Ball Helix Bioinformatics Assoc. Scientist Analyst Test

M. Ross Alexander

12/13/2019

```
## Loading required package: BiocGenerics

## Loading required package: parallel

##
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:parallel':
##
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, parApply, parCapply, parLapply,
##     parLapplyLB, parRapply, parSapply, parSapplyLB

## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs

## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, basename, cbind, colnames,
##     dirname, do.call, duplicated, eval, evalq, Filter, Find, get, grep,
##     grepl, intersect, is.unsorted, lapply, Map, mapply, match, mget,
##     order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,
##     rbind, Reduce, rownames, sapply, setdiff, sort, table, tapply,
##     union, unique, unsplit, which, which.max, which.min

## Loading required package: Biobase

## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.

## Loading required package: locfit

## locfit 1.5-9.1    2013-03-22

## Loading required package: lattice

##     Welcome to 'DESeq'. For improved performance, usability and
##     functionality, please consider migrating to 'DESeq2'.

## Loading required package: limma
```

```
##
## Attaching package: 'limma'

## The following object is masked from 'package:DESeq':
##
##     plotMA

## The following object is masked from 'package:BiocGenerics':
##
##     plotMA

## Loading required package: grid

## Loading required package: futile.logger

## Loading required package: AnnotationDbi

## Loading required package: stats4

## Loading required package: IRanges

## Loading required package: S4Vectors

##
## Attaching package: 'S4Vectors'

## The following object is masked from 'package:base':
##
##     expand.grid

##

##
```

Analyst Test data exercise

The original data frame consisted of count data produced by the tool Salmon for three replicates during the bud phase and then two days after flowering. The following document walks through the analyses used to determine which genes might be good candidates for further study.

basic folder structure figures–preliminary figures saved here input_data–original data files to be called for analyses processed_data–modified data files produced by analyses

Summary

1) This analysis identified 8245 genes that showed statistically significant changes between bud and flower phases (alpha = 0.05). From these initial 8245 genes, I identified 99 genes that were not present in the bud phase but up regulated in the flower phase and 108 genes that were present in the bud phase but then down regulated in the flower phase. These 207 genes appear to be good candidates for further investigation of the desred trait

2) The figure illustrating the position of these genes can found in figures/gene_onoff.png.

3) For this experiment three samples were taken from plants in the bud phase and then two days after flowering. Each of these samples were analyzed to identify genes of interest using the tool Salmon. Using various references, this tool identifies gene sequences of interest and allows us to count how often they are expressed in our samples. The higher the count, the more these genes are present in each the bud and the flowering phases. By tallying these counts we have identified genes that were either turned on or off as the plants went from the bud phase (closed flowers) to 2 days after the flowering phase.

Included Scripts

This summary is composed of insights cleaned after carrying out the following scripts. Not all analyses detailed in scripts are presented in this document.

1_data_walk through.R–Initial looks at the data frame. Included to show thought process; considered rough.

2_genetics_example.R–preliminary analyses to identify potential genes of interest.

Original data frame

```
# Read in data ------------------------------------------------------------

datafile = read.table("input_data/test_gene_expression_matrix.txt", header=T,
row.names = 1) # making column 1 row names
head(datafile)

##                  Bud_Rep1 Bud_Rep2 Bud_Rep3 Day2_Rep1 Day2_Rep2 Day2_Rep3
## Flower_060945-RA 10.83150  9.56753  6.24958   9.60078   6.01394   5.99760
## Flower_027927-RA  0.00000  0.00000  0.00000   0.00000   0.00000   0.00000
## Flower_027924-RA  1.56159  2.00614  1.85567   2.57317   1.27196   1.32492
## Flower_027916-RA  8.00183  7.34657  6.13769  29.93260  29.57190  29.14260
## Flower_027915-RA  0.00000  0.00000  0.00000   2.47182   1.11015   0.00000
## Flower_027910-RA  0.00000  0.00000  0.00000   0.00000   0.00000   0.00000

summary(datafile)

##     Bud_Rep1            Bud_Rep2            Bud_Rep3
##  Min.   :    0.000   Min.   :    0.000   Min.   :    0.000
##  1st Qu.:    0.000   1st Qu.:    0.000   1st Qu.:    0.000
```

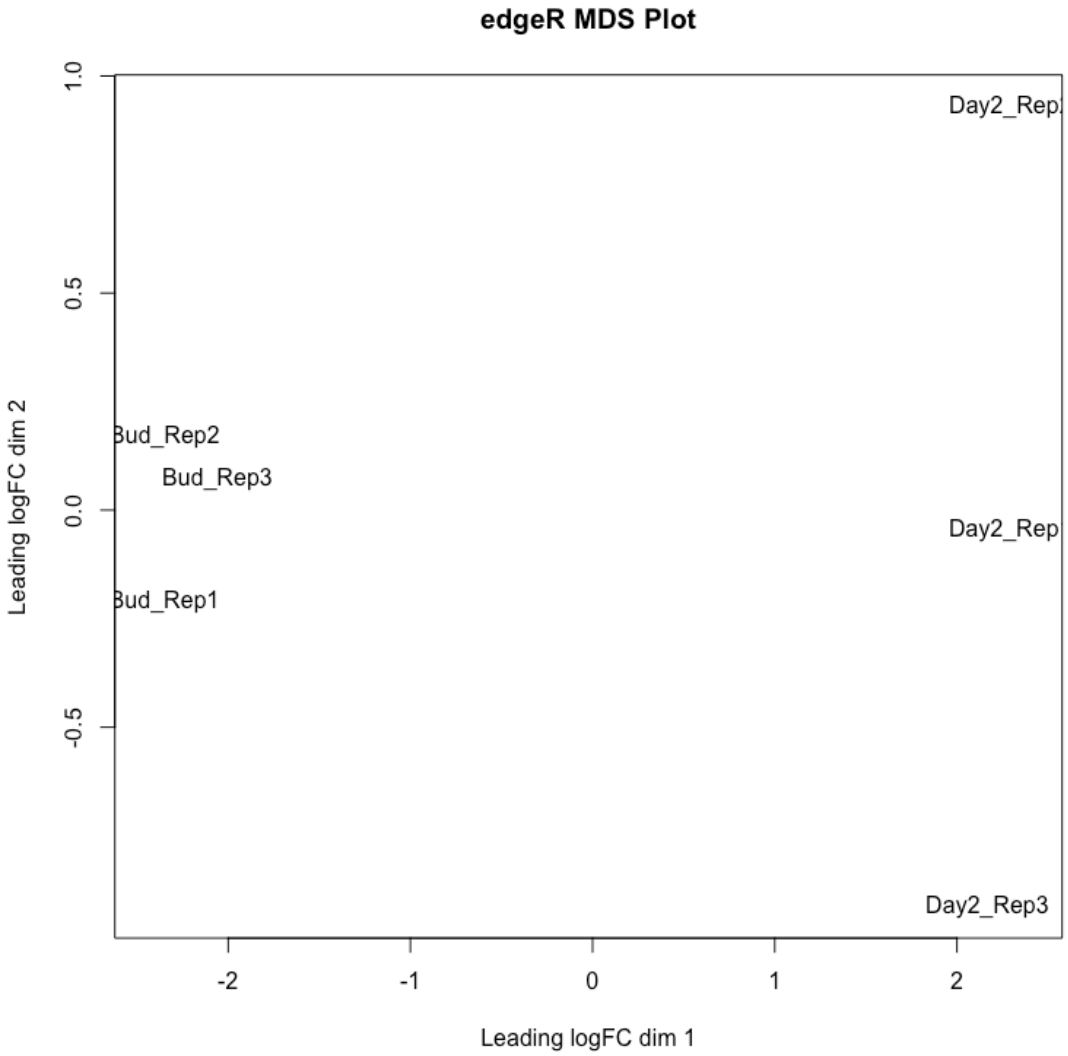
```
##  Median :    3.683   Median :    3.597   Median :    3.563
##  Mean   :   34.091   Mean   :   34.091   Mean   :   34.091
##  3rd Qu.:   23.714   3rd Qu.:   23.062   3rd Qu.:   23.128
##  Max.   :25513.500   Max.   :25231.000   Max.   :24956.200
##    Day2_Rep1           Day2_Rep2           Day2_Rep3
##  Min.   :    0.000   Min.   :    0.000   Min.   :    0.000
##  1st Qu.:    0.000   1st Qu.:    0.000   1st Qu.:    0.000
##  Median :    2.325   Median :    2.535   Median :    2.272
##  Mean   :   34.091   Mean   :   34.091   Mean   :   34.091
##  3rd Qu.:   18.955   3rd Qu.:   19.671   3rd Qu.:   19.303
##  Max.   :13699.700   Max.   :14201.600   Max.   :13860.400
```

```
## Read in the data making the row names the first column
counttable <- datafile #creating a count table for later use
head(counttable)
```

```
##                  Bud_Rep1 Bud_Rep2 Bud_Rep3 Day2_Rep1 Day2_Rep2 Day2_Rep3
## Flower_060945-RA 10.83150  9.56753  6.24958   9.60078   6.01394   5.99760
## Flower_027927-RA  0.00000  0.00000  0.00000   0.00000   0.00000   0.00000
## Flower_027924-RA  1.56159  2.00614  1.85567   2.57317   1.27196   1.32492
## Flower_027916-RA  8.00183  7.34657  6.13769  29.93260  29.57190  29.14260
## Flower_027915-RA  0.00000  0.00000  0.00000   2.47182   1.11015   0.00000
## Flower_027910-RA  0.00000  0.00000  0.00000   0.00000   0.00000   0.00000
```

Making a metadata file to run edgeR analytical stemps

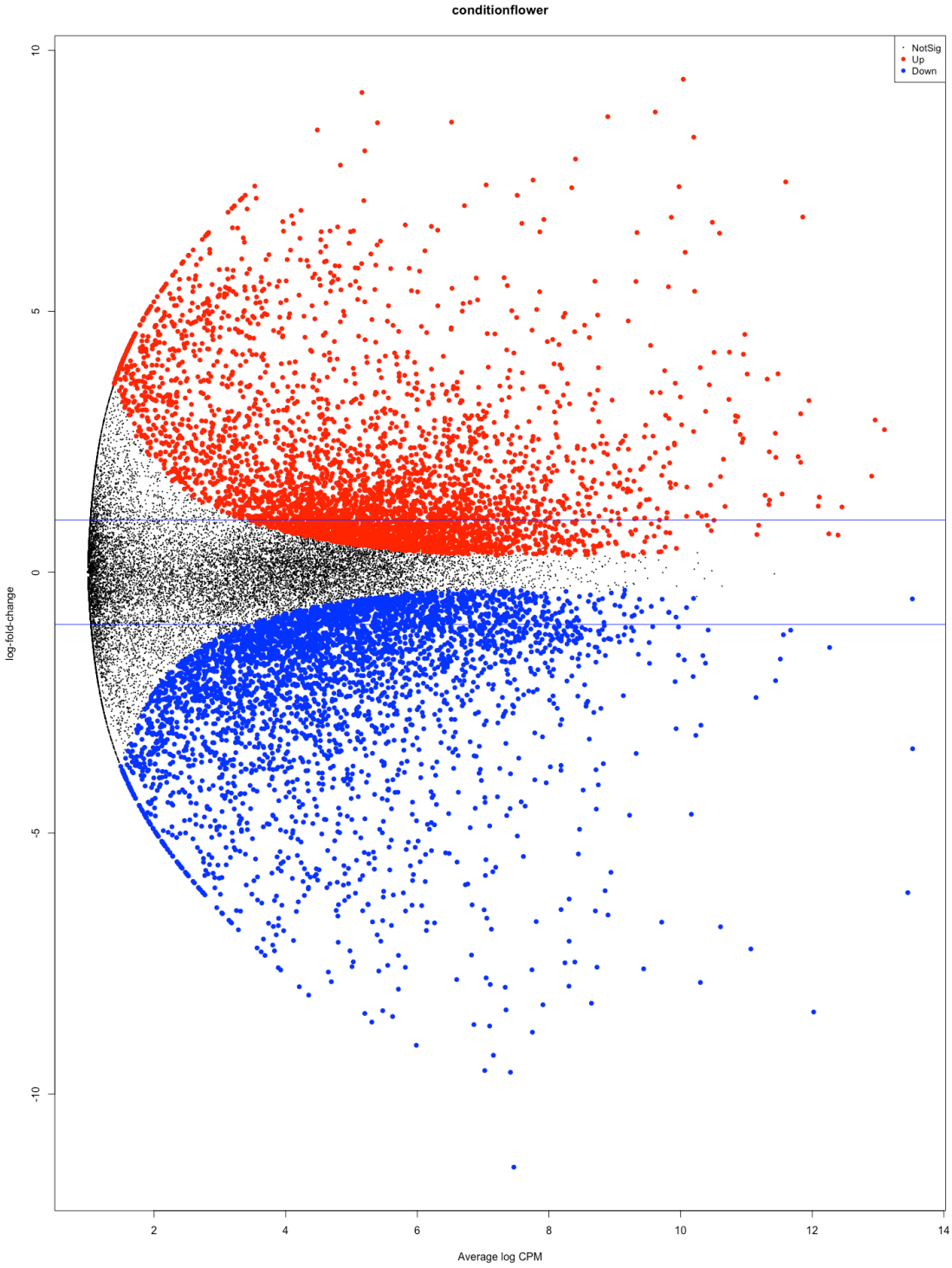
```
##           condition
## Bud_Rep1        bud
## Bud_Rep2        bud
## Bud_Rep3        bud
## Day2_Rep1    flower
## Day2_Rep2    flower
## Day2_Rep3    flower
```



The ordination plot shows us that the genes being expressed during the bud phase are more similar across replicates than the genes being expressed after flowering

```
##                       logFC    logCPM        LR        PValue           FDR
## Flower_028715-RA  8.816934  9.615373 1119.3027 2.106582e-245 6.179238e-241
## Flower_045575-RA -7.219929 11.067035  974.8029 5.386207e-214 7.899680e-210
## Flower_030986-RA  6.800501  9.858868  970.7382 4.119255e-213 4.027670e-209
## Flower_003600-RA  8.335348 10.200407  931.3438 1.507264e-204 1.105314e-200
## Flower_049949-RA -7.569132  8.729698  891.3426 7.479175e-196 4.387733e-192
## Flower_031552-RA -8.427950 12.021636  859.7831 5.428833e-189 2.654066e-185

## sig_edgeR
## FALSE  TRUE   Sum
## 21088  8245 29333

##        conditionflower
## Down              4002
```

```
## NotSig            21088
## Up                 4243
```



This plot shows the genes in which we see a significant count change that either increases (up; red) or decreases (down; blue). Blue horizontal lines show a 2-fold change in expression. Significance was determined at the alpha = 0.05 level.

```
##  conditionflower                  gene
##  Min.   :-1.000000   Flower_000007-RA:    1
##  1st Qu.: 0.000000   Flower_000008-RA:    1
##  Median : 0.000000   Flower_000009-RA:    1
##  Mean   : 0.008216   Flower_000010-RA:    1
##  3rd Qu.: 0.000000   Flower_000011-RA:    1
##  Max.   : 1.000000   Flower_000012-RA:    1
##                      (Other)         :29327

##  conditionflower               gene
##  Min.   :1       Flower_000040-RA:   1
##  1st Qu.:1       Flower_000042-RA:   1
##  Median :1       Flower_000099-RA:   1
##  Mean   :1       Flower_000118-RA:   1
##  3rd Qu.:1       Flower_000130-RA:   1
##  Max.   :1       Flower_000158-RA:   1
##                  (Other)         :4237

##  conditionflower               gene
##  Min.   :-1      Flower_000017-RA:   1
##  1st Qu.:-1      Flower_000059-RA:   1
##  Median :-1      Flower_000110-RA:   1
##  Mean   :-1      Flower_000127-RA:   1
##  3rd Qu.:-1      Flower_000134-RA:   1
##  Max.   :-1      Flower_000136-RA:   1
##                  (Other)         :3996
```

Taking the underlying data frame that made the previous figure, we can extract the names of genes that were significantly up or down regulated. This still leaves about four thousand genes each that showed significant increases or decreases in expression between buds and flowers. Although fewer than the total genes we started with (ca. 30,000), we can still reduce the number further.

```
# creating a manhattan plot to show which genes would be up regulated and
which genes are down regulated in this scenario
rna.dat <- datafile

rna.dat$gene <- as.factor(row.names(rna.dat))
summary(rna.dat)

##     Bud_Rep1            Bud_Rep2            Bud_Rep3
##  Min.   :    0.000   Min.   :    0.000   Min.   :    0.000
##  1st Qu.:    0.000   1st Qu.:    0.000   1st Qu.:    0.000
##  Median :    3.683   Median :    3.597   Median :    3.563
##  Mean   :   34.091   Mean   :   34.091   Mean   :   34.091
##  3rd Qu.:   23.714   3rd Qu.:   23.062   3rd Qu.:   23.128
##  Max.   :25513.500   Max.   :25231.000   Max.   :24956.200
##
```

```
##    Day2_Rep1           Day2_Rep2           Day2_Rep3
##  Min.   :    0.000   Min.   :    0.000   Min.   :    0.000
##  1st Qu.:    0.000   1st Qu.:    0.000   1st Qu.:    0.000
##  Median :    2.325   Median :    2.535   Median :    2.272
##  Mean   :   34.091   Mean   :   34.091   Mean   :   34.091
##  3rd Qu.:   18.955   3rd Qu.:   19.671   3rd Qu.:   19.303
##  Max.   :13699.700   Max.   :14201.600   Max.   :13860.400
##
##                gene
##  Flower_000007-RA:    1
##  Flower_000008-RA:    1
##  Flower_000009-RA:    1
##  Flower_000010-RA:    1
##  Flower_000011-RA:    1
##  Flower_000012-RA:    1
##  (Other)         :29327
```

```
# separatign bud phase from flower phase so that we can take the mean of each
of these
bud.dat <- rna.dat[,c("gene", "Bud_Rep1", "Bud_Rep2", "Bud_Rep3")]
flower.dat <- rna.dat[,c("gene", "Day2_Rep1", "Day2_Rep2", "Day2_Rep3")]

summary(bud.dat)
```

```
##                gene          Bud_Rep1            Bud_Rep2
##  Flower_000007-RA:    1   Min.   :    0.000   Min.   :    0.000
##  Flower_000008-RA:    1   1st Qu.:    0.000   1st Qu.:    0.000
##  Flower_000009-RA:    1   Median :    3.683   Median :    3.597
##  Flower_000010-RA:    1   Mean   :   34.091   Mean   :   34.091
##  Flower_000011-RA:    1   3rd Qu.:   23.714   3rd Qu.:   23.062
##  Flower_000012-RA:    1   Max.   :25513.500   Max.   :25231.000
##  (Other)         :29327
##     Bud_Rep3
##  Min.   :    0.000
##  1st Qu.:    0.000
##  Median :    3.563
##  Mean   :   34.091
##  3rd Qu.:   23.128
##  Max.   :24956.200
##
```

```
bud.mean <- data.frame(gene=bud.dat$gene,
                       value = rowMeans(bud.dat[,c("Bud_Rep1", "Bud_Rep2",
"Bud_Rep3")]),
                       type = as.factor("bud"))
summary(bud.mean)
```

```
##                gene           value            type
##  Flower_000007-RA:    1   Min.   :    0.000   bud:29333
##  Flower_000008-RA:    1   1st Qu.:    0.029
##  Flower_000009-RA:    1   Median :    3.627
```

```
##  Flower_000010-RA:    1   Mean   :   34.091
##  Flower_000011-RA:    1   3rd Qu.:   23.325
##  Flower_000012-RA:    1   Max.   :24189.300
##  (Other)         :29327
```

```
bud.mean$log2.value <- ifelse(bud.mean$value == 0, log2(bud.mean$value + 1e-
6), log2(bud.mean$value)) # adding a small value to all zero values to avoid
infinity error

flower.mean <- data.frame(gene=flower.dat$gene,
                          value = rowMeans(flower.dat[,c("Day2_Rep1",
"Day2_Rep2", "Day2_Rep3")]),
                          type = as.factor("flower"))
flower.mean$log2.value <- ifelse(flower.mean$value == 0,
log2(flower.mean$value + 1e-6), log2(flower.mean$value)) # adding a small
value to all zero values to avoid infinity error
summary(flower.mean)
```

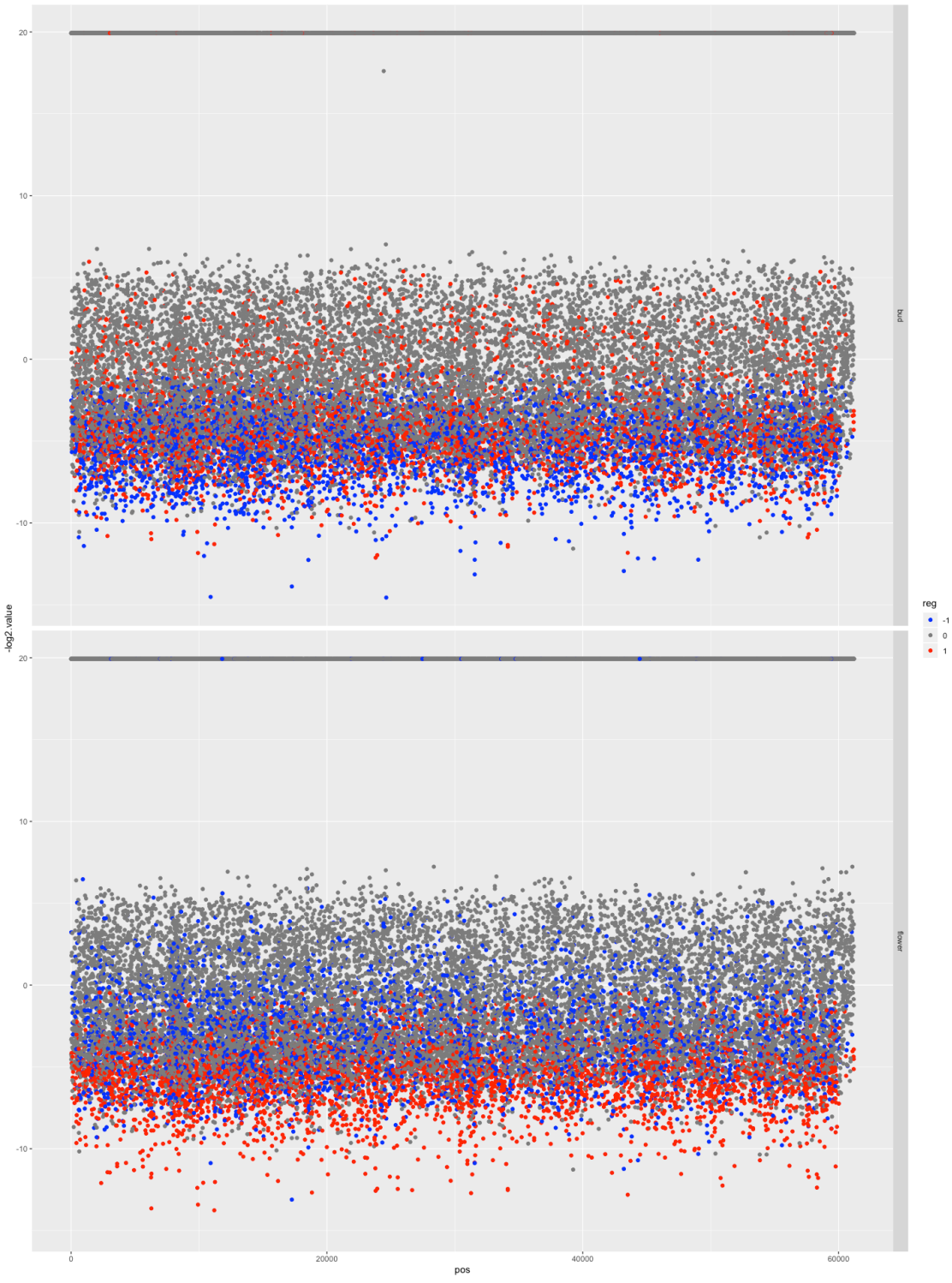
```
##                 gene           value               type         log2.value
##  Flower_000007-RA:    1   Min.   :    0.000   flower:29333   Min.   :-
19.932
##  Flower_000008-RA:    1   1st Qu.:    0.015                  1st Qu.: -
6.031
##  Flower_000009-RA:    1   Median :    2.431                  Median :
1.281
##  Flower_000010-RA:    1   Mean   :   34.091                  Mean   : -
3.114
##  Flower_000011-RA:    1   3rd Qu.:   19.340                  3rd Qu.:
4.274
##  Flower_000012-RA:    1   Max.   :13920.567                  Max.   :
13.765
##  (Other)         :29327
```

```
# stacking together for graphing purposes

rna.mean.stack <- rbind(bud.mean, flower.mean)

# adding color for the up/down regulation of genes
rna.mean.stack$reg <- as.factor(ifelse(rna.mean.stack$gene %in% up.reg$gene,
1,
                               ifelse(rna.mean.stack$gene %in% down.reg$gene,-
1,0)) )

summary(rna.mean.stack)
```

```
##                 gene           value               type         log2.value
##  Flower_000007-RA:    2   Min.   :    0.000   bud    :29333   Min.   :-
19.932
##  Flower_000008-RA:    2   1st Qu.:    0.024   flower:29333   1st Qu.: -
5.406
```

```
##  Flower_000009-RA:    2   Median :    2.971                   Median :
1.571
##  Flower_000010-RA:    2   Mean   :   34.091                   Mean   : -
2.955
##  Flower_000011-RA:    2   3rd Qu.:   21.503                   3rd Qu.:
4.426
##  Flower_000012-RA:    2   Max.   :24189.300                   Max.   :
14.562
##  (Other)         :58654
##  reg
##  -1: 8004
##  0 :42176
##  1 : 8486
##
##
##
##
```

```
gn.names <- as.factor(matrix(unlist(strsplit(paste(rna.mean.stack$gene),
"_")), nrow=nrow(rna.mean.stack), byrow=T)[,2])
rna.mean.stack$pos <- matrix(unlist(strsplit(paste(gn.names), "-")),
nrow=nrow(rna.mean.stack), byrow=T)[,1]
rna.mean.stack$pos <- as.numeric(rna.mean.stack$pos)
```


-log2.value
20
10
0
-10
bud
flower
reg
-1
0
1
0
20000
40000
60000
pos

These initial plots were used to identify genes that in the bud stage were completely turned off and then were turned on in the flowering stage, and those genes that were expressed during the bud stage and then completely turned off in the flowering stage. Note: data were -log2 transformed to assist with visualization.

```
# Pulling out the band of values at the top that are colored red in the bud
# Meaning they were turned off in the bud but were then turned on in the
flower

rna.mean.stack.bud <- rna.mean.stack[rna.mean.stack$type=="bud",]
summary(rna.mean.stack.bud)

##                gene           value               type         log2.value
##  Flower_000007-RA:    1   Min.   :    0.000   bud   :29333   Min.   :-
19.932
##  Flower_000008-RA:    1   1st Qu.:    0.029   flower:    0   1st Qu.: -
5.127
##  Flower_000009-RA:    1   Median :    3.627                   Median :
1.859
##  Flower_000010-RA:    1   Mean   :   34.091                   Mean   : -
2.796
##  Flower_000011-RA:    1   3rd Qu.:   23.325                   3rd Qu.:
4.544
##  Flower_000012-RA:    1   Max.   :24189.300                   Max.   :
14.562
##  (Other)         :29327
##  reg             pos
##  -1: 4002   Min.   :    7
##  0 :21088   1st Qu.:13453
##  1 : 4243   Median :27938
##             Mean   :28955
##             3rd Qu.:44228
##             Max.   :61202
##

rna.mean.stack.bud$gn.sig <- as.factor(ifelse(rna.mean.stack.bud$log2.value <
-10 & rna.mean.stack.bud$reg == 1,"Y","N"))
summary(rna.mean.stack.bud)

##                gene           value               type         log2.value
##  Flower_000007-RA:    1   Min.   :    0.000   bud   :29333   Min.   :-
19.932
##  Flower_000008-RA:    1   1st Qu.:    0.029   flower:    0   1st Qu.: -
5.127
##  Flower_000009-RA:    1   Median :    3.627                   Median :
1.859
##  Flower_000010-RA:    1   Mean   :   34.091                   Mean   : -
2.796
##  Flower_000011-RA:    1   3rd Qu.:   23.325                   3rd Qu.:
4.544
##  Flower_000012-RA:    1   Max.   :24189.300                   Max.   :
```

```
14.562
##  (Other)         :29327
##  reg             pos        gn.sig
##  -1: 4002   Min.   :    7   N:29234
##  0 :21088   1st Qu.:13453   Y:   99
##  1 : 4243   Median :27938
##             Mean   :28955
##             3rd Qu.:44228
##             Max.   :61202
##
```

```
gn.on <- ggplot(rna.mean.stack.bud[rna.mean.stack.bud$gn.sig=="Y",]) +
  geom_hline(aes(, yintercept=1),lwd=1000, col="grey65") +
  geom_vline(aes(xintercept=pos, y=1), col="purple") +
  labs(x="Position", y="", fill="", title="Turned On", caption = "n = 99") +
  theme(axis.line=element_line(color="black"),
        panel.grid.major=element_blank(),
        panel.grid.minor=element_blank(),
        panel.border=element_blank(),
        panel.background=element_blank(),
        axis.text.x=element_text(angle=0, color="black", size=16, vjust= 0.5,
face="bold"),
        axis.text.y=element_blank(),
        strip.text=element_text(face="bold", size=22),
        axis.line.x = element_line(color="black", size = 0.5),
        axis.line.y = element_blank(),
        axis.ticks.y = element_blank(),
        legend.position="top",
        legend.key.size = unit(0.75, "cm"),
        legend.text = element_text(size=18),
        legend.title = element_text(size=22),
        legend.key = element_rect(fill = "white")) +
  guides(fill = guide_colourbar(barwidth = 15, barheight = 3,title="")) +
  theme(axis.title.y= element_text(size=24, face="bold")) +
  theme(axis.title.x= element_text(size=24, face="bold")) +
  scale_x_continuous(limits = c(0,max(rna.mean.stack.bud$pos)),breaks  =
seq(0,max(rna.mean.stack.bud$pos), by=10000))
```

```
## Warning: Ignoring unknown aesthetics: y
```

```
# The center of the distribution is just a flat out mess, but there are blips
of blue around the 20 mark that merit investigation
# Pulling out the band of values at the top that are colored blue in the
flower meaning they were turned on in teh bud and are not turned off
rna.mean.stack.flower <- rna.mean.stack[rna.mean.stack$type=="flower",]
summary(rna.mean.stack.flower)
```

```
##                 gene           value               type         log2.value
##  Flower_000007-RA:    1   Min.   :    0.000   bud   :    0   Min.   :-
19.932
##  Flower_000008-RA:    1   1st Qu.:    0.015   flower:29333   1st Qu.: -
```

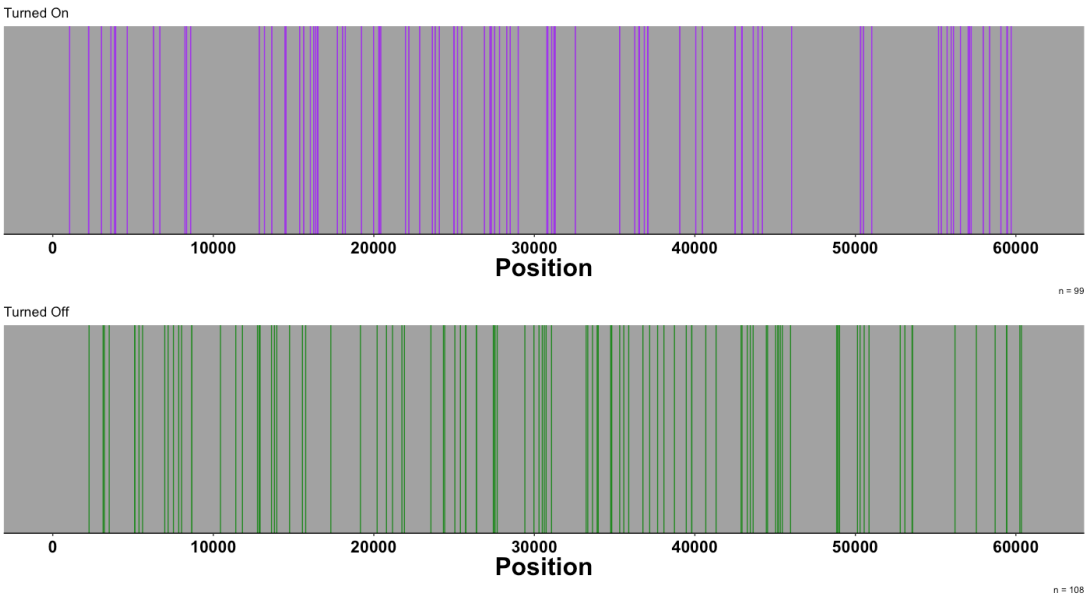
```
6.031
##  Flower_000009-RA:    1   Median :    2.431                   Median :
1.281
##  Flower_000010-RA:    1   Mean   :   34.091                   Mean   : -
3.114
##  Flower_000011-RA:    1   3rd Qu.:   19.340                   3rd Qu.:
4.274
##  Flower_000012-RA:    1   Max.   :13920.567                   Max.   :
13.765
##  (Other)         :29327
##  reg             pos
##  -1: 4002   Min.   :    7
##  0 :21088   1st Qu.:13453
##  1 : 4243   Median :27938
##             Mean   :28955
##             3rd Qu.:44228
##             Max.   :61202
##

rna.mean.stack.flower$gn.sig <-
as.factor(ifelse(rna.mean.stack.flower$log2.value < -10 &
rna.mean.stack.flower$reg == -1,"Y","N"))

summary(rna.mean.stack.flower)

##                gene           value               type         log2.value
##  Flower_000007-RA:    1   Min.   :    0.000   bud   :    0   Min.   :-
19.932
##  Flower_000008-RA:    1   1st Qu.:    0.015   flower:29333   1st Qu.: -
6.031
##  Flower_000009-RA:    1   Median :    2.431                  Median :
1.281
##  Flower_000010-RA:    1   Mean   :   34.091                  Mean   : -
3.114
##  Flower_000011-RA:    1   3rd Qu.:   19.340                  3rd Qu.:
4.274
##  Flower_000012-RA:    1   Max.   :13920.567                  Max.   :
13.765
##  (Other)         :29327
##  reg             pos        gn.sig
##  -1: 4002   Min.   :    7   N:29225
##  0 :21088   1st Qu.:13453   Y:  108
##  1 : 4243   Median :27938
##             Mean   :28955
##             3rd Qu.:44228
##             Max.   :61202
##

gn.off <- ggplot(rna.mean.stack.flower[rna.mean.stack.flower$gn.sig=="Y",]) +
  geom_hline(aes(, yintercept=1),lwd=1000, col="grey65") +
```

```
  geom_vline(aes(xintercept=pos, y=1), col="forestgreen") +
  labs(x="Position", y="", fill="", title = "Turned Off", caption ="n = 108")
+
  theme(axis.line=element_line(color="black"),
        panel.grid.major=element_blank(),
        panel.grid.minor=element_blank(),
        panel.border=element_blank(),
        panel.background=element_blank(),
        axis.text.x=element_text(angle=0, color="black", size=16, vjust= 0.5,
face="bold"),
        axis.text.y=element_blank(),
        strip.text=element_text(face="bold", size=22),
        axis.line.x = element_line(color="black", size = 0.5),
        axis.line.y = element_blank(),
        axis.ticks.y = element_blank(),
        legend.position="top",
        legend.key.size = unit(0.75, "cm"),
        legend.text = element_text(size=18),
        legend.title = element_text(size=22),
        legend.key = element_rect(fill = "white")) +
  guides(fill = guide_colourbar(barwidth = 15, barheight = 3,title="")) +
  theme(axis.title.y= element_text(size=24, face="bold")) +
  theme(axis.title.x= element_text(size=24, face="bold")) +
  scale_x_continuous(limits = c(0,max(rna.mean.stack.bud$pos)),breaks  =
seq(0,max(rna.mean.stack.bud$pos), by=10000))
```

```
## Warning: Ignoring unknown aesthetics: y
```

From the above plots, we can identify the genes that were either switched on or off after flowering present them in a more appealing way.

```
##
## ************************************************************

## Note: As of version 1.0.0, cowplot does not change the

##   default ggplot2 theme anymore. To recover the previous

##   behavior, execute:
##   theme_set(theme_cowplot())

## ************************************************************
```



From the original ca. 30,000 genes we have identified 99 that were significantly up regulated and 108 that were significantly down regulated between bud and flower stages. At this point, I would rely on the expertise of my colleagues to further refine these candidates for further analysis in a cost-effective manner.

Code was amended from examples illustrated in 'Getting Genetics Done' website
https://www.gettinggeneticsdone.com/2012/09/deseq-vs-edger-comparison.html

edgeR documentation can be found here:
https://bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf