Project 2 3/18/22

Group: Tinman

Transcriptional Profile of Mammalian Cardiac Regeneration with mRNA-Seq

Data Curator: Reshma
Programmer: Rizky
Analyst: Aneeq
Biologist: Ally

INTRODUCTION (Shared)

The limitations of the adult mammalian heart when it comes to repairing sustained injuries has been well documented (Steinhauser et al. 2011). However, cardiac myocytes in neonatal mice are able to partially undergo cell fate reversion (Porello et al. 2011). O'Meara et al. attempted to identify the mechanistic transcriptional changes on the genome level that are associated with regeneration of the neonatal mouse heart in order to "[elucidate] the molecular roadblocks" responsible for hindering an analogous response in the adult mammalian heart (2015).

In concordance with the methodology of O'Meara et al., this study attempts to replicate these findings by using publicly available mRNA-Seq datasets (2015). The primary focus will be on data from *in vivo* maturation of myocytes at different postnatal timepoints as well as the matured myocyte denoted as adult (Ad). In reproducing these results, crucial transcriptional regulators and mediators can be confirmed. In this replication analysis, results are overall comparable to that of O'Meara et al. (2015).

DATA (Data Curator)

O'Meara et al. (2015) extracted total RNA using Trizol from 0 day mouse postnatal ventricular myocardium and performed paired-end sequencing with read length of 40 base pairs using Illumina HiSeq 2000. RNA-Seq sequence-read archive (SRA) sample GSM1570702 was obtained from the Gene Expression Omnibus. The SRA file was converted to FASTQ format using the NCBI SRA Toolkit [4]. FastQC [5] was inspected for the quality of the sequencing reads.

METHODS (Programmer/Analyst)

P_0 sample alignment and QA via TopHat and RSeQC

Using the P_0 data prepared by the data curator, we aligned the extracted FASTQ files to the mouse reference genome (mm9) using *TopHat*. This reference, along with its associated FASTA file and its *Bowtie2* indexes was obtained from the cluster at /project/bf528/project_2/reference. Due to the memory intensive nature of *TopHat*'s operations, the *TopHat* alignment was run as a batch job on the cluster through the *qsub* command. In order to run the TopHat alignment as a batch job, we modified a *qsub* file template, which was obtained from the following directory: /project/bf528/project_2/scripts/qsub_skel.qsub. Within the qsub file we generated to run this alignment, the necessary utilities and their respective dependencies were loaded using the command: *module load samtools bowtie2 boost topha*t. The required arguments were obtained from /project/bf528/project_2/scripts/tophat_args.txt. Once all arguments and input/reference/index files had been specified in the *qsub* file, the job was submitted using the qsub command. Upon successful completion of the *TopHat* alignment, a file called accepted_hits.bam was generated. Using *samtools* and *flagstat* on this output file generated a summary of the bam file. Further analysis of the data was generated through three RseQC utilities, *geneBody_coverage.py*, *inner_distance.py*, and *bam_stat.py*. This suite of RseQC utilities generates a collection of plots summarizing the uniformity coverage for the alignment, the mean insert size for the paired reads, as well as a summary of the mapping statistics ("RNA-seq quality metrics with RseQC" (accessed 2022); Wang et al. 2012). In order to run these programs, the RseQC utilities were loaded by using the following command: *module load python3 samtools rseqc*. After this, the BAM file was indexed using the *samtools index* command, and the python utilities were run using accepted_hits.bam as the input, with the exception of *geneBody_coverage.py*, which used the .bai file generated after indexing the BAM file.

Using cufflinks to quantify gene expression

In order to quantify gene expression for the P_0 sample data, another batch job was submitted to the cluster, this time using *cufflinks*. A *qsub* file was created using the qsub_skel.qsub file mentioned earlier in the methods for the *TopHat* alignment. Required arguments were obtained from /project/bf528/project_2/scripts/cufflinks_args.txt. A histogram was constructed from the resulting "genes.fpkm_tracking" output file using R. For the purposes of making the dataset more presentable, samples with FPKM above 1000 were considered significant and were subsequently included in the following results and discussion sections. Finally, one last batch job was submitted to identify differentially expressed genes between the P0 file and the remaining samples provided for the assignment (P0_2, Ad_1, Ad_2). The tool used in this instance was *cuffdiff*. The differentially expressed genes were identified by filtering for all genes passing the "significant" filter in R. This output was further filtered by only including sample FPKM values above 1000 to identify the highest performing i.e. most upregulated differentially expressed genes.

Identification of DE expressed genes

The output from *cuffdiff* was then processed using R to extract the most differentially expressed genes. Cuffdiff results were sorted by smallest q-values and the top ten genes were extracted. Furthermore, significant genes were separated from the cuffdiff results and histograms plotted before and after separation to compare the data. These significant genes were then separated into up-regulated and down-regulated lists to perform functional annotation clustering analysis using DAVID.

RESULTS (Curator/Programmer/Analyst)

In order to assess RNA sequencing data quality, 2 html reports of quality metrics for the paired-end reads (P0_1_1 and P0_1_2) were generated from FastQC files. The results were almost the same for read1 and read2. The report showed that overall sequence quality was high (**Figure 1**), with the lowest per base sequence quality score being 30. GC content was 52% and normally distributed in both reads (**Figure 2**). The percentage of sequences remaining if deduplicated was found to be 50.29% and 51.82% in read1 and read2 respectively (**Figure 3**).

This indicates that there was considerable sequence duplication, which is expected for RNA-Seq experiments.

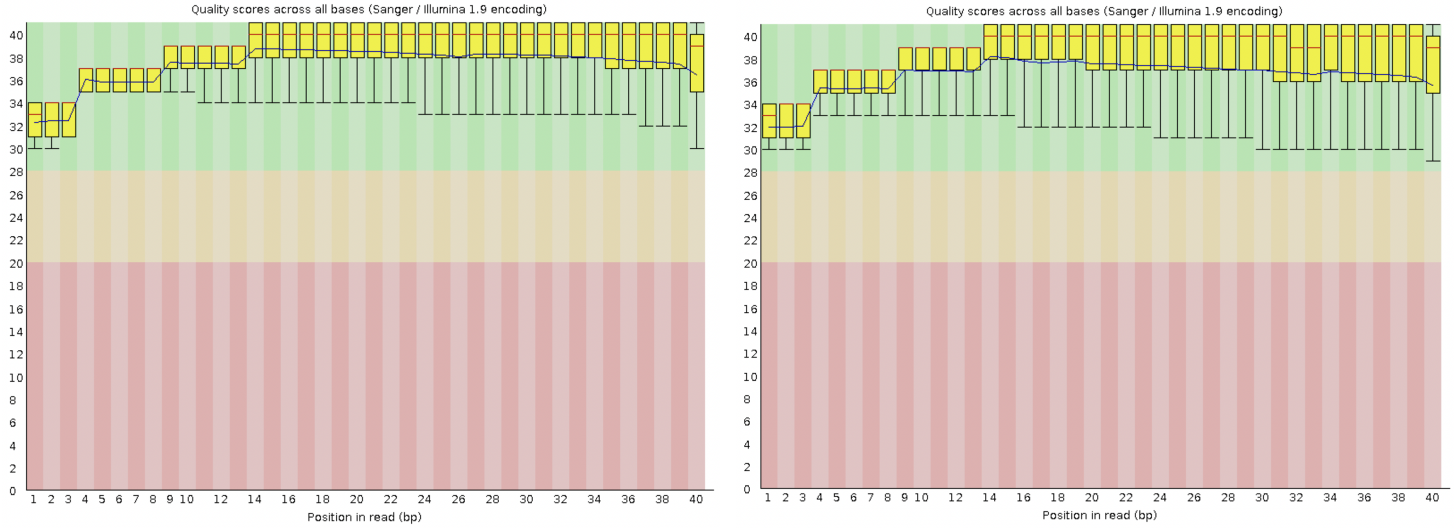


Figure 1. Per base sequence quality of P0 read1 and read2 respectively: An overview of the range of quality values across all bases at each position in the FastQ file

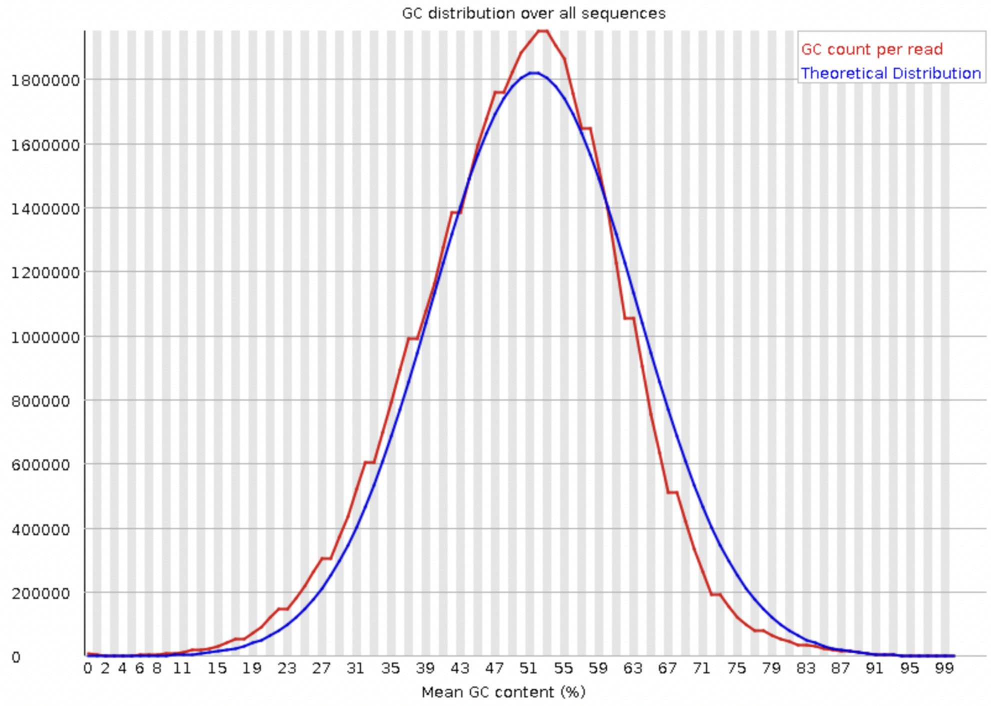


Figure 2. Per sequence GC content of P0 read1: Measures the GC content across the whole length of each sequence

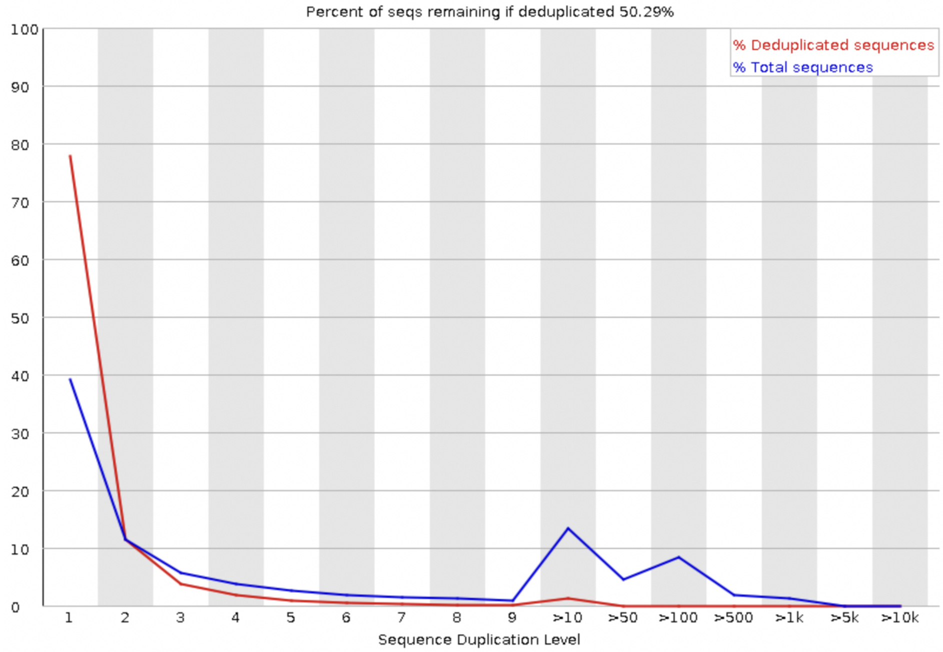


Figure 3. Sequence duplication levels of P0 read1: Counts the degree of duplication for every sequence and creates a plot showing the relative number of sequences with different degrees of duplication.

The only metric that failed was per-base sequence content (Figure 4), which indicated that sequence composition was biased at the first few bases of the reads.

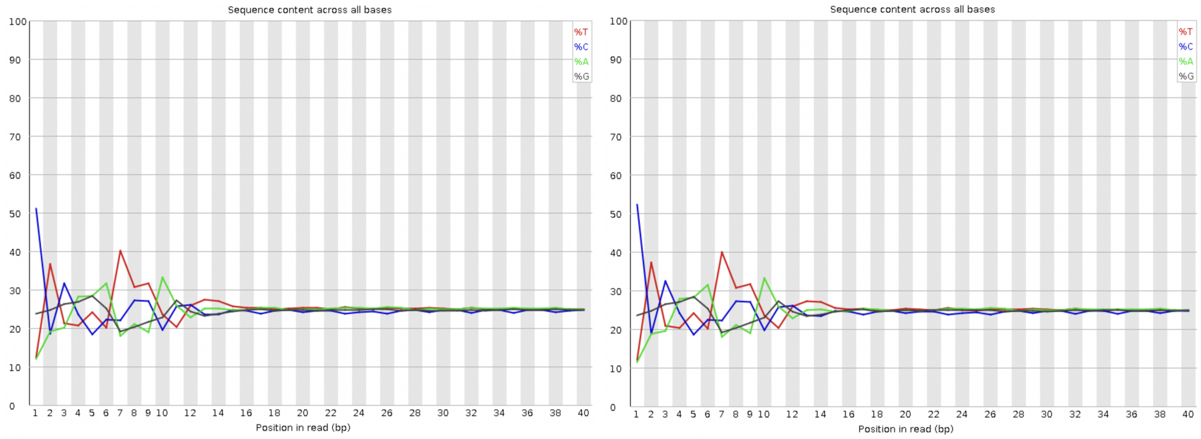


Figure 4. Per base sequence content of P0 read1 and read 2 respectively: Plots out the proportion of each base position in a file for which each of the four normal DNA bases has been called.

Inspecting the alignment summary generated using *samtools* and *flagstat*, it was revealed that the alignment between the mm9 reference genome and the P_0 sample had a 95.9% overall read mapping rate. For the left and right reads, the input was analogous (2.16e7). Looking at the left reads, 2.09e7 reads (96.8% of the total number of reads) were mapped– of which 1.47e6 (7%) had multiple alignments. In addition to this, 2.05e7 reads (95.1% of the total reads) were mapped for the right reads. 1.43e6 (7%) of the right reads had multiple alignments. Analysis of the aligned pairs revealed a total number of 2.0e7 paired reads, of which 1.38e6 (6.9% of the paired reads had multiple alignments. Furthermore 7.94e5 (4.0%) of the aligned pairs are discordant alignments. Overall, there was an 88.9% concordant pair alignment rate.

When inspecting the output of the *geneBody_coverage.py* command as illustrated by **Figure 5**, it is clear that a fair number of reads had a slight 3' bias, with close to 0 coverage at the 5' end of the reads, denoted by the 0 end of the x-axis. The majority of the reads had coverage in the 40th through 100th percentile of the gene body (**Figure 5**).

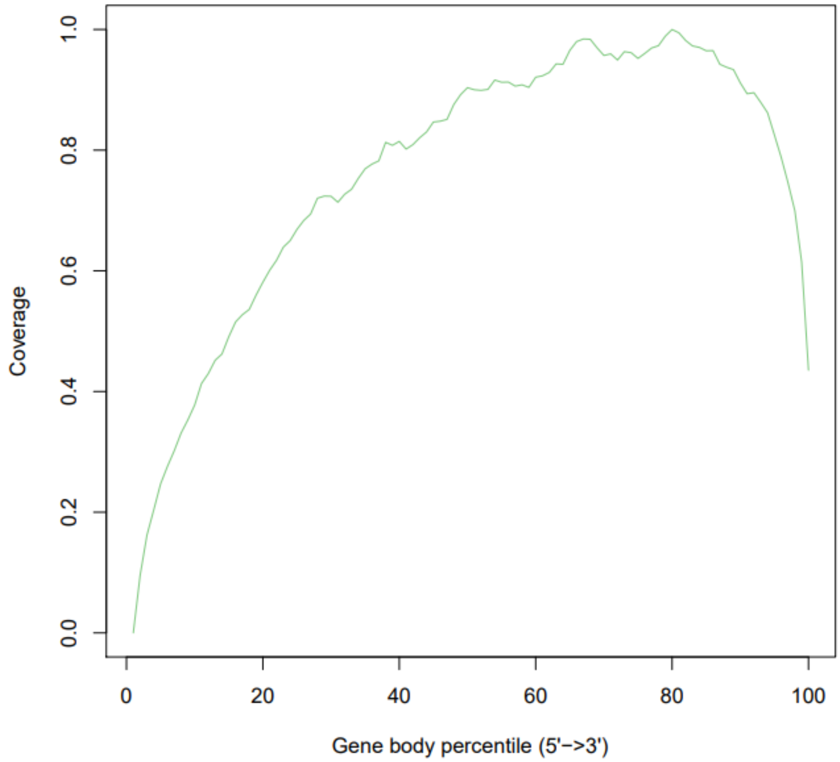


Figure 5. **Coverage distribution for the RNA-seq reads over the gene body.** This is a measure of coverage uniformity and visualizes whether or not 5' to 3' bias exists ("RNA-seq quality metrics with RseQC" (accessed 2022); Wang et al. 2012). From this figure it is evident that generally the reads mapped to the 3' end of the inserts.

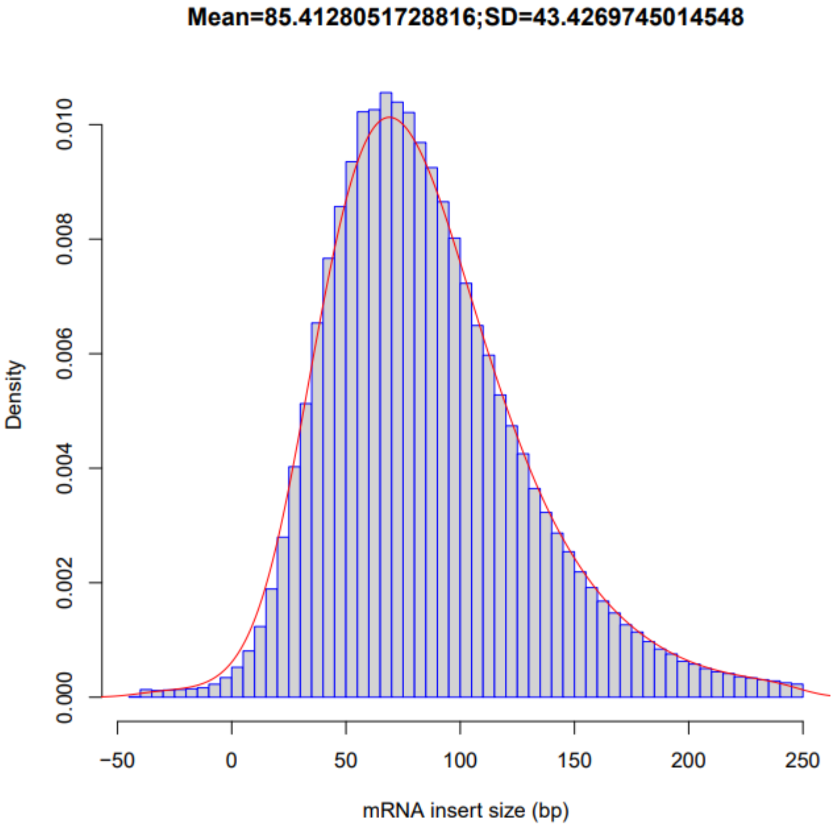


Figure 6. mRNA insert size (bp) distribution histogram generated using *inner_distance.py* ("RNA-seq quality metrics with RseQC" (accessed 2022); Wang et al. 2012). The mean insert size (MIS) was 85.4 bp and the standard deviation (SD) was 43.4. The insert size profile follows a normal distribution pattern.

Examination of the output of the *bam_stat.py* program revealed that a total of 4.97e7 reads were recorded, of which 100% passed the QC. No reads were a result of optical/PCR duplication. In addition, there were 8.32e6 non-primary hits and 0 unmapped reads. In terms of unique reads, there were 3.85e7 unique reads (77.4% of the total number of reads. Conversely, there were 2.90e5 non-unique reads i.e. 5.83% of all reads. To further summarize, there were 3.31e7 Non-splice reads (66.6% of all accepted hits) and 5.39e6 splice reads (10.8% of all accepted hits), as well as 2.80e7 reads mapped in proper pairs (56.3% of all accepted hits).

The *cufflinks* batch job output reveals that there were multiple genes with FPKM values that were several orders of magnitude bigger than the vast majority of the expression values (**Figure 7**). Namely, it is worth noting that there were 108 genes with greater than 1000 FPKM. As mentioned in the methodology section of Porello et al., only genes with a FPKM greater than 1000 will be included in our subsequent analysis and genes with lower expression values will be considered "noise" (2011). After filtering out FPKM values that were less than 1000, the average FPKM value goes from 67.695 to 39491 and the median goes from 0.022875 to 2371.25. Even after omitting >99% of the genes output in the *cufflinks* job, there is a clear left-skew to the remaining data points, with one outlier dragging the tail to the right end of the x-axis (**Figure 7**).

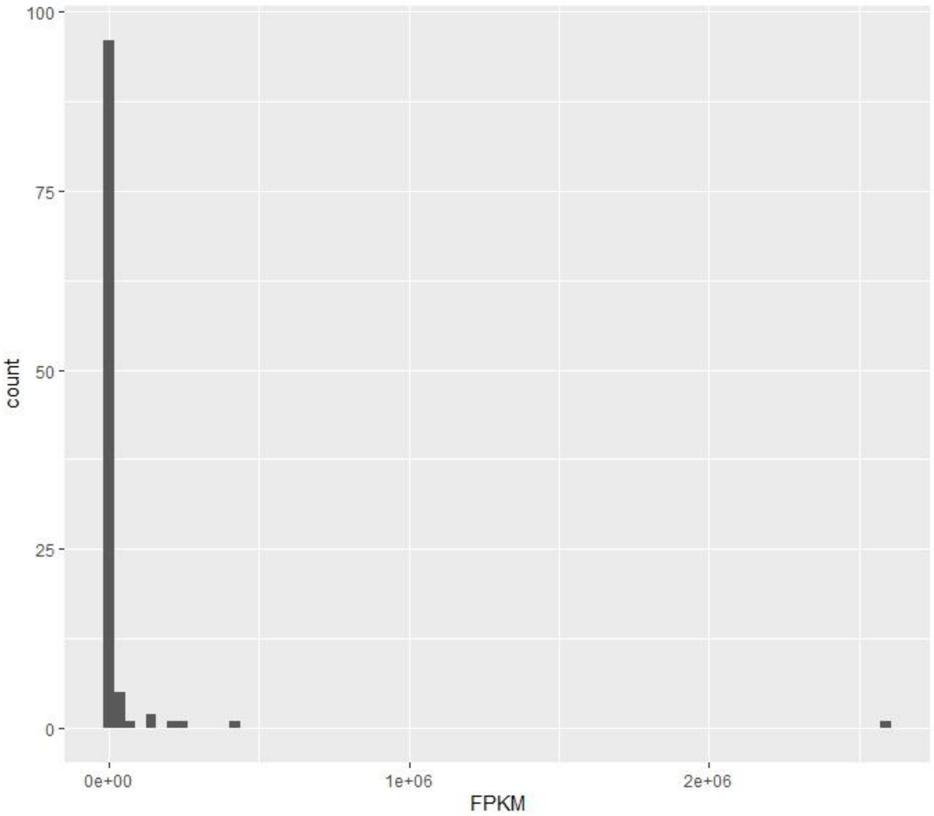


Figure 7. A histogram of FPKM values of 108 genes above a FPKM value of 1000. Note that even after filtering out 37361 samples below the FPKM threshold of 1000, genes with higher FPKM values continue to skew the axis (See **Figure S.1** to see the effect of including samples below FPKM = 1000).

Lastly, when analyzing the number of differentially expressed genes, a total of 36329 genes were generated in the cuffdiff output file. To better understand the cuffdiff results, histograms were produced before (**Figure 8**) and after (**Figure 9**) selection of the significant genes. After filtering for significance, it was found that 5427 genes were differentially expressed. Of the 5427, we determined that 2830 of them were upregulated and 2597 were downregulated. Further filtration to only include FPKM values greater than 1000 for both samples identified the 19 most upregulated differentially expressed genes of interest across the 3 samples used in the cuffdiff job. These were the highest performing up-regulated genes (**Table S.1**). **Table 1** summarizes the top differentially expressed genes before FPKM filtering. DAVID was used to perform functional annotation clustering and the top clusters identified for the upregulated and downregulated genes are listed in **Table 2 and 3** respectively.

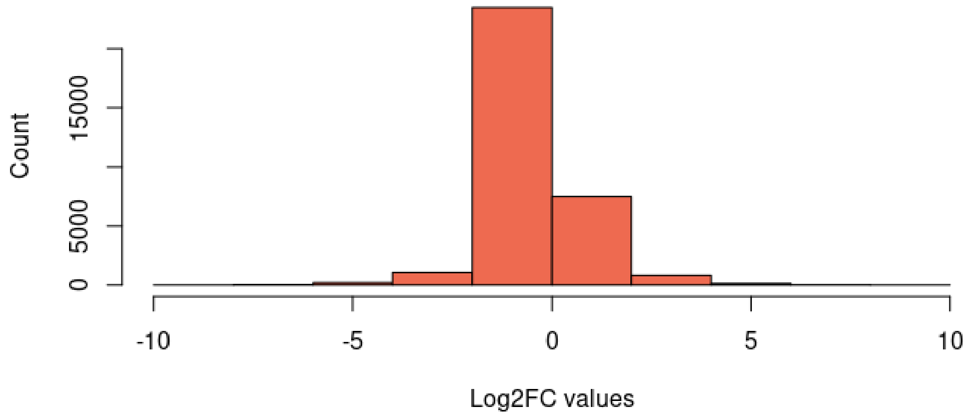


Figure 8. Histogram showing the distribution of Log 2 Fold Change values of all the genes.

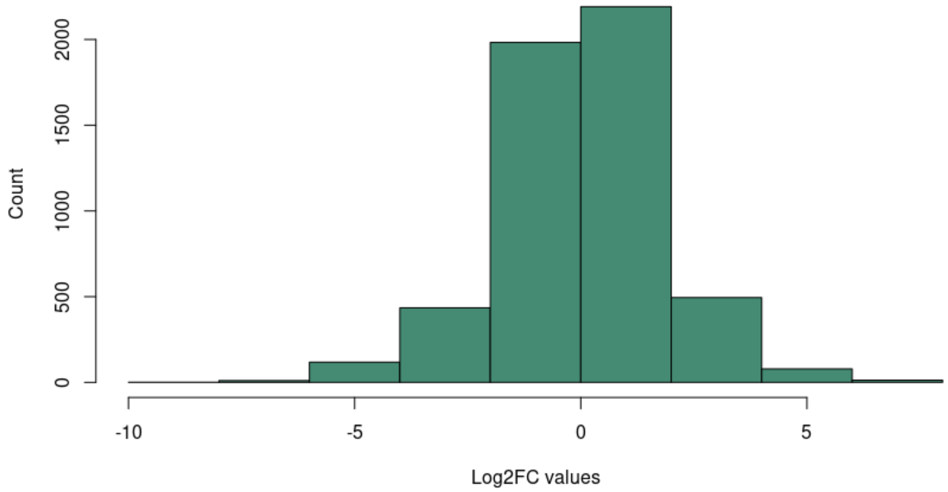


Figure 9. Histogram showing the distribution of Log2 Fold Change values of significant genes.

Genes	Q-value	Log-2 Fold Change
Adhfe1	3.2×10^{-4}	1.017
Tmem70	3.2×10^{-4}	1.131
Gsta3	3.2×10^{-4}	4.033
Lmbrd1	3.2×10^{-4}	0.944
Dst	3.2×10^{-4}	1.387
Plekhb2	3.2×10^{-4}	1.407
Cox5b	3.2×10^{-4}	0.802
Mrpl30	3.2×10^{-4}	1.144
Tmem182	3.2×10^{-4}	1.163
Nck2	3.2×10^{-4}	-0.950

Table 1. Summarized list of the top differentially expressed genes based on the q-values.

Cluster Annotation	Enrichment Score	Number of genes	P-value	Adjusted P-value
Ion Binding	118.91	933	2.1×10^{-181}	5.2×10^{-178}
Organic acid metabolic process	80.3	274	5.9×10^{-98}	4.4×10^{-95}
Mitochondrion	76.18	579	1.6×10^{-234}	1.9×10^{-231}
Intracellular signal transduction	72.12	477	1.8×10^{-113}	3×10^{-110}
Small molecule binding	68.4	473	4.5×10^{-108}	5.6×10^{-105}
Organophosphate	64.34	288	1×10^{-111}	1.5×10^{-108}

metabolic process				
Identical protein binding	57.51	427	9.2×10^{-95}	5.7×10^{-92}
Cellular Localization	55.54	456	1.6×10^{-90}	7.9×10^{-88}
Phosphorus metabolic process	54.1	642	2×10^{-183}	2.1×10^{-179}
Cellular response to chemical stimulus	52.99	586	2.2×10^{-131}	7.5×10^{-128}

Table 2. List of the top functional clusters of the up-regulated genes.

Cluster Annotation	Enrichment Score	Number of genes	P-value	Adjusted P-value
Organic cyclic compound binding	178.13	880	2.7×10^{-207}	5.4×10^{-204}
Ion binding	136.66	867	1.5×10^{-177}	9.9×10^{-175}
Cellular macromolecule biosynthetic process	103.19	802	3.9×10^{-199}	3.8×10^{-195}
Chromosome Organization	95.42	324	5.5×10^{-131}	2.4×10^{-128}
Regulation of cellular component organization	85.89	517	8.8×10^{-142}	4.3×10^{-139}
Cell proliferation	62.6	345	1.4×10^{-79}	1.9×10^{-77}
Nucleoplasm part	58.88	263	2.5×10^{-82}	5.7×10^{-80}

Enzyme binding	56.05	390	2.2 x 10^{-93}	5.6 x 10^{-91}
Regulation of signal transduction	54.13	434	6.7 x 10^{-84}	1.2 x 10^{-81}
Microtubule cytoskeleton	53.03	259	1.2 x10^{-67}	2.4 x 10^{-65}

Table 3. List of the top functional clusters of the down-regulated genes.

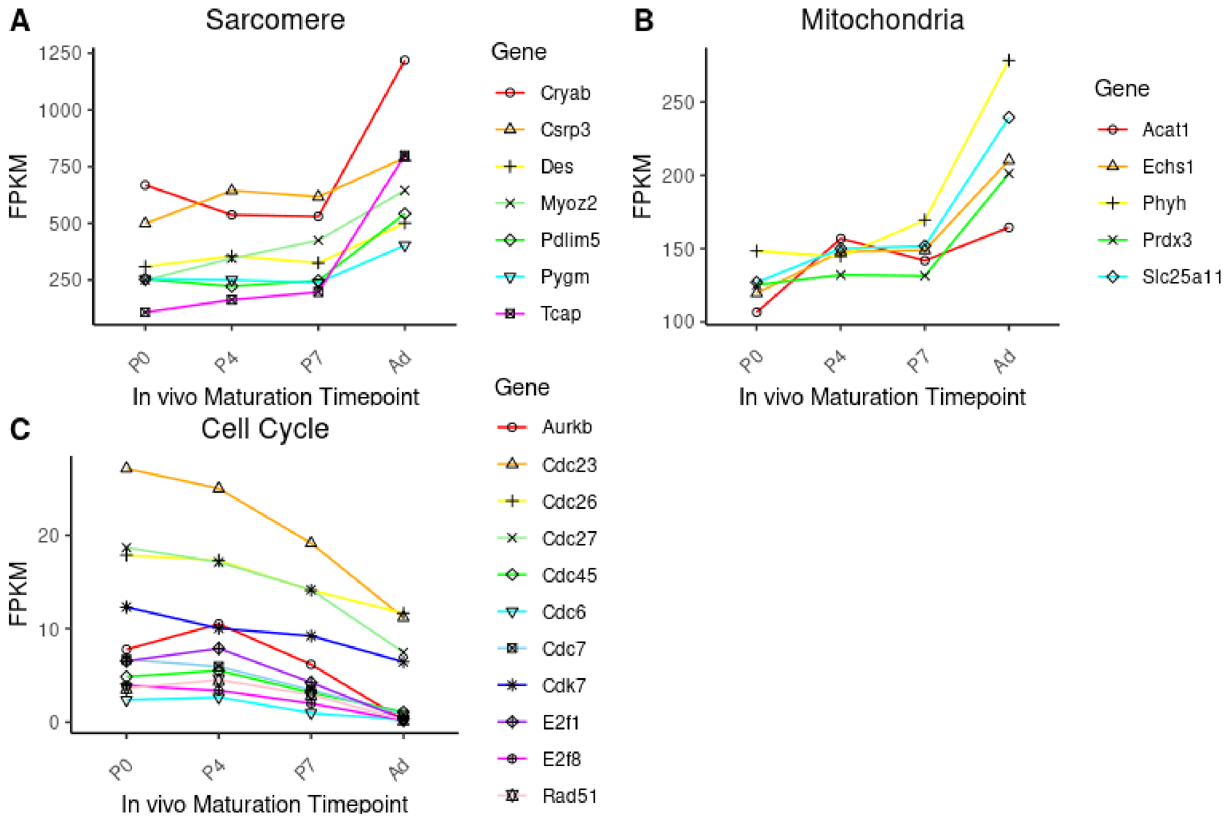


Fig. 10. FPKM values of selected genes across different in vivo maturation timepoints: This is an attempt to replicate the FPKM plots of selected genes at each timepoint in Figure 1D of the paper. Genes that were selected partakes in sarcomere (A), mitochondrial (B), and cell cycle (C) gene expression, respectively. In vivo maturation time points include postnatal Day 0 (P0), Day 4 (P4), Day 7 (P7), as well as adult (Ad) samples. The mean of duplicates for each timepoint were used.

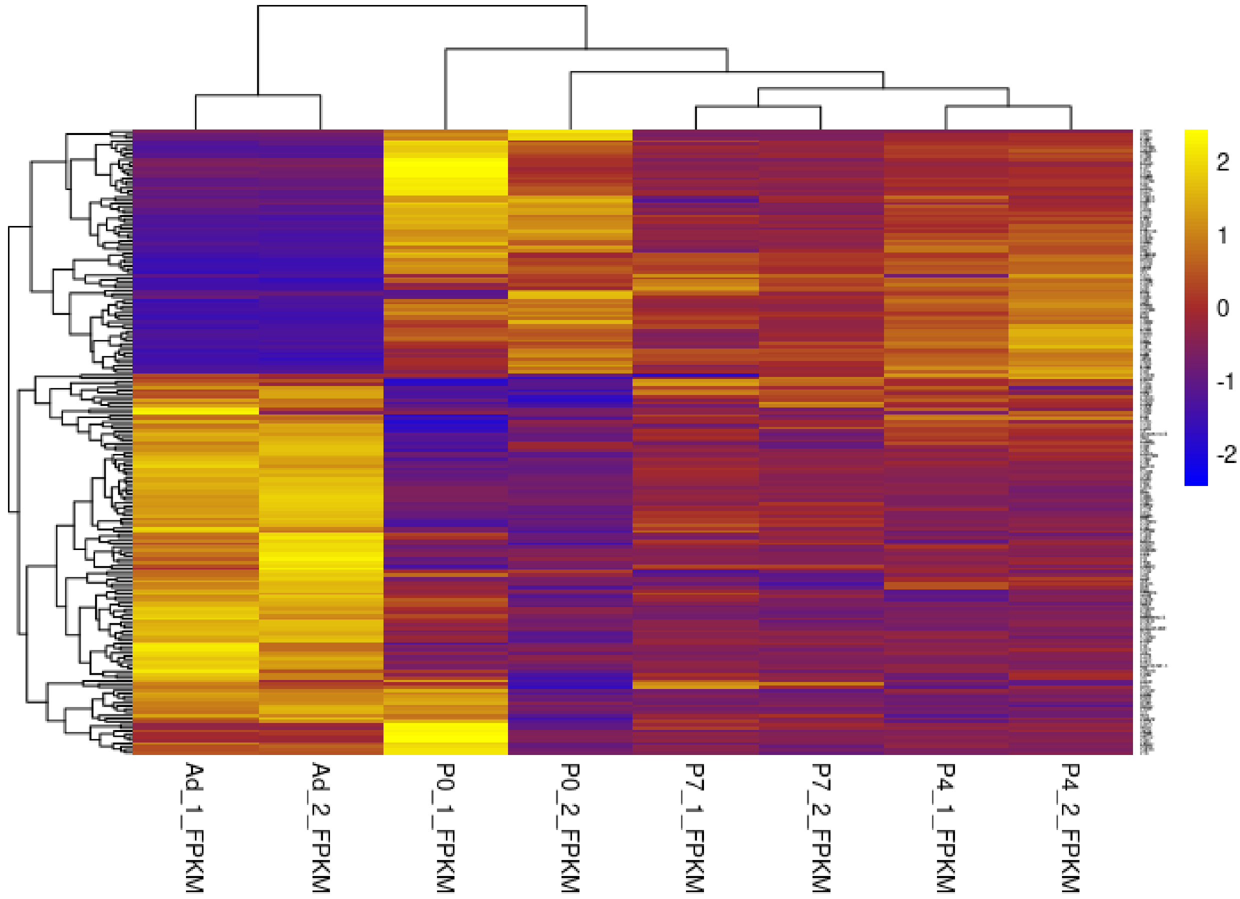
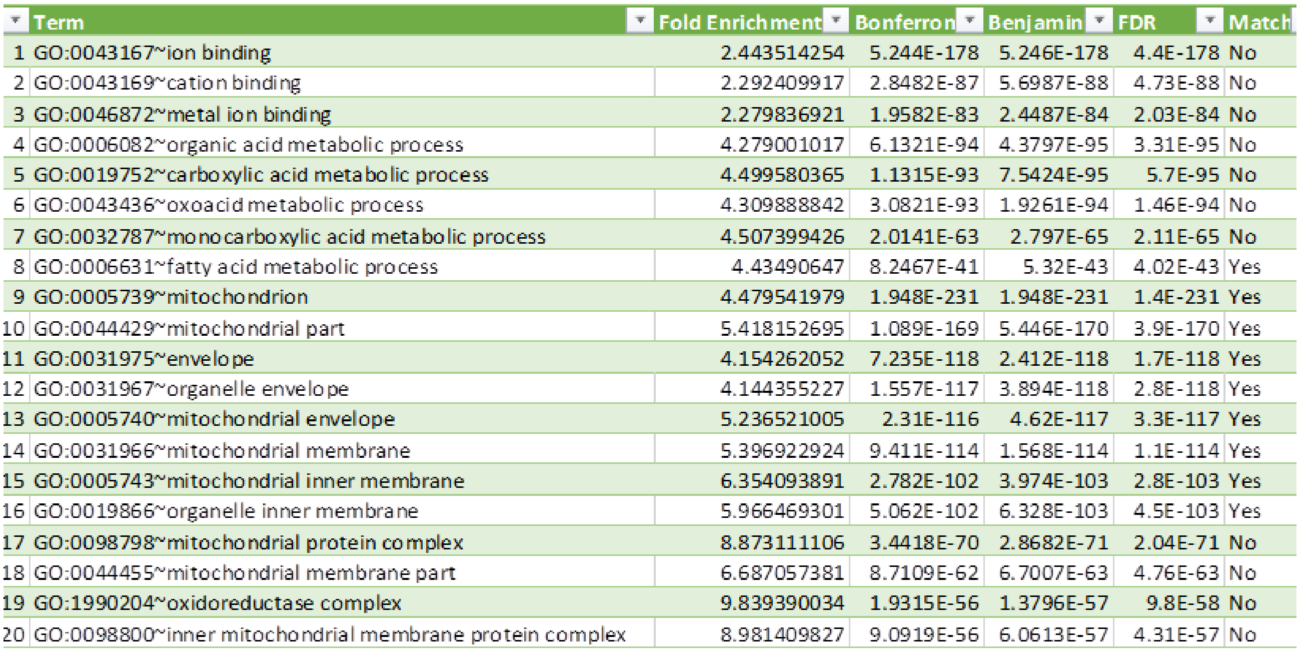
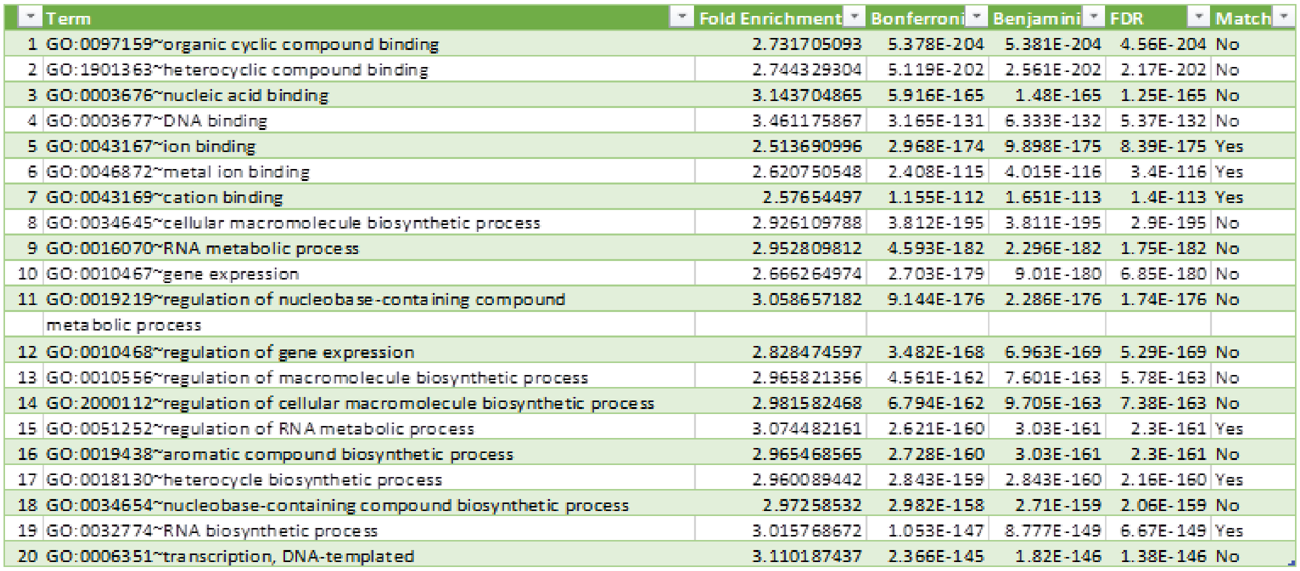


Fig. 11. Cluster Heatmap for differentially expressed genes of samples P0, P4, P7, and Ad: Plotted are 200 of the top 1000 genes that were determined to be differentially expressed when comparing postnatal Day 0 (P0) gene expression and adult (Ad) myocyte gene expression. Column names refer to samples from the different in vivo maturation timepoints with genes as rows. Data from each timepoint were obtained in duplicate, as specified by '_1_' and '_2_'. Upregulation of genes are depicted on the yellow end of the spectrum while downregulated genes are blue.



	Term	Fold Enrichment	Bonferron	Benjamin	FDR	Match
1	GO:0043167~ion binding	2.443514254	5.244E-178	5.246E-178	4.4E-178	No
2	GO:0043169~cation binding	2.292409917	2.8482E-87	5.6987E-88	4.73E-88	No
3	GO:0046872~metal ion binding	2.279836921	1.9582E-83	2.4487E-84	2.03E-84	No
4	GO:0006082~organic acid metabolic process	4.279001017	6.1321E-94	4.3797E-95	3.31E-95	No
5	GO:0019752~carboxylic acid metabolic process	4.499580365	1.1315E-93	7.5424E-95	5.7E-95	No
6	GO:0043436~oxoacid metabolic process	4.309888842	3.0821E-93	1.9261E-94	1.46E-94	No
7	GO:0032787~monocarboxylic acid metabolic process	4.507399426	2.0141E-63	2.797E-65	2.11E-65	No
8	GO:0006631~fatty acid metabolic process	4.43490647	8.2467E-41	5.32E-43	4.02E-43	Yes
9	GO:0005739~mitochondrion	4.479541979	1.948E-231	1.948E-231	1.4E-231	Yes
10	GO:0044429~mitochondrial part	5.418152695	1.089E-169	5.446E-170	3.9E-170	Yes
11	GO:0031975~envelope	4.154262052	7.235E-118	2.412E-118	1.7E-118	Yes
12	GO:0031967~organelle envelope	4.144355227	1.557E-117	3.894E-118	2.8E-118	Yes
13	GO:0005740~mitochondrial envelope	5.236521005	2.31E-116	4.62E-117	3.3E-117	Yes
14	GO:0031966~mitochondrial membrane	5.396922924	9.411E-114	1.568E-114	1.1E-114	Yes
15	GO:0005743~mitochondrial inner membrane	6.354093891	2.782E-102	3.974E-103	2.8E-103	Yes
16	GO:0019866~organelle inner membrane	5.966469301	5.062E-102	6.328E-103	4.5E-103	Yes
17	GO:0098798~mitochondrial protein complex	8.873111106	3.4418E-70	2.8682E-71	2.04E-71	No
18	GO:0044455~mitochondrial membrane part	6.687057381	8.7109E-62	6.7007E-63	4.76E-63	No
19	GO:1990204~oxidoreductase complex	9.839390034	1.9315E-56	1.3796E-57	9.8E-58	No
20	GO:0098800~inner mitochondrial membrane protein complex	8.981409827	9.0919E-56	6.0613E-57	4.31E-57	No

Table 4. Top 20 Upregulated Genes from DAVID: Aggregated table shows top enrichment score cluster of upregulated genes from DAVID analysis with a column indicating its match for similar GO Terms in O'Meara et al. (2015).



	Term	Fold Enrichment	Bonferroni	Benjamini	FDR	Match
1	GO:0097159~organic cyclic compound binding	2.731705093	5.378E-204	5.381E-204	4.56E-204	No
2	GO:1901363~heterocyclic compound binding	2.744329304	5.119E-202	2.561E-202	2.17E-202	No
3	GO:0003676~nucleic acid binding	3.143704865	5.916E-165	1.48E-165	1.25E-165	No
4	GO:0003677~DNA binding	3.461175867	3.165E-131	6.333E-132	5.37E-132	No
5	GO:0043167~ion binding	2.513690996	2.968E-174	9.898E-175	8.39E-175	Yes
6	GO:0046872~metal ion binding	2.620750548	2.408E-115	4.015E-116	3.4E-116	Yes
7	GO:0043169~cation binding	2.57654497	1.155E-112	1.651E-113	1.4E-113	Yes
8	GO:0034645~cellular macromolecule biosynthetic process	2.926109788	3.812E-195	3.811E-195	2.9E-195	No
9	GO:0016070~RNA metabolic process	2.952809812	4.593E-182	2.296E-182	1.75E-182	No
10	GO:0010467~gene expression	2.666264974	2.703E-179	9.01E-180	6.85E-180	No
11	GO:0019219~regulation of nucleobase-containing compound metabolic process	3.058657182	9.144E-176	2.286E-176	1.74E-176	No
12	GO:0010468~regulation of gene expression	2.828474597	3.482E-168	6.963E-169	5.29E-169	No
13	GO:0010556~regulation of macromolecule biosynthetic process	2.965821356	4.561E-162	7.601E-163	5.78E-163	No
14	GO:2000112~regulation of cellular macromolecule biosynthetic process	2.981582468	6.794E-162	9.705E-163	7.38E-163	No
15	GO:0051252~regulation of RNA metabolic process	3.074482161	2.621E-160	3.03E-161	2.3E-161	Yes
16	GO:0019438~aromatic compound biosynthetic process	2.965468565	2.728E-160	3.03E-161	2.3E-161	No
17	GO:0018130~heterocycle biosynthetic process	2.960089442	2.843E-159	2.843E-160	2.16E-160	Yes
18	GO:0034654~nucleobase-containing compound biosynthetic process	2.97258532	2.982E-158	2.71E-159	2.06E-159	No
19	GO:0032774~RNA biosynthetic process	3.015768672	1.053E-147	8.777E-149	6.67E-149	Yes
20	GO:0006351~transcription, DNA-templated	3.110187437	2.366E-145	1.82E-146	1.38E-146	No

Table 5. Top 20 Downregulated Genes from DAVID: Aggregated table shows top enrichment score cluster of downregulated genes from DAVID analysis with a column indicating its match for similar GO Terms in O'Meara et al. (2015).

DISCUSSION (shared)

The FastQC reports indicated that the sequencing reads were of satisfactory quality, although the distribution of per base sequence content at the 5' end appeared to be noisy (**Figures 1-4**). This may be the result of improper operation during library preparation, as well as overrepresented sequences, both of which can lead to biased composition and potentially affect downstream analysis.

The coverage for our accepted hits is predisposed to a 3' bias (**Figure 5**). It has been well documented that this uneven coverage uniformity can generally be attributed to biases stemming from multiple steps in the RNA-seq protocols used to prepare libraries for sequencing (Finotello et al. 2014). In terms of the 3' bias we observe in our reads, reverse transcription facilitated via poly-dT oligomers generally favors the 3' end of mRNA transcripts since they bind to poly-A tails (Nagalakshmi et al. 2008). While there have been multiple documented attempts to reduce bias in "transcript abundances," this issue has persisted throughout the history of the use of NGS technologies (Griebel et al. 2011). One such methodology proposed by Finotello et al. called *maxcounts* is designed to reduce such systematic errors by aligning reads to an exon and "exploit[ing] read coverage" to ascertain a count for every position in a given sequence (2014). The resulting counts are then "quantified as the [max] of the 'positional' counts" (Finotello et al. 2014). It would be interesting to do a side by side analysis of histograms generated through the *maxcounts* method and the *geneBody_coverage.py* method utilized in this paper to visually assess how our coverage uniformity analysis methodology shifts insert sizes.

When looking at the mean insert size (MIS) results from **Figure 6**, note how there were several insert size values below 0 bp. As far as we are aware, there were no steps in our methodology that involved normalizing the data input into our alignments, so our initial thought is that perhaps these negative inserts are indicative of artifacts in the mRNA seq technology used to generate the read data (namely PCR). Optical duplication is a known byproduct of PCR (Aird et al. 2011). Perhaps these negative inserts are a result of using PCR during amplification of RNA samples on sequencer to generate reads. However, this is no longer a consideration after we inspect the mapping statistics generated via both the *samtool* and *flagstat* commands and the *bam_stats.py* program. The mapping statistics summaries demonstrated that 100% of the paired reads present in the accepted hits passed the QC, and 0 reads were a result of optical/PCR duplication, which disproves our initial hypothesis on the potential cause of the reads with negative bp length reads represented in **Figure 6**. The veracity of this conclusion is supported by the fact that our alignment had a demonstrably high performing alignment, with a 95.9% read mapping rate to the mm9 reference genome, along with an 88.9% concordant pair alignment rate. Thus, we are now inclined to think that these negative reads are simply an artifact of the *inner_distance.py* command. To account for these artificially negative inserts, a better representation of the insert size data would be a histogram of the absolute values of the insert sizes.

As illustrated by **Figure 5**, it is evident that a filter of FPKM > 1000 was not enough to entirely remove the left-skewed observed in **Figure S.1**. The FPKM values in the filtered dataset used to generate **Figure 5** ranged from 1000.71 to 2604770, a 3 order of magnitude difference. Perhaps a revised FPKM threshold would yield a more informative histogram distribution. That being said, determining a FPKM threshold given the initial expression range of the dataset was difficult. Setting the FPKM threshold to 1000 already effectively removed 99% of all expression values so setting the threshold another order of magnitude or 2 higher would effectively exclude even more potential genes of interest with non-insignificant levels of gene expression.

Figure 8 and **Figure 9** illustrate histograms showing the change in the log 2 Fold change values before and after the removal of non-significant genes. This results in the drop of ~31,000 genes from the dataset and also stops the skewing of data which was observed as the massive spike in **Figure 8** due to the large number of genes (~15,000) with a Log 2 Fold change value of 0.

In the interest of identifying a handful of differentially expressed genes, we first filtered by significance, then genes with FPKM > 1000 for both samples. Through these means, we were able to isolate the 19 genes found in **Table S.1**. While this additional filtration was not really important for downstream analysis (namely the biologist portions of the paper), it was interesting to see which genes were isolated by this additional filter.

As previously mentioned, our initial cuffdiff results found a total of 5427 significant differentially expressed genes of which 2830 were upregulated. This differs from O'Meara et al. who found 2253 upregulated genes (2015). This difference could likely be due to differences in data processing. However, analysis of enriched gene ontology found results matching those of O'Meara et al. as observed in **Tables 4 and 5** (2015). For simplicity, only the top 20 Terms are shown in each table, but despite these differences, overall, the Gene Ontology (GO) terms that were present in our results correlates to the pathways that were differentially expressed during myocyte maturation. Genes found in these tables are involved in pathways found to be of transcriptional significance from the earlier DAVID results in **Tables 2 and 3**. This includes mitochondrial and cellular processes that are upregulated, while genes corresponding to regulations of gene expressions or RNA metabolic activities are observed as downregulated when comparing postnatal Day 0 (P0) and adult (Ad) level of expressions. While not all top functional clusters from our DAVID results correlate to those listed in Figure 1C of the original study, there are some clusters in our table that fall under the general umbrella of metabolism and cellular processes in theirs, so generally, these are a decent match (O'Meara et al. 2015). Possibly of note is that ion-binding was observed to be enriched in both up-regulated and down-regulated clusters, but we regarded this to be a result of cellular regulation due to ligand binding and cofactor activity. To reiterate, this is likely due to the differences in our utility of the tools provided for our analysis.

As **Figure 10** illustrates, there is a general trend of upregulation of sarcomere and mitochondrial genes as cells reach full maturity. On the contrary, cell cycle gene expressions are downregulated as cells prepare to reach senescence as it matures. Of note, however, there is one gene missing from the Mitochondria and Cell Cycle subplot, Mpc1 and Bora, respectively, but is present in the original study (O'Meara et al. 2015). While these genes are representative of their respective pathways, their data were not found in the supplemental files nor data supplied. Regardless, the pattern of expression and its magnitude closely resemble that of Figure 1D (O'Meara et al. 2015). To that end, we were successful in reproducing the trend observed in this data. We observed up-regulation of mitochondrial and sarcomere associated genes while seeing down-regulation of genes associated with cell cycle processes, which is consistent with our other findings.

As the authors of the original study performed hierarchical clustering with DAVID annotations in the form of a heatmap in Figure 2A, we also attempted to reproduce this plot, but with less success (O'Meara et al. 2015). As observed in **Figure 11**, the heatmap does not resemble that of Figure 2A in O'Meara et al.'s study, primarily due to the size of the gene set that was plotted (2015). In order to avoid having overlapping gene names in the rows, our heatmap is only a subset of 200 from our original subset of 1000 most differentiated genes from the P0 vs Ad analysis. In addition, there were more samples plotted in the heatmap in the original study. Heatmaps will also appear differently depending on how the rows of genes were clustered or if they were properly normalized, so this would not be a fair comparison. More details from the authors would be necessary in order to achieve a better reproduction of this heatmap.

When observing **Figure 11**, there is a clear difference between the adult (Ad) samples and the postnatal samples (P0,4,7). That being said, it is also worth noting that there are also minor differences observed, particularly those between the duplicates of P0_1 and P0_2, which is an indication of a difference in how the data was processed. Although very subtle, the clustering algorithm groups P0_1, the data processed separately, as farther apart from the rest of the postnatal samples. Likewise, due to having only 2 replicates, standard deviation would not be meaningful and thus was not performed on the line plots in **Figure 10**. However, the overall trend is similar between these two replicates.

Although the number of replicates is limited, the authors have for this reason included data from *in vitro* observations as well as those from an explant model, which can capture live transcriptional activities through RNA Seq data. As with many other studies in angiogenesis, obtaining *in vitro* data can be difficult due to the nature of the assay, but obtaining accurate *in vivo* data can be extremely time consuming and difficult to assess (Staton et al. 2009). This is due to the difference in the microenvironment of the cells in question. For this reason, multiple collections were made to be pooled into one replicate for our analysis for both *in vitro* and *in vivo* models, but especially for the explant model, which is a direct observation of how this would unravel in real time. The purpose of this is to accurately compare whether what is observed in the laboratory is also true in its niche environment. This point is further shown in the complete Figure 1D of the original study, which shows the exacerbated variation of gene expression levels of the same representative genes, due to the undisturbed, isolated environment of *in vitro* studies (O'Meara et al. 2015). Despite these difficulties, these studies are done in this manner to potentially improve human health.

CONCLUSION (shared)

O'Meara et al. had elucidated the transcriptional signature of cardiac regeneration using Tuxedo Suite to analyze RNA Seq datasets based on the *in vitro* differentiation of embryonic stem cells and the *in vivo* maturation of postnatal left ventricular cells to serve as a basis for comparison with the explant model (2015). In this replication study, similar bioinformatic tools were also used and results have been determined to be comparable to those with O'Meara et al. (2015).

As mentioned above, one of the difficulties we encountered is the possibility of not having the complete dataset to get reproducible results posted by the original study. As heatmaps can be misleading for many reasons, it would be beneficial to use another algorithmic tool for clustering or for other analyses to be run concurrently to confirm our replication findings. Overall, our study provides some insight into the transcriptional signature of cardiac myocytes in reversed transcriptional differentiation with implications for potential *ex vivo* therapies for cellular damage.

Surprisingly, when only working with two of the samples included in O'Meara et al.'s analysis, we obtain a similar number of upregulated genes (2015). When performing our analysis on the reads generated from the *TopHat* analysis, it became increasingly evident that there are several aspects of the benchtop workflow for RNA-seq that have a significant impact on the biases in read mapping data returned in our accepted hits. Additionally, the importance of the selection process for the computational methodologies utilized in order to create alignment files as well as to analyze the accepted reads during the experimental design phase of a study cannot be understated. The tools used to generate and process reads always have assumptions and conditions underlying the code. It is up to the user to determine whether or not the tools they have at their disposal are appropriate for tackling the dataset on their hands.

REFERENCES

1. Porrello, E. R., Mahmoud, A. I., Simpson, E., Hill, J. A., Richardson, J. A., Olson, E. N., & Sadek, H. A. (2011). Transient regenerative potential of the neonatal mouse heart. Science (New York, N.Y.), 331(6020), 1078–1080. https://doi.org/10.1126/science.1200708
2. Steinhauser, M. L., & Lee, R. T. (2011). Regeneration of the heart. EMBO molecular medicine, 3(12), 701–712. https://doi.org/10.1002/emmm.201100175
3. 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
4. SRA-Tools. (n.d.). Retrieved March 17, 2021, from Github.io website: http://ncbi.github.io/sra-tools/
5. FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2015), "FastQC,"https://qubeshub.org/resources/fastqc
6. Wang, L., Wang, S., & Li, W. (2012). RSeQC: quality control of RNA-seq experiments. Bioinformatics (Oxford, England), 28(16), 2184–2185. https://doi.org/10.1093/bioinformatics/bts356
7. RNA-seq quality metrics with RseQC. Retrieved March 17, 2022 from: https://chipster.csc.fi/manual/rseqc.html.
8. Aird, D., Ross, M. G., Chen, W. S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D. B., Nusbaum, C., & Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome biology, 12(2), R18. https://doi.org/10.1186/gb-2011-12-2-r18
9. Finotello, F., Lavezzo, E., Bianco, L. et al. Reducing bias in RNA sequencing data: a novel approach to compute counts. BMC Bioinformatics 15, S7 (2014). https://doi.org/10.1186/1471-2105-15-S1-S7
10. Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., & Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. Science (New York, N.Y.), 320(5881), 1344–1349. https://doi.org/10.1126/science.1158441
11. Griebel, T., Zacher, B., Ribeca, P., Raineri, E., Lacroix, V., Guigó, R., & Sammeth, M. (2012). Modelling and simulating generic RNA-Seq experiments with the flux simulator. Nucleic acids research, 40(20), 10073–10083. https://doi.org/10.1093/nar/gks666
12. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
13. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
14. Staton, C. A., Reed, M. W., & Brown, N. J. (2009). A critical analysis of current in vitro and in vivo angiogenesis assays. International journal of experimental pathology, 90(3), 195–221. https://doi.org/10.1111/j.1365-2613.2008.00633.

Project 2 3/18/22
Group: Tinman

SUPPLEMENTARY FIGURES

Programmer

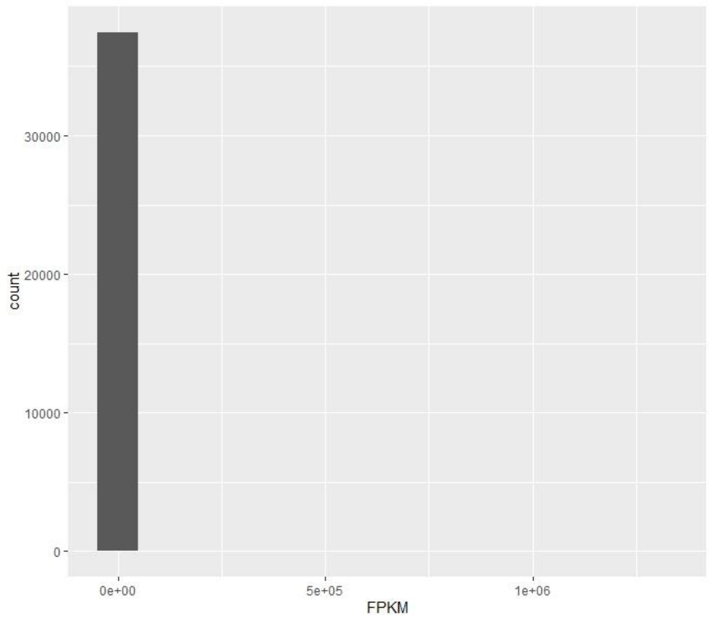


Figure S.1. The full gene.fpkm_tracking dataset (n = 37469) depicting a histogram of FPKM values with no value ranges omitted.

gene	sample_1	sample_2	value_1	value_2	Q_value (p-adj)
Pln	P0	Ad	1455.52	3254.99	0.000321
Mb	P0	Ad	1537	9934.16	0.000321
Atp5g3	P0	Ad	1151.85	2230.93	0.000321
Gm12563	P0	Ad	301714	118967	0.000321
Fabp3	P0	Ad	1006.16	1929.1	0.000321
Mir5105	P0	Ad	5709940	1940320	0.000321
Atp2a2	P0	Ad	1464.07	3641.56	0.000321
6030429G01Rik,Tnni3	P0	Ad	1668.25	4835.63	0.000321
Slc25a4	P0	Ad	1615.6	2961.93	0.000321
Myl3	P0	Ad	2615.11	5379.66	0.000321
Tpm1	P0	Ad	1827.85	3941.48	0.000321
mt-Cytb	P0	Ad	11818.3	26337.5	0.000321
Actc1	P0	Ad	5985.26	10436.6	0.003688
Gapdh	P0	Ad	1035.42	1588.79	0.006269

Atp5a1	P0	Ad	1508.97	2342.54	0.009913
Gm15662	P0	Ad	1919.47	1089.56	0.014195
Tnnc1	P0	Ad	1800.93	2736.99	0.015764
Tnnt2	P0	Ad	3003	4464.61	0.026413
Atp5b	P0	Ad	2132.13	3230.42	0.0421

Table S.1. 19 differentially expressed (and in this instance upregulated) genes between the Ad samples and P0. This list was generated using cuffdiff and was filtered for significant Q-values (p-adj values that account for FDR) and FPKM >1000 across both sample sets. Highlighted in red are the 12 lowest Q-value samples.