Whole Genome Bioinformatic Analysis of Skin from Three Immunologically Diverse Mouse Models

Kieran Bissessar

1 Abstract

In the experiment, Whole Genome Analysis of Skin from Three Immunologically Diverse Mouse Models by Luckett-Chastain and Gallucci, expression profiling by high throughput sequencing was done on three mouse models that differ immunologically. The first mouse model, C57BL/6 was used to represent the properties of Th1. The second mouse model, BALB/c, was used to demonstrate the properties of Th2. The third mouse model was in the same family as the C57BL/6 mouse model but had its IL-6 gene knocked out. The goal of profiling each mouse model's differential gene expression through RNA sequencing was to highlight possible influences of the Th1/Th2 balance and provide further insight into the role of IL-6 in overall basal immunity. The analysis involved processing the unique reads of the data to construct a DESeq dataset using DESeq2. After filtering out low read counts and normalizing the data, three comparison objects were created using DESeq2's results function which were then processed to determine lists of differentially expressed genes. After performing a GO Term Enrichment Analysis to determine the roles of each differentially expressed gene, conclusions about the three mouse models were made. There were many findings for each mouse model. The C57BL/6 mouse model showed Th1 production to be involved with the expression of genes that are involved with the regulation of leukocyte migration, regulation of neutrophil chemotaxis, and the regulation of T Cell Migration. The IL-6 KO mouse model showed that when the IL-6 gene was knocked out there was a higher expression of genes involved with the epidermis and the cell cycle. It also showed that when knocked out, muscle-related genes are expressed more as well with transport-related genes and metabolism related genes. The BALB/c mouse model showed Th2 production to be involved with the expression of genes involved with Antigen processing, T cell regulation, apoptosis regulation, and peptide crosslinking.

2 Introduction

2.1 Biological Context

2.1.1 Th1/Th2

The study revolves around the immune system. The immune system is defined to be the defense system that organisms possess to perform multifarious biological processes meant to protect the organism from illness. The immune system includes three lines of defense. The first line of defense's role is to protect the organism from foreign invaders through skin and mucous membranes. The second line of defense protects the organism by using phagocytic leukocytes, antimicrobial proteins, and inflammatory responses. The third and final line of defense protects the organism by using lymphocytes, antibodies, and memory cells. The first and second lines of defense are known as innate immunity and the third line of defense is known as adaptive immunity.

Focusing more on adaptive immunity, it is composed of vastly specialized and systemic cells and processes whose role is to remove pathogens and hinder their development. The adaptive immune system is mainly composed of B cells and T cells.

T Cells play a significant role in adaptive immunity in that they are responsible in orchestrating immune responses through the production of cytokines. Cytokines are proteins that are secreted in order to compel cells to act in a particular way. T cells can be further categorized into cytotoxic and helper T-cells. The difference between them is that cytotoxic T-cells can actively kill infected cells (such as when a tumor is present) and helper T cells just "help" other cells do that for them (through cytokines).

Determined by the cytokines that can be secreted and the immunological responses that they induce, helper T-cells can be further classified into Th1 and Th2. One of the primary differences between

them is that Th1 initiates immunological responses on intracellular parasite such as bacteria and viri. Th2, on the other hand, initiates immunological responses on extracellular parasites that include helminths and extracellular parasites. Another difference between them is that Th1 help macrophages conduct inflammatory responses while Th2 aid B-Cells to create immunoglobins. It should also be noted that each type of helper cell produce cytokines that encourage their own development and hinder the other – suggesting that they have diametrically opposing roles.

2.1.2 IL – 6

IL – 6 (Interleukin 6) is an interesting type of interleukin that possesses the unique ability to act as a pro-inflammatory cytokine as well as an anti-inflammatory myokine. IL – 6 is known to act as a mediator in fevers as well as in acute phase responses. When specific microbial molecules are present macrophages secrete the IL-6 cytokine as an immunological response. In the cases of prolonged muscle use however, IL – 6 is produced to decrease inflammation levels, increase beta oxidation processes, and increase atrophy.

The relationship between IL – 6 and muscular atrophy is also interesting. Whether IL – 6 was knocked out or had high expression levels, muscular atrophy was still common. It is usually expected that the effects of when a gene is knocked out versus when its expressed more would be contradictory. But that is not the case here.

2.2 Experimental Design

The goal of the study was to highlight possible influences of the Th1/Th2 balance and to provide further insight into the role of IL-6 in overall basal immunity. To do this, three mouse models were chosen for experimentation. The three mouse models chosen were C57BL/6, a mouse also in the C57 family but with the IL-6 gene knocked out, and BALB/c. The C57BL/6 model was chosen because it naturally produces more Th1 and was used as the control. The IL-6 KO mouse model was chosen to be compared with the other mice to highlight the different roles of IL-6. And the BALB/c mouse model was chosen because it naturally leans towards Th2 dominated immunological responses.

With the results from sequencing three mouse models with an Illumina MiSeq, different comparisons can be made during the bioinformatic analysis to learn more about the Th1/Th2 balance and IL-6 functionality. There were three comparisons performed in the analysis. The first comparison was between the C57BL/6 and IL-6 KO mouse models to help determine the different roles of IL-6. The second comparison was done to compare the C57BL-6 and BALB/c mouse models to determine differences in the functions of Th1 and Th2. And the third comparison compares the C57BL/6 model with the BALB/c model to also determine the different roles of IL-6. The third comparison can help confirm the results of the first comparison.

The analysis involves processing the unique reads of the data to construct a DESeq dataset using DESeq2. Then, low read counts were filtered out and the data was normalized by estimating size factors, estimating dispersions, and conducting a Wald test of differential gene expression. Once this was done, three comparison objects were created using DESeq2's results function. These comparison objects were then processed by first shrinking the log fold change since they are usually over estimated. Once this was done, lists of differentially expressed genes were determined by taking genes with an adjusted p-value less than 0.05 and a log change value less than -1 and greater than 1. The results were then clustered to identify substructures in the data and identify groups of genes that behave similarly. After performing a GO Term Enrichment Analysis to determine the roles of each differentially expressed gene in each cluster, conclusions about the three mouse models were made.

3 Methods

After importing all of the unique reads for each of the three replicates in each mouse model, a DESeq Dataset was constructed. The gene IDs were in Ensemble annotation and the DESeq Dataset had almost 49585 genes in it. After filtering out genes with counts lower than 10, 31526 genes were filtered out (roughly 66%). The DESeq wrapper function was then used to estimate the size factors, dispersions, and conduct a Wald gene test of differential gene

expression using the normalized count data. After plotting the data it was determined that something needed to fixed.

Because of how the dispersion plot looked, the samples were clustered hierarchically and a principal components analysis (PCA) was done to provide indication of the distances between sample gene expression profiles and to detect possible batch effects. Because the developers of DESeq2 recommend that before clustering of samples a regularized log or variance stabilizing transformation should be performed on the normalized counts, this was done before clustering and PCA. Once the samples were clustered and PCA was performed, it was determined that one of the samples was off – possibly due to a batch effect. The data from the sample that was problematic was also different in that there were duplicates of each gene.

Once the dispersion plot looked normal, DESeq2's results function was used for each comparison to obtain the information necessary to determine which genes are differentially expressed. Because DESeq2 tends to overestimate log fold change values, they were shrunk for each of the three comparison objects.

To determine which genes were differentially expressed, genes with an adjusted p-value less than 0.01 and with a log fold change value less than -1 and greater than 1 were kept. The rest were removed. With this criteria, the C57BL/6 vs IL – 6 KO comparison object had around 2771 differentially expressed genes. The C57BL/6 vs BALB/c comparison object had around 488 differentially expressed genes. And the IL – 6 KO vs BALB/c comparison object had around 3125 differentially expressed genes.

Once the differentially expressed genes were determined, clustering was performed to identify substructures in the data and identify groups of genes that behave similarly. After normalized values for each gene was retrieved, a Pearson correlation was conducted to create distance objects need to cluster. Once clustered, each cluster was split into two groups. More groups were tried, but the results from them was either redundant, showered a difference in gene expression with in mouse model samples, or filled with outliers. For this reason, 2 groups were chosen.

A GO term enrichment analysis was performed for each group in each cluster. When a list of genes are obtained that are up-regulated under specific conditions, an enrichment analysis has the ability to locate which GO terms are over-represented or under-represented utilizing the annotations for that gene list. A hypergeometric test was the statistical analysis behind deciding which genes were overrepresented. The GO term enrichment analysis was able to determine the functions of the most expressed genes in each group in each cluster. With this information, along with the heat maps for each group, conclusions can be made on each mouse model in the study.

4 Results

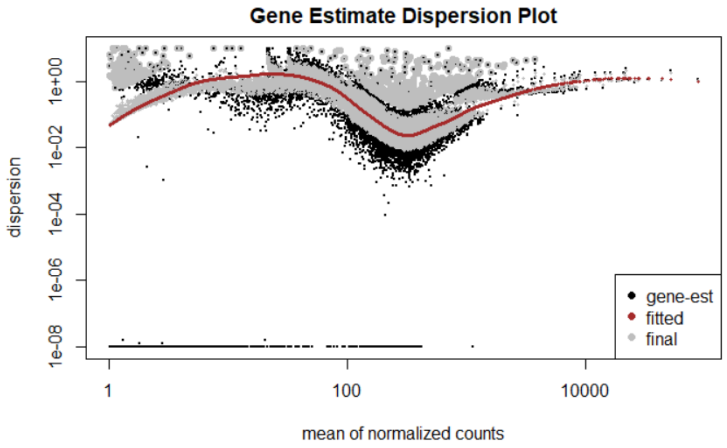

Figure 1: Original Gene Estimate Dispersion Plot – This figure depicts the gene estimate dispersion plot for the original DESeq dataset constructed from the raw unique counts provided by the study of all of the samples.

This dispersion plot is not ideal. A dispersion plot allows one to visualize how much the variance deviates from the mean. The curve enables the researcher to more accurately identify differentially expressed genes when the sample sizes are small. Ideally, the data should scatter around the curve with the dispersion decreasing as the mean expression levels rise. Because this was not the case, it was suspected that one of the samples could be causing this.

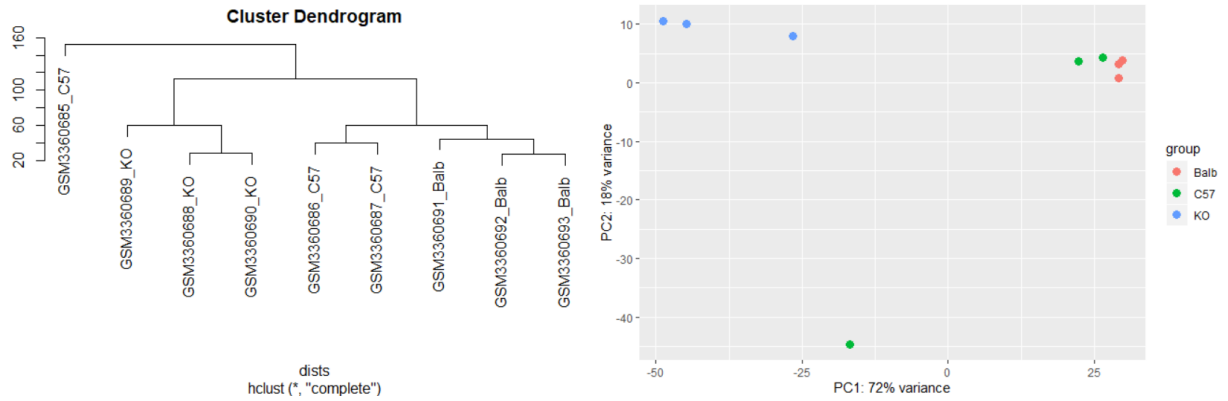

Figure 2: Hierarchical Sample Clustering [left] and PCA [right] – This figure represents two different methods to determine sample similarities. The left depicts hierarchal clustering of the samples and the right depicts principal component analysis.

Because of the dubious dispersion plot, Hierarchical Sample Clustering and PCA were done to determining the distances between sample gene expression profiles and look for indications of any batch effects. Both of the graphs indicate that one of the control samples, specifically GSM3360685_C57 is unlike the rest of the C5BL/6 samples. Interestingly enough, this sample had duplicate gene ids in its dataset. A reason for this sample's discrepancy could be because of a batch effect (such as if they were run on the same lane) or because the file got flawed before uploaded to NCBI.

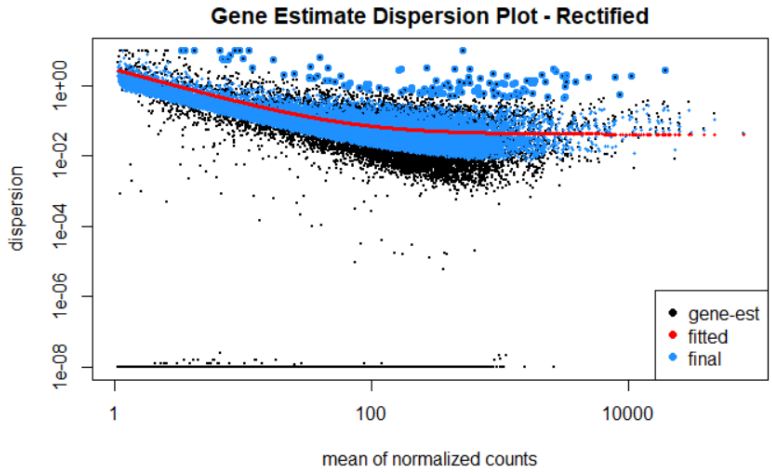

Figure 3: Rectified Gene Estimate Dispersion Plot – This figure depicts the gene estimate dispersion plot for the original DESeq dataset constructed from the raw unique counts provided by the study of all but one of the samples. The GSM3360685_C57 sample was excluded due to its dissimilarity with the other two replicates of the control mouse model.

Because the data scatters around the curve with the dispersion decreasing as the mean expression levels rise, the dispersion plot is as expected.

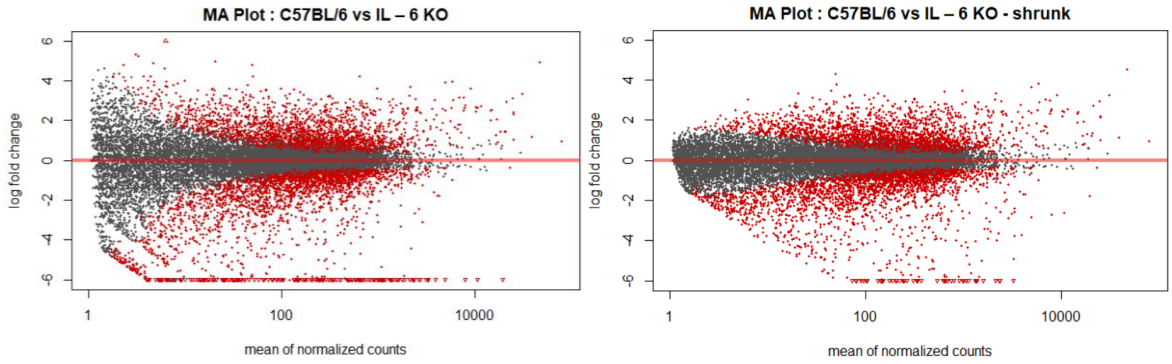

Figure 4: MA Plot of the C57BL/6 vs IL – 6 KO Comparison – This figure depicts the MA plot of the log fold change of the C57BL/6 and IL -6 KO comparison result object. The normal MA plot is shown on the left and the shrunken MA plot is on the right.

MA plots were constructed for each comparison to visualize the difference in log fold changes across the entire dataset. Because DESEq2 log2 fold-change estimates are frequently over-estimated particularly for low expression genes, to obtain better estimates of the log2 fold-change, DESeq2 recommends "shrinking" the raw estimates in the output table above with the lfcShrink() function to extract results from a DESeqDataSet object with shrunken log2 fold-change estimates.

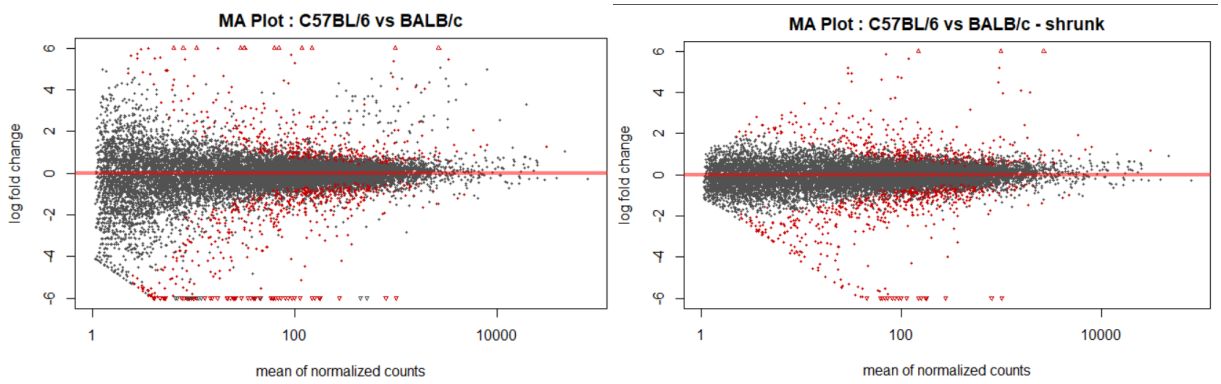

Figure 5: MA Plot of the C57BL/6 vs BALB/c Comparison – This figure depicts the MA plot of the log fold change of the C57BL/6 and BALB/c comparison result object. The normal MA plot is shown on the left and the shrunken MA plot is on the right.

Because this comparison has less differentially expressed genes than the other two comparisons, it can be suggested that there is a larger difference in the number of genes expressed when the IL-6 gene is knocked out as opposed to if Th1 or Th2 is expressed more.

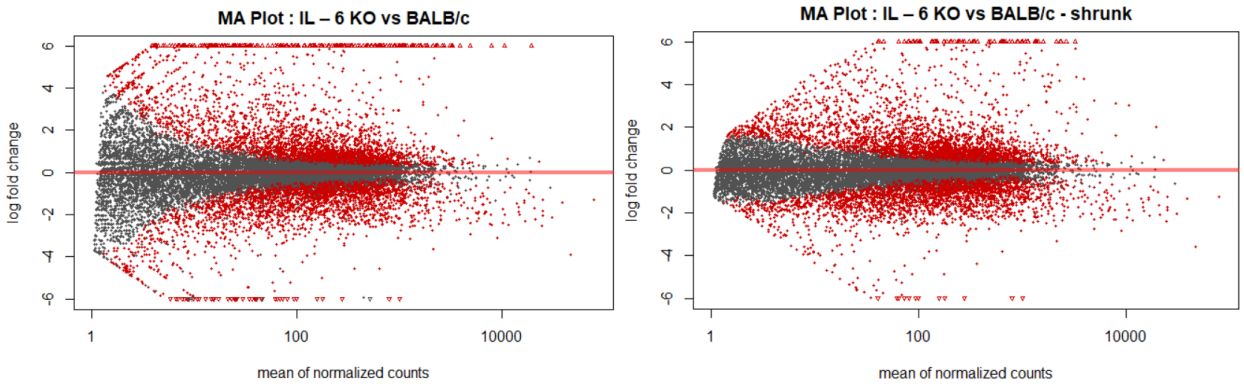

Figure 6: MA Plot of the IL – 6 KO vs BALB/c Comparison – This figure depicts the MA plot of the log fold change of the IL – 6 KO and BALB/c comparison result object. The normal MA plot is shown on the left and the shrunken MA plot is on the right.

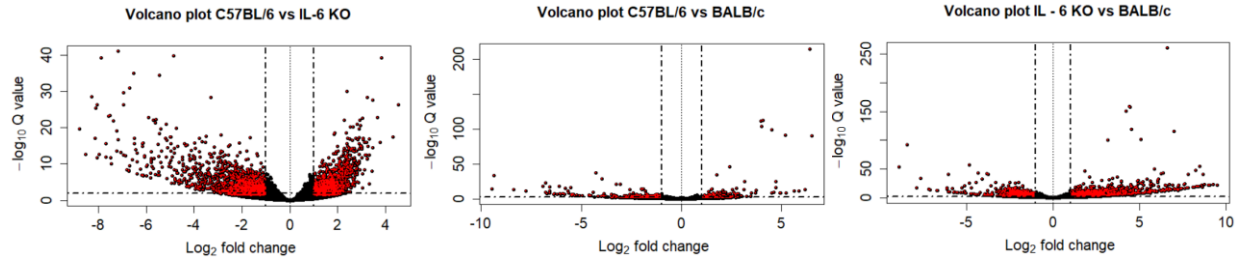

Figure 7: Volcano Plots for the Three Comparison Objects – This figure depicts the volcano plots to visualize the differentially expressed genes in each comparison. Differentially Expressed genes were determined by filtering out genes with an adjusted p-value greater than 0.05 and with a log fold change value between 1 and -1.

What can be seen in Figure 7 is that there is a large number of differentially expressed genes in each comparison. Although it may seem that the volcano plots of the C57BL/6 vs BAL/c comparison and the IL-6 KO vs BALB/c comparison are shorter than the C57BL/6 vs IL-6 KO comparison, this is not necessarily the case since they all have different y-axis scales. All in all, each of the comparisons possess a large number of statistically significant differentially expressed genes. The tow comparisons that involve the IL-6 gene have more differentially expressed genes than the other. The C57BL/6 vs IL – 6 KO comparison object had around 2771 differentially expressed genes. The C57BL/6 vs BALB/c comparison object had around 488 differentially expressed genes. And the IL – 6 KO vs BALB/c comparison object had around 3125 differentially expressed genes.

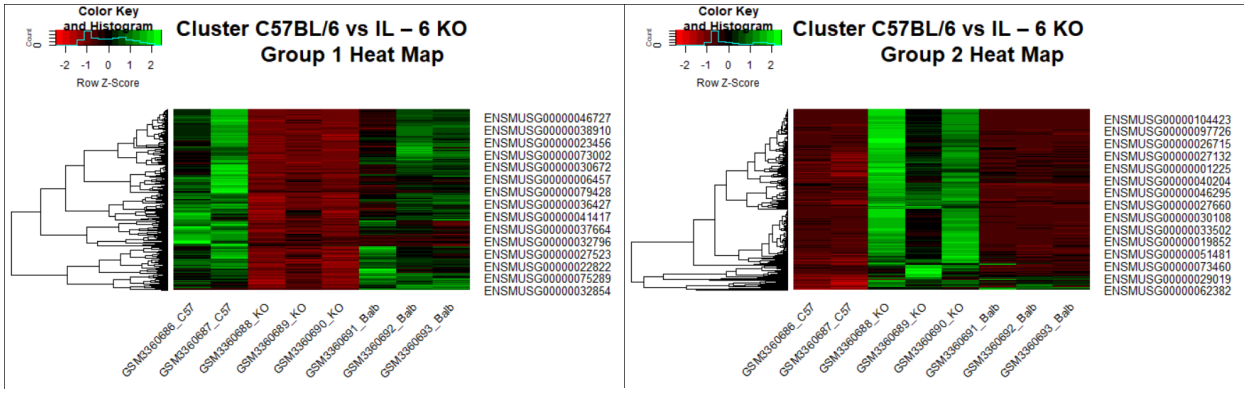

Figure 8: C57BL/6 vs IL-6 KO Cluster Heatmaps – After clustering the data (Pearson correlation used as distance objects) to determine substructures and identify groups of genes that behave similarly, the cluster was split into two groups. A heat map was constructed to visualize the gene expression data. The color and intensity of the boxes represent the changes in gene expression.

C57BL/6 vs IL – 6 KO Group 1	C57BL/6 vs IL – 6 KO Group 2
[1] "muscle cell development"	*[1] "epidermis development"*
[2] "muscle tissue development"	*[2] "nuclear division"*
[3] "striated muscle cell differentiation"	*[3] "cell division"*
[4] "sarcomere organization"	*[4] "keratinization"*
[5] "muscle organ development"	*[5] "hair cycle"*
[6] "heart process"	*[6] "peptide cross-linking"*
[7] "ion transport"	*[7] "positive regulation of cell cycle process"*
[8] "muscle contraction"	*[8] "epithelial cell differentiation"*
[9] "blood circulation"	*[9] "molting cycle process"*
[10] "regulation of muscle system process"	*[10] "nuclear chromosome segregation"*
[11] "cardiocyte differentiation"	*[11] "keratinocyte differentiation"*
[12] "muscle structure development"	*[12] "mitotic spindle organization"*
[13] "system process"	*[13] "regulation of cell cycle"*
[14] "cardiac muscle contraction"	*[14] "mitotic sister chromatid segregation"*
[15] "transmembrane transport"	*[15] "cell cycle phase transition"*
[16] "divalent inorganic cation transport"	*[16] "regulation of attachment of spindle microtubules to kinetochore"*
[17] "inorganic ion transmembrane transport"	*[17] "osteoblast differentiation"*
[18] "positive regulation of ion transmembrane transport"	*[18] "regulation of mitotic nuclear division"*
[19] "regulation of metal ion transport"	*[19] "regulation of melanin biosynthetic process"*
[20] "regulation of cation transmembrane transport"	*[20] "phenol-containing compound metabolic process"*

Table 1: GO Terms for Genes for Each Group of the C57BL/6 vs IL-6 KO Cluster – This table shows the Go terms of each group in the cluster that are expressed or not expressed in the corresponding heat map.

Based on the Heatmap and the Go term Enrichment analysis, it was found that when the control mouse model was compared with the IL-6 KO mouse model, genes involving muscle development, growth and differentiation were not expressed in the KO model as well as genes involving the transport of ions and metals. Genes involved with skin development and cell cycle expression were expressed in the KO variant however and not in the other two mouse models.

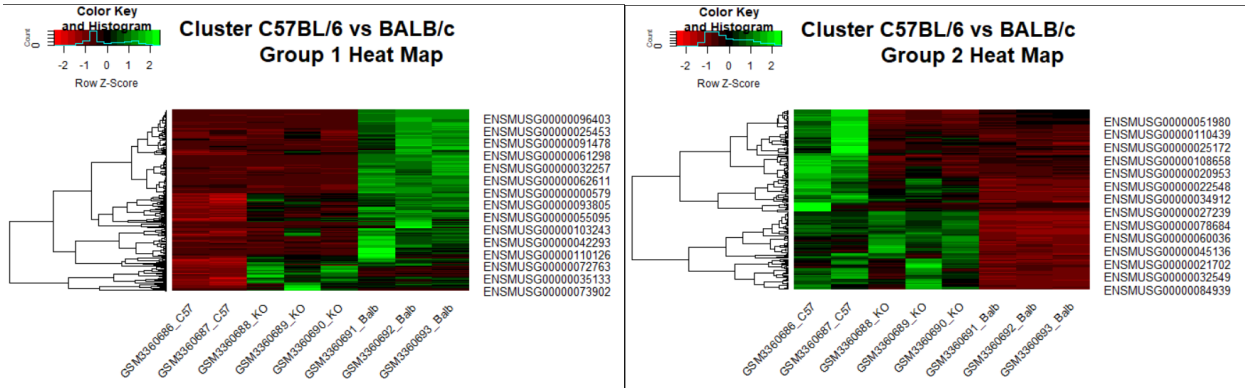

Figure 9: C57BL/6 vs BALB/c Cluster Heatmaps – After clustering the data (Pearson correlation used as distance objects) to determine substructures and identify groups of genes that behave similarly, the cluster was split into two groups. A heat map was constructed to visualize the gene expression data. The color and intensity of the boxes represent the changes in gene expression.

C57BL/6 vs BALB/c Group 1	C57BL/6 vs BALB/c Group 2
[1] "antigen processing and presentation of endogenous peptide antigen via MHC class I via ER pathway, TAP-independent"	[1] "positive regulation of leukocyte migration"
[2] "antigen processing and presentation of endogenous peptide antigen via MHC class Ib"	[2] "cytoplasmic translation"
[3] "antigen processing and presentation via MHC class Ib"	[3] "regulation of neutrophil chemotaxis"
[4] "antigen processing and presentation of endogenous peptide antigen via MHC class I"	[4] "positive regulation of chemotaxis"
[5] "antigen processing and presentation of endogenous antigen"	[5] "positive regulation of fast-twitch skeletal muscle fiber contraction"
[6] "positive regulation of T cell mediated cytotoxicity"	[6] "negative regulation of supramolecular fiber organization"
[7] "peptide cross-linking"	[7] "regulation of leukocyte chemotaxis"
[8] "antigen processing and presentation of peptide antigen"	[8] "myofibril assembly"
[9] "response to nematode"	[9] "negative regulation of protein depolymerization"
[10] "keratinocyte differentiation"	[10] "regulation of twitch skeletal muscle contraction"
[11] "regulation of T cell mediated immunity"	[11] "fat-soluble vitamin catabolic process"
[12] "positive regulation of cell killing"	[12] "muscle system process"
[13] "positive regulation of hydrogen peroxide biosynthetic process"	[13] "regulation of T cell migration"
	[14] "response to muscle stretch"
	[15] "response to chemical"
	[16] "regulation of actin filament-based process"
	[17] "lymphocyte chemotaxis"

Table 2: GO Terms for Genes for Each Group of the C57BL/6 vs BALB/c Cluster – This table shows the Go terms of each group in the cluster that are expressed or not expressed in the corresponding heat map.

When the Th1 expressing control was compared with the Th2 expressing mouse model, the genes involving antigen processing, T Cell Regulation, Apoptosis Regulation and peptide cross-linking were expressed in the BALB/c variant. Genes involving leukocyte migration, neutrophil chemotaxis regulation, and T cell migration were

expressed in the C57BL/6 variant. Genes expressing the regulation of myofibril assembly and actin filament regulation was also expressed in both the control and KO variants – suggesting that these genes may have been expressed because the IL-6 KO and C57BL/6 mouse models were related as opposed to it being because of the difference in immune responses relative to the BALB/c model.

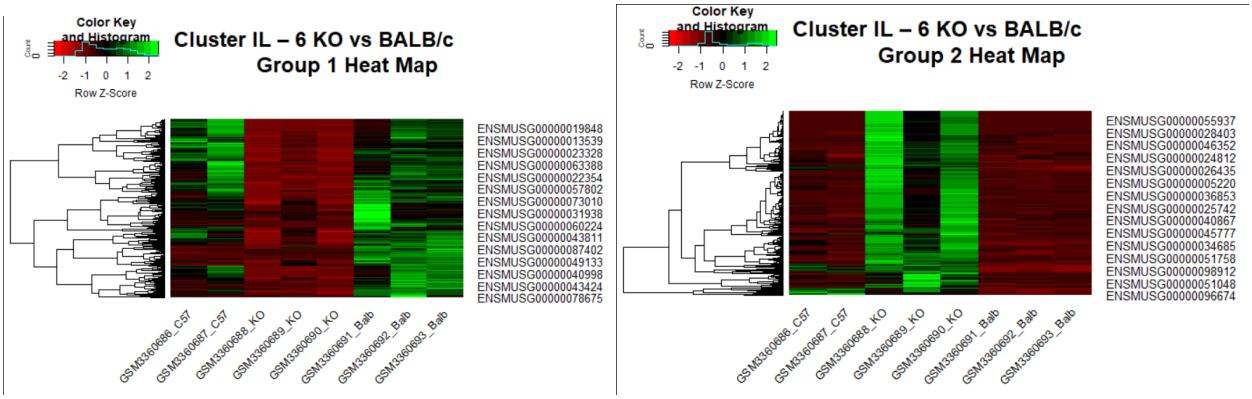

Figure 10: IL-6 KO vs BALB/c Cluster Heatmaps – After clustering the data (Pearson correlation used as distance objects) to determine substructures and identify groups of genes that behave similarly, the cluster was split into two groups. A heat map was constructed to visualize the gene expression data. The color and intensity of the boxes represent the changes in gene expression.

IL – 6 KO vs BALB/c Group 1	IL – 6 KO vs BALB/c Group 2
[1] "muscle cell development"	[1] "hair cycle"
[2] "ion transport"	[2] "molting cycle process"
[3] "muscle tissue development"	[3] "cell division"
[4] "striated muscle cell differentiation"	[4] "mitotic sister chromatid segregation"
[5] "transmembrane transport"	[5] "hair follicle development"
[6] "ribonucleotide metabolic process"	[6] "keratinization"
[7] "purine nucleotide metabolic process"	[7] "morphogenesis of an epithelium"
[8] "carboxylic acid metabolic process"	[8] "epithelium development"
[9] "regulation of muscle system process"	[9] "mitotic cell cycle"
[10] "muscle organ development"	[10] "regulation of epidermis development"
[11] "antigen processing and presentation of endogenous peptide antigen via MHC class I via ER pathway, TAP-independent"	[11] "nuclear division"
[12] "cellular respiration"	[12] "multicellular organism development"
[13] "antigen processing and presentation of endogenous peptide antigen via MHC class Ib"	[13] "skin development"
[14] "pyruvate metabolic process"	[14] "epidermis development"
[15] "heart process"	[15] "nuclear chromosome segregation"
[16] "sarcomere organization"	[16] "embryonic organ development"
[17] "purine nucleoside monophosphate metabolic process"	[17] "regulation of attachment of spindle microtubules to kinetochore"
[18] "energy derivation by oxidation of organic compounds"	[18] "positive regulation of cell cycle process"
[19] "oxidoreduction coenzyme metabolic process"	[19] "morphogenesis of a branching structure"
[20] "ATP metabolic process"	[20] "developmental process"

Table 3: GO Terms for Genes for Each Group of the IL – 6 KO vs BALB/c Cluster – This table shows the Go terms of each group in the cluster that are expressed or not expressed in the corresponding heat map.

When the IL-6 KO model was compared to the Th2 producing BALB/c model, genes involving muscle development/differentiation and transport involving ion/transmembrane were not expressed in the IL-6 KO model. Antigen Processing and Metabolism was also present in the BALB model and not the IL-6 KO model. Also like the first comparison, genes that involved the epidermis and cell cycle were expressed in the KO variant and not the other two models.

5 Discussion (1 page – what you learned, what was your analysis, what questions were answered, limitations faced and how they were tackled, etc.)

The C57BL/6 mouse model showed Th1 production to be involved with the expression of genes that are involved with the regulation of leukocyte migration, regulation of neutrophil chemotaxis, and the regulation of T Cell Migration.

It was known that helper T cells are responsible for producing cytokines to tell other cells how to act in dealing with a foreign invader. Th1 generates immune responses against intracellular parasites like bacterial and villi. They also help macrophages produce inflammatory responses. This is shown in the results of the GO term analysis for the C57BL/6 mouse model. Because the C57BL/6 mouse model was associated with the regulation of leukocyte migration, regulation of neutrophil chemotaxis, and the regulation of T Cell Migration, it can be seen that there is further support that Th1 helps macrophages produce immune responses.

The IL-6 KO mouse model showed that when the IL-6 gene was knocked out there was a higher expression of genes involved with the epidermis and the cell cycle. It also showed that when knocked out, muscle-related genes are expressed more as well with transport-related genes and metabolism related genes.

This goes in hand with what was expected. It was known both the absence and over production of IL 6 can lead to muscular atrophy. It was seen that when IL-6 was knocked out, that the other replicates of the other two mouse models expressed genes related to muscle development. Since they weren't expressed in the IL-6 KO mouse model, it provides further evidence that IL-6 needs to be present to prevent muscular atrophy.

Although not predominantly known for it, IL-6 seems to have a role in transport and metabolism. There are numerous studies that exist that also substantiate this claim. IL-6 was found to stimulate epithelial sodium channels in mouse cortical collecting duct cells as well as modulate colonic transepithelial ion transport in the stress-sensitive Wistar Kyoto rat. Other studies show IL-6 to directly increase glucose metabolism in human resting skeletal muscle. There are an extensive level of branches in the IL-6 signaling pathway that are associated with hypertrophic muscle growth and myogenesis through the regulation of the proliferation capacity of the muscle stem cells. The signal pathway is also associated with energy metabolism which also has to do with muscle contraction capacity.

The BALB/c mouse model showed Th2 production to be involved with the expression of genes involved with Antigen processing, T cell regulation, apoptosis regulation, and peptide crosslinking. Th2 was known to generate immune responses against helminths and extracellular parasites. Th2 was also known to help B-Cells produce immunoglobins. The RNA seq analysis was also able to show other roles of Th2 which include apoptosis regulation and peptide crosslinking.

The study was able to confirm what was already suspected about the three proteins in question while also determining other roles of the proteins. Also important, the study was able to record the specific genes that were differentially expressed for the three immunologically different mouse models.

The only limitation was the small number of replicates – especially since one had to be removed. If this were to be repeated, more replicates should be used. Despite the low number of replicates, differentially expressed genes were still identified which could be because how well they fit on the PCA plot [Figure 2]

6 References

https://diabetes.diabetesjournals.org/content/56/6/1630

https://hbctraining.github.io/DGE_workshop/lessons/04_DGE_DESeq2_analysis.html

https://www.frontiersin.org/articles/10.3389/fmicb.2018.00553/full

https://www.frontiersin.org/articles/10.3389/fphar.2012.00190/full

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE119202

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1782809/

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2928617/

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4163639/

https://www.ncbi.nlm.nih.gov/pubmed/9577091

https://www.the-scientist.com/technology-profile/distinguishing-th1-and-th2-cells-54709

7 Appendix (your code – knit to html and attach with paper)