**Plink Smokescreen GWAS Progress Report**

The R code and PLINK instructions used for this progress report can be found at <https://github.com/MareesAT/GWA_tutorial/blob/master/1_QC_GWAS.zip>

Legend:

1. Normal text is written in black
2. Plink commands are written and commented in green (# indicates a comment)
3. Bash commands are written in blue (# indicates a comment)

The following files were opened using PLINK software:

“Smokescreen\_NIDA\_Study60\_Grassi\_clean.fam” “Smokescreen\_NIDA\_Study60\_Grassi\_clean.bim” “Smokescreen\_NIDA\_Study60\_Grassi\_clean.bed”

The original “Smokescreen\_NIDA\_Study60\_Grassi\_clean.fam” was missing phenotype values. All the phenotype values in the original file were “-9”, indicating that the phenotype value was missing. The original “Smokescreen\_NIDA\_Study60\_Grassi\_clean.fam” file was cross referenced with the “5a\_SubjectPhenotypes\_DS\_NIDA\_Study60\_Grassi.txt” file. This way we were able to determine the phenotype for each participant in the study. The new files with the correct phenotypes (“1” = Control, “2” = Case) were named:

“run1.fam”

“run1.bim”

“run1.bed”

**STEP 1:** Delete SNPs and Individuals with high levels of missingness (See Appendix: Table 1, Box 1).

# Delete SNPs with missingness >0.02.

plink --bfile run1 --geno 0.02 --make-bed --out run1\_p1\_s1

40957 variants removed due to missing genotype data (--geno).

# Delete individuals with missingness >0.02.

plink --bfile run1\_p1\_s1 --mind 0.02 --make-bed --out run1\_p2\_s1

18 people removed due to missing genotype data (--mind).

**STEP 2:** Check for sex discrepancies (See Appendix: Table 1, Box 2).

# Create file with individuals flagged PROBLEM

plink --bfile run1\_p2\_s1 --check-sex

10813 Xchr and 0 Ychr variant(s) scanned, 5 problems detected (--check-sex)

# 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 run1\_p2\_s1 --remove sex\_discrepancy.txt --make-bed --out run1\_p1\_s2

1524 people remaining (--remove)

**STEP 3:** Generate a bfile with autosomal SNPs only and delete SNPs with a low minor allele frequency (See Appendix: Table 1, Box 3).

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

awk '{ if ($1 >= 1 && $1 <= 22) print $2 }' run1\_p1\_s2.bim > snp\_1\_22.txt

# Generate a bfile with only autosomal SNPs

plink --bfile run1\_p1\_s2 --extract snp\_1\_22.txt --make-bed –out run1\_p1\_s3

505104 variants remaining (--extract)

# Remove SNPs with a low MAF frequency.

plink --bfile run1\_p1\_s3 --maf 0.01 --make-bed --out run1\_p2\_s3

89127 variants removed due to minor allele threshold(s) (--maf)

**STEP 4:** Delete SNPs which are not in Hardy-Weinberg equilibrium (See Appendix: Table 1, Box 4).

# 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 run1\_p2\_s3 --hwe 1e-6 --make-bed --out run1\_p1\_s4

689 variants removed due to Hardy-Weinberg exact test (--hwe).

# 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 run1\_p1\_s4 --hwe 1e-10 --hwe-all --make-bed --out run1\_p2\_s4

107 variants removed due to Hardy-Weinberg exact test (--hwe-all).

**Step 5:** Remove individuals with a heterozygosity rate deviating more than 3 SD from the mean (See Appendix: Table 1, Box 5).

# Checks for heterozygosity are performed on a set of SNPs which are not highly correlated. Therefore, to generate a list of non-(highly)correlated SNPs, we exclude high inversion regions (inversion.txt [text file with High LD regions]) and prune the SNPs using the command --indep-pairwise’. The parameters ‘50 5 0.2’ stand respectively for: the window size, the number of SNPs to shift the window at each step, and the multiple correlation coefficient for a SNP being regressed on all other SNPs simultaneously.

Contents of inversion.txt:

6 25500000 33500000 8 HLA

8 8135000 12000000 Inversion8

17 40900000 45000000 Inversion17

plink --bfile run1\_p2\_s4 --exclude inversion.txt --range --indep-pairwise 50 5 0.2 --out indepSNP

Pruning complete. 244512 of 405947 variants removed.

# This file contains your pruned data set.

plink --bfile run1\_p2\_s4 --extract indepSNP.prune.in --het --out R\_check

161435 variants scanned, report written to R\_check.het (--het).

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

On R run the code found on GitHub heterozygosity\_outliers\_list.R

Output of the code above is: 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 run1\_p2\_s4 --remove het\_fail\_ind.txt --make-bed --out run1\_p1\_s5

1516 people remaining (--remove).

**STEP 6:** It is essential to check datasets you analyze for cryptic relatedness. Assuming a random population sample we are going to exclude all individuals above the pihat threshold of 0.2.

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

plink --bfile run1\_p1\_s5 --extract indepSNP.prune.in --genome --min 0.2 --out pihat\_min0.2

IBD calculations complete.

Finished writing pihat\_min0.2.genome .

# 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

# We aim to remove all 'relatedness' from our dataset. 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 run1\_p1\_s5 --filter-founders --make-bed --out run1\_p1\_s6

1516 people, 1502 founders and 14 nonfounders present

14 people removed due to founder status (--filter-founders).

1502 people, 1502 founders and 0 nonfounders present

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

plink --bfile run1\_p1\_s6 --extract indepSNP.prune.in --genome --min 0.2 --out pihat\_min0.2\_in\_founders

161435 variants remaining (--extract).

# For each pair of 'related' individuals with a pihat > 0.2, remove the individual with the highest missing call rate.

plink --bfile run1\_p1\_s6 --missing

Sample missing data report written to plink.imiss, and variant-based missing data report written to plink.lmiss (--missing).

# I used notepad to manually determine which of the individual on the 12 pairs of ‘related’ individuals had the highest missing call rate and made a new document named “0.2\_high\_missing\_call\_rate\_pihat.txt” with the FID and IID of the ‘related’ individuals with the highest missing call rates.

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

plink --bfile run1\_p1\_s6 --remove 0.2\_high\_missing\_call\_rate\_pihat.txt --make-bed --out run1\_p2\_s6

THAT’S ALL FOLKS!

**References:**

Marees, A. T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie‐Claire, C., & Derks, E. M. (2018). A tutorial on conducting genome‐wide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research*, *27*(2). https://doi.org/10.1002/mpr.1608

<https://github.com/MareesAT/GWA_tutorial/blob/master/1_QC_GWAS.zip>

**Appendix:**

![A close up of a newspaper

Description automatically generated]()