# BF528 Project 2: Transcriptional Profile of Mammalian Cardiac Regeneration with mRNA-Seq

**Authors:** Pragya Rawat (Data Curator), Pooja Savla (Programmer), Vrinda Jethalia (Analyst), Manasa Rapuru (Biologist)

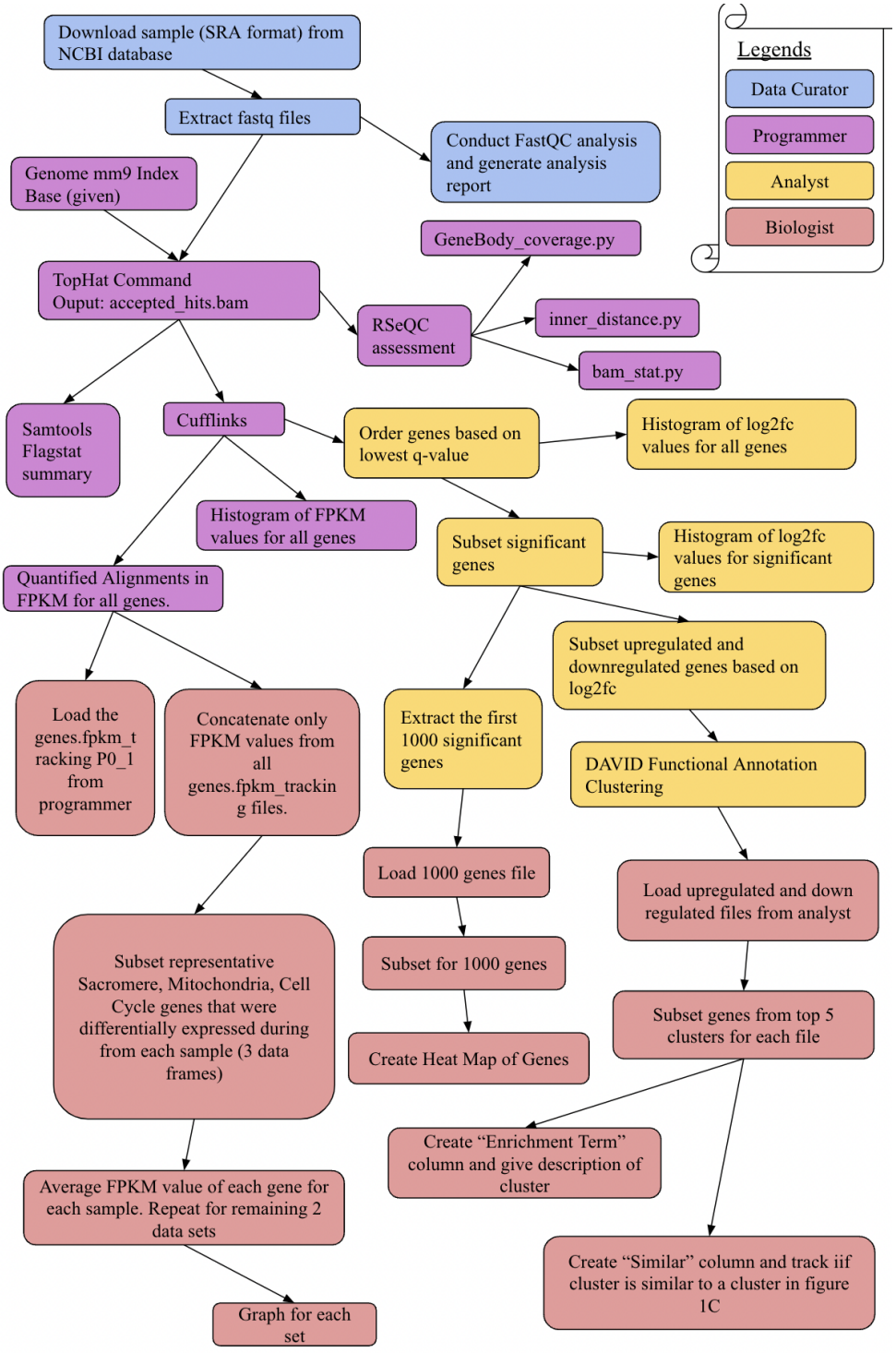
**Date:** 17th March, 2023

## Introduction:

The mammalian heart undergoes limited cardiomyocyte self-renewal throughout life and is even capable of modest regeneration early after birth (Verjans et al). It is still not understood why the cardiomyocytes only possess a limited capacity for regeneration and the aim of this paper is to underpin this process and investigate it at the molecular level. Identifying the key regulators can help develop therapeutic options for cardiac regeneration. To achieve this, the authors employed neonatal mice - who can regenerate their hearts post cardiac injury only up to the first week of their lives - as their models (O'Meara et al). The objectives of this paper are twofold - first to investigate if the myocytes revert become dedifferentiated during regeneration and secondly to analyze the transcriptional data to identify potential regulators of this process.

To explore the dedifferentiation, differentiation and regeneration of cardiomyocytes the authors profiled global gene expression patterns in various models, including in vitro differentiation of mouse embryonic stem cells to cardiomyocytes, in vivo cardiomyocyte maturation from neonate to adult, and a cardiomyocyte explant model where cells lose their differentiated phenotype (O'Meara et al). In this paper, they identified genes and gene networks that changed dynamically during these processes - including the reactivation of cell cycle genes and developmental programs during heart regeneration. Specifically, they identified interleukin 13 (IL13) as a new regulator of cardiac myocyte cell cycle entry, and found STAT6, STAT3, and periostin to be critical mediators of IL13 signaling in cardiac myocytes.

We attempted to recreate the results of this paper using a subset of the RNAseq data and conducted our analyses on that. Below is a workflow of our methods which will be described in great detail in the methods section.




**Workflow Diagram 1:** A step-by-step workflow describing our analysis to reproduce the paper results

**Data:**

For this study, the RNA sequencing data was obtained from various sources which include a combination of in vitro and in vivo models. For the in vivo experiments, the CD1 neonatal mice from the Charles River Labs (MA) were decapitated at P0, P4, and P7 stages, and by isoflurane overdose at 8–10 weeks of age (O'Meara et al). The authors performed whole heart ventricle isolation and pooled at least two heart ventricles for each replicate and used two replicates for the RNASeq analysis. Apical resection surgeries were performed on postnatal day 1 (P1) CD-1 mice and the heart apex from 3 animals pooled for each biological replicate before being processed for RNASeq analysis. The neonatal cardiac myocytes were dissociated from whole mouse hearts (P0 and P4), and also from sham and resected neonatal mouse hearts at 7 days post surgery using the Neonatal Heart Dissociation Kit (Miltenyi Biotec). Five to ten mouse pup hearts were pooled for each biological replicate and three biological replicates were generated per time point (O'Meara et al).

Total RNA was extracted from all samples using Trizol (Invitrogen) and the polyadenylated RNA was isolated using the Dynabeads mRNA purification kit (Invitrogen). The first strand was synthesized from the fragmented polyadenylated RNA using the Superscript III reverse transcription kit (Invitrogen) and the double stranded DNA was synthesized with DNA polymerase I (Invitrogen). End repair, A-tailing, adapter ligation and size selection were performed using the SRPI-Works System (Beckman Coulter) followed by minimal amplification and addition of barcodes by PCR. Paired-end 40 base pair read length sequencing was then performed on an Illumina HiSeq 2000. Sequence alignment and assembly was performed. Due to low RNA yield prior to RNASeq, For the the TrueSeq (Invitrogen) sample preparation protocol was used instead for the purified neonatal cardiac myocyte samples (O'Meara et al).

Finally, RNASeq data for differentiation of the embryonic stem cells into myocytes was obtained from Wamstad et al (2012).

**Methods:**

***Data Acquisition***

The sample SRR1727914.sra was downloaded from the NCBI database using the prefetch command in the SRA toolkit on the command line. The fastq files were extracted from the SRA files using a paired-end sequencing run, resulting in two files: SRR1727914_1.fastq and SRR1727914_2.fastq. Data quality was then analysed using FastQC.

***RNA-seq Data Analysis and Gene Enrichment Analysis***

In bioinformatics, Tophat and Bowtie are commonly used computational tools for analyzing high-throughput sequencing data, particularly for RNA sequencing (RNA-seq) data. Tophat is used for mapping RNA-seq reads to a reference genome, while Bowtie is used for ultrafast and memory-efficient alignment of short RNA sequence reads to a large reference genome. Flagstat was used and is a tool that generates statistics on the number of reads in a SAM or BAM file. It provides information on the total number of reads, the number of mapped and unmapped reads, and the number of reads that are properly paired. To analyze the RNA-seq data, Tophat was run as a batch job due to the large size of the data. Also, Cufflinks was used to quantify gene expression based on the aligned reads. The software versions and parameters used were Tophat version 2.1.1, RSeQC version 5.0.1, Cufflinks version 2.2.1, and R version 3.6.3 and Python 2.7 and Python 3.10.5. The analysis was run on a shared computer computing (SCC) cluster, with

Tophat running as a batch job. The entire analysis process required significant computational resources and took several hours to complete.

Two input fasta files were used: /projectnb/bf528/users/group_5/project_2/Samples/SRR1727914_1.fastq and /projectnb/bf528/users/group_5/project_2/Samples/SRR1727914_2.fastq. The genome index base was set as /project/bf528/project_2/reference/mm9/.

The resulting output BAM files produced using python2, bamtools and tophat is known as accepted.bam were further analyzed for quality using RSeQC. The "accepted_hits.bam" file is a binary file containing sequencing data, and the "samtools flagstat" command is a tool used to summarize information about the alignments in the file. The output provides various statistics about the sequencing data, such as the total number of reads, the number of secondary reads, the number of duplicates, and the number of mapped reads. It also provides information about the sequencing pairs, such as the number of properly paired reads and the number of singletons.

Quality assessment of the alignment was further conducted using RSeQC, python3, samtools, bowtie2, tophat which includes geneBody_coverage.py, inner_distance.py, and bam_stat.py. GeneBody_coverage.py outputs a plot showing the coverage of the genes. The plot showed that the reads were well-distributed across the gene body with a slight bias towards the 5' and 3' ends. Inner_distance.py calculates the insert size distribution of the paired-end reads. Bam_stat.py calculates various metrics such as the number of reads, the number of mapped and unmapped reads, and the read length. In the present study, no normalization or outlier detection was performed on the data.

However, a summarization method was used to obtain FPKM values using Cufflinks software, python3, samtools and RSeQC. The rationale for using Cufflinks was to quantify gene expression based on the aligned reads, as it is a popular and reliable tool for RNA-seq analysis. Cufflinks was run on the file P0_1_tophat/accepted_hits.bam, and the quantified alignments in FPKM for all genes in the file P0_1_cufflinks/genes.fpkm_tracking were obtained. Next, the file was then loaded into R, and a histogram of the FPKM values was created. Then the number of genes in the analysis, including any filtering used, was identified. The histogram of the FPKM values for all genes showed a wide distribution of gene expression levels. Genes with an FPKM value of zero were filtered out to focus on the more highly expressed genes in the analysis. Moreover, the histogram is generated using the ggplot2 package in R, and the resulting plot had several customizations. It has a title of "FPKM Values Histogram", with the x-axis labeled as "FPKM Values" and the y-axis labeled as "Frequency". The x-axis limits are set to range from 0 to 1000, and the y-axis was displayed on a log10 scale with scientific notation formatting. Outliers were removed from the histogram and other changes were incorporated: In the histogram 86 rows containing non-finite values were removed(stat_bin), transformation introduced infinite values in continuous y-axis, and lastly 16 rows containing missing values were removed (geom_bar).

The results from cufflinks differential expression tests (gene_exp.diff) were analyzed further to identify significant genes expressed between experimental conditions (P0 vs Ad) and for exploring the biological processes and pathways that are affected by these changes in gene expression.

The gene_exp.diff file was sorted based on ascending order of q-values. We use q-value since it is a modified p-value that takes in account false positives based on the number of tests. To get a visual representation of the difference in gene expression, a histogram was created for all genes using the log2foldchange values.

The significant genes were then subsetted and were used throughout the remainder of the analysis. Moreover, these significant genes were bifurcated into upregulated genes and downregulated genes based on their log2foldchange values. With a log2foldchange cutoff of 0, those genes that had a log2fc value greater than 0 were upregulated and those less than 0 were downregulated. For a visual understanding, another histogram was created for only the significant genes.

DAVID, which is a comprehensive set of functional annotation tools, was used to infer the biological significance of the significant genes. The upregulated genes and downregulated genes were uploaded on DAVID one at a time with the identifier set to Official_Gene_Symbol and species set to mus musculus. The Functional Annotation Tool was used and only three terms under Gene Ontology were selected for Functional Annotation Clustering. The three terms are ‘GOTERM_BP_FAT’, ‘GOTERM_CC_FAT’ and ‘GOTERM_MF_FAT’.

***Biological Interpretations***

This section uses the FPKM Expression gene matrices found for each sample to interpret biological significance. The analysis is split into 3 main subsections.

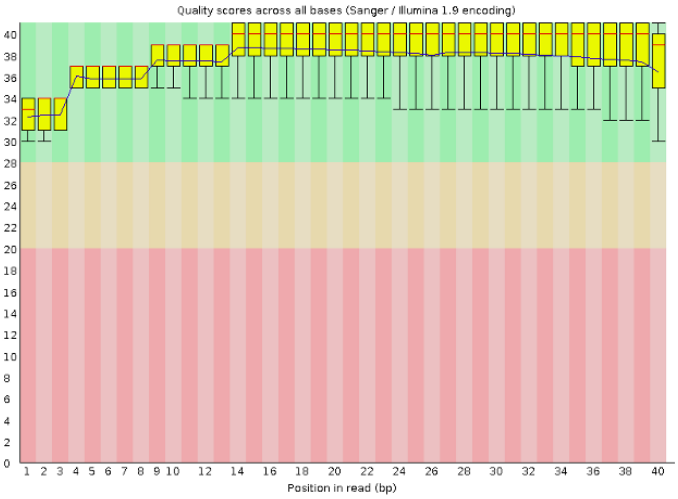
The first section is replicating Figure 1D and comparing the results with those of the paper by graphing the FPKM values of the representatively differentially expressed sarcomere genes, mitochondria genes, and the cell cycle genes for the P0, P4, P7, AD samples. The FPKM values the programer generated for P0_1 and the FPKM values of the genes from the other samples that were provided to us were all combined into a matrix. Each of the genes graphed in Figure 1D for each of the three gene types were subsetted into different matrices. An average of each of the two replicates for each sample (P0, P1,P4, and AD) was graphed, resulting in three graphs.

Next task was to identify if the list of common upregulated and down regulated genes found during differentiation that was provided by the analyst, resulted in similar gene enrichment terms found in the paper during differentiation. The top 5 clusters from each of the upregulated and down regulated files were selected. Two columns were added to the ends of each of the clusters in these files: “Gene Enrichment Term” and “Similar”. The “Gene Enrichment Term" was filled with a term that the biologist has used to describe what most of the genes in the cluster consisted of and the “Common” column denotes that the term is found in the corresponding table in Figure 1C with “yes” or " no".

The last task was to replicate the heat map of the top 1000 differentially expressed genes using the FPKM values. The names of the 1000 differentially expressed genes were provided by the analyst. The data used to generate the heat map was the matrix of the concatenated FPKM values for all eight samples for all genes. These 1000 genes were subsetted for using that initial matrix and the heat map was created.

## Results:

### *Data quality*



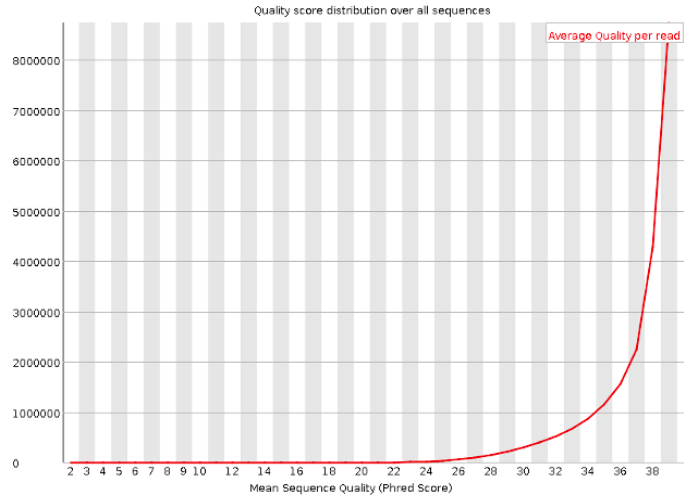
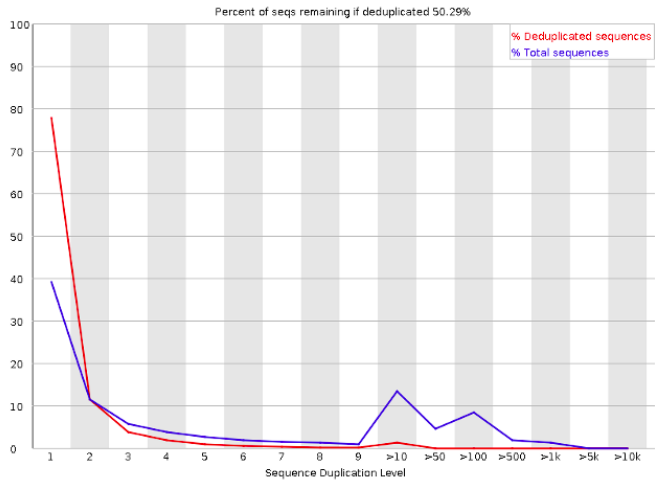




Fig 1. A, B Quality scores for SRR1727914_1.fastq



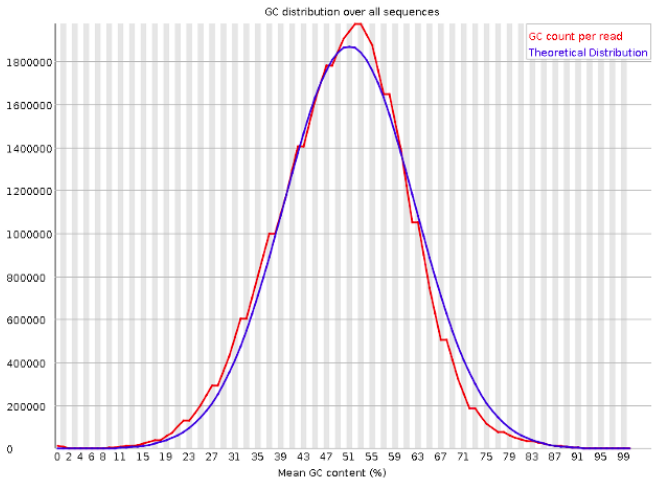




Fig 1. C: Overrepresented sequences in the SRR1727914_1.fastq samples, Fig 1. D: GC distribution over all sequences in the SRR1727914_1.fastq samples

**Fig 1:** Plots demonstrating quality of the data used in analysis. Plots obtained from FastQC Report. (png)

Per base sequence quality of our data shows very good quality calls and is stable throughout the length of the sequences (Fig1. A), the average quality per read is very high as well (Fig1. B).

There are no overrepresented sequences, and only very few sequences are duplicated (Fig1. C). GC distribution over all the sequences of the sample matches the theoretical distribution (Fig1. D). Overall, none of the sequences were flagged as poor quality exhibiting the high quality FastQ analysis.

***Samtools and RSeQC***

| | Number of Reads | Percentage of Reads |
|---|---|---|
| Mapped Reads | 49706999 | 100.00% |
| Unmapped Reads | 0 | 0.00% |
| Primary Reads | 41389334 | 83.27% |
| Secondary Reads | 8317665 | 16.73% |
| Unique Reads | 1452862 | 3.51% |
| Properly Paired Reads | 29422646 | 71.09% |
| Total Reads | 49706999 | 100.00% |

**Table 1:** Breakdown of Total Reads from Flagstat Tool

The breakdown of total reads from flagstat tool table (Table 1) produced by the Samtools tool reports several key metrics related to read alignment. Specifically, the analysis identified a total of 49,706,999 reads, with 100% of these reads passing quality control. Of the total reads, 41,389,334 (83.27%) were primary reads, while 8,317,665 (16.73%) were identified as secondary reads. Among the primary reads, 1,452,862 (3.51%) were classified as unique reads, indicating that they aligned to a single location in the reference genome. Additionally, 29,422,646 (71.09%) of the total reads were properly paired, meaning they were correctly oriented and had the appropriate insert size. Finally, there were no unmapped reads identified in the analysis.

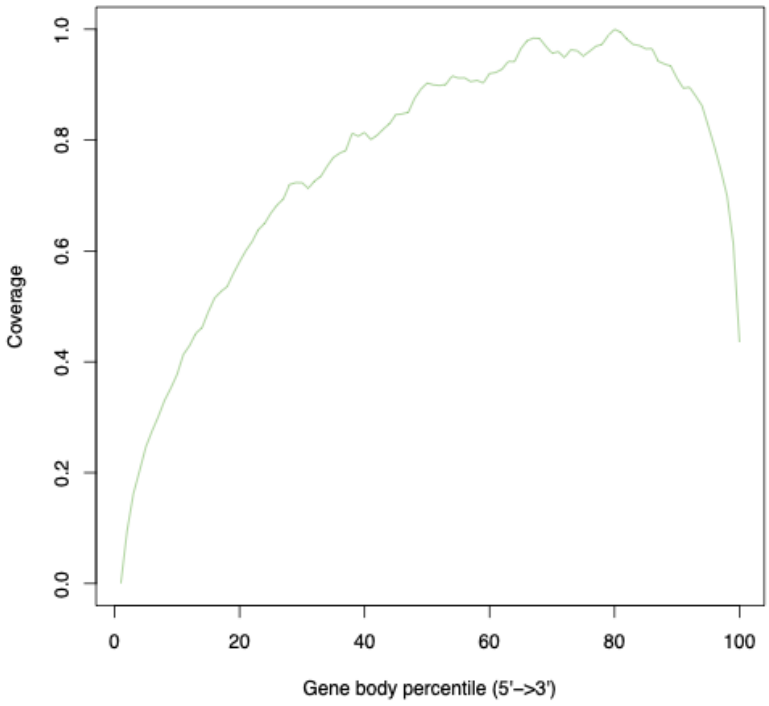



**Fig 2:** Plot demonstrating RSeQC output: Gene Body Coverage Curve (geneBody_coverage_curve.png)

In Fig. 2, A plot is produced demonstrating the RSeQC output of the geneBody_coverage.py curve. There is a high correlation between biological replicates. This indicates that the biological variation has been minimized and that the technical variation is small. There is also a high percentage of mapped reads. This high-quality dataset should have a high percentage of reads that can be mapped to the reference genome. This indicates that the sequencing reads are of high quality and that the library preparation was successful. This also indicates that there is a low percentage of reads that cannot be mapped to the reference genome or transcriptome. As seen in the curve there is some 3' bias as the curve peaks around 80 percentile when the coverage is nearing 1.0.

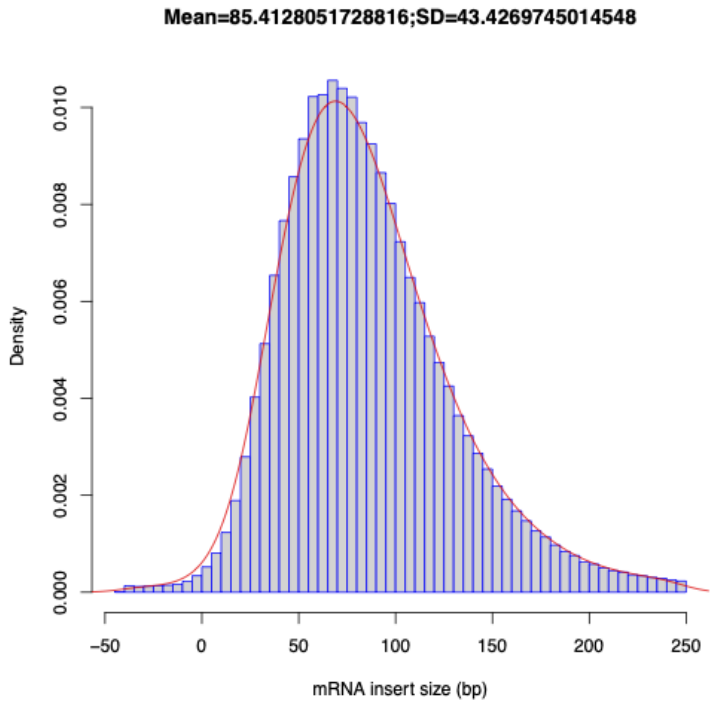



**Fig 3:**Plot demonstrating RSeQC output: Inner Distance Curve (inner_distance.png)

Fig 3 describes a RSeQC output of the inner distance curve. The plot includes two curves: a red curve that represents the distribution of inner distances for read pairs that map to the same gene/chromosome, and a blue curve that represents the distribution of inner distances for read pairs that map to different genes. This paired-end RNA-seq dataset has a well-defined peak in the red curve that corresponds to the expected fragment size of the library, and a long tail in the blue curve that indicates a high rate of read pairs spanning across different genes. The peak in the red curve at the expected fragment size is well-defined which indicates that the library preparation was successful and that the majority of the sequenced fragments are of the expected size. A long tail in the blue curve indicates a high rate of read pairs spanning across different chromosomes. This is expected in RNA-seq data, as many genes span across multiple chromosomes, and it is an indication that the library preparation and sequencing were successful in capturing these complex structures. This inner distance curve is somewhat normally distributed and has an average of 85.41 mRNA insert size (bp) and a standard deviation as +/- 43.43 SD.

| | Number of Reads | Percentage of Reads |
|---|---|---|
| Mapped Reads | 49,706,999 | 100.00% |
| Unmapped Reads | 0 | 0.00% |
| Primary Reads | 41,389,334 | 83.27% |
| Secondary Reads | 8,317,665 | 16.73% |
| Unique Reads | 38,489,380 | 93.00% |
| Non-Unique Reads | 2,899,954 | 3.51% |
| Proper Pairs | 27,972,916 | 67.58% |
| Total Reads | 49,706,999 | 100.00% |

**Table 3:** Breakdown of Total Reads from BamStat Tool

In Table 3, a breakdown of primary reads from the bamstat tool are produced. Based on the bam_stat analysis, the total number of reads was 49,706,999, and it was confirmed that all the reads passed quality control and were mapped. Of these reads, 8,317,665 were identified as non-primary hits and as secondary hits, while 2,899,954 were classified as non-unique reads, indicating low-quality alignment. The remaining 38,489,380 reads were classified as unique, indicating high alignment quality. Finally, out of the total reads, 27,972,916 were mapped in proper pairs, while only 4 reads were properly mapped to different chromosomes.

Flagstat is primarily used for generating summary statistics on a BAM file, while Samtools and Bamtools are more comprehensive suites that can perform a variety of tasks. Bam stat.py is a specific Python script that provides detailed statistics on a BAM file.

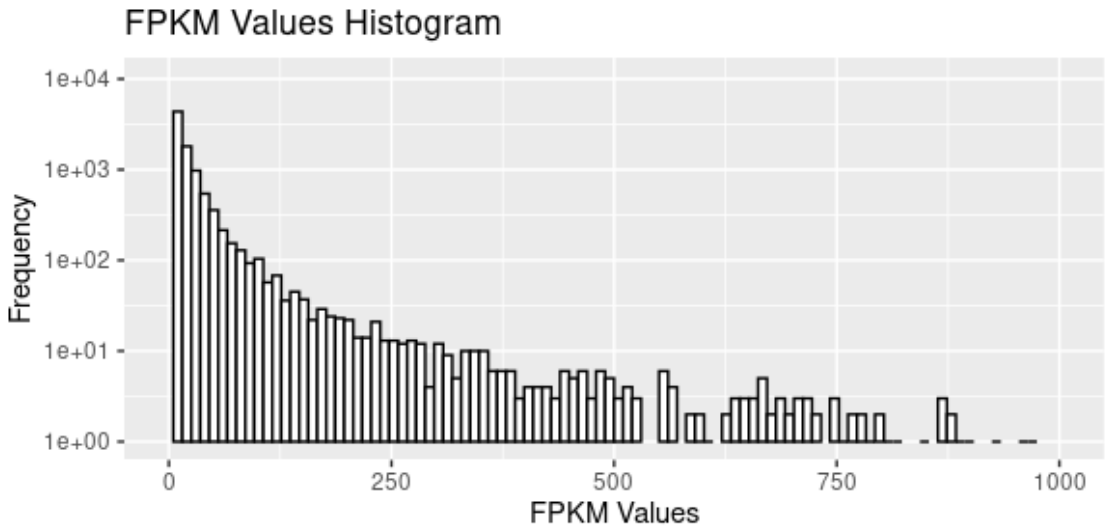



**Fig 4:** Histogram plot of the FPKM Values produced by Cufflinks Sequence Alignment (cufflinks_histogram.png)

Fig 4 is a Histogram displaying the distribution of log10 transformed FPKM normalized counts for 20,487 genes that passed the FPKM>0 filtering condition. The y axis displays gene counts. The log scaled FPKM values are normally distributed around 0. Following the quantification of gene expression for P0, differential gene expression between the P0 and adult timepoints was quantified using the Cufflinks program Cuffdiff (Trapnell et al., 2013). The analysis was performed on a 16-processor node, specifying 16 threads for analysis, and ran for approximately one hour. The output obtained from Cuffdiff was examined in R to identify the differentially expressed genes associated with myocyte differentiation. The top ten differentially expressed genes were extracted based on their q-values from Cuffdiff (Table 1). A histogram was constructed to compare the distribution frequency of all genes versus significant genes, using the log2 fold change values also obtained from Cuffdiff (Figure 4). This histogram proved to be useful in comparing the frequency of genes that were significant and all the genes combined. The up and down-regulated genes were determined by the log2 fold change values obtained from Cuffdiff. The gene sets were then organized into functionally related clusters using DAVID (Huang et al., 2008). DAVID Functional Annotation Clustering groups gene sets based on the genes they share.

***DAVID Functional Annotation Clustering***

| Gene | Value 1 | Value 2 | Log2FC | p-value | q-value |
|---|---|---|---|---|---|
| Adhfe1 | 12.71 | 25.74 | 1.017 | 5e-05 | 0.0003206 |
| Tmem70 | 36.95 | 80.96 | 1.132 | 5e-05 | 0.0003206 |
| Gsta3 | 0.413 | 6.770 | 4.036 | 5e-05 | 0.0003206 |
| Lmbrd1 | 6.588 | 12.68 | 0.9441 | 5e-05 | 0.0003206 |

| | | | | | |
|---|---|---|---|---|---|
| Dst | 19.72 | 51.60 | 1.388 | 5e-05 | 0.0003206 |
| Plekhb2 | 25.85 | 68.60 | 1.408 | 5e-05 | 0.0003206 |
| Cox5b | 505.4 | 881.8 | 0.8029 | 5e-05 | 0.0003206 |
| Mrpl30 | 56.21 | 124.3 | 1.145 | 5e-05 | 0.0003206 |
| Tmem182 | 46.22 | 103.5 | 1.163 | 5e-05 | 0.0003206 |
| Nck2 | 12.17 | 6.300 | -0.9504 | 5e-05 | 0.0003206 |

**Table 4:** Table of the top 10 differentially expressed genes with FPKM values (Value 1 and Value 2), log2FC, p-value and q-value.

Table 4 indicates the top ten differentially expressed genes which is a result of Cufflinks differential expression. Value 1 indicates the FPKM value for postnatal day 0 mice and Value 2 indicates the FPKM value for adult mice. We can see that all the genes have the same q-value indicating high significance.

From figure 6, we infer that there are many genes with a log2foldchange value at zero and just below zero. However, to get a clear understanding of the upregulated and downregulated spread, we created another histogram with only the significant genes as seen in figure 7. The total number of significant genes was 5247 with 2830 upregulated genes and 2597 downregulated genes. From figure 7, we can tell that there are slightly higher upregulated genes (bins on the positive side of zero) in comparison to downregulated genes (bins on the negative size of 0).

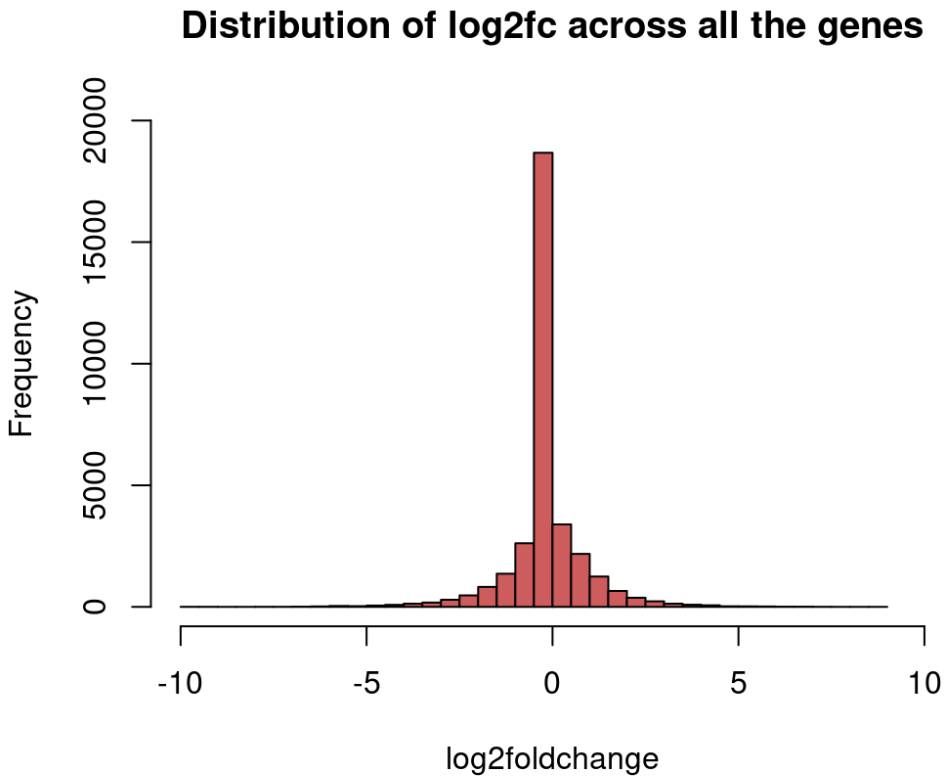



**Fig 6:** Histogram of log2foldchange values for all the differentially expressed genes. (log2foldchange.png)

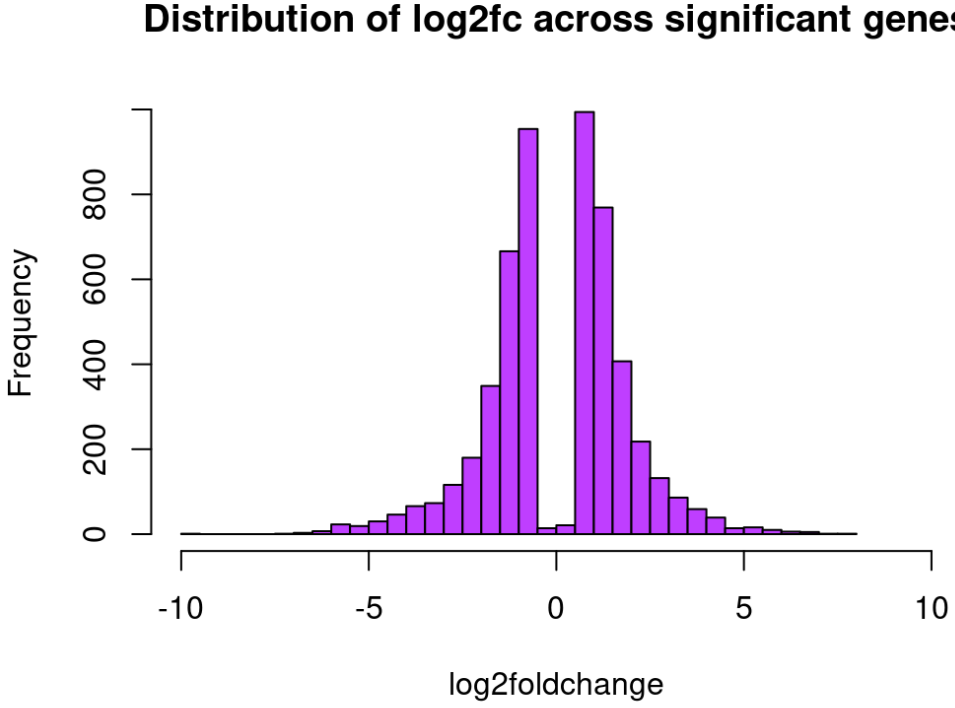


**Fig 7:** Histogram of log2foldchange values for only the significant genes.(log2fc_significant.png)

The results from the functional annotation clustering on DAVID were quite insightful as most of the top enriched terms were in concordance with the paper. In total, there were 839 clusters for upregulated genes and 739 clusters for downregulated genes. All the enrichment terms seen below in Table 5 were highlighted in the paper and showcased high enrichment scores in our results.

The enrichment analysis results, as anticipated, showed that during in vitro and in vivo differentiation, differentially expressed genes were primarily up-regulated in sarcomere and mitochondrial related functions, while down-regulated in cell cycle processes. The pathways from upregulated genes reflect sarcomere assembly and organization during cardiac myocyte differentiation and maturation. Moreover, from the down-regulated pathways, Cell cycle exit is a key feature of mature cardiac myocytes, and the failure to re-enter the cell cycle is believed to contribute to the lack of cardiac regeneration in adult mammals. Downregulation of genes associated with DNA repair and mRNA processing also fall in trend with the inability of cardiac regeneration.

| **Upregulated** | | **Downregulated** | |
|---|---|---|---|
| **GO term** | **Enrichment Score** | **GO term** | **Enrichment Score** |
| Mitochondrian | 66.63 | Nucleic Acid Binding | 39.48 |
| Oxidative Reduction | 33.74 | Chromosome | 38.54 |
| Electron Transport Chain | 33.74 | Transcription, DNA templated | 30.15 |
| Cellular Respiration | 33.74 | DNA repair | 23.83 |
| Monovalent inorganic cation transport | 13.95 | Cell Cycle | 20.36 |
| Regulation of heart contraction | 10.75 | Organelle Fission | 20.36 |
| Sarcomere | 7.77 | mRNA Processing | 9.6 |

**Table 5:** Table summarizing the top clusters with enrichment scores for up and down regulated genes.

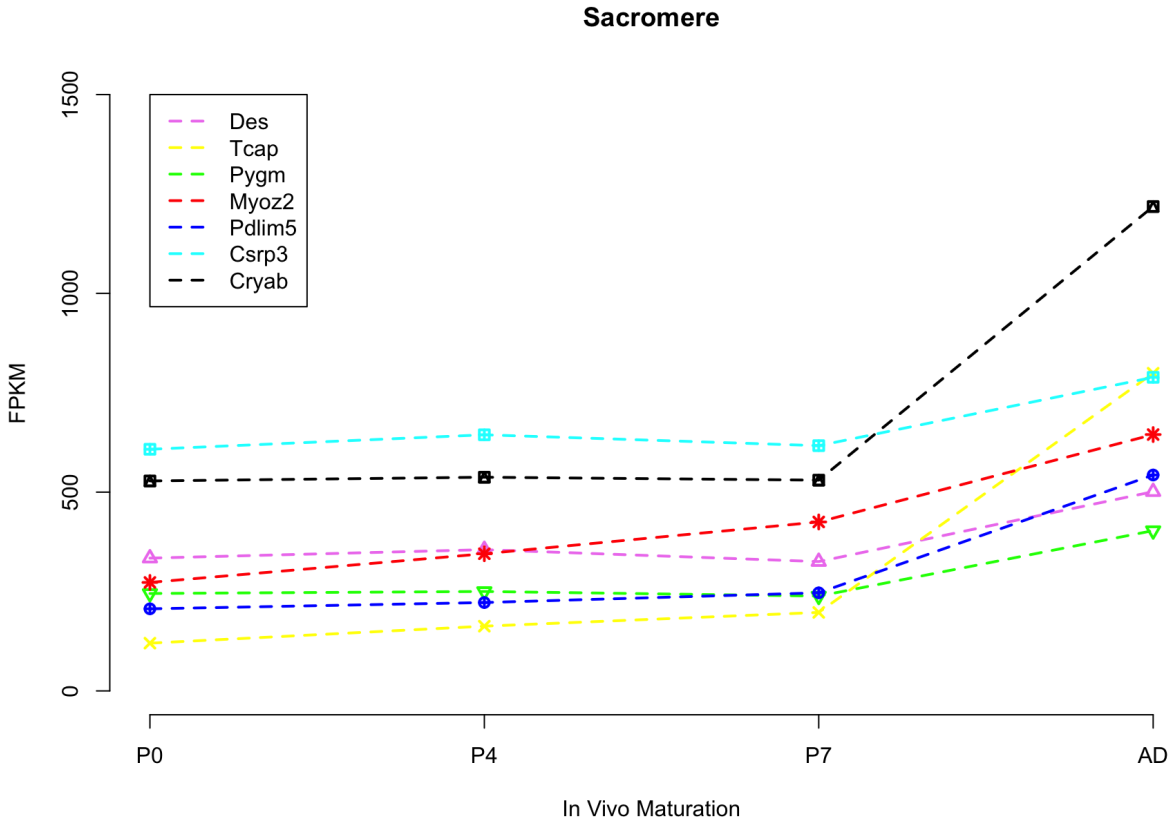



**Fig 8:** FPKM values of representative sarcomere genes that were significantly differentially expressed during in vivo maturation. (fig8.png)

A noticeable difference was that the FPKM values of Cryab in the Sarcomere gene set seemed slightly higher than the values obtained by the authors at the P0, P4, and P7 stages. And Tcap from the same graph is slightly less than the authors' results in the P4, P7, AD stages. Overall the results generated were replicated similarly to the graph by the authors.

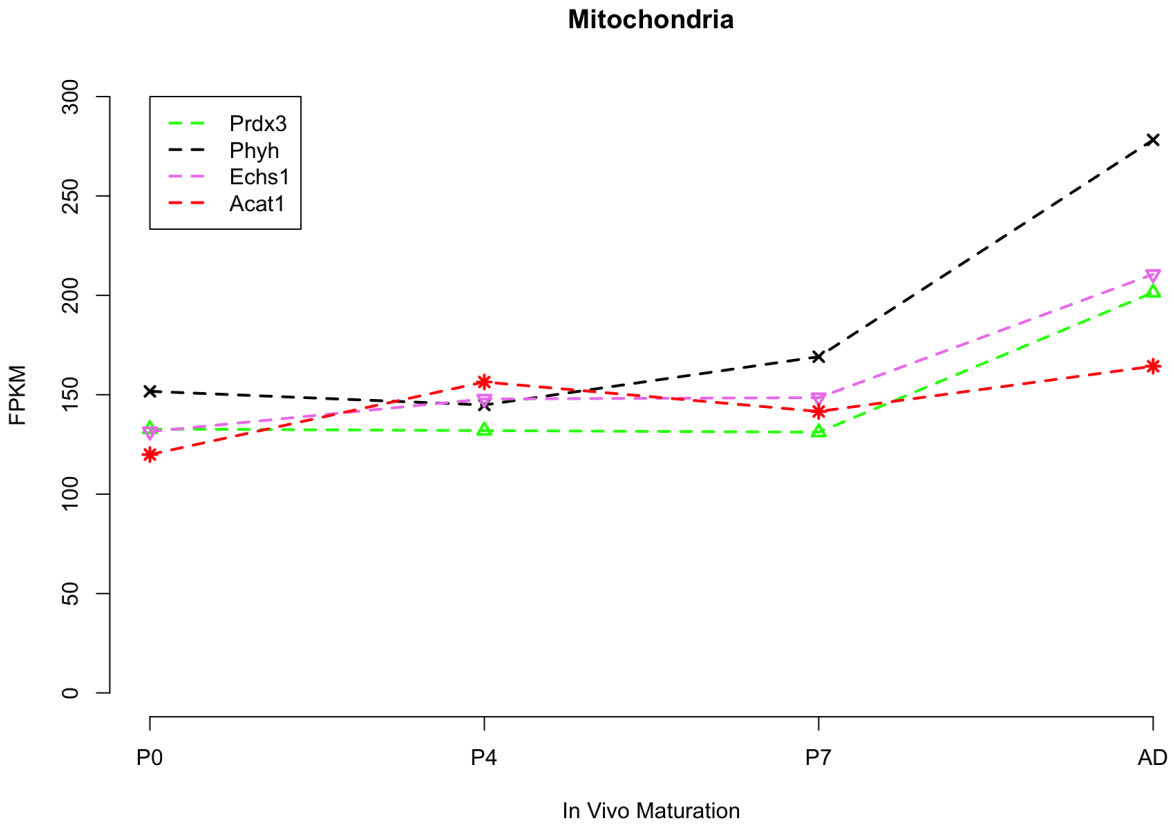



**Fig 9:**FPKM values of representative mitochondria genes that were significantly differentially expressed during in vivo maturation (fig9.png)

The trend of the genes here is similar to the one in the paper, however this graph was missing "Mpc1" and "Slc25a11". The data this graph was generated from did not note these genes as significantly differentially expressed.

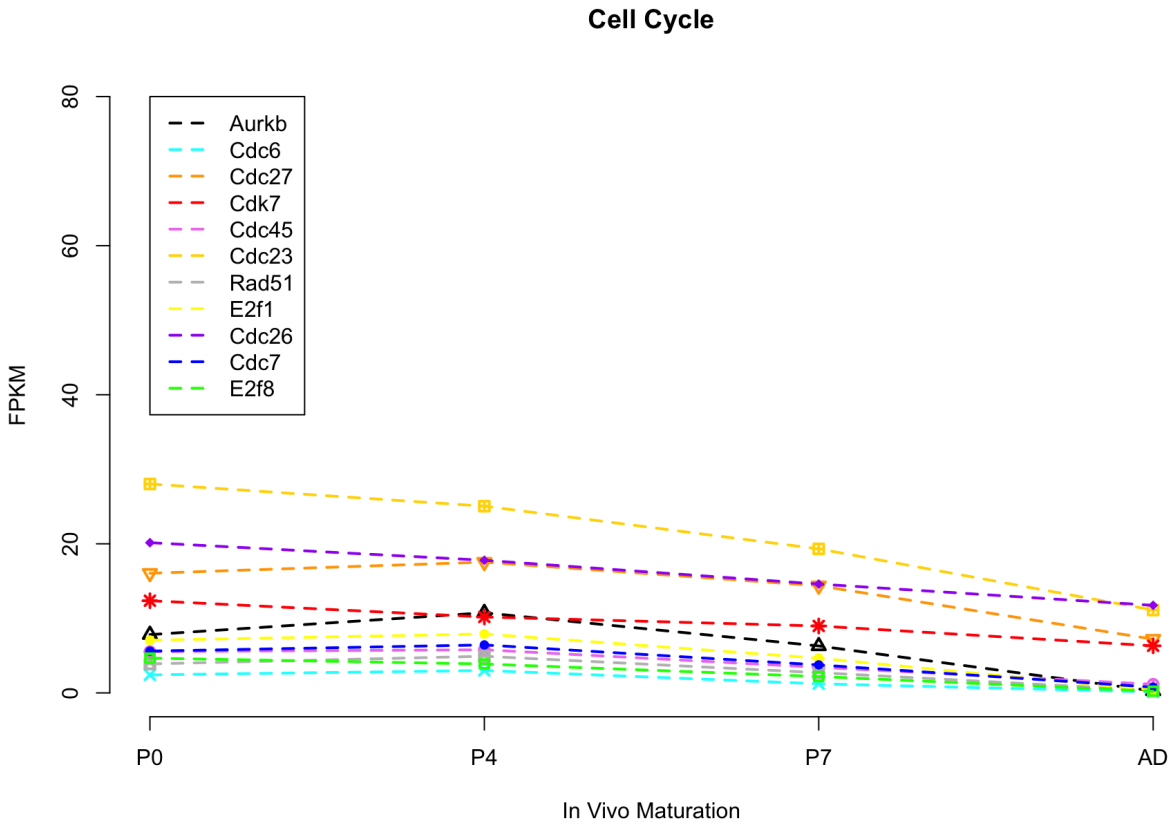



**Fig 10:** FPKM values of representative cell cycle genes that were significantly differentially expressed during in vivo maturation (fig10.png)

The Cell Cycle graph was missing the “Bora”gene. The trends among the graphs were fairly similar to what was observed in the results of the authors, for the exception of the missing gene. This further provides evidence for the higher expression of cell cycle related genes in the earlier prenatal days.

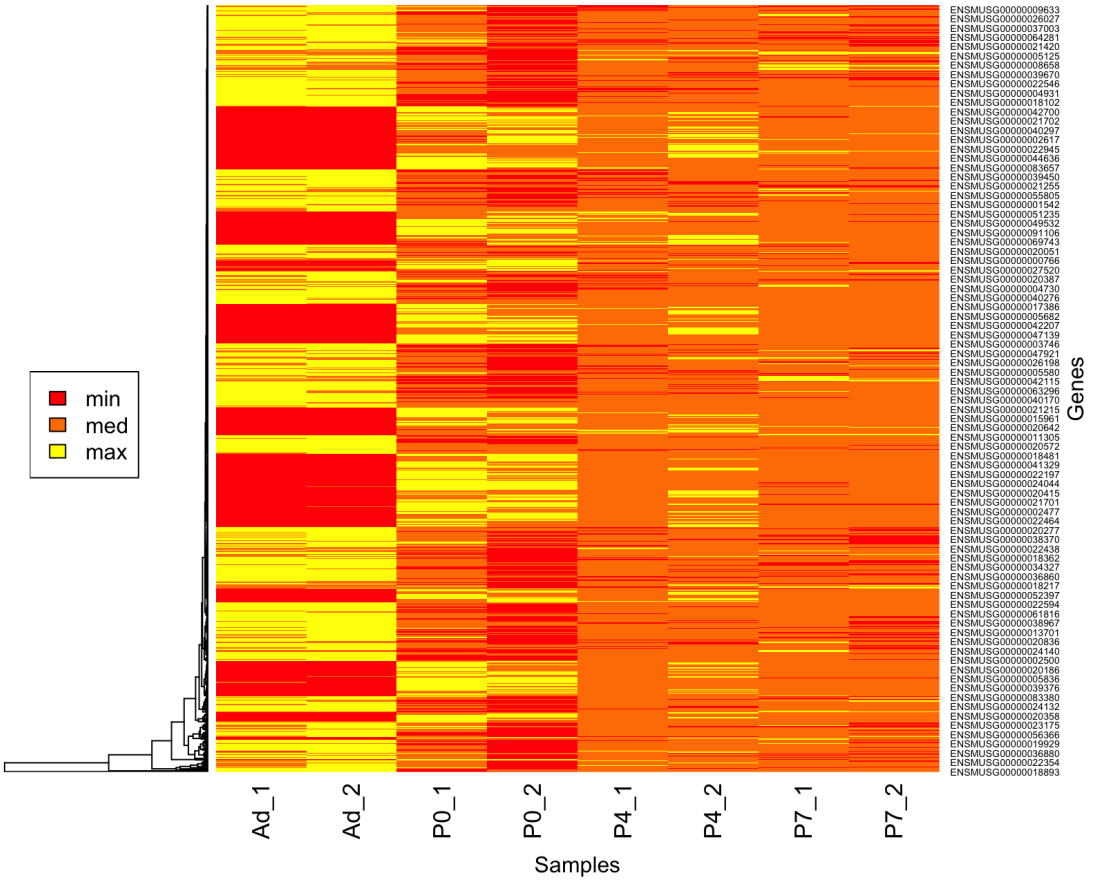



**Fig 11**: Heatmap of the top 1000 genes that were differentially expressed between the AD and P0 samples. (fig11.png)

The results of the heat map were to be compared with that of figure 2A. The input data was not provided with gene ontology to completely replicate figure 2A from the paper but the general pattern of expression levels seem to be similar among the two. For example both heat maps show to have more highly differentially expressed genes in the adult stage.

**Discussion:**

The presented results provide a comprehensive overview of the quality and characteristics of RNA-seq data obtained from the analyzed experiment. The use of different tools and metrics allowed for a thorough assessment of the sequencing quality, alignment quality, and gene expression quantification.

Our numbers are very close to the number of the significant genes mentioned in the paper. We obtain 5427 significant genes and the paper mentions 5823 significant genes for P0 vs Ad. However, the individual count of upregulated and downregulated genes vary with what the authors of the paper mentioned. From our analysis, we get 2597 downregulated genes and 2830 upregulated genes, whereas, the authors mention 4341 downregulated genes and 1482 upregulated genes.

We see this difference primarily because the authors did not mention their log2foldchange cut off which plays an important role in differentiating the upregulated and downregulated genes. We took a standard of 0 which might not be what the authors considered. Secondly, results from a complete computational pipeline are very hard to replicate so changes in the version of the packages might add to the slight discrepancy in the numbers.

However, we are of the opinion that this paper comparatively allowed for high reproducibility because our total number of significant genes were in the same range. Moreover, the Gene Ontology terms corroborated our opinion because we obtained almost all of the enrichment terms which they mentioned with significant enrichment scores. With our analysis, it was easy to understand why adult mice lose their ability for cardiac regeneration which was the focal point of the paper.

It was easy to verify if the results of the study were replicated for some parts of the study. For example, by comparing the graphs in figure 1D with the graphs produced in this study it was easy to identify if they were similar. But it was harder to evaluate the level of similarity between the heatmap from this project and the heat map in figure 2A for the annotations information was not provided. A short list of the enriched GO terms found in the study is in Figure 1C.

When comparing the DAVID results obtained with the author's results, there seemed to be quite some differences. The gene enrichment scores obtained in this project were fairly higher than the ones the authors reported for upregulated genes. And the scores from the down regulated clusters are much lower than the ones reported by the author. Though it was tough to determine the exact terms that were representative of the cluster, it seemed most of the terms were common between the corresponding clusters in the results of the project and of the authors. Mitochondria, respiration, and metabolism were common terms in the upregulated genes. Cell cycle, rna processing, terms related to organelle membrane and were common among the down regulated genes.

**Conclusion:**

Tophat and Bowtie are important computational tools used in bioinformatics for analyzing high-throughput sequencing data. Using these tools in combination with RSeQC and Cufflinks allows for efficient analysis of RNA-seq data, resulting in a better understanding of gene expression levels. Overall, the presented results demonstrate the high quality of the RNA-seq data and the successful alignment and quantification of gene expression. These findings can be used to further understand the biological processes associated with myocyte differentiation and can serve as a basis for future studies.

The paper provides detailed information on the methods used for data analysis and the software packages employed, which can aid in assessing the reproducibility of the results. Additionally, the authors could have included more information on the quality control metrics of the datasets and the raw data availability to facilitate the replication of the study. Based on the text and figures provided in the paper, it was not explicitly stated whether the reader was able to replicate the results of the study. However, the paper provides detailed information on the methods used for data analysis and the software packages employed, which can aid in assessing the reproducibility of the results.

The figures and tables for biological interpretation that were generated from the project's data were fairly similar to the results attained by the authors, with the exception of Fig 11, which could not be well determined.

Works Cited


Verjans R, van Bilsen M, Schroen B. Reviewing the Limitations of Adult Mammalian Cardiac Regeneration: Noncoding RNAs as Regulators of Cardiomyogenesis. Biomolecules. 2020 Feb 10;10(2):262. doi: 10.3390/biom10020262. PMID: 32050588; PMCID: PMC7072544.

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.

SRA Toolkit Development Team, SRA Toolkit, https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software

Wamstad JA, Alexander JM, Truty RM, Shrikumar A, Li F, Eilertson KE, Ding H, Wylie JN, Pico AR, Capra JA, Erwin G, Kattman SJ, Keller GM, Srivastava D, Levine SS, Pollard KS, Holloway AK, Boyer LA, Bruneau BG. Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage. Cell. 2012 Sep 28;151(1):206-20. doi: 10.1016/j.cell.2012.07.035. Epub 2012 Sep 12. PMID: 22981692; PMCID: PMC3462286.

Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation Cole Trapnell, Brian Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Jeltje van Baren, Steven Salzberg, Barbara Wold, Lior Pachter. Nature Biotechnology, 2010, doi:10.1038/nbt.1621(cufflinks)

TopHat tool, Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics doi:10.1093/bioinformatics/btp120(Tophat) https://ccb.jhu.edu/software/tophat/index.shtml

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

Samtools, Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H, Twelve years of SAMtools and BCFtools, GigaScience (2021) 10(2) giab008 [33590861] (SAM tools)

Python2, Van Rossum, G., & Drake Jr, F. L. (1995). Python reference manual. Centrum voor Wiskunde en Informatica Amsterdam. (python)

Python3, Van Rossum, G., & Drake, F. L. (2009). Python 3 Reference Manual. Scotts Valley, CA: CreateSpace.

R, R Core Team. (2016). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from https://www.R-project.org/(R)