GENOME-WIDE ARRAY DATA VARIANT AND INDIVIDUAL FILTERING

After “genotype calling” from CEL files to PLINK format files (e.g. Axiom Analysis Suite 2.0 if HumanOrigins platform)

0. INITIAL VARIANT FILTERS (VCFtools/0.1.14) *Perform independently on each dataset*

0.1. Exclude sites with genotype missingness > 5%.

0.2. Exclude individuals with genotype missingness > 10%

0.3. Filter SNPs in HW disequilibrium pvalue 10^{-6}

1. RELATEDNESS ESTIMATION (PLINK/1.9b) *Perform independently on each dataset*

1.1. Remove sites with minor-allele frequency (MAF) < 1%

1.2. Estimate relatedness

1.3. Remove related individuals (the one with less missing sites)

2. MERGING WITH REFERENCE PANELS (PLINK/1.9b) *Perform independently on each dataset*

2.1. Merging datasets from step 0.3 (without related individuals): keeping only overlapping SNPs

2.2. Remove sites with minor-allele frequency (MAF) < 1%

EXTRA filtering: Linkage disequilibrium pruning (for allele-frequency-based analyses)

window = 200; step = 25; r^2 = 0.5 | window = 200; step = 25; r^2 = 0.5