Introduction

In their 2015 paper, O'Meara et. al. sought to understand the transcriptional changes that mammalian cardiac myocytes (CMs) undergo as they mature and differentiate and to identify regulators of this process. In essence, if this process was well understood, it might have important medical applications and could be utilized to induce greater regeneration in the adult heart. By comparing the gene expression profiles of neonatal and adult CMs, mouse embryonic cells in the process of differentiation into CMs, and adult CMs induced to revert to a more embryonic state, the authors were able to identify a number of genes and pathways associated with differentiation and regeneration.

RNA-Seq data was generated using an Illumina Hi-Seq . Reads were aligned to the mouse reference genome using tophat and transcriptional abundance was determined using cufflinks. Results were then filtered using the calculated q-value (and there was no minimum cutoff value set for fold-change). Pearson's test was used to determine consistency between two replicate experiments and those values were high (0.92-0.96). The authors then used a number of bioinformatics techniques to analyze the DEGS, including hierarchical clustering (to visualize and cluster differences in expression between subtypes), principal component analysis (to visualize the clusters of subtypes with similar gene expression profiles) and DAVID functional annotation (to determine the most prevalent up- or downregulated function).

Data

To reproduce the results from O'Meara et. al. paper, we first retrieved their sample, GSM1570702 from NCBI GEO Series GSE64403. This sample consists of cardiac myocytes from neonatal and adult mice at different cell stages; embryonic stem cells (ESC), mesoderm (MES), cardiac progenitor (CP), and cardiac myocyte (CM). Neonatal cells were collected on postnatal days 0,4, and 7 (P0,P4,P7) and from 8-10 week old adult mice. O'Meara et. al., conducted three replicates per time point and treatment, exposure to injury, on approximately five to ten young mice. These cells were prepared for high-throughput sequencing by Illumina HiSeq 2000, through; RNA extraction, isolation and fragmentation of polyadenylated RNA, reverse transcription of the first strand. Following, double stranded DNA were synthesized along with end repair, A-tailing, adapter ligation, and size selection. Lastly, amplification and barcodes were added by PCR.

GSM1570702 short reads (sra) file was uploaded to the remote SCC server and renamed as P0_1.sra. From these short reads FASTQ files needed to be generated. The modules sratoolkit and fastqc were submitted within a script as a qsub to produce two FASTQ files. These two files were produced as a result because this data involves paired end reads. Within both files, we observed a total of 21,577,562 paired end reads with a read length of 40. Ensuring this was of best quality for both FASTQ files, we reviewed the FASTQC reports. Both reports demonstrated good per base sequence qualities (Supplemental Image 1A, 1B) and sequence qualities scores(Supplemental Image 2A, 2B). In addition no sequence was flagged as a concern based on these reports. We are able to conclude that there was no contamination or error concerns

during preparation of RNA-seq and during sequencing. Therefore, the FASTQ files produced can be passed for alignments.

Methods

All programs are operated and run through the Linux command line using BU's Shared computing cluster. Once the Fastq data of samples are obtained, various software programs were used to alignment and read mapped RNA-seq data. This method was designed with the following programs: TopHat version, RseQC, and Cufflinks. TopHat, RSeQC, and Cufflinks were used for various computational challenges in our RNA-seq alignment and quantifying gene expression.

TopHat is one of a bioinformatics tool that has been used in our experiment to align RNA-seq reads to a mouse reference genome mm9 FASTA file (mm9.fa) using short read aligner Bowtie. This programming required a number of software and parameters in order to perform the alignment. The tophat argument command module loaded Samtools-0.0.19, bowtie2,boost, tophat version was described in the run_tophat.qsub file that was created to run the argument. Running tophat in the command line provides the fundamental format of the tophat usage. TopHat argument was develop based on the description below:

```
Tophat -r 200 -G <Reference genome> --segment-length=20
--segment-mismatches=1 --no-novel-juncs <Reference genome> <read_1>
<read_2>
```

Mouse genome mm9 FASTA file used as reference genome, read_1 and read_2 are the FASTQ file of the paired end reads. The run_tophat.qsub file takes over one an hour to run.
After running this argument, it creates a new file named accepted_hits.bam which contains a list of read alignment. Using the accepted_hits.bam file, RseQC tool was used for quality control metric and to create mapping statistics plot images. RseQC run file includes the format description below:

```
geneBody_coverage.py -i accepted_hits.bam<reference genome> -oRsgenebody
Inner_distance.py -i accepted_hits.bam <reference genome> -o Rsdistance
Bam_stat.py -i accepted_hits.bam
```

The RseQC run file output brings two plot images that provide plots about transcript coverage and inner distance between paired RNA reads (Supplementary image 3 and 4). Using the accepted_hits.bam file, Cufflinks and cuffdiff tools were generated in order to analyze gene differentially expressed and transcription. Cufflinks is another tool that was used to calculate gene expression levels analysis of RNA-seq data in our experiment. After running the cufflinks qsub, we get P0_1_cufflinks/genes.fpkm_tracking file which provides the FPKM (Fragments Per Kilobase Million) of all genes (Supplementary 5). Cuffdiff is a tool used to calculate the expression levels. The cuffdiff output provides normalized gene and transcription expression levels file.

Results

Differentially expressed genes (identified through the computational methods detailed above) were further processed in order to elucidate any overall trends. First, the top 10 differentially expressed genes (based on q-value) were identified, and are presented below along with their FPKM values, log2 fold change, p-value, and q-values.

Gene ID	Gene Name	FPKM (P01)	FPKM (Adult)	log2 fold change	p-value	q-value
XLOC_000106	Plekhb2	22.56790	73.568300	1.70481	5e-05	0.00106929
XLOC_000127	Mrpl30	46.45470	133.038000	1.51794	5e-05	0.00106929
XLOC_000199	Coq10b	11.05830	53.300000	2.26901	5e-05	0.00106929
XLOC_000214	Aox1	1.18858	7.091360	2.57682	5e-05	0.00106929
XLOC_000221	Ndufb3	100.60900	265.235000	1.39851	5e-05	0.00106929
XLOC_000398	Sp100	2.13489	100.869000	5.56218	5e-05	0.00106929
XLOC_000454	Cxcr7	4.95844	32.275300	2.70247	5e-05	0.00106929
XLOC_000459	Lrrfip1	118.99700	24.640200	-2.27184	5e-05	0.00106929
XLOC_000461	Ramp1	13.20760	0.691287	-4.25594	5e-05	0.00106929
XLOC_000477	Gpc1	51.20620	185.329000	1.85570	5e-05	0.00106929

Table 1. Top 10 differentially expressed genes P01 vs. Adult

This set of 36,329 genes contained a large number results with log2 fold changes at or approximately at zero. The gene set was further filtered to remove these, using the significance attribute (yes/no) assigned by cufflinks (see Methods), resulting in exclusion of >90% of the genes initially listed. The significance value establishes a cutoff at 0.0075, as opposed to a more routine $p < 0.01$ value. This stricter cutoff removes an extra 237 genes, but because cufflinks is a software built for this application, we can rely on its determination of a reasonable cutoff value, and assume that the 237 genes we lose do not represent the most significant results, or a result that is not captured in other, more significantly differentially expressed genes.

The remaining 2139 genes were separated into upregulated and downregulated gene sets, containing 1084 and 1055 DEGs, respectively.

These two gene lists were uploaded to DAVID for functional annotation and analysis. The most significant annotations among the upregulated set were those related to mitochondria and cell metabolism, while the downregulated gene set aligned mostly to a cell cycle and cell growth profile. This makes sense in the context of comparing young, undifferentiated cells to adult cells, and the details and implications of this study will be discussed further below.

Category	Term	Count	PValue	Fold Enrichment	Bonferroni	Benjamini	FDR
GOTERM_CC_FAT	GO:0005739~mitochondrion	263	1.91E-50	2.52422528	1.32E-47	1.32E-47	1.18E-47
GOTERM_CC_FAT	GO:0044429~mitochondrial part	166	6.52E-45	3.28170247	4.52E-42	2.26E-42	2.02E-42
GOTERM_CC_FAT	GO:0005740~mitochondrial envelope	123	2.77E-32	3.2446502	1.92E-29	4.73E-30	4.23E-30
GOTERM_CC_FAT	GO:0005743~mitochondrial inner membrane	96	2.80E-32	3.97862955	1.95E-29	4.73E-30	4.23E-30
GOTERM_CC_FAT	GO:0031966~mitochondrial membrane	118	3.41E-32	3.34351814	2.37E-29	4.73E-30	4.23E-30
GOTERM_CC_FAT	GO:0019866~organelle inner membrane	98	2.32E-29	3.60927159	1.61E-26	2.68E-27	2.40E-27
GOTERM_CC_FAT	GO:0044455~mitochondrial membrane part	60	1.37E-26	5.18571377	9.50E-24	1.36E-24	1.21E-24
GOTERM_CC_FAT	GO:0098798~mitochondrial protein complex	54	1.93E-26	5.76870395	1.34E-23	1.67E-24	1.50E-24
GOTERM_CC_FAT	GO:1990204~oxidoreductase complex	43	7.62E-26	7.32246745	5.29E-23	5.87E-24	5.25E-24
GOTERM_CC_FAT	GO:0031967~organelle envelope	151	1.57E-25	2.4294592	1.09E-22	1.09E-23	9.71E-24
GOTERM_CC_FAT	GO:0031975~envelope	151	2.55E-25	2.41814887	1.77E-22	1.61E-23	1.44E-23
GOTERM_BP_FAT	GO:0006091~generation of precursor metabolites and energy	73	5.53E-27	4.41339306	3.60E-23	1.80E-23	1.71E-23
GOTERM_BP_FAT	GO:0006082~organic acid metabolic process	121	5.01E-25	2.78283251	3.26E-21	1.09E-21	1.04E-21
GOTERM_BP_FAT	GO:0043436~oxoacid metabolic process	113	2.23E-24	2.86296696	1.45E-20	3.63E-21	3.45E-21
GOTERM_BP_FAT	GO:0019752~carboxylic acid metabolic process	112	4.11E-24	2.85915535	2.67E-20	5.35E-21	5.09E-21
GOTERM_CC_FAT	GO:0098800~inner mitochondrial membrane protein complex	43	9.53E-22	5.96426784	6.61E-19	5.51E-20	4.92E-20
GOTERM_CC_FAT	GO:0070469~respiratory chain	36	5.54E-21	7.03607073	3.84E-18	2.96E-19	2.64E-19
GOTERM_BP_FAT	GO:0015980~energy derivation by oxidation of organic compounds	55	1.62E-21	4.64687513	1.05E-17	1.75E-18	1.67E-18
GOTERM_CC_FAT	GO:0005746~mitochondrial respiratory chain	33	2.18E-19	7.09470465	1.52E-16	1.08E-17	9.67E-18

Table 2. Summary of top upregulated gene functions from DAVID

Category	Term	Count	PValue	Fold Enrichment	Bonferroni	Benjamini	FDR
GOTERM_BP_FAT	GO:0007049~cell cycle	167	3.48E-30	2.536237451	2.32E-26	2.32E-26	2.14E-26
GOTERM_BP_FAT	GO:0051301~cell division	91	1.44E-26	3.592258323	9.58E-23	4.79E-23	4.43E-23
GOTERM_BP_FAT	GO:0000278~mitotic cell cycle	109	1.69E-25	3.028919267	1.12E-21	3.75E-22	3.47E-22
GOTERM_BP_FAT	GO:0022402~cell cycle process	133	3.95E-25	2.622744753	2.63E-21	6.57E-22	6.08E-22
GOTERM_BP_FAT	GO:1903047~mitotic cell cycle process	100	2.63E-24	3.115292432	1.75E-20	3.51E-21	3.24E-21
GOTERM_BP_FAT	GO:0051128~regulation of cellular component organization	197	2.22E-17	1.795857655	1.48E-13	2.46E-14	2.28E-14
GOTERM_BP_FAT	GO:0007067~mitotic nuclear division	63	1.26E-16	3.296098638	7.40E-13	1.20E-13	1.11E-13
GOTERM_BP_FAT	GO:0007399~nervous system development	187	2.05E-16	1.80874347	1.48E-12	1.71E-13	1.58E-13
GOTERM_BP_FAT	GO:0000280~nuclear division	75	6.18E-16	2.847571988	4.44E-12	4.57E-13	4.23E-13
GOTERM_BP_FAT	GO:0000280~nuclear division	75	6.18E-16	2.847571988	4.44E-12	4.57E-13	4.23E-13
GOTERM_BP_FAT	GO:0010564~regulation of cell cycle process	70	7.56E-16	2.966772676	5.18E-12	5.04E-13	4.66E-13
GOTERM_BP_FAT	GO:0051726~regulation of cell cycle	91	1.76E-14	2.386224354	1.18E-10	9.78E-12	9.05E-12
GOTERM_BP_FAT	GO:0008283~cell proliferation	158	1.63E-14	1.84384085	1.09E-10	9.78E-12	9.05E-12
GOTERM_CC_FAT	GO:0005694~chromosome	109	1.48E-14	2.161987686	1.06E-11	1.06E-11	9.58E-12
GOTERM_BP_FAT	GO:0048285~organelle fission	75	2.02E-14	2.662664716	1.35E-10	1.04E-11	9.60E-12
GOTERM_BP_FAT	GO:0048285~organelle fission	75	2.02E-14	2.662664716	1.35E-10	1.04E-11	9.60E-12
GOTERM_BP_FAT	GO:0009893~positive regulation of metabolic process	221	5.83E-14	1.608897132	3.88E-10	2.77E-11	2.57E-11
GOTERM_BP_FAT	GO:0022008~neurogenesis	141	1.32E-13	1.882526712	8.82E-10	5.51E-11	5.10E-11

Table 3. Summary of top downregulated gene functions from DAVID

Discussion

Once we determined the differentially expressed genes in our analysis of P0 vs. Ad, we were able to further analyze the results to draw biological conclusions. The O'Meara et. al. paper was able to set a framework to better understand CM repair through the changes of transcriptional expression. Our reproduction of the paper produced similar results and we were ultimately able to interpret some of the same biological conclusions.

To determine the biological meaning from our findings, we analyzed the FPKM (Fragments per kilobase of exon per million fragments) values of our significantly differentially expressed genes. Using the FPKM values of genes representing sarcomere, mitochondria, and the cell cycle, we created several plots to show the overall trends in the samples (Figures 1, 2, and 3). Our FPKM plots show positive directions in both the sarcomere and mitochondria gene samples, which is very comparable to the paper's FPKM plot trends. Furthermore, the cell cycle plot has more of a negative direction through the samples, also comparable to the paper. These comparisons lead us to conclude that as CMs mature, from P0 to Ad, there are increased sarcomere structures and metabolic processes. Though, the CMs also begin to exit the cell cycle, since there is a decrease in cell cycle genes through increased time.

As we continued our analysis, we compared our DAVID results to the O'Meara et. al. paper's analysis of up and down regulated genes, found in the online supplementary material. Specifically comparing the gene ontology (GO) terms we were able to find significant similarities between the paper's findings and ours. To compare the paper's reference tables to our DAVID tables, we added an asterisk indicating the processes they had in common. The top 20 GO terms are shown in the tables in the supplementary images (Supplementary Images 6 and 7). Although, we did perform a further comparison of the top 50 GO terms of our DAVID results (located in our GitHub Repository). Our findings showed that several upregulated GO terms in common regarded mitochondria as well as metabolic processes. The downregulated GO terms in common mostly dealt with cell cycle processes. This further suggests our finding that the cell cycle genes are downregulated in the maturing CMs.

Lastly, we created a clustered heatmap to compare our reproduction to the paper. Our heatmap (Figure 4) used the top 500 differentially expressed genes. We compared our heatmap to the patterns found in the paper's heatmap found in their Figure 2A. Our heatmap results are not the exact same as the papers. O'Meara et. al. were able to use their heatmap to find gene clusters repressed and expressed over the maturation of the CMs (O'Meara et. al.). Unfortunately, we were unable to draw the same conclusions from our own heat map creation.

All in all, we were able to somewhat reproduce the results of the O'Meara et. al. paper. Many of our findings lead us to the same biological conclusions the paper found. Though, some of our results may differ due to the methods and tools we used while reproducing the paper.

Figure 1: FPKM value plot of Sarcomere genes from samples 1 & 2

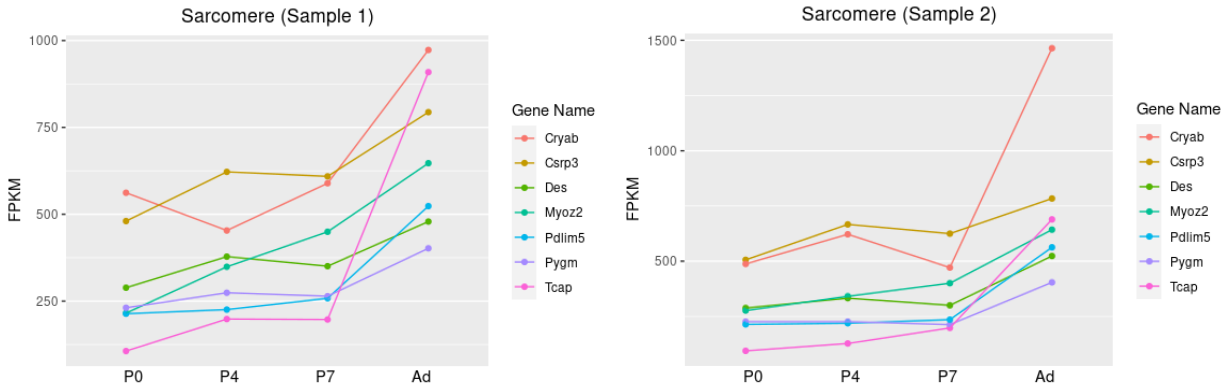


Figure 2: FPKM value plot of Mitochondria genes from samples 1 & 2

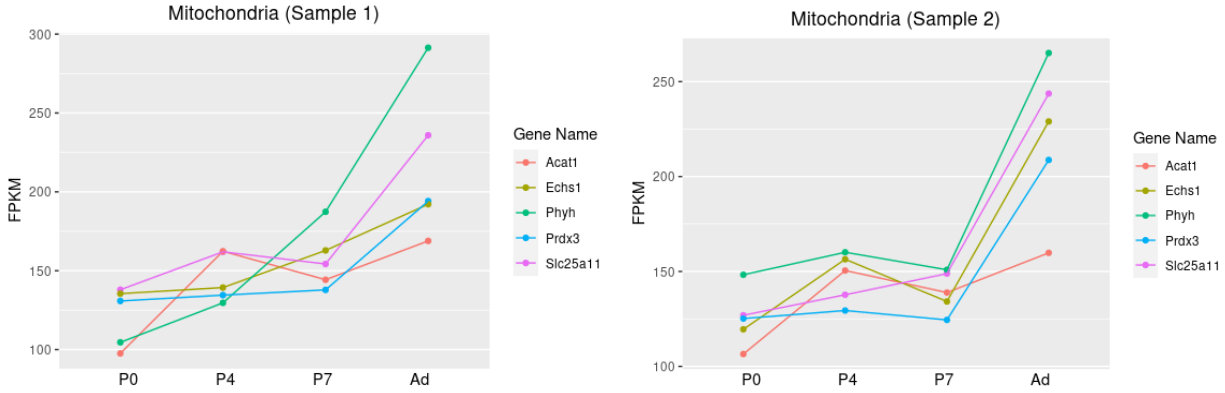


Figure 3: FPKM value plot of Cell Cycle genes from samples 1 & 2

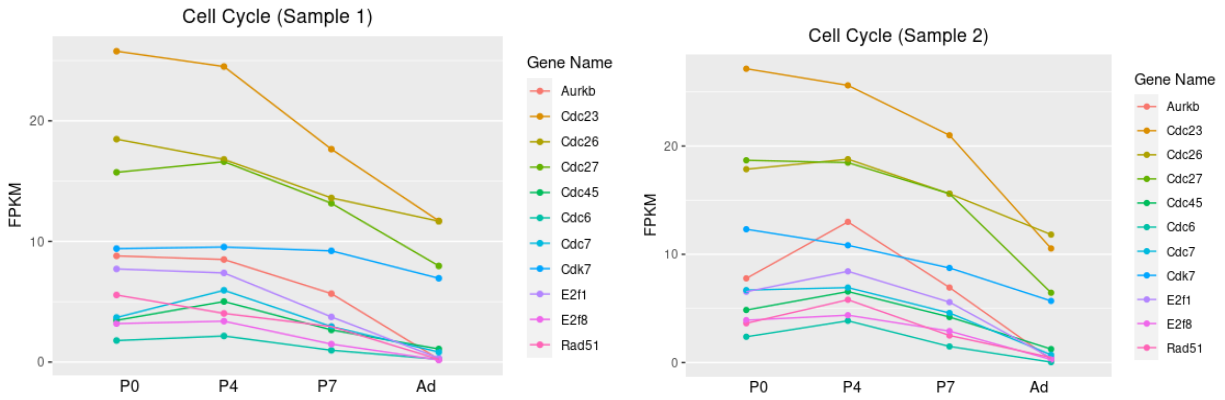
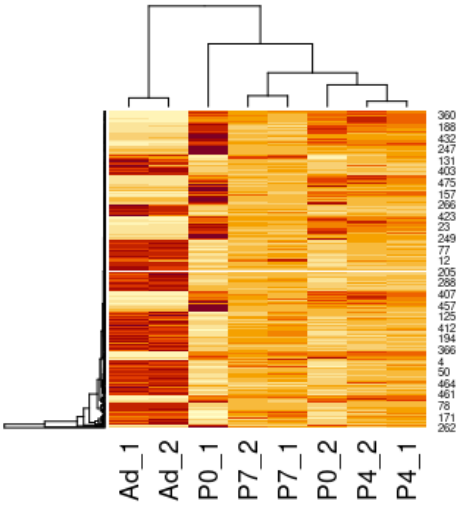


Figure 4: Clustered heatmap of FPKM values of top 500 differentially expressed genes



Conclusion

To conclude, our reproduction of the O'Meara et. al. paper shows comparable findings to explain transcriptional expression through CM maturation. The paper ultimately showed that cardiac regression is a direct transcriptional reversion of the differentiation process (O'Meara et. al.). Through our results, we found the same trends in the transcriptional differentiation process. The CM upregulated genes are incorporated with the sarcomere and mitochondria, whereas the cell cycle genes are downregulated. The significantly differentially expressed genes we found are the guiding factor to further understanding cardiac myocytes processes.

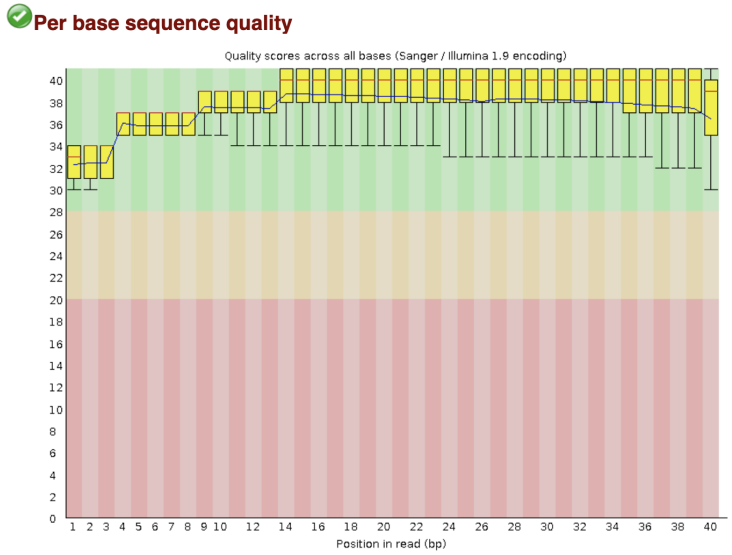
Through our reproduction of the paper, we did encounter several challenges. For example, when creating the FPKM plots, we ran into issues changing the axis titles to create an easily interpretable plot. Through several R and ggplot tutorials, along with guidance from our TA, we were able to overcome this issue. Another example encountered was during data curation. During this step the qsub used to generate the FASTQ files would unexpectedly fail due to fastq-dump command. We were able to overcome this issue by re-downloading the sample and re-running the qsub run_extract.qsub until the files were able to generate. Ultimately, we were able to work through our issues to produce our analysis.

References

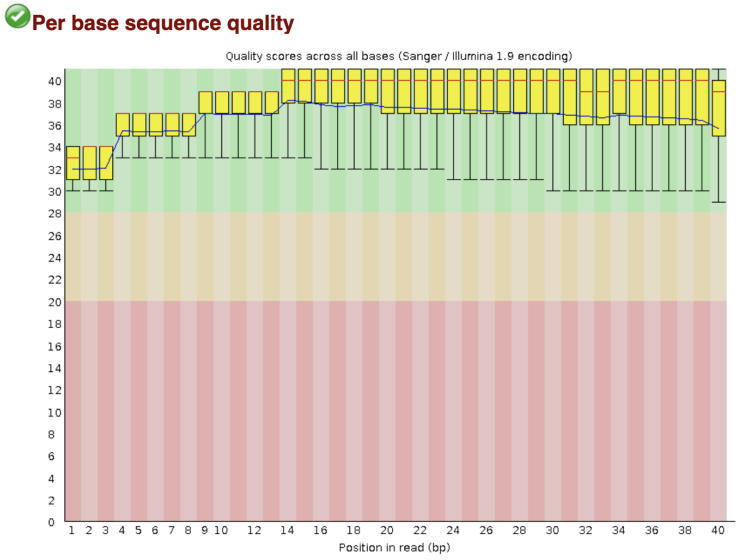
O'Meara, Caitlin C., et al. "Transcriptional Reversion of Cardiac Myocyte Fate During Mammalian Cardiac Regeneration." *Circulation Research*, vol. 116, no. 5, 2015, pp.

804–815., doi:10.1161/circresaha.116.304269.

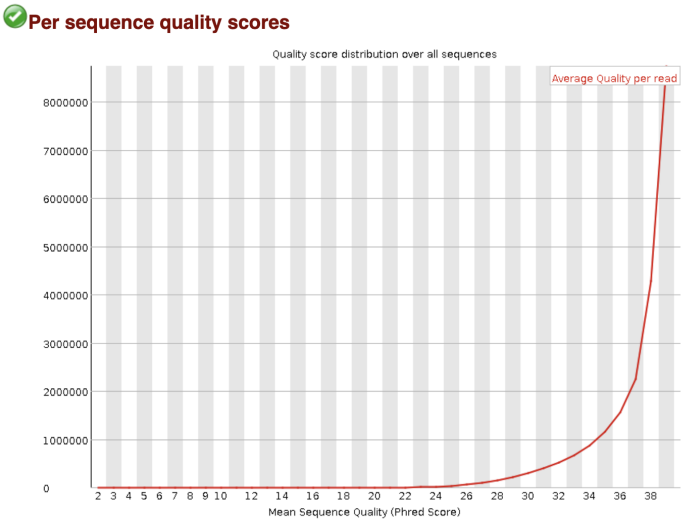
Supplementary Images



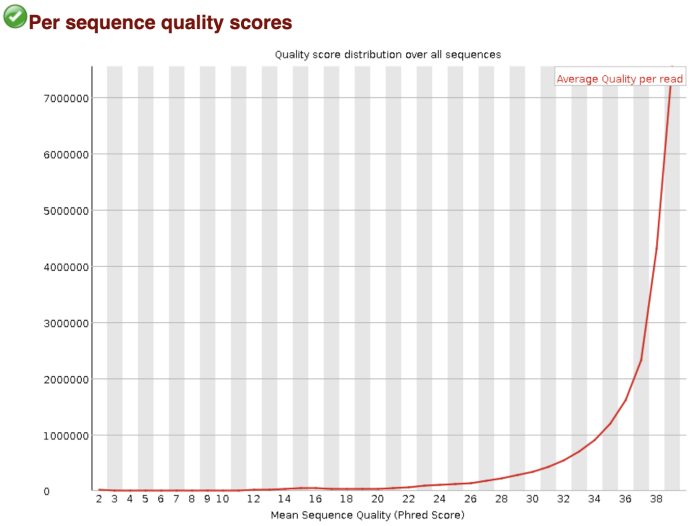
Supplementary Image 1A: FASTQC per base sequence quality reports from P0_1_1.fastq



Supplementary Image 1B: FASTQC per base sequence quality report from P0_1_2.fastq

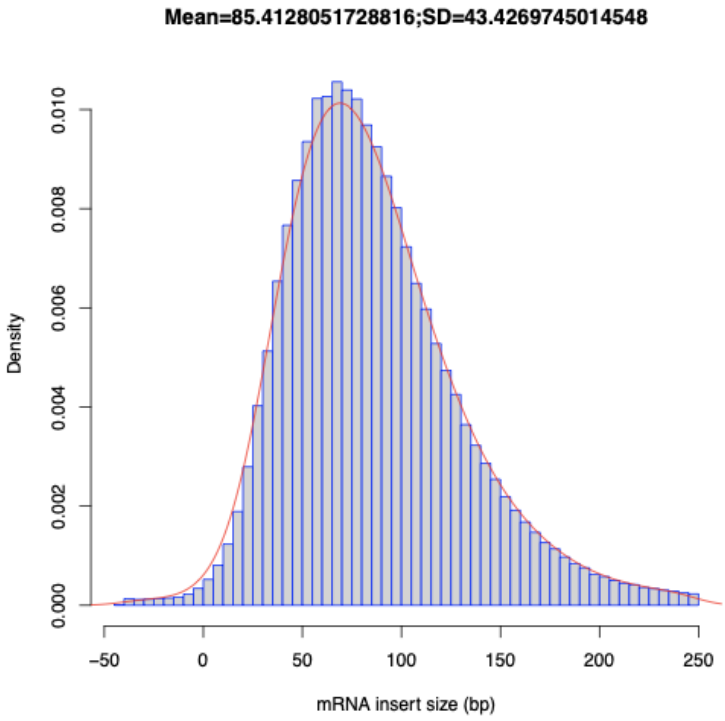


Supplementary Image 2A: FASTQC per sequence quality reports from P0_1_1.fastq

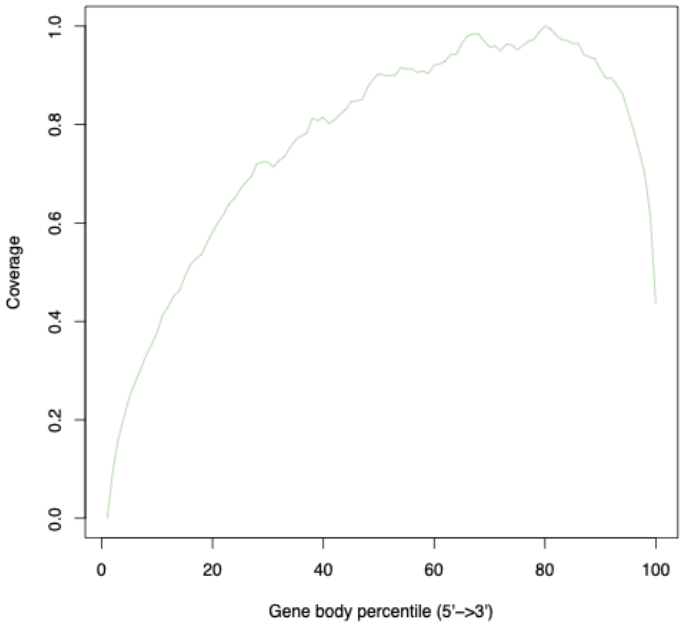


Supplementary Image 2B: FASTQC per sequence quality report from P0_1_2.fastq

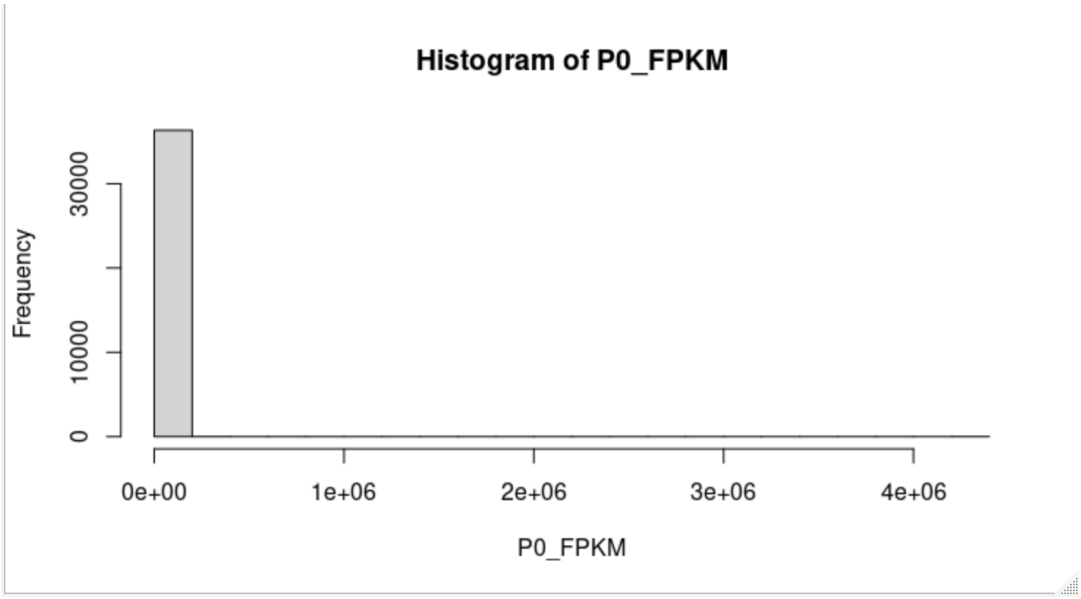
Supplementary Image 3: mRNA inner distance between two paired reads.



Supplementary Image 4: A whole Transcriptom body coverage. Transcriptome body coverage was generated using `geneBody_coverage.py` tool in RseQC.



Supplementary Image 5: Gene FPKMs (Fragments Per Kilobase Million) from P0_1_cufflinks/genes.fpkm_tracking



Supplementary Image 6:

Down Regulated DAVID GO Terms

	Term	Bonferroni	Benjamini	FDR	Comparison
1	GO:0007049~cell cycle	2.32E-26	2.32E-26	2.14E-26	*
2	GO:0051301~cell division	9.58E-23	4.79E-23	4.43E-23	*
3	GO:0000278~mitotic cell cycle	1.12E-21	3.75E-22	3.47E-22	*
4	GO:0022402~cell cycle process	2.63E-21	6.57E-22	6.08E-22	*
5	GO:1903047~mitotic cell cycle process	1.75E-20	3.51E-21	3.24E-21	NA
6	GO:0051128~regulation of cellular component organization	1.48E-13	2.46E-14	2.28E-14	NA
7	GO:0007067~mitotic nuclear division	7.4E-13	1.2E-13	1.11E-13	NA
8	GO:0007067~mitotic nuclear division	7.4E-13	1.2E-13	1.11E-13	NA
9	GO:0007399~nervous system development	1.48E-12	1.71E-13	1.58E-13	NA
10	GO:0000280~nuclear division	4.44E-12	4.57E-13	4.23E-13	*
11	GO:0000280~nuclear division	4.44E-12	4.57E-13	4.23E-13	*
12	GO:0010564~regulation of cell cycle process	5.18E-12	5.04E-13	4.66E-13	*
13	GO:0051726~regulation of cell cycle	1.18E-10	9.78E-12	9.05E-12	*
14	GO:0008283~cell proliferation	1.09E-10	9.78E-12	9.05E-12	NA
15	GO:0005694~chromosome	1.06E-11	1.06E-11	9.58E-12	*
16	GO:0048285~organelle fission	1.35E-10	1.04E-11	9.6E-12	*
17	GO:0048285~organelle fission	1.35E-10	1.04E-11	9.6E-12	*
18	GO:0009893~positive regulation of metabolic process	3.88E-10	2.77E-11	2.57E-11	NA
19	GO:0022008~neurogenesis	8.82E-10	5.51E-11	5.1E-11	NA
20	GO:0010604~positive regulation of macromolecule metabolic process	1.81E-09	1.06E-10	9.86E-11	*

Supplementary Image 7:

Up Regulated DAVID GO Terms

	Term	Bonferroni	Benjamini	FDR	Comparison
1	GO:0005739~mitochondrion	1.32E-47	1.32E-47	1.18E-47	*
2	GO:0044429~mitochondrial part	4.52E-42	2.26E-42	2.02E-42	*
3	GO:0005740~mitochondrial envelope	1.92E-29	4.73E-30	4.23E-30	*
4	GO:0005743~mitochondrial inner membrane	1.95E-29	4.73E-30	4.23E-30	*
5	GO:0031966~mitochondrial membrane	2.37E-29	4.73E-30	4.23E-30	*
6	GO:0019866~organelle inner membrane	1.61E-26	2.68E-27	2.4E-27	*
7	GO:0044455~mitochondrial membrane part	9.5E-24	1.36E-24	1.21E-24	NA
8	GO:0098798~mitochondrial protein complex	1.34E-23	1.67E-24	1.5E-24	NA
9	GO:1990204~oxidoreductase complex	5.29E-23	5.87E-24	5.25E-24	NA
10	GO:0031967~organelle envelope	1.09E-22	1.09E-23	9.71E-24	*
11	GO:0031975~envelope	1.77E-22	1.61E-23	1.44E-23	*
12	GO:0006091~generation of precursor metabolites and energy	3.6E-23	1.8E-23	1.71E-23	*
13	GO:0006082~organic acid metabolic process	3.26E-21	1.09E-21	1.04E-21	NA
14	GO:0043436~oxoacid metabolic process	1.45E-20	3.63E-21	3.45E-21	NA
15	GO:0019752~carboxylic acid metabolic process	2.67E-20	5.35E-21	5.09E-21	NA
16	GO:0098800~inner mitochondrial membrane protein complex	6.61E-19	5.51E-20	4.92E-20	NA
17	GO:0070469~respiratory chain	3.84E-18	2.96E-19	2.64E-19	*
18	GO:0015980~energy derivation by oxidation of organic compounds	1.05E-17	1.75E-18	1.67E-18	*
19	GO:0005746~mitochondrial respiratory chain	1.52E-16	1.08E-17	9.67E-18	*
20	GO:0032787~monocarboxylic acid metabolic process	3.16E-16	4.51E-17	4.29E-17	NA