Transcriptional Profile of Mammalian Cardiac Regeneration with mRNA-Seq

Group: Hedgehog
Teaching Assistant: Joey Orofino

Data Curator: Merai Dandouch
Programmer: Rojashree Jayakumar
Analyst: Dylan Beeber
Biologist: Qinrui Wu

INTRODUCTION

Neonatal mice have proliferative capacity to repair heart cells, and this ability is mostly lost after birth. For example, findings show that cardiomyocytes in neonatal mice retain maximum capacity for cell repair abilities following an injury such as tissue damage to the left ventricular apex (Porrello et al., 2011). In contrast, it has been heavily documented that invertebrate species retain this ability throughout their adult lives (Oberpriller JO and Oberpriller JC, 1974; Poss et al., 2002). Proliferation of cardiac myocytes is induced through pre-existing cardiac myocytes and not stem cells. During the regeneration phase, cardiac myocyte cells lose distinct sarcomere structures and enter the cell cycle as indicated by phosphorylated histone (H3) expression and high expression of aurora B kinase, which is revealed to participate in the reversion of cell differentiation into stem cells (Porrello et al., 2011). The stress response of cardiac myocytes post heart injury remains ambiguous and in this paper we will be reproducing findings from O'Meara et al., and finding molecular differential expression patterns through varying time periods of mouse heart ventricles from postnatal day P0, P4, P7 and 8-10 week adult mice Ad_1, Ad_2. These time periods represent myocyte cell maturation from neonatal to adult. The authors of this paper collected RNA gene transcripts and found upstream regulators such as interleukin 13 (IL 13), STAT6, STAT3 to be involved in the cell cycle state of cardiac myocytes. Moreover, O'Meara et al. findings elucidate specific signaling pathways involved in the reversion process from differentiated cells to stem cells through cardiac cell regeneration.

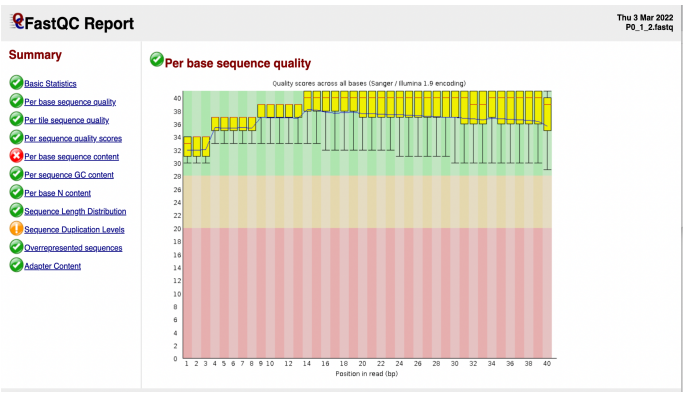
DATA

Whole heart ventricles were collected from Cr1: CD1 (CD-1) neonatal mice at P0, P4, P7, Ad_1, Ad_2 growth stages where heart ventricles were homogenized and prepared for total RNA extractions. The samples consisted of two replicates for each experimental condition. Total RNA was extracted from all samples using a phase separation Trizol method, which was followed by cDNA library preparation from the total RNA using end repair, a-tailing, adapter ligation. A paired-end 40 base pair read length sequencing was then sent for RNASeq sequencing and analysis using Illumina HiSeq 2000. Data analysis was performed on the reads using one-way ANOVA or paired t-test. We used P0_1, P0_2, Ad_1, Ad_2 samples to replicate results from O'Meara et al. The P0_1 sample was a 1.1 Gb SRA file collected from **SRR1727914** belonging to GEO Series GSE64403. The remaining files were given to us. The SRA file was converted into a FastQC file using the FastQC package from SRA tools package. The PHRED score for each base pair in the reads is between 32 - 40 as shown in report 1. The mean GC content is between 45-51% which means the GC content for these sequences have high thermal stability and low probability of folding into a secondary structure. In contrast, the duplication levels for the paired end reads is high around 50%, occurring around positions 9-1000. RNA-Seq data is dominated by high numbers of transcripts for a few genes, meaning duplication levels of 50% or higher are easily obtained from higher expressed transcripts with

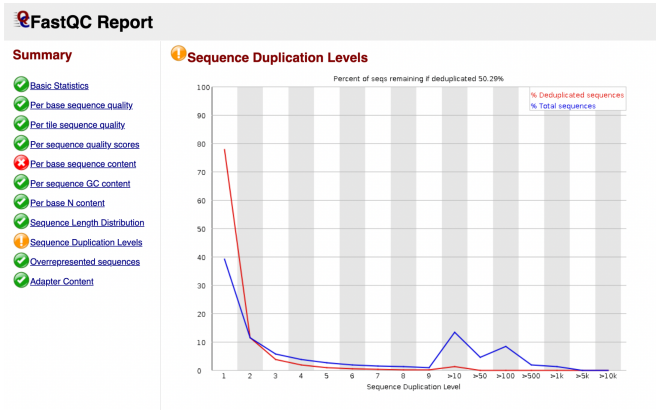
high coverage. Plots e and f in report 1 show high variability in base composition averages near the start of the sequence data which might be a result of the a-tailing end repair adapter ligation during cDNA preparation creating sequence-specific biases.

A

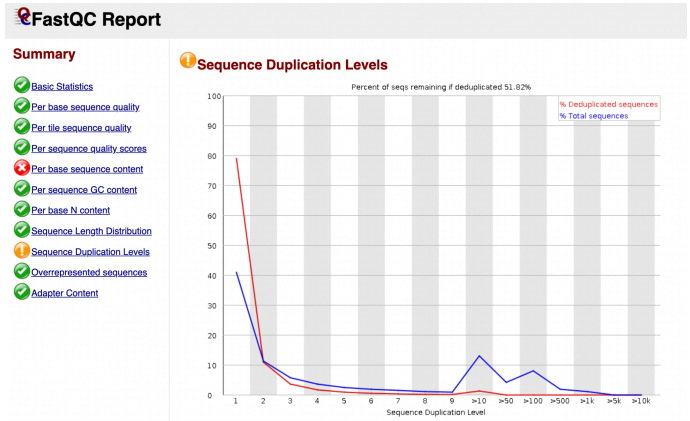
B



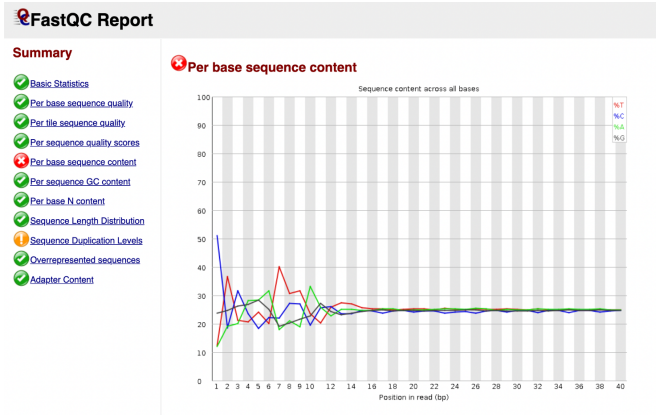
C



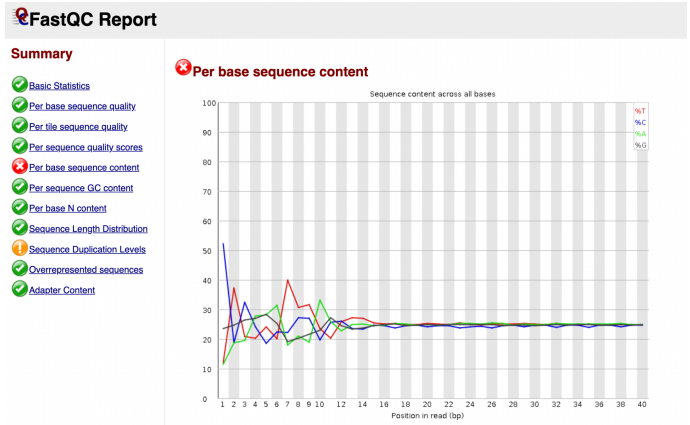
D



E



F



Report 1: FastQC Report obtained from a paired-end sequencing run displaying the quality measures (using PHRED scores), per base sequence content, and sequence duplication levels. Plots A and B show per base sequence quality starting with P0_1 on the left and P0_2 on the right. Plots C and D summarize sequence duplication events. Finally, plots E and F display per base sequence content graphs.

METHODS

Aligning and Quality assurance (QA) using tophat and RSeQC:

The sequenced Fastq reads of P0_1 sample that were extracted from SRA format, were aligned using TopHat 2.1.1 on BU's Shared Computing Cluster (SCC). TopHat is a fast splice junction mapper for RNA-seq reads and it aligns using its utilities Bowtie2 2.4.2, a short read aligner to a mammalian reference genome and here mm9 of *Mus musculus* was used (Kim D. et al., 2013). For alignment Samtools 1.10 and Boost 1.69 were loaded with TopHat as dependencies. The options for aligning the reads were formatted from both the O'Meara et al. and project description. The expected (mean) inner distance between mate pairs was set at 200 (-r 200), only one segment mismatch (--segment-mismatches=1) of length 20 (--segment-length=20) using the mm9.gtf file was allowed. Only junctions indicated in the GTF file (--no-novel-juncs) were used for alignment. The command was submitted as a batch job as alignment using TopHat has a time-complexity of nearly an hour for alignment.

The accepted_hit.bam file was created following alignment, which has the reads and alignment in binary format which was indexed using Samtools index accepted_hit.bam command. To assess the alignment quality the Samtools flagstat argument on the acceped_hits.bam file was run which gave alignment statistics. QA of the alignment was done using RSeQC 3.0.0. Python3/ 3.8.10 was also loaded to make the RSeQC utilities available. QA involved calculation of RNA-reads coverage over gene body and inner distance (insert size) of RNA-seq fragments and summarization of mapping statistics in BAM file on accepted_hits.bam file which were done using the geneBody_coverage.py, inner_distance.py and bam_stat.py modules in RSeQC respectively. geneBody_coverage.py and inner_distance.py required a bed file (mm9.bed) as additional options. QA using RSeQC was also submitted as a batch job as the time-complexity was nearly 2 hours.

Quantifying gene expression with cufflinks

Cufflinks 2.2.1 (Wang et al., 2012), a tool which takes in RNA-Seq reads and assembles them into sets of transcripts and counts the reads and differential expression by mapping to transcripts using GTF file (mm9.gtf) was loaded (Kim D. et al., 2013). The options for quantifying gene expression were formatted from both the O'Meara et al. and project description. The options to run cufflinks made sure to include only fragments which are compatible with reference transcript file, mm9.gtf (--compatible-hits) and the fragments were

corrected for fragment bias (-b) and multi-read mapping (-u). The job was submitted as a batch job and it took nearly 3 hours to complete. When cufflinks was completed, the genes.fpkm_tracking contains the quantified alignments in FPKM for all genes was loaded into Rstudio and a histogram of the relative abundance transcripts in FPKM matrix for all FPKM values greater than 0 in log scale was plotted. Since most of the reads were greater than 1, it was efficient to filter values below 0. The cuffdiff package in cufflinks was used for differential expression analysis, between the neonatal (P0_1 and P0_2) and adult (Ad_1 and Ad_2) mice RNA sequencing data. The options to run cuffdiff included correction for fragment bias (-b) and multi-read mapping (-u) with 16 threads to align reads (-p) using the accepted_hits.bam file and mm9.gtf file and it took nearly an hour. The output from cuffdiff was used for downstream analysis.

Differential Expression and Gene Clustering Analysis

Following mapping via cuffdiff, patterns of differential expression were analyzed in R version 4.1.2, utilizing the tidyverse (Wickham et al., 2019) software package. The top ten differentially expressed genes as determined by the cuffdiff q-value were extracted for further analysis. Distributions of $\log_2$ fold change values for all genes vs. for only significant genes were compared via histogram. Number and direction of significant differentially expressed genes were obtained from the cuffdiff $\log_2$ fold change values and reported cuffdiff significance threshold.

Functional annotation clustering was performed by running separate clustering jobs through DAVID (Jiao et al., 2012) for the upregulated and downregulated genes. The raw output of the DAVID functional annotation clustering were then summarized into five explanatory categories for each of the upregulated and downregulated gene sets.

FPKM Expression Matrices Analysis

To compare the trend of three selected genes differentially expressed between our study and O'Meara, we combined genes.fpkm_tracking and fpkm_matrix.csv tables using gene_short_name. The average FPKM value of replicated genes and the mean of two replicates for P0, P4, P7, Ad were calculated, and extracted genes values in Sarcomere, Mitochondria, and Cell Cycle respectively, then generated three line plots.

Further developed the table contains the top five clusters for each up and down regulated genes. Compared this table to figure 1C in paper, and annotated which enrichment terms were the same as that in paper.

An FPKM matrix with all eight samples was created for plotting a heatmap with the top 1000 genes. In order to find the top 1000 genes that were most differentially expressed between P0 and Ad, we filtered the matrix with values in the significant column and arranged the matrix in descending order by q_value. Then visualized the top 1000 differentially expressed genes over the course of in vivo maturation as heatmap, genes along rows and samples along columns.

RESULTS

Following alignment, the quality metrics of each category was assessed using Samtools flagstat and the reads that passed and failed QC were characterized. The quality of alignment was overall good, as none of the categories had any failed reads and the mapping was 100%, with 71.09% of reads properly paired and there were no duplicates.

Category	QC-passed reads	QC-failed reads
Total	49706999	0
Secondary	8317665	0
Supplementary	0	0
Duplicates	0	0
Mapped	49706999 (100%)	0
Paired in sequencing	41389334	0
Read1	20878784	0
Read2	20510550	0
Properly paired	29422646 (71.09%)	0
With itself and mate mapped	39936472	0
Singletons	1452862 (3.51%)	0
With mate mapped to a different chr	1387382	0
With mate mapped to a different chr (mapq>=5)	704916	0

Table1: The summary statistics of the alignment using TopHat. Each category in the output is broken down into QC pass and QC fail, to assess the quality of alignment.

The RSeQC package gave a plot (Fig1), which showed the gene body coverage across the entire alignment. The coverage was higher in the 3' end, indicating a bias in the coverage. However since RNA starts degrading from 5' end, this bias is expected. However, the RNA sequence quality was accepted as the overall coverage was greater than 0.8. The inner_distance.py generated a plot which showed the mean insert size as 85.41($\pm$ 43.2). The plot approximately followed a normal distribution and hence most of the reads have an insert size around the mean. Further an insert size around 80 base pairs which is usually expected in paired

end sequencing, indicating that not much information is lost. Though there are some negative insert sizes which indicate some degree of overlap, due to its low density, this quality metric is sufficient for further analysis.

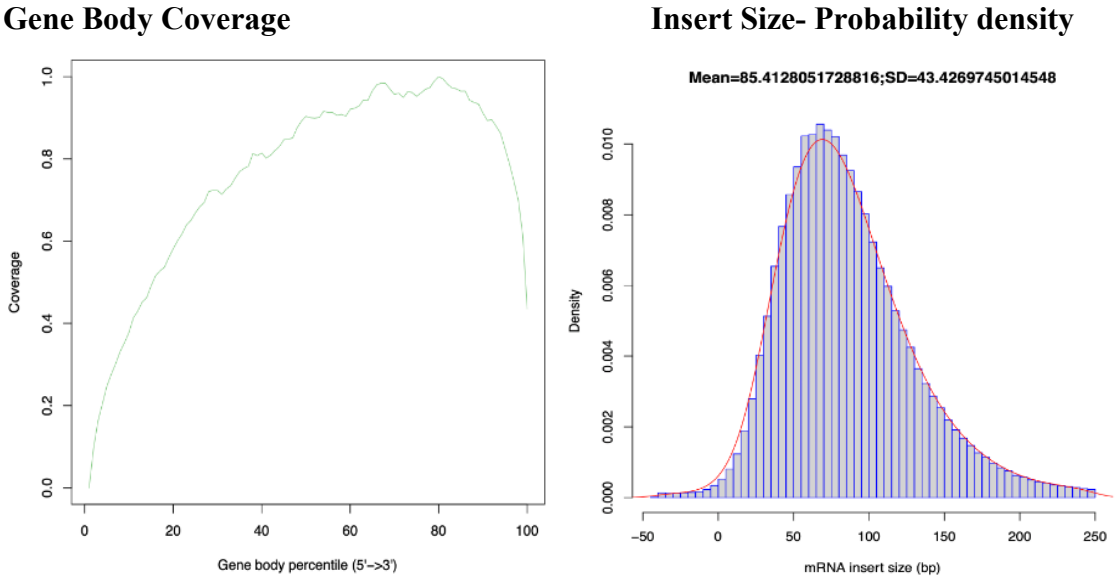


Fig1: RNA reads coverage over the gene body (left) and probability distribution of the mRNA insert size in RNA sequencing (right)

The bam_stat.py output (Table2) showed that the total number of records was 49706999 of which none of the reads failed quality control and all reads mapped to the reference indicating robustness of the RNA sequencing alignment using TopHat. It also gave the read counts mapping to the positive and negative strands and splicing junctions. The number of reads properly mapped were 27972916. The number of uniquely mapped reads was 38489380.

Category	**Read Counts**
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

Table2: Output of the Bam_stats.py.

The histogram of relative abundance of transcripts in fragments per kilobase of exon per million mapped fragments (FPKM) values for all genes in P0 sample was plotted (Fig2). The FPKM values were log 10 transformed. Before transformation, all FPKM values less than 1 were removed as the majority of the values were between 0 and 2. The number of genes prior to filtering was 37469 and after filtering genes with FPKM counts less than 1, the total number of genes was 14205.

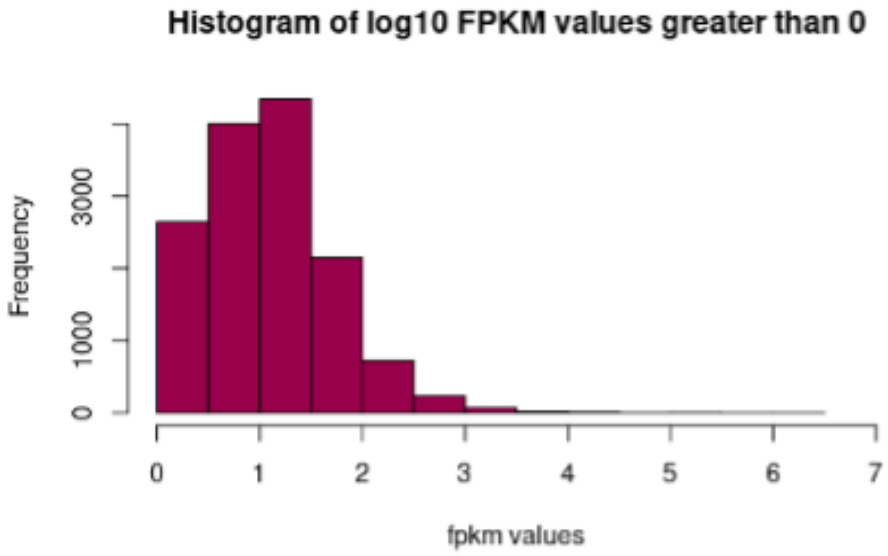


Fig 2: Histogram of relative abundance of transcripts in FPKM values for all genes in P0 sample.

The top ten differentially expressed genes provided by the cuffdiff output are listed in table 3. Rankings of differentially expressed genes were determined using the cuffdiff generated q-value, however, 666 genes in total were reported as having the same minimum q-value of 0.001069. The genes reported in table 3 are a subset of these genes with the minimum q-value, and it is important to note that there are many other differentially expressed genes that are equally significant but not included in this table. Log_2(fold change), a measurement of the magnitude and direction of differential expression is also reported in this table. In this study a positive log_2(fold change) indicates that a gene is upregulated in neonatal mice (P0) as compared to Adult (Ad) mice, whereas a negative log_2(fold change) indicates downregulation.

Gene	P0 FPKM	Ad FPKM	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 3: Contains ten of the most highly differentially expressed genes as determined by the q-value. FPKM values for each of the samples, log_2(fold change), and p-values are also reported. Note that a q-value of 0.001069 was obtained for many genes and this table does not include many other genes that are determined to be equally significant based on the q-values.

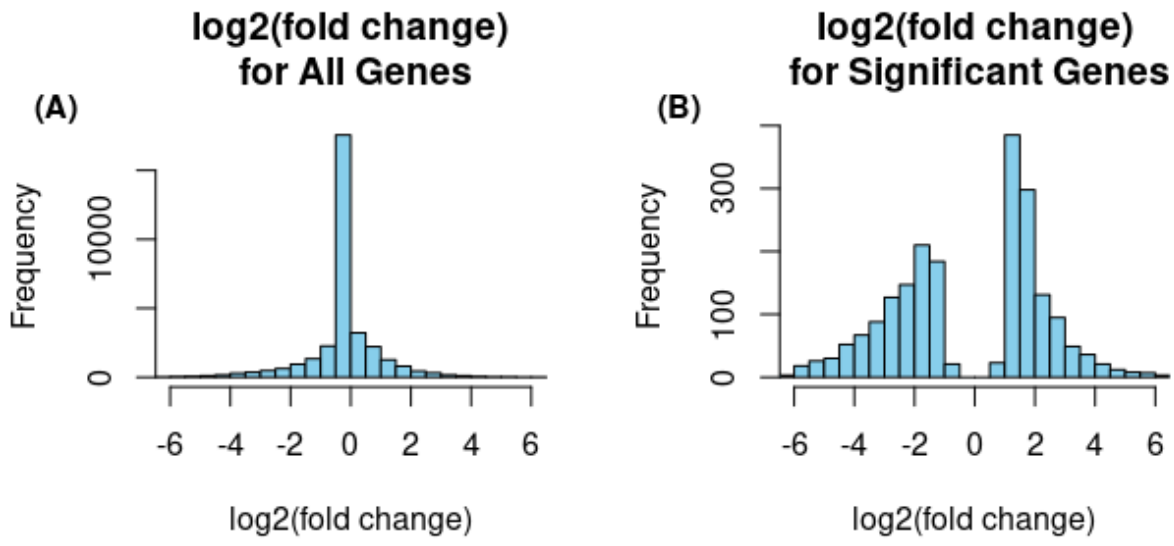


Figure 3: (A) Displays the distribution of the $\log_2$ fold changes across all genes, with no filtering performed. (B) Displays the distribution of the $\log_2$ fold changes across only genes that are determined to be significantly differentially expressed.

The distribution of $\log_2$(fold change) values in figure 3A demonstrate that a significant majority of the genes had relatively low $\log_2$(fold change) values. However, when considering only the significantly differentially expressed genes in figure 3B, we found that 1084 genes were upregulated in neonatal (P0) mice, and 1055 genes were downregulated in neonatal (P0) mice. Limiting our analysis to only significantly differentially expressed genes changes the distribution of $\log_2$(fold change) values by removing values close to zero. This relationship is expected as the absolute value of $\log_2$(fold change) directly affects the magnitude of the p-value.

Table 4 contains the results of DAVID clustering analysis for upregulated and downregulated genes that were considered significantly differentially expressed as determined by cufflinks. DAVID reported clusters as lists of individual gene ontology terms which we summarized into explanatory terms in the table below. The enrichment score (ES) provided in table 4 is equal to the $-\log_{10}$ of the average p-value of the cluster. We also checked for the presence of significant differential expression for three key genes listed in the O'Meara et al. paper: Stat3, Postn, and Il13. All three of these genes were present in the initial list of genes returned from cufflinks, however only Stat3 and Postn were found to be significantly differentially expressed. Stat3 was found to be upregulated in neonatal (P0) mice, while Postn was found to be downregulated.

ES	Upregulated	ES	Downregulated
25.4	Organic acid metabolism, including carboxylic and fatty acid	20.4	Cell cycle regulation
21.3	Cellular respiration and ATP synthesis	10.8	Post-translational peptide modification
21.1	Mitochondrial proteins	6.74	Extracellular matrix structural components
13.04	Metabolic process	6.32	High mobility group (HMG) box domain
8.44	Cellular response	5.8	Chromosome

Table 4: A summary of the top five clusters that were obtained for upregulated and downregulated genes through DAVID functional annotation. The Enrichment Score (ES) provided is equal to the $-\log_{10}$ of the average p-value of terms within the cluster.

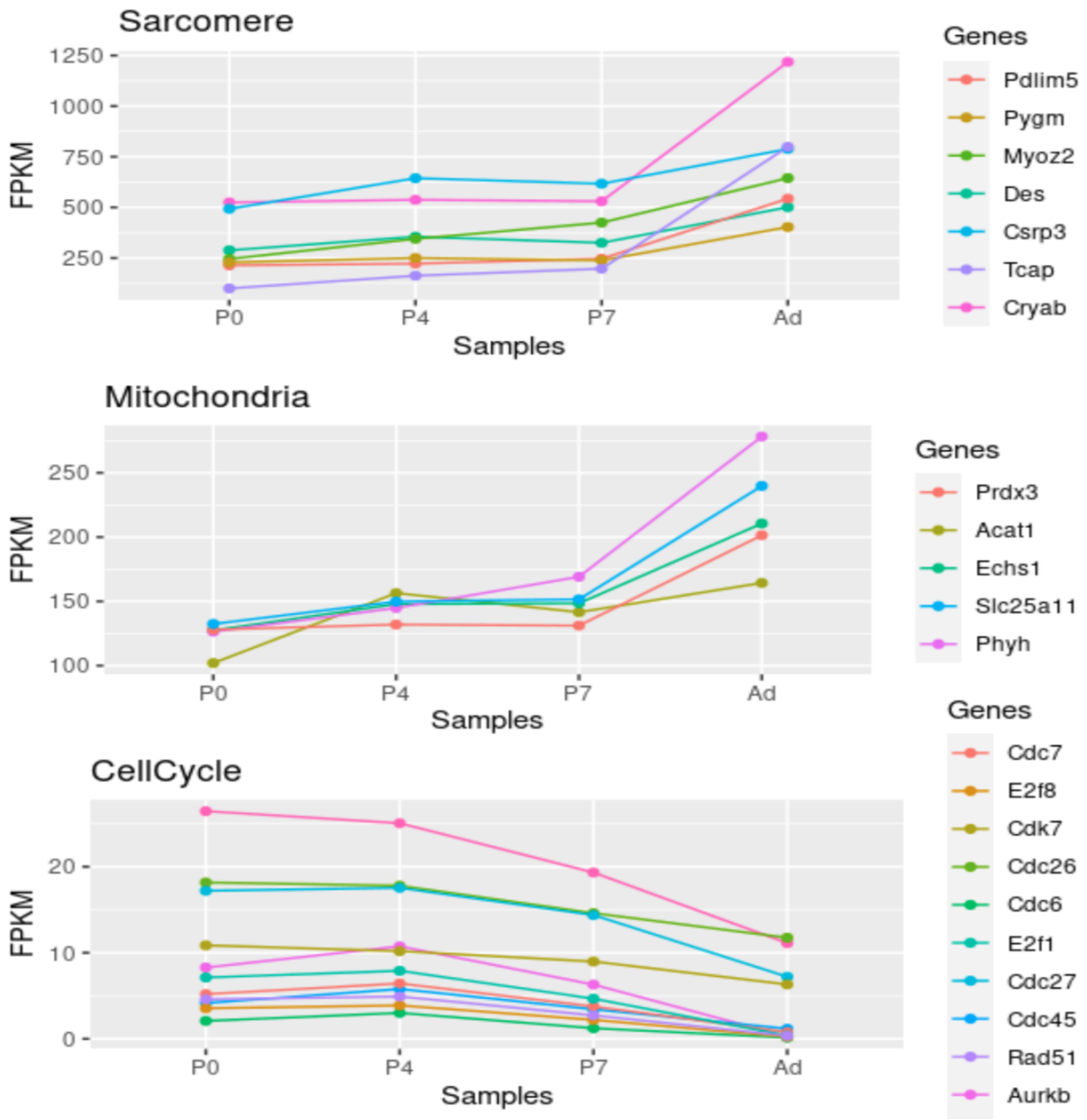


Figure 4: In vivo maturation models indicates a core transcriptional signature

There are three line plots included in figure 6 that demonstrated FPKM values of representative sarcomere, mitochondrial, and cell cycle significantly differentially expressed during in vivo maturation.

Enrichment Term	Score (Our study)	Overlapped
Upregulated		
Organic acid metabolism, including carboxylic and fatty acid	25.4	Yes
Cellular respiration and ATP synthesis	21.3	Yes
Mitochondrial proteins	21.1	Yes
Metabolic process	13.04	Yes
Cellular response	8.44	
Downregulated		
Cell cycle regulation	20.4	Yes
Post-translational peptide modification	10.8	
Extracellular matrix structural components	6.74	
High mobility group (HMG) box domain	6.32	
Chromosome	5.8	

Table 5: An annotated result that lists overlapped GO terms between our top five clusters that were obtained for upregulated and downregulated genes through DAVID analysis and the O'Meara et al. paper

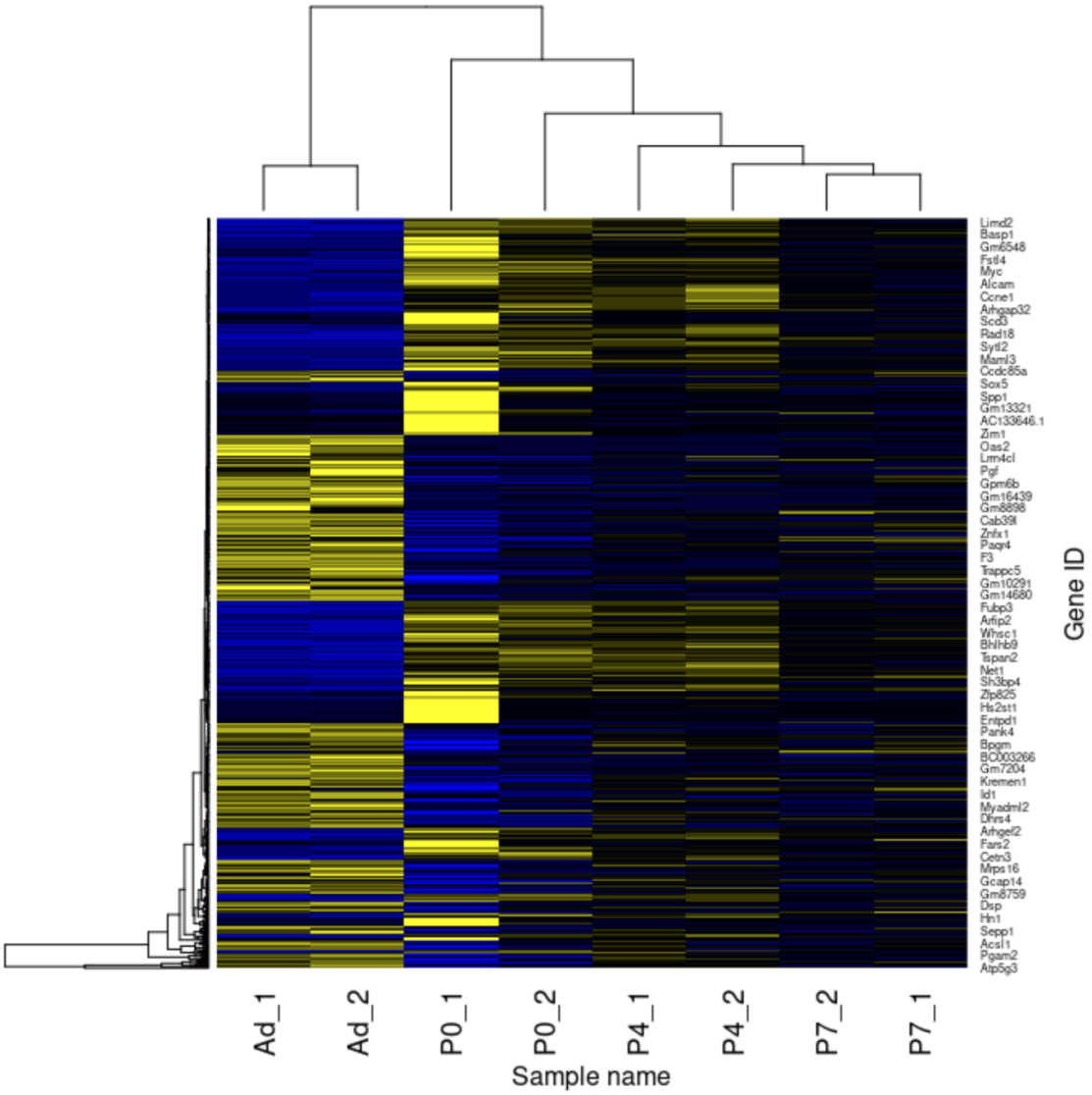


Figure 5: Hierarchical clustering of the top 1000 differentially expressed genes over the course of in vivo maturation (P0 vs Ad analysis).

DISCUSSION

In comparing the differential expression results between our study and the original O'Meara study, we find significantly fewer differentially expressed genes in our study to the number reported in figure 1B of the O'Meara paper. We found a roughly equal number of upregulated and downregulated genes, while O'Meara found many more downregulated genes than upregulated genes. This difference may be due to the selection of the threshold that we used to determine significance. Our significance threshold was a q-value less than or equal to 0.05, it is unknown what threshold the O'Meara study used to determine significance in differential expression.

Our clustering analysis of differentially expressed genes returns similar results to the analysis performed by O'Meara et al. particularly given that we started with a different list of differentially expressed genes. We found that genes related to mitochondria and metabolic processes were commonly upregulated in the neonatal mice (table 4). O'Meara found evidence of upregulation for mitochondrial genes as well, which are crucial to meeting the energy demands for cell proliferation. However, O'Meara et al. also found evidence of upregulation for sarcomere related genes. We did not find specific evidence for upregulation in these specific genes, but that may be a limitation of the methods we used to perform clustering or a consequence of our choice of significance threshold, as figure 4 shows that we found a similar relationship between the FPKM results of sarcomere genes as O'Meara.

Concerning the clustering results for downregulated genes, we found that cell cycle regulation genes were downregulated with a particularly high enrichment score. One of the key findings in the O'Meara paper was that cell cycle regulation genes were downregulated in neonatal mice. The reasoning behind this being that cell cycle regulation is most active in fully differentiated cells and likely prevents the kind of regeneration seen in neonatal mice. Additionally, our analysis reports that post-translational peptide modifications were determined to be downregulated. O'Meara et al. notes that post-translational peptide modifications are important parts of cell type commitment and differentiation, and also found that these genes tended to be downregulated.

The O'Meara et al. paper suggests that the relationship between IL13 and STAT3 may play a large role in heart regeneration of neonatal mice. Interleukin-13 (IL13), a cytokine secreted by various immune cells, has been found to play roles in response to inflammation and injury (Marone et al. 2019). Stat3 is a transcription factor and signaling molecule that participates in protecting the heart from a variety of pathologies (Harhous et al. 2019). The O'Meara paper posits Il13 and Stat3 may be regulatory elements that induce cardiac regeneration. Specifically, they suggest that Stat3 mediates the expression of periostin (Postn), which has been shown to stimulate the production of muscle cells and heal injuries (Kuhn et al. 2007). In our study, we followed up on these findings by attempting to determine whether there was a clear pattern of differential expression for these genes in P0 vs. Ad mice. We found that Il13 was not significantly differentially expressed, Stat3 was upregulated in P0 mice, and Il13 was downregulated in P0 mice. Thus no clear pattern in the expression of these genes emerged. We note that this does not directly affect the findings of the O'Meara study, as their evidence for the role that these genes play in regeneration is based on experiments not replicated by this study.

After tracking the FPKM values of sarcomere, mitochondria, and cell cycle genes (Figure 4), we observed that all three plots were consistent with the original paper. Furthermore, the Mitochondria plot lacks gene Mpc1 and the Cell Cycle plot lacks gene Bora due to no FPKM values associated with these genes. Depending on the FPKM value shown in the paper's figure 1D, the loss of data on these two genes may be due to low signal during the analysis. Based on the results of our hierarchical clustering for P0 and Ad analysis (Figure 5), we observed that genes were upregulated in P0 but downregulated in Ad. If genes appeared downregulated in P0,

they were shown to be upregulated in Ad. In addition, most genes in P4 and P7 were determined to be less upregulated than in the previous time. For example, P4 and P7 are light yellow or dark blue if genes were shown to be yellow at the P0 stage, that is, P4 and P7 have slightly less upregulation compared to P0. Regarding the comparison of overlapped GO terms between our result from David analysis and that in paper (Table 5), the majority of the top five upregulated clusters were the same as O'Meara, however, downregulated only has few enrichment terms that match to paper. The GO terms we compared to were common up and down regulated gene enrichment terms that were common in both in vitro differentiation and in vivo maturation models' datasets. That might be one of the possibilities that made the overlapped differences.

CONCLUSION

Overall, in our study, we were able to reproduce several of the results of the O'Meara paper using data from the in vivo maturation model. One of the main differences between our results was the number of differentially expressed genes we found, likely due to differing significance thresholds. This then affected downstream analysis, as the list of differentially expressed genes was later used for clustering analysis. Other differences may be due to limitations of the dataset that we used. O'Meara used both in vitro differentiation and in vivo maturation models to perform analysis, and some results were generated based on common genes in both datasets, however, all of our analyses only used the in vivo maturation dataset. Based on the visualizations of the content of FPKM expression matrices, we can conclude that sarcomere and mitochondria were indicated as upregulated and cell cycle was downregulated which were highly correlated with the O'Meara et al. study.

REFERENCES

Harhous Z, Booz GW, Ovize M, Bidaux G, Kurdi M. An Update on the Multifaceted Roles of STAT3 in the Heart. Frontiers in Cardiovascular Medicine. 2019; 6:150.

Jiao X, Sherman BT, Stephens R, Baseler MW, Lane HC, Lempicki RA. DAVID-WS: a stateful web service to facilitate gene/protein list analysis. Bioinformatics. 2012; 28(13):1805-1806.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. . Genome Biology 2013, 14:R36

Kuhn B, del Monte F, Hajjar RJ, Chang YS, Lebeche D, Arab S, Keating MT. Periostin induces proliferation of differentiated cardiac myocytes and promotes cardiac repair. Nature medicine. 2007; 13:962–969.

Marone G, Granata F, Pucino V, Pecoraro A, Heffler E, Loffredo S, Scadding GW and Varricchi G. The Intriguing Role of Interleukin 13 in the Pathophysiology of Asthma. Front. Pharmacol. 2019; 10:1387.

Oberpriller JO, Oberpriller JC. Response of the adult newt ventricle to injury. J Exp Zool. 1974;187:249–253. [PubMed: 4813417]

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. [PubMed:21350179]

Poss KD, Wilson LG, Keating MT. Heart regeneration in zebrafish. Science. 2002; 298:2188–2190.[PubMed: 12481136]

Wang L, Wang S, Li W. RSeQC: quality control of RNA-seq experiments. Bioinformatics. 2012 Aug 15;28(16):2184-5. doi: 10.1093/bioinformatics/bts356. Epub 2012 Jun 27. PMID: 22743226.

Wickham H et al. Welcome to the tidyverse. Journal of Open Source Software. 2019;4(43):1686.