BF528 Individual Project

Transcriptional Profile of Mammalian Cardiac Regeneration with mRNA-Seq

(Programmer & Biologist)

Mrugakshi Chidrawar

(Group - Dreadlocks)

INTRODUCTION

Neonatal mice have remarkable plasticity in organ tissue during early development. Following traumatic damage, they can regenerate their hearts and restore normal function, although this capability only lasts into the first week of infancy. As of the 2015 publication of O’Meara et al’s paper “Transcriptional reversion of cardiac myocyte fate during mammalian cardiac regeneration”, however, the transcriptional machinery underlying this phenomenon had yet to be elucidated. A fundamental question that must be answered prior to the development of an understanding of this phenomenon is whether myocytes revert the transcriptional phenotype to a less differentiated state during regeneration and further systematically interrogate the transcriptional data to identify and validate potential molecular pathways that may be involved in that process. These questions can be answered through the exploration and analysis of transcriptional data. O’Meara et al. use high-throughput RNA sequencing to address these questions. This approach allows the authors to detect almost any mRNA molecule concentrated above the limit of detection in a sample. This contrasts with older methods, such as microarrays, that only measure molecules with specific sequences.

This project aims to explore the differences in the transcriptional profiles of neonatal and adult mice using high throughput sequencing mRNA abundance data from O'Meara et al. 2015. Unlike the original study, which takes several stages of development during the healing process after an injury into account, this study only focuses on RNA-seq data from neonatal mice at day 0 and adult mice aged 8 to 10 weeks to duplicate the first section of the O'Meara et al's research by identifying differentially expressed genes using the transcriptional data used in the original study.

DATA

The sample file SRR1727914.sra is downloaded in the form of a short read archive (SRA) file from the accession number GSM1570702 (vP0_1). This file contains the sequencing data uploaded to the public database Gene Expression Omnibus from GEO Series GSE64403. For alignment and analysis, pre-processed data files from previously completed project are used for this study, including tables 2 and 3 that display the gene enrichment analysis results for upregulated and downregulated genes using the *DAVID* Functional Annotation Clustering tool. These are used to evaluate the overlapping biological pathways with those from the original study.

METHODS

Alignment and Quality checks:

TopHat, a rapid splice junction mapper for RNA-Seq reads, is used to align two fastq files derived from P0 samples to the mm9 mouse reference genome. It uses the ultra high-throughput short read aligner Bowtie to align RNA-Seq reads to mammalian-sized genomes, then analyzes the mapping findings to find splice junctions between exons. For this analysis, mouse genome mm9 is used as reference. The predicted inner distance between mate pairs is set to 200bp, and each read is divided into pieces of at least 20bp in length. This study only looks for reads across junctions mentioned in the given junctions file, allowing for one mismatch in each segment alignment. The original reads are combined with any alignments detected by TopHat in a BAM encoded file, which is a binary version of the SAM (Sequence Alignment/Map) format. Because TopHat is a computationally intensive tool, it is necessary to run it on an HPC (High Performance Cluster), which takes around an hour to finish the alignment and mapping. The *SAMTools (v1.10)* flagstat tool is used to assess whether alignment readings passed or failed, primarily to determine whether any erroneous mapping to alternate chromosomes, reads, or duplication has happened.

After that, using *RseQC (v3.0.0) Utilities*, the previously stated BAM file is indexed, and quality control metrics are obtained. This is done to confirm that there were no alignment or mapping issues, as well as to ensure the original dataset's integrity. To perform quality checks, three modules from this package are utilized: *geneBody_coverage.py, inner_distance.py, and bam_stat.py*. Over the gene, the geneBody_coverage.py module estimates RNA seq read coverage, inner_distance.py calculates inner distance between read pairs, and bam_stat.py summarizes the mapping statistics of the BAM file.

Cufflinks (v2.2.1) is used to count how many reads mapped to annotated regions after the FastQ files were mapped with quality control checks and thus translating the RNA seq data into gene-based form. The gene annotation file used is based on the mice genome (mm9.gtf), and it is combined with the reference genome and the indexed BAM file created in the previous phase to create a gene tracking file with FPKM (Fragments per Kilobase of transcript per Million mapped reads) values which reflects the total number of fragments for all genes for a specific region. Next, a graphical representation of the distribution of log10 FPKM values is constructed, with genes with an FPKM value larger than zero being filtered out (Figure 3). There are, in total, 23264 genes with FPKM values greater than zero that are dropped. The last component of the analysis is performed using *Cuffdiff*, a Cufflinks tool, which is utilized to find genes that are differentially expressed. Its output is then evaluated in order to derive biological conclusions.

Biological Interpretation:

Tables 2 and 3, which summarize the top cluster results of the differential expression statistics generated for P0 vs Ad, are updated to add annotations for biological pathways that demonstrate overlap with O'Meara et al’s findings. The file *NIHMS647083-supplement-2.xlsx* from the original study's supplementary data is used for this assessment. The enriched gene ontology terms are searched for in the file using R.

The FPKM values of the 3 lists of genes specific to the most prominent GO terms discovered in the analysis, Sarcomere, Mitochondria and Cell cycle, that are significantly expressed in mice on postnatal day 0 (P0), day 4 (P4), day 7 (P7), and adult (Ad) are plotted against the biological age of the sample using the *ggplot* package in Rstudio (Figure 4). For this purpose, the P0_1

FPKM, extracted from the differentially expressed genes file generated by cufflinks, is merged with the other 7 samples provided in the project sample files.

Finally, the p*heatmap* function in R is used to create a clustered heatmap of the top 1000 genes discovered to be differentially expressed between P0 and Ad.

RESULTS

Each QC category from the samtools flagstat output is provided as pass or fail followed by a category description. The properly matched flag is set after all of the alignments pass QC. The output of the flagstat tool is summarized in the form of a table displaying statistics for important read categories (Table 1). As shown in the align summary table, there are a total of 49,706,999 reads with 49,706,999 mapped reads, resulting in an overall mapping rate of 100 percent. The total number of unique mapped reads is 41,389,334 (83.27%), which is made up of 20,878,784 left reads and 20,510,550 right reads. Multi-mapped reads account for 8,317,665 (16.73%) of the total. Singletons, or reads that are mapped but do not have mates, account for 1,452,862 reads (3.51 percent).

Category	Statistic	Percentage
Total number of reads	49706999	100%
Number of mapped reads	49706999	100%
Number of unique reads	41389334	83.27%
Number of multi-mapped reads	8317665	16.73%
Number of unaligned reads	0	0%

Table 1. Summary statistics of counts for various read categories from samtools flagstat tool

To find the bias in read values, the geneBody coverage graph (Figure 1) is mapped between nucleotide position and number of reads using RSeQC. It is observed that the 3' end of genes has better coverage than the 5' end, indicating a possible 3' bias, which suggests that a higher proportion of reads are concentrated in that area. The coverage map shows a drop off in coverage near the ends, which is expected and is typical of high-throughput sequencing. The distance between two paired reads, measured in mRNA length, is determined using the Insert Size or inner distance graph (Figure 2). The mean of the distribution and the standard deviation are 85.4 base pairs and 43.42 base pairs respectively. The limited number of negative values suggest that there is little overlap between two reads, implying that there is little to no error.

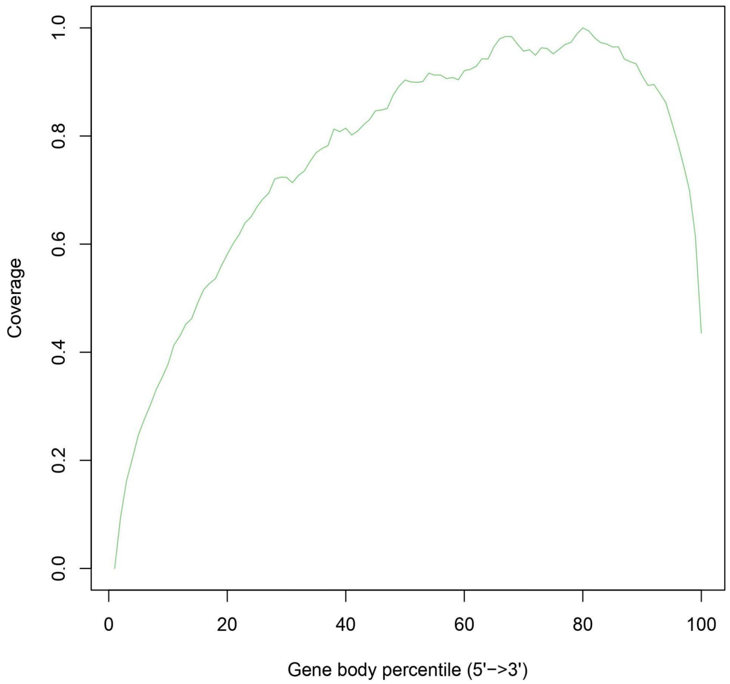

Figure 1. Gene body coverage plot showcasing the relation between coverage and gene body percentile

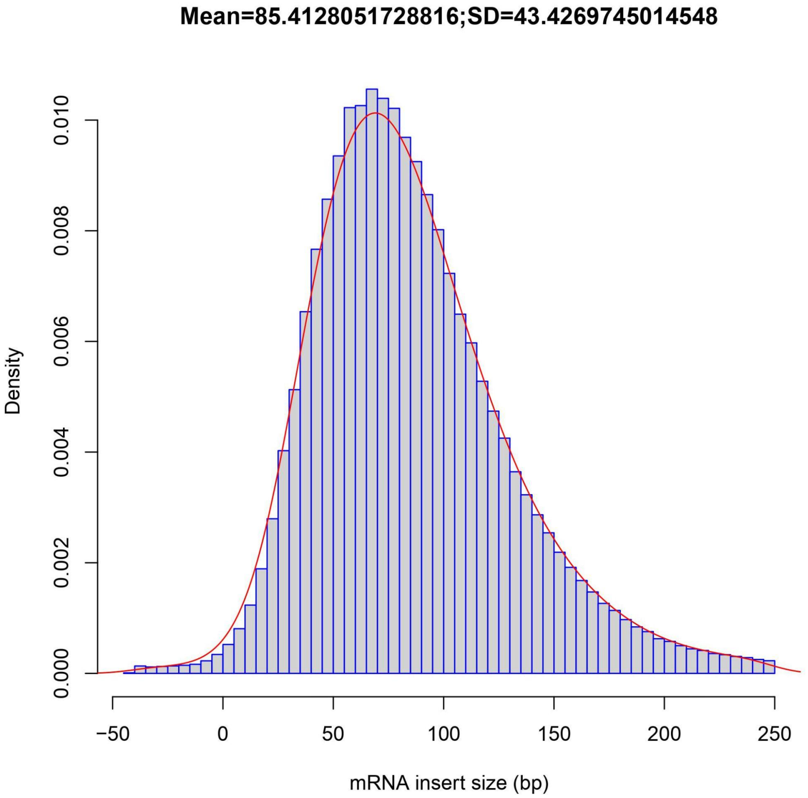

Figure 2. Distribution of inner distance between two paired ends

FPKM is a gene expression unit and therefore, the higher a gene's FPKM value, higher is the gene's expression. There are a total of 14,205 genes after dropping those with zero FPKM values. We can infer about the intricacies of expression profiles by looking at the distribution of the expression levels (Figure 3). Most of the FPKM values are concentrated around 1 with a slightly skewed tail at the right.

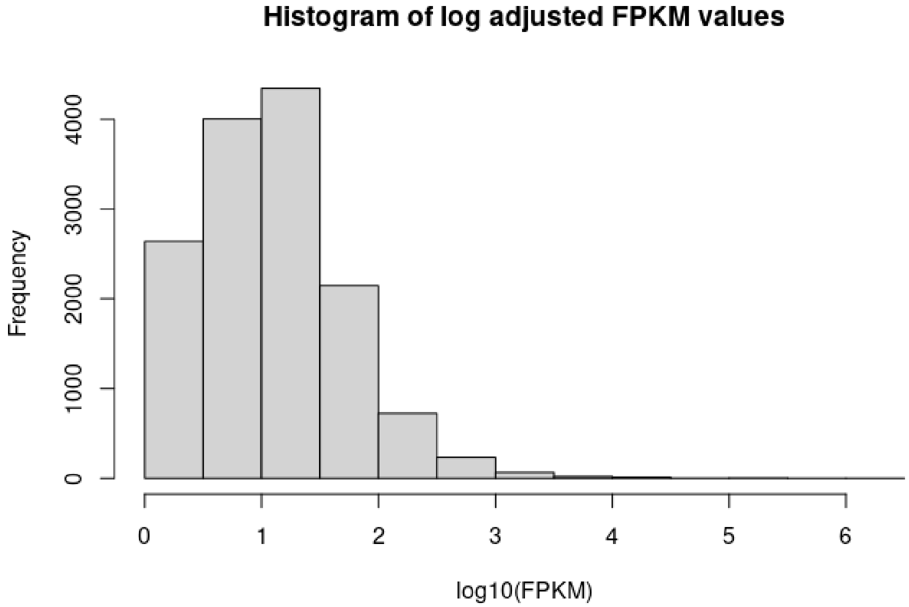

Figure 3. Distribution of log adjusted FPKM values for genes with FPKM values greater than 0

For differentially expressed genes identified during in vivo maturation, the FPKM values fold changes of representative genes of Sarcomere, Mitochondria, and Cell Cycle are displayed against the biological age of the sample (postnatal days 0(P0), 4(P4), 7(P7), and adult(Ad)) in figure 4. None of the figures are identical to those in the publication [O'Meara, C.C. et al], but they indicate similar tendencies throughout the in vivo development phase. Across the maturation period, the FPKM values of the sarcomere and mitochondrial genes have increased, while the FPKM values of the cell cycle genes have decreased. This shows that from postnatal to adult maturation, sarcomere and mitochondrial genes are upregulated whereas cell cycle genes are downregulated.

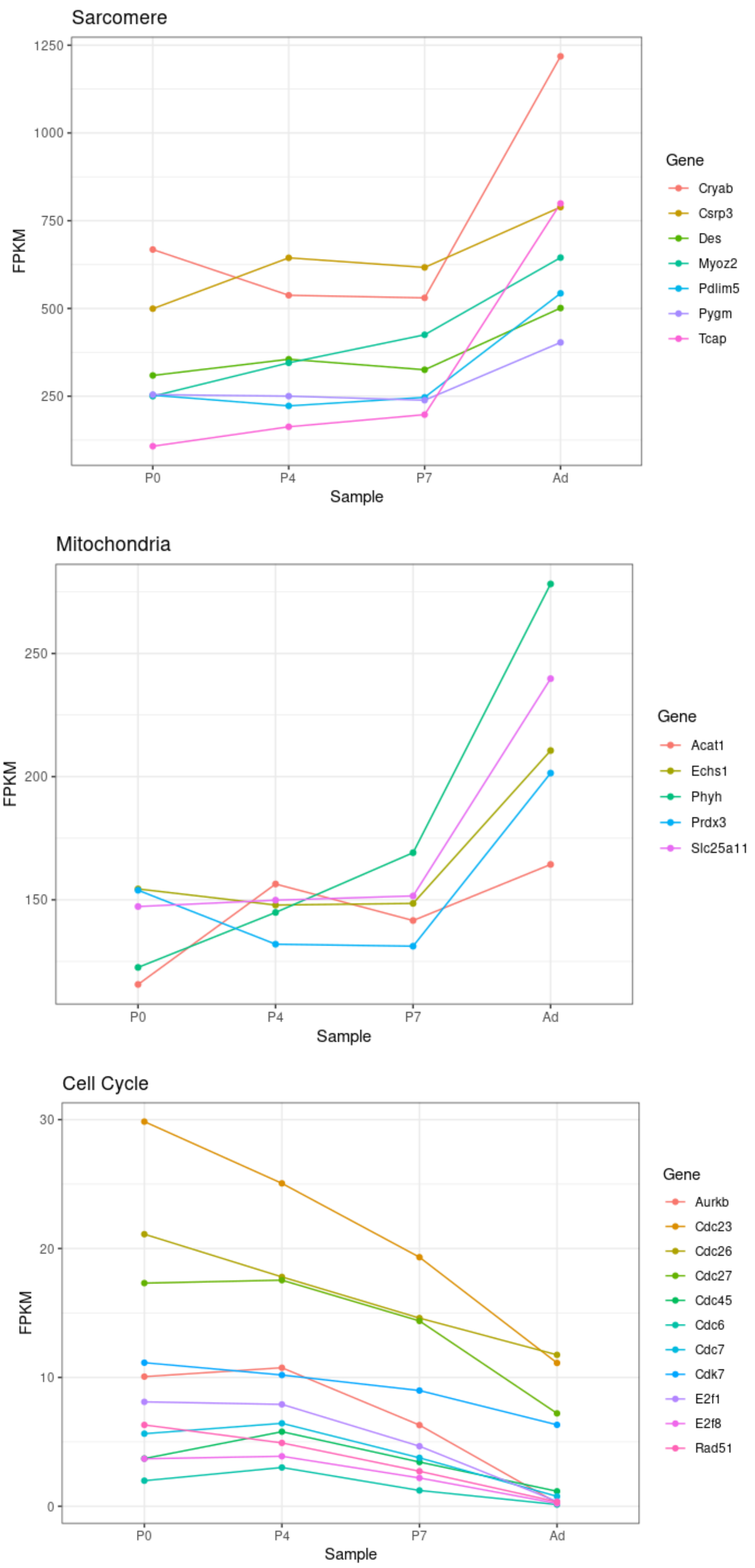

Figure 4. FPKM values of representative Sarcomere, Mitochondrial, and Cell Cycle genes significantly differentially expressed during in vivo maturation

Cluster	Enrichment Term	Enrichment score	p-value	Count
1	Ion binding*	43.82	1.6E-66	354
2	Mitochondrion*	32.59	4.0E-62	190
3	Organic acid metabolic process	29.58	1.1E-58	133
4	Phosphorous metabolic process	27.1	4.5E-62	235
5	Identical protein binding*	25.67	3.8E-41	173

Table 2. Summary of top 5 gene clusters for up-regulated genes using DAVID Functional Annotation Clustering tool

Cluster	Enrichment Term	Enrichment score	p-value	Count
1	Ion binding*	65.69	1.2E-78	367
2	Regulation of macromolecule biosynthetic process	56.12	3.0E-96	321
3	Regulation of cellular component organization	30.09	4.3E-60	216
4	Chromosome*	28.2	1.5E-33	111
5	Chromosome organization*	25.49	4.8E-49	129

Table 3. Summary of top 5 gene clusters for down-regulated genes using DAVID Functional Annotation Clustering tool

The top cluster results obtained from the DAVID Functional Annotation Clustering Groups gene sets analysis are summarized in Tables 2 and 3. These tables include annotations for biological pathways marked with an asterisk that show overlap with the data presented in the publication. The bulk of the top gene enrichment phrases found in this study are comparable to those reported by O'Meara et al. From the top 5 enriched GO terms from the upregulated genes, mitochondrion, ion binding and identical protein binding show overlap. Ion binding, chromosome and chromosome organization from the top 5 gene clusters for down-regulated genes in Table 3 appear to overlap with the O'Meara, et al.'s dataset.

Figure 5 displays a clustered heatmap of the top 1000 differentially expressed genes based on gene expression of log fold change over in vivo maturation in postnatal vs adult phase. Negative

FPKM values in the blue areas indicate downregulation or low gene expression, whereas positive FPKM values in the orange sections suggest high gene expression or upregulation. The heatmap demonstrates that the P0 and Ad samples have significantly different gene expression patterns. The generated heatmap differs greatly from the one published in the original study.

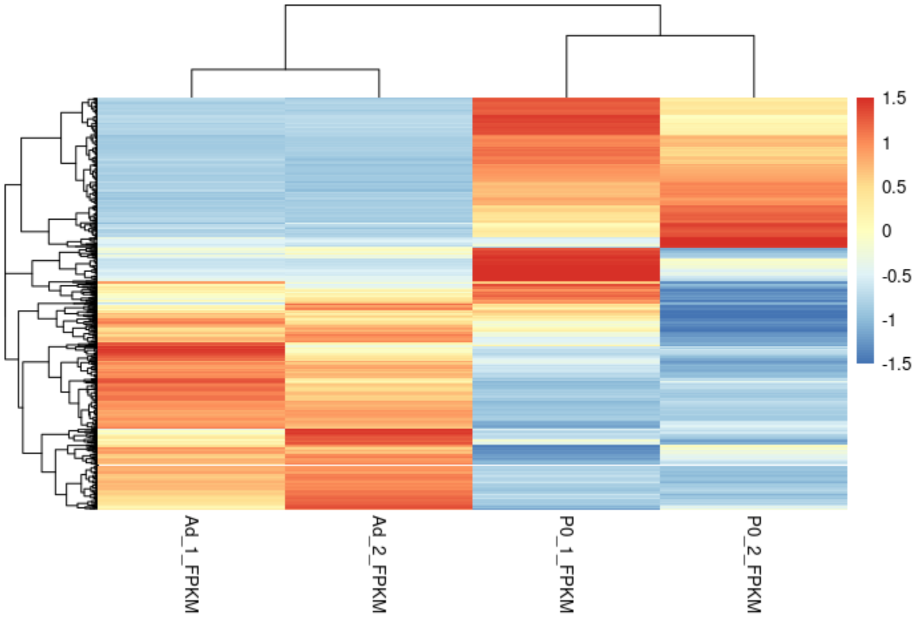

Figure 5. Clustered heatmap of FPKM values using the top 1000 DE genes from the P0 vs Ad analysis

DISCUSSION

RSeQC geneBody coverage results (Figure 1) reveals that a higher percentage of reads are found around the 3' end. Polyadenylated RNA (RNA with several adenine bases at the 3'end) found in the samples could be one reason for this, as it favors readings towards the sequence's 3'end. Another possibility is that the samples have deteriorated. The low number of negative values in the insert size graph (Figure 2) indicates that there is very little overlap between 2 reads and therefore implies the lack of error.

The bulk of the top gene enrichment phrases found in DAVID are comparable to those reported by O'Meara et al., except for 2 up-regulated and down-regulated terms that do not overlap. Our findings corroborate those of O'Meara et al., who also found an overrepresentation of mRNA transcripts associated to catabolism and protein binding that are upregulated, as well as transcripts related to regulation and organization that are downregulated, at time point P0 compared to time point Ad.

Based on Figure 5, the gene expression patterns of the P0 and Ad groups show significant variations. On comparing the two timepoints, it may be determined that genes that are strongly expressed in the postnatal period are downregulated in the adult phase, while genes that are upregulated in the adult phase are expressed less in the postnatal period. The heatmap produced in this study significantly differs from O'Meara et al’s study. A contributing factor to this could be that this project only focuses on *in vivo* maturation and specifically on the P0 and Ad timepoints, unlike the original study which generated the heatmap based on datasets for the *in vitro* differentiation, *in vivo* maturation and adult cardiac myocyte (CM) explants.

CONCLUSION

This study provides a useful foundation for studying adult cardiac regeneration and identifying possible stimulators by understanding transcriptional alterations for cardiac myocyte repair. In conclusion, this project's findings and inferences are similar to those of O'Meara et al to a large extent. The observed discrepancies are most likely attributable to tool selection, tool version differences, and analytic parameters.

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