BF528 Project 2: Group Frizzled

Camilla Belamarich - *Data Curator*
Yashrajsinh Jadeja - *Programmer*
Zhuorui Sun - *Analyst*
Janvee Patel - *Biologist*

Transcriptional Profile of Mammalian Cardiac Regeneration with mRNA-seq

Introduction

Although there is evidence that some adult vertebrates can fully regenerate their hearts after injury, adult mammals lack this bewildering ability shortly after birth.[1-3] However, it was found that neonatal mice can fully regenerate their heart after injury.[4] The reason why mammals cannot fully regenerate their heart after an injury as an adult is due to cardiac myocytes (CM) leaving the cell cycle.[5] This occurs shortly after birth and heart growth persists. [5] When the left ventricle apex was resected from neonatal mice, genetic fate mapping showed that already established CMs gave rise to the new and regenerating CMs, which produced a fully regenerated heart.[4] More broadly, this study found that identifying the mechanisms in which myocytes undergo cell cycle activity during regeneration will give a deeper understanding of why heart regeneration is limited in adult hearts.[4] The transcriptional changes that cause a regenerative heart phenotype were generally unknown until O'Meara et al. (2015) aimed to identify the genes and gene networks that change throughout the regeneration process. Specifically, researchers global gene expression patterns at different time points of mouse CM differentiation in both in vivo and in vitro. Additionally in neonatal mice, researchers observed global gene expression patterns in whole heart ventricles and purified CMs after apical resection. This revealed that heart regeneration can be characterized by a transcription reversion of the CM differentiation process.[6] Using RNA-sequencing datasets, researchers were able to accurately predict regulators and associated pathways that control the cell cycle state of CMs.[6] RNA-seq analysis is essential in observing gene expression patterns and provides a deep understand of the transcriptome.[7]

From this analysis, O'Meara et al. (2015) found a significant regulator of the CM cell cycle and linked it to the regulation of differential gene expression during heart regeneration in zebrafish.[8] Overall, researchers concluded that cardiac regeneration is a regulated process in which transcriptional reversion of the differentiation process occurs. Our goal for this project is to reanalyze just one sample of the data and replicate the results found in the paper using the same bioinformatics tools. Reanalyzing and focusing on just one sample of the data instead of the 36 examined in the paper could provide different insights as to how neonatal mice are able to regenerate their heart tissue but lose the ability later in life at a transcriptional level.

Data

O'Meara et al. (2015) obtained RNA-seq data of differentiation of mouse embryonic stem cells into cardiac myocytes from Wamstad et al. (2012).[9] The raw data contained samples from each time point of development. These time points include embryonic stem cell, mesoderm, cardiac progenitor, and cardiac myocytes. For RNA sequencing and analysis, researchers used Thermo Fisher's Invitrogen technology to extract RNA from the samples using Trizol. Using the Dynabeads mRNA purification kit, polyadenylated RNA was isolated from RNA samples and then fragmented. The first strand was then synthesized using Superscript III reverse transcription kit. Double-stranded DNA was synthesized with DNA polymerase I. Researchers then performed a paired-end 40 base pair read length sequencing using the Illumina HiSeq 2000 technology. Due to a low RNA yield, they used TrueSeq for sample preparation. The raw RNA-Seq we used for our analysis is available at NCBI Gene Expression Omnibus, GEO accession number GSE64403. This dataset contains RNA-Seq samples of *Mus musulus* from different time points. In this project we use the sample GSM1570702**,** which is postnatal day 0 from the ventricular myocardium.[10]

We downloaded the sample GSM1570702 from NCBI and used the SRAToolkit to extract SRA format to FASTQ format.[11] In total, there were 49,706,999 reads that passed quality control with 20,878,784 for read one and 20,510,550 for read two. We named the file P0_1.sra before converting to FASTQ files. To convert the files, we used the "fastq-dump" command. This left us with P0_1_1.fastq and P0_1_2.fastq, which represents the paired-end reads. In order to double check that it extracted the information and reformatted it correctly, we examined the header line and length of the reads were exactly the same between the two files. These files will be used in the FastQC analysis.

After processing the FASTQ files, we extracted the quality metrics using the FastQC package on the command line.[12] This package is used as a quality control tool for high throughput sequence data. Using the "fastqc" command and inputting the fastq files we processed, we directed the results into a directory called fastqc_results using the output parameter. This produced html and image files with read statistics. In order to determine the overall quality of the reads, we must analyze each of the FastQC modules to see if they passed or failed, and if they failed, what might be the reason. The first quality metric, per base read quality, displayed good read quality on the basis of their phred score (Fig. S1). Both reads had a phred score of 30 or higher across all positions in the read. Additionally, every position across the read stayed in the green zone of the graph, which represents high quality reads. Per base sequence content is the next quality metric we analyzed (Fig. S2). This is the only metric that failed during the FastQC analysis due to the nucleotide distribution being non-uniform across all read positions. Non-uniform distribution of nucleotides can be seen in the fluctuations in the graph generated by FastQC. The uneven distribution of nucleotides across the read could indicate

issues with library preparation or contaminated samples. Next, the per sequence GC content showed the fraction of Guanine and Cytosine across the read (Fig. S3). The GC content found in our reads were distributed normally (Gaussian) around the target mean. Our data looks good and matches the theoretical distribution as we would expect. The sequence duplication levels module displayed a warning sign that this metric should be carefully considered (Fig. S4). This metric revealed that there were 50.4% and 51.82% unique sequences in read one and read two, respectively. Furthermore, about half of the sequences were unique in the read. A lower percentage of unique sequences could indicate possible PCR bias or contamination in the samples. Lastly, there were no overrepresented sequences found in either of the reads. Overrepresented sequences could be caused by a variety of reasons such as contamination, issues with library preparation, or having a low complexity library. No overrepresented sequences found in these reads was slightly shocking due to the non-uniform nucleotide distribution and lower percentage of unique sequences, which could be explained by contamination or issues in library preparation. Overall we determined the data to be of high quality; however, we should consider sources of error, which were possible contamination and library preparation challenges.

Methods

The sequenced paired-end reads from the postnatal day 0 (P0) sample were aligned against the mouse reference genome mm9 (MGSCv37) using TopHat v. 2.1.1 as it is a splicing-aware tool that allows for transcripts to be aligned to reference genomes based on knowledge of splicing junctions that results in better alignments.[13,14] The alignment was performed using the parameters detailed in the paper by O'Meara et al. (2015) where the expected inner distance between mate pairs parameter was set to 200, the read segment length parameter to 20, allowing for 1 mismatch in each segment alignment, and looking for reads across known junctions in the mm9 gene model annotation.[6] SAMtools v. 1.10 was used to evaluate the alignment quality of the genome-mapped reads generated by TopHat.[14,15] The mapped file was then indexed using SAMtools to facilitate quicker access to the alignments of the reads overlapping particular genomic regions.[15]

RSeQC package v. 3.0.0 was used to obtain various quality control metrics to further assess the quality of the mapping and obtain detailed information about the coverage uniformity over gene body, insert size, and read mapping statistics.[16] RSeQC module 'geneBody_coverage.py' was utilized to assess the uniformity of read coverage and the presence of 5'/3' bias. This module scales all transcripts to 100 nt and calculates the number of reads covering each nucleotide position to estimate read depth. A plot was generated by the module that illustrated the coverage profile along the gene body. The RSeQC module 'inner_distance.py' was utilized to calculate the inner distance (or insert size) between two paired RNA reads and visualize the results. Another RSeQC module 'bam_stat.py' was used to obtain detailed statistics like the number of unique reads, paired reads etc.[16]

Cufflinks v. 2.2.1 was used to estimate how the reads mapped to the genomic region defined by the mm9 gene annotation and to quantify their expression.[17] The expression level of each full-length transcript fragment was calculated by Cufflinks in terms of FPKM (Fragments Per Kilobase of transcript per Million mapped reads) after the sequence reads were normalized and assembled. Cuffdiff tool in the Cufflinks suite was used to calculate the differential gene expression.[18] The neonatal day 0 sample was compared to neonatal day 2, adult day 1 and 2 (P0_2, Ad_1, and Ad_2) samples to find differentially expressed genes at the transcriptional or post-transcriptional level. The differential expression values were obtained in terms of log normalized fold change values.

To obtain the difference between the samples of P0 mice and Adult mice in gene level, we did a gene differentially expressed analysis based on the gene difference table we generated before. The top 10 differentially expressed genes were selected by sorting the table with q-value. We compared two selected methods by using p-value or labeled 'significant' to filter out the significant gene and report the numbers of significant genes. We used two histograms to visualize the distribution of log2_fold_change across all genes and significant genes. By the value of log2_fold_change, the gene set was splitted to a up-regulated gene set and a down-regulated gene set. The names of these genes were written in two .csv files for the following enrichment analysis.

An enrichment analysis was performed to cluster genes in up-regulated gene set and down-regulated gene set with similarly functional roles based on the DAVID Functional Annotation Tool.[18,19] With the DAVID Functional Annotation Tool, OFFICIAL_GENE_SYMBOL was selected as the grouping identifier and Mus Musculus was selected as the species. The GOTERM_BP_FAT, GOTERM_MF_FAT and GOTERM_CC_FAT were selected to obtain Gene Ontology (GO) terms results in the GO group. The boxes we used here BP indicated for biological processes, MF indicated for molecular function and CC indicated for cellular components. These functions in DAVID helped to group genes in functionally related clusters and to analyze the functional roles. The GO terms from the annotation clusters that showed overlap with the common GO terms identified in O'Meara et al. (2015) were determined using the results from Supplemental 2 Tables 1D and 1E.

To compare the trends between our results and O'Meara et al. (2015), the genes.fpkm_tracking tables for the replicates of the samples were utilized. For the Gene Ontology terms identified in Figure 1D in the paper (Sarcomere, Mitochondria, and Cell Cycle), the FPKM values for representative genes from the replicates of the samples were extracted, and the average values were computed between the two replicates for P0, P4, P7, and Ad, and visualized as a line plot. The sarcomere-specific genes were Pdlim5, Pygm, Myoz2, Des, Csrp3, Tcap, and Cryab. The mitochondria-specific genes were Prdx3, Acat1, Echsw1, Slc25a11, and

Phyh. The cell cycle-specific genes were Cdc7, E2f8, Cdk7, Cdc26, Cdc6, E2f1, Cdc27, Cdc45, Rad51, Aurkb, and Cdc23. Gene Mpc1 was omitted from the mitochondria plot and gene Bora was omitted from the cell cycle plot since there were no FPKM values associated with these genes which indicated that these genes were possibly filtered out due to low signal.

An FPKM matrix was generated from the FPKM columns of each of the replicates for P0, P4, P7, and Ad. Duplicated Ensembl tracking ids within the matrix were addressed so that the corresponding FPKM values were appropriately aligned. From the Cuffdiff output file gene_exp.diff, the top 1000 significantly differentially expressed genes between P0 and Ad were selected by q-value. Within the top 1000 significantly differentially expressed genes, there were entries specifying more than one gene which possibly indicated that there were multiple overlapping splicing transcripts mapping to approximately the same region. For these situations, the first gene symbol listed was utilized. In order to map the gene symbol to the Ensembl tracking id, Bioconductor packages such as biomaRt v.2.46.3, EnsDb.Mmusculus.v79 v.2.99.0, and org.Mm.eg.db v.3.12.0 were utilized.[27,28,29] However, these methods did not return all Ensembl ids for the 1000 gene symbols that were inputted due to instances such as the gene symbol was the gene synonym for that given id, and the presence of different or deprecated ids for those gene symbols in the Ensembl database. Therefore, the corresponding gene symbol was used to subset the FPKM matrix. The FPKM values of the top 1000 differentially expressed genes from the P0 and Ad analysis were visualized as a clustered heatmap.

Results

Table 1: Summary statistic output from Samtools v. 1.10 'flagstat'. The counts for each read category are classified into the QC pass and QC fail subcategories.

Samtools 'flagstat' Summary		
Category	**QC Passed (reads)**	**QC Failed (reads)**
Total	**49,706,999**	0
Secondary	**8,317,665**	0
Supplementary	0	0
Duplicates	0	0
Mapped	49,706,999 **(100%)**	0
Paired in Sequencing	41,389,334	0
Read 1	20,878,784	0
Read 2	20,510,550	0
Properly Paired	29,422,646 (71.09%)	0
With Itself and Mate Mapped	39,936,472	0
Singletons	1,452,862 (3.51%)	0
With Mate Mapped to a Different Chr	1,387,382	0
With Mate Mapped to a Different Chr (mapQ>=5)	704,916	0

The summary statistics generated by Samtools for all the reads that were mapped to the genome is illustrated in Table 1. All the reads from the P0 sample mapped to the reference mm9 genome successfully. None of the reads in the sample were flagged as low quality as illustrated in Table 1. This is crucial as low quality reads could interfere with the interpretation of results and may impact the strength of the results obtained.[20] The reads marked as secondary are multi-mapping reads that could potentially align to multiple regions in the genome.[21]

Multi-mapping reads occur due to gene duplication which is a common occurrence in mammalian genomes. This could be due to repeating elements like retro-transposons, chimeric sequences, recombination and other factors. These reads could potentially cause trouble during further analysis by skewing expression values.[22] However, we account for those multi-mapped reads during further analysis in Cufflinks by equally splitting the multi-mapped reads between all their alignments.[23] 71.09% of reads are labelled as properly paired that are categorized as reads that fall in line with the expected insert size however that number could be influenced by the insert size parameters being in TopHat.[24]

Table 2: Summary of the read mapping statistics by the RseQC 'bam_stat.py' module. Uniquely mapped reads are determined from mapping quality that details the probability of a read being misaligned/misplaced.

RSeQC 'bam_stat.py' Summary	
Total records (reads)	**49,706,999**
QC Failed	0
Optical/PCR Duplicate	0
Non Primary Hits	**8,317,665**
Unmapped Reads	0
mapq < mapq_cut (Non-Unique)	2,899,954
mapq >= mapq_cut (Unique)	**38,489,380**
Non-splice Reads	33,099,839
Splice Reads	5,389,541
Reads Mapped in Proper Pairs	27,972,916
Proper-paired Reads Map to Different Chromosome	4

The summary statistics generated by RSeQC 'bam_stat.py' module are illustrated in Table 2.[25] Mapped reads were further summarised by the RseQC 'bam_stat.py' module that gave detailed statistics on the unique mapped reads. 77.4% of the reads were uniquely mapped. The non-primary hits were the same as observed in the previous Samtools flagstat output. Detailed statistics about the splice and non-splice reads were also obtained from the 'bam_stat.py' module.[25]

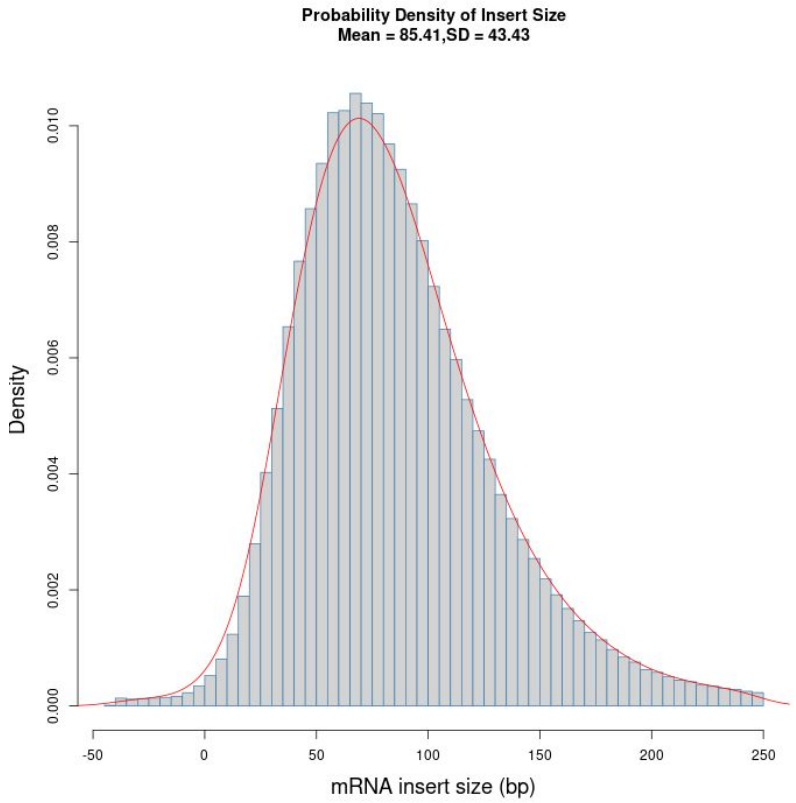

Figure 1: Probability density (y-axis) plot of the mRNA insert size (x-axis) from the P0 Sample generated by the RSeQC 'inner_distance.py' module. The mean length between two paired fragments is 85.41 base pairs and the standard deviation is 43.42.

Using the RseQC package, coverage uniformity over the gene body and insert size were assessed. The mRNA insert size had a mean of 85.41 and a standard deviation of 43.42 as detailed in Figure 1 where the data follows a normal distribution. There were some negative insert sizes observed which could be a result of overlapping due to the inner mate distance being shorter than the fragment lengths on either side of the reads. This could indicate over-fragmentation or degradation of the sample however, on some occasions it could be intentional as shorter overlapping reads could be sequenced as one combined longer read.[26] This would be more relevant for single-end sequencing however, one rationale behind it could be researchers intentionally choosing some reads to overlap to obtain greater coverage for certain regions of interest.[6]

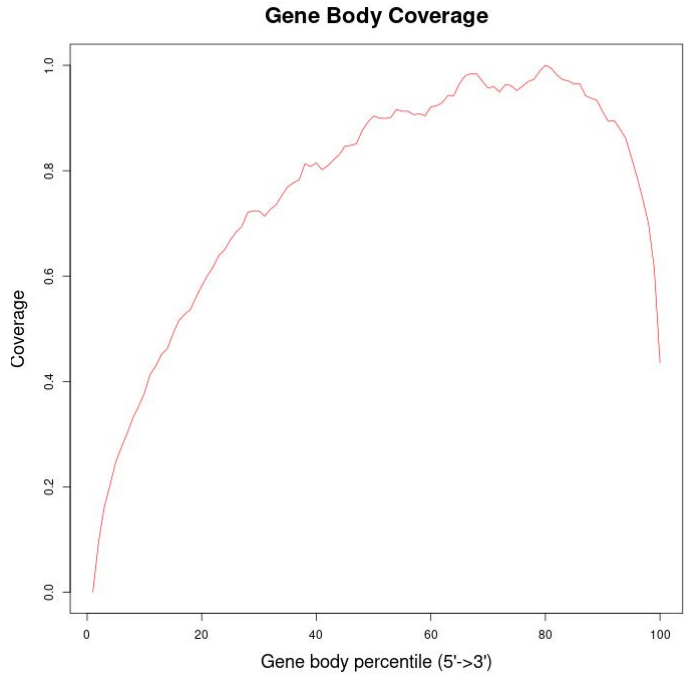

Figure 2: Line plot of the P0 reads coverage (y-axis) over gene bodies (x-axis) generated by RSeQC 'geneBody_coverage.py' module.

The gene coverage along the gene body (full transcript length) analysis showed lower coverage at the 5' end, an increased coverage in the middle, and peak coverage towards the 3'end in Figure 2. The high coverage at the 3' end of the gene body indicates a 3' bias, which is expected from the mRNA sequencing method O'Meara et al. (2015) utilized for their study. In the study, polyadenylated RNA were fragmented, resulting in the observed 3'end bias that could be a result of poly-A enrichment or RNA degradation.[6]

a) **b)**

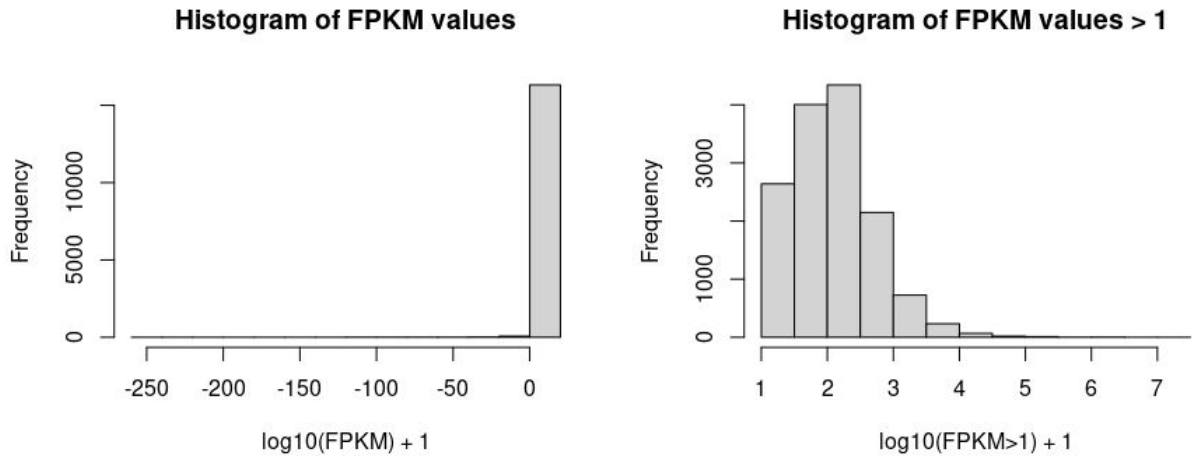

Figure 3: Histogram of log 10 normalized FPKM values before (a) and after (b.) removing FPKM values less than 1. The log 10 normalized FPKM values have a pseudocount of 1 added to them.

The log 10 normalized FPKM values of genes are visualized in Figure 3 where the FPKM values less than 1 were removed in accordance with the filtering criteria used by O'Meara et al.[6] The values were normalized to facilitate clear visualization with respect to the scale of the data and a pseudocount of 1 was added to prevent negative FPKM values. As expected most genes (23,264 out of total 37,469) had normalized FPKM values that were less than 1. This is due to the null hypothesis where transcripts are expected to follow a baseline range where most genes aren't over/under expressed while differentially expressed genes have transcripts that deviate from the normal range. We look at the transcripts deviating from the normal distribution to estimate the differential expression of genes. A total of 14,204 genes were observed with FPKM values greater than 1.

During gene differentially expressed analysis, we sorted the gene.diff table by q-value and the top 10 differentially expressed genes between P0 and Ad were shown in Table 3. From the table, all these 10 genes had the smallest q-value for 0.00106929 and p-value for 5e-5 which were much smaller than 0.05. The range of FPKM value_1 was from about 1.19 to 119.00, the range of FPKM value_2 was from about 0.69 to 185.33 Two genes Cspp1 and Pi15 had negative log2 fold change values and the other eight genes had positive values.

Table 3: Top 10 differentially expressed genes select from gene.diff table sorted by q-value. The table includes the Gene_name, FPKM value_1 and value_2, log fold change, p-value and q-value as the statistics.

Gene name	FPKM value 1	FPKM value 2	Log2 fold change	p-value	q-value
Plekhb2	22.56790	73.568300	1.70481	5e-05	0.00106929
Mrpl30	46.45470	133.038000	1.51794	5e-05	0.00106929
Coq10b	11.05830	53.300000	2.26901	5e-05	0.00106929
Aox1	1.18858	7.091360	2.57682	5e-05	0.00106929
Ndufb3	100.60900	265.235000	1.39851	5e-05	0.00106929
Sp100	2.13489	100.869000	5.56218	5e-05	0.00106929
Cxcr7	4.95844	32.275300	2.70247	5e-05	0.00106929
Lrrfip1	118.99700	24.640200	-2.27184	5e-05	0.00106929
Ramp1	13.20760	0.691287	-4.25594	5e-05	0.00106929
Gpc1	51.20620	185.329000	1.85570	5e-05	0.00106929

Two different methods were used to filter out the significant genes among the total 36329 genes. The first one was using P-value. Since small p-value indicates more differences by using p-value less than 0.01 as a condition, we got 2376 significant genes. By using the 'significant label' in the table, using 'significant == yes' as a condition, we got 2139 significant genes. To prepare for the following enrichment analysis, we splitted the gene set into an up-regulated set and a down-regulated set by the value of Log2_fold_change. The gene was grouped in the up-regulated gene set with a positive log2 fold value and grouped in the down-regulated gene set if the log2 fold change value was negative. By 'p-value' method, we got 1187 genes in the up-regulated gene set and 1189 genes in the down-regulated gene set. By 'labeled significant' method, we got 1084 genes in the up-regulated gene set and 1055 genes in the down-regulated gene set. The summary of the number of genes were shown in Table 4. Based on the results, there are about 2000+ genes among 36329 genes shows a clear difference between the P0 sample and the adult. In these 2000+ genes, about half of them are up-regulated and other half of them are down-regulated. Compared to the results in paper by O'Meara et al (2015), both our up-regulated gene set and down-regulated gene set were smaller. Based on our data, the number

of up-regulated genes and almost the same with the number of down-regulated genes with these two select methods.

Table 4: Summary of the number of significant genes based on different select methods. By using p-value less than 0.01 or Labeled significant to filter out genes. Genes were divided into up-regulated genes and down-regulated genes by log2_fold_change value.

	Total number of genes	**Up-regulated genes**	**Down-regulated genes**
P-value < 0.01	2376	1187	1189
Labeled significant	2139	1084	1055

To visualize the distribution of Log2_fold_change, the histograms of Log2_fold_change for genes were shown in Figure.4 and Figure.5. Figure.4 showed the distribution of log2 fold change for all genes (36329 genes in total) and Figure.5 showed the distribution of log2 fold change for the significant genes we selected by 'labeled significant' with 2139 genes. Most of the genes in the significant gene set had log2 fold change value around -4 to -2 and 1 to 2. To visualize the distribution of log2 fold change value for genes, we obtained two histograms. From these two histograms, there was an obvious peak in Figure.4 around 0. However, in Figure.5, there were almost no genes around 0 which indicates that the genes with log2 fold change value equal to 0 were not selected as significant genes. This difference could be explained. The values of log2_fold_change were calculated by log2 (FPKM value 2/ FPKM value 1). If the FPKM value 1 and FPKM value 2 are relatively close, the log2_fold_change will be close to log2(1) which is 0. Thus, if the log2_fold_change value of a gene is near 0, this gene has closely FPKM value 1 and FPKM value 2, which indicates that there are no significant differences between P0 and Ad sample for this gene.

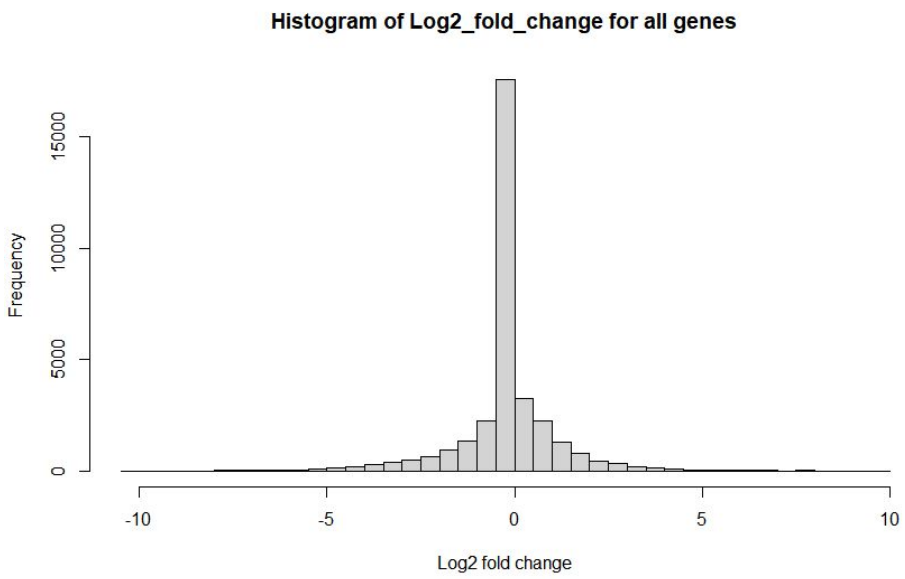

Figure 4: Histogram of the distribution of Log2_fold_change for all genes (36329 genes in total)

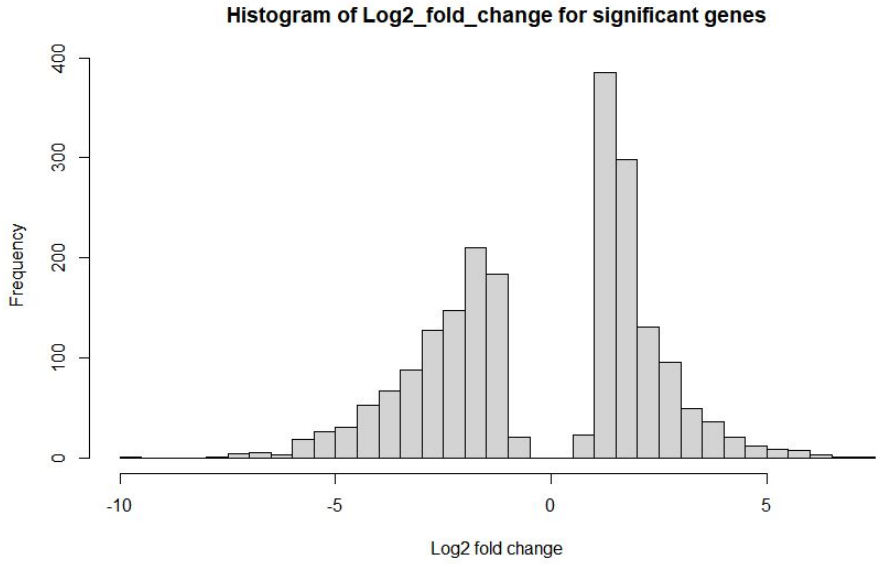

Figure 5: Histogram of the distribution based on Log2_fold_change for significant genes(2139 genes in total)

We did an enrichment analysis based on DAVID Functional Annotation Tool for further functional annotation clustering. By using the up-regulated gene set and down-regulated gene set we generated before, we got 411 annotation clusters for up-regulated gene set and 421 annotation clusters for down-regulated gene set. The summaries of these two tables are shown in Table 5 and Table 6. In each table, we selected the top five clusters for each gene set and the top three GO terms as the examples for each cluster. For each cluster, the cluster id, enrichment score, some examples of GO terms are included. In Table 5 and Table 6, some of the terms found

within the annotation cluster that overlapped with O'Meara et al (2015) were depicted by an asterisk. We also found some clusters not included in the paper by O'Meara et al (2015).

Table 5: Top five cluster results of up-regulated genes by DAVID analysis. Terms overlapping with O'Meara et al. (2015) denoted by " * "

DAVID Analysis Result on Up-Regulated Genes			
Cluster id	Enrichment Score	Example Terms in Clusters	Terms Overlap
Annotation Cluster 1	21.93	GO:0005739~mitochondrion GO:0044429~mitochondrial part GO:0005740~mitochondrial envelope	*
Annotation Cluster 2	16.81	GO:0006082~organic acid metabolic process GO:0043436~oxoacid metabolic process GO:0019752~carboxylic acid metabolic process	
Annotation Cluster 3	15.31	GO:0006091~generation of precursor metabolites and energy GO:0015980~energy derivation by oxidation of organic compounds GO:0046128~purine ribonucleoside metabolic process	*
Annotation Cluster 4	11.8	GO:0043230~extracellular organelle GO:1903561~extracellular vesicle GO:0070062~extracellular exosome	
Annotation Cluster 5	7.15	GO:0030016~myofibril GO:0043292~contractile fiber GO:0030017~sarcomere	*

Table 6: Top five cluster results of down-regulated genes by DAVID analysis. Terms overlapping with O'Meara et al. (2015) denoted by " * "

DAVID Analysis Result on Down-Regulated Genes			
Cluster id	Enrichment Score	Example Terms in Clusters	Terms Overlap
Annotation Cluster 1	11.13	GO:0007049~cell cycle GO:0051301~cell division GO:0022402~cell cycle process	*
Annotation Cluster 2	9.71	GO:0031012~proteinaceous extracellular matrix GO:0031012~extracellular matrix GO:0044420~extracellular matrix component	*
Annotation Cluster 3	9.61	GO:0008283~cell proliferation GO:0042127~regulation of cell proliferation GO:0008284~positive regulation of cell proliferation	
Annotation Cluster 4	8.55	GO:0051128~regulation of organelle organization GO:0033043~positive regulation of cellular component GO:0051130~positive regulation of organelle organization	*
Annotation Cluster 5	8.04	GO:0009887~organ morphogenesis GO:0060429~epithelium development GO:0048729~tissue morphogenesis	*

Within the DAVID analysis results for up-regulated genes, three of the top five annotation clusters contained Gene Ontology (GO) terms that overlapped with O'Meara et al. (2015). These included terms pertaining to mitochondria, generation of precursor metabolites, energy derivation, and components of the myocyte such as myofibrils, contractile fibers, and sarcomere. These results correspond to the process of *in vivo* maturation of the cardiomyocyte from neonatal day 0 (P0) to adult stage (Ad) since as the cardiomyocyte matures, there will be organization of the compositional units of the cardiomyocyte such as sarcomere assembly and an increase in energy demand as well. The terms that did not show overlap for the up-regulated genes such as organic acid metabolic processes and extracellular organelles appear to have biological relevance to the overall process. For instance, metabolic processes function in generating metabolites that can impact downstream pathways, and in generating sufficient energy supply as well.

Within the DAVID analysis results for down-regulated genes, four of the top five annotation clusters contained overlapping GO terms. These included terms pertaining to cell cycle, extracellular matrix, regulation of organelle organization, epithelium development, and tissue morphogenesis. As cardiomyocytes mature, there is commonly an exit of the cell cycle, so there will be down-regulation of genes associated with the cell cycle. Terms that did not show overlap in the down-regulated genes such as cell proliferation have biological relevance to the maturation process. For instance, as the cardiomyocyte matures, there will be decreased expression of genes associated with cell proliferation.

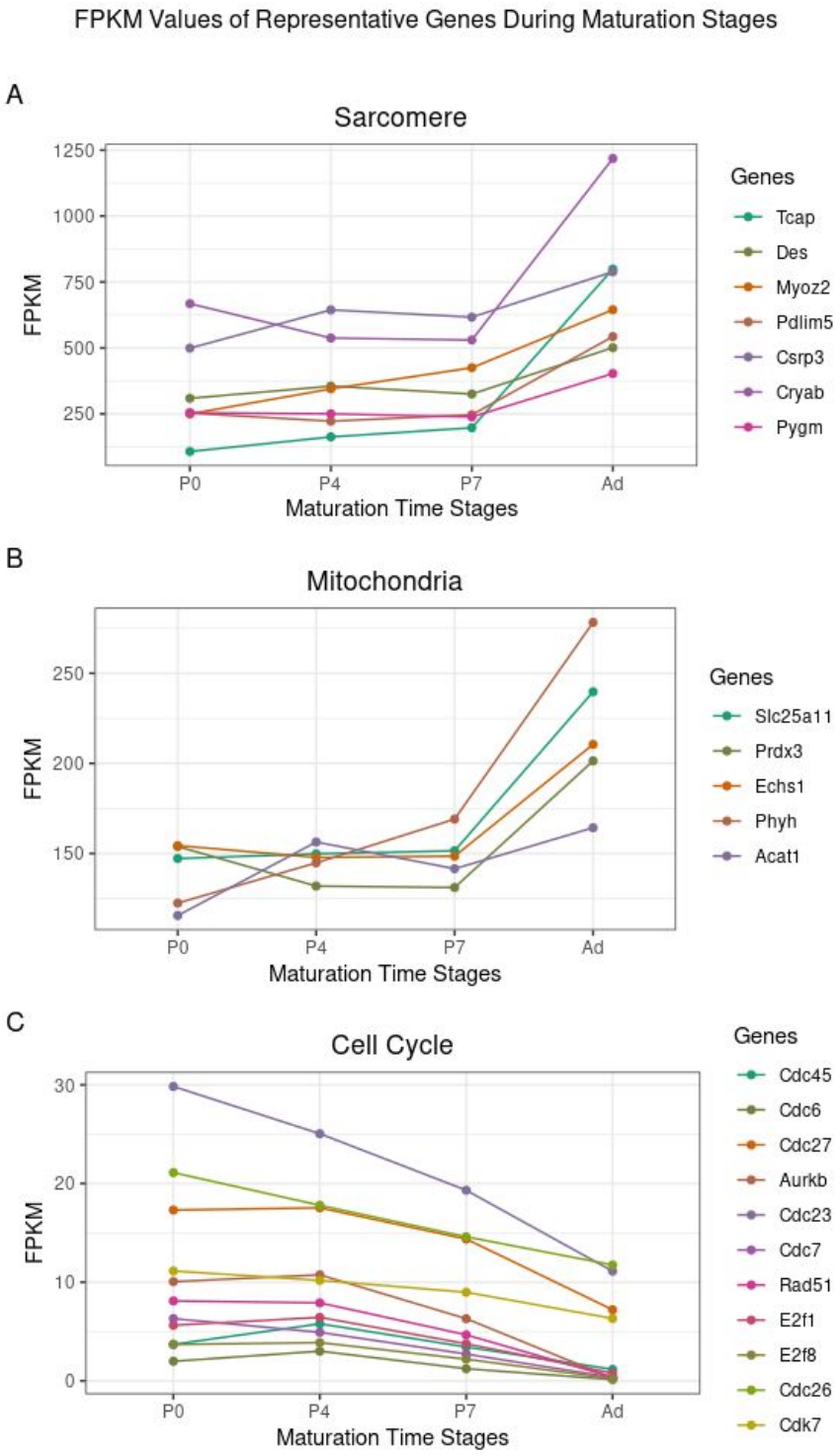

Figure 6. FPKM values of specific sarcomere, mitochondria, and cell cycle genes during the time points of cardiomyocyte maturation (P0, P4, P7, Ad). Figure 6B omits gene Mpc1 due to no FPKM value associated with this gene possibly due to low signal during the analysis. Figure 5C omits gene Bora due to no FPKM value associated with this gene possibly due to low signal during the analysis. Values are expressed as the averages of the two replicates of a given sample.

The results in Figure 6 show the same general trend in direction and magnitude as compared to the Figure 1D in O'Meara et al (2015). The samples P0, P4, P7, and Ad represent days or stages of maturation. For instance, P0 represents postnatal day 0, while Ad represents the adult stage. In Figure 6A, the FPKM values for seven genes specific to the sarcomere GO term are shown for each of the samples with an overall upward directional trend. The FPKM values are fragments per kilobase of transcript per million mapped reads. For P0, the averaged FPKM values range from approximately 110 to 670, while for Ad, the averaged FPKM values span a greater range from approximately 400 to 1,220. For instance, gene Tcap showed a change in FPKM values from P0 to Ad, and Tcap had an averaged FPKM value of 107.4795 for P0 and 799.0830 for Ad which indicated a fold change of approximately 7.43 (log2 fold change: 2.89) from P0 to Ad. As cardiomyocyte maturation proceeded, the results displayed an increase in expression for the sarcomere-related genes. This indicated the progression toward sarcomere assembly as cardiomyocyte maturation occurred[6].

In Figure 6B, the FPKM values for five genes specific to the mitochondria GO term are shown. This figure shows a similar overall trend compared to Figure 1D in O'Meara et al (2015) which is an upward directional trend among P0 to Ad. For P0, the averaged FPKM values range from approximately 115 to 155, while for Ad, the values range from approximately 165 to 280. The gene Phyh showed a change in FPKM values from P0 to Ad, and this gene had an averaged FPKM value of 122.5511 for P0 and 278.2150 for Ad which indicated a fold change of approximately 2.27 (log2 fold change: 1.18) from P0 to Ad. As the cardiomyocyte maturation occurred, the results reflected an increase in expression for the mitochondria-related genes which can be due to the greater energy demands.

In Figure 6C, the FPKM values for eleven genes specific to the cell cycle GO term are shown with an overall downward trend being displayed. This figure shows a similar downward trend that is displayed in Figure 1D in O'Meara et al (2015). For P0, the averaged FPKM values range from approximately 2 to 30, while for Ad, the values range from approximately 0.13 to 11. However, there are 7 genes (Aurkb, Rad51, Cdc7, E2f1, Cdc45, E2f8, and Cdc6) with an FPKM value less than 10 for P0 and almost close to 0 for Ad. As maturation occurred, the results displayed a decrease in expression for the cell cycle-related genes which reflects the process of exiting of the cell cycle.

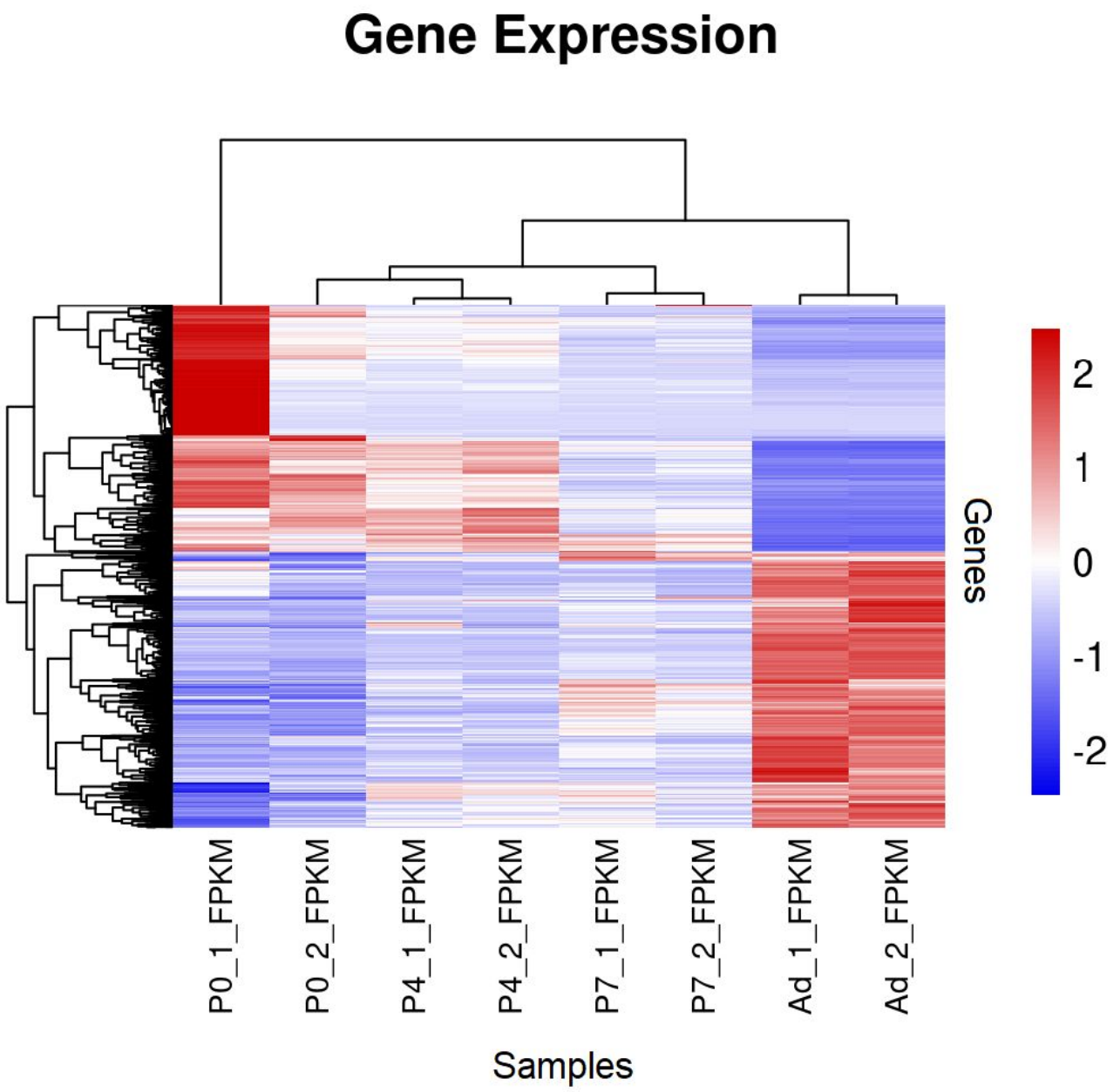

Figure 7. Clustered heatmap of the top 1000 differentially expressed genes of the 8 samples from P0, P4, P7, and Ad stages. The differentially expressed genes are along the rows and samples along the columns. The intensity of the colors shown in the color key represent the gene expression level with red indicating up-regulation and blue indicating down-regulation.

The results in Figure 7 appear to be divided into 2 primary clusters of genes in which there are multiple clusters that subdivide the genes further. The results show that gene clusters that were up-regulated in P0 (postnatal day 0) appeared to be down-regulated in Ad (adult stage). This is shown by the darker red color for the P0 replicates in comparison to the darker blue color for the Ad replicates. In addition, the same genes within this cluster in the P4 replicates have slightly less up-regulation in comparison to the P0 replicates. The P7 replicates have slightly less down-regulation in comparison to the Ad replicates. Similarly, gene clusters that were down-regulated in the P0 replicates appeared to be up-regulated in the Ad replicates. This is shown by the darker blue in the P0 replicates in comparison to the darker red colors in the Ad replicates. In comparison to the P0 replicates in this cluster, P4 replicates exhibit slightly less down-regulation of these genes, and the P7 replicates exhibit a further increase in expression in comparison to the P4 replicates. In addition, the replicates of P4, P7, and Ad in Figure 7 serve as a reasonability check since there appears to be similar expression results in terms of color intensity for up-regulation and down-regulation among these 3 replicates. Between P0_1 and P0_2, there is a distinct difference in the color intensity of the topmost gene cluster. This could be due to the processing steps within Cufflinks and Cuffdiff, or possible batch effects between the replicates

Discussion

The top 10 differential expression genes were shown in Table 2 based on the smallest q-value. By expanding the table, we were able to see the top 680 genes had the same q-value. All of these genes were significant since lower q-value indicated obvious difference. By filtering out the significant genes, we had 2139 labeled significant genes in total. And in this significant gene set we got 1084 up-regulated genes and 1055 down-regulated genes based on log2 fold change value. Compared to the original paper by O'Meara et al (2015), they had 2409 genes in up-regulated gene set and 7570 genes in down-regulated gene set. One possible reason was they used different select methods to filter out the significant genes. We also tried to use p-value less than 0.01 as a condition, and we got 2376 genes based on this method. If they used p-value less than 0.05 or other values as the condition, they could have a larger significant gene set.

We performed an enrichment analysis based on DAVID Functional Annotation Tool for further functional annotation clustering. There were some annotation clusters with overlapping results with O'Meara et al (2015), such as the clusters pertaining to mitochondrion, respiration and metabolic processes, and myofibril and sarcomere in the up-regulated genes. For these genes, these terms reflect the processes that increase in expression as cardiomyocytes mature such as increase in metabolic demands within the cell and organization into a more rigid structure. The terms that showed overlap with O'Meara et al (2015) such as cell cycle process, extracellular matrix, regulation of organelle organization, and tissue morphogenesis in the down-regulated genes, for instance, exhibit the characteristics of cell cycle exit in mature

cardiomyocytes. The paper states that loss of cardiac regeneration in complex mammals is thought to arise from the failure to re-enter the cell cycle[6]. Adult cardiomyocytes are terminally differentiated, and terms such as tissue morphogenesis reflect how there would be decrease in this process and expression levels as cardiomyocytes matured. In addition, for the up-regulated gene set, we got a highest enrichment score of 21.93, and for the down-regulated gene set, we got a highest enrichment score of 11.13. The enrichment scores were different with the results from O'Meara et al (2015). This difference can be explained because the enrichment scores were calculated by average p-value. Since they had a larger gene set to cluster, the std of p-value of the gene set were greater than a smaller gene set.

Although the most common gene ontology terms were similarly reproduced, we also observed some different GO terms with O'Meara et al. (2015). For the up-regulated genes, we observed a gene cluster for GO terms such as extracellular organelle and organic acid metabolic processes have biological relevance as the metabolic processes function in generating metabolites and chemical energy for the cell. For the down-regulated genes, we observed a cluster of genes for cell proliferation with a high enrichment score of 9.61. One possible reason for this difference is the different gene set. As we discussed above, different sizes of the gene set cause the different enrichment score. In this study, we selected the top 5 clusters based on enrichment score. In fact, among 400+ clusters, the top of them are not very different in enrichment score. Even though terms pertaining to cell proliferation did not overlap with O'Meara et al. (2015), there is biological relevance because there is a shift from a proliferative state to cell cycle exit state within the differentiation and maturation processes, and adult cardiomyocytes are generally also terminally differentiated cells.[6]

Based on Figure 6, there was an overall upward trend for the representative sarcomere-related genes, and this indicated the progression of sarcomere assembly that occurs during cardiomyocyte maturation. One of the sarcomere-related genes was Tcap, and the protein encoded by this gene Telethonin functions in the assembly of titin at the Z-disk in the sarcomere in which the titin is important for the elasticity component during contraction of the cardiomyocyte.[27] Another sarcomere-related gene was Des, and the protein Desmin encoded by this gene is important for maintaining the structure of the sarcomere and interconnecting myofibrils of the myocyte with cellular organelles such as mitochondria[27]. Consistent with the findings in O'Meara et al (2015), the representative sarcomere-related genes had an increase in expression as cardiomyocyte maturation occurred due to the organization of the compositional units within the myocyte such as the interconnection of myofibrils and the assembly of sarcomeres. Within the cardiomyocytes, mitochondria are present as well to support the energy needs of the cell. One of the mitochondria-related genes identified was Slc25a11 which has functions pertaining to metabolic processes and transport activity within the mitochondria.[27] To supply the energy demands needed for contraction of cardiac myocytes in the form of ATP,

mitochondria and thus, also mitochondria-related genes, have an increase in expression as cardiomyocyte maturation occurs.

Most of the genes depicted in Figure 6C come from the cell-division-cycle (cdc) genes family, and for instance, gene Cdc6 has functions pertaining to DNA replication initiation within the cell cycle steps.[27] Genes relating to the cell cycle showed decrease in expression due to the exit of the cell cycle as cardiomyocyte maturation occurred. Overall, based on Figure 6 and consistent with the results in O'Meara et al (2015), there was a general increase of expression of sarcomere and mitochondria-related genes, and a general decrease in expression of cell cycle-related genes from P0 to Ad stages. This is due to the observation that as cardiac myocytes mature from P0 to Ad, there will be a drive towards organization of the sarcomere structure, increase in energy demand, and cell cycle exit. In addition, as is shown in Figure 7, as cardiac myocytes matured, there were clusters of genes that were up-regulated within P0 and down-regulated with Ad, and similarly, clusters that were down-regulated within P0 and up-regulated within Ad which reiterates the point that certain genes with specific functions will show these patterns.

In our project specifically, we analyzed the stages from P0 to Adult which represent *in vivo* maturation. We analyzed how the expression of genes change as maturation occurred from the P0 to the Ad stages. O'Meara et al. (2015) conducted multiple experiments in addition such as analyzing gene expression in *in vitro* differentiation and adult cardiac myocytes explanted and cultured. They developed a transcriptional signature by determining differentially expressed genes during the processes of *in vitro* differentiation and *in vivo* maturation to identify the molecular processes that function in shifting from proliferation to exit of the cell cycle. O'Meara et al. (2015) also studied how the adult cardiomyocyte explant would have a loss in the differentiated state which could be used to determine transcriptional changes. One of their findings was that there was reversion indicated by a decrease in expression of sarcomere-related components and increase in expression of cell cycle components.[6] They found that genes that were up-regulated during the differentiation process were down-regulated during the explant and cultured process such as ones that function in metabolic processes, sarcomere organization, and heart function, and similarly, down-regulated genes during the differentiation process were up-regulated during the explant and cultured process such as ones that function in cell cycle.[6] Therefore, in their results, they determined a transcriptional reversion process in cardiomyocyte regeneration.

Based on the results we observed after conducting the analysis for P0 (postnatal day 0) versus Ad (adult stage) we can conclude that overall, most genes determined to be up-regulated within the P0 stage corresponded to being down-regulated within the Ad as cardiomyocyte maturation occurred. Similarly, most genes determined to be down-regulated within the P0 stage corresponded to being up-regulated within the Ad stage. We could potentially leverage this

information to conduct further studies with gene-knockout experiments to validate the findings and observe how the expressed phenotype changes and the impact it has on the regeneration capacity of the cardiomyocyte. This is also of great importance in healthcare as more studies could potentially reveal similar mechanisms in humans. Implementing this insight in application to cardiovascular diseases such as myocardial infarction and congestive heart failure could progress research into therapies centered on regeneration of damaged cardiac tissue. In addition, expanding research into regenerative therapies that can target patients with acute rejection events post-heart transplant or patients who are on heart transplant waiting lists can be beneficial to the progression of cardiac regenerative medicine and the healthcare industry as well. If a crucial breakthrough in the discovery of genes responsible for regeneration in more complex mammals occurs, it could revolutionize the healthcare industry.

Conclusion

Overall, in this study, we reproduced some similar results from O'Meara et al (2015) using data from *in vivo* maturation models. We observed some explained differences due to the different size of the gene data set and the inclusion of data pertaining to only P0 through Ad maturation stages. As our study focused on the analysis of the stages of *in vivo* maturation from P0 (postnatal day 0) to Ad (adult stage) with P4 and P7 stages included, we were able to see genes that had an increase in expression pertaining to metabolic processes and sarcomere and cardiomyocyte organization, and decrease in expression pertaining to cell cycle. Utilizing this transcriptional signature, further research was done in O'Meara et al (2015) to determine a transcriptional reversion process during cardiac myocyte regeneration. This information is vital to research on therapies focused on cardiac regeneration for damaged cardiac tissue from cardiovascular conditions such as myocardial infarction and congestive heart failure. Applications for this research in cardiac regenerative medicine can hopefully in the future address the complications of cardiovascular disease that affect a great portion of the world's population.

References

1. Oberpriller JO, Oberpriller JC. Response of the adult newt ventricle to injury. *J Exp Zool*. 1974;187:249–253. doi: 10.1002/jez.1401870208.
2. Poss KD, Wilson LG, Keating MT. Heart regeneration in zebrafish. *Science*. 2002;298:2188–2190. doi: 10.1126/science.1077857.
3. Steinhauser ML, Lee RT. Regeneration of the heart. *EMBO Mol Med*. 2011;3:701–712. doi: 10.1002/emmm.201100175.
4. Porrello ER, Mahmoud AI, Simpson E, Hill JA, Richardson JA, Olson EN, Sadek HA. Transient regenerative potential of the neonatal mouse heart. *Science*. 2011;331:1078–1080. doi: 10.1126/science.1200708.
5. Bicknell KA, Coxon CH, Brooks G. Can the cardiomyocyte cell cycle be reprogrammed? *J Mol Cell Cardiol*. 2007;42:706–721. doi: 10.1016/j. yjmcc.2007.01.006.
6. 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
7. Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, *10*(1), 57–63. https://doi.org/10.1038/nrg2484
8. Fang Y, Gupta V, Karra R, Holdway JE, Kikuchi K, Poss KD. Translational profiling of cardiomyocytes identifies an early Jak1/Stat3 injury response required for zebrafish heart regeneration. *Proc Natl Acad Sci U S A*. 2013;110:13416–13421. doi: 10.1073/pnas.1309810110.
9. Wamstad JA, Alexander JM, Truty RM, et al. Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage. *Cell*. 2012;151:206–220. doi: 10.1016/j.cell.2012.07.035.
10. Boyer, L. A. (2014, December 19). Geo accession viewer. Retrieved March 17, 2021, from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE64403
11. http://ncbi.github.io/sra-tools/
12. Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
13. MGSCv37 - mm9 - genome - assembly - NCBI. (2010, October 21). Retrieved March 17, 2021, from https://www.ncbi.nlm.nih.gov/assembly/GCF_000001635.18/
14. Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, 25(9), 1105-1111.
15. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... & Durbin, R. (2009). The sequence alignment/map format and SAMtools. Bioinformatics, 25(16), 2078-2079.
16. Wang, L., Wang, S., & Li, W. (2012). RSeQC: quality control of RNA-seq experiments. Bioinformatics, 28(16), 2184-2185.

17. Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., ... & Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology, 28(5), 511-515.
18. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nature Protoc. 2009;4(1):44-57
19. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37(1):1-13.
20. Romero, I. G., Pai, A. A., Tung, J., & Gilad, Y. (2014). Impact of RNA degradation on measurements of gene expression. bioRxiv, 002261.
21. http://www.htslib.org/doc/samtools-flagstat.html
22. Deschamps-Francoeur, G., Simoneau, J., & Scott, M. S. (2020). Handling multi-mapped reads in RNA-seq. Computational and Structural Biotechnology Journal. 3
23. Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., ... & Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology, 28(5), 511-515.
24. Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, 25(9), 1105-1111.
25. Wang, L., Wang, S., & Li, W. (2012). RSeQC: quality control of RNA-seq experiments. Bioinformatics, 28(16), 2184-2185.
26. Turner, F. S. (2014). Assessment of insert sizes and adapter content in fastq data from NexteraXT libraries. Frontiers in genetics, 5, 5.
27. Bateman, A., Martin, M. J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., ... & UniProt Consortium. (2020). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*. Chicago
28. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Steffen Durinck, Paul T. Spellman, Ewan Birney and Wolfgang Huber, Nature Protocols 4, 1184-1191 (2009).
BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Steffen Durinck, Yves Moreau, Arek Kasprzyk, Sean Davis, Bart De Moor, Alvis Brazma and Wolfgang Huber,Bioinformatics 21, 3439-3440 (2005). 2
29. Rainer J (2017). *EnsDb.Mmusculus.v79: Ensembl based annotation package*. R package version 2.99.0. 3
30. Carlson M (2019). *org.Mm.eg.db: Genome wide annotation for Mouse*. R package version 3.8.2. 4

Supplemental Material

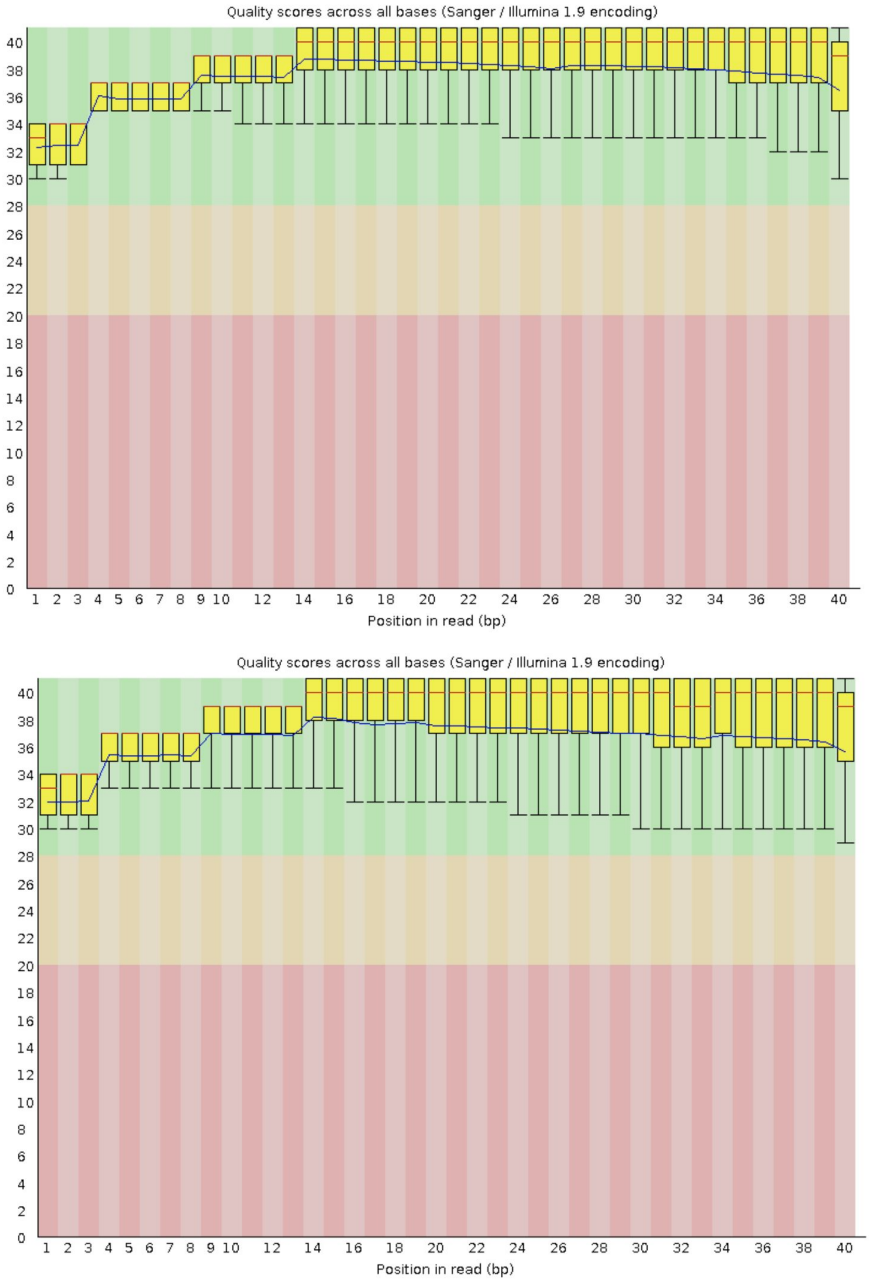

Figure S1. Per Base Sequence Quality. Phred score across every position in the read. Top and bottom figure displays the FastQ paired-end reads, read one and read two respectively.

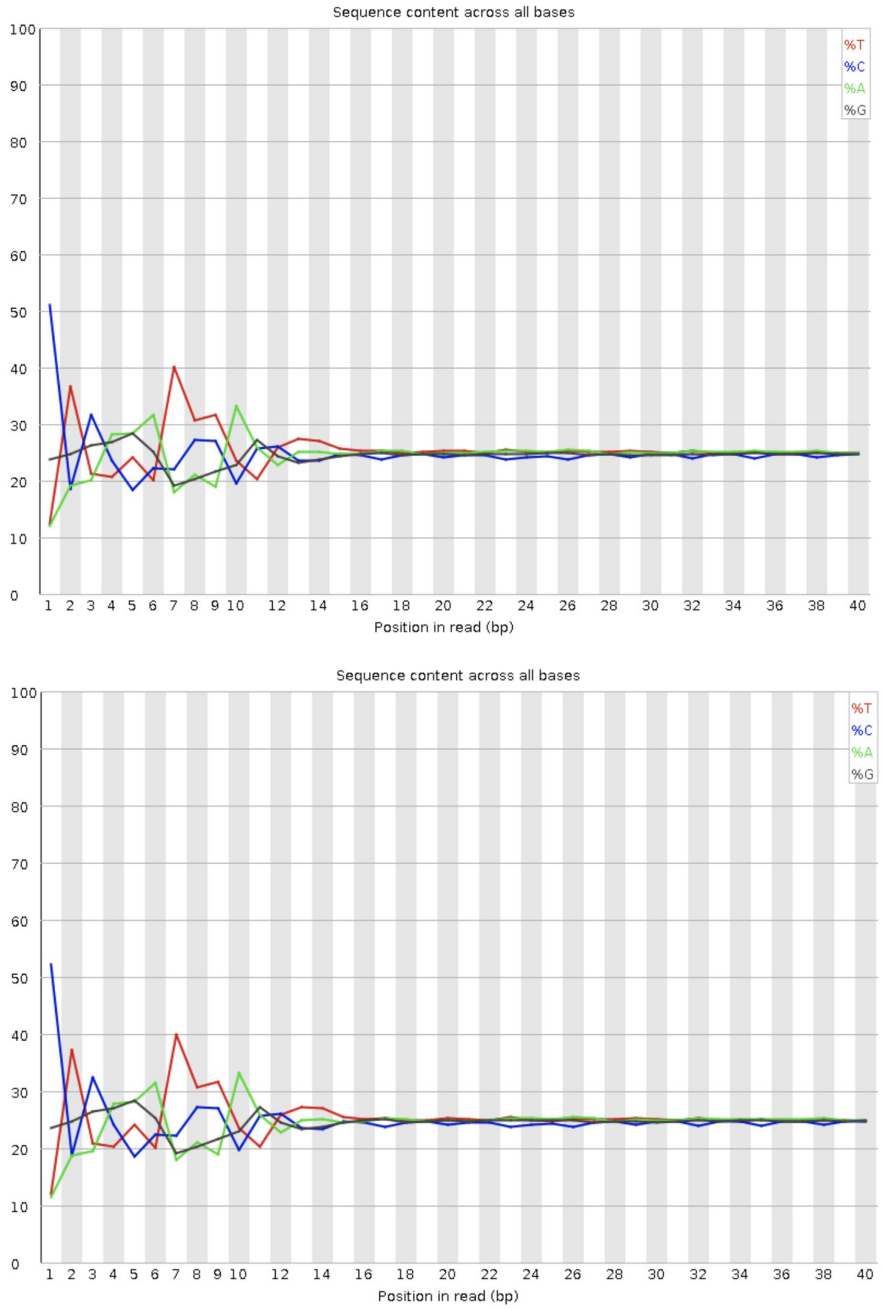

Figure S2. Per Base Sequence Content. Percentage of each nucleotide across each position in the read. Top and bottom figure displays the FastQ paired-end reads, read one and read two respectively.

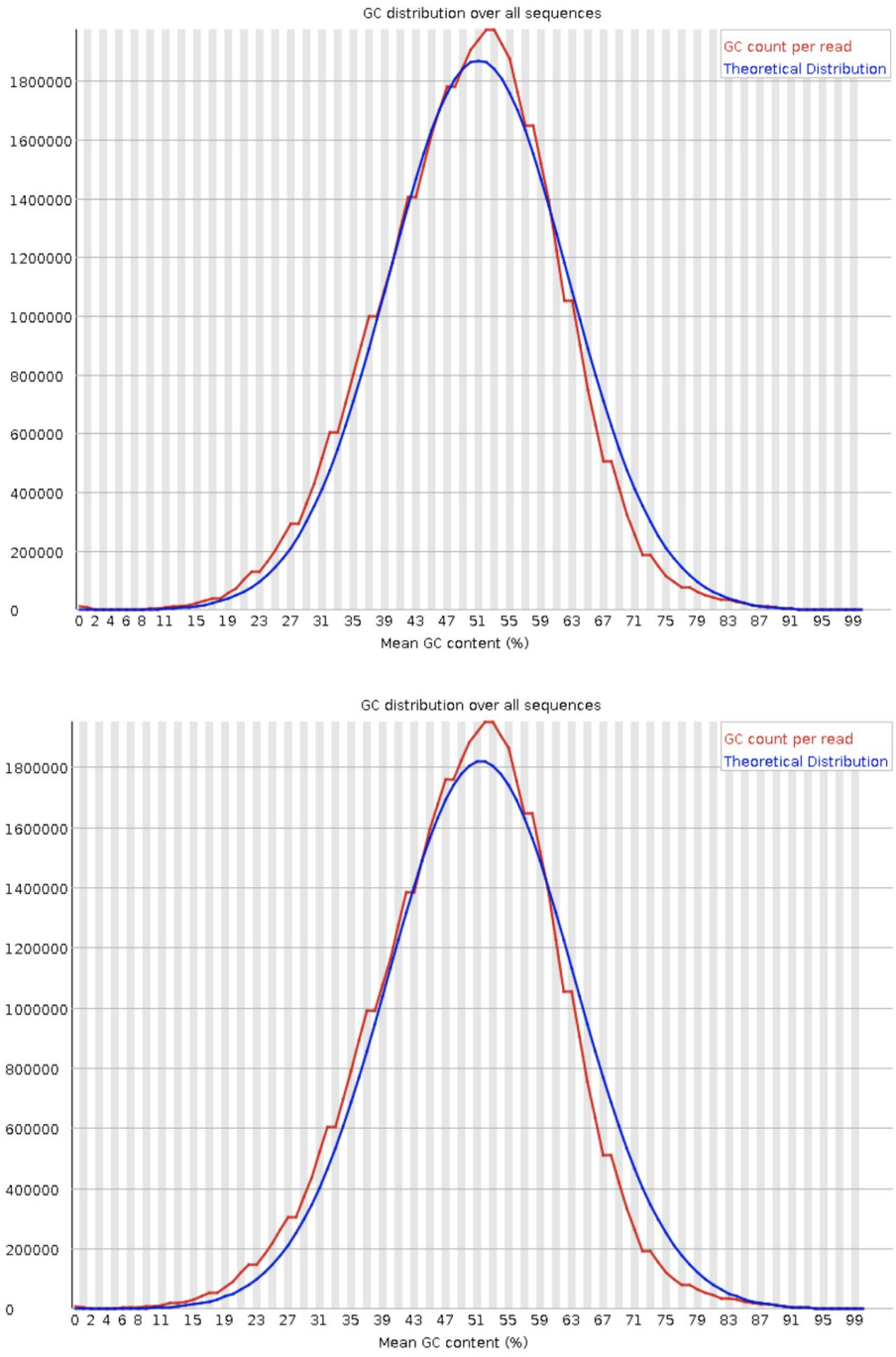

Figure S3. Per Sequence GC Content. Red line represents the GC content across every position of the read and the blue line represents the theoretical distribution of GC content. Top and bottom figure displays the FastQ paired-end reads, read one and read two respectively.

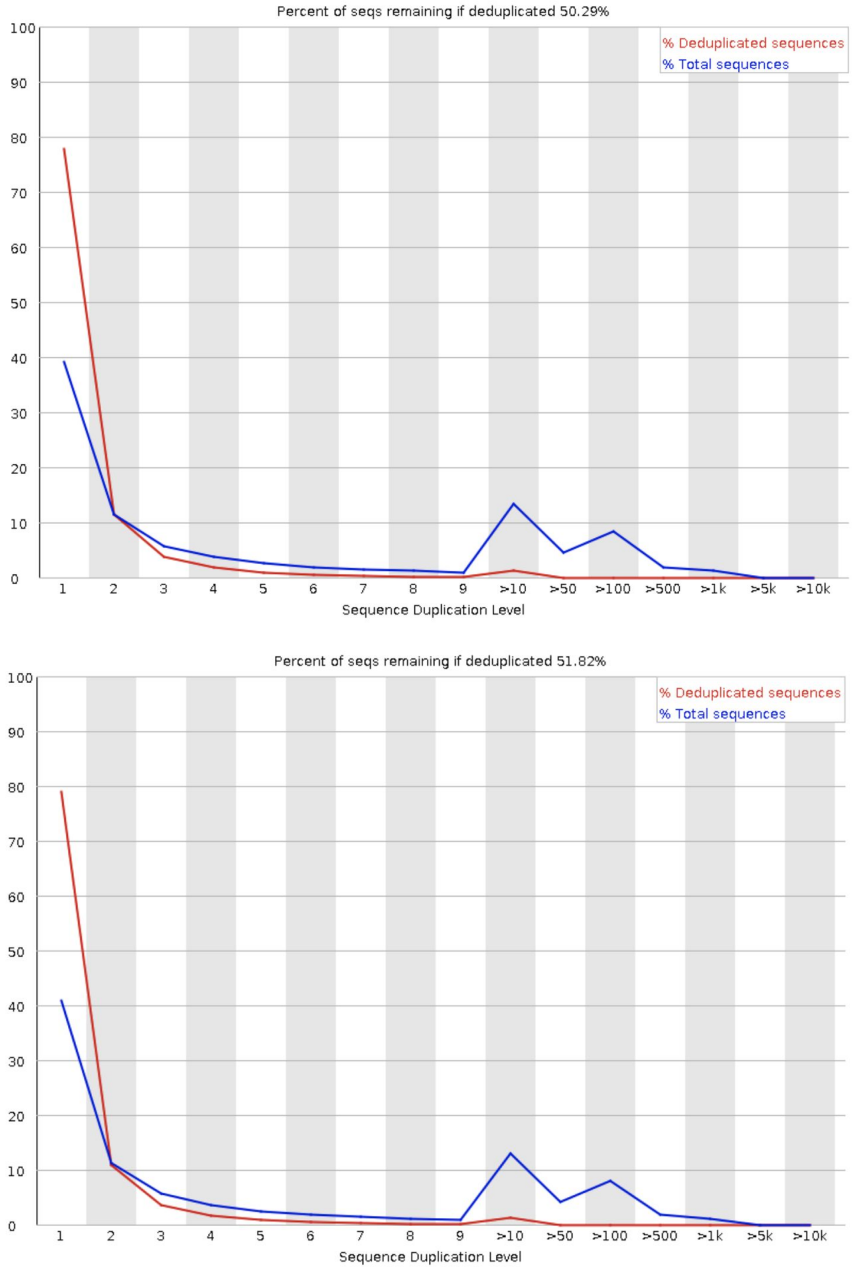

Figure S4. Sequence Duplication Levels. The red line represents the percent of duplicated sequences in the read and the blue line represents the percent of unique sequences in the read. Top and bottom figure displays the FastQ paired-end reads, read one and read two respectively.

Supplemental Citations:

1. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
2. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

3. Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.1. https://CRAN.R-project.org/package=dplyr
4. Erich Neuwirth (2014). RColorBrewer: ColorBrewer Palettes. R package version 1.1-2. https://CRAN.R-project.org/package=RColorBrewer
5. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
6. H. Wickham. Reshaping data with the reshape package. Journal of Statistical Software, 21(12), 2007.
7. Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra
8. Alboukadel Kassambara (2020). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0. https://CRAN.R-project.org/package=ggpubr
9. Tal Galili (2015). dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering. Bioinformatics. DOI: 10.1093/bioinformatics/btv428