Julie Garcia
October 9, 2017

Lab #5
Differential expression

In this lab, we will be conducting a two-sample test for each gene/probe on the array to identify differentially expressed genes/probes between ketogenic rats and control diet rats. This small data set was run on the rat RAE230A Affymetrix array. The objective of the study was to determine differences in mRNA levels between brain hippocampi of animals fed a ketogenic diet (KD) and animals fed a control diet. "KD is an anticonvulsant treatment used to manage medically intractable epilepsies", so differences between the 2 groups of rats can provide biological insight into the genes that are regulated due to the treatment.

We are going to identify those genes/probes that are differentially expressed between the 2 rat diet groups and plot the results with a couple of different visual summaries.

1.) Download the GEO rat ketogenic brain data set and save as a text file.

2.) Load into R, using read.table() function and header=T/row.names=1 arguments.

```
data = read.table("c:\\temp\\rat_KD\\rat_KD.txt", header=T,
row.names=1)
```

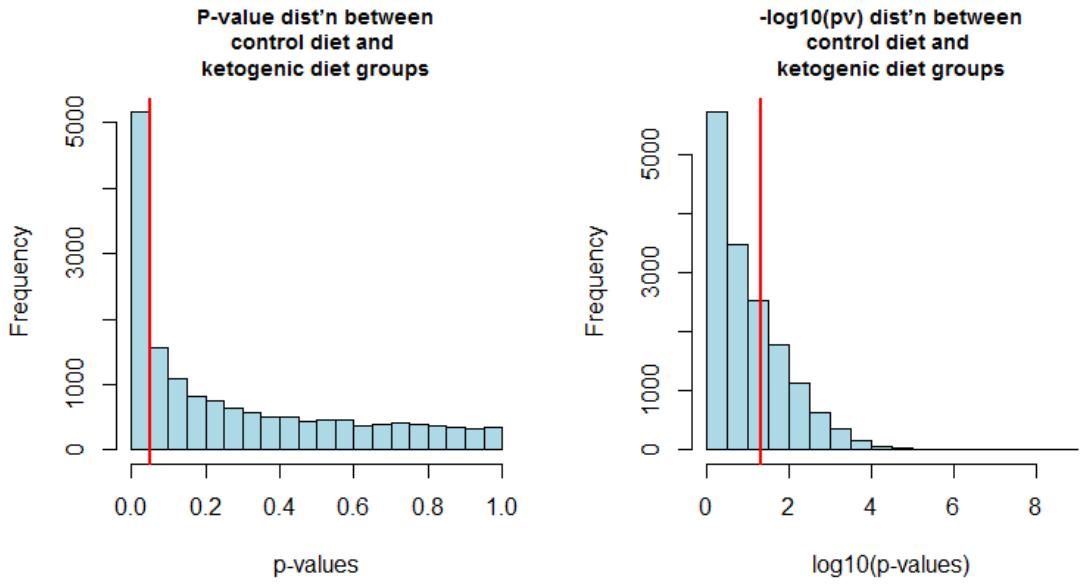
3.) First $\log_2$ the data, then use the Student's t-test function in the notes to calculate the changing genes between the control diet and ketogenic diet classes. (Hint: use the names() function to determine where one class ends and the other begins).

```
t.test.all.genes <- function(x,s1,s2) {
  x1 <- x[s1]
  x2 <- x[s2]
  x1 <- as.numeric(x1)
  x2 <- as.numeric(x2)
  t.out <- t.test(x1,x2, alternative="two.sided",var.equal=T)
  out <- as.numeric(t.out$p.value)
  return(out)
}

names(data.log2)
dim(data.log2)
gc <- 1:6
act <- 7:11
pv <- apply(data.log2,1,t.test.all.genes,s1=gc,s2=act)
```

4.) Plot a histogram of the p-values and report how many probesets have a p<.05 and p<.01. Then divide an alpha of 0.05 by the total number of probesets and report how many probesets have a p-value less than this value. This is a very conservative p-value thresholding method to account for multiple testing called the Bonferroni correction that we will discuss in upcoming lectures.

```
par(mfrow=c(1,2))
hist(pv,col="lightblue",xlab="p-values",
     main="P-value dist'n between\ncontrol diet and \nketogenic
diet groups",
     cex.main=0.9)
abline(v=.05,col=2,lwd=2)
hist(-log10(pv),col="lightblue",xlab="log10(p-values)",
     main="-log10(pv) dist'n between\ncontrol diet and
\nketogenic diet groups",cex.main=0.9)
abline(v=-log10(.05),col=2,lwd=2)
```




```
length(pv)
[1] 15923

pv.lessthan05 <- pv[pv < .05]
length(pv.lessthan05)
[1] 5160

pv.lessthan01 <- pv[pv < .01]
length(pv.lessthan01)
[1] 2414
```

```
alpha <- 0.5/dim(data.log2)[1]
alpha
pv.lessthanalpha <- pv[pv < alpha]
length(pv.lessthanalpha)
[1] 52
```

So probesets with a p-value less than .05 are 5160, less than .01 are 2414 and less than .05/(num probesets) are 52.

5.) Next calculate the mean for each gene, and calculate the fold change between the groups (control vs. ketogenic diet). Remember that you are on a $\log_2$ scale. (don't divide)

```
control.m <- apply(data.log2[,control],1,mean,na.rm=T)
keto.m <- apply(data.log2[,keto],1,mean,na.rm=T)
fold <- control.m-keto.m
```

6.) What is the maximum and minimum fold change value, please report on the linear scale? (retranspose) Now report the probesets with a p-value less than the Bonferroni threshold you used in question 4 **and** |fold change|>2. Remember that you are on a $\log_2$ scale for your fold change and I am looking for a linear |fold| of 2. (convert it)

```
fold.max <- 2^max(fold)
fold.max

[1] 55.15521

fold.min <- 2^min(fold)
fold.min

[1] 0.08240443

# find p-values less than threshold, and |fold change| > 2
pv.sig <- pv.lessthanalpha
fold <- 2^fold
fold.sig <- fold[abs(fold)>2]
significant <- intersect(names(fold.sig), names(pv.sig))
significant

[1] "1367553_x_at" "1387011_at"   "1387091_at"   "1370239_at"   "1370240_x_at"
[6] "1371102_x_at" "1371245_a_at" "1372053_at"   "1388608_x_at"
```

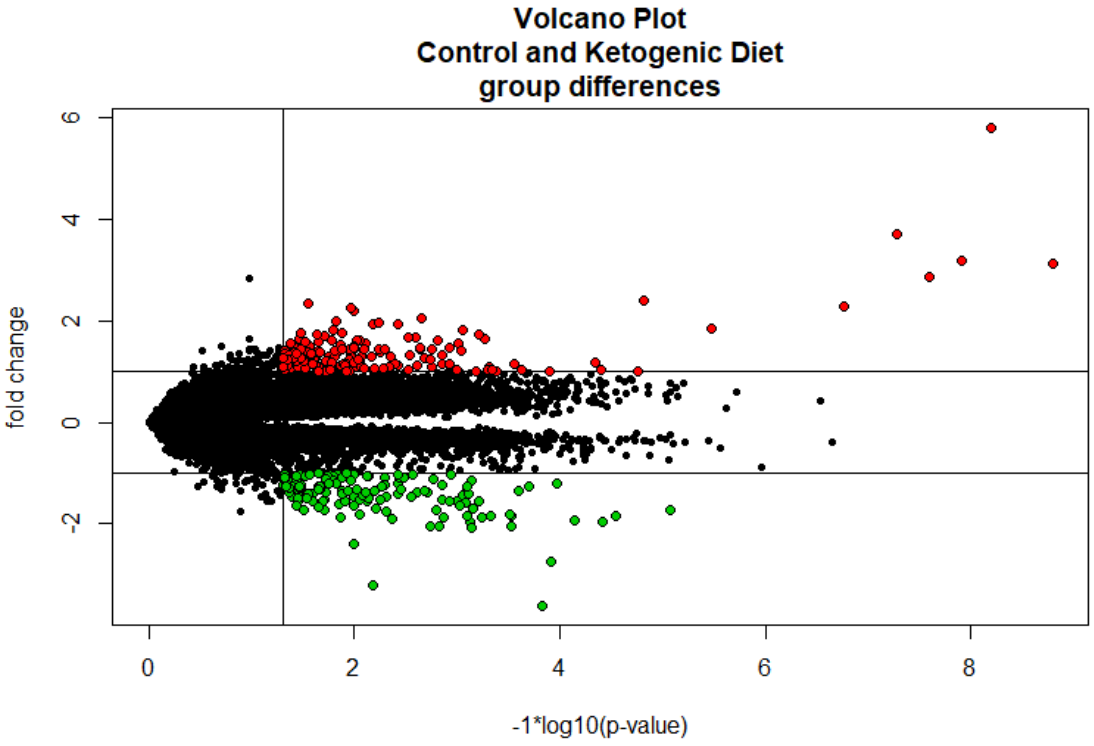
7.) Go to NetAffx or another database source if you like and identify gene information for the probesets that came up in #6. What is the general biological function that associates with these probesets?

I searched for the above probe sets in NetAffx and found that 5 out of the 9 are genes that express various domains of the

hemoglobin complex, which is the protein complex that carries out oxygen transport throughout the bloodstream. Another probe set was lipocalin 2 (LCN2), which involved in immune system processes and one was a peptidyl arginine deaminase (PADI2). After further investigation on DAVID and GeneCards using the gene names, I found that PADI2 has been associated with Alzheimer's disease and is involved in the pathways of the innate immune system and chromatin organization. LCN2 is involved in pathways that transport small molecules such as glucose, lipids, and steroids. Overexpression of LCN2 has been associated with kidney disease.

8.) Transform the p-value (-1*log10(p-value)) and create a volcano plot with the p-value and fold change vectors (see the lecture notes). Make sure to use a $\log_{10}$ transformation for the p-value and a $\log_2$ (R function log2()) transformation for the fold change. Draw the horizontal lines at fold values of 2 and -2 ($\log_2$ value=1) and the vertical p-value threshold line at p=.05 (remember that it is transformed in the plot).

```
# volcano plot pv.trans vs. log2(fold)
plot(range(pv.trans),range(fold),type='n',
     xlab='-1*log10(p-value)',ylab='fold change',
     main='Volcano Plot\nControl and Ketogenic Diet\ngroup
differences')
points(pv.trans,fold,col='black',pch=20,bg=1)
points(pv.trans[(pv.trans> -
log10(.05)&fold>log2(2))],fold[(pv.trans> -
log10(.05)&fold>log2(2))],col=1,bg=2,pch=21)
points(pv.trans[(pv.trans> -log10(.05)&fold< -
log2(2))],fold[(pv.trans> -log10(.05)&fold< -
log2(2))],col=1,bg=3,pch=21)
abline(v= -log10(.05))
abline(h= -log2(2))
abline(h=log2(2))
```


Volcano Plot
Control and Ketogenic Diet
group differences
fold change
6
4
2
0
-2
0
2
4
6
8
-1*log10(p-value)