**Contractile genes enrichment in differential gene expression analysis between heart and lymph node samples**

Rebecca Fuchs Final Report - May 5

**Table of Contents**

Abstract Page 1

Introduction Page 1

Methods Page 2

Results Page 3

Discussion Page 12

References Page 13

Supplementary Data Page 14

**Abstract**

Differential gene expression analysis allows researchers to explore cell biological processes. By comparing cell gene expression using statistics, the researcher is able to find out what is different between tissues. Learning how cells differ can help us learn about each sample. In this research paper I will compare gene expression from heart and lymph node tissue. We hoped to learn what pathways are enriched in these cells. We expected to find contractile related pathways to be significantly differentially expressed.

I used globus genomics to check the quality from the samples, map them, and get gene expression counts. Then I used R to run the differential gene expression and make tables to visualize the results. Finally I used DAVID and enrichr to run pathway enrichment on the most differentially expressed genes. The results found cytoskeletal related pathways were expressed in heart tissue, like muscles and micro fibril pathways. There were also a few pathways relating to the immune system were over expressed, like B cell signaling pathway and immunodeficiency.

**Introduction**

In this report I explore the differential gene expression from tissues in the lymph node and heart. The heart contains cardiomyocytes and cardiac pacemaker cells (wiki:cardiac\_muscles). These cells specialize in contracting; they require high amounts of energy and usage of cytoskeleton. The heart is a crucial to the human body, pumping blood containing molecules and oxygen to the rest of the cells.

There are multiple lymph nodes throughout the body; they are involved in function the immune system. They contain B and T cells, to combat infection and cancer (wiki:lymph\_node). Immune cells develop and are activated in the lymph nodes; these cells are essential to human survival by combating infections, parasites, and cancer. An NCBI article explains the lymph nodes relevance in this quote “T-cells move continuously between lymph nodes and the blood, testing AOCs for signs of infection”. (Nicholson)

Differential gene expression compares gene expression from one group of samples to another. After normalizing gene expression, a statistical test is used to find significantly expressed genes. These genes are ranked based on P value, the value describes how likely the gene is to not be significantly differentially expressed. There are thousands of tests run, so a multiple testing correction must be implemented. This corrects for the increased likelihood that there will be a significant gene found when many tests are run. The list of significant genes will be input into gene enrichment tools to learn what pathways have increased expression.

I will examine the differential gene expression between these two organs. I hypothesize there will be a higher expression of cytoskeletal proteins and genes in the heart because the cells contract while the lymph nodes don’t contract.

**Methods**

There are three human samples from the heart, and three from the lymph nodes. I used Globus Genomics for quality control, trimming, mapping, and acquiring gene counts for each gene in each sample. R studio ran the differential expression analysis, using packages edgeR, statmod, org.Mm.eg.db and org.Hs.eg.db.

Globus Genomics

First, FastQC was run on each fastq file to check the quality of each sample, parameters were set as default. Trim Galore was next run on each sample fastq file to remove adapters and any reads that are low quality. Next I used Tophat2 to map the rnaseq data to reference genome hg38. Finally, featureCounts was used to input all tophat2 files to create a table of counts for each gene, again genome hg38 was used as a reference genome.

Table 1.

*This table shows the tools and parameters used in the workflow of this gene expression analysis.*

|  |  |
| --- | --- |
| Globus Genomics Tools | Parameters |
| FastGC | Version FASTQC: 0.11.3  Standard output and error |
| TrimGalore | Library mate paired? Single  Default advanced settings, fastq files input |
| TopHat2 | Galaxy Tool Version: SAMTOOLS: 1.2; BOWTIE2: 2.1.0; TOPHAT2: 2.0.14  Library mate paired? Single |
| featureCounts | GFF/GTF Source- indexed\_filtered  Reference Gene set- GRCh38.77  GFF feature type filter- exon  Unstranded |

R studio (Important code from R is attached in the supplementary data)

I installed the necessary packages and read in the counts. Genome annotation was run with the package “org.Hs.eg.db” version 3.10, and “statmod” was used version 1.4.34. . I used the function DGEList to separate the genes and the counts. Next I used calcNormFactors to normalize the counts. I labeled the samples “H” for heart, and “L” for lymph nodes. Finally I was able to run edgeR (version 3.28.1), with the function exactTest, and look at the summary statistics to see how many genes were up regulated versus down. Finally I used a function named topTags to order the top p value genes into a file. We also plotted a MDS plot, a BCV plot, and a smear plot.

Enrichr

We used the online enrichment tool “Enrichr” to run gene enrichment on the differentially expressed genes. P values were all very small; I ran the enrichment tool on a gene list of 500 genes and 1000 genes. I looked at KEGG pathways, GO biological process, and GO cellular component.

**Results**

FastQC provided warnings for duplication and adapter content. The samples are mRNA so duplication warning shouldn’t be worried about because mrna can have high levels and low levels depending on its expression. There are a few quality warnings, we’ll use a trimming tool to remove bad quality reads. Below is a chart (table2) summarizing the quality of each sample.

Table2. Quality score for each sample

*This table shows per quality warning output from FastQC from each sample. FastQC gave warnings for all samples in seq duplication, overrepresented seq, and kmer content. These all have warnings because the data is mrna sequencing, and some mrna’s have more then others. We are measuring the amount of mrna in the samples.*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Check | Lymph 4a | Lymph 4b | Lymph 5a | Heart 5a | Heart 5b | Heart 6a |
| Basic stat | Ok | Ok | Ok | Ok | Ok | Ok |
| Per base seq quality | Ok | Ok | Ok | Ok | Ok | Ok |
| Per tile seq quality | Ok | Ok | Ok | ! | Ok | Ok |
| Per base seq content | X | X | X | X | X | X |
| Seq duplication | ! | ! | ! | X | X | X |
| Overrepresented seq | ! | X | X | X | ! | ! |
| Adapter content | X | X | Ok | Ok | ! | ! |
| Kmer concent | X | X | X | X | X | X |

After tophat2 was run on each sample, featureCounts brought all the samples gene counts together. Columns 2-7 represent each samples counts, column 1 is the gene name. Column 8 is the total. A snapshot of the first few rows of this table is shown below.

Table3. Feature counts

*This table is an example of the first few rows, showing the counts for each gene in each sample. The last column is the total.*

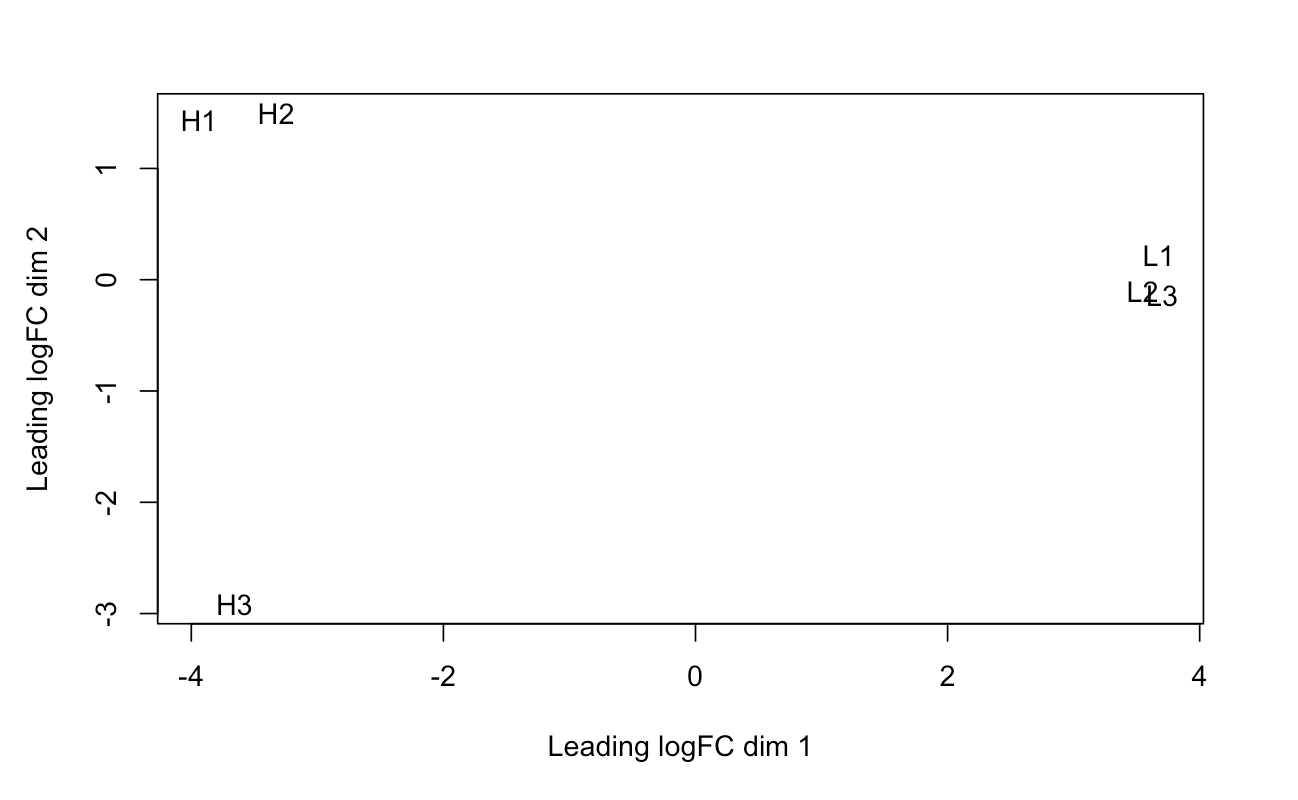
| **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| RP11-34P13.14 | 0 | 0 | 0 | 0 | 0 | 0 | 323 |
| RP11-34P13.9 | 0 | 0 | 0 | 0 | 0 | 0 | 457 |
| FO538757.3 | 4 | 0 | 0 | 2 | 4 | 1 | 718 |
| FO538757.2 | 14 | 17 | 20 | 36 | 32 | 35 | 1982 |
| AP006222.2 | 4 | 4 | 10 | 1 | 4 | 5 | 4039 |
| RP5-857K21.15 | 0 | 0 | 0 | 0 | 0 | 0 | 676 |
| RP4-669L17.2 | 0 | 0 | 0 | 0 | 0 | 0 | 1095 |

Plots from R studio

MDS plot (multidimensional scaling)

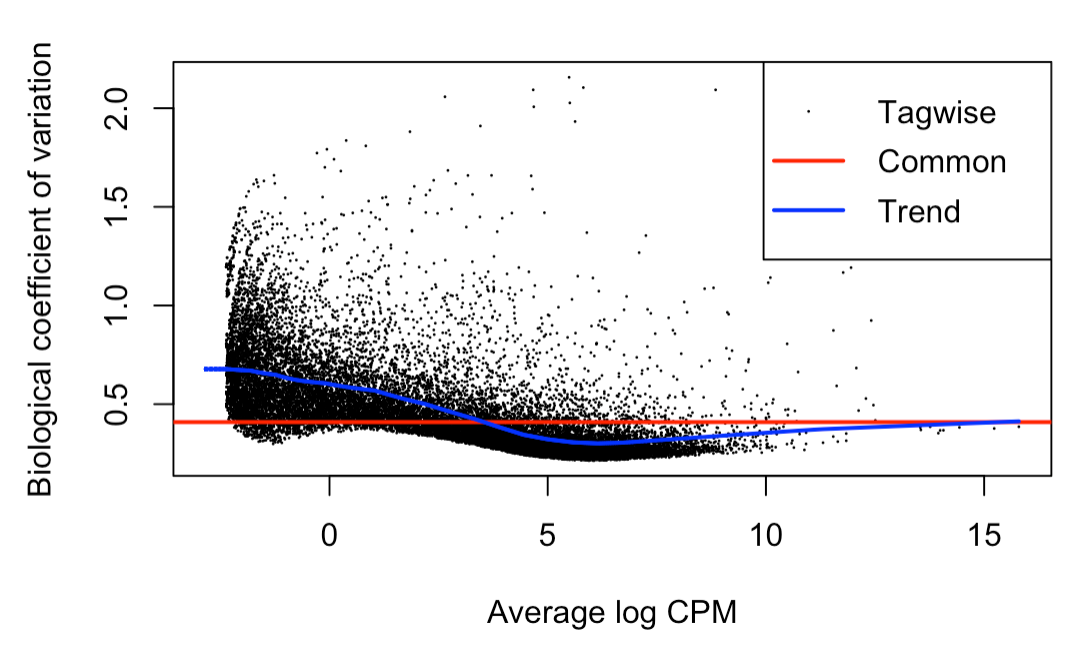
This plot below shows samples visualized by plotting them with using there most variable attributes. We can see samples H1-3 are on the same side of the plot, while samples L1-3 are clumped together. It is surprising to see H3 is not clumping right with H1 and H2. Perhaps heart sample 3 was from a different age then the other samples or perhaps the patient was sick. There is also a chance this is just due to random variation or error.

Figure1. MDS plot Heart & Lymph Node MDS

*This figure shows similar samples plotted, H are heart samples, L are lymph node samples. All samples plot close to each to the same tissue on the x axis. Heart samples do not plot close to the other samples on the y axis. This tells us the lymph node samples are all every similar but there is some variation among the heart samples, specifically samples 3, while samples 1 and 2 plot close together.*

BCV plot (Biological Coefficient of Variation)

The below BCV plot (figure2) measures variation. We can see there is an area of high variation but it drops to lower then 0.4. We know there are some variables genes but the trend lines drops below the red common threshold of 0.4, where the variation is the same as other human variation.

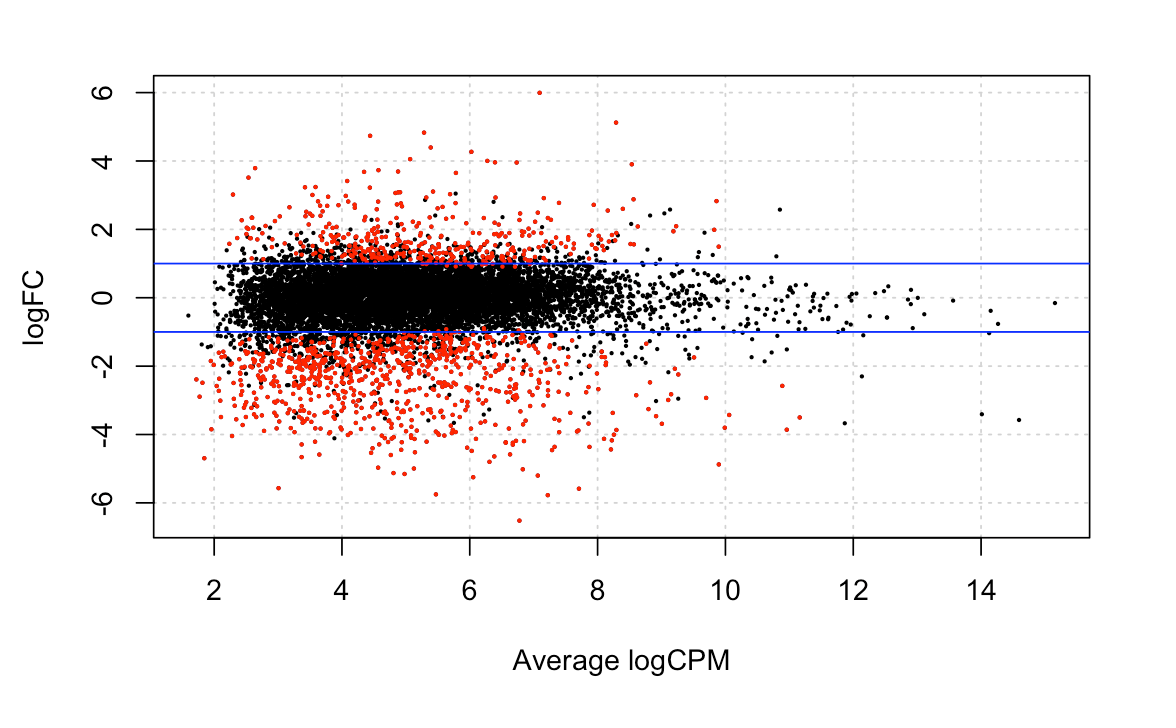
Figure 2. BCV plot BCV plot Heart & Lymph Nodes

*The BCV plot shows the general trend of the samples variation. The trend line shows it starting about normal human variation, but going lower with the average log CPM increasing.*

Figure 3. Smear plot

*In the smear plot we can see the genes in red that have significant log fold changes. This was made using r studio’s edgeR package, and significant genes found from the exact test are in red. The blue line marks 1 and -1.*

Smear Plot Heart & Lymph Nodes



The smear plot above (figure 3) gives a good visual representation of how many significantly differentially expressed genes there are out of the whole genome. You can see most genes fall between the blue lines but there are many outside marked in red.

Table 4. Results from edgeR

*There were 4412 up regulated genes and 4832 down regulated genes comparing Lymph node samples to heart samples.*

|  |
| --- |
| Results from edgeR L-H |
| Down Regulated 4832 |
| Not Significantly changed 24411 |
| Up Regulated 4412 |

Table5. Top P values genes (showing the top 20)

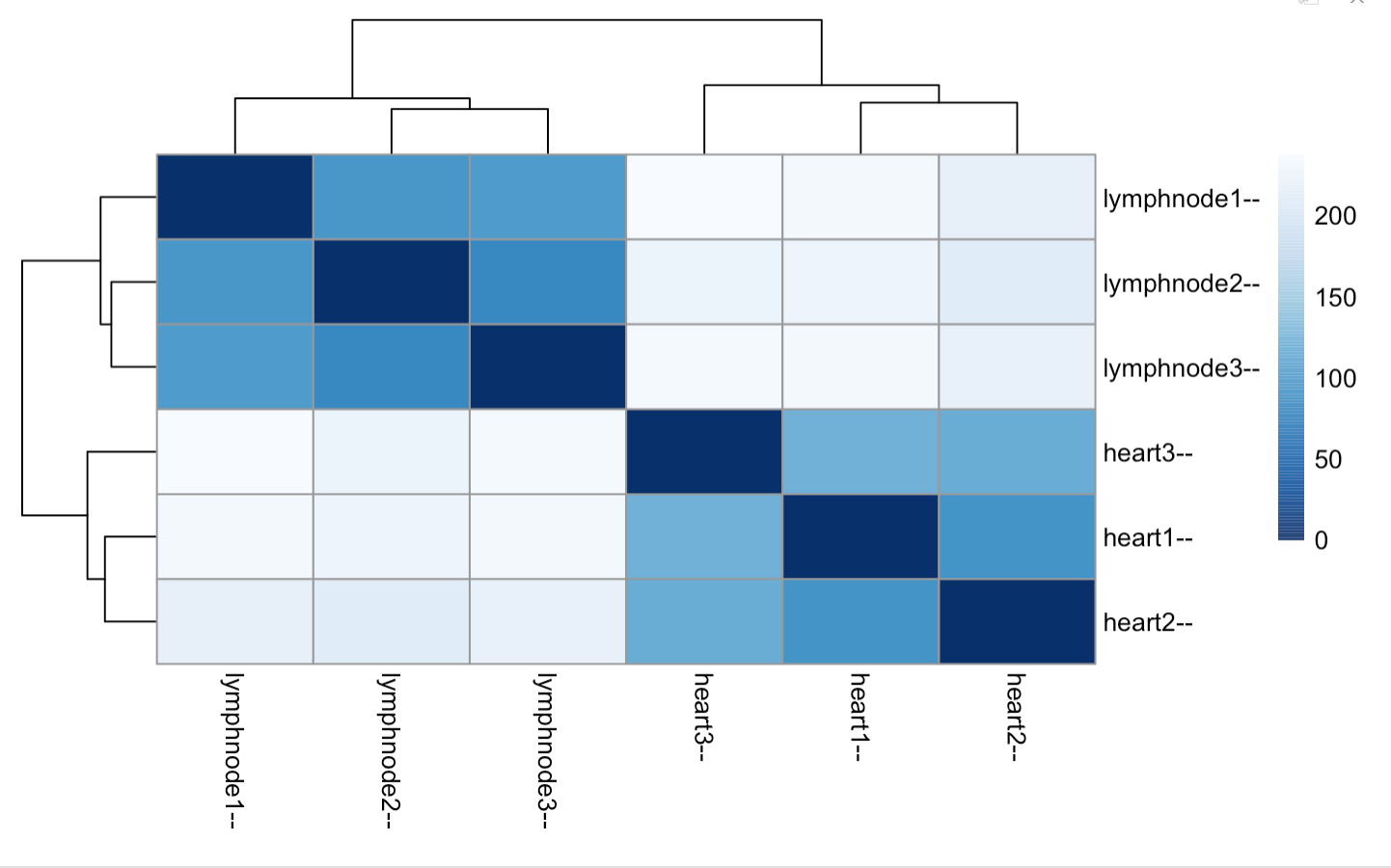
*This table is output from edgeR, it is a sorted list of the lowest P value genes. The logFC score in this top 25 table are all negative, but within the top 100 genes there are some positive logFC genes.*

Lowest P value Genes



The P values are extremely low and even after correction, FDR is still at 10^-91. This is highly significant and merits looking into these specific genes. Some of these genes in table 5 have their function defined later into this paper in table 6. There are many genes with these very low P values seen in table 4.

Figure 4. Heat mapHeat map of heart and Lymph node samples

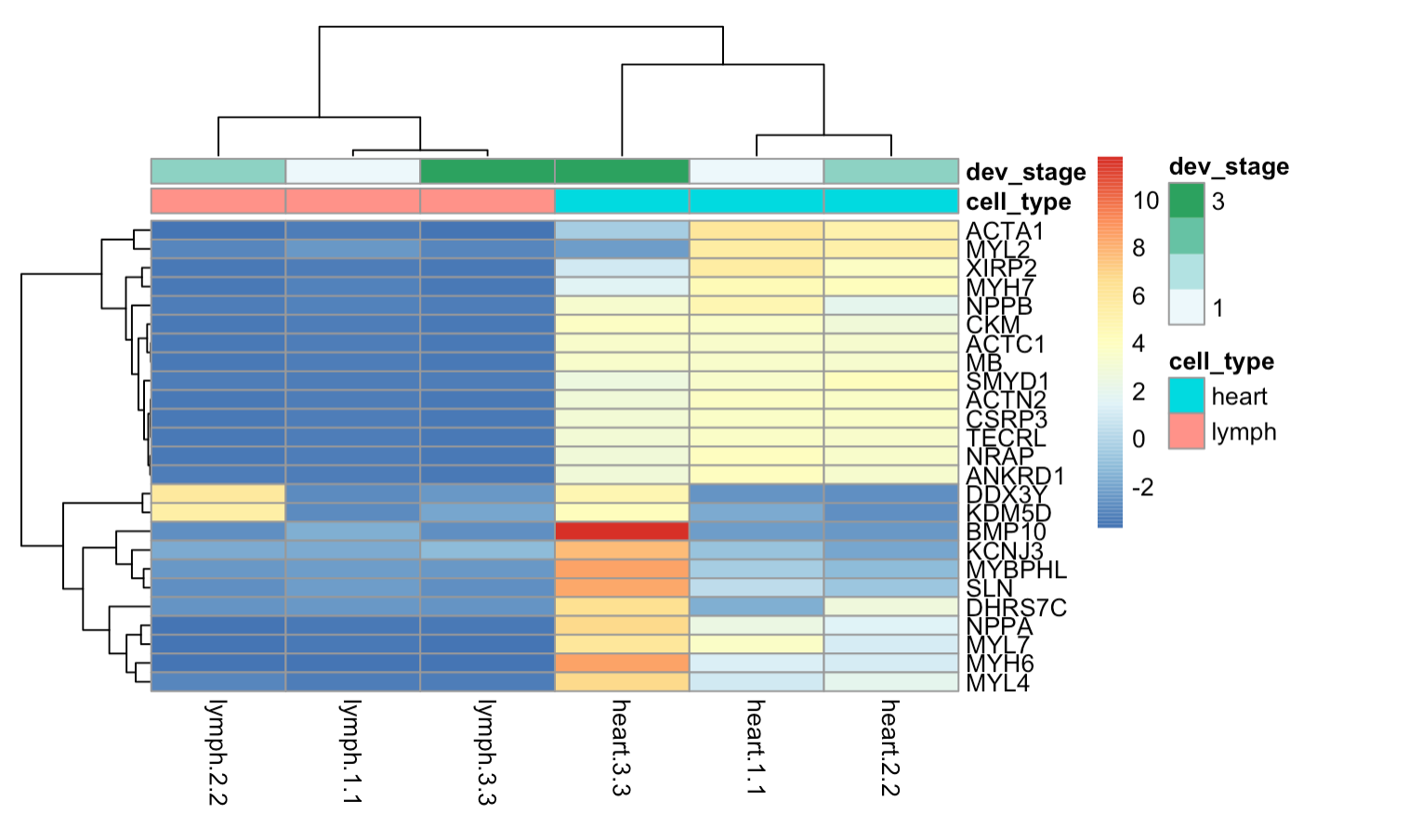
*This heat map was made with R studio. It compares gene expression between samples. We can see the dark blue is most similar, and light blue/white is most variability.*

We can see in Figure 4 the lymph node samples are most similar to each other. The heart samples are most similar to one another too. Lymph node samples and heart samples are not similar and don’t cluster together.

Figure5. Heat map

*This heat map shows the expression of the overall top 25 differentially expressed genes for each sample. Light blue means the gene is down regulated, and red means the gene is up regulated. The genes are on the y axis and the samples are on the bottom of the graph.*

Heat map of Samples and Genes



In figure 5, is a heatmap of top 25 genes in all 6 samples. There is very clear difference in clustering of genes from heart and lymphnodes samples. Heart sample 3.3 has unusually high levels in some genes compared to all other cells. Perhaps there is a difference in the health condition of the sample. The variation in heart samples can also be seen in the MDS plot. One gene that is bright red, showing its high expression in heart sample 3 is BMP10, full name Bone Morphogenic Protein 10. Otherwise heart samples 2.2 and 1.1 have very similar gene expression.

Table6. Gene Definitions

*Defined top genes, pink was related to cardiac related, blue to muscular relation.*

|  |  |  |
| --- | --- | --- |
| gene | defined | reference |
| TNNT2 | Among its related pathways are [Cardiac conduction](http://pathcards.genecards.org/card/cardiac_conduction" \o "See Cardiac conduction at Pathcards" \t "_blank) and [Dilated cardiomyopathy (DCM)](http://pathcards.genecards.org/card/dilated_cardiomyopathy_(dcm)" \o "See Dilated cardiomyopathy (DCM) at Pathcards" \t "_blank). Gene Ontology (GO) annotations related to this gene include actin binding and structural constituent of cytoskeleton. An important paralog of this gene is [TNNT1](https://www.genecards.org/cgi-bin/carddisp.pl?gene=TNNT1" \t "_blank). | <https://www.genecards.org/cgi-bin/carddisp.pl?gene=TNNT2#summaries> |
| HHATL | HHATL (Hedgehog Acyltransferase Like) is a Protein Coding gene. Diseases associated with HHATL include [Skin Squamous Cell Carcinoma](http://www.malacards.org/card/skin_squamous_cell_carcinoma" \o "See Skin Squamous Cell Carcinoma at Malacards" \t "_blank) and [Cardiomyopathy, Dilated, 1E](http://www.malacards.org/card/cardiomyopathy_dilated_1e" \o "See Cardiomyopathy, Dilated, 1E at Malacards" \t "_blank). An important paralog of this gene is [HHAT](https://www.genecards.org/cgi-bin/carddisp.pl?gene=HHAT" \t "_blank). | <https://www.genecards.org/cgi-bin/carddisp.pl?gene=HHATL&keywords=HHATL> |
| MYBPC3 | Among its related pathways are [Cardiac conduction](http://pathcards.genecards.org/card/cardiac_conduction" \o "See Cardiac conduction at Pathcards" \t "_blank) and [Dilated cardiomyopathy (DCM)](http://pathcards.genecards.org/card/dilated_cardiomyopathy_(dcm)" \o "See Dilated cardiomyopathy (DCM) at Pathcards" \t "_blank). Gene Ontology (GO) annotations related to this gene include identical protein binding and structural constituent of muscle. An important paralog of this gene is [MYBPC2](https://www.genecards.org/cgi-bin/carddisp.pl?gene=MYBPC2" \t "_blank). | <https://www.genecards.org/cgi-bin/carddisp.pl?gene=MYBPC3&keywords=MYBPC3> |
| MLIP | Required for precocious cardiac adaptation to stress through integrated regulation of the AKT/mTOR pathways and FOXO1. Regulates cardiac homeostasis and plays an important role in protection against cardiac hypertrophy. | genecards.org/cgi-bin/carddisp.pl?gene=MLIP&keywords=MLIP |
| SPHKAP | Diseases associated with SPHKAP include [Renal Cell Carcinoma, Papillary, 1](http://www.malacards.org/card/renal_cell_carcinoma_papillary_1" \o "See Renal Cell Carcinoma, Papillary, 1 at Malacards" \t "_blank). Gene Ontology (GO) annotations related to this gene include protein kinase A binding. An important paralog of this gene is [AKAP11](https://www.genecards.org/cgi-bin/carddisp.pl?gene=AKAP11" \t "_blank). | <https://www.genecards.org/cgi-bin/carddisp.pl?gene=SPHKAP&keywords=SPHKAP> |
| MYPN | Among its related pathways are [Focal Adhesion](http://pathcards.genecards.org/card/focal_adhesion" \o "See Focal Adhesion at Pathcards" \t "_blank). Gene Ontology (GO) annotations related to this gene include actin binding and cytoskeletal protein binding. An important paralog of this gene is [PALLD](https://www.genecards.org/cgi-bin/carddisp.pl?gene=PALLD" \t "_blank). | <https://www.genecards.org/cgi-bin/carddisp.pl?gene=MYPN#summaries> |
| SN5A | The protein encoded by this gene is an integral membrane protein and tetrodotoxin-resistant voltage-gated sodium channel subunit.  Diseases associated with SCN5A include [Sudden Infant Death Syndrome](http://www.malacards.org/card/sudden_infant_death_syndrome" \o "See Sudden Infant Death Syndrome at Malacards" \t "_blank) and [Long Qt Syndrome 3](http://www.malacards.org/card/long_qt_syndrome_3" \o "See Long Qt Syndrome 3 at Malacards" \t "_blank). | <https://www.genecards.org/cgi-bin/carddisp.pl?gene=SCN5A&keywords=SCN5A#summaries> |
| PPP1R3A | The glycogen-associated form of protein phosphatase-1 (PP1) derived from skeletal muscle is a heterodimer composed of a 37-kD catalytic subunit and a 124-kD targeting and regulatory subunit. T | <https://www.genecards.org/cgi-bin/carddisp.pl?gene=PPP1R3A&keywords=PPP1R3A#summaries> |
| MYOM2 | Diseases associated with MYOM2 include [Rheumatic Fever](http://www.malacards.org/card/rheumatic_fever" \o "See Rheumatic Fever at Malacards" \t "_blank) and [Blood Protein Disease](http://www.malacards.org/card/blood_protein_disease" \o "See Blood Protein Disease at Malacards" \t "_blank). Gene Ontology (GO) annotations related to this gene include structural constituent of muscle | <https://www.genecards.org/cgi-bin/carddisp.pl?gene=MYOM2&keywords=MYOM2> |

Table 7 Enrichment of top genes

*Enrichr was used to find enriched pathways from significantly differentially expressed genes. Using a list of the top 500 genes and a list of the top 1000 genes. I looked at KEGG pathways, gene ontology Biological process, and gene ontology cellular component.*

|  |  |
| --- | --- |
| Criteria | Enriched pathways |
| Enrichr  P value < 4.0 E-23  500 genes | |  | | --- | |  |   KEGG  Macintosh HD:Users:rebeccafuchs:Desktop:Screen Shot 2020-03-24 at 7.11.07 PM.png  GO Biological Process  Macintosh HD:Users:rebeccafuchs:Desktop:Screen Shot 2020-03-24 at 7.12.12 PM.png  GO cellular componentMacintosh HD:Users:rebeccafuchs:Desktop:Screen Shot 2020-03-24 at 7.13.44 PM.png |
| Enrichr  P value <  3.35E-16  1000 genes | Macintosh HD:Users:rebeccafuchs:Desktop:Screen Shot 2020-03-24 at 7.21.29 PM.pngMacintosh HD:Users:rebeccafuchs:Desktop:Screen Shot 2020-03-24 at 7.22.05 PM.pngMacintosh HD:Users:rebeccafuchs:Desktop:Screen Shot 2020-03-24 at 7.18.54 PM.png |

Table 7 shows the enriched pathways of the lowest p value lists. With both gene lists, pathways related to muscles, the heart, cytoskeleton, and some immune related pathways are found.

**Discussion**

After looking up functions and associations of the top differently expressed genes. I found most were related to the heart and muscle contraction. The pathway enrichment highlights similar related parts of the cell. Some immune system pathways were enriched.

Generally, there were more then 4000 genes up regulated and more then 4000 genes down regulated. There is clearly a very big difference in gene expression between cell in the heart and lymph nodes. This many genes show how these tissues are specialized to perform their function in the body. Their gene expression gives light to the cellular processes that are occurring to make these tissues function in there unique way.

The muscle related genes are involved in the cytoskeleton. The cytoskeleton can move and shapes each cell. Top lowest P value genes, PPP1R3A, MYPN, and MYOM2 are all highly involved it the cytoskeleton. Pathways found such as striated muscle contractions, cardiac muscle contraction, and actin cytoskeleton are all involved in the muscle contraction in the heart.

MYBPC3, MILP, TNNT2, and HHATL are all genes associated with the heart. Some are also involved in cytoskeleton and some are associated to heart diseases. Some pathways associated to the heart are dilated cardiomyopathy, heart contractions, and cardiac muscle cell development. These enriched pathways and association give evidence to the differently expressed genes between the heart and lymph node tissues.

There were specific cytoskeleton related pathways in the shorter 500 genes list analysis like actin cytoskeleton and microfibrils. While the longer gene list captured some immune pathways. The long gene list of 1000 genes found pathways that seem less involved in the heart or lymph node tissues like insulin resistance and sarcoplasm. It is possible these pathways were enriched by chance or maybe there is an underlying pathway of genes that are involved and enriched in these tissues.

There are some immune related pathways found to be enriched in enrichr, like B cell receptor signaling pathways, and primary immunodeficiency. These were found in when using longer list of differently expressed genes. This means there were more genes to look for an association, but they were less significantly differentially expressed. There were not any top genes associated with the immune system

Figure 5. shows a few genes that don’t follow the same pattern as other samples. One gene in particular, BMP10, Bone Morphogenetic Protein 10. This gene is involved in transforming growth factor beta superfamily proteins. It is known to play an important role in cardiovascular development, regulation of heart side, and other heart functions. Perhaps this sample is from a person’s heart that is still developing or had a developmental problem. This could explain the significantly different expression of BMP10.

My hypothesis was correct in that we found genes and pathways related to contractions and the cytoskeleton. There is very differently expression between the heart and lymph nodes because they function in very different ways. The heart must contract constantly to keep blood pumping through the body. The lymph nodes are involved in housing and activating the immune system. The heart cells have a very specific job so there gene expression allows them to contract properly and create the machinery in the cell to do so.

We can see how the heart cells have a very different gene expression then other cells to be able to contract constantly. An interesting follow up to this study would be to compare other muscle tissues to see how the gene expression changes. There could be specific genes only the heart uses for contractions, or maybe all muscle tissues use similar genes and cellular processes.

**References**

*Cariac muscle cell*. (n.d.). Retrieved from wikipedia: https://en.wikipedia.org/wiki/Cardiac\_muscle\_cell

*Lymph node*. (n.d.). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Lymph\_node

T[he Globus Galaxies platform: delivering science gateways as a service](https://doi.org/10.1002/cpe.3486), Ravi Madduri, Kyle Chard, Ryan Chard, Lukasz Lacinski, Alex Rodriguez, Dinanath Sulakhe, David Kelly, Utpal Dave, Ian Foster. Concurrency and Computation: Practice and Experience 29 April 2015 doi: /10.1002/cpe.3486

Nicholson, Lindsay B. “The immune system.” *Essays in biochemistry* vol. 60,3 (2016): 275-301. doi:10.1042/EBC20160017

Robinson MD, McCarthy DJ, Smyth GK (2010). “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.” Bioinformatics, **26**(1), 139-140. doi: [10.1093/bioinformatics/btp616](https://doi.org/10.1093/bioinformatics/btp616).

Stelzer G, Rosen R, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, Iny Stein T, Nudel R, Lieder I, Mazor Y, Kaplan S, Dahary D, Warshawsky D, Guan - Golan Y, Kohn A, Rappaport N, Safran M, and Lancet D. *The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analysis*, [Current Protocols in Bioinformatics(2016), 54:1.30.1 - 1.30.33.doi: 10.1002 / cpbi.5.](http://onlinelibrary.wiley.com/doi/10.1002/cpbi.5/abstract" \t "_blank)[[PDF]](https://cdn.genecards.org/papers/Protocols-2016-Stelzer-GCSuite.pdf" \t "_blank)

**Supplementary data**

R code:

#seperate counts and gene names

y <- DGEList(counts=cancer\_counts[,2:7],genes=cancer\_counts[,1])

#normalize

y <- calcNormFactors(y)

#name samples

y$samples$group = c("H", "H", "H", "L", "L", "L")

#MDS plot

plotMDS(y)

et <- exactTest(y, pair=c("H","L"))

summary(de <- decideTestsDGE(et))

detags <- rownames(y)[as.logical(de)]

#for smear plot

plotSmear(et, de.tags=detags)

abline(h=c(-1, 1), col="blue")

#top p value genes

diffExpGenes <- topTags(et, n=1000, p.value = 0.05)

head(diffExpGenes$table)