Project 2 Write-up

Abhishek Thakar (Biologist), Allison Nau (Data Curator), Mae Rose Gott (Programmer), and Sheila Yee (Analyst)

Introduction

The leading cause of death in the United States is heart disease according to the CDC. Factors such as high cholesterol, smoking, high blood pressure and others can potentially lead to cardiovascular disease, strokes, or heart attacks (Fryar 2012). The prevalence of such a major cause of death necessitates an intervention or path to medical regeneration of heart tissue to aid in rehabilitation. Research into repairing the human heart after injury indicates minimal capacity to heal damaged areas itself. However, in studying neonatal mice, which have the capability to fully regenerate their heart at an early stage of life would provide molecular signatures or medical information to develop life-extending discoveries in humans.

Neonatal mice possess the remarkable ability to restore their hearts following trauma, provided it is in the first week of infancy. The transcriptional system behind the cardiac healing has not yet been fully analyzed and validated. It is important to determine whether the cardiac myocytes possibly revert and return to a less differentiated phase for this occurrence at a transcriptional level. Utilizing transcriptional data, it is possible to identify molecular regulators within this pathway.

Genetic fate mapping is a method that is used to identify the origin of adult tissues and cells. In neonatal mice, the cardiac myocytes present in the left ventricular apex originate from prior cardiac myocytes rather than stem cells. The expression of phosphorylated histone H3 (pH3) combined with the up regulation of aurora B kinase indicate potential for cell regeneration (Porello et al. 2011).

The global gene expression patterns are modeled in vitro and in vivo, from embryonic stem cells to differentiation into cardiomyocytes, and neonate cardiac myocytes to adult, respectively. The comparison of those signatures to an explant model and a tissue resection model, provided the basis to observe the transcriptional reversal process in cardiac myocytes. The RNAseq datasets helped identify regulators such as interleukin 13 (IL13) as well as mediators such as STAT6, STAT3, and periostin in the cell cycle phase within these myocytes.

In order to elucidate the different transcriptional profiles between newborn mice (P0) and adults (8-10 weeks old), O'Meara et al analyzed short mRNA sequences to compare differences in gene expression level and identify significantly differentially expressed genes. These genes were subjected to functional annotation clustering through DAVID (Database for Annotations, Visualization, and Integrated Discovery). The authors also created a clustered heatmap of differentially expressed genes comparing in vivo, in vitro, and explant cardiac myocytes, which utilizes hierarchical clustering to better visualize the data matrix of results.

Data

We looked at 8 cardiomyocyte samples isolated from whole heart ventricles, 2 samples each from postnatal day 0 mice (P0), postnatal day 4 mice (P4), postnatal day 7 mice (P7), and 8-10 week old adult mice (Ad) (O'Meara, 2015; NCBI GEO accension GSE64403 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE64403). Each replicate was pooled from at least two heart ventricles for adults; five to ten pups were pooled for each postnatal sample. For mouse pups, cardiac myocytes were isolated by using the Neonatal Heart Dissociation Kit and Neonatal Cardiac Myocyte Isolation kit from Miltenyi. Adults were processed as described in Liao and Jain 2007. Total RNA was extracted using Trizol (Invitrogen), Polyadenylated RNA was fragmented, then reverse transcription was performed and double stranded DNA synthesized. Beckman Coulter's SRPI-Works system was used to finish processing the RNA followed by the addition of barcodes. 40 base pair long paired-end sequences were sequenced on an Illumina HiSeq 2000. Sequences were aligned to the mm9 mouse reference genome. For the 8 samples we looked at, the library size was between 16,136,245 and 25,665,192, average 21,609,428.9. Quality control measures taken prior to downstream analysis were not explicitly stated by O'Meara et al.

Methods

We first took a closer look at one of the RNAseq samples, GSM1570702 (vP0_1), a cardiomyocyte sample isolated from whole heart ventricles from postnatal day 0 mice. We extracted the paired FASTQ files from the Short Read Archive file using fastq-dump in the SRAtoolkit module on SCC (sratoolkit v 2.10.5). We then assessed the quality of the sequence using FastQC on SCC (fastqc v 0.11.7). FastQC is a quality analysis tool for unmapped FASTQ sequencing files that performs a quick check on common areas of concern, and then presents those checks graphically. These checks do not consider the specific details of a given experiment (such as the expected GC content for that species' genome), so the results must be interpreted critically. These two steps together took less than 10 minutes.

After sequence quality was assessed, the reads were aligned via the TopHat v2.0.0 algorithm using the mm9 mouse genome reference, which took 1 hour to complete. The accepted hits were then indexed using Samtools v1.10, Boost v1.75.0, and Bowtie2 v2.4.2. Indexing took less than 10 minutes. Once indexed, three quality control metrics were performed within the RSeQC v4.0.0 module using Python v3.9.0, geneBody_coverage.py, inner_distance.py, and ban_stat.py, taking 2 hours to complete. These metrics visualized the coverage of the short reads on the genome, measured and graphed the distance between the paired-end reads on the genome, and gave statistics on the reads.

Module Cufflinks v2.2.1 was then used to quantify FPKM (fragment per million mapped reads per kilobase exon) values for the genes, which took about 14 minutes. Cuffdiff, a tool within Cufflinks v 2.2.1, was then run to identify differentially expressed genes between the samples P0_1, P0_2, Ad_1, and Ad_2 and took 2.5 hours to finish. The FPKM values generated by Cufflinks was then loaded into R 4.0.3 and used to generate a series of histograms.

The file generated by cuffdiff containing differentially expressed genes and gene statistics between postnatal day 0 (P0) and adult (Ad) mice was subject to further analysis to determine relationships with myocyte differentiation. This file was loaded onto R 4.0.3 and sorted such that the genes with the smallest q-values were at the top of the gene list. A table of the top 10 differentially expressed genes was generated. Additionally, a histogram of the log2 fold change of all genes was also created. The list of genes was further subdivided to create a new list of genes that was denoted as "significant" in the dataframe. Significance was determined by identifying p-values greater than FDR after Benjamini-Hochberg correction for multiple hypothesis testing (Trapnell et al., 2012). From this, another histogram was produced that displayed the log2 fold change of only these significant genes. This list of significant genes was further subsetted into up-regulated and down-regulated genes and saved into two separate csv files.

The up- and down-regulated genes were separately organized into functionally related clusters using DAVID (Database for Annotations, Visualization, and Integrated Discovery) Functional Annotation Clustering version 6.8 (Huang et al. 2009). For both sets of analysis, the species of interest was *Mus musculus*. Gene ontology groups of interest were GOTERM_BP_FAT, GOTERM_MF_FAT, and GOTERM_CC_FAT. Through DAVID, enriched GO terms organized into clusters based on functional relatedness were returned for both up- and down-regulated genes.

The reproduction of Figure 1D in vivo maturation used the FPKM tables from the samples (P0, P4, P7, Adult) where P0 was the dataset that was generated from previous steps and processed for data manipulation. The specific genes mapped to the GO terms presented in the study were plotted to compare to the original plots. The sample genes.fpkm_tracking tables were provided in project directories (Ad_1, Ad_2, P0_2, P4_1, P4_2, P7_1, P7_2) while one replicate was provided through the cuffdiff output from the programmer. Using ggplot2 and tidyverse packages the plots were created for the respective sarcomere, mitochondrial, and cell cycle genes presented in Figure 9A-C. To create a heatmap on the top 100 differentially expressed genes, an FPKM matrix was created using the same tracking tables from the previous analysis. A data frame was created by filtering on the significant column and the log2.fold_change so that a clustered heatmap could be created based on the largest positive and negative log fold change values, for unique genes. The top 100 were selected to help draw better conclusions as seen in Figure 9D and compare with Figure 2A from the original article.

Results

The sample we took a closer look at, GSM1570702 (P0_1), passed most of the FastQC measures and the metrics were generally well aligned for both directions (P0_1_1 and P0_1_2). There were 21,577,562 sequences in this sample. The results for the sequence pair were generally well aligned, with P0_1_2 having slightly lower quality scores and a bit more "streaking" in the per tile sequence quality. These two metrics were still in the high quality range, and may indicate that the sequencer was performing (slightly) sub-optimally for P0_1_2. Quality scores for each base were greater than or equal to 30 in P0_1_1, and greater than or equal to 29 in P0_1_2, and therefore in the high quality range (Figure 1A). Per tile sequence quality was good for both directions, with a few tiles as an outlier in P0_1_1 and some minor streaking in P0_1_2 (Figure 1B). Per sequence quality scores were good (Figure 1C). Per base sequence content failed, with the first 14 bases having some sort of nucleotide biases; the same biases were observed for sequences in the pair (Figure 1D). Bases 15-40 do not show these types of biases. Since this was observed at the start of the sequence, it could indicate biases in the library preparation, which is a known phenomenon in RNA-seq library preparation (HBC, 2018). Sequence duplication was also flagged as a potential concern, with some sequences duplicated anywhere from >10 duplicates to >1k duplicates (Figure 1E). If deduplicated, 50.29% and 51.82% of sequences would remain for P0_1_1 and P0_1_2, respectively. Given this was an RNA-seq experiment looking at gene expression levels, and not a whole genome sequencing experiment, some sequence duplication is expected. FastQC did not list any overrepresented sequences, indicating that no sequence represented 0.1% or more of the sample (Michigan State Tutorial). GC content was 49%, the mouse genome has a GC content of 42% (Ruvinsky, 2005; BNID 102409). The GC distribution over all sequences was similar to the theoretical normal distribution, with a slight shift (Figure 2). In RNA-seq, a GC content distribution that does not perfectly match a normal distribution centered on the mean GC content for the species' genome does not necessarily indicate an issue, but does warrant closer evaluation (Galaxy Training). No sharp peaks in the GC content distribution were observed that would indicate potential contamination. Per base N content and adapter content were both negligible. All sequences were 40bp long.

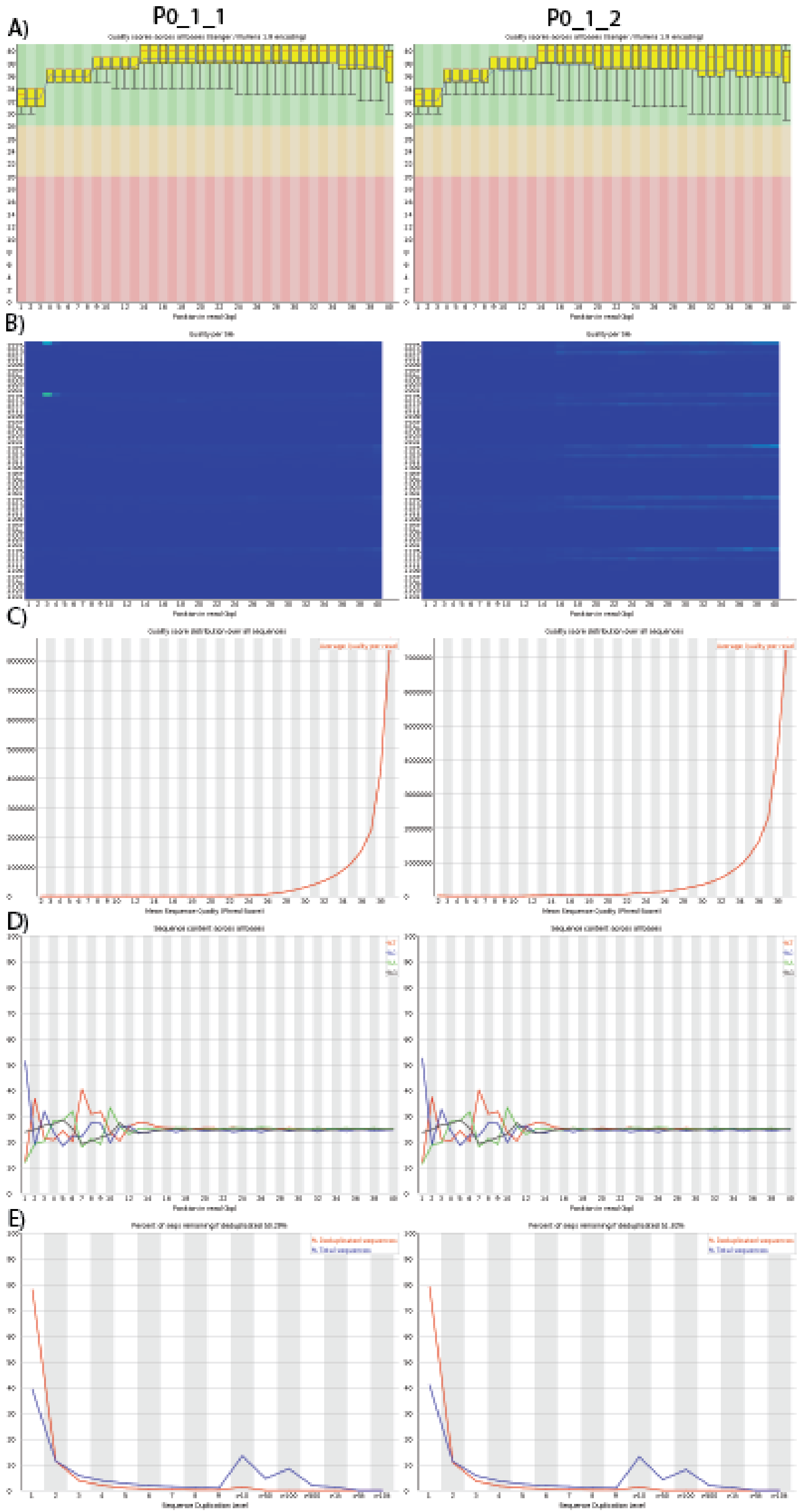


Figure 1. FastQC metrics for P0_1, from both members of the sequencing pair. P0_1_1 is on the left; P0_1_2 is on the right. A) Quality scores across all bases. B) Quality per tile. C) Quality

score distribution over all sequences. D) Sequence content biases across all bases. E) Sequence duplication level.

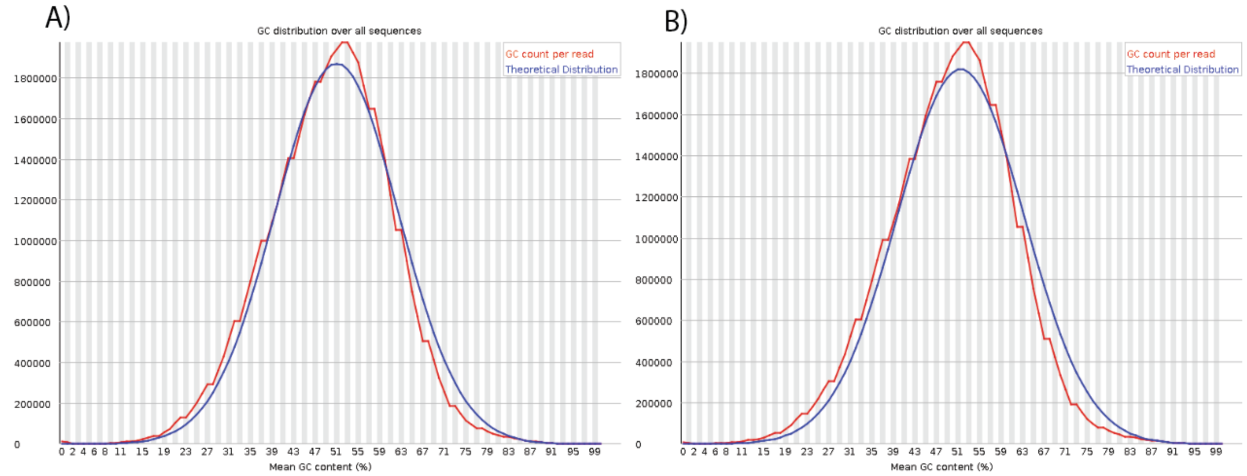


Figure 2. GC content distribution in P0_1. A) GC content distribution in P0_1_1. B) GC content distribution in P0_1_2.

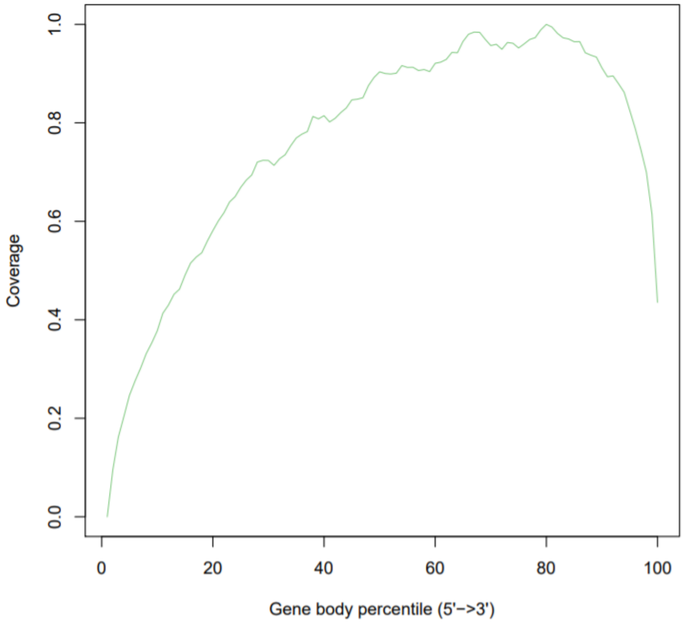


Figure 3. Output from geneBody_coverage.py run on sample P0_1.

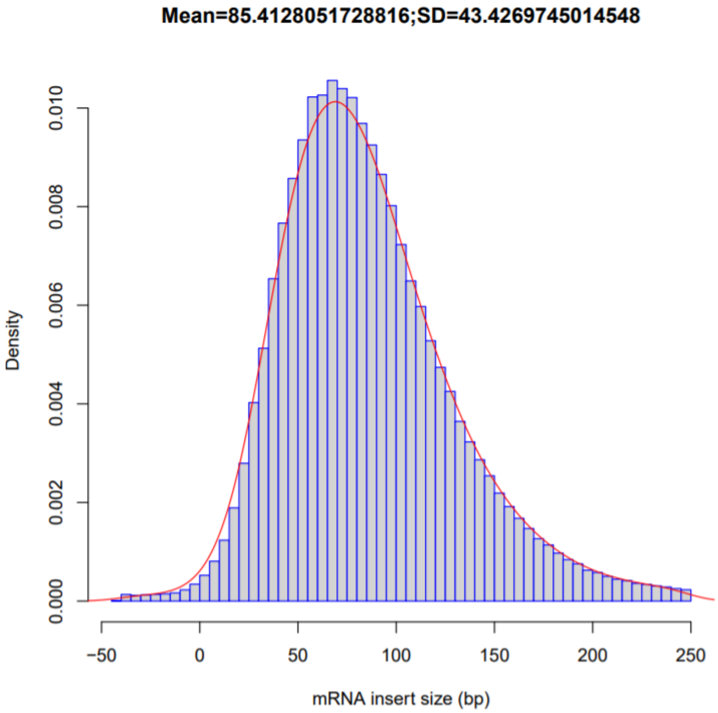


Figure 4. Output from inner_distance.py run on P0_1 short reads to show the density of the mRNA insert size in number of base pairs. Mean size reported was 85.4 base pairs with a standard deviation of 43.4

Quality control metrics run on RSeQC for the P0_1 sample recorded 49,706,999 total records in the bam_stat.py output. QC failed for 0 reads, with 0 reads having Optical/PCR duplicate and 0 unmapped reads. Of the reads, 77.43% (38,489,380) were unique. 38.70% (19,236,824) reads were mapped to '+,' 38.74% (19,252,556) reads were mapped to '-,' and 56.28% (27,972,916) reads were mapped in proper pairs.

Output from geneBody_coverage.py run on sample P0_1 skewed towards 3', though the experimental data generally covered the gene body well (Figure 3). Average size of the paired-end reads was 85 base pairs, with a standard deviation of 43 (Figure 4).

FPKM values had a wide range of values, from 0 to over 15000. Most FPKM values were on the lower-end of each spectrum (Figure 5, 6A, 6B, 6C, 6D). To break down the distribution of FPKM values, data was then graphed by order of magnitude. Figure 6A displays the first order of magnitude, 6B displays the second, 6C displays the third, and 6D displays everything larger than three orders of magnitude.

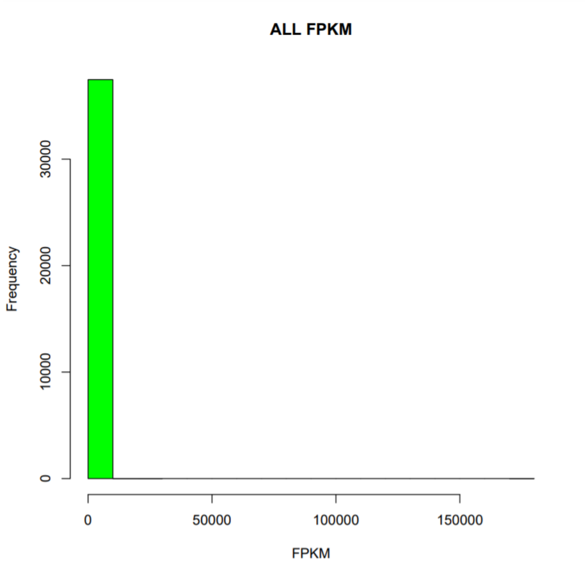


Figure 5. Table of the frequency of the FPKM values from running cuffdiff of P0_1, P0_2, Ad_1, and Ad_2.

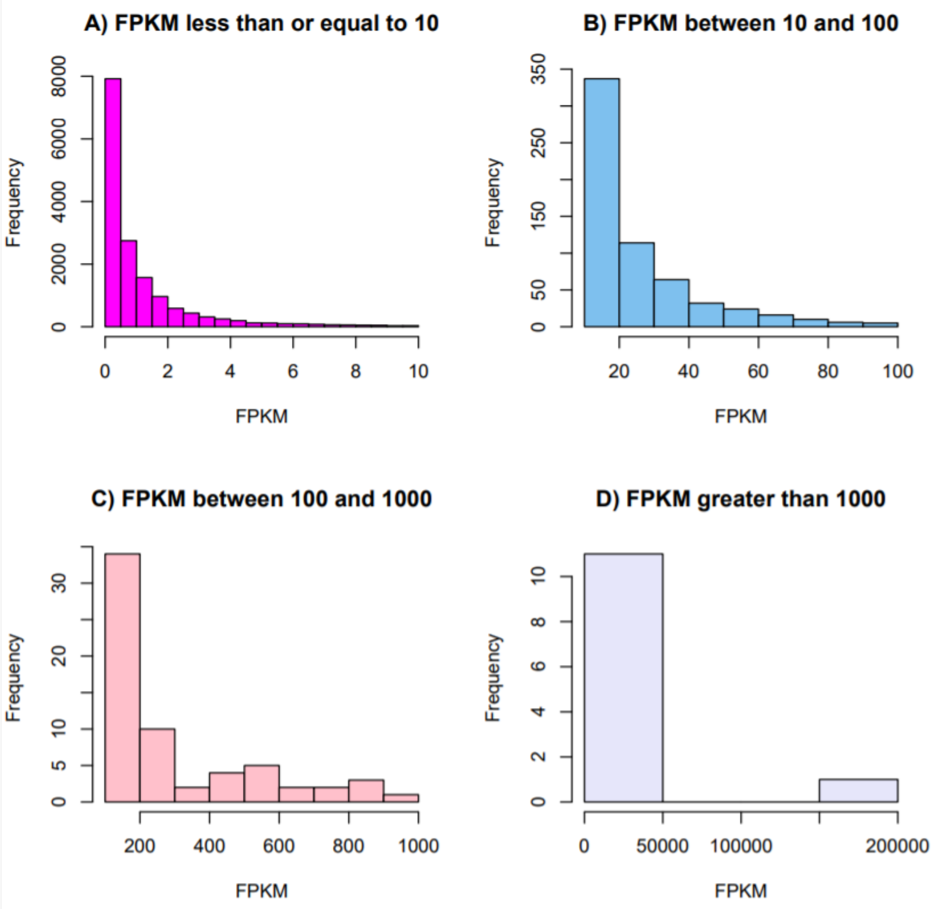


Figure 6. FPKM values broken down by order of magnitude. A) reports FPKM values up to 10e1, B) reports FPKM values between 10e1 and 10e2, B) reports FPKM values between 10e2 and 10e3, and D) reports FPKM values greater than 10e3.

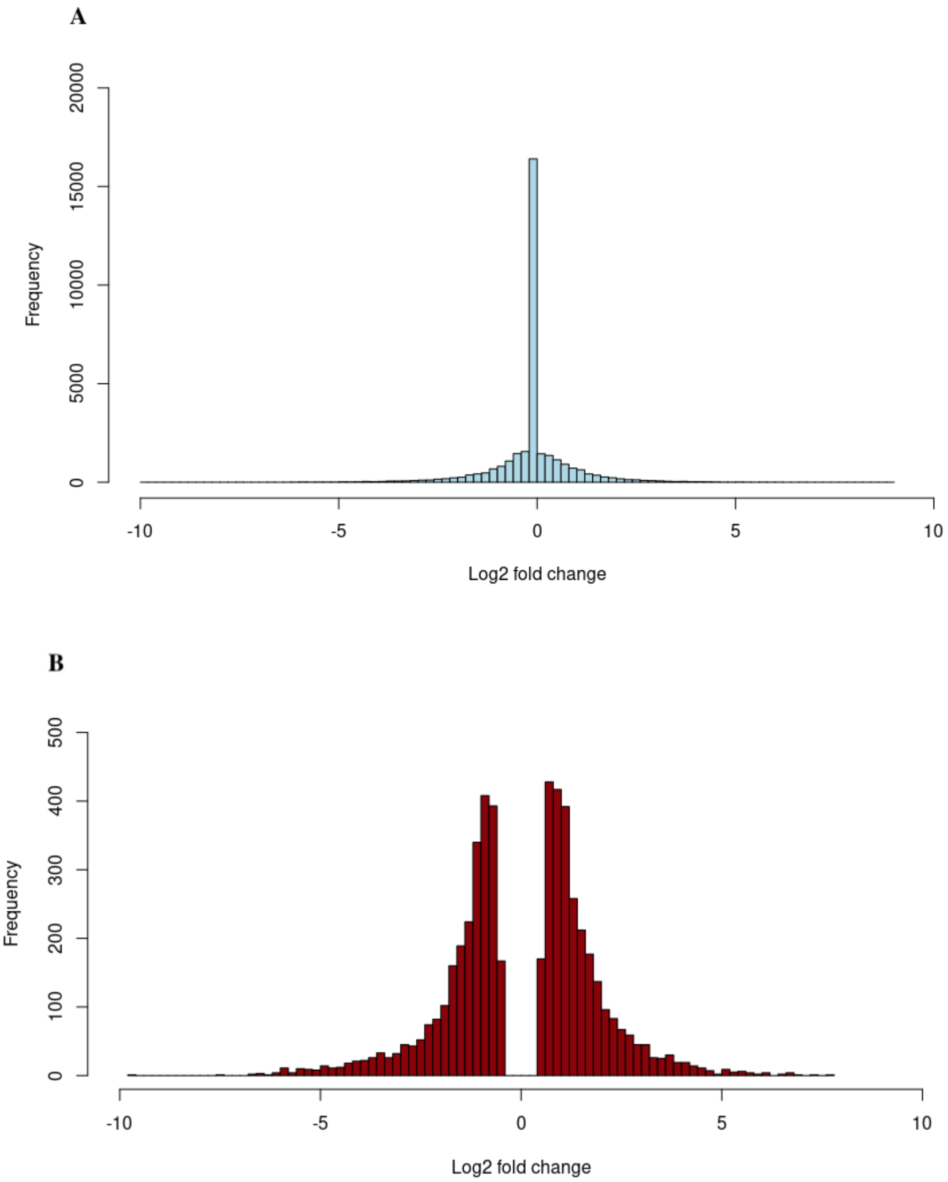


Figure 7. Log2 fold distribution. A) Distribution of all differentially expressed genes between P0 and adult mice. B) Distribution of all significantly differentially expressed genes between P0 and adult mice.

Log2 fold distribution is a measure that quantifies how a gene's expression changes between two conditions. Therefore, it measures the under and over-expression of a gene, and in the context of this study, it measures the change in gene expression between P0 and adult mice in myocyte differentiation. A large distinct peak of a log2 fold distribution of zero for all differentially expressed genes was observed and demonstrated that a large proportion of differentially expressed genes were not significant (Figure 7A). However, after removal of insignificant genes, this peak was not observed for significantly differentially expressed genes (Figure 7B). Instead, we can see that there are distinct genes with a log2 fold change of less than zero, indicating down-regulation, and distinct genes with a log2 fold change of greater than zero, indicating up-regulation.

Table 1. Top ten differentially expressed genes between P0 and adult mice as determined by the lowest q-values. P0 FPKM is fragment per kilobase million for postnatal day 0, adult FPKM is fragment per kilobase million for adults, Log2 fold change is log2(Adult FPKM/P0 FPKM).

	Gene	**P0 FPKM**	**Adult FPKM**	**Log2 fold change**	**p-value**	**q-value**
1	Adhfe1	12.71	25.74	1.02	5e-05	0.000320557
2	Tmem70	36.94	80.96	1.13	5e-05	0.000320557
3	Gsta3	0.41	6.77	4.04	5e-05	0.000320557
4	Lmbrd1	6.59	12.68	0.95	5e-05	0.000320557
5	Dst	19.72	51.60	1.39	5e-05	0.000320557
6	Plekhb2	25.85	68.60	1.41	5e-05	0.000320557
7	Cox5b	505.41	881.80	0.80	5e-05	0.000320557
8	Mrpl30	56.21	124.29	1.14	5e-05	0.000320557
9	Tmem182	46.22	103.52	1.16	5e-05	0.000320557
10	Nck2	12.17	6.30	-0.95	5e-05	0.000320557

Statistics for differentially expressed genes were sorted by ascending q-value. The top ten differentially expressed genes with the smallest q-value are listed in table 1.

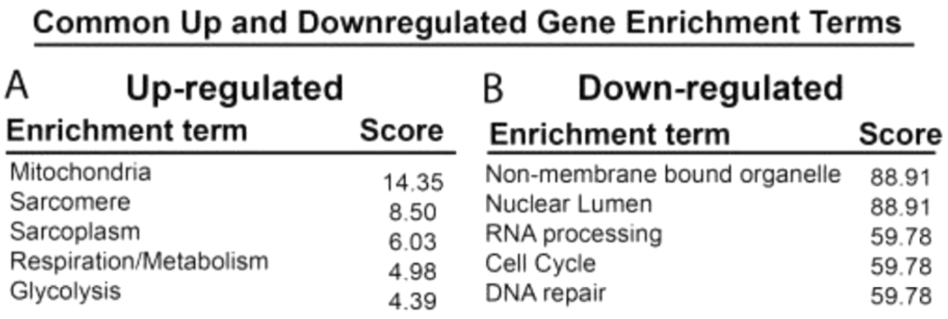
Table 2. The number of genes that were significantly up- and down-regulated ($p < 0.01$) in adult mice relative to P0.

Differentially expressed genes (p < 0.01)	
Up-regulated genes	2,830
Down-regulated genes	2,597
Total genes	5,427

Out of 36,329 differentially expressed genes, the total number of significantly differentially expressed genes ($p < 0.01$) determined by our analyses was 5,427 genes (Table 2). Out of this total, 2,830 genes were found to be up-regulated and therefore, with a log2 fold change greater than zero while 2,597 genes were found to be down-regulated and therefore, with a log2 fold change less than zero.

Table 3. The top enriched Gene Ontology (GO) terms for significant up- and down-regulated genes associated with myocyte differentiation. These genes were clustered based on functional relatedness. Terms highlighted in green indicate overlap with GO terms reported by O'Meara et al.

	Up-regulated genes		**Down-regulated genes**	
Cluster	**GO enrichment term**	**Enrichment score**	**GO enrichment term**	**Enrichment score**
1	Mitochondrion	52.75	Cell cycle	38.22
2	Purine nucleoside metabolic process	23.77	Nucleic acid binding	28.12
3	Mitochondria protein complex	23.24	chromosome	25.02
4	Organic acid metabolic process	19.49	Chromosome organization	20.21
5	Extracellular organelle	16.10	DNA metabolic process	18.97



Common Up and Downregulated Gene Enrichment Terms

A Up-regulated

Enrichment term	Score
Mitochondria	14.35
Sarcomere	8.50
Sarcoplasm	6.03
Respiration/Metabolism	4.98
Glycolysis	4.39

B Down-regulated

Enrichment term	Score
Non-membrane bound organelle	88.91
Nuclear Lumen	88.91
RNA processing	59.78
Cell Cycle	59.78
DNA repair	59.78

Figure 8. The top up- and down-regulated enrichment terms in Figure 1C from O'Meara et al.

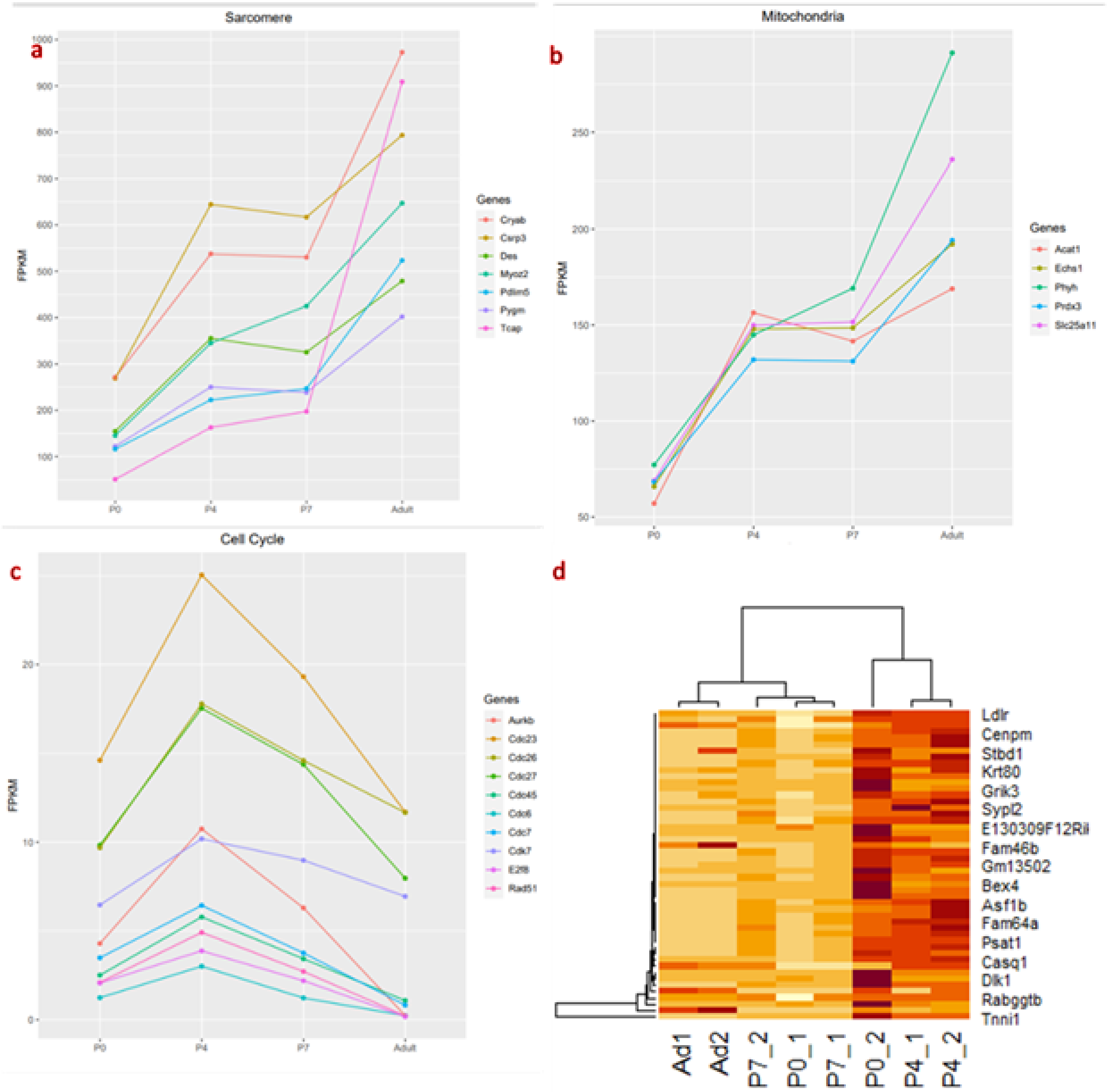


Figure 9. FPKM values of genes found in a) sarcomere, b) mitochondria, and c) cell cycle for in vivo maturation replicated to compare between Figure 1D from the article. d) is a clustered heatmap to visualize the top 100 differentially expressed genes compared to Figure 2A from the article.

Discussion

Differential gene expression profiles between P0 neonatal and adult mice were examined using DAVID. Regarding up-regulated genes, GO enrichment terms that overlapped with O’Meara et al. were mitochondrion and mitochondria protein complex, enrichment terms important for myocyte differentiation and maturation (Table 3). Some other observed GO enrichment terms for up-regulated genes were purine nucleoside metabolic process and organic acid metabolic process. While these terms were not observed by the original authors, it could potentially be reasonably inferred that they fall under the original author's umbrella enrichment term of respiration/metabolism. Regarding down-regulated genes, the GO enrichment term that overlapped with O'Meara et al was cell cycle (Table 3). Again, it can be inferred that some of our GO enrichment terms for certain down-regulated genes can be related to those in O'Meara et al. such as DNA metabolic process with the author's reporting of DNA repair as a down-regulated gene enrichment term. The enrichment scores determined by our analysis were also not the same as those reported by O'Meara et al. (Table 3 and Figure 8). However, in short, some GO enrichment terms were found to overlap with O'Meara et al. as observed with the up-regulation of mitochondrial genes and down-regulation of cell cycle genes. It is reflective of the inability of adult mycotes' limited regenerative capacity and neonatal mycotes' ability to re-enter the cell cycle to proliferate.

The discrepancies observed between these two DAVID analyses can be due to several reasons. First, the authors did not report GO IDs in their main findings (Figure 8). It appears that either the GO terms were collapsed into keywords, the authors manually curated their results, or only the broadest category GO terms were reported. In contrast, our GO terms were much more specific and descriptive when describing each enrichment term. Additionally, the authors used genes from both in vitro and in vivo differentiation whereas we only used two of the samples that they produced from the in vivo maturation model. While the version of DAVID that the authors used was not provided, given that the original study was published in 2014, the version of DAVID since then has been updated resulting in loss of access to older versions. Our statistical thresholds were also likely different for certain steps, especially when the O'Meara et al did not explicitly describe their methodology. For example, to determine differentially expressed genes, the authors used a q-value of 0.05 while we used a cut-off based on p-values greater than FDR after Benjamini-Hochberg correction for multiple hypothesis testing.

Furthermore, there were observed differences between the number of up- and down-regulated genes determined by O'Meara et al. and our reproduction. In O'Meara et al, there were 2,409 up-regulated and 7570 down-regulated genes. In contrast, we discovered 2,830 up-regulated and 2,597 down-regulated genes (Table 2). Given that there is almost a three-fold difference between the number of down-regulated genes, this may explain why there is less agreement with the GO enrichment terms for down-regulated genes than up-regulated genes. However, even with all of these discrepancies, it is worth highlighting that our GO enrichment terms trend in a similar direction as the author's, particularly with significantly up-regulated genes.

Looking at Figures 9A-C we are able to distinctly see an upregulation of the identified genes in sarcomere and mitochondrial regions between stage P0 and Adult. These plots are similarly modeled to ones observed in Figure 1D of the original article, which plotted genes representative of sarcomere, mitochondria, and cell cycle in both in vivo and in vitro experiments. The major noticeable trend is that we see an increase in all genes representing sarcomere and mitochondria from P0 to Adult. However, we note the opposite trend in the cell cycle plot (Figure 9C), where looking at only P0 to Adult, there is a decreasing trend. While we see an increase in cell cycle genes at P4, we may attribute to the analysis method that one of our replicates was not clustered with the 7 other samples. The down regulation of cell cycle genes signifies a potential 'loss of ability' to re-enter the cell cycle and initiate cell regeneration. The authors of the article were primarily interested in identifying these factors by examining differentially expressed genes and mapping these genes to GO terms through DAVID analysis for confirmation in particular pathways. An increase in sarcomere genes indicates that as cardiac myocytes grow over time, increased organization and sarcomere assembly processes have become initiated. To potentially meet this energy demand within cardiac myocytes as well as the fact that hearts are high energy organs, mitochondrial genes will likely show an increase in gene expression.

The clustered heatmap provides the ability to visualize the differences in gene expression between different stages of in vivo maturation. We note that compared to the heatmap created in the article (Figure 2A), our heatmap (Figure 9D) does show that the two replicates P0_1 and P0_2 show differences in gene expression. This is likely due to the fact that the samples were not clustered together leading to potentially different results. In addition, our main gene expression differences we see are between P0_2, and the P4 replicates against the adult replicates. We see that in both adult replicates (Ad_1/Ad_2) the Plekha7, Stbd1, and Myl4 bands are heavily expressed and are involved in sarcomeric assembly and organization (adherens junctions, myosin chains).

Additional details would have aided reproducing the author's analysis and interpreting our results. First, we would like to know what quality control measures they took before analyzing their samples. More explicit descriptions of how they decided what GO terms to report would have enabled a more direct comparison between our results and theirs. Were the GO terms collapsed into keywords, or was some other curation method performed?

O'Meara et al looked conducted many different experiments in this study, including in vitro, in vivo, and explant experiments. By looking at their question from many different directions, they helped reduce methodological biases. In vitro experiments enabled more direct manipulation to experimentally test what could trigger regeneration. The authors use of in vivo and in vitro experiments is to be able to holistically approach the question and confirm possibly identifying factors in cardiac myocyte regeneration. The use of in vitro and in vivo experiments allows us to study the effects in both environments and establish significant results. The use of an explant model allowed for practical visualization of sarcomere assembly and to help identify reasons for the differentiated state of a cardiac myocyte. This use of an ex vivo model allows for comparison between in vivo models while allowing for potential stimuli in a controlled approach.

Conclusion

We sought to compare the transcriptional profile of neonatal and adult mice in order to identify significantly differentially expressed genes related to the cardiac myocyte differentiation process. In comparison to O'Meara et al., we were only able to partially reproduce their results albeit with a high degree of similarity. Possible explanations for this discrepancy could be that O'Meara et al used a larger gene data set, they did not include more specific GO enrichment terms, and they did not include all the details necessary to replicate their results. Regardless, our overall biological findings corroborate with those of O'Meara et al. While the overall difficulty to replicate results from the author can be mainly attributed to personal unfamiliarity with R software and difficulties coding, the data that was provided as well as the importance of the analysis was clear, even when some minor differences exist. A future project that might aid visualizations would be to try conduct a network analysis with these identified genes and help narrow down potential areas to induce regeneration or delay organizational processes.

References

Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Data

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., & Soboleva, A. (2012). NCBI GEO: Archive for functional genomics data sets—update.

Batut, B. Galaxy training: Quality control. Retrieved March 13, 2021, from https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html

Edgar, R. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, *30*(1), 207–210. https://doi.org/10.1093/nar/30.1.207

Fryar CD, Chen T-C, Li X. (2012). Prevalence of uncontrolled risk factors for cardiovascular disease: United States, 1999–2010 pdf icon [PDF-494K]. NCHS data brief, no. 103. Hyattsville, MD: National Center for Health Statistics; Accessed February 27, 2021.

Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2018). dplyr: A Grammar of Data Manipulation. R package version 0.7.6. https://CRAN.R-project.org/package=dplyr

Harvard Chan Bioinformatics Core Training (HBC). (2018). Quality control: Assessing FASTQC results. Retrieved March 13, 2021, from https://hbctraining.github.io/Intro-to-rnaseq-hpc-salmon/lessons/qc_fastqc_assessment.html

Huang, D., Sherman, B. T., & Lempicki, R. A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, *37*(1), 1–13. https://doi.org/10.1093/nar/gkn923

Huang, D., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols*, *4*(1), 44–57. https://doi.org/10.1038/nprot.2008.211

Langmead B, Salzberg S. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods. 9:357-359.

Li H.*, Handsaker B.*, Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9. [PMID: 19505943]

Liao, R., & Jain, M. (2007). Isolation, culture, and functional analysis of adult mouse cardiomyocytes. *Methods in molecular medicine*, *139*, 251–262. https://doi.org/10.1007/978-1-59745-571-8_16

Michigan State University College of Natural Science Research Technology Support Facility. FastQC tutorial & FAQ. Retrieved March 13, 2021, from https://rtsf.natsci.msu.edu/genomics/tech-notes/fastqc-tutorial-and-faq/

Milo, R., Jorgensen, P., Moran, U., Weber, G., & Springer, M. (2010). BioNumbers--the database of key numbers in molecular and cell biology. *Nucleic acids research*, *38*(Database issue), D750–D753. https://doi.org/10.1093/nar/gkp889

Nucleic Acids Research, *41*(D1), D991–D995. https://doi.org/10.1093/nar/gks1193

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

Porrello ER, Mahmoud AI, Simpson E, Hill JA, Richardson JA, Olson EN, Sadek HA. (2011) Transient regenerative potential of the neonatal mouse heart. Science.331:1078–1080.

Roberts, A., Trapnell, C., Donaghey, J. et al. (2011). Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biol 12, R22. https://doi.org/10.1186/gb-2011-12-3-r22

Ruvinsky, A., Marshall, G. J. (2005). Mammalian genomics. In Mammalian genomics (p. 92). Wallingford, Oxfordshire, UK: CABI Pub.
(Bionumber ID 102409, https://bionumbers.hms.harvard.edu/bionumber.aspx?id=102409&ver=9&trm=mouse+mRNA+GC+content&org=)

SRA Toolkit Development Team. SRA-Tools. Retrieved March 13, 2021, from http://ncbi.github.io/sra-tools/

Trapnell, C., Hendrickson, D., Sauvageau, M. et al. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol 31, 46–53. https://doi.org/10.1038/nbt.2450

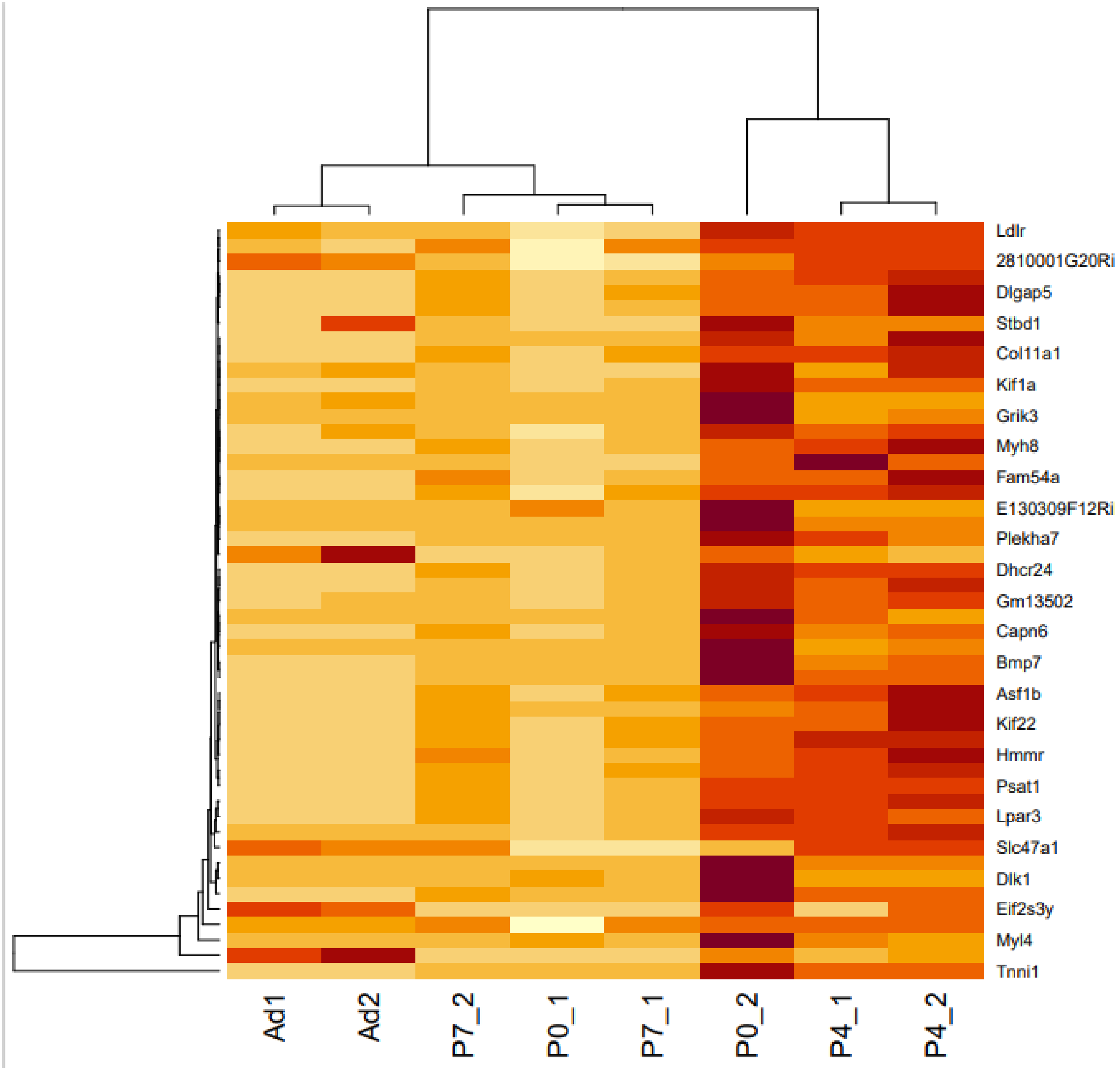
Trapnell, C., Williams, B., Pertea, G. et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28, 511–515. https://doi.org/10.1038/nbt.1621

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., & Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols*, *7*(3), 562–578. https://doi.org/10.1038/nprot.2012.016

Wamstad, J. A., Alexander, J. M., Truty, R. M., Shrikumar, A., Li, F., Eilertson, K. E., Ding, H., Wylie, J. N., Pico, A. R., Capra, J. A., Erwin, G., Kattman, S. J., Keller, G. M., Srivastava, D., Levine, S. S., Pollard, K. S., Holloway, A. K., Boyer, L. A., & Bruneau, B. G. (2012). Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage. *Cell*, 151(1), 206–220. https://doi.org/10.1016/j.cell.2012.07.035 (Data accessible at NCBI GEO database accession GSE64403 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE64403)

Wang, L., Wang, S., Li, W. (2012). RSeQC: quality control of RNA-seq experiments, *Bioinformatics*, 28(16), 2184–2185, https://doi.org/10.1093/bioinformatics/bts356

Wickham H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.



Expanded view of heatmap (Figure 9d).