CISC875/BMED866: PLINK Tutorial

PLINK v. 1.9: https://www.cog-genomics.org/plink/1.9/

1. File Formats

Prior to using PLINK, experimental data is usually kept in two separate files - one for genotype and phenotype data and another file with genetic map information. The file with genotype data is called a PEDFILE, short for pedigree file, and follows a standard format in which the first six variables represent the following:

ˆ *Family ID Number*
ˆ *Individual ID Number*
ˆ *ID of the Father of this Individual, if Included in the Study*
ˆ *ID of the Father of this Individual, if Included in the Study*
ˆ *Gender*
ˆ *Phenotype, 1=Unaffected, 2=Affected*

Let's take a look at an actual pedfile using R. First we need to get the file in the right place.

1.1. Access your CAC student accounts:

$ ssh –X yourusername@login.cac.queensu.ca

Copy all GWAS files into your home directory:

$ mkdir GWAS
$ scp /global/project/hpcg1554/PLINKtutorial/GWAS4* /global/home/yourusername/GWAS
$ cd GWAS

1.2. Load R:

$ salloc -c 1 --mem=48g

$module load gcc/7.3.0
$module load r/3.5.1

$ module load R
$ R

Optional: use RStudio:
https://login.cac.queensu.ca/ (click RStudio in top right corner)
login using CAC username and password
click RStudio once again

1.3. Read in a pedfile to R:

> **simped <- read.table("GWAS4.ped", header = F, stringsAsFactors = F)**

To show dimensions of the dataset and display values for the first five individuals and first ten variables.

```
> nrow(simped)
[1] 244
> ncol(simped)
[1] 12782
> simped[1:5, 1:10]
   V1      V2 V3 V4 V5 V6 V7 V8 V9 V10
1 1328 NA06989  0  0  2 -9  C  C  A   A
2 1377 NA11891  0  0  1 -9  C  C  A   A
3 1349 NA11843  0  0  1 -9  T  T  A   A
4 1328 NA06984  0  0  1 -9  C  C  A   A
5 1418 NA12275  0  0  2 -9  T  C  A   A
```

The first six columns are:

V1 Family ID
V2 Individual ID
V3 Paternal ID
V4 Maternal ID
V5 Sex (1=male; 2=female; other=unknown)
V6 Phenotype (-9=missing; 0=missing; 1=unaffected; 2=affected)

Genotypes (column 7 onwards) can be any character (e.g. 1,2,3,4 or A,C,G,T or anything else) except 0 which is, by default, the missing genotype character. All SNPs (whether haploid or not) must have two alleles specified, otherwise it should be missing (i.e. 0).

1.4. Read in the .map file:

```
> simmap<- read.table("GWAS4.map", header=F, stringsAsFactors=F)
> nrow(simmap)
[1] 6388
> ncol(simmap)
[1] 4
> simmap[1:5, ]
  V1       V2 V3      V4
1  1 rs2843403  0 2518957
2  1 rs4648462  0 3155127
3  1 rs7410846  0 3926588
4  1 rs1490413  0 4267183
5  1 rs1878052  0 4452662
```

By default, each line of the MAP file describes a single marker and must contain exactly 4 columns:

chromosome (1-22, X, Y or 0 if unplaced)
rs# or snp identifier
Genetic distance (morgans)
Base-pair position (bp units)

>**nrow(simmap[which(siimmap$V1==1),])**

[1] 501

1.5. Read in the phenotype file:

> **simpheno<-read.table("GWAS4pheno.txt", header=T, sep="\t")**
> **dim(simpheno)**
[1] 253 3
> **head(simpheno)**
FID IID Aff
1 1328 NA06989 1
2 1377 NA11891 2
3 1349 NA11843 1
4 1330 NA12341 2
5 1444 NA12739 1
6 1328 NA06984 2

Let's see how many cases and controls are in the dataset.
> **table(simpheno$Aff)**
1 2
162 91

Questions:
1. Why are there 12,782 columns in the PEDFILE, which consists of 6,388 SNP genotypes?

2: How many SNPs are located on chromosome 3 in the MAPFILE?

3: How many individuals have Affected status in the PHENOFILE?

For the rest of the exercise, we will be using both PLINK and R, so it will be convenient to open another CAC session in a separate tab or use RStudio.

We will refer to this as the **PLINK window** and the previous window as the **R window**.

2. GWAS Time!

Now we're going to do a simple GWAS. In truth this isn't a GWAS because the genotype data is not genome-wide, but the mechanics of the analysis are the same.

In the PLINK window, type the following at the UNIX prompt to launch PLINK:

$ module load plink

Use the following PLINK command to perform a basic association test comparing allele frequencies between cases and controls for the 6388 SNPs in our sample. In the PLINK window, type:

$ plink --file GWAS4 --pheno GWAS4pheno.txt --assoc --out GWAS1

In the R window. . .
Read in the results file
Look at the first five rows of results
Order the results by p-value

```
> res1 <- read.table("GWAS1.assoc", header = T, stringsAsFactors = F)
> res1[1:5, ]
> res1 <- res1[order(res1$P), ]
> res1[1:5, ]
> head(res1)
```

The p-values are very low!

```
> res1b<-res1[res1$P<=0.05,]
```

Questions:

4: How many SNPs give a P value of <= 0.05?

5: Why do you think there so many SNPs with such low P values?

6: Derive the Bonferroni-adjusted p-values by correcting for 6388 independent statistical tests. How many SNPs have a Bonferroni P value <=0.05? Please list the SNP IDs (rs#).

3. Visualization of Results

3.1 Simple Manhattan Plot

GWAS results are often displayed in a Manhattan plot, which depicts the strength of association (-log10 transformed P values) at each SNP across the genome. One of the nice things about R is

that it provides an easy way to utilize code written by others. In this case, we are installing a R package from the CRAN website that consists of all the source code for making a manhattan plot:

```
> install.packages("qqman", repos = "http://cran.utstat.utoronto.ca/")
> library("qqman")

> forplot <- res1[, c("SNP", "CHR", "BP", "P")]
> pdf("GWAS1_Manhattan.pdf")
> manhattan(forplot)
> dev.off()
```

If using R Studio, you should be able to view this pdf in the bottom right pane.

If using linux R, you can switch into the Plink window and type the following to view the pdf:

```
$ evince GWAS1_Manhattan.pdf
```

Question 7: For your manhattan plot, are there any SNPs that are statistically significant? If yes, how many SNPs and on which chromosomes are these located?

3.2 Quantile-Quantile (Q-Q) Plot

GWAS results are also displayed in a Q-Q plot, which is a plot of the observed P value distribution compared to the expected (null) distribution. To create a Q-Q plot of the GWAS results, type:

```
> pdf("GWAS1_QQ.pdf")
> qq(res1$P)
> dev.off()
```

If using R Studio, you should be able to view this pdf in the bottom right pane.

If using linux R, you can switch into the Plink window and type the following to view the pdf:

```
$ evince GWAS1_QQ.pdf
```

Question 8: Based on the Q-Q plot, how many SNPs deviate from the null distribution with –log10(P)> 8? Please submit your QQ plot with the assignment.