INDIVIDUAL PROJECT

Transcriptional Profile of Mammalian Cardiac Regeneration with mRNA-Seq

Vishwa Talati (Programmer and Biologist)
Group Name: Glass Bottom Boat
Github Repository Link

INTRODUCTION

An adult mammalian heart has a limited capacity for self-renewal following injury. Shortly after birth, mammalian cardiac myocytes exit the cell cycle, and subsequent heart growth is achieved primarily by hypertrophy of existing cardiac myocytes. Substantial evidence suggests that even these terminally differentiated adult cardiac myocytes retain some limited ability for cell division. However, the innate ability of the adult mammalian heart to repair itself following injury such as myocardial infarction is inadequate to replace the loss of functional myocardium. Where adult mammalian hearts fail to regenerate after injury, neonatal mice can fully regenerate their heart following the resection of left ventricular apex. Evidence from genetic mapping demonstrate that this was derived from pre-existing cardiac myocytes and not stem cells. Cardiac myocytes in the regenerating neonatal mouse heart demonstrate loss of distinct sarcomere structures and a significant proportion of these cells enter the cell cycle, as indicated by phosphorylated histone H3 (pH3) expression and up-regulation of aurora B kinase, suggestive of cell fate reversion. Thus, identifying transcriptional changes that underpin mammalian cardiac regeneration at the molecular level is essential to understand what prevents cell and tissue regeneration in adult hearts.

The objectives of this study were to determine if myocytes revert the transcriptional phenotype to a less differentiated state during regeneration and to systematically interrogate the transcriptional data to identify and validate potential regulators of this process. A global gene expression pattern is profiled over the course of mouse cardiac myocyte differentiation both in vitro and in vivo to compare this transcriptional signature of differentiation to a cardiac myocyte explant model whereby the cardiac myocytes loose fully differentiated phenotype to identify genes and gene network that changed dynamically during these processes. The RNA seq datasets are interrogated to predict and validate the upstream and downstream regulators along with its associated pathways that can modulate cell cycle state of cardiac myocytes. The aim of this project was to replicate the findings of O'Meara et al using similar tools and methods.

DATA

The sequencing data was uploaded to the public database Gene Expression Omnibus with accession number GSM1570702 (vP0_1) from GEO Series GSE64403. The sample fileSRR1727914.sra was downloaded from this accession number. This was further processed via fastq-dump into separated sequence files in fastq format. For further alignment and analysis, pre-processed data files from prior project were used.

METHODS

Aligning and Quality Analysis:

To further pursue the study objectives, it was necessary to realign the RNA seq data and evaluate it both qualitatively and quantitatively before further analysis. TopHat (v2.1.1), a fast splice junction mapper for RNA reads was used to align multiple reads against mouse reference genome (mm9) with Bowtie2 (v2.4.2) indexes. For this purpose, a batch job was submitted that took almost an hour to run.

The SAMTools (v1.10) flagstat tool was used to evaluate the passing or failing of alignment reads mainly to indicate whether any improper mapping to alternate chromosomes, reads or duplicates has occurred or not.

RSeQC is an RNA seq quality control package which provides several useful modules that can comprehensively evaluate high throughput sequencing data. Three modules of this package namely geneBody_coverage.py, inner_distance.py and bam_stat.py were used for quality check of our data. The geneBody_coverage.py module calculated RNA seq reads coverage over the gene body, inner_distance.py calculated inner distance between read pairs and bam_stat.py summarized the mapping statistics of the BAM file. For this purpose, SAMtools index function was first used to organize the data by position and then it was feed to RSeQC package. A batch job was submitted with necessary commands which took almost 2 hours to run.

The final step for evaluation of the input RNA seq datasets was done by using Cufflinks Data package. It is a tool that counts how reads map to genomic regions defined by annotation thus translating the RNA seq data into gene-based form. This was done by running a batch job with required commands on the BAM file created by TopHat which thereafter gave a file containing quantified alignments in FPKM (Fragments per Kilobase of transcript per Million mapped reads) which indicated number of fragments for particular a region for all genes. A histogram of FPKM frequency distribution filtering FPKM with zero for all genes was thus plotted (Figure 3). Cuffdiff, a tool in the suite of Cufflinks was used to identify differentially expressed genes.

Functional annotation and Cell clustering:

Using the data table generated by cuffdiff, genes under 0.01 threshold were identified as differentially expressed genes and DAVID tool was used to perform functional annotation clustering analysis. Top 10 gene set clusters were identified along with their enrichment scores. For this project, the top 10 clusters list was taken from previous group's analysis on the same project and compared with that of the paper.

Biological Interpretation:

Using the differentially expressed genes file generated by cufflinks, the FPKM values of representative genes from sarcomere, mitochondria and cell cycle that were significantly expressed in mice on postnatal day 0 (P0), day 4 (P4), day 7 (P7) and adult (Ad) were plotted against the biological age of sample. (Figure 4). Furthermore, a clustered heatmap was generated based on the

gene expression of log fold change over the course of in vivo maturation in postnatal to adult maturation.

RESULTS

Number of total reads	49706999	100%
Number of mapped reads	49706999	100%
Number of unique mapped reads	41389334	83.27%
Number of multi-mapped reads	8317665	16.73%
Number of unaligned reads	0	0.00%

Table 1: Report of the total number of reads, number of mapped, unique, multi-mapped, and unaligned reads with percentages of total reads for each

After performing TopHat, a BAM file was created which is a binary version of SAM (Sequence Alignment/Map). This file contained all the original reads with the alignments discovered by TopHat. On evaluation of the BAM file using SAMtools flagstat tool, zero reads failed the quality standards for which 49706999 reads were mapped in some fashion. 8317665 reads were considered secondary i.e., mapped multiple times. 41389334 reads were mapped uniquely without repeats where 20878784 were for read1 and 20510550 were for read2. 1452862 reads (3.51%) were considered singletons i.e., reads that mapped but whose mates did not.

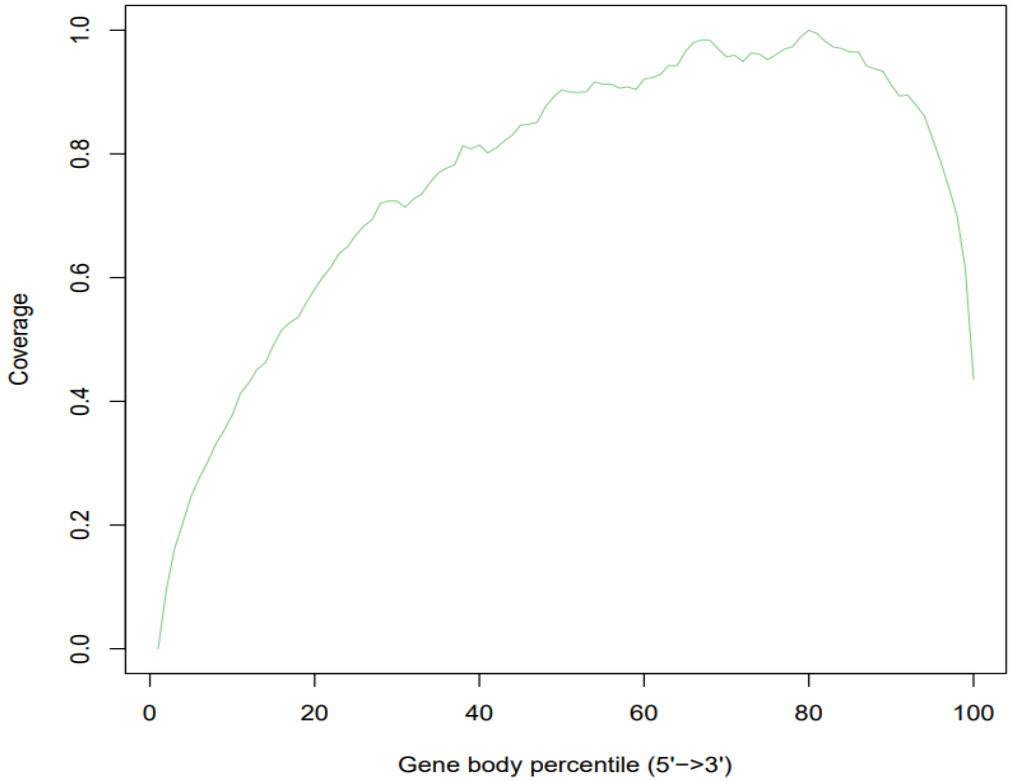

Figure 1: Coverage graph of RNA reads with 3' bias

From RSeQC, the geneBody_coverage graph (Figure 1) was mapped between nucleotide position and number of reads to locate the bias in read values. From Figure 1, a clear 3' end bias was observed in the data which meant that a greater percentage of reads were located around that region. The Insert Size or inner_distance graph (Figure 2) evaluated the distance between two paired reads. The distribution of reads was centred around 60 base pairs with a tail that pulls the distance further to the right, away from zero implying level of regularity and lack of additional factors complcating the alignments. Lack of error was also implied by the low number of negative values which indicated that there was not much overlap between two reads.

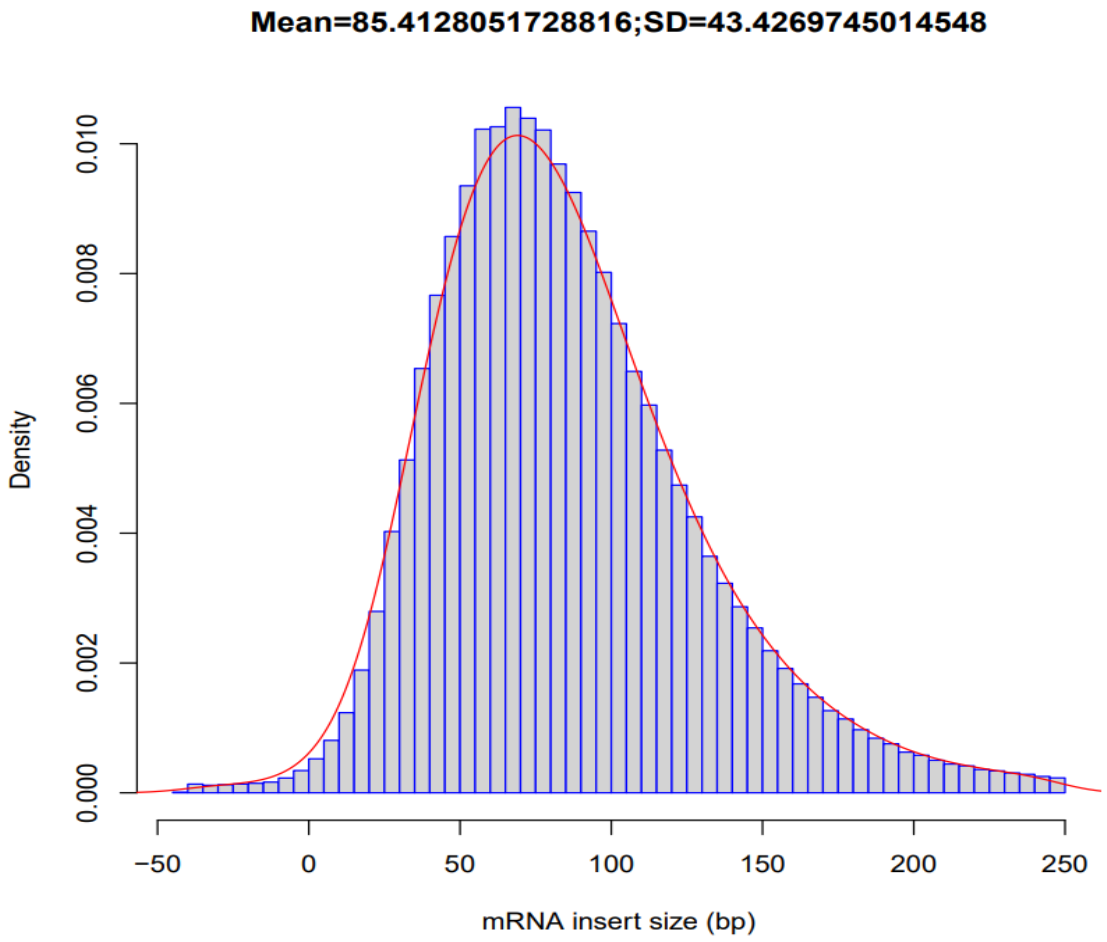

Figure 2: Plot showing inset size between two paired reads

On removing genes with zero FPKM values, there were total 20101 genes. The distribution of expression levels provides us insight into details of expression profiles (Figure 3).

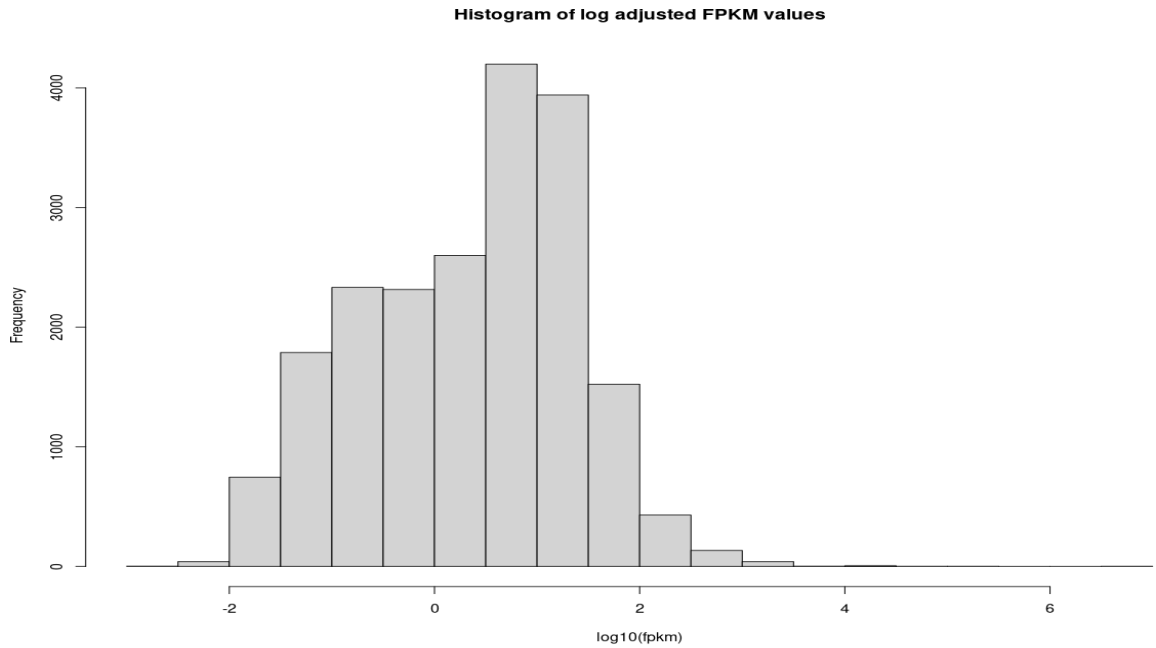

Figure 3: Histogram of FPKM values for all genes

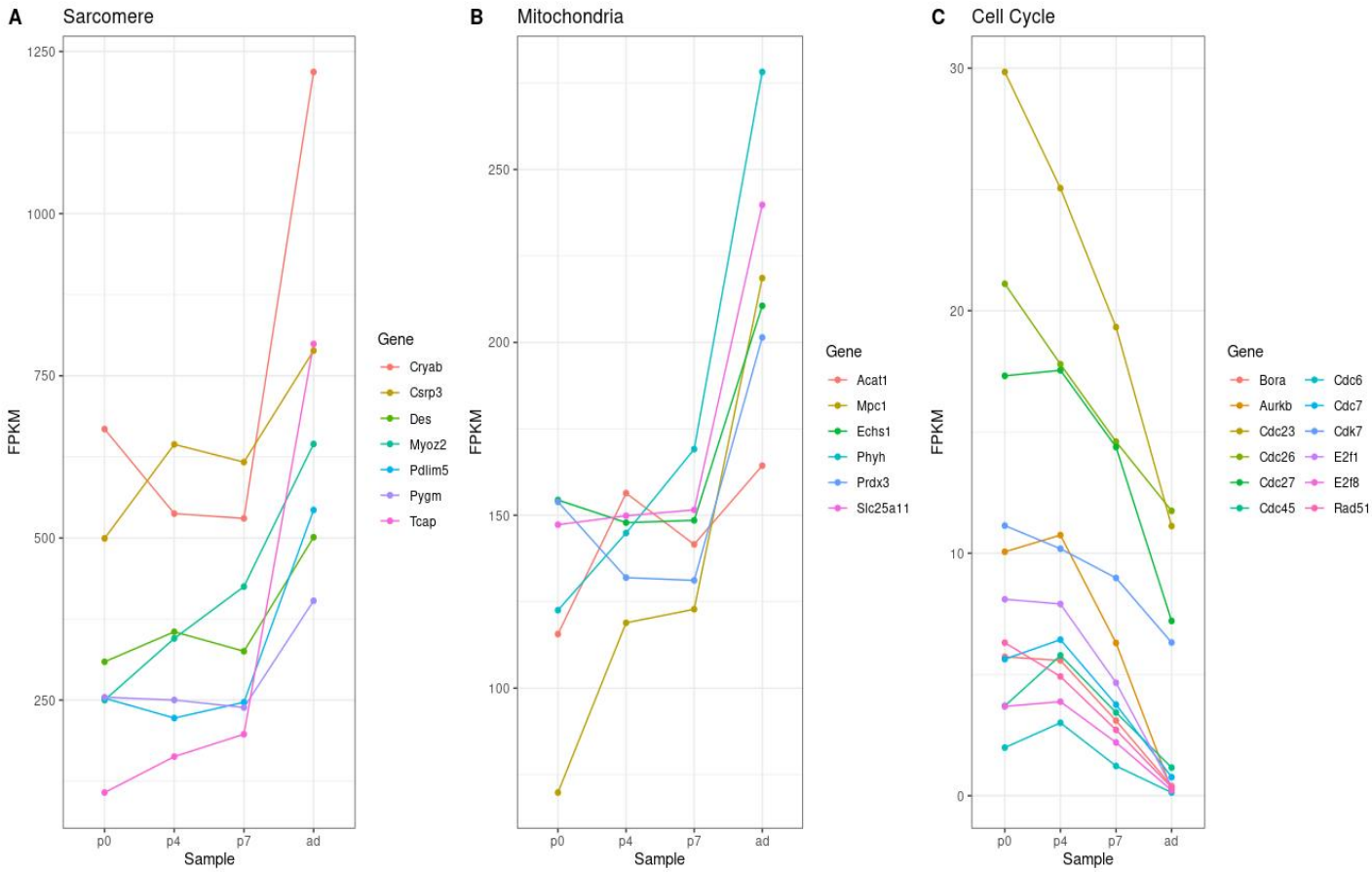

Figure 4: FPKM values of representative Sarcomere, Mitochondria and cell cycle genes which were significant and differentially expressed.

Above is a plot of FPKM values of Sarcomere, Mitochondria and cell cycle was plotted against the biological age of sample for differentially expressed genes from Cufflinks. The genes were significant and differentially expressed in mice on postnatal day 0(P0), 4(P4), 7(P7) and adult (AD). The sarcomere differentially expressed genes do not match the genes from research article. However, the plot follows a similar trend to that of paper in terms of increase in FPKM values from P7 to adult suggesting that sarcomere genes were up regulated from postnatal to adult maturation.

For Mitochondrial differentially expressed genes, the plot seems quite like that produced in the reference article where except Prdx3 all the genes showed similar trend.

The 12 differentially expressed cell cycle genes also show a trend like the plot C in figure 1D in reference paper. This trend is across in-vivo maturation. The genes are downregulated for cell cycle like that in the findings of reference paper.

Cluster	GO Term	Enrichment score
1	Mitochondrial membrane, mitochondrial envelope	21.496
2	Generation of precursor metabolites and energy/ oxidative phosphorylation	15.293
3	Lipid/ fatty acid metabolic process	13.27
4	Extracellular organelle	10.579
5	Sarcomere	6.967
6	Fatty acid/ lipid oxidation	6.104
7	Metal cluster binding	5.985
8	Glycolysis/ ATP/ carbohydrate metabolic process	5.559
9	NAD binding	5.492
10	Cellular response	5.257

Table 2: Selected top 10 up regulated gene clusters with their enrichment scores from DAVID where the highlighted ones are those that overlap with O'Meara et al. paper.

Cluster	GO Term	Enrichment score
1	Cell cycle, chromosome segregation, nuclear segregation	10.965
2	Extracellular matrix	10.799
3	Cell Proliferation	9.569
4	Regulation of cellular component/organelle organization	8.592
5	Circulatory system development	8.383
6	Cell differentiation/neuron system development/neuron differentiation	8.086
7	RNA metabolic process	7.887
8	Embryonic development	7.447
9	Chromosome	7.339
10	Regulation of Gene Expression/ DNA binding/ Nucleic Acid Binding	7.287

Table 3: Selected top 10 down regulated gene clusters with their enrichment scores from DAVID where the highlighted ones are those that overlap with O'Meara et al. paper.

UP REGULATED		DOWN REGULATED	
GO Term	**Enrichment Score**	**GO Term**	**Enrichment Score**
Mitochondria	14.35	Non-membrane bound organelle	88.91
Sarcomere	8.50	Nuclear lumen	88.91
Sarcoplasm	6.03	RNA processing	59.78
Respiration/ Metabolism	4.98	Cell cycle	59.78
Glycolysis	4.39	DNA repair	59.78

Table 4: DAVID results for up and down regulated genes from O'Meara et al. paper

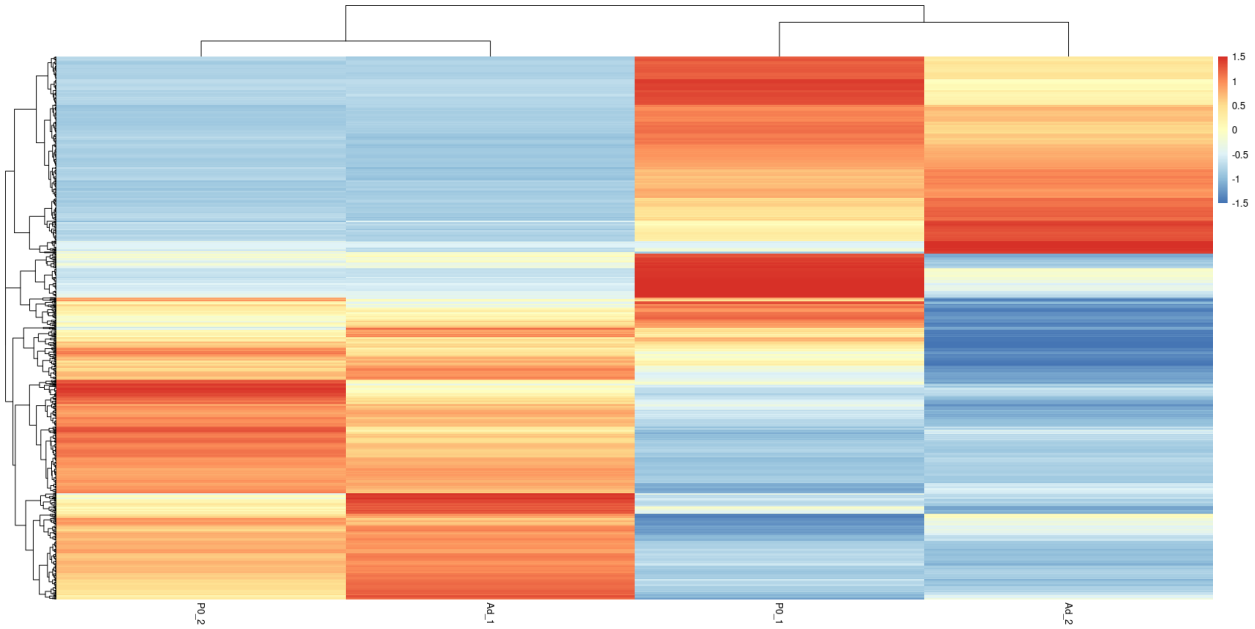

Figure 4: Clustered heatmap of top 1000 differentially expressed genes based on gene expression of log fold change over in vivo maturation in postnatal vs adult phase where the blue part represents less expressed genes and orange represents highly expressed genes and gene symbols were removed to enhance representation.

DISCUSSION

It was evident from RSeQC (Figure 1), that a greater percentage of reads were located around 3' end. One possible reason for this could be Polyadenylated RNA (RNA with multiple adenine bases at 3'end) that was isolated from the samples favoring reads around 3'end of the sequence. Another reason could be degradation of samples. From Inert Size Graph (Figure 2), the relatively low number of negative values indicated lack of error and overlaps between the two reads.

The results from up regulated DAVID analysis showed 3 out of 5 GO terms that overlapped with the reference paper namely Mitochondria (21.46), Sarcomere (6.967) and Glycolysis (5.559). For down regulated DAVID analysis, 3 out 5 GO terms overlapped which were chromosome (7.339), RNA process (7.887) and cell cycle (10.965). However, the enrichment scores did not overlap may

be due to differences in the version of DAVID that the author used or the difference in the gene terms used for DAVID analysis. It is worth to mention that there is more than 5 years of gap between the paper analysis and our analysis which means DAVID must have been altered with many genes added and pathways edited. Getting exact results might want us to question the validity of DAVID. Another possible reason for difference in results could be the difference in number of up and down regulated genes used for DAVID analysis.

On comparing the heatmap of expression level of P0 vs Ad in this project with that generated in the reference paper, too much information was not obtained. This is because the heatmap generated in reference paper included datasets of all experiments and thus possess larger scale of normalization. Due to different scales, color in cells between two heatmaps was not much informative. However, the cluster of genes well represented the changes of expression levels between stages which indirectly supported the rationale of the reference paper that groups of genes with common function were identified in ex vivo cultured cardiac myocytes via hierarchical clustering.

CONCLUSION

This study provides a critical framework for understanding the transcriptional expression changes required for cardiac myocyte repair in response to injury that will be invaluable for ultimately guiding efforts to promote adult mammalian cardiac regeneration. The analysis done in this project was done correctly and it replicated the findings of O'Meara et al. to great extent. However, the differences risen were likely due to the choice of tool, differences in versions of tools and parameters for analysis.

REFERENCES

1. O'Meara CC, Wamstad JA, Gladstone RA, Fomovsky GM, Butty VL, Shrikumar A, Gannon JB, Boyer LA, Lee RT. Transcriptional reversion of cardiac myocyte fate during mammalian cardiac regeneration. Circ Res. 2015 Feb 27;116(5):804-15. doi: 10.1161/CIRCRESAHA.116.304269. Epub 2014 Dec 4. PMID: 25477501; PMCID: PMC4344930.
2. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013 Apr 25;14(4):R36. doi: 10.1186/gb-2013-14-4-r36. PMID: 23618408; PMCID: PMC4053844.
3. TopHat. (2016). Retrieved May 8, 2021, from Jhu.edu website: http://ccb.jhu.edu/software/tophat/index.shtml
4. Langmead B, Wilks C, Antonescu V, Charles R. Scaling read aligners to hundreds of threads on general-purpose processors. Bioinformatics. 2018 Jul 18. doi: 10.1093/bioinformatics/bty648.
5. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012 Mar 4;9(4):357-9. doi: 10.1038/nmeth.1923
6. Bowtie 2: fast and sensitive read alignment. (2012). Retrieved May 8, 2021, from Sourceforge.net website: http://bowtie-bio.sourceforge.net/bowtie2/index.shtml

7. Trapnell, C., Williams, B., Pertea, G. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28, 511–515 (2010). https://doi.org/10.1038/nbt.1621
8. Cufflinks. (2014, December 10). Retrieved May 8, 2021, from Cufflinks website: http://cole-trapnell-lab.github.io/cufflinks/
9. RSeQC: An RNA-seq Quality Control Package — RSeQC documentation. (2020). Retrieved May 8, 2021, from Sourceforge.net website: http://rseqc.sourceforge.net/
10. Wang L, Wang S, Li W. RSeQC: quality control of RNA-seq experiments. Bioinformatics. 2012 Aug 15;28(16):2184-5. doi: 10.1093/bioinformatics/bts356. Epub 2012 Jun 27. PMID: 22743226.
11. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
12. 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]
13. Li H A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011 Nov 1;27(21):2987-93. Epub 2011 Sep 8. [PMID: 21903627]