Transcriptional Profiling of Mammalian Cardiac Regeneration with mRNA-seq

David Lenci, Nikita Tomar, and Daniel Gealow

Introduction

Heart failure is one of the leading causes of deaths worldwide with 26 million patients affected according to a 2020 study [2]. One of the major causes of heart failure is the death of cardiac myocytes (CM), which diminishes heart pump functionality. Currently, there are no therapies that allow for the regeneration of these adult cardiac mammalian myocytes [4]. That being said, it is known that CMs are able to self-regenerate shortly after birth, and this functionality is lost in mammalian myocytes at adulthood [1]. In O'Meara et al, the researchers sought to identify the causes of the loss of this functionality by comparing the CM gene expression profile of juvenile and adult mice, hypothesizing that during regeneration the myocytes revert the translational phenotype to a less differentiated state [3].

Data

The data used in this analysis consists of high throughput RNA sequencing of injury induced cardiac myocyte cells from mice generated by O'Meara et al in *Transcriptional Reversion of Cardiac Myocyte Fate During Mammalian Cardiac Regeneration*, and accessed by the Gene Expression Omnibus (GEO) series GSE64403 [3][5]. RNA was isolated from embryonic stem cells (in vitro) and from neonate and adult mice (in vivo). For this analysis we focused mainly on the in vivo data generated in O'Meara et al. Using the SRA toolkit, the SRA file obtained from the GEO series mentioned above was converted to fastq format, and then fastqc was performed to ensure quality of the generated short reads.

From the fastqc results, the generated reads passed most of the quality tests performed including *per sequence base quality*, *per sequence GC content*, and *overrepresented sequences*. The *sequence duplication levels* test returned a warning, and the *per base sequence content* test failed. The figures generated from the fastqc report can be seen in the supplemental figures section. For each figure there are two images, as the reads used were paired end meaning there were two fastq files.

The per sequence base quality test shows us the distribution of quality scores at each position in the read across all reads, and can be seen in figure S1. For per sequence quality all reads obtained a quality score of at least 30 meaning a probability of incorrect base call of 1 in 1000. Next, the per sequence GC content test shows us the GC distribution across all reads, as well as the theoretical GC distribution [6]. Similarly, for the per sequence GC content, the test passed with a GC content of 49% and can be seen in Figure S2. Lastly, the overrepresented sequences test informs us if there are any sequences that appear more than 0.1% of the time, which could be a sign of contamination [6]. From Figure S3 we can see that there are no overrepresented sequences from our data.

The sequence duplication test aims to inform us if a large percentage of our reads represent non-unique reads [6]. Specifically, this test will return a warning if at least 20% of our sequences are not unique. For this test, as seen in figure S4, fastqc gave us a warning, but this could be because of the context in which the data was generated. Since the data was generated when the CMs were undergoing regeneration it makes sense that certain gene sets important to this process would be highly enriched. The last aspect of the fastqc report worth mentioning is the per base sequence content test, which was given a failed status and can be seen in figure S5. This is the case for all RNA sequencing data, as the first 10-12 bases result from the random hexamer priming that occurs during RNA-seq preparation [6].

Methods

We used tophat to align the FASTQ files against the mm9 mouse reference genome. The program was run with a mate inner distance of 200, segment length of 20, 1 allowed segment mismatch, and no novel junctions allowed. This took about an hour to run with 16 threads on SCC.

We used samtools for an overview of the alignment results:

```
49706999 + 0 in total (QC-passed reads + QC-failed reads)
8317665 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
49706999 + 0 mapped (100.00% : N/A)
41389334 + 0 paired in sequencing
20878784 + 0 read1
20510550 + 0 read2
29422646 + 0 properly paired (71.09% : N/A)
39936472 + 0 with itself and mate mapped
1452862 + 0 singletons (3.51% : N/A)
1387382 + 0 with mate mapped to a different chr
704916 + 0 with mate mapped to a different chr (mapQ>=5)
```

All of the reads were mapped to somewhere in the genome, and 71.09% were successfully mapped with their paired read.

Next, we used RseQC utilities to perform quality control verification. Figure 1 shows the plot created by geneBody_coverage.py, which maps the coverage of each part of every gene on the same 0 to 100% scale to detect whether there are any systematic biases toward the 5’ or 3’ end. This plot does reveal some bias towards higher coverage of the 3’ side of genes. This is likely an artifact of the sequencing methodology, and we decided that it was not drastic enough to discredit the data.

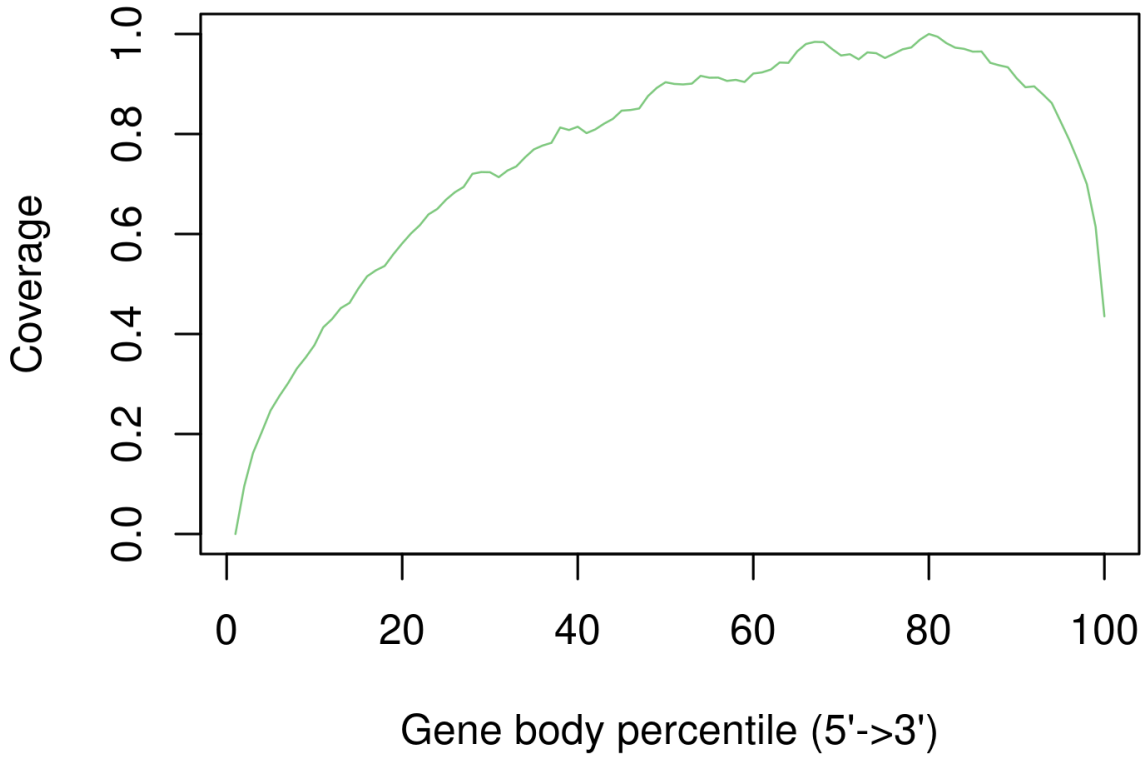

Figure 1. Gene body coverage. The dropoffs in coverage towards 0% and 100% are typical. There is some skewness towards higher coverage near the 3’ end.

The inner_distance.py utility measured the distribution of gaps between each of the aligned paired ends (Figure 2). While the center of the distribution (around 85 bp) is definitely lower than the expected 200 bp gaps, the insert sizes do still appear to be on a reasonable order of magnitude, with a smooth unimodal distribution and very few below zero or far above 200. This indicates that the RNA fragments sequenced may have been somewhat shorter on average than intended.

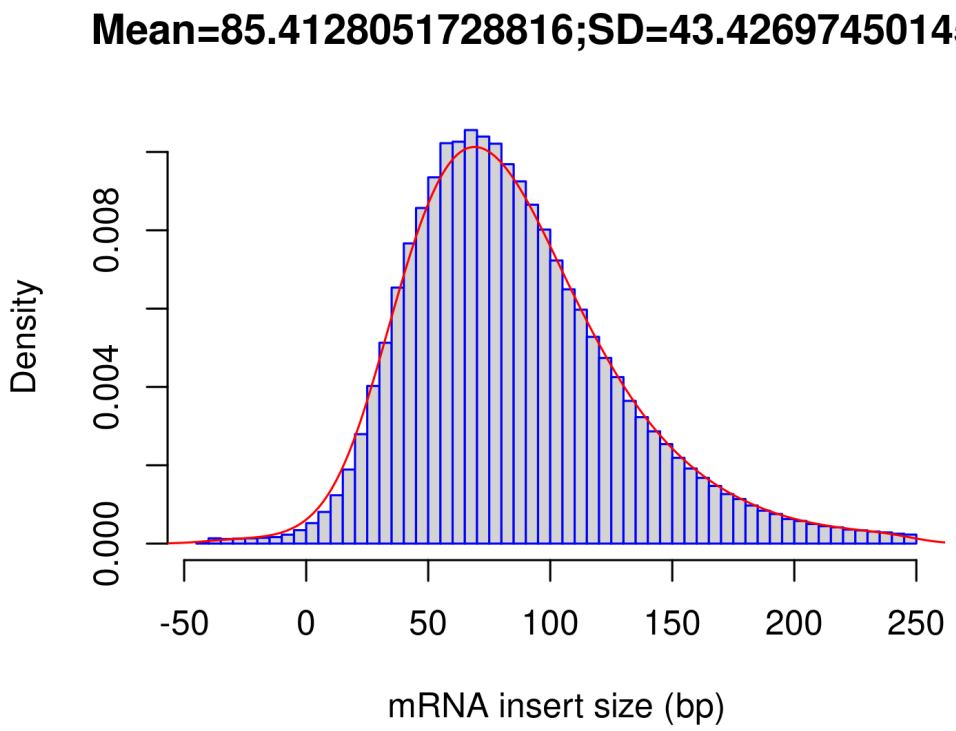

Figure 2. Inner distance histogram. Blue bars are the actual histogram of inner distances sampled from the alignment; red line is the smoothed approximation to the distribution.

Finally, bam_stat.py had the following output:

```
Total records:                          49706999

QC failed:                              0
Optical/PCR duplicate:                  0
Non primary hits                        8317665
Unmapped reads:                         0
mapq < mapq_cut (non-unique):           2899954

mapq >= mapq_cut (unique):              38489380
Read-1:                                 19409941
Read-2:                                 19079439
Reads map to '+':                       19236824
Reads map to '-':                       19252556
Non-splice reads:                       33099839
Splice reads:                           5389541
Reads mapped in proper pairs:           27972916
Proper-paired reads map to different chrom:4
```

None of the reads failed bam_stat's quality control metrics. About half mapped to each strand of the chromosome, as expected. There were not as many mapped in proper pairs as we would hope, but there were still enough to be reasonably confident in the quality of the data. We ran all the RNA seQC tools in parallel on SCC, with the longest taking about 40 minutes to finish.

Next, we ran cufflinks, with the "rescue methods" for multi-reads activated, and only counting hits compatible with the reference. This ran in about an hour using 16 threads on SCC. To visually inspect the FPKM values (a count of the RNA abundance), we plotted a histogram of the logarithm of FPKM, since values spanned many orders of magnitude. From a visual inspection, it was clear that the primary mass of nonzero genes detected was above an FPKM value of 0.01. Therefore, we filtered out the genes with an expression value below that, and plotted the histogram of the 16337 genes that were expressed with FPKM > 0.01 (Figure 3).

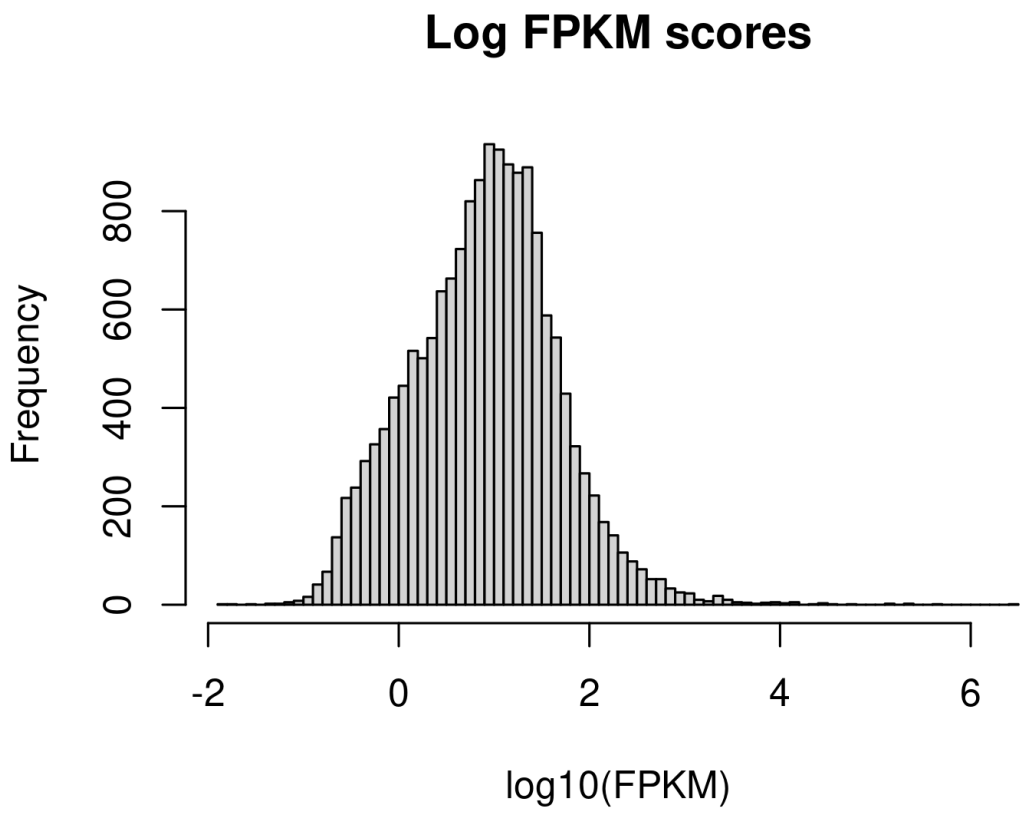

Figure 3. Distribution of log-transformed FPKM scores. Most nonzero scores were between 0.1 and 1000, though a few were as high as 1,000,000.

To find the genes which were differentially expressed between juveniles and adult mice, we ran cuffdiff, again with “rescue method” active and 16 threads, on SCC. It took about two hours to complete.

Identifying differentially expressed genes associated with myocyte differentiation:

The result of cuffdiff produced a file that contained 36329 differentially expressed genes statistics. This file was read into R studio. The data was first sorted in increasing order of the q-values. We subset the sorted data to only include relevant columns of gene name, FPKM values, Log2 fold change, p_value, q_value. From this a list of top 10 gene expression rows were assessed. A new data frame was created to include only significant genes i.e. genes with value of significant as ‘yes’. We also found the number of genes with p-values < 0.01. Next, we plotted histograms for the distribution of Log2 Fold Change values in the original data frame as well as for the Log2 Fold Change values in the significant data frame. The difference between these two plots were observed and analyzed. The significant data frame was also used to find the list of up-regulated and down-regulated genes using the log2(fold_change) column. These lists were stored in separate text files.

Functional Annotation Clustering Analysis:

The list of upregulated genes and downregulated genes were used to do an analysis in DAVID (Database for Annotations, Visualization and Integrated Discovery) 6.8 using the Function Annotation Clustering tool. This tool helped us obtain enriched gene sets which were organized into clusters ranked in the decreasing order of their enrichment scores. The Mus Musculus was selected as the species. The GOTERM_BP_FAT, GOTERM_MF_FAT and GOTERM_CC_FAT Gene Ontology groups were selected to perform the Function Annotation Clustering. The clusters were analyzed on their GO terms and enrichment scores.

Results

We highlighted the top 10 differentially expressed genes in Table 1. We see that all of them are up-regulated genes with p-values and q-values less than 0.01 indicating that these genes are of great significance. Moreover, out of the total 36329 differentially expressed genes, 2139 genes were significant. And a total of 2376 genes had their p-values < 0.01. There were a total of 1084 up-regulated genes and 1055 down-regulated genes from the 2139 significant genes.

We also plotted two histograms to visualize the frequency distribution of log2 Fold Change values across all genes (Figure 4) and across the significant genes (Figure 5) respectively. The distribution peaked at Log2 Fold Change value of zero in the histogram plot of all genes. However, the histogram for significant genes did not have any distribution at zero value of Log2 Fold Change. The peak in the former plot is a representation of the insignificant genes which were removed in the latter plot.

Gene	FPKM_1	FPKM_2	log2(fold_change)	p_value	q_value
Plekhb2	22.5679	73.5683	1.70481	5.00E-05	0.001069
Mrpl30	46.4547	133.038	1.51794	5.00E-05	0.001069
Coq10b	11.0583	53.3	2.26901	5.00E-05	0.001069
Aox1	1.18858	7.09136	2.57682	5.00E-05	0.001069
Ndufb3	100.609	265.235	1.39851	5.00E-05	0.001069
Sp100	2.13489	100.869	5.56218	5.00E-05	0.001069
Cxcr7	4.95844	32.2753	2.70247	5.00E-05	0.001069
Lrrfip1	118.997	24.6402	-2.27184	5.00E-05	0.001069
Ramp1	13.2076	0.691287	-4.25594	5.00E-05	0.001069
Gpc1	51.2062	185.329	1.8557	5.00E-05	0.001069

Table 1: The top ten differentially expressed gene associated with myocyte differentiation with respect to lowest q-values, where FPKM_1 is the fragment per kilobase million for Postnatal Day 0, FPKM_2 is the fragment per kilobase million for Adult sample, and log2fold_change is the log2(FPKM_Adult/FPKM_P0).

Differentially expressed genes at p_value<0.01 significant genes	
Up regulated genes	1084
Down regulated genes	1055
Total	2139

Table 2: The number of up and down regulated genes out of the significantly expressed genes.

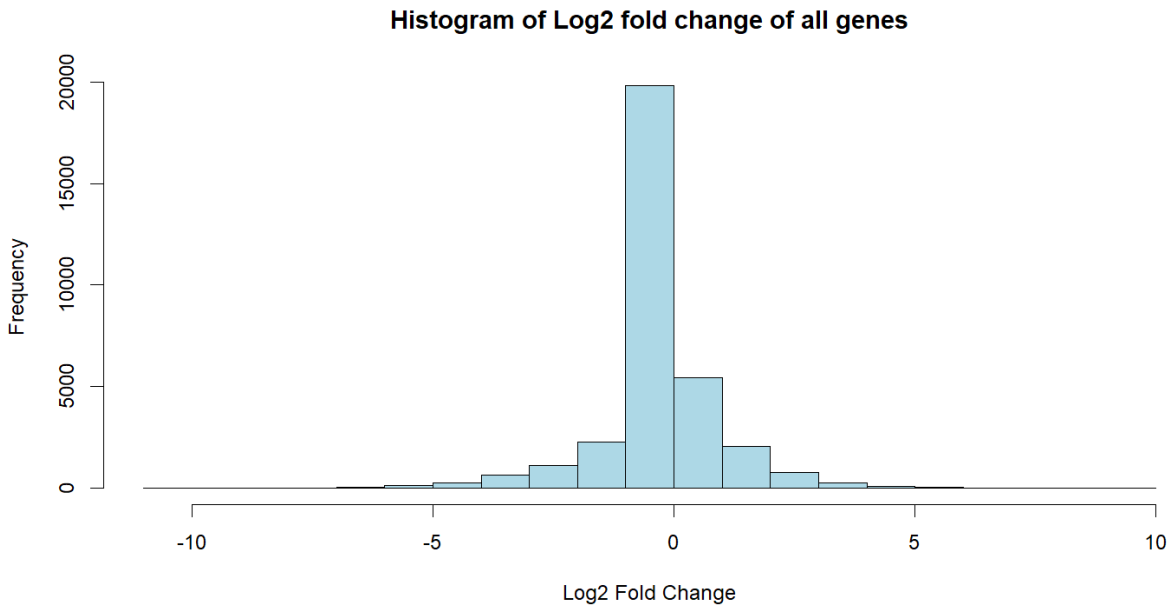

Figure 4: A histogram showing the distribution of Log2 Fold Change values across all genes.

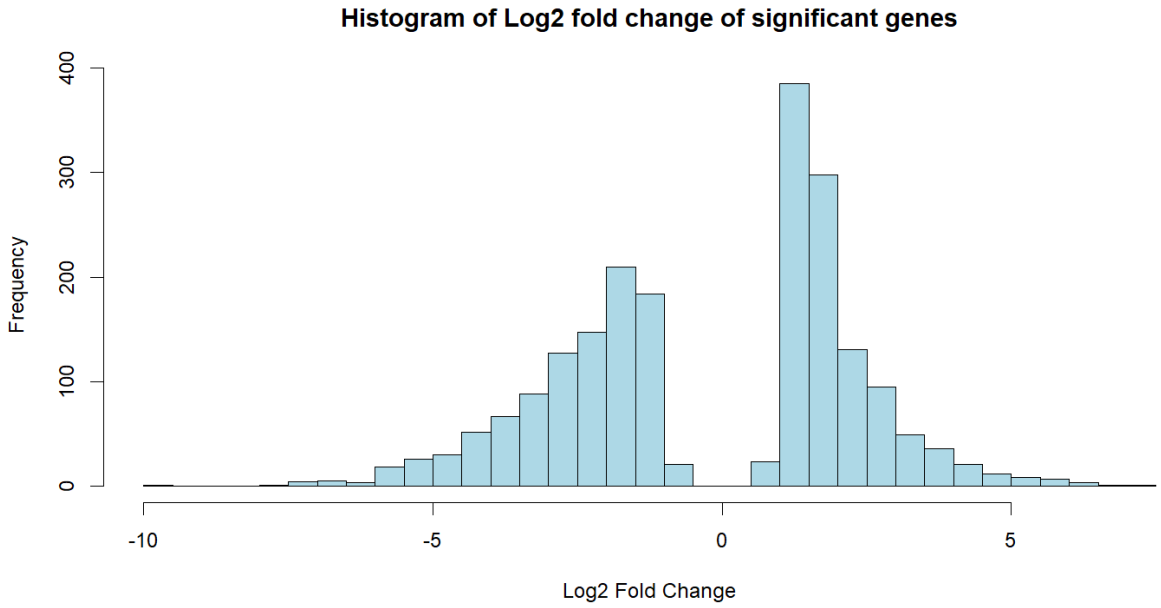

Figure 5: A histogram showing the distribution of Log2 Fold Change values across significant genes.

In order to organize the enriched genes into functionally related clusters we used the DAVID 6.8 tool to perform a Function Annotation Clustering. Table 3 summaries the top clusters for up and down regulated gene sets according to the enrichment scores of each cluster. The top GO (Gene Ontology) terms for up-regulated genes were metabolic process, mitochondrion, cellular response, catabolic process, ion binding, biosynthetic process and cellular binding. The top GO terms for down-regulated genes were ion binding, cell proliferation, cytoskeleton organization, metabolic/biosynthetic process, cellular component, cell development and cell death. The highest enrichment score for up-regulated genes was 28.24 and for the down-regulated genes was 19.65.

	Up-regulated		Down-regulated	
Cluster No.	**GO Term**	**Enrichment Score**	**GO Term**	**Enrichment Scores**
1	metabolic process	28.24	ion binding	19.65
2	mitochondrion	18.84	cell proliferation	17.11
3	cellular response	18.61	cytoskeleton organization	13.67
4	catabolic process	12.23	metabolic/biosynthetic process	13.61
5	ion binding	11.92	cellular component	13.33
6	biosynthetic process	11.45	cell development	12.31
7	cellular respiration	10.36	cell death	12.2

Table 3: The top 7 clusters from the result of DAVID 6.8 Function Annotation Clustering for significant up-regulated genes and down-regulated genes.

Discussion

From the histograms that were plotted for Log2 Fold Change values of all genes (Figure 4) and significant genes (Figure 5) we found that for the plot of all genes was also centered on 0 with almost equal distribution of positive and negative Log2 Fold Change values. For the plot of significant genes we found that it was not centered on 0. In fact, the distribution was more in the positive values indicating more number of up-regulated genes than down-regulated genes. O'meara et al [3] showed a total of 2760 up-regulated and 7570 down-regulated genes from a total of 10000 significant genes which is very different from our study in Table 2 with 1084 up-regulated genes and 1055 down-regulated genes. A possible reasoning may be that the author might have used a higher threshold than the 0.01 we used for p-value to find the significant genes.

Conclusion

In analyzing the significant genes, the total number of up and down regulated differentially expressed genes did not match accurately with the original O'meara et al. In functional annotation clustering the observation was that the GO Enrichment Terms were related to each other, but not exactly the same as the top clusters in O'meara et al. We believe that our analysis was correctly done, and the differences arisen are likely the result of different parameters or different versions of the tools used for the analyses. In conclusion, although the results of analyses may not be exact, the same overall biological functions were able to be observed between this study's groups and the original O'meara et al. paper.

References

[1] Bicknell, Katrina A., et al. "Can the Cardiomyocyte Cell Cycle Be Reprogrammed?" *Journal of Molecular and Cellular Cardiology*, vol. 42, no. 4, Apr. 2007, pp. 706–21. *ScienceDirect*, https://doi.org/10.1016/j.yjmcc.2007.01.006.

[2]Bowen, Robert E. S., et al. "Statistics of Heart Failure and Mechanical Circulatory Support in 2020." *Annals of Translational Medicine*, vol. 8, no. 13, July 2020, p. 827. *PubMed Central*, https://doi.org/10.21037/atm-20-1127.

[3]O'Meara, Caitlin C., et al. "Transcriptional Reversion of Cardiac Myocyte Fate During Mammalian Cardiac Regeneration." *Circulation Research*, vol. 116, no. 5, Feb. 2015, pp. 804–15. *ahajournals.org (Atypon)*, https://doi.org/10.1161/CIRCRESAHA.116.304269.

[4]Steinhauser, Matthew L., and Richard T. Lee. "Regeneration of the Heart." *EMBO Molecular Medicine*, vol. 3, no. 12, Dec. 2011, pp. 701–12. *DOI.org (Crossref)*, https://doi.org/10.1002/emmm.201100175.

[5] GEO, GSE64403, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE64403

[6] Harvard Chan Bioinformatics Core, *Introduction to RNA-seq Using High Performance Computing*, https://hbctraining.github.io/Intro-to-rnaseq-hpc-salmon/lessons/qc_fastqc_assessment.html

Supplemental Figures

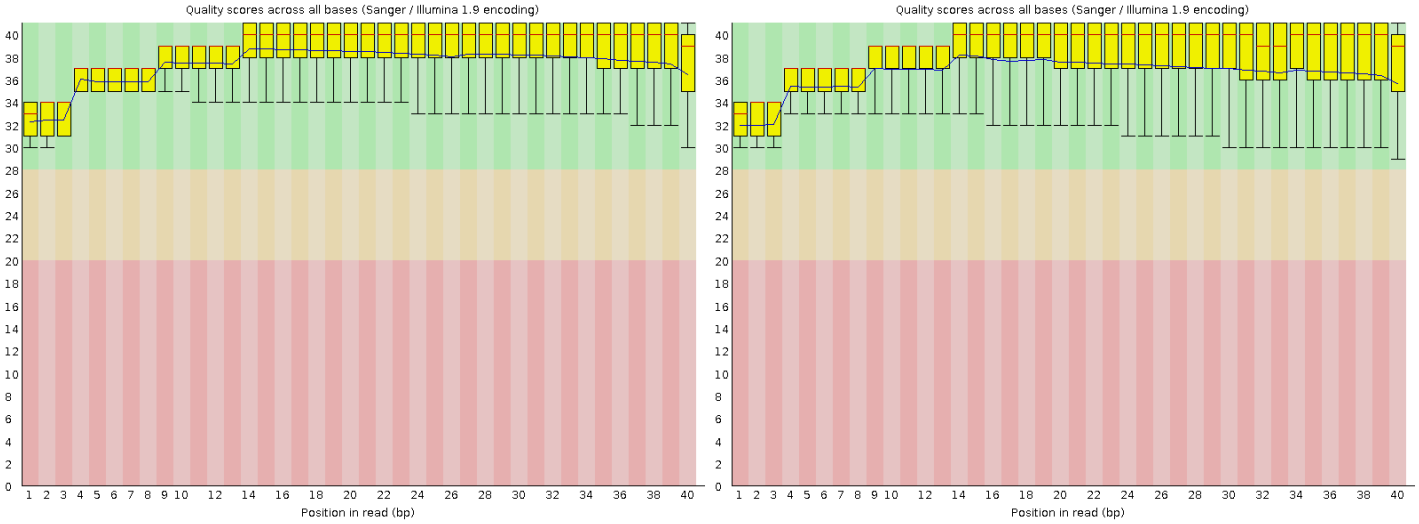

Figure S1 Per Base Quality Score: Quality score tells us the likelihood of incorrectly calling the base at a specific location. Each bar represents the distribution of quality scores. From the above figures we can see that for both fastq files all of our reads maintained a score of at least 30 across all locations, meaning a likelihood of an incorrect call of 1 in 1000.

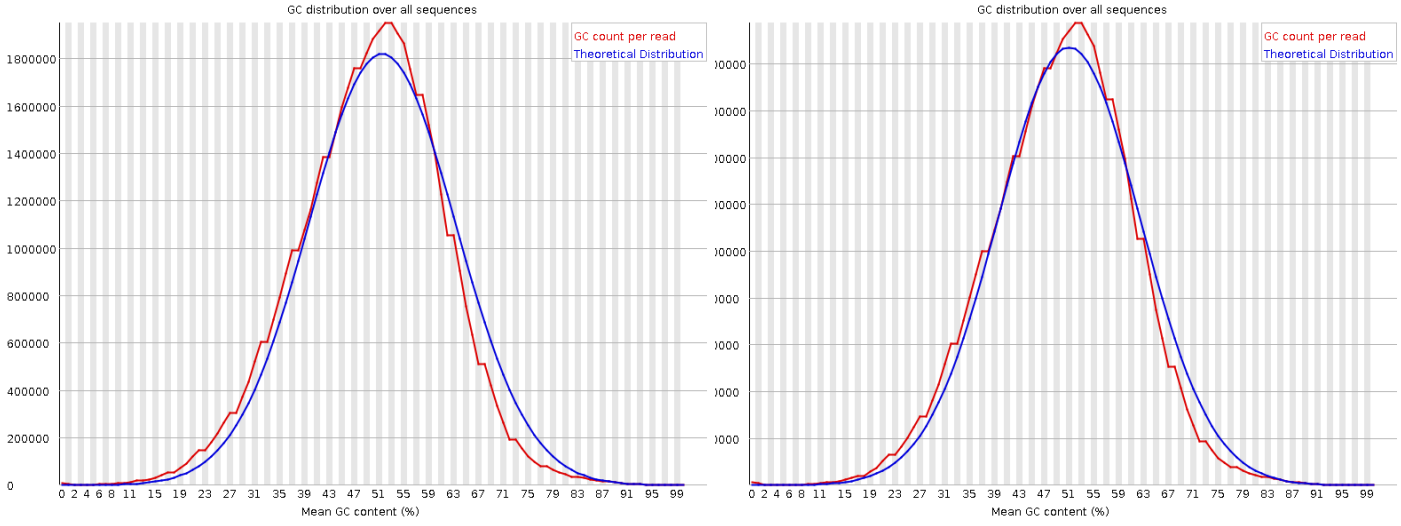

Figure S2 Per Sequence GC Content: These figures show us the GC content distribution across all reads (red) as well as the theoretical GC distribution (blue). As we can see, our read GC content aligns with the theoretical GC content fairly well.

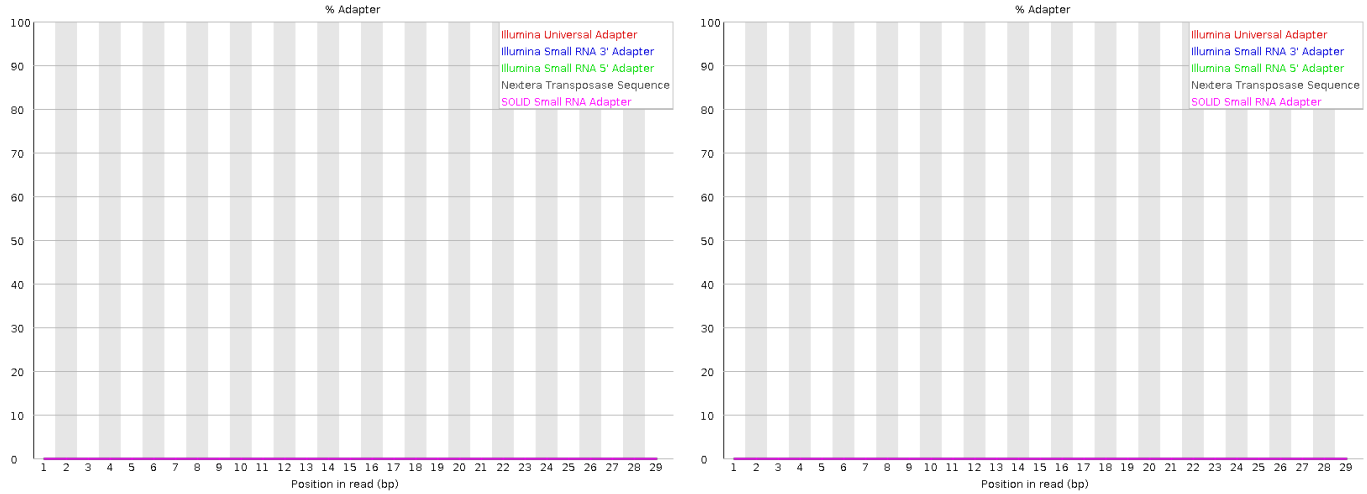

Figure S3 Overrepresented Sequences: These figures indicate whether there were any sequences that made up more than 0.1% of all sequences in our fastq files. If this were to happen it would likely be due to contamination, usually by adapters used in the sequencing process.

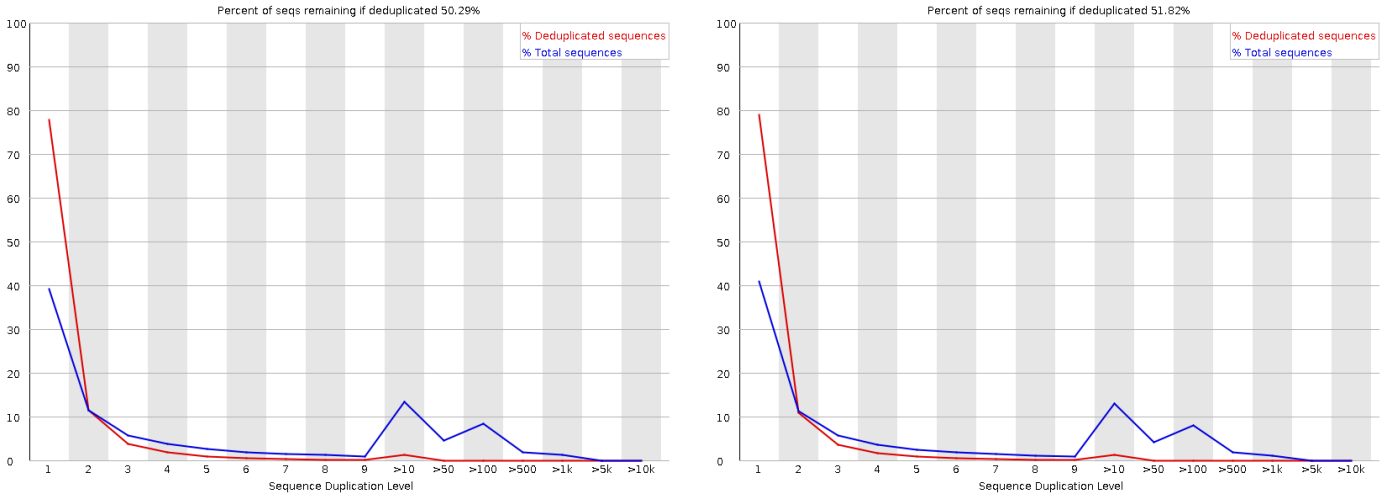

Figure S4 Duplicated Sequences: The above figure shows us how much of our sequence data is made up of non-unique sequences. Low levels of duplicated sequences is an indication of good coverage, but in the instances of certain gene sets being highly enriched it is expected that there would be large amounts of non-unique sequences. Since our data was generated in the context of CMs undergoing regeneration it is logical that gene sets and molecular pathways related to this process would show up in our data repeatedly.

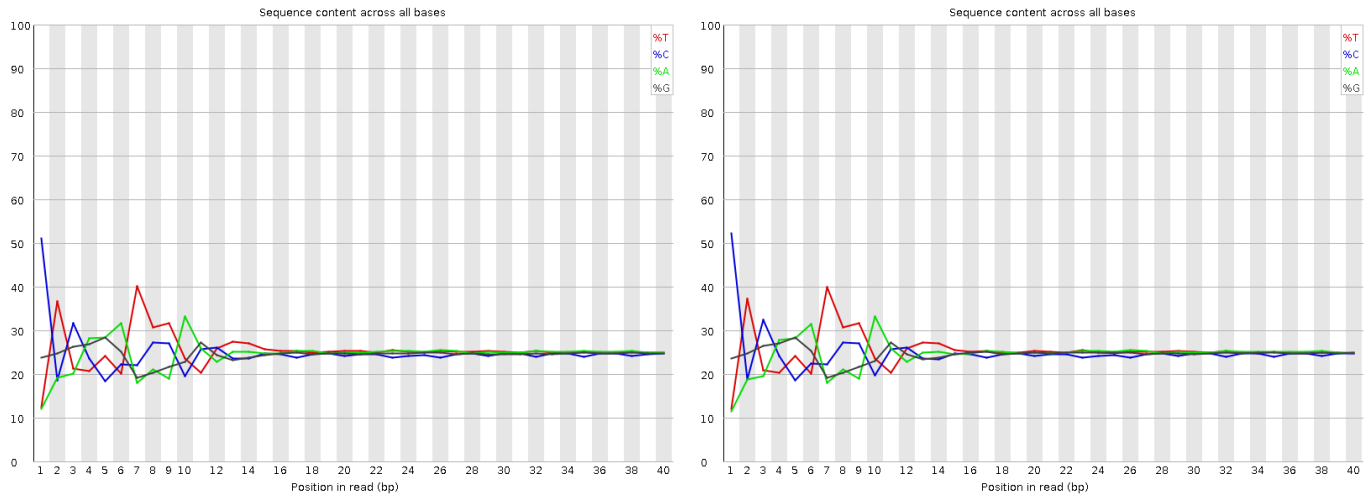

Figure S5 Per Base Sequence Content: The above figures show us the percent representation of each base (A, T, C, and G) at each position across all reads. During the preparation of RNA-seq data each sequence consists of random hexamer primers in the first 10-12 bases. This can be seen on the left side of each figure, and is why fastqc always gives this test a *failed* status for RNA-seq data.