Transcriptional Profile of Mammalian Cardiac Regeneration with mRNA-Seq

Data Curator- pre computed
Programmer- **Neha Gupta**
Analyst – pre computed
Biologist- **Neha Gupta**
TA- Jackie Turcinovic
Group – Swiss cheese
Github Repo Link

Introduction

The adult hearts of all the mammalian have a limited capacity for self-rehabilitation and regenration. Following birth, mammalian heart growth is carried out primarily via hypertrophy of existing cardiac myocytes [1]. Evidence has shown that although adult cardiac myocytes are terminally differentiated, they do retain a limited ability for cell division. However, this cell division capacity is insufficient to replace the functional tissue lost due to injury. Neonatal mice are able to fully regenerate cardiac tissue following the resection of the left ventricular apex. Further investigation through genetic fate mapping showed that the cardiac myocytes involved in heart regeneration were derived from pre-existing cardiac myocytes, not stem cells.

These cells exhibited loss of sarcomere structures and a large majority of cells involved in regeneration re-entered the cell cycle [1]. Therefore, identifying the mechanisms by which myocytes naturally undergo cell cycle activity during regeneration is fundamental to understanding what prevents cell and tissue regeneration in adult hearts.

The objective of this study was to determine if myocytes revert the transcription phenotype to a less differentiated state during regeneration [1] and to systematically examine 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 (mouse embryonic stem cells differentiated to cardiac myocytes) and in vivo (cardiomyocyte maturation from neonate to adult) and compared this transcriptional signature of differentiation to a cardiac myocyte explant model whereby cardiac myocytes lose the fully differentiated phenotype to identify genes and gene networks that changed dynamically during these processes [1]. The RNA sequencing (RNAseq) datasets are interrogated as well to predict and validate upstream regulators and associated pathways that can modulate the cell cycle state of cardiac myocytes.

Data

The sequencing data was upload to the public database Gene Expression Omnibus (GEO) with accession number GSM1570702. The SRA file was download from this accession number via fastq-dump into paired end fastq file format. For the further alignment and analysis of the project, the pre computed fastq files from the prior project were use.

Methods

For further analysis, RNA seq data was realigned and evaluated in both qualitatively and quantitavely for error sources. The reads in the fastq files were mapped with the reference genome using TopHat(v2.1.1)[3] with Bowtie(v2.4.2)[5] indexes. To handle these huge sequenecing data, powerful cluster computers were required. Batch jobs created with a specific of 16 threads to operate on a single node, in order to allocate enough memory for this time consuming and intensive computational tasks.

The RSeQC(v3.0)[8, 9] package was further utilized to evaluate the RNA-seq data. To get the read sequences sorted and organized by position, SAMtools[4] sort and index functions used. The sorted data was direct to RSeQC package. Three modules of this package namely geneBody_coverage.py; inner_distance.py and bam_stat.py 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. The SAMTools (v1.10) [4] flagstat tool was used to evaluate the passing or failing of alignment reads to several categories, mainly in order to indicate whether improper mapping to alternate chromosomes, reads, or duplicates occurred. Zero reads failed the quality control standards.

The final step to quality check the input RNAseq datasets were based on the Cufflinks data package. Cufflinks can be used to map sequence data to genomic annotations thus transferring the RNAseq data into a 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. FPKM is another rough normalization metric for evaluating the RNAseq data, indicating the expected number of fragments for a particular genomic region. Part of the calculation for this value involves scaling up the values for readability. A histogram of FPKM frequency (Figure 3) was plotted below in order to roughly evaluate the distribution of FPKM values.

Using the data table generated by cufflink[7], 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 list was taken from previous group's analysis and compared with that of the paper.

Cuffdiff package was used to evaluate differentially expressed genes. The output file was imported into the RStudio. 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 5). 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

Discussion

From RSeqc, it was observed that a greater percentage of reads were located around 3' end. This could be because of 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 Insert Size graph (Figure 1), the relatively low number of negative values indicated lack of error and overlaps between the two reads.This was less computationally intensive.

The results from up regulated DAVID analysis showed 2 out of 5 clusters showed overlapping with reference paper namely Mitochondria, Sarcomere and Glycolysis. For down regulated clusters, 1 out of 5 showed overlapping with reference paper like Cell cycle and RNA process. However, the enrichment score did not overlap which may be because of difference in the version of DAVID that was used. It is worth noting that with so many years gap between the article and this project, it is very certain DAVID must have altered with many genes and pathways edited. The other reason for the difference could be the difference in the number of up and down regulated genes for DAVID analysis.

Concerning the clustered heatmaps when compared with Figure 13 of reference paper, the expression of PO Vs Ad did not give much information. This is because the heatmap generated in the research article included datasets of all experiments and thus possess larger scale of normalization. Dues to different scales, color in cells between two heatmaps, not much information was extracted. Nonetheless, the cluster of genes well represented the changes of expression levels between stages that indirectly supported the rationale of the reference paper that groups of genes with common function were identified I ex-vivo cultured cardiac myocytes via hierarchical clustering.

Results

Feature	**read length (basepairs)**	**Reads %**
Number of total reads	49706999	100%
Number of mapped reads	49706999	100%
Number of multi mapped reads	8317665	16.73%
Number of unique mapped reads	41389334	83.2%
Number of unaligned reads	0	0.00%

Table 1: Table for total number of reads, mapped reads, unique mapped reads and unaligned reads with percentage of total reads for each.

The SAMTools (v1.10) [4] flagstat tool was used to evaluate the passing or failing of alignment reads to several categories, mainly in order to indicate whether improper mapping to alternate chromosomes, reads, or duplicates occurred. Zero reads failed the quality control standards; 49706999 reads were mapped in some fashion. A total of 8317665 (16.73% of total) reads were considered secondary, or mapped multiple times, meaning 41389334 (83.27% of total) were mapped uniquely without repeats. Of the ones left, 1452862 (2.922%) were considered singletons, reads that mapped whose mates did not. As an additional note for reproduction of results, percentages displayed by SAMtools were based on the unique mappings.

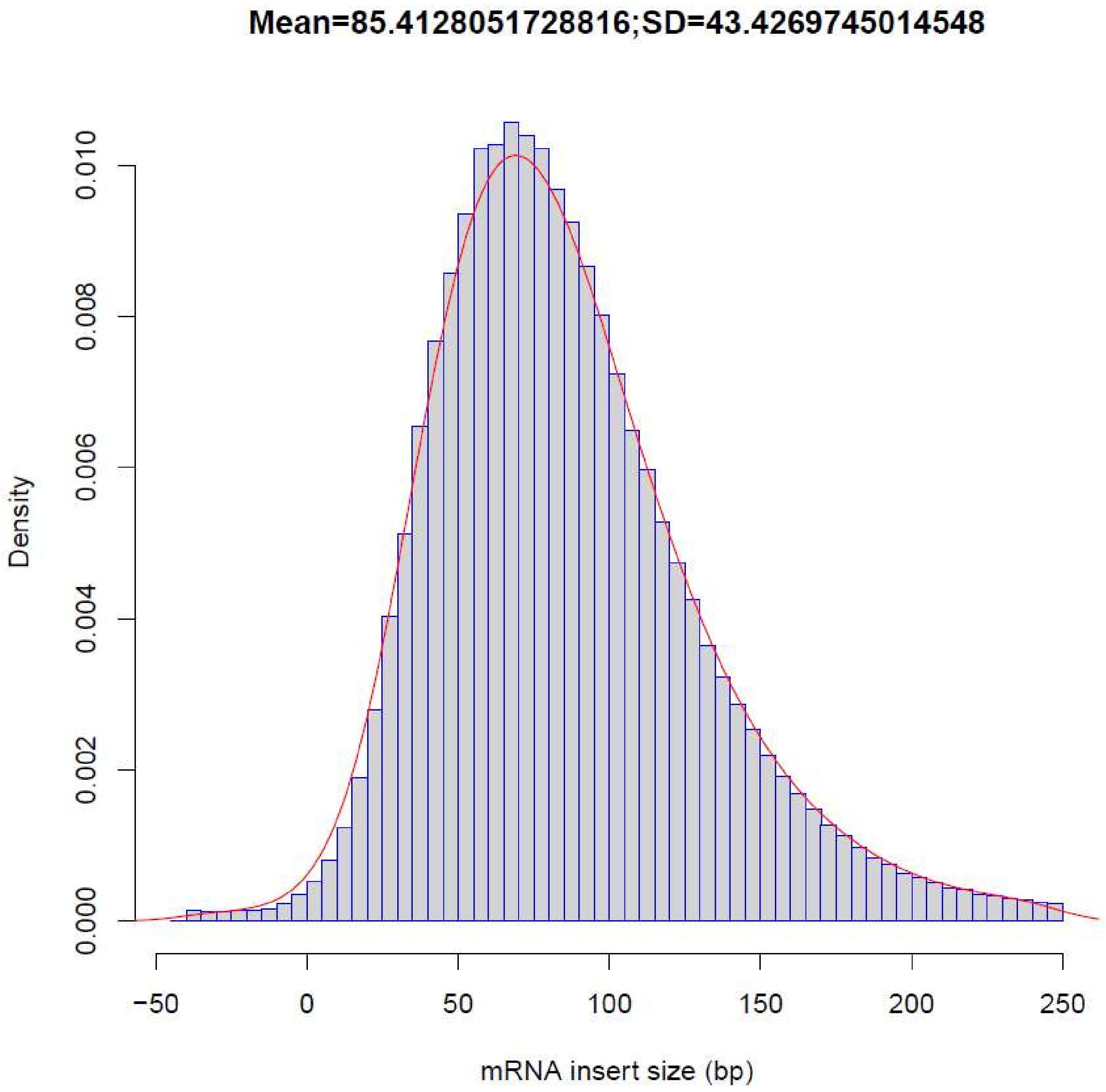

Figure 1 The plot evaluate distance between two paired reads.

The distance distribution was more around 60 base pairs with tail that pulls the distance to right away from 0. This is because a level of regularity for distance and a lack of additional factors complicating the alignments**.**

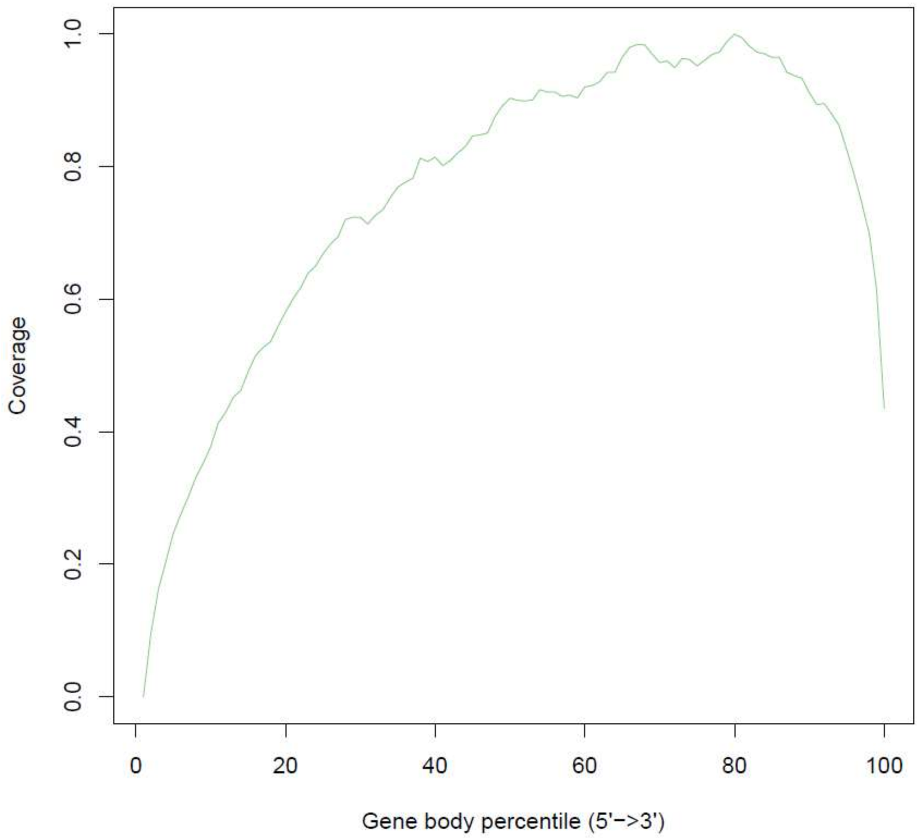

Figure 2: The graph mapped between nucleotide position and number of reads to locate bias in read values.

For the gene body coverage graph a clear 3’ end bias was noted in the data because a greater percentage of reads located around that region. This was expected as polyadenylated RNA (RNA with multiple adenine bases at the 3’ end) which was isolated for the samples favoring reads around the 3’ end of the sequences.

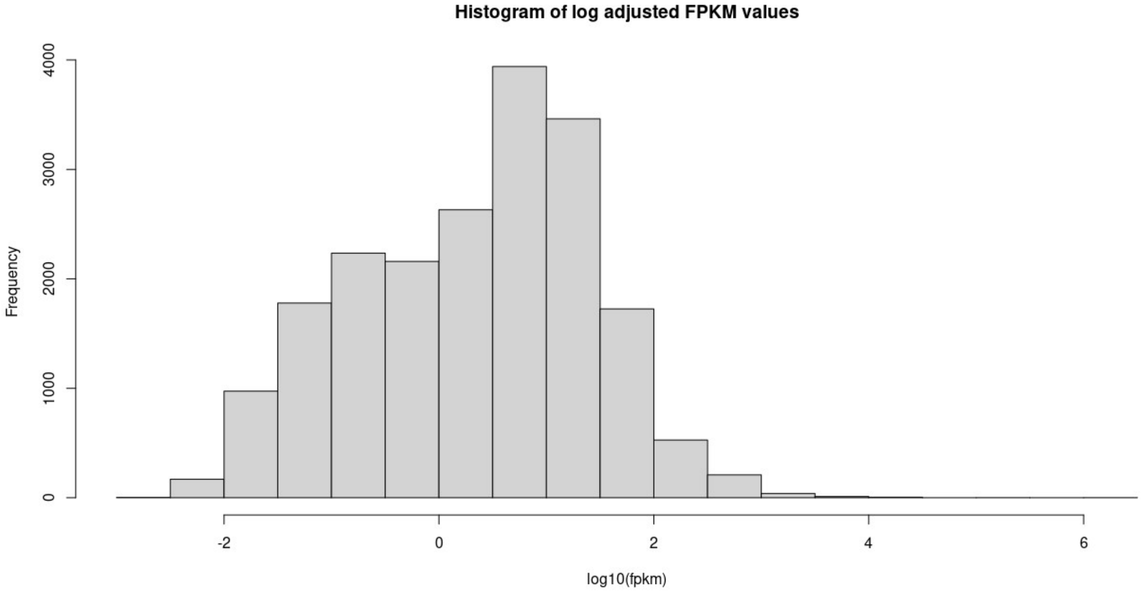

Figure 3 Histogram of FPKM values for all the genes

Based on the quantification of gene expression, in total 19870 genes remained after removing ones having FPKM of zero values. The distribution of expression levels provides insight into details of expression profiles.

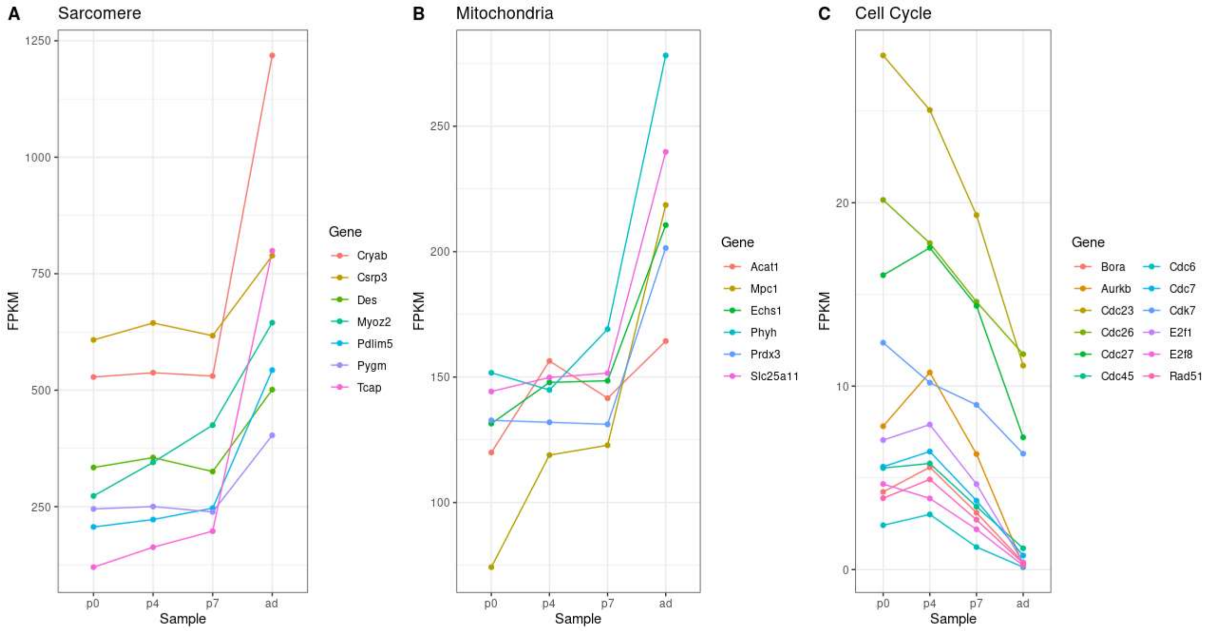

Figure 4: FPKM values of representative Sarcomere, Mitochondria and cell cycle genes which were significant and differentially expressed.

Using differentially expressed genes received from Cufflinks, the FPKM values for Sarcomere, Mitochondiral and cell cycle were plotted respectively against the biological age of sample. The genes were significant and differentially expressed in mice on postnatal day 0(P0), 4(P4), 7(P7), and adult (Ad).

The Sarcomere DE genes, do not match the reference figure from the research article. Nevertheless, the plot exhibits an increase in FPKM values from P7 to adult suggesting that those sarcomere genes were up-regulated from postnatal to adult maturation, which is consistent with the reference study.

The mitochondrial DE genes, looks very similar to the plots produced in reference article. The Apcat1, Mpc1, Phyh, Slc25a11 and Echs1 show very similar trend as compared to reference paper but Prdx3 shows some dissimilarity in Adult points.

The 12 DE genes also show similar trend as represented in figure 1D plot C of reference paper. The plot show trend across in-vivo maturation. The cell cycle DE genes demonstrate down-regulation of cell cycle gene during in-vivo maturation, which is consistent with the findings in the reference article.

Cluster	GO Term	Enrichment Score
1	Mitochondrion	53.28
2	Ribonucleotide metabolic process	23.6
3	Mitochondrial protein complex	22.5
4	Lipid metabolic process	21.79
5	Sarcomere	10.93

Table 2: The result from DAVID for significantly up regulated gene clusters using Gene ontology terms. Highlighted in blue are the genes that were present in both the experimental study and cited in the reference paper.

Cluster	GO Term	Enrichment Score
1	Cell cycle	32.79
2	Nuclear chromosome	21.56
3	Regulation of RNA metabolic process	20.94
4	Chromatin organization	17.83
5	Regulation of cell cycle	16.77

Table 3: Results from DAVID for significantly down regulated gene clusters using Gene ontology terms. Highlighted in blue are the genes that were present in both the experimental study and cited in the reference 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: The DAVID results from the reference paper.
(*) signifies that the genes were found in both the experimental study and the reference paper

The up and down regulated differentially expressed genes were used to get the top gene enriched cluster from DAVID 6.8. The table 4 summarized the top such clusters. The top GO terms for up regulated genes were mitochondrion, ribonucleotide metabolic process, mitochondrial protein complex, lipid metabolic process and sarcomere. Similarly for down regulated genes were cell cycle, nuclear chromosome, regulation of RNA metabolic process, chromatin organization and regulation of cell cycle. These up and down regulated tables were referred from the prior study done on the same project.

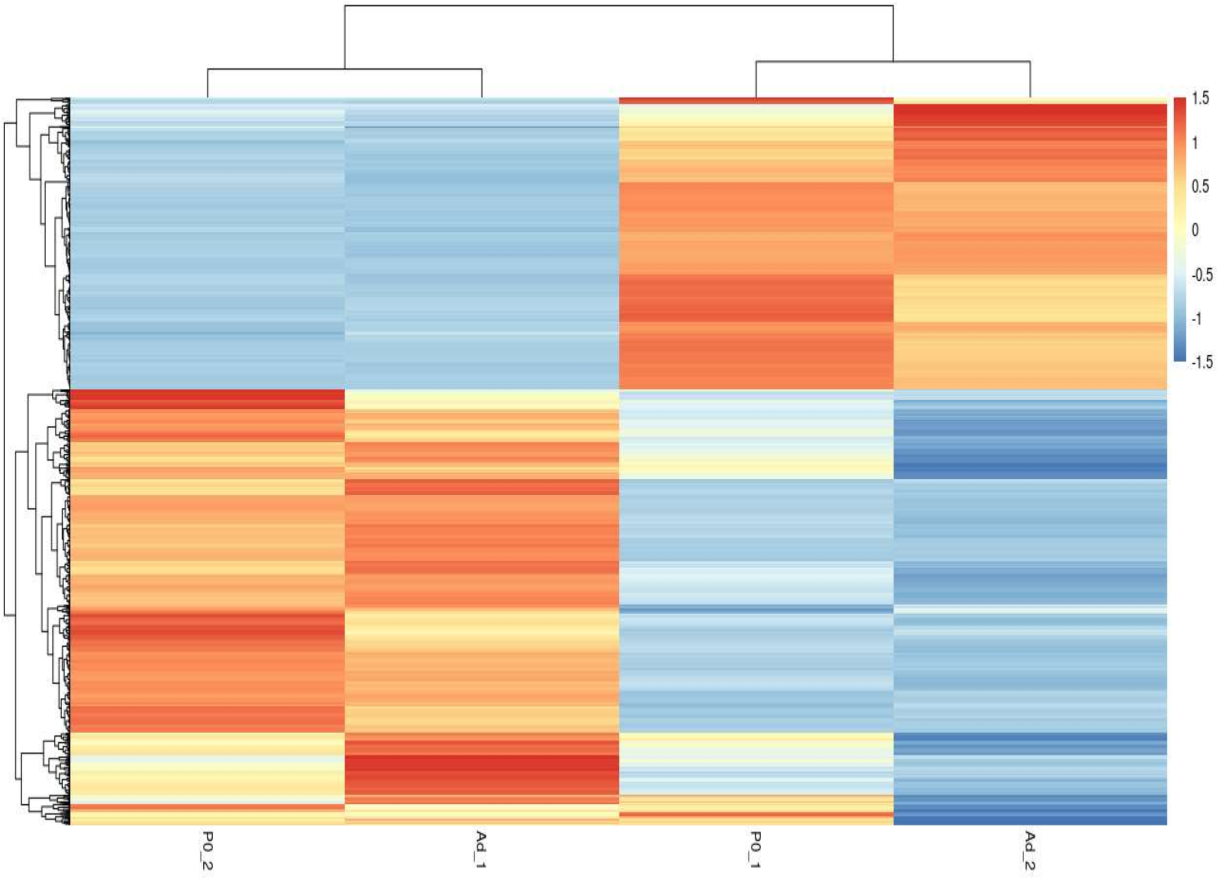

Figure 5 Clustered heatmap of top 1000 differentially expressed genes (P0 Vs Adult).

This heatmap was generated based on the gene expression of log fold change over in vivo maturation in postnatal to adult phase. The blue part represents genes, which were less expressed whereas orange represents higher expressed genes. For better visualization, gene symbols were removed.

Conclusion

The analysis in this study was done correctly and it replicated the findings of O'Mera et al. to great extent. The differences risen were likely due to the choice of tool, difference in versions of computational packages and parameters for analysis.

References

1- O'Meara, Caitlin C et al. “Transcriptional reversion of cardiac myocyte fate during mammalian cardiac regeneration.” Circulation research vol. 116, 5 (2015): 804-15. doi:10.1161/CIRCRESAHA.116.304269

2- Senyo, Samuel E et al. “Mammalian heart renewal by pre-existing cardio myocytes.” Nature vol. 493, 7432 (2013): 433-6. Doi:10.1038/nature11682

3- 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.

4- Li, Heng et al. “The Sequence Alignment/Map format and SAMtools.” Bioinformatics (Oxford, England) vol. 25, 16 (2009): 2078-9. doi:10.1093/bioinformatics/btp352

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- Cufflinks. (2014, December 10). Retrieved May 8, 2021, from Cufflinks website: http://cole-trapnell-lab.github.io/cufflinks/

8- RSeQC: An RNA-seq Quality Control Package — RSeQC documentation. (2020). Retrieved May 8, 2021, from Sourceforge.net website: http://rseqc.sourceforge.net/

9- 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.

10- 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].