Project 2 BF528

Transcriptional Profile of Mammalian Cardiac Regeneration with mRNA-Seq

Kyrah Kotary, Brad Fortunato, Vishwa Talati, Marina Natividad

Introduction

Within their first week of life, neonatal mice have the ability to regenerate their cardiac tissue in the event of an injury. This phenomenon was investigated by O'meara *et al.* (2015) in the study Transcriptional Reversion of Cardiac Myocyte Fate During Mammalian Cardiac Regeneration. The researchers analyzed RNAseq data for adult and early postnatal mice, to investigate through transcriptional profiling whether mouse myocytes revert to a less differentiated state upon injury. Our objective was to use similar analytical tools and methods to reproduce results from the study.

Data

For this project, analysis was conducted on eight samples from the O'Meara *et al.* (2015) study: P0_1, P0_2, P4_1, P4_2, P7_1, P7_2, Ad_1, and Ad_2. All but one sample was provided and processed. The remaining sample P0_1 (accession number GSM1570702) was manually downloaded from GEO Series GSE64403 from the O'Meara *et al.* (2015) repository on the NCBI Gene Expression Omnibus. In the study by O'Meara *et al.* the samples were prepared using a protocol that involved embryonic stem cell differentiation, whole heart ventricle isolation, adult cardiomyocyte isolation, neonatal mouse apical resection, and neonatal cardiomyocyte purification. Complete descriptions of each of these steps can be found in the paper.

FASTQ files from both read directions were analyzed and inspected for various quality control measures. Aside from some slight variations, the quality control metrics from both reads were in agreement with each other, and no sequences were flagged as poor quality in either file.

As shown in Figure 1, the per base sequence quality is high across the entire read length, and highest from position 14 onward. The consistently high scores indicate high quality base calls, and that there are no concerns about sample degradation. This information is confirmed by the per sequence quality score plot in Figure 2, which shows a very high average quality score for all reads, with no indication of lower quality subsets causing concern. The only FastQC metric to fail was the per base sequence content. However, this is expected for RNA-seq data, and is not cause for concern (Figure 3). The GC content across the whole length of each sequence is compared to a modelled normal GC content distribution in Figure 4. As shown by these plots, our results have the expected normal GC content distribution.

The FastQC report included a warning for our sequence duplication levels (Figure 5), which indicates that non-unique sequences make up more than 20% of the total. A high level of duplication could indicate some kind of enrichment bias, such as PCR over amplification. There were no overrepresented sequences according to the FastQC report.

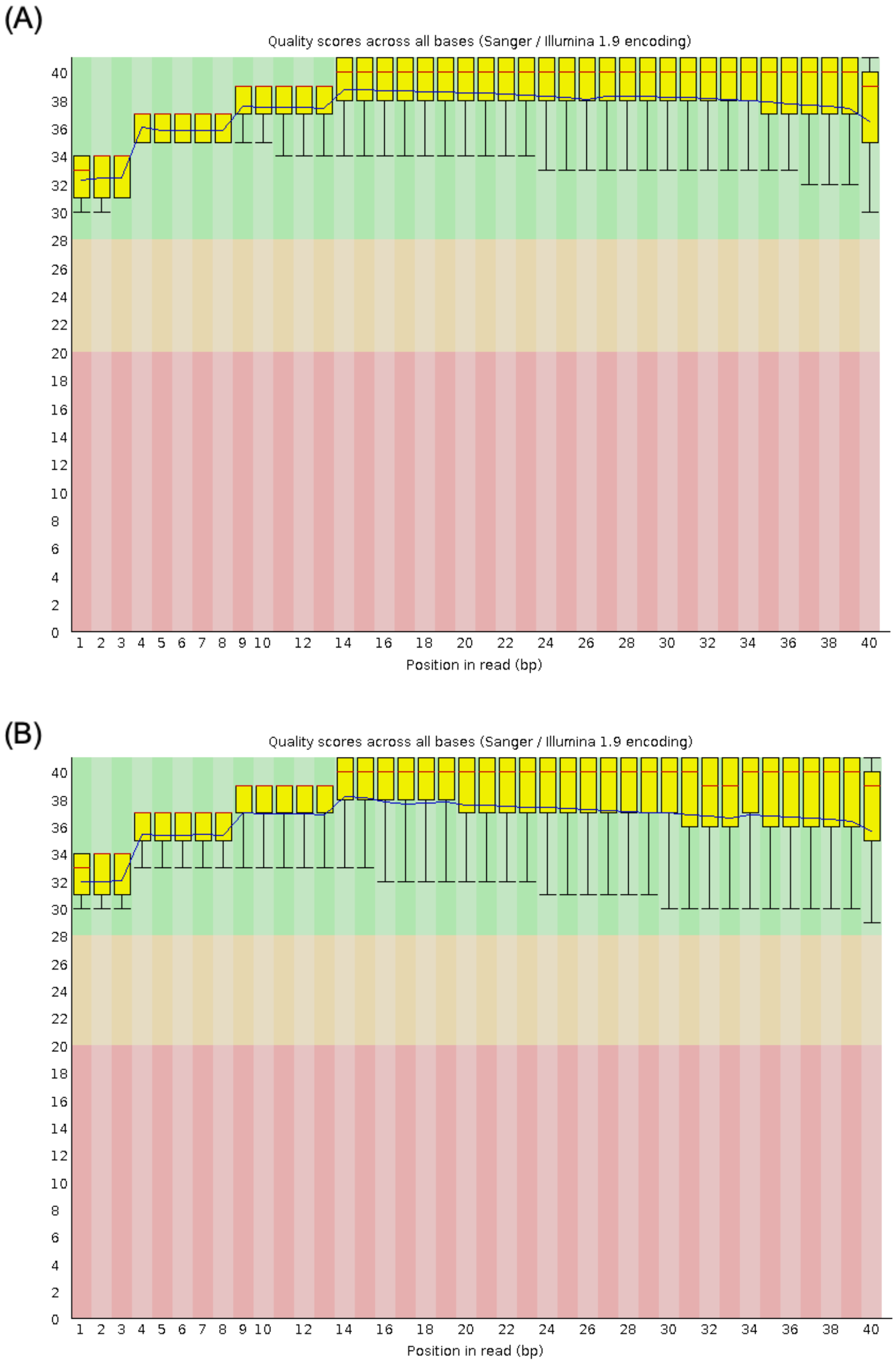

Figure 1. Per base sequence quality plots for paired ends 1 (A) and 2 (B)

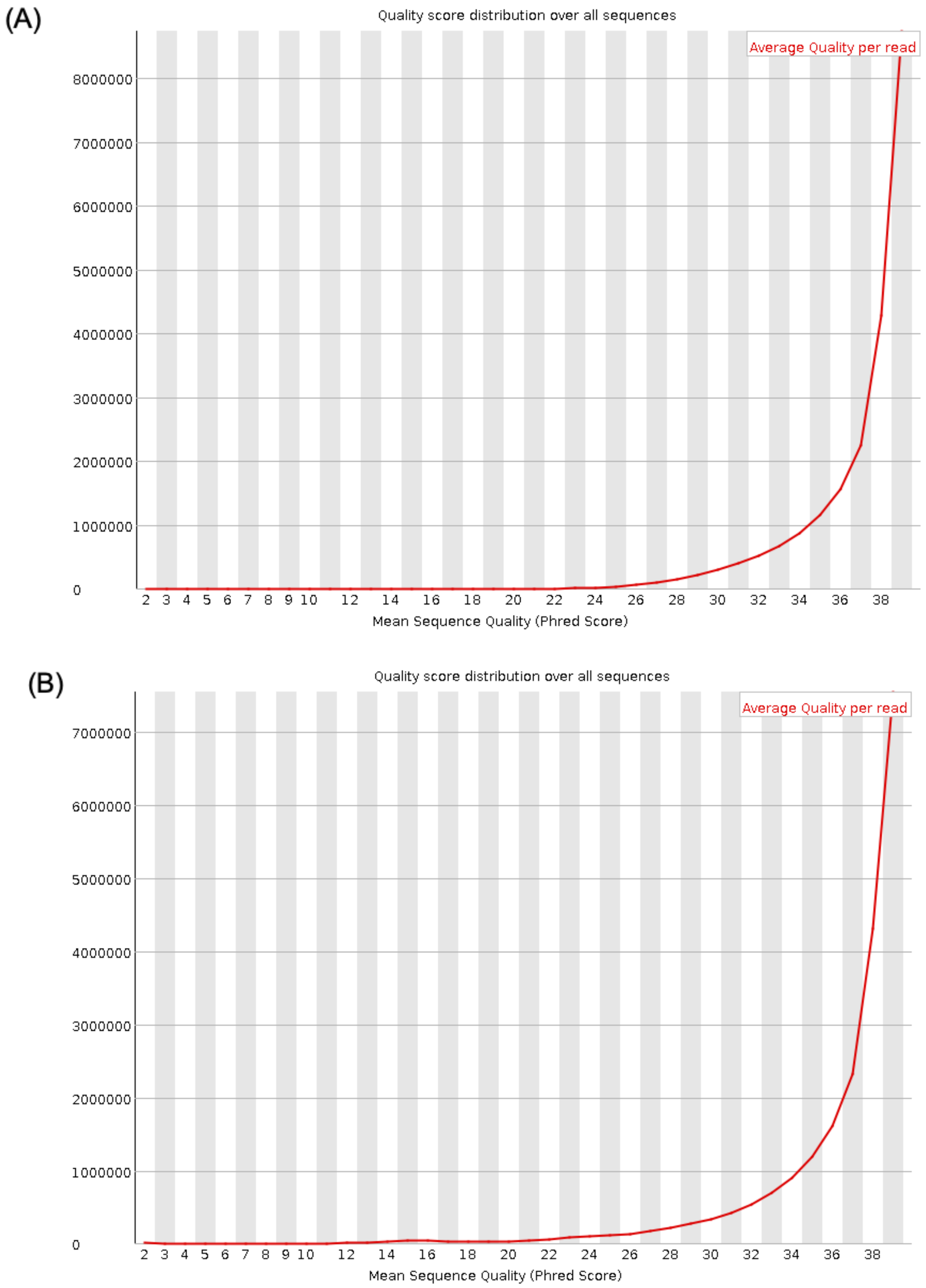

Figure 2. Per sequence quality scores for paired ends 1 (A) and 2 (B).

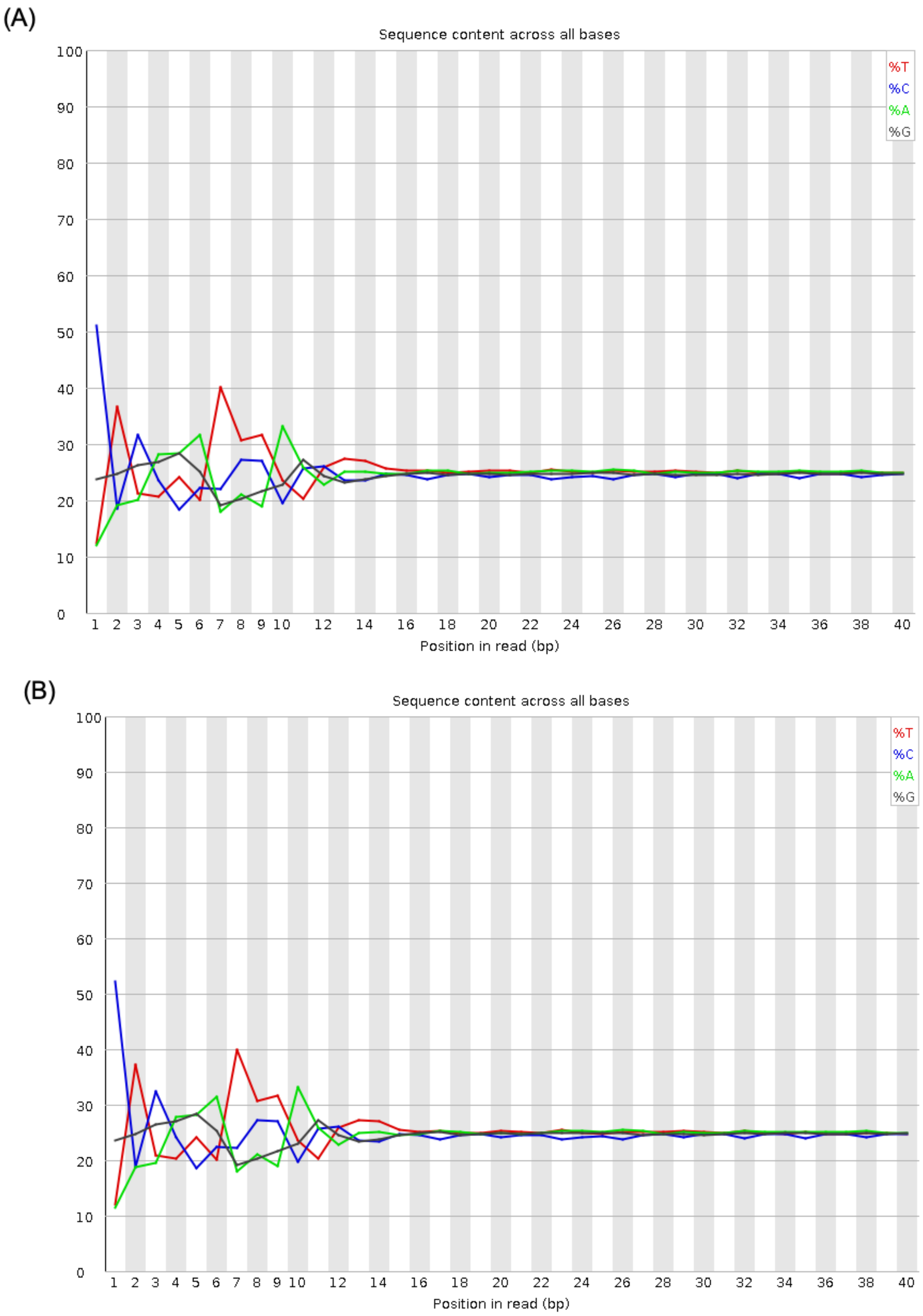

Figure 3. Per base sequence content for paired ends 1 (A) and 2 (B).

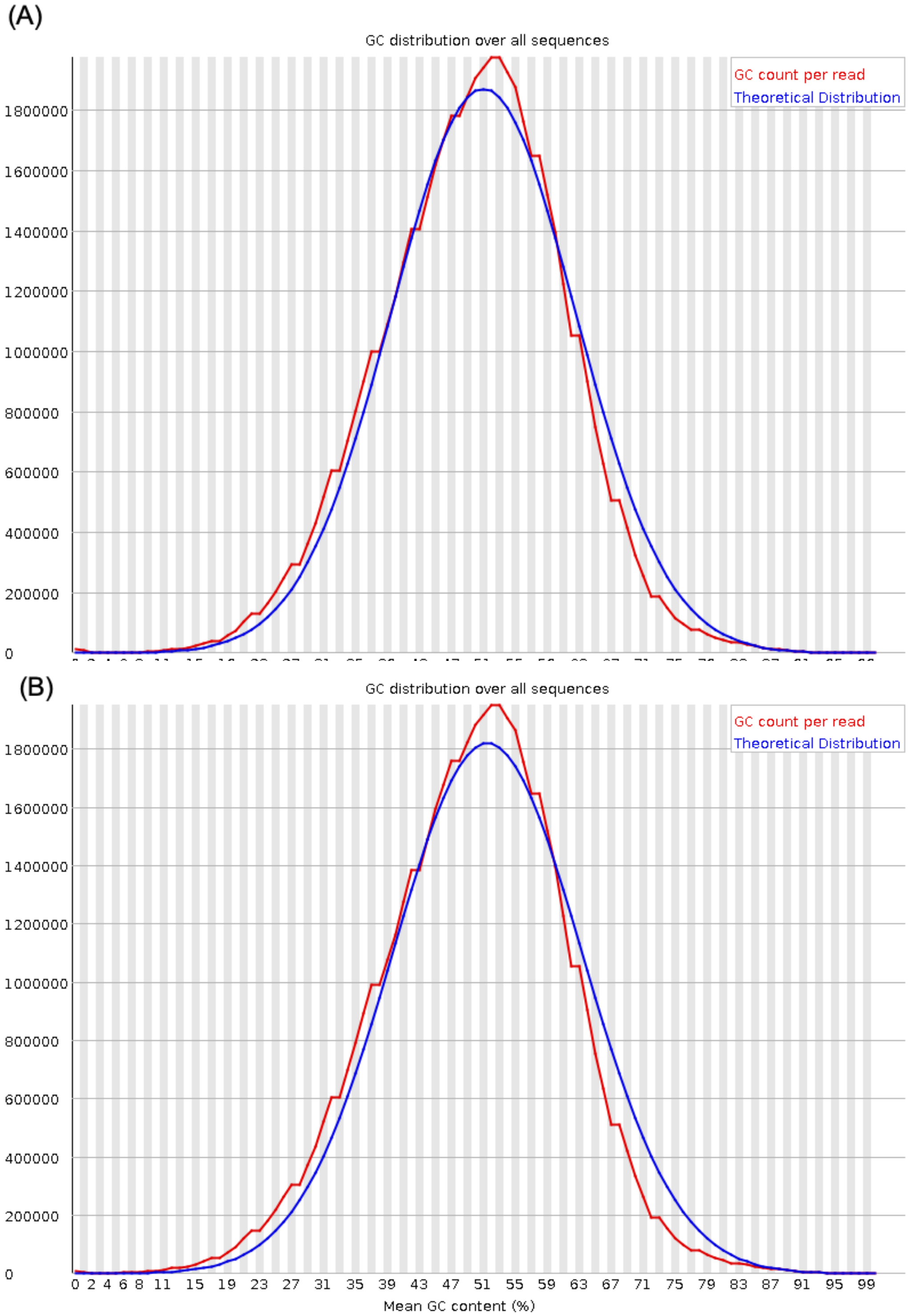

Figure 4. Per sequence GC content for paired ends 1 (A) and 2 (B).

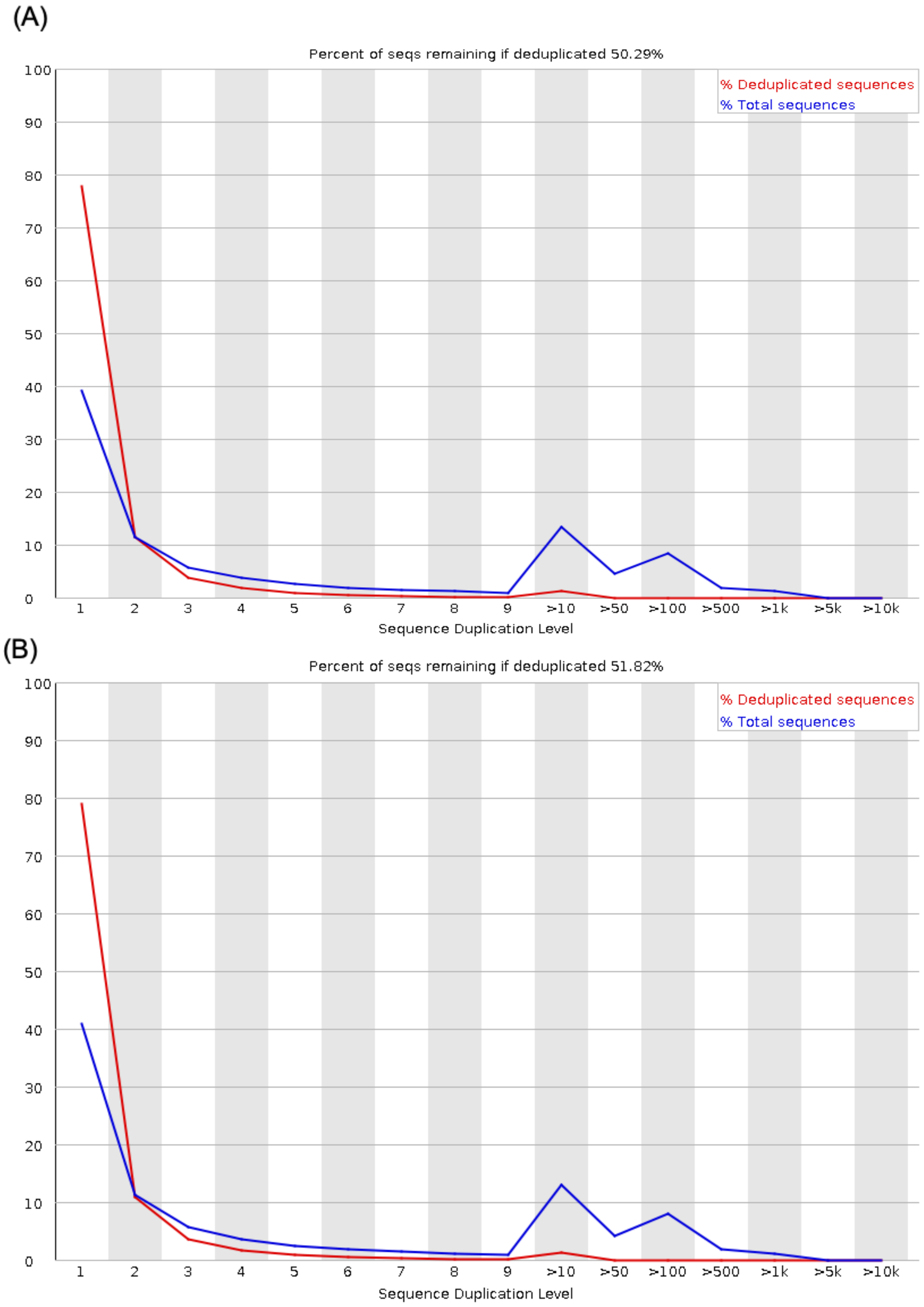

Figure 5. Sequence duplication levels for paired ends 1 (A) and 2 (B)

Methods

Alignment of the mouse genome reads was performed against the reference sample genome mm9 found on the BU SCC. Using MobaXterm to interface with the BU SCC, the first step of alignment involved loading in several modules (it should be noted that this aspect of the project was performed through linux, and that unless stated otherwise it should be assumed that other steps were also executed through linux). TopHat ("*TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons.*"[2]) and its dependencies SAMtools (*SAMtools is a set of utilities for interacting with and post-processing short DNA sequence read alignments in the SAM, BAM and CRAM formats."[4]),* BOOST ("*BOolean Operation based Screening and Testing (BOOST) is a method for detecting gene-gene interactions. It allows examining all pairwise interactions in genome-wide case-control studies in a remarkably fast manner."[5])* along with BowTie2 ("*Bowtie 2 is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences."*[3]) were loaded in, and next was the creation of the batch file run_tophat.qsub where all code for running our TopHat query was placed. As TopHat requires a large amount of memory to execute,and due to the size of our reads, a batch file was necessary to carry out our analysis. Run time of the query was around 1 hour.

Upon completion of our query, a file named accepted_hits.bam was created. Statistical analysis of this file was performed using SAMtools flagstat along with the RseQC command bam_stat.py. The RseQC tools geneBody_coverage.py and inner_distance.py were utilized to determine the read coverage over the gene body and the calculated inner distance (or insert size) between the two paired RNA reads, respectively. Graphs were produced displaying the outputted results (Figures 6, 7, & 10).

Analysis proceeded with the Cufflinks tool ("*Cufflinks assembles transcripts, estimates their abundances, and tests for differential expression and regulation in RNA-Seq samples."*[6]). The batch file run_cufflinks.qsub was created to rub our commands on accepted_hits.bam through the SCC. From this job, the file genes.fpkm_tracking was created, containing the quantified alignments in FPKM for all genes. This file was loaded into the statistical analysis program R and a bar graph was formed to compare the FPKM values of all the genes against one another (Figure 8). This first iteration of the bar graph was unreadable; only after filtering out FPKM values that were less than 2500 (Figure 9) was the graph deemed complete. Finally, we utilized the cuffdiff tool in the cufflinks suite to identify differentially expressed genes.

The differential expression of Postnatal mice Day 0 to Adult mice (Week 8-10) was loaded in R Studio and the data frame was sorted based on increasing q-value. A subset having only significant genes was then made. Two histograms were plotted based on log2 fold change for all genes and significantly expressed genes respectively. The differentially expressed genes were filtered based on p-values<0.01 and the number of up and down regulated genes was noted. Further the significant genes were subset into up and down regulated genes and written into .csv files which were then upload on DAVID(Database for Annotation, Visualisation and Integrated Discovery) version 6.8 for functional annotation clustering to identify enriched terms using GOTERM_BP_FAT, GOTERM_MF_FAT and GOTERM_CC_FAT gene ontology terms for Mus musculus species. The top annotations identified were used for comparison.

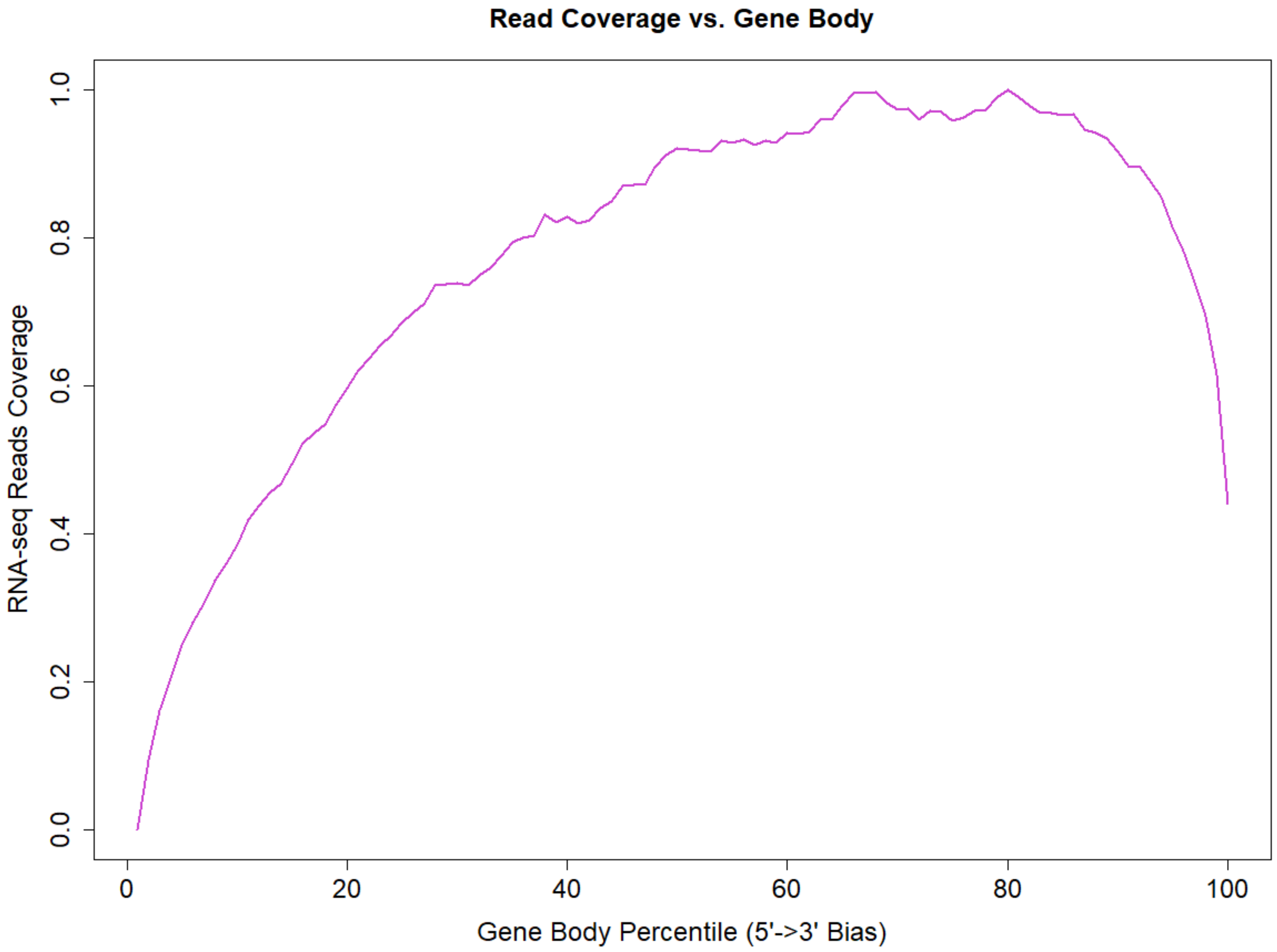

Figure 6: Read coverage over the gene body, showing coverage uniformity and 5'/3' bias.

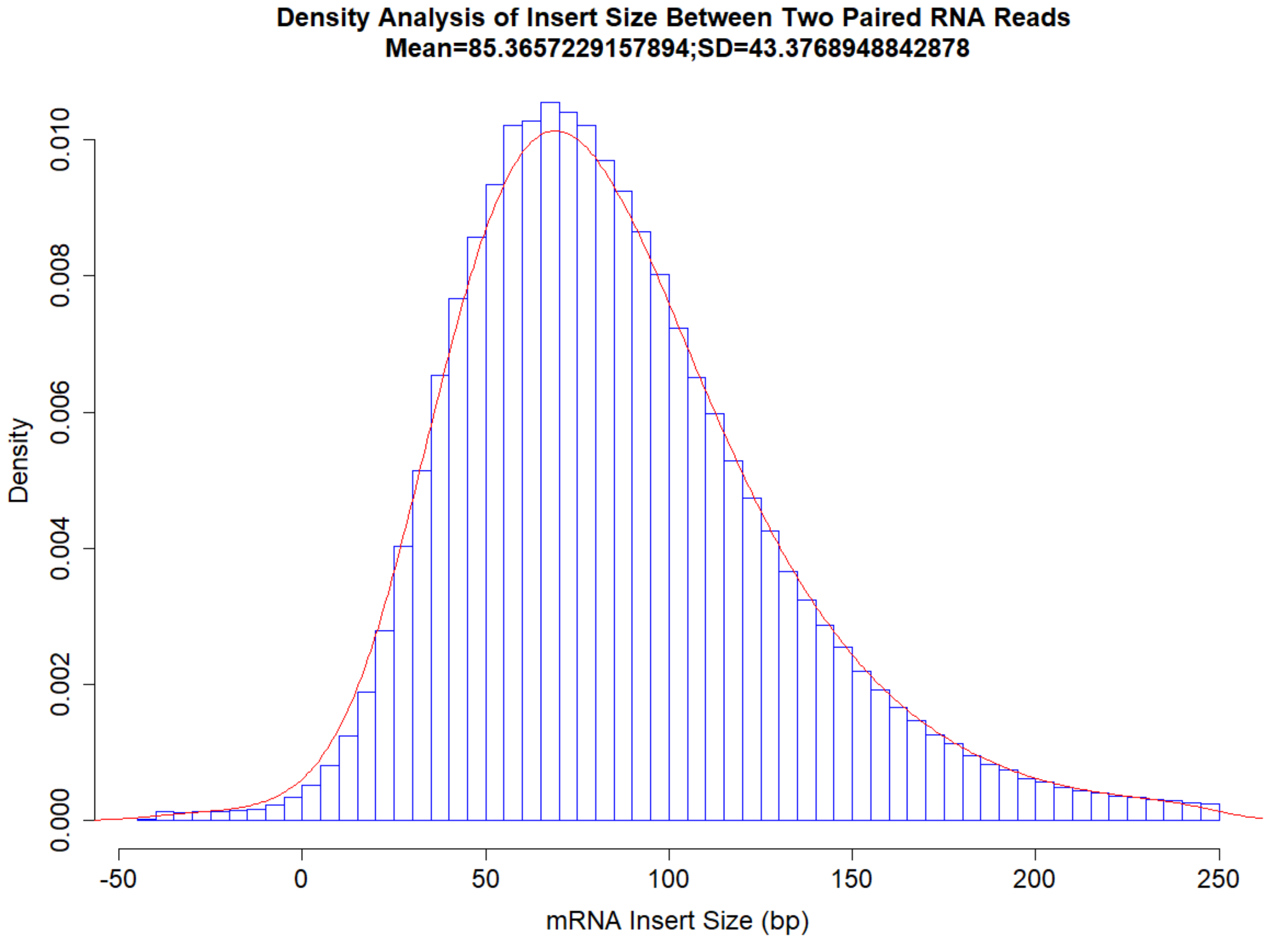

Figure 7: Density figure of the calculated inner distance (or insert size) between the two paired RNA reads.

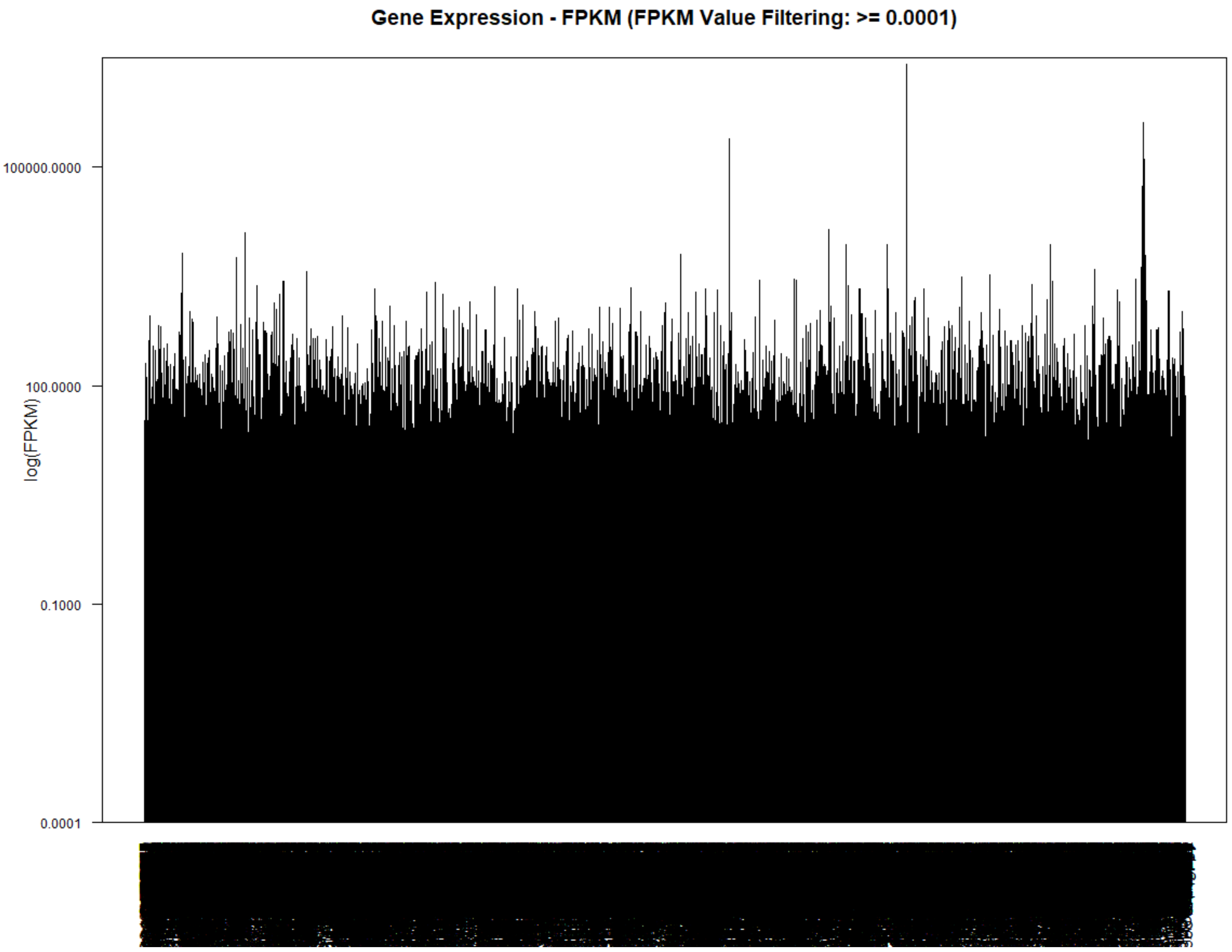

Figure. 8: First iteration of FPKM value chart for each gene, containing nearly all FPKM values (excluding any values >= 0.0001). Obviously completely unreadable.

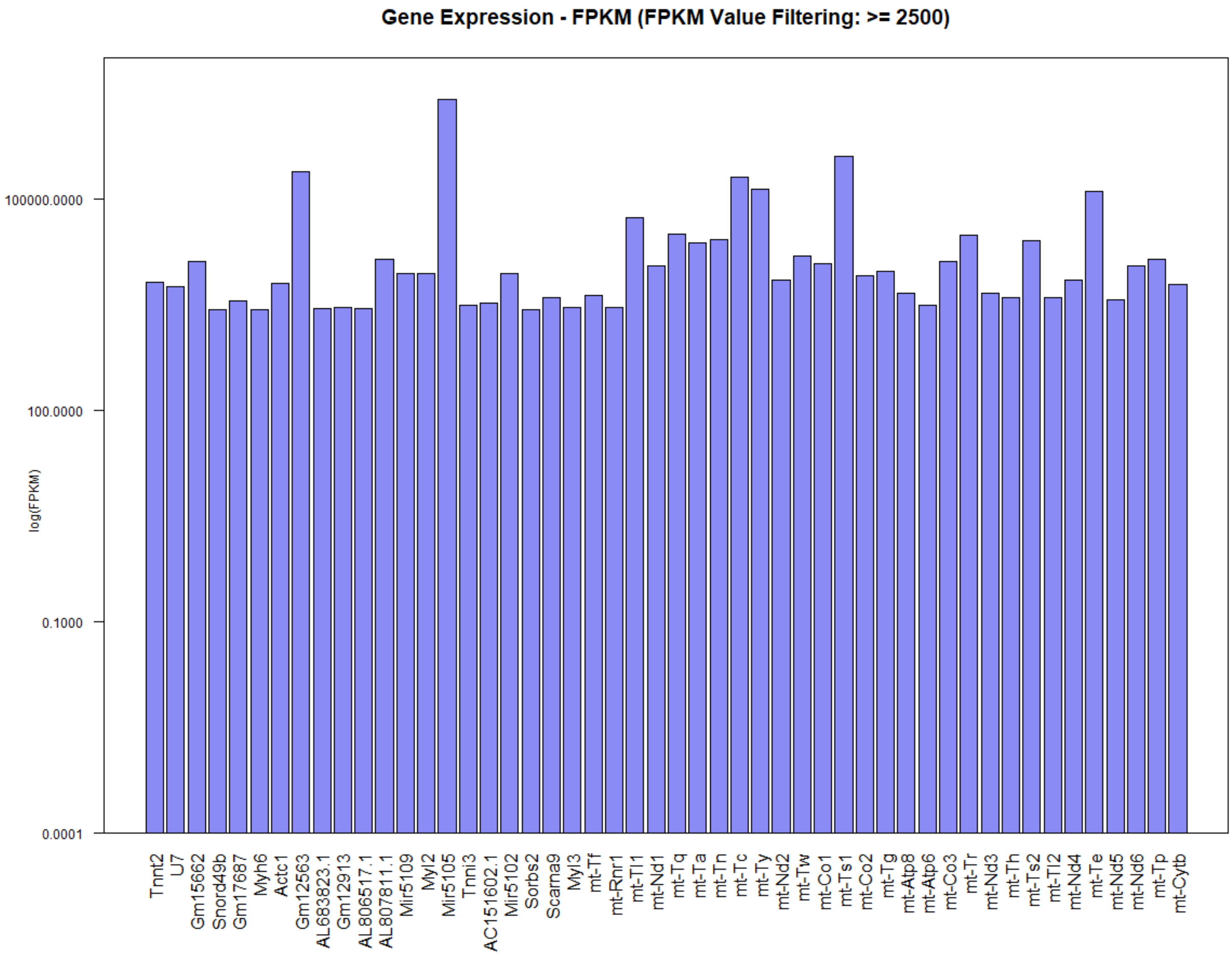

Figure. 9: Second iteration of FPKM value chart for each gene, containing all FPKM values that are >= 2500 and the genes those values correspond to.

Results

Table 1 shows Top 10 differentially expressed genes of Postnatal and Adult with its statistics. The p_values of all the top 10 differentially expressed genes is less than 0.05 and q_values are equal to 0.00107252 which is again less than 0.05. This indicates the genes are of great significance.

Number of Total Reads	49767295	
Number of Mapped Reads	49767295	100% of Total Reads
Number of Unique Reads	38542738	77.4% of Total Reads
Number of Multimapped Reads	8377961	16.8% of Total Reads
Number of Unaligned Reads	0	0.00% of Total Reads

Figure 10: Report of the total number of reads, number of mapped, unique, multimapped, and unaligned reads with percentages of total reads for each.

Gene	FPKM_PO	FPKM_ADULT	Log2 fold change	p_value	q_value
Plekhb2	22.5881	73.5662	1.70348	5.00E-05	0.00107252
Mrpl30	46.4658	133.033	1.51754	5.00E-05	0.00107252
Tmem182	39.1222	110.932	1.50361	5.00E-05	0.00107252
Coq10b	11.0911	53.3001	2.26474	5.00E-05	0.00107252
Aox1	1.19013	7.09119	2.57491	5.00E-05	0.00107252
Sp100	2.13855	100.856	5.55952	5.00E-05	0.00107252
Cxcr7	4.96472	32.2744	2.70061	5.00E-05	0.00107252
Lrrfip1	118.97	24.6387	-2.2716	5.00E-05	0.00107252
Gpc1	51.0971	185.324	1.85874	5.00E-05	0.00107252
Slc41a1	5.98702	20.4103	1.76939	5.00E-05	0.00107252

Table 1: Top 10 differentially expressed genes associated with myocyte differentiation sorted by smallest q-values where FPKM_PO is the fragment per kilobase million for Postnatal Day 0, FPKM_ADULT is the fragment per kilobase million for Adult sample (8 to 10-Week-old mice), log2 fold change is log2(FPKM_ADULT/FPKM_PO) and p-value

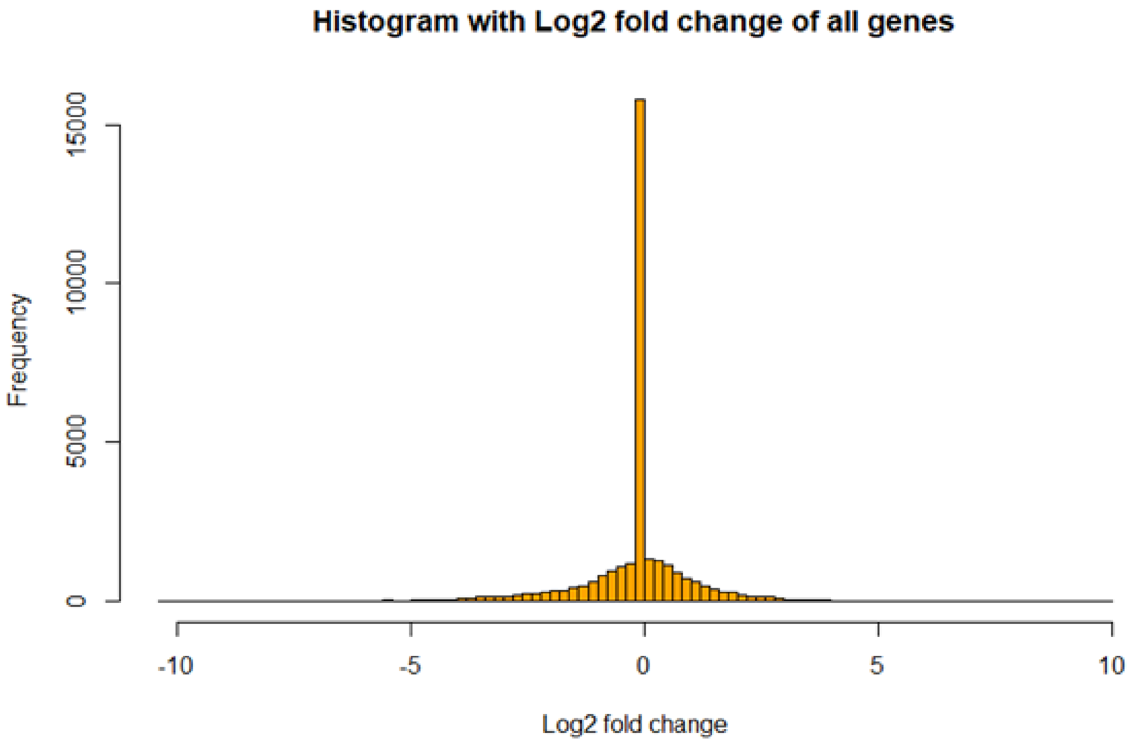

Figure 11: Histogram showing the frequency of log2 fold change of all differentially expressed genes.

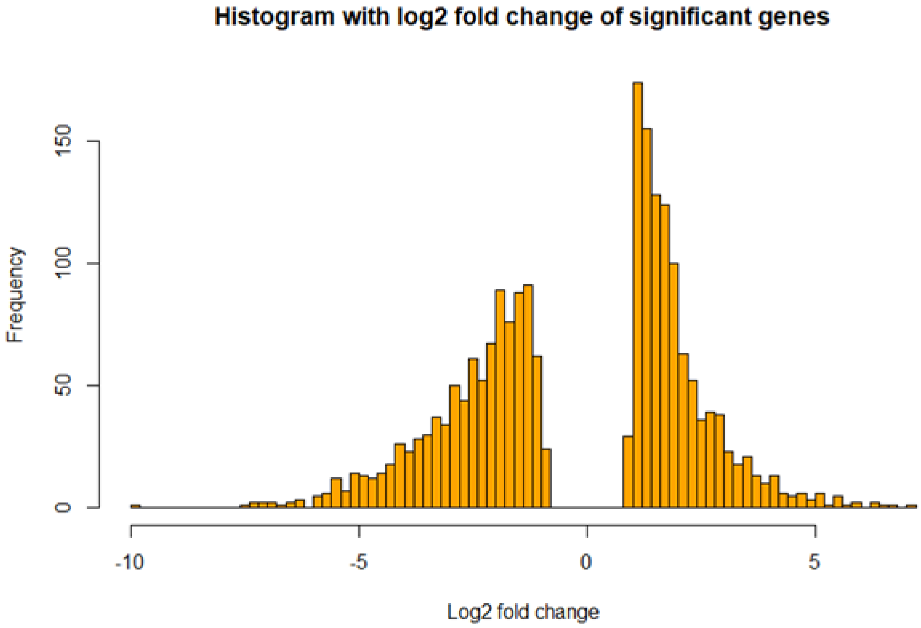

Figure 12: Histogram showing frequency of log2 fold change of significantly expressed genes.

The histogram of Log2 fold change of differentially expressed all genes and significant genes is shown in Figure11 and Figure 12. For all genes, the histogram was centered to 0 with almost equal number of positive and negative values whereas for significant genes that was not the case. Here the distribution was towards the positive and negative values with greater frequency of positive values.

Differentially expressed genes at p_value<0.01 for all genes	
Up regulated genes	1185
Down regulated genes	1189
Total	**2374**

Table 2: Number of differentially expressed genes from all genes at p_value<0.01

Differentially expressed genes at p_value<0.01 significant genes	
Up regulated genes	1091
Down regulated genes	1075
Total	**2166**

Table 3: Number of significantly expressed genes at p_value<0.01

Table 2 and 3 show the number of up and down regulated genes at p_value<0.01 where log2 fold change above 0 means up regulated and log2 fold change below 0 means down regulated for all differentially expressed genes and significant genes respectively. Total number of genes at p_value<0.01 were 2374 for all genes with 1185 up regulated and 1189 down regulated genes whereas for significant genes, there were 1091 up regulated genes and 1075 down regulated genes with a total of 2166 significant genes.

The top 10 selected clusters from DAVID analysis for down and up regulated genes along with their enrichment scores is shown in Table 4 and Table 5 respectively. The highest enrichment score for down regulated genes was 10.965 which was for cell cycle, chromosome segregation and nuclear segregation. The highest enrichment score for up-regulated genes was 21.496 which was for mitochondrial membrane and envelope.

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

Table 4: Selected top gene clusters for down regulated genes along with their enrichment scores. Clusters that overlap with the O'Meara *et al.* (2015) results are underlined.

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

Table 5: Selected top gene clusters for up regulated genes along with their enrichment scores. Clusters that overlap with the O'Meara *et al.* (2015) results are underlined.

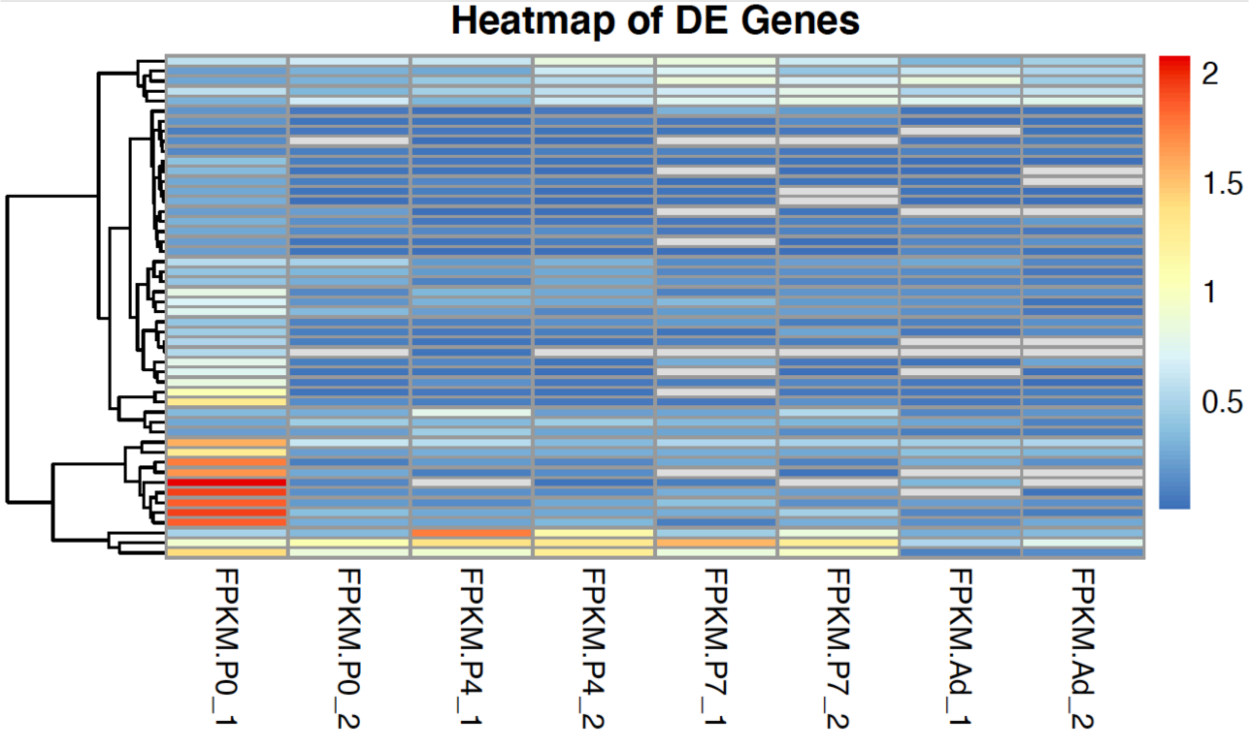

Figure 13: Heatmap of FPKM of the top50 DE genes between P0 and Ad across all samples. Each column represents a sample and they are clustered by replicate followed by age. Genes were clustered using euclidean distance metric.

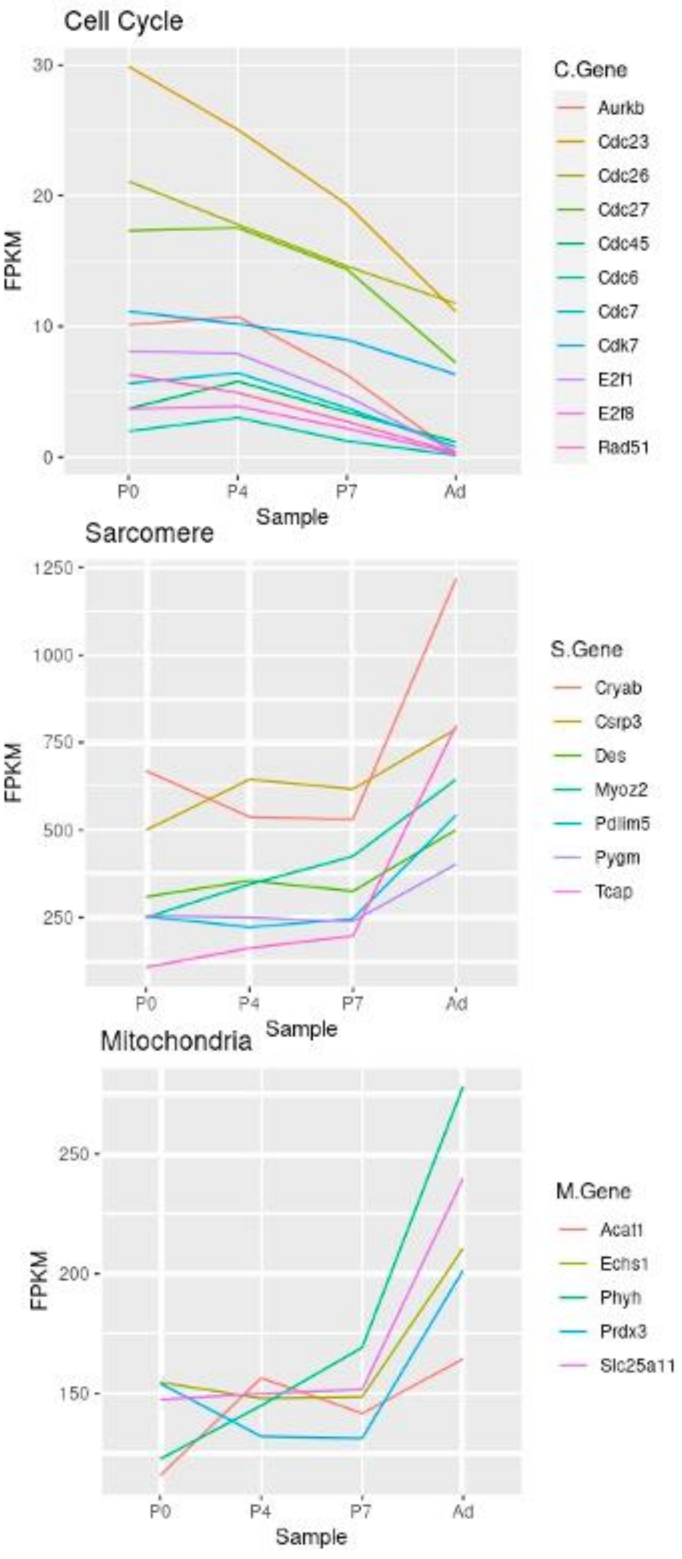

Figure 14. Recreation of O´Meara's Figure 1D for datasets P0, P4, P7 and Ad.

Discussion

Figure 6 shows an overall increase in the gene body percentile in accordance with the RNAseq reads coverage, with a sharp dropoff occurring around the 80th percentile. Figure 7 displays a bell curve of the mRNA insert size, with the majority of the distribution being between 50-100 base pairs long. Figure 9 shows the 53 genes resulting after filtering took place and their FPKM values along a logarithmic scale on the y axis.

The output of cuffdiff produces a list of 36329 differentially expressed genes along with their gene symbol, FPKM values, Log2 fold change, p_value, q_value and significance (yes/no). Log2 fold change was evaluated as log2 of (FPKM_ADULT/FPKM_PO) which measures the change of expression level between given samples. A subset of genes that were significant was made which had 2166 genes in total. Histograms were plotted for log2 fold change of all genes and significant genes (Figure 11 and Figure 12). We found that for all genes, values were centered at 0 whereas for significant genes, distribution was towards the extremes with more positive values than the negative values which meant that the up regulated genes were more than the down regulated genes. Based on p_values, we filtered all the genes and found that there were 1185 up regulated genes and 1189 down regulated genes. The number of up-regulated genes and down regulated genes in the O'meara et al study was 2409 and 7570 respectively with a total of 10000 significant genes which was quite different from the results we got in Table 3 and 4. One possible reason for that might be that the author must have applied different thresholds than what we used (p_value<0.01). However, in paper the number of down regulated genes were greater than up regulated genes which was similar to what we found.

From DAVID results for up-regulated genes, we found Mitochondria(21.46), sarcomere(6.967) and Glycolysis(5.559) which overlapped with the results in the paper. For down regulated genes, we found RNA process(7.887) and cell cycle(10.965) that overlapped with down regulated gene results from the paper. However, despite overlapping the enrichment scores for these enrichment terms differed which might be due to the differences in the version of DAVID that the author used or difference in the gene terms used for DAVID analysis. It is important to mention that with 5+ years between the paper and our current analysis, DAVID is bound to have been altered with many genes added and pathways edited. Getting exact results would make us question the validity of DAVID. Differences might also be due to the difference in the number of up and down regulated genes used for DAVID tests.

A visual representation of DE genes is shown in Figure 13. There is a significant difference between the heatmaps on the original paper and the one in our results. It is clear that there was a difference in the processing of the P0_1 sample as it differs from the rest. It would not be wise to compare Figure 11 to the authors' similar Figure since they clustered according to their DAVID results and not only did we plot a smaller percentage of genes, but we also clustered according to a subset of the results. Heatmaps can look completely different based on the clustering method of the rows. Therefore, by selecting the same genes we used for our analysis in the same order, it is possible we would be able to replicate our results from those the author´s data. It is also worth noting that IL13 does not appear in the heatmap.

Contrary to Figue 13, our recreation of the author´s Figure 14 could be considered exact. Differences in the behavior of some genes could be due to the fact that we are dealing with one less sample in our datasets, but they all follow the same trend that was seen by the authors. It should be noted that the gene Bora was filtered out from our analysis in the Cell cycle graph.

Conclusion

The fact that the author´s Figure 1D was accurately replicated shows that O'Meara *et al.* (2015) analysis is reproducible, our analysis was correctly done and that the differences risen are likely the result of the choice of tool or parameters for our analysis. Although heatmaps are visually attractive, they are harder to reproduce since there different methods for gene clustering can lead to misleading results. The contrasting results in the pathway enrichment analysis on DAVID suggest that previously published data could potentially benefit from reruns of pathway enrichment databases in later years for more updated results. These could lead to increased p-value of certain relationships or the appearance of new ones.

References

1) O'Meara, C. C., Wamstad, J. A., Gladstone, R. A., Fomovsky, G. M., Butty, V. L., Shrikumar, A., Gannon, J. B., Boyer, L. A., & Lee, R. T. (2015). Transcriptional reversion of cardiac myocyte fate during mammalian cardiac regeneration. *Circulation research*, *116*(5), 804–815. https://doi.org/10.1161/CIRCRESAHA.116.304269
2) TopHat. Ccb.jhu.edu. Retrieved from https://ccb.jhu.edu/software/tophat/index.shtml
3) Bowtie 2: Manual. Bowtie-bio.sourceforge.net. Retrieved from http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml#introduction
4) SAMtools. Wikipedia, the Free Encyclopedia. Retrieved from https://en.wikipedia.org/wiki/SAMtools
5) BOOST Main Page. Bioinformatics.ust.hk. Retrieved from http://bioinformatics.ust.hk/BOOST.html
6) Cufflinks (30 Nov. 2010.). Cufflinks. Cufflinks. Retrieved from http://cole-trapnell-lab.github.io/cufflinks/
7) DAVID: Functional Annotation Result Summary. (2021). Retrieved from Ncifcrf.gov website: https://david.ncifcrf.gov/summary.jsp