Quality Control - Part 3 - Phenotype Quality Control

KidneyGenAfrica Course – January 2026

January 21, 2026

Contents

1 Context and Motivation

Phenotype quality control is a critical step in GWAS to ensure that clinical measurements are biologically plausible, internally consistent, and suitable for association testing. In this practical, we focus on cleaning kidney-related phenotypes, with a particular emphasis on **estimated glomerular filtration rate (eGFR)**.

eGFR values depend on biological factors such as **sex** and **ancestry**. Failure to properly clean and model these effects can introduce bias and confounding in downstream GWAS analyses.

2 Phenotype File and eGFR

The phenotype file `afreur_pheno.csv` contains demographic and clinical information:

- **FID**: Family ID
- **IID**: Individual ID
- **Sex**: Biological sex (Men / Women)
- **Superpopulation**: Broad ancestry group (e.g., EUR, AFR)
- **Age**: Age at the time of measurement
- **Sc**: Serum creatinine (mg/dL)
- **eGFR**: Estimated glomerular filtration rate (computed using the 2009 equation, mL/min/1.73 m^2)

eGFR is derived from serum creatinine and has commonly accepted physiological values between approximately **10 and 150**. eGFR distributions are known to differ according to **sex** and **ancestry**.

For this exercise, we use the dataset:

- `3_Data_qc_admixture/afreur_pheno.csv`

The objective is to detect outliers, assess potential confounders, and generate a final cleaned phenotype file for GWAS.

Exercise 7: Phenotype Outlier Detection

Using **R**, load the phenotype file and inspect the distribution of eGFR.

Tasks

- Identify individuals with eGFR values < 10 or > 150
- Exclude these outlier individuals from the phenotype file

Exercise 7 – Questions

- Why can extreme eGFR values indicate data quality issues?
- How might measurement or data-entry errors affect GWAS results?

Exercise 8: Consistency With Genotype QC

Individuals excluded during genotype QC should also be removed from the phenotype dataset to maintain consistency.

Tasks

- Use the PLINK `.fam` file `3_Data_qc_admixture/genotyped_qc.fam`
- Remove individuals not present in the final genotype dataset

Exercise 8 – Question

- Why is it important that genotype and phenotype files contain identical sets of individuals?

Exercise 9: Assessing Sex and Ancestry as Confounders

eGFR is influenced by biological sex and ancestry. These variables may act as confounders in association analyses.

Tasks

- Fit a generalized linear model (GLM) with eGFR as the outcome
- Include **sex** and **ancestry** as predictors
- Assess whether sex and ancestry are significantly associated with eGFR

Exercise 9 – Questions

- Do sex and ancestry significantly explain variation in eGFR?
- Should these variables be included as covariates in GWAS? Why?

Exercise 10: Final Phenotype and Genotype Cleaning

Based on the previous steps, generate final cleaned datasets for GWAS.

Tasks

- Create a final cleaned phenotype file
- Update the PLINK files to reflect the same set of individuals
- Ensure that IDs are consistent between genotype and phenotype data

Exercise 10 – Final Question

- How can inadequate phenotype QC impact statistical power and false-positive rates in GWAS?