Julie Garcia
January 29, 2018

Lab 3
Advanced Genomics and Genetics Analyses

This lab introduces concepts for analyzing case-control whole-genome association studies. Similar to the previous lab, we will be implementing more than a single software tool in this lab, to better understand both the tools and the application to whole-genome association studies. This data set includes 43 pediatric patients that are considered controls for the outcome under study. We are comparing these control children to a much larger cohort of 400 control children. We would expect few differences between these 43 'cases' and 400 controls.

Like the beginning of any SNP analysis, we need to QC the data and conduct data cleansing, similar to what we have done previously. Then we will construct a vector to help control for population stratification using MDS and compare the output between case-control associations controlling for population stratification and not controlling for this.

1.) First, get the case-control_SNP_files.zip file from the website and unzip it. We will be working with the ped and map files from this zip. Isolate the nr_no_443.ped pedigree file and the snps_1809.map map file.

2.) Run analysis to filter out SNPs based on the following criteria
    a. <5% missingness rate per SNP
    b. MAF > 5%
    c. <10% missingness rate per subject
    d. HWE significance at p<.001

    Make sure to create a new pedigree and map file from this analysis. How many SNPs remain? How many subjects remain?

    I renamed the files to snps_lab3.ped and snps_lab3.map.

```
plink --noweb --file snps_lab3 --geno 0.05 --maf 0.05 --mind 0.2
--hwe 0.001 --recode
```

    After filtering, 1668 SNPs remain, and 42 subjects and 400 controls remain.

3.) Using this newly filtered ped and map file, calculate a simple association test using the Fisher's exact test method. Plot the results in R in a $-\log_{10}$(p-value) plot similar to how you have before to visualize the major significant SNPs that show association. How many SNPs are significant at p<.01 and p<.05?

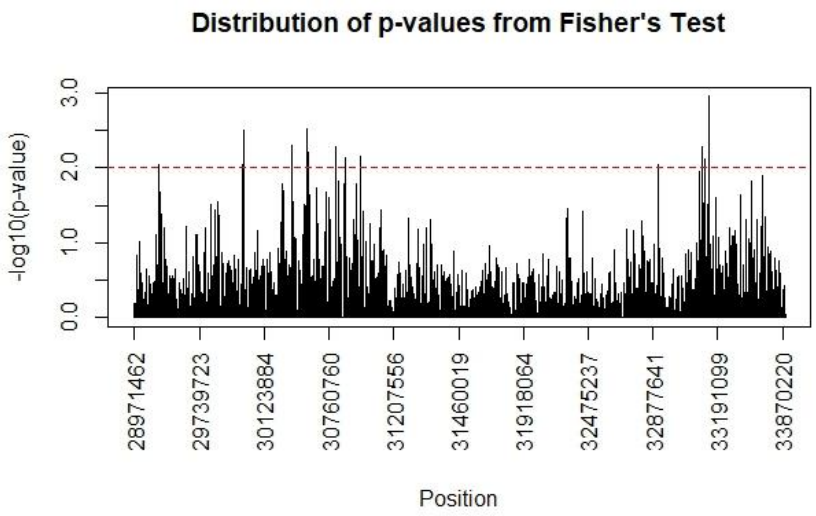
    In plink:

```
plink --noweb --ped plink.ped --map plink.map –fisher –ci 0.95
```

In R:

```
#read in Fisher's test results from plink output
data = read.table("C:\\TEMP\\datasets\\plink.assoc.fisher",
header=T)
dim(data)
data[1:4,]
colnames(data)


# plot
par(mar=c(8,5,5,5))
plot(-log10(data$P), type="n",
     xaxt="n", xlab="", ylab="-log10(p-value)",
     main="Distribution of p-values from Fisher's Test",
     col = "black")
xtick<-seq(1, 1668, by=166)
axis(side=1,at=xtick,labels=data$BP[xtick], las=2)
lines(-log10(data$P),
      type = "h", col = "black")
abline(2.0,0,col="red",lty="dashed")
mtext("Position", side=1, line=6)
```



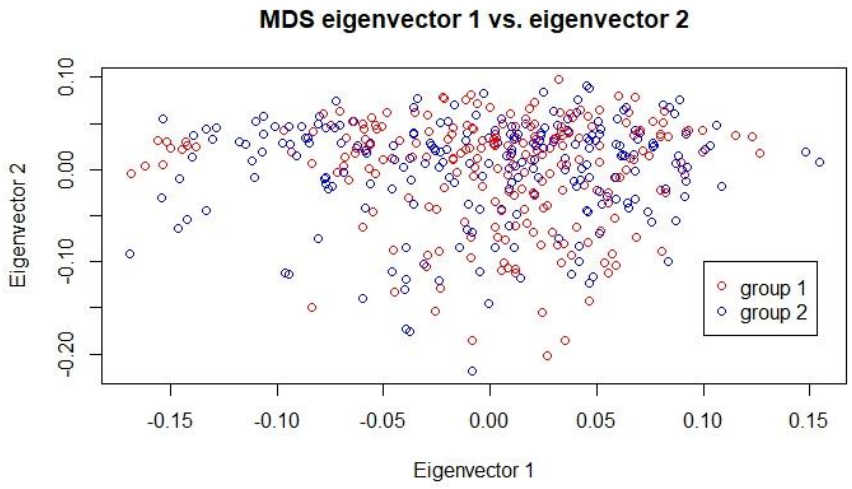

```
plessthan01 <- data[data$P < 0.01,]
dim(plessthan01)
plessthan05 <- data[data$P < 0.05,]
dim(plessthan05)
```

There are 13 with p-values less than .01 and 60 with p-values less than .05.

4.) Now calculate MDS on this pedigree file in Plink and plot the first 2 eigenvectors against each other in R. Make sure to color the 2 groups different colors and add a legend.

```
plink --noweb --ped plink.ped --map plink.map --genome
plink --noweb --ped plink.ped --map plink.map --read-genome
plink.genome --cluster --mds-plot 2

#read MDS results from plink and plot
mds = read.table("C:\\TEMP\\datasets\\plink.mds", header=T)
dim(mds)
colnames(mds)
plot.df <- data.frame(pc1=mds$C1, pc2=mds$C2)
plot(plot.df, col=c(2,4), xlab="Eigenvector 1",
     ylab="Eigenvector 2", main="MDS eigenvector 1 vs.
eigenvector 2")
legend(0.1, -0.1, c("group 1", "group 2"), col = c(2,4),pch =
c(1,1))
```



5.) Write these first 2 eigenvectors out in a covariate file using the correct Plink format (Family ID, Individual ID, covariate 1, covariate 2) and run the linear regression association test in Plink. In R, plot the same -$\log_{10}$ plot as before. Make sure to only extract that ADD column since this indicates the coefficient in the model that we are testing for. How many SNPs are significant at p<.01 and p<.05?

In R:

```
mycov <- mds[,c(1,2,4,5)]
write.table(mycov,file="C:\\TEMP\\datasets\\mycov.txt",
row.names=FALSE)
```

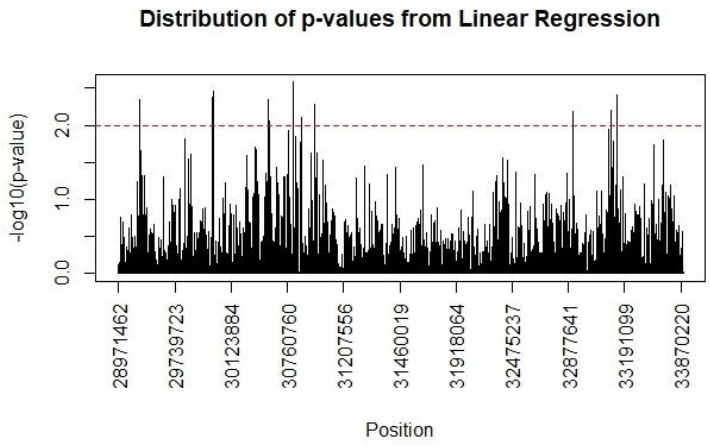
In plink:

```
plink --noweb --ped plink.ped --map plink.map --linear --covar
mycov.txt
```

In R:

```
# read in covar file
covar = read.table("C:\\TEMP\\datasets\\plink.assoc.logistic",
header=T)
dim(covar)
colnames(covar)
covar.add <- covar[covar$TEST=="ADD",]
dim(covar.add)

# plot
par(mar=c(8,5,5,5))
plot(-log10(covar.add$P), type="n",
     xaxt="n", xlab="", ylab="-log10(p-value)",
     main="Distribution of p-values from Linear Regression",
     col = "black")
xtick<-seq(1, 1668, by=166)
axis(side=1,at=xtick,labels=covar.add$BP[xtick], las=2)
lines(-log10(covar.add$P),
      type = "h", col = "black")
abline(2.0,0,col="red",lty="dashed")
mtext("Position", side=1, line=6)
```




```
#results
plessthan01.covar <- covar.add[covar.add$P < 0.01,]
dim(plessthan01.covar)
plessthan05.covar <- covar.add[covar.add$P < 0.05,]
dim(plessthan05.covar)
```

There are now 12 with p-values less than .01 and 57 with p-values less than .05.