs with no association to disease
    2. Highly differentiated SNPs with no association to disease
    3. Causal SNPs associated with a disease
- Results were compared with:
    - Armitage trend test statistic (uncorrected for stratification)
    - Genomic control (corrects for stratification using a uniform inflation factor)

Advantages of Eigenstrat

- Eigenstrat corrected for stratification better than the uncorrected and genomic control-corrected test statistics in all simulation scenarios.
    - Fewer spurious associations in non-causal SNPs.
    - More powerful when detecting true associations at causal SNPs.
- Computationally tractable

Section 4

Example - Bovine Data

microbov data

- PCA can correct for population stratification in bovines using data from the “adegenet” package in R.
- microbov: sample of 704 cattle from Africa and France genotyped at 373 SNPs.

R Code

```r
dim(data)
```

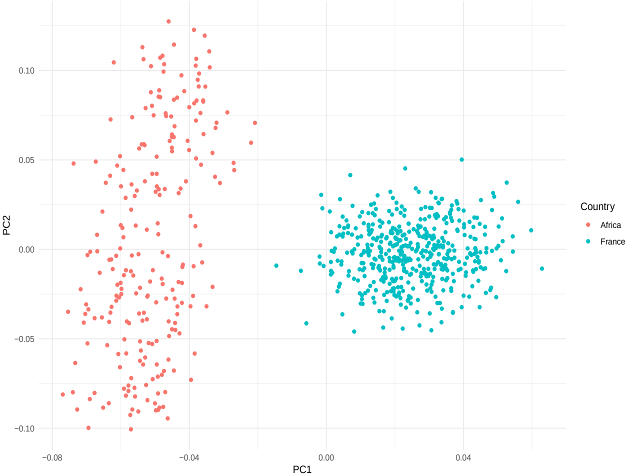
```
## [1] 373 704
```

```r
data[1:3,1:3]
```

```
##              AFBIBOR9503 AFBIBOR9504 AFBIBOR9505
## INRA63.167             0           0           0
## INRA63.171             0           0           0
## INRA63.173             0           0           0
```

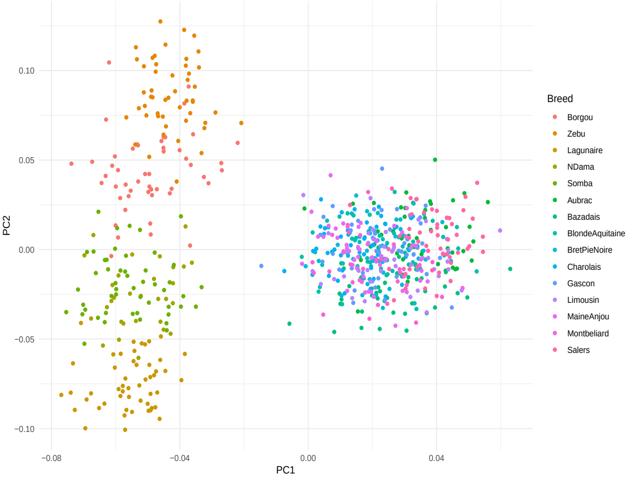
Bovines by Country

- There is a clear separation between the cattles' country of origin with the first principal component.



Bovines by Breed

- The breed of each bovine is also included in the data. There is some evident separation with the first two principal components.



Simulation Setup

- 10,000 candidate SNPs with no association to disease are created using highly differentiated allelic frequencies between countries:
  - 0.8 for Africa
  - 0.2 for France
- The case-control simulation study will include:
  - 100 cases from Africa and 50 from France
  - 50 controls from Africa and 100 from France

Simulation Results

Using 10,000 candidate SNPs and a significance level of 0.0001:

- There were 6743 spurious associations detected using the Cochran-Armitage trend test statistic (Type I error rate $= 0.6743$)
- There were 23 spurious associations detected using the Eigenstrat test statistic (Type I error rate $= 0.0023$)

Section 5

Conclusion

Conclusion

- Principal components analysis plays an important role in detecting and correcting for population stratification.
- Eigenstrat outperformed the alternatives at the time of publication and continues to be one of the most widely used methods of correction today
- "Eigenstrat is not a panacea". Association studies should still be designed properly.
  - Poor designs may result in a loss of power with Eigenstrat

References

- Balding, D. A tutorial on statistical methods for population association studies. Nat Rev Genet 7, 781-791 (2006).
- Novembre, J., Johnson, T., Bryc, K. et al. Genes mirror geography within Europe. Nature 456, 98-101 (2008).
- Price, A., Patterson, N., Plenge, R. et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38, 904-909 (2006).