Quality Control - Part 2 - population

KidneyGenAfrica Course – January 2026

January 21, 2026

Contents

1 Context and Motivation

Population structure and admixture can bias association results and generate spurious signals in GWAS. The objective of this practical is to detect population outliers or potential sample swaps that may affect downstream analyses.

2 Objectives

1. Visualize population structure using Principal Component Analysis (PCA)
2. Estimate ancestry proportions using ADMIXTURE
3. Identify and remove population outliers to ensure homogeneous study groups

3 Data

We use genotype data obtained from the previous exercise (genotype QC). Alternatively, you may directly use the prepared dataset:

- Genotype data: `Data_qc_genotype/afreur_qc_rel` (filtered for HWE, MAF, and missingness)
- Phenotype file: `Data_qc_genotype/qc_pheno.tsv` (contains ancestry-related information)

Exercise 3: Preparation of Independent SNPs

Before PCA and ADMIXTURE, we select approximately independent SNPs to avoid bias due to linkage disequilibrium.

```
mkdir -p admixture

../bin/plink \
  --bfile 2_Data_qc_genotype/afreur_qc_rel \
  --indep-pairwise 50 10 0.1 \
  --out admixture/afreur_pihat

../bin/plink \
  --bfile 2_Data_qc_genotype/afreur_qc_rel \
  --extract admixture/afreur_pihat.prune.in \
  --make-bed \
  --out admixture/afreur_pihat_indep
```

Exercise 3 – Question

- Why is LD pruning necessary before PCA and ADMIXTURE analyses?

Exercise 4: Principal Component Analysis (PCA)

PCA helps visualize genetic similarity between individuals and detect population stratification or mislabeled samples.

3.1 Key PLINK Arguments

- `--bfile`: input PLINK binary fileset
- `--pca`: compute principal components (default = 20)
- `--out`: output prefix

Exercise 4 – Questions

- How many principal components should be retained for population structure correction?
- What might explain clusters or outliers observed in the PCA plot?
- How could individuals deviating strongly from the main cluster be identified and excluded?
- How can principal components be integrated as covariates in a GWAS?

Exercise 5: ADMIXTURE Analysis

ADMIXTURE estimates the proportion of ancestry components (K) for each individual. It is useful for detecting admixture, substructure, or potential sample contamination.

3.2 Example Command

```
admixture admixture/afreur_pihat_indep.bed --cv 3
```

Here, $K = 3$ corresponds to the assumed number of ancestral populations.

3.3 Key Concepts

- `--cv`: performs cross-validation to assess model fit
- Input must be in PLINK `.bed` format
- Output files:
  - `.Q`: ancestry proportions per individual
  - `.P`: ancestral allele frequencies

Exercise 5 – Questions

- How do you determine the optimal number of clusters (K)?
- What patterns of admixture are observed across individuals?
- How can ancestry proportions be used as covariates in GWAS?

Exercise 6: Detecting and Handling Population Outliers

Population outliers can be detected by combining PCA and ADMIXTURE results.

3.4 Typical Strategy

1. Visualize PCA results, coloring individuals by ADMIXTURE ancestry proportions
2. Identify individuals whose genetic ancestry does not match their reported population group

3.5 Using PCA

- Plot PC1 versus PC2 in R
- Identify PCs that separate major ancestry groups
- Define thresholds to exclude individuals falling outside the main cluster

3.6 Using ADMIXTURE

- Load ADMIXTURE results for $K = 2$
- Merge `.Q` files with `.fam` and phenotype data
- Identify ancestry components corresponding to African and European ancestry
- Flag individuals whose ancestry proportions are inconsistent with their labels

3.7 Clean Dataset Definition

Individuals are retained only if their main ancestry component is at least **0.7**. Individuals with ancestry proportion < 0.7 for the target population are excluded.

Exercise 6 – Questions

- How can the removal of population outliers influence GWAS results?
- Could sample swaps or mislabeling explain the presence of outliers?