Association Analysis

KidneyGenAfrica Course – January 2026

January 21, 2026

Contents

1 Context and Objectives

The objective of this practical is to understand how to perform a **genetic association analysis**, how to define **inputs and outputs**, and how to specify appropriate **phenotypes and covariates**.

We compare association results obtained:

- Before quality control (QC)
- After genotype, population, and phenotype QC

For simplicity, association tests are performed using **PLINK 2**. Note that alternative software packages such as **GEMMA** or **BOLT-LMM**, which explicitly model relatedness, may provide more accurate results depending on study design and population structure.

2 Data

Two datasets are used:

- **Before QC**: `1_Data_beforeqc/afreur`
- **After QC**: `4_Data_qc_admixture_pheno/genotyped_qc`

Each dataset is analyzed using an appropriate phenotype file and covariate specification.

3 PLINK 2 Association Models

3.1 Key PLINK 2 Options

- `--bfile`: input PLINK binary fileset
- `--pheno`: phenotype file
- `--pheno-name`: phenotype to test (e.g., `egfr`)
- `--covar`: covariate file
- `--covar-name`: list of covariates to include
- `--glm`: run linear or logistic regression
  - `allow-no-covars`: allow models without covariates
  - `hide-covar`: suppress covariate output
- `--out`: output prefix

3.2 Example Without Covariates

```
../bin/plink2 \
  --bfile headerplk \
  --pheno pheno_file \
  --pheno-name egfr \
  --glm allow-no-covars \
  --out egfr_raw
```

Exercise 11: Association Analysis Before QC

Model Specification

- Genotype data: `1_Data_beforeqc/afreur`
- Phenotype file: `1_Data_beforeqc/afreur_pheno.tsv`
- Phenotype: `egfr`
- Covariates: `Sex`, `Age`

Tasks

- Run a linear association analysis using PLINK 2
- Inspect the resulting output files

Exercise 11 – Questions

- How many variants are tested?
- Do you observe inflation or unexpected signals?
- What sources of bias might still be present at this stage?

Exercise 12: Association Analysis After QC

Model Specification

- Genotype data: `4_Data_qc_admixture_pheno/genotyped_qc`
- Phenotype file: `4_Data_qc_admixture_pheno/qc_pheno.tsv`
- Phenotype: `egfr`
- Covariates: `sex`, `age`, `Superpopulation`

Tasks

- Run the association analysis including covariates
- Compare results with the pre-QC analysis

Exercise 12 – Questions

- How do effect sizes and p-values differ compared to the pre-QC analysis?
- Does inclusion of ancestry reduce spurious associations?
- Why is it important to match covariates to the QC steps performed earlier?

Exercise 13: Understanding and Interpreting Output

PLINK 2 outputs association results in files such as:

`egfr_raw.egfr.glm.linear`

To inspect the file header:

```
head egfr_raw.egfr.glm.linear
```

Key Output Columns

Column	Description
`#CHROM`	Chromosome number
`POS`	Base-pair position
`ID`	Variant identifier (rsID)
`REF`	Reference allele
`ALT`	Alternative allele
`A1`	Tested allele
`BETA`	Estimated effect size
`SE`	Standard error
`T_STAT`	Test statistic
`P`	P-value

Exercise 13 – Questions

- Which column is used to assess statistical significance?
- How should effect size and direction be interpreted for eGFR?
- Why is multiple-testing correction required in GWAS?