**BF528 Individual Project**

# TRANSCRIPTIONAL PROFILING OF MAMMALIAN CARDIAC REGENERATION WITH mRNA-Seq

(By O'Meara et. al)

## Introduction

I chose to perform the programmer and analyst steps from Project 2, "Bioinformatics Reanalysis of: Transcriptional Profile of Mammalian Cardiac Regeneration with mRNA- Seq." As the project's original biologist, I was interested in comparing the original outputs I worked on for the project to the work I could perform as a programmer and analyst.

The paper aimed to discover why adult mouse hearts fail to regenerate fully after injury such as a heart attack, as opposed to neonatal hearts, which can fully repair, and which transcriptional factors are responsible for regulating heart cell regeneration and repair. Repeating analyses is good scientific practice, not only for validating experimental results, but also as a hands-on approach to analyzing project processes. So, I decided to compare my results to those of my teammates and to the final results attained in the original study, O'Meara et al., by repeating the programmer and analyst procedures. The initial researchers wanted to evaluate the transcriptional state of neonate and adult mouse heart tissue because neonatal and pregnant mammal hearts are known to have specific regeneration pathways following injury that growing and adult animals do not have. Both this and our previous analysis replicated a comparison of cardiac myocyte tissue from an adult (Ad) and a zero-day postnatal mouse (P0).

## Methods

I used the files generated from our data curator. The files are under the repository (/projectnb/bf528/users/frizzled/project_2/data/)

**Programmer:** The two FASTQ data were aligned to the reference mouse genome, mm9. I performed the alignment using TopHat version 2.1.1 and 16 cores on the SCC. TopHat is a program that demands a significant amount of memory to execute. As a result, rather than executing TopHat interactively, it is frequently essential to run it as a batch job on the cluster. When you've finished writing your tophat command into the qsub file, execute the TopHat job using qsub run tophat.qsub and wait for the alignment to finish.

I then used RSeQC 3.0.0 to do quality control analysis on the TopHat alignment's acceptable hits, including gene coverage and statistics on the produced BAM file. Cufflinks 2.2.1 was then utilized with 16 processing cores to map the aligned reads to Mus musculus genes of areas. The flagstat command in Samtools version 0.1.19 was used to perform some quality analysis. I loaded the Cufflinks result into R version 4.0.2 and plotted the quantified gene fragments per kilobase per million mapped fragments (FPKM). Finally, we utilized cuffdiff from the same Cufflinks version to find genes that were differentially expressed between P0 and Ad, the cardiac myocytes from adult Mus musculus.

| | | |
|---|---|---|
| **Number of total reads** | 49706999 | 100 % |
| **Number of mapped reads** | 49706999 | 100 % |
| **Number of unique mapped reads** | 41389334 | 83.27 % |
| **Number of multi-mapped reads** | 8317665 | 16.73 % |
| **Number of unaligned reads** | 0 | 0 % |

**Table 1:** Report of the total number of reads, number of mapped, unique, multi-mapped, and unaligned reads with percentages of total reads for each

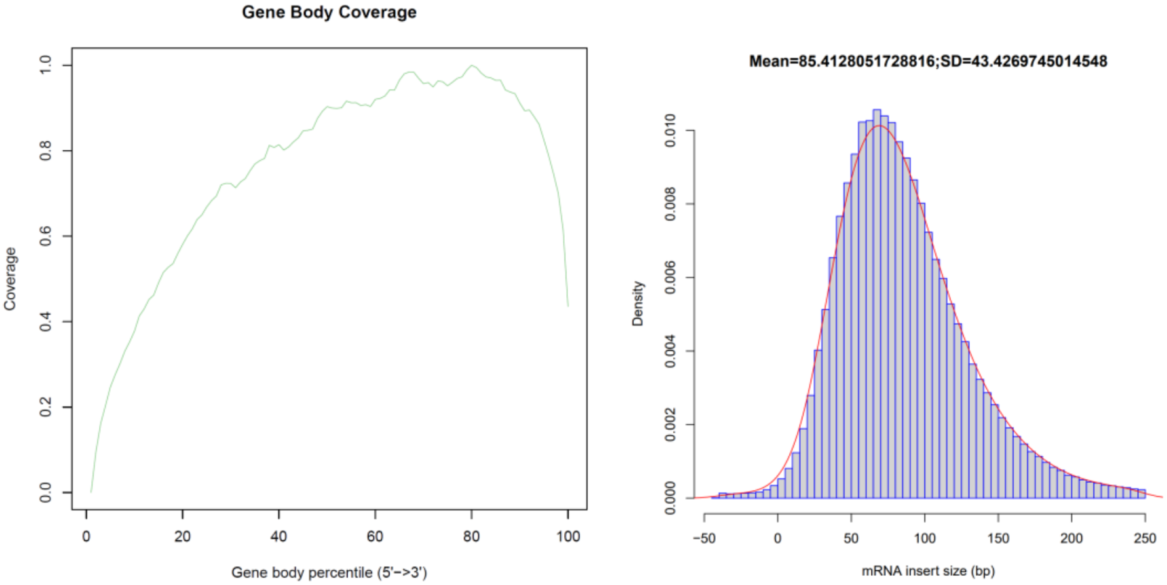


**Figure 1:** RseQC results: Gene Body Coverage and Inner Distance.

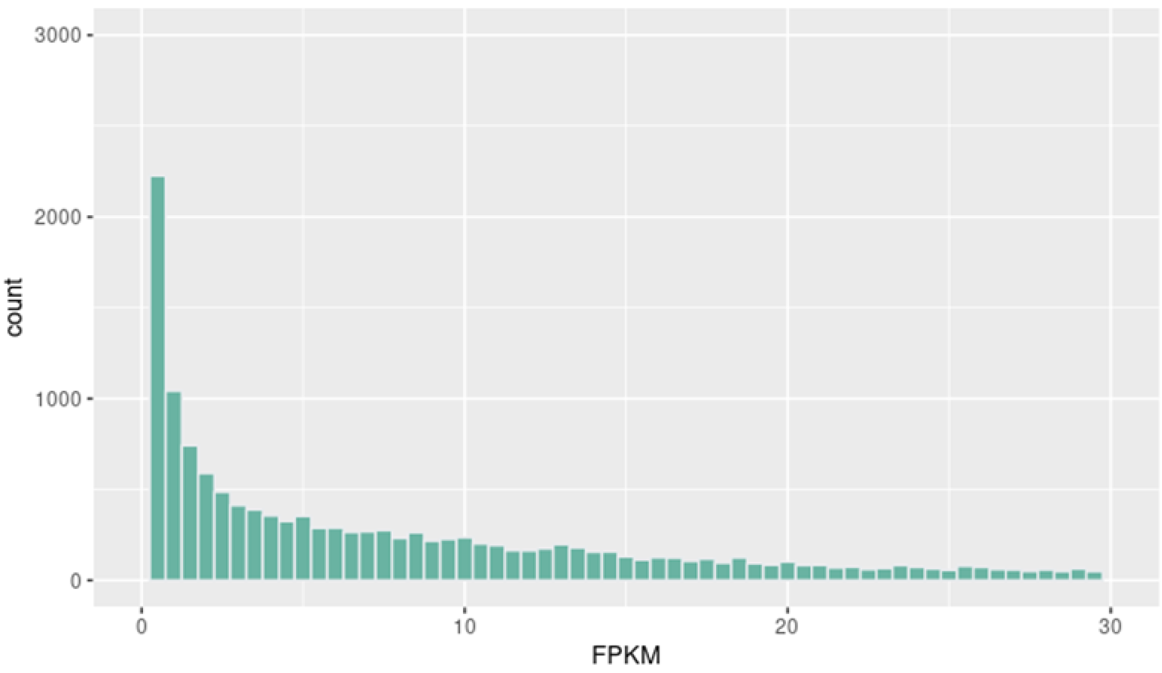


**Figure 2:** Histogram of FPKM values for all genes

Finally, I reproduced the data analyst technique of comparing and functionally annotating the cuffdiff-identified differentially expressed genes. Using the same R version as before, I identified the top ten differentially expressed genes based on the lowest q-value and showed the distribution of all and significant differentially expressed genes based on log2 fold change. I then downloaded the gene names as down and up-regulated genes after filtering the negatively and positively expressed significant genes to exclude smaller log2 fold change values. The 500 or so genes in each list were then imported into DAVID independently for functional annotation.

While it was not within the scope of this particular project, I did replace some of the original biologist role's inputs with my freshly generated data, and I plotted some of the heat maps once again to compare findings visually. The original GitHub https://github.com/BF528/individual-project-Shrishtee-kandoi contains the details for creating these heat maps.

**Results**

I utilized some quality control scripts to check how well the mapping performed after aligning the acceptable hits to the reference mouse genome. There were no failed reads, and more than 80% of the reads mapped to the referred genome. RSeQC plots were utilized to measure coverage and fragment insert sizes to further quantify the alignment, with no notable divergence from the first instance of this project.

After finishing the Cufflinks run for P0, I plotted the distribution of non-zero FPKM numbers. The distribution is very skewed, with the vast majority of FPKM values falling below 100 fragments. Of the 37,469 total genes discovered, 16,453 had non-zero FPKM values (43.9 percent). I used cuffdiff to identify the ten most differentially expressed genes between P0 and Ad, ranked by q-value.

After visualizing the log2 fold change for all and then significant genes, I filtered the gene set by removing those that were not significant (p 0.01). Those genes that were not discovered to be significant were deleted in order to choose the most impacting genes. There were 2,139 significant genes discovered, with 1,084 being up-regulated and 1,055 being down-regulated.

In order to find a more differentially expressed subset of genes, I filtered the up- and down-regulated genes to those with a log2 fold change greater than 2 and less than -2, respectively. There are 378 up-regulated genes and 640 down-regulated, important genes in the final set I utilized for functional annotation (of the original 36,329 total). I discovered 194 clusters in the up-regulated gene set and 270 clusters in the down-regulated gene set after functionally annotating these gene sets with DAVID.

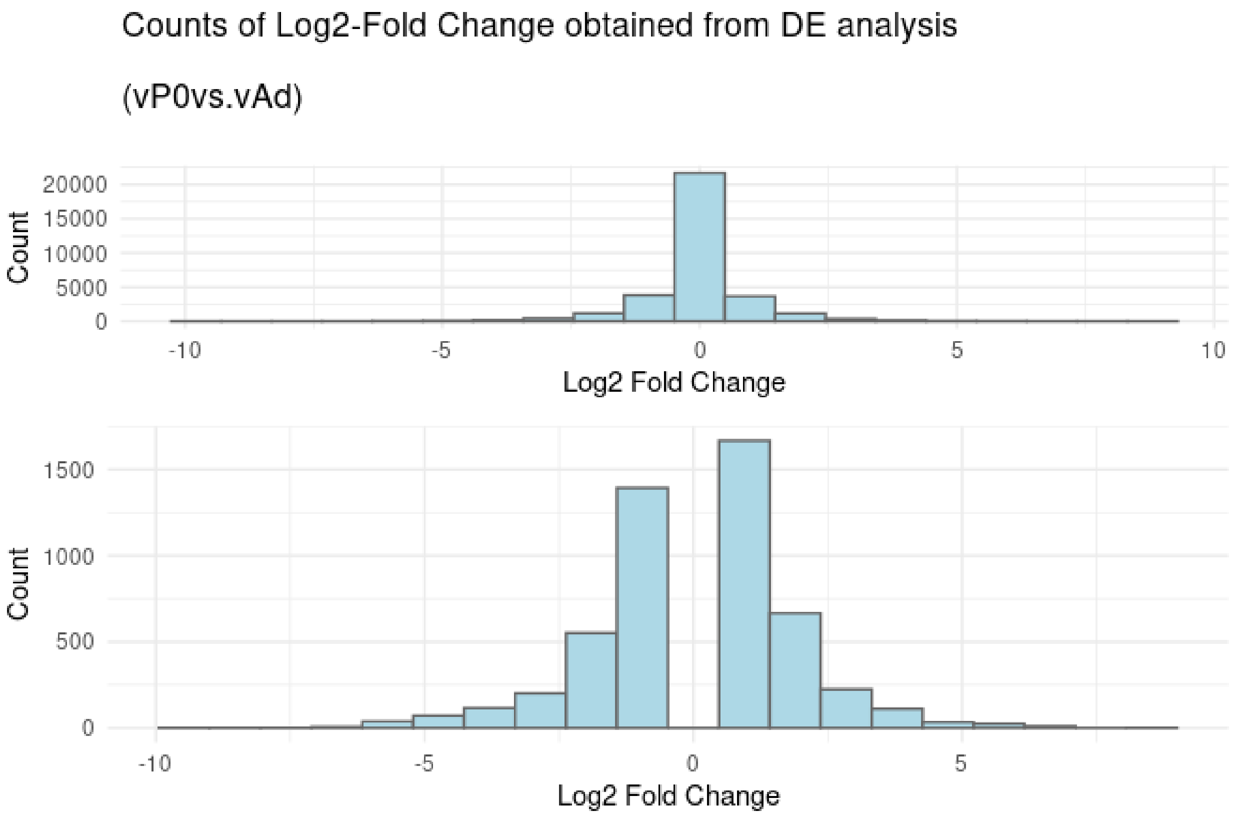
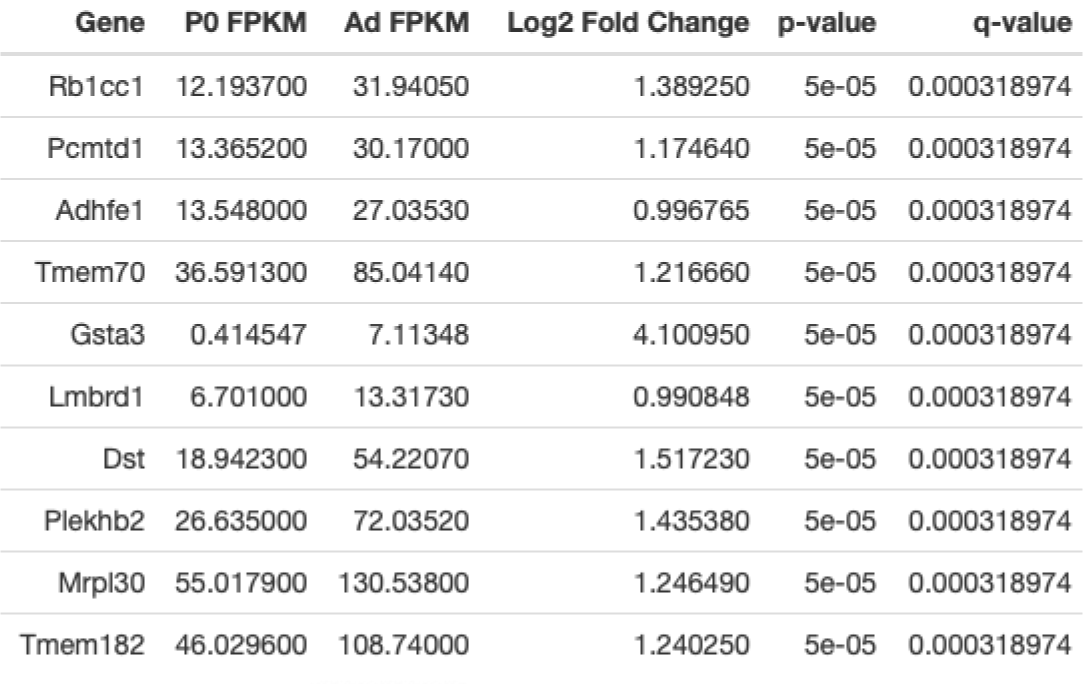


**Figure 3:** Distribution of log2 fold change values. Note the log10-transformed y-axes. A) The number of genes for each log2 fold change level. B) The same set of genes, with those not meeting the significance p-value (0.01) removed.



| Gene | P0 FPKM | Ad FPKM | Log2 Fold Change | p-value | q-value |
|---:|---:|---:|---:|---:|---:|
| Rb1cc1 | 12.193700 | 31.94050 | 1.389250 | 5e-05 | 0.000318974 |
| Pcmtd1 | 13.365200 | 30.17000 | 1.174640 | 5e-05 | 0.000318974 |
| Adhfe1 | 13.548000 | 27.03530 | 0.996765 | 5e-05 | 0.000318974 |
| Tmem70 | 36.591300 | 85.04140 | 1.216660 | 5e-05 | 0.000318974 |
| Gsta3 | 0.414547 | 7.11348 | 4.100950 | 5e-05 | 0.000318974 |
| Lmbrd1 | 6.701000 | 13.31730 | 0.990848 | 5e-05 | 0.000318974 |
| Dst | 18.942300 | 54.22070 | 1.517230 | 5e-05 | 0.000318974 |
| Plekhb2 | 26.635000 | 72.03520 | 1.435380 | 5e-05 | 0.000318974 |
| Mrpl30 | 55.017900 | 130.53800 | 1.246490 | 5e-05 | 0.000318974 |
| Tmem182 | 46.029600 | 108.74000 | 1.240250 | 5e-05 | 0.000318974 |

**Table 2:** Top Differentially expressed genes ordered by ascending *q*-values for postnatal day 0 (P0) versus Adult (Ad) mice.

**Discussion**

In the Tophat part, I used flagstat and RseQC bamstat to have quality control on Tophat results. Obviously, the QC results are not precisely identical. The reason might be the setting of mapq_cut, which is the threshold of mapping quality scores. For RseQC, the default mapq_cut is 30 while it is 5 for flagstat. Nevertheless, the difference is acceptable, and both results show a perfect quality of the sequences. Meanwhile, the inner distance is an ideal Gaussian Distribution indicating the sequencing process is fast and efficient. In most cases, inner distance is in multivariate Gaussian Distribution because of the batch effect. On the other hand, the batch effect is under control in the data.

Differentially expressed gene sets from postnatal day 0 (P0) versus Adult (vAd) mice were analyzed in an attempt to partially replicate the findings in O'Meara et al. 1B and 1C. Table A displays that the data provided in the paper lead to valid differential expression results.

**Conclusion**

This study provides a critical framework for understanding the transcriptional expression changes required for cardiac myocyte repair in response to injury that will be invaluable for ultimately guiding efforts to promote adult mammalian cardiac regeneration. The analysis done in this project was done correctly and it replicated the findings of O'Meara et al. to great extent. However, the differences risen were likely due to the choice of tool, differences in versions of tools and parameters for analysis.

## References

- Caitlin C. O'Meara et al. "Transcriptional Reversion of Cardiac Myocyte Fate during Mammalian Cardiac Regeneration". eng. In: Circulation Research 116.5 (Feb. 2015), pp. 804–815. issn: 1524-4571. doi: 10.1161/CIRCRESAHA.116.304269.
- Cole Trapnell et al. "Transcript Assembly and Quantification by RNA-Seq Reveals Unannotated Tran- scripts and Isoform Switching during Cell Differentiation". eng. In: Nature Biotechnology 28.5 (May 2010), pp. 511–515. issn: 1546-1696. doi: 10.1038/nbt.1621.