

# Polygenic Risk Score Analyses Workshop 2022



Day 1: Introduction to  
PLINK

# Day 1 Timetable

Time	Title	Presenter
9:00 - 9:15	Welcome Address	Dr Daneshwar and Dr Baichoo
9:15 - 9:30	Opening Speech from Organisers	Dr Segun Fatumo and Dr Nicki Tiffin
9:30 - 10:30	<u>Lecture</u> : Background to PRS: GWAS & relevant Statistics	Dr Paul O'Reilly
10:30 - 11:00	Coffee Break and Q&A	-
11:00 - 12:00	<u>Practical</u> : Introduction to Bash and R	Dr Paul O'Reilly & Tutors
12:00 - 13:30	Lunch	-
13:30 - 15:00	<u>Practical</u> : Introduction to PLINK I - Basics	Dr Conrad Iyegbe & Tutors
15:00 - 15:30	Coffee Break and Q&A	-
15:30 - 16:30	<u>Practical</u> : Introduction to PLINK II - QC & GWAS	Dr Conrad Iyegbe & Tutors

# Contents

<b>Day 1 Timetable</b>	<b>1</b>
Day 1 Timetable . . . . .	1
<b>1 Introduction to PLINK I: Basics</b>	<b>3</b>
1.1 Key Learning Outcomes . . . . .	3
1.2 Introduction . . . . .	3
1.3 Command line basics . . . . .	3
1.4 Exploring Data Sets . . . . .	4
1.5 Recoding alleles as counts . . . . .	5
1.6 PLINK website . . . . .	6
1.7 Write SNP list and extract SNPs . . . . .	6
<b>2 Introduction to PLINK II: Performing QC &amp; GWAS</b>	<b>8</b>
2.1 Key Learning Outcomes . . . . .	8
2.2 Generate summaries to perform QC . . . . .	8
2.2.1 Individual missingness . . . . .	8
2.2.2 SNP Missingness . . . . .	9
2.2.3 Hardy-Weinberg Equilibrium . . . . .	9
2.2.4 Allele frequencies . . . . .	9
2.3 Apply QC filters . . . . .	10
2.3.1 Apply individual missingness thresholds . . . . .	10
2.3.2 Apply SNP missingness and MAF thresholds . . . . .	11
2.3.3 Apply Hardy-Weinberg thresholds . . . . .	11
2.4 Perform GWAS . . . . .	12
2.4.1 Case/Control GWAS - no covariates . . . . .	12
2.4.2 Case/Control GWAS - with covariates . . . . .	12
<b>License</b>	<b>14</b>

# 1 Introduction to PLINK I: Basics

## 1.1 Key Learning Outcomes

After completing this practical, you should be able to:

1. Explore and generate genetic data sets needed for GWAS
2. Recode and reorder allelic data
3. Use the PLINK website
4. Select and exclude lists of samples and SNPs



All data used in this workshop are **simulated**. They have no specific biological meaning and are for demonstration purposes only.

## 1.2 Introduction

PLINK is the most popular software program for performing genome-wide association analyses. It is extremely extensive, allowing a huge number of analyses to be performed. It also includes many options for reformatting your data and provides useful data summaries. Software packages are usually best learnt by having a go at running some of their basic applications and progressing from there (rather than reading the entire user manual first!) - so we begin by running some basic PLINK commands and then work steadily towards performing more sophisticated analyses through these PLINK tutorials.

## 1.3 Command line basics

In all of the instructions below, *italics* indicate commands - they can be directly copy and pasted. Anything in between the symbols <> needs to be changed in some way. For example, <file\_name> indicates that you should replace that entire statement (including the <> symbols) with the appropriate file name. **Bold** indicates non-command-line instructions (e.g. **right-click**)

1. Open up a terminal

2. Navigate to the Practical/ folder and then the Day 1 working directory  
(`cd <directory_name>`)
3. List all files in this directory by typing `"ls"`
4. Test PLINK with no input by typing `./Software/plink`
5. Note that you can see lots of PLINK options by using the built-in help function:

```
1 | ./Software/plink --help
```



Calling PLINK with no output will test if PLINK is installed and available in directory, because you should see some output showing the PLINK license and some commands. If you do not see this then please ask for help now!

### 1.4 Exploring Data Sets

1. Open an Explorer window ('Finder' on a Mac) and navigate to your PLINK working directory.



An explorer window should show same files as the 'ls' command

2. Open the file called 'D1D.map' with a Text Editor e.g. by typing **right-click** > **Open**.
3. Open the file 'D1D.ped'. Note this is a large file - if it will not open or is very slow, skip this step.
4. Go to the PLINK website and investigate the format of the MAP/PED files (<http://zzz.bwh.harvard.edu/plink/download.shtml>)  
(Look in the blue column on the left side)



What are the 4 columns in the map file?

What are the first 6 columns of ped file?

What information is in the remaining columns of the ped file?

5. Create 'binary' format PLINK files using the recode command:

```
1 | ./Software/plink
2 | --file Data/D1D
3 | --make-bed
```

4  --out Data/D1D

6. List files (ls) and check which new files have appeared
7. Open and examine files ending .bim and .fam. Do not open the .bed file.
8. Open and skim the '.log' file.







How is the fam file similar to the ped file? How is it different? Use the PLINK website if necessary.

How is the bim file similar to the map file? How is it different?

## 1.5 Recoding alleles as counts

Genotype data in allele count format is very useful, for example to use in regression modelling in statistical software such as R.

1. Generate the D1D data in allele count format:

```
1  ./Software/plink
2  --bfile Data/D1D
3  --recodeA
4  --out Data/D1D_AC
```



There are several options for recoding SNPs in different ways - more information on the PLINK website (see next section).

Again note that a log file was generated - skim the log file or screen output.



Look inside the .raw file. What do you think the 0/1/2 represent?

Do there appear to be more 0s or more 2s? Why might this be?

## 1.6 PLINK website

Go to <http://zzz.bwh.harvard.edu/plink/download.shtml> and skim through the front page to get an idea of PLINK's functionality. Note the list of clickable links on the left side of the website.

Under 'Data Management' (click the heading on the left) and read the list of the different ways you may want to recode and reorder data sets. Don't attempt to read much further as this is a very large and detailed section - a useful future resource but too much for today.

Under 'Data Management', click 'Write SNP list' and read the instructions there to write SNP lists.

## 1.7 Write SNP list and extract SNPs

You will now use the information that you found on the PLINK website to create a command to extract a list of SNPs. Below is a list of requirements - try to do this before you go to the end of this section, where the full command is given and explained.

1. Set the D1D binary file as input
2. Set MAF threshold to 0.05
3. Set SNP missingness threshold to 0.05
4. Add the appropriate command to write out a snp list containing only those SNPs with MAF above 0.05 and missingness below 0.05
5. Use 'D1D\_snps' as the output file name
6. After the command has run, check the output for your SNP list and look at it with the default viewer.

You will now use the SNP list that you have created to extract those SNPs and create a new set of data files in a single command.

1. Use the D1D binary file set as input
2. Find the command for extracting a set of SNPs listed in a file (hint: Data Management section) and combine it with a command that you learned above to create binary files
3. Use the output file name 'D1D\_MAF\_MISS'



Log files are useful to check that the number of SNPs and samples is as expected. Always check your log files to be sure that they are sensible.

SNP lists can also be used to EXCLUDE SNPs - select 'exclude' above instead of

‘extract’.

Sample ID lists can also be used to ‘keep’ or ‘remove’ individuals in the same ‘Filter’ window. Note that BOTH sample IDs (FID IID, separated by a space) are required in the sample list file.

Solutions:

```
1 ./Software/plink
2 --bfile Data/D1D
3 --maf 0.05
4 --geno 0.05
5 --write-snpList
6 --out Data/D1D_snps
```

```
1 ./Software/plink
2 --bfile Data/D1D
3 --extract Data/D1D_snps.snpList
4 --make-bed
5 --out Data/D1D_MAF_MISS
```



## 2 Introduction to PLINK II: Performing QC & GWAS

### 2.1 Key Learning Outcomes

After completing this practical, you should be able to:

1. Generate summaries of the data needed for QC
2. Apply QC thresholds
3. Perform GWAS

### 2.2 Generate summaries to perform QC

There are many kinds of summaries of the data that can be generated in PLINK in order to perform particular quality control (QC) steps, which help to make our data more reliable. Some of these involve summaries in relation to the individuals (e.g. individual missingness, sex-check) and some relate to summaries of SNP data (e.g. MAF, Hardy-Weinberg Equilibrium). Over the next few sub-sections you will go through some examples of generating summary statistics that can be used to perform QC.

#### 2.2.1 Individual missingness

1. Use the D1D binary files to generate files containing missingness information (--missing). Use the output file name 'D1D\_miss'
2. Open the 2 files that were generated (lmiss & imiss).



What do the two output files contain?

In the imiss file, what is the meaning of the data in the column headed "F\_MISS"?

### 2.2.2 SNP Missingness

1. Use the D1D binary files to generate files containing missingness information (`--missing`). Use the output file name 'D1D\_miss'
2. Look inside the file containing SNP missingness information: D1D\_miss.lmiss.



What is the meaning of the value under the heading F\_MISS?

What does the command '`--test-missing`' do, and why might it be useful?

### 2.2.3 Hardy-Weinberg Equilibrium

1. Generate HWE statistics using the `--hardy` option. Use output file name D1D\_hardy.
2. Open and examine results.



Why are there multiple rows for each SNP, and what does each mean?

Which of the rows do you think should be used to exclude SNPs from the subsequent association analysis (if any) for failing the HWE test? Why?

### 2.2.4 Allele frequencies

1. Generate allele frequencies using the command `--freq`. Use D1D\_freq as the output name.
2. Examine the output.



What is the heading of the column that tells you which nucleotide is the minor allele?

\*Note - this is important to remember as many PLINK files use this notation - the minor allele is always labelled the same way

## 2.3 Apply QC filters

**There are different strategies for performing QC on your data:**

- (a) create lists of SNPs and individuals and use `--remove`, `--extract`, `--exclude`, `--include` to create new file sets (good for documentation, collaboration)
- (b) apply thresholds one at a time and generate new bed/bim/fam files (good for applying sequential filters)
- (c) use options (e.g. `--maf`) in other commands (e.g. `--assoc`) to remove SNPs or samples at required QC thresholds during analysis.



We have already seen how to select or exclude individuals or SNPs by first creating lists (a), so in this section we will set thresholds to generate new file sets in a single command. However, it is useful to have lists of all SNPs and individuals excluded pre-analysis, according to the reason for exclusion, so generating and retaining such files using the techniques that we used before is good practice.

### 2.3.1 Apply individual missingness thresholds

1. Generate new binary file sets (`--make-bed`) from the 'D1D' binary file set, removing individuals with missingness greater than 3% using a single command (hint: In the 'Inclusion thresholds' section, see the 'Missing/person' sub-section). Use the output file name 'D1D\_imiss3pc'
2. Examine the output files (no need to open, and remember the bed file can not be read) and the log file



How many individuals were in the original file?

How many individuals were removed?

How many males and females were left after screening?

### 2.3.2 Apply SNP missingness and MAF thresholds

1. Create new binary file sets from the 'D1D\_imiss3pc' binary file set (NOT the original D1D files) by setting MAF threshold to 0.05 and SNP missingness threshold to 0.02 (See 'Inclusion thresholds' to obtain the correct threshold flags). Use the output file name 'D1D\_imiss3pc\_lmiss2pc\_maf5pc'
2. Examine the output files and the log file



How many SNPs were in the original file?

How many SNPs were removed for low minor allele frequency?

How many SNPs were removed for missingness?

### 2.3.3 Apply Hardy-Weinberg thresholds

1. Generate a new binary file set called 'D1D\_QC' from the D1D\_imiss3pc\_lmiss2pc\_maf5pc file, applying a HWE threshold of 0.0001.
2. This is our final, QC'ed file set.
3. Examine log and output files.

## PRACTICAL 2. INTRODUCTION TO PLINK II: PERFORMING QC & GWAS



How many SNPs were removed for HWE  $P$ -values below the threshold?

NOTE - it is useful to know how to do this but be careful about setting this threshold - strong association signals can cause departures from HWE and you may remove great results! Use a lenient threshold and apply to controls only to avoid this problem. HWE can also be checked post-hoc for each SNP.

## 2.4 Perform GWAS

### 2.4.1 Case/Control GWAS - no covariates

Run the following code, which performs a genetic association study using logistic regression on some case/control data:

```
1 ./Software/plink
2 --plink
3 --bfile D1D_QC
4 --logistic
5 --adjust
6 --pheno D1D.pheno1
7 --out Results/D1D_CC
```



What are the raw and Bonferonni-adjusted  $P$ -values for the top hit? What does this mean - is there a significant association?

Are there any other significant associations?

### 2.4.2 Case/Control GWAS - with covariates

Here we repeat the previous analysis but this time including some covariates. The file D1D.pcs1234 contains the first 4 principal components from a PCA on the genetic data.

1. Run the analysis specifying the covariates file:

```
1 ./Software/plink
2 --plink
3 --bfile D1D_QC
4 --logistic
5 --adjust
6 --pheno D1D.pheno1
7 --covar D1D.pcs.1234
8 --out Results/D1D_CC_PCadj
```



What are the raw and Bonferonni-adjusted P-values for the top hit? What does this mean - is there a significant association?

Suggest a reason for the different results when adjusting for the 4 PCs.

# License

This work is licensed under Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International Public License and the below text is a summary of the main terms of the full Legal Code (the full licence) available at <https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode>.

## You are free to:

- **Share** — copy and redistribute the material in any medium or format
- **Adapt** — remix, transform, and build upon the material

The licensor cannot revoke these freedoms as long as you follow the license terms.

## Under the following terms:

- **Attribution** — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
- **NonCommercial** — You may not use the material for commercial purposes.
- **ShareAlike** — If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.

No additional restrictions — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

## Notices:

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable exception or limitation.

No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as publicity, privacy, or moral rights may limit how you use the material.