

RNA-Seq Differential Expression

Jessica Murphy

March 29, 2019

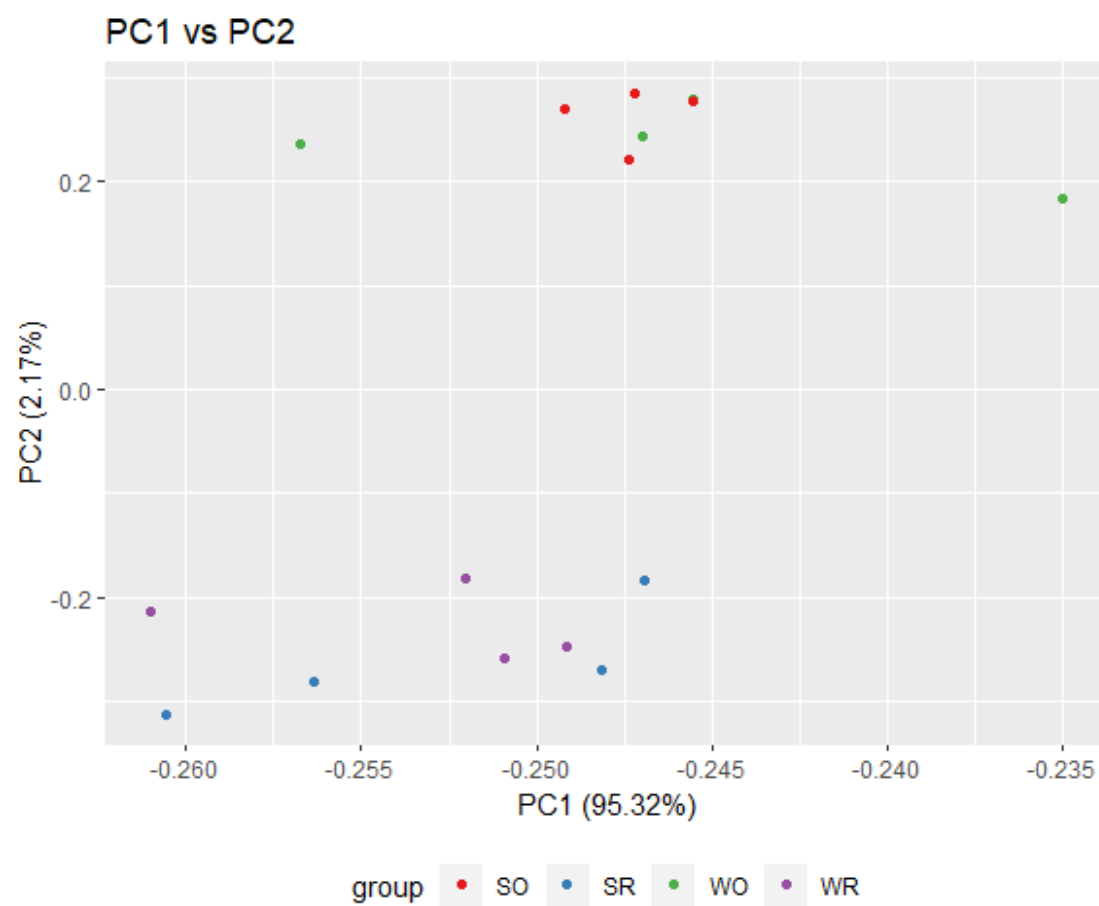
Overview

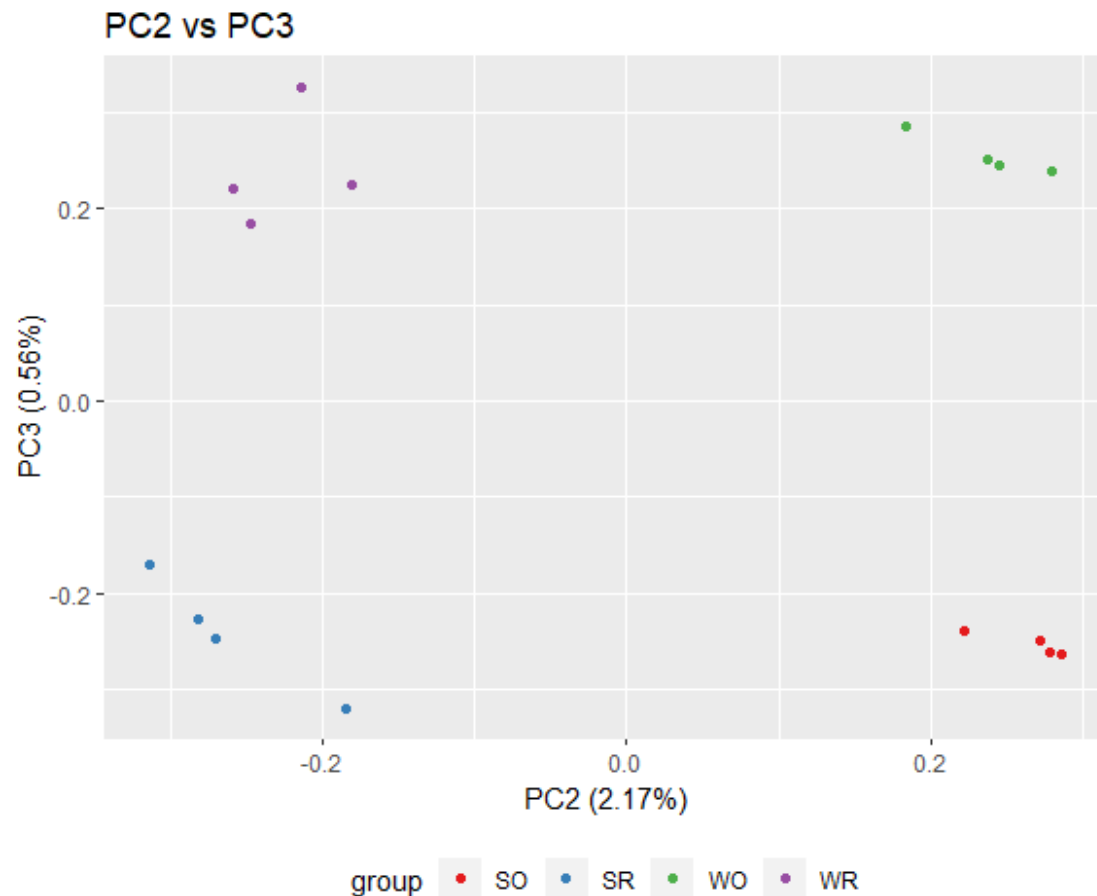
The following data was obtained from a 2x2 factorial design using fly models. The two variables of interest are strain and tissue. The strains are white eyed (W) and sevenless (S) and the tissues are optic lobe (O) and retina (R). There are four samples in each tissue/strain combination.

The dataset contains a matrix of RSEM expected counts for 17,453 genes and a matrix of phenotypes to accompany the counts.

Visualize Data

Principal component analysis (PCA) was initially performed on the data and the PC plots of the 16 samples are shown below.





The PC1 vs PC2 plot shows a large separation in tissue type on PC2. Also, a very large percent of the variance is explained by PC1. The PC2 vs PC3 plot shows a great separation in groups. PC2 describes tissue differences (2nd letter in group) and PC3 describes strain differences (1st letter in group).

Perform Statistical Test

A likelihood ratio test (LRT) was then performed using DESeq2 to determine which genes have a significant strain effect regardless of dependence/independence on tissue.

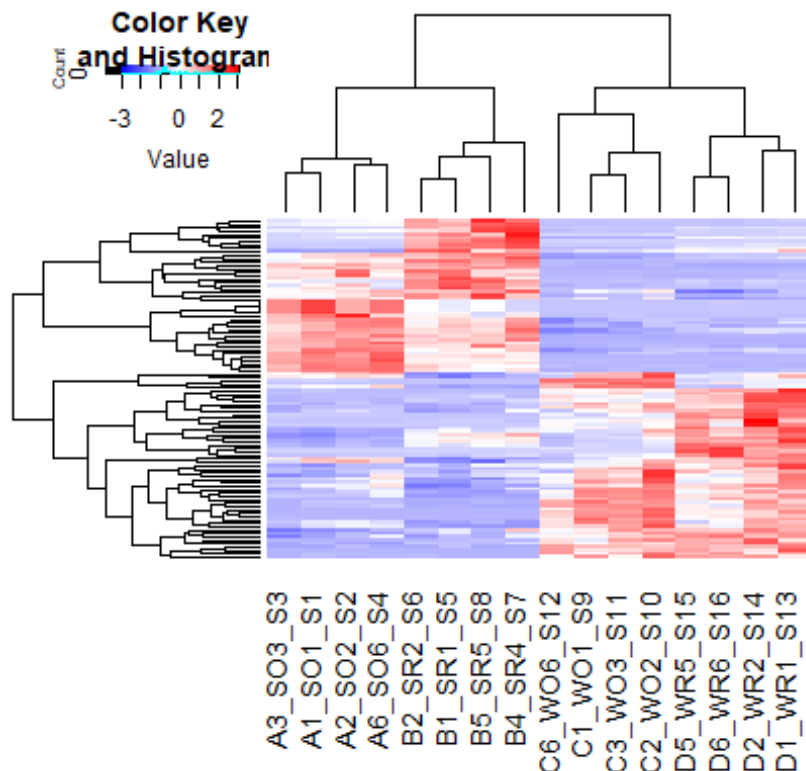
Based on an adjusted p-value < 0.05, there were 2,098 gene candidates. Since this is such a large amount, a stricter criteria was applied. Based on an adjusted p-value < 0.001, there are now 878 gene candidates. The top 10 candidates are listed below.

	Base Mean	log2 Fold Change	Adjusted p-value
FBgn0052475	846.31527	2.0922266	0.000000e+00
FBgn0267160	770.14610	1.4761384	4.408069e-289
FBgn0050428	177.59063	-3.6259762	8.239672e-174
FBgn0051865	86.33801	0.5608518	9.706659e-145
FBgn0052011	159.10743	1.4197483	1.499746e-132

FBgn0051676	2087.10410	0.4941369	6.383283e-109
FBgn0037228	1543.28030	-0.0275406	2.984941e-99
FBgn0040524	85.36571	3.3180130	4.973883e-91
FBgn0015714	66.40389	2.3278783	3.621200e-88
FBgn0058469	4589.87052	0.4486818	1.558954e-83

Visualize Candidates

The following heatmap shows the top 100 candidates with the smallest adjusted p-values. This is good to see large trends, but difficult to see individual genes.



Perform Enrichment Analysis

Enrichment analysis was then performed using enrichR to determine if the previous 100 candidates are part of any known biological pathway or system.

The flybase gene IDs were first converted to 64 human associated genes. Using the GO Biological Process 2018 database, the top 10 processes with the smallest adjusted p-values are listed below.

Process	Overlap	Adjusted p-value
transmembrane transport (GO:0055085)	19/378	2.334847e-14
cholesterol efflux (GO:0033344)	3/23	8.993247e-02

sterol transport (GO:0015918)	3/21	1.017712e-01
cellular response to ketone (GO:1901655)	3/28	1.233515e-01
short-chain fatty acid catabolic process (GO:0019626)	2/8	2.845073e-01
cholesterol transport (GO:0030301)	3/44	3.205358e-01
drug transmembrane transport (GO:0006855)	2/11	3.967169e-01
ER-associated misfolded protein catabolic process (GO:0071712)	2/13	4.902621e-01
cellular response to glucocorticoid stimulus (GO:0071385)	2/17	7.535940e-01
drug transport (GO:0015893)	2/19	8.492720e-01