Project 2: Transcriptional Profile of Mammalian Cardiac Regeneration with mRNA-Seq

Alec Jacobsen, Daisy Wenyan Han, Divya Sundaresan, Emmanuel Saake

Analyst Data Curator Biologist Programmer

ENG BF528, Spring 2021

Introduction

The mammalian heart is known to exit the cell cycle shortly after birth, with the majority of further growth resulting from cell hypertrophy rather than proliferation. These cardiac cells, however, are believed to retain their innate ability to regenerate, to some degree. In previous studies, it had been discovered that while adult mammalian hearts were unable to regenerate following injury, neonatal mice appeared to retain such ability, following resection of the left ventricular apex (Porrello et al., 2011). O'Meara et al. therefore set out to characterize the transcriptional changes that define these two distinct phenotypic outcomes in their article, *Transcriptional Reversion of Cardiac Myocyte Fate During Mammalian Cardiac Regeneration.*

The objective of O'Meara's study was to determine whether the cardiac myocytes of neonatal mice revert to a less differentiated state during cardiac regeneration, and to characterize the transcriptional changes that occur during this time. Similarly, in our study, we attempted to replicate the first portion of this analysis by determining the differentially expressed genes, via interrogation of the transcriptional data used in the original analysis, to validate the findings regarding a single subset of neonatal mice.

The authors interrogated the RNA sequencing datasets in order to identify genes and gene networks that changed during the differentiation process, and explain the transcriptional changes required for a fully differentiated phenotype to re-enter the cell cycle. The use of RNA-sequencing data gave the researchers the opportunity to identify key transcriptional changes that occurred during this process, as well as predict upstream and downstream changes that may occur. The identification of the biological pathways in which these genes play a significant role then provided significant insight into the signaling pathways of the myocyte cell cycle, as well as those required specifically for cardiac regeneration.

Data

For the project, we analyzed the GSM1570702 sample from the GEO Series GSE64403 data set. This dataset, made public in December of 2014, contains thirty-six Mus musculus RNA-Seq samples, and was used by O'Meara et al. for their analysis of transcriptional reversion of cardiac myocyte fate. While various samples, taken at differing times and under differing circumstances, are available in this dataset, we downloaded the GSM1570702 sample (labelled P0_1) for our analysis.

The GSM1570702 sample came from the day zero, postnatal ventricular myocardium of a mus musculus. The RNA was extracted using Trizol, with RNA quality determined by Agilent Bioanalyzer. The RNA libraries were then prepared using an Illumina TruSeq kit, on the Illumina HiSeq 2000. Basecalling was performed using the Illumina Offline Basecaller software, with Bowtie, Tophat and Cufflinks used to determine gene expression levels, with the mm9 genome used as reference, in the original study. These methods will therefore be repeated in our own analysis. The unprocessed sample was downloaded as an SRA file (SRR1727914) - with further details regarding quality control outputs and download link found in the *Data Availability* section.

The SRA file SRR1727914 contains 21.6 million paired-end reads, across 1.7 giga-base-pairs. Each read is forty base pairs long. The SRA file was then extracted using the sratoolkit package to produce two paired-end FASTQ files, which were examined for quality using the fastqc package.

The fastqc analysis of the two FASTQ files generated HTML reports on data quality. Quality scores were acceptable in both files, with most per-base sequence quality scores above the cut-off of 28. Therefore, zero sequences were flagged as poor quality and requiring removal, with the majority of per-sequence quality scores between the range of 32 to 38. The per-sequence GC content was as expected, and followed the theoretical distribution well. Low per-base N content was observed throughout. No overrepresented sequences were noted, with little adapter content.

Of note, a warning message was generated regarding the per-base sequence content for both files. However, libraries produced by priming are known to inherit intrinsic bias at the start of the read positions. This is common in nearly all RNA-Seq libraries, and is not believed to affect downstream analysis. Additionally, a warning was generated for duplicate sequences, as the percent of sequences remaining if deduplicated was 50.29%. This warning is also known to be common for RNA-Seq libraries. In order to observe lowly expressed transcripts, it is common to over-sequence highly expressed transcripts, thereby potentially creating a large number of duplicates. This is not thought to influence downstream analysis.

A note of interest during the data acquisition process was the discrepancy between the submission and update dates for this particular SRA file. The initial data was submitted on December 19, 2014, which is inline with the original publication date of the O'Meara et al. paper. However, the data appears to have been updated on May 15, 2019, and was labelled "SRR1727914.1". While it is unclear whether this updated dataset will have a significant impact on the final results of our analysis, this fact may be important to keep in mind when evaluating our ability to replicate the findings of the original paper.

Methods

RNA-Seq Alignment

Tophat maps sequences from spliced transcripts to genomes. Using Tophat, we aligned paired reads (In fastq format: *P0_1_1.fastq* and *P0_1_2.fastq*) to the mm9 reference genome. The process took approximately 54mins on a shared cluster using 16 processing cores. All reads and alignment discovered by tophat was written to a bam file, *accepted_hits.bam*. Below is the alignment summary:

```
Left reads:
      Input    : 21577562
       Mapped   : 20828891 (96.5% of input)
        of these:  1467266 ( 7.0%) have multiple alignments (47541 have >20)
Right reads:
      Input    : 21577562
       Mapped   : 20429297 (94.7% of input)
        of these:  1426406 ( 7.0%) have multiple alignments (46449 have >20)
95.6% overall read mapping rate.
Aligned pairs:  19851441
   of these:  1370320 ( 6.9%) have multiple alignments
            787327 ( 4.0%) are discordant alignments
88.4% concordant pair alignment rate
```

Mapping-Quality Check and Summary of Flags in Aligned Sequence

The mapping quality of the tophat alignment is checked using "samtools flagstat". This analysis is based on the flag fields of the *accepted_hits.bam* file. Each flag field is examined category by category, and graded as QC passed or QC failed. The mapping quality summary is as follows:

49558971 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 duplicates
49558971 + 0 mapped (100.00%:-nan%)
49558971 + 0 paired in sequencing
25033981 + 0 read1
24524990 + 0 read2
32290540 + 0 properly paired (65.16%:-nan%)
47575446 + 0 with itself and mate mapped
1983525 + 0 singletons (4.00%:-nan%)
5069694 + 0 with mate mapped to a different chr
698456 + 0 with mate mapped to a different chr (mapQ>=5)

Quality Assurance and Evaluation of Aligned RNA-Seq data

The aligned RNA-seq is comprehensively evaluated using the RSeQC RNA-seq Quality Control Package. The bam file is first optimized by indexing the aligned RNA-seq data (accepted_hits.bam) to generate a *bai* file, *accepted_hits.bam.bai*. Three RSeQC utilities were executed in this process: *geneBody_coverage.py*, *inner_distance.py*, *bam_stat.py*.

The geneBody_coverage.py utility was used in capturing the RNA-seq reads over genebody, and a plot **Figure 2.1** was obtained. *Inner_distance.py* calculated the inner distance between paired reads of RNA-seq fragments and generated a plot, **Figure 2.2**. *Bam_stat.py* summarized mapping statistics of a BAM or SAM file. **Figure 2.1** illustrates the uniform coverage over genebody after transcripts were scaled to 100 nucleotides.The bam file statistics shown in **Figure 2.3** indicates a 0.0 QC fails.

Quantifying Gene Expression

We accounted for how the reads mapped to genomic regions by running *cufflinks* on the annotations file *mm9.gtf*, the index mm9.fa and the accepted_hits.sam file. The output of the cufflinks execution included a *gene.fpkm_tracking* file. The Fpkm values of 0.0 in the data of *gene.fpkm_tracking* file were dropped and the fpkm stats were plotted as shown in **Figure 2.4**.

Identifying Differentially Expressed Genes

Running the cuffdiff utility of cufflinks on the indexed *accepted_hits.bam*, P0_2, Ad_1, and Ad_2 data, the differentially expressed genes were identified and summarized into the *gene_exp.diff* file.

Table 1.1 : Modules and data files used for aligning and quality control

Modules	Data files used with modules
Tophat : v2.1.1/ Bowtie 2.4.2.0	Gene Model Annotation data: mm9.gtf, reads in fastq format, mm9.fa index

Samtools/0.1.19: *flagstat *index *view	accepted_hits.bam
python3, rseqc/3.0.0 *geneBody_coverage.py *inner_distance.py *bam_stat.py	mm9.bed file, accepted_hits.bam
cufflinks/3.0.0 *cufflinks *cuffdiff	mm9.bed file, Accepted_hits.bam,
* subtool or sub-utility used	

Differentially Expressed Genes

The cufflinks output was imported into R and genes with a significant difference in expression were subset, i.e. genes with a false discovery rate less than 0.05. Those genes were then further separated by the directionality of their difference, either being upregulated in adult mouse cardiac myocytes as compared to neonatal myocytes, or downregulated in the adult myocytes. Separate lists of official gene symbols for upregulated and downregulated genes were then imported into the Gene Functional Classification Tool of the Database for Annotation, Visualization, and Integrated Discovery (DAVID) for functional cluster analysis (Huang et al. 2009). Genes were clustered based on their gene ontology (GO) terms, with GO terms in the biological processes, molecular function, and cellular components categories all being included in the analysis. All plotting was done with ggplot2 in R (Wickham 2016).

Results

RSeQC Output : Gene_body coverage and Insert_distance analysis of aligned P0 reads

The gene body coverage analysis of the aligned paired read from the mRNA sequence of the postnatal day 0 heart sample shows quite a good coverage from 5' to 3' ends. No noise in the plot but a little bias to the 3' end in Figure 2.1.

Figure 2.2 shows estimations for the inner distance between the paired reads, P0_1_1 rand P0_1_2. This mRNA distance apart is negative when two fragments overlap; and this number is 0. The histogram shows a uniform distribution, and the distance between the two fragments is averagely 85bp and a standard deviation of 43.3..

From the bam stats plot in Figure 2.3 we observer a 0 QC-fail, and 0 unmapped reads,

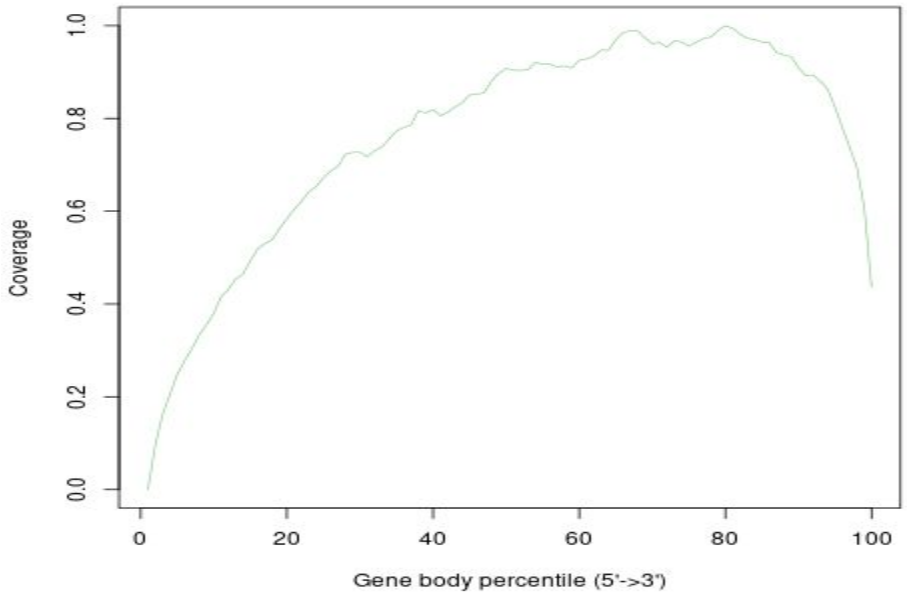


Figure 2.1: Coverage uniformity over gene body after all transcripts were scaled to 100 nucleotides.

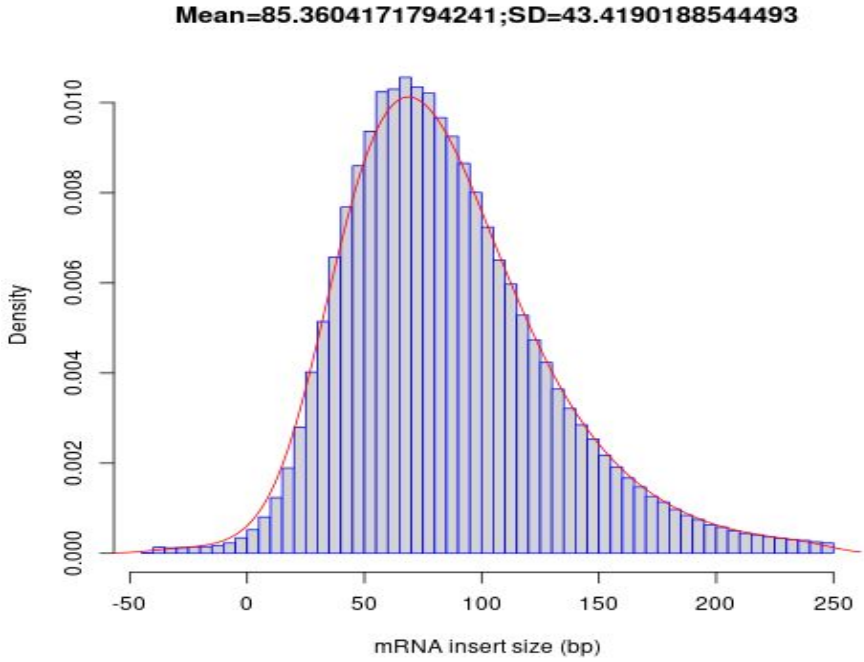


Figure 2.2: Inner distance(insert distance) between paired reads

Figure 2.3: Bam_stat dot plot showing mapping statistics of reads that are QC failed, unique mapped, splice mapped, mapped in proper pair

Quantified Alignment in FPKM

The quantified alignments in FPKM for all genes were illustrated in two plots; Figure 2.4A without normalization and 2.4B after normalizing the FPKM values. The single bar in Figure 2.4A captures where the FPKM values are mostly centered and the red overlay breaks down the quantification of the FPKM values further. In Figure 2.4A we observe few genes with outlier FPKM. After normalizing we get a finer illustration of gene quantification in 2.4B which shows most of FKPM values are a little above 0, specifically with a mean of 0.008. Eliminating genes with FPKM values of 0, a total of 16527 genes were identified...filtering out unique genes based on gene_short_name resulted in 16400 genes.

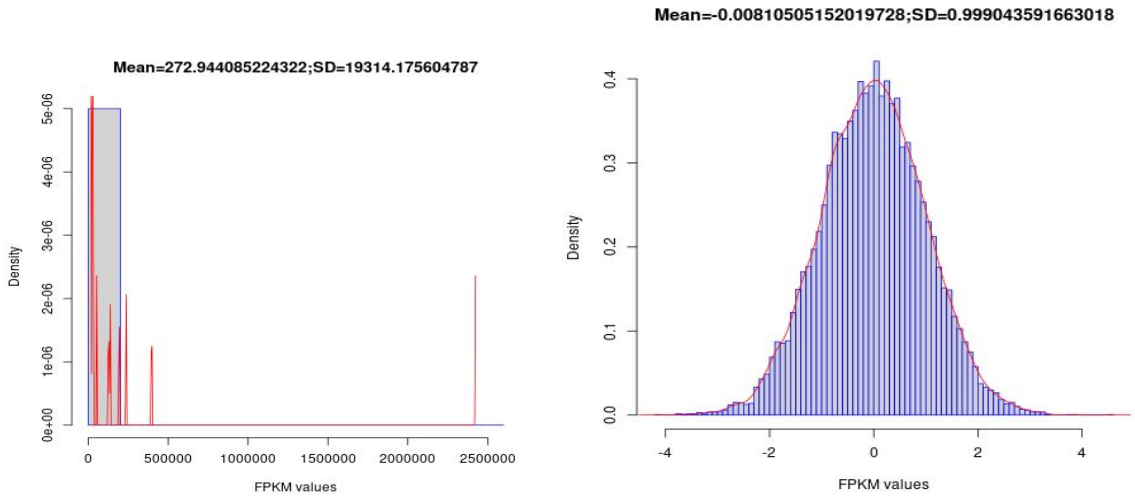


Figure 2.4: A plot (histogram) of genes.fpkm_tracking values. 2.4A) A density-based histogram plot of the fpkm values without normalization 2.4B) A histogram illustrating the fpkm_tracking values of genes after the values have been normalized.

Differentially Expressed Genes

The histogram of the number of significant and non-significant differentially expressed genes had a distribution of values centered around 0, tapering exponentially to either side. The vast majority of mRNA transcripts sequenced had either no change in expression, or a change not significantly associated to the age of the mouse from which the samples were taken. For those genes whose difference in expression was found to be significant, the majority had a Log2 fold change greater/less than ±0.15, with a distribution centered around zero and tapering exponentially to either side. 1091 genes were found to be significantly upregulated in adult mouse cardiac myocytes as compared to neonatal myocytes, and 1032 genes were found to be significantly downregulated in adult myocytes as compared to neonatal myocytes.

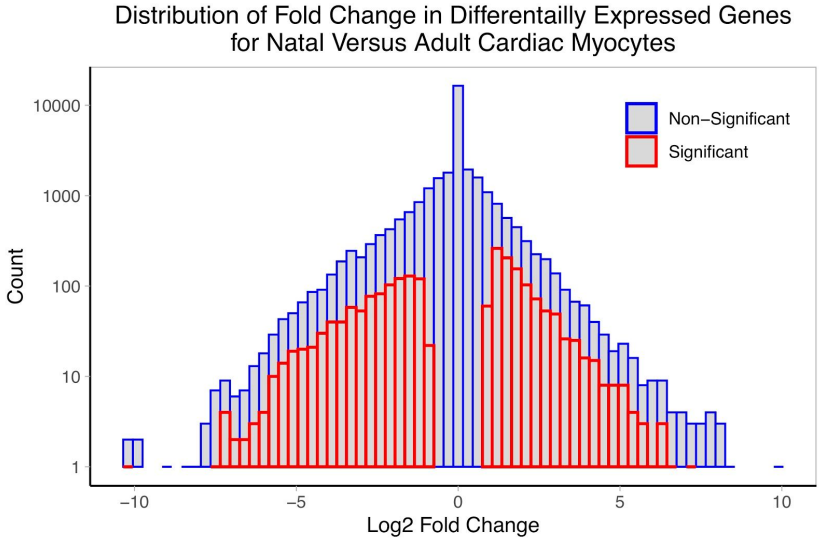
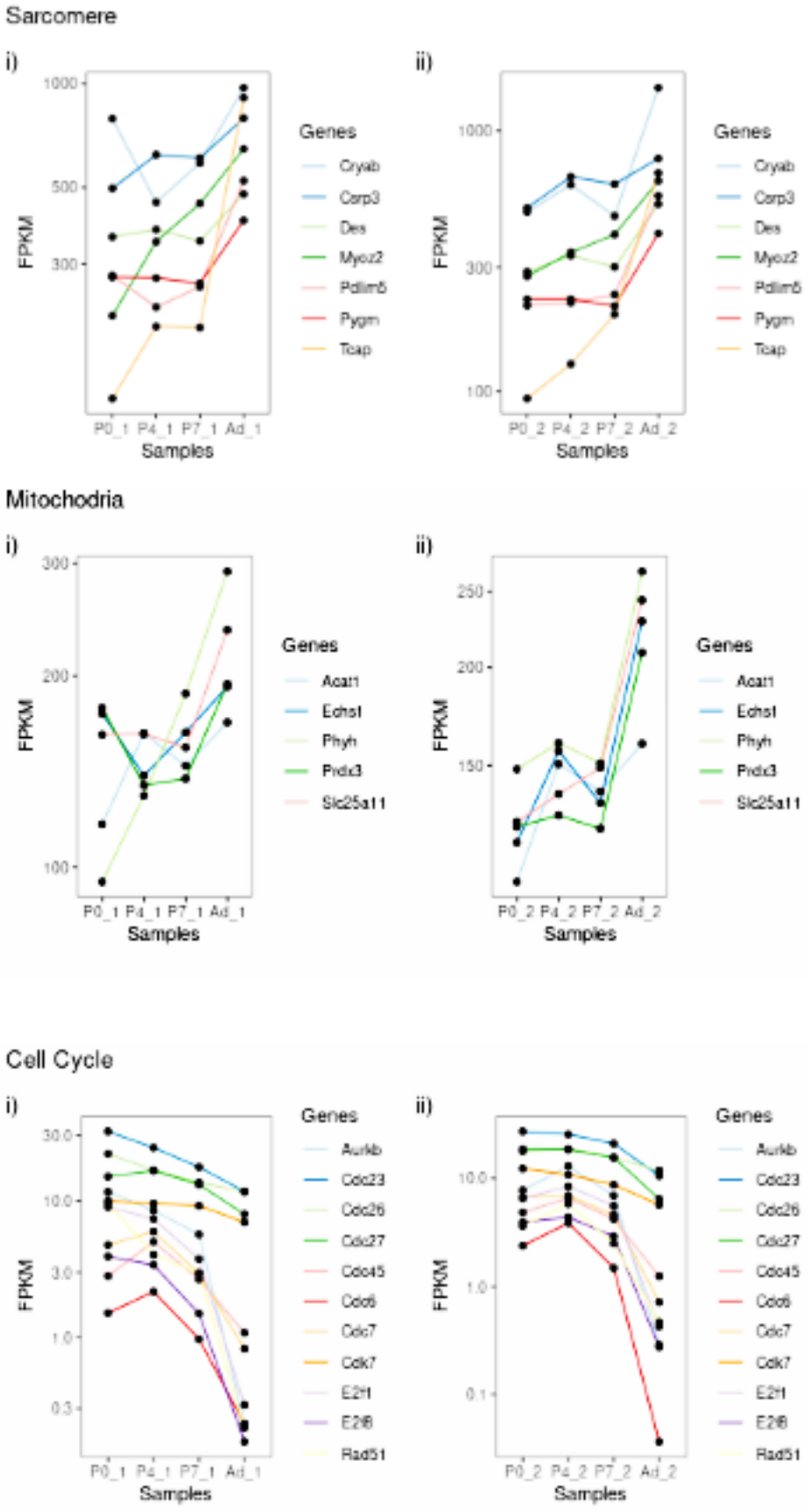


Figure 3: The distribution of the Log2 Fold changes for significantly and non-significantly differentially expressed genes between neonatal and adult mouse cardiac myocytes.

The Functional Gene Classification tool from DAVID revealed that, for genes up-regulated in adult mice, GO terms generally fell into clusters associated with mitochondrial function, metabolic processes for ATP generation, the production of extracellular vesicles, and myofibril/sarcomere maintenance. For downregulated genes, the GO terms were associated with the cell cycle, mitosis, and cellular division processes such as chromosome condensation. The top five annotation clusters are summarized in table 2 by the common cellular component, molecular function, or biological process to all GO terms per cluster, as well as the enrichment score.

Table 2: Common elements between GO terms within the top five enriched clusters output from the DAVID Functional Gene Classification tool for up-regulated and downregulated genes in neonatal and adult myocytes.

Common Up and Downregulated Gene Enrichment Terms			
Up-regulated		**Down-regulated**	
Enrichment Term	Score	Enrichment Term	Score
Mitochondrion	21.15	Cell cycle	16.90
Respiration/Metabolism	16.88	Mitosis	10.21
Organic Acid/Lipid Metabolic Processes	14.03	Extracellular Matrix	10.17
Extracellular vesicles	10.93	Cell Proliferation	9.79
Myofibril/Sarcomere	7.20	Chromosome Condensation	8.29



P0 vs. AD line Plots

Figure 4

These are line plots generated from samples trying to replicate Figure 1D in the paper.As there were only two replicates I decided to plot both, sample 1 on the left and and sample 2 on the right. The x-axis are the days and the y-axis is the FPKM counts. We can see they both sample 1 and 2 generally agree with each other so I will combine their results. Specifically we are trying to compare the Fragments Per Kilobase of transcript per Million mapped reads for P0 vs. Ad. P0 referring to postnatal day 0 and Ad referring to Adult mice 8–10 weeks of age, We can see a general upward trend for selected genes for the sarcomere as well as the mitochondria. The cell cycle we see a decline in FPKM from P0 to Ad. Range of FPKM for each category: Sarcomere is ~ 100-1000, Mitochondria ~100-300. Cell cycle ~ 0.1 -30.

** The genes picked for the line plots were selected based on the genes selected from the original paper.*

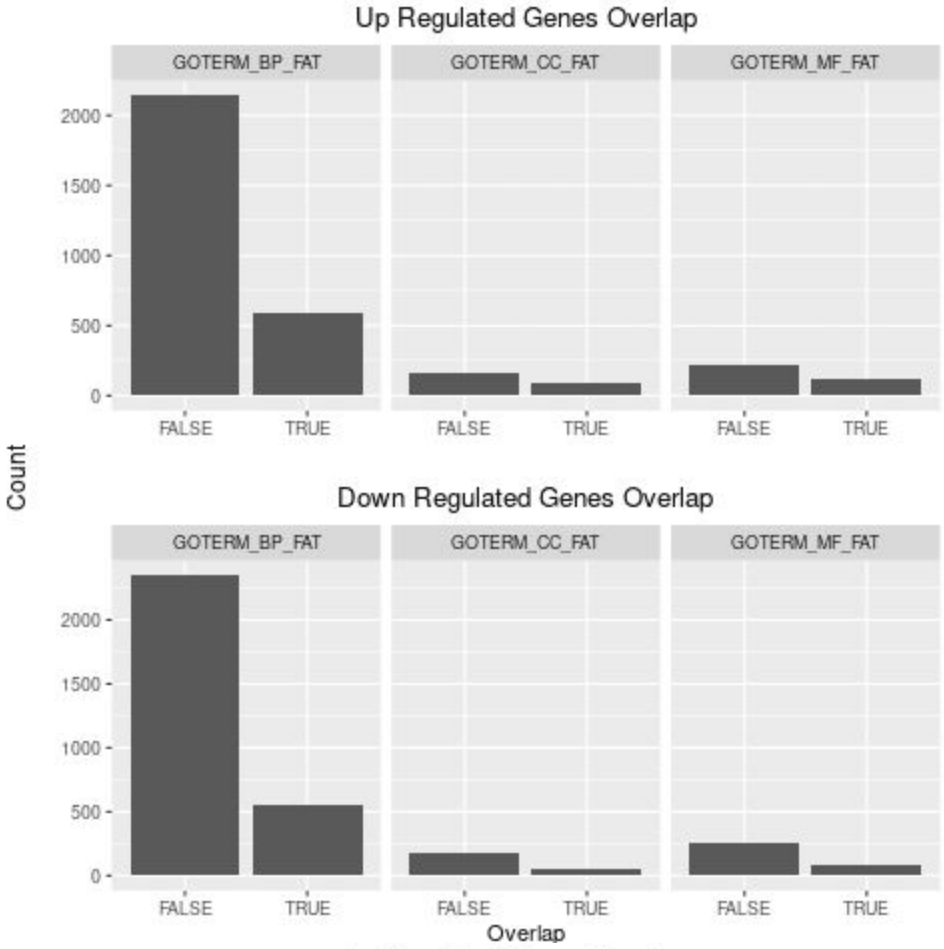


Figure 5.1
These barplots are visualizing the overlapped GO terms for each sub ontologies: Biological processes (BP), Cellular component (CC), and Molecular function (MF). False means no overlap whereas True means the term was found in our analysis as well as the paper. We can see there is a much higher non overlap in each category and the biological processes seems to have the most GO terms for up and down regulated.

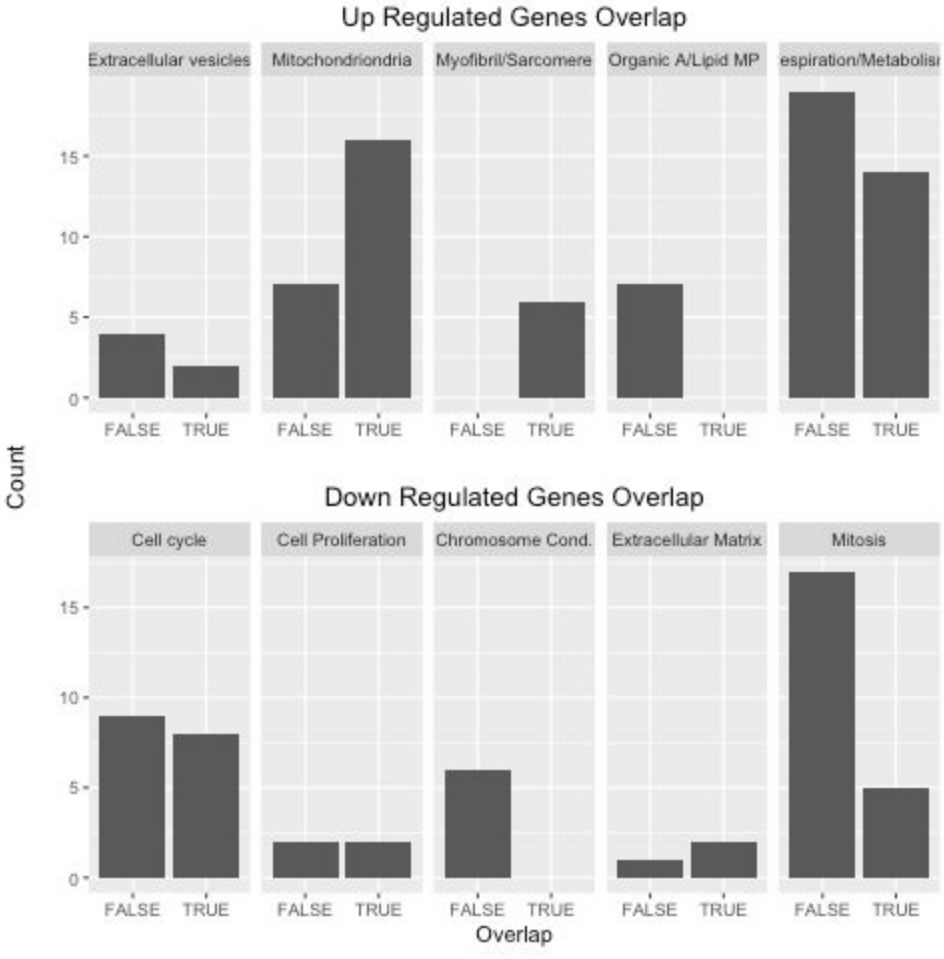


Figure 5.2
The bottom two graphs are grouped by the top five clusters and see if they have a different pattern than the just looking at sub ontologies. We can see that there is not much of a pattern; some groups such as Sarcomere have all overlaps whereas Organic Acid/Lipid MP and Chromosome Condensation have no overlaps. The rest are ~40-50% overlap.

Extended DAVID tables can be found below: ***Deliverables 7.2***
Upregulated: https://github.com/BF528/project-2-lava-lamp/blob/master/Up_regulated_extended.csv
Downregulated https://github.com/BF528/project-2-lava-lamp/blob/master/Down_regulated_extended.csv

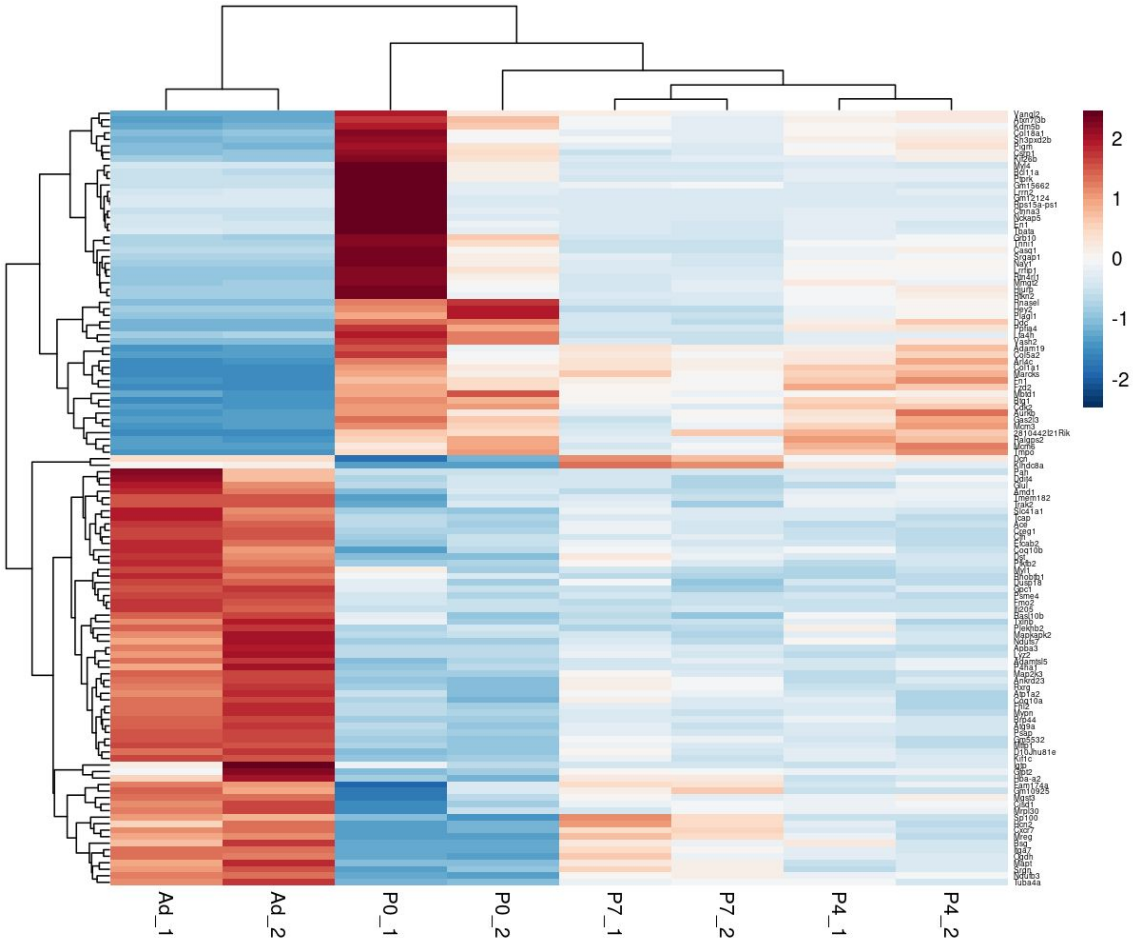


Figure 6: Top 120 differentially expressed genes
This is a clustered heatmap using all samples visualizing 120 of the top differentially expressed genes. With the dendrogram we can see that P7, P4 are the closest with then P0 following, then Ad is clustered by itself, and is the least related. The list of differentially expressed genes are on the right however due to the size it is tough to see the more important takeaway is the clustering which can easily be seen through the colors. Most samples on the same day seem to align with each other, interestingly our sample that we ran P0_1 seems to not be aligning with p0_2.

Discussion

Primary Results Summary

Functional gene classification of differentially expressed genes in neonatal and adult cardiac myocytes show that for the adult myocytes, upregulated genes are generally involved in structures that would be expected for cardiac tissues, such as sarcomeres and myofibrils. Additionally, genes corresponding to energy production such as mitochondrial genes and genes associated in the respiratory metabolic

pathways were upregulated, as would again be expected for cardiac tissues with higher energy requirements.

For downregulated genes, functional gene classification found that nearly all of the substantially enriched clusters were associated with various processes of the cell cycle, cellular division, and growth. This suggests that cell cycle processes are significantly reduced in adult myocytes as compared to the neonatal myocytes. These transcriptional differences between neonatal and adult cardiac myocytes are reflective of the phenotypic behaviour of the myocytes at the different stages of development, with the adult myocytes being highly differentiated and unable to regenerate, while the neonatal myocytes are still capable of growth and mitosis.

The findings of our analysis therefore corroborate those of O'Meara et al., in that it appears the neonatal cardiac myocytes are able to revert to a less differentiated state in order to repair the inflicted injuries. It is this state that corresponds to the downregulation of mitotic, cellular proliferation and cell cycle pathways that were identified in adult cardiac myocytes, in relation to that of the neonatal cardiac myocytes, as illustrated in *Table 2*. Biologically, this suggests that with further research, it may soon be possible to identify key transcriptional regulators and selectively induce the proliferation of cardiac myocytes. This ability would be especially useful in cases where significant damage has been done to the heart tissue, such as following a myocardial infarction.

On the other hand, the number of genes output by our analysis was significantly less than the number reported in O'Meara et al., with only 1091 upregulated genes discovered by our analysis as compared to 1482, and 1032 downregulated genes in our analysis as compared to 4341. This could be due to different parameters used during alignment in tophat and gene expression quantification in cufflinks, or simply due to more recent versions of both tools being used in our analysis than in O'Meara et al.

Comparison of DE Genes to O'Meara et al.

Figure 4 took genes that were important to specific processes in the original paper and measured their FPKM. If we take a look at figure 1d in the paper we see that we were actually able to replicate the same trends with both the Sarcomere and Mitochondria increasing and the Cell cycle decreasing for these specific genes; the ranges of FPKM are also similar. FPKM normalizes the reads to the length so in RNA-seq the relative expression of a transcript is proportional to the number of cDNA fragments that originate from it therefore we can use this to compare gene A to gene B at different postnatal days. As mentioned in the paper these specific genes have important functions and as postnatal day increases, over the course of differentiation we see increases in expression in Sarcomere which potentially reflects sarcomere assembly and organization during cardiac myocyte differentiation and maturation (O'Meara, 2015) .

Comparison of GO Terms to O'Meara et al.

We similarly did a comparison of GO terms found in the paper vs. the ones that were found in our DAVID analysis **Figure 5.1, 5.2**. Our clusters ended up in a different order than the paper but did have some overlap such as the top upregulated cluster being the mitochondria. Also among these clusters there was around 40-50% overlap of specific GO terms that occurred. This is interesting because considering we used the same samples it seems that we are getting slightly different terminology. In this case we were

not able to reproduce the exact results from the paper. It would be interesting to compare all the different teams GO terms against each other, as it seems to vary quite a bit. As the GO clustering is based on an algorithm there must be something from the input, the genes list that is varying and that list goes back to our original analysis. So knowing this we can see we would have to further test if this is normal or if something went wrong in our original analysis. Based on our analysis though it would be interesting to find important genes related to respiration and create line plots such as in **Figure 4** to gain more insight into specific genes and pathways related to respiration that are important during mouse development.

Comparison of Heatmap to O'Meara et al.

The final portion left to compare to the original paper is **Figure 6** this is a clustered heatmap looking at the top 120 differentially expressed genes for P0 vs. Ad, we can compare this to figure 2A from the original paper. It seems that our clustering overall matches the paper with the P0, P4, P7 clustering together and Ad being quite different than those. I am also not sure why the p0_1 sample has quite a bit of higher expression in regions that do not align with the p0_2 sample. This could be due to the sample itself or some error in the analysis that occurred. However, despite that region of higher expression it still is clustering with the other postnatal samples though perhaps the most distant. It would be interesting to further develop this heatmap by having a bar which clustered the genes into groups with common functions and then we would be able to visually tell more about which specific genes groups are more active or repressed in Adult vs. Postnatal day 0. For now I was able to pick out genes such as Tcap and a couple of others that were mentioned in the **Figure 4** line plots and see that the heatmap and line plots align with these specific genes either being expressed at higher/lower values depending on which sample they are in. In our case because we got higher GO terms dealing with respiration it would be interesting to take those genes dealing with respiration that are clustering together and create some protein protein interaction maps through Stringdb to further see how these specific gene processes are related to each other and which pathways that they are a part of could be regulating part of the mouse development.

Conclusion

[Collectively] While we were unable to replicate the exact results from the O'Meara et al. study, the final conclusion we hope to convey to our readers remains the same: there is a significant difference in the transcriptional regulation of neonatal and adult cardiac myocytes following injury. Our results were able to reaffirm that the neonatal cells undergo significant dedifferentiation following injury, allowing them to re-enter the cell cycle and proliferate. The ability of cells to retain this ability following injury opens the door to incredible possibilities in the biological and medical fields, which we hope will one day translate to new, innovative diagnostic tools and treatments for currently known cardiac diseases.

[Data Curator] As the data curator for this project, there were few, if any, problems or challenges encountered in this project for my portion of the analysis. One aspect I did notice was that we were possibly using an updated version of the SRA file, which could possibly lead to differing final results. In order to resolve this issue, the original version of the SRA file would be required, but it did not seem to be available on the GEO website. Additionally, while not necessarily pertaining to my portion of the analysis, it came to my attention that the TopHat and Cufflinks tools appear to be deprecated at this time. While these analysis methods align with those used in the original paper, the updated SRA file was released after these tools had stopped being updated, potentially leading to discrepancies further down in

the analysis. Alternatively, the same analysis could be repeated using more updated tools. This would provide an interesting comparison of both the genes themselves, as well as elucidate possible biases implicit in particular softwares.

[Programmer] Using tophat pipeline with bowtie2, paired reads from the heart samples of mice were aligned to the mm9 reference genome. Quality control was undertaken with the RSeQC utility package, and Cufflinks was used for quantifying gene expression. Each of these executions were few lines of instruction without challenges. The key challenges were module load errors which were resolved initially by direct module load from terminal, coupled with line by line module load in bash scripts.

[Analyst] The analysis section was fairly straight forward, with the majority of the work involving simple subsetting of the data using features present in the dataset. Running the Functional Gene Classification tool on DAVID was also fairly straightforward. The challenge was interpreting the output, as DAVID does not label the clusters, so the clustering must be inferred by the GO terms present in each cluster. This was resolved by searching the unknown terms on google to understand how they related to the other terms in each cluster, and the gist of each cluster could be summarized in table 2.

[Biologist] From the RNA-seq analysis we see how specific groups of genes related to important developmental processes in mice are being expressed. This information allows us to narrow down which genes are important to consider for conducting further research in relation to myocyte regeneration. This would be of specific interest as the transcriptional changes that occur during mammalian cardiac regeneration have not been fully characterized at the molecular level (O'Meara, 2015). So understanding this in mice would lead to further developments in heart injury research. For specific challenges I encountered, I had a couple problems with duplicate genes which I was able to average for the heatmap. The other challenge was pinpointing why analyses of the same sample data are leading to slightly different results with GO clustering.

References

Hadley Wickham et al., (2021). dplyr: A Grammar of Data Manipulation. R package version 1.0.4. https://CRAN.R-project.org/package=dplyr

Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*. 2009;37(1):1-13. doi:10.1093/nar/gkn923

Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. 2009;4(1):44-57. doi:10.1038/nprot.2008.211

Kolde, R. (2013). pheatmap: Pretty Heatmaps. R package version 0.7.7. http://CRAN.R-project.org/package=pheatmap.

O'Meara CC, Wamstad JA, Gladstone RA, et al. Transcriptional reversion of cardiac myocyte fate during mammalian cardiac regeneration. *Circ Res*. 2015;116(5):804-815. doi:10.1161/CIRCRESAHA.116.304269

Porrello ER, Mahmoud AI, Simpson E, et al. Transient regenerative potential of the neonatal mouse heart. *Science*. 2011;331(6020):1078-1080. doi:10.1126/science.1200708

RCoreTeam(2013).R:A language and environment for statistical computing.Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Trapnell C, Williams BA, Pertea G, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*. 2010;28(5):511-515. doi:10.1038/nbt.1621

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Data Availability

All code used to generate our analyses and quality control metrics can be found at:
https://github.com/BF528/project-2-lava-lamp
Data is accessible at NCBI GEO database (O'Meara et al., 2013), accession GSE64403.