Genetic Quality Control and Imputation procedure

Natalia Vilor Tejedor
BarcelonaBeta Brain Research Center

Genotype Quality Control

Sample QC

Steps:

- **Sample call rate** (Proportion of genotypes successfully called (non-missing) for each individual.). **Individuals with a high percentage of missing genotypes (e.g., >5%) may indicate poor DNA quality or technical issues during genotyping.**

- **Sex check.** Compare genetically inferred sex (based on sex chromosome data) with the sex reported in metadata. **Discrepancies might indicate sample mix-ups or data labeling errors.**

- **Heterozygosity outliers** (detects contamination or inbreeding). High heterozygosity = sample contamination (mixed DNA). **Low heterozygosity = inbreeding or population-specific features.**

- **Duplicates or relatedness** (IBD/kinship analysis)

Marker-Level QC

Steps:

- **SNP call rate.** The proportion of samples for which a SNP was successfully genotyped (i.e., not missing; remove SNPs with >5% missing data). **Poorly genotyped SNPs may introduce noise and bias downstream analyses.**

- **Minor Allele Frequency (MAF) threshold.** **Frequency of the less common allele at a given SNP.** SNPs with MAF below 1–5% are often excluded (depends on study goals).

- **Hardy-Weinberg Equilibrium (HWE) test.** A test to check if genotype frequencies at a SNP fit expected proportions under random assignation. **Common cutoff is p < 1e-6 for controls only (not cases, due to potential disease association).**

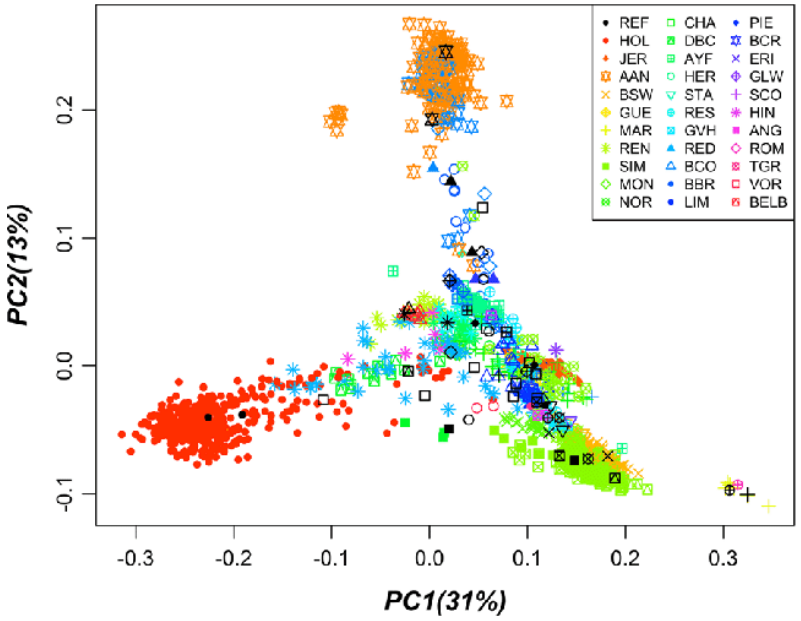
Population Structure and Ancestry

Population stratification refers to ancestral differences in allele frequencies.

- Principal Component Analysis (PCA) helps identify clusters of individuals with similar ancestry.
- PCA can reveal samples that deviate from expected population structure.

  These may be:
  - Sample **contamination, Mislabeled ancestry, Technical artifacts** (e.g., batch effects from different genotyping centers or chips)

Projects like the **1000 Genomes Project** provide genotypes from known global populations. You can project your study samples onto this reference PCA space.



Tools Commonly Used

- PLINK – for most QC steps
- KING or REAP – relatedness checks
- EIGENSOFT – PCA
- R packages – visualization and custom filtering

Hands-on
Pipeline Genetic QC

Genotype Imputation

Genotype Imputation Overview

- Increases genomic coverage
- Improves power and resolution of association studies
- TOPMed reference panel offers high-quality, diverse haplotypes (other panels => HRC)
- **Requires pre-imputation QC**

Pre-Imputation QC Steps

- Remove individuals and SNPs failing basic QC (missingness, MAF, HWE)
- Align strand and reference alleles to reference panel
- Split autosomes and remove duplicated SNPs

Check: https://www.chg.ox.ac.uk/~wrayner/tools/

Imputation with TOPMed (or Michigan)

- Use TOPMed Imputation Server
  https://imputation.biodatacatalyst.nhlbi.nih.gov
- But also.. Michigan Imputation Server
  https://imputationserver.sph.umich.edu/#!pages/home

- Upload phased VCF files and select reference panel

- Choose appropriate phasing/imputation parameters

- Download imputed data and QC results

NIH National Heart, Lung, and Blood Institute | BioData CATALYST TOPMed Imputation Server

Michigan Imputation Server 2
Free Next-Generation Genotype Imputation Platform

Post-Imputation QC and best practices

- Filter on imputation quality (e.g., R^2 or INFO score > 0.3 or 0.8)
- Exclude poorly imputed variants/monomorphic SNPs
- Prepare formats (e.g. Plink…)
- Document imputation pipeline and settings

Hands-on
Pipeline Genetic Imputation