COBB 2010 - Foundations of Computational Biology

Instructor: T. Lezon

Homework 6 - Due Thursday, October 1, 2020 at 1:45PM EDT

Arabidopsis is a commonly used model organism. There is a preliminary result of the sequencing analysis of *Arabidopsis thaliana*, and please conduct further analysis associating the questions.

1. Please read the following data into a data frame, and get read of NA values. **Data source**: https://raw.githubusercontent.com/jmzeng1314/my-R/master/DEG_scripts/tair/DESeq2_DEG.Day1-Day0.txt
2. By applying *org.At.tair.db* package. Based on the manual, transfer the original *gene_id* into gene symbol.
   - one may notice some original gene ids in the file have no corresponding gene symbols. This is because these genes have little biological meaning or they have not been annotated yet. Try to count the number of these ids.
   - Meanwhile, some original ids have more than one gene symbol due to they were named by the different research centers. Complete the gene id convert with an output like below and export the data frame into a *.csv* file:

```
  gene_id   baseMean log2FoldChange ...           symbol
AT2G33830 1938.15972      -2.560609 ...        AtDRM2/DRM2
AT2G33750    9.78974     -19.989301 ...        ATPUP2/PUP2
AT3G54500 2238.31465       2.720430 ...               LNK2
AT2G46830  169.50508       3.527082 ...        AtCCA1/CCA1
AT1G28330 1636.02224      -1.493341 ... AtDRM1/DRM1/DYL1
AT1G48598   39.84136      -9.779030 ...           CPuORF31
```

3. Find out the top 10 significantly expressed genes by setting *padj* value less than 0.05.
   - Generate a barplot of `baseMean` of the top 10 most significantly expressed genes that have symbol names. To make it visually compared easily, one may want to use logarithmic value, and show gene symbols instead of gene ids.
   - Generate a dot plot of the `log2FoldChange` data in increasing order, and show the error bar as log folder change standard error (lfcSE).
4. Isolate all of the deferentially expressed genes (padj < 0.05) into a new data frame, and count the number of significantly up/down-regulated genes.
5. Plot the relationship between `log2FoldChange` and -log(`padj`), with highlighting the significantly up/down-regulated genes in the figure (Volcano plot).

Please complete the above tasks and generate a report with R markdown, and show all the code written in chunks (it is unnecessary to show outcomes of each chunk, one can decide to show which part). One should submit a *.rmd* file.