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.

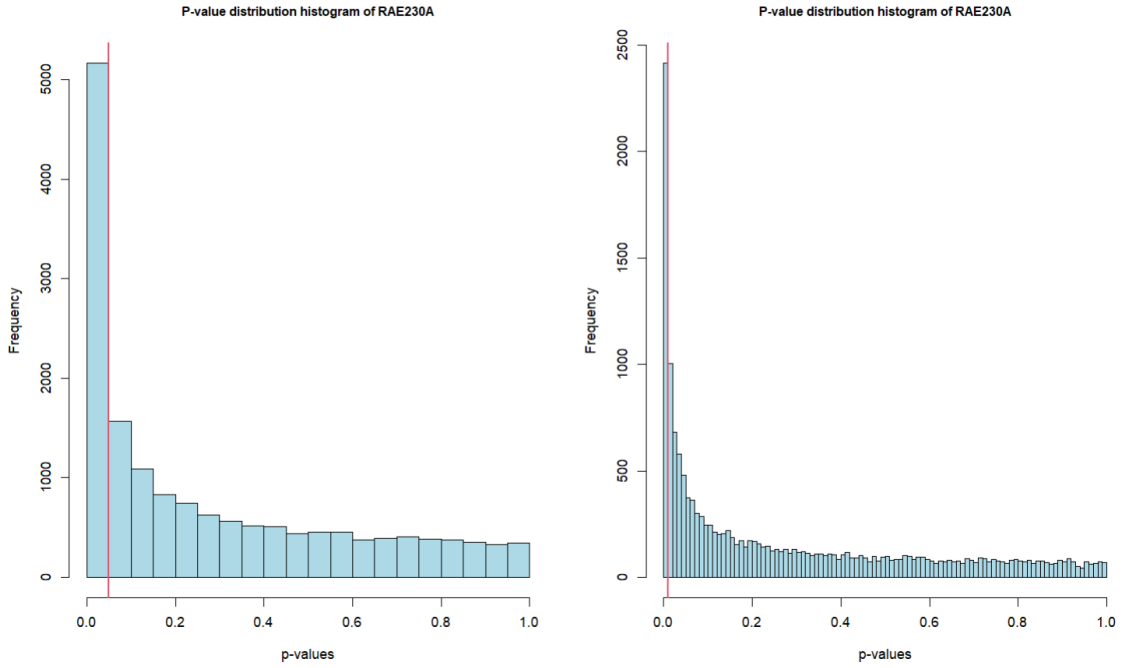
2.) Load into R, using read.table() function and header=T/row.names=1 arguments.

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).

```
library(dplyr)
library(Biobase); library(annotate); library(golubEsets);

rat_kd <- read.table('C:/Users/97481/OneDrive/JHU/gene expression/sourcedata/rat_KD.txt',header=T,row.names=1)
rat_kd <- log2(rat_kd)
names(rat_kd)
control <- as.character(names(rat_kd[,(1:6)]))
keto <- as.character(names(rat_kd[,(7:11)]))

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)
}


#t.test.all.genes(rat_kd,s1=c(1:6),s2=c(7:11))
pv <- apply(rat_kd,1,t.test.all.genes,s1=c(1:6),s2=c(7:11))
```

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 distribution histogram of RAE230A",cex.main=0.9,breaks =seq(0,1,0.05))
abline(v=.05,col=2,lwd=2)
hist(pv,col="lightblue",xlab="p-values", main="P-value distribution histogram of RAE230A",cex.main=0.9,breaks =seq(0,1,0.01) )
abline(v=.01,col=2,lwd=2)
```




P-value histogram with different breaks of p-value, red line in left histogram indicate P-value == .05, right indicate p_value==.01

```
> length(pv[pv<0.05])
[1] 5160
> length(pv[pv<0.01])
[1] 2414
```

```
> bf_p <- 0.05/length(pv)
> length(pv[pv<bf_p])
[1] 12
```

So,there are 5160 probesets have a p<.05 and 2414 p<.01

And there are 12 probesets has p lower than Bonferroni correction

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.

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

6.) What is the maximum and minimum fold change value, please report on the linear scale?

```
> max(fold)
[1] 5.785425
> min(fold)
[1] -3.601134
```

In linear scale, the maximum fold change value is 5.785425, the minimum is -3.601134

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.

```
fc_p<-data.frame(fold,pv)
prob_select <- filter(fc_p,abs(fold)>2&pv<bf_p)
```

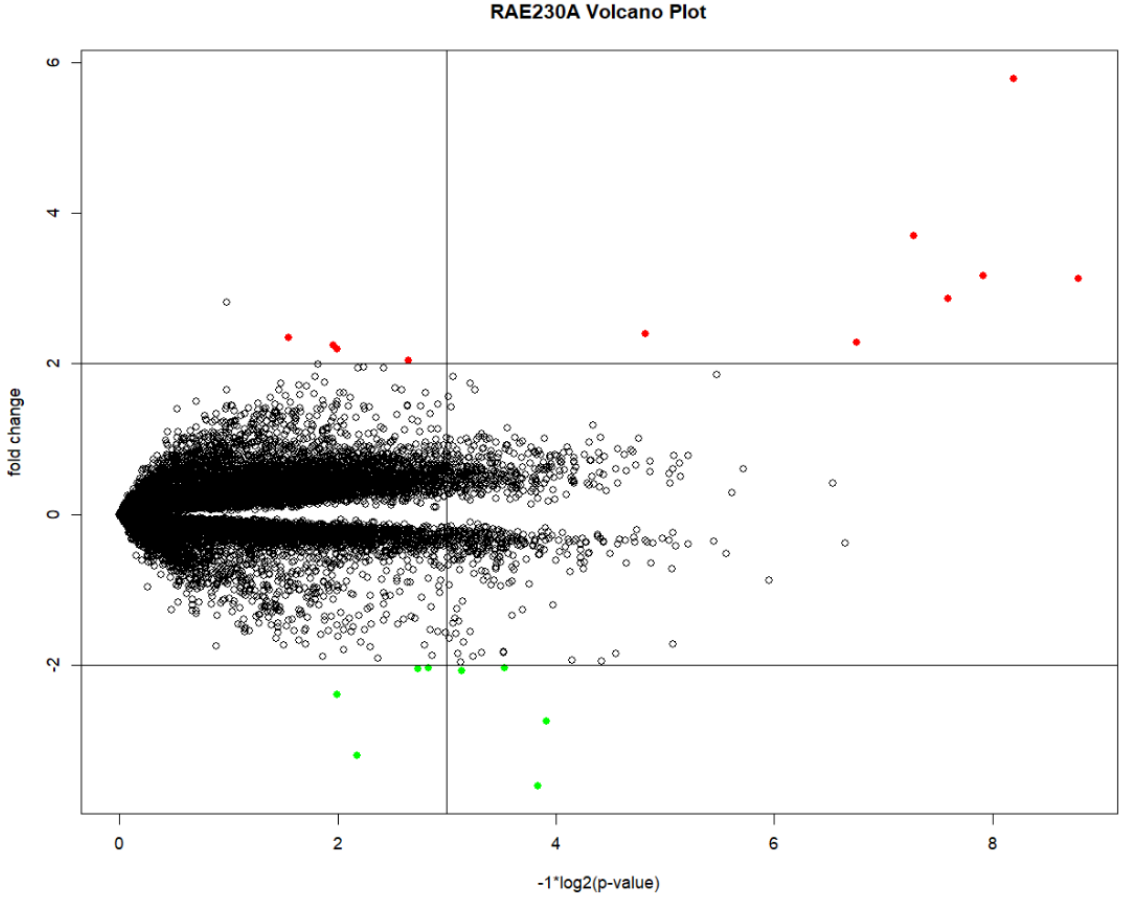
| | fold | pv |
|---|---|---|
| **1367553_x_at** | 3.171265 | 1.224053e-08 |
| **1370239_at** | 3.699730 | 5.280180e-08 |
| **1370240_x_at** | 3.132486 | 1.622293e-09 |
| **1371102_x_at** | 2.864893 | 2.583221e-08 |
| **1371245_a_at** | 5.785425 | 6.370531e-09 |
| **1388608_x_at** | 2.288613 | 1.743055e-07 |

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?

According to the search in the NetAffx, all the selected probsets in problem 5 are about hemoglobin.

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).

```
t.test.run <- apply(rat_kd,1,t.test.all.genes,s1=c(1:6),s2=c(7:11))
t.test.run[is.na(t.test.run)]<-1
p.trans <- -1 * log10(t.test.run)

plot(range(p.trans),range(fold),type='n',xlab='-1*log2(p-value)',ylab='fold change',main='RAE230A Volcano Plot')
points(p.trans,fold,col='black')
points(p.trans[(p.trans>-log2(0.5)&fold>2)],fold[(p.trans>-log2(0.5)&fold>2)],col='red',pch=16)
points(p.trans[(p.trans>-log2(0.5)&fold< -2)],fold[(p.trans>-log2(0.5)&fold< -2)],col='green',pch=16)
abline(v=3); abline(h=-2); abline(h=2);
```


RAE230A Volcano Plot
fold change
-1*log2(p-value)
6
4
2
0
-2
0
2
4
6
8