Preliminary Data Analysis:

Using Gene Expression Analysis to Understand Muscle Atrophy Following Peripheral Nerve Injury

BINF 5005

Integrative Research Project

November 7th, 2023

Suzan Ahmad

Tanuj Fernando

Karolina Urbanovich

Data description & preprocessing

The dataset we are using is GSE 45550 from the National Center of Biotechnology Information's (NCBI's) Gene Expression Omnibus (GEO). The organism that this dataset comes from is a Norway rat, or brown rat. The data profiles 36 samples of the rat soleus in three different experimental groups; control group, experimental group with spinal cord injury (SCI), and an experimental group with spinal cord injury and locomotor treadmill training (SCI, TM). The SCI group was profiled 3, 8, and 14 days following spinal cord injury and the SCI, TM group was profiled 8 and 14 days after spinal cord injury.

In order to begin looking into the trends shown by the data, 6 groups were defined in the dataset with 6 samples in each; control, 3 Days SCI, 8 Days SCI, 8 Days SCI TM, 14 Days SCI, and 14 Days SCI TM. The False Discovery Rate was controlled using the Benjamini-Hochberg method, thereby increasing the method's power (Haynes, 2013). The significance level cut off (i.e. the p value) was set to 0.05. In addition, log transformation was applied to the data to reduce the skewness.

RNA was extracted and amplified from muscle samplesm, and analyzed using Affymetrix microarray chips. The raw seq data was normalized using global scaling to ensure that data between arrays can be compared.

Link to our data: https://humberital-my.sharepoint.com/:u:/r/personal/n01619138_humber_ca/Documents/5005%20Project/GSE45550_RAW.tar?csf=1&web=1&e=bScsHT

Explanation of Pipeline

To analyze our data, we are using the *limma* (linear models for microarray data) analysis package. The analysis pipeline for our analysis may change, but it is currently based on the *limma* user manual, a step-by-step guide for *limma* found on *Medium* (https://blog.devgenius.io/differential-gene-expression-analysis-using-limma-step-by-step-

358da9d41c4e), as well as GEO2R. GEO2R also uses *limma* to analyze microarray data, therefore we aim to learn more about this process by analyzing the automatically generated R script this software outputs.

Our data arrives in Affymetrix CEL files, which contain data on the intensity of the reads from the microarray flow cells. The data is sorted into groups based on the treatment each sample received. Since our data is already normalized, we skip the normalization step straight to fitting our data to the model.

Our model returns a number of results, including the log fold change values and average expression values for our probes. We hope to visualize these results to further display the differing expression levels we've found (heatmaps, venn diagrams, volcano plots).

Code

Please see attached file, it contains the automated output from GEO2R used for analysis and graphing.

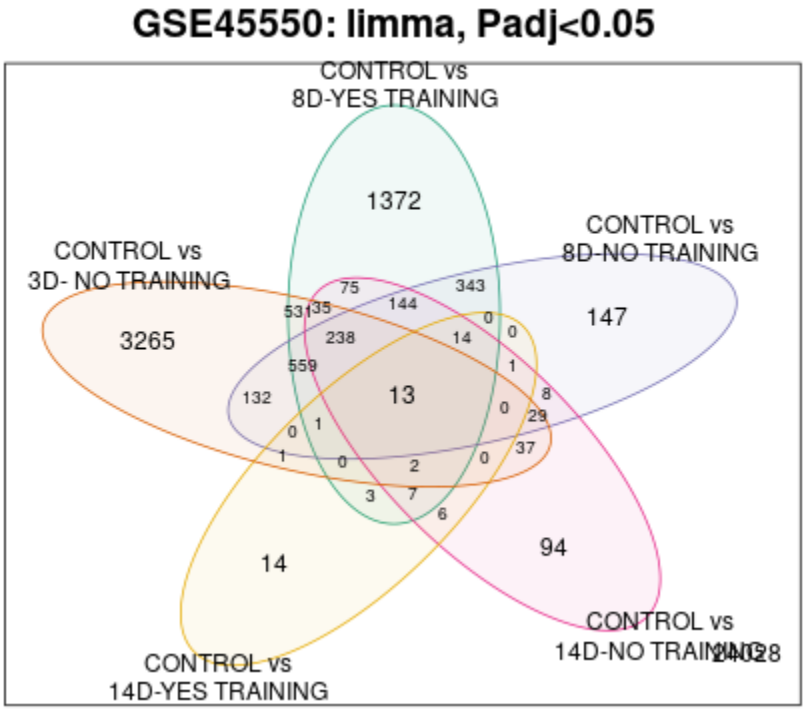
After conducting an analysis of our data (GSE45550) using GEO2R, we focused on the comparison between the following groups:

- CONTROL vs 3D- NO TRAINING
- CONTROL vs 8D- YES TRAINING
- CONTROL vs 8D- NO TRAINING
- CONTROL vs 14D- NO TRAINING
- CONTROL vs 14D- YES TRAINING

In the Venn diagram, it was clear that 13 genes displayed significance and were shared among all these comparisons. (Nnat, Slc25a25, Atf5, RGD1309676, LOC303590, Ppp3cb, G0s2, Acat1, Il17rc, Apobec3b, Gnl3l).

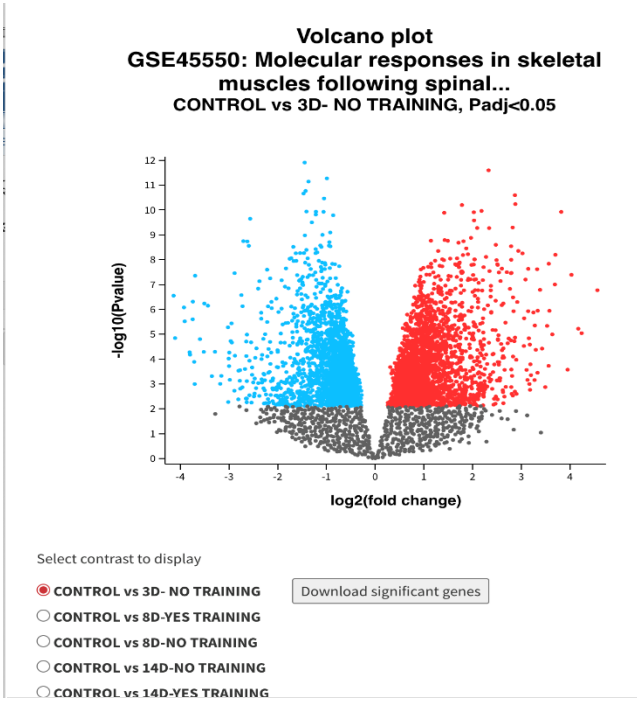
Specifically, when comparing CONTROL vs 3D- NO TRAINING, we noticed that 3,265 genes exhibited differential expression after SCI.

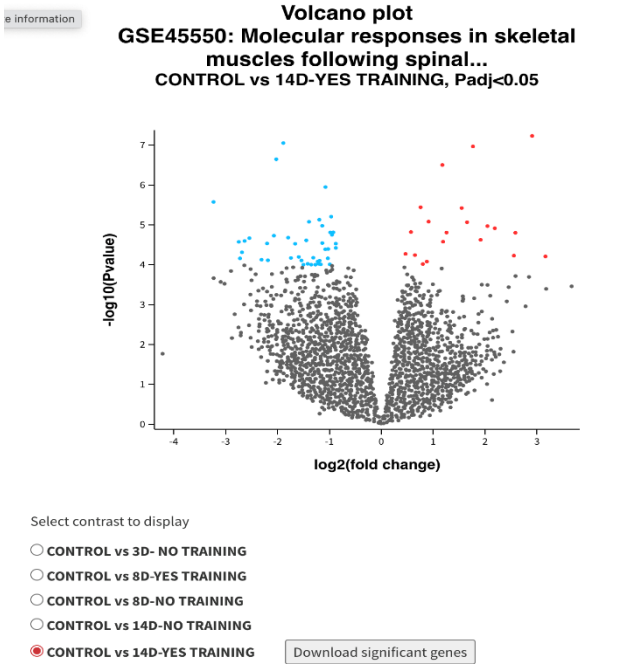
Notably, the impact of training after 8 days appeared to have a negative effect on the injury, as 1,372 genes were differentially expressed when compared to only 147 genes without training. Conversely, it was observed that training had a positive effect after 14 days, with only 14 genes being expressed with training compared to 94 genes without training.



I**n the Volcano plot**, which visualized differential gene expression, red dots represented upregulated genes (with log2(fold change > 0)), while blue dots represented downregulated genes (with log2(fold change < 0)). It was evident that genes were highly expressed immediately after the injury (3 days post-injury) when compared to the control group before the injury. However, as time progressed, gene expression increased or decreased, with the lowest expression

occurring at 14 days after the injury, and few significant changes in gene expression were observed.





Documentation of Challenges

One challenge with our project is the fact that we're using microarray data. Modern DGE studies use RNA-seq technology, which offers higher specificity and sensitivity. In turn, many modern tools are designed for this type of data. Finding tutorials and resources for dealing with *limma* was definitely hindered by its age. We also feel like we do not completely understand how to read/interpret microarray data, therefore we are unsure exactly what our models are returning to us once run our pipeplines.

For our next steps, we hope to gain a better understanding of our microarray data, run our data outside of GEO2R, and construct a clear visual explanation of our pipeline for data analysis.