Lecture 2: Introduction to the PLINK Software for GWAS & Population Structure Inference

Instructors: Joelle Mbatchou and Loic Yengo

Summer Institute in Statistical Genetics 2023

PLINK Overview

- PLINK is a free, open-source whole genome association analysis toolset, designed to perform a range of basic, large-scale analyses in a computationally efficient manner:

  https://www.cog-genomics.org/plink/1.9/
  https://www.cog-genomics.org/plink/2.0/

- PLINK has numerous useful features for managing and analyzing genetic data

PLINK Overview

- Data management
  - Read data in a variety of formats (BED, PGEN, BGEN, VCF,...)
  - Convert between different formats
  - Recode and reorder files
  - Merge multiple genetic files
  - Extracts subsets (SNPs or individuals)

PLINK Overview

- Report summary statistics for quality control
  - Allele & genotypes counts/frequencies
  - Missing genotype rates
  - Mendel error rate
  - HWE tests
  - Sample variant counts
  - Inbreeding, IBS and IBD statistics for individuals and pairs of individuals

PLINK Overview

- Perform basic association testing
  - Standard allelic test & Fisher's exact test for case-control data (PLINK1.9)
  - Linear and logistic regression
  - Dominant/recessive and general models
  - Association conditional on one or more SNPs
  - Family-based association tests (PLINK1.9)

PLINK Overview

- Additional features
  - Gene-based tests of association
  - Screen for epistasis
  - Gene-environment interaction with continuous and dichotomous environments
  - Meta-analysis (PLINK1.9)
    - Automatically combine several generically-formatted summary files, for millions of SNPs
  - Simulate genetic data with no LD (PLINK1.9)

Input Files

PLINK BED

BIM

```
Chr      ID          CM    Pos.     Ref Alt
1 1:12030946:T:C 0 12030946 T C
1 1:12032428:A:C 0 12032428 A C
1 1:12057950:C:T 0 12057950 C T
1 1:12095233:A:C 0 12095233 A C
1 1:12100532:T:C 0 12100532 T C
```

Variant info

FAM

```
FID.      IID      Fa Mo Sex Y
1432 HGDP00702 0 0 2 -9
1433 HGDP00703 0 0 1 -9
1434 HGDP00704 0 0 2 -9
1436 HGDP00706 0 0 2 -9
1438 HGDP00708 0 0 2 -9
```

Sample info

BED

Compressed binary file (bytes) storing 0/1/2/NA

```
00000000: 01101100 00011011 00000001 11111111 11111110 11111111  l.....
00000006: 11111110 11111111 11111111 11111111 11111111 11111111  ......
0000000c: 11111111 11111111 11111111 11111111 11111110 11111111  ......
00000012: 11111111 11111111 11111111 11111111 11111111 11111111  ......
```

Genotype data

Data Management

- Inclusion/Exclusion criteria options
  - `--keep incl_samples.txt, --remove excl_samples.txt`
  - `--extract incl_snps.txt, --exclude excl_snps.txt`
  - `--chr 2,6,X, --from rs273744 --to rs89883`
- Other data management options
  - `--make-bed, --export, --pmerge`
- Using files with phenotypes/covariates
  - `--pheno --pheno-name, --covar --covar-name`

Quality Control (QC)

- Summary statistics options:
  - minor allele frequency (MAF): `--freq`
  - genotype counts: `--geno-counts`
  - SNP & individual missing rate: `--missing`
  - Hardy-Weinberg: `--hardy`
- Inclusion/Exclusion filters
  - MAF: `--maf` , `--max-maf`
  - minor allele count (MAC): `--mac`,`--max-mac`
  - SNP missing rate: `--geno`
  - Individual missing rate: `--mind`
  - Hardy-Weinberg: `--hwe`

Association Analysis with PLINK

With PLINK

- Association testing: `--assoc`, `--linear`, `--logistic`
- Conditional analysis: `--condition-list`
- GxE interaction: `--gxe`

With PLINK2

- Association testing: `--glm`
- Conditional analysis: `--condition-list`
- GxE interaction: `--glm interaction`

Background: Population Structure

- **PLINK can also be used to infer population structure**
- Humans originally spread across the world many thousand years ago.
- Migration and genetic drift led to genetic diversity between isolated groups.

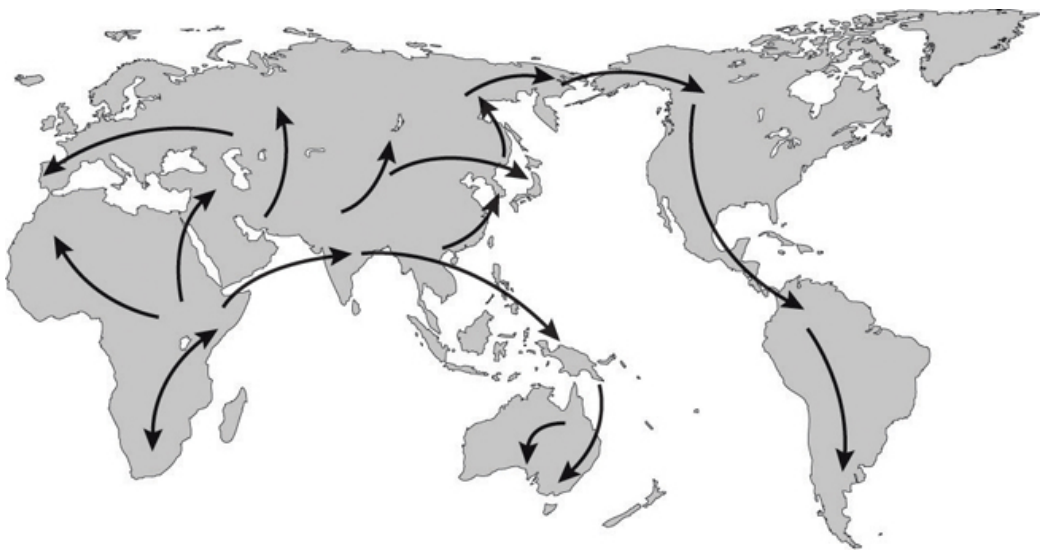


Figure: https://science.education.nih.gov

Population Structure Inference

- Inference on genetic ancestry differences among individuals from different populations, or **population structure**, has been motivated by a variety of applications:
  - population genetics
  - genetic association studies
  - personalized medicine
  - forensics
- Advancements in array-based genotyping technologies have largely facilitated the investigation of genetic diversity at remarkably high levels of detail
- A variety of methods have been proposed for the identification of genetic ancestry differences among individuals in a sample using high-density genome-screen data.

Inferring Population Structure with PCA

- Principal Components Analysis (PCA) is the most widely used approach for identifying and adjusting for ancestry difference among sample individuals
- PCA applied to genotype data can be used to calculate **principal components** (PCs) that explain differences among the sample individuals in the genetic data
- The top PCs are viewed as continuous axes of variation that reflect genetic variation due to ancestry in the sample.
- Individuals with "similar" values for a particular top principal component will have similar ancestry for that axes.

Standard Principal Components Analysis (sPCA)

- sPCA is an unsupervised learning tool for dimension reduction in multivariate analysis.
- Widely used in genetics community to infer population structure from genetic data.
  - Belief that top principal components (PCs) will reflect population structure in the sample.
- Orthogonal linear transformation to a new coordinate system
  - sequentially identifies linear combinations of genetic markers that explain the greatest proportion of variability in the data
  - these define the axes (PCs) of the new coordinate system
  - each individual has a value along each PC
- EIGENSOFT (Price et al. 2006) is a popular implementation of PCA.

Data Structure

- Sample of n individuals, indexed by $i = 1, 2, \ldots, n$.
- Genome screen data on m genetic autosomal markers, indexed by $l = 1, 2, \ldots, m$.
- At each marker, for each individual, we have a genotype value, G_{il}.
  - Here we consider SNP data, so G_{il} takes values 0, 1, or 2, corresponding to the number of minor alleles.
- We center and standardize these genotype values:

$$z_{il} = \frac{G_{il} - 2\hat{p}_l}{\sqrt{2\hat{p}_l(1-\hat{p}_l)}}$$

where $\hat{p}_l$ is an estimate of the minor allele frequency for marker l.

Genetic Correlation Estimation

- Create an n x m matrix, $\mathbf{Z}$, of centered and standardized genotype values, and from this, a n x n genetic correlation matrix (GRM):

$$\hat{\boldsymbol{\Psi}} = \frac{1}{m}\mathbf{Z}\mathbf{Z}^T$$

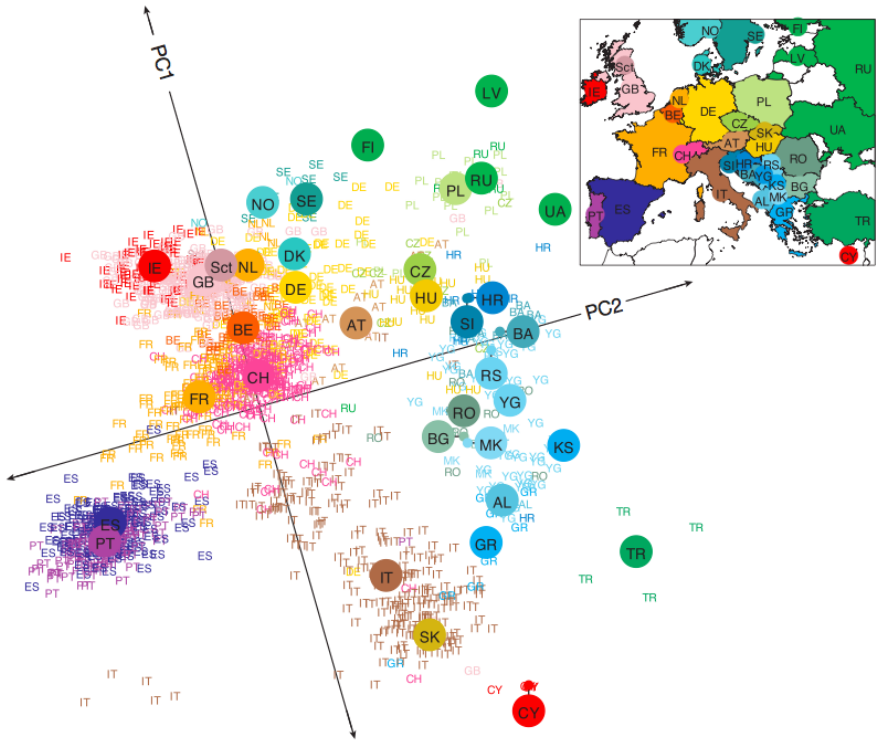
- $\hat{\boldsymbol{\Psi}}_{ij}$ is an estimate of the genome wide average genetic correlation between individuals i and j.
- PCA is performed by obtaining the eigen-decomposition $\hat{\boldsymbol{\Psi}}$

Standard Principal Components Analysis (sPCA)

- Identify orthogonal axes of variation, i.e. linear combinations of SNPs, that best explain the genotypic variability between the n sample individuals.
- The result is:
  - a set of n length n eigenvectors, $(\mathbf{V}_1, \mathbf{V}_2, \ldots \mathbf{V}_n)$, where $\mathbf{V}_d$ is a column vector of coordinates of each individual along axis d
  - each principal component is a different linear combination of the m markers
  - and a corresponding set of n eigenvalues, $(\lambda_1 > \lambda_2 > \ldots > \lambda_n)$, in decerasing order.
  - The d^{th} principal component (eigenvector) corresponds to eigenvalue λ_d, where λ_d is proportional to the percentage of variability in the genome-screen data that is explained by $\mathbf{V}_d$.
- These eigenvectors (PCs) are used as surrogates for population structure

PCA of Europeans

- Application of PCA in European samples (Novembre et al., *Nature* 2008)
- Among Europeans for whom all four grandparents originated in the same country, the first two PCs computed using 200k SNPs could map their country of origin quite accurately



Relatedness Confounds sPCA

- Recall that the GRM used by sPCA, $\widehat{\mathbf{\Psi}}_{ij}$, and is an estimate of the genome wide average genetic correlation between individuals i and j.
- It can be shown:

$$\mathbf{\Psi}_{ij} = 2\left[\phi_{ij} + (1 - \phi_{ij})A_{ij}\right]$$

  - ϕ_{ij}: kinship coefficient - a measure of familial relatedness
  - A_{ij}: a measure of ancestral similarity
- PCA is an unsupervised method; in related samples we don't know the correlation structure each eigenvector is reflecting
  - If the only genetic correlation structure among individuals is due to ancestry, $\mathbf{\Psi}$ and the top PCs will capture this.
  - If there is relatedness in the sample, the top PCs may reflect this or some combination of ancestry and relatedness.
- Association studies have known or cryptic relatedness!

sPCA: Best practices

- Apply QC to variants & samples:
  - Restrict to common variants (e.g. MAF ≥ 0.01)
  - Remove variants with high missing genotypes rates (e.g. ≥ 0.01)
  - Remove variants which fail HWE test (e.g. p-value $\leq 10^{-10}$)
  - Remove samples with high missing genotypes rates (e.g. ≥ 0.1)
  - Keep only variants on autosomal chromosomes
- Remove related individuals (e.g. 3rd degree related or closer)
- Prune variants in linkage disequilibrium (LD) (e.g. $r^2 \geq 0.2$) include long-range LD regions (Price et al., *AJHG*, 2008)

R package bigsnpr

- Apply QC to variants & samples (relies on PLINK2)

```
snp_plinkQC(plink.path, prefix.in,
file.type="--bfile", maf = 0.01, geno = 0.1,
mind = 0.1, hwe = 1e-10, autosome.only = TRUE )
```

- Remove related individuals (e.g. 3rd degree related or closer)

```
extra.options = "--king-cutoff 0.0442"
```

- Compute PCs
  - Prune variants in linkage disequilibrium (LD) (e.g. $r^2 \geq 0.2$)
  - Removes long-range LD regions

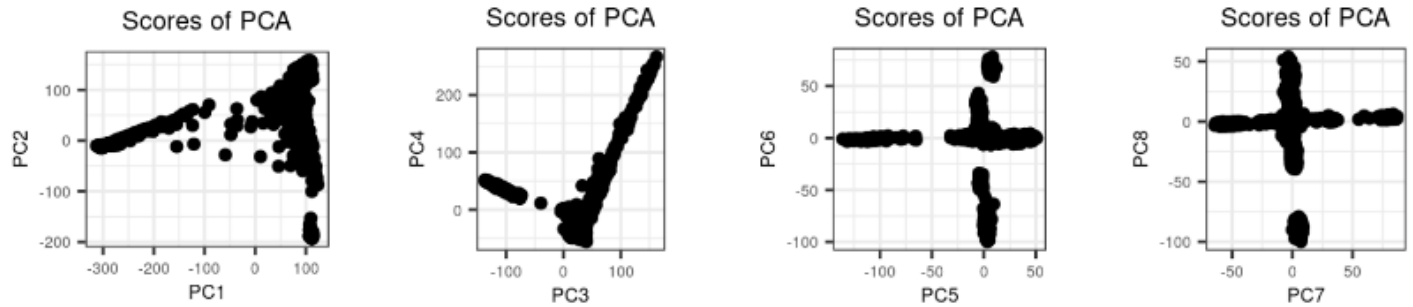
```
pca <- bed_autoSVD(obj.bed, thr.r2 = 0.2, k = 20)
predict(pca)
```

- Project related samples (excluded from training model)

```
bed_projectSelfPCA(object.svd, obj.bed, ind.row)
```

R package bigsnpr

```
plot(obj.svd2, type = "scores", scores = 1:20, coeff = 0.4)
```




```
plot(obj.svd2, type = "loadings", loadings = 1:20, coeff = 0.4)
```

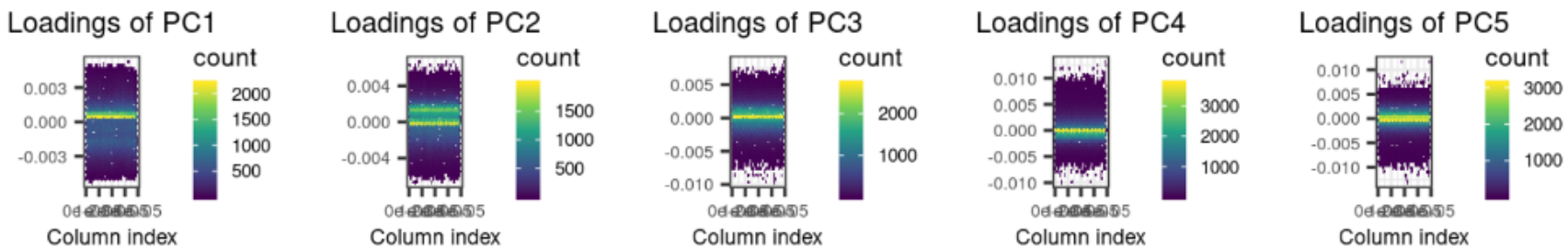


Figure: https://privefl.github.io/bigsnpr/articles/bedpca.html

Summary

- PLINK is a versatile software widely used for managing and QC of genetic data
  - Commonly used tool for VCF format : `bcftools`
- It can also be used for association testing as well as interaction analyses
- Important use of PLINK is in population structure inference
- It can be a major source of confounding in GWAS if unaccounted for
- Quite effective at capturing genetic differences between individuals due to ancestry

References

- Patterson,N., Price, A.L., Reich, D. (2006) Population structure and eigenanalysis. *PLoS Genet.* **2**, e190.
- Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergmann, S., Nelson, M.R. (2008). Genes mirror geography within Europe. *Nature* **456**, 98-101.
- Alexander, D.H., Novembre, J., Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**,1655-1664.

References

- Price, Alkes L, Michael E Weale, Nick Patterson, Simon R Myers, Anna C Need, Kevin V Shianna, Dongliang Ge, et al. (2008). Long-Range LD Can Confound Genome Scans in Admixed Populations. *The American Journal of Human Genetics*, **83**(1), 132-135.
- Conomos MP, Miller M, Thornton T (2015). Robust Inference of Population Structure for Ancestry Prediction and Correction of Stratification in the Presence of Relatedness. *Genetic Epidemiology* **39**, 276-93
- Privé, F., Luu, K., Blum, M. G., McGrath, J. J., Vilhjálmsson, B. J. (2020). Efficient toolkit implementing best practices for principal component analysis of population genetic data. *Bioinformatics*, 36(16), 4449-4457.