Transcriptional Profile of Mammalian Cardiac Regeneration with mRNA-Seq: Analyst and Biologist

Elysha Sameth
Group Van Gogh
Github

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 injury to postnatal hearts form scar tissue that can impede proper functioning[1]. Identifying the mechanism of cell cycle activity in myocytes during this process is therefore important in understanding the molecular hurdles that prevent regeneration in the adult heart[2].

To understand the transcriptional changes that accompany myocyte regenerative response to injury, O'Meara et al[1] analyzed gene expression patterns of mouse hearts at different stages of postnatal development (P0, P4, P7, and Ad). By examining genes that were differentially expressed over the course of CM differentiation in neonatal mice and loss of differentiation in adult mice, key regulators and mediators were identified. In doing so, it was concluded that cardiac regeneration is a regulated process in which transcriptional reversion of the differentiation process occurs.

This 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. In particular, the analytical and biological aspects of the study were focused on. By identifying differentially expressed genes across samples and performing functional annotation clustering of genes, insight was gained on the processes involved in myocyte regeneration.

DATA

Within O'Meara et al, the RNAseq data for the differentiation of mouse embryonic stem cells into cardiac myocytes was obtained from Wamstad et al[3]. Heart ventricles from mice for each replicate were excised and processed for RNA sequencing, with raw data from the replicates at each time point (ESC, MES, CP and CM) processed according to the Invitrogen protocol. This included total RNA extraction from all samples, isolation and purification of

polyadenylated RNA, poly-A RNA fragmentation, and double-stranded DNA synthesis. End repair, A-tailing, adaptor ligation and size selection was then done using SPRI-Works System (Beckman Coulter). Paired-end 40 base pair read length sequencing was then performed on an Illumina HiSeq 2000, however for the purified neonatal CM samples, the TrueSeq (Invitrogen) sample preparation protocol was performed because of low RNA yield.

Within this study, postnatal day 0 (P0) data from O'Meara et al (GSE64403) processed in a former analysis was used. Previously, sequenced paired-end reads from the P0_1 sample were aligned to the mouse reference genome mm9 using *TopHat*[4], and assessed for quality using *RseQC*[5]. Gene expression levels for the replicate were then quantified using *Cufflinks*[6] in fragments per kilobase per million (FPKM), and compared to the remaining pre-prepared samples (P0_2, Ad_1, and Ad_2) using the *Cuffdiff*[7] package to determine differentially expressed genes used in this analysis.

METHODS

Gene Expression Analysis

Given the previously performed gene-level differential expression from *Cuffdiff* comparing postnatal day 0 (P0) and adult (Ad) samples, the *gene_exp.diff* output was analyzed using *R*[8]. This contains results of the differences in the summed FPKM of transcripts sharing each gene ID. Using genes with test status "OK" (successful), the top 10 differentially expressed genes were identified using the FDR-adjusted p-value (q-value). The distribution of the $\log_2$FC of all genes were then visualized to understand the differences in gene expression levels, and was compared to the distribution of genes considered significant by *Cuffdiff*. This significance was determined by whether p-value is greater than the FDR after Benjamini-Hochberg correction for multiple-testing[9].

To determine up- and down-regulated genes, genes with a successful test status and considered significant according to *Cuffdiff* were subsetted according to $\log_2$FC. Those with a $\log_2$FC greater than 0 were considered up-regulated while those lower than 0 were down-regulated. The genes of each category were outputted to files and then further analyzed for gene set enrichment.

Gene Set Enrichment Analysis

Using DAVID Functional Annotation Clustering[10][11], each set of up- and down-regulated genes were organized into functionally related clusters. To observe information about the biological process (BP), molecular function (MF), and cellular component (CC) of the genes, Gene Ontology (GO) terms used for the analysis were GOTERM_BP_FAT, GOTERM_MF_FAT, and GOTERM_CC_FAT. The top 5 clusters with the highest enrichment scores ($-\log_{10}$(average p-value)) were then summarized, and resulting terms were compared to O'Meara et al.

FPKM Expression Matrix

To determine biological patterns and validate results, the contents of the FPKM expression matrices from *Cufflinks* was further analyzed as according to O'Meara et al. Figure 1D of the original study, containing line graphs of genes specific to the most prominent GO terms (Sarcomere, Mitochondria, and Cell Cycle) discovered in the analysis, was replicated by extracting the associated genes in the FPKM matrices for the replicates. For each replicate of a sample, the FPKM values for the gene were averaged. The observed FPKM values within this study were then plotted and compared to the results of O'Meara et al.

Finally, a FPKM matrix of all replicates from the P0, P4, P7, and Ad samples was created. However, there were 19 duplicated genes in P0_1 due to overlapping splicing transcripts mapping to the same region. To account for this, these genes were removed from further analysis. The top 1000 differentially expressed genes, according to q-value, between P0 and Ad from this filtered dataset were then used for hierarchical clustering using *heatmap.2()* to see how well the samples group together while visualizing differences in gene expressions that could define these clusters.

RESULTS

Differentially Expressed Genes Associated with Myocyte Differentiation

The test status from *Cuffdiff* of each gene showed that of the 36329 genes, only 14243 had sufficient number of alignments and fragments to be considered successful (Table 1). There were a high number of genes with insufficient number of alignments for testing and few with too many fragments in the locus, however none that were considered too complex or shallowly sequenced. This makes sense as the results from samtools-1.10[12] and *RseQC* suggest that 71% of the reads were properly paired, with 77.43% uniquely mapped, 16.7% aligned equally well at more than one location, and 3.51% being singletons. Since there are reads presented exactly

once, this may represent genes considered NOTEST. Similarly, the HIDATA genes may be indicative of the multi-mapped reads as these are genes with transcripts/regions that have > 1 million reads mapped to them.

Test Status	**Number of Genes**
HIDATA	7
NOTEST	22079
OK	14243
Total: 36329	

Table 1. Summary of the number of genes for each test status from *Cuffdiff*. HIDATA represents that there are too many fragments in the locus, NOTEST represents that there are not enough alignments for testing, and OK represents a successful test.

An analysis of the genes with successful tests show that the top differentially expressed genes between P0 and Ad according to q-value have the same p- (0.0005) and q-values (0.00106929). This suggests that the top genes are equally statistically significant, which may be due to the limited number of genes being tested in the permutation test. It was found that 666 genes share this same significance, with Gm2078,Mir1895 being the most up-regulated gene and Tnni1 being the most down-regulated (Table 2).

Gene	**P0 FPKM**	**Ad FPKM**	**$\log_2$FC**	**p-value**	**q-value**
Gm2078,Mir1895	2.840640	370.10700	7.02558	0.00005	0.00106929
Tuba8	1.007430	81.60070	6.33983	0.00005	0.00106929
Rpl3l	4.65505	295.65900	5.98899	0.00005	0.00106929
Slc38a3	0.626069	36.18870	5.85308	0.00005	0.00106929
Csdc2	1.07584	53.65010	5.64004	0.00005	0.00106929
Xirp2	2.921580	140.60700	5.58877	0.00005	0.00106929
Sp100	2.13489	100.86900	5.56218	0.00005	0.00106929
Trim72	6.936190	323.06800	5.54155	0.00005	0.00106929

Bdh1	2.64365	111.63200	5.40007	0.00005	0.00106929
Rxrg	1.331430	48.64330	5.19119	0.00005	0.00106929
Tnni1	1440.20000	1.5043700	-9.90290	0.00005	0.00106929
Xist	41.32750	0.2516690	-7.35943	0.00005	0.00106929
Ptn	136.81900	0.9123610	-7.22844	0.00005	0.00106929
H19,Mir675	1556.82000	11.3891000	-7.09481	0.00005	0.00106929
Myl4	248.46600	2.6424300	-6.55504	0.00005	0.00106929
Ncrna00085	48.17380	0.6131420	-6.29588	0.00005	0.00106929
Fbn2	18.48180	0.3133800	-5.88205	0.00005	0.00106929
Top2a	28.44900	0.5575220	-5.67321	0.00005	0.00106929
Ncam1	126.64000	2.5708000	-5.62237	0.00005	0.00106929
Col12a1	5.39827	0.1138850	-5.56684	0.00005	0.00106929

Table 2. Top ten differentially expressed genes between postnatal day 0 (P0) and adult (Ad) mice outputted from *Cuffdiff* according to q-value and p-value. In yellow are the top up-regulated genes according to $\log_2 FC > 0$ and in blue are down-regulated genes according to $\log_2 FC < 0$.

Of the genes considered to be successful, 924 were significant as according to *Cuffdiff*. Distributions of the $\log_2 FC$ show that before filtering, a majority of genes are not differentially expressed between the P0 and Ad samples or have an insignificant difference in expressions, i.e. the $\log_2 FC \cong 0$ (Figure 1A). However, when filtering for significant genes there is a clear removal of genes with $\log_2 FC$ of 0 and the distribution becomes bimodal rather than normal (Figure 1B). This represents the distributions of up-regulated ($\log_2 FC > 0$) and down-regulated genes ($\log_2 FC < 0$), where we see a majority of genes having approximately a two to four fold ($\log_2 FC$ between 1 and 2) difference in expression.

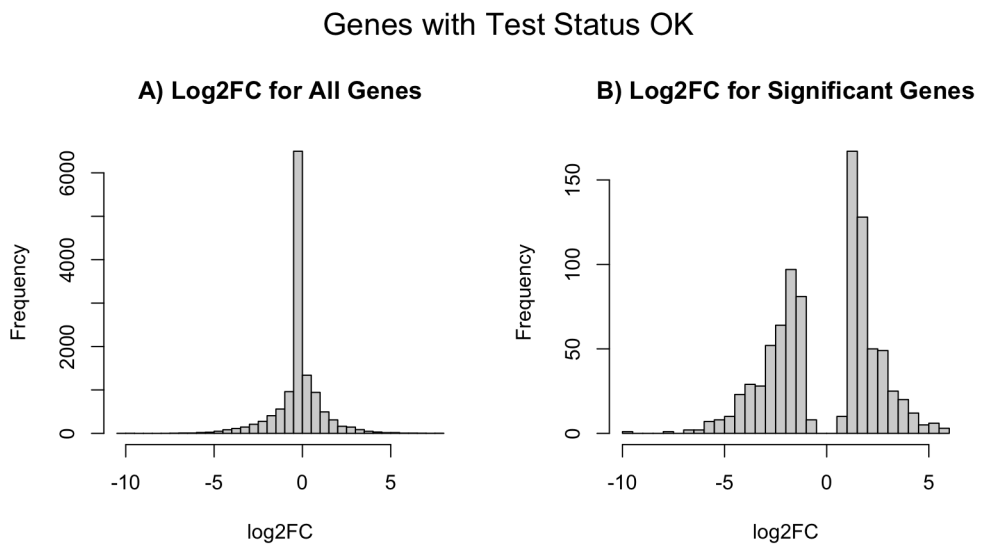

Figure 1. Histogram of the **A)** $\log_2$FC for all of the genes with a successful test status vs. **B)** genes with a successful test status and deemed significant by *Cuffdiff*. The x-axis represents the $\log_2$FC scale while the y-axis represents the frequency.

Of the genes considered successful and significant, 483 (52.27%) were identified as down-regulated while 441 (47.72%) were up-regulated (Table 3), which significantly differs from the original paper's report of 2409 and 7570 genes respectively. This may be due to differences in filter methods, in which this analysis used a more stringent filter dependent on p-value and FDR. However overall, we can confirm that genes considered significant are differentially expressed with clear separation of up- and down-regulated genes by $\log_2$FC.

Category	**# Genes (All)**	**# Genes (OK)**	**# Genes (OK, Significant)**	**# Genes (OK, p<0.01)**
Down-Regulated	11898	4886	483	498
Not DE	14450	5244	0	0
Up-Regulated	9981	4113	441	518
Total	36329	14243	924	1016

Table 3. Comparison and summary of the number of down-regulated, un-differentially expressed, and up-regulated genes for each subset: all genes, genes with a successful test status, genes with a successful test status and considered significant according to *Cuffdiff*, and genes with a successful test and p-value < 0.01.

Gene Ontology Enrichment

Despite differences in differentially expressed gene numbers, gene set enrichment analysis was consistent with the results of O'Meara et al. Top Gene Ontology (GO) terms (Figure 2) showed biological processes, molecular functions, and cellular components associated with mitochondrial and respiration/metabolism up-regulation, and down-regulation of cell cycle genes, which were all highly enriched in the original study. Furthermore, all terms in the top five clusters overlapped with O'Meara et al (Table 4), as well as the top enrichment terms of their study with ours, however with a lower enrichment score. An observation of the enrichment scores of the two groups suggest that the up-regulated genes are better clustered than the down-regulated genes based on GO analysis, which may be due to the down-regulated genes being significantly enriched in the CM (cardiomyocyte) vs ESC (embryonic stem cell) dataset rather than the P0 vs Ad. However, such down-regulated terms like cardiovascular system development and cell proliferation are expected since we observe an inability to regenerate myocytes in adult mice.

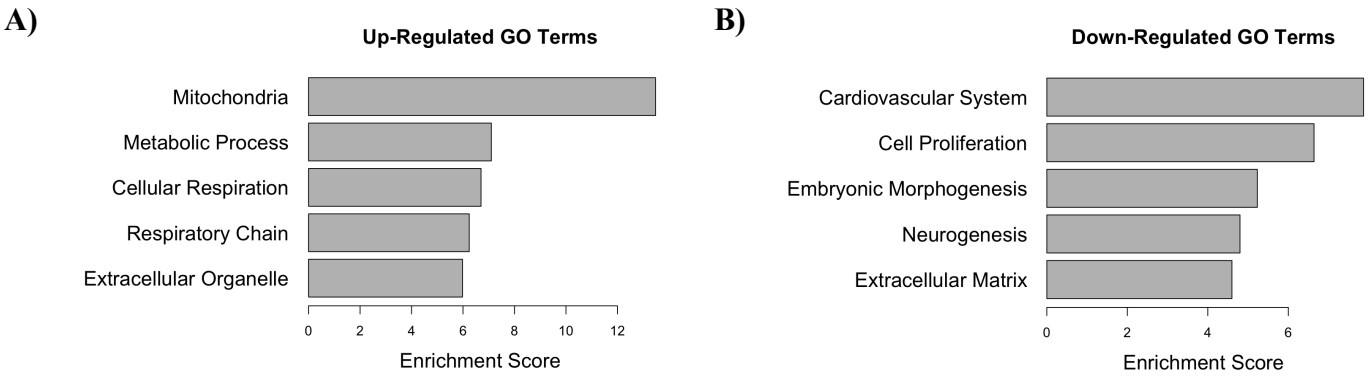

Figure 2. The top 5 Gene Ontology clusters (y-axis) according to enrichment score (x-axis) of the most significant genes from postnatal P0 and adult datasets as determined by DAVID for **A)** genes that are up-regulated and **B)** genes that are down-regulated.

A) DAVID Gene Set Enrichment of Up-Regulated Genes			
Cluster	**Score**	**Gene Ontology Terms**	**Overlap**
1	13.48	Mitochondrial part Mitochondrion Mitochondrial inner membrane	*

2	7.1	Organic acid metabolic process Oxoacid metabolic process Carboxylic acid metabolic process	*
3	6.7	Energy derivation by oxidation of organic compounds Cellular respiration Respiratory chain	*
4	6.24	Oxidoreductase complex Inner mitochondrial membrane protein complex Mitochondrial protein complex	*
5	5.98	Generation of precursor metabolites and energy Purine ribonucleoside monophosphate metabolic process Purine nucleoside monophosphate metabolic process	*

B) DAVID Gene Set Enrichment of Down-Regulated Genes			
Cluster	**Score**	**Gene Ontology Terms**	**Overlap**
1	7.87	Cardiovascular system development Circulatory system development Blood vessel development	*
2	6.64	Cell proliferation Regulation of cell proliferation Negative regulation of cell proliferation	*
3	5.23	Tube development Tube morphogenesis Epithelium development	*
4	4.8	Neurogenesis Generation of neurons Cell development	*
5	4.6	Extracellular matrix organization Extracellular structure organization Proteinaceous extracellular matrix	*

Table 4. Top Gene Ontology clusters of **A)** up-regulated genes and **B)** down-regulated genes from DAVID with terms overlapping with O'Meara et al denoted with an asterisk

Gene Trends

A comparison of FPKM values of the genes within the top clusters (Sarcoplasm, Mitochondria, and Cell Cycle) seen in O'Meara et al suggest that there are consistencies in

direction and magnitude of effect across samples. The regulation of the sarcomere and mitochondria genes were significantly increasing between the postnatal P0 to the adult phase while the genes in the cell cycle decreased (Figure 3). This is consistent with the observed trends in Figure 1D of the original paper since it was found that the former were up-regulated while the latter was down-regulated. However, it can be seen that Mpc1 (Mitochondria) and Bora (Cell Cycle) have no FPKM values associated with them, which may be due to low signal resulting in these genes being filtered. Overall, the results are consistent with O'Meara et al and is sound as expression of cell cycle genes decreases over the course of cell maturation and development.

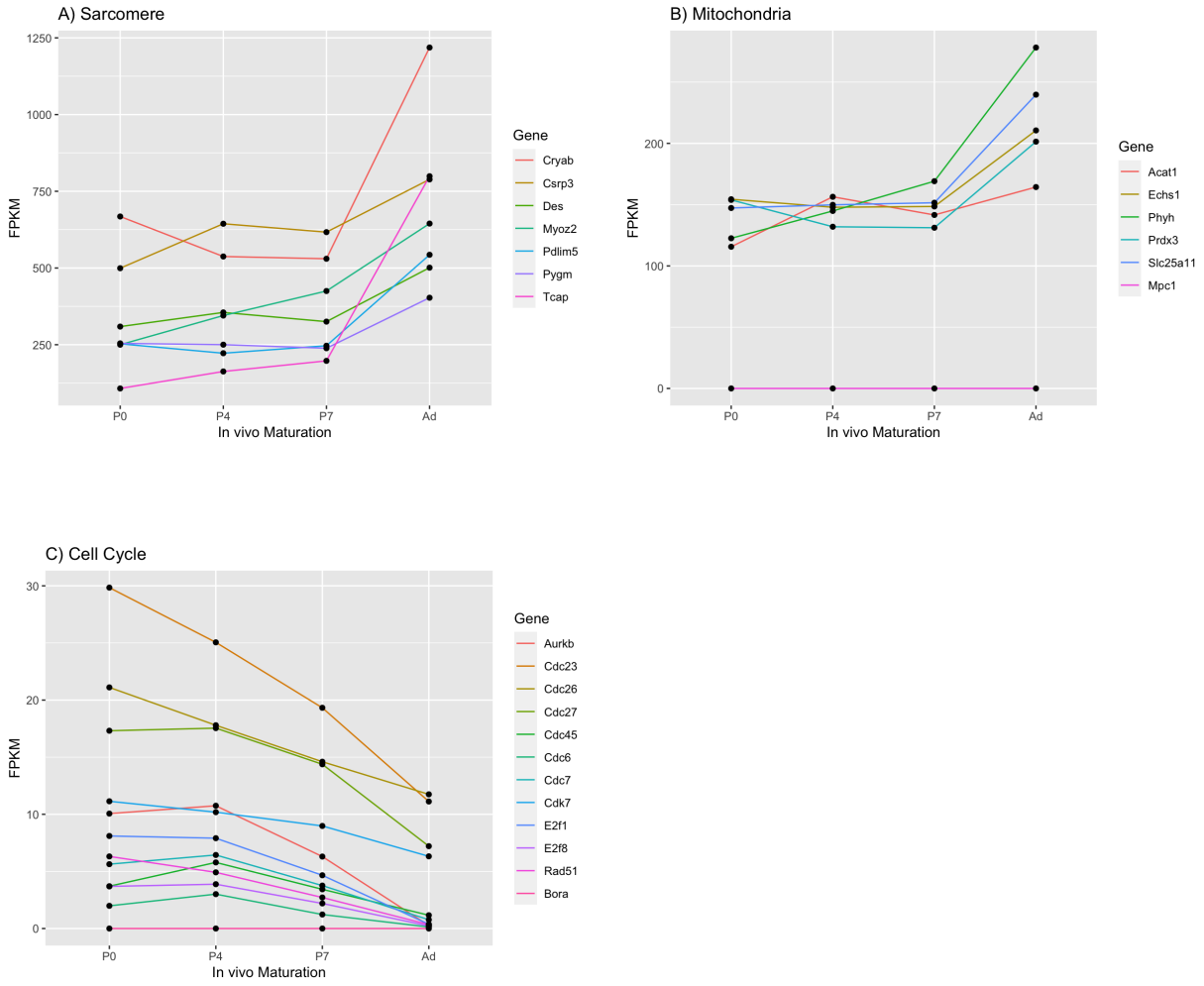

Figure 3. FPKM expression values for **A)** Sarcomere, **B)** Mitochondria and **C)** Cell Cycle genes at different maturation time stages. The x-axis represents the time stages while the y-axis represents FPKM values.

Hierarchical Clustering

The hierarchical clustering according to the top 1000 differentially expressed genes between P0_1 and Ad show clear gene expression signatures that define each sample. It can be observed that the genes of P0_1, Ad_1, and Ad_2 replicates have increased expression while the others have lower expression (Figure 4). However, due to this, P0_1 was incorrectly clustered with the Ad replicates while its counterpart replicate was clustered with the P4 samples. This could be due to batch effects, poor sample quality, or contamination in the P0 sample. Despite this, the P4, P7 and Ad samples correctly cluster with each other and are consistent with the results of O'Meara et al. Averaging the FPKM values for the replicates of each sample may produce a more accurate clustering with P0 expected to be more closely related to the P4 sample.

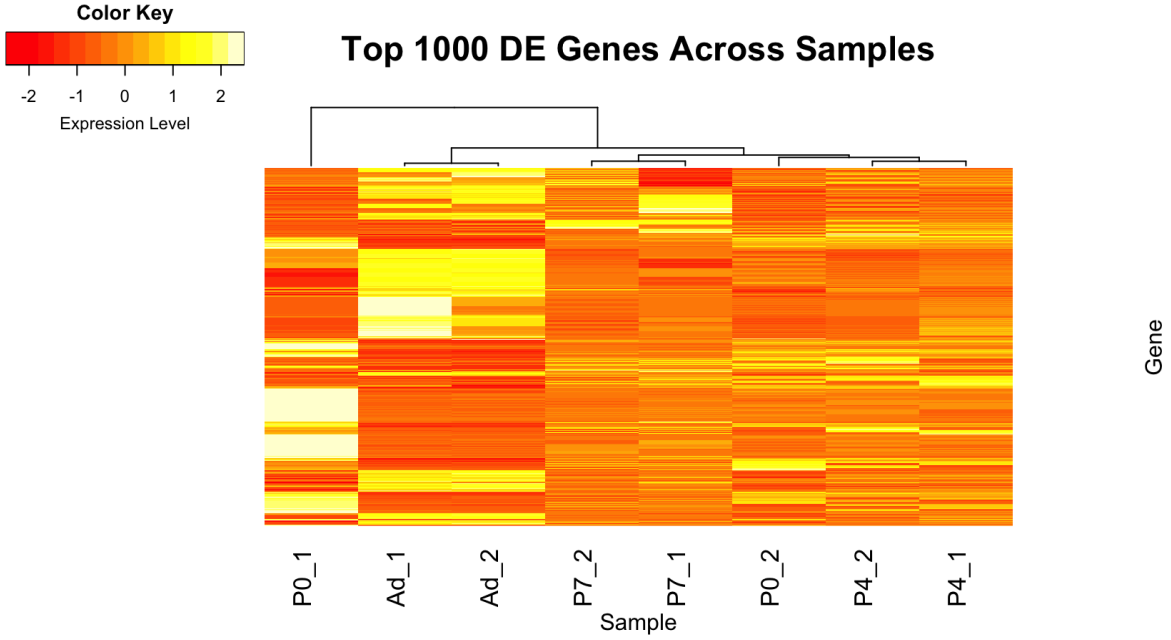

Figure 4. Heatmap hierarchical clustering of samples P0, P4, P7, and Ad for the top 1000 differentially expressed genes between P0_1 and Ad. The rows are the genes and columns are the samples, with color intensity dependent on FPKM.

DISCUSSION

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 suggests that cardiac myocytes exit the cell cycle in the adult stage, thus resulting in a failure to reactivate myocyte proliferation-related genes after injury. To assess this further, future studies may perform gene-knockout experiments to evaluate how the expressed phenotype changes and impacts CM

regeneration. Furthermore, identifying novel cytokines and growth factors that could initiate cell cycle entry of cardiomyocytes may help researchers understand how to reactivate proliferation.

Although our results are similar to O'Meara et al., inconsistencies have been found in the number of significant genes, enrichment scores, and clustering. We found notably less significant genes between the P0 and adult groups than reported in the literature. However, this may be due to differences in filter methods and gene set enrichment analysis being performed on both *in vivo* and *in vitro* cardiac myocyte differentiation, whereas this study focused on *in vivo* (P0 and Ad) cardiomyocyte maturation. Furthermore, a P0 replicate is incorrectly clustered, which may be due to the sample processing or the clustering method used. Overall, despite these differences, this study has demonstrated trends and enrichment terms that are consistent with the study, and may be deemed successful.

CONCLUSION

Through the replication of the analytical and biological aspects of the O'Meara et al study, P0 gene expression was profiled, and hierarchical clustering and enriched gene ontology analysis was performed. In doing so, key regulators and mediators were identified, and a core transcriptional signature of cardiac myocyte differentiation for different developmental stages was determined with high accuracy. As a result, we have successfully demonstrated that adult cardiac myocytes exhibit a transcriptional reversion of the differentiation process using bioinformatic tools.

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] Wamstad JA, Alexander JM, Truty RM, Shrikumar A, Li F, Eilertson KE, Ding H, Wylie JN, Pico AR, Capra JA, Erwin G, Kattman SJ, Keller GM, Srivastava D, Levine SS, Pollard KS, Holloway AK, Boyer LA, Bruneau BG. Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage. Cell. 2012;151:206–220.

[4] 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.

[5] 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.

[6] 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.

[7] 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

[8] “The R Project for Statistical Computing.” R, www.R-project.org/.

[9] Trapnell, Cole. Cufflinks, cole-trapnell-lab.github.io/cufflinks/cuffdiff/.

[10] 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]

[11] 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]

[12] 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.