Transcriptional Profile of Mammalian Cardiac Regeneration with mRNA-Seq

Data Curator: Yichi Zhang
Programmer: Varun Raghuraman
Analyst: Neha Gupta
Biologist: Rachel Thomas

INTRODUCTION:

Adult mammalian hearts have a limited capacity for self rehabilitation. 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 [2]. 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:

- **Data Description**
  RNAseq data for the differentiation of mouse embryonic stem cells into cardiac myocytes through mesodermal and cardiac progenitor intermediates was

obtained from Wamstad et. al[3].CD1(CD-1) neonatal mice (Mus musculus, Charles Rivers Laboratories, MA) were sacrificed by decapitation at P0, P4, and P7, and by isoflurane overdose at 8–10 weeks of age. Hearts were excised, washed in ice cold PBS, and snap frozen in liquid nitrogen. Heart atria were dissected and discarded, and ventricles were processed for RNAseq. At least two heart ventricles were pooled for each replicate. Two replicates were processed for RNAseq[1].

The sequencing data used in this report comes from neonatal cardiac myocytes that were dissociated from whole mouse hearts (P0), and also from sham and resected neonatal mouse hearts at 7 days post surgery using the *Neonatal Heart Dissociation Kit* (Miltenyi Biotec). Cardiac myocytes were purified from the dissociated cell population using the *Neonatal Cardiac Myocyte Isolation Kit* (Miltenyi Biotec) according to the manufacturer's instructions. Three biological replicates were generated per time point and experimental treatment for RNAseq. Hearts from five to ten mouse pups were pooled for each biological replica[1].

Total RNA was extracted from the P0 sample. RNA was isolated using *Trizol* according to manufacturer's instructions, including optional steps in protocol. RNA quality was determined by *Agilent Bioanalyzer*. RNA-seq libraries were prepared using an *Illumina TruSeq* kit. A final round of size selection by *Agencourt AMPure XP* beads was performed to remove small fragments such as primers. Sequencing was run on either an *Illumina Hi-Seq 2000* (barcoded)[1].

The sequencing data was uploaded to the public database Gene Expression Omnibus (GEO) with accession number GSM1570702[4].

- **Data Quality Control**
  The SRA file downloaded from the GEO database was processed via *fastq-dump*[6] into separated sequence files in fastq format. Those reads were analysed through *FastQC* to inspect qualities[5]. The size of the paired-end library is 21577562 and length of each read is 40 with a %GC of 49. The overall quality of the sequencing data is good.

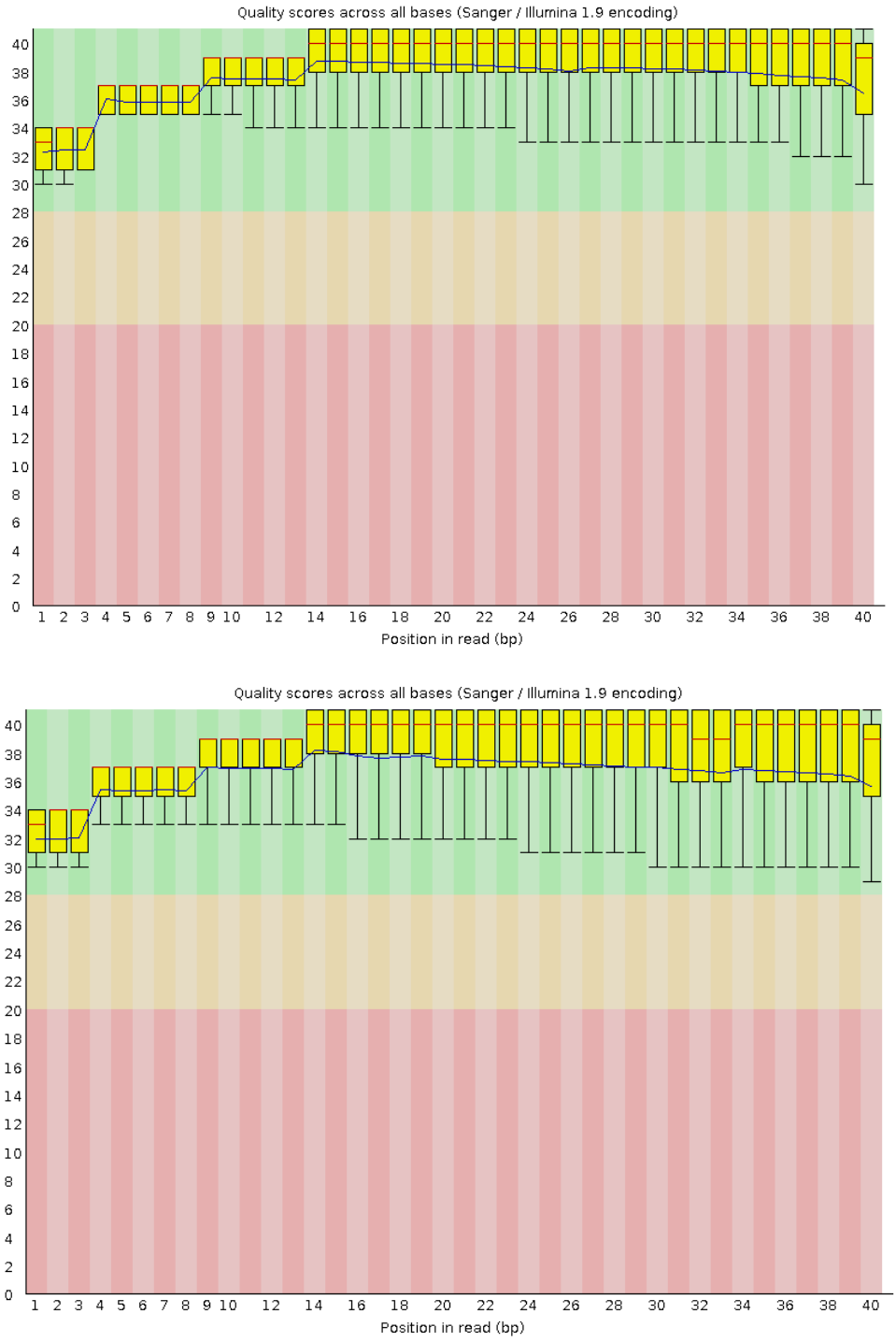


Figure 1: Per base sequence quality plots of paired-end reads

Fig 1 provides the distribution of quality scores across all bases at each position in the reads. The central red line was the median value,the yellow box represented the interquartile range, the upper and lower whiskers represented the 10% and 90% points and the blue line represented the mean quality. For both

plots, boxes were stable, and scores and reads across all bases lied in the green region indicating very good quality calls. This judgement was supported by Fig 2, the per sequence quality score report which a significant portion of the sequences in a run had an overall high quality to deliver a peak on the right.

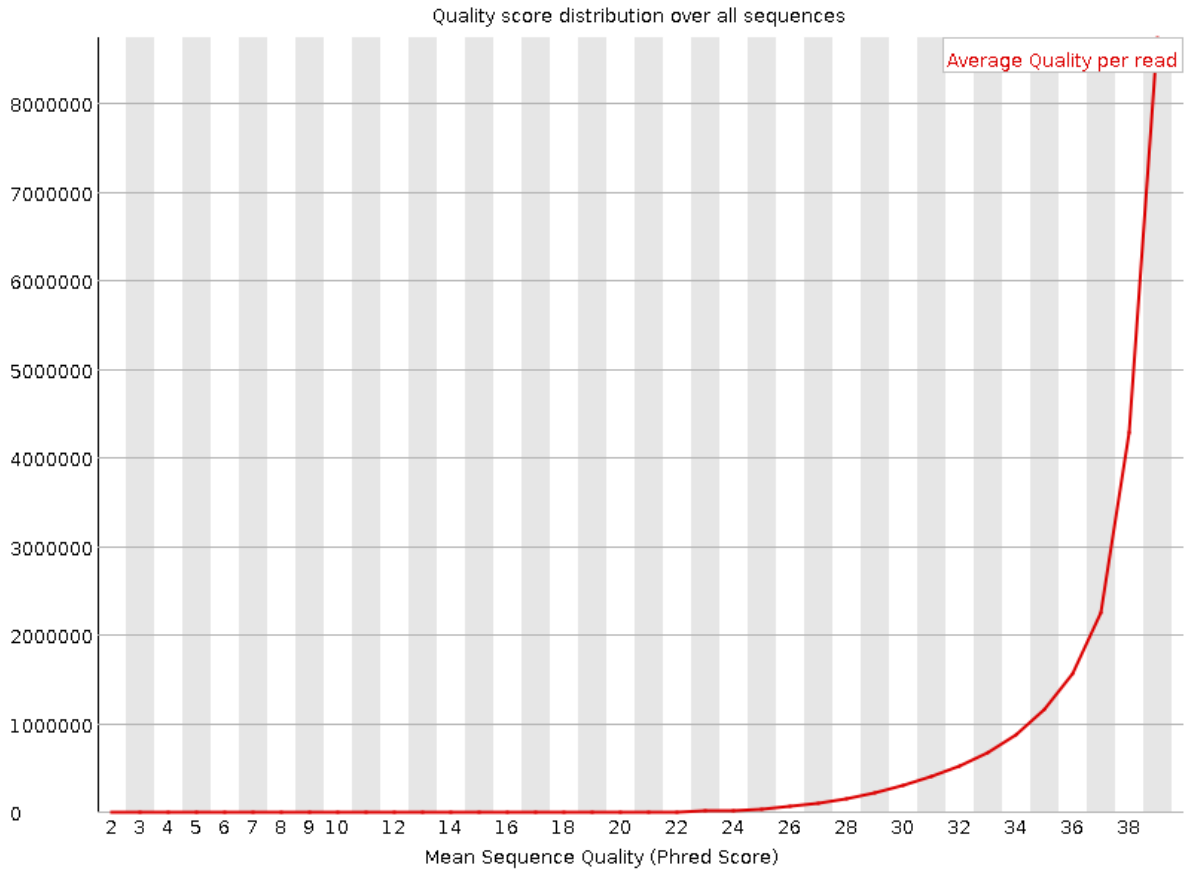


Figure 2: Per sequence quality scores of paired-end read 1.

The model failed on the per base sequence content shown in Fig 3. Per base sequence content plotted out the proportion of each base position in a file for which each of the four normal DNA bases has been called[5]. In a random library one would expect that there would be little to no difference between the different bases of a sequence run, so the lines in this plot should run parallel with each other. The relative amount of each base should reflect the overall amount of these bases in a genome, but in any case they should not be hugely imbalanced from each other.

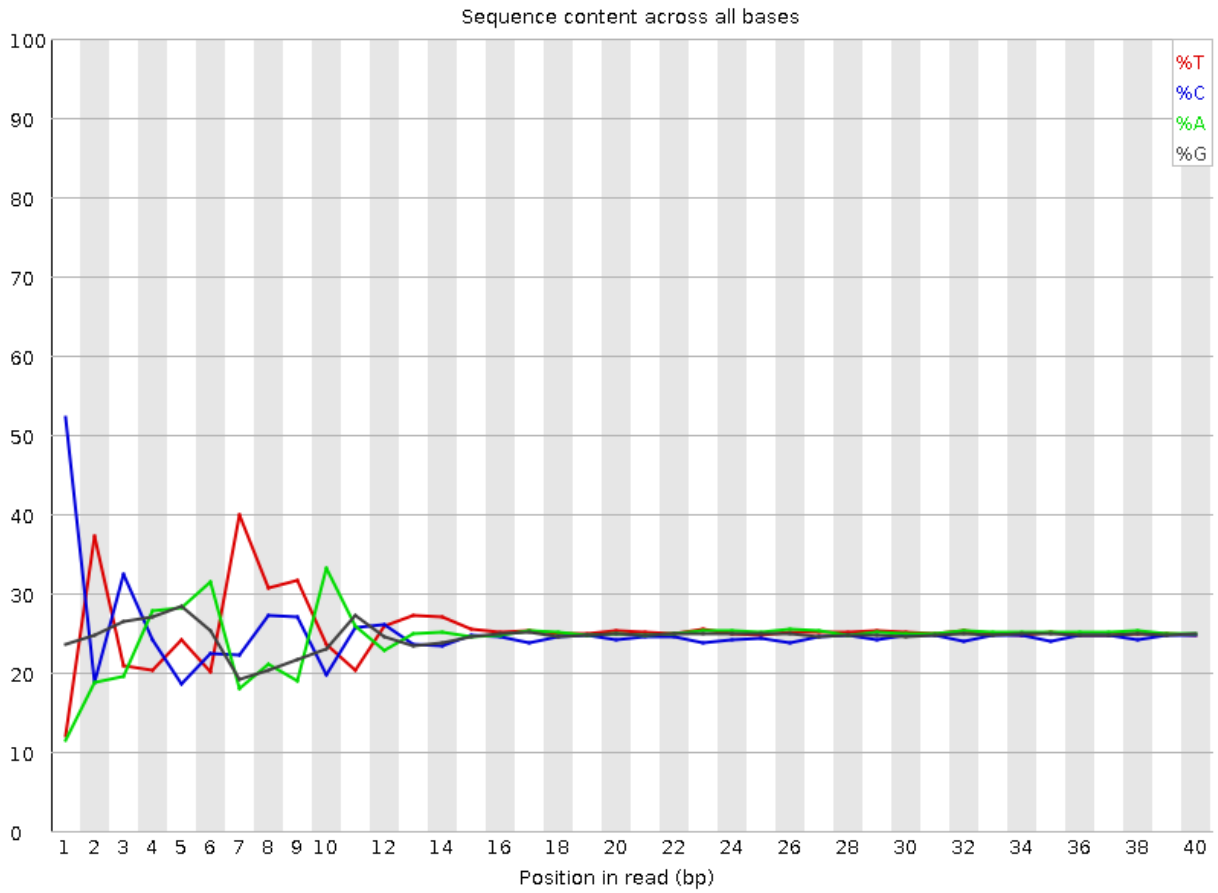


Figure 3: Per base sequence content of the paired-end read 1.

The model failed again when the difference between A and T, or G and C was greater than 20% in any position (notice first 10 positions). It is worth noting that some types of library will always produce biased sequence composition, normally at the start of the read. Libraries produced by priming using random hexamers (including nearly all RNA-Seq libraries) and those which were fragmented using transposases inherit an intrinsic bias in the positions at which reads start. This kind of bias did not concern an absolute sequence, but instead provided enrichment of a number of different K-mers at the 5' end of the reads. This technical bias cannot be corrected by trimming and in most cases doesn't seem to adversely affect the downstream analysis[5].

There was no overrepresented sequence detected in this library, adapter content was well recognized, and the per-sequence GC content showed a normal-like Gaussian distribution (Fig 4) which was expected where the central peak

corresponds to the overall GC content of the underlying genome.

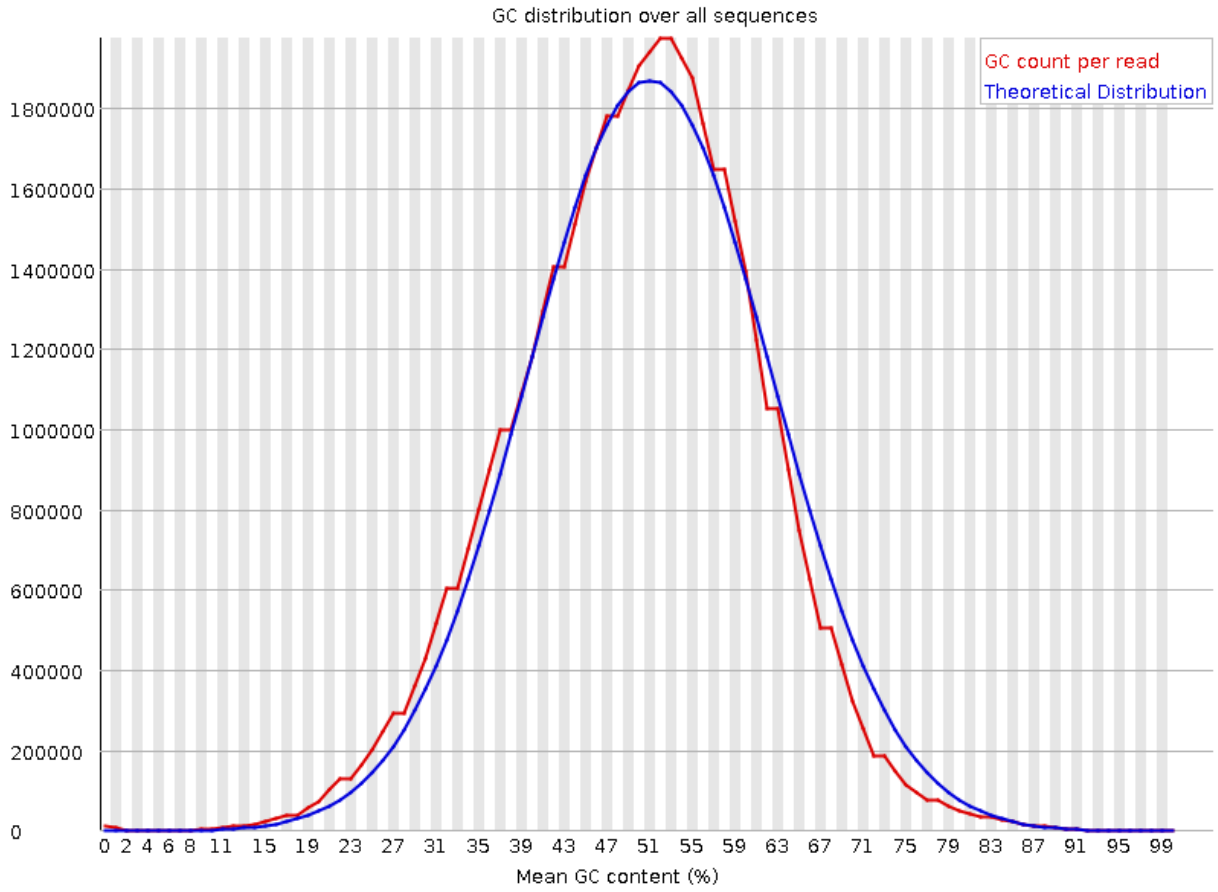


Figure 4: Per sequence GC content of the paired-end read 1.

The *FastQC* report displayed a warning on sequence duplication levels (Fig 5) as well, which indicated that non-unique sequences had made up more than 20% of the total. Fig 5 shows the proportion of the library which was made up of sequences in each of the different duplication level bins. The blue line took the full sequence set and shows how its duplication levels were distributed. In the red plot the sequences were de-duplicated and the proportions shown were the proportions of the deduplicated set which come from different duplication levels in the original data.

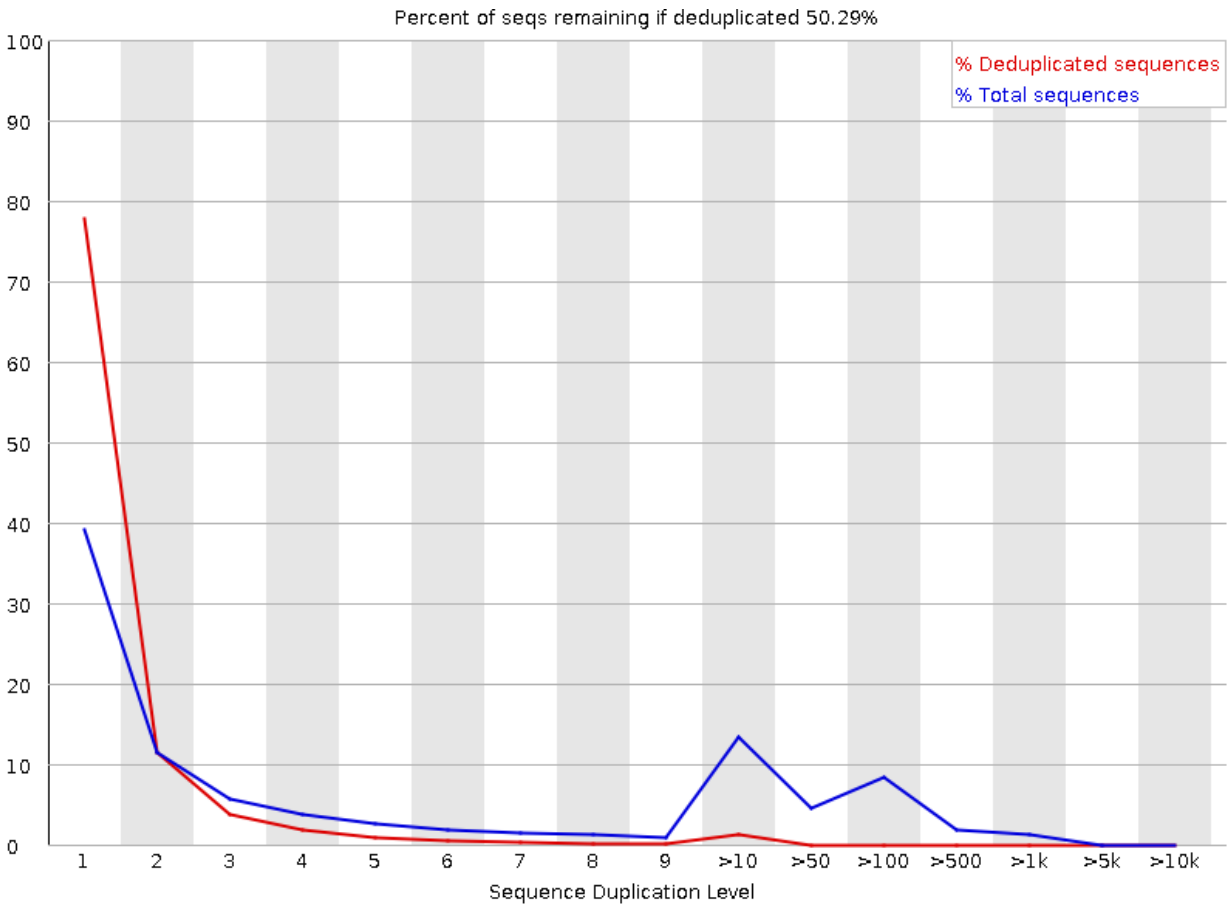


Figure 5: Sequence duplication levels of the paired-end read 1.

In a diverse library most sequences are expected only once. A low level of duplication may indicate a very high level of coverage of the target sequence, but a high level of duplication is more likely to indicate some kind of enrichment bias as we mentioned earlier for the technical bias[5]. In general there are two potential types of duplicate in a library, technical duplicates arising from PCR artefacts, or biological duplicates which are natural collisions where different copies of exactly the same sequence are randomly selected. The high, single duplication peak in blue disappears in the red line (Fig 5) suggesting the diversity has been partially or completely exhausted and a waste of sequencing capacity, instead of large numbers of different highly duplicated sequences which might indicate either a contaminant set or a very severe technical duplication.

METHODS:

In order to further pursue the study objectives, it is necessary to realign RNAseq data and evaluate it both qualitatively and quantitatively for sources of error or bias before further analysis. Multiple reads were aligned to the reference genome using *TopHat* with *Bowtie2* indexes.The *SAMTools* 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.

The *RSeQC* package was further utilized to evaluate the RNA-seq data. The *SAMtools* sort and index functions were utilized to organize the sequence by position, and feed them into the *RSeQC* package. The Gene Body Coverage graph (Figure 6) mapped between nucleotide position and number of reads to locate bias in read values. A clear 3' end bias was noted in the data, signifying 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[1]. Another source of this skew could be degradation of samples, and while that cannot be ruled out, the definite presence of bias from sample collection and the relatively "gentle" slope in terms of coverage seemed to rule imply that the data collection bias was the greater factor in production of these results.

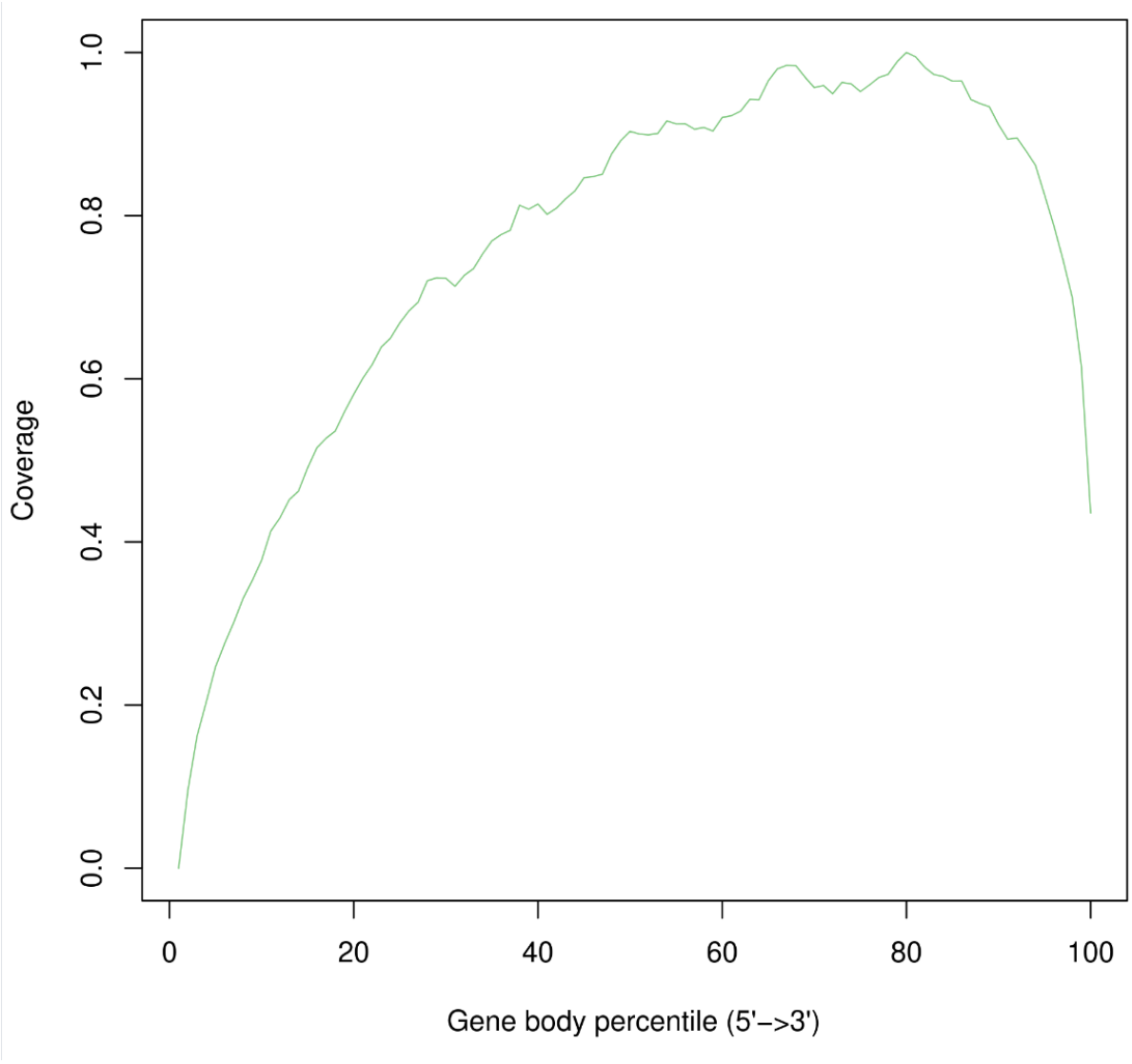


Figure 6: Coverage Graph of RNA reads with 3' bias

The next quality control measurement was the inner distance, or insert size tool from *RSeQC*. The Insert Size Graph (Figure 7) helped evaluate the distance between two paired reads. The distribution of distances was centered around 60 base pairs, with a tail that pulls the distance further to the right, away from zero, implying a level of regularity for distance and a lack of additional factors complicating the alignments. The most useful feature for indicating lack of error was the number of negative values (on the left side of the distribution); their relatively low number indicates that there was not much overlap between the two reads. The final quality control metric from *RSeQC* mirrored the *SAMtools* flagstat tool in generating statistics for the read, except using .bam files as input (Figure 8). The most notable additional characteristics were the "unique" categories and the re-calculated read counts. The unique count is 38489380, or 77.43% of the total reads, while the initial Read 1 and Read 2 were 20878784 and 20510550 respectively, in comparison to the new counts of 19409941 and 19079439. The largest change was in the Read 1 counts, which have decreased by 2.879% of the original read counts--which

is fairly reasonable considering the overall size of the datasets. On a qualitative level, it seemed as if the aligned sequence data was lacking in enough errors and unexpected bias to be processed further .

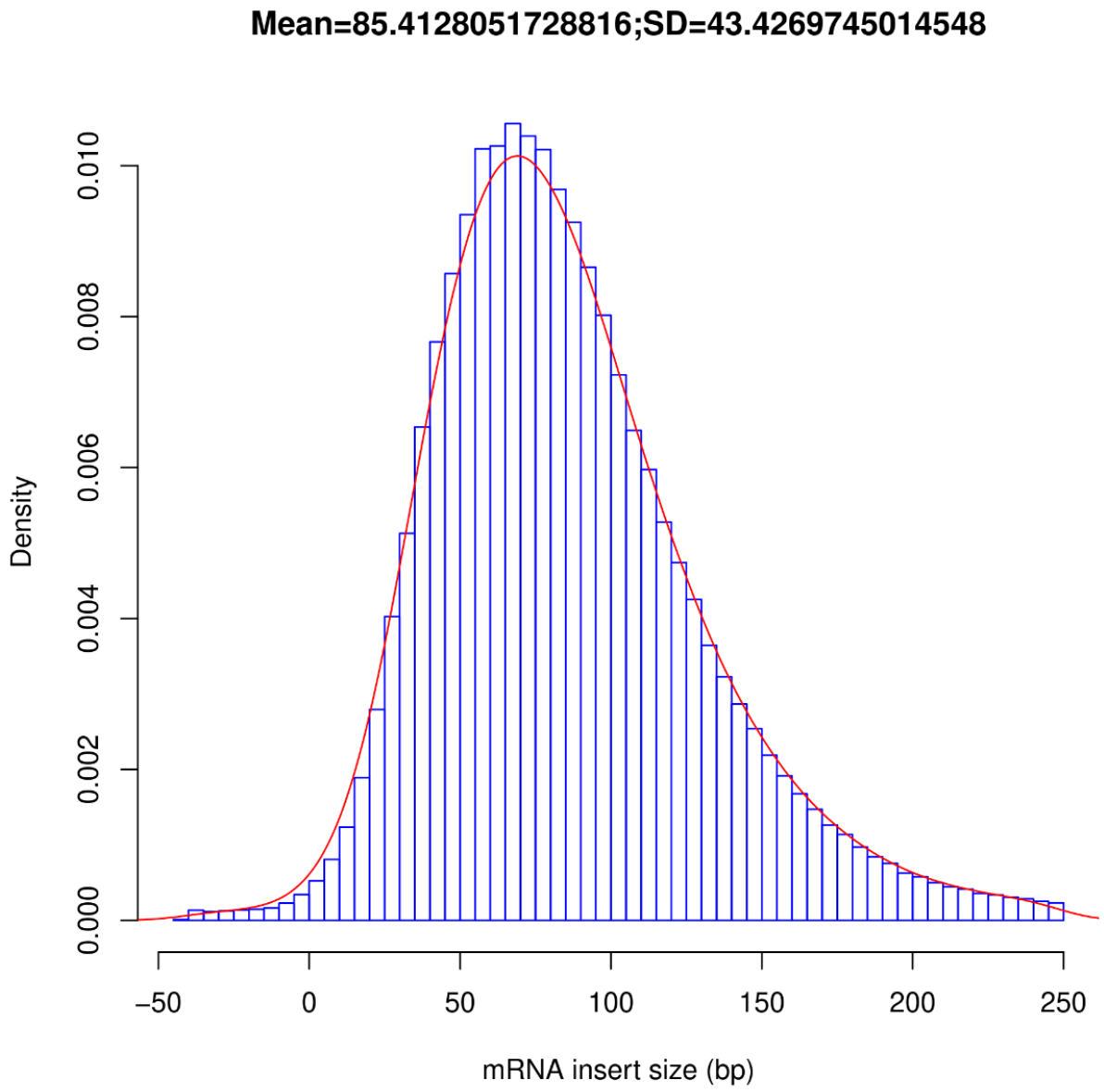


Figure 7: Insert Size between Two Paired Reads

Samtools Flags	Values
Total records:	49706999
QC failed:	0
Optical/PCR duplicate:	0

Non primary hits:	8317665
Unmapped reads:	0
mapq < mapq_cut (non-unique):	2899954
mapq >= mapq_cut (unique):	38489380
Read-1:	19409941
Read-2:	19079439
Reads map to '+':	19236824
Reads map to '-':	19252556
Non-splice reads:	33099839
Splice reads:	5389541
Samtools Flags	Values
Total records:	49706999

Table 1: Output of bam_stat.py Tool from *RSeqc*

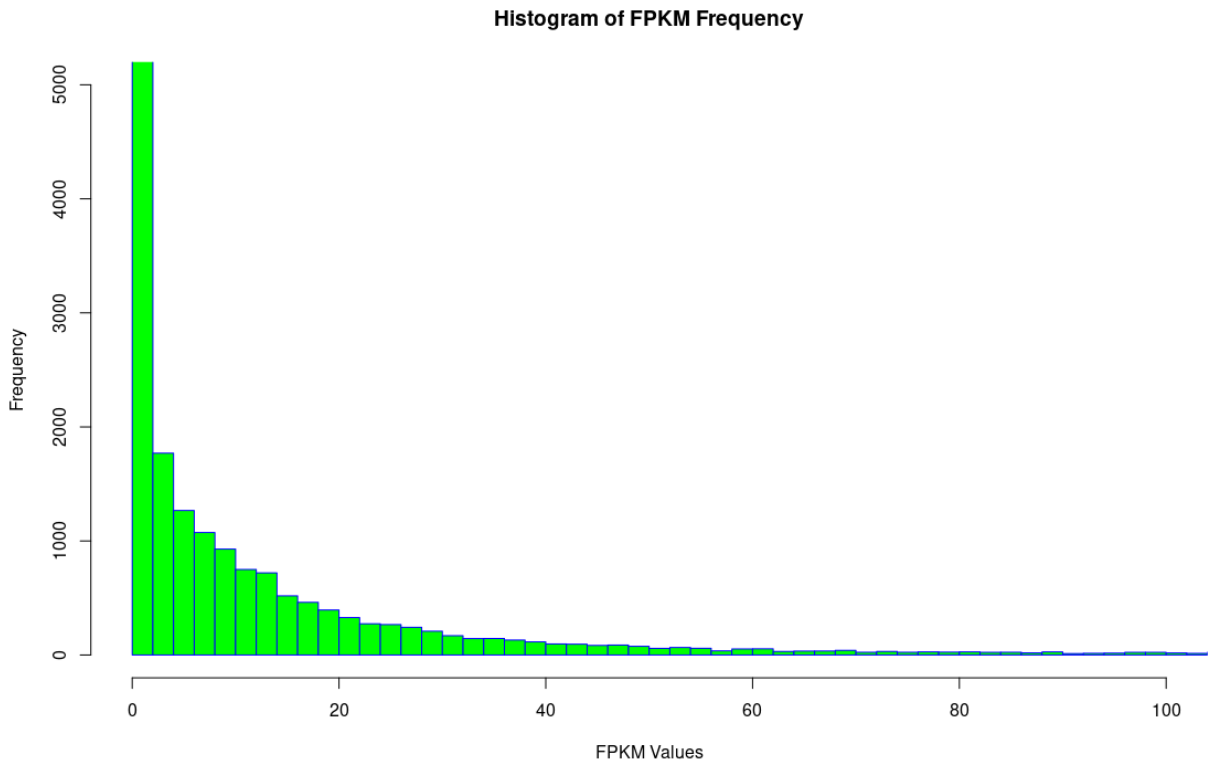


Figure 9: A Histogram of FPKM Values Produced by *CUFFLINKS.* Frequency was limited to 5000 for better visualization.

The figure was limited in both X and Y axis, and was subjected to the maximum number of breaks to show extreme clustering of values between the range of 0 and 1.

Using *Cufflinks*, a file for differentially expressed genes were produced with columns as gene id, name, log2 fold change values, p-value, q-values and significance for P0 vs. Adult. The file was imported into *Rstudio* and arranged in ascending order based on q values. A subset was created for the top 10 among those sorted values and reported in table 2 Furthermore, histograms were plotted for analysis and compared for all genes vs. significant genes on their log 2 fold change values.

The significant up-(positive values) and down-regulated(negative values) genes were stored in separate files using the log2 fold change column. These files were used as inputs for DAVID(Database for Annotation, Visualization and Integrated Discovery (DAVID) version 6.8 for identifying enriched biological terms, particularly GO terms. This resulted in gene enriched data sets organized into clusters with their enrichment scores. We customized for *Mus Musculus* species with GOTERM_BP_FAT, GOTERM_MF_FAT, and GOTERM_CC_FAT groups.

RESULTS:

- **Gene Expression Analysis**

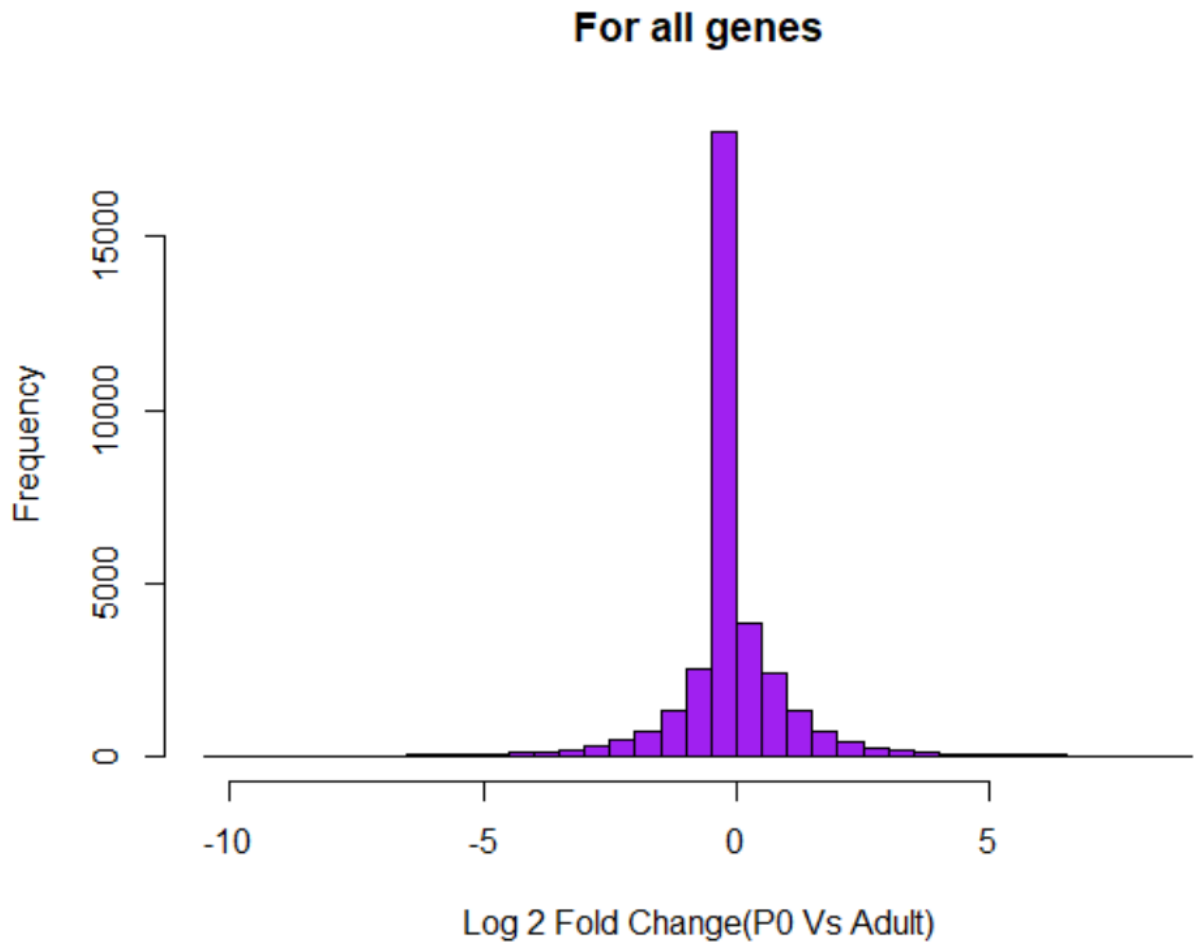
Table 1 summarizes the top 10 up-regulated genes during the differential expression analysis, in addition to p-values and q-values for confidence. They were identified by the log2 fold chain in comparison of change of expression level between two samples. Positive log2 fold change values indicated up-regulated genes and vice versa. 36329 genes in total were observed to be differentially expressed, out of which 5193 were significantly expressed genes: 2760 up-regulated and 2733 down-regulated, with a threshold of p-value < 0.01 applied.

gene	value1	value2	Log2 fold change	P-value	Q-value
Rb1cc1	12.193700	31.94050	1.389250	5e-05	0.000318974
Pcmtd1	13.365200	30.17000	1.174640	5e-05	0.000318974
Adhfe1	13.548000	27.03530	0.996765	5e-05	0.000318974
Tmem70	36.591300	85.04140	1.216660	5e-05	0.000318974
Gsta3	0.414547	7.11348	4.100950	5e-05	0.000318974

Lmbrd1	6.701000	13.31730	0.990848	5e-05	0.000318974
Dst	18.942300	54.22070	1.517230	5e-05	0.000318974
Plekhb2	26.635000	72.03520	1.435380	5e-05	0.000318974
Mrpl30	55.017900	130.53800	1.246490	5e-05	0.000318974
Tmem182	46.029600	108.74000	1.240250	5e-05	0.000318974

Table 2: Summary of Top 10 Up-regulated Genes

Histograms with fold change between the experimental and control groups (Fig 10 & Fig 11) were plotted. Fold change represents logarithm of fpkm ratios for differentially expressed genes. And log2 ratios ≥ 1.0 or log2 ratios ≤ −1.0 means two Fold change. |Fold change| ≥1 means genes differentially expressed.



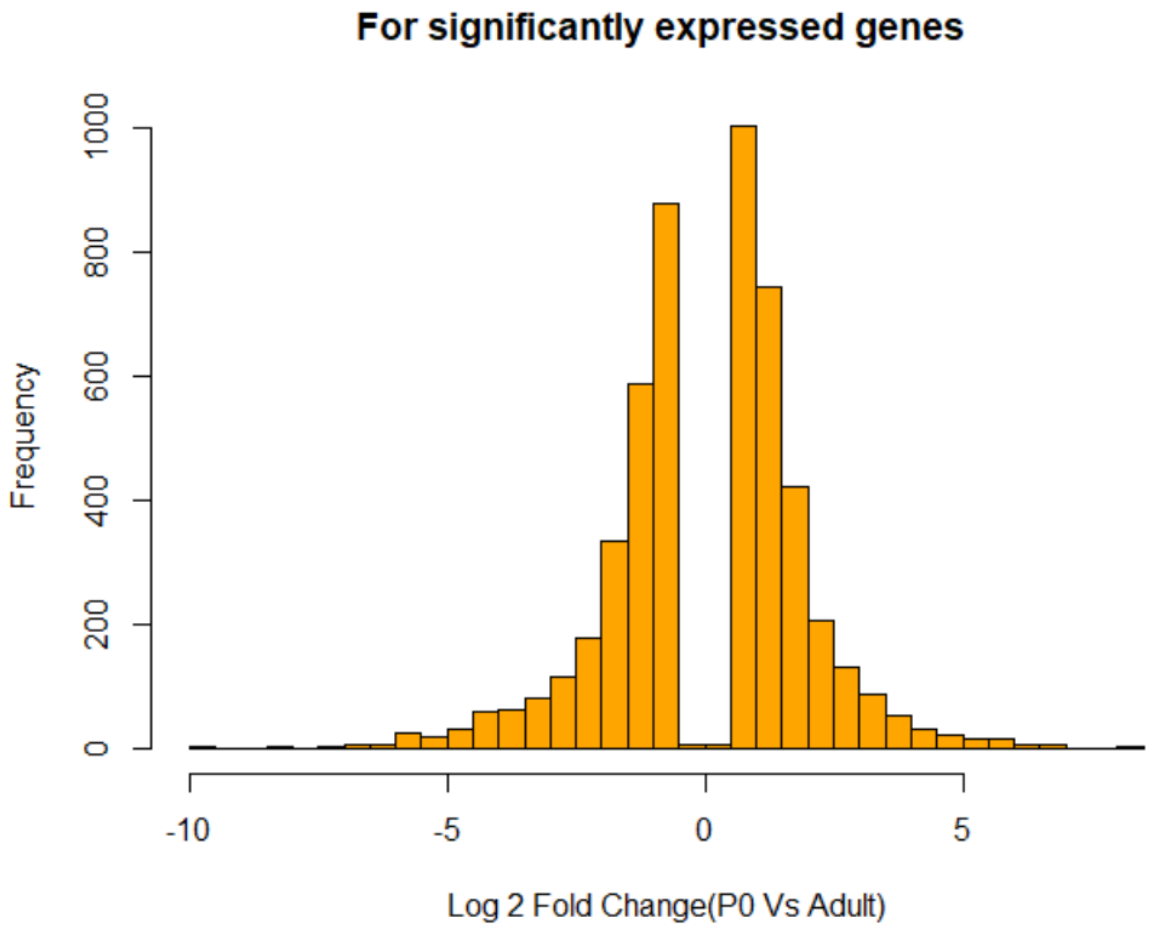


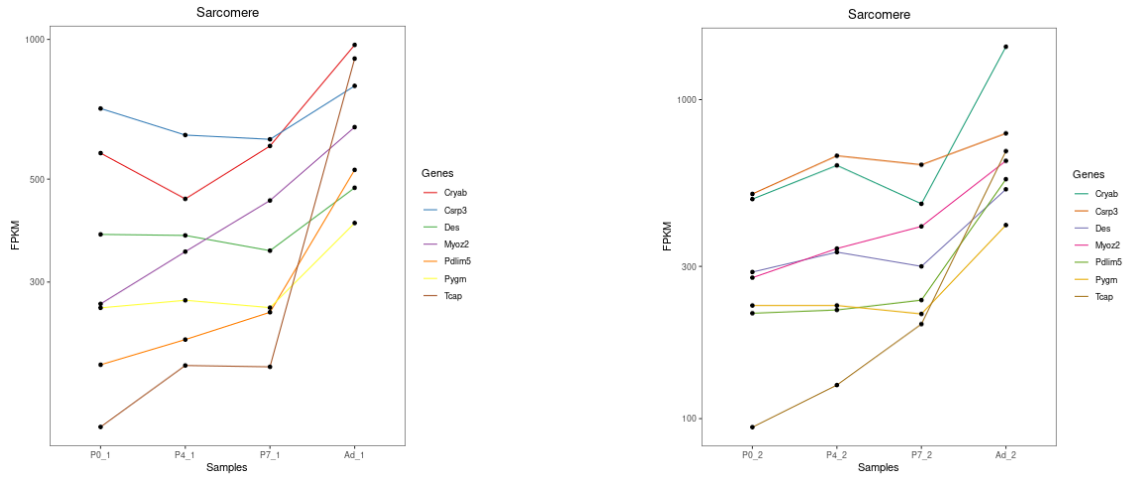
Figure 10&11: Histograms of the Log2 Fold Changes

The number of up and down regulated differentially expressed genes from our study did not align with the reference paper's results. The comparison is shown in the table below.

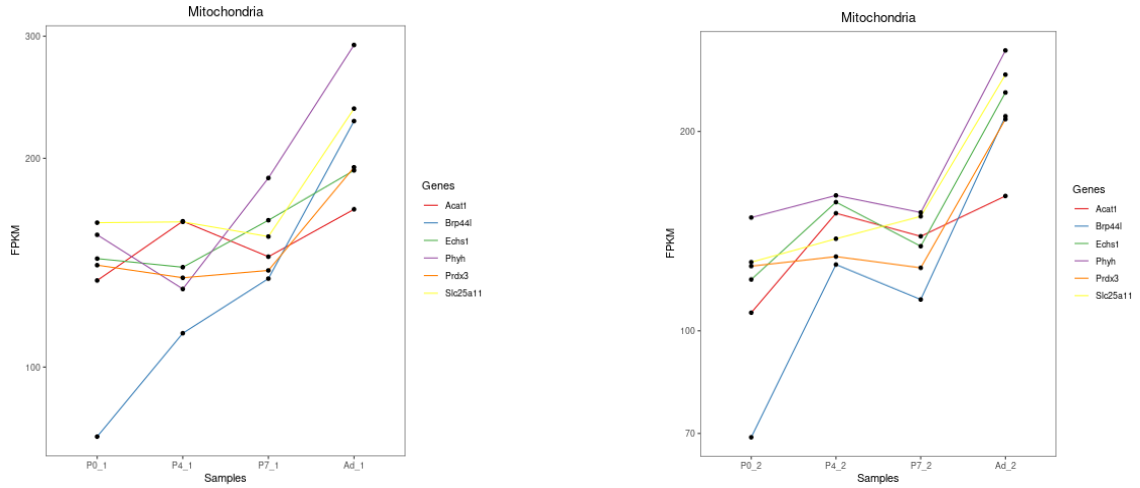
	This Study	Reference Paper
Up-regulated	2760	2409
Down-regulated	2433	7570
Total	5193	9979

Table 3: Comparison of Genes Expressed Significantly Differently Between This Study and the Reference Paper

(A)



(B)



(C)

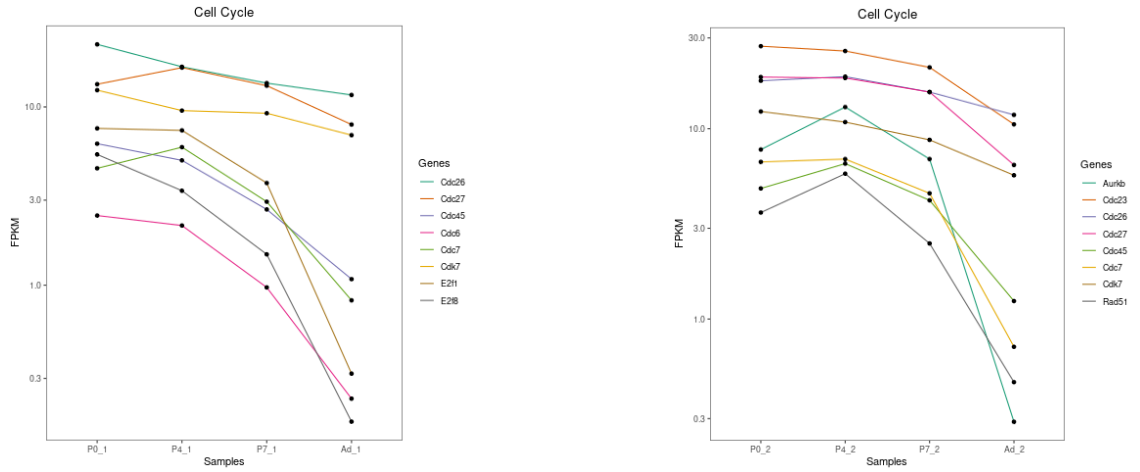


Figure 12: FPKM values of representative Sarcomere, Mitochondria and Cell Cycle genes significantly differentially expressed

Using DE genes file generated from Cufflinks, the FPKM values of representative Sarcomere (A), Mitochondria (B), and Cell Cycle (C) genes significantly differentially expressed in mice on postnatal day 0 (P0), 4 (P4), 7 (P7), and adult (Ad) were plotted against the biological age of the sample. Sample group 1 and 2 are displayed on the left and right respectively.

The results from the figure (A), Sarcomere DE genes, do not match the reference Figure 1D diagram explicitly. However, this plot does show an increase in FPKM values in P7 to Ad suggesting that the representative sarcomere genes were up-regulated from postnatal to adult maturation, which is consistent with the reference study.

The results from figure (B), Mitochondrial DE genes, look similar to those produced in the reference paper. Specifically, the Acat1, Brp44l, and the Prdx3 genes demonstrate a very similar pattern to the mitochondrial DE gene plot found in figure 1D of the reference paper. The Brain Protein 44 Like gene (Brp44l) is an alias for the Mitochondrial Pyruvate Carrier 1(Mpc1) gene that is shown in figure 1D in the reference study.

Only 8 out of 12 genes that were represented in figure 1D of the reference study are shown in figure (C) due to errors that occurred while attempting to plot all 12 representative genes. However, the 8 genes selected show a similar trend across the in-vivo maturation in figure 1D. The cell cycle DE genes demonstrate down-regulation of cell cycle genes during in-vivo maturation which is consistent with the findings in the reference paper.

- **DAVID Results:**

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.

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 4: 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 5: 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 6: The DAVID results from the reference paper.

(*) signifies that the genes were found in both the experimental study and the reference paper

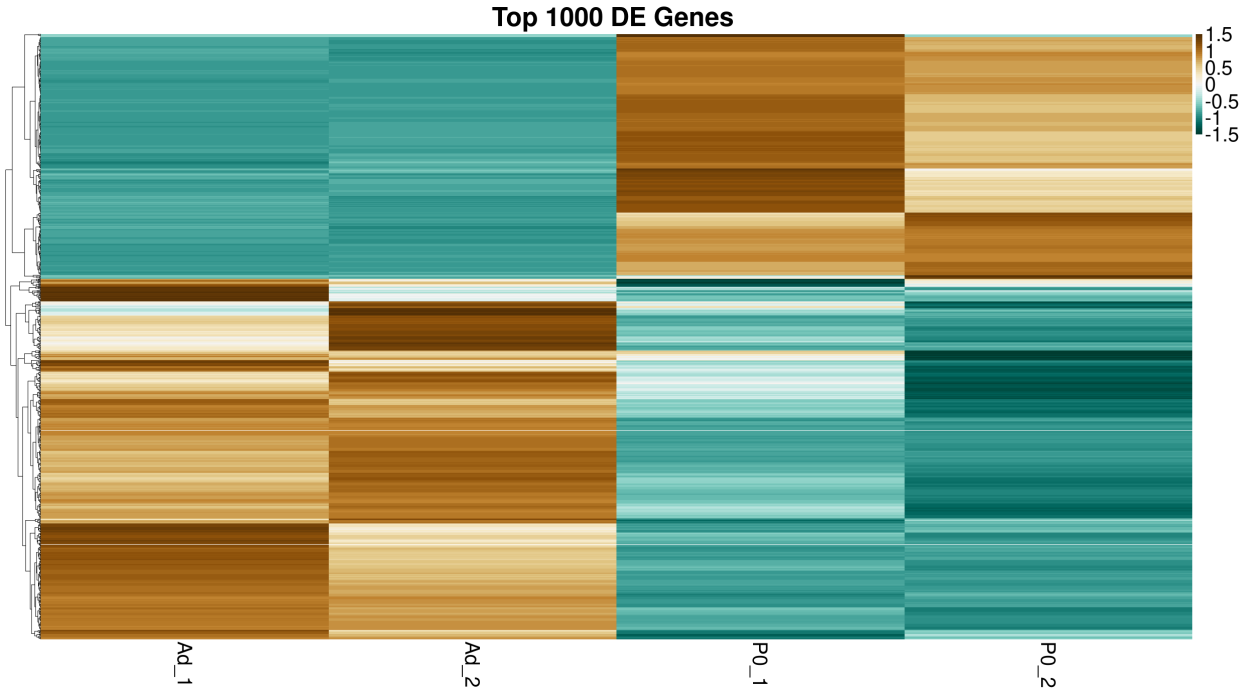


Figure 13: Clustered Heatmap of Top 1,000 Differentially Expressed Genes (Adult vs. P0). The heatmap was generated based on the gene expression of log fold change over the course of in vivo maturation in postnatal to adult maturation. The Tiffany blue represents low expression of genes while the browns represent higher expressed genes. Gene symbols were removed for better visualization. The clusters, in GO terms, were well separated so that active genes on the P0 stage became less expressed in adults, and vice versa, a clear maker of “role change” of genes in heart development.

DISCUSSION:

In the reference paper, researchers have confirmed that FPKM values of representative sarcomere, mitochondrial, and cell cycle genes significantly differentially expressed during in vitro differentiation and in vivo maturation among samples in different stages[1]. Our analysis partially agreed on findings of the reference paper. In vivo maturation samples, our analysis on both replicates captured the distinct expressions of Csrp3, Cryab, Tcap, and Myoz2 in FPKM values of representative sarcomere which agreed with the reference paper(Fig 12A) on both the trend and genes. The analysis achieved the same results on representative mitochondria and cell cycle(Fig 12B&C). In both the sarcomere and mitochondria gene representations, all genes display a vertical shift from P7 to Ad, suggesting that the sarcomere genes were primarily up-regulated during development in the postnatal to adult period. The sharp vertical increase from P7 to Ad also indicates that these genes were primarily up-regulated during development in the postnatal to adult stage. Fig 12C FPKM values displayed a decreasing trend in the

postnatal to adult maturation stage. This suggests that these cell cycle genes were down-regulated during the in-vivo maturation stage. Genes critical for sarcomere assembly, such as Titin Cap and Cardiac troponin I type 3, showed pronounced increases in expression over the course of differentiation reflecting sarcomere assembly and organization during cardiac myocyte differentiation and maturation[1]. However numerically the expression levels of genes in plots did not align with those of the reference paper, a possible reason is researchers at that time averaged expression levels of all replicates to increase the robustness of the study. As long as the topology of plots in both studies complies, numeric fluctuations would not be a concern.

The difference between regulated genes identified in two studies has been summarized by Table 3. The total number of up-regulated genes obtained from this analysis was 2760 which produced roughly 350 more up-regulated genes (2409) than those referenced in the *O'meara et al* study. In addition, the number of down-regulated genes obtained were 2433, which was vastly different from the 7570 down-regulated genes cited in the reference study. The possible explanation for this dramatic difference is researchers identified up- and down-regulated genes directly from the GO analysis. Tools like GSEA and StringDB support converting ENSEMBL IDs into gene symbols, clusters were generated first then genes were counted, where this analysis went through log2 fold change to select most/least expressed genes. The FPKM values on their own were not necessarily indicative of much without being coupled to other evaluations, so the genes were set to the analysis step input in larger quantities as well. Overall, it was easy to locate and verify results reproduced from the paper, especially for instructions of data/sample preparations. Details on the data source has significantly increased the reliability of results.

The reference paper used data from experiments done in vivo, in vitro, and in explant. In vitro experiments were performed outside of a living organism and therefore an in vitro study occurs in a controlled environment. Data collected from in vitro experiments might function as negative controls and background information. On the other hand, the reference paper used in vivo samples to observe changes in gene expressions when samples matured in natural conditions. Furthermore, explant data were used to observe the regeneration function of the heart during early stages. The comparative analysis of in vivo data and explant data allowed researchers to observe a transcriptional reversion of cardiac myocyte differentiation processes and thus drew conclusions on candidate regulators of cardiac myocyte cell state. This analysis only involves replicates of in vivo experiments, therefore it is expected that results showed a good agreement with the reference paper.

The results from the up-regulated DAVID test showed overlapping in 4 out of 5 top genes, which are mitochondria, sarcomere, sarcoplasm, and respiration/ metabolism genes. The results from the down-regulated DAVID test had overlapping in 3 out of 5 genes, which consisted of RNA processing, DNA repair, and the cell cycle. Three out of the ten top up-regulated and down-regulated genes were consistent to those reported in the reference study. Although there was overlapping of genes present in both studies, the enrichment scores and fold enrichment values did not align well.This difference in results might have been because of different versions of the DAVID used, different gene sets used for annotation or due to difference in vivo maturation condition.

Despite generating a heatmap of expression level on P0 vs. Ad in the current study, not too much information was acquired when comparing to the reference paper heatmap. The heatmap produced by the reference paper had included datasets of all experiments and therefore possessed a larger scale of normalization. In other words, colors of cells between two heatmaps were not informative due to different scales. Due to the same reason, the clear marker of margin shown in Fig 13 disappeared in the reference heatmap. What was informative is the cluster of genes well reflected the changes of expression levels between stages, which indirectly supported the claim of the reference paper that groups of genes with common functions were identified in ex vivo cultured cardiac myocytes via hierarchical clustering[1].

CONCLUSION:

Our study supports findings addressed by the reference paper, Transcriptional Profile of Mammalian Cardiac Regeneration with mRNA-Seq, that in vivo maturation models showed a core transcriptional signature[1] and differentially expressed genes in vivo differentiation did not strongly reflect hallmarks of myocyte differentiation.

Major challenges encountered lied in tracking and processing of a large amount of data, More importantly, the biological interpretation and validation compared to the reference paper. Due to the limitation of the sequencing data, this study was unable to verify all results of the reference paper, especially results including medical images. Due to the limitation of expertise, this analysis could not validate the reliability sequencing data generated from mice prepared for the reference paper researchers. This can be overcome by introducing diverse members for better intercurricular collaboration.

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 cardiomyocytes.” *Nature* vol. 493,7432 (2013): 433-6. doi:10.1038/nature11682
3. Wamstad JA, Alexander JM, Truty RM, et al. Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage. Cell. 2012;151(1):206-220. doi:10.1016/j.cell.2012.07.035
4. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A. NCBI GEO: archive for functional genomics datasets--update Nucleic Acids Res. 2013 Jan;41(Database issue):D991-5.
5. FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2015)
6. SRA Toolkit Development Team. “SRA-Tools.” NCBI, ncbi.github.io/sra-tools/.