Will Mischler - Data curator
Reva Shenwai - Programmer
Nitsueh Kebere - Analyst
Xinyu Sun - Biologist

TA: Dakota Hawkins

Introduction

Neonatal mice have the ability to regenerate their heart tissue in response to injury but lose this response within the first week of development (Bicknell et al. 2007). The purpose of this study was to identify the potential regulators for the mammalian cardiac regeneration in neonatal mice and if myocytes revert transcriptionally to a less differentiated state during regeneration. The group used RNAseq data obtained from various stages of injury-induced cardiac myocyte regeneration to analyze the transcriptional changes that may cause this regeneration. The authors discovered a transcriptional reversion of cardiac myocyte differentiation processes such as reactivation of latent developmental programs. Interleukin 13 (IL13) was identified as a potential upstream regulator that induced myocyte cell cycle entry and STAT3/6 signaling. They demonstrate that these signaling molecules are crucial to IL13 signaling in myocytes as well as modulated in regenerating cardiac tissue (O'Meara et al. 2015).

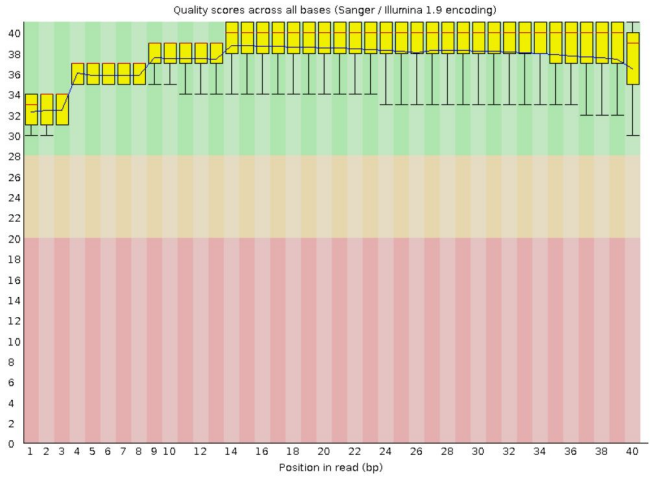
Data

Neonatal cardiac myocytes were dissociated from whole mouse hearts by authors using the Neonatal cardiac myocyte isolation Kit (Miltenyi Biotec). Total RNA was extracted from samples using Trizol (Invitrogen) and polyadenylated RNA was isolated using Dynabeads mRNA purification kit (Invitrogen). This RNA was fragmented and the first strand was synthesized with the Superscript III reverse transcription kit (Invitrogen). Double stranded DNA was synthesized with DNA polymerase I (Invitrogen). End repair, A-tailing and size selection were then performed using the SRPI-Works System (Beckman Coulter). This was followed by minimal amplification and addition of barcodes by PCR. Illumina HiSeq 2000 was used to perform paired-end 40 base pair read length sequencing of the samples. Sequence alignment and assembly was performed as described in Trapnell et al.(2009). Due to low RNA yield, purified neonatal cardiac myocyte samples underwent TrueSeq sample preparation protocol (Invitrogen) prior to RNAseq.

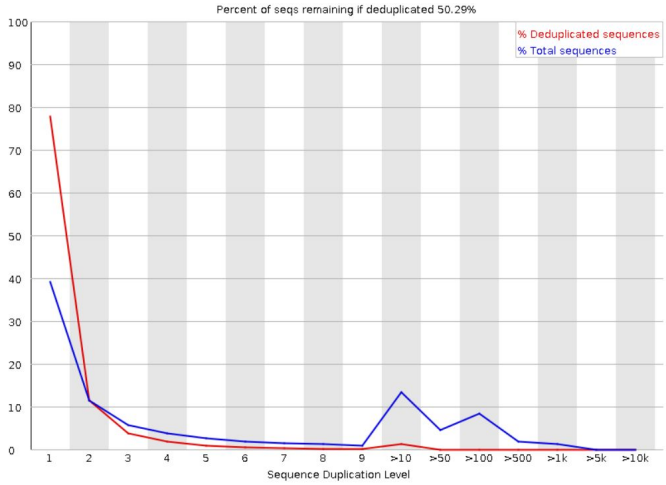
RNAseq data for this project was downloaded from Gene Expression Omnibus (GEO) Series GSE64403, specifically sample GSM1570702 (P0). After using SRA tools to convert data to FASTQ format, the data was assessed for quality using the FastQC tool on the cluster. Each read had passing overall quality scores and showed no GC content bias (Fig. 1A&C). Sequence

duplication levels showed that there may be an enrichment bias due to PCR over amplification or sample contaminants are present (Fig. 1B). The base sequence content also raises a warning during the quality check but as this occurs towards the beginning of the sequence, this should not adversely affect downstream analysis.

A



B



C

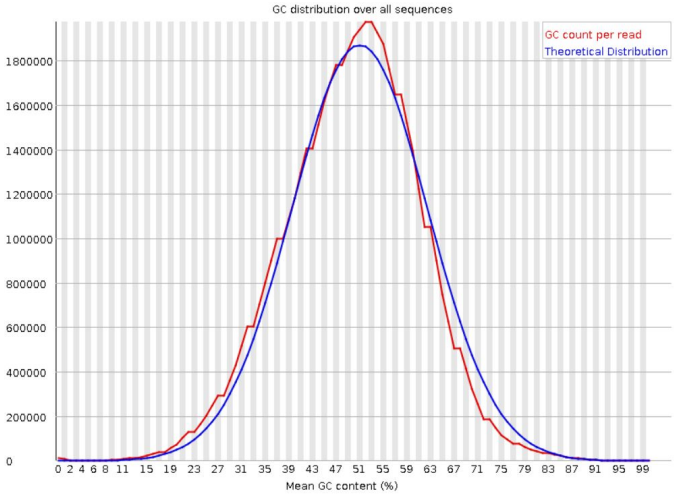


Figure 1. Quality scores, duplicates and GC content. Representative images for quality control run on sequences. A) Shows the overall quality scores which indicate that samples were of good quality. B) Indicates that there is an increase in duplicated sequences at level 10 and 100. Possibly due to contaminants or over amplification during PCR. C) GC content hovers around ~49% which passes checks.

Methods

Once the FASTQ files were quality controlled (as outlined in Data), further analyses were run using several bioinformatics tools on Linux shell. These are summarized in this section.

Tophat 2.1.1 was first used to align the FASTQ files against mm9, which is a mouse reference genome. Tophat is a fast splice junction mapper for RNA-seq reads, and aligns said reads to large mammalian genomes using the ultra high-throughput short read aligner Bowtie (John Hopkins University, 2016). It then analyzes results of this mapping to identify splice

junctions between exons. Due to the large memory requirement, it was run as a batch job, with parameters bowtie reference index and the previously obtained FASTQ files. It took about an hour to complete. The result was a BAM file which contained the original reads and any alignments discovered by TopHat. Flagstat, a tool in the Samtools 1.10 repertoire, was used to generate summary statistics for the BAM file (table 1). Then, RSeQC tools were also run on the BAM file for quality control of the data. Some key summary statistics for this RSeQC output (table 1), and graphs showing gene body coverage (fig. 2) and mRNA read inner distance (fig. 3) are described under Results.

Cufflinks 2.2.1 was then run on the BAM file to assemble the transcripts, estimate abundances, and test for differential expression/regulation (Trapnell, 2017). It was run as a batch job using reference sequence data and the BAM file as parameters, which took about an hour to complete. Cufflinks produced quantified alignments in FPKM (Fragments Per Kilobase per Million mapped reads) for all the analyzed genes. FPKM is a normalized estimation for gene expression data based on gene length and sequencing depth of reads mapped to each gene sequence (EMBL-EBI, 2020).

Figure 4 shows a graphical representation of the FPKM values; rows with FPKM = 0 were removed since these represent transcripts without any aligned reads, and to allow for log-transformation. Figure 5 shows the same data as in figure 4, but with log-transformation. The number of genes before and after filtering steps are provided in table 2 and the Results section.

Finally, cuffdiff, a tool in the cufflinks package, was used to identify differentially expressed genes (DEGs)(Trapnell, 2014). It was also run as a batch job using parameters reference FASTA for bias correction, a gene transfer format file containing transcript data, and BAM files containing hits, which took 2-3 hours to complete. The differential expression statistics file produced by this tool captured gene-level differential expression by comparing FPKM of transcripts sharing each gene ID, and was used in further analysis (Trapnell, 2017).

Results

Key summary statistics generated using Flagstat and RSeQC tools on the alignments produced using Tophat can be found in table 1. Figures 2 and 3 show graphs produced by RSeQC tools. Interpretations for these analyses are described in each figure's caption, but overall, genes appear to have a good average coverage of ~80% and average mRNA fragment lengths of 50 - 100 bp.

FPKM normalization results are summarized in figures 4 and 5. Figure 5 shows log-normalized FPKM values, and figure 4 allows visualization of the patterns in FPKM values without log-normalization. Fig 4(b) only shows 0<FPKM values<=100 to better visualize the data pattern without a log transformation, and there appears to be an exponential decrease in frequency with increasing FPKM value, thus supporting the log-scale transformation in fig 5.

Table 2 shows the number of transcripts and genes identified in Cufflinks analysis and FPKM normalization, both before and after filtering to remove values with FPKM=0.

From the differential expression statistics obtained above, the top ten DEGs (Table 3) with the smallest q values were filtered, and a histogram of their log2 fold change values was plotted as seen in (fig. 6). In figure 6, it can be seen that all of the genes have log2 fold change values greater than 0, indicating that the top ten DEGs are upregulated genes in the adult sample. Additionally, a histogram of all the significant DEGs was plotted (fig. 7), and unlike figure 6, this histogram shows log2 fold change values ranging approximately between -5 and +5, which indicates the presence of both up and down regulated genes.

Based on a traditional p value cut off of 0.05, there are 5188 differentially expressed genes; 2757 of these are upregulated, while 2431 are downregulated. However, when p value cut off of 0.01 is applied, there 4686 differentially expressed genes; 2532 of these are upregulated, while 2431 are downregulated. Overall, more genes seem to be upregulated than downregulated.

O’Meara et al. (2016) observed the differential expression of the representative genes in sarcomere, mitochondrial, and cell cycle. They then analyzed the biological pathways that are related to cardiac myocyte regeneration. According to O’Meara et al (2016), gene clusters that were repressed during differentiation were reactivated during cardiac myocyte regeneration. In general these pathways are related to RNA processing and chromatin modification. In our analysis, we used the database for annotation, visualization, and integrated discovery (DAVID), to obtain functional annotation clustering. Similar to the results obtained by O’Meara et al. (2016), our results showed that the significant upregulated enrichment terms function in mitochondria, nucleoside metabolic processes, sarcomere, and muscle cell differentiation (Table 4). We also had some common significant downregulated enrichment terms, which function in cell cycle, RNA processing, and transcription processes (Table 5).

In O’Meara et al. (2016), FPKM values of representative sarcomere, mitochondrial, and cell cycle genes (which are significantly differentially expressed during in vitro differentiation and in vivo maturation) were compared. For this project, 8 FPKM tables(P0_1, P0_2, P4_1, P4_2,

P7_1, P7_2, Ad_1, Ad_2) were used to make the same line plot as Figure 1D in O'Meara et al. to compare the FPKM values in vivo maturation. In order to replicate the line plots, FPKM columns from 8 original tables were first extracted and joined, and then significant genes listed in Figure 1D of O'Meara et al. (2016) were selected and plotted using a line plot.
The plots generated in figure 8 have the same direction and similar magnitude of effect as the ones reported in O'Meara et al, and shows that in adult mice, all signature gene expression involved in cell cycle declined (Fig. 8 C). In contrast, all signature gene expression involved in sarcomere and mitochondrial function showed an increase in expression (Fig. 8 A, Fig. 8 B).

We then plotted a heatmap (Fig. 9) to show the top 1000 differentially expressed genes between P0 and Adult groups, using FPKM values from 8 samples (P0_1, P0_2, P4_1, P4_2, P7_1, P7_2, Ad_1, Ad_2). To plot the heatmap, we first selected the top 1000 genes from the differential expression statistics data previously obtained using Cuffdiff. Then, we sorted up and down regulated genes by their log2 fold changes, and selected the top 500 up and down regulated genes respectively. Samples and genes seen in the heatmap are clustered based on hierarchical clustering (Fig. 9). Figure 9 indicates that the 1000 differentially expressed genes have significant differences between P0 and adult groups, and the difference becomes less obvious as postnatal days increase (Fig. 9).

Table 1. Key summary statistics for BAM file generated using Flagstat and RSeQC tools. Includes total number of reads, number of mapped, unique, multi-mapped, and unaligned reads, as well as percentages of total reads for each.

Statistic	**Number**	**Percent of total reads**
Total reads	49706999	100 %
Mapped reads	49706999	100 %
Unique reads	38489380	77.4 %
Multi-mapped reads	2899954	5.8 %
Unaligned reads	0	0 %

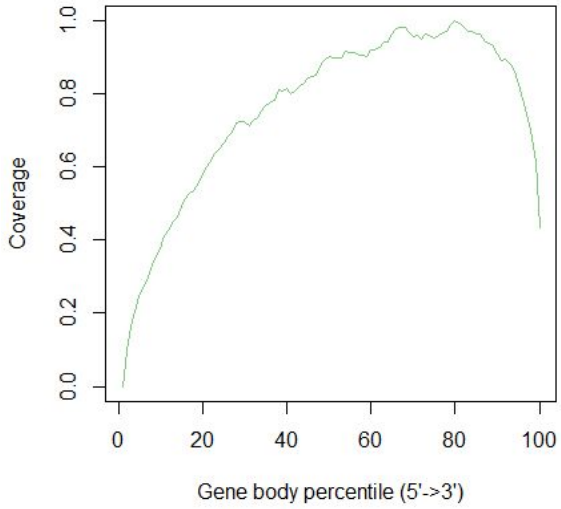


Figure 2. Graph showing Coverage vs. Gene Body percentile for transcript alignments from Tophat. Graph was produced using geneBody.py RSeQC tool. Most genes appear to have a good body coverage of ~80%, indicating most of the reads are aligned to 80% region from 5' to 3', which may indicate a 3' depletion. Overall, the graph showed that our data has high quality.

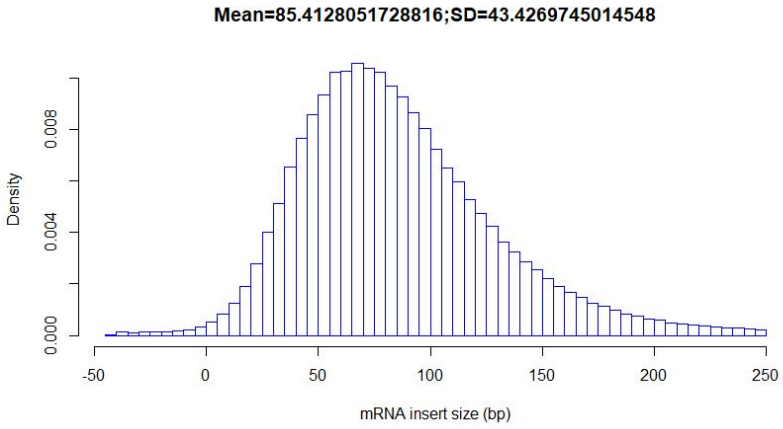
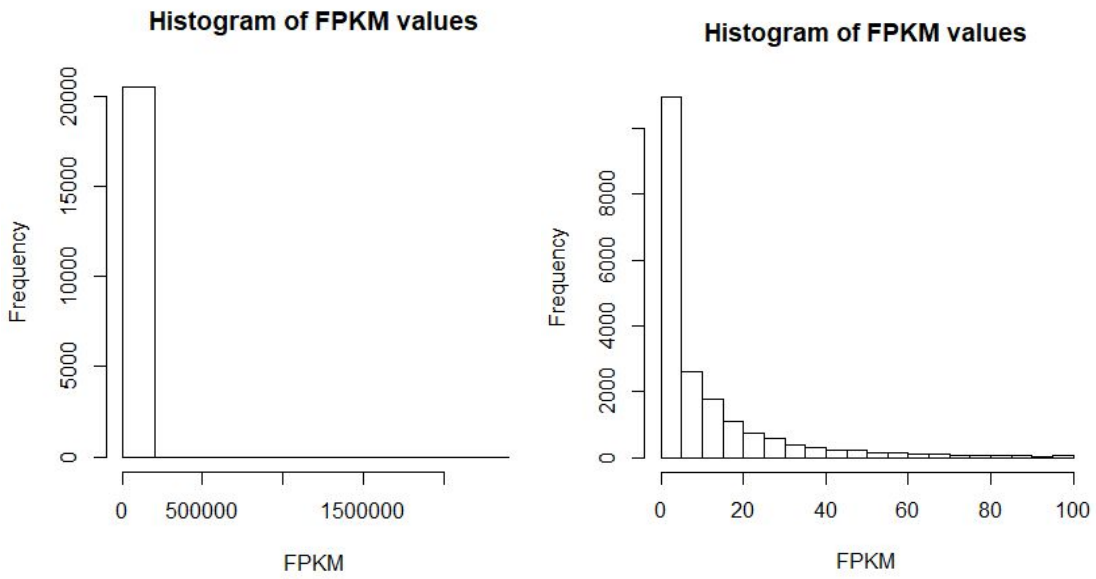


Figure 3. Graph showing density of mRNA fragments of each size for transcript alignments from Tophat. Graph was produced using inner_distance.py RSeQC tool. Most mRNA fragments had lengths of 50-100 bp, and there were few extremely short or long fragments.



(a) **(b)**

Figure 4. Histogram of FPKM < 100: Histogram showing FPKM values for genes with FPKM<100, for better visualization of patterns on a linear scale. All FPKM data points with values of 0 were removed for both 4(a) and (b), since these represent unaligned reads. Those greater than 100 were also removed for 4(b), simply to allow better data visualization.

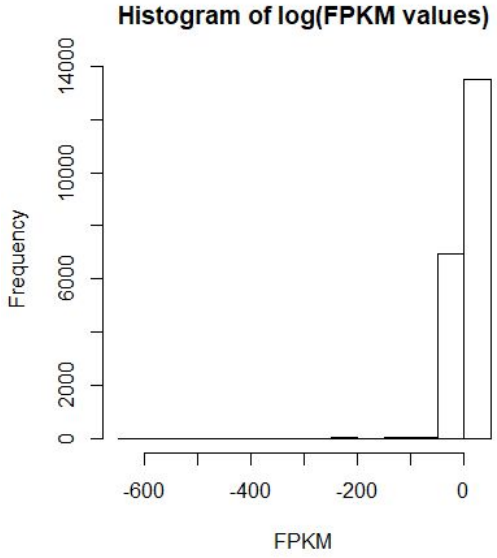


Figure 5. Histogram of log(FPKM) showing log of FPKM values for all genes, for better visualization of patterns on a log scale. All FPKM data points with values of 0 were removed since these represent unaligned reads, and then data was transformed using natural log.

Remaining FPKM values represent good quality reads. A large proportion of ln(FPKM) values seemed to be just above 0, indicating a large proportion of the FPKM values are above 1. However, a smaller but significant proportion of ln(FPKM) values also appear to be less than 1, indicating that a significant proportion of FPKM values are between 0 and 1.

Table 2. Number of genes before and after filtering identified in Cufflinks analysis and FPKM normalization. Filtering was performed to remove values with FPKM=0.

	No. of Transcripts (FPKM values)	**No. of gene_id**	**No. of gene_short_name**
Before	37469	37448	34081
After	20487	20478	20302

Table 3. Top 10 DEGs with the smallest q values. All genes have log2 fold change value > 0, thus making the genes upregulated in the adult sample.

	Gene	**log2.fold change**	**p value**	**q value**
1	Rb1cc1	1.389250	5e-05	0.000318974
2	Pcmtd1	1.174640	5e-05	0.000318974
3	Adhfe1	0.996765	5e-05	0.000318974
4	Tmem70	1.216660	5e-05	0.000318974
5	Gsta3	4.100950	5e-05	0.000318974

6	Lmbrd1	0.990848	5e-05	0.000318974
7	Dst	1.517230	5e-05	0.000318974
8	Plekhb2	1.435380	5e-05	0.000318974
9	Mrpl30	1.246490	5e-05	0.000318974
10	Tmem182	1.240250	5e-05	0.000318974

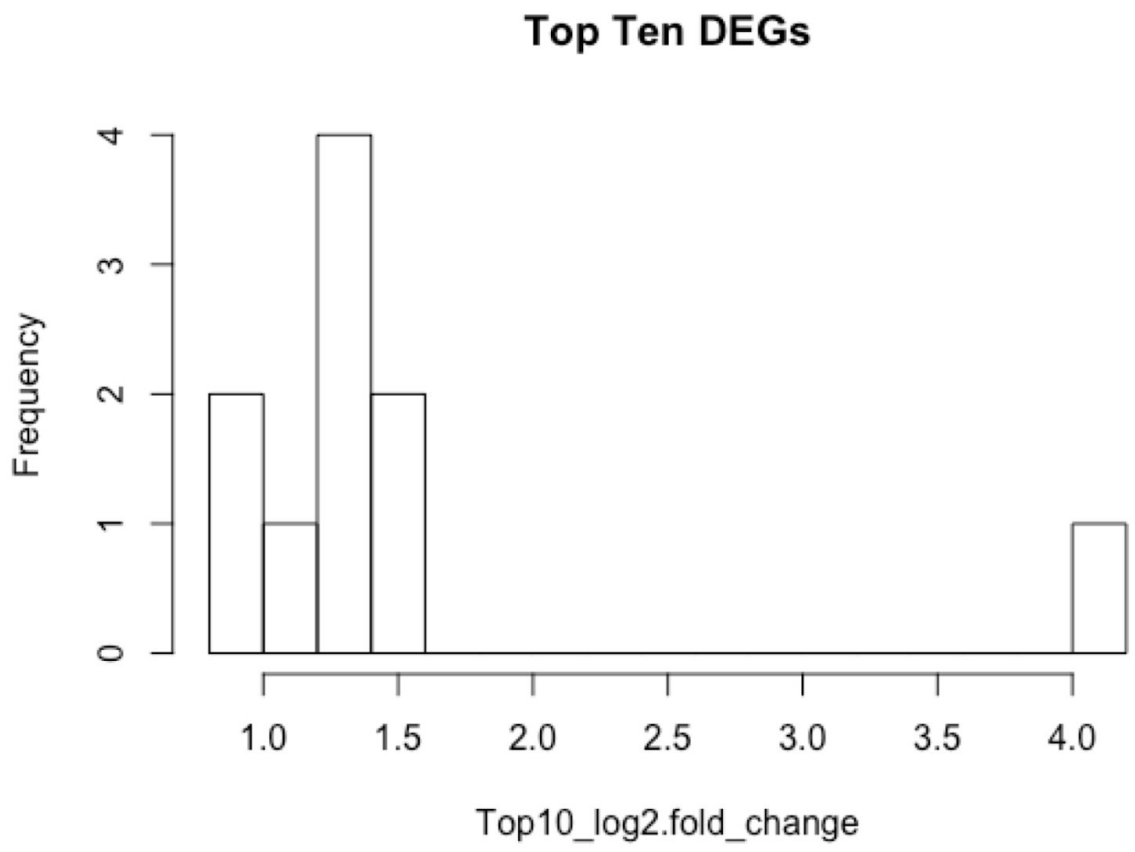


Figure 6. Histogram of top 10 DEGs. A graphical representation of table 3, where all the genes are upregulated in adult samples(log2 fold change > 0)

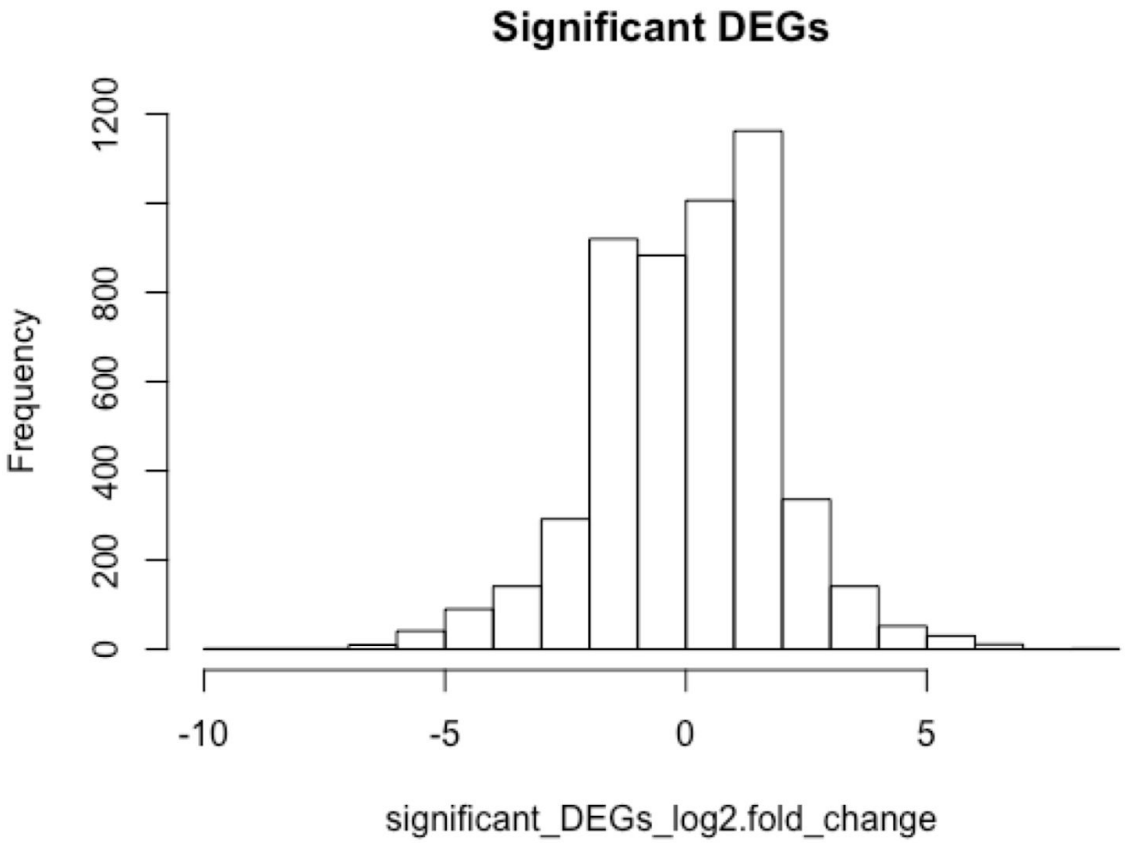
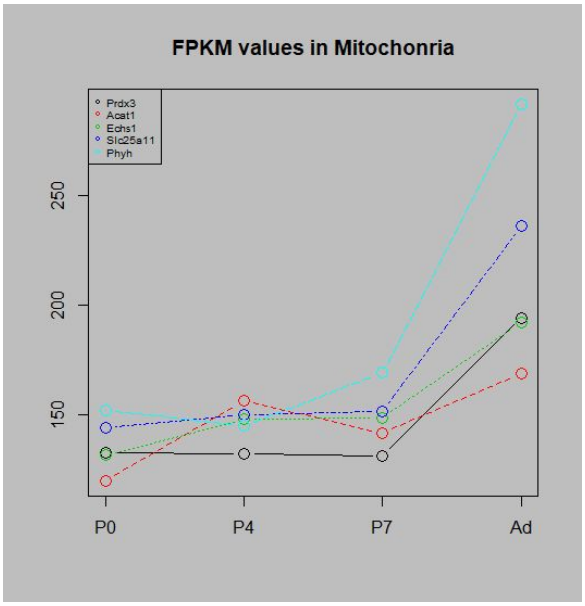
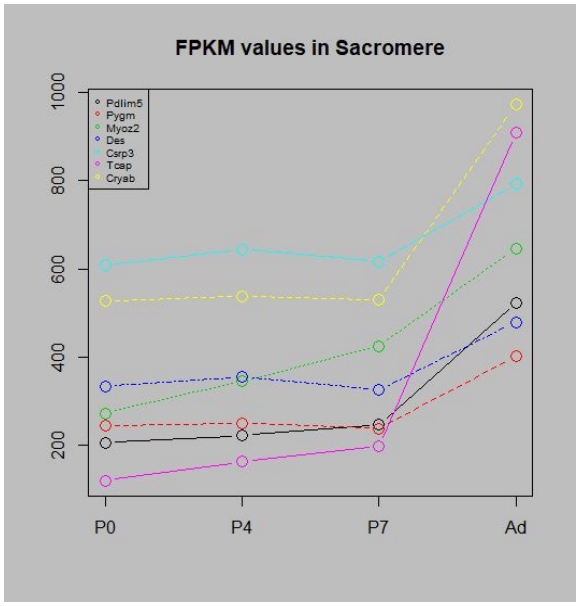


Figure 7. Histogram of all significant DEGs. Both up and down regulated genes are represented in distribution (log2 fold change values approximately ranged between -5 to +5)

A.



B.



C.

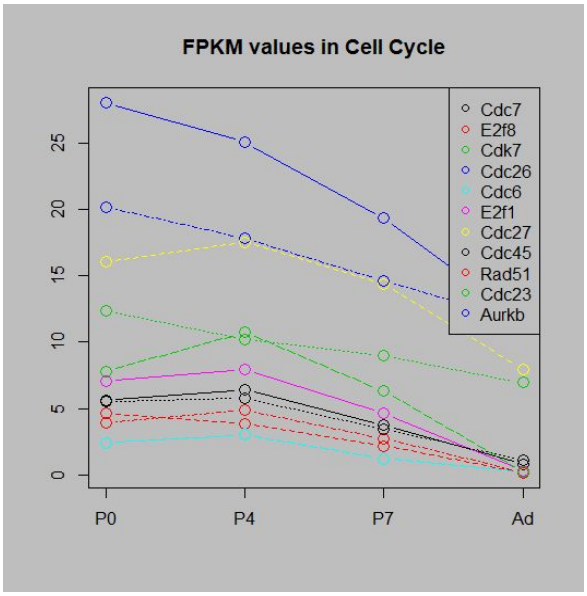


Figure. 8 Genes in sarcomere and mitochondria are upregulated, and genes in cell cycle are downregulated. FPKM values of representative sarcomere, mitochondrial, and cell cycle genes
These line plots capture FPKM values of representative sarcomere, mitochondrial, and cell cycle genes. Genes in sarcomere and mitochondria are upregulated, and genes in cell cycle are downregulated.

Table 4. Top 7 downregulated clusters have similar function groups with Caitlin's paper's result, and most of the pathways are common. A table includes enrichment score, related function of top clusters in downregulated annotation clusters. Percentage of common pathways indicates the percentage of common pathways between our upregulated cluster table and the table in Caitlin's paper.

	Function	Enrichment Score	Percentage of common pathways
Cluster 1	Cell Cycle	27.47118029	0.6
Cluster 2	Chromosome	21.62207915	1
Cluster 3	Transcription	20.96743394	0.342105
Cluster 4	Chromosome organization	17.87386595	1
Cluster 5	DNA repair	15.44591522	0.75
Cluster 6	Cytoskeleton organization	15.38789066	0.666667
Cluster 7	Regulation of transcription	15.31603743	0.434783

Table 5. Top 9 upregulated clusters have similar function groups with Caitlin's paper's result, and most of the pathways are common. A table includes enrichment score, related function of top clusters in upregulated annotation clusters. Percentage of common pathways indicates the percentage of common pathways between our upregulated cluster table and the table in Caitlin's paper.

	Function	Enrichment Score	Percentage of common pathways
Cluster 1	Mitochondria	52.93998	0.8
Cluster 2	Nucleoside metabolic process	23.7162	0.424242
Cluster 3	Mitochondria	22.5421	0.533333
Cluster 4	Carboxylic acid metabolic process	21.50029	0.125
Cluster 5	Extracellular organelle	13.53755	0.333333
Cluster 6	Sarcomere	10.95992	1
Cluster 7	Fatty acid catabolic process	9.558286	0.727273
Cluster 8	Response to lipid	9.389292	0
Cluster 9	Heart muscle contraction	8.057296	0.470588

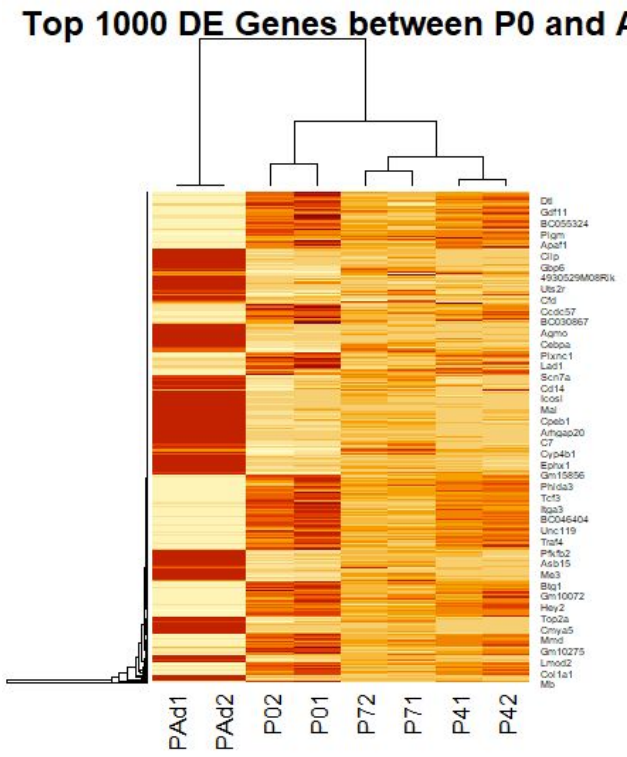


Figure. 9 A heatmap shows the 1000 DE genes expression difference from Ad are smaller from P0 to P7. Heatmap that is created by 1000 differentially expressed genes(500 upregulated, and 500 downregulated) between P0 and Ad group. There are four different groups, P0, P4, P7 and Ad. For each group, there are two replicates with suffixes of 1 and 2. It is clear that the pairs of replicates have similar expression levels.

Discussion

Overall, our replicated line plots agree with results presented in Figure 1D in O'Meara et al(2016). All signature genes in sarcomere and mitochondria generally showed a significant increase between postnatal day 7 (P7) and adult(Ad) groups. According to O'Meara et al (2016), cardiac myocytes in the regenerating neonatal mouse heart tend to have less distinct sarcomere structures, and this is depicted by the low gene expression present in P0, P4 and P7 mice. As mentioned in O'Meara et al the big difference in gene expression between P7 and Ad mice is due to aging cardiac myocytes being unable to reenter the cell cycle to regenerate their damaged cells (Fig. 8). Future research can focus on longitudinal study after P7, and a more detailed gene expression profile can be discovered. Furthermore, the signature genes that

function in cell cycle processes gradually declined from P0 to P7, making this result consistent with the initial hypothesis of O'Meara et al., that mammalian cardiac myocytes exit the cell cycle shortly after birth. A difference we noted from O'Meara et al. in our analysis is the FPKM values of some of the genes. This difference could be due to different gene annotation files. Additionally, cufflink counts how reads map to genomic regions defined by an annotation, so different annotation files can result in different FPKM values.

In gene functional annotation clustering, RNA processing and chromatin modification being two of the most significant clusters (Table 4) suggested that during cardiac myocyte regeneration, more RNA related pathways are activated, and more genes are expressed. Transcriptional alteration of a specific subset of genes that regulate sarcomere organization, RNA processing, and cell cycle progression is critical for remodeling cardiac myocytes differentiation. However, there are some pathways that are different between our analysis and the paper's result, since we used a summarized version of Gene Ontology(GOTERM_BP_FAT, GOTERM_MF_FAT, and GOTERM_CC_FAT).

In figure 9's heatmap, we observed a trend in the 1000 DEGs that have significant differences between P0 and Ad groups, and the differences between them gets smaller after 7 Days post resection(Dpr). These genes may be critical for cardiac myocyte repair in response to injury and promote adult mammalian cardiac regeneration. Similar to Figure 2A in O'Meara et al(2016), our plots also showed the transition of gene cluster expressions.

Conclusion

Overall, our project successfully replicated part of O'Meara et al. (2016). We observed that RNA processing and chromatin modulation are reactivated in neonatal mice, and the loss of sarcomere structure during the repair from line plots, GO term analysis and hierarchical clustering. This evidence supports that cardiac regeneration is a transcriptional reversion of the differentiation process.

One problem we encountered in the biologist part is regarding fastq files. In the first try, we used -I/–readids flag in fastq-dump function. This option would append read id, 1 and the other 2, after spot id, so we thought it was necessary for pair-end reads. However, the read ids option might break downstream analysis like cuffdiff which results in incorrect mapping, and produce incorrect FPKM values. Thereby, when we used the cuffdiff results, it gave us fewer significant genes(~2000) than the correct number of significant genes(~5000). Additionally, the heatmap would have shown inconsistency between P0_1 and P0_2. We solved the problem by comparing each step in our analysis with other groups, and tried to find any discrepancy.

Finally, we spotted the difference in fastq-dump function at the very beginning and re-ran the codes again.

References

Bicknell KA, Coxon CH, Brooks G. Can the cardiac myocyte cell cycle be reprogrammed? J Mol Cell Cardiol. 2007; 42:706–721. [PubMed: 17362983]

EMBL-EBI. (2020). FPKM. Retrieved March 18, 2020, from https://www.ebi.ac.uk/training/online/glossary/fpkm.

John Hopkins University. (2016). TopHat. Retrieved March 18, 2020, from https://ccb.jhu.edu/software/tophat/index.shtml.

O'Meara et al. Transcriptional Reversion of Cardiac Myocyte Fate During Mammalian Cardiac Regeneration. Circ Res. Feb 2015. [PMID: 25477501]

Trapnell, C. (2017). Cufflinks. Retrieved March 18, 2020, from http://cole-trapnell-lab.github.io/cufflinks/.

Trapnell C, Pachter L, Salzberg SL. Tophat: Discovering splice junctions with rna-seq. Bioinformatics. 2009; 25:1105–1111. [PubMed: 19289445]