Ned Brewer
STAT 555 StatGenomics
12/13/19

Final Project: Nebulin-KO mice and tissue-specific differential gene expression

Introduction

Nebulin is a filamentous protein that partners with the actin filaments of the skeletal muscle sarcomere. Nebulin mutations are the main cause of nemaline myopathy, which is a hereditary neuromuscular disorder that can cause muscle weakness, difficulty swallowing, and impaired speech. In a 2015 publication in *Human Molecular Genetics, Li et al*[1] created adult-only nebulin-KO mice lines, and performed various studies on their muscle mass, muscle function, and the makeup of the muscle. They revealed that nebulin is incredibly important for muscle force development and trophicity. However, aside from stating in the supplemental figures that a few ubiquitin ligase genes were upregulated, DEG analysis wasn't the focal point of their study. Using the data that they published on the NCBI's Gene Expression Omnibus, I dove deeper into the implications of the gene expression data, in both the quadriceps and soleus muscles, of knocking out the nebulin gene. Their data had 24 samples, consisting of 6 replicates each of controls and knock-outs for the quadriceps and soleus muscles. I performed DEG analysis of several contrasts of the sample groupings using *limma*, and used Storey's method to account for the FDR. I then created Venn diagrams that show the overlap of DEGs between the various contrasts and bar charts to show the comparative number of significant DEGs. Then, using the *goSTAG* program, I performed gene ontology analysis for molecular functions, cellular components, and biological processes to find out the types of genes that were differentially expressed. Lastly, I performed a principal components analysis (PCA) with the 24 samples in order to discover whether the muscle type or the treatment type (control/knockout) was more determinative of the sample-specific expression values.

Materials and Methods

Original Data

I obtained my Affymetrix Mouse Gene 1.0 ST Array microarray data from the NCBI's Gene Expression Omnibus, available at this link: https://www.ncbi.nlm.nih.gov/geo/download/?acc=GSE70213&format=file

Annotation Files

http://www.informatics.jax.org/downloads/reports/Affy_1.0_ST_mgi.rpt

Programs and packages in R Studio

oligo[2]	hexbin[3]	qvalue[4]	limma[5]
dplyr[6]	tidyverse[7]	statmod[8]	data.table[9]
goSTAG[10]	devtools[11]	ggbiplot[12]	

Results

Differential Gene Expression

In the original paper, *Li et al.* did a differential expression analysis and reported the results for a few genes that they were interested in the supplemental figures. I decided to re-do the analysis and look at differential expression across the various contrasts between groups of the 24 samples with the help of *limma* and *qvalue*. The contrasts included nebulin-KO vs control across all muscle types (KOvsC), KO vs control for the quadriceps samples (QuadKOvsC), the same for the soleus samples (SolKOvsC), a comparison of the quadriceps control vs the soleus control to determine DEGs between muscle types (QuadCvsSolC) and then the interaction (Interact) between muscle type and treatment in this fashion: (QuadKO - SolKO) - (QuadC - SolC), and the results are shown in ***Fig. 1***. My analysis shows that KOvsC, QuadKOvsC, and QuadCvsSolC have more than 12,000 DE probesets (out of 35,556

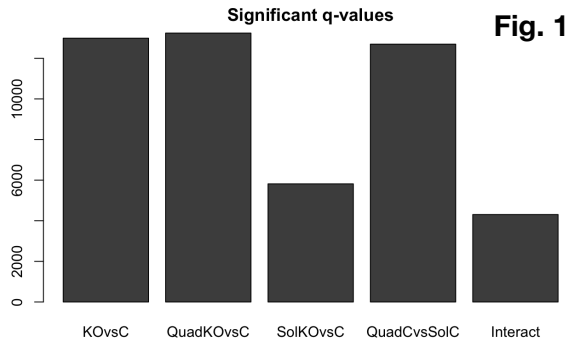


Fig. 1

total probesets). Then, I filtered the probesets for genes that were in the annotation file from the Mouse Genome Informatics server[13] in order to proceed. By using the list of probesets with gene counterparts, I could then take the p-values and create a correlation matrix for each of the five contrasts with each other. The results in ***Table 1*** show that the muscle-specific KOvsC comparisons (QuadKOvsC and SolKOvsC) each correlate highly with the global KOvsC, but the correlation is weaker with each other. Furthermore, the 3 KOvsC contrasts don't correlate very well at all with the muscle type contrast (QuadCvsSolC) or the muscle-treatment interaction contrast (Interact). The only correlations that are above .5 are the quadricep and soleus specific contrasts each with the global KOvsC contrast.

Table 1	KOvsC	QuadKOvsC	SolKOvsC	QuadCvsSolC	Interact
KOvsC	1.0000000	0.8593513	0.8030215	0.2774271	0.3029186
QuadKOvsC	0.8593513	1.0000000	0.4845495	0.3357547	0.5355737
SolKOvsC	0.8030215	0.4845495	1.0000000	0.2560008	0.2394049
QuadCvsSolC	0.2774271	0.3357547	0.2560008	1.0000000	0.4680488
Interact	0.3029186	0.5355737	0.2394049	0.4680488	1.0000000

In light of this, I decided to create a Venn diagram of DEGs between the global KOvsC contrast and the tissue-specific KOvsC contrasts. The results in ***Fig 2*** are that there are 3766 shared DEGs between all 3, 4030 between the global and quadricep contrast, 1912 between the global and soleus contrasts, and not surprisingly, given the middling correlation, that there are only 263 shared DEGs between the soleus and quadricep contrasts. Another Venn diagram that I did (in the Supplementary file) between the quadricep and soleus KOvsC contrasts and the QuadCvsSolC contrast showed that there were 1818 DEGs that were significant between the QuadKOvsC and SoleusKOvsC contrasts but not significant in the QuadCvsSolC contrast. This means that the differential expression was cause specifically by the nebulin knockout, not the tissue type. I stored the genes, probeset IDs, and p-values in the "AllP_vennQvS_therapeutic_targets" variable, but I did not have time to perform GO analysis on

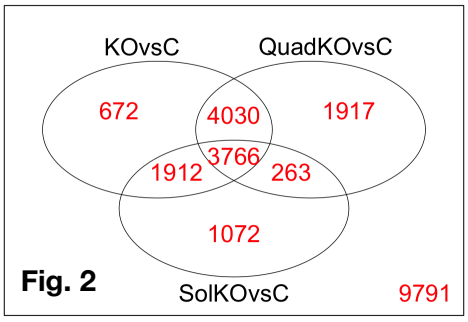


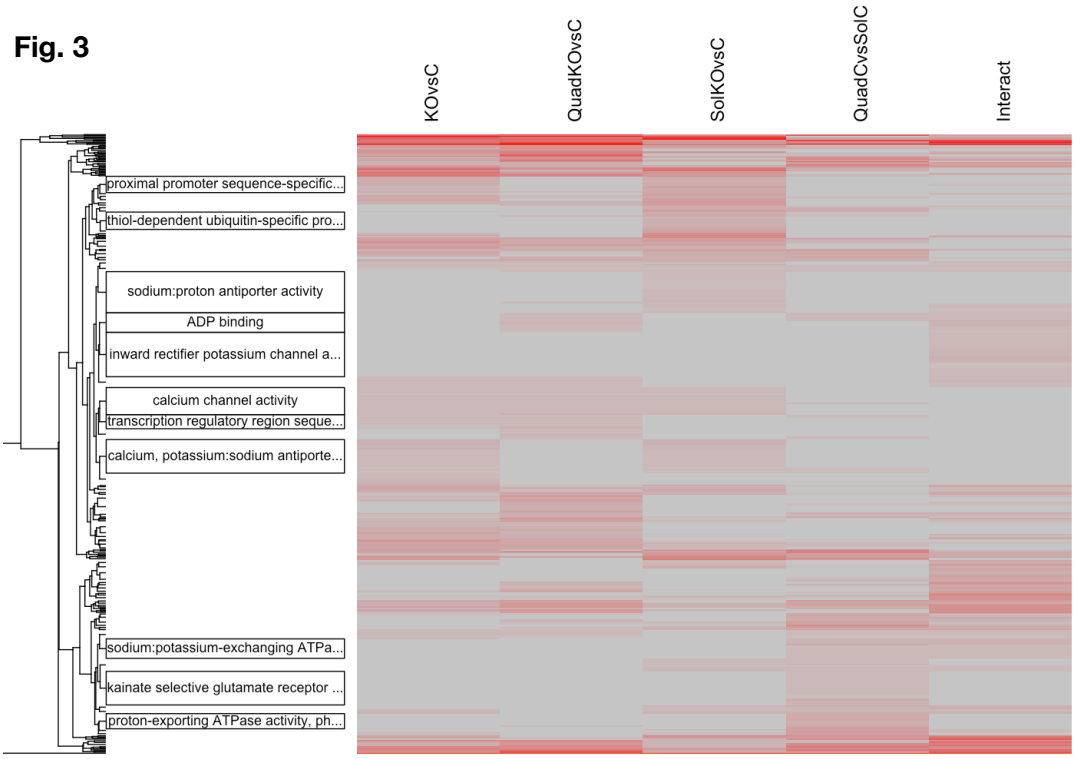
Fig. 2

them. They would seem to be interesting therapeutic targets, and they did not seem to be mentioned in the original paper.

GO Analysis

My next portion of the project was to use *goSTAG* to perform a gene ontology analysis. To do this, I ordered the genes based on qvalue (lowest first), and picked the top 1000 genes for each contrast. I then ran 3 different GO analyses, based on molecular function, biological process, and cellular components. For space saving reasons, I will only show the molecular function GO heat map as it is the most relevant (***Fig 3.***) The minimum number of terms for this graph is 13, meaning all labels correspond to GO groups that have 13 or more terms that show up in my datasets. A lot of them correspond to calcium channels and ATP/ADP activity, which makes sense given that nebulin has has also been shown to regulate actin-myosin interactions by inhibiting ATPase activity in a calcium-calmodulin sensitive manner[14]. This is exemplified by there being red areas in the calcium channel activity label under all 3 KOvsC contrasts, but not the QuadCvsSolC, which had no nebulin knockout samples included. Overall though, the GO analysis did not give me as strong of results as I was expecting.

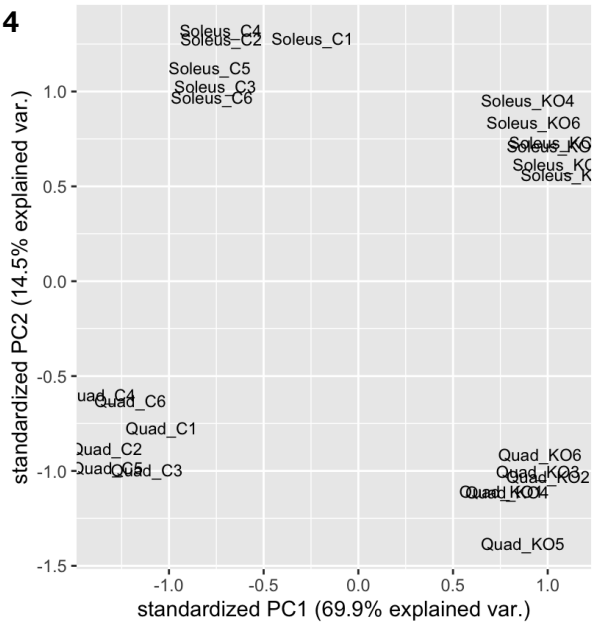
Fig. 3



Principal Components Analysis

My final analysis was a principal components analysis (PCA), using the *prcomp* function from the built-in R package *stats.* In this analysis, I wanted to see if muscle type was more predictive of gene expression variation than the presence of a nebulin knockout. PCAs reduce multi-variable datasets to the “principal components” or components that can explain most of the variation. I chose to use the first 2 components, as they explained over 84% of the variation within the dataset. To do this, I selected the 1000 most significant genes from the KOvsC contrast, and looked at their normalized expression value across the 24 samples, then ran a PCA in ***Fig 4***. The results were mostly what one would expect, with each replicate grouped very closely with other replicates. But the PC1, which explains 69.9% of the variance, seems to mostly represent the nebulin knockout, as the KO samples are on the right and the C samples on the left. The PC2, explaining 14.5% of the variance, seems to represent the muscle type, with soleus samples up top and quadricep samples on the bottom. This result is interesting, given the fact that the muscle type KOvsC contrasts in Table 1 did not correlate as much with each other as they did with the global KOvsC contrast.

Fig. 4



Discussion

These results portray a conflicted picture of what drives differential expression in this dataset, with the correlational matrix and Venn diagrams suggesting that muscle tissue determines the majority of expression, and the PCA analysis suggesting that it might be driven more by the knockout of nebulin, regardless of the muscle type. Gene expression is a complex process, and it is intuitive that the factors that control it are also complex. The therapeutic targets that I mentioned previously

are a good launch point for future investigations into the cause, mechanism, and pathways of nebulin dysfunction, and subsequently, therapeutic relief for nemaline myopathy patients.

References

1. Frank Li, Danielle Buck, Josine De Winter, Justin Kolb, Hui Meng, Camille Birch, Rebecca Slater, Yael Natelie Escobar, John E. Smith, Lin Yang, John Konhilas, Michael W. Lawlor, Coen Ottenheijm, Henk L. Granzier, Nebulin deficiency in adult muscle causes sarcomere defects and muscle-type-dependent changes in trophicity: novel insights in nemaline myopathy, *Human Molecular Genetics*, Volume 24, Issue 18, 15 September 2015, Pages 5219–5233, https://doi.org/10.1093/hmg/ddv243
2. Carvalho BS, Irizarry RA (2010). “A Framework for Oligonucleotide Microarray Preprocessing.” *Bioinformatics*, **26**(19), 2363-7. ISSN 1367-4803, doi: 10.1093/bioinformatics/btq431.
3. https://cran.r-project.org/package=hexbin
4. Storey JD, Bass AJ, Dabney A, Robinson D (2019). *qvalue: Q-value estimation for false discovery rate control*. R package version 2.18.0, http://github.com/jdstorey/qvalue.
5. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015). “limma powers differential expression analyses for RNA-sequencing and microarray studies.” *Nucleic Acids Research*, **43**(7), e47. doi: 10.1093/nar/gkv007.
6. https://www.rdocumentation.org/packages/dplyr
7. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). “Welcome to the tidyverse.” *Journal of Open Source Software*, **4**(43), 1686. doi: 10.21105/joss.01686.
8. Giner, G, and Smyth, GK (2016). statmod: probability calculations for the inverse Gaussian distribution. R Journal 8(1), 339-351. (pinvgauss, qinvgauss, dinvguass and rinvgauss functions)
9. https://rdatatable.gitlab.io/data.table/
10. Bennett BD, Bushel PR (2017). “goSTAG: Gene Ontology Subtrees to Tag and Annotate Genes within a set.” *Source Code Biol Med*.
11. https://cran.r-project.org/web/packages/devtools/index.html
12. https://www.rdocumentation.org/packages/ggbiplot/versions/0.55
13. http://www.informatics.jax.org/
14. Root DD, Wang K (Oct 1994). "Calmodulin-sensitive interaction of human nebulin fragments with actin and myosin". *Biochemistry*. **33** (42): 12581–91. doi:10.1021/bi00208a008. PMID 7918483.