Transcriptional Profile of Mammalian Cardiac Regeneration with mRNA-Seq

Group Van Gogh

Data Curator: Andrew Gjelsteen

Programmer: Elysha Sameth

Analyst: Lindsay Wang

Biologist: Monil Gandhi

INTRODUCTION

Mammals have shown to have the capacity to repair and regenerate myocardial damage following an injury. However, shortly after birth, cardiomyocytes (CM) retain a limited ability for repair and replacing the loss of functional ability is inadequate. Rather than undergoing CM proliferation, postnatal hearts form scar tissue that can impede proper functioning of the heart[1]. Up to a week after birth, neonatal mice have exhibited this phenomenon and can fully regenerate their heart following resection. New CM in the injured or excised site have been demonstrated to regenerate from present myocytes as opposed to stem or progenitor cells. As a result, identifying the mechanism of cell cycle activity in myocytes during this process is important in understanding the molecular hurdles that prevent regeneration in the adult heart[2].

Although previous studies have stated that CM undergo partial reversion of cell fate during mouse heart repair, this idea has been based primarily on observations at the ultrastructural level[2]. Within O'Meara, et al., gene expression patterns of mouse hearts at different stages of postnatal development (P0, P4, P7, Adult) were compared to understand the transcriptional changes that accompany regenerative response to injury. By examining genes that were differentially expressed over the course of mouse CM differentiation in neonatal mice and loss of differentiation in adult mice, key regulators and mediators were identified.

Our study attempts to reproduce the methods outlined by O'Meara, et al. for the P0 sample and analyze differential gene expressions from P0, P4, P7, and adult mice. Using mRNA-Seq datasets publicly published by the paper, our analyses include performing quality control metrics, quantifying gene expression for a P0 replicate, identifying differentially expressed genes across samples, and performing functional annotation clustering of genes. In doing so, we have found that our results compare to O'Meara, et al. and validate their findings.

DATA

For this study, we processed and analyzed the O'Meara, et al. cDNA read data for cardiomyocytes isolated from CD-1 neonatal mice euthanized at postnatal day 0 (P0). Within the paper, hearts were excised, washed in ice cold PBS, and snap frozen in liquid nitrogen. Heart atria were dissected and discarded, and at least two heart ventricles were pooled for each replicate. Two replicas were then processed for RNAseq. From there, total RNA was extracted from all samples using Trizol (Invitrogen), according to the manufacturer's instructions. Polyadenylated RNA was isolated and purified from total RNA and was fragmented, with the first strand being synthesized using the Superscript III reverse transcription kit (Invitrogen). Double-stranded DNA was then synthesized with DNA polymerase I (Invitrogen), with end pair, A-tailing, adaptor ligation and size selection done using SPRI-Works System (Beckman Coulter).

Polyadenylated RNA was isolated from 1 to 10μg of total RNA using Dynabeads mRNA purification kit (Invitrogen). Poly-A RNA was fragmented, and the first strand was synthesized using the Superscript III reverse transcription kit (Invitrogen). Double-stranded DNA was then synthesized with DNA polymerase I (Invitrogen). End repair, A-tailing, adaptor ligation, and size selection were then performed using the SPRI-Works System (Beckman Coulter) followed by minimal amplification and addition of barcodes by PCR. Paired-end 40 base pair read length sequencing was then performed on an Illumina HiSeq 2000, and sequenced reads were aligned to the mm9 mouse genome using TopHat. For the purified neonatal CM samples, the TrueSeq (Invitrogen) sample preparation protocol was performed because of the low RNA yield.
We obtained the postnatal day 0 cDNA library's SRA (short read archive) data file through NCBI Gene Expression Omnibus, sample GSM1570702 of GEO accession number GSE64403.

METHODS

Data Acquisition and Quality Control of FASTQ Files

A short read archive (SRA) file containing cDNA library made from an RNA-seq library for *Mus musculus* 0 day postnatal ventricular myocardium cardiomyocyte was downloaded from the NCBI Gene Expression Omnibus portal (under sample GSM1570702 vP0_1[3] from Series GSE64403[4]) and uploaded to the Shared Computing Cluster[5] (SCC). The *qsub_extract.qsub* script was submitted on the SCC, which utilized SRA toolkit's fastq-dump[6] function to extract

the paired-end read SRA format to two FASTQ files. These FASTQ files were assessed for quality using the FASTQC[7] package, which produced Figure 1 and Appendix A.

Sample Alignment to Reference Genome

The sequenced paired-end reads from the P0_1 sample were aligned to the mouse reference genome called mm9 using TopHat[8]. Using the mm9 FASTA and Bowtie2[9] indexes available on the SCC, RNA splice sites in the P0_1 sample were discovered *de novo* by performing the alignment with the arguments specified by O'Meara, et al. This included setting the expected (mean) inner distance between mate pairs to 200 (*--r 200*), cutting each read into 20 base pair segments (*--segment-length=20*), allowing 1 mismatch in each segment alignment (*--segment-mismatches=1*), and only looking for reads across junctions indicated in the mm9 gene model annotation (*--no-novel-juncs*). Since TopHat is computationally demanding and takes approximately an hour to run, a *run_tophat.qsub* script containing the command was submitted as a batch job on the SCC with the dependencies samtools-0.1.19[10], bowtie2 and boost[11] loaded. Upon completion of the alignment, the file *accepted_hits.bam* was outputted which contained the original reads and any alignments discovered.

Quality Control Metrics

The *accepted_hits.bam* file was further evaluated for flag-based mapping quality, uniform read coverage, read pair insert size, and mapping statistics. To assess quality-controlled reads, the samtools flagstat argument was applied to *accepted_hits.bam*. Results were analyzed using both samtools-0.1.19 and samtools-1.10 due to differences in output and the presence of the secondary and supplementary categories in version 1.10. Coverage uniformity over gene body, insert size calculation, and read mapping statistic calculation was done using the *geneBody_coverage.py*, *inner_distance.py*, and *bam_stat.py* (*--i accepted_hits.bam --r mm9.bed*) modules in the RseQC[12] package. In order to use this package, the BAM file was indexed using `samtools index` to create an *accepted_hits.bam.bai* for quick extraction of alignments overlapping particular genomic regions. These modules outputted the coverage profile along the gene body and inner distance distribution, which were visualized using R[13], as well as a summary of the mapping statistics. All three quality control metrics were run on the SCC from a qsub script called *run_rseqc.qsub* with python2[14] and samtools-0.1.19 loaded.

Quantifying Gene Expression

Gene expression levels were quantified using the Cufflinks[15] package (*--compatible-hit-norm -G -b -u -p 16*). Arguments provided were defined by O'Meara, et al. and were used to count only fragments compatible with some reference transcript, run bias detection, and do an initial estimation procedure to more accurately weigh multi-mapped reads. After approximately three hours, a *genes.fpkm_tracking* file containing the relative abundance of transcripts in fragments per kilobase per million (FPKM) for all genes in the P0_1 sample was outputted. The distributions of the FPKM values with FPKM of 0 removed and FPKM > 1 were then visualized using R with $\log_{10}$FPKM for visual clarity. This cutoff value was chosen due to O'Meara, et al. filtering for genes expressing > 1 FPKM in at least 1 sample in both replicates for hierarchical clustering and gene ontology enrichment. Finally, to determine differentially expressed genes, the replicate was compared to the remaining pre-prepared samples (P0_2, Ad_1, and Ad_2) located on the SCC using the Cuffdiff[16] package. The P0_1 *accepted_hits.bam* and the merged assembly were fed into Cuffdiff in the script *run_cuffdiff.qsub* and a set of outputted text files containing observed log fold changes were used in downstream analysis.

Gene Expression Analysis

Using the *gene_exp.diff* file from the Cufflinks output, genes with status "OK" and have at least one FPKM value greater than 1 were kept for analysis. After initial filtering, 14243 out of 36329 genes were kept for further analysis. By sorting the q-values, the top 10 genes with the smallest q-values were tabled. Then, using the same file, a histogram of the $\log_2$ fold-change of all genes and the genes with significant expression level were plotted. The significant genes were determined using the significant column in the *gene_exp.diff* file. Since the significant definition by Cuffdiff is whether the p-value is greater than the q-value, we also checked the p-values for the significant group. The result suggests that all the significant genes have p-values less than 0.01, but the reverse might not be true.

Based on the $\log_2$ fold-change value of the significant genes, we further separated the genes into up- or down-regulated groups using DAVID Functional Annotation Clustering. The gene ontology groups used for the analysis were GOTERM_BP_FAT, GOTERM_MF_FAT, and GOTERM_CC_FAT[17][18]. The top 5 clusters with the highest enrichment scores of up- and down-regulated genes were summarized and then compared to the original article.

RESULTS

Data Acquisition

Processing the short read archive files to extract two FASTQ files produced files containing high-quality DNA reads for each of the 40 nucleotide positions in the reads. These figures show each file's summary statistics and average Phred score for each position in the 40-base pair reads, respectively.

Measure	Value	Measure	Value
Filename	P0_1_1.fastq	Filename	P0_1_2.fastq
File type	Conventional base calls	File type	Conventional base calls
Encoding	Sanger / Illumina 1.9	Encoding	Sanger / Illumina 1.9
Total Sequences	21577562	Total Sequences	21577562
Sequences flagged as poor quality	0	Sequences flagged as poor quality	0
Sequence length	40	Sequence length	40
%GC	49	%GC	49

Figure 1. FASTQC's output summary statistics for p0_1_1.fastq and p0_1_2.fastq. Note that no sequences were flagged as poor quality for either file.

Sample Quality Control

The summary statistics of the accepted hits from samtools show that all 49,706,999 reads map to the reference genome (Table 1). Although there are some differences between samtools-0.1.19 and samtools-1.10, we can conclude that there are no duplicates and all alignments passed quality control with 65.32 - 71.09% properly paired (Table 1).

Samtools Flagstat Results				
	Samtools 0.1.19		**Samtools 1.10.0**	
Category	**QC-passed Reads**	**QC-failed Reads**	**QC-passed Reads**	**QC-failed Reads**
Total	49,706,999	0	49,706,999	0
Secondary	N/A	N/A	8,317,665	0
Supplementary	N/A	N/A	0	0
Duplicates	0	0	0	0
Mapped	49,706,999	0	49,706,999	0
Paired in Sequencing	49,706,999	0	41,389,334	0
Read1	25,089,027	0	20,878,784	0

Read2	24,517,972	0	20,510,550	0
Properly paired	32,466,938 (65.32%)	0	29,422,646 (71.09%)	0
With itself and mate mapped	47,843,662	0	39,936,472	0
Singletons	1,863,337 (3.75%)	0	1,452,862 (3.51%)	0
With mate mapped to a different chr	5,098,744	0	1,387,382	0
With mate mapped to a different chr (mapQ>=5)	704,916	0	704,916	0

Table 1. Summary statistic output from *samtools flagstat* versions 0.1.19 and 1.10. This tool provides counts for each read category and is broken down into QC pass and QC fail. In yellow are the rows/categories with differing values between samtools versions.

To assess the mapped reads further, the RseQC results were broken down as percentages of the total reads (Table 2). Of these reads, 77.43% were uniquely mapped and 16.7% aligned equally well at more than one location. These multi-mapped reads may be due to gene duplications, which are common genomic occurrences, or chimeras, which occur due to structural variations, gene fusions, RNA-seq or experimental protocols[19]. Our results from samtools-1.10 show that all of the multi-mapped reads have secondary alignments and no supplementary alignments (Table 1), therefore we can conclude that there are duplicated genes in our sample. While the presence of duplicated genes is expected, it may cause genes to be inaccurately quantified and affect the gene differential expression analysis. These genes are often removed or ignored from the analysis, however we performed multi-read correction in Cufflinks to accurately weigh the counts.

RseQC BAM Statistic Results	
Total Reads	49,706,999
Mapped	49,706,999 (100.00%)
Unique	38,489,380 (77.43%)
Multi-mapped	8,317,665 (16.7%)

Unaligned	0 (0.0%)

Table 2. Summary of the reads mapping statistics outputted by the RseQC *bam_stat.py* module. This script determines "uniquely mapped reads" from mapping quality, which quality the probability that a read is misplaced.

Using the RseQC package, coverage uniformity over the gene body and insert size were assessed. An analysis of the gene coverage along the full transcript length shows little coverage at the 5'-end, increasing coverage in the middle of the gene body, and peak coverage towards the 3'-end (Figure 2). The high coverage at the 3'-end indicates a gene coverage bias, which is as expected from the mRNA sequencing method used by O'Meara, et al. Within the study, polyadenylated RNAs were fragmented, therefore this 3'-end bias may represent RNA degradation and poly-A enrichment[20]. As a result, due to the nature of the method, we do not find the results to be of concern and conclude that the data is fit for further analysis.

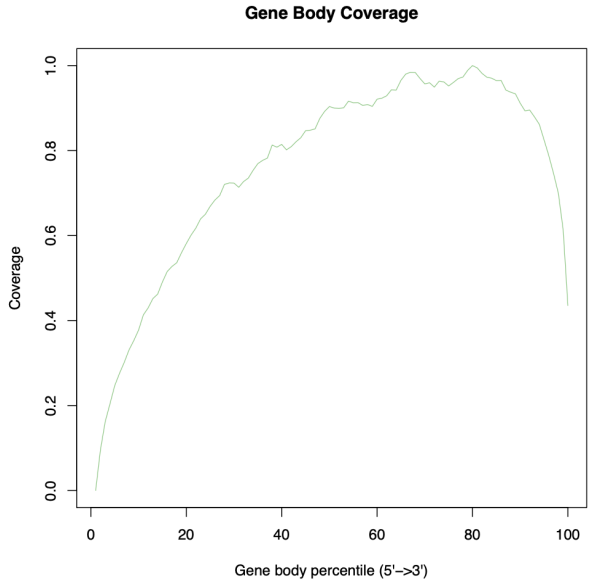


Figure 2. Line graph of the P0_1 read coverage (y-axis) over gene body (x-axis) generated using R. The number of reads covering each nucleotide position from the 5' to 3' end was calculated by passing the *accepted_hits.bam* file from TopHat into the *geneBody_coverage.py* utility in RseQC, a RNA-seq quality control package. This module scales all transcripts to 100 nt and calculates the number of reads covering each nucleotide position.

Since the fragmented RNA underwent paired-end sequencing, it is important to assess the distribution of insert size between two paired RNA reads. This can be used to provide evidence regarding the quality of the nucleic acid in the final library where short fragment sizes may indicate sample degradation or over-fragmentation[21]. Based on the inner distance plot from

RseQC (Figure 3), there is an average of 85.41 base pairs (SD = 43.42) between the alignments of the reads. However, there are negative insert sizes which indicate that some reads overlap and the fragment was shorter than 2x the read length of 40. While this may be a concern, there is a small number of reads that overlap and the insert sizes overall follow a normal distribution. Therefore, the overlapping may be due to intentionally fragmenting some reads to sequence them as a longer read rather than degradation or over-fragmentation. While this is sound for single-end read experiments, this could be valid if more depth was needed for some fragments of interest.

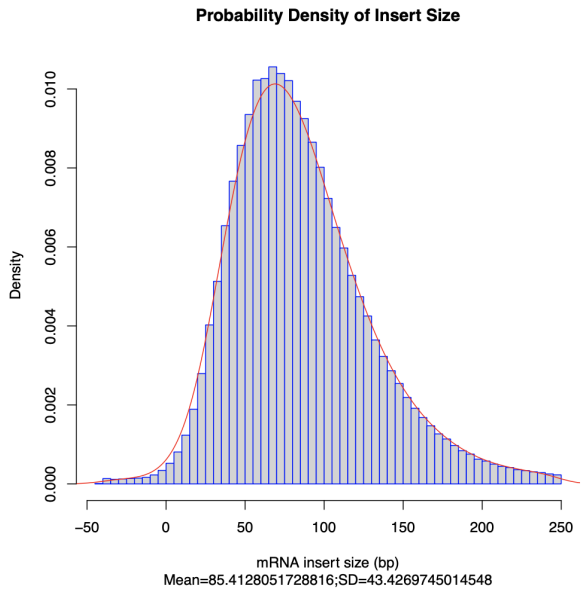


Figure 3. Probability density (y-axis) of the inner distance/insert size (x-axis) of the P0 sample fragments. These values were calculated using the inner_distance.py utility in RseQC, a RNA-seq quality control package, and visualized using R. The mean length between two paired fragments was found to be 85 base pairs (standard deviation of 43) with a minimum insert size of -50 and maximum 250. Negative insert size indicates read overlapping.

After aligning the P0_1 reads to the mm9 genome using TopHat and performing quality control metrics using samtools and RseQC, quantifying the gene expression levels was necessary to distinguish differences between neonatal and adult cardiomyocytes. Of the 37469 genes in the P0_1 sample, 16453 had a read mapping to that location and was expressed (Figure 4A). Over 85% of the genes have a $\log_{10}$FPKM > 0 (FPKM > 1) while the remaining have values that may be considered an FPKM of 0 (Figure 4B). Since 14205 genes have an FPKM > 1, genes lower than this cutoff were filtered out and it was found that ~92% fall within a $\log_{10}$FPKM of 0 and 2, or FPKM of 1 and 100 (Figure 4C). As a result, using FPKM > 1 as the cutoff for what is

considered 'expressed' is reasonable and does not remove a substantial amount of genes in our P0_1 sample when performing gene expression analysis.

.

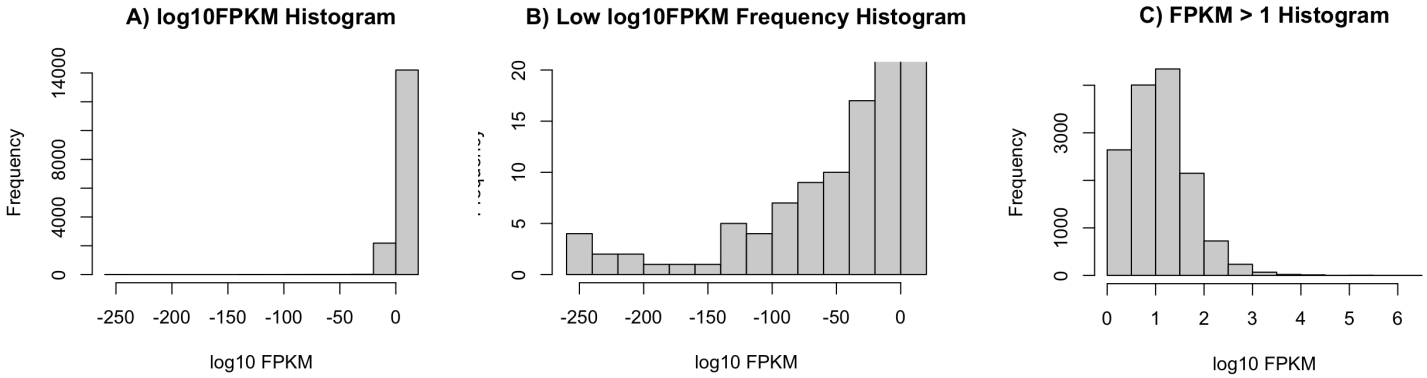


Figure 4. Histogram of the relative abundance of transcripts in fragments per kilobase of exon model per million mapped fragments (FPKM) for genes in the P0_1 sample. Due to variability within the data, the $\log_{10}$FPKM was graphed and FPKM of 0 were removed so only expressed genes were represented. **A) All $\log_{10}$FPKM with FPKM 0 removed.** Most of the $\log_{10}$FPKM values were found to be greater than 0. **B) $\log_{10}$FPKM values where the FPKM frequency is less than 20.** Of the 16453 genes that are expressed, 63 have frequency count less than 20 and cannot be seen in the overall $\log_{10}$FPKM histogram in A. **C) $\log_{10}$FPKM values where FPKM > 1.** This cutoff value was used due to the gene filter specified by O'Meara, et al. A majority of the $\log_{10}$FPKM values were found to fall between 0 and 2.

Differential Expression Analysis

Log2 Fold-Change of Top10 Genes With the Lowest q Value					
Gene	P0 FPKM	Adult FPKM	log2.fold_change.	p_value	q_value
Plekhb2	22.5679	73.5683	1.70481	5.00E-05	0.00106929
Mrpl30	46.4547	133.038	1.51794	5.00E-05	0.00106929
Coq10b	11.0583	53.3	2.26901	5.00E-05	0.00106929
Aox1	1.18858	7.09136	2.57682	5.00E-05	0.00106929
Ndufb3	100.609	265.235	1.39851	5.00E-05	0.00106929
Sp100	2.13489	100.869	5.56218	5.00E-05	0.00106929
Cxcr7	4.95844	32.2753	2.70247	5.00E-05	0.00106929
Lrrfip1	118.997	24.6402	-2.27184	5.00E-05	0.00106929
Ramp1	13.2076	0.691287	-4.25594	5.00E-05	0.00106929
Gpc1	51.2062	185.329	1.8557	5.00E-05	0.00106929

Table 3. Top 10 Genes that are most differentially expressed between postnatal P0 and adult, sorted based on ascending q value.

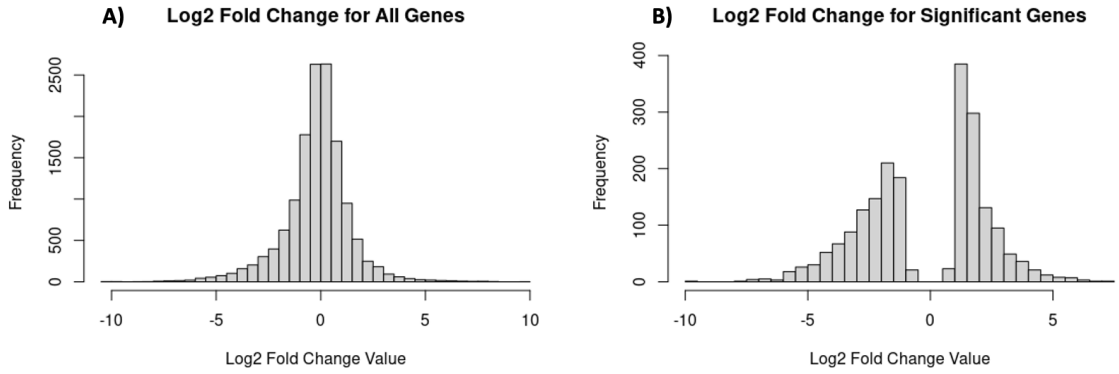


Figure 5. **A)** Histogram of Log2 fold-change for all of the genes. **B)** Histogram of Log2 fold-change differentially expressed genes (Cufflinks significant and $p < 0.01$).

The figure above summarizes the log2 fold-change score of all the genes and that for only the significant genes. From the plots, we can clearly see that the significant genes all have non-zero log2 fold-change values.

Based on the Cufflink analysis, we identified 2139 genes (15.0%) of the genes differentially expressed between postnatal P0 and adult. Among them, 1084 (50.7%) genes were up-regulated and 1055 (49.3%) genes were down-regulated. However, the article reported a total of 9779 significant genes with 2409 genes up-regulated and 7570 genes down-regulated[1].

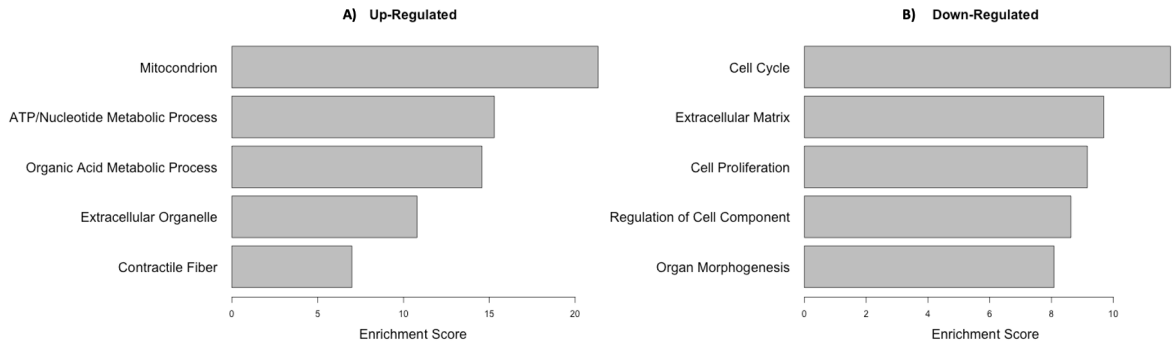
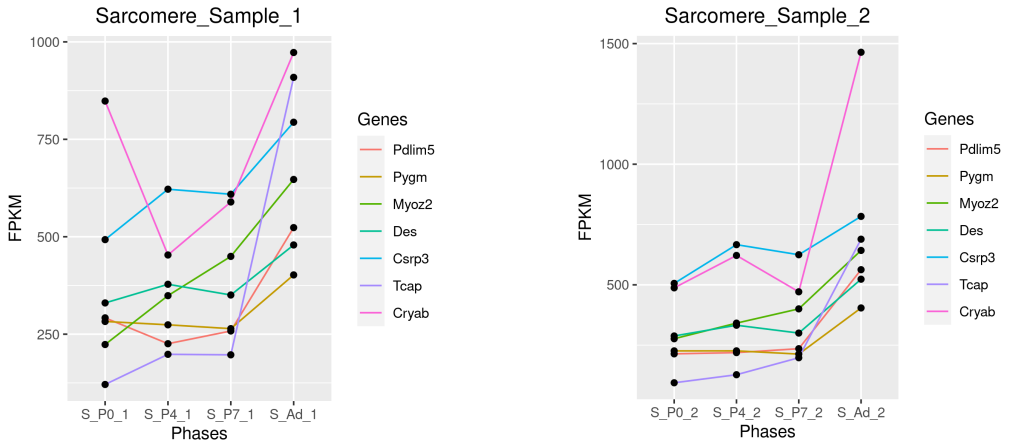


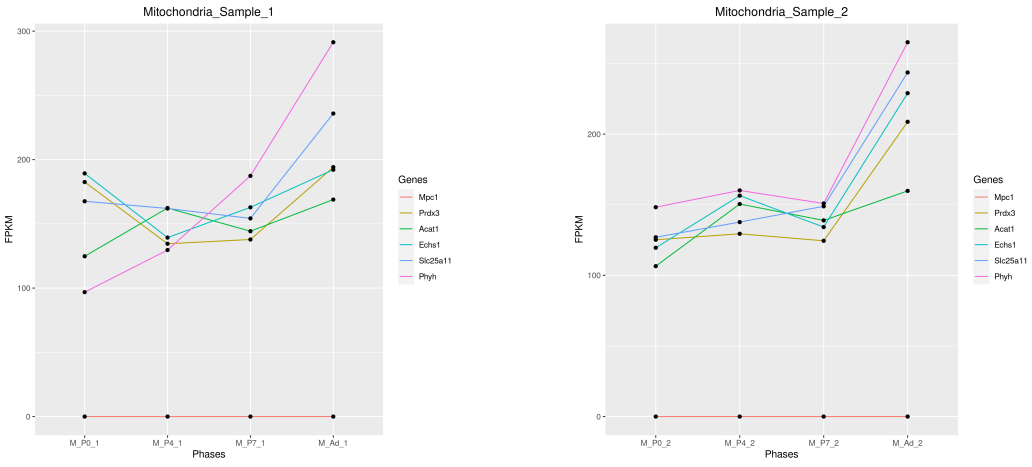
Figure 6. Gene ontology analysis result for significant gene from postnatal P0 and adult datasets. The y axis labels summarize the GO term cluster. The GO clusters with the 5 highest enrichments scores are plotted here. **A)** Barplot for genes that are up-regulated for the datasets. **B)** Barplot for genes that are down-regulated for the datasets.

Enriched gene ontology (GO) analysis between postnatal P0 and adult samples showed primarily up-regulation in mitochondrion and metabolic-related genes, and down regulation in cell cycle genes. The result for the ontology analysis is generally consistent with the original paper. The enrichment score of the two groups suggest that the up-regulated genes are better clustered than the down-regulated genes based on GO analysis. However, from Figure 1C of the original article[1], the enrichment terms of down-regulated genes should have higher enrichment scores. Since the enrichment term and enrichment scores for the up-regulated genes from our analysis are similar to the results in the paper, those for the down-regulated genes might be significantly enriched in the CM (cardiomyocyte) vs ESC (embryonic stem cell) dataset.

A)



B)



C)

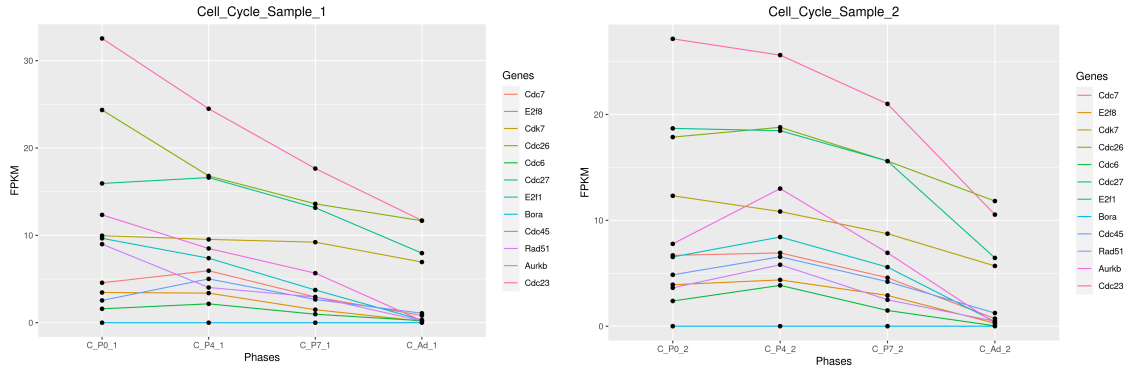


Figure 7. FPKM expression values for Sarcomere, Mitochondria and Cell Cycle genes. Each representative category was further divided into two replicate samples, sample 1 and sample 2. The horizontal axis are the phases which are divided into postnatal P0, P4, P7 and adult Ad groups. The gene categories are divided based on the replicate samples of the phases. The vertical axis are the FPKM values of the genes. **A)** Line plot for upregulated sarcomere genes group. **B)** Line plot for upregulated mitochondria genes group. **C)** Line plot for downregulated cell cycle genes group. Overall, the genes in sarcomere and mitochondria are upregulated and the genes in cell cycle are downregulated

Hierarchical Clustering and Gene Ontology Enrichment

Our results almost duplicated the reference paper in terms of the comparison analysis of the line plots in Figure 1D of O'Meara, et al. We plotted the FPKM values across several different developmental phases for sarcomere, mitochondria and cell cycle genes (Figure 7). The regulation of the sarcomere and mitochondria genes were significantly increasing between the postnatal P0 to the adult phase Ad while the genes in the cell cycle were down-regulated. The magnitude and direction of the three gene groups were similar to the paper, except for null FPKM values for two genes, one from each mitochondria (MpC1) and cell cycle (Bora) groups. As per O'Meara, et al., cardiac myocytes exit the cell cycle, which can be interpreted from the down-regulation of the cell cycle genes from P0 to Ad phase in Figure 7.

Cluster	**Biological Pathway**	**Enrichment Score**	**Ratio of similarity with the original reference**
1	Mitochondria	21.34461241497	0.6
2	Generation of precursor metabolites and energy	15.2869485107342	0.4

3	Organic acid metabolic process	14.5742562254800	0
4	Extracellular organelle	10.7892324863081	0.3
5	Sarcomere	6.99113782124350	1

Table 4. Table with top 5 upregulated GO clusters representing common biological pathways in comparison to the reference paper O'Meara, et al. The table contains cluster information with biological functions and their related enrichment scores. It also has a column for the ratio of similarity to the O'Meara, et al papers in relation to the biological pathways.

Cluster	Biological Pathway	Enrichment Score	Ratio of similarity with the original reference
1	Cell Cycle	11.85151895533	0.5
2	Extracellular matrix	9.6873935995841	0.6
3	Cell proliferation	9.1620571394321	0
4	Regulation of cellular component organization	8.6287252940002	0.7
5	Organ morphogenesis	8.0785583618391	0.6

Table 5. Table with top 5 downregulated GO clusters representing common biological pathways in comparison to the reference paper O'Meara, et al. The table contains cluster information with biological functions and their related enrichment scores. It also has a column for the ratio of similarity to the O'Meara, et al papers in relation to the biological pathways.

We compared our GO annotation clusters for both down-regulated and up-regulated genes to that of the reference paper. Above tables (Table 4 and 5) list the top five clusters based on the highest enrichment scores for the phases between P0 and Ad. The GO terms mitochondria and cell cycle overlapped with that of O'Meara, et al. for up- and down-regulated genes respectively. GO terms like Sarcomere were found to be the common up-regulated genes between the two papers. We observed that our results overlapped with that of O'Meara, et al. for the clusters with the highest enrichment scores, however some of our functional pathway clusters were found to be uncommon. This could be because we used specific terms for Gene Ontology such as GOTERM_BP_FAT, GOTERM_MF_FAT, and GOTERM_CC_FAT. For the up-regulated gene comparison, we had 1631 overlapped GO terms out of a possible of 4551,

whereas for down-regulated genes we computed a total of 1870 overlapped GO terms out of 4691.

Annotation Cluster 1	Enrichment Score: 11.851518955339927	PValue	Similarity
Category	Term		
GOTERM_BP_FAT	GO:0007049~cell cycle	7.51E-30	Yes
GOTERM_BP_FAT	GO:0051301~cell division	1.28E-26	Yes
GOTERM_BP_FAT	GO:0000278~mitotic cell cycle	3.72E-25	Yes
GOTERM_BP_FAT	GO:0022402~cell cycle process	4.84E-25	Yes
GOTERM_BP_FAT	GO:1903047~mitotic cell cycle process	1.85E-24	No
GOTERM_BP_FAT	GO:0007067~mitotic nuclear division	4.37E-17	No
GOTERM_BP_FAT	GO:0000280~nuclear division	6.27E-16	Yes
GOTERM_BP_FAT	GO:0010564~regulation of cell cycle process	2.98E-15	Yes
GOTERM_BP_FAT	GO:0048285~organelle fission	2.04E-14	Yes
GOTERM_BP_FAT	GO:0051726~regulation of cell cycle	3.58E-14	Yes
GOTERM_BP_FAT	GO:0007059~chromosome segregation	7.23E-13	No

Table 6. Aggregated table of the comparison for similar GO teams to the reference paper O'Meara, et al. The table shows the top enrichment score cluster with the corresponding p-value for each GO term. The Similarity column has a binary value for the similarity of the row to that of the reference paper O'Meara, et al.

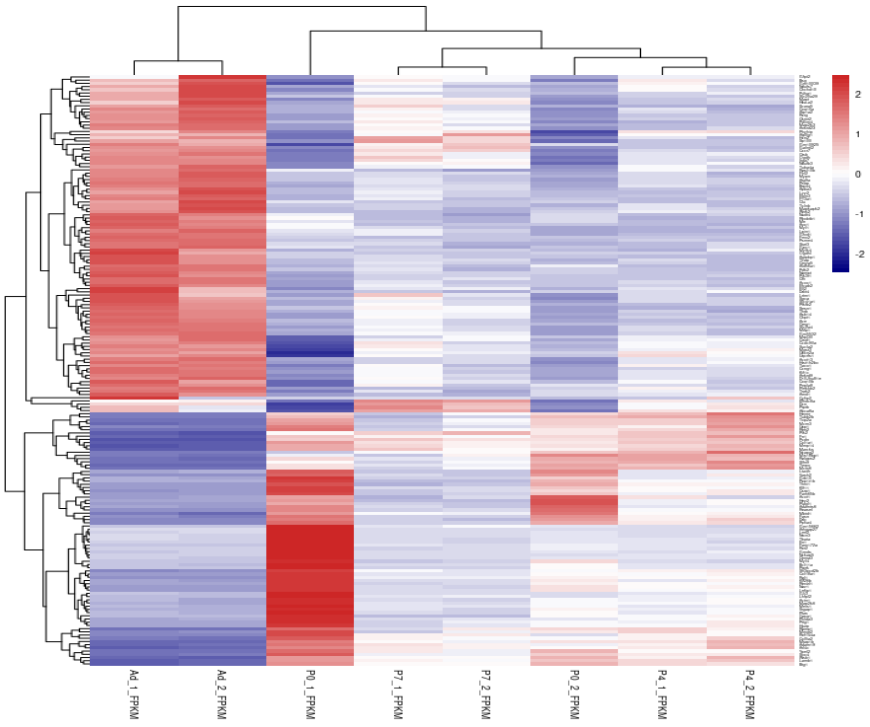


Figure 8. Clustered Heatmap for differentially expressed genes for different groups namely P0, P4, P7 and Ad. Each group is divided into replicate samples with suffixes of 1 and 2. The rows in the heatmap are the genes which are differentially expressed, whereas the columns are the samples of groups between P0 and P4, including Ad. The expression levels across the genes can be interpreted by the darkness of the columns and the replicate samples have almost similar expression levels.

We plotted a heatmap for the top 150 most differentially expressed genes with different developmental stages (Figure 8). The significant expression differences in the stages between P0 and Ad throughout the development of mice cardiac myocytes gives us an understanding of the repair in response to an injury or insight into regeneration of mice cardiac myocytes. Our hierarchical clustering expression heatmap was similar to the O'Meara, et al. Figure 2A, as it shows differential expression patterns in the different developmental stages and also gene clusters representing biological significance.

DISCUSSION

Through our replication of O'Meara et al.'s research, we extracted P0 FASTQ files, assessed read quality, aligned the reads to a reference genome, performed quality control metrics, profiled P0 gene expression, and conducted hierarchical clustering and enriched gene ontology analysis. In doing so, we have identified a core transcriptional signature of cardiac myocyte differentiation, and found that adult cardiac myocytes demonstrate a transcriptional reversion of the differentiation process. Our results are therefore consistent with O'Meara, et al. and we conclude that we were successful in our reproduction.

Although our results are similar to O'Meara, et al., inconsistencies have been found in the number of significant genes and enrichment scores. We found notably less significant genes between the P0 and adult groups than reported in the literature. In order to explain this difference, we compared the number of significant genes filtered out by other groups, and most groups have the same results as we do. Thus, the differences may be due to the normalization method and software version used to process the data. Furthermore, the enrichment scores of the significant GO biological function clusters in both up- and down-regulated genes did not match with those of the reference paper. The reason for the dissimilar results could be due to the studies done in the paper being both *in vivo* and *in vitro* cardiac myocyte differentiation, whereas our study focused on *in vivo* cardiac myocyte maturation.

Overall, our results show that there is up-regulation in mitochondria and sarcomere-related genes, and down-regulation in cell cycle genes within adult mice. This is consistent with the results in O'Meara, et al. and suggests that cardiac myocytes exit the cell cycle in the adult stage. This results in a failure to reactivate proliferation-related genes after injury and contributes to the loss of cardiac regeneration. To assess this further, future studies may perform gene-knockout experiments to evaluate how the expressed phenotype changes and impacts CM regeneration.

Current regenerative strategies have looked to reactivate proliferation in adult myocytes. Accumulating evidence suggests that cytokines and growth factors trigger regeneration in other cell types such as skeletal muscle, optic nerve, and hepatocytes[2]. As a result, identifying novel cytokines and growth factors that could initiate cell cycle entry of cardiac myocytes may help to reactivate proliferation. O'Meara, et al. identified Nrg1 and OSM as two such factors, and since then, studies have linked Igf2bp3[23], CCL24[24], and MYDGF[25] as developmental regulators of postnatal CM proliferation. Igf2 is one cytokine predicted by O'Meara, et al. to be upstream regulators of the differentiated state of cardiac myocytes by Ingenuity network analysis and our results show that the Igf2 gene family, including Igf2bp3, is highly expressed in P0 samples.

Recent studies have also shed light on the mechanism of CM regeneration in neonatal mice through immune cell analysis. While CMs account for the majority of myocardial mass, other cell types contribute to transcriptional profiles and changes [1], and their distinct gene expression profiles may affect heart repair. Studies have since been conducted to analyze the roles of fibroblasts as well as the immune, conduction, and nervous system cell populations during heart regeneration. Transcriptomic analysis of CMs and immune cells from neonatal and adult mice has revealed that adult, but not neonatal, CMs and endothelial cells fail to reactivate proliferation-related genes upon cardiac injury[24]. This contributes to the loss of cardiac regeneration in the adult stage, thus transcriptional regulation and underlying regulatory mechanisms are crucial for heart regeneration. By combining comparative genomic analysis with cross-species transgenic assays, alterations of regeneration enhancers and the effects on regenerative capacity may be assessed. In doing so, we can reveal novel gene regulatory networks underlying heart regeneration. More importantly, when applied to gene therapy, this

information can be used to drive regeneration in injured mammalian hearts and improve heart repair in humans.

Based on the concept of comparing gene expression profiles across different developmental stages, studies have also been done for other types of cells to characterize the differentiation process. For example, a research study published by Lamba and Reh[23] used microarray profiling to study human retinal gene expression pattern during development by comparing genetic profile of human retina to that of human embryonic stem cell derived retina cells. Given the work done by O'Meara, et al. and Lamba and Reh, it might also be possible in the future to study the genetic expression pattern during development of human cardiac myocyte, and compare the genetic profile of human cardiac myocyte to stem cell derived myocyte. With the understanding of gene expression patterns and possible checkpoint inducers, we could potentially culture artificial cardiac myocytes to repair heart defects.

CONCLUSION

The research by O'Meara, et al. identified upstream regulators for the core cardiac myocyte regeneration network. We attempted to reproduce these results by processing mRNA-Seq data from P0 and adult mice. Despite some error between the results, we were able to generate similar gene ontology results and feel that we were successful in our reproduction. By adding data from postnatal day 4 and postnatal day 7, we were able to replicate the gene expression analysis in key regulatory events for different developmental stages with relatively high accuracy. Comparison analysis of the GO terms indicated that the higher enrichment scores are more likely to overlap with the results in the paper for both up- and down-regulated genes. These results provided us with insight into the transcriptional reversion of CM differentiation and identifies a core transcriptional signature of CM.

REFERENCES

[1] O'Meara, Caitlin C et al. “Transcriptional reversion of cardiac myocyte fate during mammalian cardiac regeneration.” Circulation research vol. 116,5 (2015): 804-15. doi:10.1161/CIRCRESAHA.116.304269

[2] Porrello ER, Mahmoud AI, Simpson E, Hill JA, Richardson JA, Olson EN, Sadek HA. Transient regenerative potential of the neonatal mouse heart. Science. 2011;331:1078–1080.

[3] NCBI Gene Expression Omnibus. “Transcriptional Reversion of Cardiac Myocyte Fate During Mammalian Cardiac Regeneration.” Boyer, Laurie A, 20 December 2014, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE64403. Accessed 1 March 2021.

[4] NCBI Gene Expression Omnibus, “ vP0_1.” Boyer, Laurie A,, 20 December 2014, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1570702. Accessed 1 March 2021.

[5] “SCC Quick Start Guide.” BU TechWeb RSS, www.bu.edu/tech/support/research/system-usage/scc-quickstart/.

[6] SRA Handbook [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2010-. Download Guide. 2009 Sep 9 [Updated 2016 Jan 14]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK242621/

[7] Babraham Bioinformatics. FastQC (Version 0.11.9) [Program documentation]. Retrieved March 3, 2021, from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[8] Cole Trapnell, Lior Pachter, Steven L. Salzberg, TopHat: discovering splice junctions with RNA-Seq, Bioinformatics, Volume 25, Issue 9, 1 May 2009, Pages 1105–1111, https://doi.org/10.1093/bioinformatics/btp120

[9] Langmead, Ben, and Steven L Salzberg. “Fast gapped-read alignment with Bowtie 2.” Nature methods vol. 9,4 357-9. 4 Mar. 2012, doi:10.1038/nmeth.1923

[10] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and

SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9. doi: 10.1093/bioinformatics/btp352. Epub 2009 Jun 8. PMID: 19505943; PMCID: PMC2723002.

[11] Boost C++ Libraries, www.boost.org/.

[12] Liguo Wang, Shengqin Wang, Wei Li, RSeQC: quality control of RNA-seq experiments, Bioinformatics, Volume 28, Issue 16, 15 August 2012, Pages 2184–2185, https://doi.org/10.1093/bioinformatics/bts356

[13] "The R Project for Statistical Computing." R, www.R-project.org/.

[14] "Python 2.0." Python.org, www.python.org/download/releases/2.0/.

[15] 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 28, 511–515 (2010). https://doi.org/10.1038/nbt.1621

[16] Trapnell, C., Hendrickson, D., Sauvageau, M. et al. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol 31, 46–53 (2013). https://doi.org/10.1038/nbt.2450

[17] Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using david bioinformatics resources. Nat Protoc. 2009; 4:44–57. [PubMed: 19131956]

[18] Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009; 37:1–13. [PubMed: 19033363]

[19] Deschamps-Francoeur G, Simoneau J, Scott MS. Handling multi-mapped reads in RNA-seq. Comput Struct Biotechnol J. 2020 Jun 12;18:1569-1576. doi: 10.1016/j.csbj.2020.06.014. PMID: 32637053; PMCID: PMC7330433.

[20] Abugessaisa I, Noguchi S, Cardon M, et al. Quality assessment of single-cell RNA sequencing data by coverage skewness analysis. bioRxiv; 2019. DOI: 10.1101/2019.12.31.890269.

[21] Sheng, Quanhu et al. “Multi-perspective quality control of Illumina RNA sequencing data analysis.” Briefings in functional genomics vol. 16,4 (2017): 194-204. doi:10.1093/bfgp/elw035

[22] Judy R Sayers, Paul R Riley, Heart regeneration: beyond new muscle and vessels, Cardiovascular Research, Volume 117, Issue 3, 1 March 2021, Pages 727–742, https://doi.org/10.1093/cvr/cvaa320

[23] Deepak A. Lamba, Thomas A. Reh; Microarray Characterization of Human Embryonic Stem Cell–Derived Retinal Cultures. Invest. Ophthalmol. Vis. Sci. 2011;52(7):4897-4906. doi: https://doi-org.ezproxy.bu.edu/10.1167/iovs.10-6504.

[24] Wang Z, Cui M, Shah AM, Ye W, Tan W, Min YL, Botten GA, Shelton JM, Liu N, Bassel-Duby R, Olson EN. Mechanistic basis of neonatal heart regeneration revealed by transcriptome and histone modification profiling. Proc Natl Acad Sci U S A. 2019 Sep 10;116(37):18455-18465. doi: 10.1073/pnas.1905824116. Epub 2019 Aug 26. PMID: 31451669; PMCID: PMC6744882.

[25] Begeman, Ian J, and Junsu Kang. “Transcriptional Programs and Regeneration Enhancers Underlying Heart Regeneration.” Journal of cardiovascular development and disease vol. 6,1 2. 22 Dec. 2018, doi:10.3390/jcdd6010002.

Appendix

A. Phred quality scores across each base position each of the FASTQ files. Outputted by fastqc package.