Bioinformatics Course

Lab 4: PLINK Setup

Name: Maria Bassem Emil ID: 20011141

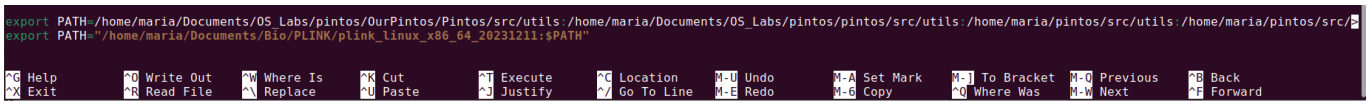
Objective

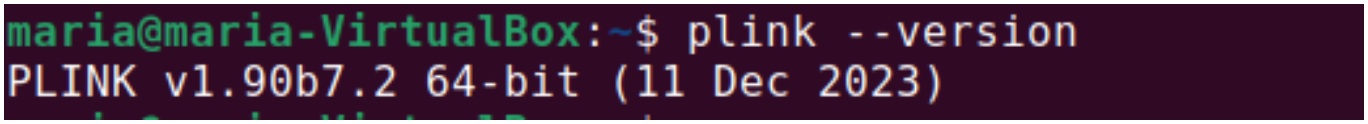
Introduction to the PLINK CLI tool.

Part 1: Installation (10 Points)

➔ Add Plink to your Path environment variable. It is found in ~/.bashrc in Linux-based platforms.

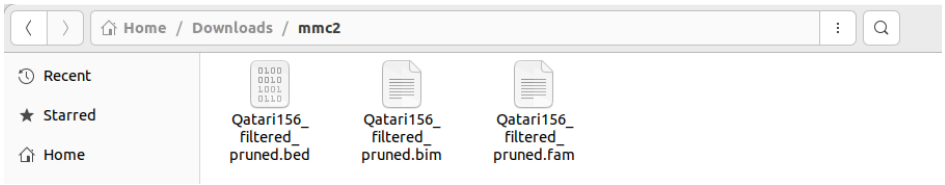
Adding the path via typing “nano ~/.bashrc” then





Part 2: Basic Commands (15 Points)

➔ Extract the dataset. What formats do you see?



.bed, .bim, .fam

- BED (.bed):
  - The .bed file contains the genotype data in a binary format, where each SNP is represented by two bits (e.g., 00 for homozygous reference, 01 for heterozygous, and 11 for homozygous variant).
- BIM (.bim):

- The BIM (or .bim) file contains information about the genetic variants (e.g., SNPs) in the dataset.
- It is composed of 6 columns which are the same as the map format but we have two extra columns for the alleles (A1, A2) i.e: reference allele and alternate allele.
    - Chromosome Number (here it is 1)
    - Marker ID (SNP rs ID)
    - Genetic distance
    - Physical position (Base-pair position)
    - Allele 1
    - Allele 2

◆ FAM (.fam):

- The FAM (or .fam) file contains information about the individuals (samples) in the dataset.
- Each row in the FAM file represents an individual and includes information such as family ID, individual ID, paternal ID, maternal ID, sex (coded as 1 for male and 2 for female), and phenotype information (e.g., disease status).

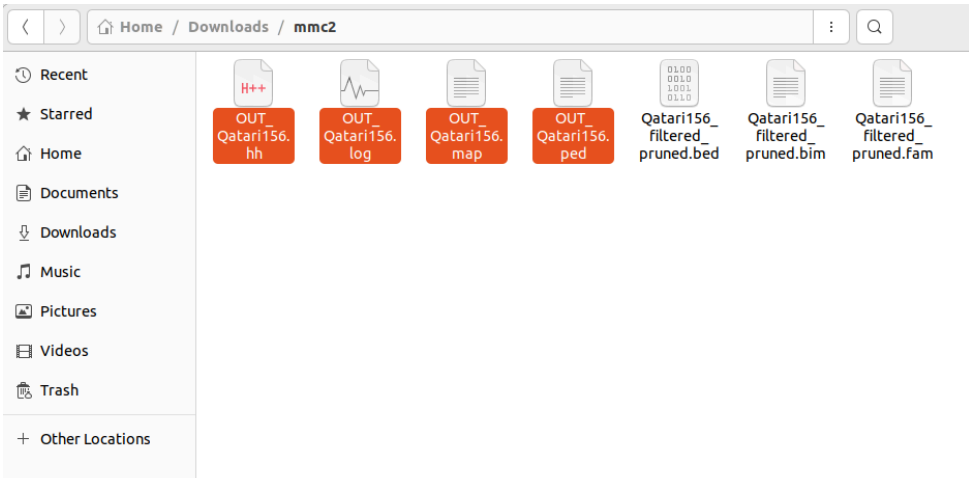
◆ These file formats, collectively known as PLINK binary formats, are widely used in genome-wide association study (GWAS), and other genetic analyses.

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

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

```
maria@maria-VirtualBox:~/Downloads/mmc2$ plink --bfile Qatari156_filtered_pruned --recode --out OUT_Qatari156
PLINK v1.90b7.2 64-bit (11 Dec 2023)           www.cog-genomics.org/plink/1.9/
(C) 2005-2023 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to OUT_Qatari156.log.
Options in effect:
  --bfile Qatari156_filtered_pruned
  --out OUT_Qatari156
  --recode

4429 MB RAM detected; reserving 2214 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 OUT_Qatari156.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 OUT_Qatari156.ped + OUT_Qatari156.map ... done.
maria@maria-VirtualBox:~/Downloads/mmc2$
```



➔ Upon executing the previous command, observe the terminal output. Specify the number of variants and the number of samples

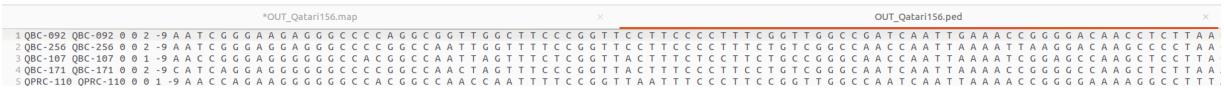
- Number of variants: 67735 variants were loaded from the .bim file.
- Number of samples: 156 people were loaded from the .fam file.

```
67735 variants loaded from .bim file.
156 people (49 males, 107 females) loaded from .fam.
```

➔ Describe the columns of the ped and map files while exploring each file's head (first 5 rows).

- This link was very helpful in knowing the format of each file

1. .ped file:




```
maria@maria-VirtualBox:~/Downloads/mmc2$ awk '{print substr($0, 1, 50) "..." substr($0, length($0)-49)}' OUT_Qatari156.ped | head -n 5
QBC-092 QBC-092 0 0 2 -9 A A T C G G G A A G A G G... T T T C C C C G G A A C C A A A G C C A G A A G G
QBC-256 QBC-256 0 0 2 -9 A A T C G G G A G G A G G... C C C C T C C G G A A T T G A G G C C G G A A A A
QBC-107 QBC-107 0 0 1 -9 A A C C G G G A G G G G G... C C C T T C C G G A A C C A A G G C C G G A A G G
QBC-171 QBC-171 0 0 2 -9 C A T C A G G A G G G G G... C T C T T A C A G T A C T A A G G C C A G G A G G
QPRC-110 QPRC-110 0 0 1 -9 A A C C A G A A G G G G... T T T C C C C G G A A C C G G G G C C G G A A G G
```

The columns in a PED file are

- Family ID (FID)
- Sample/Individual ID
- Paternal ID [It appears to be set to "0" in the provided data, suggesting that paternal information may not be available or necessary for these individuals].
- Maternal ID [It appears to be set to "0" in the provided data, suggesting that paternal information may not be available or necessary for these individuals].
- Sex (1=male; 2=female)

- Phenotype (“-9” means that the phenotype information may not be available, as mentioned by the professor).
- Genotypes ('A', 'C', 'G', 'T', or ‘-9’ = missing)

2. .map file:

OUT_Qatari156.map

```
1 1       rs10907175      1.12059 1120590
2 1       rs7519837       1.500664        1500664
3 1       rs10907187      1.748914        1748914
4 1       rs6603803       1.802548        1802548
5 1       rs6688000       1.813782        1813782
```

```
maria@maria-VirtualBox:~/Downloads/mmc2$ head -n 5 OUT_Qatari156.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
```

The columns in a MAP file are:

- Chromosome Number (here it is 1)
- Marker ID (SNP rs ID)
- Genetic distance
- Physical position (Base-pair position)

➔ Perform one of the quality controls, Missing Call Rate, found in this link.

Try different thresholds and report the number of variants removed based on the thresholds used.

- The --bfile means that the given file is in .bed, .bim, .fam format, while --file means that the file is given in the .ped, .map format.
- The --geno <threshold> parameter specifies genotype filtering, where variants (SNPs) with missing genotype rates greater than <threshold> are excluded. This step helps remove low-quality variants that have a high proportion of missing genotypes, which could be due to genotyping errors.
- The --recode parameter indicates that the data will be written into a ped/map format after filtering.
- The --make-bed command is used to convert a ped format to binary bed format. The bed format is more efficient for certain operations and typically used for large-scale genetic data analysis.

Missing rate per SNP

Subsequent analyses can be set to automatically exclude SNPs on the basis of missing genotype rate, with the `--geno` option: the default is to include all SNPS (i.e. `--geno 1`). To include only SNPs with a 90% genotyping rate (10% missing) use

```
plink --file mydata --geno 0.1
```

As with the `--maf` option, these counts are calculated after removing individuals with high missing genotype rates.

1. .bed,.bim,.fam:

plink --bfile <prefix> --geno <threshold> --recode --make-bed -out <filename>

- Threshold = 0.00003

```
maria@maria-VirtualBox:~/Downloads/mmc2$ plink --bfile Qatari156_filtered_pruned --geno 0.00003 --recode --make-bed -out Qatari156_filtered_pruned0.00003
PLINK v1.90b7.2 64-bit (11 Dec 2023)           www.cog-genomics.org/plink/1.9/
(C) 2005-2023 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to Qatari156_filtered_pruned0.00003.log.
Options in effect:
  --bfile Qatari156_filtered_pruned
  --geno 0.00003
  --make-bed
  --out Qatari156_filtered_pruned0.00003
  --recode

4429 MB RAM detected; reserving 2214 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
Qatari156_filtered_pruned0.00003.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 Qatari156_filtered_pruned0.00003.bed +
Qatari156_filtered_pruned0.00003.bim + Qatari156_filtered_pruned0.00003.fam ...
done.
--recode ped to Qatari156_filtered_pruned0.00003.ped +
Qatari156_filtered_pruned0.00003.map ... done.
```

Total genotyping rate is 0.998816.
12509 variants removed due to missing genotype data (--geno).
55226 variants and 156 people pass filters and QC.

- Threshold = 0.005

```
maria@maria-VirtualBox:~/Downloads/mmc2$ plink --bfile Qatari156_filtered_pruned --geno 0.005 --recode --make-bed -out Qatari156_filtered_pruned0.005
PLINK v1.90b7.2 64-bit (11 Dec 2023)           www.cog-genomics.org/plink/1.9/
(C) 2005-2023 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to Qatari156_filtered_pruned0.005.log.
Options in effect:
  --bfile Qatari156_filtered_pruned
  --geno 0.005
  --make-bed
  --out Qatari156_filtered_pruned0.005
  --recode

4429 MB RAM detected; reserving 2214 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
Qatari156_filtered_pruned0.005.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 Qatari156_filtered_pruned0.005.bed +
Qatari156_filtered_pruned0.005.bim + Qatari156_filtered_pruned0.005.fam ...
done.
--recode ped to Qatari156_filtered_pruned0.005.ped +
Qatari156_filtered_pruned0.005.map ... done.
```

Total genotyping rate is 0.998816.

12509 variants removed due to missing genotype data (--geno).
55226 variants and 156 people pass filters and QC.

- Threshold = 0.05

Total genotyping rate is 0.998816.
0 variants removed due to missing genotype data (--geno).
67735 variants and 156 people pass filters and QC.

- Threshold = 0.1

Total genotyping rate is 0.998816.
0 variants removed due to missing genotype data (--geno).
67735 variants and 156 people pass filters and QC.

2. .ped, .map:

plink --file <prefix> --geno <threshold> --recode --make-bed -out <filename>

- Threshold = 0.0003

```
maria@maria-VirtualBox:~/Downloads/mmc2$ plink --file OUT_Qatari156 --geno 0.00003 --recode --make-bed -out OUT_Qatari156_0.00003
PLINK v1.90b7.2 64-bit (11 Dec 2023)           www.cog-genomics.org/plink/1.9/
(C) 2005-2023 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to OUT_Qatari156_0.00003.log.
Options in effect:
  --file OUT_Qatari156
  --geno 0.00003
  --make-bed
  --out OUT_Qatari156_0.00003
  --recode

4429 MB RAM detected; reserving 2214 MB for main workspace.
.ped scan complete (for binary autoconversion).
Performing single-pass .bed write (67735 variants, 156 people).
--file: OUT_Qatari156_0.00003-temporary.bed +
OUT_Qatari156_0.00003-temporary.bim + OUT_Qatari156_0.00003-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 OUT_Qatari156_0.00003.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 OUT_Qatari156_0.00003.bed + OUT_Qatari156_0.00003.bim +
OUT_Qatari156_0.00003.fam ... done.
--recode ped to OUT_Qatari156_0.00003.ped + OUT_Qatari156_0.00003.map ... done.
```

Total genotyping rate is 0.998816.
12509 variants removed due to missing genotype data (--geno).
55226 variants and 156 people pass filters and QC.

- Threshold = 0.005

```
maria@maria-VirtualBox:~/Downloads/mmc2$ plink --file OUT_Qatari156 --geno 0.005 --recode --make-bed -out OUT_Qatari156_0.005
PLINK v1.90b7.2 64-bit (11 Dec 2023)           www.cog-genomics.org/plink/1.9/
(C) 2005-2023 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to OUT_Qatari156_0.005.log.
Options in effect:
  --file OUT_Qatari156
  --geno 0.005
  --make-bed
  --out OUT_Qatari156_0.005
  --recode

4429 MB RAM detected; reserving 2214 MB for main workspace.
.ped scan complete (for binary autoconversion).
Performing single-pass .bed write (67735 variants, 156 people).
--file: OUT_Qatari156_0.005-temporary.bed + OUT_Qatari156_0.005-temporary.bim +
OUT_Qatari156_0.005-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 OUT_Qatari156_0.005.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 OUT_Qatari156_0.005.bed + OUT_Qatari156_0.005.bim +
OUT_Qatari156_0.005.fam ... done.
--recode ped to OUT_Qatari156_0.005.ped + OUT_Qatari156_0.005.map ... done.
```

Total genotyping rate is 0.998816.
12509 variants removed due to missing genotype data (--geno).
55226 variants and 156 people pass filters and QC.

- Threshold = 0.05

```
maria@maria-VirtualBox:~/Downloads/mmc2$ plink --file OUT_Qatari156 --geno 0.05 --recode --make-bed -out OUT_Qatari156_0.05
PLINK v1.90b7.2 64-bit (11 Dec 2023)           www.cog-genomics.org/plink/1.9/
(C) 2005-2023 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to OUT_Qatari156_0.05.log.
Options in effect:
  --file OUT_Qatari156
  --geno 0.05
  --make-bed
  --out OUT_Qatari156_0.05
  --recode

4429 MB RAM detected; reserving 2214 MB for main workspace.
.ped scan complete (for binary autoconversion).
Performing single-pass .bed write (67735 variants, 156 people).
--file: OUT_Qatari156_0.05-temporary.bed + OUT_Qatari156_0.05-temporary.bim +
OUT_Qatari156_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 OUT_Qatari156_0.05.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 OUT_Qatari156_0.05.bed + OUT_Qatari156_0.05.bim +
OUT_Qatari156_0.05.fam ... done.
--recode ped to OUT_Qatari156_0.05.ped + OUT_Qatari156_0.05.map ... done.
```

Total genotyping rate is 0.998816.
0 variants removed due to missing genotype data (--geno).
67735 variants and 156 people pass filters and QC.

- Threshold = 0.1

```
maria@maria-VirtualBox:~/Downloads/mmc2$ plink --file OUT_Qatari156 --geno 0.1 --recode --make-bed -out OUT_Qatari156_0.1
PLINK v1.90b7.2 64-bit (11 Dec 2023)           www.cog-genomics.org/plink/1.9/
(C) 2005-2023 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to OUT_Qatari156_0.1.log.
Options in effect:
  --file OUT_Qatari156
  --geno 0.1
  --make-bed
  --out OUT_Qatari156_0.1
  --recode

4429 MB RAM detected; reserving 2214 MB for main workspace.
.ped scan complete (for binary autoconversion).
Performing single-pass .bed write (67735 variants, 156 people).
--file: OUT_Qatari156_0.1-temporary.bed + OUT_Qatari156_0.1-temporary.bim +
OUT_Qatari156_0.1-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 OUT_Qatari156_0.1.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 OUT_Qatari156_0.1.bed + OUT_Qatari156_0.1.bim +
OUT_Qatari156_0.1.fam ... done.
--recode ped to OUT_Qatari156_0.1.ped + OUT_Qatari156_0.1.map ... done.
```

Total genotyping rate is 0.998816.
0 variants removed due to missing genotype data (--geno).
67735 variants and 156 people pass filters and QC.

Comments:

- Regarding the high rate of genotyping in the quality control problem, it suggests that the dataset is of high quality with very few or no missing genotypes. This is generally a good sign for genetic analysis, as it ensures that the data is reliable and can provide accurate results.

- The data quality control steps performed using PLINK help improve the reliability and usability of the genetic dataset by removing low-quality variants with high rates of missing genotypes. This ensures that upcoming analyses are based on high-quality data, leading to more accurate and meaningful results in genetic association studies or other analyses.