Transcriptional Reversion of Cardiac Myocyte Fate During Mammalian Cardiac Regeneration

Data Curator: Maha Naim
Programmer: Arushi Shrivastava
Analyst: Teresa Rice

INTRODUCTION

Mammalian cardiac regeneration is a widely studied topic within the realms of stem cell biology. In particular, the capacity for cardiac regeneration is a remarkable phenomenon noted among neonatal mice. A number of studies have examined the potential mechanisms underlying this biological process, however, questions remain with regards to the molecular regulators that govern the capacity for cardiac regenerative response following a mouse's first week of life [1]. The study, *Transcriptional Reversion of Cardiac Myocyte Fate During Mammalian Cardiac Regeneration*, approaches this anomaly with the aim of determining the transcriptional changes by which the process of cardiac regeneration is characterized. To examine the transcriptional phenotype and identify its key regulators, three experimental models were profiled to explore the molecular changes: in vitro and in vivo cardiac myocyte differentiation, and in vitro adult cardiac myocyte explantation [2]. The authors found that cardiac regeneration is a transcriptional reversion of the differentiation process observed in immature neonatal cardiac myocytes. Additionally, interleukin 13 was identified as a novel regulator involved in cardiac regeneration and cell cycle activity of cardiac myocytes. The bioinformatic techniques used in the study allowed for the identification of core molecular changes that take place during mammalian cardiac myocyte differentiation and regeneration via interrogation of high throughput RNA expression data [2]. Comparative analysis using multi-model profiling methods grants further verification of the representativity of the observed transcriptional changes. Consequently, our project aims to reproduce a portion of the results using the tools referenced in the study.

METHODS

The sequencing data utilized in the study was accessed from Gene Expression Omnibus (GEO) Series GSE634403, a publicly available genomics data repository, and prepared for the *Mus musculus* genome.

RNA sequencing data was collected at each timepoint during the differentiation of mouse embryonic stem cells into cardiac myocytes via mesodermal and cardiac progenitor intermediates from Wamstad et al [2, 3]. There were various samples analyzed by the researchers over the course of the study, however our project looked at a single P0 sample, sourced from postnatal day zero ventricular myocardium. Paired-end reads, each forty nucleotides in length, were generated using an Illumina HiSeq 2000 instrument. Upon access to the GEO repository, the SRA (Sequence Read Archive) file was downloaded and extracted to FASTQ format using the SRA toolkit. Since the study performed paired-end sequencing, two FASTQ files were generated: PO_1_1 and PO_1_2. To confirm the data was consistent, a head command was used to evaluate that the headers of both files were the same.

To assess the quality of the FASTQ files, the FASTQC package was utilized to begin processing and perform quality control measurements. Using FASTQC, various tests were conducted on the samples to determine the quality of the dataset. The data was assessed to be high quality per the basic statistics included in the FASTQC report.

After receiving the two FASTQ files of the sample, P0_1, the mouse genome reference, mm9, was used along with the Bowtie2 indexes to align these reads. TopHat, an open-source alignment tool, was used to perform the alignment on the reads generated from the RNA-Seq data with the reference genome using Bowtie. With the help of this tool, the two FASTQ files were aligned against the reference genome which generated the accepted_hits.BAM file containing all the original reads and the alignment by TopHat in a binary format.

After receiving the BAM file, an additional step to visualize the information was performed using the integrated genome viewer (fig.3) with mm9 as the reference genome. The summary statistic was created using Samtools that shows that 49,706,999 QC-passed reads, there were no duplicates, all reads mapped and paired in sequencing. There were 25,089,027 reads that belong to read1 and 24,617,972 reads belong to read2. 32,466,938 reads got properly paired, 47,843,662 reads with itself and mate mapped, 1,863,337 were singletons, 5,098,744 with mate mapped to a different chromosome. The accepted_hits.BAM.bai file was generated after completing the indexing of the BAM file.

With the help of RseQC Packages, that comprehensively evaluates high throughput sequence data especially RNA-seq data. Therefore, first geneBody_coverage.py was used to calculate the RNA-seq reads coverage over the gene body. The x-axis contains the list of base pairs of genes and the y-axis shows the coverage. The plot seems to be reasonable as the coverage increases with increasing the Gene body percentile (5'−>3') and at the end it shows degradation which can be due to some artifact as this is usually the case seen most of the time. This can be due to the technical or non-technical error during the process.(fig.4a)

Using inner_distance.py, the inner distance (or insert size) between two paired RNA reads was then calculated. The distance is the mRNA length between two paired fragments. It seems that the plot follows the normal distribution with Mean=85.4;SD=43.4.(fig.4b)

Using the bam_stat.py file for summarizing mapping statistics of a BAM file gives a detailed understanding of the mapping statistics. The total record was 49,706,999 read counts. There were no PCR duplicates and no reads failed during quality control. There were 8,317,665 non primary hits, 0 unmapped reads, 33,099,839 non-splice reads, 5,389,541 splice reads, and 19,236,824 reads map to the positive strand (‘+’), whereas 19,252,556 reads map to the reverse strand (‘-’).

After using TopHat to align the reads to the reference genome, inference of the relative abundance of transcripts was performed using the tool, Cufflinks. Cufflinks assembles transcripts, estimates their abundances, and tests for differential expression and regulation in RNA-Seq samples. Cufflinks produced the four files including genes.fpkm_tracking which contains the quantified alignments in FPKM for all genes. This file was utilized for further analysis. With the help of R programming (R version 4.0.3) the histogram was generated (fig.5). Originally there were 37,469 genes, however after filtering, 16,453 genes remained. The filter was applied on the FPKM values and all genes with zero FPKM value were removed.

Finally, another package of Cufflinks called Cuffdiff was utilized at the gene and transcript levels in more than one condition, and then tested for the significance differences. There were four samples being used for the study (each of P0 and Ad have two replicates). All commands were runned on the SCC server using qsub, as a batch job. It took ~more than an hour to complete.

For TopHat, the module ‘samtools/0.1.19 bowtie2 boost tophat’ was used with the following parameters: -r 200 -G mm9.gtf --segment-length=20 --segment-mismatchs=1 --no-novel-juncs -o P0_1_tophat -p 16 mm9 P0_1_1.fastq P0_1_2.fastq. To examine the result ‘samtools flagstat accepted_hits.bam’ was used, as it counts the number of alignments for each FLAG type. It provides

counts for each of 13 categories based primarily on bit flags in the FLAG field. Each category in the output is broken down into QC pass and QC fail. Afterwards, 'module load python3 samtools rseqc' was used to analyze the RNA-Seq data. BAM file indexing was done using 'samtools index accepted_hits.bam'. The parameters used for geneBody_coverage.py -r mm9.bed -i accepted_hits.bam -o output. For inner_distance.py -r mm9.bed -i accepted_hits.bam -o output1 and for bam_stat.py -i accepted_hits.bam. For cufflinks, the cufflinks module was loaded and the parameters used for cufflinks --compatible-hits-norm -G mm9.gtf -b mm9.fa -u -o P0_1_cufflinks -p 16 accepted_hits.bam. For cuffdiff, the following parameters were: cuffdiff -p 16 -L $LABEL -u -b $FASTA -o $OUTDIR $SAMPLEDIR/merged_asm/merged.gtf $P0REPS $ADREPS.

R version 4.0.3 was used to analyze the data. R packages tidyverse and dplyr were used to easily manage and manipulate the data. R package ggplot2 was used to create plots from the data and RDAVIDWebService was used in order to easily maneuver data created from the DAVID database.

RESULTS

A number of quality control tests were performed on both of the extracted FASTQ files, PO_1_1 and PO_1_2, by the FASTQC tool. Results are shown for each file in Figures 1 and 2, respectively. Both samples passed every quality control measure, except for the per base sequence content test. All passed control measures are depicted for PO_1_1 in Figure 1A to 1I and for PO_1_2 in Figure 2A to 2I, excluding 1E and 2E which display the per base sequence content test results. The detection of irregular base content above a difference of 20% between bases constitutes as a failed test and may be indicative of a source of contamination [5]. Though the sequence duplication quality test was passed, a warning was issued for both PO_1_1 and PO_1_2 and is indicated in Figures 1B and 2B. Warnings for sequence duplication tests are issued when non-unique sequences make up more than 20% of the total amount [5]. This error may also indicate a source of contamination. Despite these errors, the data appears to be of good quality given the overall statistics computed by FASTQC and posed no implications in later downstream alignment and analysis.

A table of the top ten differentially expressed genes was produced using the cuffdiff output gene expression dataframe (table 1). The entire data-frame revealed there were 628 genes which all shared the lowest q-value (these all shared a p-value of 0.00005 as well). An updated two tables were made utilizing log2FC (table 2). These tables more closely resemble the lists from the supplement but still have significant variation. This could be due to working with a modified dataset as well as the supplement determining the up-regulated genes from P0 vs Adult Log2FC.

A histogram created of all log2FC values of all genes reveals a normal distribution centered over 0 (fig 6). A second histogram of significantly differentially expressed genes shows a histogram with a gap in the center. Any log2FC value close to zero is not seen as significantly differentiated and therefore removed (fig 7). The second histogram revealed 1084 genes were found to be up-regulated and 1055 genes were found to be down-regulated for a total of 2139 differentially expressed genes. The paper states "In total, 927 genes were commonly up-regulated in both datasets"[2]. From figure 1B it can be seen that there were 3229 genes commonly down-regulated in both datasets[2], my results indicate a more balanced expression of each side.

Two csv files containing the up and down regulated significantly expressed genes are named upreg.csv and downreg.csv respectively and can be found in our github repository.

DAVID Functional Annotation Clustering tool was used to interpret the up and down-regulated

GO terms. At the bottom of this paper there are two tables containing summarized data from our team and from the paper (tables 3&4). The first, most noticeable difference is the enrichment values for the down-regulated GO terms are significantly higher than the ones produced by our team. These are coupled with much larger gene counts and p-values smaller by about one factor of ten. Due to the fact the gene count for down-regulated terms is about four fold higher than ours, this likely is affecting the enrichment, p-value, Benjamini and other statistical tests were they performed. The up-regulated terms produced by our team more closely reflect those from the paper. Enrichment scores are not identical but very similar, our top enriched GO term is the same as the one produced by the paper and even has the same number of members. Other differences in our data is likely stemmed from the difference in gene count, our gene count for up-regulated GO-terms is overall higher than those from the paper and our p-values are more significant.

DISCUSSION

To perform downstream analysis, cardiac myocyte samples from neonatal mice (P0) were downloaded, extracted, and quality controlled. Generally, the resulting data met good quality standards for both P0_1_1 and P0_1_2, as evaluated by FASTQC. No GC content bias or overrepresented sequences were detected. Contrastingly, however, our data failed the per base sequence content test and received a warning by the sequence duplication test. The per base sequence content error may be due to the selection of biased primers or biased sequence composition as indicated by the spike in Thymine and Adenine nucleotides in Figure 1E and 2E towards the start of the read. Despite this technical bias, it is unlikely that this error affected our downstream analysis [5]. In terms of the sequence duplication warning, it indicates the presence of either technical or biological duplicates. If technical, these duplicates may arise from PCR artifacts [6]. If biological, it is implied that our duplicates are a result of natural collisions where different copies of exactly the same sequence were randomly selected [6]. Given that the test did not fail, it is unlikely that our data retained a significant contaminant set. It is possible, however, that if there were tight constraints placed on our sequence start points, our start sites may have generated concerning duplication levels detected by FASTQC which do not need to be treated as a problem for later downstream analysis [6]. With that in mind, we believe our data reaches a sufficient standard which did not interfere with the following sequence alignment and gene expression quantification.

A table of the top ten differentially expressed genes was originally produced by arranging the cuffdiff output gene expression dataframe (table.1) by q-value and looking only at the top ten items in the dataframe. Upon expanding the data-frame this revealed there were 628 genes which all shared the lowest q-value (these all shared a p-value of 0.00005 as well). This top ten did not match the list from supplement 2 1B_CommonUP or 1C_CommonDown, another two tables were produced by adding additional filters to remove log2 fold change (log2FC) values of either positive or negative infinity then arranging the tables by logFC ascending and descending respectively. The new tables have the top 10 differentially expressed genes for both positive and negative log2FC(table 2). These tables more closely resemble the lists from the supplement but still have significant variation. This could be due to working with a modified dataset as well as the supplement determining the up-regulated genes from P0 vs Adult Log2FC.

A histogram was made of the log2FC for all genes, instances where log2FC=0 or +/- infinity have been removed to produce a more informative plot; abundance/ count is on the x-axis (fig 6) From this histogram another histogram was produced, now of the differentially expressed genes. In filtering this

data I found that looking at genes where the column significant = yes is a slightly more stringent filter than setting $p<0.01$ and includes all of the same genes, the more stringent filter was used (fig 7). Both histograms were made using R package ggplot2 and large bin sizes of 300.

After all advised steps were followed using the DAVID Functional Annotation Clustering tool the clustering data was exported into a text file and read into R utilizing RDAVIDWebService. This R package more easily manipulated the data from the DAVID output and could provide consistent results without needing to rely on copy/paste[4]. The table(s) created using the DAVID output and R package only details information of the top 6 clusters of both up regulated and down regulated results, this is to simplify the upwards of 400 clusters each analysis created (table 3). The table will only give data on the top GO from each cluster, however each cluster in the raw data have varying numbers of GO and each GO a varying number of genes (see analyst.Rmd or the upregcluster.txt and downregcluster.txt from our GitHub page).

CONCLUSION

The capacity for cardiac regeneration in neonatal mice is modulated by characteristic transcriptional changes and molecular regulators. A single sample from the study was utilized in our project and obtained from the GEO repository. Our samples passed every quality control measurement, except the per base sequence content test. Despite this technical bias, it is unlikely that this error affected our downstream analysis [5].

The FASTQ file containing the reads from the sample, P0_1, were aligned to the mouse reference genome, mm9, and its Bowtie2 index. By using Tophat, the use of samtools and boost were also included as they are dependencies of TopHat. After successful completion of the TopHat tasks, the BAM file was generated, with which indexing was performed. This resulted in the bam.bai file. The BAM file was used to visualize the reads to the reference genome. Afterwards, the RSeQC package was used to understand the data more precisely using three open-source scripts: geneBody_coverage.py, inner_distance.py, bam_stat.py. The reads were also visualized with the reference genome in the IGV. Cufflinks was then used for understanding how reads map to genomic regions defined by an annotation. With the data available for other samples, Cuffdiff was utilized to identify differentially expressed genes.

Analysis of our data revealed significant differences from the results form the paper in some areas but this is likely due to the modified data-set our team worked with. Much of our data revealed normally distributed gene differentiation and similar up and down-regulated genes to what was expected. Problems with viewing DAVID Functional Annotation Clustering data were resolved by utilizing the R package RDAVIDWebService [4]. Had there been further analysis and interpretation of the FPKM expression matrices, additional meaningful biological interpretations could have been made.

REFERENCES

1. Evidence that human cardiac myocytes divide after myocardial infarction. (2001). *New England Journal of Medicine, 345*(15), 1130-1131. doi:10.1056/nejm200110113451513
2. O'Meara, C. C., Wamstad, J. A., Gladstone, R. A., Fomovsky, G. M., Butty, V. L., Shrikumar, A., . . . Lee, R. T. (2015). Transcriptional reversion of cardiac myocyte fate during mammalian cardiac regeneration. *Circulation Research, 116*(5), 804-815. doi:10.1161/circresaha.116.304269
3. Wamstad, J., Alexander, J., Truty, R., Shrikumar, A., Li, F., Eilertson, K., . . . Bruneau, B. (2012).

Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage. *Cell, 151*(1), 206-220. doi:10.1016/j.cell.2012.07.035

4. Fresno C, Fernández EA (2013). "RDAVIDWebService: a versatile R interface to DAVID." *Bioinformatics*, **29**(21), 2810–2811. http://bioinformatics.oxfordjournals.org/content/29/21/2810.
5. Per Base Sequence Content. (n.d.). Retrieved March 17, 2021, from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/4%20Per%20Base%20Sequence%20Content.html
6. Duplicate Sequences. (n.d.). Retrieved March 17, 2021, from http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/8%20Duplicate%20Sequences.html

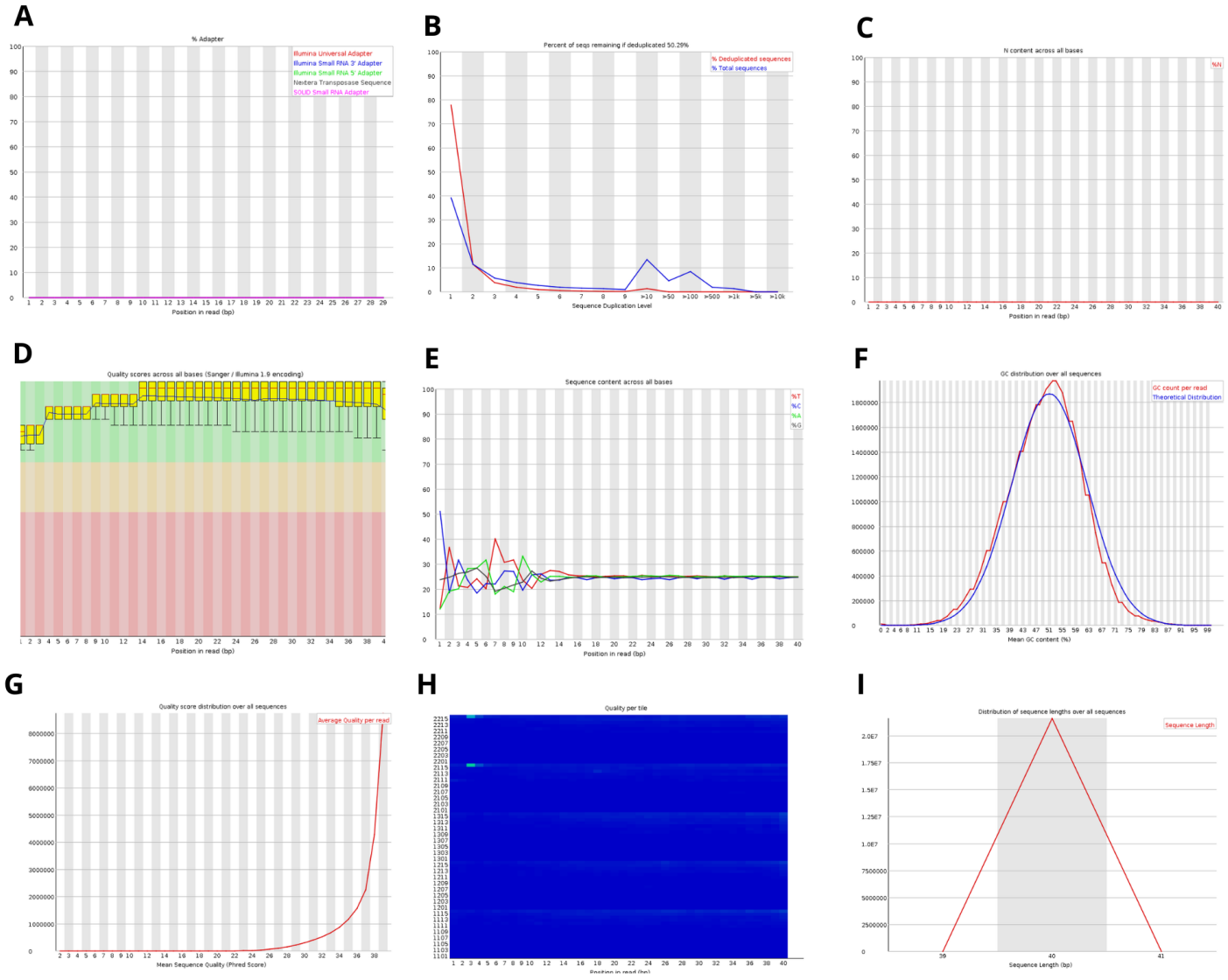

Figure 1. FASTQC Test Results for PO_1_1. FASTQC generated 9 graphs assessing the quality of the first FASTQ file, PO_1_1, against multiple parameters. Graph A displays adapter content. Graph B

displayed duplication levels. Graph C displays per base nucleotide content. Graph D displays per base quality content. Graph E displays per base sequence content. Graph F displays per sequence GC content. Graph G represents per sequence quality. Graph H displays per tile quality. Graph J displays sequence length distribution.

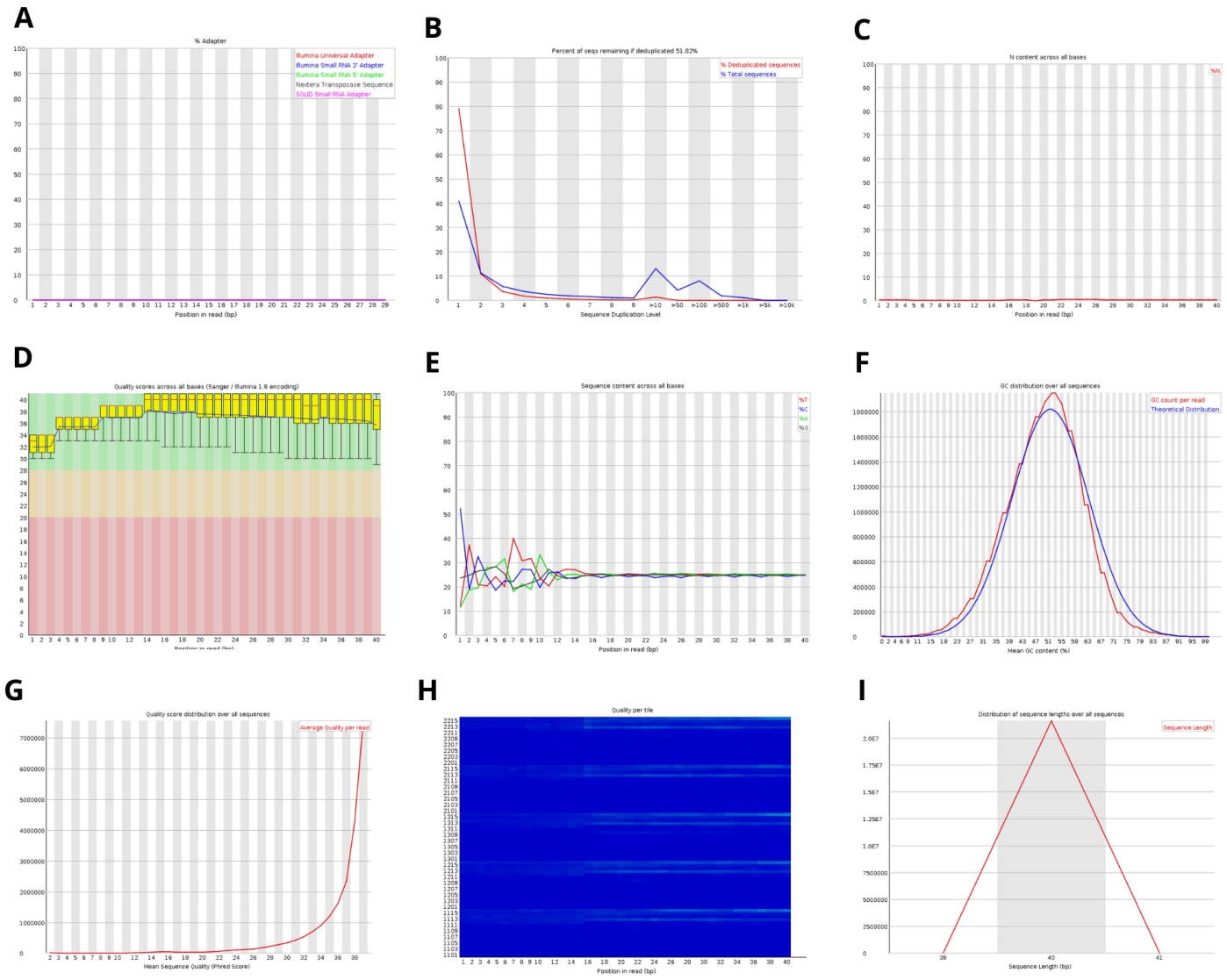

Figure 2. FASTQC Test Results for PO_1_2. FASTQC generated 9 graphs assessing the quality of the first FASTQ file, PO_1_2, against multiple parameters. Graph A displays adapter content. Graph B displayed duplication levels. Graph C displays per base nucleotide content. Graph D displays per base quality content. Graph E displays per base sequence content. Graph F displays per sequence GC content. Graph G represents per sequence quality. Graph H displays per tile quality. Graph J displays sequence length distribution.

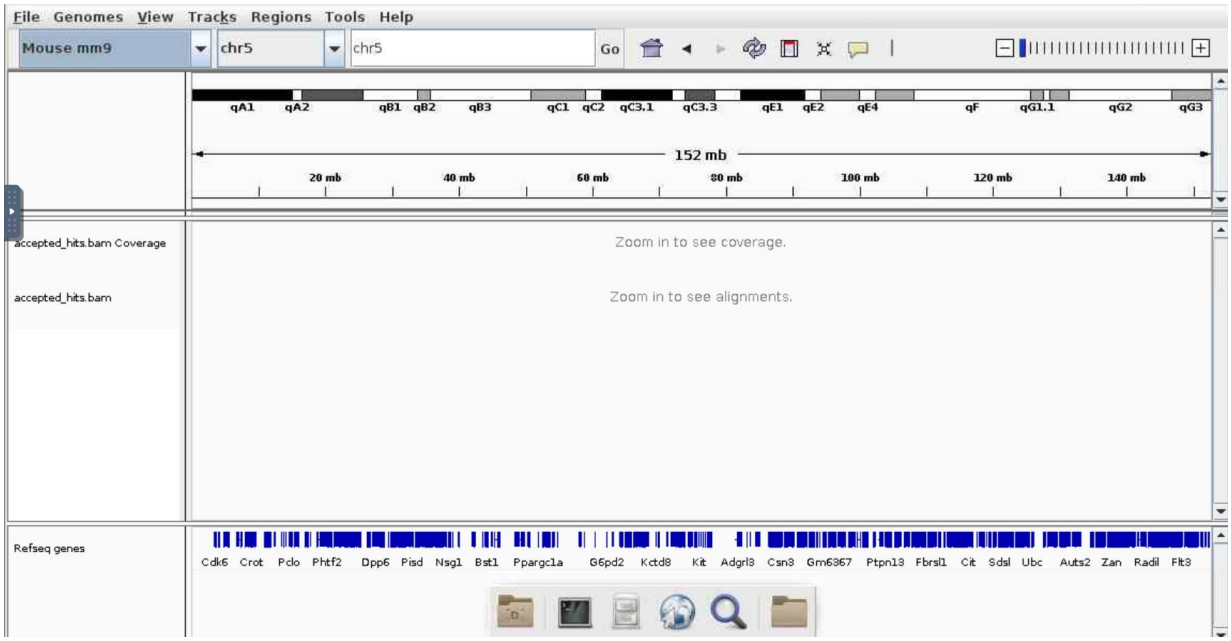

Figure 3: The graphical representation of the aligned read to the reference genome using IGV. Used the BAM file from the sample1 and the reference genome was mm9 from the mouse.

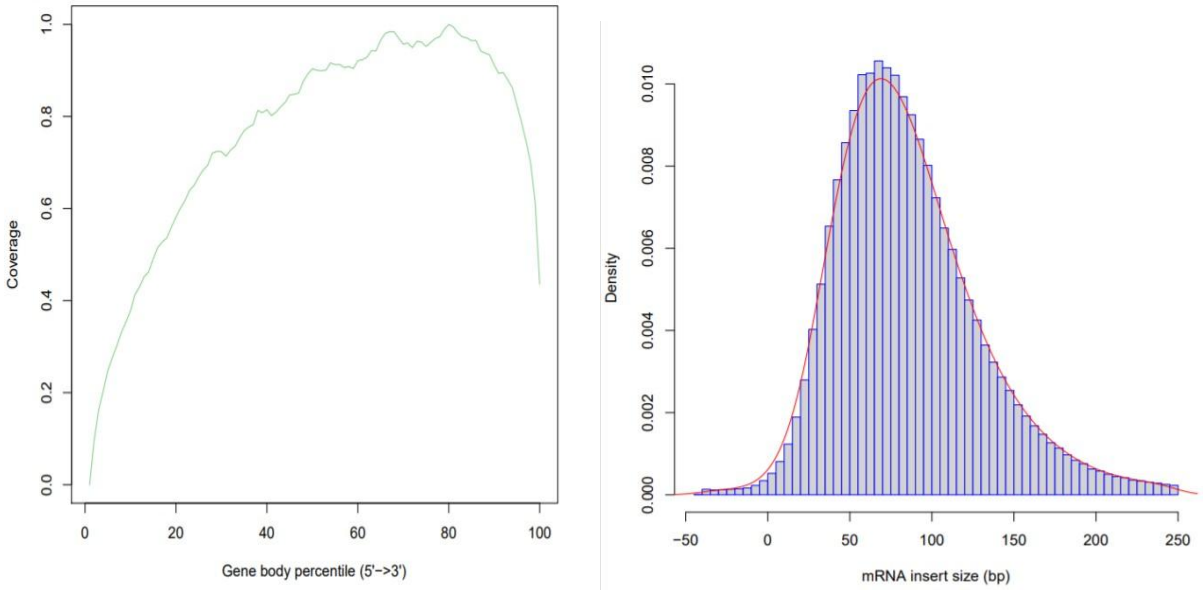

Figure 4: Analysing the data using RSeQC package: a) geneBody_Coverage helps in understanding the RNA-seq reads coverage over the gene body. According to the figure as the Gene body percentile

increases the coverage also increases but it degraded after some point. This might be due to some technical error. However, our data lies at the ideal condition where the coverage is maximum in between. i.e.(between 30-90 of gene_body percentile).b) inner_distance it helps in understanding the inner distance between read pairs. So, the statistics from this plot generated was the mean of 85.1 and standard deviation of 43.4.

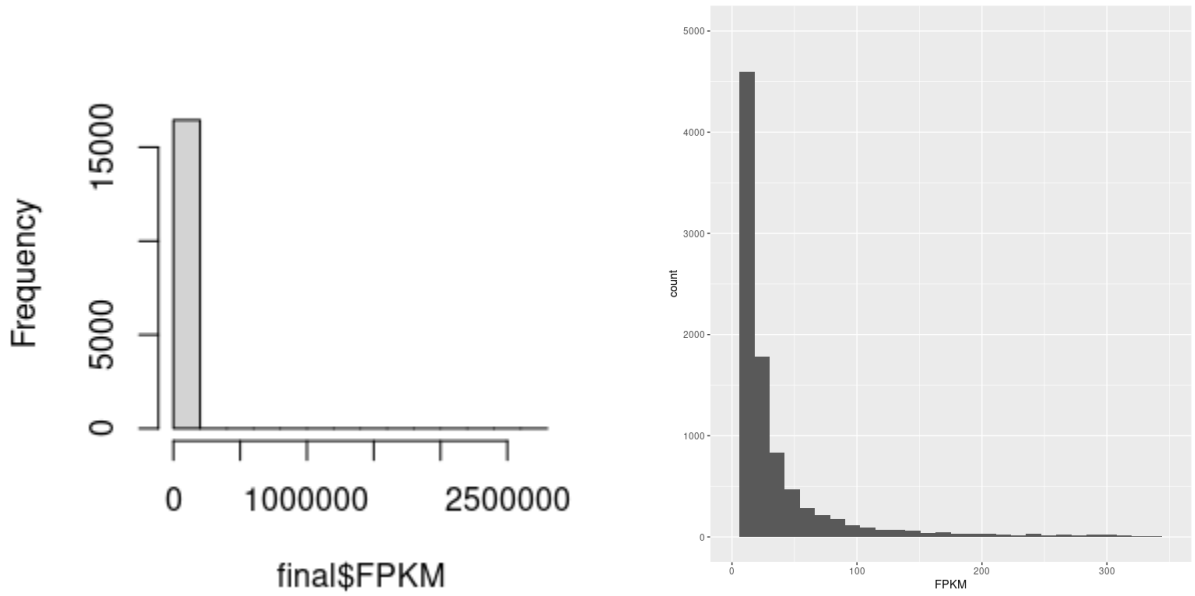

Figure 5: Highlighting the histogram generated from cufflinks output using the genes.fpkm_tracking file.

a) This plot was generated using the filtered data set. The filter was applied on the FPKM column.
b) This graph was depicting the same information as (a) but here the axis is different as compared to that. So as to visualize the data better.

Top 10 Differentially Expressed Genes - Original

gene	FPKM 1	FPKM 2	log2.FC	p_value	q_value
Plekhb2	22.568	73.568	1.705	5.00E-05	0.00106929
Mrpl30	46.455	133.038	1.518	5.00E-05	0.00106929
Coq10b	11.058	53.300	2.269	5.00E-05	0.00106929
Aox1	1.189	7.091	2.577	5.00E-05	0.00106929
Ndufb3 *	100.609	265.235	1.399	5.00E-05	0.00106929
Sp100	2.135	100.869	5.562	5.00E-05	0.00106929
Cxcr7	4.958	32.275	2.702	5.00E-05	0.00106929

Lrrfip1	118.997	24.640	-2.272	5.00E-05	0.00106929
Ramp1	13.208	0.691	-4.256	5.00E-05	0.00106929
Gpc1 *	51.206	185.329	1.856	5.00E-05	0.00106929

Table 1. Original table created of the top 10 differentially expressed genes. Genes with an asterisk were found in the supplement but at positions much greater than the top 10. Ndufb3 is found at position 413, Gpc1 is found at position 226, both found in the up-regulated genes from supplement 2 1B_CommonUP. No other genes were identified in the differentially expressed genes list.

Top 10 Differentially Expressed Genes - Modified					
gene	**FPKM 1**	**FPKM 2**	**log2.FC**	**p_value**	**q_value**
Gm2078,Mir1895	2.84	370.11	7.03	5.00E-05	0.001069
Tuba8	1.01	81.60	6.34	5.00E-05	0.001069
Rpl3l	4.66	295.66	5.99	5.00E-05	0.001069
Slc38a3 *	0.63	36.19	5.85	5.00E-05	0.001069
Csdc2	1.08	53.65	5.64	5.00E-05	0.001069
Xirp2	2.92	140.61	5.59	5.00E-05	0.001069
Sp100	2.13	100.87	5.56	5.00E-05	0.001069
Trim72 *	6.94	323.07	5.54	5.00E-05	0.001069
Bdh1	2.64	111.63	5.40	5.00E-05	0.001069
Rxrg	1.33	48.64	5.19	5.00E-05	0.001069
Tnni1	1,440.20	1.50	-9.90	5.00E-05	0.001069
Xist *	41.33	0.25	-7.36	5.00E-05	0.001069
Ptn	136.82	0.91	-7.23	5.00E-05	0.001069
H19,Mir675	1,556.82	11.39	-7.09	5.00E-05	0.001069
Myl4	248.47	2.64	-6.56	5.00E-05	0.001069
Ncrna00085	48.17	0.61	-6.30	5.00E-05	0.001069
Fbn2	18.48	0.31	-5.88	5.00E-05	0.001069
Top2a *	28.45	0.56	-5.67	5.00E-05	0.001069
Ncam1	126.64	2.57	-5.62	5.00E-05	0.001069
Col12a1	5.40	0.11	-5.57	5.00E-05	0.001069

Table 2. Modified table of top 10 differentially expressed genes. Genes with an asterisk were found in the supplement. Slc38a3 is found at position 2, Trim72 is found at position 4 in the up-regulated genes from

supplement 2 1B_CommonUP. Xist is found at position 3129, Top2a is found at position 506 in the down-regulated genes from supplement 2 1C_CommonDwn.

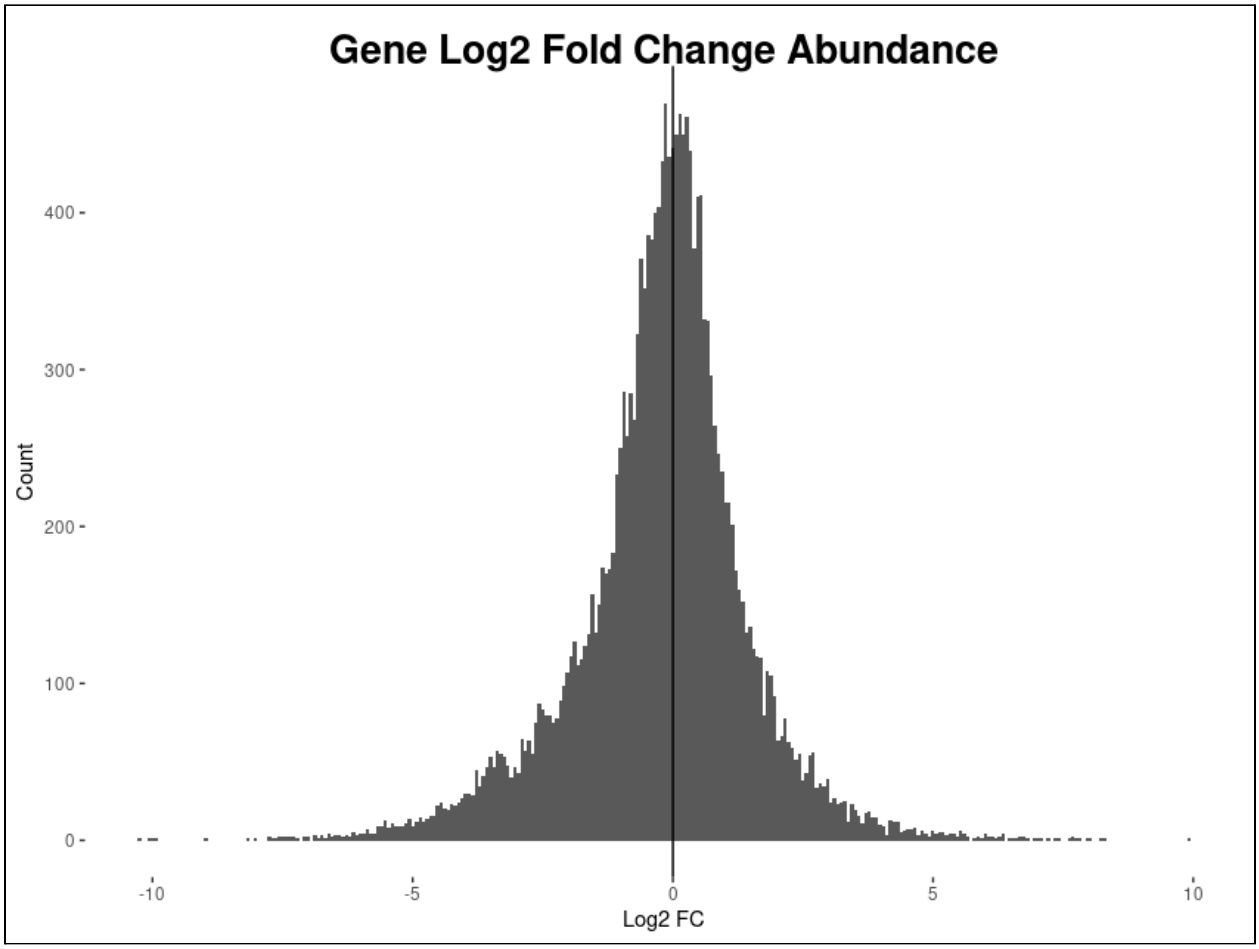

Figure 6. Histogram of all non-zero or infinite log2FC values of all genes. Log2FC of zero were in the greatest abundance, provided little information, and disrupted visualization of more informative values, therefore these have been removed. Histogram reveals most genes had insignificant log2FC as they are centered around zero. A vertical line is placed at x=0 for visualization.

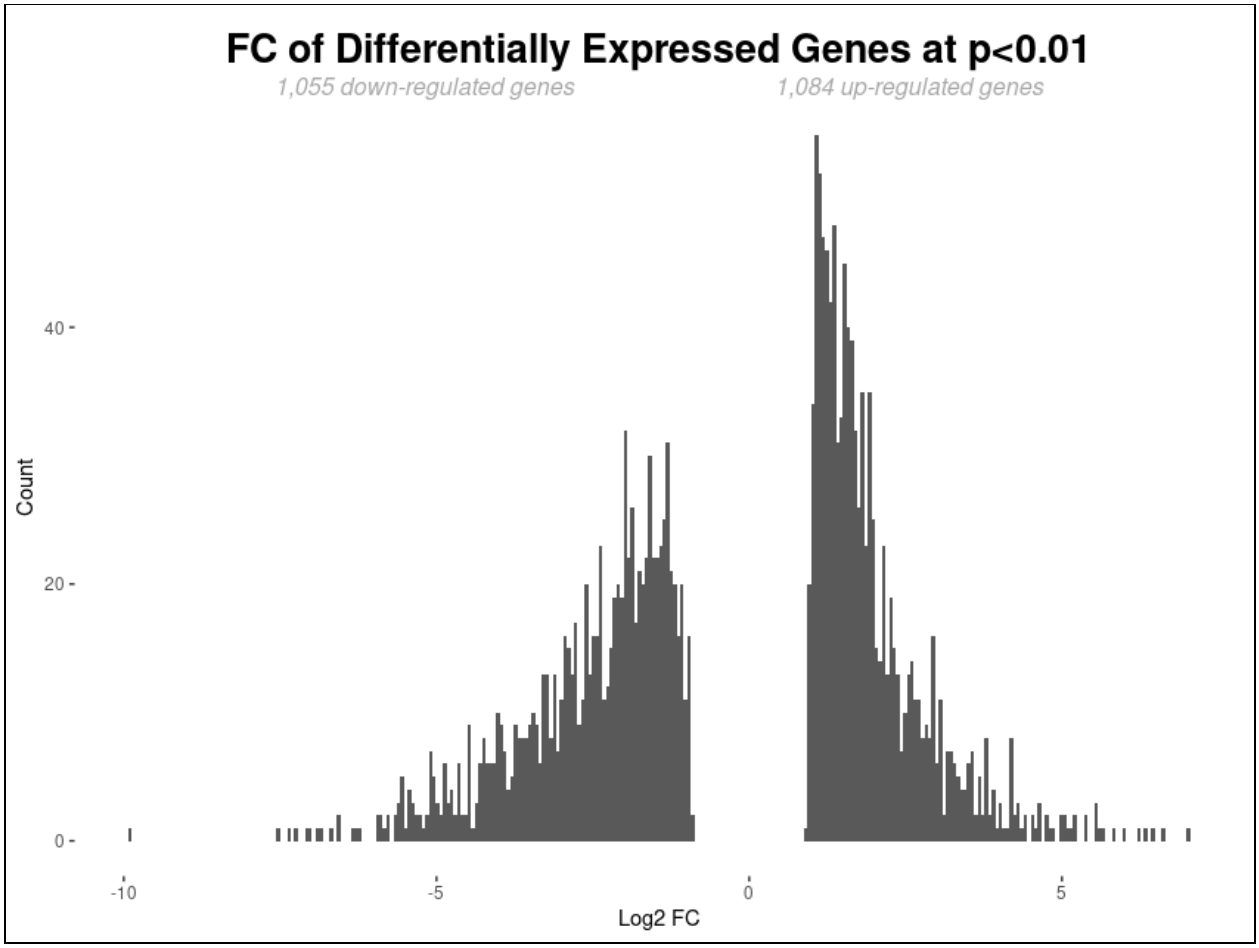

Figure 7. Significant log2FC values of all genes. Modified version of fig6. Only significant fold change values are included, this essentially removes values around zero. The histogram reveals slightly more up-regulated (1084) than down-regulated (1055) genes.

Produced by Team tinman						
Cluster	**Enrichment**	**Members**	**Top Up-Regulated Term**	**p-value**	**Gene Count**	**Benjamini**
1	21.927	23	GO:0005739~mitochondrion *	1.91E-50	263	1.32E-47
2	16.810	7	GO:0006082~organic acid metabolic process	5.01E-25	121	1.09E-21
3	15.309	33	GO:0006091~generation of precursor metabolites and energy	5.53E-27	73	1.80E-23
4	11.804	6	GO:0043230~extracellular organelle	8.75E-18	256	3.80E-16
5	7.147	6	GO:0030016~myofibril	1.31E-08	37	3.67E-07
6	6.460	11	GO:0044282~small molecule catabolic process	6.87E-10	41	8.60E-08
Cluster	**Enrichment**	**Members**	**Top Down-Regulated Term**	**p-value**	**Gene Count**	**Benjamini**
1	11.111	32	GO:0007049~cell cycle **	3.48E-30	167	2.32E-26
2	9.694	3	GO:0005578~proteinaceous extracellular matrix	1.22E-11	53	2.91E-09
3	9.580	4	GO:0008283~cell proliferation	1.63E-14	158	9.78E-12
4	8.520	4	GO:0051128~regulation of cellular component organization	2.22E-17	197	2.46E-14
5	8.025	15	GO:0009887~organ morphogenesis	2.92E-13	102	1.08E-10
6	7.816	39	GO:0007399~nervous system development	2.05E-16	187	1.71E-13

Table 3. Results from DAVID Functional Annotation Clustering tool produced by our team. Top 6 clusters have been selected for both up-regulated and down-regulated GO terms, top clusters are determined by Enrichment score. Members refer to the number of GO terms in the specified cluster, only the top GO term is provided for in the Top Term. Gene count refers to genes related to the top GO term only. Terms in common with the paper have been identified with (*, **).

Produced by Paper						
Cluster	**Enrichment**	**Members**	**Top Up-Regulated Term**	**p-value**	**Gene Count**	**Benjamini**
1	14.35	23	GO:0005739~mitochondrion *	1.71E-25	157	5.87E-23
2	8.50	6	GO:0044449~contractile fiber part	1.18E-09	22	3.71E-08
3	6.03	3	GO:0016528~sarcoplasm	1.02E-08	14	2.06E-07
4	5.00	12	GO:0015980~energy derivation by oxidation of organic compounds	2.39E-07	19	9.09E-05
5	4.39	15	GO:0006006~glucose metabolic process	4.60E-08	24	2.62E-05
6	4.11	7	GO:0005739~mitochondrion	1.65E-16	365	1.58E-14
Cluster	**Enrichment**	**Members**	**Top Down-Regulated Term**	**p-value**	**Gene Count**	**Benjamini**
1	88.90	3	GO:0043232~intracellular non-membrane-bounded organelle	2.79E-126	566	1.50E-123
2	81.30	7	GO:0031981~nuclear lumen	1.19E-119	358	3.22E-117
3	59.78	10	GO:0006396~RNA processing	3.13E-90	237	9.86E-87
4	49.89	13	GO:0007049~cell cycle **	6.47E-70	260	1.02E-66
5	41.29	6	GO:0006259~DNA metabolic process	1.07E-63	201	8.41E-61
6	36.00	8	GO:0005694~chromosome	1.20E-70	180	9.26E-69

Table 4. Paper results from DAVID. Same values chosen from Table 3 for accurate comparison, table put together from paper supplementary by team.