Transcriptional Profile of Mammalian Cardiac Regeneration with mRNA-Seq

BF528 Individual Project by Reina Chau

Project 2 - Programmer and Analyst

Github: https://github.com/RC-88/bf528_project_5

Introduction

The neonatal mice can regenerate their heart tissues in response to injury but lose this ability after their first week of life. The transcriptional changes that are responsible for the mammalian cardiac regeneration have not been fully characterized at the molecular level. Therefore, the goal of this study was to identify gene expression patterns and pathways that associated with the cardiac regeneration of the neonatal mice.

To assess the underlying genes and pathways that supported the cardiac regeneration of the neonatal mice, TopHat was used to align the sequencing reads to the mm9 mouse genome, and Cufflinks was used to assemble the transcripts, determine relative abundances of the transcripts measured in fragments per kilobase per million fragments mapped, and identify differentially expressed genes that passed the FDR adjusted p-value < 0.01 to account for multiple testing. Lastly, DAVID was used to identify the GO category enrichment for each cluster of genes that differed between the postnatal day 0 and adult mouse samples.

Methods

Data samples

Both in vitro and in vivo models were utilized to identify the transcriptomic pathways associated with the switch from a proliferative cardiomyocyte to a post-replicative cardiomyocyte and included the use of murine embryonic stem cells differentiated to cardiomyocytes, cardiac ventricular tissue isolated from neonatal mice (P0, P4, and P7 time points) and adult mice (8-10 weeks of age), cardiomyocytes isolated from adult mice (cultured for 0, 24, 48, and 72 hours), and isolated cardiomyocytes from sham surgery and apical resected murine hearts. In this analysis, the paired-end reads of the postnatal day 0 mice samples (GSM1570702) were obtained from GEO Series GSE64403 and used for sequencing alignment.

Aligning RNA-Seq to the Mouse Genome Reference

TopHat was utilized to align the paired-end reads of the neonatal samples to the mm9 mouse genome. TopHat is a robust aligning program that is based on the ultrafast short read mapping program of Bowtie, and it is ideal to be used in this study as it aligns the paired RNA-Seq reads with the reference annotation and identify exon-exon splice junctions. There is a total of 21,577,562 input reads obtained from the left and right read

files. Out of the total reads, the left read had 20,878,784 (~96.8% of input) mapped, including 1,468,843 (~7.0%) reads with multiple alignments. The right read had 20,510,550 (~95.1% of input) mapped, including 1,431,111 (~7.0%) reads with multiple alignments. A summary of the alignment is shown in **Table 1**. The overall mapping rate of the reads is ~95.9% and the alignment rate of the paired-end reads is ~88.9% which indicates a good quality alignment.

Table 1: Summary of the sequencing alignment obtained from TopHat.

Read files	Total mapped reads	% of mapped reads	Multiple alignments	% of multiple alignment
Left	20,878,784	96.8%	1,468,843	7.0%
Right	20,510,550	95.1%	1,431, 111	7.0%

Performing Quality Control on the Mapped Reads

There are three python scripts, geneBody_coverage.py, inner_distance.py, and bam_stat.py that were used to perform different quality control metrics on the mapped reads. These scripts are parts of the RSeQC package in python3. To run the geneBody_coverage.py and inner_distance.py scripts, a sorted and indexed BAM file generated from TopHat and a reference gene model of the mouse (mm9.bed) must be provided as inputs. The geneBody_coverage.py is a script that was used to calculate the RNA-seq reads coverage over gene body and check if any 5'/3' bias was present. The result in **Figure 1** shows the RNA-Seq reads are slightly skewed towards the 3' end of the gene body. The skewness towards increased 3' coverage could be attributed to the technical or biological processes during any of the experimental steps, such as reaction failure, or cell death, triggering mRNA degradation.

Inner_distance.py was used to calculate the inner distance or insert size between two paired-end RNA reads. The result in **Figure 2** shows a rough belled curve skewed slightly to the left with a mean of 85.4bp and standard deviation of 43.4bp. This skewness often indicates the inconsistency of inner distance between the DNA fragments and should be analyzed to detect structure variation or aberrant splicing.

Lastly, bam_stat.py was used to calculate the mapping statistics from a BAM file including QC failed, unique mapped, splice mapped, mapped in proper pair, etc. In a total of 49,706,999 reads, 2,899,954 reads are determined to be non-uniquely mapped while 38,489,380 reads are uniquely mapped. The ratio of non-uniquely mapped reads as related to uniquely mapped reads is ~0.08% which indicates the amount of non-uniquely reads are low and the majority of reads were mapped to the reference genome.

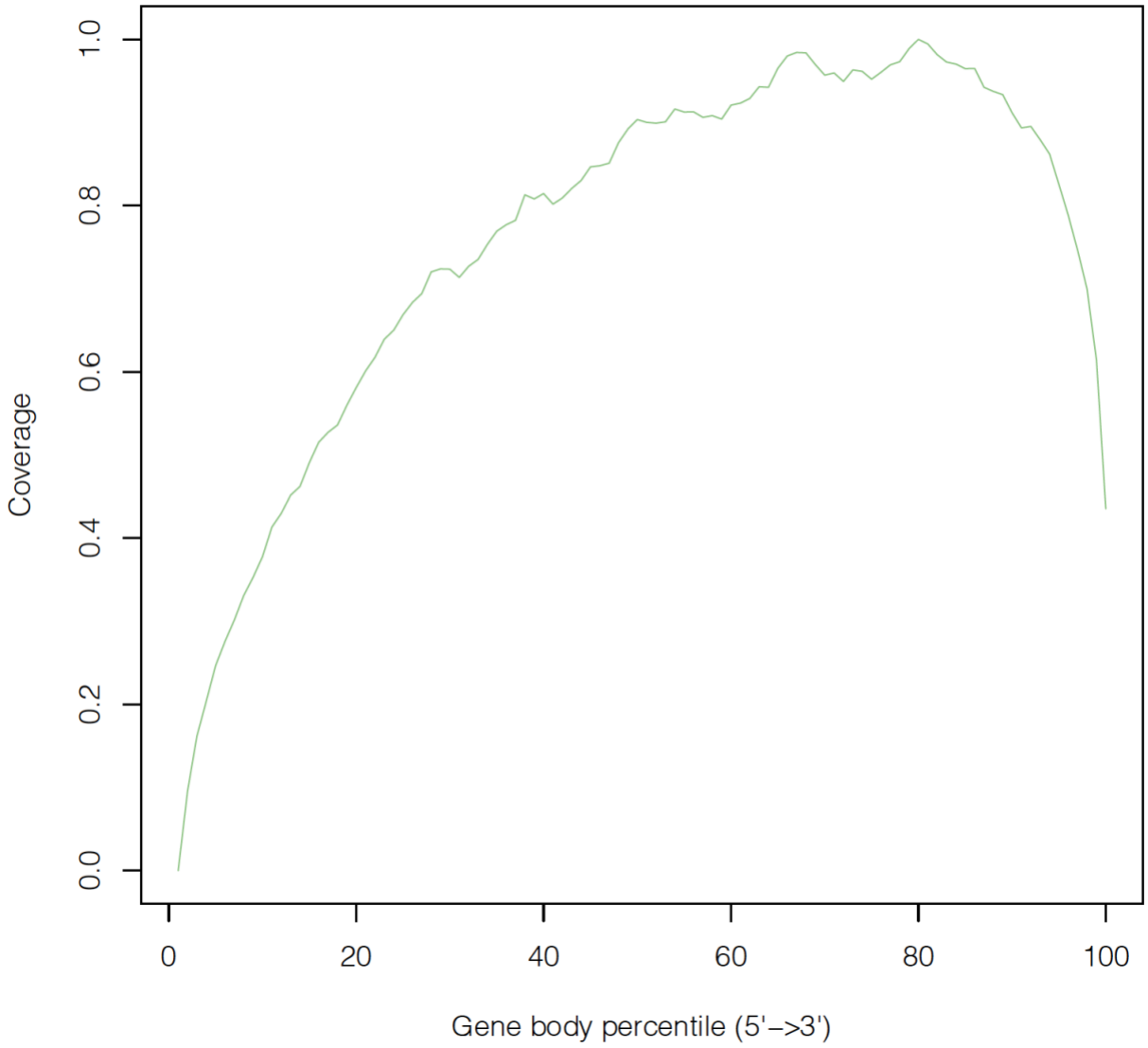


Figure 1: **RNA-Seq Reads Coverage for the Entire Gene Body**. The RNA-Seq reads are slightly skewed towards the 3' end of the gene body. This skewness could be attributed to the technical or biological processes during any of the experimental steps, such as reaction failure, or cell death, triggering mRNA degradation.

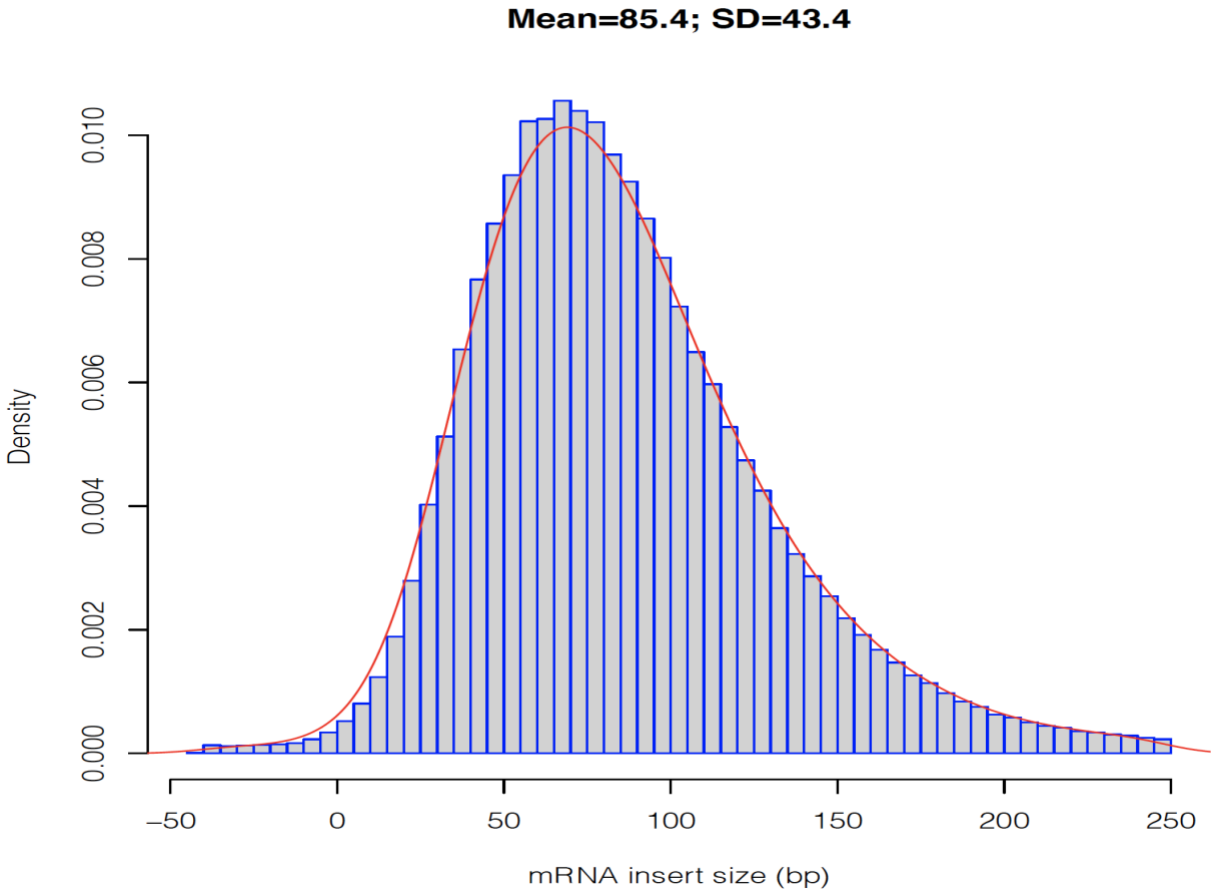


Figure 2: **Inner distance distribution between paired reads**. The distribution of the insert size is slightly skewed to the left with a mean of 85bp and standard deviation of 43bp. This skewness often indicates structure variation or aberrant splicing in the RNA-seq samples.

Assembling Transcriptomes

Cufflinks was used to assemble transcripts and estimates their abundances. The algorithm of Cufflinks accepts aligned RNA-seq reads and assembles the alignments into a parsimonious set of transcripts. Lastly, Cufflinks estimates the relative abundances of these transcripts and produces quantified alignments in FPKM for all genes. After removing the genes with an FPKM value of zero, the FPKM values had a range value of 0 to 2,604,770. This high variation in FPKM values indicates the quantified alignments have greater variability than it was expected. To eliminate the high variability in FPKM values, a log transformation was performed and a distribution of log FPKM values is shown in **Figure 3**. It seems that the distribution of log FPKM values is still slightly skewed to left even after the log transformation. This persisted skewness indicates additional normalization methods can be performed to further smooth out the distribution.

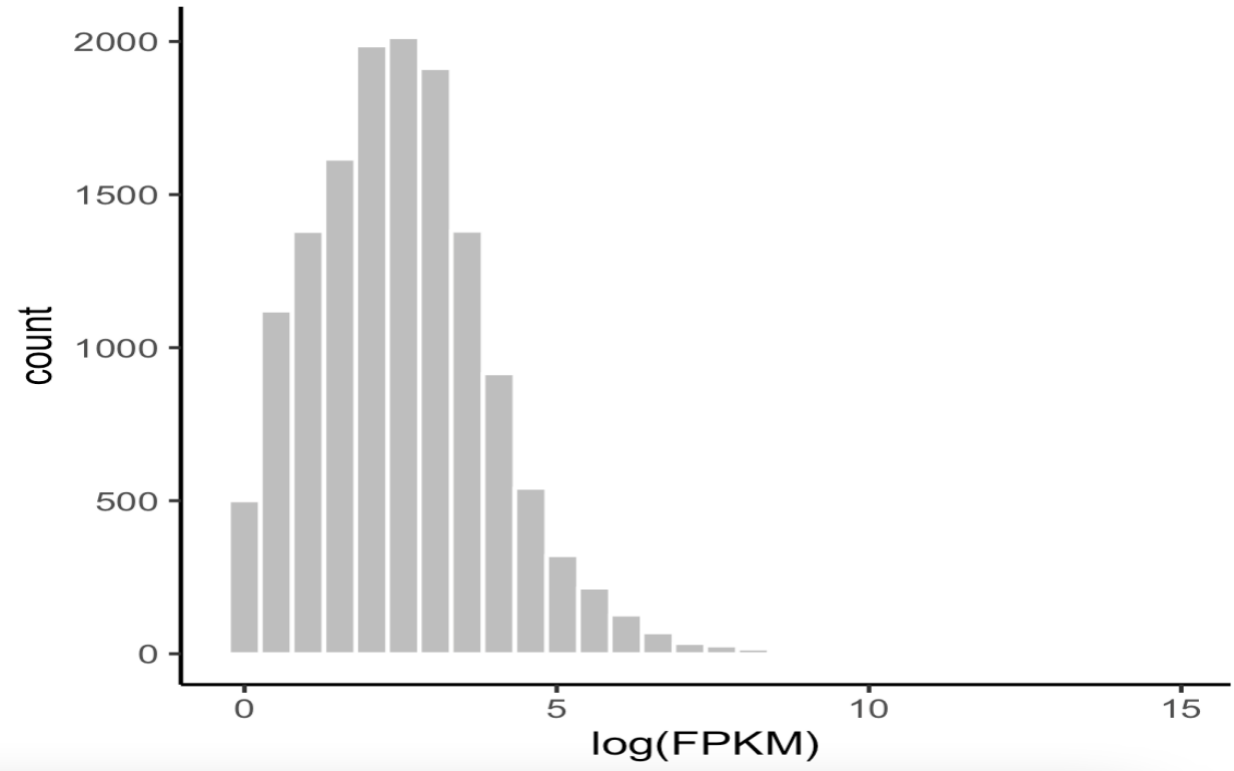


Figure 3: Distribution of log-transformed FPKM values. The log-transformation was performed to eliminate the high variability in the FPKM values as they are ranged from 0 to 2,604,770.

Identifying Differentially Expressed Genes and Gene Ontology Enrichment

In Cufflinks, there is a tool suite called cuffdiff that was used to identify differentially expressed genes between the postnatal day 0 versus adult samples. Cuffdiff tests the observed log fold change of the two samples against the null hypothesis of no change (i.e., is the true log fold change equaled to zero?). If the gene transcripts that passed the threshold of significance of p-value < 0.05 and adjusted FDR p-value < 0.01, they are considered differentially expressed. The significant genes were further subset into up and down-regulated genes and utilized by DAVID, a functional annotation clustering tool, to identify the GO category enrichment for each cluster of genes that differed between the postnatal day 0 versus adult samples.

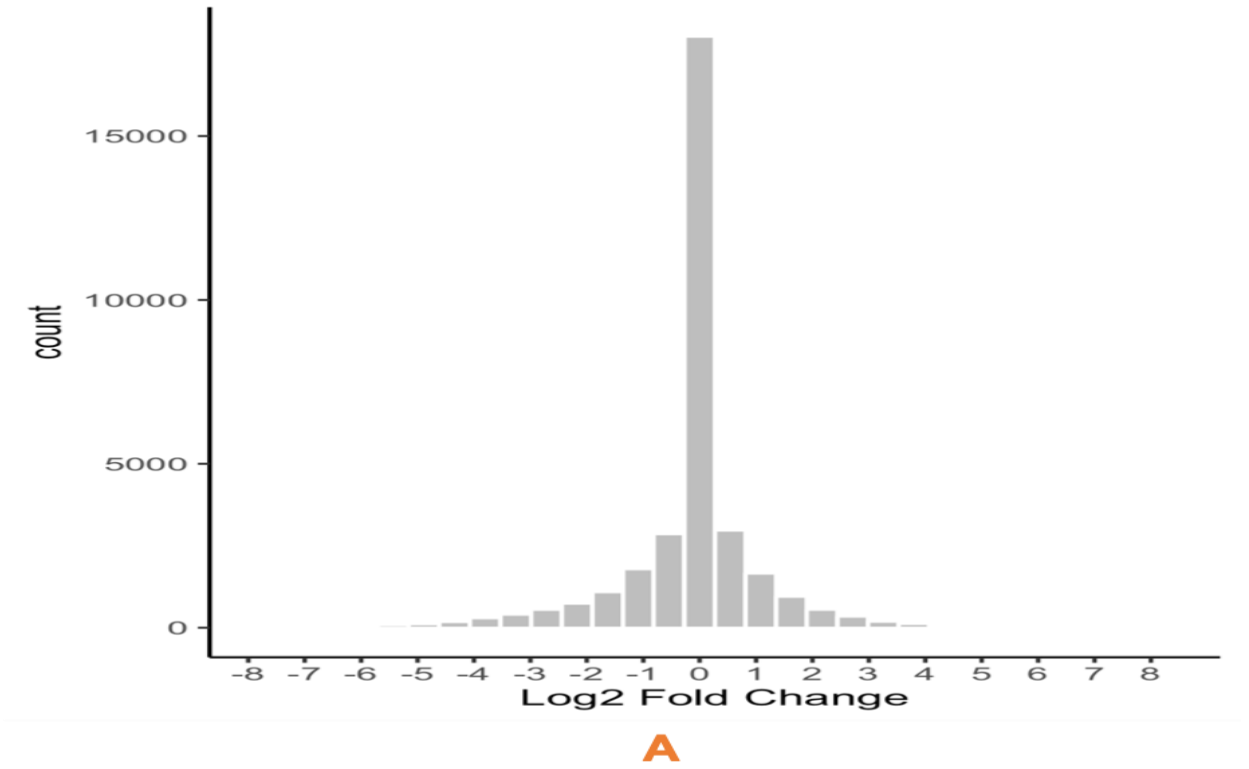
Results

With the use of cuffdiff, we identified 2,139 genes that were differentially expressed between the two samples, cardiac tissue from postnatal day 0 versus adult mice. **Table 2** shows the top 10 differentially expressed genes between the two samples, and **Figure 4A** shows the distribution of log2 fold change of all the significant genes. Out of the 2,139 differentially expressed genes, 1,084 are up-regulated genes while 1,055 are down-regulated genes. A histogram of the log2 fold changes for both up- and down-

regulated genes is illustrated in **Figure 4B**. The left side of the histogram represents the down-regulated genes while the right side of the histogram shows the up-regulated genes.

Table 2: Top 10 genes that are differentially expressed between the neonatal and adult mouse samples.

Gene	Locus	Log2 Fold Change	P-value	Q-value
PLEKHB2	chr1:34906803-34936425	1.70481	0.00005	0.00106929
MRPL30	chr1:37947321-37955380	1.51794	0.00005	0.00106929
COQ10B	chr1:55109613-55129545	2.26901	0.00005	0.00106929
AOX1	chr1:58086774-58163257	2.57682	0.00005	0.00106929
NDUFB3	chr1:58643227-58652808	1.39851	0.00005	0.00106929
SP100	chr1:87496952-87606573	5.56218	0.00005	0.00106929
CXCR7	chr1:92100554-92113326	2.70247	0.00005	0.00106929
LRRFIP1	chr1:92895303-93025521	-2.27184	0.00005	0.00106929
RAMP1	chr1:93076398-93121773	-4.25594	0.00005	0.00106929
GPC1	chr1:94728221-94756775	1.8557	0.00005	0.00106929



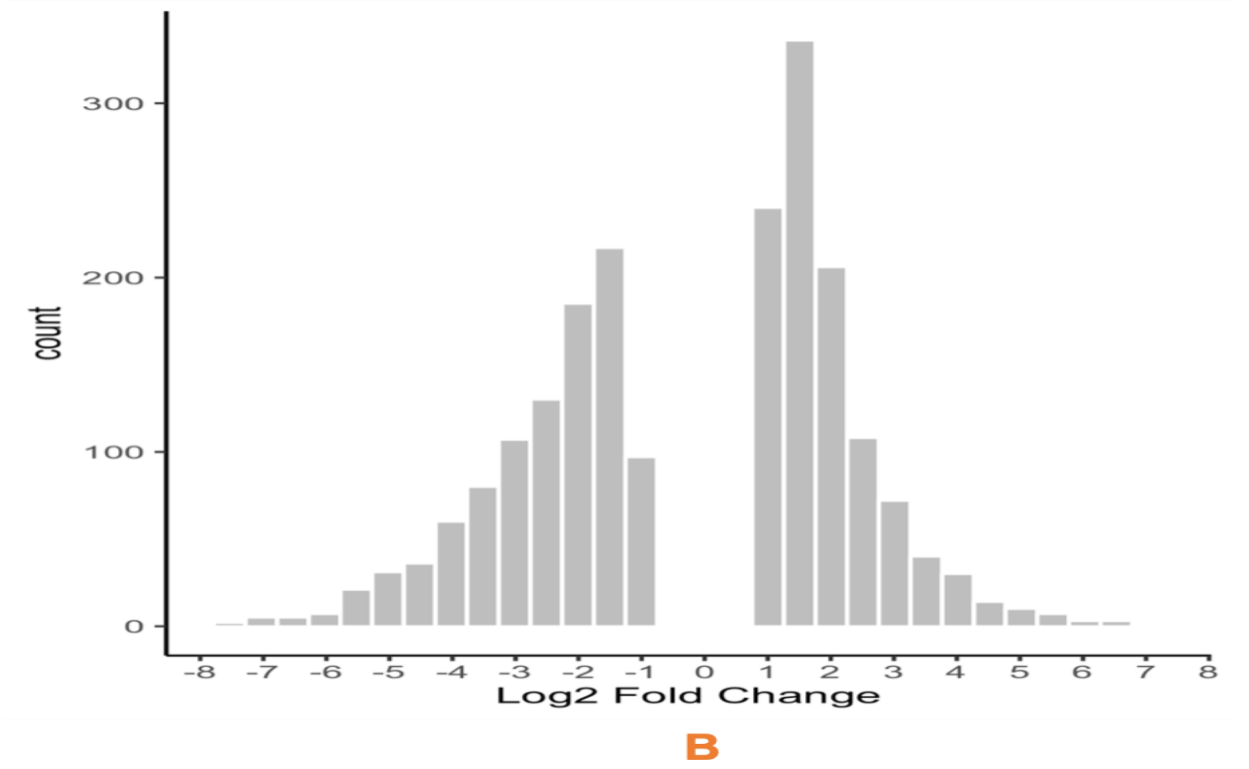


Figure 4: Distribution of log2 fold change of all significant genes (**A**) and distribution of log2 fold change for both up- and down-regulated genes (**B**). In Figure **4B**, the left

handed-side are the down-regulated genes while the right-handed side are the up-regulated genes.

To identify biological pathways that associated with the cardiac generation of neonatal mice as opposed to adult mice, DAVID was used to provide annotations of the genes regarding their roles in biological processes, cellular components, and molecular functions. The top 3 functional clusters obtained from DAVID analysis are shown in **Table 3**. The enrichment terms for the top three clusters shown similar results as reported in the original paper (O'Meara et al., 2015, Figure 1D). For instance, the enrichment terms associated with mitochondrion (*PRDX3, ACAT1, ECHS1, SLC25A11, PHYH*) and sarcomere (*PDLIM5, PYGM, MYOZ2, DES, CSRP3, TCAP, CRYAB*) were highly expressed in the cardiac tissue of adult mice as opposed to tissue from neonatal mice. Reversely, genes that involved in cell cycle (*CDC7, E2F8, CDK7, CDC26, CDC6, E2F1, CDC27, CDC45, RAD51, AURKB, CDC23*) were highly expressed in neonatal mice but down-regulated in adult mice (**Figure 5**).

Table 3: Gene enrichment terms associated with the up- and down- regulated genes obtained from DAVID analysis. *** indicates similar results were reported in original paper.

	Up-regulated		Down-regulated	
	Enrichment Term	**Score**	**Enrichment Term**	**Score**
Cluster 1	Mitochondrion *** Mitochondrial part Mitochondrial envelope Oxidoreductase activity	21.93	Cell cycle *** Cell cycle process Mitotic cell cycle Mitotic nuclear division	11.11
Cluster 2	Organic acid metabolic process *** Carboxylic acid metabolic process *** Fatty acid metabolic process *** Lipid metabolic process ***	16.81	Proteinaceous extracellular matrix Extracellular matrix Extracellular matrix component	9.69
Cluster 3	Generation of precursor metabolites and energy Purine ribonucleotide metabolic process *** Purine nucleoside metabolic process *** Cellular respiration ***	15.31	Cell proliferation Regulation of cell proliferation Positive regulation of cell proliferation Negative regulation of cell proliferation	9.58

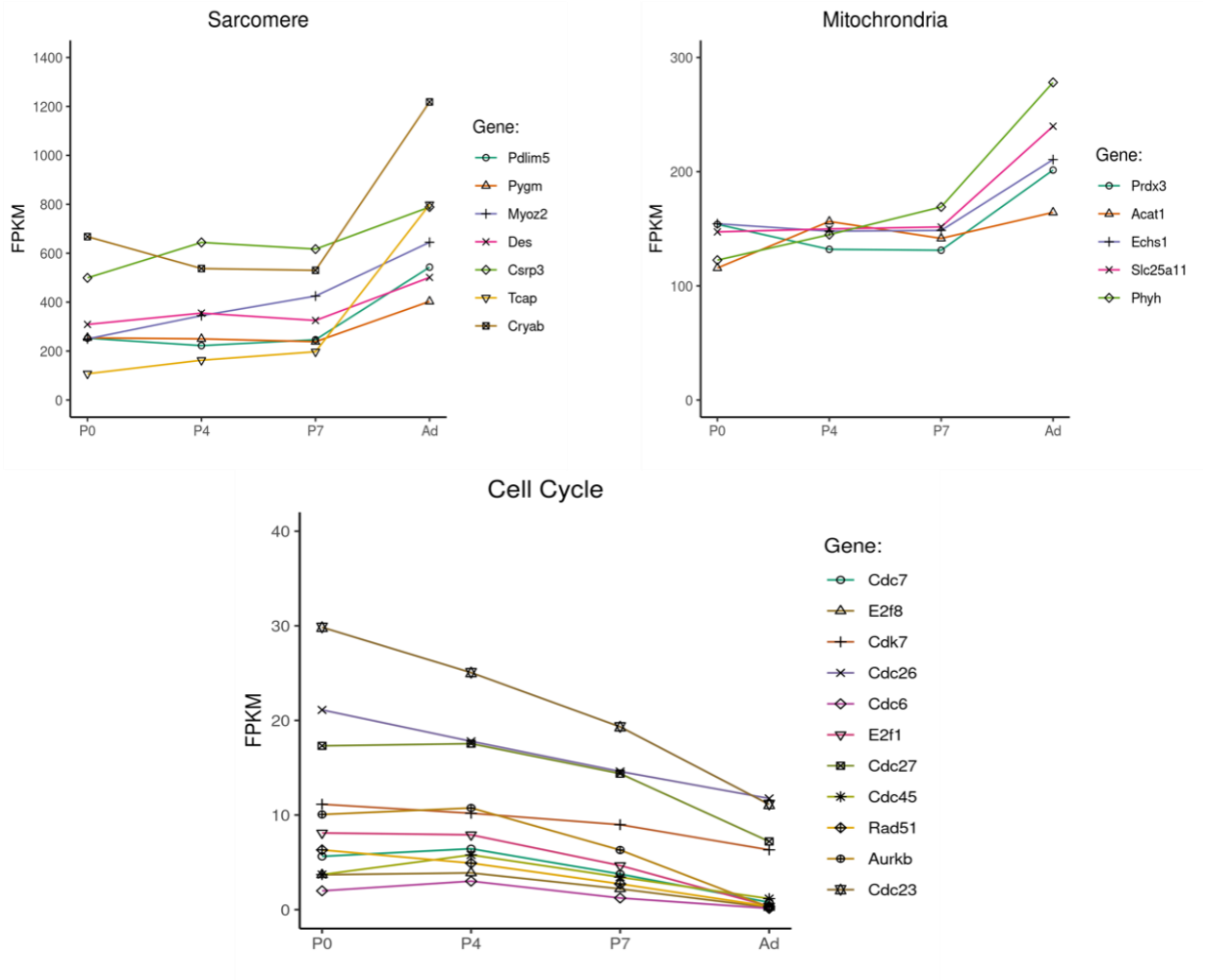


Figure 5: FPKM values of representative sarcomere, mitochondrial, and cell cycle related genes that were differentially expressed in heart tissue from postnatal day 0 (P0), 4 (P4), and 7 (P7) and adult (Ad) animals.

Discussion

The neonatal mice have the capability of regenerating their heart tissue if injured. However, this ability was lost after their first week of life. In order to identify regulators of the regenerative process, the gene expression patterns were compared between the cardiac tissue of the neonatal mice as opposed to post-replicative cardiac tissue of adult mice.

In this analysis, similar results were found as in the publication by O'Meara *et al*. Genes that involved in sarcomere assembly and organization, and mitochondrial-related genes were more abundant in cardiac tissue from adults compared to post-natal day 0 mice. Reversely, genes that involved in cell cycle regulation were less abundant in the heart tissue of adult animals compared to tissue from the neonatal animals. The loss in ability

to undergo replication was likely contribute to the exit of cardiomyocytes from the cell cycle and becoming non-replicative cells.

Cardiovascular disease is one of the major causes of death in the United States. During a myocardial infarction, the blood supply to regions of the heart is cut off and results in apoptotic and necrotic cell death of the cardiomyocytes. As a result of the loss of cardiomyocytes, the heart is not able to function as effectively, and is associated with poorer quality of life and a high risk of mortality. Therefore, by understanding the pathways and mediators of cardiac regeneration, we can identify potential therapeutic implications with such pathways and manipulated them to repair the human heart following injury, such as post-myocardial infarction.

References

1. O'Meara CC, Wamstad JA, Gladstone RA, Fomovsky GM, Butty VL, Shrikumar A, Gannon JB, Boyer LA and Lee RT. Transcriptional reversion of cardiac myocyte fate during mammalian cardiac regeneration. *Circ Res*. 2015;116:804-15.

2. Piquereau J and Ventura-Clapier R. Maturation of Cardiac Energy Metabolism During Perinatal Development. *Front Physiol*. 2018;9:959.

3. Cole Trapnell, Lior Pachter, Steven L. Salzberg. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009 May 1; 25(9): 1105–1111.

4. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research.* 2008; **18**:1851-8.

5. Trapnell, C., Williams, B., Pertea, G. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol.* 2010; 28, 511–515.