Transcriptional Profile of Mammalian Cardiac Regeneration with mRNA-Seq

Yueh-Ting Wang, Cody Webb, and Chris Lin

Introduction

In O'meara et al. [1], the authors investigated the ability of neonatal mice to regenerate cardiac tissue. However, this potential is lost after the first week of life. One key step in further understanding this phenomenon was the analysis of RNAseq data in adult and early postnatal mice to investigate the differences in transcriptional profiles through analysis of enriched gene ontology (GO) terms. In this project, we sought to reproduce their results using similar methods as in O'meara et al. First, the publicly available dataset from the authors was acquired, followed by quality control and alignment to the reference mouse genome. The differentially expressed genes were then extracted and investigated through GO term functional analysis in an effort to determine if our enrichment terms were similar to those of O'meara et al.

Data

The raw data was collected from the whole heart ventricle cells of (CD-1) neonatal mice at postnatal day 0, 4, and 7 (P0, P4, and P7) and from 8–10-week-old male CD-1 mice. The cells were washed in ice-cold PBS, and snap-frozen in liquid nitrogen. The atria parts of the heart were dissected and discarded, and ventricles were processed for RNAseq. At least two heart ventricles were pooled for each replicate and then the data were processed for RNAseq. The raw data of explanted adult mouse cardiac myocytes were collected after 0, 24, 48, and 72 hours in culture and then processed under the same protocol described previously. There were no samples eliminated due to contamination. The data used in the O'meara et al. (2014) research can be found on the NCBI Gene Expression Omnibus with accession number GSE64403. (https://www.ncbi.nlm.nih.gov/geo/)

For this project, there was only one remaining sample needed to be downloaded and processed. The majority of the samples were previously curated and downloaded into the group's repository. The sample GSM1570702 (0-day postnatal ventricular myocardium) was manually downloaded from GEO Series GSE64403 of the O'meara et al. (2014) repository on the NCBI Gene Expression Omnibus. The RNAseq data was conducted using Illumina HiSeq 2000 (Mus musculus) and the reads are paired-end. The length of each read is 40 base pairs and the number of reads is 21577562.

Methods

In order to see which genes were being differentially expressed, the mRNA data needed to be aligned to the mouse genome. This would allow for the ability to quantify which genes were being differentially expressed in each scenario. The mRNA was aligned to the mm9 mouse genome using tophat[2] (-r 200 -G --segment-length 20 --segment-mismatches 1 --no-novel-juncs -p 16) on a shared computing cluster (SCC). This took around two hours to

complete. After alignment was performed, the samtools program flagstat [3] was used on the alignment bam file to determine the quality of the reads, and it found that all reads were within its quality control standards. As such, no reads were removed. In order to further confirm that our data was of sufficient quality, the samtools utilities in RseQC [4] were used. This showed no abnormalities in either gene body coverage or in the paired read mRNA (Fig. 1). Since this was the case, it was determined to be appropriate to continue on with the differential expression analysis.

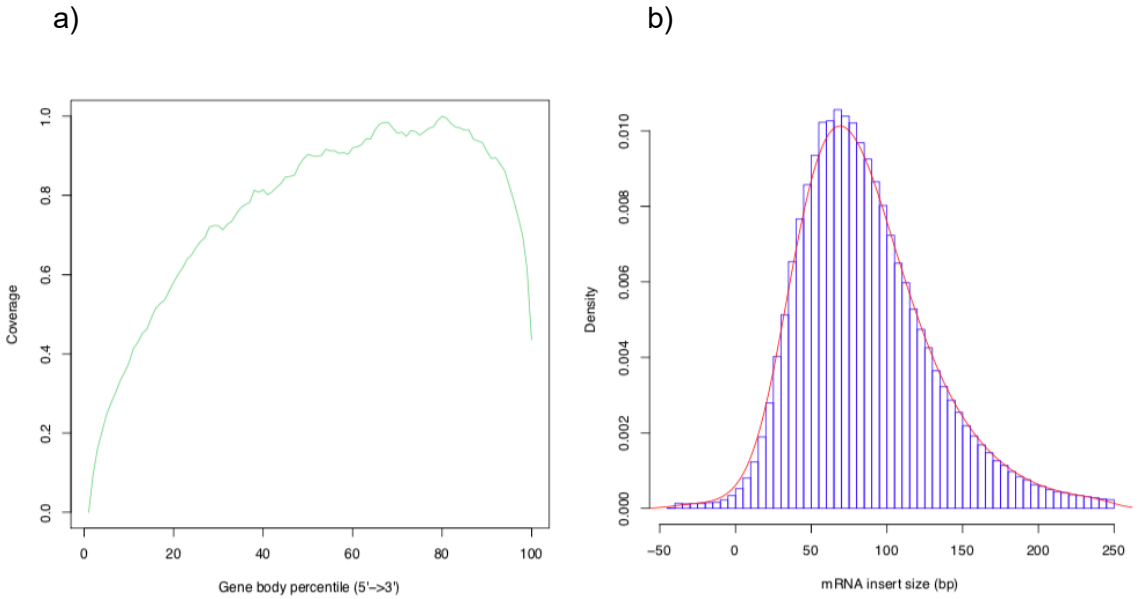


Figure 1. a) Gene body coverage information from RseQC. b) The inner distance of the mRNA segment pairs. Nearly all of the values were greater than zero, with an average distance of 85 base pairs and a standard deviation of 43 base pairs.

cufflinks [5] (--compatible-hits-norm -G -b -u -p 16) was then used to assemble the transcripts, determine the transcript abundance of our sample, and determine which genes were differentially expressed. mRNA levels were determined using the measurement of fragments per kilobase per million fragments mapped (FPKM). Then, using the cufflinks utility cuffdiff (-p 16 -u -b), it was determined which genes were differentially expressed. Both of these steps took place on the SCC and took between one and three hours to complete.

Differential expression analysis was performed using genes and statistics from the cufflinks pipeline using R version 3.5.1. The list of differentially expressed genes was first sorted by q-value, then subsetted to contain only genes that had p is greater than the FDR after Benjamini-Hochberg correction for multiple-testing [6]. Using our subsetted gene list, we further divided our data into two separate dataframes based on whether each gene was up-regulated or down-regulated as determined by the log-fold change. Each list of gene names was then exported as a .csv file, and analysis of enriched GO terms was performed on each individual list using Database for Annotation, Visualization and Integrated Discovery (DAVID) [7].

Results

The fast QC report of sample GSM1570702 showed the per-base sequence quality is acceptable, however, the plot of per base sequence content showed a non-uniform distribution of bases with %A not equal to %T and %G not equal to %C for the first 10-12 nucleotides(Fig. 2). For the plot of sequence duplication levels (Fig. 3), we also found that about 80% of the duplicated sequences are at the duplication level of ">10".

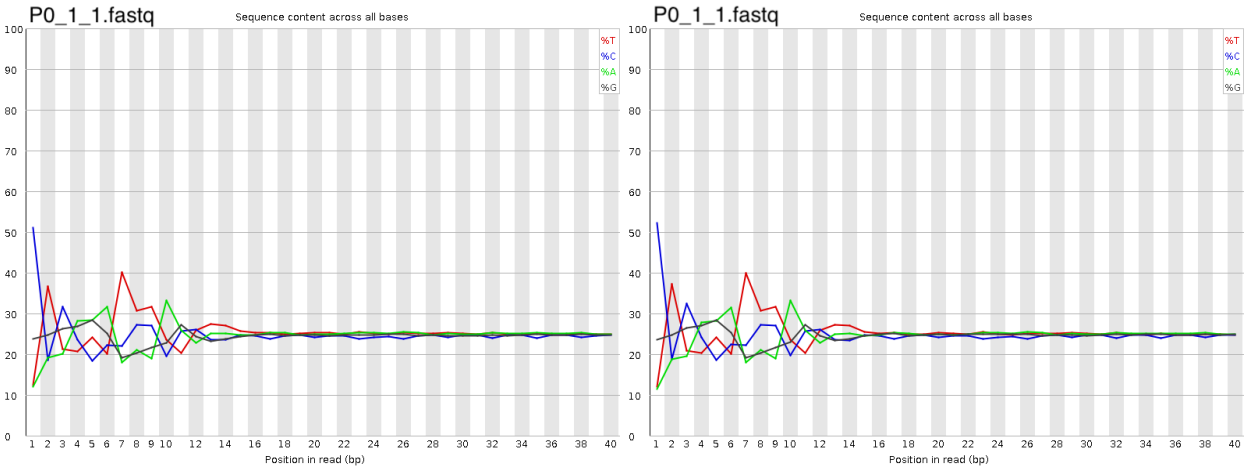


Figure 2. Per base sequence quality from FastQC output with 2 fastq files.

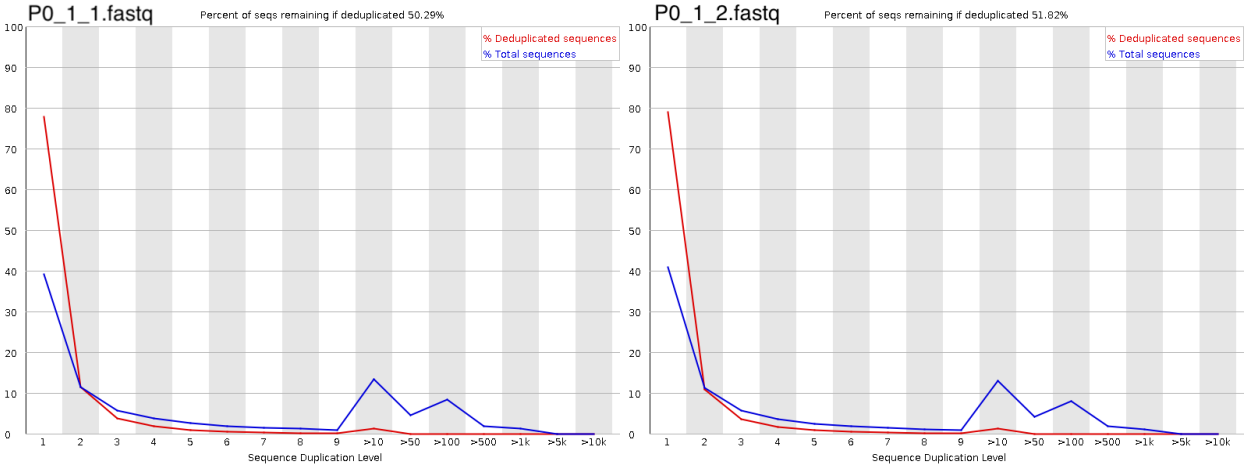


Figure 3. Sequence duplication levels from FastQC output with 2 fastq files.

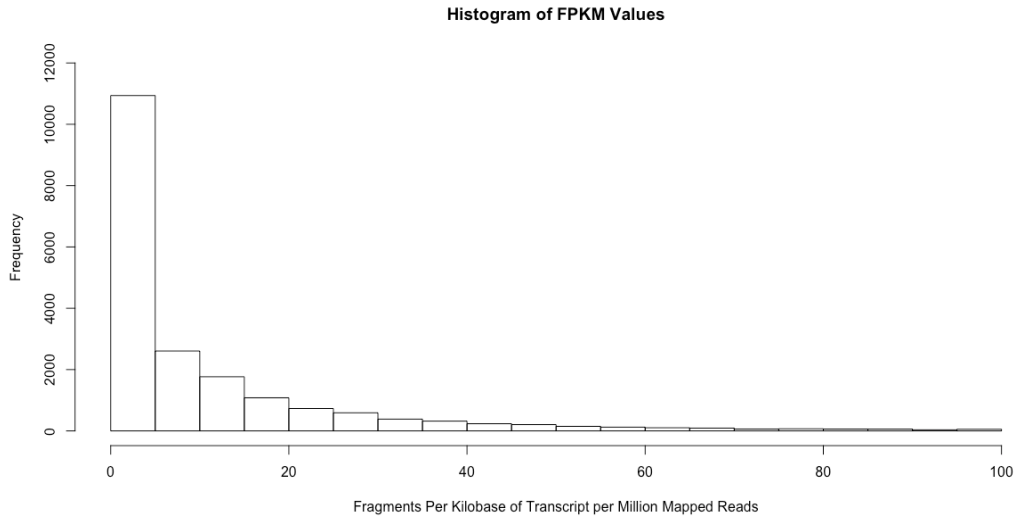


Figure 4. After removing all zero values from the dataset, a significant portion of the reads were only found less than four times. This histogram does not include the FPKM values over 100, which only consisted of 826 reads.

In order to reproduce the findings of O'meara et al., a list of differentially expressed genes between P0 and adult mice, including statistics, was obtained from the programmer. This gene list was then sorted by q-value, identifying the top ten differentially expressed genes (Fig. 5A). The majority of our gene list had a log-fold change between -0.4 and 0.4, with many of these values being non-significant (Fig. 5B), thus for our analysis we only took into account genes that had p greater than the FDR after Benjamini-Hochberg correction for multiple-testing (Fig. 5B). After filtering for these significant genes, there were in total 2,757 genes that were upregulated and 2,431 genes that were downregulated, with a total of 5,188 significant genes (Fig. 5A). Compared to O'meara et al., we had a larger number of total up-regulated genes, and a smaller number of down-regulated genes when comparing the Adult v P0 datasets.

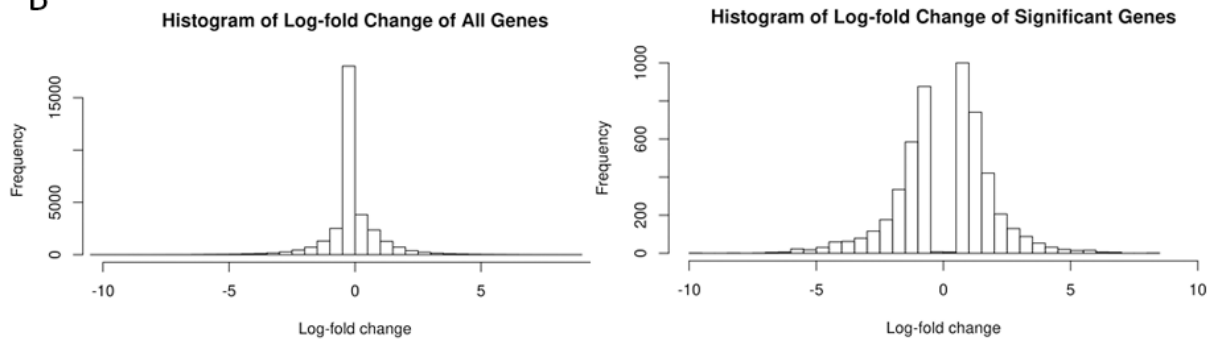
Interpretation of these differentially expressed genes was then determined by analysis of GO terms [7]. Differentially expressed genes showed primarily upregulation of mitochondrial, metabolic process, and respiratory related genes. Overall, compared to O'meara et al., many of our clustered enrichment terms showed overlap, such as with mitochondria, sarcomere, and metabolism. Although our reproduction also had mitochondria as the enrichment term with the highest score, other terms did not correspond to the exact order of enrichment terms in O'meara et al. Differentially expressed genes showed primarily downregulation of cell cycle, chromosomal, and transcription regulation genes. These downregulated enrichment clusters showed less overlap with that of O'meara et al., with only cell cycle genes being included in both our reproduction and the original author's work.

A

Gene Name	FPKM	Log-fold change	P-value	Q_value
Rb1cc1	12.193700	1.389250	5e-05	0.000318974
Pcmtd1	13.365200	1.174640	5e-05	0.000318974
Adhfe1	13.548000	0.996765	5e-05	0.000318974
Tmem70	36.591300	1.216660	5e-05	0.000318974
Gsta3	0.414547	4.100950	5e-05	0.000318974
Lmbrd1	6.701000	0.990848	5e-05	0.000318974
Dst	18.942300	1.517230	5e-05	0.000318974
Plekhb2	26.635000	1.435380	5e-05	0.000318974
Mrpl30	55.017900	1.246490	5e-05	0.000318974
Tmem182	46.029600	1.240250	5e-05	0.000318974

Upregulated Genes	Downregulated Genes	Total Genes
2757	2431	5188

B



C

Common up and Downregulated Gene Enrichment Terms

Up-regulated

Enrichment Term	Score
Mitochodrion	54.67
Metabolic process	24.48
Respiration	23.42
Extracellular	14.92
Myofibril/Sarcomere	11.2

Down-regulated

Enrichment Term	Score
Cell Cycle	21.96
Chromosome	21.06
Metabolic/Biosynthetic process	20.57
Transcription Regulation	15.45
Microtubule/Centrisome	13.15

Figure 5. Differentially Expressed Genes Associated with Myocyte Differentiation

A) Table of top ten differentially expressed genes from P0 vs Adult mice as defined by descending Q-value; total number of upregulated and downregulated genes that were considered significant after filtering by log-fold change.

B) Distribution of the log-fold change of all differentially expressed genes between P0 and Adult mice; distribution of the log-fold change of only the differentially expressed genes that were significant in regards to P-value.

C) Table of top gene enrichment terms for common up- or down-regulated genes during differentiation.

Discussion

The fast QC report of sample GSM1570702 failed to pass the “Per base sequence content” (Fig. 2) in both 2 fastq files. There is about a 20% difference between A & T at the 7th position of the read and a >20% difference between G & C at the beginning position. This might be due to the type of library kit used or other issues. The report also showed warnings for the “Sequence duplication levels” in both 2 fastq files. According to (Fig. 3), about 80% of the duplicated sequences are at the duplication level of ">10" and there is a fluctuation of duplication levels across the reads. This may have resulted from the PCR duplication in which library fragments have been over-represented due to biased PCR enrichment or truly over-represented sequences such as very abundant transcripts in an RNA-Seq library.

Using data from O’meara et al., we were able to interpret differential expression profiles between neonatal and adult mice using DAVID. Overall, the enrichment terms that we observed in our reproduction were similar to those of the original authors. For upregulated genes, enrichment terms associated or important to cardiac muscle division [8], such as mitochondrion, metabolic processes, and myofibril/sarcomere were seen in both our analysis and that of the original authors. For downregulated genes, in both our reproduction and the work of the authors, cell division was an overlapping enrichment term. As expected, the difference in cardiac muscle regeneration between adult and neonatal mice are most likely related to these down-regulatory gene changes in cell division [9]. Since the authors did not provide specific GO term IDs, it was difficult to confirm whether our results matched that of the authors exactly. Rather than providing general ‘enrichment terms’, such as ‘Metabolism’, the authors should have also included specific GO term IDs in addition to cluster enrichment score.

Additionally, the authors had access to a variety of both in-vivo and in-vitro samples to perform rnaSeq and GO enrichment analysis on, while we were limited to only two of the samples that they produced. This difference in samples is a possible explanation to why our reproduction differs with that of the original authors. However, even with only two samples, our reproduction of GO enrichment analysis still had strong agreement in regards to the up-regulated genes.

Furthermore, looking at the overall number of up- and down-regulated genes, there is a discrepancy in our reproduction and the original authors’ work. O’meara et al. uses 2,409 upregulated genes and 7,570 down-regulated genes for their GO analysis, while we used 2,757 and 2,431 genes respectively. It is possible that they used different quality assurance methods or different values for key filtering steps such as filtering by FPKM. Such a large difference in the number of down-regulated genes used in GO analysis may explain why our down-regulated enrichment terms show less agreement than our up-regulated terms.

Conclusion

Given the lack of concrete GO ID terms and a reduced dataset, we are unable to say with certainty that we were able to reproduce the complete results obtained by O'meara et al. However, from our individual analysis, we obtained many similar if not identical enrichment terms as that of the original authors. This seems to corroborate the findings of the original paper's claims of the distinct expression profiles between adult and neonatal mice, even with the restrictions imposed on our analysis. Given both specific GO IDs from O'meara et al.'s clustering results and the full dataset used, a more accurate reproduction would be possible.

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: Trapnell, C & Salzberg, S. (2016). TopHat (Version 2.1.1). https://ccb.jhu.edu/software/tophat/index.shtml

3: Li, H.. (2009). SAMtools (Version 0.1.19). http://www.htslib.org/

4: Wang, L. (2012). RseQC (Version 3.0.1). http://rseqc.sourceforge.net/

5: Pachter, L. (2009). Cufflinks (Version 2.2.1). http://cole-trapnell-lab.github.io/cufflinks/

6. Trapnell, C., Roberts, A., Goff, L. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7, 562–578 (2012).

7. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using david bioinformatics resources. Nat Protoc. 2009;4:44–57

8. Claycomb, W C. "Cardiac-muscle hypertrophy. Differentiation and growth of the heart cell during development." The Biochemical journal vol. 168,3 (1977): 599-601.

9. Hassan, Narmeen et al. "Concise review: skeletal muscle stem cells and cardiac lineage: potential for heart repair." Stem cells translational medicine vol. 3,2 (2014): 183-93.