Julie Garcia
January 23, 2018

Lab 2
Advanced Genomics and Genetics Analyses

This lab introduces concepts for analyzing family-based SNP data. We will be implementing multiple software tools in this lab, to better understand both the tools and the application to family-based SNP data. First, we will begin with some important data cleansing concepts such as SNP filtering based on HWE criterion, missingness, and MAFs. Then we will run an IBD analysis for verification of parental genotypes. This is important because sometimes the parental information is not from a biological parent, so this subject needs to be removed from the analysis since this person did not transmit the genotype information to the offspring under study. We will then calculate a TDT and plot the results.

1.) First, get the family_SNP_files.zip file from the website and unzip it. We will be working with the ped and map files from this zip.

2.) Run analysis to filter out SNPs based on the following criteria
   a. <5% missingness rate per SNP
   b. MAF > 10%
   c. <20% 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 ran plink from the command line with the following commands and exported the new files with filtered data.

```
plink –noweb –file ped_final_re_MHC_mod –geno 0.05 –-maf 0.1 –
mind 0.2 –hwe 0.001 --recode
```

   Before filtering there were 201 subjects with 21813 SNPs. After filtering, 194 (30 case and 164 control) subjects and 1422 SNPs remain.

3.) Using this newly filtered ped and map file, run IBD to identify any inconsistencies in parental genotypes using a p<.01 threshold. **How many paternal or maternal subjects have a p<.01 when compared to their offspring? How many paternal or maternal subjects have p<.01 compared to each other?** We would expect paternal and maternal subjects to not be 'genetically related' and thus have a significant p-value, but for many paternal and maternal subjects, this is not the case, **why do you think this is not occurring using this data set?**

   I ran the following command in plink to get the IBD results.

```
plink –noweb –file plink –genome –rel-check
```

The results showed no paternal or maternal subjects that had a p-value of less than .01. When compared to each other there were 4 paternal and maternal p-values < .01, which means only those 4 sets of parents are significantly unrelated. The other parents don't seem to be very closely related, but genetically related enough that it may have come from a community where inbreeding is common.

4.) For those parents that differ significantly from their child (from above, **if any**), remove them (using the –-remove flag and instructions on either the plink website or the lecture notes) and run the TDT with output of 95% confidence intervals.

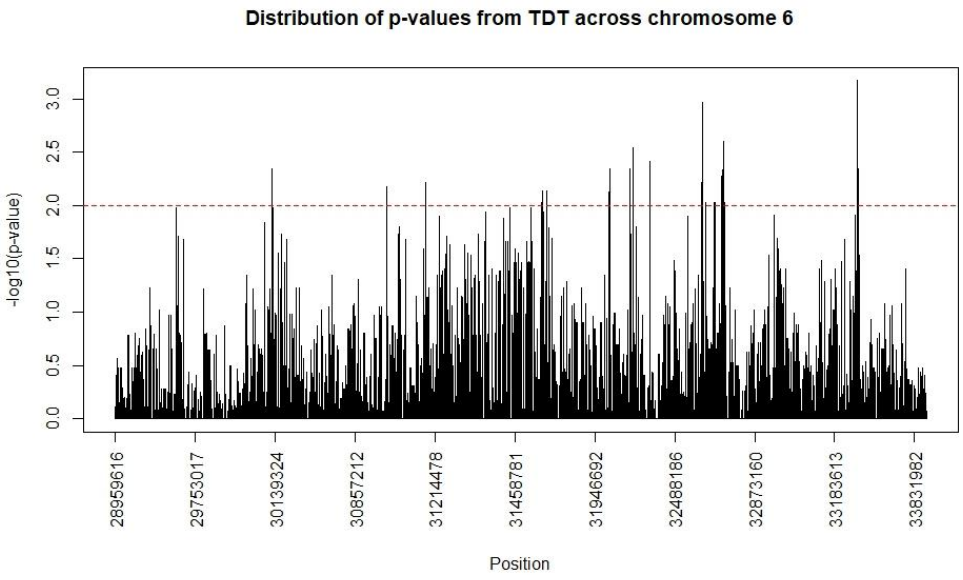
Since I didn't find any parents that differ significantly from their child, I skipped the first step. Then I ran the TDT in plink:

```
plink –noweb –-file plink –-tdt –-perm –-ci 0.95
```

5.) Now we want to visualize the regions along chromosome 6 where there are significant p-values representing significant transmission of alleles from parent to offspring. Read in the TDT results and create the plot below applying your knowledge of the plot function in R. Be sure to add the axes after you plot the lines and plot structure. You may also have to use a for loop to add the lines to the plot. **How many SNPs have p<.05?** Load the qvalue R package and compute false discovery rate adjusted p-values using fdr.level=0.05. **How many SNPs have FDR adjusted p-values of p<.05? What does this latter result indicate to you?**

```
# code to generate the plot
par(mar=c(8,5,5,5))
plot(-log10(tdt.chr6$P), type="n",
     xaxt="n", xlab="", ylab="-log10(p-value)",
     main="Distribution of p-values from TDT across chromosome 6",
     col = "black")
xtick<-seq(1, 1422, by=140)
axis(side=1,at=xtick,labels=tdt.chr6$BP[xtick], las=2)
lines(-log10(tdt.chr6$P),
        type = "h", col = "black")
abline(2.0,0,col="red",lty="dashed")
mtext("Position", side=1, line=6)
```

Here is my plot, it took a bit of coercing to get margins and axes correct:

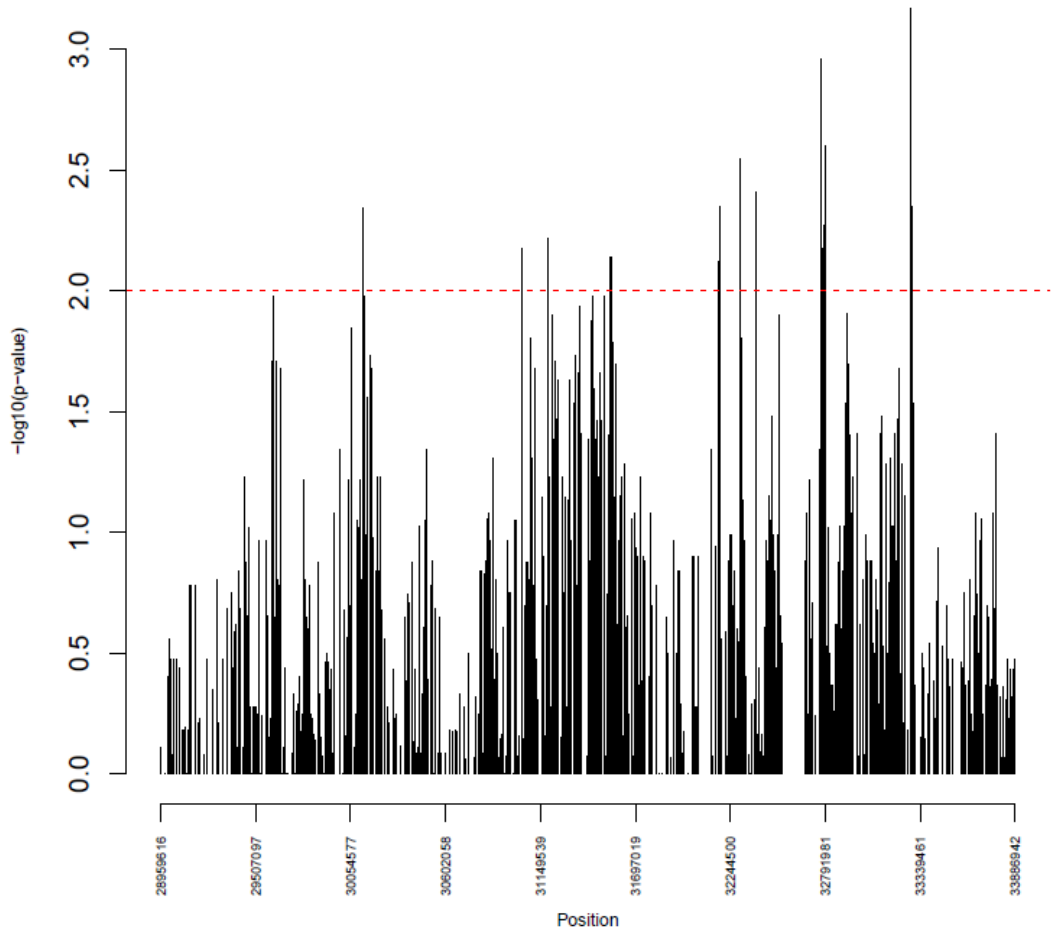



```
plessthan05 <- tdt.chr6[tdt.chr6$P < 0.05,]
dim(plessthan05)
```

121 SNPs have p<.05.

```
source("https://bioconductor.org/biocLite.R")
biocLite("qvalue")
browseVignettes("qvalue")
library("qvalue")
qobj <- qvalue(p = tdt.chr6$P, fdr.level=0.05)
qlessthan05 <- qobj$qvalues[qobj$qvalues < 0.05]
```

0 adjusted p-values < 0.05. Setting a p-value of 0.05 means that we are accepting a 5% rate of false positives. This qvalue result indicates that the original p-value calculation had a number of false positives and that there are no truly significant SNPs in this dataset.


Distribution of p−values from TDT across chromosome 6
−log10(p−value)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
28959616
29507097
30054577
30602058
31149539
31697019
32244500
32791981
33339461
33886942
Position