Genotype Quality Control for GWAS - Part 1 : genetics

KidneyGen Africa Course 2026 – Wits University

Jean-Tristan Brandenburg

January 2026

Contents

1 Genotype Quality Control

1.1 GWAS Quality Control (QC): Context and Objectives

The objective of this exercise is to understand the **basics of genotype quality control (QC)** and how QC affects **bias in GWAS results**. Most genotype QC steps are performed **before imputation** to ensure that only high-quality data are carried forward for analysis.

In this session, we will perform a subset of key QC procedures, focusing on:

- Missingness
- Allele frequency
- Hardy–Weinberg equilibrium (HWE)
- Relatedness

1.2 Missingness

Missingness can reflect **genotyping problems**. Variants or individuals with a high proportion of missing genotype calls often indicate:

- Poor DNA quality or contamination
- Technical issues during array hybridization or scanning
- Systematic batch effects

By filtering out SNPs and individuals with high missingness rates (e.g., $> 5\%$), we reduce technical noise and improve the reliability of downstream analyses.

1.3 Frequency (Minor Allele Frequency, MAF)

Purpose: To remove unreliable or non-informative variants.

Variants with very low frequency ($< 1\%$) are harder to genotype accurately and tend to:

- Have higher genotyping error rates
- Fail imputation or have low INFO/R^2 scores
- Be poorly represented in reference panels

Filtering variants with MAF < 0.01 reduces noise and improves imputation accuracy by relying on variants with stable linkage disequilibrium structures.

1.4 Hardy–Weinberg Equilibrium (HWE)

Deviation from Hardy–Weinberg expectations can indicate **technical or biological problems**, including:

- Allele miscalls (strand flips)
- Poor cluster separation

- Batch effects
- Population stratification
- Sample contamination or sex mislabeling

Removing SNPs that strongly deviate from HWE improves the quality of variants used in GWAS and imputation.

1.5 Relatedness

GWAS assumes that samples are **independent**. Related individuals (e.g., siblings or parent–child pairs) can inflate association signals due to shared ancestry.

Why check relatedness:

- Identify duplicates or sample mix-ups
- Avoid overrepresentation of families
- Preserve independence between samples

Relatedness is commonly measured using the **pi-hat** statistic (identity-by-descent), estimated by **PLINK** or **PLINK 2**. PLINK 2 implements the **KING** algorithm for efficient detection of related individuals.

1.5.1 Typical Relatedness Thresholds

Table 1: Common pi-hat thresholds used for relatedness filtering

Relationship	**pi-hat range**	**Action**
Duplicate / same individual	> 0.9	Remove one
First-degree relatives	0.35 – 0.9	Remove one per pair
Second-degree relatives	0.125 – 0.35	Often acceptable

1.6 Other Possible QC Steps

To explore more comprehensive genotype QC pipelines:

- H3A GWAS QC Pipeline — a reproducible workflow developed by H3ABioNet covering missingness, MAF, HWE, relatedness, population structure, and sex checks.
- GWAS Tutorial: Quality Control and Statistical Analysis — a tutorial covering GWAS QC and analysis.
- QCGWAS (R package) — an R-based framework for QC diagnostics and visualization.

These tools help standardize QC procedures and ensure reproducibility before imputation or association testing.

2 Practical Steps

2.1 Data Used

The dataset used for this practical session is located in the following directory:

`1_Data_beforeqc/`

plink file

`1_Data_beforeqc/afreur`

This dataset contains genotype data **prior to any quality control steps** and will be used as the starting point for all QC procedures described in this exercise.

2.2 Genotype Cleaning: MAF, GENO, HWE

The objective is to exclude SNPs and individuals failing basic QC:

- SNPs with MAF < 0.01
- SNPs with HWE $p < 1 \times 10^{-6}$
- SNP missingness > 0.05
- Individual missingness > 0.005

All filters can be applied using **PLINK** in a single command.

2.2.1 Common PLINK Arguments

- `--bfile`: Input binary PLINK files
- `--maf`: Minor allele frequency filter
- `--hwe`: Hardy–Weinberg equilibrium filter
- `--geno`: SNP missingness filter
- `--mind`: Sample missingness filter
- `--make-bed`: Create new binary files
- `--out`: Output prefix

```
plink --help bfile
```

2.2.2 Questions

- How many SNPs and individuals were excluded by each filter?
- Which filter had the largest impact on the dataset?

2.3 Relatedness Filtering

Individuals with high relatedness are removed using the **KING** algorithm implemented in PLINK 2.

2.3.1 Key PLINK2 Arguments

- `--bfile`: Input binary files
- `--king-cutoff`: Relatedness threshold
- `--make-bed`: Create new dataset
- `--out`: Output prefix

2.3.2 Question

- How many individuals were removed due to relatedness?