## Transcriptional Profile of Mammalian Cardiac Regeneration with mRNA-Seq

### Introduction

Cardiac myocytes(CM) are responsible for heart growth in mammals. An adult mammalian heart has a limited potential for its cells to regenerate after injury, due to CMs leaving the cell cycle shortly after birth. Genetic fate mapping has shown that new CMs in a regenerating heart derive from preexisting CMs, rather than from a stem cell or progenitor population.

Partial reversion of cell fate in mouse heart repair has only been observed at the structural level[1]. This study was performed to replicate the in vivo part of the study of O'Meara, C.C. et al. to look at molecular roadblocks that could prevent regeneration in an adult mammalian heart.

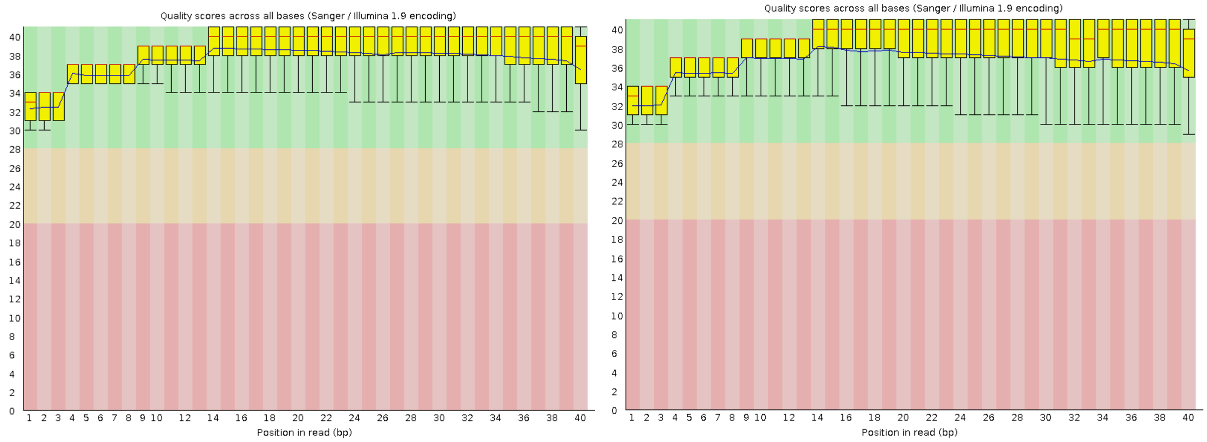
The bioinformatics methods performed included RNA paired-end read sequencing, sequence alignment, and assembly. This was needed to identify a common set of differentially expressed genes during in vivo CM maturation. The gene expressions of the in vivo maturation were used to find transcriptional changes that occur when the state of the CM destabilizes, and also during regeneration.

### Data

A gene sequence sample, which contained RNA-seq data for differentiation of stem cells into CMs, was downloaded from the GEO database as a short read archive (SRA) format file. The RNA was extracted by trizol reagent. Due to paired end sequencing, the SRA file was extracted to two FASTQ files.

With the two generated FASTQ files, the FASTQC package was used on the SCC server to generate visualizations of important statistics such as sequence duplication levels, per base read quality, per sequence content, and per base GC content, as seen in the figures below.
There were over 21.5 million total sequence reads, with read lengths of 40 nucleotides.

To ensure high quality base calling, a quality score of each sequence position is determined by the Read Quality plots, with each position containing a range represented by a box and whisker plot, the central red line being the median value, and the blue line being the mean[2]. The 10th to 90th percentile range of the scores for each position fell within the green range, indicating very good quality calls.



**1** **2**

**Figure 1.** Read Quality Table Metrics for each end of the sequencing run.

For each position, Per Base Sequence Content plots show the percentage of each of the four nucleotides. Each position should show little difference between each nucleotide. Due to there being a more than 20 percent difference between nucleotides at some positions, this is considered a failed QC module. RNAseq libraries usually have biased fragmentation, resulting in ore variability at the start due to priming of random hexamers.

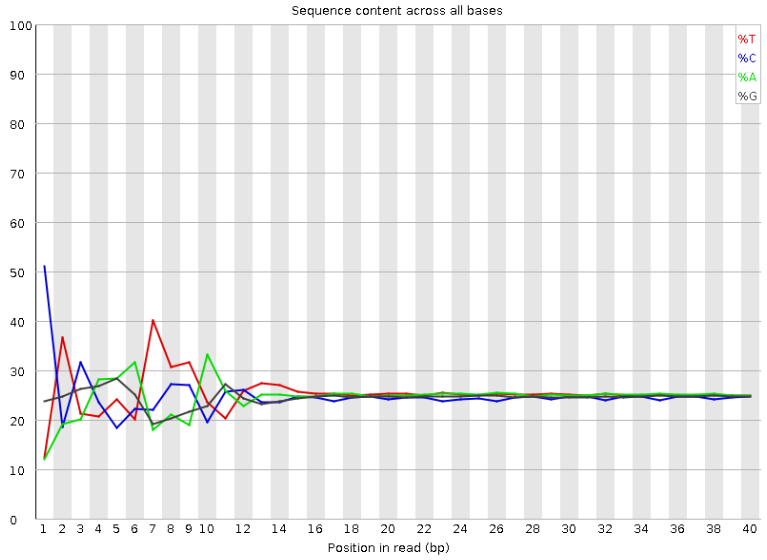


**Figure 2.** The Distribution of Nucleotide Sequences across all bases.

Figure 3 below measures the extent of duplication for every sequence as a percentage of all sequences, shown by the blue line. The red line represents the percentage of total sequences after deduplication.

This module was listed as a warning in the QC report due to having more than 20% of total sequences duplicated. Duplication can be caused by the RNA sample not having enough fragments. The other end of the sequencing read would have 51.82% of sequences left after deduplication.

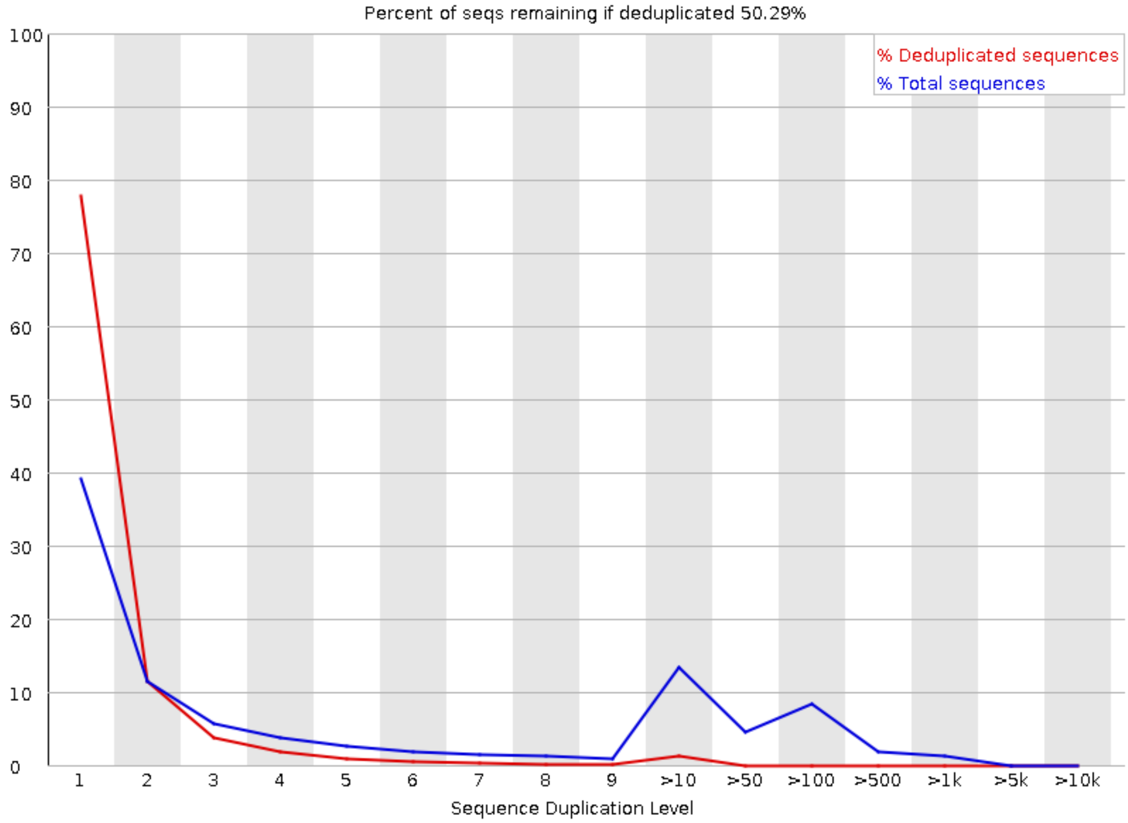


**Figure 3.** Percent of nucleotide sequences that remain if deduplicated

For both reads, the GC distribution followed a normal distribution. The central peak represents the overall GC content, which is usually expected to be 50% for genes in a human genome. The y-axis represents the number of sequences that contain GC content at a percentage. The normal distribution also indicates that the library is not contaminated and free of any biased subset.

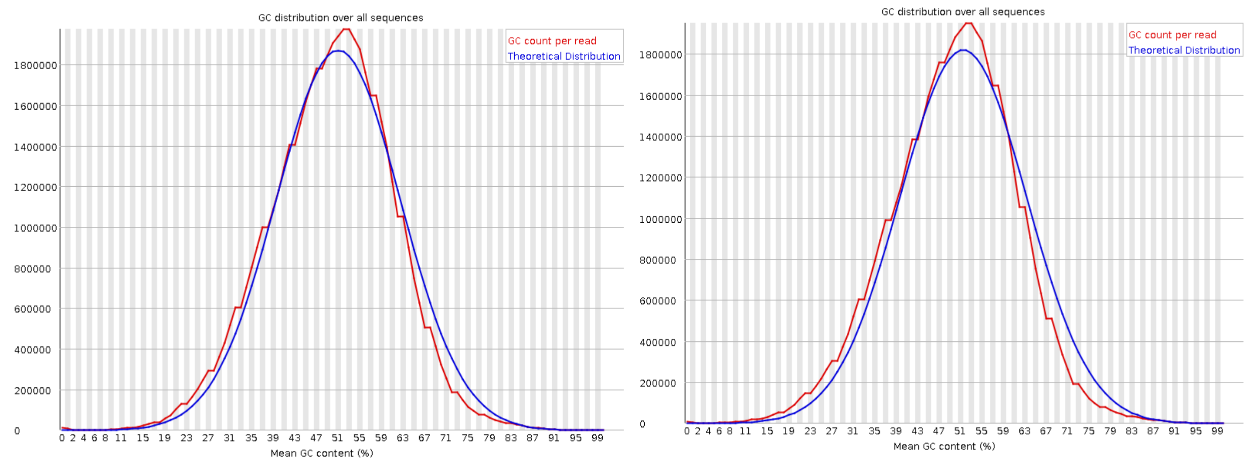


**Figure 4.** Concentration of GC over all sequences. The independent variable is Mean GC content percentage

## Methods

The sample GSM1570602 from GEO series GSE64403 was used for this analysis, which needed to be converted from an archived .SRA format to .fastq format using SRAtoolkit (v2.11.1). This yielded a pair of paired-end FASTQ files that were processed using FastQC (v0.11.7), with read quality, GC content bias, sequence duplication levels, and overrepresented sequences noted. The result of this initial data curation was a summary of the dataset quality as well as two FASTQ files used for downstream analysis.

The next step was aligning and quality assurance of the FASTQ files obtained in the earlier step using TopHat (v2.1.1), a fast splice junction mapper for RNA-Seq reads. TopHat required Python 2, Boost, and Samtools to run, and required both the FASTQ files as well as a reference genome to accurately align the two aforementioned sequences. For this analysis mouse genome mm9 and its indexes were used as the reference, and the end result was a .bam file with successfully aligned pairs. Since TopHat is computationally intensive running the program required the use of a HPC (High Performance Cluster), which took approximately an hour to complete alignment and mapping. The aforementioned BAM file was then indexed, and quality control metrics were then retrieved using RseQC Utilities (v3.0.0). This was done in order to ensure that there were no errors in alignment and mapping, as well as to guarantee the integrity of the original dataset.

After the FastQ files were mapped with quality control checking, Cufflinks (v2.2.1) was used in order to count how many reads mapped to annotated regions. The gene annotation file used was based on the mice genome (mm9.gtf), and inputted in conjunction with the reference genome as well as the indexed BAM file produced in the earlier step in order to produce a gene tracking file which listed FPKM values for all genes. FPKM values are used to quantify gene expression as in RNA-seq the relative expression of the transcript is proportional to the number of cDNA which originates from it. A graphical representation of the distribution of log10 FPKM values was then created, filtering out genes which have an FPKM value greater than one (Figure 8). 23264 genes were observed with FPKM values greater than 1. Afterwards, Cuffdiff(v2.2.1) was run on four additional samples, which reported on differentially expressed gene values.

Furthermore, it is important to note that a cuffdiff file that contains all of the differentially expressed gene information was read into R studio as a CSV file. Next, a table with the top ten differentially expressed genes, their FPKM values, log fold change, and p and q-values were all gathered. The hist function was utilized in order to produce a histogram of the log2.fold_change for all of the non-significant and significant genes. The subset function was used in order to produce an additional data frame with only the significant genes. The dataframe with the significant genes was then additionally subsetted into Up-regulated and Down-regulated genes and saved as two separate files. In total, there were 1084 genes that were Up-Regulated, 1055 genes that were Down-Regulated, and 36,329 total genes.

Onwards, the gene sets were then grouped into functionally related clusters via the DAVID software. The Up-Regulated and Down-Regulated gene lists were uploaded to the software and the Mus Musculus species was selected. Afterwards, under the gene ontology section of the software, GOTERM_BP_FAT, GOTERM_MF_FAT, and GOTERM_CC_FAT were all selected. A window that contains a list of the gene ontology terms were categorized into clusters based on how functionally related that they are. Then the DAVID results were compared to the David results obtained by O'Meara, C.C. et al.

Using the 3 lists of genes specific to the most prominent GO terms discovered in the analysis(Sarcomere, Mitochondria and Cell cycle) , 3 line charts of the FPKM data generated in this project were plotted using the ggplot package in Rstudio.. Before the plot, the P0_1 FPKM were combined with the other 7 samples provided in the project sample folders using gene_short_name. Mean was calculated for samples with same gene short name but different tracking IDs as the value for those specific genes.Also, the mean between the 2 duplicates were calculated as the data for the final FPKM table (P0,P4,P7,Ad).

Finally, a clustered heatmap of the top 1k genes found to be differentially expressed between P0 and Ad was generated using the heatmap function in R. Before plotting the heatmap, the 8 samples were also reduced to 4 by using mean value for the 2 duplicates of the 4 groups (P0,P4,P7,Ad).

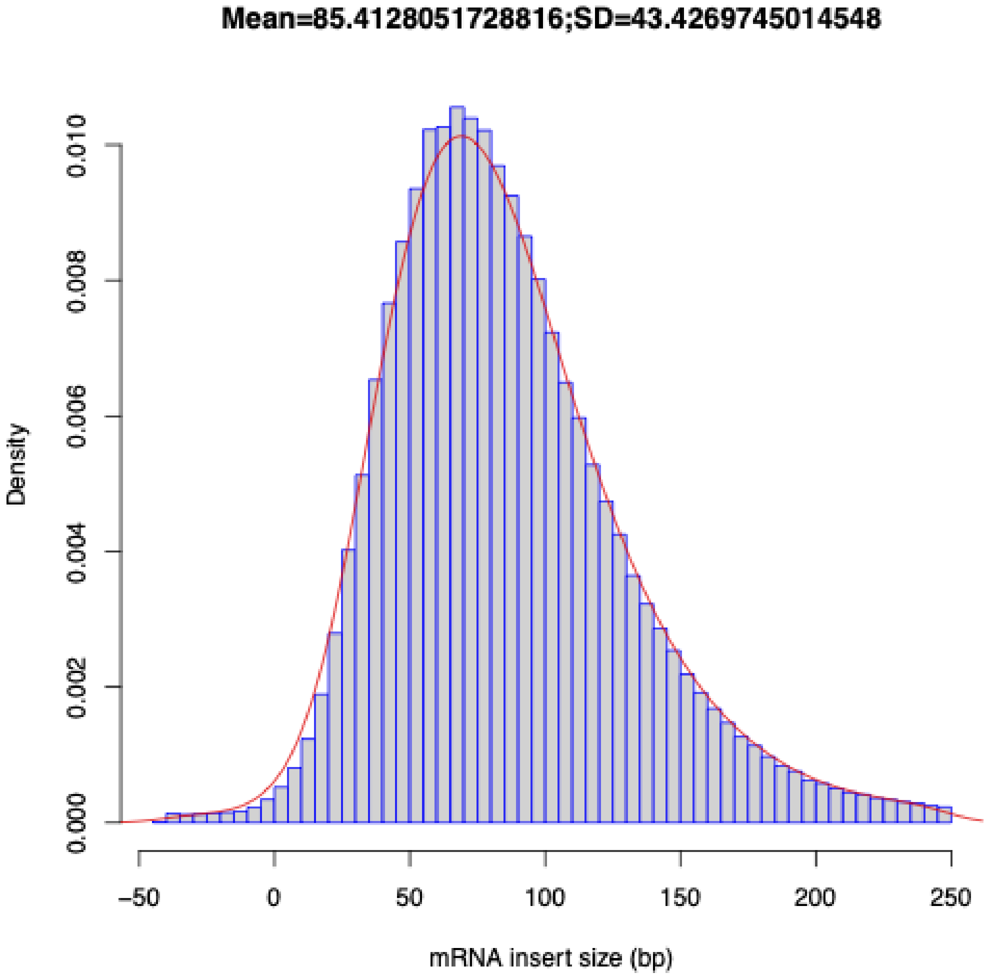


**Figure 6.** This figure describes the distribution of mRNA insert sizes in terms of base pairs, and the respective density of each size. The distribution shown here has a mean of insert size of 85.4 bps, with a standard deviation of 43.42 bps.

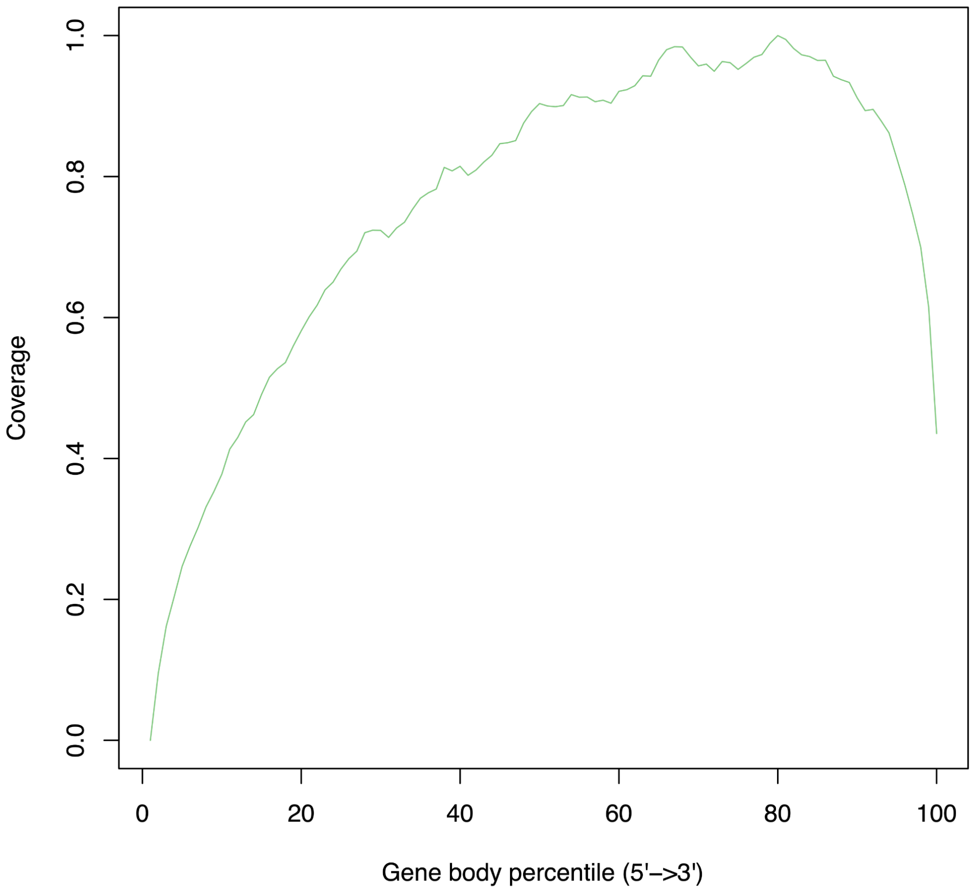


**Figure 7.** The Gene Body Coverage graph is used to determine whether there is 5' to 3' coverage bias, displaying coverage of reads between each respective end of the gene. This figure shows a slight 3' bias as the peaks are between 60-80, closer to the 3' end.

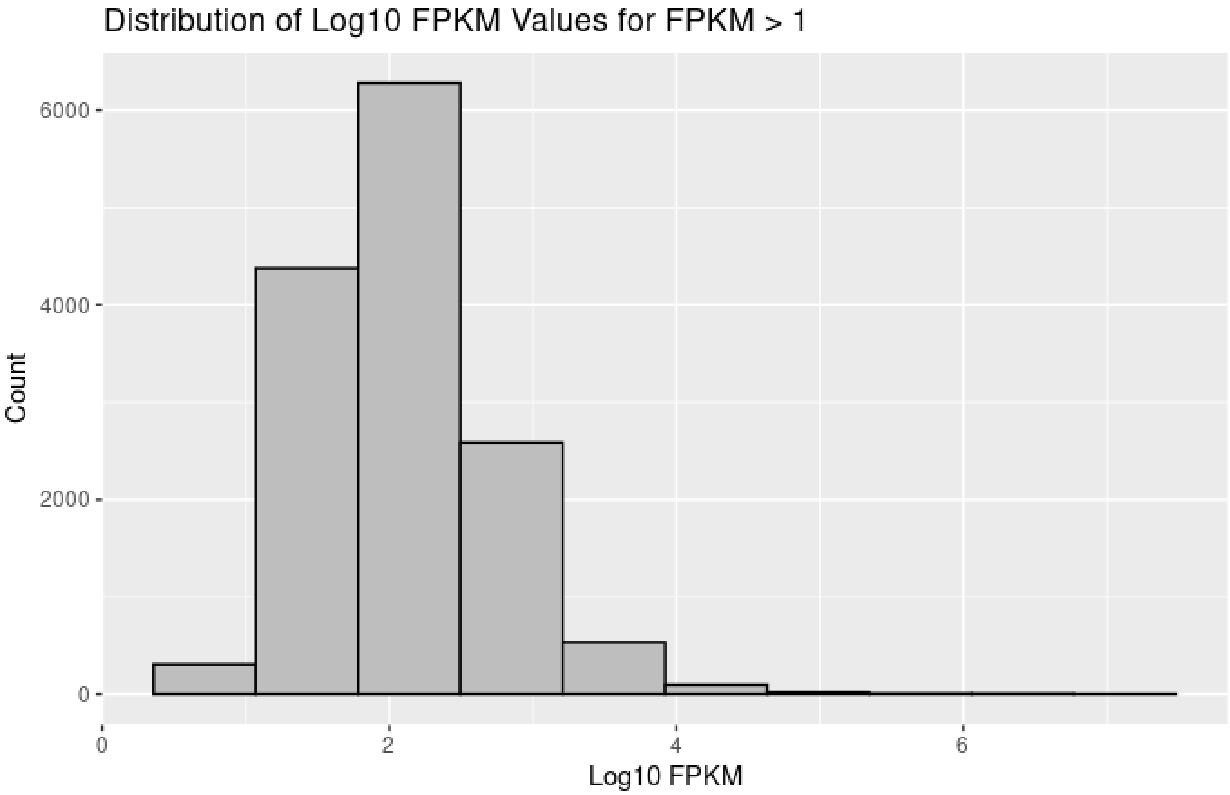


**Figure 8.** This graph illustrates the distribution of Log10 FPKM values when the FPKM is greater than 1. As seen, a Log10 FPKM value of 2 has the highest count of the number of genes.

## Results

The resulting QC on the FastQ files reported 49706999 total reads, 100% mapping rate, 77.43% unique reads, 5.83% multi-mapped reads, and 11.1% reads unaligned. The figures generated by RseQC also found a mean mRNA insert size of 85.41 base pairs with a standard deviation of 43.43 base pairs (Figure 6), as well as a slight 3’ bias based on the gene body coverage graph (Figure 7).

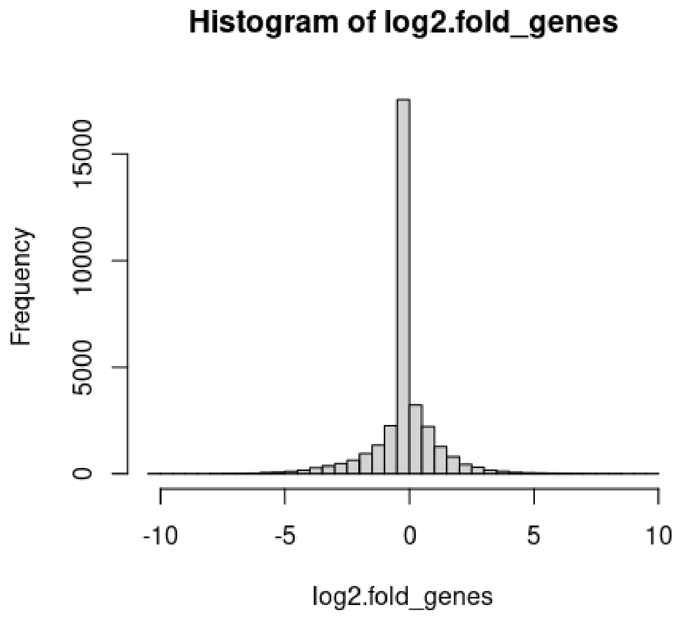


**Figure 9.** This graph illustrates the log2.fold change of the differentially expressed genes

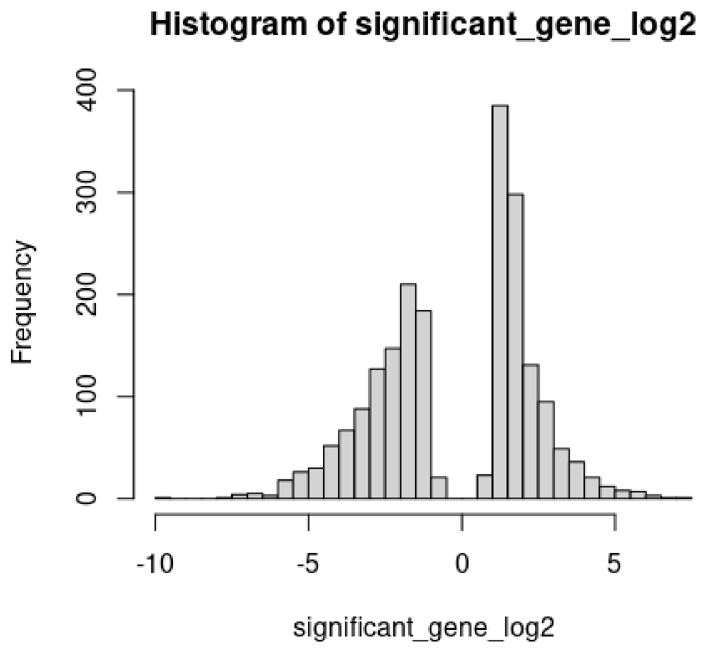


**Figure 10.** This graph illustrates the log2.fold change of the differentially expressed genes

The two log2 Fold distribution graphs above measure the frequency of the differentially expressed genes in two conditions: one when the genes are non_significant and the other when the genes are significant. The distribution, in general, measures the gene expression change of myocyte differentiation that occurs between both the P0 and the adult mice. In Figure 9, it is clear that there is a tall, unique peak that occurs at the distribution level of 0. More than likely, this distinct peak signifies that a majority of the expressed genes are not significant, which is why the subset function was used to create a dataframe with just the significant values.

Furthemore, Figure 10 shows no outlying peak near 0, which signifies that expressed genes are significant. All of the genes under a distribution level of 0 signify down-regulated genes, while on the other hand, all of the genes that are over the distribution level of 0 are the up-regulated genes. In total, there were 1084 Up-Regulated genes and 1055 Down-Regulated genes, and 36,329 total genes.

| Gene | 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 |

**Table 1.** These are the Top 10 differentially expressed genes with their FPKM data, Log2 Fold Change, p-value, and q-value

| **Differentially Expressed Genes** | **Quantity** |
|---|---|
| Up-Regulated Genes | 1084 Genes |
| Down-Regulated Genes | 1055 Genes |
| Total Genes | 36,329 Genes |

**Table 2.** Table summarizing the total Up-Regulated Genes, Down-Regulated Genes, and the Total Observable Genes

| DAVID Analysis on Up-Regulated Genes | | | |
|---|---|---|---|
| Cluster ID | Enrichment Score | Gene Ontology (GO) Enrichment Examples in Cluster | Overlap with the results of O'Meara, C.C. et al. |
| Annotation Cluster 1 | 52.28 | GO:0043436 ~ oxoacid metabolic process<br>GO:0006082 ~ organic acid metabolic process | YES |
| Annotation Cluster 2 | 42.7 | GO:0043167 ~ ion binding<br>GO:0043169 ~ cation binding | NO |

| | | | |
|---|---|---|---|
| Annotation Cluster 3 | 40.11 | GO:0070887 ~ cellular response to chemical stimulus<br>GO:0010033 ~ response to organic substance | NO |
| Annotation Cluster 4 | 37.53 | GO:0006091 ~ generation of precursor metabolites and energy<br>GO:0019637 ~ organophosphate metabolic process | YES |
| Annotation Cluster 5 | 36.86 | GO:0005739 ~ mitochondrion<br>GO:0031966 ~ mitochondrial membrane | YES |

**Table 3.** The table summarizes each annotation cluster, their enrichment scores, and sample Gene Ontology Enrichment examples for the Up-Regulated Genes with comparison to the results of O'Meara, C.C. et al.

| DAVID Analysis on Down-Regulated Genes | | | |
|---|---|---|---|
| Cluster ID | Enrichment Score | Gene Ontology (GO) Enrichment Examples in Cluster | Overlap with the results of O'Meara, C.C. et al. |
| Annotation Cluster 1 | 55.91 | GO:0043167 ~ ion | NO |

| | | | |
|---|---|---|---|
| | | binding<br>GO:0046872 ~ metal ion binding | |
| Annotation Cluster 2 | 38.44 | GO:0008283 ~ cell population proliferation<br>GO:0042127 ~ regulation of cell population proliferation | NO |
| Annotation Cluster 3 | 30.9 | GO:0010646 ~ regulation of cell communication<br>GO:0023051~ regulation of signaling | NO |
| Annotation Cluster 4 | 30.02 | GO:0009893 ~ positive regulation of metabolic process<br>GO:0034645 ~ cellular macromolecule biosynthetic process | NO |
| Annotation Cluster 5 | 28.34 | GO:0006996 ~ organelle organization<br>GO:0007010 ~ cytoskeleton | NO |

|  |  |  |  |
|---|---|---|---|
|  |  | organization |  |

**Table 4.** The table summarizes each annotation cluster, their enrichment scores, and sample Gene Ontology Enrichment examples for the Down-Regulated Genes,with comparison to the results of O'Meara, C.C. et al.

Plekhb2, Mrpl30, Coq10b, Aox1, Ndufb3, Sp100, Cxcr7, Lrrfip1, Ramp1, Gpc1 were Top 10 differentially expressed genes found(Table 1). A total of 1084 up-regulated genes and 1055 down-regulated genes were found. (Table 2). The top five up-regulated genes clusters generated by DAVID were identified to have the GO terms of  oxoacid/ organic acid metabolic process, cell population proliferation, cell communication, generation of precursor metabolites and energy as well as mitochondrion (Table 3). The top five down regulated genes clusters generated by DAVID were identified to have the GO terms of ion binding, cell population proliferation, regulation of signaling, regulation of metabolic process and cytoskeleton organization (table 4). Three of the five top up-regulated genes clusters overlapped with the DAVID results of O'Meara, C.C. et al. None of the down-regulated genes clusters identified in this project overlapped with the DAVID results of O'Meara, C.C. et al.

Figure 11 showed the FPKM fold change of significantly differentially expressed genes identified during in vivo maturation. Mpc1 from the Mitochondrial gene list were missing in the FPKM table generated in this project, compared to  the results of O'Meara, C.C. et al. Bora from the Cell Cycle gene lists were missing in the FPKM table generated in this project, compared to the results of O'Meara, C.C. et al. These values may be excluded because they were pretty low. Except for these missing values,  the FPKM values and fold change in FPKM seemed to have exactly the same trend and magnitude for all rest of the genes. The Sarcomere genes and Mitochondrial genes all show an overall  up-regulated trend from P0 to Ad while the Cell cycle genes all show an overall down regulated trend.

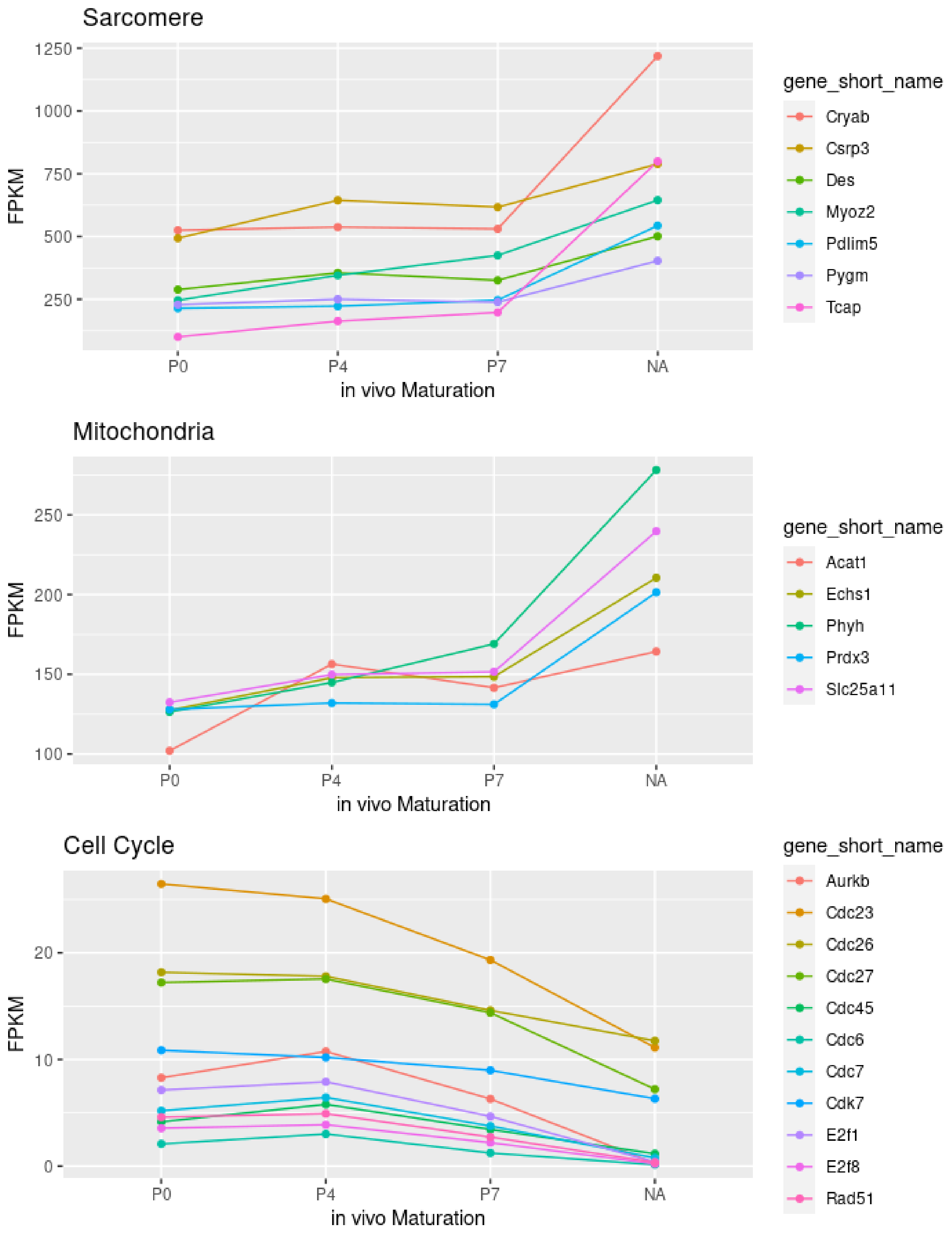


**Figure 11.** Line charts showing the FPKM fold change of significantly differentially expressed genes during in vivo maturation

The heatmap in Figure 12 was based on the top 1000 DE genes from P0 vs Ad. The result clearly showed that these 1k genes have completely opposite expression patterns . The P4 and P7 almost  have all the fpkm values in the middle, indicating that  the trends of up-regulation or down-regulation are all continuous.The clusters were a little messy on the row side. The blue(low) and yellow(high) parts of the heatmap are not so separate. The clustering did not find a clear pattern on the row(genes) side.

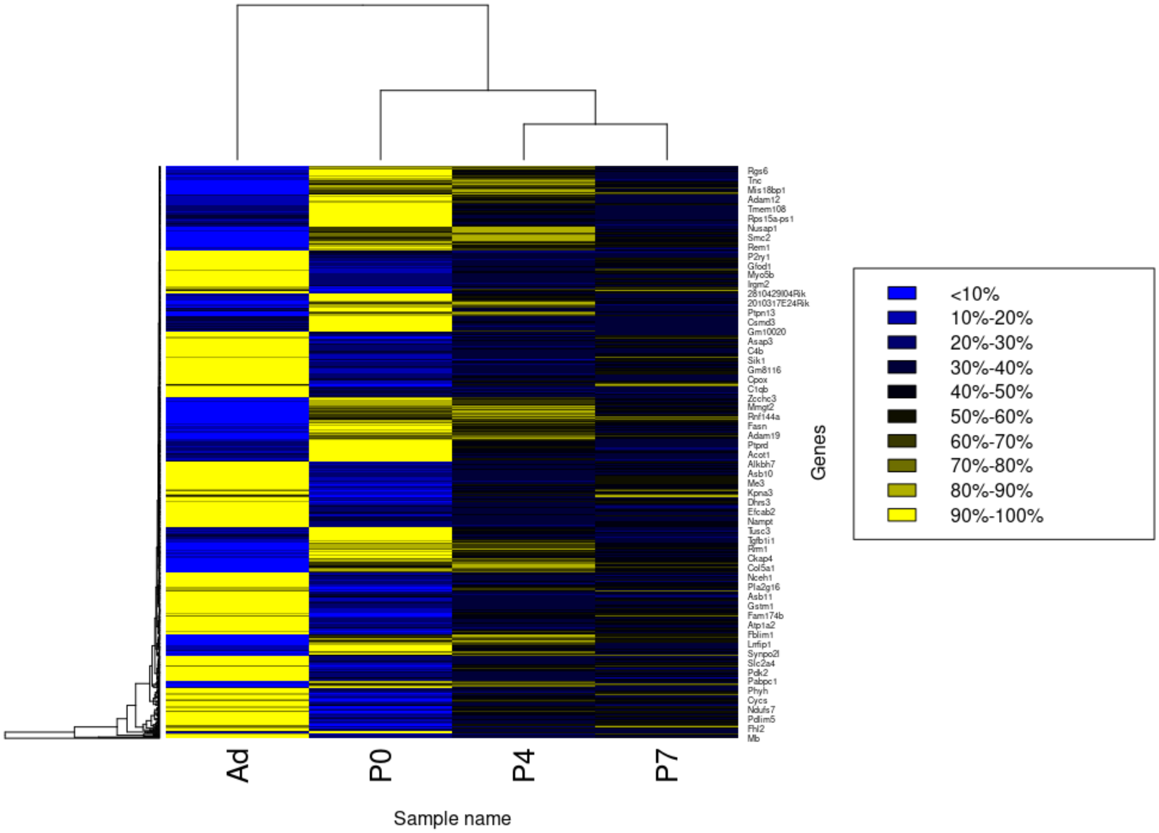


**Figure 12.** Heatmap of FPKM values of the top 1000 DE genes found in the P0 vs Ad Analysis

## Discussion

The overall methodology of this study begins with data acquisition, downloading the sample from the GEO series and converting the .SRA files to FastQ format. The FastQ files were then aligned with Tophat, with quality control metrics produced by RseQC, and the resulting

BAM file indexed. The indexed BAM file was then inputted into Cufflinks to produce quantifiable gene expression data, and run with other samples (post-natal vs. adult) using CuffDiff in order to identify differentially expressed genes. Genes which were observed to have significant log2 fold change values were extracted, as well as a list of up and down regulated genes. These genes and their respective enriched gene clusters were then identified by DAVID. The authors of the paper made clear that transcription factors such as STAT3/periostin play a critical role in the mediation of interleukin 13 signaling, which is highly important in cardiac myocyte (CM) regeneration. In particular, STAT3 was found on the list of up-regulated genes, which further supports the authors' prediction that STAT3 has a key importance in the cm regeneration process. Analysis of the FPKM values for samples in different stages of development, created by Cufflinks, was also further analyzed.

It is discovered in this project that the Sarcomere genes and Mitochondrial genes all show an overall  up-regulated trend from P0 to Ad while the Cell cycle genes all show an overall down regulated trend. These trends agree with the trends found in the paper by O'Meara, C.C. et al. Sarcomere would grow with the age increasing, thus an upregulation trend is not surprising. Mitochondria genes are related to metabolisms, which also increases with age increasing..  Cell cycle genes are down-regulated, which can be due to they are linked with differentiation and thus depressed with age increasing.

The heatmap from this project was not similar to the paper's heatmap. It was more messy compared to the paper's as well. One reason is that the heatmap only has the in vivo part while the paper's heat map also includes the in vitro differentiation part and Adult CM Explant part. The paper's heatmap had GO terms on the side instead of gene IDs. Another reason is 1k genes were selected as the top 1k DE genes between Ad and P0. None of the other groups were taken into consideration. This also made the heatmap look very different to the paper's. Unlike the line charts, due to the two reasons mentioned above,it is very hard to verify this plot using the papers.

The top 5 up-regulated genes clusters are mostly related to metabolic process. Mitochondria and other metabolism related genes increase expression with age increasing .3 of the clusters also appeared in the results by O'Meara, C.C. et al. These results also agree with the trend discovered by the line charts. The two missing clusters from the paper are about Sarcomere and Sarcoplasm. This may be due to the dataset the project worked on only looked at in vivo maturation and those two clusters were not in the top 5.

The top 5 down-regulated genes clusters do not show a clear common pattern as the up-regulated genes clusters. The results also did not overlap with the 5 top clusters identified in the paper by O'Meara, C.C. et al. This project found cell proliferation, ion binding and regulation of metabolism and signaling as well as cytoskeleton organization to be mostly down -regulated gene functions with age increasing. However the paper found cell cycle regulation, differentiation regulation and RNA processing, which also related to differentiation, to be mostly down regulated. These results found by th.e paper could give a reason to the loss of the regeneration function – differentiation. By contrast, the findings in this project could not give a good reason. This contrast can also be due to the dataset the project worked on only looked at in vivo maturation.

From the previous results discussion, it could be found that the strength of the paper is that the paper profiled in vitro systems that specifically model a transient cardiac myocyte state (in vitro differentiation and adult cardiac myocyte explants) to address confounding factor that numerous additional cell types may also contribute to the transcriptional profiles obtained via RNA profiling and some changes in cell type composition may change transcription. By profiling and analyzing multiple models, they could thus identify the expression change during the differentiation and loss of differentiation processes specific to cardiac myocytes. Since the project only looked at the in vivo part, it is not so surprising to find huge differences between the project results and the paper results.

## Conclusion

In conclusion, the analysis of genes responsible for cardiac regeneration and destabilization in mammals during various stages of development provides an interesting insight into cardiac myocytes. It was found that the most up-regulated gene clusters were related to the metabolic process, while the most down-regulated genes did not show any discernible biological pattern. Furthermore, all Sarcomere and Mitochondrial genes were found to be more up-regulated throughout development, which was expected as these continue growing with age.

The main challenges faced during the QC and quantifying gene expression step would be the runtime involved, as each process took at least an hour to run, requiring the use of a high performance cluster in order due to computational constraints. This made the process of debugging any runtime errors extremely inefficient since there was not a way to run the code piecewise. In order to rectify this inefficiency, detailed runtime logs were kept in order to identify exactly where the process was breaking, and commands were first on the command line to find any syntax or value passing errors.

Challenges faced during analysis were mainly induced when trying to match the results of this project to that of the original paper's, as there were large discrepancies with the results. This could be explained with the fact that this project only worked with samples in vivo, while the original study worked with both in vivo and in vitro samples, drawing upon a much larger pool of data and thus more genes to be examined.

## References

1. O'Meara CC, Wamstad JA, Gladstone RA, Fomovsky GM, Butty VL, Shrikumar A, Gannon JB, Boyer LA, Lee RT. Transcriptional reversion of cardiac myocyte fate during mammalian cardiac regeneration. Circ Res. 2015 Feb 27;116(5):804-15. doi: 10.1161/CIRCRESAHA.116.304269. Epub 2014 Dec 4. PMID: 25477501; PMCID: PMC4344930.

2. Per base sequence quality. (n.d.). Retrieved March 17, 2022, from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/2%20Per%20Base%20Sequence%20Quality.html