Training Exercise for Genome-Wide Association Studies (GWAS)

KidneyGenAfrica Course – January 2026

January 21, 2026

Contents

1 GWAS Context

Genome-Wide Association Studies (GWAS) are powerful research approaches used to identify genetic variants associated with specific traits or diseases. They involve scanning the entire genome of many individuals to detect differences in allele frequencies between affected (cases) and unaffected (controls) individuals.

These differences, often single-nucleotide polymorphisms (SNPs), can reveal genetic loci linked to disease risk, biological pathways, and potential therapeutic targets.

GWAS rely on large, well-characterized cohorts and careful data processing to ensure reliable results. A typical GWAS workflow includes defining the study sample, genotyping, rigorous quality control (QC), genotype imputation, association testing, and biological interpretation of significant findings.

2 Main Steps of a GWAS

2.1 Sample Definition and Collection

- Define clear inclusion and exclusion criteria.
- Ensure balanced representation to avoid bias (e.g., sex, ancestry, case/control ratio).
- Record relevant covariates (age, sex, recruitment site, lifestyle factors).

2.2 Genotyping

- Select an appropriate genotyping platform (e.g., H3Africa array, GWAS arrays).
- Consider cost, genomic coverage, and population relevance.

2.3 Genotype Quality Control

- Remove poorly genotyped variants and samples.
- Identify genotyping errors and batch effects.
- Detect duplicate or related individuals using genetic data.

2.4 Population Structure and Admixture

- Assess genetic ancestry using PCA or ADMIXTURE.
- Detect population outliers or hidden relatedness.
- Control for population stratification in association models.

2.5 Phenotype Quality Control

- Identify missing values and implausible measurements.
- Detect and handle outliers.
- Ensure phenotype consistency across datasets.

2.6 Imputation

- Infer untyped variants using reference panels (e.g., 1000 Genomes, H3Africa, TOPMed).
- Choose reference data that best match the ancestry of the study population.

2.7 Association Analysis

- Run appropriate statistical models (linear or logistic regression).
- Include relevant covariates (age, sex, ancestry PCs, site).
- Evaluate test statistics and genomic inflation factor (λ).

2.8 Post-Association Interpretation

- Visualize results using Manhattan and QQ plots.
- Extract independent loci using clumping and regional plots.
- Annotate variants and assess biological relevance.
- Perform replication or fine-mapping when possible.

3 Objective of the GWAS Exercise

Your collaborator, Prof. Nephro Logist, conducted a study aimed at identifying **common genetic variants** associated with kidney function, using **estimated glomerular filtration rate (eGFR)** as a quantitative phenotype.

You are provided with:

- Raw genotype data in PLINK format
- Phenotype information

However, potential data issues are suspected:

- Implausible or incorrect phenotype values
- Mismatches between genotype and phenotype files

4 Overall Goals of the Exercise

The goals of this training exercise are to:

1. Perform quality control on genotype and phenotype data
2. Conduct a GWAS for eGFR
3. Identify significant and independent SNPs
4. Generate and interpret summary plots (QQ, Manhattan, and regional plots)

5 Overview of the Practical Sessions

Part 0 — Understanding the Data and Software

- `Data_beforeqc`: dataset provided by collaborators (genotype and phenotype data)
- Overview of software: `plink`, `plink2`, `R`, `admixture`

Part 1 — Genotype Quality Control

- Missingness, MAF, and HWE filtering
- Detection of duplicates and related individuals
- Sex concordance checks
- Heterozygosity and missingness outlier detection

Part 2 — Population Quality Control

- PCA to visualize population structure
- ADMIXTURE to estimate ancestry proportions
- Identification and exclusion of population outliers

Part 3 — Phenotype Quality Control

- Plausibility checks and outlier detection
- Assessment of covariates (age, sex, ancestry)

Part 4 — Association Analysis

- Linear regression for eGFR
- Inclusion of covariates
- Comparison of models before and after QC

Part 5 — Post-Association Analysis

- Identification of independent loci
- Manhattan and QQ plots
- Regional association plots