**Plink Smokescreen GWAS Progress Report**

Files below 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:

“smokescreen.fam”

“smokescreen.bim”

“smokescreen.bed”

These files were subjected to two different quality controls, the first and less stringent one is described in detail by the steps below

Before moving on: See color code legend at end of document for better understanding

Disclaimer: 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>

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

# Delete SNPs with missingness >0.2.

plink –bfile smokescreen --geno 0.2 --make-bed --out smokescreen\_s1\_p1

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

# Delete individuals with missingness >0.2.

plink --bfile smokescreen\_s1\_p1 --mind 0.2 --make-bed --out smokescreen\_s1\_p2

0 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 smokescreen\_s1\_p2 --check-sex

11397 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 smokescreen\_s1\_p2 --remove sex\_discrepancy.txt --make-bed --out smokescreen\_s2\_p1

1542 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 }' smokescreen\_s2\_p1.bim > snp\_1\_22.txt

# Generate a bfile with only autosomal SNPs

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

545368 variants remaining (--extract).

# Remove SNPs with a low MAF frequency.

plink --bfile smokescreen\_s3\_p1 --maf 0.05 --make-bed --out smokescreen\_s3\_p2

183082 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 smokescreen\_s3\_p2 --hwe 1e-6 --make-bed --out smokescreen\_s4\_p1

1218 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 smokescreen\_s4\_p1 --hwe 1e-10 --hwe-all --make-bed --out smokescreen\_s4\_p2

114 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 smokescreen\_s4\_p2 --exclude inversion.txt --range --indep-pairwise 50 5 0.2 --out indepSNP

Pruning complete. 229260 of 353170 variants removed.

# This file contains your pruned data set.

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

123910 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 smokescreen\_s4\_p2 --remove het\_fail\_ind.txt --make-bed --out smokescreen\_s5\_p1

1529 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 smokescreen\_s5\_p1 --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 smokescreen\_s5\_p1 --filter-founders --make-bed --out smokescreen\_s6\_p1

1529 people, 1515 founders and 14 nonfounders present

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

1515 people, 1515 founders and 0 nonfounders present.

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

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

123910 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 smokescreen\_s6\_p1 --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 smokescreen\_s6\_p1 --remove 0.2\_high\_missing\_call\_rate\_pihat.txt --make-bed --out smokescreen\_s6\_p2

***STEP 6 Concludes the Quality Control Steps using less stringent filters.***

The quality control with more stringent filters only differed in STEP 1. On the less stringent quality control we deleted SNPs with missingness >0.2 and individuals with missingness > 0.2. On the more stringent quality control we deleted SNPs with missingness >0.02 and individuals with missingness > 0.02. STEP 2-6 were carried on the same manner as described above for the more stringent quality control process. It is also worth noting that the mother files for the more stringent quality control were named “smokescreen\_stringent.fam”, “smokescreen\_stringent.bim”, “smokescreen\_stringent.bed”.

**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]()

**Legend:**

Normal Text

# Comment to describe line of PLINK command

Plink code typed in Windows Command Prompt

# Comment to describe Linux command

Linux Command in Bash