

Bioinformatics Reanalysis of : *Transcriptional Profile of Mammalian Cardiac Regeneration with mRNA-Seq*¹

B. Cole^{**}, T. Falk^{**}, M. Knox^{**}, and S. Pandit ^{**}

*Boston University Bioinformatics Program, Boston, MA
+Lead co-authors

Abstract- Researchers O'Meara, et. al. were attempting to evaluate which transcriptional factors were responsible for cardiac repair when there is cardiac injury. They further wanted to explore differences between neonatal mice hearts less than one-week versus older mice hearts in which cardiac function could not be fully restored [1].

Introduction

The researchers were essentially trying to answer two specific questions:

1. Why do adult mouse hearts fail to regenerate fully after injury such as heart attack opposite neonatal hearts which can fully repair?
2. Which transcriptional factors are responsible for regulating heart cell regeneration and repair?

Many newborn mammals have the ability to regenerate cardiac tissue after damage. Newborn mice can repair and completely restore their heart as well as its functionality after cardiac injury. Other species maintain this ability to repair damaged heart tissues past the early stages in their lifespan, one example being the zebrafish. However, unlike zebrafish, previous research has shown that newborn mice lose the ability to restore their hearts after one week. Bicknell et. al. [2] show that this is due to an expansion of a mature mouse cell's existing cell pool instead of continual cellular regeneration. O'Meara et. al. suggest that their work provides a "critical framework" in understanding cardiac myocyte response to injury, which will be useful in responding to heart injury in adult mammals, including humans[1].

Researchers O'Meara, et. al. reviewed myocardium within the mouse, *Mus musculus*, genome [3]. The samples included comparisons of neonatal (3, pooled) and adult hearts (2, 8-10 week old males). We used transcription data from postnatal day 0 mouse cardiac myocytes to recreate select analyses and figures from O'Meara et. al., often mentioned as the original or reference paper. They ultimately concluded that a transcriptional reversion takes place during myocyte regeneration.

Data

For this project, we utilized NIH's Gene Expression Omnibus (GEO) to download GSM1570702 dataset procured during this study [6]. During the data procurement process, we were able to discern several key foundational factors utilized for the transcriptional analysis. The Illumina HiSeq 2000 was utilized for the sequencing portion of the study in which the Trizol protocol was used [1]. The complete steps for sample preparation were as follows:

1. RNA extraction (Trizol by Invitrogen)
2. Polyadenylated RNA isolation (Dynabeads mRNA purification kit by Invitrogen)
3. Polyadenylated RNA fragmentation & synthesis (Superscript III reverse transcription kit by Invitrogen)
4. DNA synthesis (DNA polymerase I by Invitrogen)
5. Adapter to join DNA strands with end-repair, sizing and a-tail procedures (SPRI-Works System by Beckman Coulter)
6. PCR was performed to amplify and bar code sample
7. Paired-end sequencing was performed (on Illumina HiSeq 2000)
8. Sequence alignment
9. Additional protocol for lower volume specimen in neonate heart muscle cells (TruSeq by Invitrogen)

After downloading the RNA data, we then ran the dataset through a strict quality control method, FASTQC to assess the quality of the transcriptional data collected [14]. We observed 21,577,562 paired-end sequences with, read length size of 40 each. The FASTQ files were contained in 2 sets (PO_1_1.fastq and PO_1_2.fastq) due to it being a paired-end sequence. Tables 1 through 9 (Supplementary Figures & Tables section) review the quality of the reads which collectively represent the entire sequence for the *M. musculus* sample.

We see that the dataset is of high quality due to passing the FASTQC measures. Some highlights were that the quality scores across all bases (Table 1) were in the green, meaning no major quality issue across all bases [4]. Issues within this module can be indicative of instrument issues, adapter problems and/or downward slope for lengthy runs.



In addition, Table 2 showed that there were almost no instrumentation issues as the majority of the table is all blue “cold”, where red “hot” would be the opposite and indicative of a major instrument issue. Each quality per tile represents a base within the run.

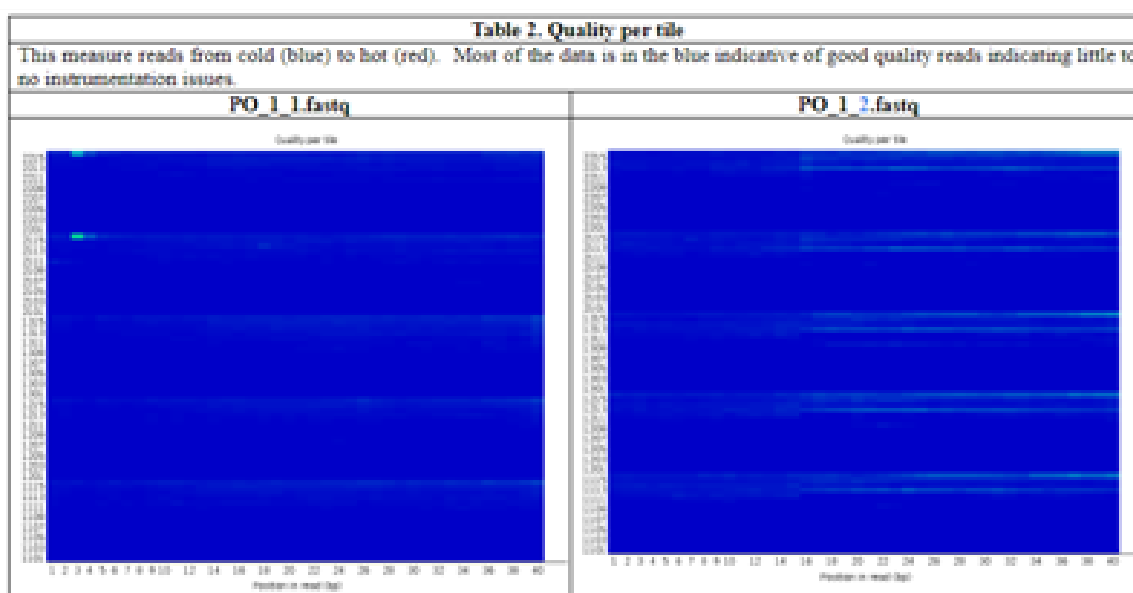


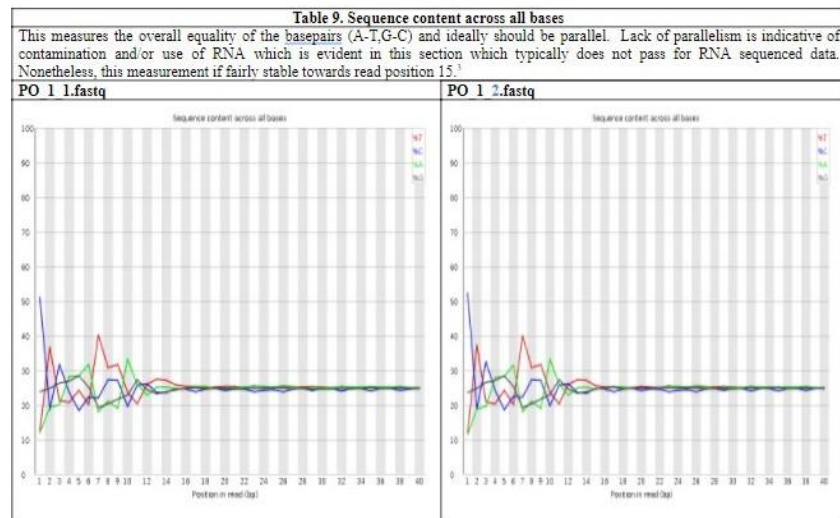
Table 3, measured the Phred score at high, about 38.5 representing high confidence in identifying nucleotides. Loss of quality within this module might come from lengthy runs which should be evaluated and possibly trimmed for down-stream data analysis.

Both, Tables 4 and 5 show passing quality scores which are indicative of low evidence of contamination.

Table 6 identifies overrepresented bases due to adapter biases and should be evaluated and possibly trimmed for down-stream data analysis. Within our sample, there was close to 0% adapter bias.

Table 8. This measure reveals how well the base pairs were called, with “N” meaning NO confidence. Our value was close to 0% revealing almost perfect confidence in calling the base pairs [5].

While over QC passed for this dataset, there was a failure for per Base Content (Table 9). Within this module, the base pair combination is evaluated and ideally should smooth out to a straight line representing about equal amount of nucleotide content. A failure could be indicative of contamination and/or use specific genomic data such as RNA. Although fairly stable around position 15, this section typically does not pass for RNA sequenced data evidenced by failure within this module [4].



Methods

For this project, we utilized NIH’s Gene Expression Omnibus (GEO) to download GSM1570702 dataset procured during this study [6].

Once the FASTQ files had been generated and analyzed for quality, they were then aligned to a *M. musculus* reference sequence. We used TopHat to align the reads from the FASTQ files to the reference genome, as had been done by the authors in the paper. We used TopHat version 2.1.1 [9], which is the version available on BU’s Shared Compute Cluster (SCC). This version of TopHat was not released until 2016, two years after the original publication date of the paper. The TopHat alignment was submitted as a batch job to SCC, running for approximately 1 hour. A quality control analysis of this alignment was then performed using the RSeQC package [10], the results of which are summarized in Figure 3.

After the FASTQ reads were aligned to the reference genome, the output BAM file was then subjected to further analysis through the CuffLinks package [8]. CuffLinks contains many functions that can be used to analyze gene expression values from RNA-seq experiments. We used CuffLinks to generate estimates of transcript abundances, as shown in Figure 4. The CuffLinks calculation was submitted as a batch job to

SCC and ran for approximately 13 minutes. We also used CuffDiff [7], a tool within the CuffLinks suite, to determine the genes that were differentially expressed between samples. Specifically, we compared gene expression values from two P0 samples to that of two Adult (Ad) samples. The CuffDiff calculation was also submitted as a batch job to SCC and ran for approximately 3.75 hours.

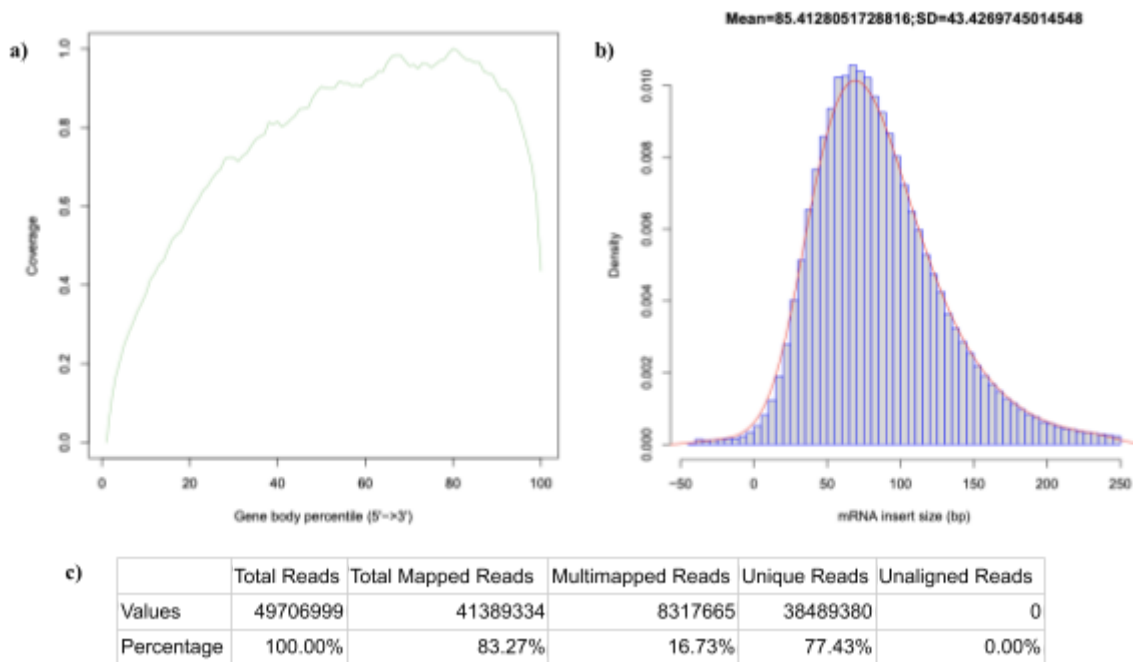


Figure 3:

a) Gene Body coverage, plotted with position within the genome (5'→3') on the x axis, and with coverage on the y axis. The coverage plot approximates a uniform distribution, indicating that alignment quality remains consistent across all positions.

b) Distribution of inner distance estimates, centered at $\mu=85.41$ and $\sigma=43.43$. The distribution of inner distances is relatively tight and approximates a normal distribution. This suggests that the distribution originates from a single population, which is consistent with the fact that this RNA-seq analysis was only performed on one sample.

c) Overall alignment statistics. These also show that the alignment was of good quality, as the number of unique reads relative to the total number of mapped reads is high, as well as having 0 unaligned reads.

With the files of CuffDiff differentially expressed genes, top expressed genes were identified by q-value, with smallest q-values indicating higher expression (Table 1). Genes were then plotted in histograms of log 2 fold change, one containing all values (Figure 5A) and the other of significant values only (Figure 5B). The significantly expressed genes were then separated into up-regulated and down-regulated genes and clustered using the DAVID Functional Annotation Clustering tool.

In order to compare differential expression levels in genes, we plotted the FPKM (Fragments Per Kilobase of exon per Million fragments) of three groups of genes: those relating to sarcomere, mitochondria, and cell cycle functions (Figure 6). Using the fragment data generated above for sample P0_1, and combining with available data for samples P0, P4, P7, and Ad, we used the 25 selected genes to subset the available FPKM data. This data was adjusted into a ggplot2 friendly format using the reshape2 package [11, 13].

Next, we annotated the DAVID results by comparing to Online Table IIA in O'Meara, et. al.. We combined the up and down regulated gene ontology (GO) terms from our above analysis, and selected the top 25 terms based on their fold enrichment score. We then merged the terms from Online Table IIA to compare the overlap of terms in our analysis and O'Meara, et. al. (Table 2).

We created a heatmap of the FPKM values of the 100 most differentially expressed genes, as described above. First we merged the FPKM data from all eight samples (sample 1 and 2 for P0, P4, P7, and Ad),

and subset these values by selecting the 100 most differentially expressed genes between P0 and Ad with the lowest q-value. This FPKM matrix was plotted using `gplots heatmap.2()` function [12]. A second heatmap was created (Figure S10) in order to compare the similarity among different mouse samples.

All of the scripts used in this analysis can be found on our GitHub page, <https://github.com/BF528/project-2-hedgehog>.

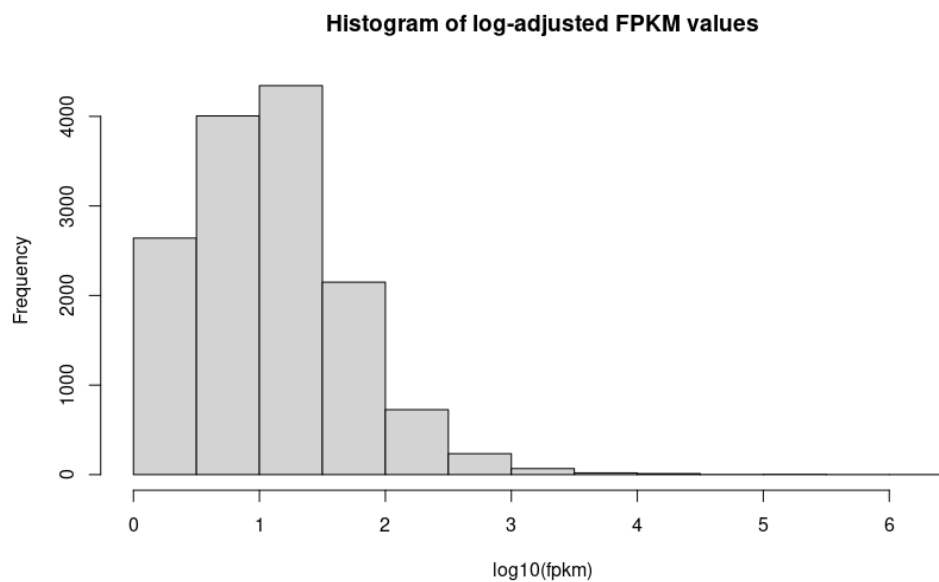


Figure 4: Histogram of transcript abundances, only included genes with FPKM values ≥ 1 , with the x-axis being on the \log_{10} scale. There were 37,469 genes in the original data, and after applying the filter there were 14,205 genes left. The left-skewed distribution of log-adjusted FPKM values are to be expected; many genes have low expression levels, but there are few genes that have high levels.

Results

	gene_id	gene	value_1	value_2	log2.fold_change.	p_value	q_value
106	XLOC_000106	Plekha2	22.56790	73.568300	1.70481	5e-05	0.0010693
127	XLOC_000127	Mrpl30	46.45470	133.038000	1.51794	5e-05	0.0010693
199	XLOC_000199	Coq10b	11.05830	53.300000	2.26901	5e-05	0.0010693
214	XLOC_000214	Aox1	1.18858	7.091360	2.57682	5e-05	0.0010693
221	XLOC_000221	Ndufb3	100.60900	265.235000	1.39851	5e-05	0.0010693
398	XLOC_000398	Sp100	2.13489	100.869000	5.56218	5e-05	0.0010693
454	XLOC_000454	Cxcr7	4.95844	32.275300	2.70247	5e-05	0.0010693
459	XLOC_000459	Lrrfip1	118.99700	24.640200	-2.27184	5e-05	0.0010693
461	XLOC_000461	Ramp1	13.20760	0.691287	-4.25594	5e-05	0.0010693
477	XLOC_000477	Gpc1	51.20620	185.329000	1.85570	5e-05	0.0010693

Table 1 List of the top 10 differentially expressed genes and their FPKM values, log fold change, p-values, and q-values.

Table 1 shows the top 10 differentially expressed genes, sorted by q-value ascending, and includes identifying information, FPKM values, log 2 fold change, p-value, and q-value. Two histograms of the log 2 fold change were produced using data from a differential expression analysis of postnatal day 0 vs adult genes. The first histogram (Figure 5A) shows a normal distribution of the fold change for all genes,

while the second histogram (Figure 5B) has a bimodal distribution when only significant genes are selected.

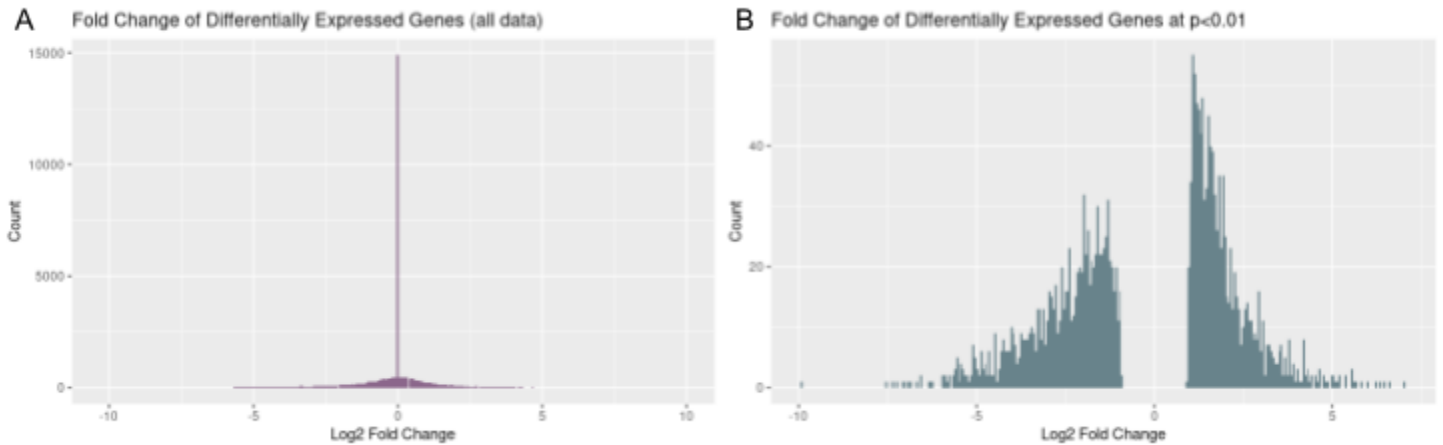


Figure 5 A) Histogram of log 2 fold changes for all genes. A) Histogram of log 2 fold changes for only significant genes. There are a total of 1084 significant upregulated genes and 1055 significant down regulated genes.

Both up- and down-regulated genes were clustered using DAVID Functional Annotation Clustering. The 1055 significantly differentially expressed, down-regulated genes (Table S1) were clustered into 420 functional clusters, while the 1084 significantly differentially expressed, up-regulated genes (Table S2) were clustered into 411 functional clusters.

Our Replicated Data								O'Meara et. al. data		
GO Term	GO Name	GO Category	p-value	Fold Enrichment	Bonferroni	Benjamini	Regulated	GO Name	Fold Enrichment	Benjamini
GO:0005739	mitochondrion	GOTERM_CC_FAT	1.91E-50	2.52	1.32E-47	1.32E-47	up	Mitochondrion	1.66	9.30E-06
GO:0005739	mitochondrion	GOTERM_CC_FAT	1.91E-50	2.52	1.32E-47	1.32E-47	up	Mitochondrion	2.02	1.30E-07
GO:0005739	mitochondrion	GOTERM_CC_FAT	1.91E-50	2.52	1.32E-47	1.32E-47	up	Mitochondrion	3.52	2.40E-77
GO:0044429	mitochondrial part	GOTERM_CC_FAT	6.52E-45	3.28	4.52E-42	2.26E-42	up			
GO:0005740	mitochondrial envelope	GOTERM_CC_FAT	2.77E-32	3.24	1.92E-29	4.73E-30	up			
GO:0005743	mitochondrial inner membrane	GOTERM_CC_FAT	2.80E-32	3.98	1.95E-29	4.73E-30	up	Mitochondrial inner membrane	3.07	6.50E-05
GO:0031966	mitochondrial membrane	GOTERM_CC_FAT	3.41E-32	3.34	2.37E-29	4.73E-30	up			
GO:0019866	organelle inner membrane	GOTERM_CC_FAT	2.32E-29	3.61	1.61E-26	2.68E-27	up			
GO:0007049	cell cycle	GOTERM_BP_FAT	3.48E-30	2.54	2.32E-26	2.32E-26	down	Cell cycle	2.5	7.40E-31
GO:0044455	mitochondrial membrane part	GOTERM_CC_FAT	1.37E-26	5.19	9.50E-24	1.36E-24	up			
GO:0098798	mitochondrial protein complex	GOTERM_CC_FAT	1.93E-26	5.77	1.34E-23	1.67E-24	up			
GO:1990204	oxidoreductase complex	GOTERM_CC_FAT	7.62E-26	7.32	5.29E-23	5.87E-24	up			
GO:0031967	organelle envelope	GOTERM_CC_FAT	1.57E-25	2.43	1.09E-22	1.09E-23	up			
GO:0031975	envelope	GOTERM_CC_FAT	2.55E-25	2.42	1.77E-22	1.61E-23	up	Envelope	1.88	3.10E-02
GO:0006091	generation of precursor metabolites and energy	GOTERM_BP_FAT	5.53E-27	4.41	3.60E-23	1.80E-23	up			
GO:0051301	cell division	GOTERM_BP_FAT	1.44E-26	3.59	9.58E-23	4.79E-23	down			
GO:0000278	mitotic cell cycle	GOTERM_BP_FAT	1.69E-25	3.03	1.12E-21	3.75E-22	down			
GO:0022402	cell cycle process	GOTERM_BP_FAT	3.95E-25	2.62	2.63E-21	6.57E-22	down			
GO:0006082	organic acid metabolic process	GOTERM_BP_FAT	5.01E-25	2.78	3.26E-21	1.09E-21	up			
GO:1903047	mitotic cell cycle process	GOTERM_BP_FAT	2.63E-24	3.12	1.75E-20	3.51E-21	down			
GO:0043436	oxoacid metabolic process	GOTERM_BP_FAT	2.23E-24	2.86	1.45E-20	3.63E-21	up			
GO:0019752	carboxylic acid metabolic process	GOTERM_BP_FAT	4.11E-24	2.86	2.67E-20	5.35E-21	up			
GO:0098800	inner mitochondrial membrane protein complex	GOTERM_CC_FAT	9.53E-22	5.96	6.61E-19	5.51E-20	up			
GO:0070469	respiratory chain	GOTERM_CC_FAT	5.54E-21	7.04	3.84E-18	2.96E-19	up	Respiratory chain	4.51	2.50E-02
GO:0015980	energy derivation by oxidation of organic compounds	GOTERM_BP_FAT	1.62E-21	4.65	1.05E-17	1.75E-18	up			

Table 2 An annotated list of up and down regulated gene ontology terms found. Terms found in common with the O'Meara paper are denoted (from Online Table IIA).

Table 2 compares the results of our DAVID analysis with those from O’Meara et. al.’s Online Figure IIA. 7 of the top 25 gene ontology terms, ordered by increasing p-value, were found to overlap with the original work. It was difficult to compare to the original paper’s gene ontology terms as they were only included by name, and not by ID. This might have contributed to the lack of correlation between gene ontology terms, along with any upstream issues in our analysis or some of the issues highlighted by the FASTQC quality control. We can also compare this table to original paper Figure 1C, which lists some common gene ontology terms between up and down regulated genes. In this case, we do see some strong overlap between terms related to mitochondria and respiration in the up-regulated genes, and the cell cycle terms in the down regulated genes.

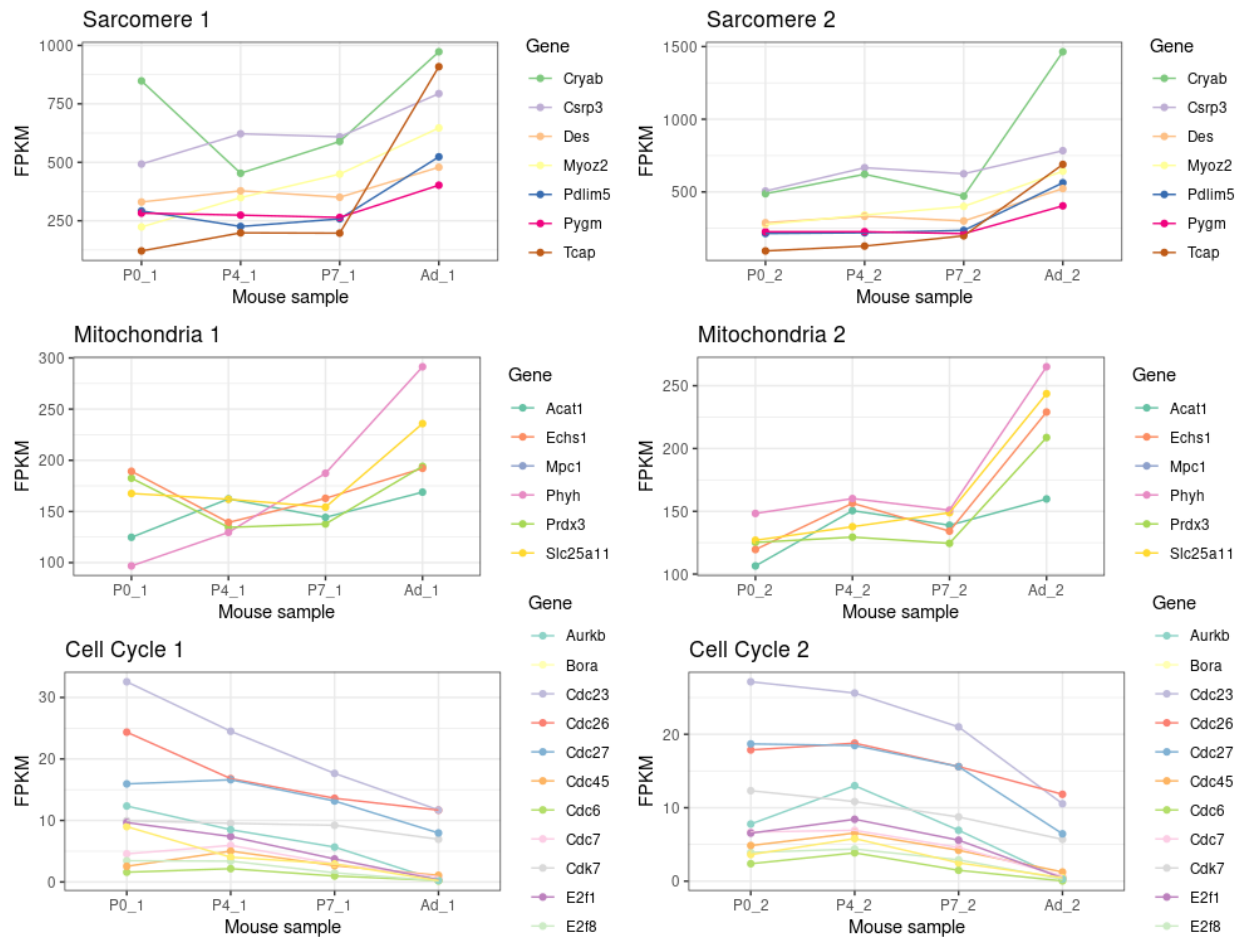


Figure 6 FPKM values plotted for three select groups of genes representing sarcomere, mitochondria, and the cell cycle. Each column represents one set of replicates.

In Figure 6 we include our reproduced example, P0 replicate 1, with the FPKM counts of the other 8 provided samples. We see a strong correlation with the original paper’s trends for this figure, O’Meara et. al. Figure 1D. Finally, we generated two heatmaps to compare the FPKM values associated with the top 100 differentially expressed genes.

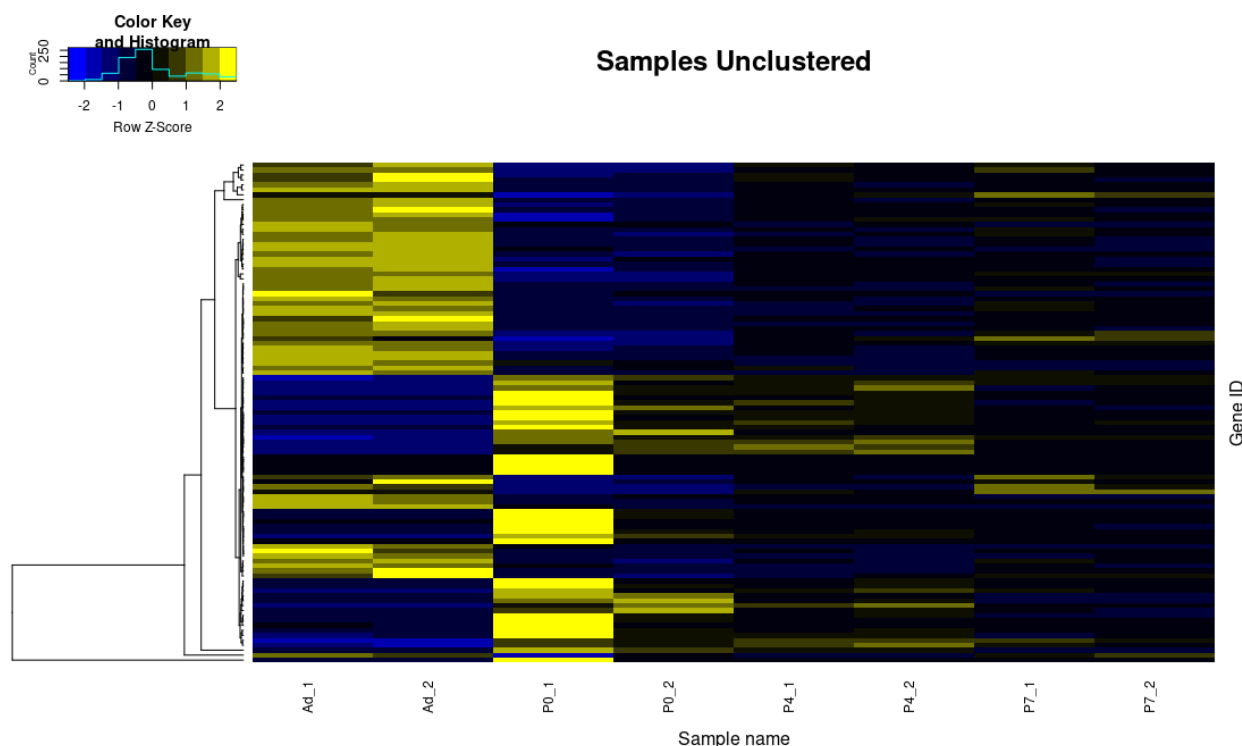


Figure 7: Heatmap of FPKM values with subset by the top 100 differentially expressed genes found in our P0 vs Ad analysis. Samples columns here are unclustered to preserved comparisons between sample replicates.

Our first heatmap, Figure 7, shows that genes with high FPKM counts in P0_1 have lower FPKM counts in the Ad columns. There is also correlation between our reproduced sample, P0_1, and the provided sample data at the same age cohort, P0_2. Concerningly, the second heatmap (Figure S10), indicates that sample P0_1 is the furthest removed from the available sample data. This dissimilarity likely stems from the earlier issues we encountered while processing this sample, and while there is similarity in which genes have high and low FPKM values, the magnitude of those values may not be consistent with the remaining samples.

Discussion

We sought to replicate the results in O'Meara et. al. using their described methods and the available RNASeq data. Of the multiple experiments conducted (in vivo, in vitro, explant), we focused on replicating the *in vivo* heart maturation. Specifically, we compared cardiac myocyte transcription levels between mice heart tissues resected on postnatal day 0 (P0) and adult (Ad) mice. In our analysis, we performed the following steps: analyzed FASTQ files generated by the original experiment, and performed a quality control analysis on these files; aligned the reads to a reference genome and determined gene expression levels; compared P0 and Ad samples to determine genes that were differentially expressed between the two life stages; and recreated some of the figures from O'Meara et. al.

After downloading the data, we conducted a quality control measure utilizing FASTQC to look for possible sources of contamination or bias. Our basic statistics passed with no flags for quality. Only one module received a failure due to the use of RNA data for the Per base sequence content (Table 9), which measures equality of G-C, T-A content. Due to there being close to 0% adapter bias (Table 6), trimming was not necessary to further analyze the data.

We then used TopHat to align the FASTQ files to a reference genome. Quality control analysis of the alignment confirmed that the resulting BAM file was valid. Transcript abundance and differential gene expression were determined using CuffLinks and CuffDiff, respectively. The results of these computations were then handed off for DAVID analysis.

CuffDiff data was filtered for significance and separated into up- and down-regulated gene sets and clustered by function. The up-regulated genes resulted in 411 functional clusters, the top 4 having enrichment scores of between 11.8 and 21.93 (Table S2). The down-regulated genes resulted in 420 clusters with enrichment scores between 8.52 and 11.11 (Table S1). The enrichment score is the rank of the biological significance of the gene group via the geometric mean of the p-values [15, 16]. Compared to O'Meara et al., our enrichment scores for the down-regulated genes are very far off (e.g. for Cluster 1, our enrichment score was 11.11 while theirs was 21.00; for Cluster 2, our score was 9.6 while theirs was 16.22). The enrichment scores for the up-regulated genes were still off target, but not quite so dramatically (e.g. looking at the same clusters, for Cluster 1 we had a score of 21.93 while they had a score of 27.81; Cluster 2, we had 16.81 while they had 18.96). Possible explanations for the difference in enrichment scores could include a difference in filtering the data for significance, or if O'Meara filtered their data based on something other than the log2 fold change for up- and down-regulated genes. Differences in what genes were selected as differentially expressed and significant could also have affected the enrichment scores as the clustered groups may not have included the same genes between our results and the original results.

Our main goal was to recreate the methods and findings of O'Meara et. al., namely that there is a transcriptional reversion observed in injured heart cells depending on the age of the mouse. The key areas that we sought to identify were the following: alterations in genes regulating sarcomeres, RNA processing, and the cell cycle. In some cases, we were able to replicate the paper's original findings with relative accuracy, such as our Figure 6 that mirrors figure 1D in the paper. In other cases, such as Table 2, there is only some overlap with the results of O'Meara et. al. Inaccuracies here are likely two-fold: inadequacies in the original methods further lead to inadequacies in our attempts at faithful reproduction. The lack of explicit code and specific software versions likely contribute to these inadequacies, such as the version of TopHat used.

In terms of the biological interpretation, some of our results align neatly to the reasoning originally offered in O'Meara et. al.. FPKM values in Figure 6 align with the trend established in the original paper, with sarcomere and mitochondria genes having more fragments as the mouse ages, and cell cycle genes having fewer fragments over the same period. However, there is an interesting difference in the heatmaps originally presented (O'Meara et. al. Figure 2A) and the one we attempted to recreate (Figure 7 and Figure S10).

One issue we found was the selected gene ontology terms represented were not cataloged in any supplemental documentation, and we were left to compare what terms were made available using gene ontology term names instead of IDs. One can also see the original Figure 2A features some difference of expression between the "Adult" column and the "P0" column, the comparison we sought to recreate.

The problem is that the granularity of ontology terms represented by each row is severely lacking, making it difficult to indicate any trends. By contrast, our heatmap shows distinct differences between adult and P0 levels. Not only is there a clearer distinction, those genes that have high FPKM values in adults have lower values in P0 mice, and vice versa, potentially painting a clearer indication of the transcriptional reversion the paper concluded was taking place in these myocytes.

We observed a difference between the enriched GO terms that our analysis yielded versus those presented in the paper. However, upon closer inspection we found that the gene ontology analysis performed in the paper was only done on a select few clusters of genes as identified in Figure 2A. Thus, while our results may be slightly different from those arrived at by the authors of the paper, this may be because the underlying data used to find gene ontology enrichment was fundamentally different between the two analyses. We were unable to tell what the true number of uniquely regulated genes were for each set from the paper or its figures. The numbers in common were made clear, however we were unable to determine if the number listed in O'Meara et al.'s Figure 1B on the venn diagram was for all genes up-regulated either *in vivo* or *in vitro*, or if it was the number of genes unique. Differences in differentially expressed genes between our results and O'Meara et al.'s would potentially explain the difference further down the process in the GO results.

One area of the original study we were unable to replicate was the variety of types of mouse cardiac myocyte tissue researchers in O'Meara et. al. were able to analyze. Namely, *in vitro*, *in vivo*, and explant. In this case, the authors likely introduced a number of different sample environments in order to control for the gene expression levels they were examining. If only *in vivo* heart tissue was examined, for instance, it might be particular to that environment rather than a consistent phenomena in mouse hearts. Researchers might also want to study expression levels in these different environments to increase the breadth of applicability this study has to further research. While we only compared a subset of their total data set, we did compare two fundamentally different cohorts of mouse cardiac myocytes. The researchers sought to compare myocytes capable of regeneration (neonates) to those incapable of regeneration (adults). Since our data also encompasses these two areas, some degree of agreement among results is not unexpected.

While we did find some agreement to O'Meara et. al., there was difficulty in determining what correlation we had truly achieved. As has been already discussed, the heatmap labels some areas with their associated ontology terms, but it is not clear which row corresponds to which term. A better method to compare results might be to compare our gene expression levels to the supplemental data, such as those in O'Meara's et. al.'s Online Table II, in order to compare exact numerical results. This would allow us to compare the supplemental results to ours on a statistical level, instead of relying on largely qualitative differences in heatmaps or line graphs.

Conclusion

We ultimately generated mixed results in our effort to replicate the methods from "Transcriptional Reversion of Cardiac Myocyte Fate During Mammalian Cardiac Regeneration." While some of our results align nicely with those of the original document, such as Figure 6, we encountered a number of issues and are not entirely confident with the degree of similarity of our results present. A more descriptive and in-depth discussion of methods, along with code used, would enable more concrete replication of this

novel study of mouse cardiac myocytes. Efforts to clarify results were also hindered by inconsistent supplemental information.

One specific issue that we ran into involves the coordination of results between roles. Due to the interdependence of each role's analysis on that of another role, downstream analyses suffer disproportionately from any upstream delays. For example, the results from the CuffDiff analysis is used as the basis of the results produced by subsequent analyses. This specific workflow gave us some trouble, as we ran into technical difficulties in generating the differential expression analysis. This led to confusion regarding the validity of the results of subsequent analyses. The way that we solved this problem was through clear communication about what specifically went wrong in our analysis. This gave us clear goals for effective troubleshooting.

References

1. O'Meara, Caitlin C., et al. "Transcriptional Reversion of Cardiac Myocyte Fate During Mammalian Cardiac Regeneration." *Circulation Research*, vol. 116, no. 5, Feb. 2015, pp. 804–15. DOI.org (Crossref), doi:10.1161/CIRCRESAHA.116.304269.
2. Bicknell, Katrina A., et al. "Can the Cardiomyocyte Cell Cycle Be Reprogrammed?" *Journal of Molecular and Cellular Cardiology*, vol. 42, no. 4, Apr. 2007, pp. 706–21. PubMed, doi:10.1016/j.yjmcc.2007.01.006.
3. *Taxonomy Browser (Mus Musculus)*.
<https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=10090>. Accessed 4 Mar. 2021.
4. The Eukaryotic Pathogen, Vector and Host Informatics Resource (VEuPathDB) Workshop: BoP bioinformatics module. FASTQC https://workshop.eupathdb.org/bop/pdfs/fastqc_output.pdf (accessed, March 2 2021)
5. Labadorf, Adam. "Sequence Analysis Fundamentals." BF528 - Applications in Translational Bioinformatics- (12, Feb 2021), Boston University, Boston, MA. Lecture
6. *GEO Accession Viewer*. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1570702>. Accessed 4 Mar. 2021.
7. Trapnell, Cole, David G. Hendrickson, et al. "Differential Analysis of Gene Regulation at Transcript Resolution with RNA-Seq." *Nature Biotechnology*, vol. 31, no. 1, 1, Nature Publishing Group, Jan. 2013, pp. 46–53. www.nature.com, doi:10.1038/nbt.2450.
8. Trapnell, Cole, et al. "Transcript Assembly and Quantification by RNA-Seq Reveals Unannotated Transcripts and Isoform Switching during Cell Differentiation." *Nature Biotechnology*, vol. 28, no. 5, May 2010, pp. 511–15, doi:10.1038/nbt.1621.
9. Trapnell, Cole, Lior Pachter, et al. "TopHat: Discovering Splice Junctions with RNA-Seq." *Bioinformatics (Oxford, England)*, vol. 25, no. 9, May 2009, pp. 1105–11. PubMed, doi:10.1093/bioinformatics/btp120.
10. Wang, Liguang, et al. "RSeQC: Quality Control of RNA-Seq Experiments." *Bioinformatics (Oxford, England)*, vol. 28, no. 16, Aug. 2012, pp. 2184–85. PubMed, doi:10.1093/bioinformatics/bts356.
11. Wickham, Hadley. *Ggplot2. Elegant Elegant Graphics for Data Analysis*. 2nd ed. Springer International Publishing, 2016. DOI.org (Crossref), doi:10.1007/978-3-319-24277-4.
12. Warnes, Gregory R., et al. *Gplots: Various R Programming Tools for Plotting Data*. 3.1.1, 2020. R-Packages, <https://CRAN.R-project.org/package=gplots>.
13. Wickham, Hadley. *Reshape2: Flexibly Reshape Data: A Reboot of the Reshape Package*. 1.4.4, 2020. R-Packages, <https://CRAN.R-project.org/package=reshape2>.
14. Babraham Bioinformatics - FastQC A Quality Control Tool for High Throughput Sequence Data. <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>. Accessed 4 Mar. 2021.

15. Huang, Da Wei et al. "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources." *Nature protocols* vol. 4,1 (2009): 44-57. doi:10.1038/nprot.2008.211
16. Huang, Da Wei et al. "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists." *Nucleic acids research* vol. 37,1 (2009): 1-13. doi:10.1093/nar/gkn923

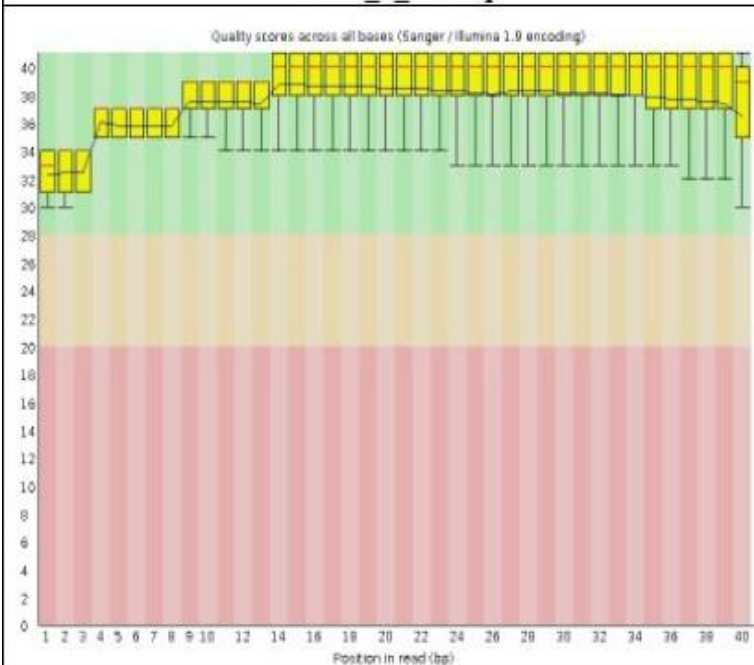
Supplementary Figures and Tables

FASTQC (Quality Results)^{4,5,14:}

Table 1. Quality scores across all bases

Overall quality is in the green representing good quality overall with the median value of 40 read for 25-75% (denoted in yellow) of the base-pair (bp) reads. Even so, the lowest quality score for both pairs is at 30, considered good (green).

PO_1_1.fastq



PO_1_2.fastq

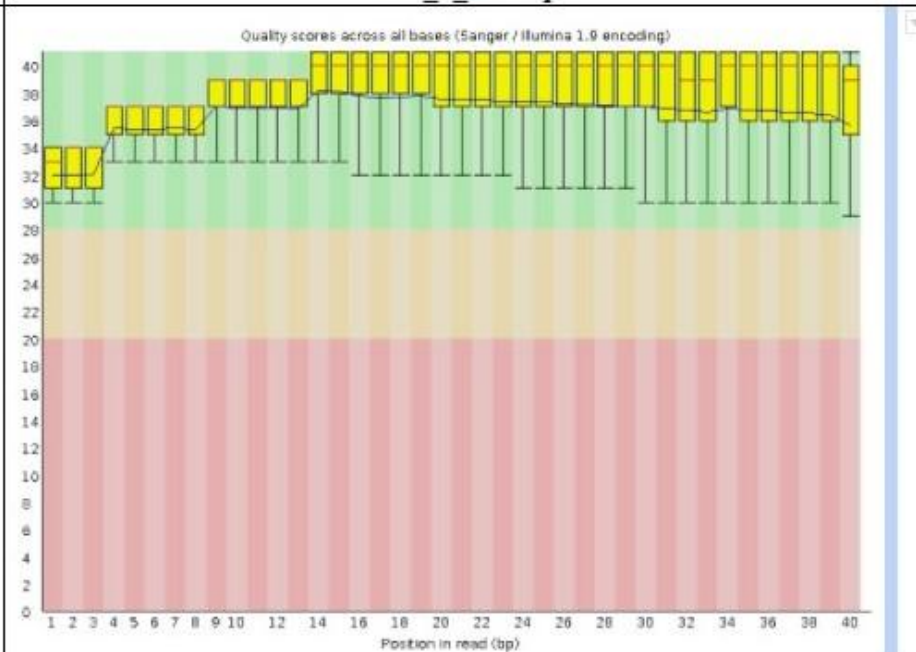
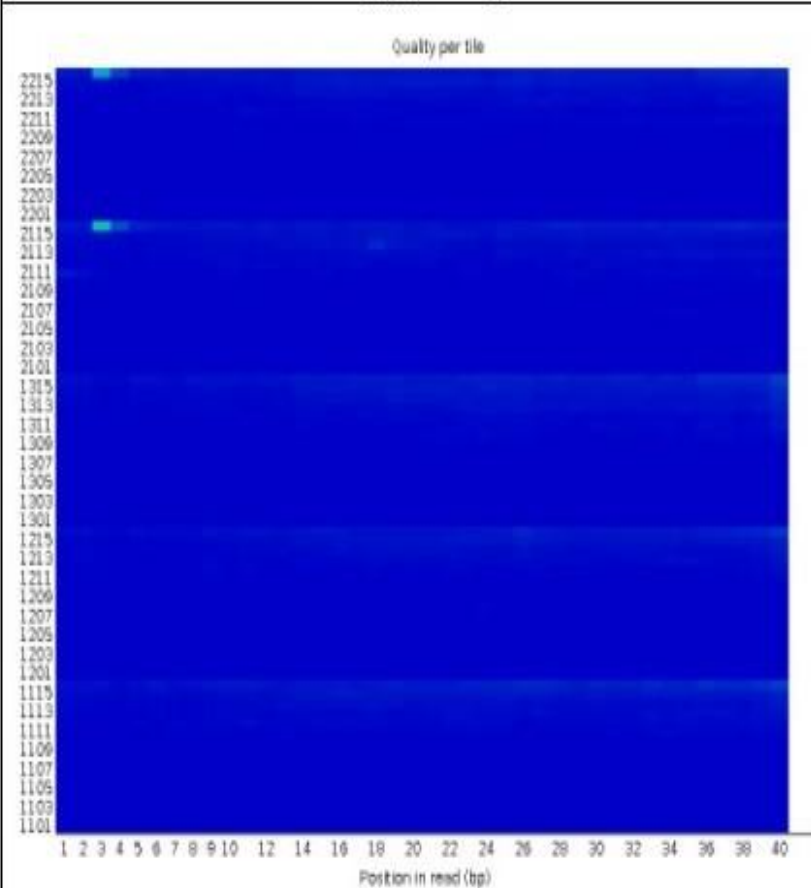
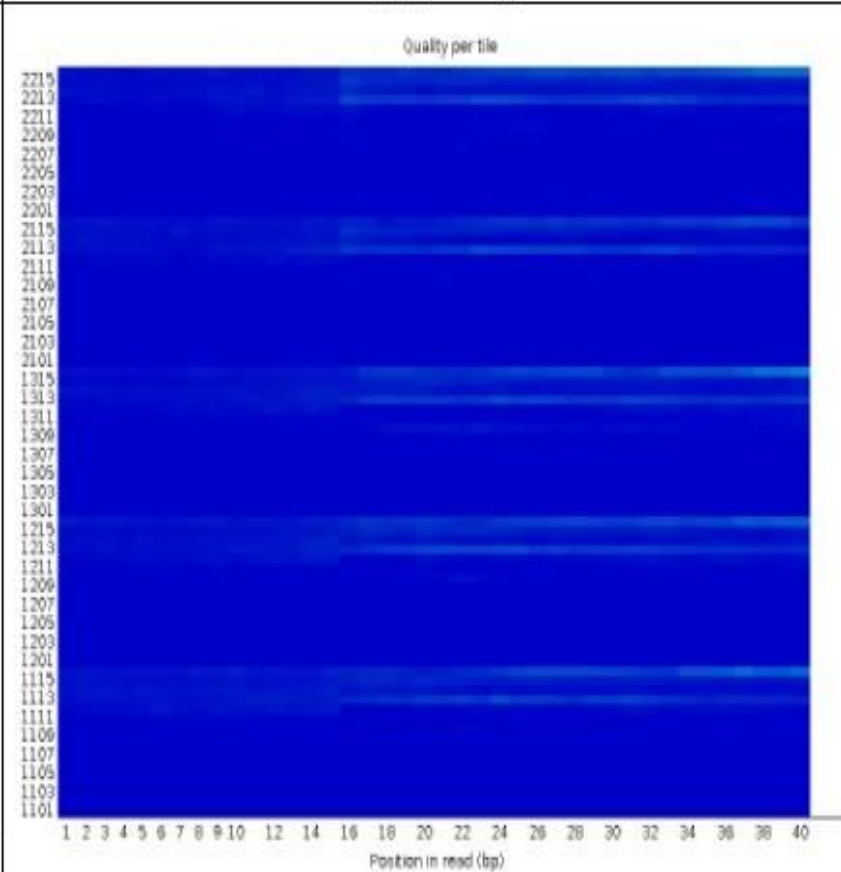


Table 2. Quality per tile

This measure reads from cold (blue) to hot (red). Most of the data is in the blue indicative of good quality reads indicating little to no instrumentation issues.

PO_1_1.fastq**PO_1_2.fastq**

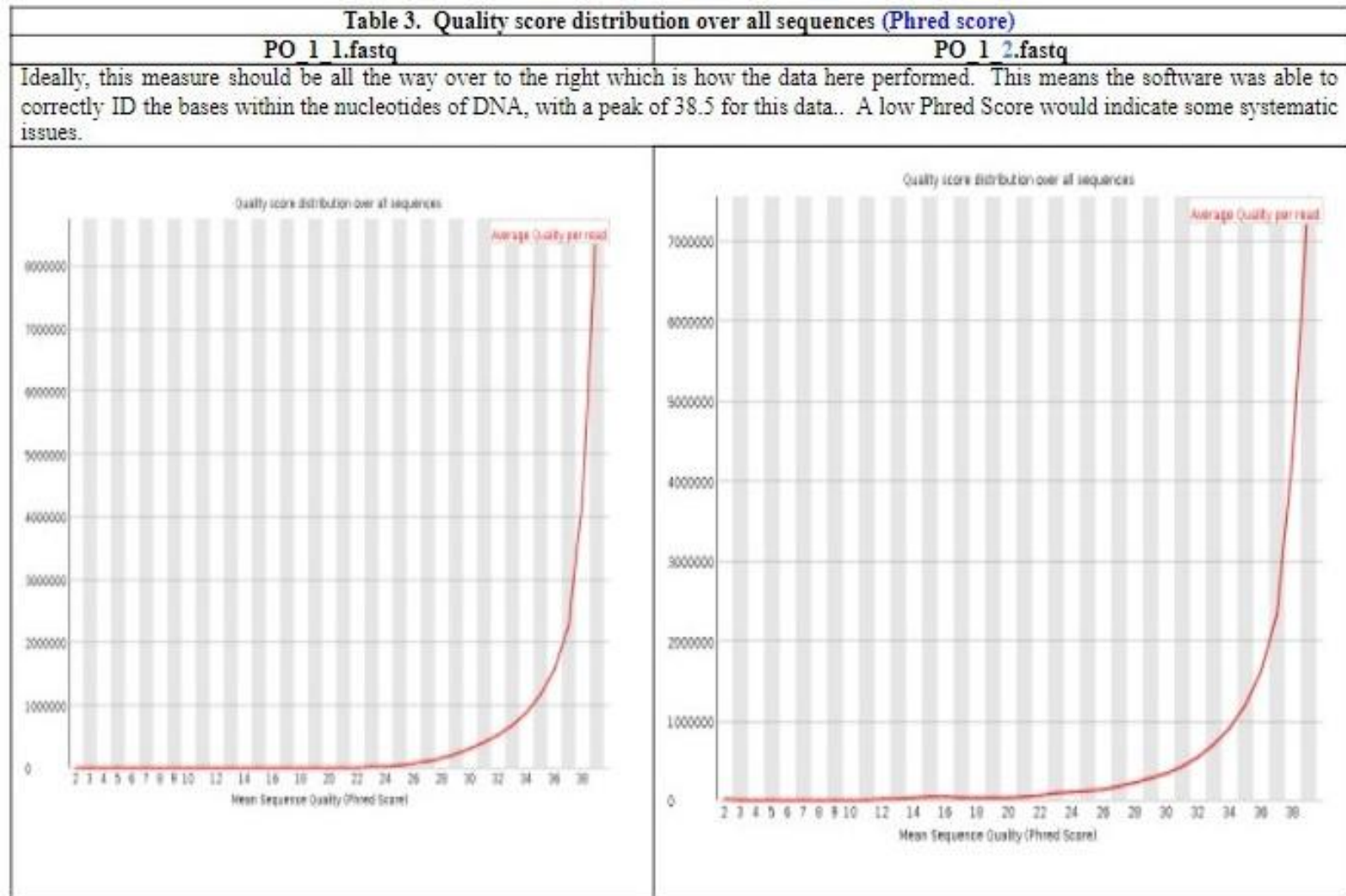
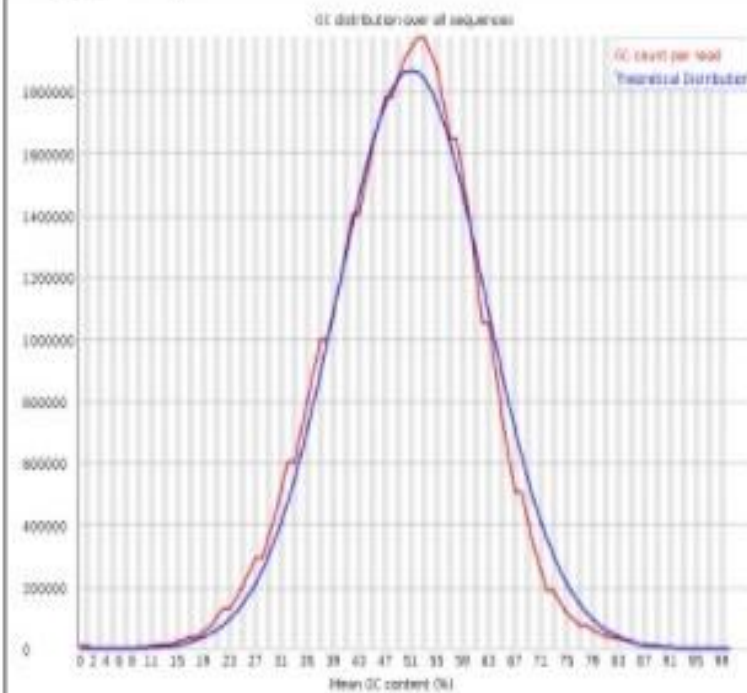


Table 4. GC distribution over all sequences

GC content for both files seem to almost match the theoretical distribution with a mean GC content which peaks at 53%, about 2% more than the theoretical curve. Large variances could be indicative of contamination and/or over sequencing of G-C content.

PO_1_1.fastq



PO_1_2.fastq

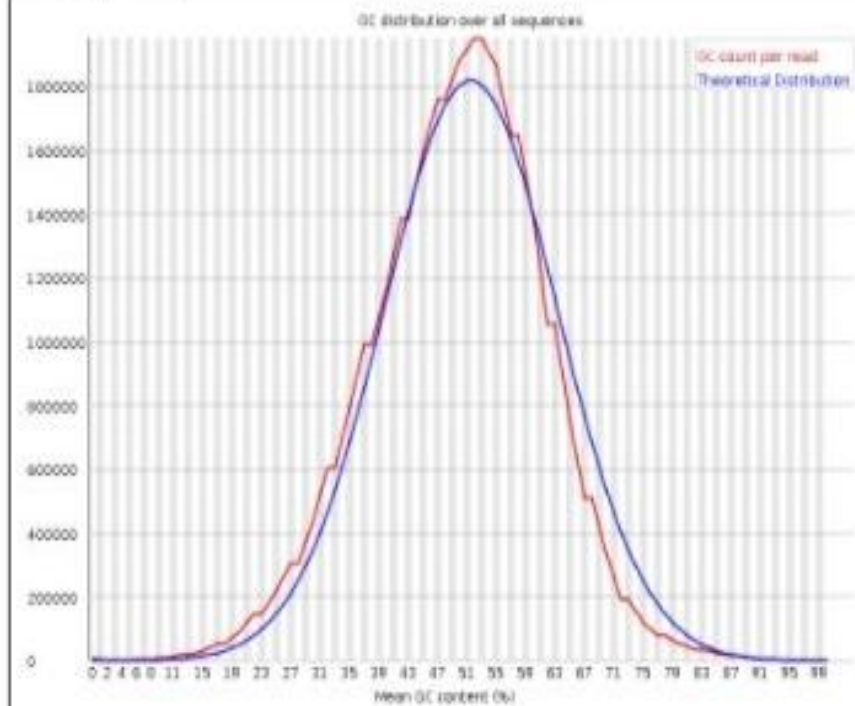


Table 5 - % of sequence remaining if deduplicated

The module passed with warnings but it generally left-shifted to pass the duplicate sequence module. There could be some sequencing bias or contamination.

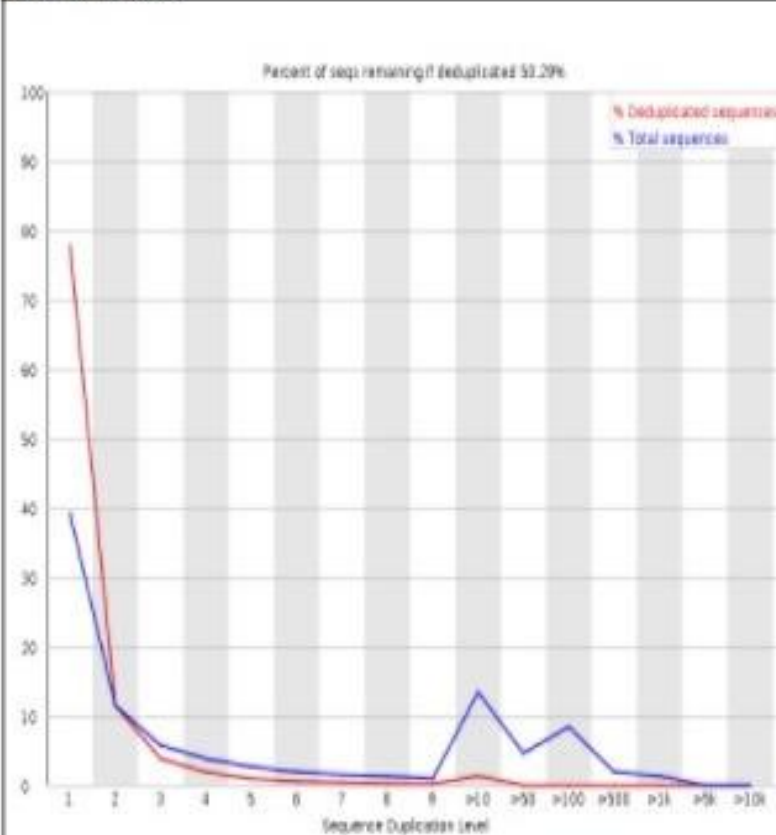
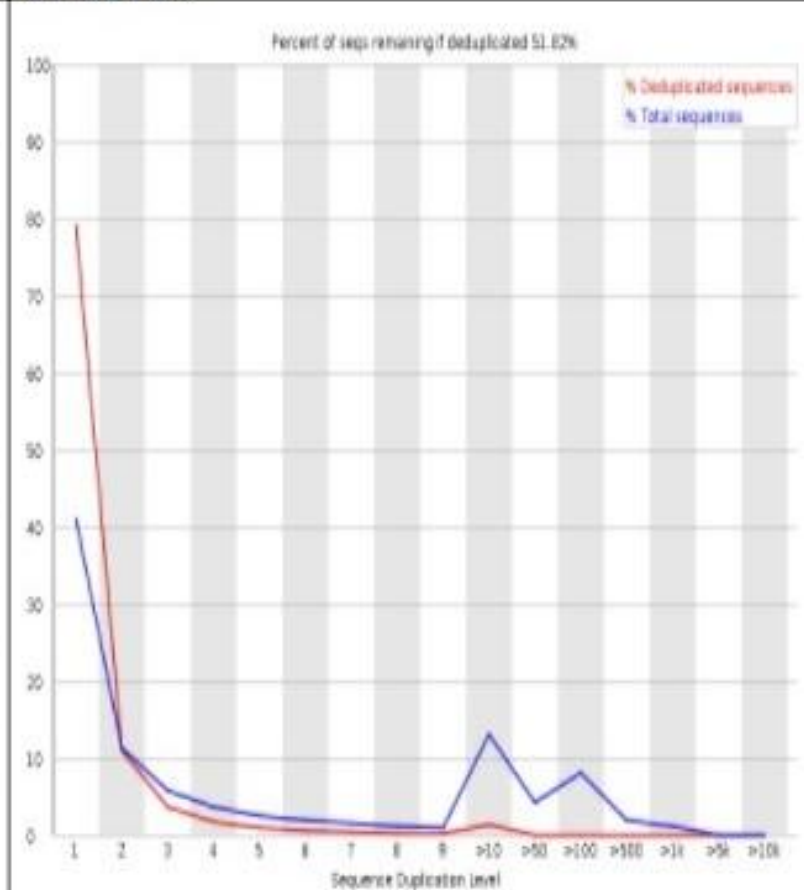
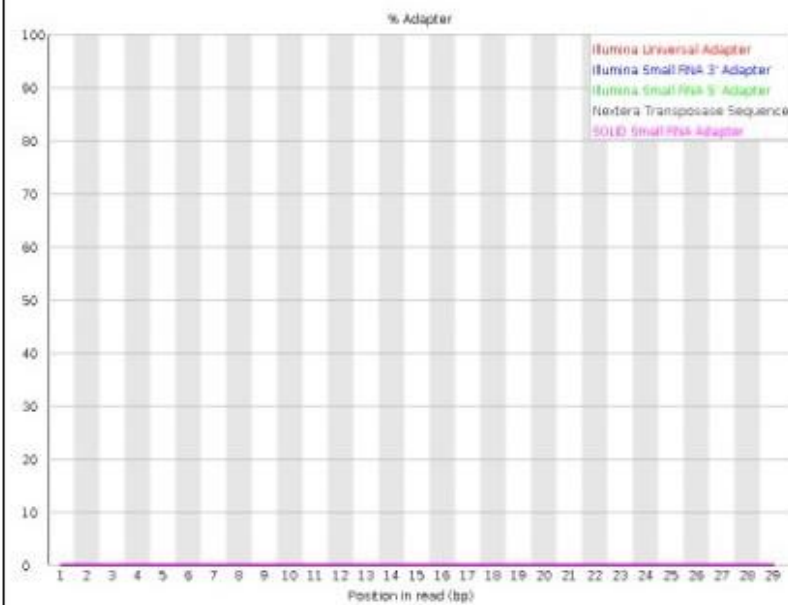
PO_1_1.fastq**PO_1_2.fastq**

Table 6. % Adapter

Adapters are needed to connect sample DNA with the DNA library to complete the sequence. The moduled passed making it unnecessary to trim data for analysis.¹

PO_1_1.fastq



PO_1_2.fastq

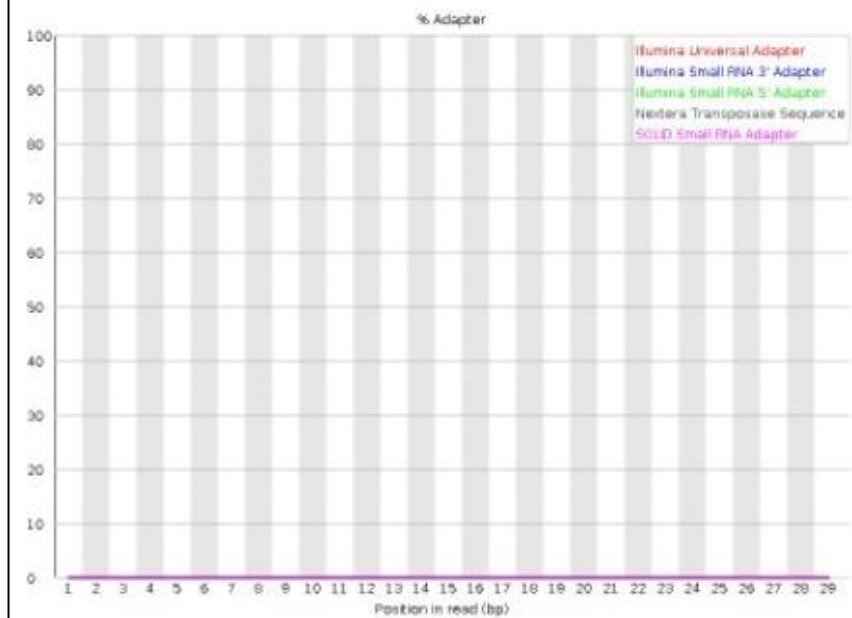
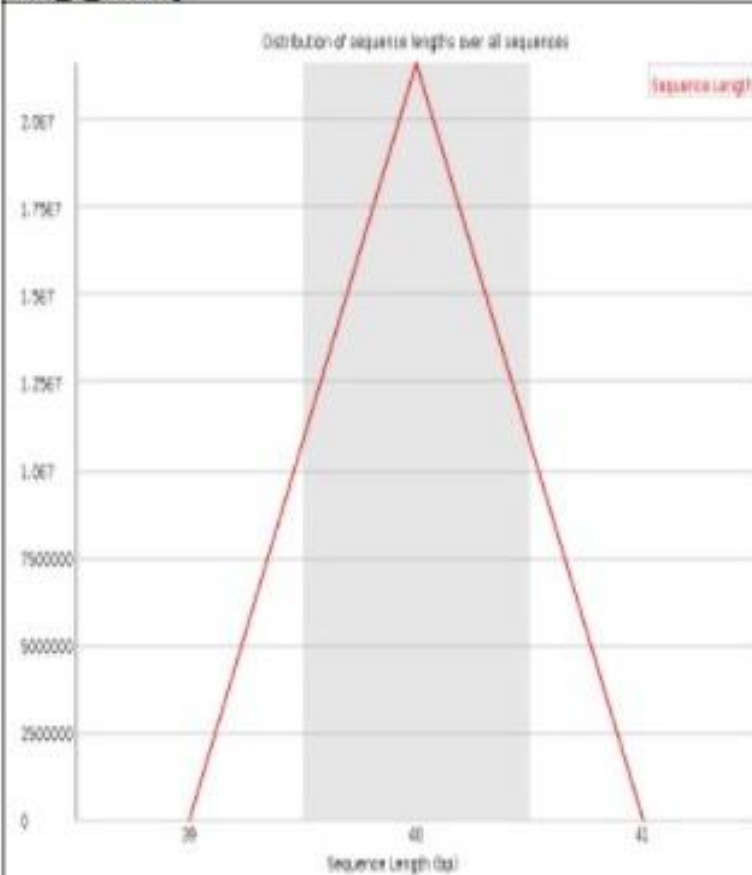


Table 7. - Distribution of sequence lengths over all sequences

The average read length was uniform about about 40 per base pair.

PO 1 1.fastq



PO 1 2.fastq

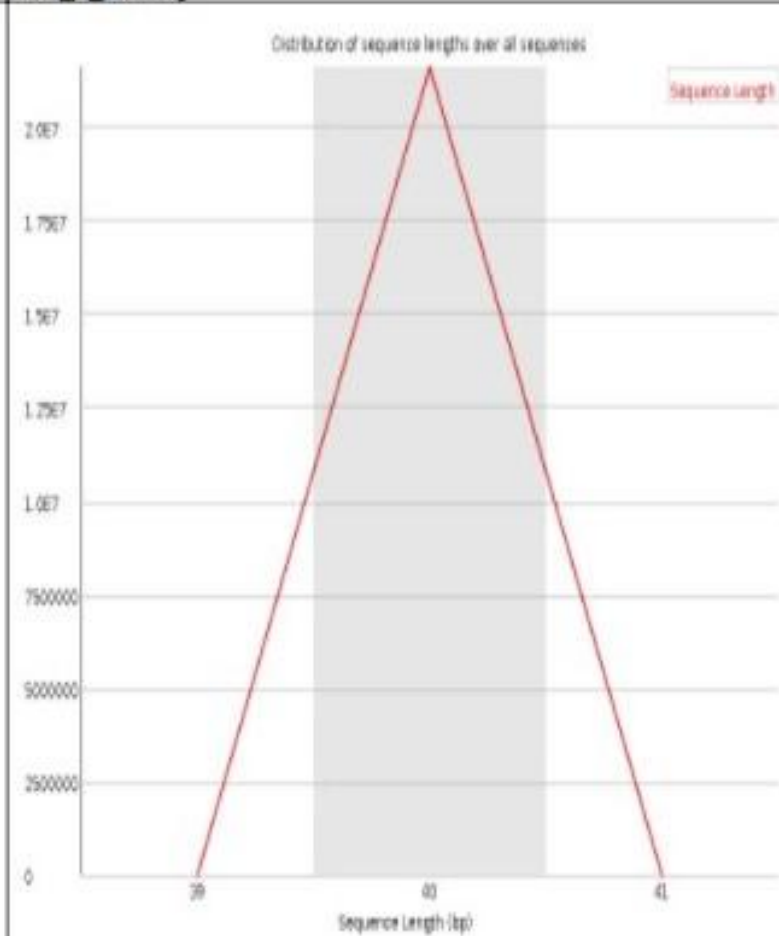


Table 8. N content across all bases

This measure reveals how well the base pairs were called, with “N” meaning NO confidence. Our value was close to 0% revealing almost perfect confidence in calling the base pairs⁴

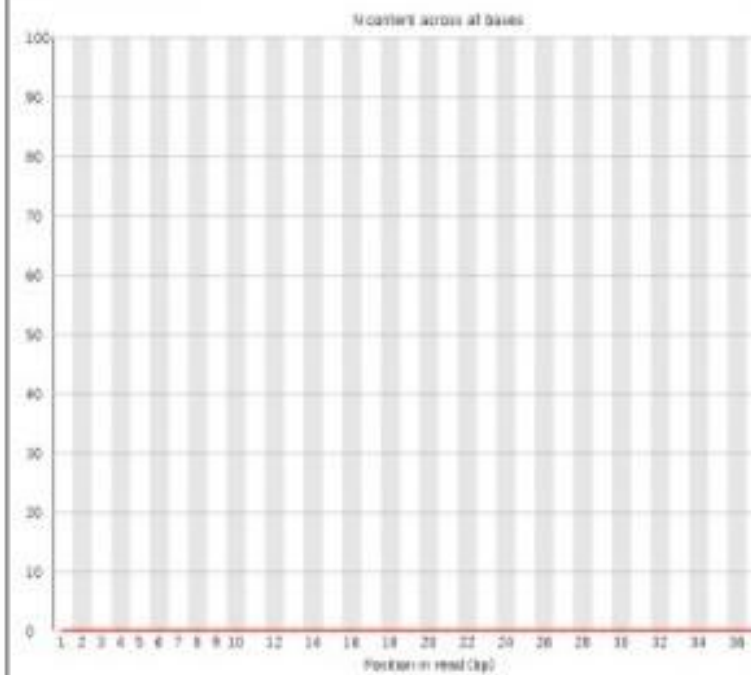
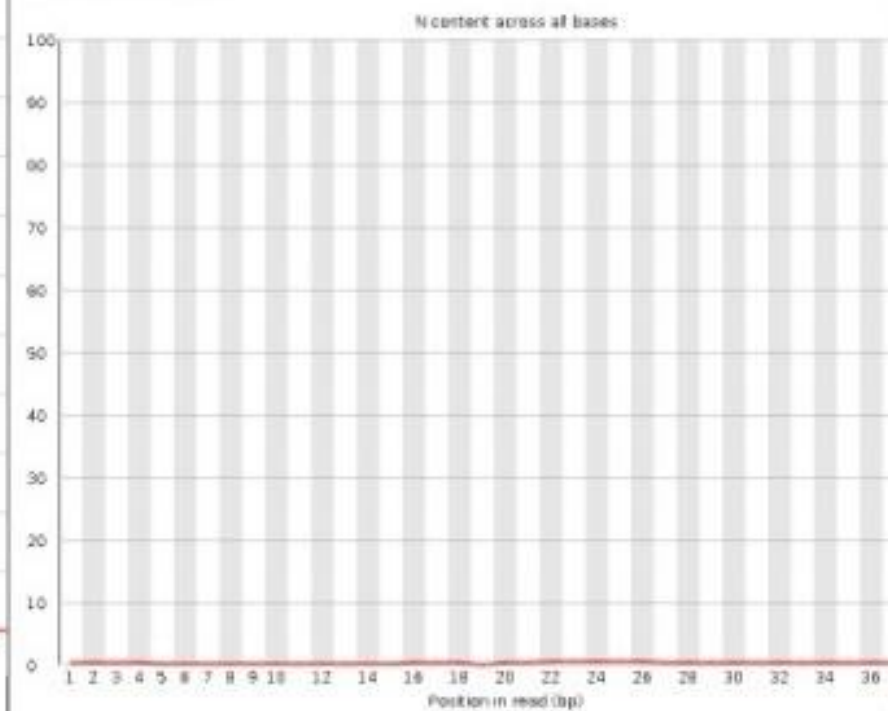
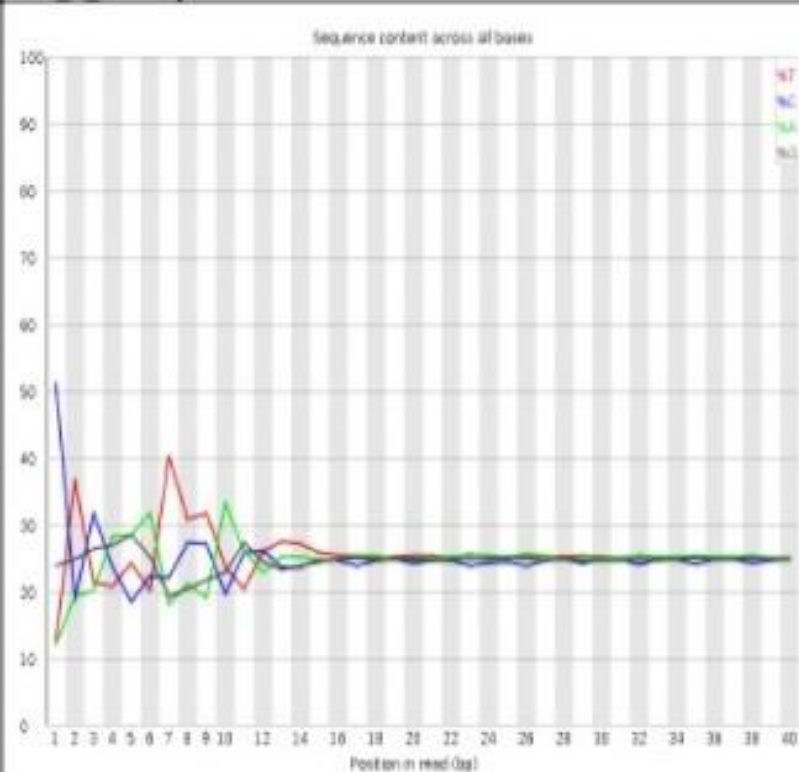
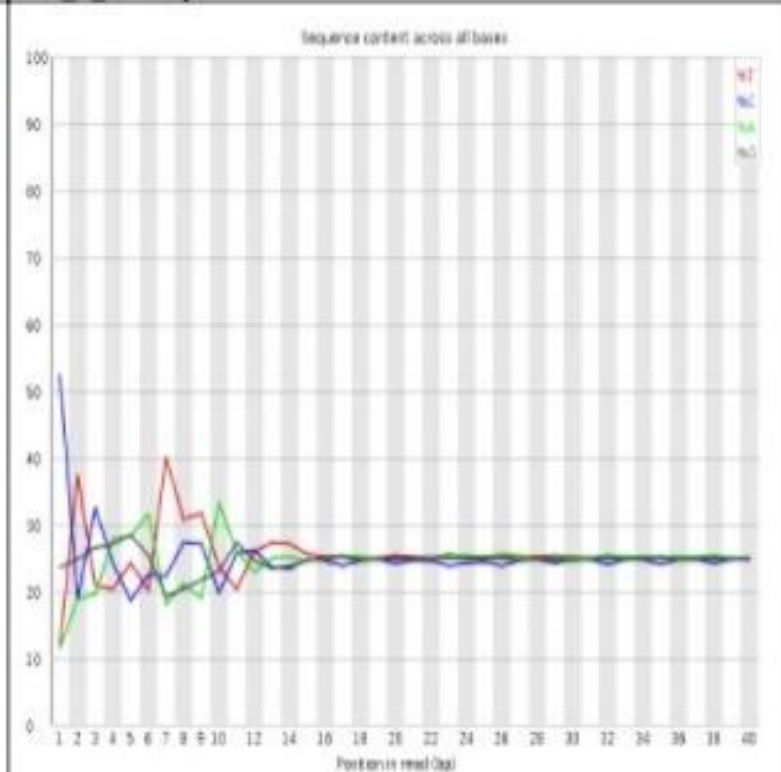
PO 1 1.fastq**PO 1 2.fastq**

Table 9. Sequence content across all bases

This measures the overall equality of the basepairs (A-T,G-C) and ideally should be parallel. Lack of parallelism is indicative of contamination and/or use of RNA which is evident in this section which typically does not pass for RNA sequenced data. Nonetheless, this measurement is fairly stable towards read position 15.³

PO 1 1.fastq**PO 1 2.fastq**

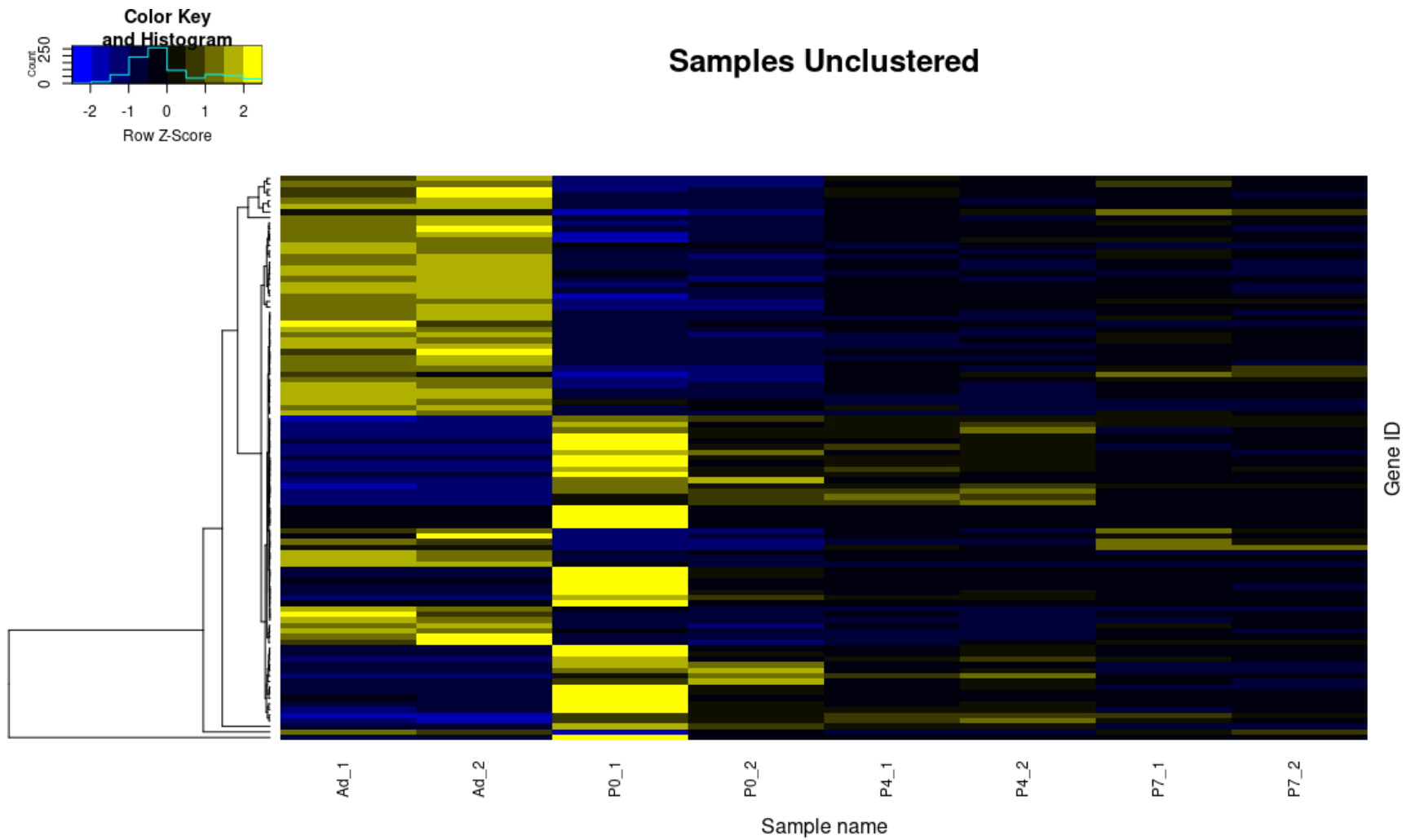


Figure S10: Heatmap of FPKM values with top 100 differentially expressed genes found in P0 vs Ad analysis. Sample columns here are also hierarchically clustered, indicating P0_1 as an outgroup to the other samples.




















































Annotation Cluster 1		Enrichment Score: 11.11			Count	P_Value	Benjamini
<input type="checkbox"/>	GOTERM_BP_FAT	cell cycle	RT		167	3.5E-30	2.3E-26
<input type="checkbox"/>	GOTERM_BP_FAT	cell division	RT		91	1.4E-26	4.8E-23
<input type="checkbox"/>	GOTERM_BP_FAT	mitotic cell cycle	RT		109	1.7E-25	3.7E-22
<input type="checkbox"/>	GOTERM_BP_FAT	cell cycle process	RT		133	3.9E-25	6.6E-22
<input type="checkbox"/>	GOTERM_BP_FAT	mitotic cell cycle process	RT		100	2.6E-24	3.5E-21
<input type="checkbox"/>	GOTERM_BP_FAT	mitotic nuclear division	RT		63	1.3E-16	1.2E-13
<input type="checkbox"/>	GOTERM_BP_FAT	nuclear division	RT		75	6.2E-16	4.6E-13
<input type="checkbox"/>	GOTERM_BP_FAT	regulation of cell cycle process	RT		70	7.6E-16	5.0E-13
<input type="checkbox"/>	GOTERM_BP_FAT	regulation of cell cycle	RT		91	1.8E-14	9.8E-12
<input type="checkbox"/>	GOTERM_BP_FAT	organelle fission	RT		75	2.0E-14	1.0E-11
<input type="checkbox"/>	GOTERM_BP_FAT	chromosome segregation	RT		46	1.6E-12	4.4E-10
<input type="checkbox"/>	GOTERM_BP_FAT	regulation of mitotic cell cycle	RT		55	5.7E-12	1.5E-9
<input type="checkbox"/>	GOTERM_BP_FAT	cell cycle phase transition	RT		49	8.8E-12	2.0E-9
<input type="checkbox"/>	GOTERM_BP_FAT	mitotic cell cycle phase transition	RT		45	4.4E-11	8.7E-9
<input type="checkbox"/>	GOTERM_BP_FAT	positive regulation of cell cycle	RT		43	2.4E-10	3.6E-8
<input type="checkbox"/>	GOTERM_BP_FAT	sister chromatid segregation	RT		30	2.8E-10	3.9E-8
<input type="checkbox"/>	GOTERM_BP_FAT	mitotic sister chromatid segregation	RT		26	2.0E-9	2.2E-7
<input type="checkbox"/>	GOTERM_BP_FAT	nuclear chromosome segregation	RT		35	2.8E-9	2.9E-7
<input type="checkbox"/>	GOTERM_BP_FAT	positive regulation of cell cycle process	RT		32	4.8E-9	4.6E-7
<input type="checkbox"/>	GOTERM_BP_FAT	regulation of cell cycle phase transition	RT		35	1.4E-8	1.1E-6
<input type="checkbox"/>	GOTERM_BP_FAT	regulation of mitotic cell cycle phase transition	RT		32	4.4E-8	3.2E-6
<input type="checkbox"/>	GOTERM_BP_FAT	regulation of chromosome segregation	RT		18	2.8E-7	1.5E-5
<input type="checkbox"/>	GOTERM_BP_FAT	cell cycle checkpoint	RT		26	1.1E-6	4.9E-5
<input type="checkbox"/>	GOTERM_BP_FAT	negative regulation of cell cycle	RT		40	2.0E-6	8.1E-5
<input type="checkbox"/>	GOTERM_BP_FAT	negative regulation of cell cycle process	RT		25	1.6E-5	5.0E-4
<input type="checkbox"/>	GOTERM_BP_FAT	positive regulation of mitotic cell cycle	RT		19	3.9E-5	1.1E-3
<input type="checkbox"/>	GOTERM_BP_FAT	negative regulation of mitotic cell cycle phase transition	RT		17	1.1E-4	2.7E-3
<input type="checkbox"/>	GOTERM_BP_FAT	negative regulation of cell cycle phase transition	RT		18	1.3E-4	3.1E-3
<input type="checkbox"/>	GOTERM_BP_FAT	negative regulation of mitotic cell cycle	RT		22	1.9E-4	4.3E-3
<input type="checkbox"/>	GOTERM_BP_FAT	regulation of sister chromatid segregation	RT		12	2.7E-4	5.7E-3
<input type="checkbox"/>	GOTERM_BP_FAT	mitotic cell cycle checkpoint	RT		16	2.8E-4	5.9E-3
<input type="checkbox"/>	GOTERM_BP_FAT	chromosome separation	RT		12	5.1E-4	9.6E-3
Annotation Cluster 2		Enrichment Score: 9.69			Count	P_Value	Benjamini
<input type="checkbox"/>	GOTERM_CC_FAT	proteinaceous extracellular matrix	RT		53	1.2E-11	2.9E-9
<input type="checkbox"/>	GOTERM_CC_FAT	extracellular matrix	RT		64	1.7E-10	2.7E-8
<input type="checkbox"/>	GOTERM_CC_FAT	extracellular matrix component	RT		27	3.9E-9	3.5E-7
Annotation Cluster 3		Enrichment Score: 9.58			Count	P_Value	Benjamini
<input type="checkbox"/>	GOTERM_BP_FAT	cell proliferation	RT		158	1.6E-14	9.8E-12
<input type="checkbox"/>	GOTERM_BP_FAT	regulation of cell proliferation	RT		136	6.8E-12	1.7E-9
<input type="checkbox"/>	GOTERM_BP_FAT	positive regulation of cell proliferation	RT		80	1.2E-7	7.2E-6
<input type="checkbox"/>	GOTERM_BP_FAT	negative regulation of cell proliferation	RT		60	3.7E-7	1.9E-5
Annotation Cluster 4		Enrichment Score: 8.52			Count	P_Value	Benjamini
<input type="checkbox"/>	GOTERM_BP_FAT	regulation of cellular component organization	RT		197	2.2E-17	2.5E-14
<input type="checkbox"/>	GOTERM_BP_FAT	regulation of organelle organization	RT		99	1.1E-9	1.4E-7
<input type="checkbox"/>	GOTERM_BP_FAT	positive regulation of cellular component organization	RT		93	3.6E-6	1.3E-4
<input type="checkbox"/>	GOTERM_BP_FAT	positive regulation of organelle organization	RT		46	9.2E-4	1.6E-2

Table S1: Top 4 down regulated gene clusters, clustered into functional groups by the DAVID Functional Clustering tool





Annotation Cluster 1		Enrichment Score: 21.93		Count	P_Value	Benjamini
<input type="checkbox"/>	GOTERM_CC_FAT	mitochondrion	RT	263	1.9E-50	1.3E-47
<input type="checkbox"/>	GOTERM_CC_FAT	mitochondrial part	RT	166	6.5E-45	2.3E-42
<input type="checkbox"/>	GOTERM_CC_FAT	mitochondrial envelope	RT	123	2.8E-32	4.7E-30
<input type="checkbox"/>	GOTERM_CC_FAT	mitochondrial inner membrane	RT	96	2.8E-32	4.7E-30
<input type="checkbox"/>	GOTERM_CC_FAT	mitochondrial membrane	RT	118	3.4E-32	4.7E-30
<input type="checkbox"/>	GOTERM_CC_FAT	organelle inner membrane	RT	98	2.3E-29	2.7E-27
<input type="checkbox"/>	GOTERM_CC_FAT	mitochondrial membrane part	RT	60	1.4E-26	1.4E-24
<input type="checkbox"/>	GOTERM_CC_FAT	mitochondrial protein complex	RT	54	1.9E-26	1.7E-24
<input type="checkbox"/>	GOTERM_CC_FAT	oxidoreductase complex	RT	43	7.6E-26	5.9E-24
<input type="checkbox"/>	GOTERM_CC_FAT	organelle envelope	RT	151	1.6E-25	1.1E-23
<input type="checkbox"/>	GOTERM_CC_FAT	envelope	RT	151	2.6E-25	1.6E-23
<input type="checkbox"/>	GOTERM_CC_FAT	inner mitochondrial membrane protein complex	RT	43	9.5E-22	5.5E-20
<input type="checkbox"/>	GOTERM_CC_FAT	respiratory chain	RT	36	5.5E-21	3.0E-19
<input type="checkbox"/>	GOTERM_CC_FAT	mitochondrial respiratory chain	RT	33	2.2E-19	1.1E-17
<input type="checkbox"/>	GOTERM_CC_FAT	respiratory chain complex	RT	32	1.6E-18	7.2E-17
<input type="checkbox"/>	GOTERM_CC_FAT	NADH dehydrogenase complex	RT	24	1.1E-16	3.9E-15
<input type="checkbox"/>	GOTERM_CC_FAT	mitochondrial respiratory chain complex I	RT	24	1.1E-16	3.9E-15
<input type="checkbox"/>	GOTERM_CC_FAT	respiratory chain complex I	RT	24	1.1E-16	3.9E-15
<input type="checkbox"/>	GOTERM_MF_FAT	oxidoreductase activity, acting on NAD(P)H, quinone or similar compound as acceptor	RT	16	1.8E-9	5.9E-7
<input type="checkbox"/>	GOTERM_MF_FAT	NADH dehydrogenase activity	RT	13	7.6E-8	1.4E-5
<input type="checkbox"/>	GOTERM_MF_FAT	oxidoreductase activity, acting on NAD(P)H	RT	19	8.4E-8	1.4E-5
<input type="checkbox"/>	GOTERM_MF_FAT	NADH dehydrogenase (ubiquinone) activity	RT	12	5.3E-7	5.9E-5
<input type="checkbox"/>	GOTERM_MF_FAT	NADH dehydrogenase (quinone) activity	RT	12	5.3E-7	5.9E-5
Annotation Cluster 2		Enrichment Score: 16.81		Count	P_Value	Benjamini
<input type="checkbox"/>	GOTERM_BP_FAT	organic acid metabolic process	RT	121	5.0E-25	1.1E-21
<input type="checkbox"/>	GOTERM_BP_FAT	oxoacid metabolic process	RT	113	2.2E-24	3.6E-21
<input type="checkbox"/>	GOTERM_BP_FAT	carboxylic acid metabolic process	RT	112	4.1E-24	5.3E-21
<input type="checkbox"/>	GOTERM_BP_FAT	monocarboxylic acid metabolic process	RT	84	4.9E-20	4.5E-17
<input type="checkbox"/>	GOTERM_BP_FAT	fatty acid metabolic process	RT	51	1.1E-10	1.5E-8
<input type="checkbox"/>	GOTERM_BP_FAT	cellular lipid metabolic process	RT	88	2.8E-9	3.2E-7
<input type="checkbox"/>	GOTERM_BP_FAT	lipid metabolic process	RT	108	3.1E-9	3.5E-7
Annotation Cluster 3		Enrichment Score: 15.31		Count	P_Value	Benjamini
<input type="checkbox"/>	GOTERM_BP_FAT	generation of precursor metabolites and energy	RT	73	5.5E-27	1.8E-23
<input type="checkbox"/>	GOTERM_BP_FAT	energy derivation by oxidation of organic compounds	RT	55	1.6E-21	1.8E-18
<input type="checkbox"/>	GOTERM_BP_FAT	purine ribonucleoside metabolic process	RT	58	1.2E-19	9.9E-17
<input type="checkbox"/>	GOTERM_BP_FAT	purine nucleoside monophosphate metabolic process	RT	52	2.1E-19	1.5E-16
<input type="checkbox"/>	GOTERM_BP_FAT	purine nucleoside metabolic process	RT	58	2.4E-19	1.6E-16
<input type="checkbox"/>	GOTERM_BP_FAT	ATP metabolic process	RT	48	6.8E-19	3.7E-16
<input type="checkbox"/>	GOTERM_BP_FAT	purine ribonucleoside monophosphate metabolic process	RT	51	9.4E-19	4.7E-16
<input type="checkbox"/>	GOTERM_BP_FAT	cellular respiration	RT	41	1.5E-18	6.8E-16
<input type="checkbox"/>	GOTERM_BP_FAT	glycosyl compound metabolic process	RT	62	5.5E-18	2.3E-15
<input type="checkbox"/>	GOTERM_BP_FAT	ribonucleoside metabolic process	RT	58	5.8E-18	2.3E-15
<input type="checkbox"/>	GOTERM_BP_FAT	nucleobase-containing small molecule metabolic process	RT	89	5.9E-18	2.3E-15
<input type="checkbox"/>	GOTERM_BP_FAT	nucleoside monophosphate metabolic process	RT	52	6.7E-18	2.4E-15
<input type="checkbox"/>	GOTERM_BP_FAT	ribonucleoside monophosphate metabolic process	RT	51	8.4E-18	2.9E-15
<input type="checkbox"/>	GOTERM_BP_FAT	purine ribonucleoside triphosphate metabolic process	RT	48	2.3E-17	7.4E-15
<input type="checkbox"/>	GOTERM_BP_FAT	ribonucleoside triphosphate metabolic process	RT	48	8.0E-17	2.5E-14
<input type="checkbox"/>	GOTERM_BP_FAT	nucleoside metabolic process	RT	58	1.1E-16	3.2E-14
<input type="checkbox"/>	GOTERM_BP_FAT	purine nucleoside triphosphate metabolic process	RT	48	1.1E-16	3.2E-14
<input type="checkbox"/>	GOTERM_BP_FAT	nucleoside phosphate metabolic process	RT	82	2.6E-16	6.7E-14
<input type="checkbox"/>	GOTERM_BP_FAT	nucleotide metabolic process	RT	81	2.9E-16	7.3E-14
<input type="checkbox"/>	GOTERM_BP_FAT	purine-containing compound metabolic process	RT	73	8.2E-16	1.9E-13
<input type="checkbox"/>	GOTERM_BP_FAT	organophosphate metabolic process	RT	108	9.5E-16	2.1E-13
<input type="checkbox"/>	GOTERM_BP_FAT	nucleoside triphosphate metabolic process	RT	49	1.6E-15	3.4E-13
<input type="checkbox"/>	GOTERM_BP_FAT	electron transport chain	RT	26	1.6E-14	3.3E-12
<input type="checkbox"/>	GOTERM_BP_FAT	purine ribonucleotide metabolic process	RT	64	4.8E-14	9.4E-12
<input type="checkbox"/>	GOTERM_BP_FAT	purine nucleotide metabolic process	RT	66	6.2E-14	1.2E-11
<input type="checkbox"/>	GOTERM_BP_FAT	respiratory electron transport chain	RT	24	1.9E-13	3.6E-11
<input type="checkbox"/>	GOTERM_BP_FAT	ribonucleotide metabolic process	RT	64	2.4E-13	4.4E-11
<input type="checkbox"/>	GOTERM_BP_FAT	ribose phosphate metabolic process	RT	64	7.1E-13	1.2E-10
<input type="checkbox"/>	GOTERM_BP_FAT	mitochondrial ATP synthesis coupled electron transport	RT	18	3.7E-11	5.6E-9
<input type="checkbox"/>	GOTERM_BP_FAT	ATP synthesis coupled electron transport	RT	18	2.3E-10	3.1E-8
<input type="checkbox"/>	GOTERM_BP_FAT	oxidative phosphorylation	RT	19	4.2E-9	4.5E-7
<input type="checkbox"/>	GOTERM_BP_FAT	carbohydrate derivative metabolic process	RT	98	6.2E-9	6.6E-7
<input type="checkbox"/>	GOTERM_BP_FAT	mitochondrial electron transport, NADH to ubiquinone	RT	9	1.6E-7	1.5E-5
Annotation Cluster 4		Enrichment Score: 11.8		Count	P_Value	Benjamini
<input type="checkbox"/>	GOTERM_CC_FAT	extracellular organelle	RT	256	8.8E-18	3.8E-16
<input type="checkbox"/>	GOTERM_CC_FAT	extracellular vesicle	RT	253	4.9E-17	2.0E-15
<input type="checkbox"/>	GOTERM_CC_FAT	extracellular exosome	RT	251	2.0E-16	6.5E-15
<input type="checkbox"/>	GOTERM_CC_FAT	membrane-bounded vesicle	RT	282	6.4E-11	1.9E-9
<input type="checkbox"/>	GOTERM_CC_FAT	extracellular region part	RT	297	5.8E-8	1.5E-6
<input type="checkbox"/>	GOTERM_CC_FAT	extracellular region	RT	314	4.6E-5	8.6E-4

Table S2: Top 4 upregulated gene clusters, clustered into functional groups by the DAVID Functional Clustering tool