Julie Garcia
February 20, 2018

HW 2
Advanced Genomics and Genetics Analyses

In this assignment, you will be working with a genome-wide association study (GWAS). The disease indications are bipolar disorder, depression, and schizophrenia. This is a rather modest study with only 153 total subjects and 443,816 SNPs using the Affymetrix 5.0 SNP array. For each subject, we have collected a set of highly useful clinical and demographic variables. This information will be utilized to strengthen any association that we identify. Note that in the pedigree file, both gender and affected status has not been designated since we will be using the clinical table to do this.

1.) Go to the course website and download the HW #2 data file.

2.) We will first want to filter out SNPs and subjects that have issues of missingness and other factors. Use the following criteria below to filter the ped and map files:

   a. MAF>10%
   b. Missing rate per SNP < 10%
   c. Missing rate per subject < 30%
   d. HW significance < 0.001

   Report the number of both SNPs and subjects remaining. What is the reasoning for using such a high minor allele frequency threshold?

   I ran plink from the command line with the following command line options and exported the new files with filtered data.

```
plink --noweb --file subjects_153 --geno 0.1 --maf 0.1 --mind 0.3
--hwe 0.001 --recode
```

   We started with 443,816 SNPs and 153 subjects and, after frequency and genotyping pruning, ended with 282,157 SNPs. After filtering, 153 subjects remained. A high MAF threshold is chosen because the sample size is small, at least 16 of the 153 individuals will need to have the minor allele. This should reduce the number false positives.

3.) Now we need to assess outliers in the data or see if there are any major population stratification issues. Run MDS using the newly created ped/map files and plot the first 2 eigenvectors. Color the samples based on the disease indication that can be extracted from the "Profile" column. Looking at this plot and the various variables in the clinical table, is there any obvious pattern of either outlier subjects or population stratification? You can do this one of two ways: 1) replot the MDS graph with each plot having the samples colored by the different levels of the sample annotation table, or 2) running a t-test or ANOVA for each column

variable using the first eigenvector from the MDS plot as the continuous variable and the different levels of the column variable as the grouping variable. If you identify a variable that seems to separate the samples fairly well, provide the MDS coloring by the factor, if not, no additional plot is required.

I ran the following command in plink to get the MDS eigenvectors:

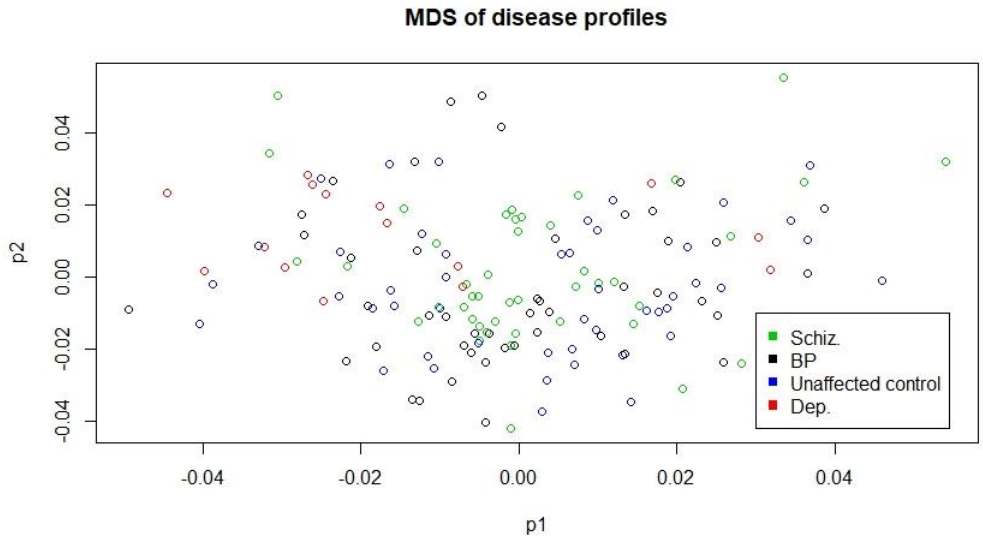
```
plink --noweb --file plink --genome
plink --noweb --file plink --read-genome plink.genome --cluster
--mds-plot 2
```

In R:

```
#read in mds data from plink
mds = read.table("C:\\TEMP\\datasets\\hw2\\plink.mds", header=T)
colnames(mds)

#read in clinical data
data.clinical <-
read.delim("C:\\temp\\datasets\\hw2\\clinical_table.txt",
header=T)
diseases <- data.clinical,"Profile"]

#plot mds with disease profiles
plot.df <- data.frame(pc1=mds$C1, pc2=mds$C2)
plot(plot.df, col=as.numeric(diseases), xlab="p1",
     ylab="p2", main="MDS of disease profiles")
legend(0.075, -0.1,
unique(diseases),pch=15,col=unique(as.numeric(diseases)))
```



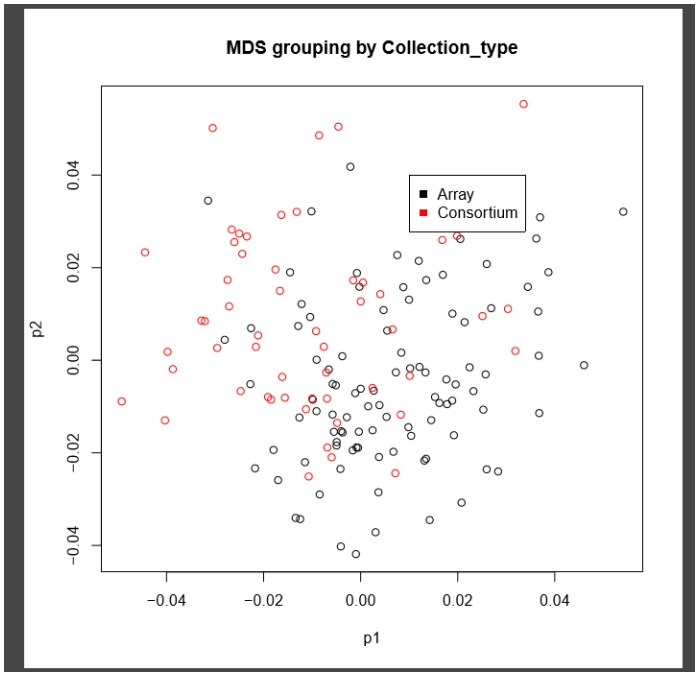
In this plot I do not see any obvious population stratification.

After manually replotting MDS by some of the other factors, I decided to write a loop to create MDS plots for all other factors and write them to a pdf file. I attached the pdf file to my submission. I did not see any obvious patterns of population stratification. The legends are a bit off in these plots, but if I would have found an obvious grouping, I would have recreated the plots to look nice.

```
# Loop to run MDS on each parameter and write to pdf
pdf("MDS_Plots.pdf")
for(factor in colnames(data.clinical)) {
       groups <- data.clinical[,factor]

       plot(plot.df, col=as.numeric(groups), xlab="p1",
                  ylab="p2", main=paste("MDS grouping by",
                  factor))
       legend(0.01, 0.04, unique(groups), pch=15,
            col=unique(as.numeric(groups)))
}
dev.off()
```

The only category that looked like it might show a slight grouping was collection type, shown in the plot below. Consortium looks like it trends to the upper left and Array to the lower right, but I wouldn't consider this significant population stratification.



4.) Now we would like to run an association analysis on this dataset using linear regression. We would first like to identify SNPs associated with bipolar disorder, so to do this, we need to compare the bipolar subjects to the control subjects. We would also like to include the following covariates: Primary eigenvector from the MDS calculation, Gender, and Left Brain (fixed/frozen). The disease status (Profile column) should be output in a separate file. In R, you will need to create one file for the appropriate individuals to include (use ***--keep*** argument in Plink), another file for the covariates to include (use ***–covar*** argument in Plink), and a third file for the affected status (use ***–pheno*** argument in Plink). Note that there are no hyphens in the Family IDs in the pedigree file. Also remember to code binary categorical variables in the covariate file to 0/1 and affected status as 1/2. Keep continuous variables as they are.

In R to create the appropriate files:

```
# --keep file
data.clinical$Profile
bip_control <- data.clinical[data.clinical$Profile=="BP" |
              data.clinical$Profile=="Unnaffected
                  control",]$Database_ID
bip_control <- cbind(as.character(bip_control),
as.character(mds[mds$FID %in% bip_control,]$IID))
write.table(bip_control,"bipolar_control.txt",sep="
      ",row.names=F,col.names=F,quote=F)

#--pheno file (1=unaffected, 2=affected)
pheno <- cbind(as.character(mds$FID),mds$IID)
temp <- data.clinical
temp$Profile
temp$Profile <- as.integer(temp$Profile)
temp[!temp$Profile==4,] <- 2
temp[temp$Profile==4,] <- 1
pheno <- cbind(pheno, temp$Profile)
pheno
write.table(pheno,"pheno.txt",sep=" ",row.names=F, col.names=F,
      quote=F)
```

In plink, I ran the linear regression with the covariates stated above, only on the bipolar vs controls. Plink code:

```
plink –noweb --ped plink.ped --map plink.map –linear --keep
bipolar_control.txt --covar mycov.txt --pheno pheno.txt
```

5.) Now extract the covariate summary statistics that correspond to the disease status in the output file (ADD coefficient) from Plink and identify the 1.) number of SNPs with p<.01 and, 2.) most prevalent chromosome among the top 100 SNPs.

I read the results back into R:

```
#read in assoc
bip_summary <-
read.table("C:\\TEMP\\datasets\\hw2\\plink.assoc.logistic.bip",
header=T)
plessthan01.bip <- bip_summary[which(bip_summary$TEST=="ADD" &
      bip_summary$P<.01),]
plessthan01.bip
```

As of right now, I am seeing p-values of 1 in every column.

6.) Repeat the analysis for the comparison between Schizophrenia and Control subjects.

```
# --keep file for schiz
data.clinical$Profile
schz_control <- data.clinical[data.clinical$Profile=="Schiz." |
                 data.clinical$Profile=="Unnaffected
                  control",]$Database_ID
schz_control <- cbind(as.character(schz_control),
as.character(mds[mds$FID %in% schz_control,]$IID))
schz_control
write.table(schz_control,"schz_control.txt",sep="
            ",row.names=F,col.names=F,quote=F)
```

In plink, I ran the linear regression with the covariates stated above, only on the bipolar vs controls. Plink code:

```
plink –-noweb --ped plink.ped --map plink.map –-linear --keep
schiz_control.txt --covar mycov.txt --pheno pheno.txt

schiz_summary <-
read.table("C:\\TEMP\\datasets\\hw2\\plink.assoc.logistic.schiz",
      header=T)
plessthan01.schiz <-
schiz_summary[which(schiz_summary$TEST=="ADD" &
      schiz_summary$P<.01),]
plessthan01.schiz
```

As of right now, I am seeing p-values of 1 in every column.