Lab3

Omar Aldawy Ibrahim Aldawy 21010864

2025-03-10

Introduction

PLINK is a widely used open-source command-line tool designed for genome-wide association studies (GWAS) and population genetics research. It enables efficient analysis of large-scale genetic data, supporting tasks such as quality control, data manipulation, and statistical association testing. PLINK is optimized for performance, allowing researchers to handle massive datasets with millions of genetic variants and thousands of individuals. Its compatibility with various file formats and integration with other bioinformatics tools make it a valuable resource for geneticists, epidemiologists, and computational biologists. Originally developed for human genetics, PLINK is also applied in animal and plant genomics, making it a versatile tool in the field of genetics research.

Part 1: Plink Walkthrough

Task 1.1: Installation

Run this to get the directory of the plink executable after installation

```
which plink
```

Use the path retrived (if it dummy/plink just use dummy) in this command to add the plink executable to the PATH

```
echo 'export PATH=dummy:$PATH' >> ~/.bashrc
```

Source the bashrc file to apply the changes

```
source ~/.bashrc
```

Run this to check if plink is in the PATH

```
echo $PATH | tr ':' '\n'
```

Run this to check if plink is installed

```
plink --version
```

Now plink is installed and ready to use

```
omar-aldawy@omar-aldawy-ASUS-TUF-Dash-F15-FX516PM-FX516PM:~$ echo 'export PATH=/home/omar-aldawy/Programs/plink_linux_x86_64_20241022:$PATH' >> ~/.bashrc
omar-aldawy@omar-aldawy-ASUS-TUF-Dash-F15-FX516PM-FX516PM:~$ source ~/.bashrc
omar-aldawy@omar-aldawy-ASUS-TUF-Dash-F15-FX516PM-FX516PM:~$ which plink
/home/omar-aldawy/Programs/plink_linux_x86_64_20241022/plink
omar-aldawy@omar-aldawy-ASUS-TUF-Dash-F15-FX516PM-FX516PM:~$ plink --version
PLINK v1.9.0-b.7.7 64-bit (22 Oct 2024)
```

Figure 1: Plink setup

Task 1.2: Basic Commands

File formats are .bim, .fam and .bed.

Convert the files in the current format to PED/MAP format using:

```
plink --bfile your_input_filename --recode --out your_output_filename
```

```
omar-aldawy@omar-aldawy-ASUS-TUF-Dash-F15-FX516PM-FX516PM:~/University/Third year/second semester/Bio-Informatics/Labs/Bio-Informatics-Labs/Lab 3/Plink operations$ plink --bfile Qatari
156_filtered_pruned --recode --out outPut
PLINK v1.9.0-b.7.7 64-bit (22 Oct 2024)           cog-genomics.org/plink/1.9/
(C) 2005-2024 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to outPut.log.
Options in effect:
  --bfile Qatari156_filtered_pruned
  --out outPut
  --recode

15685 MB RAM detected; reserving 7842 MB for main workspace.
67735 variants loaded from .bim file.
156 people (49 males, 107 females) loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 156 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Warning: 1388 het. haploid genotypes present (see outPut.hh ); many commands
treat these as missing.
Total genotyping rate is 0.998816.
67735 variants and 156 people pass filters and QC.
```

Figure 2: file-conversion

The number of variants = 67735.

The number of samples = 156 people (49 males, 107 females).

Columns of .ped file are:

- Family ID
- Individual ID
- Paternal ID (0 if unknown)
- Maternal ID (0 if unknown)
- Sex (1 = Male, 2 = Female, 0 = Unknown)
- Phenotype (1 = Unaffected, 2 = Affected, -9 = Missing)
- Genotype data (Two alleles per SNP, space-separated)

```
omar-aldawy@omar-aldawy-ASUS-TUF-Dash-F15-FX516PM-FX516PM:~/University/Third year/second semester/Bio-Informatics/Labs/Bio-Informatics-Labs/Lab 3/Plink operations$ head -n 5 outPut.ped

QBC-092 QBC-092 0 0 2 -9 A A T C G G G A A G A G G G C C C C A G G C G G T T G G C T T C C C G G T T C C T T C C C C T T T C G G T T G G C C G A T C A A T T G A A A C C G G G G A C A A
 C C T C T T A A G G G G T T C C C C C C C G C G G G G T T T T G G T T C C T T C C C C G G T T C T A G T T C C A A G A G A T T C T G A C C T T C C A A T T T T T T C A A G C C C C C C G G
 A G T T C T T T C C T C G A C C A G T T G G T C C C C C C T G G T C C C A G T C A G C C G A C C C C G G C C T T A A A A A A C C A A C C C T A G G G T G C C T T T T A G T T C C G G G G
 T T T T A A C T A A A A G G C C T T C C G C C C T T T T G G G A T T G A T A G G C C A A T T C C G G G G T C G G T T G C A A A A C C C C T T C C C C T C C G G G G G G G A A G G C T C C
 A G A G C C A T G G G G C C T A A A G G C C T T C C G G G G T T C C A A C C C C A A C C T T T C A G C C A A A G T C A A A A A A C T C T A A A G T T A A C C C C T T C C G G C T T T C C
```

Figure 3: ped-header

```
head -n 5 your_output_filename.ped
```

Columns of .map file are:

- Chromosome Number

- SNP ID
- Genetic Distance (cM) (Can be 0 if unknown)
- Physical Position (bp) (Base pair position in the genome)

Figure 4: map-header

```
head -n 5 your_output_filename.map
```

Performing the Missing Call Rate with different thresholds:

```
plink --file data_file --geno threshold_value --recode --out output_file
```

Threshold = 1e-1

- Number of variants removed = 0

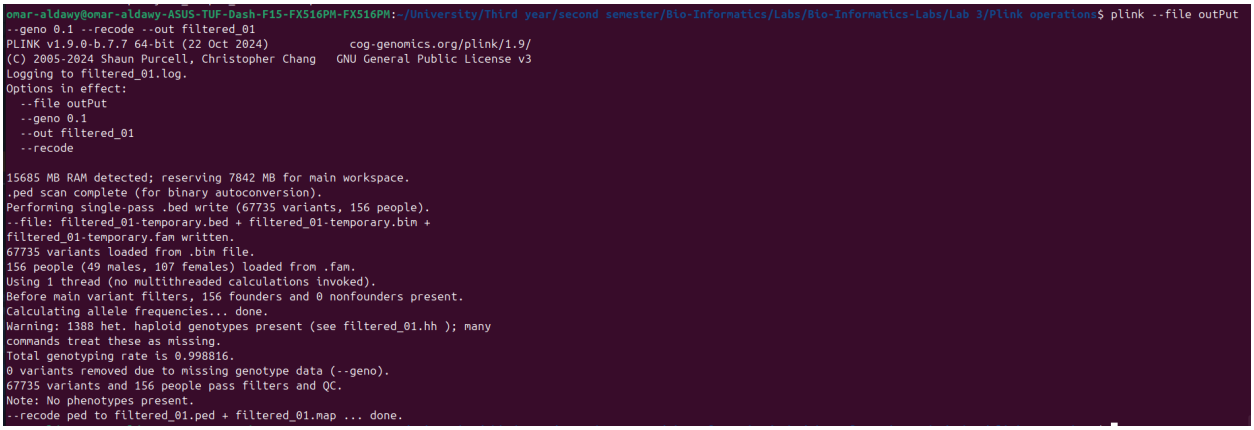

Figure 5: threshold 0.1

Threshold = 1e-2

- Number of variants removed = 0

Threshold = 1e-3

- Number of variants removed = 12509

Threshold = 1e-4

- Number of variants removed = 12509

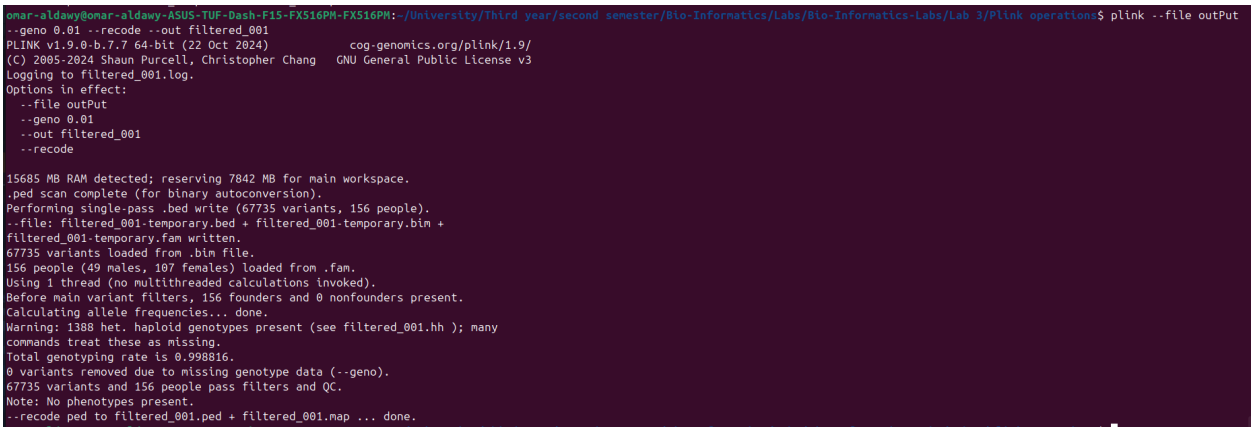

Figure 6: threshold 0.01

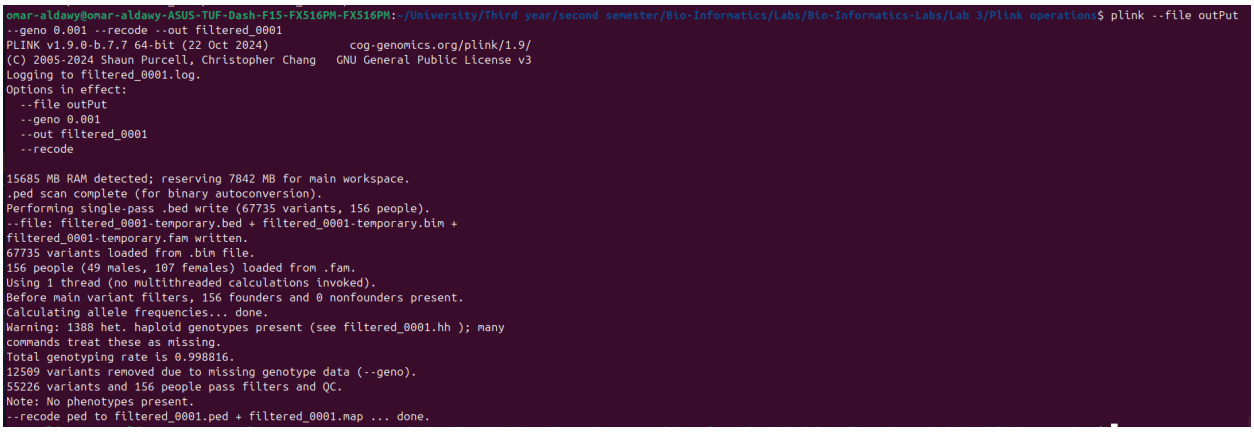

Figure 7: threshold 0.001

Figure 8: threshold 0.0001

Threshold = 1e-5

- Number of variants removed = 12509

Figure 9: threshold 0.00001

While missing call rate filtering significantly improves dataset reliability, it also reduces the total number of SNPs available for analysis. Choosing an appropriate threshold is crucial to balance quality control and retaining sufficient genetic markers for robust downstream analysis.

Part 2: Quality Control using PLINK

Running Minor Allele Frequency count on dataset to create .frq file

Minor Allele Frequency (MAF) is a key metric in genetics that measures how common the less frequent allele (minor allele) is in a given population.

```
plink --file filtered_000001 --freq --out maf_output
```

sample of .frq file

```
head maf_output.frq
```

- CHR:Chromosome number where the SNP is located
- SNP:SNP ID (rs number)
- A1:Minor allele (less frequent)
- A2:Major allele (more frequent)
- MAF:Minor Allele Frequency
- NCHROBS:Number of observed chromosomes (2 × sample size for autosomes)

Running QC on dataset using PLINK

Minor Allele Frequency (MAF) filtering

Minor Allele Frequency (MAF) filtering helps remove rare variants that might introduce noise into genetic studies. By setting a threshold, you keep only common variants, ensuring robust statistical power.

```
omar-aldawy@omar-aldawy-ASUS-TUF-Dash-F15-FX516PM-FX516PM:~/University/Third year/second semester/Bio-Informatics/Labs/Bio-Informatics-Labs/Lab 3/Plink operations$ plink --file filtere
d_000001 --freq --out maf_output
PLINK v1.9.0-b.7.7 64-bit (22 Oct 2024)           cog-genomics.org/plink/1.9/
(C) 2005-2024 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to maf_output.log.
Options in effect:
  --file filtered_000001
  --freq
  --out maf_output

15685 MB RAM detected; reserving 7842 MB for main workspace.
.ped scan complete (for binary autoconversion).
Performing single-pass .bed write (55226 variants, 156 people).
--file: maf_output-temporary.bed + maf_output-temporary.bim +
maf_output-temporary.fam written.
55226 variants loaded from .bim file.
156 people (49 males, 107 females) loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 156 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Warning: 1118 het. haploid genotypes present (see maf_output.hh ); many
commands treat these as missing.
Total genotyping rate is exactly 1.
--freq: Allele frequencies (founders only) written to maf_output.frq .
omar-aldawy@omar-aldawy-ASUS-TUF-Dash-F15-FX516PM-FX516PM:~/University/Third year/second semester/Bio-Informatics/Labs/Bio-Informatics-Labs/Lab 3/Plink operations$ head maf_output.frq
 CHR          SNP   A1   A2          MAF  NCHROBS
   1   rs10907175    C    A      0.08974      312
   1   rs10907187    A    G       0.2596      312
   1    rs6603803    G    A       0.3141      312
   1    rs6688000    A    G       0.1346      312
   1    rs7513222    A    G       0.3013      312
   1    rs3128309    A    G      0.05449      312
   1   rs12084736    T    C       0.1763      312
   1   rs12045693    A    C       0.2564      312
   1    rs2843130    G    C       0.0609      312
```

Figure 10: maf-output

```
plink --file your_output_filename --maf threshold --recode --out filtered_maf_01
```

```
omar-aldawy@omar-aldawy-ASUS-TUF-Dash-F15-FX516PM-FX516PM:~/University/Third year/second semester/Bio-Informatics/Labs/Bio-Informatics-Labs/Lab 3/Plink operations$ plink --file outPut
--maf 0.05 --recode --out maf_005
PLINK v1.9.0-b.7.7 64-bit (22 Oct 2024)           cog-genomics.org/plink/1.9/
(C) 2005-2024 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to maf_005.log.
Options in effect:
  --file outPut
  --maf 0.05
  --out maf_005
  --recode

15685 MB RAM detected; reserving 7842 MB for main workspace.
.ped scan complete (for binary autoconversion).
Performing single-pass .bed write (67735 variants, 156 people).
--file: maf_005-temporary.bed + maf_005-temporary.bim + maf_005-temporary.fam
written.
67735 variants loaded from .bim file.
156 people (49 males, 107 females) loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 156 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Warning: 1388 het. haploid genotypes present (see maf_005.hh ); many commands
treat these as missing.
Total genotyping rate is 0.998816.
0 variants removed due to minor allele threshold(s)
(--maf/--max-maf/--mac/--max-mac).
67735 variants and 156 people pass filters and QC.
Note: No phenotypes present.
--recode ped to maf_005.ped + maf_005.map ... done.
```

Figure 11: maf-filter-0.05

MAF threshold = 0.05

- Number of variants removed = 0

MAF threshold = 0.01

- Number of variants removed = 0

MAF threshold = 0.3

- Number of variants removed = 48833

```
omar-aldawy@omar-aldawy-ASUS-TUF-Dash-F15-FX516PM-FX516PM:~/University/Third year/second semester/Bio-Informatics/Labs/Bio-Informatics-Labs/Lab 3/Plink operations$ plink --file outPut
--maf 0.01 --recode --out maf_001
PLINK v1.9.0-b.7.7 64-bit (22 Oct 2024)           cog-genomics.org/plink/1.9/
(C) 2005-2024 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to maf_001.log.
Options in effect:
  --file outPut
  --maf 0.01
  --out maf_001
  --recode

15685 MB RAM detected; reserving 7842 MB for main workspace.
.ped scan complete (for binary autoconversion).
Performing single-pass .bed write (67735 variants, 156 people).
--file: maf_001-temporary.bed + maf_001-temporary.bim + maf_001-temporary.fam
written.
67735 variants loaded from .bim file.
156 people (49 males, 107 females) loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 156 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Warning: 1388 het. haploid genotypes present (see maf_001.hh ); many commands
treat these as missing.
Total genotyping rate is 0.998816.
0 variants removed due to minor allele threshold(s)
(--maf/--max-maf/--mac/--max-mac).
67735 variants and 156 people pass filters and QC.
Note: No phenotypes present.
--recode ped to maf_001.ped + maf_001.map ... done.
```

Figure 12: maf-filter-0.01

```
omar-aldawy@omar-aldawy-ASUS-TUF-Dash-F15-FX516PM-FX516PM:~/University/Third year/second semester/Bio-Informatics/Labs/Bio-Informatics-Labs/Lab 3/Plink operations$ plink --file outPut
--maf 0.3 --recode --out maf_03
PLINK v1.9.0-b.7.7 64-bit (22 Oct 2024)           cog-genomics.org/plink/1.9/
(C) 2005-2024 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to maf_03.log.
Options in effect:
  --file outPut
  --maf 0.3
  --out maf_03
  --recode

15685 MB RAM detected; reserving 7842 MB for main workspace.
.ped scan complete (for binary autoconversion).
Performing single-pass .bed write (67735 variants, 156 people).
--file: maf_03-temporary.bed + maf_03-temporary.bim + maf_03-temporary.fam
written.
67735 variants loaded from .bim file.
156 people (49 males, 107 females) loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 156 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Warning: 1388 het. haploid genotypes present (see maf_03.hh ); many commands
treat these as missing.
Total genotyping rate is 0.998816.
48833 variants removed due to minor allele threshold(s)
(--maf/--max-maf/--mac/--max-mac).
18902 variants and 156 people pass filters and QC.
Note: No phenotypes present.
--recode ped to maf_03.ped + maf_03.map ... done.
```

Figure 13: maf-filter-0.3

Missing Genotype Filter

Genotype missingness filtering helps ensure high-quality genetic data by removing SNPs or individuals with too many missing genotypes.

```
plink --file your_output_filename --geno threshold --recode --out filtered_geno_10
```

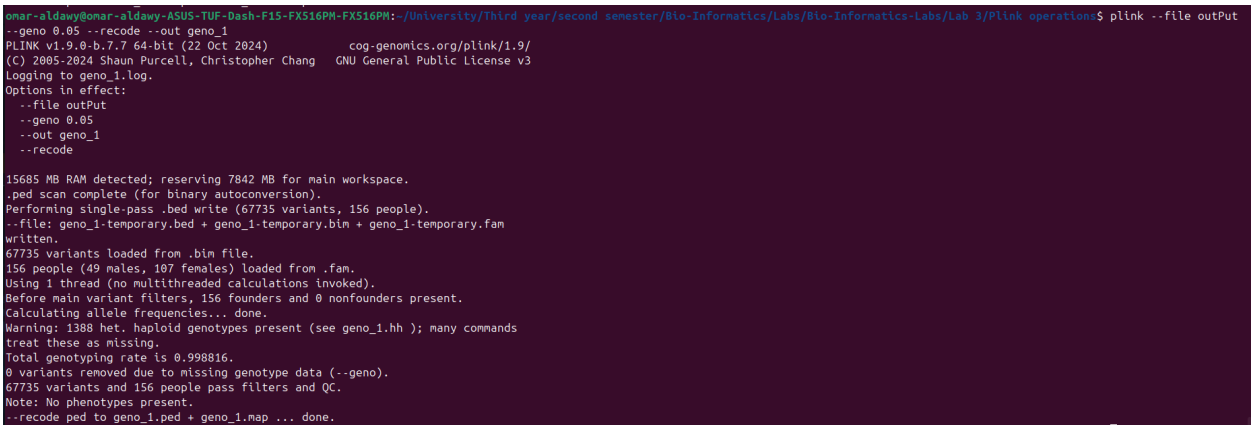

Figure 14: missing-genotype-0.05

Missing Genotype threshold = 0.05

- Number of variants removed = 0

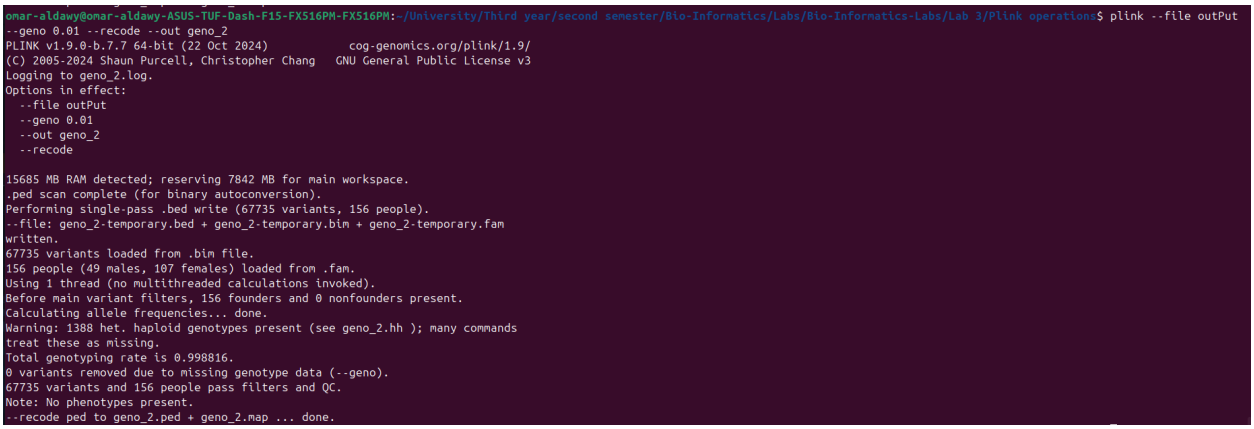

Figure 15: missing-genotype-0.01

Missing Genotype threshold = 0.01

- Number of variants removed = 0

Missing Genotype threshold = 0.0001

- Number of variants removed = 12509

```
omar-aldawy@omar-aldawy-ASUS-TUF-Dash-F15-FX516PM-FX516PM:~/University/Third year/second semester/Bio-Informatics/Labs/Bio-Informatics-Labs/Lab 3/Plink operations$ plink --file outPut
--geno 0.0001 --recode --out geno_3
PLINK v1.9.0-b.7.7 64-bit (22 Oct 2024)           cog-genomics.org/plink/1.9/
(C) 2005-2024 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to geno_3.log.
Options in effect:
  --file outPut
  --geno 0.0001
  --out geno_3
  --recode

15685 MB RAM detected; reserving 7842 MB for main workspace.
.ped scan complete (for binary autoconversion).
Performing single-pass .bed write (67735 variants, 156 people).
--file: geno_3-temporary.bed + geno_3-temporary.bim + geno_3-temporary.fam
written.
67735 variants loaded from .bim file.
156 people (49 males, 107 females) loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 156 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Warning: 1388 het. haploid genotypes present (see geno_3.hh ); many commands
treat these as missing.
Total genotyping rate is 0.998816.
12509 variants removed due to missing genotype data (--geno).
55226 variants and 156 people pass filters and QC.
Note: No phenotypes present.
--recode ped to geno_3.ped + geno_3.map ... done.
```

Figure 16: missing-genotype-0.0001

Hardy-Weinberg Equilibrium (HWE) Filter

Hardy-Weinberg Equilibrium (HWE) filtering removes SNPs that show significant deviation from expected allele frequencies, which can indicate genotyping errors or population structure issues.

```
plink --file your_output_filename --hwe threshold --recode --out filtered_hwe_3
```

```
omar-aldawy@omar-aldawy-ASUS-TUF-Dash-F15-FX516PM-FX516PM:~/University/Third year/second semester/Bio-Informatics/Labs/Bio-Informatics-Labs/Lab 3/Plink operations$ plink --file outPut
--hwe 1e-1 --recode --out hwe_5
PLINK v1.9.0-b.7.7 64-bit (22 Oct 2024)           cog-genomics.org/plink/1.9/
(C) 2005-2024 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to hwe_5.log.
Options in effect:
  --file outPut
  --hwe 1e-1
  --out hwe_5
  --recode

15685 MB RAM detected; reserving 7842 MB for main workspace.
.ped scan complete (for binary autoconversion).
Performing single-pass .bed write (67735 variants, 156 people).
--file: hwe_5-temporary.bed + hwe_5-temporary.bim + hwe_5-temporary.fam
written.
67735 variants loaded from .bim file.
156 people (49 males, 107 females) loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 156 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Warning: 1388 het. haploid genotypes present (see hwe_5.hh ); many commands
treat these as missing.
Total genotyping rate is 0.998816.
Warning: --hwe observation counts vary by more than 10%, due to the X
chromosome.  You may want to use a more stringent (i.e. less extreme) --hwe
p-value threshold for X chromosome variants: male samples are ignored there, so
the same degree of HWE violation corresponds to a less-extreme p-value than it
does elsewhere in the genome.
--hwe: 9083 variants removed due to Hardy-Weinberg exact test.
58652 variants and 156 people pass filters and QC.
Note: No phenotypes present.
--recode ped to hwe_5.ped + hwe_5.map ... done.
```

Figure 17: HWE-0.1

HWE threshold = 0.1

- Number of variants removed = 9083

HWE threshold = 0.01

- Number of variants removed = 1346

HWE threshold = 1e-5

```
omar-aldawy@omar-aldawy-ASUS-TUF-Dash-F15-FX516PM-FX516PM:~/University/Third year/second semester/Bio-Informatics/Labs/Bio-Informatics-Labs/Lab 3/Plink operations$ plink --file outPut
--hwe 1e-2 --recode --out hwe_3
PLINK v1.9.0-b.7.7 64-bit (22 Oct 2024)           cog-genomics.org/plink/1.9/
(C) 2005-2024 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to hwe_3.log.
Options in effect:
  --file outPut
  --hwe 1e-2
  --out hwe_3
  --recode

15685 MB RAM detected; reserving 7842 MB for main workspace.
.ped scan complete (for binary autoconversion).
Performing single-pass .bed write (67735 variants, 156 people).
--file: hwe_3-temporary.bed + hwe_3-temporary.bim + hwe_3-temporary.fam
written.
67735 variants loaded from .bim file.
156 people (49 males, 107 females) loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 156 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Warning: 1388 het. haploid genotypes present (see hwe_3.hh ); many commands
treat these as missing.
Total genotyping rate is 0.998816.
Warning: --hwe observation counts vary by more than 10%, due to the X
chromosome.  You may want to use a more stringent (i.e. less extreme) --hwe
p-value threshold for X chromosome variants: male samples are ignored there, so
the same degree of HWE violation corresponds to a less-extreme p-value than it
does elsewhere in the genome.
--hwe: 1346 variants removed due to Hardy-Weinberg exact test.
66389 variants and 156 people pass filters and QC.
Note: No phenotypes present.
--recode ped to hwe_3.ped + hwe_3.map ... done.
```

Figure 18: HWE-0.01

```
omar-aldawy@omar-aldawy-ASUS-TUF-Dash-F15-FX516PM-FX516PM:~/University/Third year/second semester/Bio-Informatics/Labs/Bio-Informatics-Labs/Lab 3/Plink operations$ plink --file outPut
--hwe 1e-5 --recode --out hwe_1
PLINK v1.9.0-b.7.7 64-bit (22 Oct 2024)           cog-genomics.org/plink/1.9/
(C) 2005-2024 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to hwe_1.log.
Options in effect:
  --file outPut
  --hwe 1e-5
  --out hwe_1
  --recode

15685 MB RAM detected; reserving 7842 MB for main workspace.
.ped scan complete (for binary autoconversion).
Performing single-pass .bed write (67735 variants, 156 people).
--file: hwe_1-temporary.bed + hwe_1-temporary.bim + hwe_1-temporary.fam
written.
67735 variants loaded from .bim file.
156 people (49 males, 107 females) loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 156 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Warning: 1388 het. haploid genotypes present (see hwe_1.hh ); many commands
treat these as missing.
Total genotyping rate is 0.998816.
Warning: --hwe observation counts vary by more than 10%, due to the X
chromosome.  You may want to use a more stringent (i.e. less extreme) --hwe
p-value threshold for X chromosome variants: male samples are ignored there, so
the same degree of HWE violation corresponds to a less-extreme p-value than it
does elsewhere in the genome.
--hwe: 0 variants removed due to Hardy-Weinberg exact test.
67735 variants and 156 people pass filters and QC.
Note: No phenotypes present.
--recode ped to hwe_1.ped + hwe_1.map ... done.
```

Figure 19: HWE-0.00005

- Number of variants removed = 0

Running the final version of QC using all the flags combined and reporting the final number of variants. Using the following thresholds (hwe: 0.01, maf: 0.1, geno: 0.001)

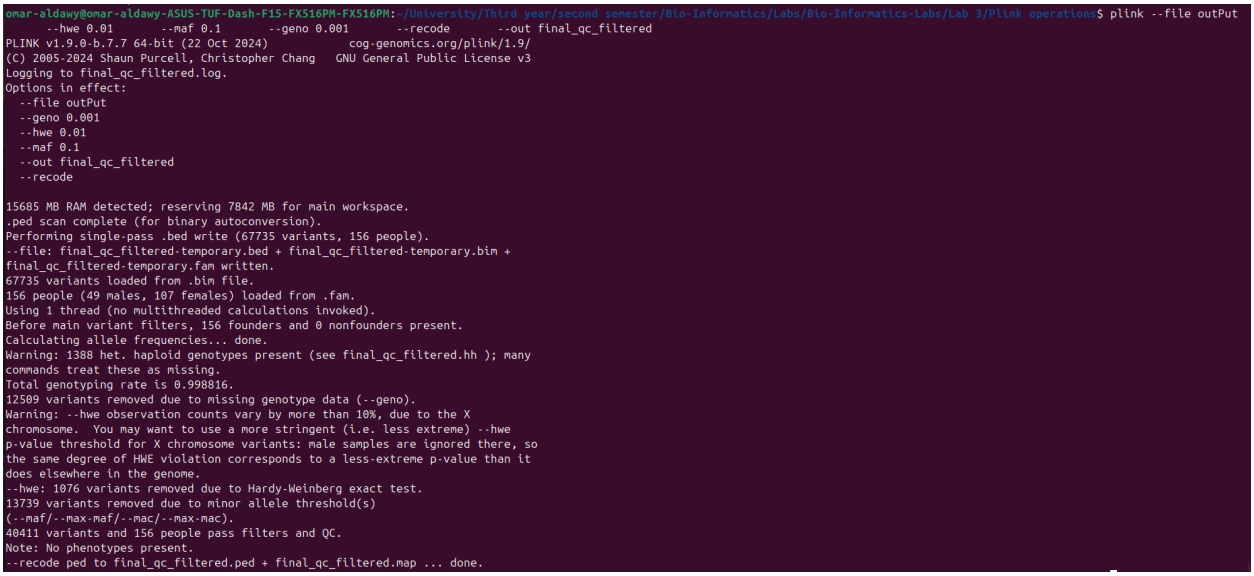

Figure 20: final-QC

- 12509 variants removed due to missing genotype data.
- 1076 variants removed due to Hardy-Weinberg exact test.
- 13739 variants removed due to minor allele threshold.
- 40411 variants and 156 people pass filters and QC.

Loading the final dataset in R

Install the required packages

```
install.packages("data.table")
```

```
## Installing package into '/home/omar-aldawy/R/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
install.packages("tidyverse")
```

```
## Installing package into '/home/omar-aldawy/R/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
library(data.table)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
```

```
## v purrr     1.0.4
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::between()     masks data.table::between()
## x dplyr::filter()      masks stats::filter()
## x dplyr::first()       masks data.table::first()
## x lubridate::hour()    masks data.table::hour()
## x lubridate::isoweek() masks data.table::isoweek()
## x dplyr::lag()         masks stats::lag()
## x dplyr::last()        masks data.table::last()
## x lubridate::mday()    masks data.table::mday()
## x lubridate::minute()  masks data.table::minute()
## x lubridate::month()   masks data.table::month()
## x lubridate::quarter() masks data.table::quarter()
## x lubridate::second()  masks data.table::second()
## x purrr::transpose()   masks data.table::transpose()
## x lubridate::wday()    masks data.table::wday()
## x lubridate::week()    masks data.table::week()
## x lubridate::yday()    masks data.table::yday()
## x lubridate::year()    masks data.table::year()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Use fread to load the .ped and .map files

```
filtered_ped <- fread("Plink operations/final_qc_filtered.ped", header = FALSE)

filtered_map <- fread("Plink operations/final_qc_filtered.map", header = FALSE)

original_ped <- fread("Plink operations/outPut.ped", header = FALSE)

original_map <- fread("Plink operations/outPut.map", header = FALSE)
```

Check the dimensions of the filtered and original datasets

```
filtered_samples <- nrow(filtered_ped)
filtered_snps <- (ncol(filtered_ped) - 6) / 2  # Each SNP has two columns (alleles)

original_samples <- nrow(original_ped)
original_snps <- (ncol(original_ped) - 6) / 2
```

Print the results of the filtering process and compare the original and filtered datasets

```
snps_removed <- original_snps - filtered_snps
samples_removed <- original_samples - filtered_samples

cat("Original SNP count:", original_snps, "\n")
```

```
## Original SNP count: 67735
```

```
cat("Filtered SNP count:", filtered_snps, "\n")
```

```
## Filtered SNP count: 40411
```

```
cat("SNPs removed:", snps_removed, "\n\n")
```

```
## SNPs removed: 27324
```

```r
cat("Original sample count:", original_samples, "\n")
```

```
## Original sample count: 156
```

```r
cat("Filtered sample count:", filtered_samples, "\n")
```

```
## Filtered sample count: 156
```

```r
cat("Samples removed:", samples_removed, "\n")
```

```
## Samples removed: 0
```