MIS-DESeq2

Sunit Jain

January 6, 2015

Contents

Dependencies

If you're unsure that you have all the pacakges required to run this workflow. Open the `Rmd` file in your favorite text editor (I used RStudio) and change the next line from `eval=FALSE` to `eval=TRUE`. Now, when you run this workflow, the dependencies should be installed first.

Generate a read count matrix using `htseq-count`

Sample command:

```
htseq-count -f bam -r name -t CDS -o scaffold.htseq.sam -i ID -q scaffold_sortedByName.bam
all_combined.gff
```

This command was run for each sample individually.

Merging duplicate genes

I performed a self blast and looked at results that had a percent identity greater than 98%, query coverage greater than 96% and a minimum alignment length of 500 bases. Once I had this subset, I screened out the hits to exons since we won't be considering them for this experiment anyway. I was left with the following two gene pairs:

- scaffold_344578___MIS_1109813.1 scaffold_133898___MIS_10093600.14
- scaffold_219988___MIS_10179608.12 scaffold_555373___MIS_1172265.1

that had high enough similarity based on the thresholds mentioned above that their count data needed to be merged. The perl script `mergeCounts.pl` was run on each htseq-count output individually in order to accomplish this. Here is a sample command used for one of the htseq-count outputs:

```
perl mergeCounts.pl -l realDuplicateGenes.list -tsv Day_1.htseqCount.tsv -o Day_1.htseqCount.merged.tsv
```

where, realDuplicateGenes.list contains the two gene pairs mentioned above.

Import Counts into DESeq2

Once we were satisfied with the genes and their counts. We imported the count data into DESeq2.

Reads per Sample

```
##   Day_1   Day_2   Day_3 Night_4 Night_5 Night_6
## 2739735  492104  691689 1105737 1587917  969992
```

Filtering the data

Get rid of genes which did not occur frequently enough. Here we say, lets get rid of genes with counts >=1 in at least 2 samples.

```
##   Day_1   Day_2   Day_3 Night_4 Night_5 Night_6
## 2716981  485111  683644 1096670 1561645  952662
```

How many reads were removed when Min Raw Count = 1?

```
##   Day_1   Day_2   Day_3 Night_4 Night_5 Night_6
##   22754    6993    8045    9067   26272   17330
```

This reduces the dataset from 1464832 tags to about 26544. For the filtered tags, there is very little power to detect differential expression, so little information is lost by filtering.

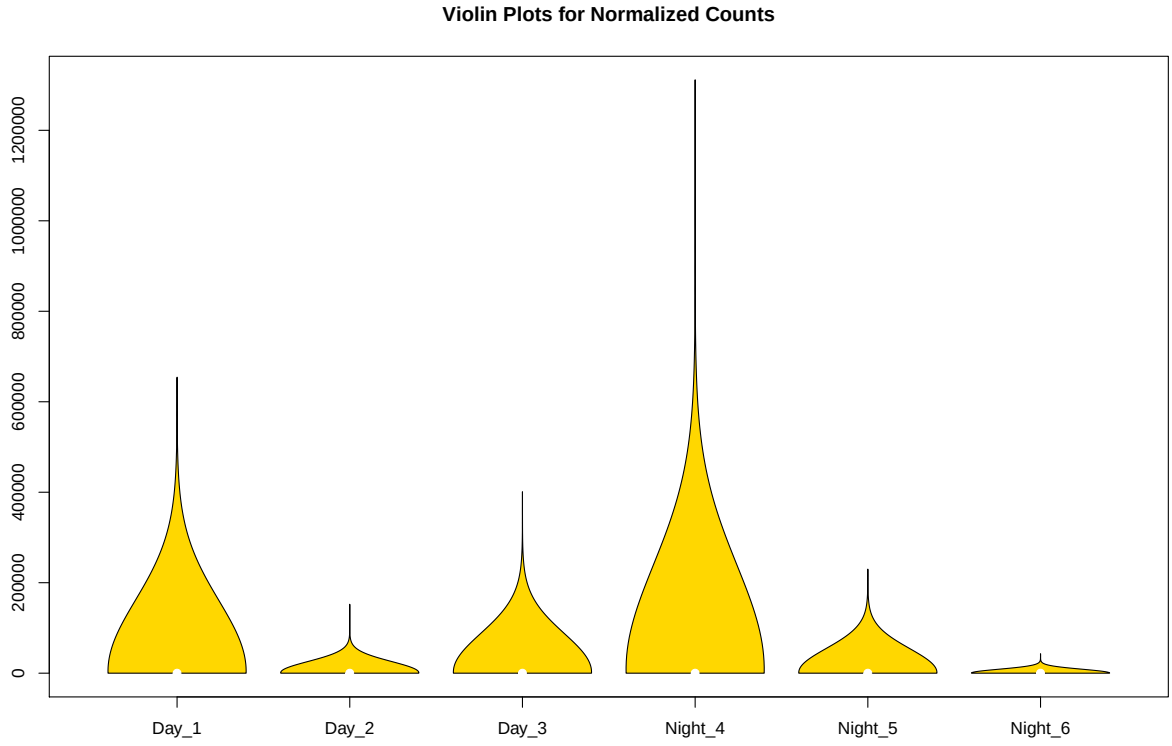
Counts

In order to normalise the raw counts we will start by determining the relative library sizes, or size factors for each library. For example, if the counts of the expressed genes in one sample are, on average, twice as high as in another, the size factor for the first sample should be twice as large as the one for the other sample. These size factors can be obtained with the function `estimateSizeFactors`:

```
##     Day_1     Day_2     Day_3   Night_4   Night_5   Night_6
## 1.6219205 0.6645299 0.5923596 0.3384058 2.2753012 2.5523719
```

Normalized Counts

```
##                                Day_1    Day_2    Day_3  Night_4   Night_5
## scaffold_0__MIS_10000001.1   0.000000 1.504823 0.000000 2.955032 0.0000000
## scaffold_0__MIS_10000001.118 0.000000 1.504823 0.000000 0.000000 0.4395022
## scaffold_0__MIS_10000001.165 3.699318 1.504823 8.440818 8.865096 8.3505426
## scaffold_0__MIS_10000001.169 0.616553 0.000000 0.000000 0.000000 4.3950224
## scaffold_0__MIS_10000001.177 1.233106 0.000000 0.000000 0.000000 3.5160179
## scaffold_0__MIS_10000001.207 0.000000 0.000000 0.000000 0.000000 0.4395022
##                                Night_6
## scaffold_0__MIS_10000001.1   3.5261319
## scaffold_0__MIS_10000001.118 1.1753773
## scaffold_0__MIS_10000001.165 1.9589622
## scaffold_0__MIS_10000001.169 0.0000000
## scaffold_0__MIS_10000001.177 0.0000000
## scaffold_0__MIS_10000001.207 0.3917924
```

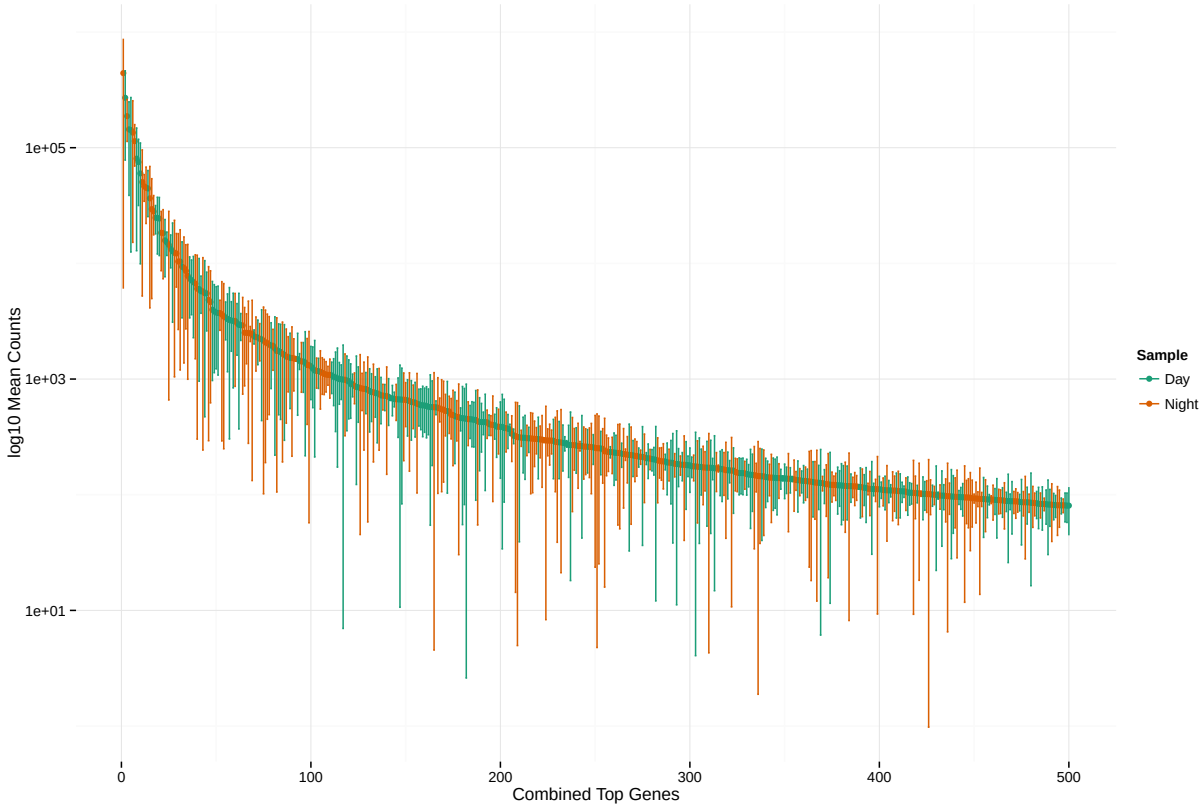


```
##      Day_1              Day_2              Day_3          
##  Min.   :     0.0   Min.   :     0.0   Min.   :     0.0  
##  1st Qu.:     0.0   1st Qu.:     0.0   1st Qu.:     0.0  
##  Median :     0.6   Median :     0.0   Median :     0.0  
##  Mean   :    63.1   Mean   :    27.5   Mean   :    43.5  
##  3rd Qu.:     1.8   3rd Qu.:     1.5   3rd Qu.:     1.7  
##  Max.   :654306.4   Max.   :152366.4   Max.   :400933.8  
##    Night_4            Night_5            Night_6        
##  Min.   :      0.0   Min.   :     0.00   Min.   :    0.00  
##  1st Qu.:      0.0   1st Qu.:     0.00   1st Qu.:    0.00  
##  Median :      0.0   Median :     0.44   Median :    0.39  
##  Mean   :    122.1   Mean   :    25.86   Mean   :   14.06  
##  3rd Qu.:      3.0   3rd Qu.:     1.32   3rd Qu.:    1.18  
##  Max.   :1311227.4   Max.   :230015.70   Max.   :43083.85
```

Rank Abundance

Plotting Rank Abundance for top 500 genes.

Day vs Night



Differential Expression

Removing Batch Effects

Differential expression was calculater using the DESeq2 wrapper function over 4 processors.

```
## Number of significant surrogate variables is:  2
## Iteration (out of 5 ):1  2  3  4  5
```

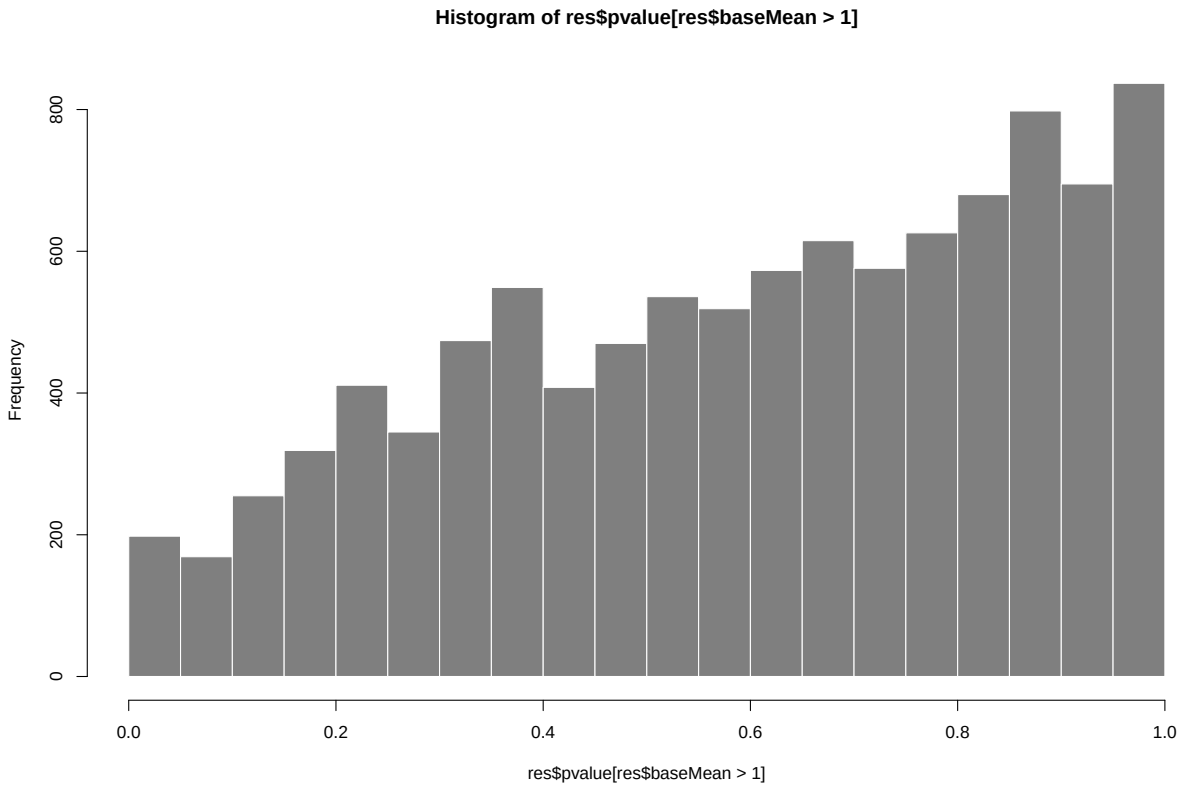
Results after removing batch effects

```
## DataFrame with 6 rows and 2 columns
##                         type                                     description
##                  <character>                                     <character>
## baseMean       intermediate      mean of normalized counts for all samples
## log2FoldChange      results log2 fold change (MAP): condition Day vs Night
## lfcSE               results         standard error: condition Day vs Night
## stat                results         Wald statistic: condition Day vs Night
## pvalue              results      Wald test p-value: condition Day vs Night
## padj                results                           BH adjusted p-values
```

```
##
## out of 26544 with nonzero total read count
## adjusted p-value < 0.1
## LFC > 0 (up)     : 53, 0.2%
## LFC < 0 (down)   : 6, 0.023%
## outliers [1]     : 0, 0%
## low counts [2]   : 25216, 95%
## (mean count < 9.9)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

P-Value Histogram

Another useful diagnostic plot is the histogram of the p values.



Significant Genes

Number of genes found to have significant differential expression:

```
## [1] 59
```

We subset the results table to these genes and then sort it by the log2 fold change estimate to get the significant genes with the strongest down-regulation.

```
## log2 fold change (MAP): condition Day vs Night
## Wald test p-value: condition Day vs Night
## DataFrame with 6 rows and 6 columns
##                                   baseMean log2FoldChange     lfcSE
##                                  <numeric>      <numeric> <numeric>
## scaffold_247603__MIS_1019830.5    48.50609      -4.654618 1.2542686
## scaffold_762984__MIS_10200131.2 7639.41777      -4.050471 1.0906161
## scaffold_39942__MIS_10004005.1   276.30102      -3.494161 1.0314812
## scaffold_83360__MIS_10039751.14  183.11743      -2.972512 0.9878613
## scaffold_762984__MIS_10200131.6  180.99322      -2.867104 0.9681510
## scaffold_55518__MIS_10017391.9   420.66374      -2.208370 0.7095335
##                                       stat       pvalue        padj
##                                  <numeric>    <numeric>   <numeric>
## scaffold_247603__MIS_1019830.5  -3.711022 0.0002064242 0.009452804
## scaffold_762984__MIS_10200131.2 -3.713929 0.0002040664 0.009452804
## scaffold_39942__MIS_10004005.1  -3.387518 0.0007052806 0.025313855
## scaffold_83360__MIS_10039751.14 -3.009037 0.0026207678 0.071028157
## scaffold_762984__MIS_10200131.6 -2.961422 0.0030622166 0.077879124
## scaffold_55518__MIS_10017391.9  -3.112425 0.0018555693 0.052429703
```

. . . and with the strongest upregulation.

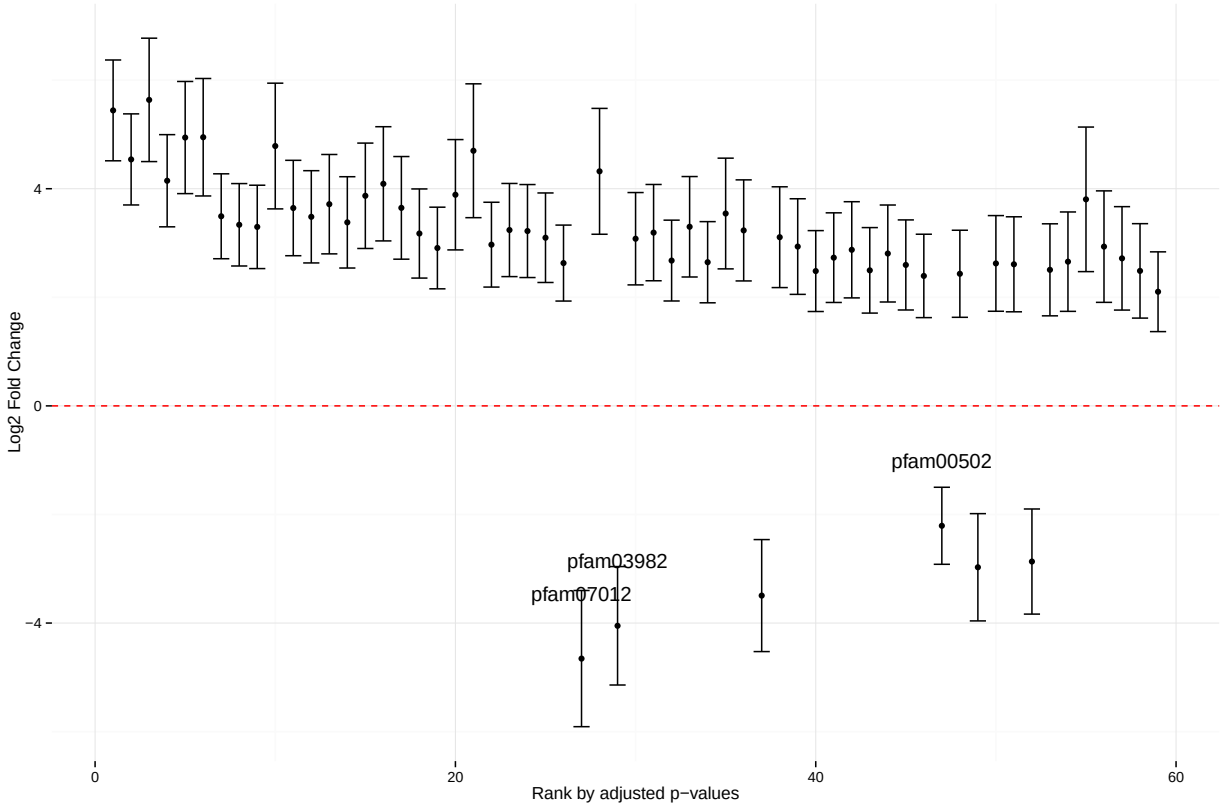
```
## log2 fold change (MAP): condition Day vs Night
## Wald test p-value: condition Day vs Night
## DataFrame with 6 rows and 6 columns
##                                    baseMean log2FoldChange     lfcSE
##                                   <numeric>      <numeric> <numeric>
## scaffold_200891__MIS_10160517.7 202.26893       5.636189 1.1364677
## scaffold_12010__MIS_10012011.12 158.18084       5.442530 0.9284052
## scaffold_12010__MIS_10012011.1   45.37752       4.948319 1.0818309
## scaffold_181056__MIS_10140692.2 747.87516       4.942481 1.0328143
## scaffold_42417__MIS_10006030.1   34.06929       4.786238 1.1576507
## scaffold_34818__MIS_10034794.1   13.99243       4.699287 1.2326452
##                                       stat       pvalue         padj
##                                  <numeric>    <numeric>    <numeric>
## scaffold_200891__MIS_10160517.7  4.959392 7.071417e-07 3.130281e-04
## scaffold_12010__MIS_10012011.12  5.862236 4.566760e-09 6.064657e-06
## scaffold_12010__MIS_10012011.1   4.574023 4.784483e-06 1.058966e-03
## scaffold_181056__MIS_10140692.2  4.785450 1.706050e-06 4.531270e-04
## scaffold_42417__MIS_10006030.1   4.134441 3.558200e-05 4.295718e-03
## scaffold_34818__MIS_10034794.1   3.812360 1.376461e-04 8.696156e-03
```

Summary Table

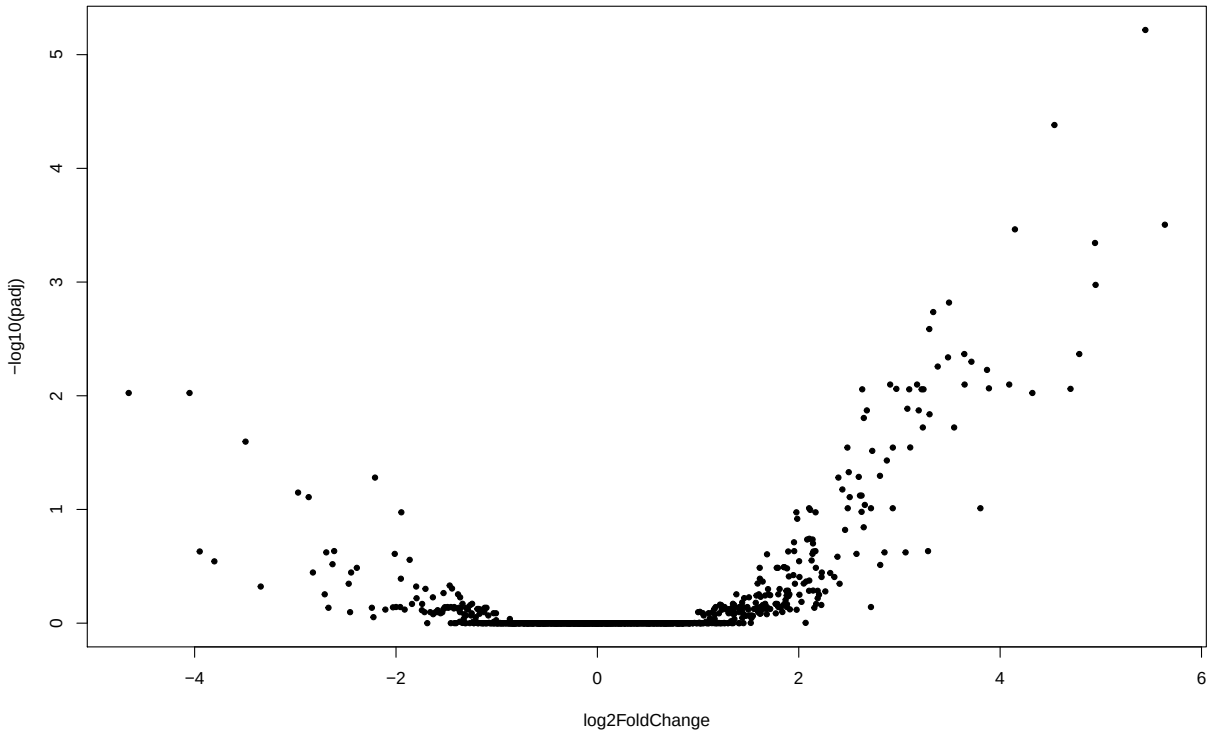
Name	log2FoldChange	padj	IMG_Product	IMG_Source
scaffold_12010___MIS_10012011.12	5.442530	0.0000061	NA	NA
scaffold_7608___MIS_10007609.1	4.539691	0.0000416	NA	NA
scaffold_200891___MIS_10160517.7	5.636189	0.0003130	NA	NA
scaffold_356736___MIS_10195009.1	4.146921	0.0003440	NA	NA
scaffold_181056___MIS_10140692.2	4.942481	0.0004531	NA	NA
scaffold_12010___MIS_10012011.1	4.948319	0.0010590	NA	NA
scaffold_356736___MIS_10195009.4	3.492412	0.0015148	NA	NA
scaffold_430758___MIS_10197677.1	3.335744	0.0018350	NA	NA
scaffold_19008___MIS_10019009.27	3.296748	0.0025895	NA	NA
scaffold_12010___MIS_10012011.3	3.644443	0.0042957	NA	NA
scaffold_42417___MIS_10006030.1	4.786238	0.0042957	NA	NA
scaffold_195616___MIS_10155245.4	3.482549	0.0045960	NA	NA
scaffold_41583___MIS_10036250.27	3.715213	0.0050183	NA	NA
scaffold_356736___MIS_10195009.2	3.380150	0.0055344	NA	NA
scaffold_12010___MIS_10012011.7	3.869782	0.0059244	NA	NA
scaffold_162064___MIS_10121714.1	4.090441	0.0079665	NA	NA
scaffold_19008___MIS_10019009.26	3.647369	0.0079665	NA	NA
scaffold_19008___MIS_10019009.28	2.907916	0.0079665	NA	NA
scaffold_232359___MIS_10185986.7	3.174997	0.0079665	NA	NA
scaffold_163615___MIS_10123264.4	3.888927	0.0086048	NA	NA
scaffold_242878___MIS_10186954.9	2.969437	0.0086962	NA	NA
scaffold_34818___MIS_10034794.1	4.699287	0.0086962	NA	NA
scaffold_133743___MIS_10093445.6	3.220367	0.0087786	NA	NA
scaffold_19008___MIS_10019009.29	2.630435	0.0087786	NA	NA
scaffold_6544___MIS_10006545.32	3.239039	0.0087786	NA	NA
scaffold_748781___MIS_10200104.4	3.097421	0.0087786	NA	NA
scaffold_23846___MIS_10023847.20	4.320381	0.0094528	NA	NA
scaffold_247603___MIS_1019830.5	-4.654618	0.0094528	Curlin associated repeat	pfam07012
scaffold_762984___MIS_10200131.2	-4.050471	0.0094528	Diacylglycerol acyltransferase	pfam03982
scaffold_12010___MIS_10012011.8	3.078971	0.0129999	NA	NA
scaffold_12010___MIS_10012011.4	3.191382	0.0134504	NA	NA
scaffold_140713___MIS_10100395.5	2.676440	0.0134504	NA	NA
scaffold_36112___MIS_10001001.11	3.299193	0.0145517	NA	NA
scaffold_425595___MIS_1167750.1	2.646232	0.0156805	NA	NA
scaffold_142012___MIS_10040323.5	3.543329	0.0190062	NA	NA
scaffold_200195___MIS_10159822.1	3.232287	0.0190062	NA	NA
scaffold_39942___MIS_10004005.1	-3.494161	0.0253139	NA	NA
scaffold_232359___MIS_10185986.6	3.107126	0.0284879	NA	NA
scaffold_276564___MIS_10189820.1	2.934461	0.0285565	NA	NA
scaffold_407786___MIS_10197014.2	2.483188	0.0285565	NA	NA
scaffold_138956___MIS_10098641.7	2.730239	0.0305307	NA	NA
scaffold_12010___MIS_10012011.15	2.874866	0.0371092	NA	NA
scaffold_230594___MIS_10185835.2	2.496583	0.0470046	NA	NA
scaffold_622706___MIS_1174630.1	2.806595	0.0506007	NA	NA
scaffold_113303___MIS_10073104.2	2.596039	0.0516996	NA	NA
scaffold_55518___MIS_10017391.9	-2.208370	0.0524297	Phycobilisome protein	pfam00502
scaffold_778707___MIS_1178057.1	2.393790	0.0524297	NA	NA
scaffold_47856___MIS_10037233.11	2.433313	0.0667004	NA	NA
scaffold_83360___MIS_10039751.14	-2.972512	0.0710282	NA	NA
scaffold_564989___MIS_10199597.1	2.607946	0.0754694	NA	NA

Name	log2FoldChange	padj	IMG_Product	IMG_Source
scaffold_654999___MIS_10199933.4	2.624083	0.0754694	NA	NA
scaffold_34818___MIS_10034794.3	2.506195	0.0778791	NA	NA
scaffold_762984___MIS_10200131.6	-2.867104	0.0778791	NA	NA
scaffold_92040___MIS_10052095.5	2.656372	0.0912172	NA	NA
scaffold_34818___MIS_10034794.2	2.486534	0.0975790	NA	NA
scaffold_407786___MIS_10197014.1	2.102636	0.0975790	NA	NA
scaffold_5754___MIS_10005755.3	3.804303	0.0975790	NA	NA
scaffold_65654___MIS_10026723.10	2.717280	0.0975790	NA	NA
scaffold_74783___MIS_10035344.4	2.933591	0.0975790	NA	NA

Log2Fold vs Rank(p-value adjusted) Plot



Volcano (log10(padj) vs log2FoldChange)



Session Info

```
## R version 3.2.1 (2015-06-18)
## Platform: x86_64-apple-darwin13.4.0 (64-bit)
## Running under: OS X 10.10.4 (Yosemite)
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] parallel  stats4    stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] knitr_1.10.5              sva_3.14.0
##  [3] genefilter_1.50.0         mgcv_1.8-7
##  [5] nlme_3.1-121              BiocParallel_1.2.9
##  [7] ggplot2_1.0.1             dplyr_0.4.2
##  [9] tidyr_0.2.0               vioplot_0.2
## [11] sm_2.2-5.4                DESeq2_1.8.1
## [13] RcppArmadillo_0.5.200.1.0 Rcpp_0.12.0
## [15] GenomicRanges_1.20.5      GenomeInfoDb_1.4.1
## [17] IRanges_2.2.5             S4Vectors_0.6.2
## [19] BiocGenerics_0.14.0
##
## loaded via a namespace (and not attached):
##  [1] locfit_1.5-9.1       lattice_0.20-33      assertthat_0.1
##  [4] digest_0.6.8         R6_2.1.0             plyr_1.8.3
##  [7] futile.options_1.0.0 acepack_1.3-3.3      RSQLite_1.0.0
## [10] evaluate_0.7         highr_0.5            lazyeval_0.1.10
## [13] annotate_1.46.1      rpart_4.1-10         Matrix_1.2-2
## [16] rmarkdown_0.7        proto_0.3-10         labeling_0.3
## [19] splines_3.2.1        geneplotter_1.46.0   stringr_1.0.0
## [22] foreign_0.8-65       munsell_0.4.2        htmltools_0.2.6
## [25] nnet_7.3-10          gridExtra_2.0.0      Hmisc_3.16-0
## [28] XML_3.98-1.3         MASS_7.3-43          grid_3.2.1
## [31] xtable_1.7-4         gtable_0.1.2         DBI_0.3.1
## [34] magrittr_1.5         formatR_1.2          scales_0.2.5
## [37] stringi_0.5-5        XVector_0.8.0        reshape2_1.4.1
## [40] latticeExtra_0.6-26  futile.logger_1.4.1  Formula_1.2-1
## [43] lambda.r_1.1.7       RColorBrewer_1.1-2   tools_3.2.1
## [46] Biobase_2.28.0       survival_2.38-3      yaml_2.1.13
## [49] AnnotationDbi_1.30.1 colorspace_1.2-6     cluster_2.0.3
```