Project 2 - Transcriptional Profile of Mammalian Cardiac Regeneration with mRNA-Seq Replicating O’Meara et al.

Group members: Emily Hughes, Simran Makwana, Sumiti Sandhu, and Michiel Smit
TA: Nick

Introduction

In their first week of life, neonatal mice can regenerate their hearts in the event of injury.[1] Adult mammals lack this ability because their cardiac myocytes exit the cell cycle, preventing further replication and growth.[2] The authors of the selected study aimed to characterize differences in gene expression between the neonatal mice, who retain the ability to regenerate their heart, and adult mice, who have lost it. Investigating these differentially expressed genes could illuminate potential clinical targets in combating heart disease.

O’Meara et al. profiled the gene expression patterns of these different types of neonatal mice to determine any differentially expressed genes or pathways using RNASeq data.[3] This was accomplished using the TopHat and Cufflinks tools to map the reads to the genome and conduct differential expression analysis on the data, respectively. GO terms were used to compare and contrast the top upregulated and downregulated genes between the adult and neonatal samples.

Data

Samples were prepared using the following protocols: Embryonic stem cell differentiation, Whole heart ventricle isolation, Adult cardiac myocyte isolation, Neonatal mouse apical resection, and Neonatal cardiac myocyte purification. The total RNA was extracted from all of these samples using Trizol (Invitrogen). Paired end 40 base pair read length sequencing was performed using Illumina HiSeq 2000. In our replication of this study, the following eight samples were analyzed: P0_1, P0_2, P0_4, P0_4, P0_7, P0_7, Ad_1, Ad_2. P0 indicates a neonatal mouse and Ad indicated an adult male mouse sacrificed at eight to ten weeks old. These samples were downloaded from the NCBI functional genomics data repository (GEO). The additional sample (P0_1, accession number GSM1570702) was downloaded in SRA format from the following link: https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR1727914.

SRA-toolkit was used to extract the downloaded SRA sample into two FASTQ files (one for each paired-end). FastQC was then run on each of these files to determine data quality. Zero sequences were flagged as poor quality from either file and both had a %GC content of 49. The per base sequence quality plots produced by this analysis (Figure 1) demonstrate the general high quality score across all bases. Another metric that indicates the quality of the data is the per base sequence content (Figure 2). Although the sequence content across all bases varies over the first 15 base pair positions, the distribution of nucleotides then plateaus, indicating that the sequence is high quality. No samples were eliminated due to low quality and no sources of error or contamination were detected.

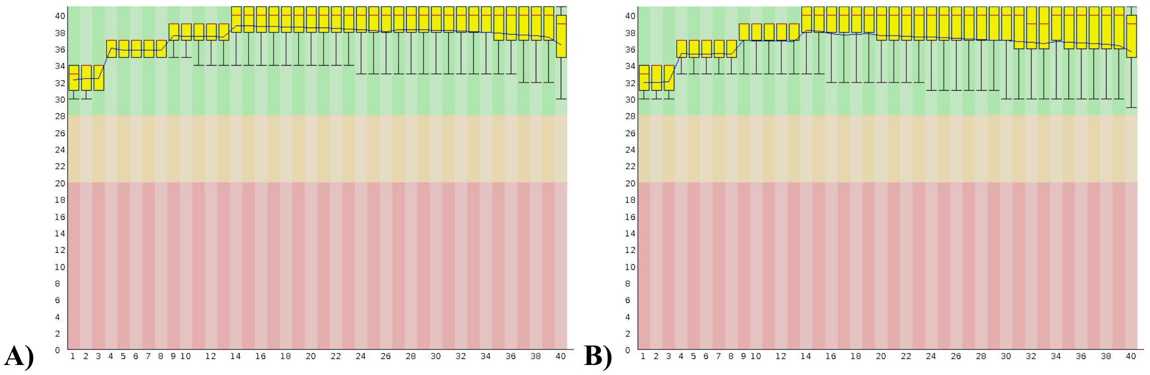

Figure 1. Per base sequence quality of **(A)** paired end 1 and **(B)** paired end 2, with position in read (bp) on the x axis.

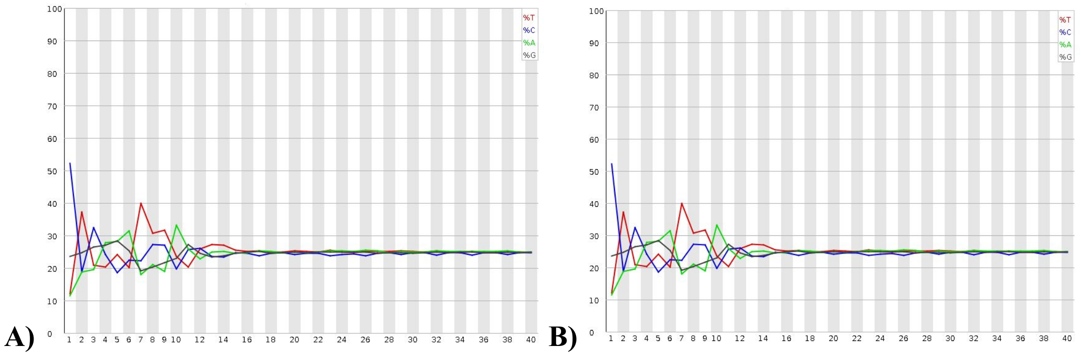

Figure 2. Per base sequence content of **(A)** paired end 1 and **(B)** paired end 2, with position in read (bp) on the x axis. Red represents thymine, Blue represents cytosine, green represents adenine, and black represents guanine.

Methods

Splice junctions from the P0 read files were mapped to the mm9 mouse reference genome using TopHat. The mean inner distance between mate pairs was set to be 200 base pairs. Each read was split into a segment read of 20 base pairs. Segments were aligned independently with up to one mismatch in each segment alignment. The results of the alignment were summarized using samtools and RSeQC. Specifically, the summary statistics, measurement of the inner distance between reads, and the calculated read coverage over the gene body were obtained. Multimapped reads were included in the analysis. Cufflinks was used to perform differential expression analysis of the P0 data. This specific tool was used in the initial study and was chosen to maintain consistency in this replication of the analysis. No outliers were removed from the data. Cuffdiff was used to perform differential expression analysis comparing the neonatal and adult data.

Results

The results from the calculation of the RNA-seq reads coverage over the gene body can be viewed in Figure 3. This graph is consistent with 3’ bias since the coverage on the 3’ end of the gene body is quite greater than the 5’end and center of the gene body.

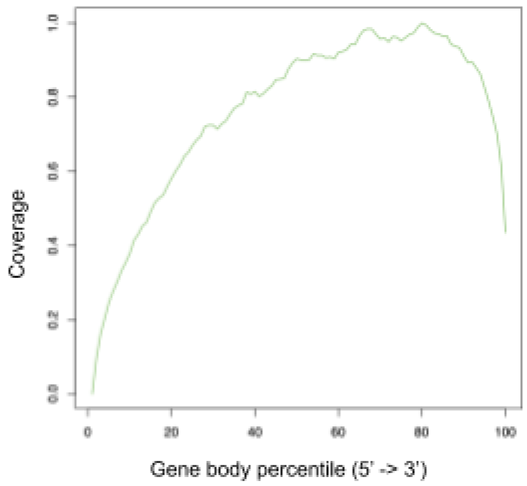

Figure 3. RNA-Seq read coverage of the gene body from a 5’ to 3’ direction. Due to the maintained high coverage at the 3’ end of the gene, these results indicate 3’ coverage bias in this analysis.

During the differential expression analysis, the read counts were normalized to fragments per kilobase of transcript per million mapped reads, or FPKM. The analysis of the P0 data resulted in the majority of the FPKM values being exactly or extremely close to 0. To visualize the other values, Figure 4 displays the data greater than 0.0001 FPKM. Based on this plot, the counts seem to be normally distributed on log scale and centered on 10 FPKM. This analysis was also performed on the other datasets provided to compare the differential expression between prenatal and adult groups.

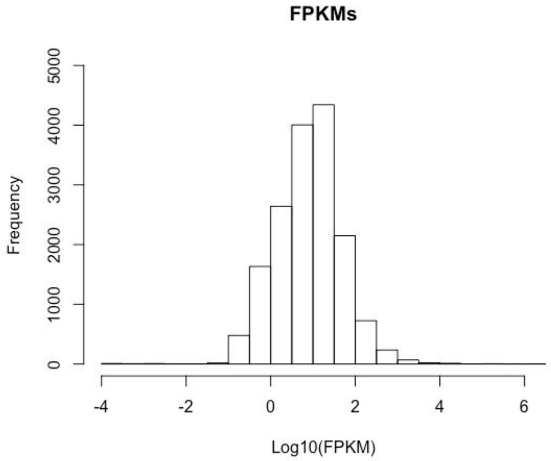

Figure 4. Frequency of the fragments per kilobase of transcript per million mapped reads (FPKM) that were greater than 0.0001. This displays the log-transformation of the FPKMs.

The summary results from the splice junction mapping are displayed in Table 1. Based on these alignment results, the insert size between the reads of the mRNA generally follows a normal distribution, as shown in Figure 5. The mean insert size was about 85.4 base pairs with a standard deviation of 43.4 base pairs.

	Number of reads	**Percentage of total reads**
Mapped reads	49,706,999	100%
Unique reads	38,489,380	77.4%
Multimapped reads	2,899,954	5.8%
Unaligned reads	0	0%

Table 1. Summary of the results from the alignment of splice read junctions from postnatal day 0 data.

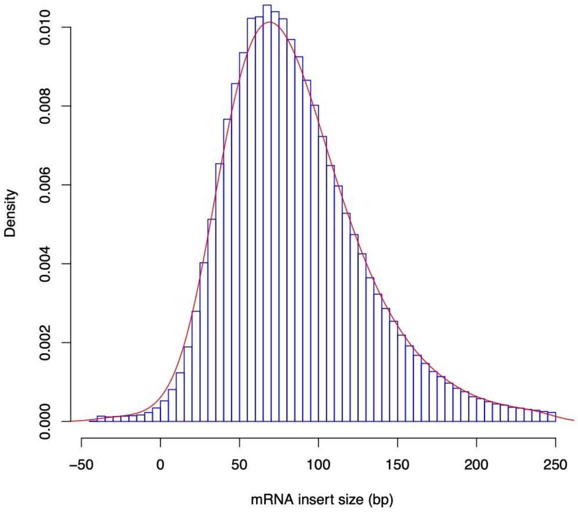

Figure 5. Measured insert size from the splice junction alignment. The insert size is defined as the distance between reads that were paired in the analysis. The mean insert size was 85.4 base pairs. The standard deviation was 43.4 base pairs.

The differential statistics data of neonatal versus adult state based on their q-value were sorted. The top ten differentially expressed genes with the smallest q-values are listed in Table 2.

S.No.	Gene	FPKM value 1	FPKM value 2	Log-2 fold change	P-Value	Q-Value
1	Plekhb2	22.57	73.57	1.70	5e-05	0.001
2	Mrpl30	46.46	133.04	1.52	5e-05	0.001
3	Coq10b	11.06	53.30	2.27	5e-05	0.001
4	Aox1	1.19	7.09	2.58	5e-05	0.001
5	Ndufb3	100.61	265.24	1.40	5e-05	0.001
6	Sp100	2.14	100.87	5.56	5e-05	0.001
7	Cxcr7	4.96	32.28	2.70	5e-05	0.001
8	Lrrfip1	119.00	24.64	-2.27	5e-05	0.001
9	Ramp1	13.21	0.69	-4.26	5e-05	0.001
10	Gpc1	51.21	185.33	1.86	5e-05	0.001

Table 2. Differential expression statistics comparing the two conditions (P0 versus Adult)

The log2 fold change values for all differentially expressed genes are displayed in Figure 6(A), the histogram shows that most of the genes have values ranging from -1 to zero. For significant genes, the histogram in Figure 6(B) displays that most of the genes had log2 fold change values ranging between -2 to -1 and 1 to 2.

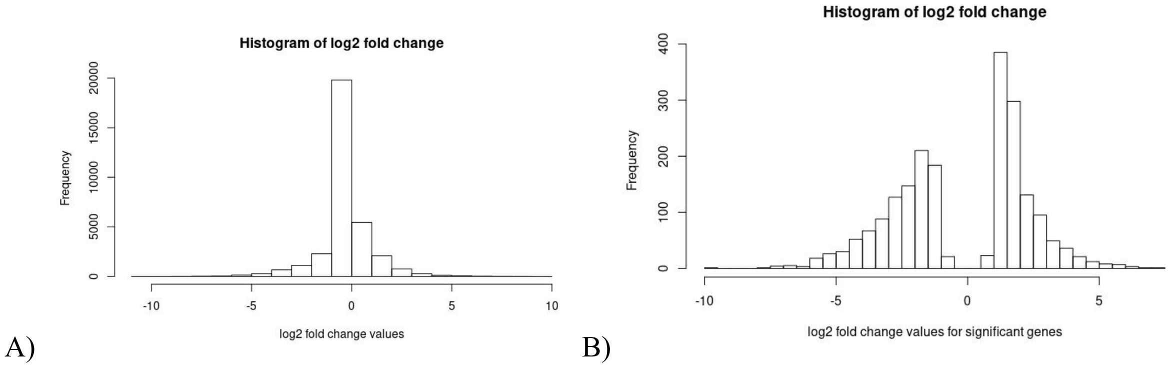

Figure 6. Histogram of **(A)** log2 fold change for postnatal day 0 (P0) versus Adult state, **(B)** log2 fold change of significant genes for postnatal day 0 (P0) versus Adult state.

The significant genes were filtered based on log2 fold change values to get up-regulated genes (log2 fold change >1) and down-regulated genes (log2 fold change <1). Out of 36,329 differentially expressed genes, a total of 1,061 up and 1,078 down-regulated genes were observed. Functional annotation clustering, performed using DAVID, on up and down-regulated genes resulted in a total of 404 annotation clusters for up-regulated genes (summarized in Table 3) and 420 annotation clusters for down-regulated genes (summarized in Table 4).

Gene set analysis on significantly up-regulated genes			
Annotation Cluster	**Gene Ontology (GO) enrichment terms**	**P-Value**	**Enrichment Score**
1	GO:0005739~mitochondrion	3.11E-48	20.89
2	GO:0006082~organic acid metabolic process	6.29E-25	16.48
3	GO:0015980~energy derivation by oxidation of organic compounds	3.06E-21	15.09
4	GO:0043230~extracellular organelle	6.00E-19	12.87
5	GO:0043292~contractile fiber	6.97E-09	7.386
6	GO:0006631~fatty acid metabolic process	4.50E-10	6.84
7	GO:0051186~cofactor metabolic process	3.39E-19	6.56
8	GO:0006805~xenobiotic metabolic process	3.67E-11	5.45
9	GO:0042802~identical protein binding	1.58E-10	5.25
10	GO:0051537~2 iron, 2 sulfur cluster binding	4.68E-09	5.24

Table 3. Enriched GO terms associated with the up-regulated genes (log2 fold change >1) organized into clusters based on functional relatedness where the terms in green are the ones that are similar to the terms reported in the paper.

Gene set analysis on significantly down-regulated genes			
Annotation Cluster	**Gene Ontology (GO) enrichment terms**	**P-Value**	**Enrichment Score**
1	GO:0008283~cell proliferation	2.71E-14	9.47
2	GO:0005578~proteinaceous extracellular matrix	2.81E-11	9.37
3	GO:0051128~regulation of cellular component organization	6.61E-18	8.85
4	GO:0007399~nervous system development	4.54E-16	7.94
5	GO:0072358~cardiovascular system development	2.03E-12	7.64
6	GO:0006259~DNA metabolic process	1.77E-11	7.55
7	GO:0005694~chromosome	1.70E-11	7.20
8	GO:0007049~cell cycle	1.71E-11	7.19
9	GO:0007399~nervous system development	4.54E-16	7.05
10	GO:0000793~condensed chromosome	2.50E-09	6.45

Table 4. Enriched GO terms associated with the down-regulated genes (log2 fold change <1) organized into clusters based on functional relatedness.

Discussion

Comparing differences in gene expression (log2 fold change values) between postnatal day 0 and adult stages showed that out of 36,329 genes, 31,789 were under-expressed (down-regulated) as compared to the 4,540 over-expressed (up-regulated) genes. There was an equal number of under and over-expressed genes (1078 and 1061 genes) with significant differential expression.

The functional annotation clustering on up-regulated genes resulted in clusters that had genes that function in the mitochondrion, organic acid metabolic, extracellular organelle processes, among others. For down-regulated genes, the annotation clusters had genes that function in cell proliferation, cardiovascular system development, cell cycle processes, among others.

As determined by gene ontology (GO) terms, the differentially expressed genes showed up-regulation of mitochondrion, organic acid metabolic process etc, related genes and down-regulation of genes related to cell proliferation, proteinaceous extracellular matrix etc. DAVID analysis using functional annotation clustering on up-regulated genes found GO terms associated with following processes (Table 3): mitochondrion (GO:0005739), energy derivation by oxidation of organic compounds (GO:0015980), fatty acid metabolic process (GO:0006631), cofactor metabolic process (GO:0051186., 2 iron, 2 sulfur cluster binding (GO:0051537) were found to be similar to the GO terms reported in the original paper for P0 versus Ad state. No similarities were found between the GO terms from our analysis (Table 4) and the original paper for down-regulated genes but some of the terms we got do make sense biologically. For example, cell proliferation (GO:0008283), especially cardiomyocyte proliferation, and cardiovascular system development (GO:0072358) are down-regulated processes that should occur mainly during the fetal stage.

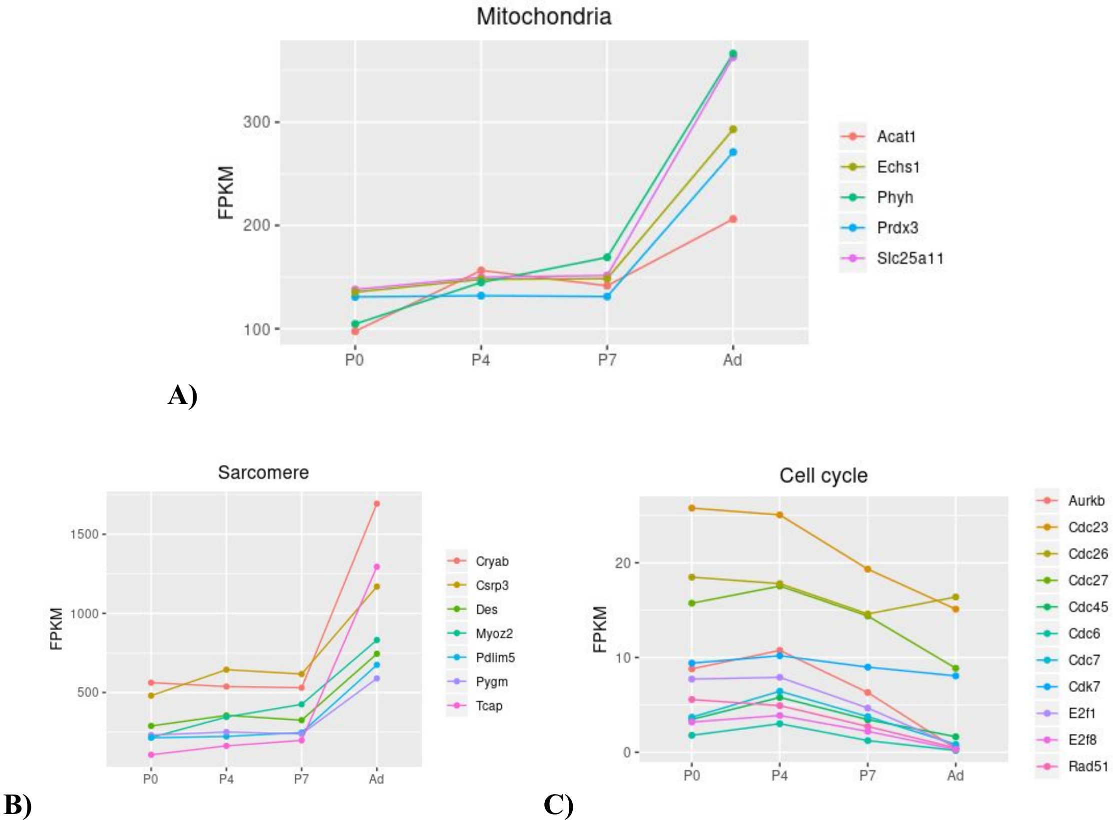

Figure 7. FPKM values of representative **(A)** sarcomere, **(B)** mitochondrial, and **(C)** cell cycle genes significantly differentially expressed during in vivo maturation.

Figure 7 appears to be identical to Figure 1D in the original study.[3] The FPKM values of the chosen genes show the exact same trend across the in vivo maturation period from P0 to adult, as we can see an overall increase in the values of sarcomere and mitochondrial genes. The genes involved in the cell cycle, on the other hand, have FPKM values that decrease slightly across the maturation period and most of these genes have FPKM values less than 10.

Figure 7B shows that genes critical for sarcomere assembly showed pronounced increases in expression in the adults, reflecting sarcomere assembly and organization during cardiac myocyte maturation. The same trend is seen with mitochondrial related genes (Figure 7A), though the FPKM values are less than those of the sarcomere related genes. The FPKM values of those related to mitochondria ranges from about 100 to 350, while the ones related to sarcomere ranges from about 200 up to 1,500. Unlike the genes related to sarcomere and mitochondria, there was a decrease in gene expression for genes related to cell cycle (Figure 7C). These results are expected since mature cardiac myocytes exit the cell cycle and a failure to re-enter the cell cycle is likely to contribute to the lack of cardiac regeneration in adults.[3] Given the trend, the authors were interested in identifying factors that could initiate cell cycle entry of cardiac myocytes by examining differentially expressed genes.

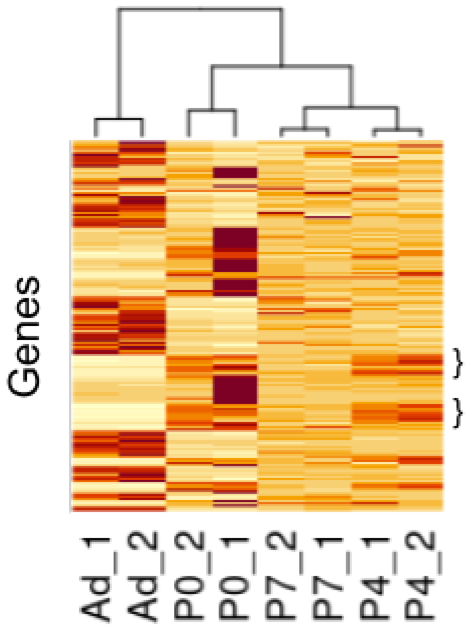

Figure 8. Heatmap of top 1,000 gene expression based on log fold change over the course of in vivo maturation. The dark red color represents higher expression while the light yellow color represents lower expression. The "}" symbol distinguishes a particular gene group of interest.

The heatmap displayed in Figure 8 shows the similarity of expression values between sample replicates (e.g. P0_1 and P0_2). This serves as a sanity check since the two replicates have similar expression. Perhaps the most obvious feature is the patterns of expression between the P0 and adult samples. Often, where the expression values of P0 are high, those of the adult are low and vice versa. This trend is also seen in Figure 7A and 7B with the genes related to the mitochondria and sarcomere.

Since the P4 and P7 samples represent the time between P0 and adult, it would be expected that the gene expression of these groups is between the neonatal and adult mice. One particular group of genes (Figure 8) appears to be strongly expressed in P0 and P4, mildly expressed in P7, and not expressed in Adult. This group of genes may be those associated with the cell cycle, demonstrating the shift from neonatal regeneration ability to adult inability.

Clustering the genes clearly shows differential expression between the P0 and adult samples. In the paper, the heatmap was simply clustered from the highest to lowest gene expression values and there were no distinctive features. Furthermore, it is not possible for further analysis of the gene expression values given by comparing this heatmap to the one produced in the paper since only the top 1,000 genes were extracted based on the log fold change. The heatmap in the paper shows a hierarchical clustering of all expressed genes while the samples were not clustered.

Conclusion

Using the top 1,000 genes based on log fold change, we found marked changes in gene expression between the adult and neonatal mouse samples. One specific challenge of the project was creating a graph that accurately displayed the FPKM values for the P0 sample. This was difficult because the majority of the FPKM values were at or close to zero. To overcome this issue, an arbitrary cutoff of 0.0001 was used to only select the non-zero values. This subset was log-transformed to make the visualization more meaningful and was shown in Figure 4.

It was not possible for us to make any meaningful comparison of the heatmap we created to the one in the paper for multiple reasons. First, we extracted only the top 1,000 genes based on the expression level while the paper used all the genes. Second, the paper did not cluster the samples as they compared cardiac myocyte differentiation in vitro and in vivo while looking at the cardiac myocyte explant model as well. It would then make sense for them not to cluster the samples, but since we are only looking at in vivo maturation, we chose to cluster the samples as well. As mentioned, we were able to see certain interesting features in the heatmap by doing so.

References

1. Steinhauser, M. L. & Lee, R. T (2011). Regeneration of the heart. *EMBO Mol Med. 3*(701-702).

2. Walsh, S., Ponten, A., Fleishmann, B. K., Jovinge, S. (2010). Cardiac myocyte cell cycle control and growth estimation in vivo--an analysis based on cardiac myocyte nuclei. *Cardiovasc. Res. 86*(365-373).

3. O'Meara, C. M. et al. (2015). Transcriptional Reversion of Cardiac Myocyte Fate During Mammalian Cardiac Regeneration. *Circ. Res. 116*(804-815)