Camilla Belamarich
May 9, 2021
BF 528: Project 5

Transcriptional Profile of Mammalian Cardiac Regeneration with mRNA-seq

Introduction

O'Meara et al (2015) provided significant findings on the transcriptional regulation of cardiac mammal heart regeneration and the biological processes that regulate heart regeneration. Researchers specifically wanted to understand why the ability of heart regeneration in neonatal mice is lost after the first week of life.[1] To gain a better understanding of transcriptional regulation of heart regeneration, researchers observed global gene expression patterns in Cardiac Myocyte (CM) differentiation in mice over time .[1] In my analysis, I will focus on comparing transcriptional changes in CM differentiation in samples from postnatal day zero (P0) versus Adult (Ad). I will first be analyzing the differential expression results and performing a functional analysis on the most significant up- and down-regulated genes using DAVID. Next, I will be using the FPKM expression matrices to observe biological patterns and compare my results to those reported in the paper. Following a similar workflow as O'Meara et al (2015) for the first and last time point of CM differentiation *in vivo* heart maturation will reveal the biological processes that regulate heart regeneration; essential for understanding why mice lose the ability to regenerate their heart later in development.

Methods

Identifying Differentially Expressed Genes in Myocyte Differentiation

The results from cuffdiff analysis contained differential expression data of all replicates of samples from timepoints P0 to Adult. Using these differential expression results, the data was sorted so that the smallest q-value was at the top. The top 10 genes from this sorted data were extracted along with the gene name, both FPKM values, log fold change, p-value, and q-value (Table 1). The significant genes were extracted in two ways. First, if the genes were labeled "significant" in the data frame, these genes were counted and reported (Table 2). Second, if the genes had a p-value less than 0.01, the genes were also considered significant and reported (Table 2). Histograms were produced for both sets of significant genes and used to display the distribution of log2 fold change values. The up- and down-regulated genes were extracted and reported using a log2 fold change threshold of zero. The significant genes using p-value were used to identify up- and down-regulated genes. A log2 fold threshold bigger than zero and a log2 fold threshold less than zero represented the up- and down-regulated genes, respectively.

DAVID Functional Analysis

In order to determine the function of these differentially expressed genes, the gene names were inserted into the DAVID Functional Annotation Tool database and used the functional

annotation clustering tool.[2] Since these genes represent the official gene symbol, that parameter was chosen along with the *Mus musculus* species option. Gene Ontology (GO) terms were selected for further analysis. Specifically, GOTERM_BP_FAT, GOTERM_MF_FAT and GOTERM_CC_FAT were selected from the list of features. These GO terms represent Biological Processes, Molecular Function, and Cellular Components, respectively. DAVID groups the functions of the inserted genes into functionally related clusters, which efficiently displays results.[2] These results were then compared to the findings in O'Meara et al (2015) Supplemental 2 Tables 1D and 1E. Common GO terms that were found in my analysis and the papers' analysis were marked with an asterisk (Table 3 and 4).

Comparing Biological Trends Using FPKM Tracking IDs

The Fragments per Kilobase of Transcript per Million (FPKM) tracking tables were used to form biological interpretations of trends in the data. The FPKM tracking table for each replicate from each timepoint (P0, P4, P7, and Adult) sample was used to identify these trends. To replicate the results from Figure 1D in the paper, we extracted representative genes from the Sarcomere, Mitochondria, and Cell Cycle.[1] The representative genes for Sarcomere were Pdlim5, Pygm, Myoz2, Des, Csrp3, Tcap, and Cryab. For Mitochondria: Prdx3, Acat1, Echs1, Slc25a11, and Phyh. Lastly for the Cell Cycle: Cdc7, E2f8, Cdk7, Cdc26, Cdc6, E2f1, Cdc27, Cdc45, Rad51, Aurkb, and Cdc23. There were two genes that were left out of this analysis since there was no FPKM value associated with them. These genes were Mpc1 in Mitochondria and Bora in the Cell Cycle. These genes could have been filtered out due to a low signal not being detected. The average FPKM values between the two replicates for each sample were computed and displayed in line plots (Figure 2). Additionally, a clustered heatmap displaying expression levels for all eight samples was generated using ggplot2 (Figure 3). In order to generate this plot, all FPKM values needed to be concatenated into one FPKM matrix. The top 750 differentially expressed genes in P0 compared to Ad were used for the heatmap.

When concatenating the eight samples into one FPKM matrix and subsetting the data, duplicated Ensembl tracking IDs were identified and removed from the matrix. Additionally, there were rows identified that had more than one gene listed, which were introduced by multiple overlapping splicing transcripts that mapped to approximately the same region. For these instances, the first gene symbol listed was used for the analysis. When mapping the gene symbols to Ensembl tracking IDs using various Bioconductor packages, not all gene symbols could be mapped due to instances of multiple gene synonyms for the same tracking ID or some Ensembl tracking IDs were out of date and could not be mapped to gene symbols. Therefore, the gene symbol in the FPKM matrix was used to subset the matrix for the clustered heatmap.

Results

Differential Expression Analysis

The q-value was ordered and the top ten differentially expressed genes between timepoints P0 and Adult were extracted from that subset of data (Table 1). All ten genes had the same q-value of 00106929 and p-value of 5e-05., which indicates that the level of significance (p-value < 0.05) was not exceeded. The FPKM values ranged from a low of 0.69129 to a high of 265.235.

Table 1. Top 10 Differentially Expressed Genes. Top differentially expressed genes between timepoints P0 and Adult sorted on q-value. The gene names, FPKM values, log fold change, and p-values were also reported.

Gene Name	FPKM Value_1	FPKM Value_2	log_fold change	P-Value	Q-Value
Plekhb2	22.5679	73.5683	1.7048	5E-05	0.00106929
Mrpl30	46.4547	133.038	1.5179	5E-05	0.00106929
Coq10b	11.0583	53.3	2.2690	5E-05	0.00106929
Aox1	1.18858	7.09136	2.5768	5E-05	0.00106929
Ndufb3	100.609	265.235	1.3985	5E-05	0.00106929
Sp100	2.13489	100.869	5.5622	5E-05	0.00106929
Cxcr7	4.95844	32.2753	2.7025	5E-05	0.00106929
Lrrfip1	118.997	24.6402	-2.2718	5E-05	0.00106929
Ramp1	13.2076	0.69129	-4.2559	5E-05	0.00106929
Gpc1	51.2062	185.329	1.8557	5E-05	0.00106929

Before filtering for significance, there were a total of 36,259 genes. The total number of significant differentially expressed genes using both methods were compared (Table 2). Using the significance column in the starting differential expression dataset, it was determined that there were a total of 2,139 genes found. Conversely, when using the p-value to filter, there were a total of 2,376 genes found. Overall, there were more significantly differentially expressed genes when using the p-value significance method.

Table 2. Number of Significantly Expressed Genes. Using p-value and the “significance” column in the differential expression dataset to filter statistically significant genes. Log2 fold change threshold of zero was applied to determine up- and down-regulated genes.

	Total	Up-Regulated	Down_regulated
P < 0.01	2,376	1,187	1,189
Labelled Significant in Data	2,139	1,084	1,055

Additionally, setting the log2 fold change threshold for each significance method was applied and compared (Table 2). Using zero as the log2 fold change threshold roughly filtered half the total amount of genes as up-regulated and half as down-regulated. The distribution of log2 fold change values and frequency can be observed in Figure 1. The top histogram shows the distribution across all genes in the differential expression dataset (Fig. 1). Before significance thresholds were applied, there was a large spike in frequency at a log2 fold change value at around -0.5. After the significance threshold (labeled column method) was applied, there are more definite up- and down-regulated groups, which is displayed in the lower image of Figure 1.

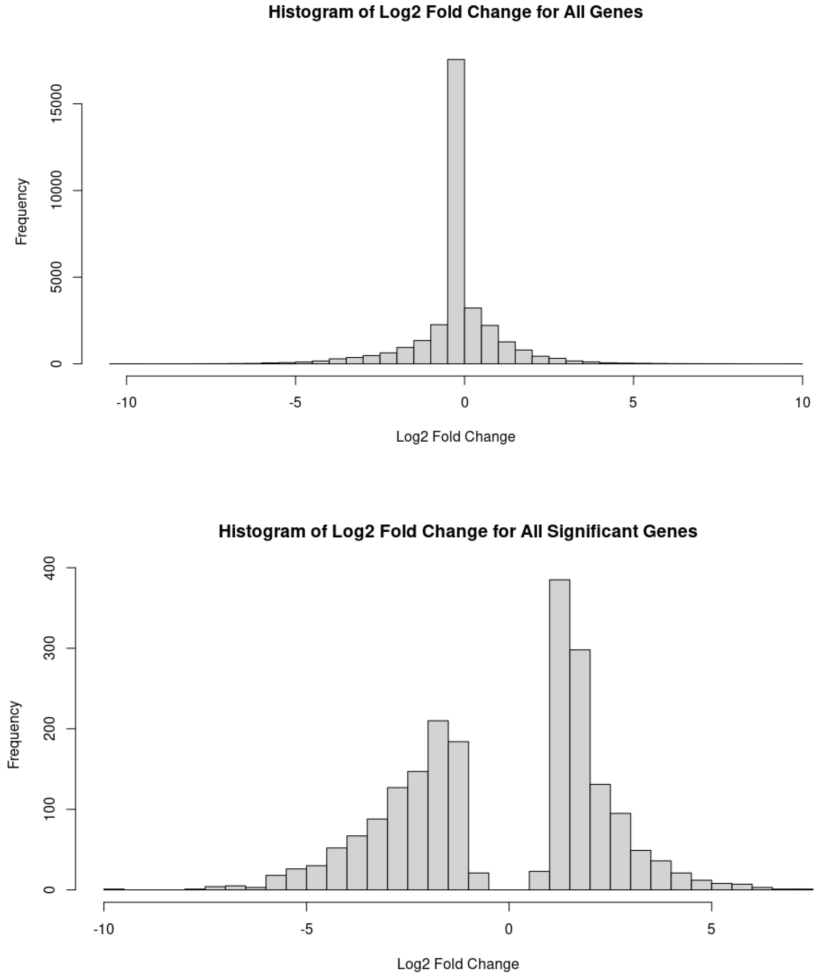

Figure 1. Distribution of Log Fold Change for All Genes and All Significant Genes. Top image displays the distribution of log2 fold change values across all genes found in the differential expression dataset (n = 36,259 genes). Lower image shows distribution of log2 fold change values in only significant genes (n = 2,139 genes).

DAVID Functional Annotation Analysis

_____Significant up- and down-regulated genes were inputted into the DAVID Annotation database and clustered based on function. The top five clusters and top three GO terms were extracted for each group (Table 3 and 4). The top five clusters were determined based on enrichment score and the top three GO terms were determined based on p-value. The top three GO terms were compared to the results in Supplemental 2 Tables 1D and 1E in the paper.[1] The overlapping GO terms are indicated with three asterisks.

Table 3. Top Five Up-Regulated Clusters from DAVID Analysis. Top five clusters based on enrichment score and top three GO terms based on p-value. (***) indicates common GO terms when compared to results from O'Meara et al (2015).

Cluster Number	Enrichment Score	Category	GO Term and Function
Cluster 1	24.36	GOTERM_CC_FAT	GO:0005739 - mitochondrion ***
			GO:0044429 - obsolete mitochondrial part
			GO:0031966 - mitochondrial membrane
Cluster 2	17.15	GOTERM_BP_FAT	GO:0006091 - generation of precursor metabolites and energy ***
			GO:0046128 - purine ribonucleoside metabolic process
			GO:0042278 - purine nucleoside metabolic process
Cluster 3	16.72	GOTERM_BP_FAT	GO:0006082 - organic acid metabolic process
			GO:0043436 - oxoacid metabolic process
			GO:0019752 - carboxylic acid metabolic process
Cluster 4	12.05	GOTERM_CC_FAT	GO:0043230 - extracellular organelle
			GO:1903561 - extracellular vesicle
			GO:0070062 - extracellular exosome
Cluster 5	8.32	GOTERM_BP_FAT	GO:0009056 - catabolic process
			GO:0006631 - fatty acid metabolic process ***
			GO:0044282 - small molecule catabolic process

In the up-regulated clusters from DAVID enrichment analysis, there were two biological processes and one cellular component that were also found in O'Meara et al (2015). These biological processes were found to be associated with a generation of precursor metabolites and energy and fatty acid metabolite processes. The common cellular component found was mitochondrion. Not listed in the top five clusters was sarcomere as a common cellular component. As CM maturation proceeds, there is a need for increased energy and organization of compositional units in the cells. Furthermore, our results correspond with biological processes and cellular components in CM maturation.

Table 4. Top Five Down-Regulated Clusters from DAVID Analysis. Top five clusters based on enrichment score and top three GO terms based on p-value. (***) indicates common GO terms when compared to results from O'Meara et al (2015).

Cluster Number	Enrichment Score	Category	GO Term and Function
Cluster 1	11.57	GOTERM_BP_FAT	GO:0007049 - cell cycle ***
			GO:0000278 - mitotic cell cycle ***
			GO:0022402 - cell cycle process ***
Cluster 2	11.12	GOTERM_CC_FAT	GO:0031012 - extracellular matrix ***
			GO:0044420 - obsolete extracellular matrix component
Cluster 3	10.31	GOTERM_BP_FAT	GO:0008283 - cell population proliferation
			GO:0042127 - regulation of cell population proliferation
			GO:0008285 - negative regulation of cell population proliferation
Cluster 4	9.18	GOTERM_BP_FAT	GO:0051128 - regulation of cellular component organization
			GO:0033043 - regulation of organelle organization
			GO:0051130 - positive regulation of cellular component organization
			GO:0072359 - circulatory system development

Cluster 5	8.96	GOTERM_BP_FAT	GO:0001568 - blood vessel development ***
			GO:0001944 - vasculature development ***

In the down-regulated clusters, there were five biological processes and one cellular component found in common with O'Meara et al (2015). The five biological processes were found to be associated with the cell cycle, mitotic cell cycle, cell cycle process, blood vessel development, and vasculature development. The one common cellular component was extracellular matrix. The reason why mammals cannot fully regenerate their heart after an injury as an adult is due to cardiac myocytes (CM) leaving the cell cycle.[3] This occurs shortly after birth and heart growth persists. This is consistent with the down-regulated biological processes that were found in this functional analysis.

Biological Trends Using FPKM Values

FPKM values were plotted against in vivo maturation timepoints to observe biological trends in this process. When using the FPKM values from each replicate of each timepoint, there was generally a similar trend of representative genes for sarcomere, mitochondria, and cell cycle (Figure 2). Two genes important to sarcomere assembly are Titin Cap (tcap) and Cardiac Troponin (cryab). O'Meara et al (2015) saw a pronounced increase; however, tcap and cryab did not show a similar increase (Fig. 2A). For Sarcomere, the range of FPKM values for P0 were found to be around 100-650, and for Adults the FPKM values were found to be 200-600. O'Meara et al (2015) had higher FPKM values for mitochondria than what was found in this analysis. Figure 2B had similar trends as O'Meara et al (2015); however, the similar issue with lower FPKM values would also affect the overall results in this analysis. For mitochondria, the range of FPKM values for P0 were found to be around 50-200, and for Adults the FPKM values were found to be 130-200. Gene Slc25a11 had an almost linear increase from P0 to Adult. Lastly, the trends for FPKM values in the cell cycle matched exactly what was found in this analysis, which decreased as in vivo maturation decreased. This was expected due to the same reasons as why cell cycle biological processes were found to be down regulated.

To compare differences in gene expression, a clustered heatmap was produced. The clustered heatmap in Figure 3 reveals that clusters found highly expressed (purple color) in P0 were faintly expressed (yellow color) in the Adult mice. Similarly, other clusters found faintly expressed in P0 were highly expressed in Adult mice (Fig. 3). Timepoints P4 and P7 had mildly expressed genes. Overall, gene expression was shown to decrease across maturation timepoints in up-regulated genes, but increased in expression across maturation timepoints in down-regulated genes. Compared to the heatmap in O'Meara et al (2015) (Fig. 2A), there were some similarities. All four time points had high expression in the first half of their gene, low expression in the middle of the second half of their genes, and high expression again towards the

bottom of their genes. Genes in P0 and P4 had high expression at the top of their genes, but low expression towards the bottom. Where our results differed most was during P7 and Adult time points. In this analysis, these time points were observed to have low expression at the top of their genes, but high expression towards the bottom of the genes list. This was almost opposite to what O'Meara et al (2015) found in their clustered heat map for in vivo heart maturation.

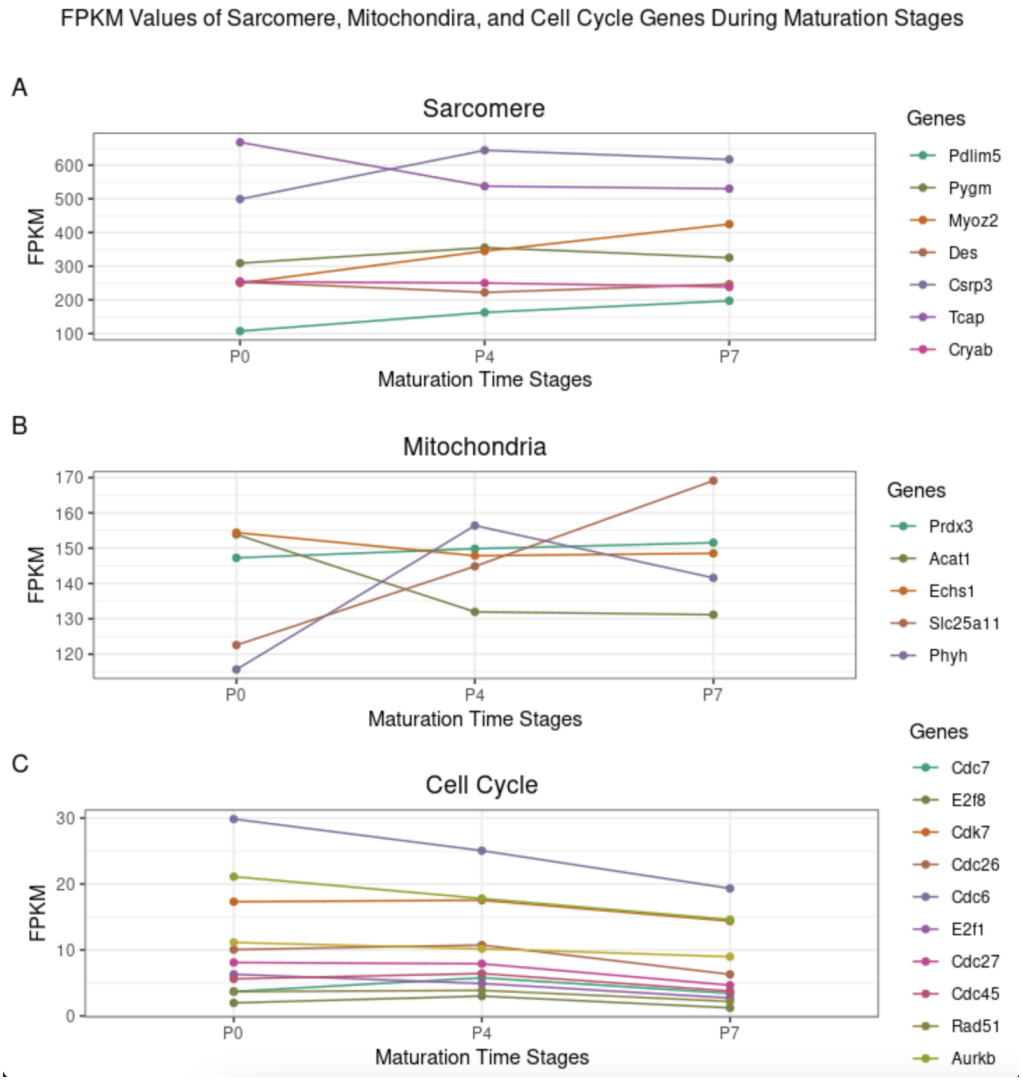

Figure 2. FPKM Values of Representative Sarcomere, Mitochondria, and Cell Cycle Genes Differentially Expressed in Vivo Maturation. A. genes from sarcomere maturation, B. genes from mitochondria maturation, and C. genes from cell cycle maturation. Genes are labelled by color.

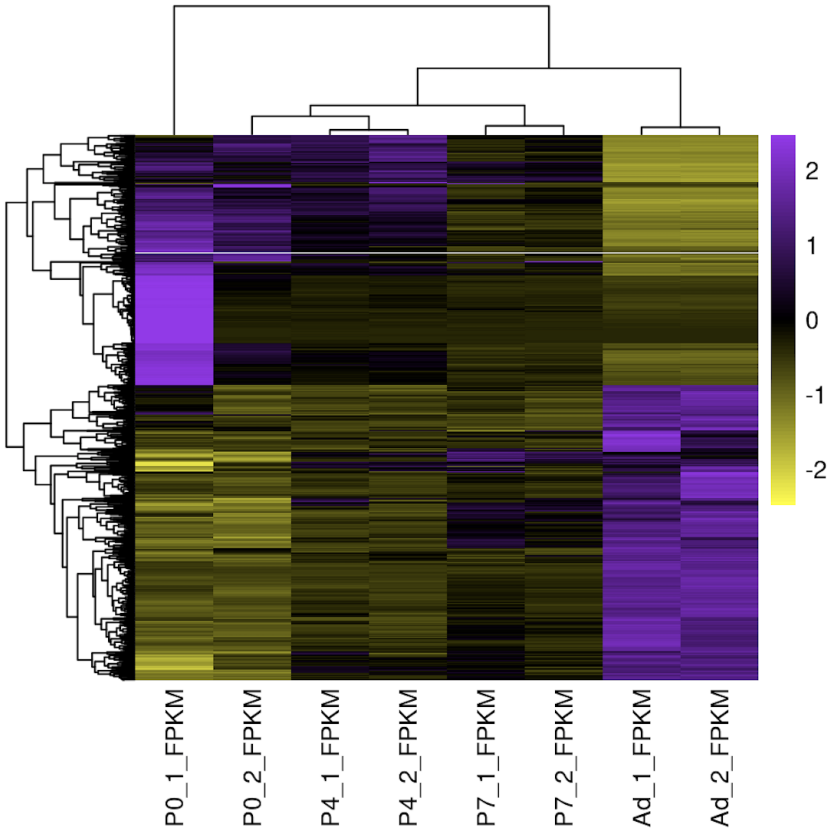

Figure 3. Clustered Heatmap of FPKM Values Using Top 750 Differentially Expressed Genes in P0 and Ad Analysis. Gene expression levels across in vivo maturation time points. Purple-colored clusters represent high expression and yellow-colored clusters represent low expression levels.

Discussion

Utilizing similar Bioinformatics tools as O'Meara et al (2015) to analyze RNAseq differential expression data to observe gene expression patterns in CM differentiation across four time points revealed fundamental biological processes and trends in the regulation of heart regeneration. While O'Meara et al (2015) looked at *in vivo*, *in vitro*, and explant experiments, I focused only on *in vivo* heart maturation between P0 and Adult time points. Due to a much smaller subset of data in this analysis, overall trends and some biological processes were different than the findings in the paper.

My main goal was to replicate similar findings to O'Meara et al (2015) that shows evidence of transcriptional reversion of injured heart cells during CM differentiation. We were able to reproduce some findings in up- and down-regulated biological processes, FPKM trends in sarcomere, mitochondria, and cell cycle, and overall expression levels in genes across all four time points. There were some inaccuracies detected in my analysis. One main possibility for the discrepancies could be the size of the data set. Since I was only working on a small portion of the

overall data, many genes were left out of this analysis, which affects all the results downstream. Genes could have been missing during the DAVID function analysis, which loses significant biological functions associated with heart regeneration. Additionally, the trends in FPKM values were significantly different in Sarcomere results, which contained tcap and cryab genes that are important for sarcomere assembly. Some of these trends were different from the trends found in the paper. Mitochondria was most notably different.

A second possibility for discrepancies in results from my analysis and from those in the paper could be explained by my groups' upstream analysis in the Data Curator or Programmer steps. Since I used my groups data for this project, the same discrepancies we had in project 2, I also had in this project. Although O'Meara et al (2015) did an exceptional job providing comprehensive evidence of transcriptional reversion of heart regeneration, it was challenging to reproduce results exactly without extensive documentation of their analysis. However, a main finding in my analysis that was also similar to O'Meara et al (2015) was the down-regulation of biological processes and downward trend in FPKM values of the cell cycle. Previously stated, it is known that hearts lose the ability to regenerate due to cardiac myocytes (CM) leaving the cell cycle.[3]

Overall, I was able to reproduce some of the results that O'Meara et al (2015) had in their analysis. If I had managed to successfully reproduce the findings in the paper, these findings can be used for future research in cardiovascular diseases, which plague the human population. Since there is evidence that neonatal mice can regenerate their heat after injury in the first week of life, understanding the transcriptional expression changes provides a concrete framework for digging deeper into this field of study.[4-6]

References

1. O'Meara, C. C., Wamstad, J. A., Gladstone, R. A., Fomovsky, G. M., Butty, V. L., Shrikumar, A., Gannon, J. B., Boyer, L. A., & Lee, R. T. (2015). Transcriptional reversion of cardiac myocyte fate during mammalian cardiac regeneration. *Circulation research*, *116*(5), 804–815. https://doi.org/10.1161/CIRCRESAHA.116.304269
2. DAVID Functional Annotation Tool: https://david.abcc.ncifcrf.gov/summary.jsp
3. Bicknell KA, Coxon CH, Brooks G. Can the cardiomyocyte cell cycle be reprogrammed? *J Mol Cell Cardiol*. 2007;42:706–721. doi: 10.1016/j. Yjmcc.2007.01.006.
4. Oberpriller JO, Oberpriller JC. Response of the adult newt ventricle to injury. *J Exp Zool*. 1974;187:249–253. doi: 10.1002/jez.1401870208.
5. Poss KD, Wilson LG, Keating MT. Heart regeneration in zebrafish. *Science*. 2002;298:2188–2190. doi: 10.1126/science.1077857.
6. Steinhauser ML, Lee RT. Regeneration of the heart. *EMBO Mol Med*. 2011;3:701–712. doi: 10.1002/emmm.201100175.

Supplemental References - R Packages used in Biologist

1. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
2. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
3. Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.1. https://CRAN.R-project.org/package=dplyr
4. Erich Neuwirth (2014). RColorBrewer: ColorBrewer Palettes. R package version 1.1-2. https://CRAN.R-project.org/package=RColorBrewer
5. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
6. H. Wickham. Reshaping data with the reshape package. Journal of Statistical Software, 21(12), 2007.
7. Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra
8. Alboukadel Kassambara (2020). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0. https://CRAN.R-project.org/package=ggpubr
9. Tal Galili (2015). dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering. Bioinformatics. DOI: 10.1093/bioinformatics/btv428