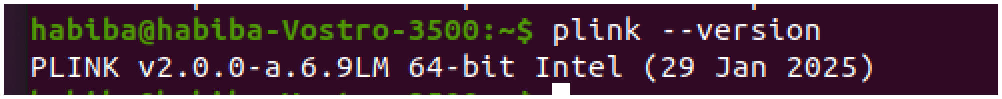
# Assignment 3: Plink

Habiba Tarek Ramadan Ali (ID: 21010445)

2025-03-09

# Part 1: Plink Walkthrough

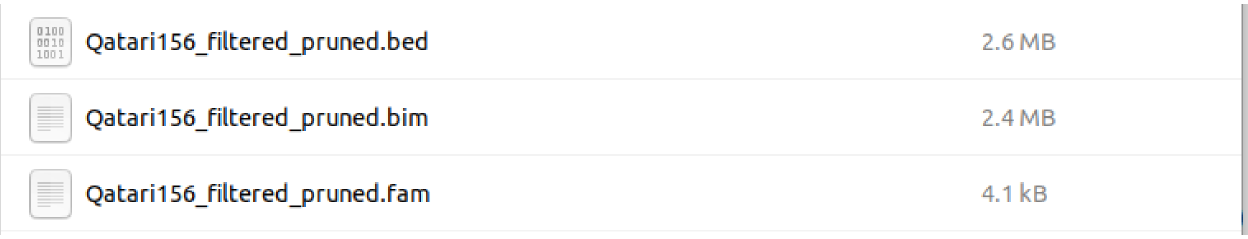
## Task 1.1: Installation



```
habiba@habiba-Vostro-3500:~$ plink --version
PLINK v2.0.0-a.6.9LM 64-bit Intel (29 Jan 2025)
```

## Task 1.2: Basic Commands

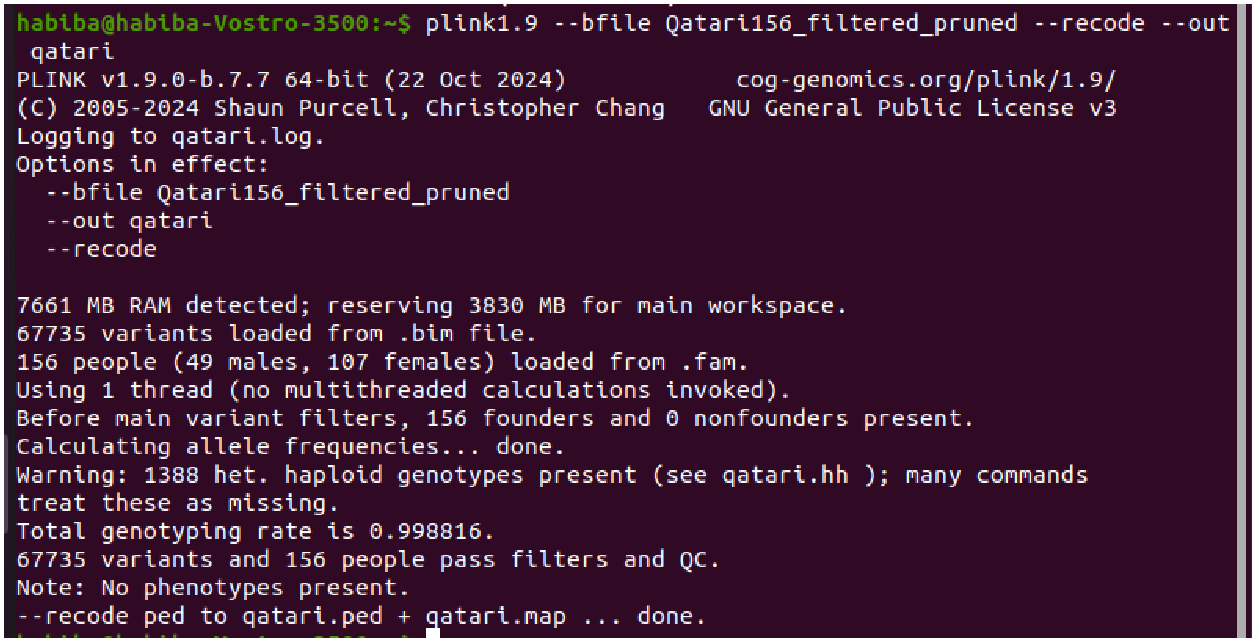
**Formats of files:** binary files (.bed , .bim and .fam)



| | |
|---|---|
| Qatari156_filtered_pruned.bed | 2.6 MB |
| Qatari156_filtered_pruned.bim | 2.4 MB |
| Qatari156_filtered_pruned.fam | 4.1 kB |

Number of Variants: 67735.

Number of Samples: 156.

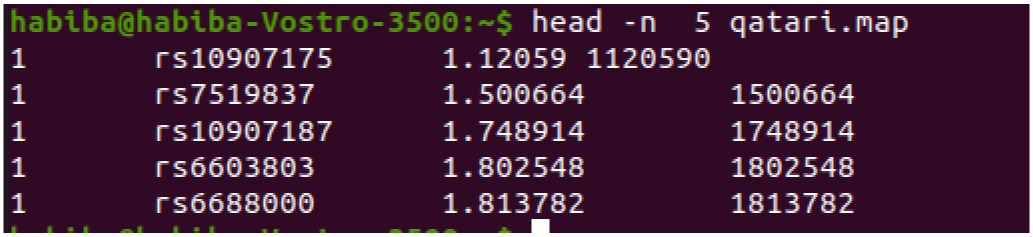


```
habiba@habiba-Vostro-3500:~$ plink1.9 --bfile Qatari156_filtered_pruned --recode --out
 qatari
PLINK v1.9.0-b.7.7 64-bit (22 Oct 2024)            cog-genomics.org/plink/1.9/
(C) 2005-2024 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to qatari.log.
Options in effect:
  --bfile Qatari156_filtered_pruned
  --out qatari
  --recode

7661 MB RAM detected; reserving 3830 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 qatari.hh ); many commands
treat these as missing.
Total genotyping rate is 0.998816.
67735 variants and 156 people pass filters and QC.
Note: No phenotypes present.
--recode ped to qatari.ped + qatari.map ... done.
```

### PLINK ‘.map‘ File Structure

Each row in a `.map` file represents a SNP and consists of four columns:

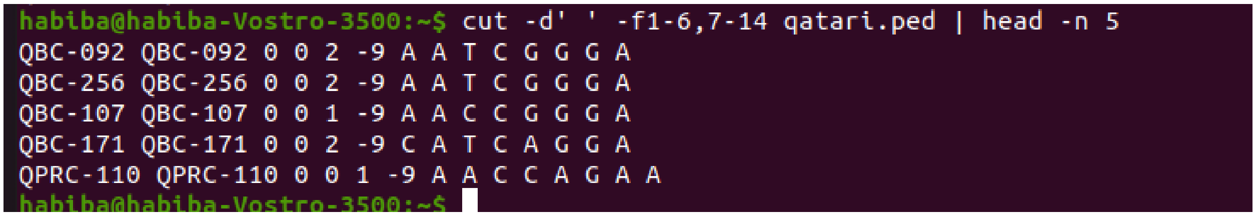
| Column | Description |
|---|---|
| 1 | **Chromosome Number** – The chromosome on which the SNP is located |
| 2 | **SNP ID** – Unique identifier for the SNP. |
| 3 | **Genetic Distance** – Position in centimorgans (cM). |
| 4 | **Base-pair Position** – Physical position on the chromosome. |



```
habiba@habiba-Vostro-3500:~$ head -n  5 qatari.map
1       rs10907175      1.12059 1120590
1       rs7519837       1.500664        1500664
1       rs10907187      1.748914        1748914
1       rs6603803       1.802548        1802548
1       rs6688000       1.813782        1813782
habiba@habiba-Vostro-3500:~$
```

### PLINK ‘.ped‘ File Structure

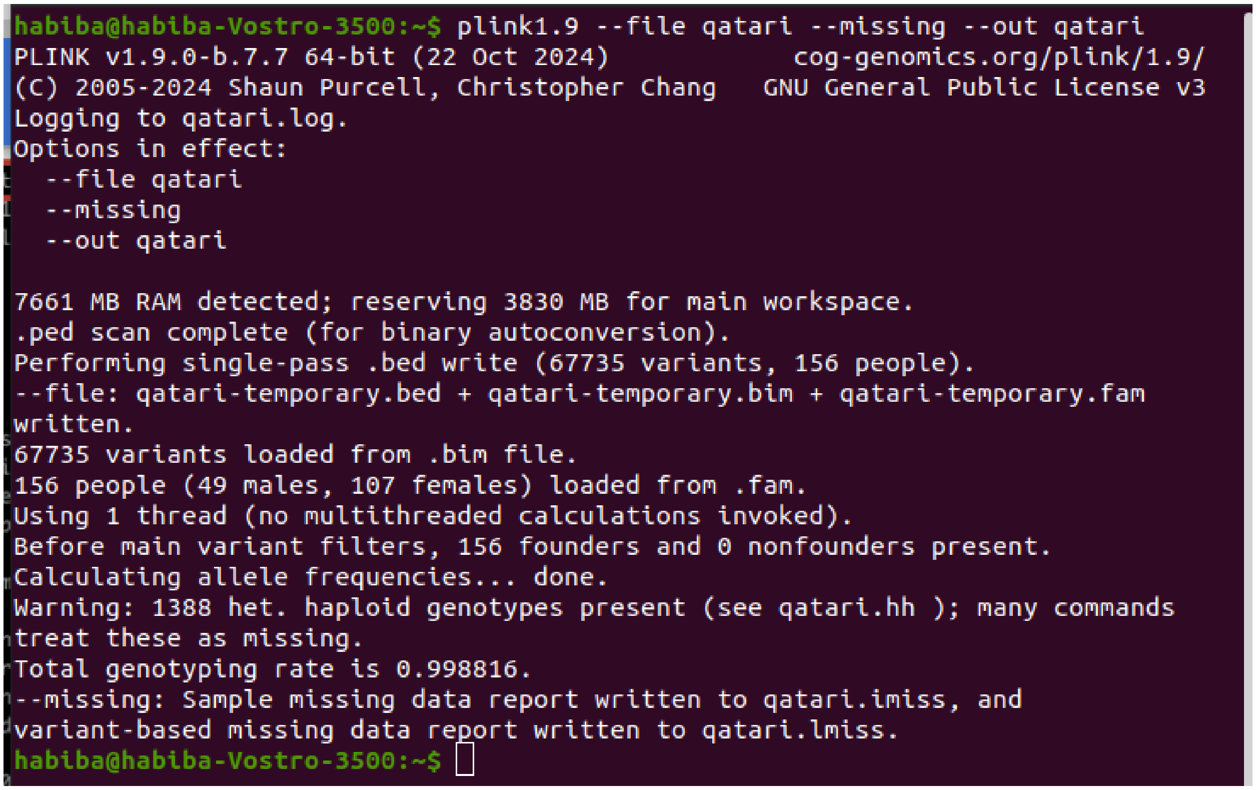
| Column | Description |
|---|---|
| 1 | **FID** – Family ID. |
| 2 | **IID** – Individual ID. |
| 3 | **F** – Father ID. |
| 4 | **M** – Mother ID. |
| 5 | **SEX** – Sex. |
| 6 | **PHENOTYPE** – 1 , 2 or -9 for unknown. |
| 7+ | **SNPs** – SNP genotype data (two columns per SNP). |



```
habiba@habiba-Vostro-3500:~$ cut -d' ' -f1-6,7-14 qatari.ped | head -n 5
QBC-092 QBC-092 0 0 2 -9 A A T C G G G A
QBC-256 QBC-256 0 0 2 -9 A A T C G G G A
QBC-107 QBC-107 0 0 1 -9 A A C C G G G A
QBC-171 QBC-171 0 0 2 -9 C A T C A G G A
QPRC-110 QPRC-110 0 0 1 -9 A A C C A G A A
habiba@habiba-Vostro-3500:~$
```

**R**un "**Missing out**" command to find missing Rate per individual "**imiss file**" and per snp "**lmiss file**"

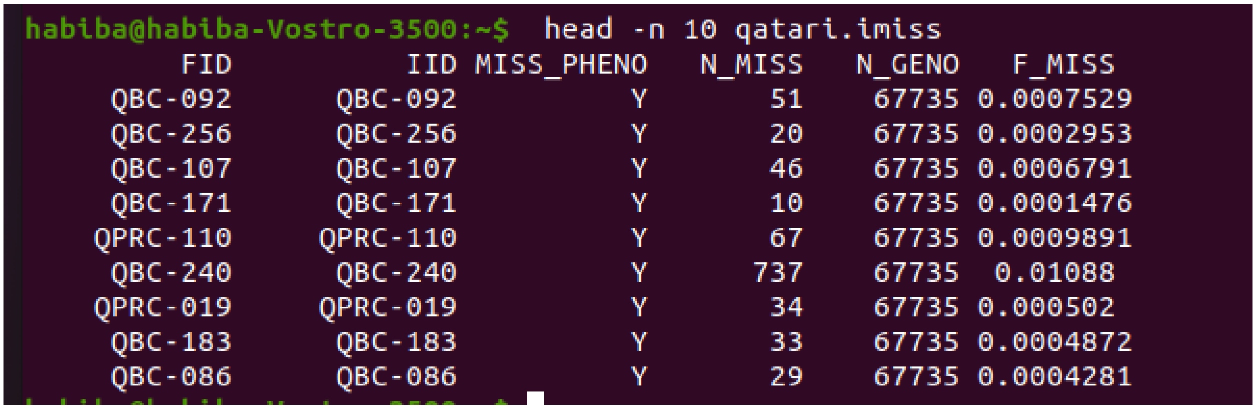
| **Statistic** | **Value** |
|---|---|
| Total Variants | 67,735 |
| Total Individuals | 156 |
| Males | 49 |
| Females | 107 |
| Total Genotyping Rate | 0.998816 $\left(\frac{156\times67735-1388}{156\times67735}\right)$ |
| Sample Missing Data Report | `qatari.imiss` |
| Variant-Based Missing Data Report | `qatari.lmiss` |



```
habiba@habiba-Vostro-3500:~$ plink1.9 --file qatari --missing --out qatari
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 qatari.log.
Options in effect:
  --file qatari
  --missing
  --out qatari

7661 MB RAM detected; reserving 3830 MB for main workspace.
.ped scan complete (for binary autoconversion).
Performing single-pass .bed write (67735 variants, 156 people).
--file: qatari-temporary.bed + qatari-temporary.bim + qatari-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 qatari.hh ); many commands
treat these as missing.
Total genotyping rate is 0.998816.
--missing: Sample missing data report written to qatari.imiss, and
variant-based missing data report written to qatari.lmiss.
habiba@habiba-Vostro-3500:~$ 
```

### PLINK ‘.imiss‘ File Structure

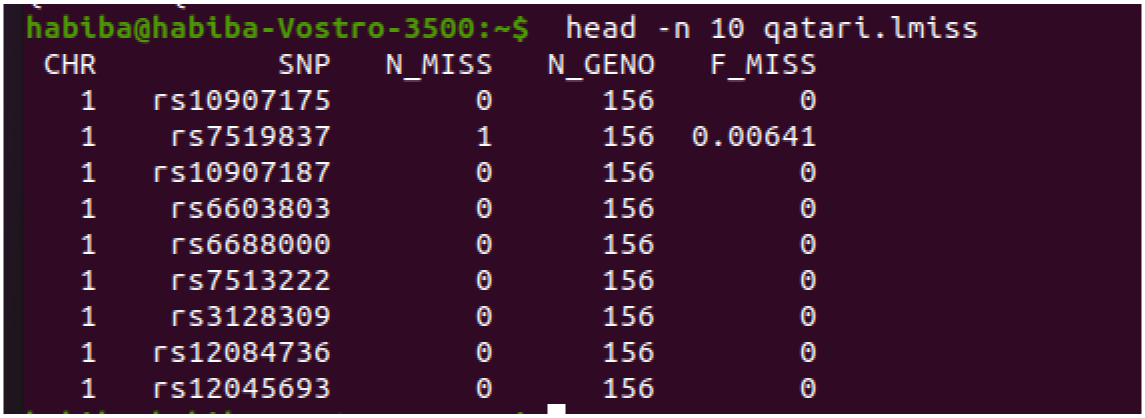
| Column | Description |
|---|---|
| 1 | **FID** – Family ID. |
| 2 | **IID** – Individual ID. |
| 3 | **MISS_PHENO** – Y == Missing Phenotype. |
| 4 | **N_MISS** – Number of missing genotype calls. |
| 5 | **N_GENO** – Total number of genotype calls. |
| 6 | **F_MISS** – Fraction of missing genotypes ($\frac{N\_MISS}{N\_GENO}$). |



```
habiba@habiba-Vostro-3500:~$  head -n 10 qatari.imiss
        FID          IID MISS_PHENO   N_MISS   N_GENO   F_MISS
    QBC-092      QBC-092          Y       51    67735 0.0007529
    QBC-256      QBC-256          Y       20    67735 0.0002953
    QBC-107      QBC-107          Y       46    67735 0.0006791
    QBC-171      QBC-171          Y       10    67735 0.0001476
   QPRC-110     QPRC-110          Y       67    67735 0.0009891
    QBC-240      QBC-240          Y      737    67735  0.01088
   QPRC-019     QPRC-019          Y       34    67735 0.000502
    QBC-183      QBC-183          Y       33    67735 0.0004872
    QBC-086      QBC-086          Y       29    67735 0.0004281
```

### PLINK ‘.lmiss‘ File Structure

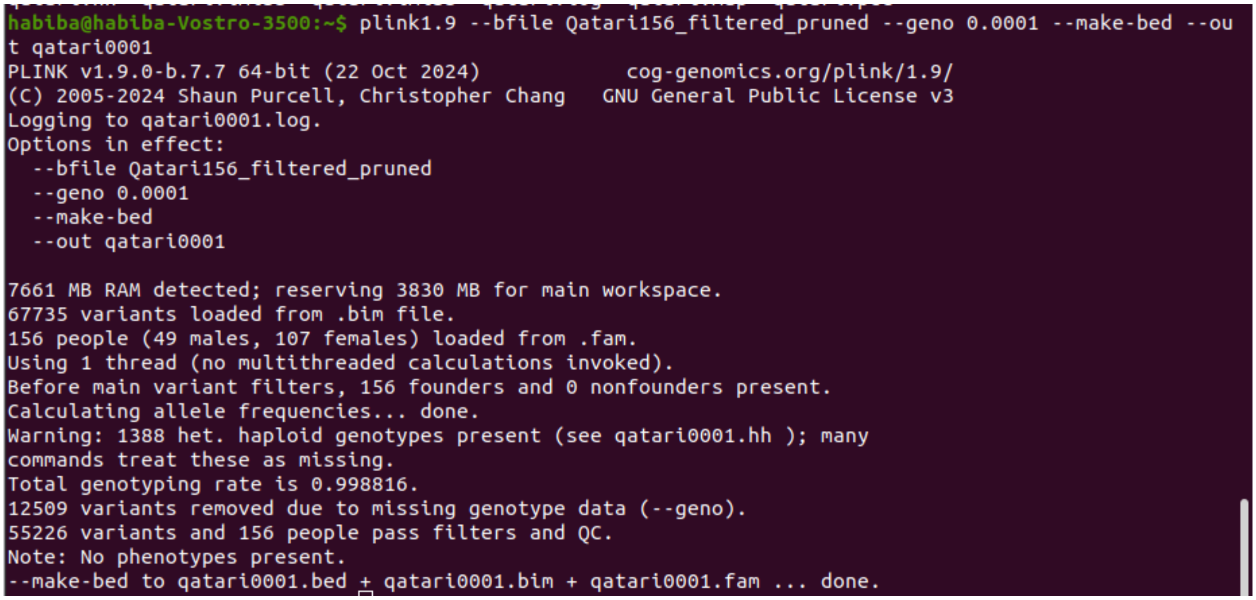
| Column | Description |
|---|---|
| 1 | **CHR** – Chromosome number. |
| 2 | **SNP** – SNP ID. |
| 3 | **N_MISS** – Number of missing genotype calls. |
| 4 | **N_GENO** – Total number of genotype calls. |
| 5 | **F_MISS** – Fraction of missing genotypes ($\frac{N\_MISS}{N\_GENO}$). |



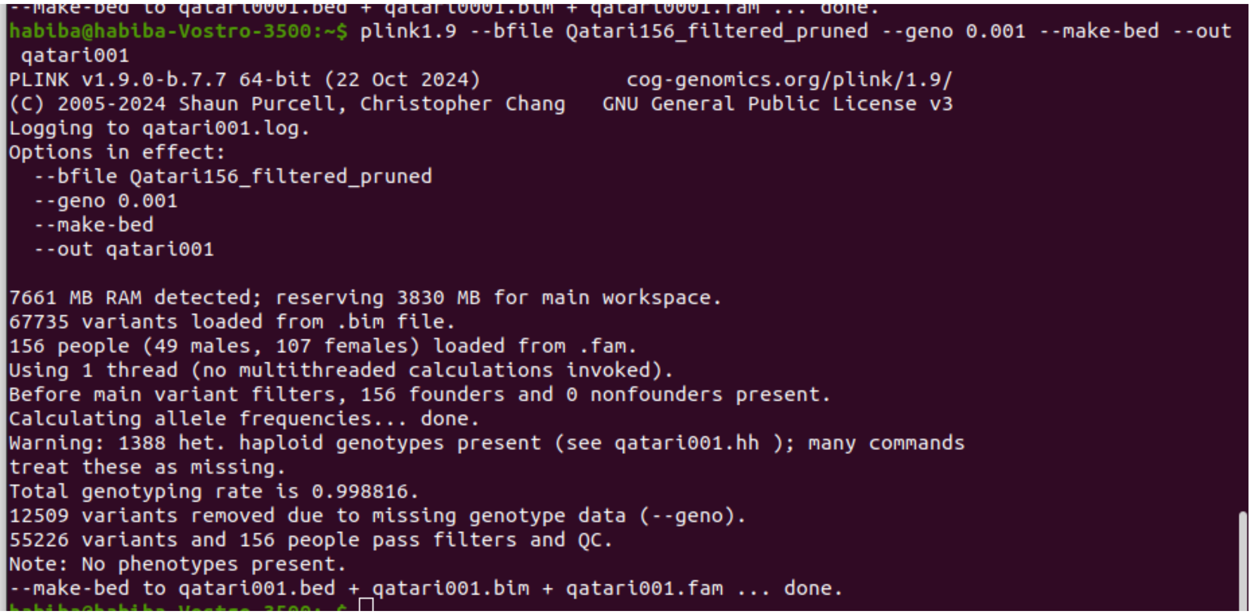
```
habiba@habiba-Vostro-3500:~$  head -n 10 qatari.lmiss
 CHR          SNP   N_MISS   N_GENO   F_MISS
   1   rs10907175        0      156        0
   1    rs7519837        1      156  0.00641
   1   rs10907187        0      156        0
   1    rs6603803        0      156        0
   1    rs6688000        0      156        0
   1    rs7513222        0      156        0
   1    rs3128309        0      156        0
   1   rs12084736        0      156        0
   1   rs12045693        0      156        0
```

**Conduct the Missing Call Rate Analysis** 5 different thresholds Working with binary dataset files (.bed, .bim , .fam) Working on SNPs missing rates Using “plink –bfile dataset –geno threshold –make.bed –out outputFile” command

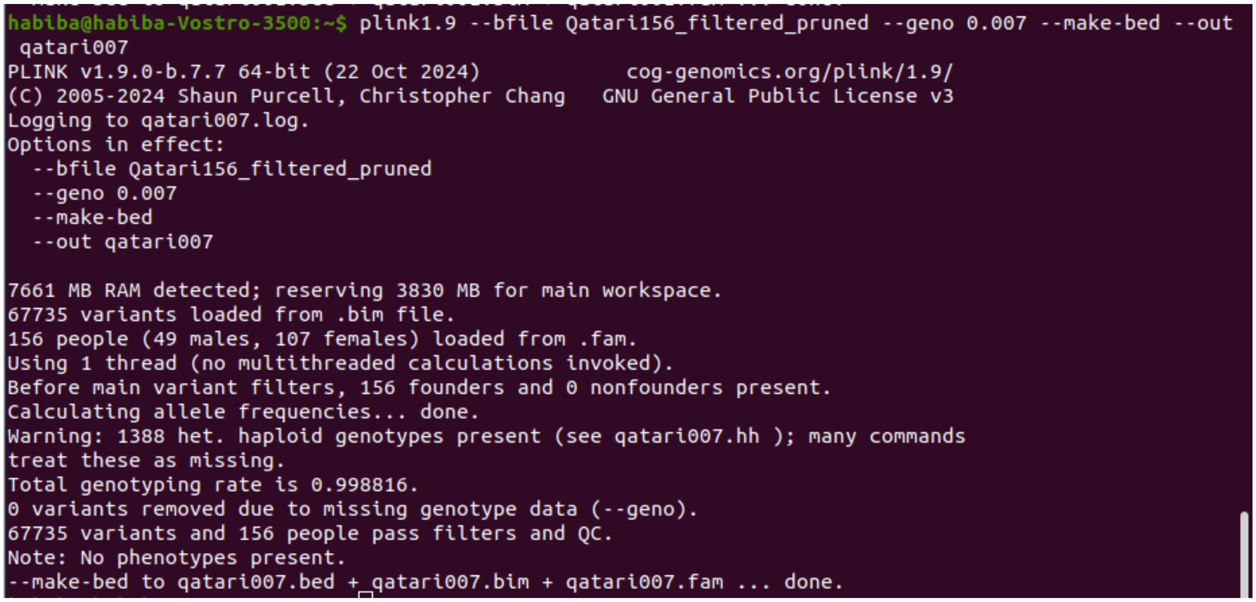
| part | description |
|---|---|
| –bfile | to use a dataset in binary format |
| –geno | filters SNPs based on missing genotype rates |
| –make-bed | This outputs the filtered data in binary PLINK format".bed , .bim and .fa |
| –out | prefix for the output files |



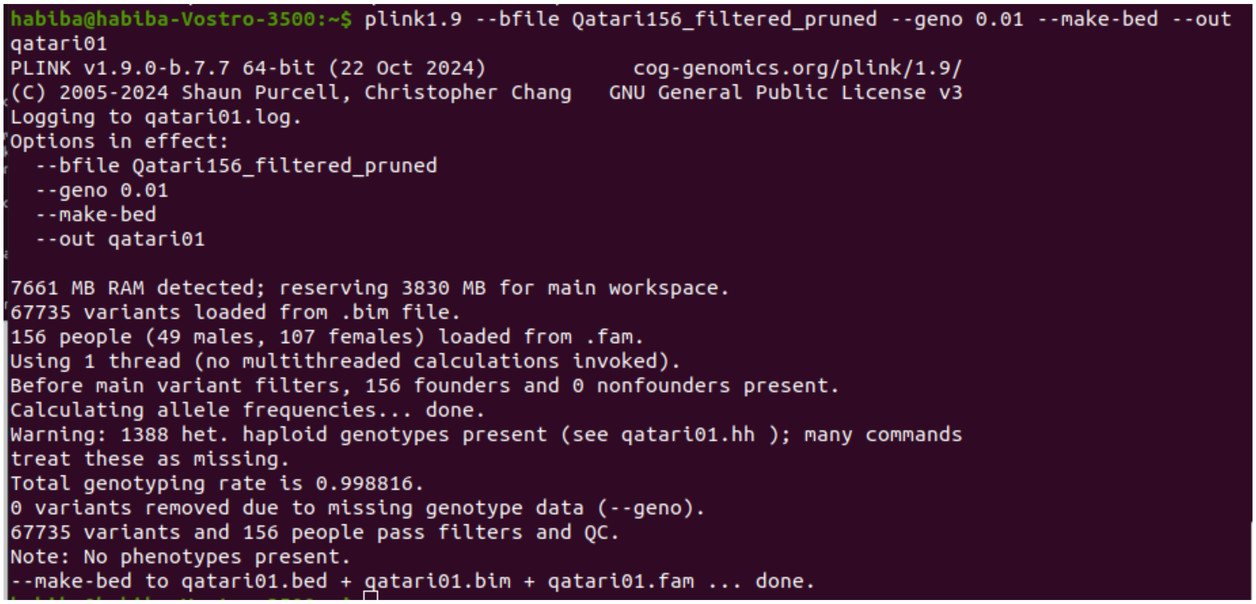
```
habiba@habiba-Vostro-3500:~$ plink1.9 --bfile Qatari156_filtered_pruned --geno 0.0001 --make-bed --ou
t qatari0001
PLINK v1.9.0-b.7.7 64-bit (22 Oct 2024)           cog-genomics.org/plink/1.9/
(C) 2005-2024 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to qatari0001.log.
Options in effect:
  --bfile Qatari156_filtered_pruned
  --geno 0.0001
  --make-bed
  --out qatari0001

7661 MB RAM detected; reserving 3830 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 qatari0001.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.
--make-bed to qatari0001.bed + qatari0001.bim + qatari0001.fam ... done.
```



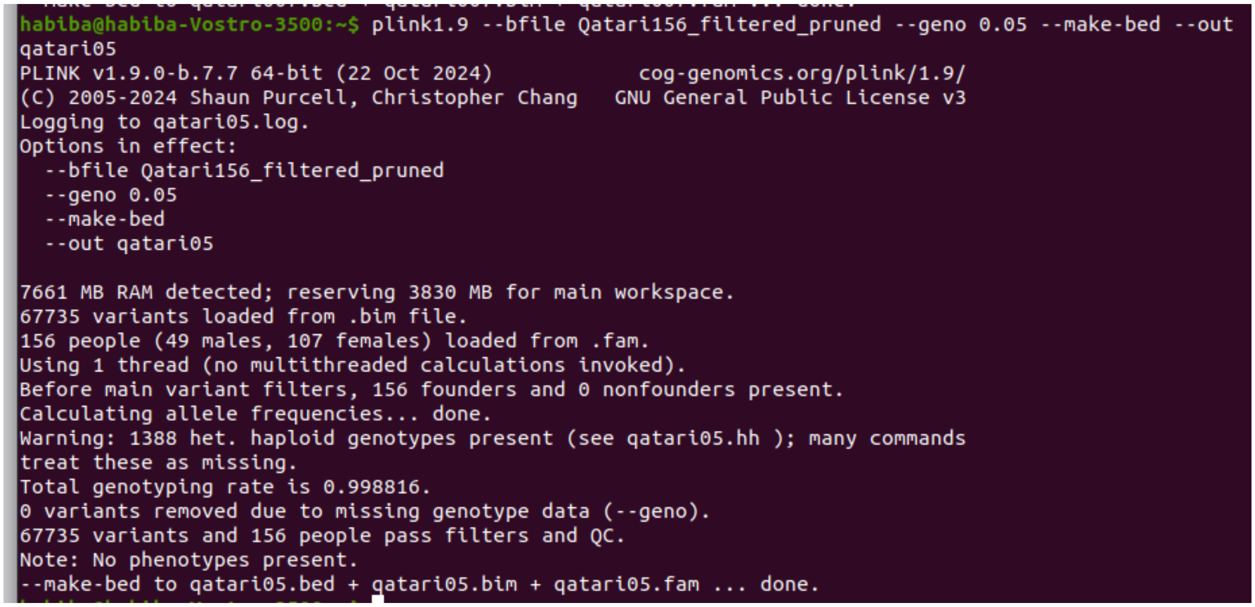
```
--make-bed to qatari0001.bed + qatari0001.bim + qatari0001.fam ... done.
habiba@habiba-Vostro-3500:~$ plink1.9 --bfile Qatari156_filtered_pruned --geno 0.001 --make-bed --out
 qatari001
PLINK v1.9.0-b.7.7 64-bit (22 Oct 2024)           cog-genomics.org/plink/1.9/
(C) 2005-2024 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to qatari001.log.
Options in effect:
  --bfile Qatari156_filtered_pruned
  --geno 0.001
  --make-bed
  --out qatari001

7661 MB RAM detected; reserving 3830 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 qatari001.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.
--make-bed to qatari001.bed + qatari001.bim + qatari001.fam ... done.
habiba@habiba-Vostro-3500:~$
```



```
habiba@habiba-Vostro-3500:~$ plink1.9 --bfile Qatari156_filtered_pruned --geno 0.007 --make-bed --out
 qatari007
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 qatari007.log.
Options in effect:
  --bfile Qatari156_filtered_pruned
  --geno 0.007
  --make-bed
  --out qatari007

7661 MB RAM detected; reserving 3830 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 qatari007.hh ); many commands
treat these as missing.
Total genotyping rate is 0.998816.
0 variants removed due to missing genotype data (--geno).
67735 variants and 156 people pass filters and QC.
Note: No phenotypes present.
--make-bed to qatari007.bed + qatari007.bim + qatari007.fam ... done.
```



```
habiba@habiba-Vostro-3500:~$ plink1.9 --bfile Qatari156_filtered_pruned --geno 0.01 --make-bed --out
qatari01
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 qatari01.log.
Options in effect:
  --bfile Qatari156_filtered_pruned
  --geno 0.01
  --make-bed
  --out qatari01

7661 MB RAM detected; reserving 3830 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 qatari01.hh ); many commands
treat these as missing.
Total genotyping rate is 0.998816.
0 variants removed due to missing genotype data (--geno).
67735 variants and 156 people pass filters and QC.
Note: No phenotypes present.
--make-bed to qatari01.bed + qatari01.bim + qatari01.fam ... done.
```



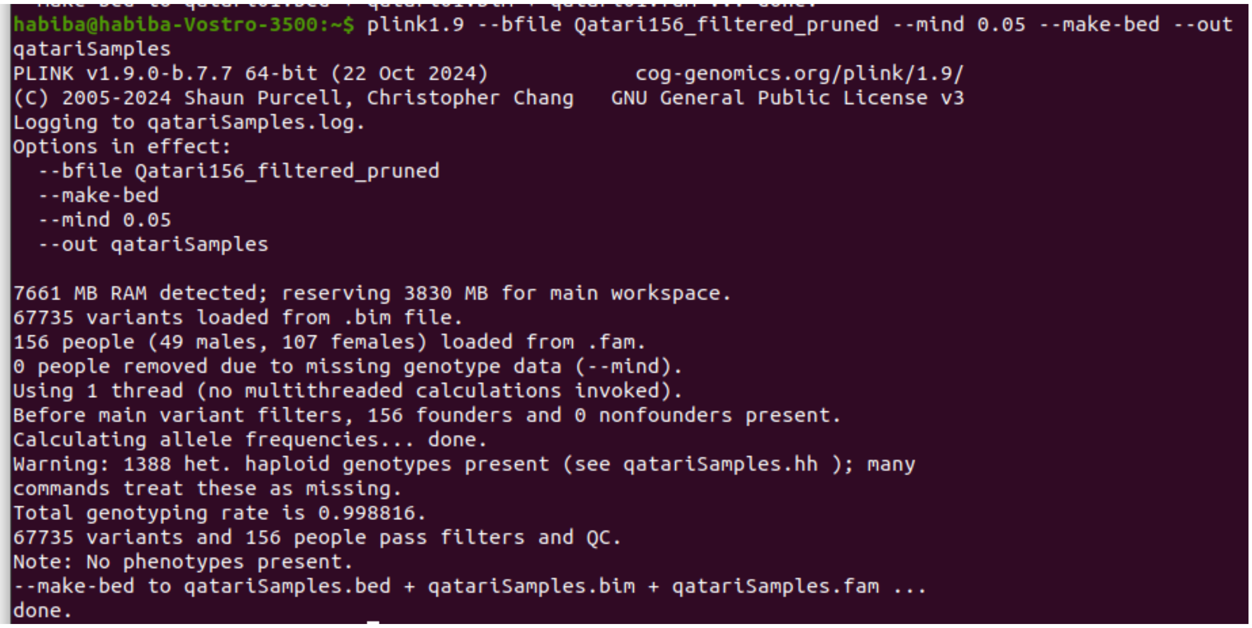
```
habiba@habiba-Vostro-3500:~$ plink1.9 --bfile Qatari156_filtered_pruned --geno 0.05 --make-bed --out
qatari05
PLINK v1.9.0-b.7.7 64-bit (22 Oct 2024)            cog-genomics.org/plink/1.9/
(C) 2005-2024 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to qatari05.log.
Options in effect:
  --bfile Qatari156_filtered_pruned
  --geno 0.05
  --make-bed
  --out qatari05

7661 MB RAM detected; reserving 3830 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 qatari05.hh ); many commands
treat these as missing.
Total genotyping rate is 0.998816.
0 variants removed due to missing genotype data (--geno).
67735 variants and 156 people pass filters and QC.
Note: No phenotypes present.
--make-bed to qatari05.bed + qatari05.bim + qatari05.fam ... done.
```

| Parameter | Description |
|---|---|
| **Threshold** | 0.0001 - 001 |
| **Total SNPs Before Filtering** | 67735 |
| **Filtered SNPs** | 12509 |
| **Remaining SNPs** | 55226 |

| Parameter | Description |
|---|---|
| **Threshold** | 0.007 - 0.01 - 0.05 - + |
| **Total SNPs Before Filtering** | 67735 |
| **Filtered SNPs** | 0 |
| **Remaining SNPs** | 67735 |

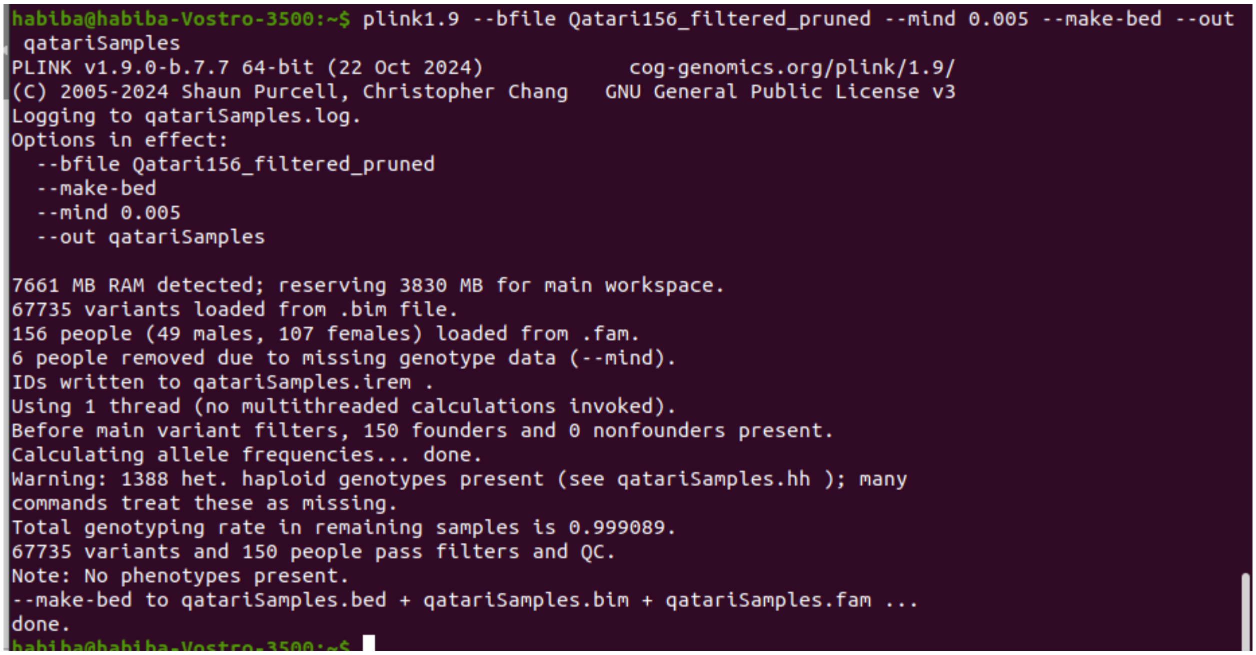
Working on samples missing rates Using “plink –bfile dataset –mind threshold –make.bed –out outputFile” command

| part | description |
|---|---|
| –bfile | to use a dataset in binary format |
| –mind | filters Samples based on missing genotype rates |
| –make-bed | This outputs the filtered data in binary PLINK format".bed , .bim and .far |
| –out | prefix for the output files |



```
habiba@habiba-Vostro-3500:~$ plink1.9 --bfile Qatari156_filtered_pruned --mind 0.05 --make-bed --out
qatariSamples
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 qatariSamples.log.
Options in effect:
  --bfile Qatari156_filtered_pruned
  --make-bed
  --mind 0.05
  --out qatariSamples

7661 MB RAM detected; reserving 3830 MB for main workspace.
67735 variants loaded from .bim file.
156 people (49 males, 107 females) loaded from .fam.
0 people removed due to missing genotype data (--mind).
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 qatariSamples.hh ); many
commands treat these as missing.
Total genotyping rate is 0.998816.
67735 variants and 156 people pass filters and QC.
Note: No phenotypes present.
--make-bed to qatariSamples.bed + qatariSamples.bim + qatariSamples.fam ...
done.
```

| Parameter | Description |
|---|---|
| **Threshold** | 0.05 |
| **Total samples Before Filtering** | 156 |
| **Filtered samples** | 0 |
| **Remaining samples** | 156 |



```
habiba@habiba-Vostro-3500:~$ plink1.9 --bfile Qatari156_filtered_pruned --mind 0.005 --make-bed --out
 qatariSamples
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 qatariSamples.log.
Options in effect:
  --bfile Qatari156_filtered_pruned
  --make-bed
  --mind 0.005
  --out qatariSamples

7661 MB RAM detected; reserving 3830 MB for main workspace.
67735 variants loaded from .bim file.
156 people (49 males, 107 females) loaded from .fam.
6 people removed due to missing genotype data (--mind).
IDs written to qatariSamples.irem .
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 150 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Warning: 1388 het. haploid genotypes present (see qatariSamples.hh ); many
commands treat these as missing.
Total genotyping rate in remaining samples is 0.999089.
67735 variants and 150 people pass filters and QC.
Note: No phenotypes present.
--make-bed to qatariSamples.bed + qatariSamples.bim + qatariSamples.fam ...
done.
habiba@habiba-Vostro-3500:~$
```

| Parameter | Description |
|---|---|
| **Threshold** | 0.005 |
| **Total samples Before Filtering** | 156 |
| **Filtered samples** | 6 |
| **Remaining samples** | 150 |

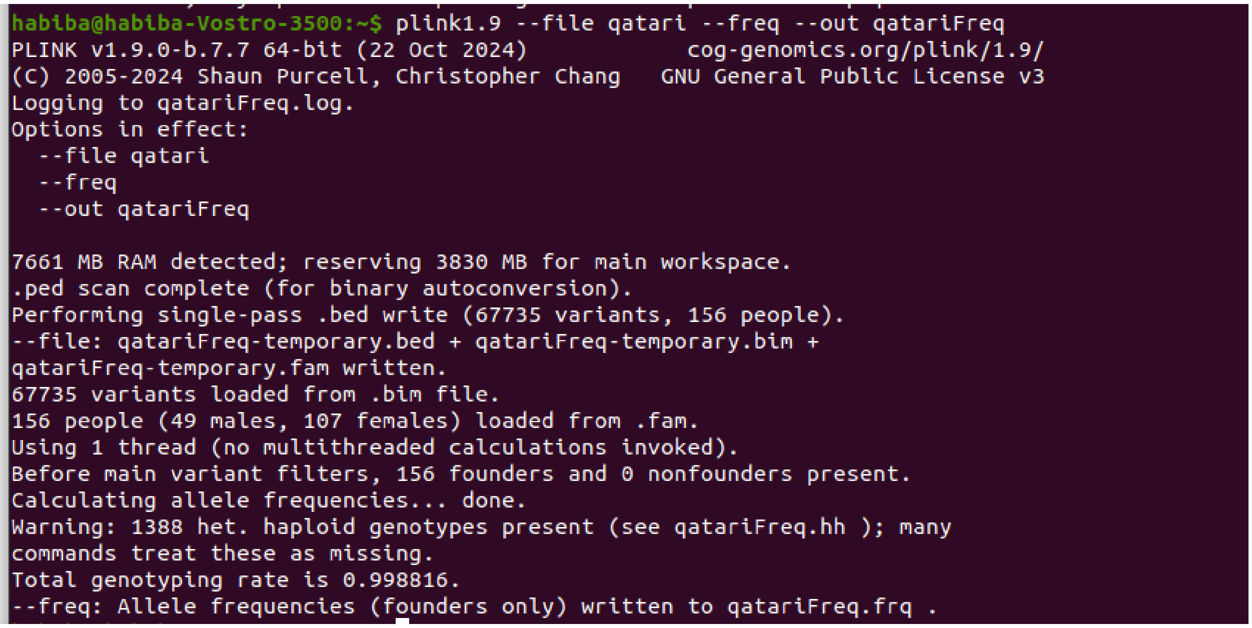
### How the data Quality control affected the dataset:

1- Reduced Number of SNPs:
Reduces noise by excluding unreliable or uninformative SNPs.

2- Reduced Number of Individuals: Improves sample-level quality, ensuring only high-fidelity samples remain.

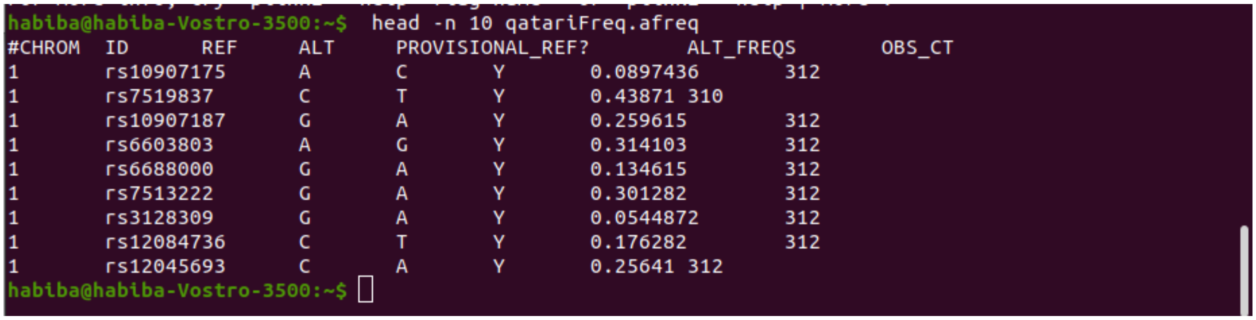
3- Improved Genotyping Rate: Higher genotyping rate reflects cleaner, more complete data.

# Part 2: Quality Control using PLINK



```
habiba@habiba-Vostro-3500:~$ plink1.9 --file qatari --freq --out qatariFreq
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 qatariFreq.log.
Options in effect:
  --file qatari
  --freq
  --out qatariFreq

7661 MB RAM detected; reserving 3830 MB for main workspace.
.ped scan complete (for binary autoconversion).
Performing single-pass .bed write (67735 variants, 156 people).
--file: qatariFreq-temporary.bed + qatariFreq-temporary.bim +
qatariFreq-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 qatariFreq.hh ); many
commands treat these as missing.
Total genotyping rate is 0.998816.
--freq: Allele frequencies (founders only) written to qatariFreq.frq .
```

## Minor Allele Frequency



```
habiba@habiba-Vostro-3500:~$  head -n 10 qatariFreq.afreq
#CHROM  ID      REF     ALT     PROVISIONAL_REF?        ALT_FREQS       OBS_CT
1       rs10907175      A       C       Y       0.0897436       312
1       rs7519837       C       T       Y       0.43871 310
1       rs10907187      G       A       Y       0.259615        312
1       rs6603803       A       G       Y       0.314103        312
1       rs6688000       G       A       Y       0.134615        312
1       rs7513222       G       A       Y       0.301282        312
1       rs3128309       G       A       Y       0.0544872       312
1       rs12084736      C       T       Y       0.176282        312
1       rs12045693      C       A       Y       0.25641 312
habiba@habiba-Vostro-3500:~$ 
```

| Column | Description |
|---|---|
| **ID** | SNP identifier |
| **REF** | Reference allele (major allele) |
| **ALT** | Alternative allele (minor allele) |
| **ALT_FREQS** | Frequency of the minor allele |
| **OBS_CT** | Number of observed alleles |

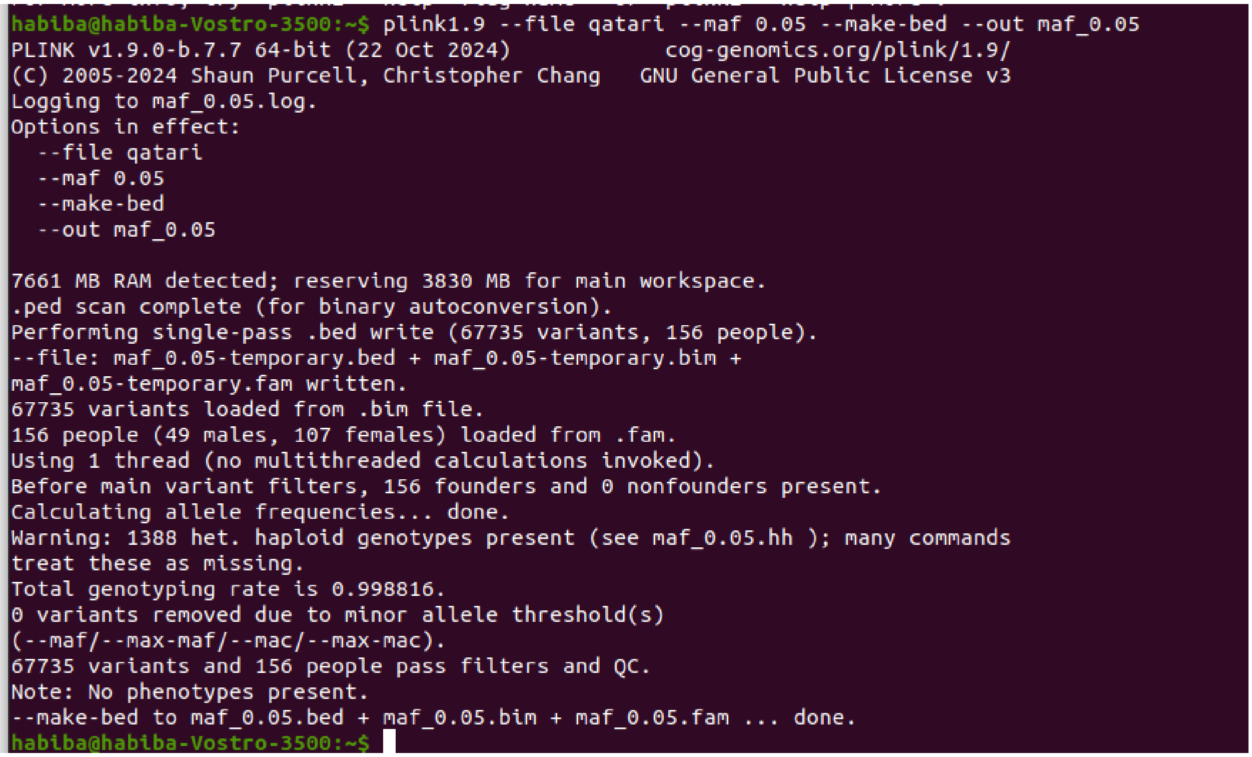
## QC using different tests with different threshold:

**1. MAF:** removes variants with a minor allele frequency below the specified threshold

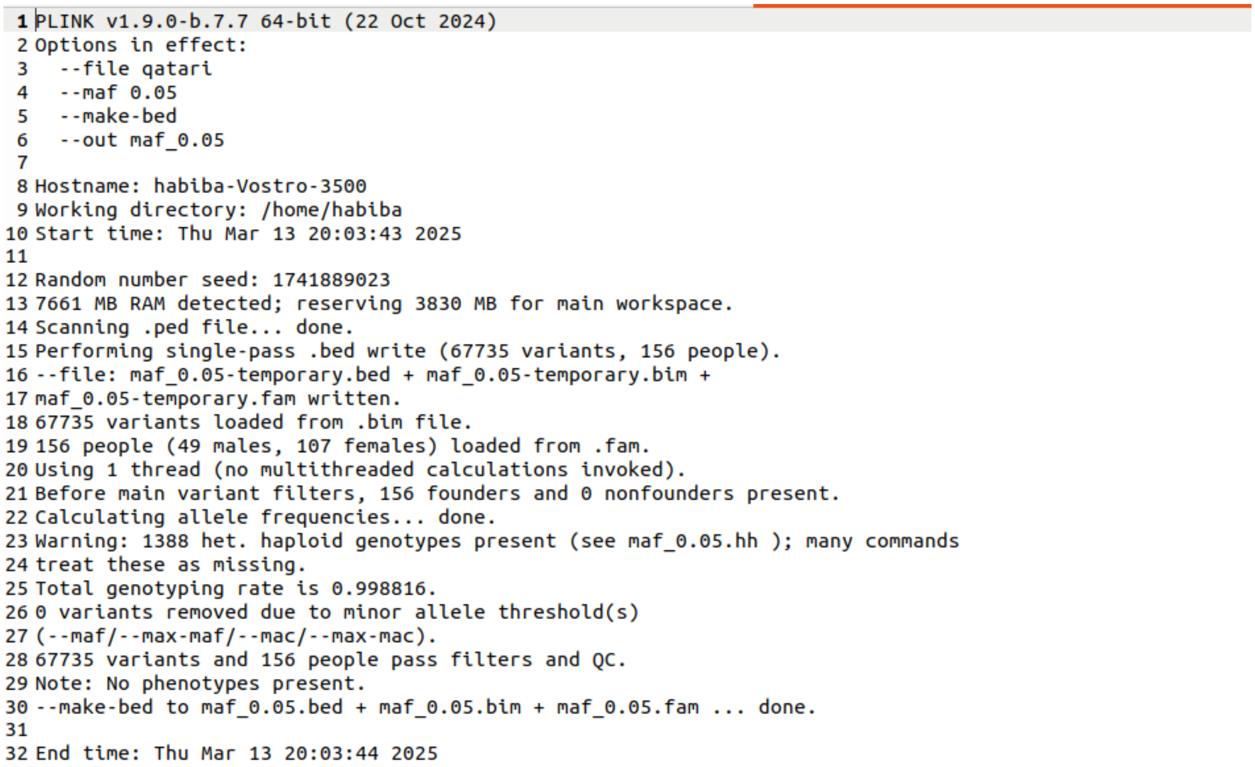
### Variants Removed vs. MAF Threshold

| MAF Threshold on Missing rate | Variants Removed |
|---|---|
| 0.05- | 0 |
| 0.06 | 4067 |
| 0.1 | 16606 |

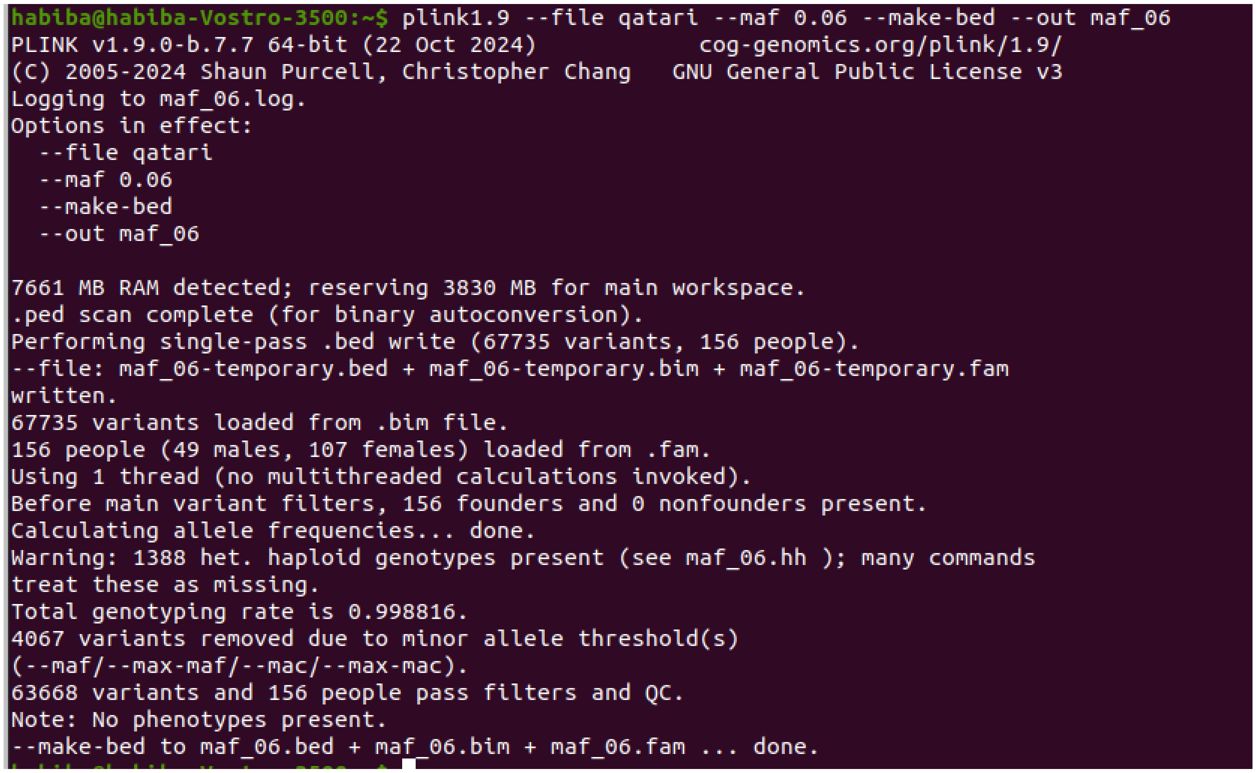
- Thresholds $\leq 0.05$ will not remove any variants since the data is already filtered.
- Thresholds $> 0.5$ are invalid because they do not represent a minor allele.



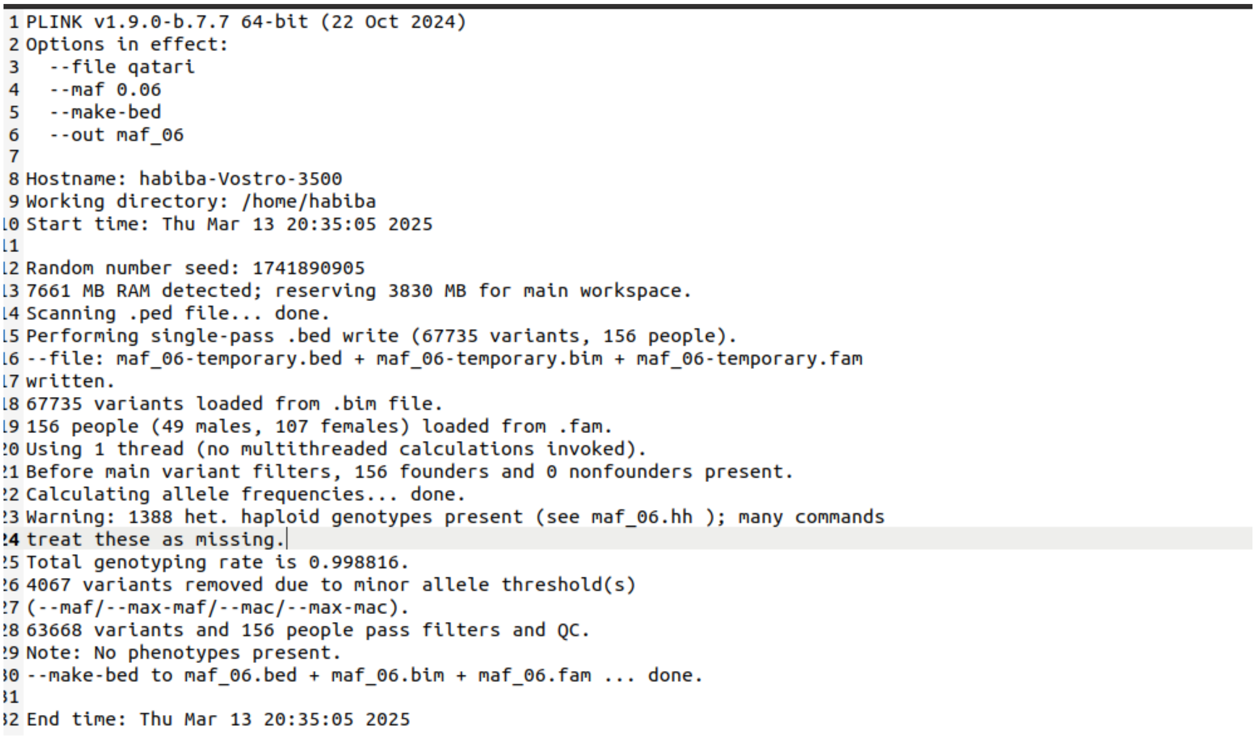
```
habiba@habiba-Vostro-3500:~$ plink1.9 --file qatari --maf 0.05 --make-bed --out maf_0.05
PLINK v1.9.0-b.7.7 64-bit (22 Oct 2024)          cog-genomics.org/plink/1.9/
(C) 2005-2024 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to maf_0.05.log.
Options in effect:
  --file qatari
  --maf 0.05
  --make-bed
  --out maf_0.05

7661 MB RAM detected; reserving 3830 MB for main workspace.
.ped scan complete (for binary autoconversion).
Performing single-pass .bed write (67735 variants, 156 people).
--file: maf_0.05-temporary.bed + maf_0.05-temporary.bim +
maf_0.05-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_0.05.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.
--make-bed to maf_0.05.bed + maf_0.05.bim + maf_0.05.fam ... done.
habiba@habiba-Vostro-3500:~$
```



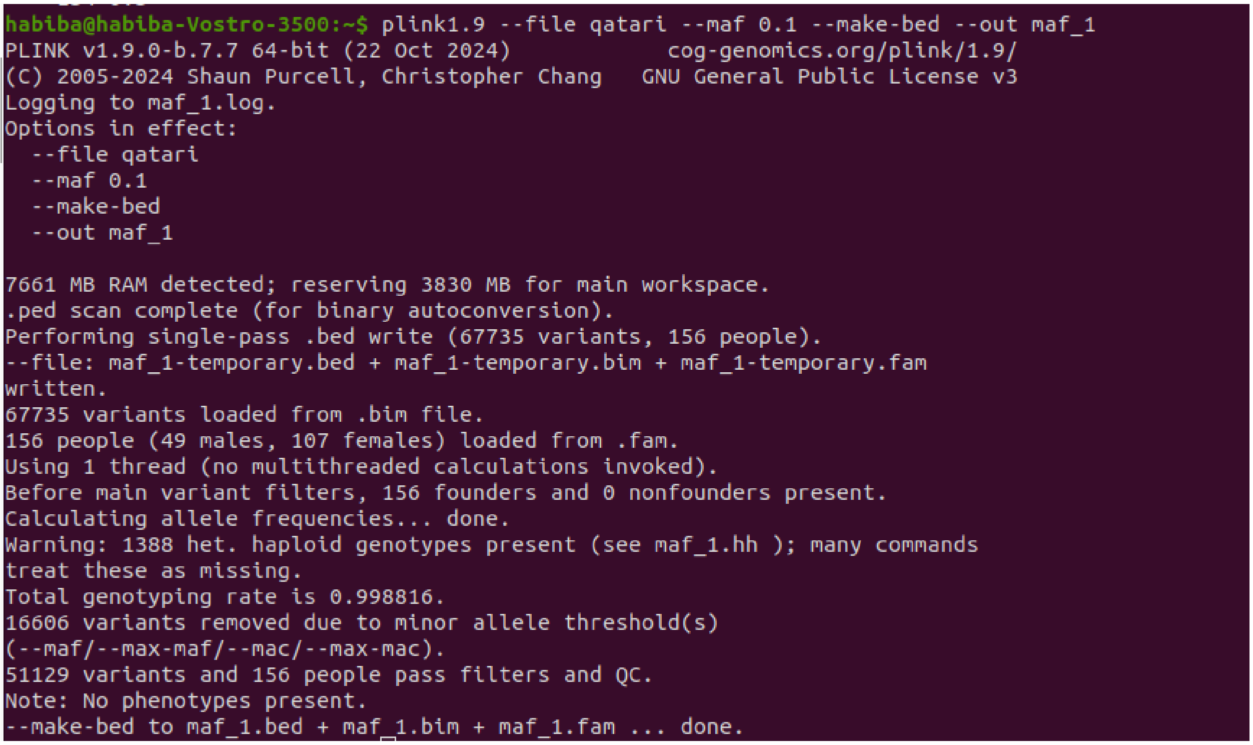
```
PLINK v1.9.0-b.7.7 64-bit (22 Oct 2024)
Options in effect:
  --file qatari
  --maf 0.05
  --make-bed
  --out maf_0.05

Hostname: habiba-Vostro-3500
Working directory: /home/habiba
Start time: Thu Mar 13 20:03:43 2025

Random number seed: 1741889023
7661 MB RAM detected; reserving 3830 MB for main workspace.
Scanning .ped file... done.
Performing single-pass .bed write (67735 variants, 156 people).
--file: maf_0.05-temporary.bed + maf_0.05-temporary.bim +
maf_0.05-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_0.05.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.
--make-bed to maf_0.05.bed + maf_0.05.bim + maf_0.05.fam ... done.

End time: Thu Mar 13 20:03:44 2025
```



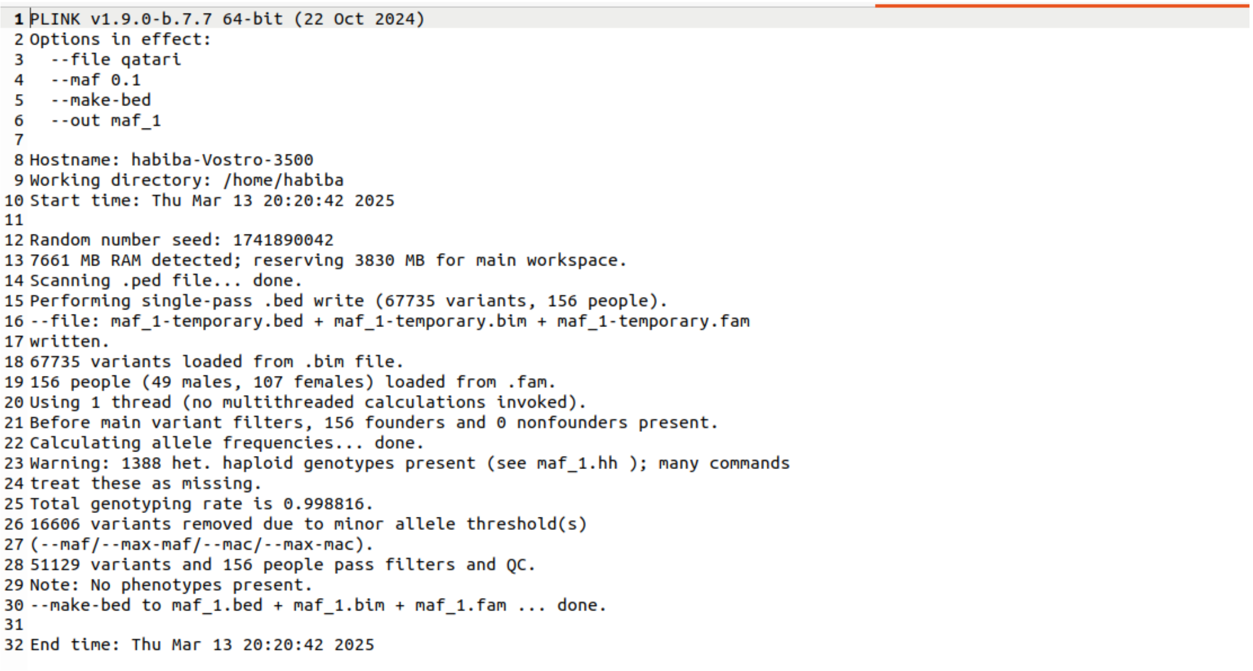
```
habiba@habiba-Vostro-3500:~$ plink1.9 --file qatari --maf 0.06 --make-bed --out maf_06
PLINK v1.9.0-b.7.7 64-bit (22 Oct 2024)            cog-genomics.org/plink/1.9/
(C) 2005-2024 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to maf_06.log.
Options in effect:
  --file qatari
  --maf 0.06
  --make-bed
  --out maf_06

7661 MB RAM detected; reserving 3830 MB for main workspace.
.ped scan complete (for binary autoconversion).
Performing single-pass .bed write (67735 variants, 156 people).
--file: maf_06-temporary.bed + maf_06-temporary.bim + maf_06-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_06.hh ); many commands
treat these as missing.
Total genotyping rate is 0.998816.
4067 variants removed due to minor allele threshold(s)
(--maf/--max-maf/--mac/--max-mac).
63668 variants and 156 people pass filters and QC.
Note: No phenotypes present.
--make-bed to maf_06.bed + maf_06.bim + maf_06.fam ... done.
```



```
PLINK v1.9.0-b.7.7 64-bit (22 Oct 2024)
Options in effect:
  --file qatari
  --maf 0.06
  --make-bed
  --out maf_06

Hostname: habiba-Vostro-3500
Working directory: /home/habiba
Start time: Thu Mar 13 20:35:05 2025

Random number seed: 1741890905
7661 MB RAM detected; reserving 3830 MB for main workspace.
Scanning .ped file... done.
Performing single-pass .bed write (67735 variants, 156 people).
--file: maf_06-temporary.bed + maf_06-temporary.bim + maf_06-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_06.hh ); many commands
treat these as missing.
Total genotyping rate is 0.998816.
4067 variants removed due to minor allele threshold(s)
(--maf/--max-maf/--mac/--max-mac).
63668 variants and 156 people pass filters and QC.
Note: No phenotypes present.
--make-bed to maf_06.bed + maf_06.bim + maf_06.fam ... done.

End time: Thu Mar 13 20:35:05 2025
```



```
habiba@habiba-Vostro-3500:~$ plink1.9 --file qatari --maf 0.1 --make-bed --out maf_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 maf_1.log.
Options in effect:
  --file qatari
  --maf 0.1
  --make-bed
  --out maf_1

7661 MB RAM detected; reserving 3830 MB for main workspace.
.ped scan complete (for binary autoconversion).
Performing single-pass .bed write (67735 variants, 156 people).
--file: maf_1-temporary.bed + maf_1-temporary.bim + maf_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 maf_1.hh ); many commands
treat these as missing.
Total genotyping rate is 0.998816.
16606 variants removed due to minor allele threshold(s)
(--maf/--max-maf/--mac/--max-mac).
51129 variants and 156 people pass filters and QC.
Note: No phenotypes present.
--make-bed to maf_1.bed + maf_1.bim + maf_1.fam ... done.
```



```
PLINK v1.9.0-b.7.7 64-bit (22 Oct 2024)
Options in effect:
  --file qatari
  --maf 0.1
  --make-bed
  --out maf_1

Hostname: habiba-Vostro-3500
Working directory: /home/habiba
Start time: Thu Mar 13 20:20:42 2025

Random number seed: 1741890042
7661 MB RAM detected; reserving 3830 MB for main workspace.
Scanning .ped file... done.
Performing single-pass .bed write (67735 variants, 156 people).
--file: maf_1-temporary.bed + maf_1-temporary.bim + maf_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 maf_1.hh ); many commands
treat these as missing.
Total genotyping rate is 0.998816.
16606 variants removed due to minor allele threshold(s)
(--maf/--max-maf/--mac/--max-mac).
51129 variants and 156 people pass filters and QC.
Note: No phenotypes present.
--make-bed to maf_1.bed + maf_1.bim + maf_1.fam ... done.

End time: Thu Mar 13 20:20:42 2025
```

## 2. HWE:

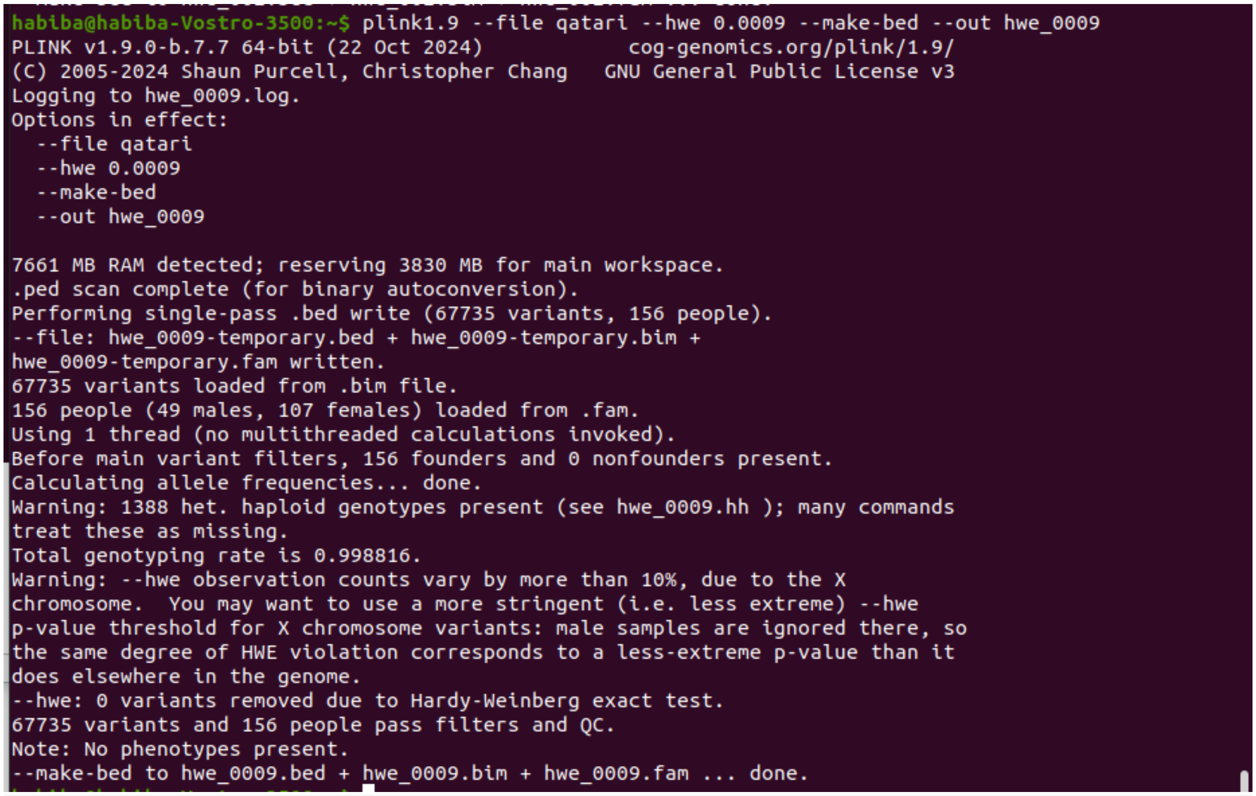
remove SNPs with significant deviations from HWE to prevent false positives caused by technical errors.

### Variants Removed vs. HWE Threshold

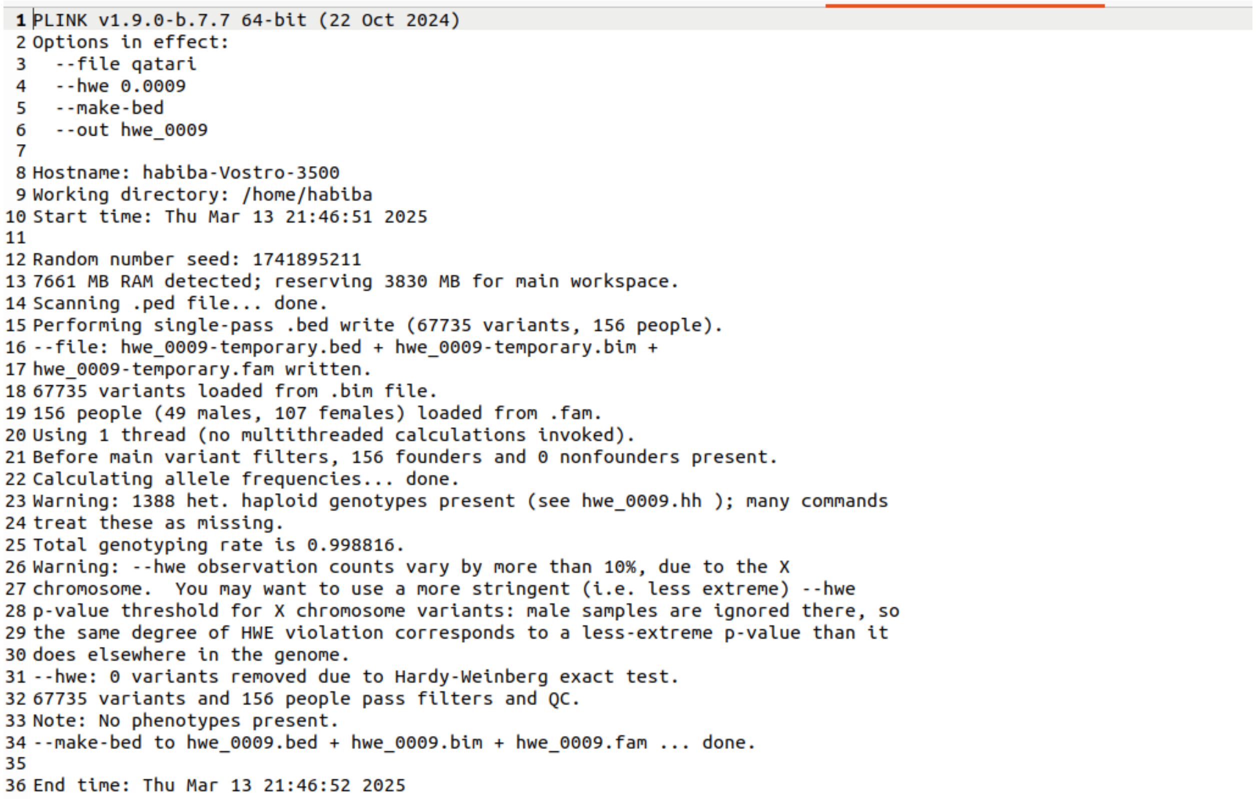
| HWE Threshold on P-value | Variants Removed |
|---|---|
| 0.001 | 0 |
| 0.0011 | 32 |
| 0.0009 | 0 |

- Thresholds $\leq 0.001$ will not remove any variants because the data has already been filtered using a threshold of approximately 0.01.

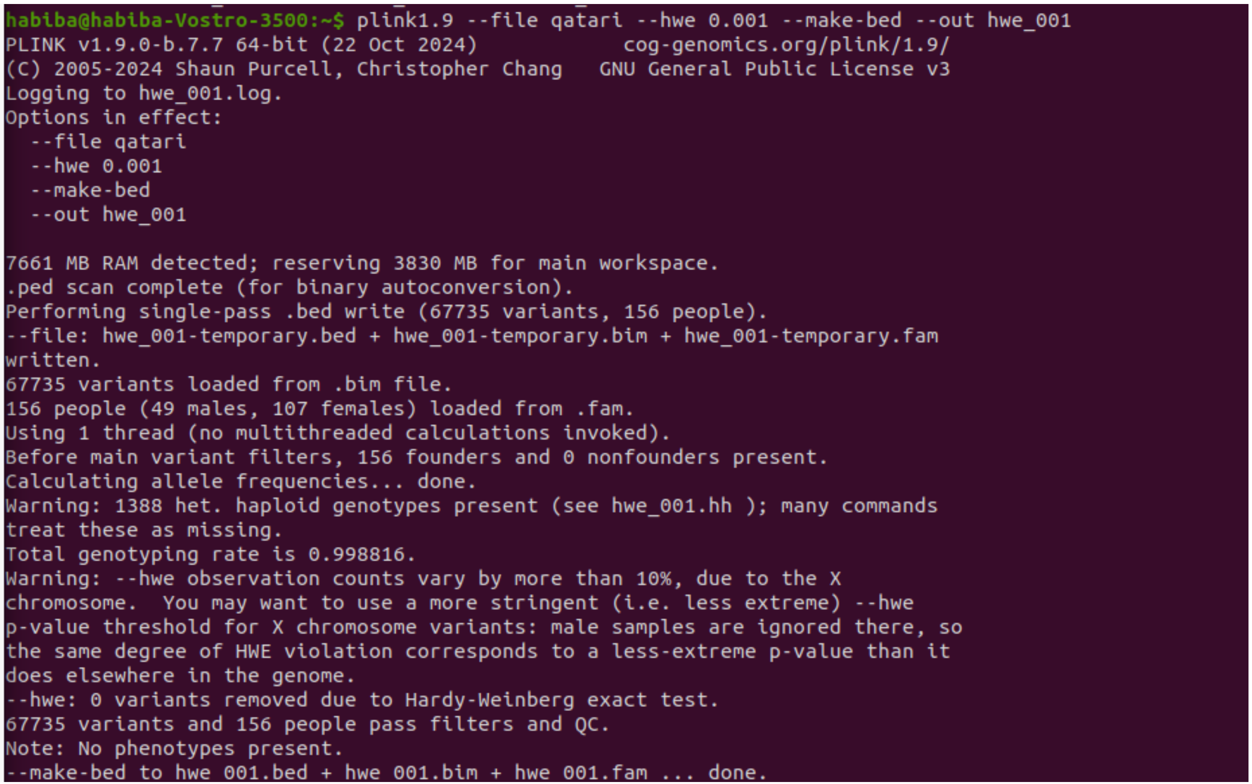
- Thresholds $> 0.001$ remove SNPs with significant deviations from HWE to prevent false positives caused by technical errors.



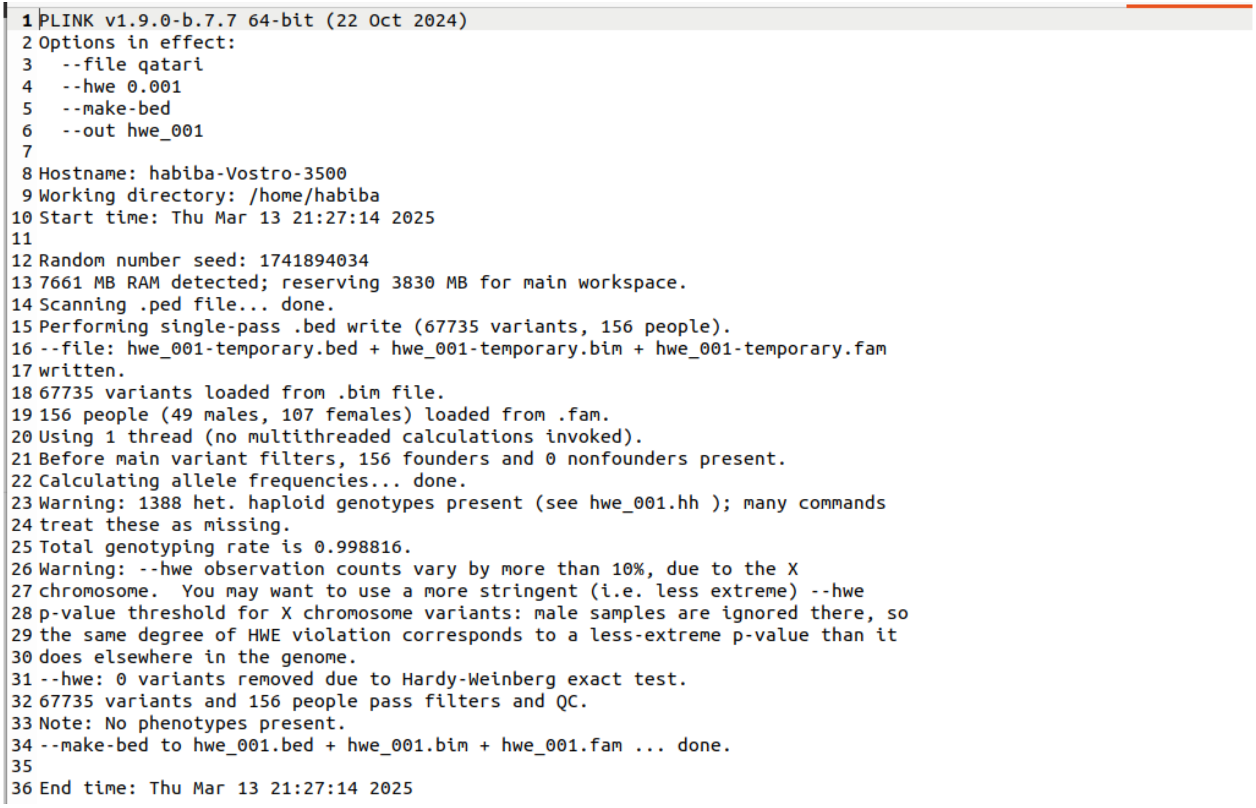
```
habiba@habiba-Vostro-3500:~$ plink1.9 --file qatari --hwe 0.0009 --make-bed --out hwe_0009
PLINK v1.9.0-b.7.7 64-bit (22 Oct 2024)            cog-genomics.org/plink/1.9/
(C) 2005-2024 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to hwe_0009.log.
Options in effect:
  --file qatari
  --hwe 0.0009
  --make-bed
  --out hwe_0009

7661 MB RAM detected; reserving 3830 MB for main workspace.
.ped scan complete (for binary autoconversion).
Performing single-pass .bed write (67735 variants, 156 people).
--file: hwe_0009-temporary.bed + hwe_0009-temporary.bim +
hwe_0009-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_0009.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.
--make-bed to hwe_0009.bed + hwe_0009.bim + hwe_0009.fam ... done.
```



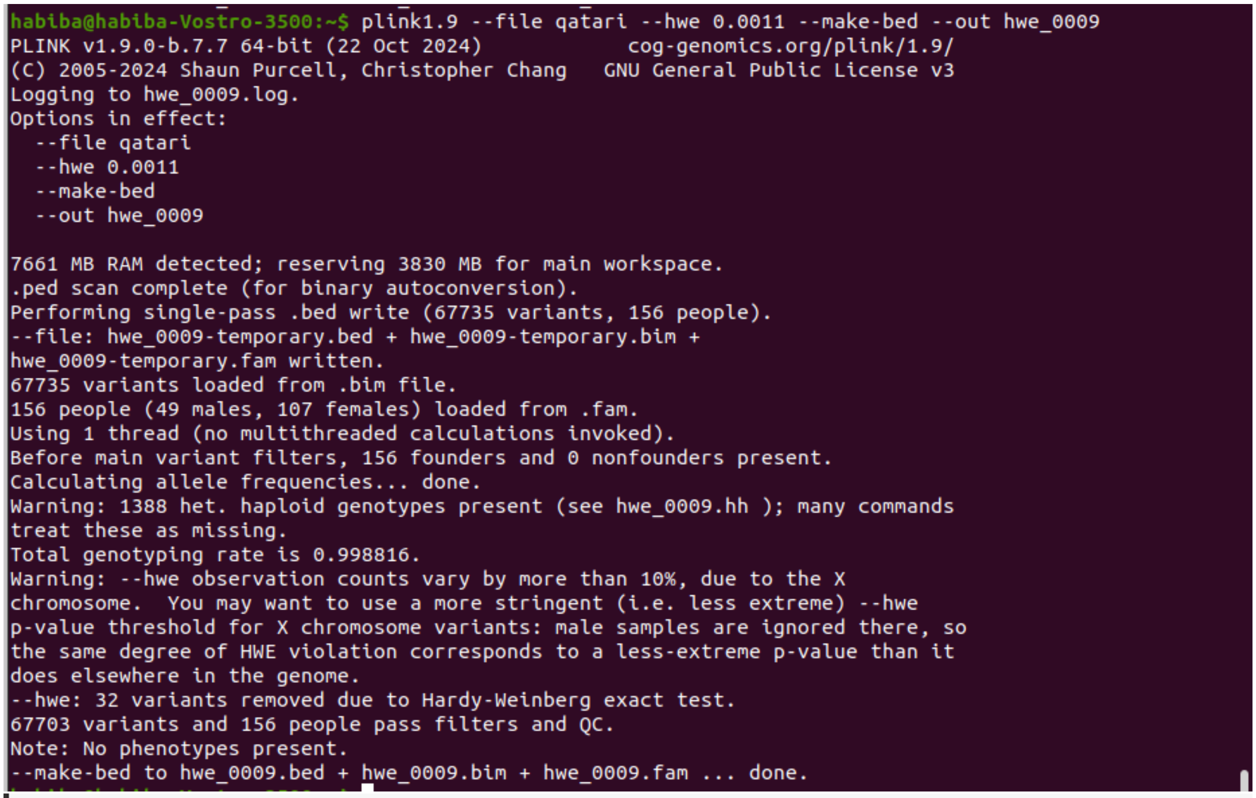
```
PLINK v1.9.0-b.7.7 64-bit (22 Oct 2024)
Options in effect:
  --file qatari
  --hwe 0.0009
  --make-bed
  --out hwe_0009

Hostname: habiba-Vostro-3500
Working directory: /home/habiba
Start time: Thu Mar 13 21:46:51 2025

Random number seed: 1741895211
7661 MB RAM detected; reserving 3830 MB for main workspace.
Scanning .ped file... done.
Performing single-pass .bed write (67735 variants, 156 people).
--file: hwe_0009-temporary.bed + hwe_0009-temporary.bim +
hwe_0009-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_0009.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.
--make-bed to hwe_0009.bed + hwe_0009.bim + hwe_0009.fam ... done.

End time: Thu Mar 13 21:46:52 2025
```



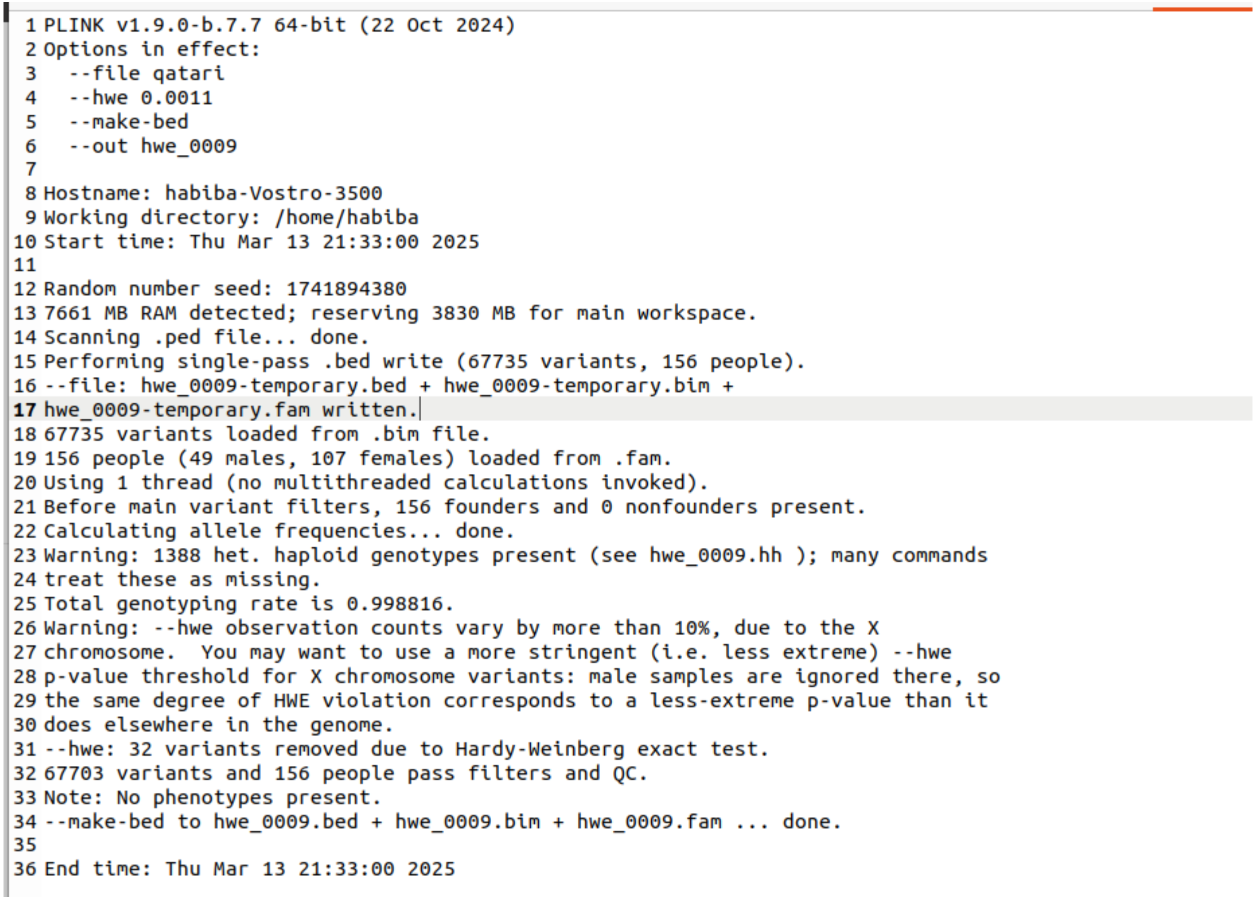
```
habiba@habiba-Vostro-3500:~$ plink1.9 --file qatari --hwe 0.001 --make-bed --out hwe_001
PLINK v1.9.0-b.7.7 64-bit (22 Oct 2024)           cog-genomics.org/plink/1.9/
(C) 2005-2024 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to hwe_001.log.
Options in effect:
  --file qatari
  --hwe 0.001
  --make-bed
  --out hwe_001

7661 MB RAM detected; reserving 3830 MB for main workspace.
.ped scan complete (for binary autoconversion).
Performing single-pass .bed write (67735 variants, 156 people).
--file: hwe_001-temporary.bed + hwe_001-temporary.bim + hwe_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 hwe_001.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.
--make-bed to hwe_001.bed + hwe_001.bim + hwe_001.fam ... done.
```



```
PLINK v1.9.0-b.7.7 64-bit (22 Oct 2024)
Options in effect:
  --file qatari
  --hwe 0.001
  --make-bed
  --out hwe_001

Hostname: habiba-Vostro-3500
Working directory: /home/habiba
Start time: Thu Mar 13 21:27:14 2025

Random number seed: 1741894034
7661 MB RAM detected; reserving 3830 MB for main workspace.
Scanning .ped file... done.
Performing single-pass .bed write (67735 variants, 156 people).
--file: hwe_001-temporary.bed + hwe_001-temporary.bim + hwe_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 hwe_001.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.
--make-bed to hwe_001.bed + hwe_001.bim + hwe_001.fam ... done.

End time: Thu Mar 13 21:27:14 2025
```



```
habiba@habiba-Vostro-3500:~$ plink1.9 --file qatari --hwe 0.0011 --make-bed --out hwe_0009
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_0009.log.
Options in effect:
  --file qatari
  --hwe 0.0011
  --make-bed
  --out hwe_0009

7661 MB RAM detected; reserving 3830 MB for main workspace.
.ped scan complete (for binary autoconversion).
Performing single-pass .bed write (67735 variants, 156 people).
--file: hwe_0009-temporary.bed + hwe_0009-temporary.bim +
hwe_0009-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_0009.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: 32 variants removed due to Hardy-Weinberg exact test.
67703 variants and 156 people pass filters and QC.
Note: No phenotypes present.
--make-bed to hwe_0009.bed + hwe_0009.bim + hwe_0009.fam ... done.
```



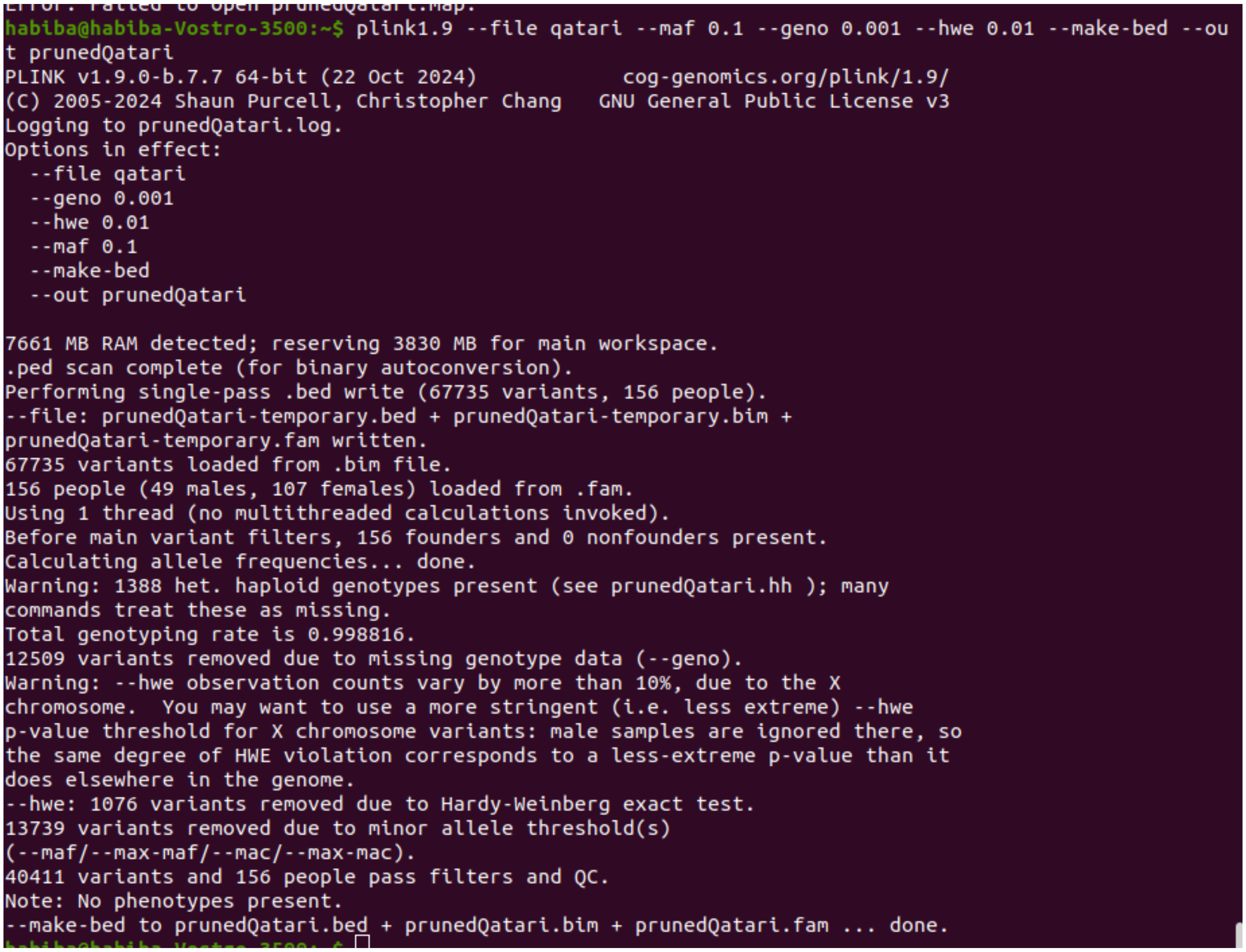
```
PLINK v1.9.0-b.7.7 64-bit (22 Oct 2024)
Options in effect:
  --file qatari
  --hwe 0.0011
  --make-bed
  --out hwe_0009

Hostname: habiba-Vostro-3500
Working directory: /home/habiba
Start time: Thu Mar 13 21:33:00 2025

Random number seed: 1741894380
7661 MB RAM detected; reserving 3830 MB for main workspace.
Scanning .ped file... done.
Performing single-pass .bed write (67735 variants, 156 people).
--file: hwe_0009-temporary.bed + hwe_0009-temporary.bim +
hwe_0009-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_0009.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: 32 variants removed due to Hardy-Weinberg exact test.
67703 variants and 156 people pass filters and QC.
Note: No phenotypes present.
--make-bed to hwe_0009.bed + hwe_0009.bim + hwe_0009.fam ... done.

End time: Thu Mar 13 21:33:00 2025
```

3. **GENO:** Go to Page 5

## The final version of your QC

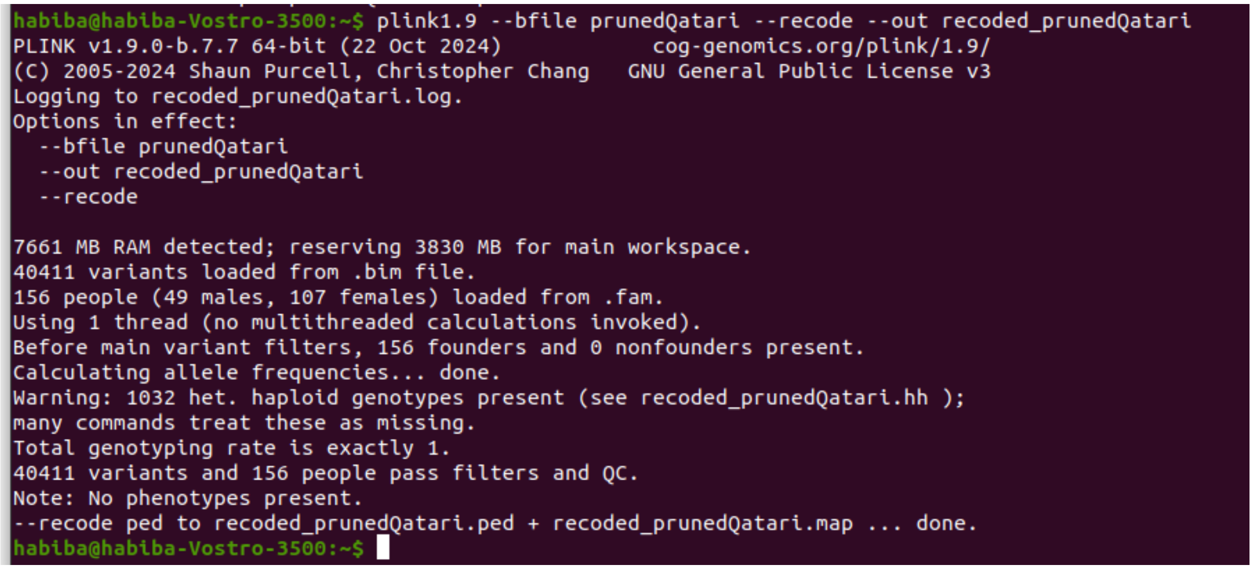
Total number of removed variants is 27324

| Test | Threshold | Variants Removed |
|---|---|---|
| HWE | 0.01 | 1076 |
| MAF | 0.1 | 13739 |
| GENO | 0.001 | 12509 |



```
Error: Failed to open prunedQatari.map.
habiba@habiba-Vostro-3500:~$ plink1.9 --file qatari --maf 0.1 --geno 0.001 --hwe 0.01 --make-bed --ou
t prunedQatari
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 prunedQatari.log.
Options in effect:
  --file qatari
  --geno 0.001
  --hwe 0.01
  --maf 0.1
  --make-bed
  --out prunedQatari

7661 MB RAM detected; reserving 3830 MB for main workspace.
.ped scan complete (for binary autoconversion).
Performing single-pass .bed write (67735 variants, 156 people).
--file: prunedQatari-temporary.bed + prunedQatari-temporary.bim +
prunedQatari-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 prunedQatari.hh ); many
commands treat these as missing.
Total genotyping rate is 0.998816.
12509 variants removed due to missing genotype data (--geno).
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: 1076 variants removed due to Hardy-Weinberg exact test.
13739 variants removed due to minor allele threshold(s)
(--maf/--max-maf/--mac/--max-mac).
40411 variants and 156 people pass filters and QC.
Note: No phenotypes present.
--make-bed to prunedQatari.bed + prunedQatari.bim + prunedQatari.fam ... done.
habiba@habiba-Vostro-3500:~$
```

### Recoding the dataset:



```
habiba@habiba-Vostro-3500:~$ plink1.9 --bfile prunedQatari --recode --out recoded_prunedQatari
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 recoded_prunedQatari.log.
Options in effect:
  --bfile prunedQatari
  --out recoded_prunedQatari
  --recode

7661 MB RAM detected; reserving 3830 MB for main workspace.
40411 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: 1032 het. haploid genotypes present (see recoded_prunedQatari.hh );
many commands treat these as missing.
Total genotyping rate is exactly 1.
40411 variants and 156 people pass filters and QC.
Note: No phenotypes present.
--recode ped to recoded_prunedQatari.ped + recoded_prunedQatari.map ... done.
habiba@habiba-Vostro-3500:~$
```

```r
ped_dataBef <- read.table("qatari.ped", header = FALSE, stringsAsFactors

cat("Number of Rows (Samples):", nrow(ped_dataBef), "\n")
```

```
## Number of Rows (Samples): 156
```

```r
cat("Number of Columns (Genotypes + Metadata):", ncol(ped_dataBef), "\n")
```

```
## Number of Columns (Genotypes + Metadata): 135476
```

```r
ped_data <- read.table("recoded_prunedQatari.ped", header = FALSE, string

cat("Number of Rows (Samples):", nrow(ped_data), "\n")
```

```
## Number of Rows (Samples): 156
```

```r
cat("Number of Columns (Genotypes + Metadata):", ncol(ped_data), "\n")
```

```
## Number of Columns (Genotypes + Metadata): 80828
```

The change in data:

- The Same number of samples
- The SNPs “variants” reduced from ($\frac{135476-6}{2} = 67735$) to ($\frac{80828-6}{2} = 40411$)