**Genome wide Association Studies Practical Session**

Monday 6th March 2023 – 11:30 – 13:00 GMT

***Instructions***

* This practical session assumes basic programming (eg bash scripting) and computing experience.
* You will need to work on a Unix/Linux environment (shell prompt).

***Software to install:***

MobaXterm (For Windows Users) (https://mobaxterm.mobatek.net/download.html)

Plink 1.9 (<https://www.cog-genomics.org/plink/>)

R (https://support--rstudio-com.netlify.app/products/rstudio/download/)

**Instruction:** In this practical, we will go through the steps in performing quality control (QC) of genotype data from a simulated genome-wide association study (HapMap\_3\_r3\_1)

Colour Interpretation:

Black : QC Steps

Blue: Description of the code

Red: Code to run

1. **Determine SNPs and Individual Missingness**

# Investigate missingness per individual and per SNP and make histograms.

plink --bfile HapMap\_3\_r3\_1 --missing

# Generate plots to visualize the missingness results.

Rscript --no-save hist\_miss.R

# Delete SNPs and individuals with high levels of missingness

# Delete SNPs with missingness >0.2

plink --bfile HapMap\_3\_r3\_1 --geno 0.2 --make-bed --out HapMap\_3\_r3\_2

# Delete individuals with missingness >0.2.

plink --bfile HapMap\_3\_r3\_2 --mind 0.2 --make-bed --out HapMap\_3\_r3\_3

# Delete SNPs with missingness >0.02.

plink --bfile HapMap\_3\_r3\_3 --geno 0.02 --make-bed --out HapMap\_3\_r3\_4

# Delete individuals with missingness >0.02.

plink --bfile HapMap\_3\_r3\_4 --mind 0.02 --make-bed --out HapMap\_3\_r3\_5

1. **Sex Discrepancy**

# Subjects who were a priori determined as females must have a F value of <0.2, and subjects who were a priori determined as males must have a F value >0.8.

plink --bfile HapMap\_3\_r3\_5 --check-sex

# Generate plots to visualize the sex-check results.

Rscript --no-save gender\_check.R

# This command generates a list of individuals with the status “PROBLEM”

grep "PROBLEM" plink.sexcheck| awk '{print$1,$2}'> sex\_discrepancy.txt

# This command removes the list of individuals with the status “PROBLEM”

plink --bfile HapMap\_3\_r3\_5 --remove sex\_discrepancy.txt --make-bed --out HapMap\_3\_r3\_6

1. **Generate autosomal SNPs and delete SNPs with low MAF**

# Select autosomal SNPs only (i.e., from chromosomes 1 to 22).

awk '{ if ($1 >= 1 && $1 <= 22) print $2 }' HapMap\_3\_r3\_6.bim > snp\_1\_22.txt

plink --bfile HapMap\_3\_r3\_6 --extract snp\_1\_22.txt --make-bed --out HapMap\_3\_r3\_7

# Generate a plot of the MAF distribution.

plink --bfile HapMap\_3\_r3\_7 --freq --out MAF\_check

Rscript --no-save MAF\_check.R

# Remove SNPs with a low MAF frequency.

plink --bfile HapMap\_3\_r3\_7 --maf 0.05 --make-bed --out HapMap\_3\_r3\_8

1. **Hardy-Weinberg equilibrium (HWE).**

# Check the distribution of HWE p-values of all SNPs.

plink --bfile HapMap\_3\_r3\_8 --hardy

# By default the --hwe option in plink only filters for controls.

# Therefore, we use two steps, first we use a stringent HWE threshold for controls, followed by a less stringent threshold for the case data.

plink --bfile HapMap\_3\_r3\_8 --hwe 1e-6 --make-bed --out HapMap\_hwe\_filter\_step1

# The HWE threshold for the cases filters out only SNPs which deviate extremely from HWE.

# This second HWE step only focusses on cases because in the controls all SNPs with a HWE p-value < hwe 1e-6 were already removed

plink --bfile HapMap\_hwe\_filter\_step1 --hwe 1e-10 --hwe-all --make-bed --out HapMap\_3\_r3\_9

1. **Heterozygosity**

# Generate a plot of the distribution of the heterozygosity rate of your subjects.

# And remove individuals with a heterozygosity rate deviating more than 3 sd from the mean.

# Checks for heterozygosity are performed on a set of SNPs which are not highly correlated.

plink --bfile HapMap\_3\_r3\_9 --exclude inversion.txt --range --indep-pairwise 50 5 0.2 --out indepSNP

plink --bfile HapMap\_3\_r3\_9 --extract indepSNP.prune.in --het --out R\_check

# Plot of the heterozygosity rate distribution

Rscript --no-save check\_heterozygosity\_rate.R

# The following code generates a list of individuals who deviate more than 3 standard deviations from the heterozygosity rate mean.

Rscript --no-save heterozygosity\_outliers\_list.R

# Output of the command above: fail-het-qc.txt .

# Adapt this file to make it compatible for PLINK, by removing all quotation marks from the file and selecting only the first two columns.

sed 's/"// g' fail-het-qc.txt | awk '{print$1, $2}'> het\_fail\_ind.txt

# Remove heterozygosity rate outliers.

plink --bfile HapMap\_3\_r3\_9 --remove het\_fail\_ind.txt --make-bed --out HapMap\_3\_r3\_10

1. **Relatedness.**

# Check for relationships between individuals with a pihat > 0.2.

plink --bfile HapMap\_3\_r3\_10 --extract indepSNP.prune.in --genome --min 0.2 --out pihat\_min0.2

# The following commands will visualize specifically these parent-offspring relations, using the z values.

awk '{ if ($8 >0.9) print $0 }' pihat\_min0.2.genome>zoom\_pihat.genome

# Generate a plot to assess the type of relationship.

Rscript --no-save Relatedness.R

# To demonstrate that the majority of the relatedness was due to parent-offspring we only include founders (individuals without parents in the dataset).

plink --bfile HapMap\_3\_r3\_10 --filter-founders --make-bed --out HapMap\_3\_r3\_11

# Now we will look again for individuals with a pihat >0.2.

plink --bfile HapMap\_3\_r3\_11 --extract indepSNP.prune.in --genome --min 0.2 --out pihat\_min0.2\_in\_founders

# For each pair of 'related' individuals with a pihat > 0.2, we recommend to remove the individual with the lowest call rate.

plink --bfile HapMap\_3\_r3\_11 --missing

# Use an UNIX text editor (e.g., vi(m) ) to check which individual has the highest call rate in the 'related pair'.

# Generate a list of FID and IID of the individual(s) with a Pihat above 0.2, to check who had the lower call rate of the pair.

# In our dataset the individual 13291 NA07045 had the lower call rate.

vi 0.2\_low\_call\_rate\_pihat.txt

i

13291 NA07045

# Press esc on keyboard!

:x

# Delete the individuals with the lowest call rate in 'related' pairs with a pihat > 0.2

plink --bfile HapMap\_3\_r3\_11 --remove 0.2\_low\_call\_rate\_pihat.txt --make-bed --out HapMap\_3\_r3\_12

################################################################################################################################