Hands-on GWAS Practical: From QC to Post-GWAS

Jean-Tristan Brandenburg

Sydney Brenner Institute for Molecular Bioscience
Wits University
KidneyGen Africa

January 26, 2026

Practical Outline

What You Will Do in This Practical

- Perform genotype quality control
- Correct for population structure
- Correct phenotype
- Run a GWAS using PLINK
- Interpret results using Post-GWAS tools

GWAS Pipeline Overview

Raw genotypes → QC Genotype Individual and SNPs → Admixture and PCA → QC phenotype → GWAS → Post-GWAS

VirtualBox

- The practical sessions will be conducted using VirtualBox.
- All software required for the exercises is pre-installed on the VirtualBox.
- The VirtualBox environment is based on Linux.

Download the Data

Open a terminal and clone the course repository:

```
git clone https://github.com/WCSCourses/KidneyGen_Africa_2026.git
```

Move to the exercise directory:

```
cd /home/manager/Desktop/KidneyGen_Africa_2026/course_modules_2026/
   exercise_gwas
```

Quality Control

In this session, we will perform a subset of key QC procedures using the following dataset:

PLINK input data

`1_Data_beforeqc/afreur`

- Excluding SNPs with a minor allele frequency (MAF) e < 0.01.
- Excluding SNPs with a missingness rate >0.05 (5%).
- Excluding individuals with a missingness rate >0.005 (0.5%).
- using PLINK 2.

QC: PLINK2 Arguments

Commonly used PLINK2 options for quality control:

- `--bfile`: Basename of the genotype files.
- `--maf`: Minor allele frequency threshold.
- `--mind`: Individual missingness rate.
- `--geno`: SNP missingness rate.
- `--make-bed`: Output files in PLINK binary format (.bed/.bim/.fam).

Sample QC

Command for quality and cleaning

```
# first step ccan be a clean with
## frequency
## missingness
## clean on maf, error genotyping less than 5 % for individual and
   positiosn,
mkdir -p genotyped_qc
# HHWE did not add to the command line, but you could add
plink2 -bfile ../1_Data_beforeqc/afreur --make-bed -maf 0.01  --mind
    0.05 --geno 0.05   -out genotyped_qc/afreur_qc
```

Relatedness

KING is a tool designed to infer family relationships using high-throughput SNP data, as commonly generated in genome-wide association studies (GWAS) or sequencing projects. It provides an efficient algorithm for detecting related individuals.
PLINK2 includes KING-based relatedness estimation via the following option:

- `--king-cutoff`

PLINK and phenotype input data

```
Genotype:  2_Data_qc_genotype/afreur_qc_rel
Phenotype:  2_Data_qc_genotype/afreur_pheno.tsv
```

Relatedness

Command to remove related individuals

```
# Second step: remove related individuals using KING
plink2 --bfile genotyped_qc/afreur_qc \
       --king-cutoff 0.17 \
       --make-bed \
       --out ../2_Data_qc_genotype/afreur_qc_rel
```

Admixture and PCs: Defining SNPs

To compute ADMIXTURE and principal components (PCs), an independent set of SNPs is required.

The first step is to extract a subset of approximately independent SNPs using PLINK2 by pruning variants in linkage disequilibrium (LD).

- `--indep-pairwise`

Genotype input data

```
Genotype:  2_Data_qc_genotype/afreur_qc_rel
```

Admixture and PCs: Defining SNPs

```
mkdir -p admixture

plink2 --bfile ../2_Data_qc_genotype/afreur_qc_rel \
       --indep-pairwise 50 10 0.1 \
       --out admixture/afreur_pihat

plink2 --bfile ../2_Data_qc_genotype/afreur_qc_rel \
       --extract admixture/afreur_pihat.prune.in \
       --make-bed \
       --out admixture/afreur_pihat_indep
```

Principal Components (PCs)

Principal components (PCs) are computed using PLINK with the `--pca n` option, where n is the number of PCs to calculate.

```
plink2 --bfile admixture/afreur_pihat_indep \
       --pca 10 \
       --out admixture/afreur_pihat_indep
```

PCs are then used as covariates in GWAS analyses.

Question for trainees: How can the inclusion of PCs improve adjustment in a GWAS?

ADMIXTURE software

ADMIXTURE is run by assuming between 1 and n ancestral populations in order to determine the optimal number of ancestral populations (K).

The best value of K is selected based on the cross-validation (CV) error. Using this optimal K, individual ancestry proportions are examined to:

- Verify consistency with self-reported ancestry.
- Identify individuals with high levels of admixture.

```
for K in 1 2 3 4 5; do
  admixture --cv admixture/afreur_pihat_indep.bed $K | tee admixture
     /log${K}.out
done
# extract best CV
grep CV admixture/*.out > cv
```

ADMIXTURE software

Using **R**, we:

- Plot the cross-validation (CV) error to identify the best K.
- Extract individual ancestry proportions for the optimal K.
- Identify individuals with high admixture or ancestry inconsistent with expectations.

Discussion points:

- How many individuals were excluded?
- How can this be explained?
- What are the most likely causes?

Phenotype Quality Control

Phenotype quality control (QC) is a critical step in genetic association studies. Errors can arise at different levels, including:

- Coding or data-entry errors during data collection.
- Values outside the biologically plausible range.
- Non-normal trait distributions that violate model assumptions.
- Missing or incorrectly specified covariates.

Important covariates to consider include age, sex, ethnicity, and principal components (PCs).

Using **R**, you will:

- Summarize phenotype values by sex and population.
- Plot trait distributions and assess normality.
- Check value ranges (e.g. eGFR outside the plausible range of 10–150).

GWAS — Association Testing

To assess the impact of quality control, you will perform several GWAS analyses using PLINK2 with the options `--glm`, `--covar`, and `--covar-name`.

- Perform a GWAS on the raw data (before QC).
- Perform a GWAS on QCed data, with and without covariates.
- Based on your understanding and the available variables, decide which covariates should be included in the phenotype file.

Genotype and phenotype input data

`Genotype after QC (with phenotype and admixture):`
`4_Data_qc_admixte_pheno/genotyped_qc`

`Phenotype after QC: 4_Data_qc_admixte_pheno/qc_pheno.tsv`

`Genotype before QC: 1_Data_beforeqc/afreur`

`Phenotype before QC: 1_Data_beforeqc/afreur_pheno.tsv`

Association Command Line

```
## Raw data, no covariates
plink2 --bfile ../1_Data_beforeqc/afreur \
       --pheno ../1_Data_beforeqc/afreur_pheno.tsv \
       --pheno-name egfr \
       --glm allow-no-covars \
       --out ../5_association/egfr_beforeqc

## QCed data, with covariates
plink2 --bfile ../4_Data_qc_admixte_pheno/genotyped_qc \
       --pheno ../4_Data_qc_admixte_pheno/qc_pheno.tsv \
       --pheno-name egfr \
       --glm hide-covar \
       --covar ../4_Data_qc_admixte_pheno/qc_pheno.tsv \
       --covar-name age,Sex,Superpopulation \
       --out ../5_association/egfr_covar_afterqc
```

GWAS — Diagnostics

- Manhattan plot
- QQ plot
- Genomic inflation factor

GWAS — Diagnostics

Using the summary statistics from the previous analyses, compute diagnostic metrics and plots to compare the models.

- Compute the genomic inflation factor (λ).
- Generate QQ plots and Manhattan plots.
- Based on these diagnostics, which model performs best?

You may use:

- The `qqman` R package for QQ and Manhattan plots using R .
- The following formula to compute the inflation factor:

```
lambda <- median(qchisq(1 - Data$P, 1), na.rm = TRUE) / qchisq(0.5,
    1)
```

Summary statistics data

`5_association/egfr_covar_afterqc.egfr.glm.linear`
`5_association/egfr_covar_beforeqc.egfr.glm.linear`

GWAS — qqplot

```
library(qqman)
library(dplyr)
library(data.table)

data_beforeqc<-fread('../5_association/egfr_beforeqc.egfr.glm.linear
    ')
alpha_beforeqc_beforeqc<-median(qchisq(1-data_beforeqc$P,1),na.rm=T)
    /qchisq(0.5,1)
qq(alpha_beforeqc_beforeqc$P)

data_afterqc<-fread('../5_association/egfr_covar_afterqc.egfr.glm.
    linear');names(Data)[1]<-'CHR'
alpha_afterqc<-median(qchisq(1-data_afterqc$P,1),na.rm=T)/qchisq
    (0.5,1)
qq(egfr_covar_afterqc$P)
```

GWAS — Manhattan plot

```
pdf('nab_afterqc.pdf')
manhattan(alpha_afterqc, chr='CHR',bp='POS', p='P',snp='ID')
dev.off()
```

Identification of Independent SNPs

LD clumping groups variants based on linkage disequilibrium and association statistics to select representative SNPs.

```
plink2 --bfile ../1_Data_beforeqc/afreur \
       --clump ../5_association/egfr_covar_afterqc.egfr.glm.linear \
       --clump-p1 5e-8 \
       --clump-p2 0.1 \
       --clump-r2 0.1 \
       --clump-kb 100000
```

Regional Plot

A regional plot is a useful way to visualize local association signals, linkage disequilibrium (LD), and nearby genes within a defined genomic region.
Using a selected lead SNP, we generate a regional plot to explore the surrounding variants and gene annotations.

```
library(locusplotr)

gg_locusplot(df = Data, lead_snp = "rs1719245", rsid = "ID", chrom =
    "CHR",
  pos = "POS", ref = "REF", alt = "ALT", p_value = P,
  plot_genes = TRUE,   genome_build = "GRCh38",   plink = "../bin/
     plink",
  bfile = "../2_Data_qc_genotype/afreur_qc_rel", compute_ld = TRUE
)
```

End of Practical

Questions?