**Project 2 - Transcriptional Profile of Mammalian Cardiac Regeneration with mRNA-Seq**

**Group Members:**

Jiaming Zhang (Data Curator)

Mikhail Kouzminov (Programmer)

Jinghan Huang (Analyst)

**TA:**

Kritika Karri

**Introduction**

In the study, *Transcriptional reversion of cardiac myocyte fate during mammalian cardiac regeneration* by O’Meara CC et al., the group investigated the phenomenon of neonatal mice’s capacity to regenerate heart tissue in response to injury in the first week of life. The objective of the team’s study is to determine whether mouse myocytes revert to less differentiated state upon receiving injury. The authors characterized the transcription profile of injury-induced cardiac myocyte with both in vitro and in vivo models using high throughput RNA sequencing. In our analysis, we performed the analysis on the expression signature of one of the samples used in O’Meara CC et al.’s study.

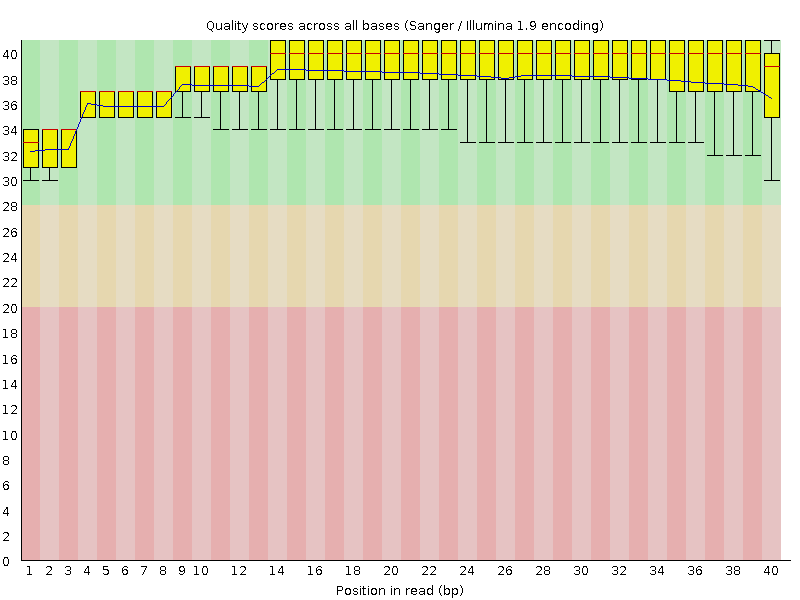
**Data**

**Data Description**

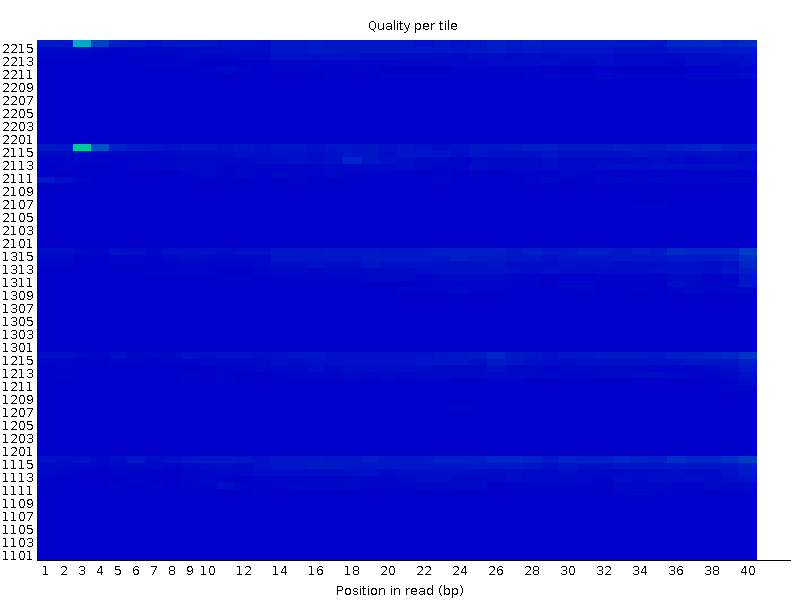
Total RNA from primal mouse primary cardiomyocytes were isolated using Trizol reagent. RNA quality was determined by Agilent Bioanalyzer. Sequencing was run on Illumina GA-2 and Illumina Offline BaseCaller1.9.3 software used for base calling. The sequencing file SRR1727914.sra for sample GSM1570702 was split into two FASTQ files for each read direction using sra toolkit. The pair end reads has a read length of 40 bp. Upon examining the FASTQ files from both read directions, the total sequence count was 21577562 and none of sequences has been flagged for poor quality.

**Data quality control**

Both FASTQ files from two read directions were examined and they agree with each other, regardless of the existence of small variations. For the purpose of examining the quality of read files, only the forward direction FastQC report images were attached. The per base sequence quality appears to be really high across the entire read length, indicating that the samples are not overly degraded, and the read length selection is adequate(Fig 1). This is confirmed by Per tile sequence quality as the color of each tile ranges from light blue to deep blue(Fig 2).

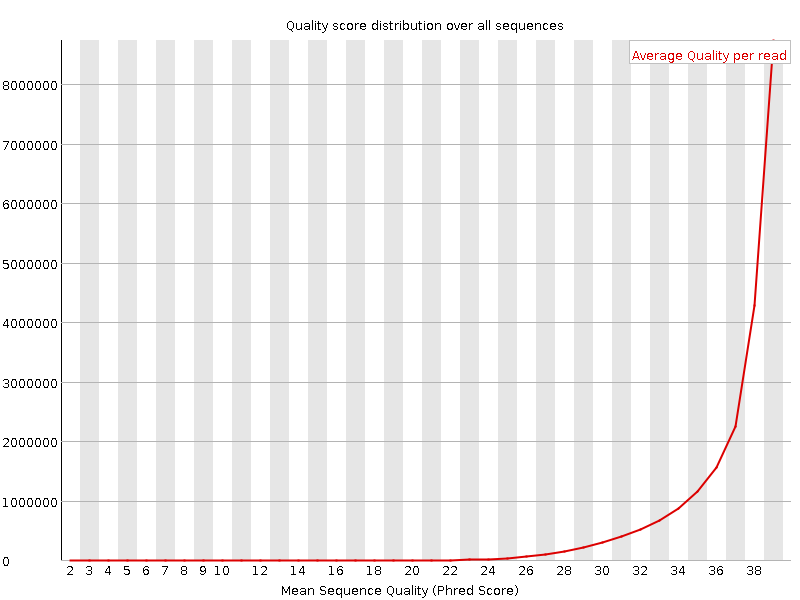


**Fig.1 Quality scores across all bases**



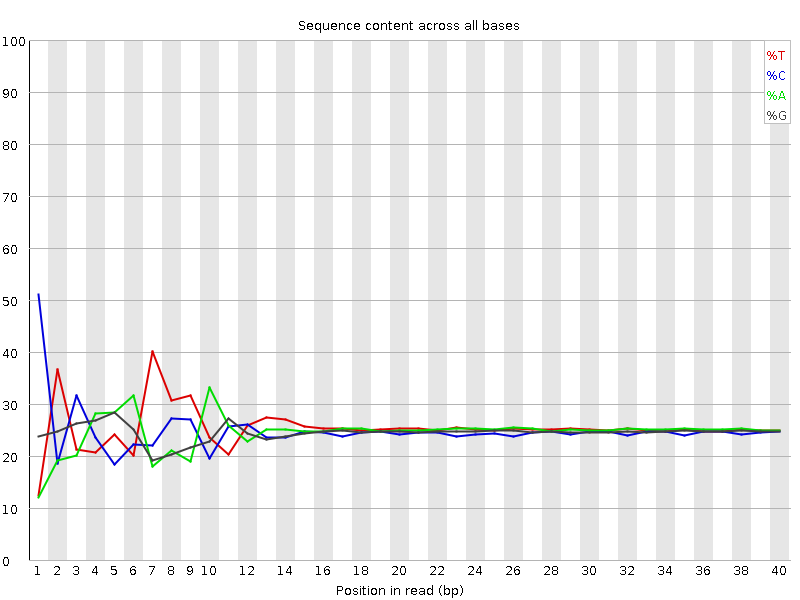
**Fig.2 Quality per tile**

The per sequence quality scores showed that the most frequently observed mean phred score is >38 indicating good quality on the majority of sequences(Fig 3).



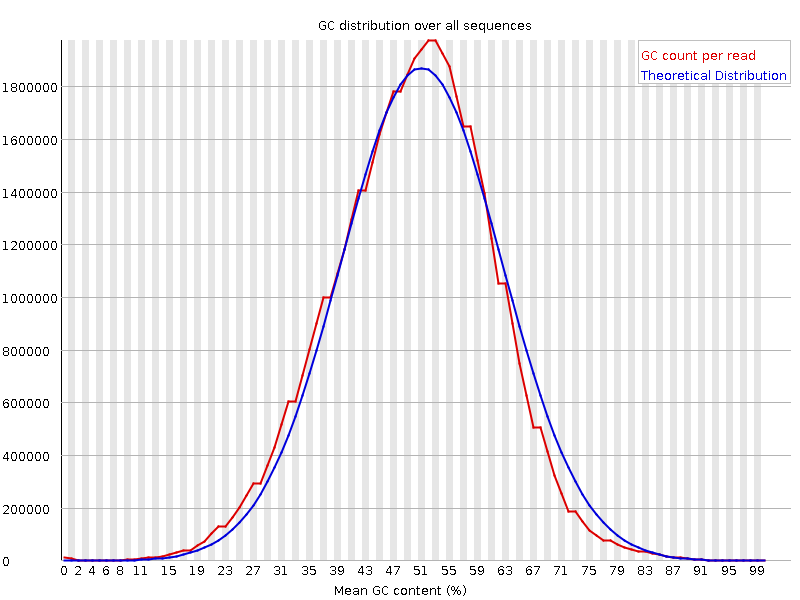
**Fig.3 Quality score distribution over all sequences**

FastQC raised a warning for per base sequence content. The failed test is a result of certain read positions having a differential base content greater than 20%. Examining the graph, read position 1 has C represented at >50% and A represented at <15%; Position 7 has T over represented at 40 % and A at <20%. From position 2 to 10, most of the read position has an equivalent representation of 4 different bases with a difference above 10%. However, since this is an RNA-Seq experiment, the libraries were most likely produced by using random hexamer priming, which introduces sequence bias at the beginning of the reads, which means this isn’t a true bias(Fig 4).



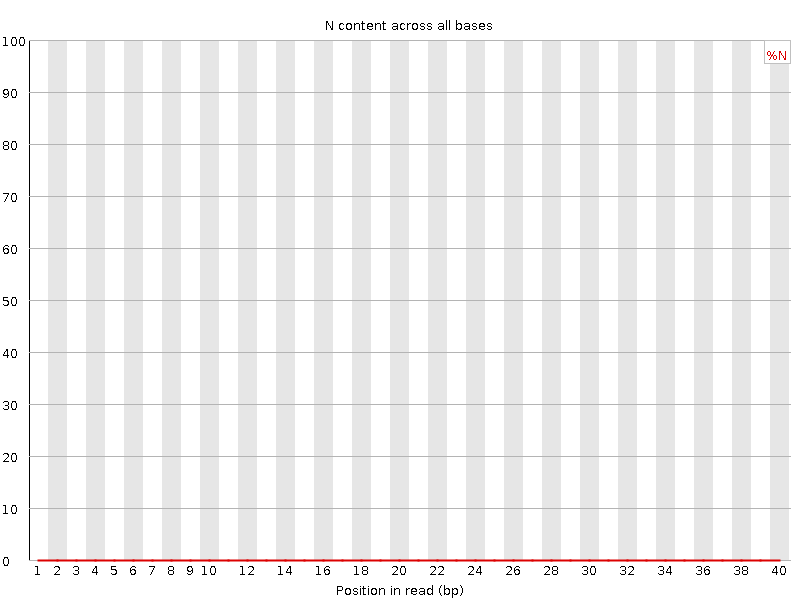
**Fig.4 Sequence content across all bases**

The per sequence GC content graph showed the GC content per read almost overlaps with the normal distribution curve indicating there is no bias in the library(Fig 5).



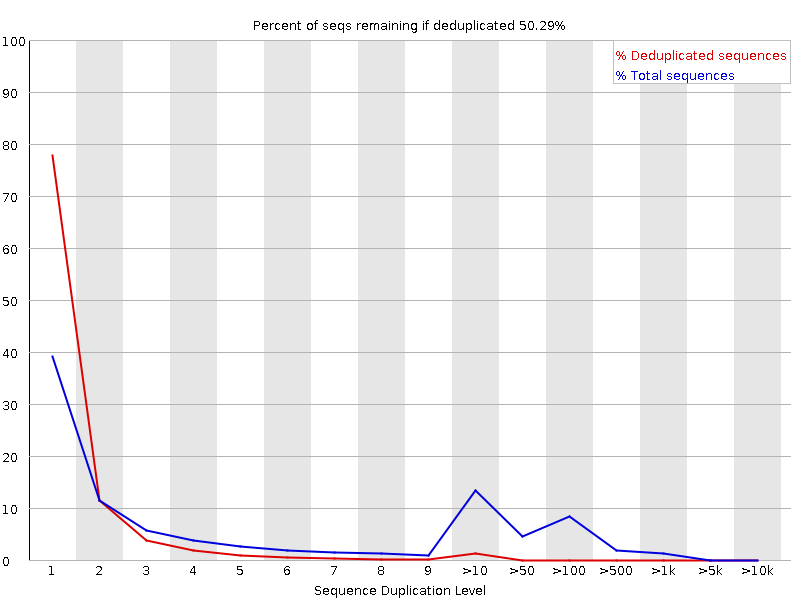
**Fig.5 GC distribution over all sequences**

The per base N content showed that %N at each read position across the sequences are close to 0, meaning the interpretation program pipeline successfully interpreted and called out the base of the short reads(Fig 6).



**Fig.6 N content across all bases**

FastQC raised a warning for sequence duplication levels. The blue line showed spikes at high duplication levels meaning the original full sequence set has significant level of duplication in a large proportion of the sequences. The red line remains flat at that position since the red line indicates the proportion if the original samples were de-duplicated. Due to the fact that transcripts levels vary greatly, in order to obtain low expression mRNAs, it’s common to over-sequence high expression transcripts in order for the low expression mRNA to be detected. Although FastQC indicated that 51.82~50.28% of the sequence will remain after deduplication, for the intent of a RNA-Seq library, the high duplication level of the transcripts is appropriate. We conclude that the RNA-Seq data is high quality and suitable for follow up analysis(Fig 7).

****

**Fig.7 Percent of seqs remaining if deduplicated: 50.29%**

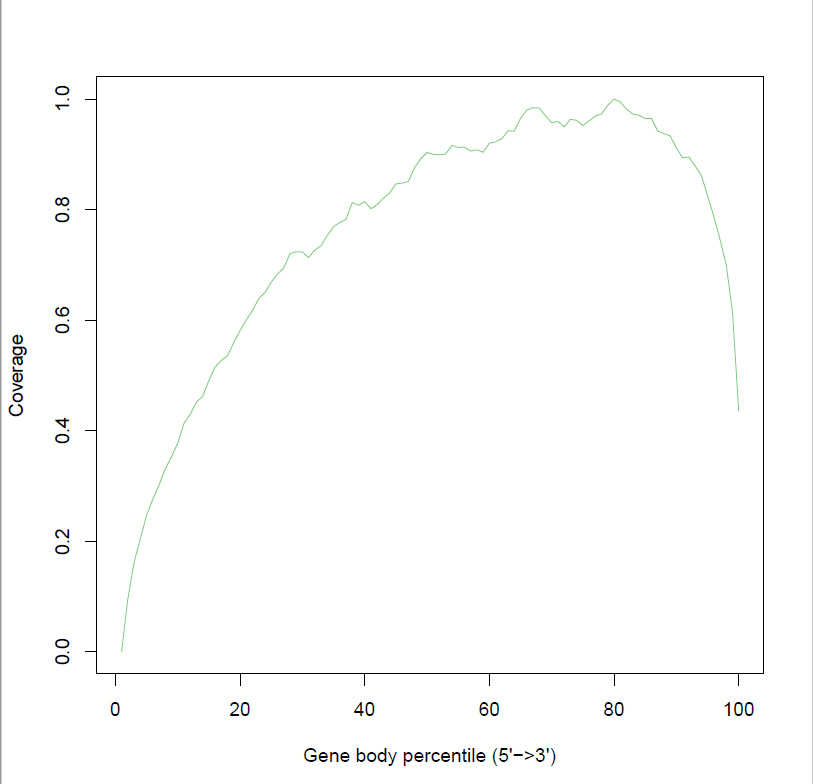
**Mapping Quality Control**

The obtained reads were aligned to the mouse reference genome mm9 using references obtained from Tophat’s website [2]. Using the samtools utility flagstat, and the RseQC utilities, we found that all 49,706,999 reads passed mapping quality control for tophat and were mapped to the reference genome. 32466938 of them, or 65.32% were properly paired with ones from the opposite strand in sequencing. Overall, we had obtained 100 % mapping quality.

|  |  |  |
| --- | --- | --- |
|  | Count: | Percent: |
| Total number of reads: | 49706999 |  |
| Mapped Reads: | 49706999 | 100 |
| Unique Reads: | 38489380 | 77 |
| Multimapped Reads: | 2899954 | 6 |
| Unaligned Reads: | 0 | 0 |

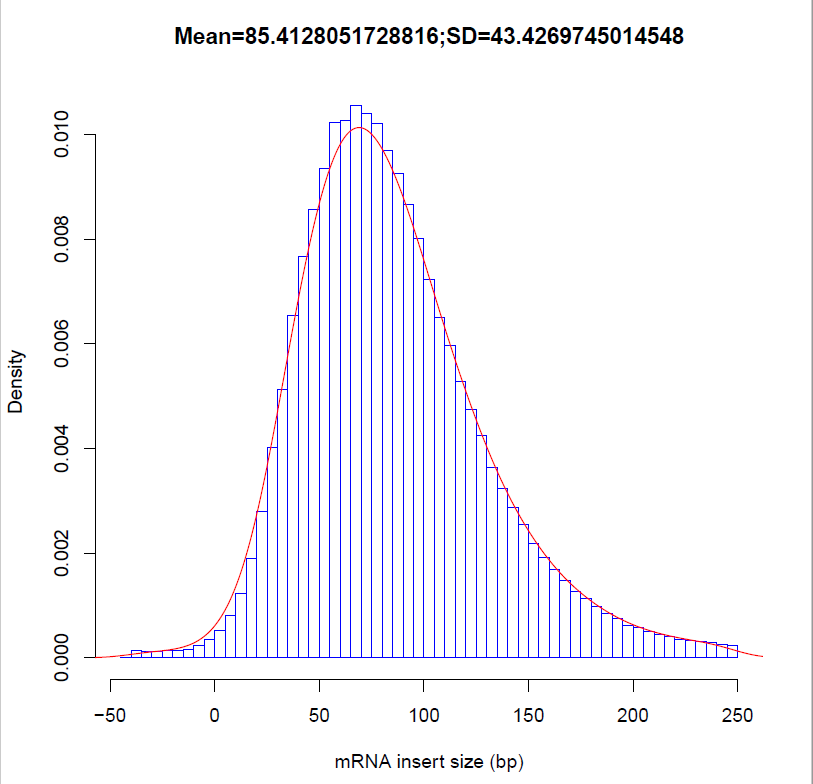
**Table 1: Counts and percentages of mapped read stats**

Running GeneBody\_coverage.py from RseQC to see if there is any 5’ or 3’ end coverage bias, we get what appears to be some 3’ coverage bias as shown by how skewed Fig.8 is. This might be due to using poly-A selection to obtain the sample.



**Fig.8 Gene Body Coverage, x axis represents location on gene, y represents coverage of region**

The mean insert size is around 85 base pairs. There is a relatively normal distribution. This is relatively small mean insert size, but they might have wanted to try and get a smaller insert size when collecting the data in order to get better coverage.



**Fig.9 inner distance plot , x axis represents insert size, y axis represents frequency**

**Methods**

**Aligning with Gene annotation**

Cufflinks was used to align with a gene annotation for the mm9 genome. A histogram of FPKM (Fragments Per Kilobase of exon per Million fragments) values was created to show a summary of how expressed certain genes are. Genes with low FPKM values were less expressed than those with high values. The histogram, and some further analysis are included in the results section.

**Differentially expressed analysis**

The differentially expressed analysis was performed to obtain the difference between the P0 and Adult clusters. The top differentially expressed genes had been selected out using the label ‘significant=yes’, p value < 0.01 and q value<0.01 respectively. The q value is another useful indicator for significance based on false discovery rate. The distribution of the log2 fold changes for all genes and the significant genes only had been plotted as histograms and compared with each other. The significant genes set was counted and subset to up-regulated gene set and down-regulated gene set according to their value of log2 fold change. Genes with positive log2 fold change represented up-regulated genes while genes with negative log2 fold change represented down-regulated genes. The symbol names of these up-regulated genes and down-regulated genes had been written into two separated files for following enrichment analysis. We reported the number of the differentially expressed genes as well as the number of the up-regulated genes and down-regulated genes in a table(Table 2) shown in Results part.

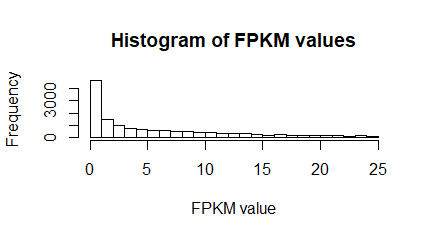
**Enrichment analysis** [3]

The enrichment analysis was performed to learn about the functional role that the up-regulated genes and down-regulated genes played. The DAVID Functional Annotation Tool was used. The symbol of the up-regulated genes and the down-regulated genes were placed in DAVID separately to obtain the GO terms results. The GOTERM\_BP\_FAT, GOTERM\_MF\_FAT, and GOTERM\_CC\_FAT boxes, which represented the summarized version of biological processes, molecular function and cellular component respectively, had been selected to conduct the Functional Annotation Clustering. This function in DAVID allowed us to organize the enriched gene sets into multiple functionally related clusters.

**Results**

**Analysis of FPKM values**

A histogram of genes with FPKM between 0.1 and 25:



**Fig.10 Histogram of FPKM values > 0.1 for all genes in this genome**

Most expressed genes were found to have very little expression and be below our chosen threshold value of 0.1, but there are still hundreds with a significantly high FPKM value. 0.1 was used as a cutoff value for the histogram because it is high enough to potentially be significant, but still low enough to show the range most genes are expressed at. Of 37469 genes identified, 18135 had over 0.1 FPKM and were potentially significant, 13501 had over 1 FPKM, 6944 had over 10 FPKM, and 826 had over 100, with the maximum value being Mir5105 with 2303120 FPKM.

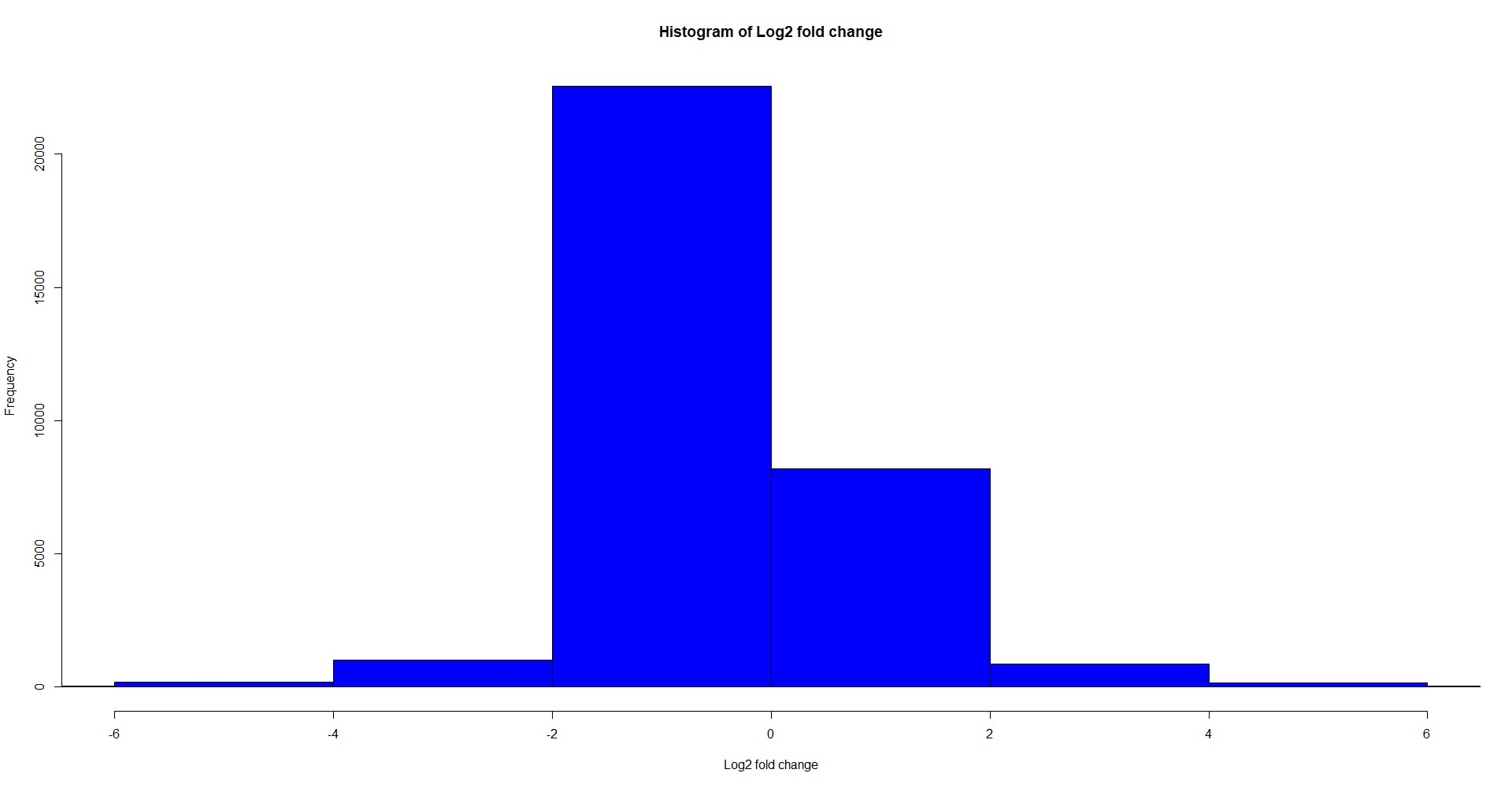
**Top 10 differentially expressed genes and statistics**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **gene** | **FPKM** | **Adjusted FPKM** | **Log2.fold change.** | **P value** | **Q value** |
| Rb1cc1 | 12.1937 | 31.9405 | 1.38925 | 5.00E-05 | 0.000319 |
| Pcmtd1 | 13.3652 | 30.17 | 1.17464 | 5.00E-05 | 0.000319 |
| Adhfe1 | 13.548 | 27.0353 | 0.996765 | 5.00E-05 | 0.000319 |
| Tmem70 | 36.5913 | 85.0414 | 1.21666 | 5.00E-05 | 0.000319 |
| Gsta3 | 0.414547 | 7.11348 | 4.10095 | 5.00E-05 | 0.000319 |
| Lmbrd1 | 6.701 | 13.3173 | 0.990848 | 5.00E-05 | 0.000319 |
| Dst | 18.9423 | 54.2207 | 1.51723 | 5.00E-05 | 0.000319 |
| Plekhb2 | 26.635 | 72.0352 | 1.43538 | 5.00E-05 | 0.000319 |
| Mrpl30 | 55.0179 | 130.538 | 1.24649 | 5.00E-05 | 0.000319 |
| Tmem182 | 46.0296 | 108.74 | 1.24025 | 5.00E-05 | 0.000319 |

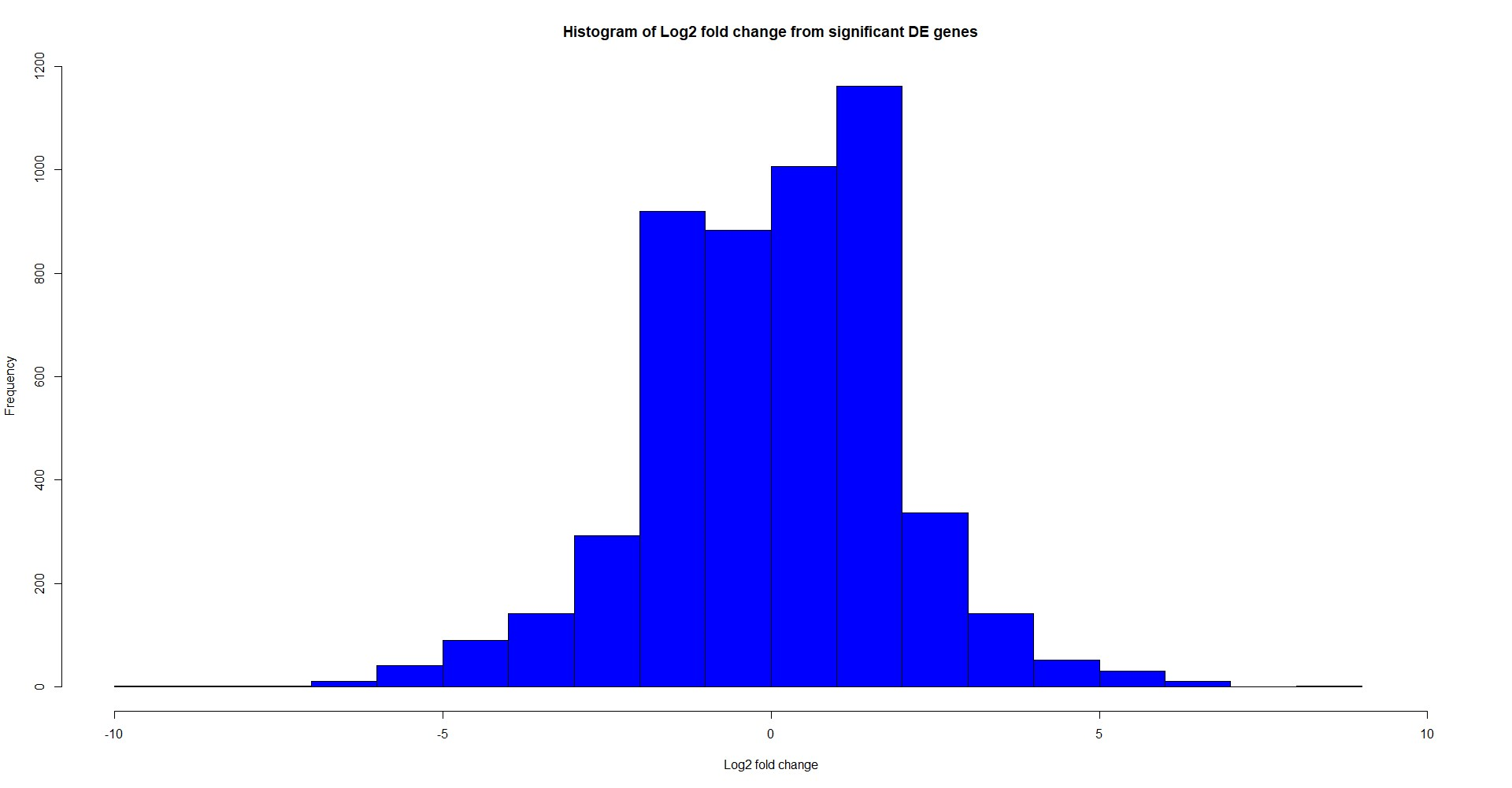
**Table.2 top 10 differentially expressed genes selected by label as ‘significant’ and their statistics (gene symbol, FPKM, adjusted FPKM, log2 fold change, p value and q value).**

The top 10 differentially expressed genes between P0 and Adult are shown as above (Table.2). The range of the adjusted FPKM is from 7.11 to 130.54. The range of the log2 fold change is from 0.991 to 1.517. The q values are all 0.000319 that are smaller than 0.05, which indicates their great significance.

**Histogram for log2 fold change**

****

**Fig.11 Histogram of log2 fold change for all genes (36329): X axe represents log2 fold change and Y axe represents the frequency of genes.**

****

**Fig.12 Histogram of log2 fold change for significant genes (5188): X axe represents log2 fold change and Y axe represents the frequency of genes.**

The histogram of log2 fold change for all genes and significant differentially expressed (DE) genes are shown above (Fig.11, Fig.12). For all genes, the distribution of the log2 fold changes are centered about 0. The number of negative log2 fold changes are obviously larger than the number of positive log2 fold changes, which indicates a higher number of down-regulated genes. Most of the genes have log2 fold changes between -2 and 2.

When it comes to the distribution for the significant DE genes, the log2 fold changes are also centered about 0. The number of positive log2 fold changes is slightly larger than the number of negative log2 fold changes, which indicates a greater number of up-regulated genes. The largest difference between the distribution for all genes and for only significant DE genes is that the log2 fold changes for significant DE genes have a larger standard deviation, which means that among the significant DE genes, there are a greater proportion of them with relatively extreme log2 fold change. From the two histograms, it could be noticed that the frequencies of the significant DE genes that have the log2 fold changes between -5 and -2 as well as between 2 and 5 is much higher and more obvious than those frequencies in the histogram for all genes.

**Number of significant genes using different selection methods**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Number of genes** | **Number of up-regulated genes (Ad)** | **Number of down-regulated genes (P0)** |
| Label as ‘significant’ | 5188 | 2757 | 2431 |
| p value < 0.01 | 4686 | 2532 | 2154 |
| q value < 0.01 | 3788 | 2101 | 1687 |

**Table.3 Number of significant genes using different selection methods: label, p value < 0.01 and q value < 0.01. The significant genes were then divided into up-regulated genes and down regulated genes.**

The number of significant DE genes as well as the number of up-regulated genes and down-regulated genes are shown above (Table.3). Among the 36329 genes, there are approximately 3800~5200 genes are significant. The q value selection is the most stringent, which results in 3788 significant genes, while label selection gives out 5188 significant genes. The range of the number of up-regulated genes is about 2100~2700 and the range of the number of down-regulated genes is about 1600~2400. Conclusively, the number of up-regulated genes is larger than the number of down-regulated genes, which means that there are more significant genes highly expressed in Ad sample than P0 sample.

**Top selected cluster results for up-regulated and down-regulated genes from the DAVID analysis**

|  |  |  |
| --- | --- | --- |
| **Summary of the cluster** | **Enrichment scores** | **Go Terms** |
| Mitochondrion, membrane and envelope | 54.76 | e.g. Mitochondrion, mitochondrial part, mitochondrial inner membrane |
| Nucleotide/ nucleoside metabolic process | 24.51 | e.g. purine nucleoside metabolic process,  organophosphate metabolic process,  nucleotide metabolic process |
| respiratory chain, dehydrogenase activity | 23.44 | e.g. mitochondrial respiratory chain,  respiratory chain complex I,  NADH dehydrogenase activity |
| Acid /lipid metabolic process | 22.98 | e.g. organic acid metabolic process,  carboxylic acid metabolic process,  cellular lipid metabolic process |
| Extracellular | 14.98 | e.g. extracellular organelle,  extracellular vesicle,  extracellular region part |
| Sarcomere | 11.22 | e.g. sarcomere,  myofibril,  contractile fiber |
| Glycolysis | 7.78 | e.g. glycolytic process,  carbohydrate catabolic process,  regulation of glycolytic process |
| Sarcoplasm | 5.40 | e.g. sarcoplasm,  sarcoplasmic reticulum,  sarcoplasmic reticulum membrane |

**Table.4 Selected top gene sets clusters for up-regulated genes. Enrichment scores and example GO terms for each cluster were also shown**

|  |  |  |
| --- | --- | --- |
| **Summary of the cluster** | **Enrichment scores** | **Go Terms** |
| Cell cycle and chromosome segregation | 29.48 | e.g. cell cycle process,  mitotic cell cycle,  nuclear chromosome segregation |
| Chromosome | 20.94 | e.g. chromosomal part,  nuclear chromosome,  chromatin |
| Regulation of metabolic/ biosynthetic process,  RNA processing | 20.82 | e.g. regulation of nitrogen compound metabolic process,  regulation of RNA metabolic process,  regulation of RNA biosynthetic process |
| Regulation of cell cycle | 20.1 | e.g. cell cycle checkpoint,  regulation of cell cycle phase transition,  regulation of mitotic cell cycle |
| Chromosome/Chromatin organization/ modification | 17.7 | e.g. chromosome organization,  covalent chromatin modification,  chromatin organization |

**Table.5 Selected top gene sets clusters for down-regulated genes. Enrichment scores and example GO terms for each cluster were also shown.**

Some selected top gene sets clusters for up-regulated genes and down-regulated genes are shown above respectively (Table.4, Table.5). For up-regulated genes, some interesting and meaningful gene set clusters include Mitochondrion, Metabolic process, Respiratory activity, Glycolic process, Sarcomere and Sarcoplasm. The highest enrichment score is 54.76 for Mitochondria cluster. The enrichment score for Sarcoplasm is relatively lower (5.40) compared with other clusters, but the GO terms, such as sarcoplasmic reticulum and sarcoplasmic reticulum membrane that are included in this cluster are still significant.

For down-regulated genes, some interesting and meaningful gene set clusters include Cell cycle, Chromosome/Chromatin, Regulation of process and RNA processing. The highest enrichment score is 29.48 for cell cycle and chromosome segregation cluster. Many of the GO terms included in these gene set clusters, such as chromosomal part, nuclear chromosome, chromatin and chromosome organization, could also be summarized as terms ‘Non-membrane bound organelle’ and ‘Nuclear Lumen’ shown in the results part of original paper. Non-membrane bound organelle [4] is defined as an organized structure of distinctive morphology and function, not bounded by a lipid bilayer membrane, while Nuclear Lumen [5] is defined as the volume enclosed by the nuclear inner membrane.

**Discussion**

The top 10 differentially expressed genes between P0 and Adult were selected out and shown as tables including their gene symbol, FPKM, adjusted FPKM, log2 fold change, p value and q value. The histogram of log2 fold change for all genes and significant differentially expressed genes were also shown and compared. The log2 fold changes for significant DE genes have a larger standard deviation, which means that among the significant DE genes, there are a larger proportion of them with relatively extreme or log2 fold change. It could be reasoned that the significant DE genes tended to have greater absolute log2 fold change. The frequencies of positive log2 fold changes of significant DE genes were slightly higher than the frequencies of negative log2 fold changes, which indicates the higher number of up-regulated genes than the down-regulated genes.

We applied three standards to select the significantly differentially expressed genes: label as ‘significant’, p value < 0.05 and q value < 0.05. The label selection provided the largest number of significant genes, while q value selection was the most stringent. The number of up-regulated genes is larger than down regulated genes using all the three methods. It can be concluded that the number of genes that are highly expressed in Adult group is greater than those in the P0 group.

However, in the original paper, the author detected about 10000 significant DE genes between P0 and Adult groups that is twice as many as we have. Among them, the number of down-regulated genes were greater than up-regulated genes, which is also different from what we observed from our data. One possible reason for the unmatched significant gene number might be that the author used other selection methods instead of label, p value or q value. Moreover, another possible reason is that they applied other thresholds such as 0.05 or 0.1 for q values. The larger number of down-regulated genes in the original paper could be explained by this greater threshold, since there were a greater number of down-regulated genes (Fig.1) among all genes and the higher threshold would include more down-regulated genes and resulted in their higher proportion among the significant DE genes. The unmatched number of significant genes is not a big concern because we still obtained similar results except the term ‘DNA repair’ using these genes in the enrichment analysis.

Some selected top gene sets clusters for up-regulated genes (Table.4) including Mitochondria, Metabolic process, Respiratory activity, Glycolic process, Sarcomere and Sarcoplasm and down-regulated genes (Table.5) including Cell cycle, Chromosome/Chromatin, Regulation of process and RNA processing were reported in tables. When comparing with the results in the paper, we successfully detected the enrichment terms Mitochondria, Sarcomere, Sarcoplasm, Respiration/Metabolism and Glycolysis for up-regulated genes. The enrichment scores for these terms in our results were higher than those in the paper. For down-regulated genes enrichment terms, the Non-membrane bound organelle, Nuclear Lumen, RNA processing and Cell cycle are successfully detected. Many GO terms shown in Table.5 such as chromosome, chromatin, nuclear chromosome and chromosome organization were all related to Non-membrane bound organelle and Nuclear lumen. The RNA processing and Cell cycle terms were also summarized as in Table.5. Nevertheless, we failed to detect the DNA repair terms for down-regulated genes as the paper suggested. One possible reason is that the number of the significant genes that we selected is not as many as the authors selected. The number of the down-regulated genes we obtained is much less than the number in paper. Therefore, we are unable to detect the DNA repair term due to the lack of some specific genes. Another possible reason is that different enrichment analysis tools were used by the authors.

According to the original paper, from the beginning of life to the adulthood, mammalian cardiac myocytes would undergo a maturation process that is characterized by development of a rigid and organized sarcomeric structure, binucleation, increased metabolic demand, and exit from the cell cycle [6]. The GO terms for up-regulated genes we obtained corresponded to this truth. For the terms for down-regulated genes, the RNA processing is necessary during the process of cardiac myocyte differentiation both in vitro and in vivo. Cell cycle exit is a hallmark of mature cardiac myocytes and a failure to re-enter the cell cycle is thought to contribute to the lack of cardiac regeneration in adult mammals [1]. We could expect to see that the P0 group would have higher expression of genes in cell cycle terms than the Adult group.

**Conclusion**

In this project, the top differentially expressed genes between P0 and Adult groups were selected out and the distribution of the log2 fold changes for all genes and significant DE genes had been plotted and compared. The significant DE genes tended to have larger absolute log2 fold changes. The number of the significant DE genes had been selected using label, p value and q value and compared with each other. The q value was the most stringent method and results in the least number of significant genes, as well as the number of up-regulated genes and down-regulated genes. The enrichment terms for significant up-regulated genes and down-regulated genes were summarized in 2 tables. All the terms in the original paper had been detected in our results except the ‘DNA repair’ term.

Another problem remained unknown is that the number of significant down-regulated genes is greater than the number of up-regulated genes in the original paper, while it is converse according to our result. We reason that the threshold used to select the significant DE genes may play a role, but it requires further verification. The failure of the detection of DNA repair term may also be explained by the different use of the threshold, which could result in different numbers of down-regulated genes.

Additionally, another challenge is to relate some of the enrichment terms in paper to the GO term clusters we detected. Some terms like Sarcomere and Sarcoplasm for up-regulated genes and Non-membrane bound organelle and Nuclear lumen for down-regulated genes require a strong background in biology to understand. We have to do more search and make efforts to find out the terms that are relevant to them.

**References**

[1] O'Meara, Caitlin C et al. “Transcriptional reversion of cardiac myocyte fate during mammalian cardiac regeneration.” Circulation research vol. 116,5 (2015): 804-15. doi:10.1161/CIRCRESAHA.116.304269

[2]http://ccb.jhu.edu/software/tophat/igenomes.shtml

[3] Nature Protocols 2009; 4(1):44 & Nucleic Acids Res. 2009;37(1):1.

[4] http://www.informatics.jax.org/vocab/gene\_ontology/GO:0043228.

[5] http://www.informatics.jax.org/vocab/gene\_ontology/GO:0031981.

[6] Soonpaa MH, Kim KK, Pajak L, Franklin M, Field LJ. Cardiac myocyte DNA synthesis and binucleation during murine development. Am J Physiol. 1996;271:H2183–H2189.