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

Group: Saxophone
TA: Jing Zhang

Yilin Yang: Data Curator
Daniel Goldstein: Programmer
Jason Rose: Analyst
Sooyoun Lee: Biologist

INTRODUCTION

Cardiac myocytes (CMs) in the mammalian heart are responsible for contractile forces [1]. Over time, hypertrophy occurs causing the inability to meet contractile force demands. This leads to heart failure as the CM has limited ability to divide in adults eventually leading to death [2]. It has been found that neonatal mice have the ability to regenerate their heart tissue after resection of the left ventral apex [3]. Using high throughput sequenced mRNA abundance data generated from O'Meara et al. 2015 [4], this study seeks to understand the changes in transcriptional phenotype of neonatal and adult mice. While O'Meara et al. look at several stages of development during injury-induced regeneration, this study focuses on the day zero neonatal and 8-10 week adult mice. mRNA abundance data during these stages is used to find the differentially expressed genes (DEGs).

The purpose of this study was to determine the potential regulators for the mammalian cardiac regeneration for neonatal mice and determine if the myocytes reverse the transcriptional phenotype to a less differentiated state during the regeneration process.

DATA

Large quantities of sequences are generated by high throughput sequencers, but prior to analyzing these sequences, it is necessary to perform quality control (QC) checks to ensure that there are no biases in the data that could impact the success of the results. Potential problems and biases that arise due to defects in the sequencer or starting materials are identified by using FASTQC. FASTQC was run on the command line to extract quality measures.

Raw data collected from the whole heart ventricle cells of (CD-1) neonatal mice at postnatal day 0, 4, 7(P0, P4, P7) and from 8 to 10 week old male CD-1 mice. The hearts were dissected to obtain the ventricles for RNAseq. The raw data of explanted adult mouse cardiac myocytes were collected after 0, 24, 48, and 72 hours and processed under the previously described protocol. No samples were eliminated due to contamination. All the samples used in this study, except for one, were downloaded and processed prior to the start of the project. The one remaining sample GSM1570702 (vP0_1) was downloaded from NCBI GEO Series GSE64403. The sample was stored in a file named SRR1727914.sra with a size of 1.1Gb, and the file is in SRA (short read archive) format.

METHODS

Data extraction:

Downloaded file SRR1727914.sra onto the SCC server and renamed before converting to FASTQ file:

```
wget https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1570702

mv SRR1727914.1 P0_1.sra
```

Made SRA tools available and extracted the SRA formatted file to FASTQ file using qsub:

module load sratoolkit

run_extract.qsub file content:

fastq-dump --split-files -O /projectnb/bf528/users/saxophone/project2/samples

qsub run_extract.qsub

Quality Control:

After the SRA file was extracted into the two FASTQ files, they were processed, and the quality measures were obtained using the FASTQC package available on SCC:

module load fastqc

fastqc -o

/projectnb/bf528/users/saxophone/project2/samples/P0_1_1.fastq

/projectnb/bf528/users/saxophone/project2/samples/P0_1_2.fastq

Alignment and Quality Analysis

Paired-end reads for the P0_1 sample were aligned to the mm9 mouse reference genome and analyzed using the Tuxedo package (e.g., TopHat, Cufflinks, and Cuffdiff) [5]. TopHat utilizes Bowtie2 and samtools as dependencies to map the short sequences to the reference genome and store the aligned sequence. Mouse reference files and Bowtie2 indexes were provided by the developers of TopHat (http://ccb.jhu.edu/software/tophat/igenomes.shtml). TopHat was run as a batch job for three hours on the SCC using the following arguments [6]: (1) mean inner distance (-r) of 200, (2) gene annotations file location (-G) at (/projectnb/bf528/project_2/reference/mm9.gtf), (3) minimum length of read segments (--segment-length) of 20, (4) number of mismatches allowed per segment (--segment-mismatches), (5) only look for reads across junctions in the annotation file (--no-novel-juncs), (6) a number of cores dedicated to run TopHat (-p) equal to 16, and (7) output file directory (P0_1_tophat). Aligned reads were assigned FLAG attributes and quantified and annotated with the flagstat utility of samtools.

Quality control analysis of the alignment output file (accepted_hits.bam) was performed using three RseQC utilities, which assessed different quality control metrics. Firstly, the BAM_stat.py module determined the number of uniquely mapped reads from mapping quality. Secondly, the inner_distance.py module calculated inner distance between paired RNA reads. Lastly, geneBody_coverage.py measured the number of reads covering each nucleotide to

determine if coverage was uniform and if there were any 5' or 3' biases in a sample. GeneBody_coverage was run as a batch job on the SCC for about two hours. These RseQC utilities required the input file (-i) saved in BAM format, the indexed BAM file, the reference file (-r) saved in BED format, and the output directory (P0_1_genebody).

Gene Expression Analysis

Using the Cufflinks tool, read count was mapped to genomic regions using the gene annotation file (/project/bf528/project_2/reference/annot/mm9.gtf) with the following arguments [7,8]: (1) multi-read-correct mapping (-u), which more accurately weighs reads to multiple locations, (2) gene annotation file location (-G) as mentioned previously, (3) number of cores needed to run Cufflinks (-p) equal to 16, (4) reference genome file in fasta format for bias detection and correction (-b), (5) only counts reads compatible with the reference (--compatible-hits-norm), and (6) output file directory (P0_1_cufflinks). Cufflinks was run as a batch job on the SCC for about two hours. The cufflinks output (genes.fpkm_tracking) included gene alignments in the number of fragments per kilobase of exon per million reads mapped (FPKM) for all genes. These alignments were read into R and a histogram of FPKM values per gene was generated to determine which genes were expressed in the P0_1 sample. Filtering of FPKM value was applied at a threshold FPKM of 250, which provided the top 34 genes that were expressed in the P0_1 sample.

Differential gene expression was assessed using the Cuffdiff tool from the Cufflinks suite. Gene expression of the P0_1 sample was compared with samples P0_2, Ad_1, and Ad_2. The Cuffdiff script was run on the SCC as a batch job for about two hours with the following arguments [9]: (1) P0_1, P0_2, Ad_1, Ad_2 labels for each sample (-L), (2) number of cores dedicated to Cuffdiff (-p) of 16, (3) multi-read-correct mapping (-u), (4) reference genome for bias detection and correction (-b), and (5) the output directory (cuffdiff_out).

Adult (AD) versus Neonatal 0 day P(0) Significance and Annotation Clustering:

After identifying DEGs, R statistical software (v4.0.4) [10] was used to analyze those of significance in the Cuffdiff output. All genes were sorted by q-value to determine lowest probability of false positives. Sorting was run several times to confirm consistent ordering to assure reproducibility and consistency. A histogram of the log2 fold change between P(0) and AD was created for all DEGs and again for only those that are significant. The significant genes were sorted into up and down-regulated bins based on log2 fold change.

For discovery of gene functionality in differentiation, DAVID v6.8 [11] functional annotation clustering was used. Analysis were conducted for both up and down-regulated significant genes using the Official Gene Symbol identifier for species *Mus musculus*.

RESULTS

The extraction of the SRA formatted file resulted in two FASTQ files in the paired-end sequence: forward and reverse reads. The format of the files was inspected by accessing the first few lines of both files using the head function and ensuring that the texts in the header files matched exactly. In addition, the outputs from the FASTQ tool included several HTML and image files that were downloaded and processed as well.

The FASTQC report of sample GSM1570702 showed the per-base sequence quality is acceptable. However, the plot of Per Base Sequence content illustrated a non-uniform distribution of bases with %A not equal to %T, and %G not equal to %C for the first 10 to 12 nucleotides(Figure 1).

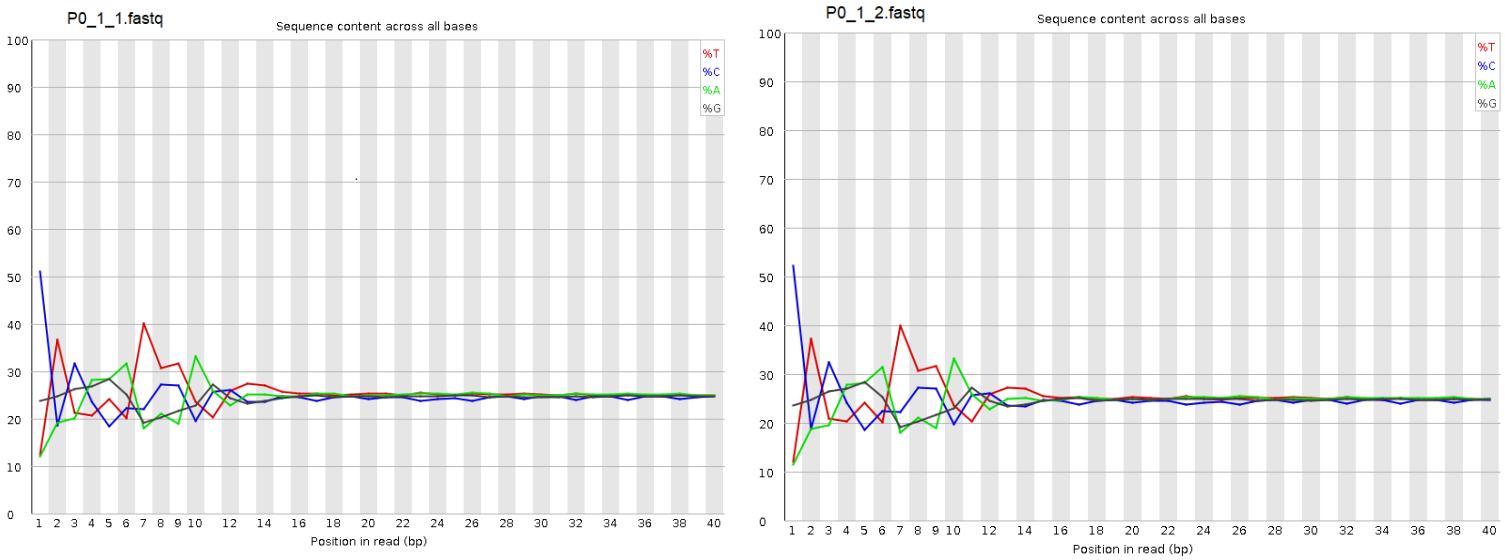


Figure 1. "Per base Sequence Content" FASTQ output for the two fastq files

Figure 2 demonstrates the sequence duplication levels, which also showed warnings in the report. In Figure 2, it is apparent that many reads are duplicated from 10 to 500 times, with percent of sequences remaining being approximately 52%. Overall, the levels are relatively low, suggesting a good diversity in the library, so no correction is needed.

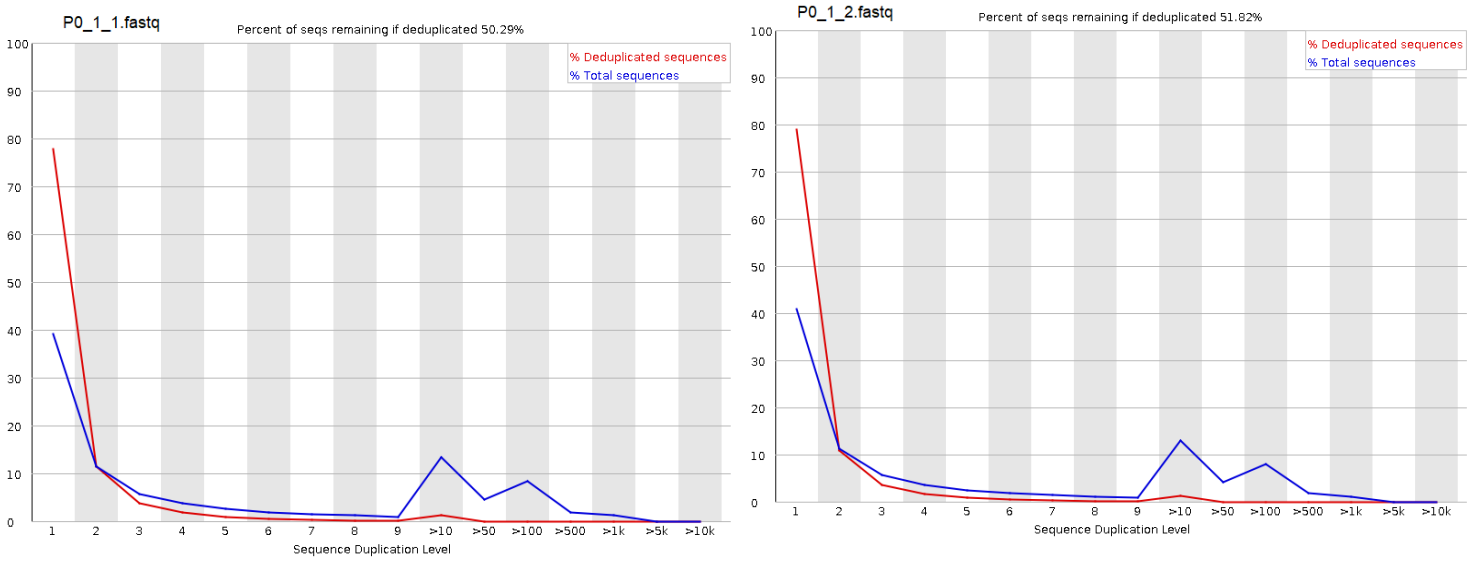


Figure 2. "Sequence Duplication Levels" FastQ output for the two fastq files

Quantification analysis from flagstat determined counts of reads with several attributes, including total number of reads as well as number of mapped and unaligned reads (Figure 3). The total number of reads that passed quality control analysis was 49,706,999. Of this total count, 100% of the reads were mapped to the reference genome and therefore none of the reads were unaligned. The number of unique reads was 77.4% of the total number of reads, or 38,489,380 unique reads (Figure 4).

```
49706999 + 0 in total (QC-passed reads + QC-failed reads)
8317665 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
49706999 + 0 mapped (100.00% : N/A)
41389334 + 0 paired in sequencing
20878784 + 0 read1
20510550 + 0 read2
29422646 + 0 properly paired (71.09% : N/A)
39936472 + 0 with itself and mate mapped
1452862 + 0 singletons (3.51% : N/A)
1387382 + 0 with mate mapped to a different chr
704916 + 0 with mate mapped to a different chr (mapQ>=5)
```

Figure 3. Flagstat output with counts for total, mapped, and unaligned reads, as well as other FLAG attributes.

```
Total records:                              49706999

QC failed:                                  0
Optical/PCR duplicate:                      0
Non primary hits                            8317665
Unmapped reads:                             0
mapq < mapq_cut (non-unique):               2899954

mapq >= mapq_cut (unique):                  38489380
Read-1:                                     19409941
Read-2:                                     19079439
Reads map to '+':                           19236824
Reads map to '-':                           19252556
Non-splice reads:                           33099839
Splice reads:                               5389541
Reads mapped in proper pairs:               27972916
Proper-paired reads map to different chrom:4
```

Figure 4. BAM stat output with counts for unmapped reads and unique reads.

From the size distribution of inner distance from Figure 5, the mean inner distance measured 85 bp long with a standard deviation of 43 bp. Therefore roughly 95%, or 2 standard deviations, of the paired-end reads in sample P0_1 fall within the 200 bp inner distance specified in the TopHat arguments. Genebody coverage analysis showed a 3' bias in the P0_1 sample, which indicates that there was some RNA degradation present; however, the distribution of genebody percentile is relatively uniform.

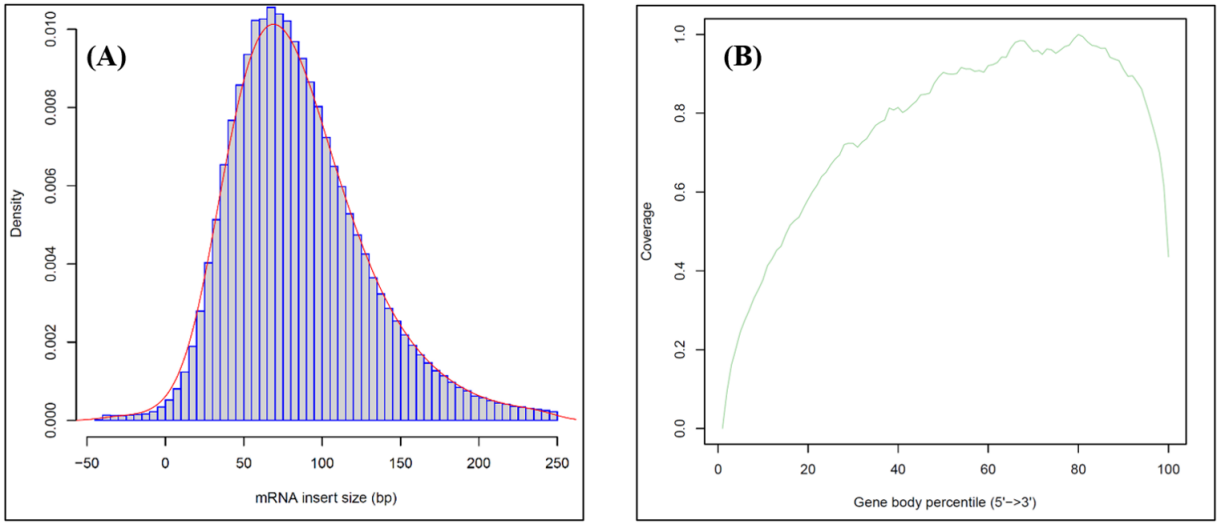


Figure 5. RSeQC Analysis Plots. (A) Size distribution of the inner distance between paired-end reads. Mean inner distance = 85 bp ± 43 bp. (B) Genebody coverage distribution of the location of reads from 5' to 3' of reference.

Gene expression of the P0_1 sample using FPKM normalization found 34 genes that were expressed out of 37,469 total genes after the FPKM threshold of 250 counts was applied (Figure 6). From Table 1, the Mir5105 gene had an FPKM value of 170,722, which is about an order of magnitude greater expression than the subsequent genes: mt-Ts1 with an FPKM of 24,679.2, Gm12563 with an FPKM of 15,836.5, and mt-Tc with an FPKM of 12,505.3.

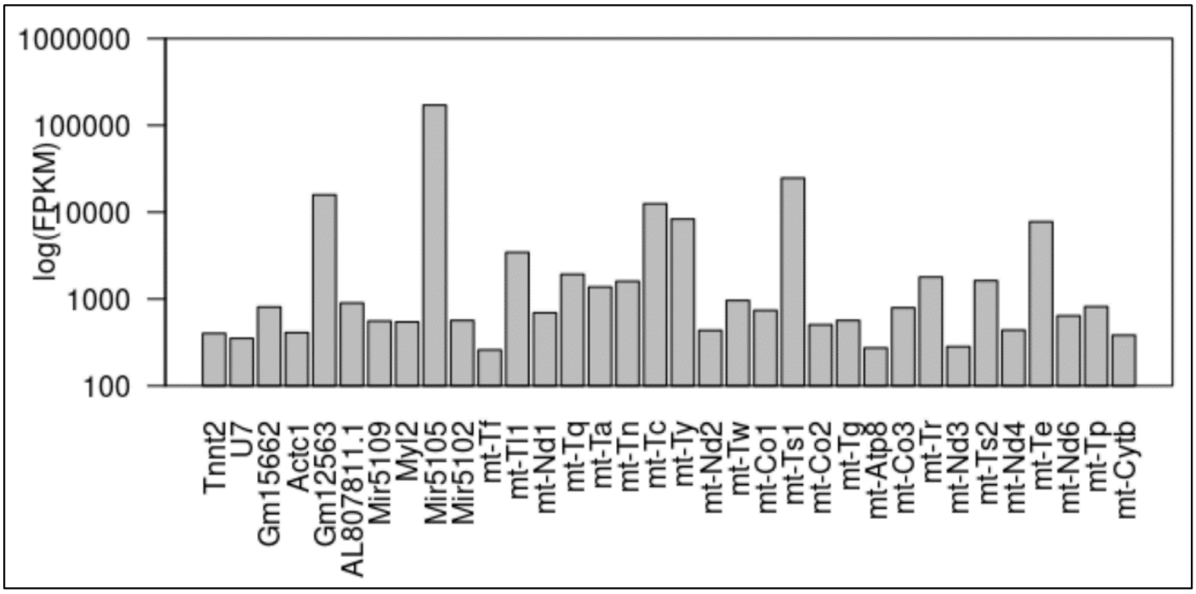


Figure 6. Histogram of the most expressed genes in the P0_1 sample, plotted as log (FPKM) vs. gene name. The ten most expressed genes are provided in Table 1.

Top Ten Expressed Genes in P0_1		
Gene ID	**Gene Name**	**FPKM**
ENSMUSG00000093077	Mir5105	170722
ENSMUSG00000064352	mt-Ts1	24679.2
ENSMUSG00000083696	Gm12563	15836.5
ENSMUSG00000064349	mt-Tc	12505.3
ENSMUSG00000064350	mt-Ty	8326.3
ENSMUSG00000064369	mt-Te	7756.2
ENSMUSG00000064340	mt-Tl1	3437.3
ENSMUSG00000064343	mt-Tq	1925.8
ENSMUSG00000064361	mt-Tr	1784.7
ENSMUSG00000064365	mt-Ts2	1622.8

Table 1. The ten most expressed genes in sample P0_1.

A total of 36329 genes were found to be differentially expressed between P(0) and AD. After trimming (Figure 7A) 2139 significant DEGs remain (Figure 7B). Of those, 1084 are up-regulated with 1055 being down-regulated. The most significant are tied among 666 genes by q-value with the top ten (Table 2) being: PLEKHB2, MRPL30, COQ10B, AOX1, NDUFB3, SP100, CXCR7, LRRFIP1, RAMP1, and GPC1.

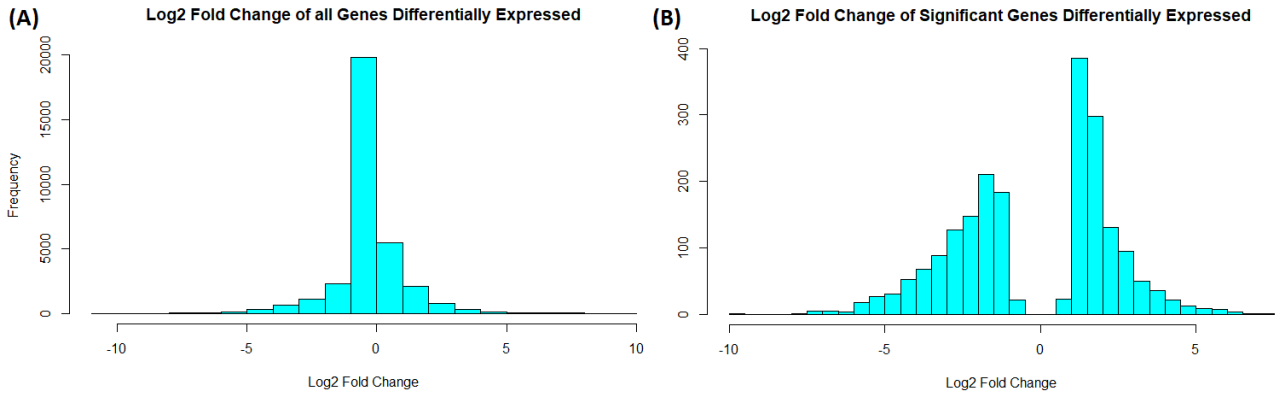
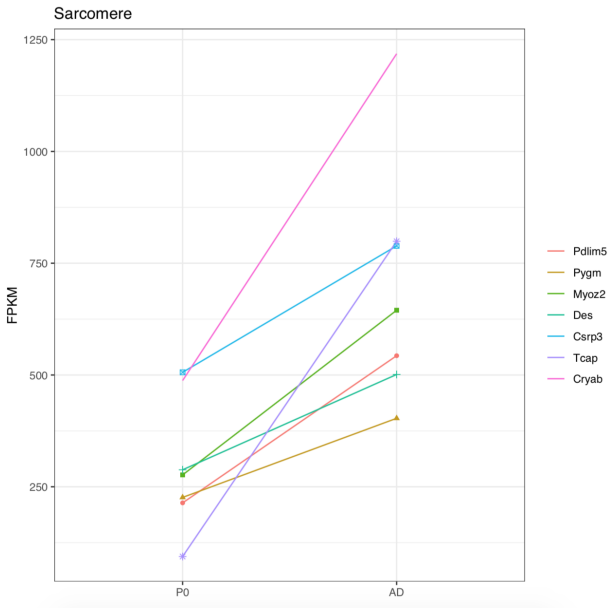


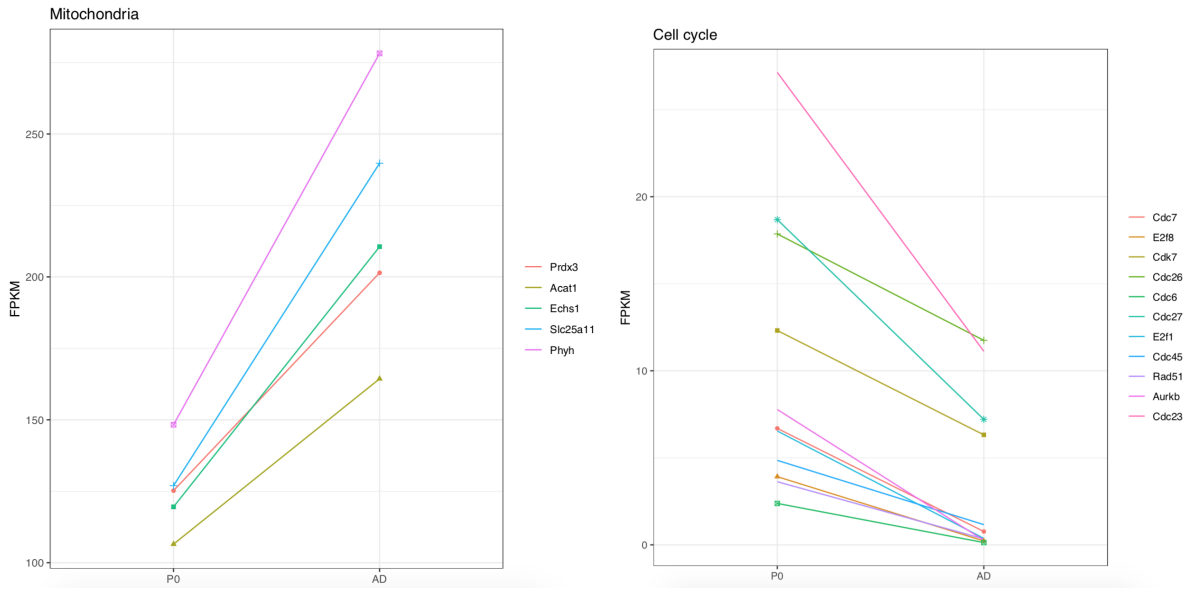
Figure 7. *Log2 fold change of* **(A)** all expressed genes and **(B)** only significantly expressed genes during differentiation.

Top Ten Differentially Expressed Genes					
Gene	FPKM P(0)	FPKM AD	Log2 Fold Change	P - Value	Q - Value
PLEKHB2	22.5679	73.5683	1.70481	5.00E-05	0.001069
MRPL30	46.4547	133.038	1.51794	5.00E-05	0.001069
COQ10B	11.0583	53.3	2.26901	5.00E-05	0.001069
AOX1	1.18858	7.09136	2.57682	5.00E-05	0.001069
NDUFB3	100.609	265.235	1.39851	5.00E-05	0.001069
SP100	2.13489	100.869	5.56218	5.00E-05	0.001069
CXCR7	4.95844	32.2753	2.70247	5.00E-05	0.001069
LRRFIP1	118.997	24.6402	-2.27184	5.00E-05	0.001069
RAMP1	13.2076	0.691287	-4.25594	5.00E-05	0.001069
GPC1	51.2062	185.329	1.8557	5.00E-05	0.001069

Table 2. *The top ten differentially expressed genes in P(0) versus AD.* Of 36329 genes, 666 tied for lowest q-value and p-values. Provided is R's interpretation of the most significant among them.

(A)





(B) **(C)**

Figure 8*. The change of expression among the highlighted genes.* FPKM values of representative (A) sarcomere, (B) mitochondria, and (C) cell cycle genes that are differentially expressed during the vivo maturation

Figure 8 compared the trends with the reference paper by capturing the FPKM values of different genes. We have focused on the highlighted genes in figure 1D from the original study and as a result, figure 8 appears to be identical to figure 1D. In this figure, the genes are divided into three groups such as sarcomere (Figure 8A), mitochondria (Figure 8B), and cell cycle genes (Figure 8C). The sarcomere and the mitochondrial genes are up-regulated while the cell cycle is down-regulated which is the same result with the reference paper. To reproduce the line plot, the tidyverse was installed since it is the package that helps to perform and interact with the data [12]. The reshape2 function was also installed to make it easy to transform data between wide and long formats [13].

Enrichment Terms			
Up-Regulated		Down-Regulated	
Enrichment Term	**Cluster Score**	**Enrichment Term**	**Cluster Score**
Mitochondrion*	21.34	Cell Cycle*	11.85
Metabolic Energy Generation	15.28	Extracellular Matrix	9.68
Organic Acid Metabolism*	14.57	Cell Proliferation	9.16
Extracellular Organelle	10.78	Regulation of Cell Organization	8.62
Contractile Fiber	6.99	Organogenesis	8.07
Small Molecule Catabolism	6.19	Cardiovascular System Development	7.59

Table 3. *The result of the most common up and down-regulated genes during the differentiation when compared with the results that are obtained from the DAVID analysis with their associated cluster scores.* *Upregulated and downregulated genes overlap with the results reported in the reference paper [4].

Table 3 depicts the highest enrichment score for both up-regulated and down-regulated genes. The result of the DAVID analysis that is obtained in 6.7 was compared with the data from the reference paper. The overlap enrichment genes are classified with the asterisk mark.

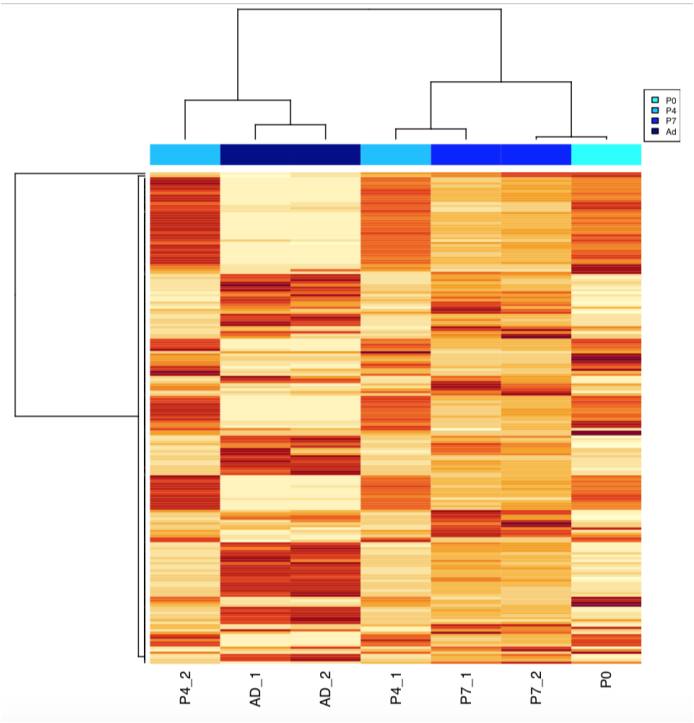


Figure 9. *Heatmap of all expressed genes over different development stages.* The dark red color represents a higher expression while the light yellow color represents a lower expression.

The FPKM matrix of different samples by concatenating the FPKM columns from each of the tracking tables from 7.1 together into a single data frame. The matrix of the top most significant is differentially expressed between the P0 and the Adults. There are some patterns represented in this heatmap which will be discussed further in the discussion section. In this figure, a total of 7 samples (P0, P4_1, P4_2, P7_1, P7_2, AD_1, and AD_2) from the FPKM values were used. All the samples in different developmental stages are well paired in the heatmap which is very similar to O'Meara et al's paper in Figure 1D [4].

DISCUSSION

A total of 21,577,562 sequences with a length of 40 were processed from the FASTQC run according to the metrics read quality. None of these sequences were overrepresented nor flagged for poor quality, and both had an overall GC content of 49. However, both files failed in "Per Base Sequence Content" (Figure 1) and showed warnings for "Sequence Duplication Levels" (Figure 2) while passing the rest of the quality matrices.

"Per Base Sequence Content" illustrates the relative amount of A, T, C, and G at each position across reads in the genome. This failed in both files because there is >20% difference between A and T, and G and C towards the beginning read positions as shown in Figure 1. A difference in proportion between G and C is seen at the first position, while a difference between A and T is at the 7^{th} position. In both files, the lines start to parallel starting at the 14^{th} position.

The report also showed warnings for the "Sequence Duplication Levels" in both FASTQ files. There is a fluctuation of duplication levels across the reads, and this may have resulted from the PCR duplication in which library fragments have been over-represented due to bias in PCR enrichment or abundant transcripts in an RNA-seq library.

The RNA-Seq alignment and quality analysis of P0_1 showed that all of the reads passed quality control and were mapped to the reference genome. Furthermore, there were 0 duplicate reads and 77.4% of the reads were unique. The inner distance of the paired-end reads followed a normal distribution that was skewed to the right. In addition, there was noticeable RNA degradation in the sample, as shown in Figure 5, where there is a 3' bias in the genebody coverage distribution.

In gene expression analysis, 34 genes were deemed to be highly expressed in the P0_1 sample after FPKM filtering. Among these expressed genes, Mir5105 (ENSMUSG00000093077) was expressed significantly more than the subsequent genes by an order of magnitude. However, Mir5105 as well as the other highly expressed genes in P0_1 were not differentially expressed when compared to Ad samples. Therefore these genes are not believed to be relevant to cellular regeneration.

In our analysis in identifying DEGs between P0 and AD, we found 36329 genes. Prior to removing non-significant genes, the frequency of genes is greatest with the smallest log2 fold change (Figure 3A) with a large number of genes experiencing little differentiation. After removal of the non-significant genes we are left with DEGs that are more well expressed (Figure

3B) during injury-induced regeneration of CMs. Of those 36329, only 1084 up-regulated 1055 down-regulated DEGs are significant. With the same data, the reference paper found 1482 up-regulated and 4341 down-regulated DEGs.

Of the significant DEGs, PLEKHB2 (Pleckstrin homology domain-containing family B member 2) was found to be the most significant among the top ten genes (Table 1). The function of PLEKHB2 is in retrograde transport of recycling endosomes that can target the heart [14] and is known to be critical in cellular homeostasis [15]. As the study induced injury to the left ventricle it is not surprising to see a gene involved in cell homeostasis be significantly up-regulated.

Figure 1D in the reference paper [4] was reproduced by comparing the expression level among the genes between P0 and the adult samples from our data. There are commonalities between the genes, for example, sarcomere and the mitochondria groups have up-regulated genes while the genes in the cell cycles are down-regulated. For the sarcomere genes, a distinct increase with the expression in the adults is represented. This increase in expression reflects the sarcomere assembly and organization during cardiomyocyte maturation. A similar trend was observed with the mitochondria gene however compared to the FPKM values of the sarcomere, the mitochondria had smaller FPKM values. The FPKM values of the sarcomere is between 200 to 1,200 while the mitochondria range from 100 to 300. In the cell cycle gene, a decrease in gene expression was observed due to the cell cycle exit is a hallmark of mature cardiac myocytes, and a failure to re-enter the cell cycle is likely to contribute to the lack of cardiac regeneration in adults mammals [4].

In table 2, the top up-regulated genes and the down-regulated genes gathered from the DAVID were different from the reference paper [4]. For the up-regulated genes, the highest enrichment score is mitochondrion with a score of 21.34. In the reference paper, the mitochondrion was also the highest enrichment gene with a score of 14.35 [4]. For the down-regulated genes, only one gene is overlapping which is the cell cycle. Genes whose expression decreased in both differentiation datasets function primarily in the cell cycle [4] which is shown as the most common down-regulated gene in our data. The discrepancy in score and clusters could be attributed to versions of DAVID used. The version we used (v6.8) is newer than what was used in the reference paper which the most recent version was v6.7. In the updated version the database was rebuilt along with added annotations categories that could have contributed to a difference in results. This could also be a result of the difference in the number of significant up and down-regulated DEGs found as our data. In comparison with the reference paper, we were unable to reproduce their exact results.

The heatmap displays in figure 5 show the trend of 1000 differentially expressed genes between the P0 and Adults. This heatmap showed an interesting pattern such as when the expression value of P0 is high, the value of adult is low. This similar trend was also shown in figure 5A and B when the genes related to the mitochondria and sarcomere. The P4 and P7 samples are representing the time between the P0 and adult. This could be because the gene expression of these groups of samples is between the neonatal and the adult mice. In P0 and P4 a

particular group of the gene was strongly expressed which is represented in dark red color, while the adult gene did not express it at all which is represented in light yellow color. We predict that these genes could be associated with the cell cycle showing the regeneration of the mice cardiac myocyte. In the reference paper, Figure 2A represents the heatmap that is clustered from the highest to the lowest gene expression values [4]. Since the heatmap in the paper only showed the hierarchical clustering of all expressed genes over the course of in vivo maturation with no samples being clustered, there was no distinct feature found which made it difficult for further analysis of these gene expression values.

CONCLUSION

Our project has successfully determined the potential regulation for mammalian cardiac regeneration for neonatal mice and well determined whether the myocytes reverse the transcription phenotype to a less differentiated state during the regeneration process.

There were certainly some reproducibility issues which likely affected results to compare with O'Meara et. al., but the tools being used have been updated over the several years that's passed since its release. For a future study based on a comparison, it would be beneficial to run the analysis with modern/updated tools along with the same databases used to assure reproducibility in methods. It is also important to note that while exact results varied, there were also similarities (such as the strong up-regulation of mitochondrion).

When comparing the results that we have obtained from the DAVID analysis in 6.7, we encountered some difficulties analyzing and interpreting the .csv file in order to find out the cluster scores for the different enrichment terms in both up and down-regulated genes. We solved the problem by closely comparing the contrasting the data we have obtained and the reference, we were able to observe the differences and the similarities.

REFERENCES

[1] Woodcock, E.A., Matkovich, S.J., 2005. Cardiomyocytes structure, function and associated pathologies. Int. J. Biochem. Cell Biol. 37, 1746–1751. https://doi.org/https://doi.org/10.1016/j.biocel.2005.04.011

[2] Beltrami, A.P., Urbanek, K., Kajstura, J., Yan, S.-M., Finato, N., Bussani, R., Nadal-Ginard, B., Silvestri, F., Leri, A., Beltrami, C.A., Anversa, P., 2001. Evidence That Human Cardiac Myocytes Divide after Myocardial Infarction. N. Engl. J. Med. 344, 1750–1757. https://doi.org/10.1056/NEJM200106073442303

[3] Porrello, E.R., Mahmoud, A.I., Simpson, E., Hill, J.A., Richardson, J.A., Olson, E.N., Sadek, H.A., 2011. Transient Regenerative Potential of the Neonatal Mouse Heart. Science (80-.). 331, 1078 LP – 1080. https://doi.org/10.1126/science.1200708

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

[5] Trapnell, C. et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 7(3): 562-578.
[6] Kim, D. *et al.* (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions, and gene fusions. *Genome Biol*. 14(4): 1-13.
[7] Trapnell, C. *et al.* (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology.* 28: 511-515.
[8] Roberts, A., Trapnell, C., *et al.* (2011). Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol.* **12**: 1-14.
[9] Trapnell, C., Hendrickson, D., *et al.* (2012). Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology* **31**: 46-53.
[10] R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
[11] Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nature Protoc. 2009;4(1):44-57.
[12] Huang DW, 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.
[13] Analytics Vidhya This is the official account of the Analytics Vidhya team. "What Is Tidyverse: Tidyverse Package in R." *Analytics Vidhya*, 14 June 2020, www.analyticsvidhya.com/blog/2019/05/beginner-guide-tidyverse-most-powerful-collection-r-packages-data-science/.
[14] Anderson, Sean. "An Introduction to reshape2." *Reshaping Data Easily with the reshape2 R Package.*,seananderson.ca/2013/10/19/reshape/#:~:text=reshape2%20is%20an%20R%20package,between%20wide%20and%20long%20formats.
[15] The UniProt Consortium, 2021. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 49, D480–D489. https://doi.org/10.1093/nar/gkaa1100