Differential Expression

Lavinia Carabet

Differential Expression

Load GEO rat ketogenic brain data

Differential gene expression between rats given a control diet and rats given a ketogenic diet (which prevents epileptic seizures)

(https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE1155)

```
dat <- read.table('./rat_KD.txt', header=T, row.names=1)
dim(dat)
```

```
## [1] 15923    11
```

```
dat[1:5,]
```

```
##                control.diet.19300 control.diet.19301 control.diet.19302
## AFFX-BioB-5_at            76.4424            86.0648            80.6466
## AFFX-BioB-M_at            94.7098            73.4476            88.6791
## AFFX-BioB-3_at            27.8835            44.5481            33.8824
## AFFX-BioC-5_at           174.3390           151.8240           167.4290
## AFFX-BioC-3_at            87.0285            94.0345           120.2830
##                control.diet.19303 control.diet.19304 control.diet.19305
## AFFX-BioB-5_at            93.8439            73.1219            97.6946
## AFFX-BioB-M_at           111.5530            92.1317            96.4250
## AFFX-BioB-3_at            60.0250            39.2463            37.6365
## AFFX-BioC-5_at           200.4780           170.7280           196.7960
## AFFX-BioC-3_at           114.5800           100.1920            88.3586
##                ketogenic.diet.19306 ketogenic.diet.19307 ketogenic.diet.19308
## AFFX-BioB-5_at              82.4622              77.2199             120.2040
## AFFX-BioB-M_at             131.2820             114.9030             156.7290
## AFFX-BioB-3_at              42.7933              50.0889              78.2358
## AFFX-BioC-5_at             192.0890             206.3390             236.0370
## AFFX-BioC-3_at             122.3530             130.9680             157.4380
##                ketogenic.diet.19309 ketogenic.diet.19310
## AFFX-BioB-5_at              98.9692              88.2618
## AFFX-BioB-M_at             117.2050             119.6470
## AFFX-BioB-3_at              47.8521              36.9666
## AFFX-BioC-5_at             202.8170             185.8010
## AFFX-BioC-3_at             110.3880             117.7130
```

```
#classes
control.diet <- names(dat[,grep('control', names(dat))])
control.diet
```

```
## [1] "control.diet.19300" "control.diet.19301" "control.diet.19302"
## [4] "control.diet.19303" "control.diet.19304" "control.diet.19305"
```

```
ketogenic.diet <- names(dat[, grep('ketogenic', names(dat))])
ketogenic.diet
```

```
## [1] "ketogenic.diet.19306" "ketogenic.diet.19307" "ketogenic.diet.19308"
## [4] "ketogenic.diet.19309" "ketogenic.diet.19310"
```

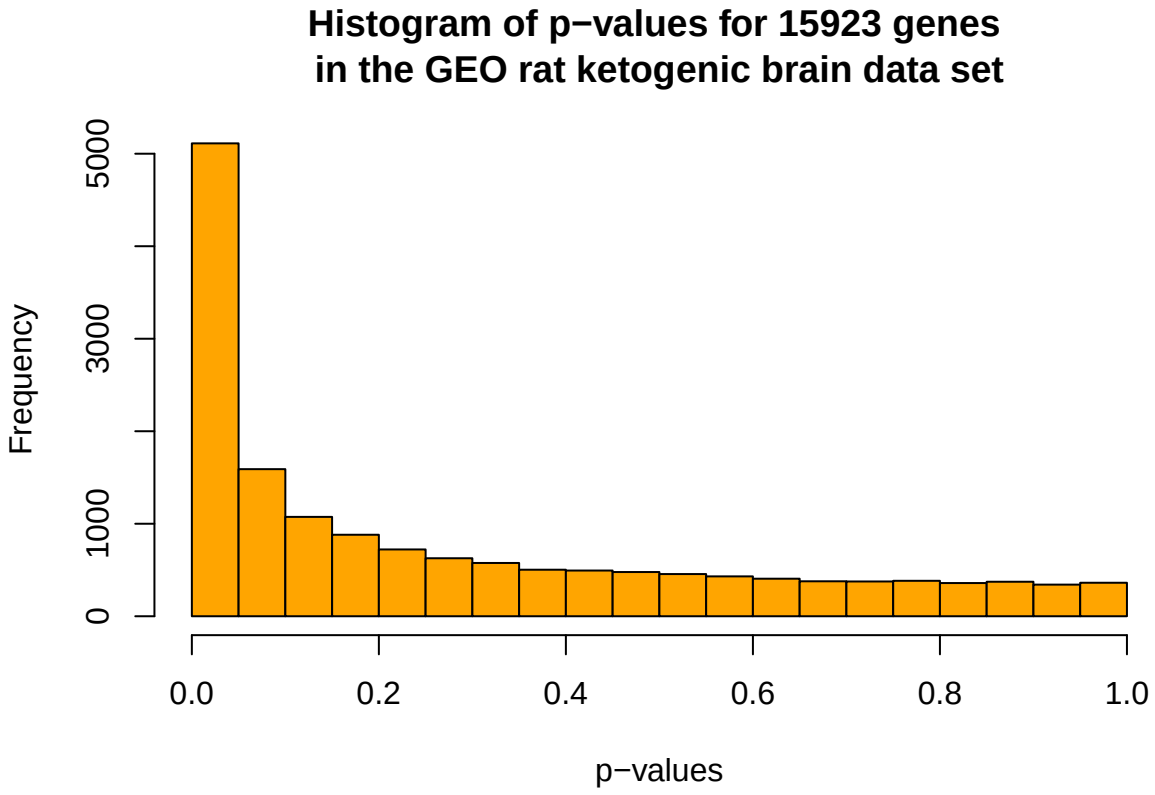
Calculate the changing genes between the control diet and ketogenic diet classes

Significance tests determine differential expression between means as a function of variance

```
# function to calculate Student's two-sample t-test on all genes at once
# function returns the p-value for the test
# NAs are removed for each test
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 = TRUE)
    out <- as.numeric(t.out$p.value)
    return(out)
}

t.test.run <- apply(dat,1,t.test.all.genes,s1=control.diet,s2=ketogenic.diet)
```

Plot a histogram of the p-values

```
xname <- "p-values"
hist(t.test.run, col='orange',
     main=paste("Histogram of" , xname, "for", nrow(dat),
                "genes \nin the GEO rat ketogenic brain data set"),
     xlab=xname)
```



Calculate fold change between the groups

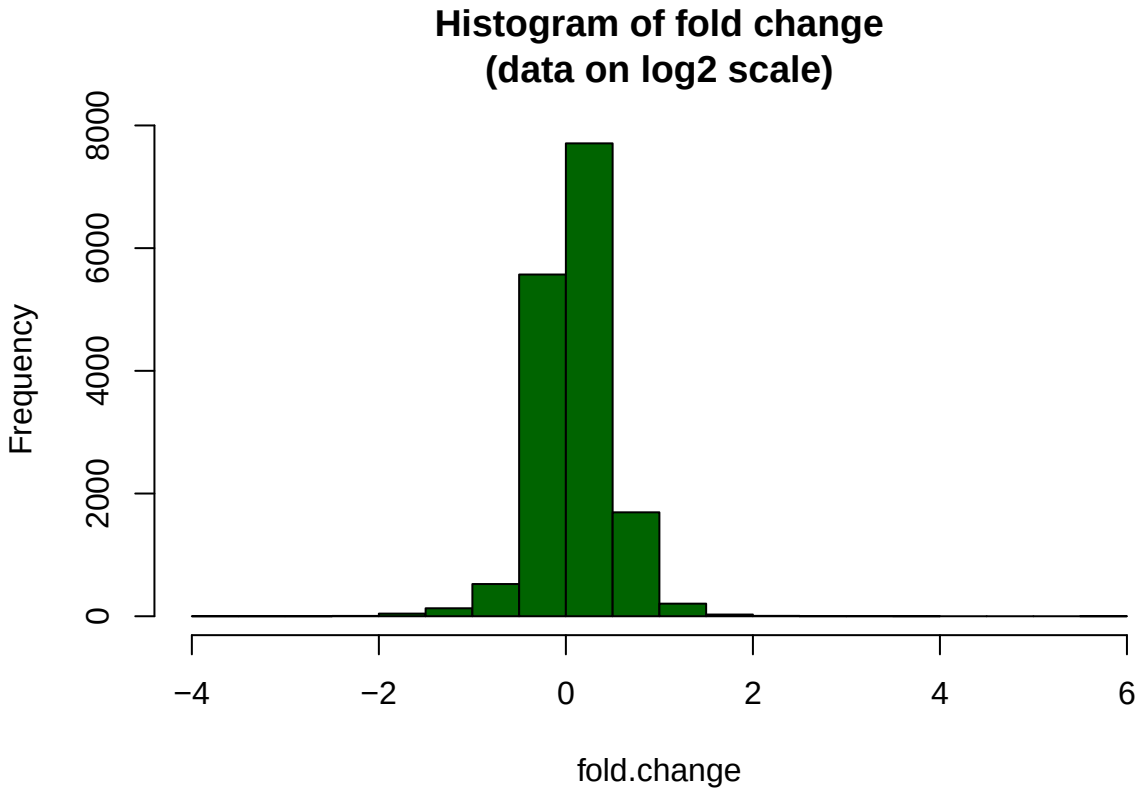
Fold change is a relative measure of the magnitude of difference between means

```
#calculate means of the groups
control.diet.mean <- apply(log2(dat[,control.diet]), 1, mean, na.rm = T)
ketogenic.diet.mean <- apply(log2(dat[,ketogenic.diet]), 1, mean, na.rm = T)

#calculate fold change
fold.change <- control.diet.mean - ketogenic.diet.mean
range(fold.change)
```

```
## [1] -3.601134  5.785425
```

Plot a histogram of the fold change values

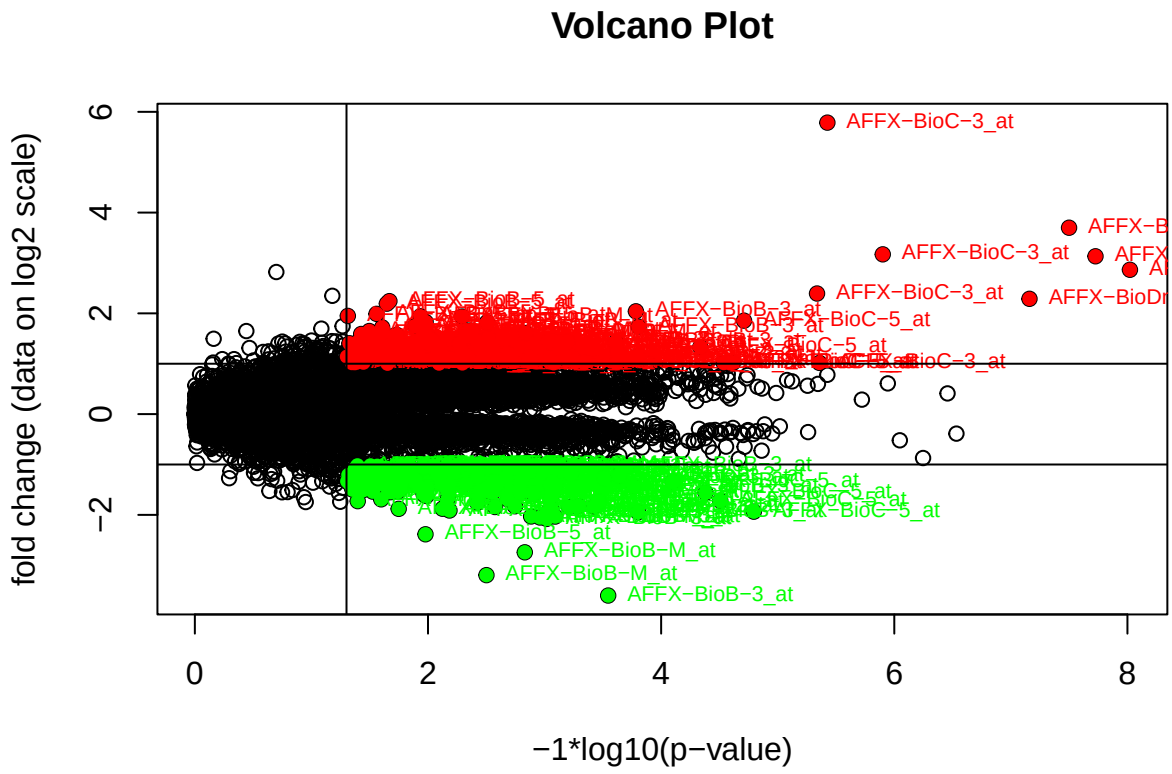
```
hist(fold.change, col='dark green', main=paste("Histogram of fold change\n(data on log2 scale)"))
```



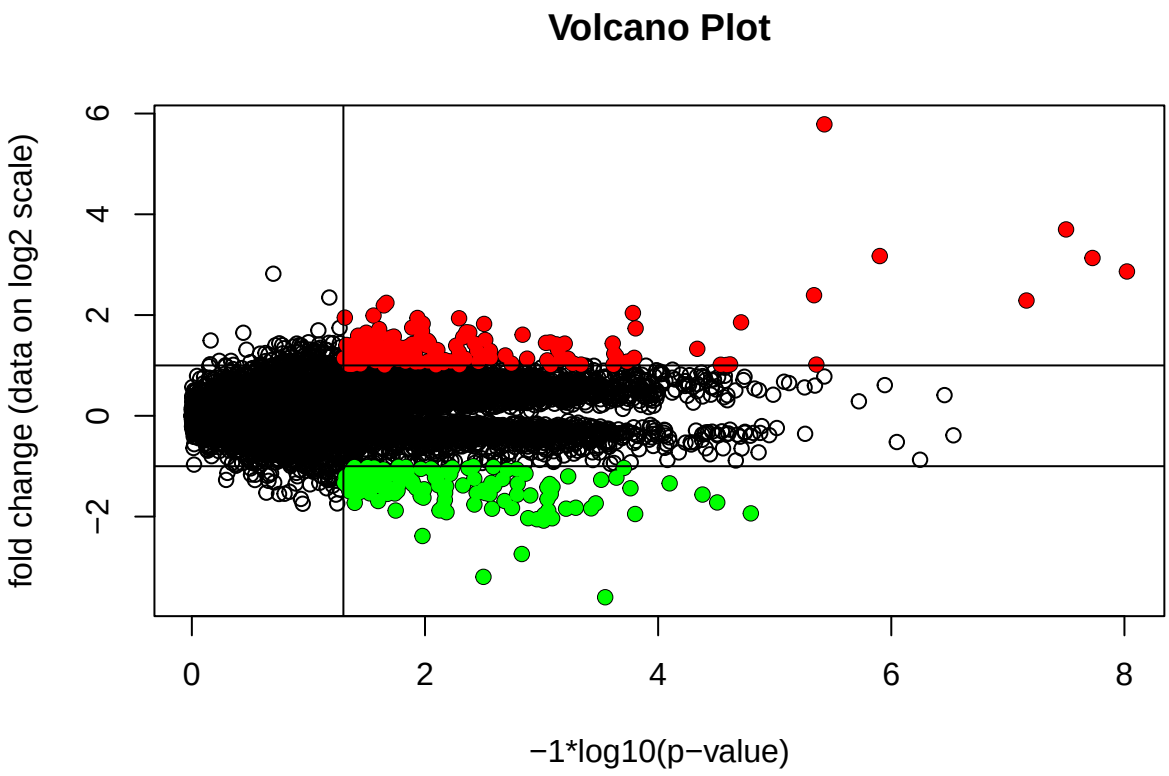
Volcano plot combining the fold.change and transformed p-values to determine the most significantly differentially expressed genes

```
p.trans <- -1 * log10(t.test.run)

x.line <- -log10(.05)    #p-value=0.05
y.line <- log2(2)        #fold change=2

plot(range(p.trans),range(fold.change),type='n',
     xlab='-1*log10(p-value)',ylab='fold change (data on log2 scale)',
     main='Volcano Plot')

points(p.trans,fold.change,col='black')
points(p.trans[(p.trans>x.line&fold.change>y.line)],fold.change[(p.trans>x.line&fold.change>y.line)],
       col='red',pch=16)
points(p.trans[(p.trans>x.line&fold.change< -y.line)],fold.change[(p.trans>x.line&fold.change< -y.line)]
       col='green',pch=16)

text(p.trans[p.trans>x.line&fold.change>y.line], fold.change[p.trans>x.line&fold.change>y.line],
     labels=dimnames(dat)[[1]][p.trans[p.trans >x.line&fold.change>y.line]],
     cex=0.65, col='red', pos=4)
text(p.trans[p.trans>x.line&fold.change< -y.line], fold.change[p.trans>x.line&fold.change< -y.line],
     labels=dimnames(dat)[[1]][p.trans[p.trans>x.line&fold.change< -y.line]],
```

```
     cex=0.65, col='green', pos=4)

abline(v=x.line)
abline(h=-y.line)
abline(h=y.line)
```




```
p.trans <- -1 * log10(t.test.run)

plot(range(p.trans),range(fold.change),type='n',
     xlab='-1*log10(p-value)',ylab='fold change (data on log2 scale)',
     main='Volcano Plot')
points(p.trans,fold.change,col='black')
points(p.trans[(p.trans>1.3&fold.change>1)],fold.change[(p.trans>1.3&fold.change>1)],
       col='red',pch=16)
points(p.trans[(p.trans>1.3&fold.change< -1)],fold.change[(p.trans>1.3&fold.change< -1)],
       col='green',pch=16)
abline(v=1.3)
abline(h=-1)
abline(h=1)
```



Gene filtering

```
# Select genes with significance alpha=0.05
diff.exp.genes.t.test <- t.test.run[t.test.run < 0.05]

# Select log2 values with significance and filter data
diff.exp.genes.fold <- c(fold.change[fold.change < -log2(2)], fold.change[fold.change > log2(2)])

# Filters original data set
dat.fs = dat[names(diff.exp.genes.t.test),]
dat.fs = dat.fs[names(diff.exp.genes.fold),]

# handle missing values (NA's in gene names) created by second filter
dat.fs = dat.fs[!is.na(dat.fs[,1]),]
dim(dat.fs)
```

```
## [1] 279  11
```

Top 5 most significantly differentiated genes (up or down-regulated)

```
#Builds data frame with p-value and fold information
ds.featured.genes <- cbind(dat, t.test.run, fold.change)
```

```
#Filters data frame with featured selection genes found
ds.featured.genes <- ds.featured.genes [rownames(dat.fs),]
#Orders data frame by fold change
ds.order <- ds.featured.genes[order(ds.featured.genes[,13]),]
dim(ds.order[,12:13])
```

```
## [1] 279   2
```

Top 5 down-regulated genes

```
head(ds.order, n=5)[,12:13]
```

```
##              t.test.run fold.change
## 1375758_at 0.0002845608   -3.601134
## 1388358_at 0.0031508060   -3.196297
## 1375213_at 0.0014776675   -2.743375
## 1387408_at 0.0105059264   -2.385415
## 1372087_at 0.0009591481   -2.079921
```

Top 5 up-regulated genes

```
tail(ds.order, n=5)[,12:13]
```

```
##                t.test.run fold.change
## 1371102_x_at 9.513546e-09    2.864893
## 1370240_x_at 1.878808e-08    3.132486
## 1367553_x_at 1.255229e-06    3.171265
## 1370239_at   3.166750e-08    3.699730
## 1371245_a_at 3.749191e-06    5.785425
```

Functional annotations can then be obtained using DAVID

DAVID bioinformatics database and analysis web resource https://david.ncifcrf.gov/tools.jsp