analyzing RNAseq for differential expression of LCM data

script modified from a script given to me by Aashish Ranjan called `edgeR_DE.R`

Ciera Martinez

Install

```
source("http://bioconductor.org/biocLite.R") biocLite("edgeR")
```

```
library(edgeR)
```

Read in Data

Read in raw count data per gene. Add checknames to FALSE because it was making the columns unique.

```
counts <- read.delim("../sam2countsResults.tsv",row.names=1)

#check the file
head(counts)
summary(counts)
colnames(counts)
#need to convert NA to 0 counts
counts[is.na(counts)] <- 0
```

Subset per DE expirement

I am going to start by subsetting the particular treatments I am looking at.

WT

Marginal Blastozone vs Other

```
wtcregion <- counts[,40:48]
head(wtcregion)
```

```
##                      wtcmbr10 wtcmbr1.4.6 wtcmbr2 wtcmbr3 wtcmbr7 wtcmbr9
## Solyc00g005040.2.1        0           9       3       1       0       0
## Solyc00g005050.2.1        5          38      21      11       4       7
## Solyc00g005060.1.1        1           3       0       0       1       0
## Solyc00g005070.1.1        5          12       7       4       6       1
## Solyc00g005080.1.1        0           7      19      45       4       7
## Solyc00g005150.1.1        0           1       3       3       2       1
##                      wtcother1.3.4 wtcother2 wtcother6
## Solyc00g005040.2.1             0          0         12
## Solyc00g005050.2.1             2         6        37
## Solyc00g005060.1.1            13         0         0
## Solyc00g005070.1.1           169         6        24
## Solyc00g005080.1.1            11        26        35
## Solyc00g005150.1.1             2         1         5
```

```
#convert data to a form that edgeR wants
group <- c(rep("wtcmbr", 6), rep("wtcother",3))
d <- DGEList(counts=wtcregion,group=group)
d$samples
```

```
##                  group lib.size norm.factors
## wtcmbr10        wtcmbr   459717            1
## wtcmbr1.4.6     wtcmbr  1158809            1
## wtcmbr2         wtcmbr  1130695            1
## wtcmbr3         wtcmbr  1560130            1
## wtcmbr7         wtcmbr   374882            1
## wtcmbr9         wtcmbr   386974            1
## wtcother1.3.4 wtcother   197345            1
## wtcother2     wtcother   319043            1
## wtcother6     wtcother  1525172            1
```

Computes counts per million (CPM) then, Filter to exclude genes that have <2 counts in (N Rep)-1

```
cpm.d<- cpm(d)
d <- d[rowSums(cpm.d>2)>=3,]
```

Estimate Common Negative Binomial Dispersion by Conditional Maximum Likelihood. Maximizes the negative binomial conditional common likelihood to give the estimate of the common dispersion across all tags.

```
d <- estimateCommonDisp(d,verbose=T)
```

```
## Disp = 0.3408 , BCV = 0.5838
```

Normalize library

```
d <- calcNormFactors(d)
```

Estimate overdispersion Important so that the correct model is fit

```
d <- estimateCommonDisp(d)
```

Calculate DE genes

```
DEtest <- exactTest(d,pair=c("wtcmbr","wtcother"))
head(DEtest$table)
```

```
##                       logFC logCPM    PValue
## Solyc00g005040.2.1 0.8537  2.558 5.520e-01
## Solyc00g005050.2.1 0.2292  4.360 7.426e-01
## Solyc00g005060.1.1 3.9245  2.478 3.197e-05
## Solyc00g005070.1.1 5.3610  6.189 6.618e-16
## Solyc00g005080.1.1 1.9088  4.688 2.075e-03
## Solyc00g005150.1.1 1.1149  2.369 2.580e-01
```

Create a table of the results, with multiple testing correction.

```
results <- topTags(DEtest) #removed #n=Inf not sure what that means ask Aashish.
head(results$table)
```

```
##                     logFC logCPM    PValue       FDR
## Solyc10g052420.1.1 7.993  9.375 4.739e-30 8.754e-26
## Solyc08g023400.1.1 8.040  8.503 2.454e-29 2.267e-25
## Solyc10g050260.1.1 7.640 10.044 5.369e-29 2.999e-25
## Solyc07g039270.2.1 7.695  9.333 6.494e-29 2.999e-25
## Solyc01g028970.1.1 7.603  9.391 2.268e-28 8.379e-25
## Solyc11g020560.1.1 7.357 11.309 3.482e-28 9.974e-25
```

These are the topTags, but I want to continue with all the DE genes. Right? How many genes are DE?

```
dim(DEtest$table)
```

```
## [1] 18470     3
```

```
sum(DEtest$table$PValue<.05)  #changed from sum(results$table$adj.P.Val<.01) which resulted in 0
```

```
## [1] 2078
```

How many genes in each direction?

```
summary(decideTestsDGE(DEtest,p.value=.05))
```

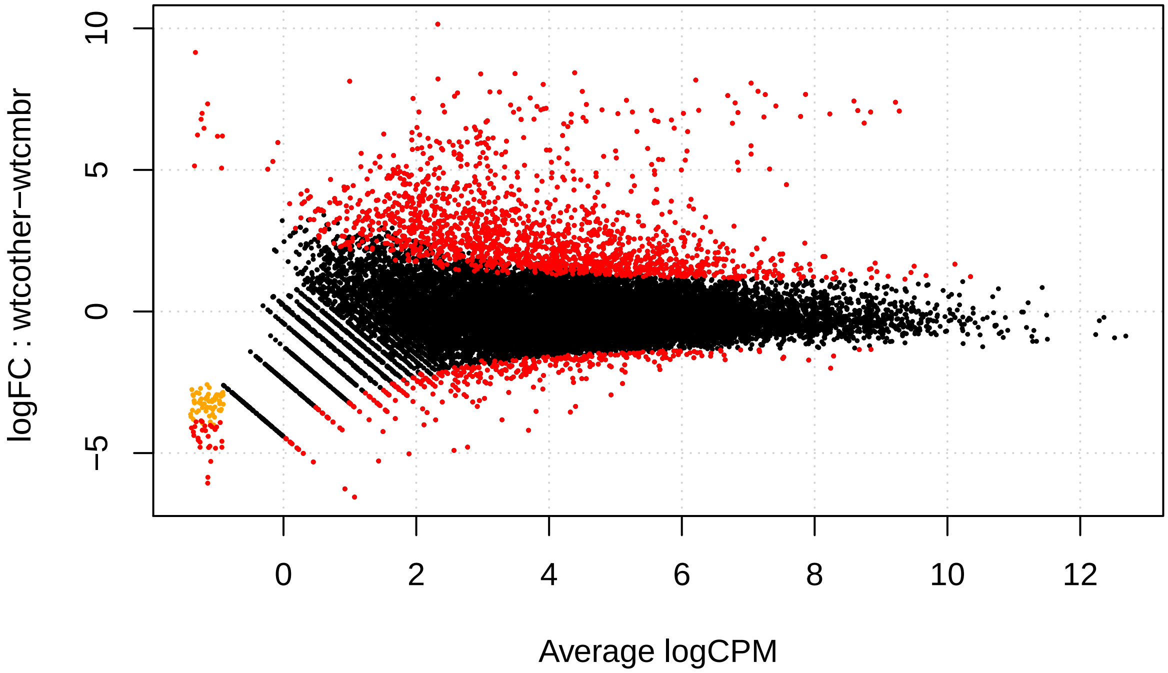
```
##    [,1]
## -1    26
## 0  17756
## 1    688
```

Plot the results First create a table of DE to highlight those with p < 0.01

```
sig.genes <- rownames(DEtest$table[DEtest$table$PValue<0.05,])
#sig.genes <- rownames(results$table[results$table$adj.P.Val<0.01,]) #original which returned 0 charact
```

Visualize with smear plot

```
plotSmear(d,de.tags=sig.genes)
```



Subset by significant score

```
results.sig <- subset(DEtest$table, DEtest$table$PValue < 0.01)
```

What are the genes that are misexpressed? For this we need to add some annotation

```
annotation1<- read.delim("../ITAG2.3_all_Arabidopsis_ITAG_annotations.tsv", header=FALSE)  #Changed to
colnames(annotation1)<- c("ITAG", "SGN_annotation")
annotation2<- read.delim ("../ITAG2.3_all_Arabidopsis_annotated.tsv")
annotation <- merge (annotation1,annotation2, by =1,1, all.x=TRUE)
head(annotation)
```

```
##                  ITAG
## 1 Solyc00g005000.2.1
## 2 Solyc00g005020.1.1
## 3 Solyc00g005040.2.1
## 4 Solyc00g005050.2.1
## 5 Solyc00g005060.1.1
## 6 Solyc00g005070.1.1
##
## 1 Aspartic proteinase nepenthesin I (AHRD V1 **-- A9ZMF9_NEPAL); contains Interpro domain(s)  IPR0014
## 2                                                                                            Unknown
## 3    Potassium channel (AHRD V1 ***- D0EM91_9ROSI); contains Interpro domain(s)  IPR000595  Cyclic nu
## 4                                                                     Arabinogalactan protein (AHRD V1 *
## 5                                                                                            Unknown
## 6                                                                                            Unknown
##         AGI symbol
## 1 AT3G20015   <NA>
## 2      <NA>   <NA>
## 3 AT5G46240   KAT1
## 4 AT5G11680   <NA>
## 5      <NA>   <NA>
## 6      <NA>   <NA>
##
## 1 pepsin A; similar to aspartyl protease family protein [Arabidopsis thaliana] (TAIR:AT3G18490.1); s
## 2
## 3
## 4
## 5
## 6
##   X..identity alignment.length e.value bit.score percent.query.align
## 1       63.76              447  7e-148       520               89.94
## 2          NA               NA      NA        NA                  NA
## 3       66.02              103   2e-37       150               85.71
## 4       76.96              204   1e-88       322               98.98
## 5          NA               NA      NA        NA                  NA
## 6          NA               NA      NA        NA                  NA
```

```
#head(results) This returns "Error: Two subscripts required"

DEtest.annotated <- merge(results.sig,annotation,by.x="row.names",by.y="ITAG",all.x=T,sort=F)

#was
#results.annotated <- merge(results$table,annotation,by.x="row.names",by.y="ITAG",all.x=T,sort=F)

head(DEtest.annotated,n=30)
```

```
##              Row.names logFC logCPM    PValue
## 1  Solyc00g005060.1.1 3.925  2.478 3.197e-05
## 2  Solyc00g005070.1.1 5.361  6.189 6.618e-16
## 3  Solyc00g005080.1.1 1.909  4.688 2.075e-03
## 4  Solyc00g006840.2.1 3.207  2.384 6.309e-04
## 5  Solyc00g009130.2.1 3.304  2.702 1.151e-04
## 6  Solyc00g014790.2.1 2.255  3.648 1.114e-03
## 7  Solyc00g014800.1.1 6.239  7.310 6.829e-21
## 8  Solyc00g026160.2.1 2.325  4.042 7.339e-04
## 9  Solyc00g036520.1.1 3.151  3.313 4.016e-05
## 10 Solyc00g042840.1.1 2.110  2.825 6.479e-03
## 11 Solyc00g094550.1.1 2.862  3.339 6.286e-05
## 12 Solyc00g095760.1.1 5.931  4.099 1.222e-14
## 13 Solyc00g105750.1.1 6.459  4.300 1.125e-16
## 14 Solyc00g112180.1.1 3.287  2.476 4.725e-05
## 15 Solyc00g112190.2.1 1.951  3.541 3.318e-03
## 16 Solyc00g121730.1.1 7.270  8.733 1.498e-26
## 17 Solyc00g131710.1.1 6.911 10.076 1.010e-25
## 18 Solyc00g195360.1.1 4.978  7.518 2.331e-15
## 19 Solyc00g206460.1.1 5.244  7.636 1.173e-16
## 20 Solyc00g212260.1.1 2.728  2.951 5.618e-04
## 21 Solyc00g227860.1.1 2.298  3.048 3.305e-03
## 22 Solyc00g272810.1.1 3.700  2.676 1.841e-04
## 23 Solyc00g281110.1.1 2.007  4.873 1.355e-03
## 24 Solyc00g313030.1.1 2.089  2.847 5.707e-03
## 25 Solyc00g323130.2.1 1.704  4.390 6.522e-03
## 26 Solyc01g005000.2.1 4.429  2.758 3.364e-07
## 27 Solyc01g005010.2.1 2.003  3.611 4.103e-03
## 28 Solyc01g005080.2.1 2.178  4.149 9.361e-04
## 29 Solyc01g005130.2.1 2.803  2.953 2.085e-03
## 30 Solyc01g005450.2.1 2.431  3.264 7.153e-04
##
## 1
## 2
## 3
## 4
## 5                Dehydrogenase/reductase SDR family member 12 (AHRD V1 ***- COHAGO_SALSA); contains
## 6
## 7                                                        Zinc-finger protein 1 (AHRD V1 *-*- D7KX25_
## 8                                              Ferric reductase oxidase (AHRD V1 **** D6RVS5_HORVU);
## 9
## 10                                 F-box domain containing protein (AHRD V1 ***- Q2QXK7_ORYSJ); conta
## 11
## 12                        1-aminocyclopropane-1-carboxylate synthase (AHRD V1 ***- Q96580_SOLLC); co
## 13                                                  Mutator-like transposase (AHRD V1 *--- Q94HK4_OR
## 14
## 15
## 16                             &aposchromo&apos domain containing protein (AHRD V1 *--- Q6L3Q3_SOLDE
## 17                                                             Pol polyprotein (AHRD V1 *-*- POL_MLV
## 18 Serine/threonine-protein phosphatase 7 long form homolog (AHRD V1 **-- PPP7L_ARATH); contains Inte
## 19                       Os06g0220000 protein (Fragment) (AHRD V1 **-- Q0DDJ2_ORYSJ); contains Interp
## 20                                                   GH3 family protein (AHRD V1 *-*- B9GQG9_POPTR);
## 21                                   UDP-glucosyltransferase (AHRD V1 ***- B3VI56_STERE); contains I
## 22                                               N-acetyltransferase (AHRD V1 ***- B6SUK9_MAIZE); co
## 23
## 24
## 25                                                          Major latex-like protein (AHRD V1 **-- B
## 26                                                    Glutamate decarboxylase (AHRD V1 ***- Q1I1D8_C
## 27
## 28                                                 Microtubule-associated protein MAP65-1a (AHRD V1
## 29                                                       Zinc finger protein 7 (AHRD V1 ***- B6U8J3_
## 30                                                F-box protein family-like protein (AHRD V1 ***- Q7
##          AGI    symbol
## 1       <NA>      <NA>
## 2       <NA>      <NA>
## 3  ATCG01280    YCF2.2
## 4       <NA>      <NA>
## 5  AT4G09750      <NA>
## 6       <NA>      <NA>
## 7       <NA>      <NA>
## 8  AT5G23980    ATFRO4
## 9       <NA>      <NA>
## 10      <NA>      <NA>
## 11      <NA>      <NA>
## 12 AT3G61510      ACS1
## 13      <NA>      <NA>
## 14      <NA>      <NA>
## 15      <NA>      <NA>
## 16      <NA>      <NA>
## 17      <NA>      <NA>
## 18      <NA>      <NA>
## 19      <NA>      <NA>
## 20      <NA>      <NA>
## 21 AT2G22590      <NA>
## 22      <NA>      <NA>
## 23      <NA>      <NA>
## 24      <NA>      <NA>
## 25      <NA>      <NA>
## 26 AT2G02010      GAD4
## 27      <NA>      <NA>
## 28 AT2G01910 ATMAP65-6
## 29      <NA>      <NA>
## 30      <NA>      <NA>
##
## 1
## 2
## 3
## 4
## 5
## 6
## 7
## 8
## 9
## 10
## 11
## 12
## 13
## 14
## 15
## 16
## 17
## 18
## 19
## 20
## 21
## 22
## 23
## 24
## 25
## 26 glutamate decarboxylase, putative; similar to GAD (Glutamate decarboxylase 1), calmodulin binding
## 27
## 28
## 29
## 30
##    X..identity alignment.length e.value bit.score percent.query.align
## 1           NA               NA      NA        NA                  NA
## 2           NA               NA      NA        NA                  NA
## 3        91.25               80   2e-38       153               79.80
## 4           NA               NA      NA        NA                  NA
## 5        75.76               66   4e-27       115               79.27
## 6           NA               NA      NA        NA                  NA
## 7           NA               NA      NA        NA                  NA
## 8        58.46              455  3e-140       495               99.55
## 9           NA               NA      NA        NA                  NA
## 10          NA               NA      NA        NA                  NA
## 11          NA               NA      NA        NA                  NA
## 12       58.76              291   2e-95       345               83.53
## 13          NA               NA      NA        NA                  NA
## 14          NA               NA      NA        NA                  NA
## 15          NA               NA      NA        NA                  NA
## 16          NA               NA      NA        NA                  NA
## 17          NA               NA      NA        NA                  NA
## 18          NA               NA      NA        NA                  NA
## 19          NA               NA      NA        NA                  NA
## 20          NA               NA      NA        NA                  NA
## 21       55.68              458  2e-138       489               96.39
## 22          NA               NA      NA        NA                  NA
## 23          NA               NA      NA        NA                  NA
## 24          NA               NA      NA        NA                  NA
## 25          NA               NA      NA        NA                  NA
## 26       82.42              495   0e+00       816               99.59
## 27          NA               NA      NA        NA                  NA
## 28       67.20              567   0e+00       732               99.64
## 29          NA               NA      NA        NA                  NA
## 30          NA               NA      NA        NA                  NA
```

Write table with results

write.table(DEtest.annotated,“wtcmbr_wtcother_DE.txt”,sep=“”,row.names=F)