Project 5

Sheila Yee

Group: Dachshund

GitHub Repo: https://github.com/sheilayee/Project-5

May 9, 2021

Introduction

Cardiac myocytes are specialized contractile muscles that are critical to the generation of force to pump blood throughout the body's circulatory system. As such, the death of myocardial cell is a leading cause of heart failure. The adult mammalian heart possesses limited regenerative ability upon acute and chronic cardiac injuries like myocardial infarctions (Doppler et al., 2017). Identifying the biological pathways and signatures of cardiac myocyte development and maturation has profound clinical significance regarding better targeted therapeutic treatments to improve heart repair.

While neonatal cardiac myocytes have the ability to regenerate in response injury, this unique function is limited after the first week of life. In neonatal mice, cardiac myocytes are known to regenerate mainly due to proliferation from pre-existing cardiomyocytes with little hypertrophy or fibrosis, thus rendering the mechanism distinct from repair processes associated with mature adult myocytes (Porrello et al., 2011). The varying molecular mechanisms and changes in gene expression pattern by which this heart regeneration occurs has not yet been fully identified and remains poorly understood. O'Meara et al. (2014) sought to answer this by elucidating the transcriptional signatures of neonatal and adult mice that underlie cardiac regeneration upon apical resectioning. Results demonstrated that cardiac myocytes revert to a transcriptionally less differentiated phenotype upon regeneration in which the regenerating mouse heart adopts a global gene expression profile similar to that of immature neonatal myocytes. This is indicated by loss of sarcomere structures and re-entry into the cell cycle.

This project seeks to reproduce some of the findings of O'Meara et al. by determining the transcriptional regulators that oversee cardiac myocyte regeneration and if regenerating cardiac myocytes undergo a transcriptional reversion to a less differentiated phenotype. Gene expression patterns from RNA-seq data originating from neonatal mice of different developmental stages (P0, P4, P7) and adult mice (8-10 weeks old) during in vivo maturation were analyzed. The identification of major transcriptional changes and the biological pathways in which these genes are involved during the cardiac myocyte differentiation process will provide insight in the signaling pathways that mediate myocyte cell cycle activity and regeneration. In addition, it will further understanding about the molecular mechanisms that prevent cardiac myocyte regeneration in adult hearts.

Methods

A short read archive (SRA) file for postnatal day 0 (P0) replicate one had already been previously obtained and processed, producing two FASTQ files. These paired-end reads from the P0 sample were aligned against the *Mus musculus* (mm9) mouse musculus reference genome using TopHat v2.1.1, a program that aligns RNA-seq data against a reference genome without relying on known splice junction sites (Trapnell et al., 2009). To help perform this task, Bowtie2 v2.4.2, a high-throughput short read

aligner based on the FM index algorithm, was implemented along with Boost v1.69, Python2 v2.17.16, and SAMtools v0.1.19 (Langmead et al., 2012; Danecek et al., 2021). This step took around one hour to run. The output of this command produced a BAM file containing details about the original sample as well as alignment information. From the SAMtools v1.10 utilities, the flagstat tool was used to generate statistics on this BAM file, taking less than ten minutes to complete. This permitted quality control evaluation on the alignment - the success of the reads alignment, the number of paired reads, improper mapping of a mate to a different chromosome.

In order to further assess the quality of the RNA-seq data and the alignment, several RseQC v3.0 utilities were used: geneBody_coverage, inner_distance, and bam_stat (Wang et al., 2012). Prior to running these commands, the BAM file was indexed using samtools sort and index functions. This step, taking ten minutes to run, sorts reads by coordinates on each chromosome and indexes the sorted file such that one can quickly look up alignments in specific genomic regions. After running RseQC which took around three hours, summary statistics were produced detailing the percentage of the gene body covered by reads, the distance between paired-end reads, and mapping statistics (unique mapped reads, unmapped reads, failed quality control, etc.). While geneBody_coverage, inner_distance were submitted as batch jobs, bam_stat was run directly on the command line.

After alignment, Cufflinks v2.2.1was implemented to quantify the number of reads mapped to the genomic regions that were defined by a gene annotation found in a provided gtf file (Trapnell et al., 2010). Execution of Cufflinks took around fifteen minutes. This tool produced quantified alignments containing the fragments per kilobase per million (FPKM) mapped reads for all genes of interest. FPKM is a normalized measure of the expression level of a gene which depends on the number of reads mapped to a gene while also taking into account the gene length and sequencing depth. Using RStudio v1.3.1073, the FPKM values were plotted on a histogram using a log10 scale to better visualize the distribution. FPKM values less than one were filtered out and not visualized. Lastly, Cuffdiff, a tool within the cufflinks suite, was used to identify and determine differentially expressed genes (Trapnell et al., 2013). This step took around two hours to complete. It generated a file with information pertaining to differentially expressed genes between P0 and adult samples: gene id's, gene names, p-values, q-values, log2 fold change values, and significance. This file was further used to compare differences in expression level between neonatal and adult sample data.

Due to the computational demand required by these tasks, TopHat, RseQC, Cufflinks and Cuffdiff were all run separately as a batch job on a shared computing cluster. For each job, 16 batch threads were allocated in order to decrease runtime.

To further analyze and compare the FPKM data to the reference paper, expression levels of the significantly differentially expressed sarcomere, mitochondrial, and cell cycle genes of interest highlighted in Figure 1D by O'Meara et al. were analyzed between P0, P4, P7, and adult time points during in vivo maturation. In order to do so, the average FPKM value between the two samples for each time point for each gene was calculated. While the P0_1 FPKM table was produced by Cufflinks, the remaining FPKM tables (P0_2, P4_1, P4_2, P7_1, P7_2, Ad_1, and Ad_2) were provided in a project directory by the instructor. The figure was reproduced using ggplot2 and tidyverse packages in RStudio v1.3.1073.

Lastly, a heatmap of the top 1000 differentially expressed genes was created from P0 to adult. To select genes for this clustered heatmap, the differential expression statistics produced by Cuffdiff were rearranged in descending q-value, filtered for significance, and then sorted by log2 fold changes. For

genes with multiple occurences, the average FPKM was taken for that gene within a sample. The top 500 most up and downregulated genes were selected. The genes and samples were hierarchically clustered.

Results

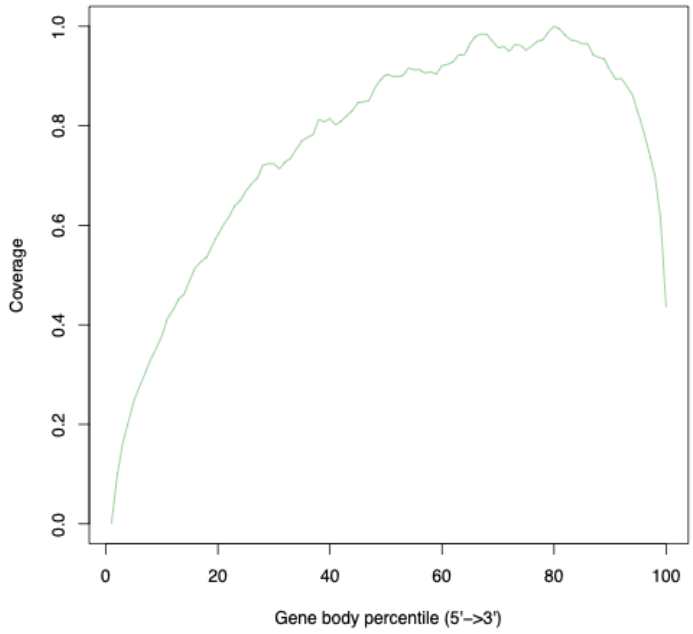

Figure 1. Coverage of RNA-sequencing reads across the gene body. The percentage of reads that covers the gene body is shown from the 5' to 3' end.

RseQC was used to assess the quality of the RNA-seq data. The first quality control metric was RNA-seq reads coverage across the gene body that was determined using the gene_Bodycoverage.py utility in RseQC. The resulting plot can indicate if there is any 3' or 5' bias (Figure 1). It is observed that the majority of genes had coverage of at least 80% with increasing coverage towards the 3' end, signifying that more reads were found in that region. Overall, this results demonstrates that there was good uniform coverage; however, 80% coverage was not achieved until the 40th percentile of the gene body, demonstrating that there is a slight 3' bias possibly due to RNA degradation as well as poly(A) enrichment of mRNA (Zhao et al., 2014). This outcome is to be expected with RNA seq data and therefore, does not require further analysis.

The mRNA length between two read pairs, otherwise known as the insert size or inner distance, was calculated by the inner_distance.py utility in RseQC. Figure 2 shows the distribution of the density of each insert size between the paired reads of P0_1. Most of the RNA reads had an inner distance of 50 - 100 base pairs. Few paired reads experienced overlap with most of the distribution greater than 1 bp. This feature indicates relative stable quality between the paired reads.

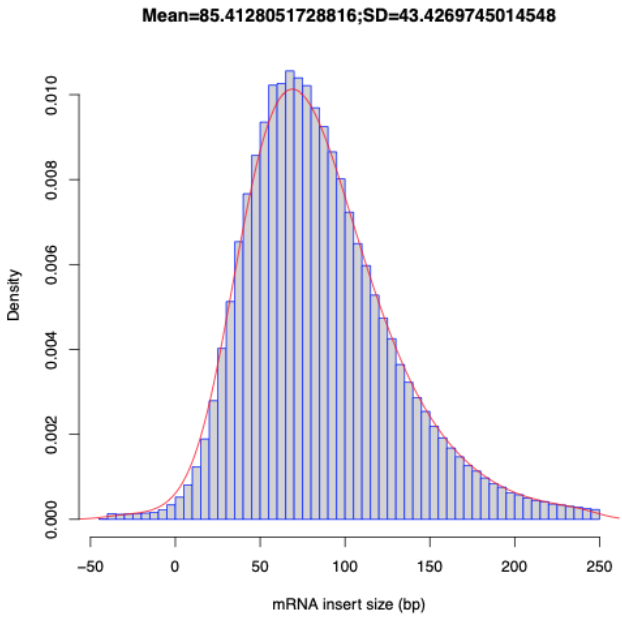

Figure 2. The density of mRNA insert size (inner distance) between read pairs. Insert size is measured in base pairs (bp).

The last quality control measure generated by RseQC were quality control statistics of the P0_1 sample (Table 1). The results were in concordance with those from SAMtools flagstat. In total, there were 49,706,999 reads in which 100% of these reads were successfully mapped. No reads failed the quality control parameters. Out of this, there were 38,489,380 uniquely mapped reads, constituting 77.4% of the total. There were 2,899,954 reads that were multi-mapped, constituting 5.8% of the total. This can occur in genomes that have multiple copies of the same gene and as such, the reads can align equally well to each of these locations. These results provide further evidence that the aligned sequence data was of good quality and did not need further processing.

Table 1. Summary statistics for BAM file generated by RseQC and SAMtools flagstat.

	Number of reads	Percentage of total reads
Total reads:	49,706,999	100%
QC failed:	0	0%
Mapped reads:	49,706,999	100%
Unmapped reads:	0	0%
Unique reads:	38,489,380	77.4%
Multi-mapped reads:	2,899,954	5.8%
Unaligned reads:	0	0%

Due to the variability of FPKM values, the FPKM values produced by Cufflinks were plotted on a log10 scale (Figure 3). FPKM is necessary to quantify differential levels of gene expression between neonatal and mature cardiac myocytes. For the P0_1 sample, FPKM values of less than one were filtered out such that only genes that are expressed are represented in the plot. Before this filter was applied, there were 37,469 FPKM values. After filtering, there were 14,205 FPKM values. It can be observed that there are few genes with distinct outlier FPKM values. Most of the data have a FPKM value between log10 base of 0 and 2.

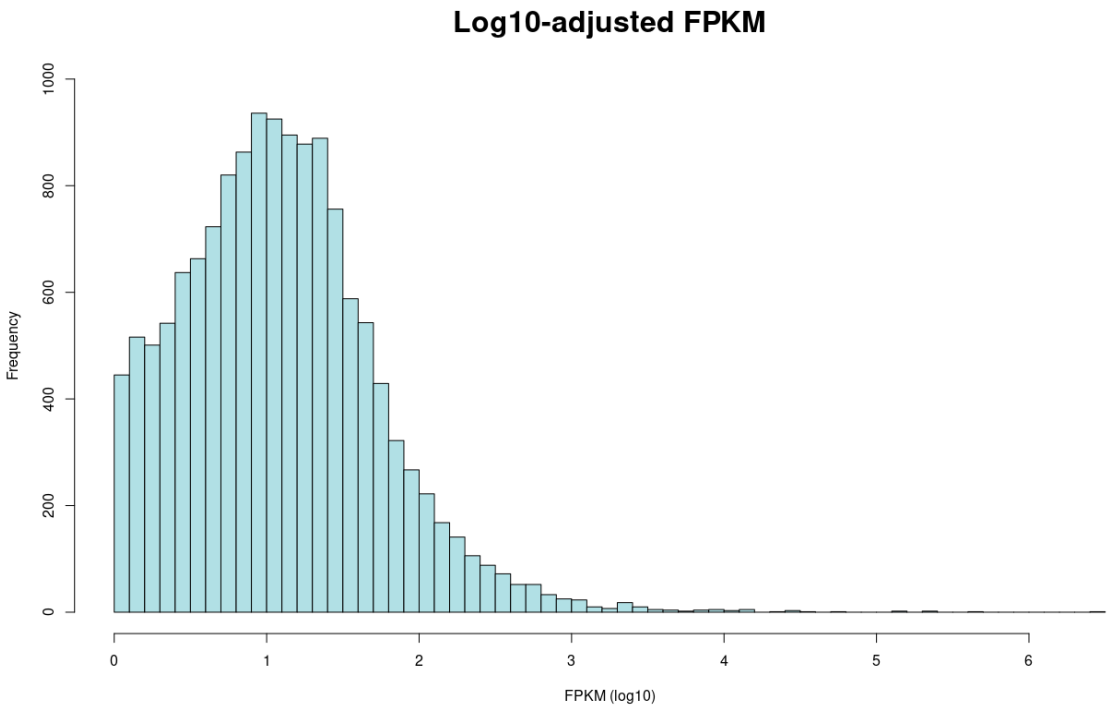

Figure 3. Relative abundance of transcripts as determined by fragments per kilobase of transcript per million (FPKM) mapped reads in the P0_1 sample. FPKM values are plotted on a log10 scale.

With the differentially expressed genes determined by Cufflinks, the FPKM values of sarcomere, mitochondrial, and cell cycle genes of interest were found to undergo different levels of regulation throughout the neonatal and adult stages (Figure 4). These genes were found to be enriched by gene ontology (GO) terms by O'Meara et al. From postnatal day 0 to adult, sarcomere genes and mitochondrial genes generally were up regulated while cell cycle genes were down regulated, consistent with findings from Figure 1D of the reference study. Throughout the different timepoints, sarcomere genes had FPKM values ranging from 0 to 1,300; mitochondrial genes, while also subjected to up regulation, had FPKM values only ranging 100 to 300 which is around up to four-fold less than sarcomere FPKM values.

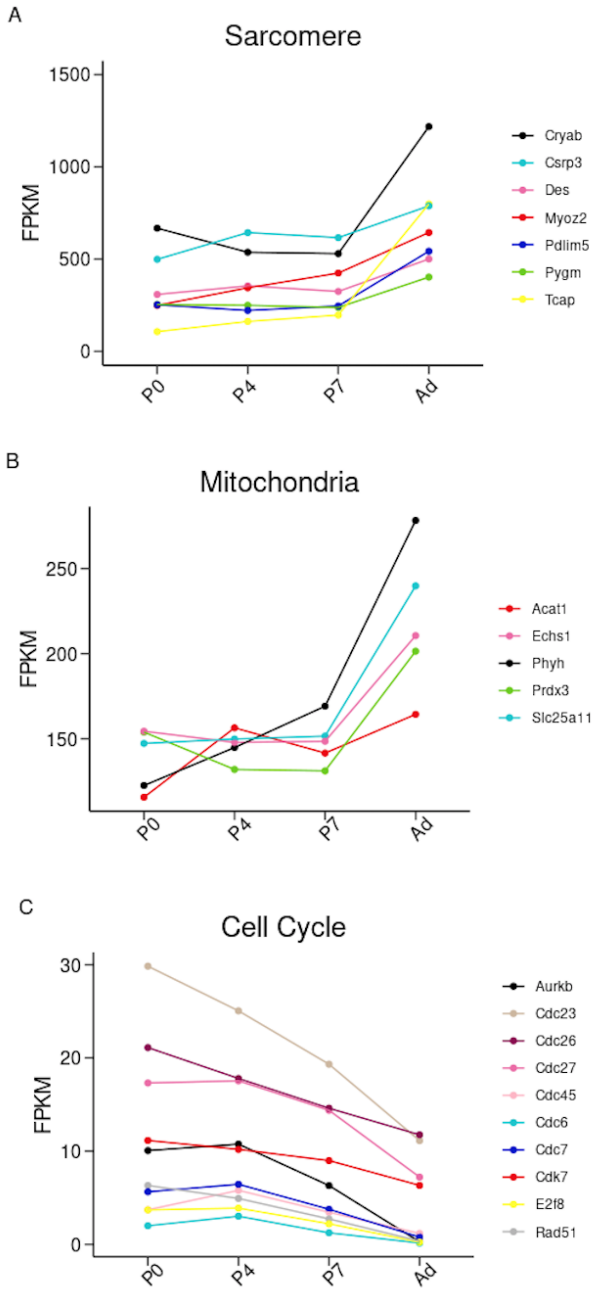

Figure 4. FPKM values of A) sarcomere, B) mitochondrial, and C) cell cycle genes of interest that are differentially expressed across different developmental stages, P0, P4, P7, and adult (Ad) during in vivo maturation.

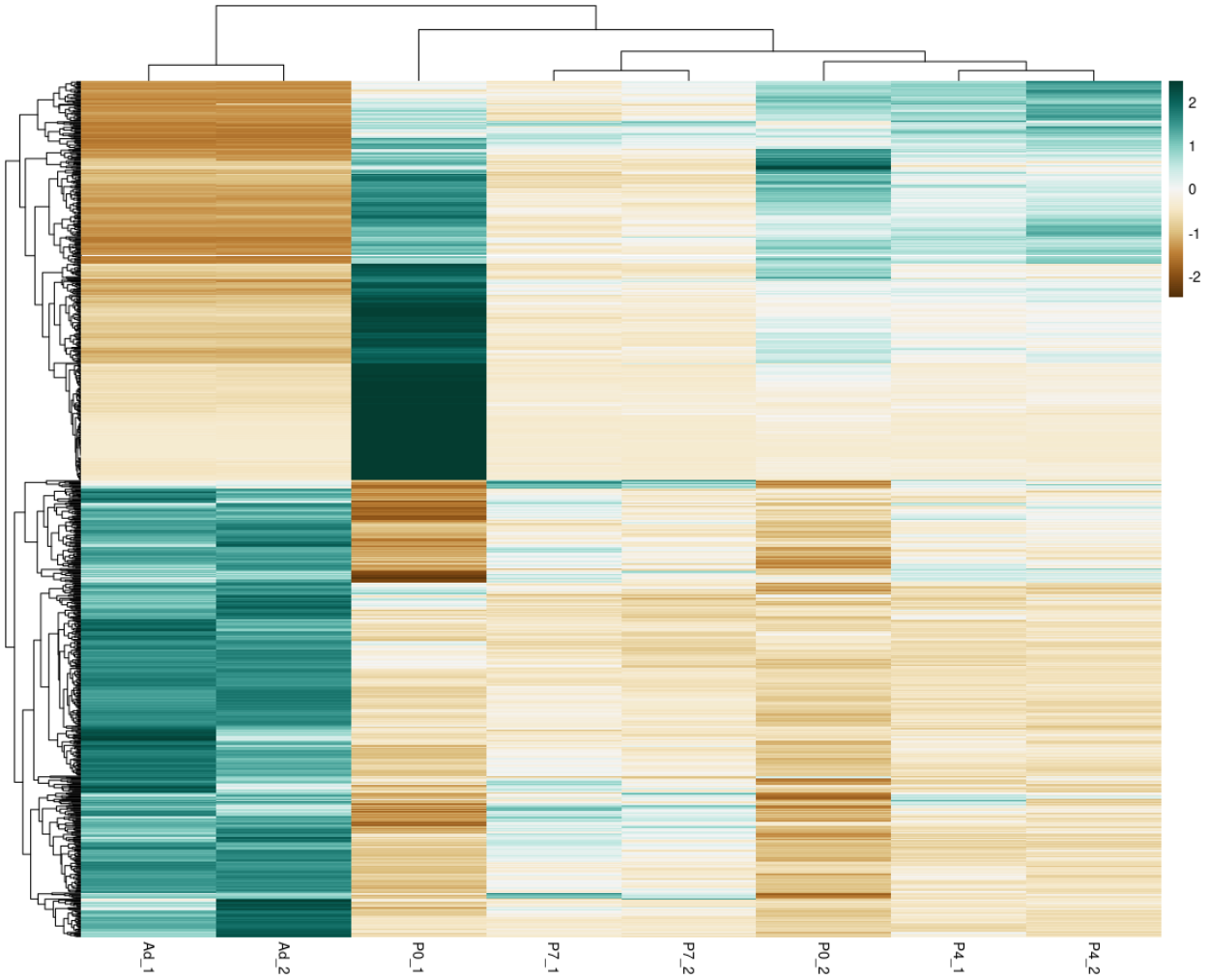

Figure 5. Clustered heatmap for the top 1,000 differentially expressed genes between P0 and adult (Ad) across all samples. Each row represents a unique gene while each column represents a replicate sample (indicated by a suffix of either 1 or 2) of each age. The level of differential expression is visualized by the color scale with darker colors indicating higher expression level. Clustering was determined by the Euclidean distance.

The top 1,000 most differentially expressed genes during the neonatal and adult stages were visualized on a heatmap (Figure 5) to provide insight into differential expression levels The most distinct patterns of gene expression can be observed between the P0 and Ad samples with genes that were markedly either up or down regulated, perhaps indicating that these genes are potential signatures of differentiation or regeneration of cardiomyocytes. Most of the replicate samples are well-paired.

Discussion

FPKM values of representative sarcomere, mitochondrial, and cell cycle genes were analyzed and compared to Figure 1D from O'Meara et al. These genes were determined by the authors to be enriched with gene ontology (GO) terms. From the P0 to adult stage, sarcomere and mitochondrial genes

experienced an increase in expression while cell cycle genes experienced a decrease in expression (Figure 4). These trends and magnitude of expression levels of the genes of interest corroborate with those identified by the original authors.

Up regulation in sarcomere-specific genes are indicative of cardiac myocyte growth and differentiation (Figure 4A). The most highly up-regulated gene, crystallin alpha B (Cryab), is involved in cytoskeletal stability of developing muscles (Maloyan et al., 2005; Wójtowicz et al., 2015). This reflects an increase in sarcomere assembly and organization during maturation of cardiac myocytes. Other genes displaying sustained and increased levels in expression are telethonin (Tcap) and cysteine and glycine-rich protein 3 (Csrp3) that are responsible for sarcomere development, myofibrillogenesis, and myogenesis (Mayans et al., 1998; Rashid et al., 2015). The mitochondrial gene (Figure 4B) displaying the most pronounced increase across the in vivo maturation stages is phytanoyl-CoA dioxygenase (Phyh) which is involved in fatty acid oxidation (Jansen et al., 1999). This up regulation can be reasoned with the growing energy demand that is required by increased saromeric organizational processes. Additionally, the decreasing trend identified in Figure 4C suggests the down regulation of cell cycle genes from the postnatal period to adult maturation. Cell cycle and DNA division regulators (E2f8) and cell cycle division genes (Cdc 6, 7, 23, 26, 27, 45) that are responsible for facilitating progression through various phases and checkpoints of the cell cycle were found to be down regulated (Dimova et al., 2005; Alberts et al., 2005). This supports cell cycle exit as a hallmark of adult cardiac myocytes.

The clustered heatmap visualizes differences in expression levels of the top 1,000 differentially expressed genes between the P0 and adult samples over all samples representing different developmental stages during in vivo maturation (Figure 5). With the exception of the P0 replicates, each pair of replicate samples at each time point have similar levels of expression across all genes. P0_1 generally has higher levels of expression than P0_2 and does not align well to each other relative to the other replicates. Regardless, the most distinct difference can be observed between the P0 and Ad samples in which genes that had high expression levels in the P0 samples were found to have low expression levels in the Ad samples and vice versa. Since the P4 and P7 developmental times points are between P0 and Ad, it can be reasoned that expression level of genes are at an intermediate level and represent a transitional shift.

Analysis of the top 10 differentially expressed genes provided a more in-depth look at the most differentially expressed genes (Figure S1). Comparing adult samples relative to P0 samples, one of the most significantly differentiated genes that is up regulated is the mitochondrial ribosomal protein L30 (MRPL30) which plays a role mitochondrial protein synthesis (Goldschmidt-Reisin et al., 1998) while one that is the most down regulated is Leucine-rich repeat flightless-interacting protein 1 (Lrrfip1) which is involved in muscle cell proliferation (Labbe et al., 2017). The direction of these transcriptional changes are in agreement with the results observed in Figure 4.

It is difficult to directly compare the heatmap in Figure 5 to the reference heatmap in Figure 2A in O'Meara et al. The authors included data from the in vivo maturation and adult cardiac myocyte explant models. By having to encompass more datasets, this likely introduced a greater range of differential expression values; the authors would have had to perform normalization on a larger scale. As such, the scaling between the two heatmaps are different and therefore, the quantitative coloring of each cell (low, medium, high) of each cell are not directly comparable. Despite this, the general changes in gene expression level throughout each developmental stage from P0 to adult still lends evidence to the directional reversion of gene expression observed between the early neonatal period and the adult stage

A potential contribution to the discrepancies observed between the heatmap produced by this current study and to that of O'Meara et al. is that the clustering may have been impacted by a fewer number of up and down regulated differentially expressed genes than those reported by the authors. This current analysis identified a total of 2,139 significantly differentially expressed genes with 1,084 genes being up-regulated and 1,055 genes being down-regulated. In contrast, O'Meara et al. produced 2,409 and 7,570 up and down regulated genes, respectively. Better reporting of parameters used to choose and curate the GO terms may result in easier reproducibility of the heatmap results.

Further plausible sources of differences is that O'Meara et al. used all samples from both the in vivo maturation and in vitro differentiation models while this current project only examined in vivo maturation samples to perform hierarchical clustering and analyze changes in regulation of sarcomere, mitochondrial, and cell cycle genes. O'Meara et al. performed experiments using in vivo, in vitro, and explant models. In vivo models can better identify changes in transcriptional changes in gene expression within a natural environment. In vitro models, while performed outside of a living organism, enable closer analysis of cardiac myocyte differentiation in a controlled environment without confounding effects. The adult cardiomyocyte explant model permitted the authors to observe dynamic changes that take place upon loss of the mature cardiomyocyte phenotype. Collectively, the usage of these different models provide a more comprehensive and multi-disciplinary understanding of various candidate regulators and molecular signatures of transcriptional reversion. Even though this current study performed analyses using data only from the in vivo model, the results generally demonstrated robust concordance with those of O'Meara et al.

Conclusion

This current analysis sought to determine differences in the transcriptional signatures between neonatal and adult mice in order to identify regulators in heart regeneration and cardiac differentiation. Overall, the results of this project generally corroborate with those of O'Meara et al., suggesting an exit of the cell cycle by mature adult cardiac myocytes. The opposing directional changes observed by neonatal and adult samples provide further evidence of a transcriptional reversion experienced by cardiac myocytes to a more immature and less differentiated state upon injury through re-entry into the cell cycle. Reproducibility could have been facilitated by having access to data from all of the models used in the study along with more detailed and transparent descriptions of tools and parameters used for processing the RNA-seq data. Ultimately, these findings have great clinical potential to further current understanding of cardiac repair and help provide better targeted solutions for diagnosing and treating known cardiac diseases.

References

Alberts B, Johnson A, Lewis J, et al. Molecular Biology of the Cell. 4th edition. New York: Garland Science; 2002. An Overview of the Cell Cycle. Available from: https://www.ncbi.nlm.nih.gov/books/NBK26869/

Dimova, D., Dyson, N. The E2F transcriptional network: old acquaintances with new faces. *Oncogene* 24, 2810–2826 (2005). https://doi.org/10.1038/sj.onc.1208612

Goldschmidt-Reisin S, Kitakawa M, Herfurth E, Wittmann-Liebold B, Grohmann L, Graack HR. Mammalian mitochondrial ribosomal proteins. N-terminal amino acid sequencing, characterization, and identification of corresponding gene sequences. J Biol Chem. 1998 Dec 25;273(52):34828-36. doi: 10.1074/jbc.273.52.34828. PMID: 9857009.

Jansen, G. A., Ofman, R., Denis, S., Ferdinandusse, S., Hogenhout, E. M., Jakobs, C., & Wanders, R. J. (1999). Phytanoyl-CoA hydroxylase from rat liver. Protein purification and cDNA cloning with implications for the subcellular localization of phytanic acid alpha-oxidation. Journal of lipid research, 40(12), 2244–2254.

Labbé, P., Faure, E., Lecointe, S., Le Scouarnec, S., Kyndt, F., Marrec, M., Le Tourneau, T., Offmann, B., Duplaà, C., Zaffran, S., Schott, J. J., & Merot, J. (2017). The alternatively spliced LRRFIP1 Isoform-1 is a key regulator of the Wnt/β-catenin transcription pathway. *Biochimica et biophysica acta. Molecular cell research*, *1864*(7), 1142–1152. https://doi.org/10.1016/j.bbamcr.2017.03.008

Langmead, B., Salzberg, S. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9, 357–359 (2012). https://doi.org/10.1038/nmeth.1923

Maloyan, A., Sanbe, A., Osinska, H., Westfall, M., Robinson, D., Imahashi, K., Murphy, E., & Robbins, J. (2005). Mitochondrial dysfunction and apoptosis underlie the pathogenic process in alpha-B-crystallin desmin-related cardiomyopathy. *Circulation*, *112*(22), 3451–3461. https://doi.org/10.1161/CIRCULATIONAHA.105.572552

Mayans, O., van der Ven, P. F., Wilm, M., Mues, A., Young, P., Fürst, D. O., Wilmanns, M., & Gautel, M. (1998). Structural basis for activation of the titin kinase domain during myofibrillogenesis. *Nature*, *395*(6705), 863–869. https://doi.org/10.1038/27603

Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, Heng Li, Twelve years of SAMtools and BCFtools, *GigaScience*, Volume 10, Issue 2, February 2021, giab008, https://doi.org/10.1093/gigascience/giab008

Porrello, E. R., Mahmoud, A. I., Simpson, E., Hill, J. A., Richardson, J. A., Olson, E. N., & Sadek, H. A. (2011). Transient regenerative potential of the neonatal mouse heart. *Science (New York, N.Y.)*, *331*(6020), 1078–1080. https://doi.org/10.1126/science.1200708

Rashid, M., Runci, A., Russo, M. *et al.* Muscle Lim Protein (MLP)/CSRP3 at the crossroad between mechanotransduction and autophagy. *Cell Death Dis* 6, e1940 (2015). https://doi.org/10.1038/cddis.2015.30

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., & Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology, 28(5), 511–515. https://doi.org/10.1038/nbt.1621

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., & Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. Nature biotechnology, 31(1), 46–53. https://doi.org/10.1038/nbt.2450

Wang L, Wang S, Li W. RSeQC: quality control of RNA-seq experiments. Bioinformatics (Oxford, England). 2012 Aug;28(16):2184-2185. DOI: 10.1093/bioinformatics/bts356.

Wójtowicz, I., Jabłońska, J., Zmojdzian, M., Taghli-Lamallem, O., Renaud, Y., Junion, G., Daczewska, M., Huelsmann, S., Jagla, K., & Jagla, T. (2015). Drosophila small heat shock protein CryAB ensures structural integrity of developing muscles, and proper muscle and heart performance. *Development (Cambridge, England)*, *142*(5), 994–1005. https://doi.org/10.1242/dev.115352

Zhao, W., He, X., Hoadley, K.A. *et al.* Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. *BMC Genomics* 15, 419 (2014). https://doi.org/10.1186/1471-2164-15-419

Supplementary Materials

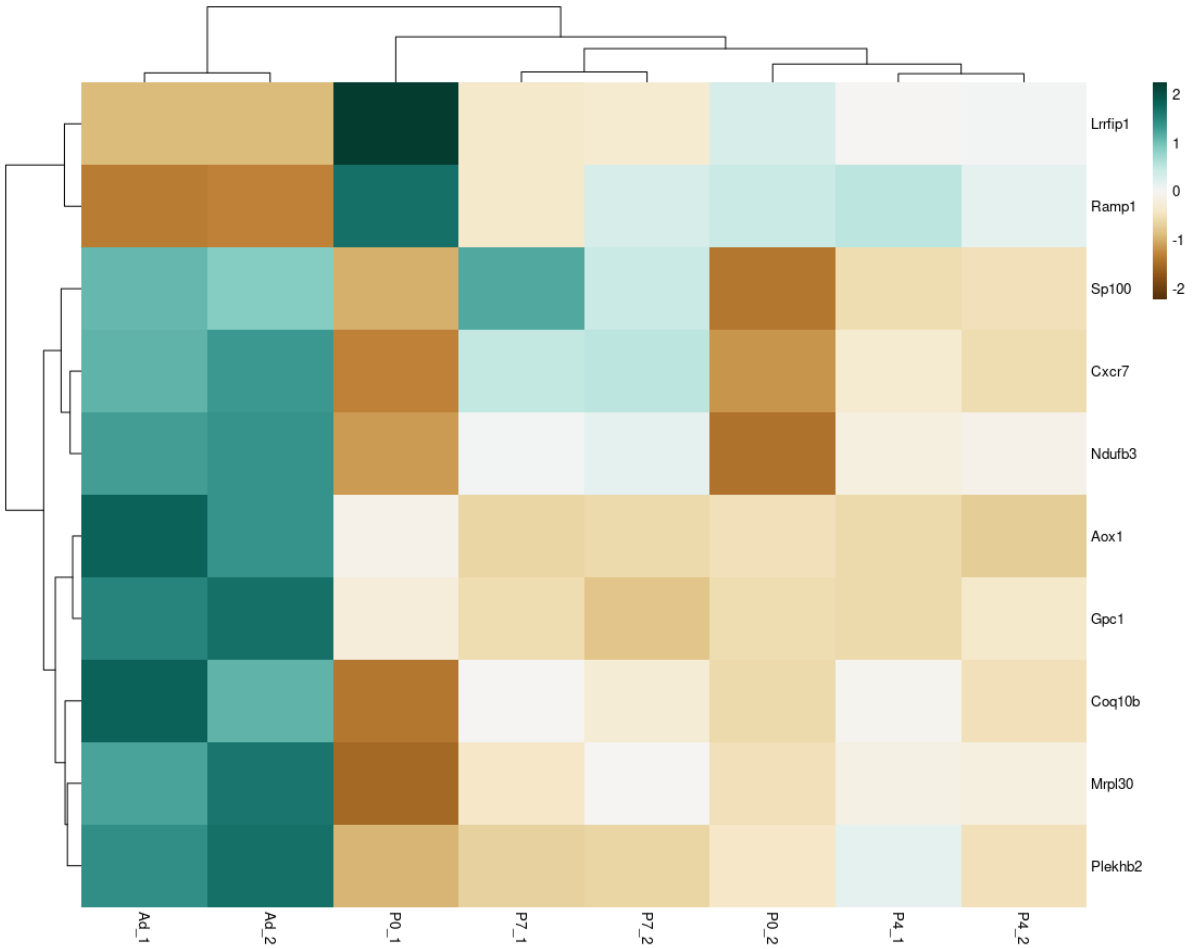

Figure S1. Hierarchically clustered heatmap for the top 10 differentially expressed genes between P0 and adult (Ad) across all samples.