Transcriptional Profile of Mammalian Cardiac Regeneration with mRNA-Seq

BF528 Project 2
Yu Zhong, Zhiyu Zhang, Jianfeng Ke, Huisiyu Yu

Introduction

The capacity of cardiac regeneration is retained through life for certain fish and amphibian species, but the same is not true for adult mammals. Previous study had demonstrated that the hearts of 1-day-old neonatal mice can regenerate after partial surgical resection, but the capacity was lost after a week of growth, where cardiac myocytes (CMs) exit the cell cycle[1]. In order to better understand the transcriptional changes that underlie mammalian cardiac regeneration on a molecular level, O'Meara et al examined global expression changes in the neonatal mouse after heart injury and analyzed the transcriptional signature of mouse CMs at different stages of postnatal development as well as mouse embryonic stem cells differentiated to CMs[2].

In this project we reimplemented the bioinformatics analysis pipeline of O'Meara et al's study, and reproduced their results in order to gain a better understanding of RNA-Seq analysis and the various bioinformatics tools implemented along the way. Our results revealed key transcriptional signatures during mammalian cardiac myocyte differentiation and was largely consistent with the original study.

Methods

Data Description & Quality Control

The original authors extracted total RNA using Trizol from 0 day mouse postnatal ventricular myocardium and performed paired-end sequencing with read length of 40 base pairs using Illumina HiSeq 2000. RNA-Seq sequence read archive (SRA) sample GSM1570702 was obtained from GEO (Gene Expression Omnibus). The SRA file was converted to FASTQ format using the NCBI SRA Toolkit[3]. FastQC[4] was used to inspect the quality of the sequencing reads.

Aligning & Quantifying gene expression

TopHat[5] is a fast splice junction mapper for RNA-Seq reads. It was used to align FASTQ files against mouse genome reference (mm9). The output of TopHat was a BAM format file, which is a binary version of the SAM (Sequence Alignment/Map) format, and contained all of the original read plus any alignments discovered by TopHat.

RSeQC[6] is an RNA-seq quality control package which provides a number of useful modules that can comprehensively evaluate high throughput sequence data. Three modules were used to explain and interpret the BAM output. The module "geneBody_coverage.py" calculated the

RNA-seq reads coverage over the gene body; “inner_distance.py” was used to calculate the inner distance between read pairs; “bam_stat.py” summarized the mapping statistics of the BAM file.
Cufflinks is a tool that counts how reads map to genomic regions defined by annotation. It was run on the BAM file generated by TopHat and created a file containing the quantified alignments in FPKM (Fragments Per Kilobase of transcript per Million mapped reads) for all genes. Cuffdiff is a tool in the cufflinks suite and it is used to identify differentially expressed genes.

Functional Annotation Clustering Analysis

By using Cuffdiff, genes under the threshold of 0.01 were identified as differentially expressed genes. DAVID were used to perform functional annotation clustering analysis.

Results

Quality Control

In order to assess the quality of the RNA sequencing data, we used FastQC to generate two html reports of quality metrics for the paired-end read (P0_1_1 and P0_1_2). The results were the same for read1 and read2. The report showed that overall sequence quality was high (Figure1), with lowest per base sequence quality score being 30. GC content was 49% and normally distributed. Percent of sequences remaining if deduplicated was 50.29%, indicating that there was considerable sequence duplication, which is expected for RNA-Seq experiments.

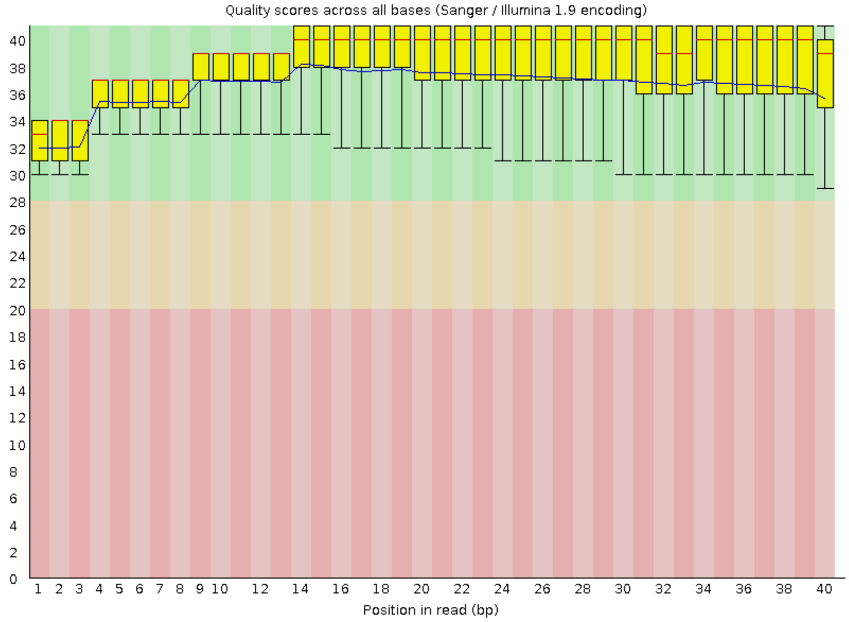


Figure 1. Per base sequence quality of P0 (Read1)
An overview of the range of quality values across all bases at each position in the FastQ file

The only metric that failed was per-base sequence content (Figure2), which indicated that sequence composition was biased at the first few bases of the reads.

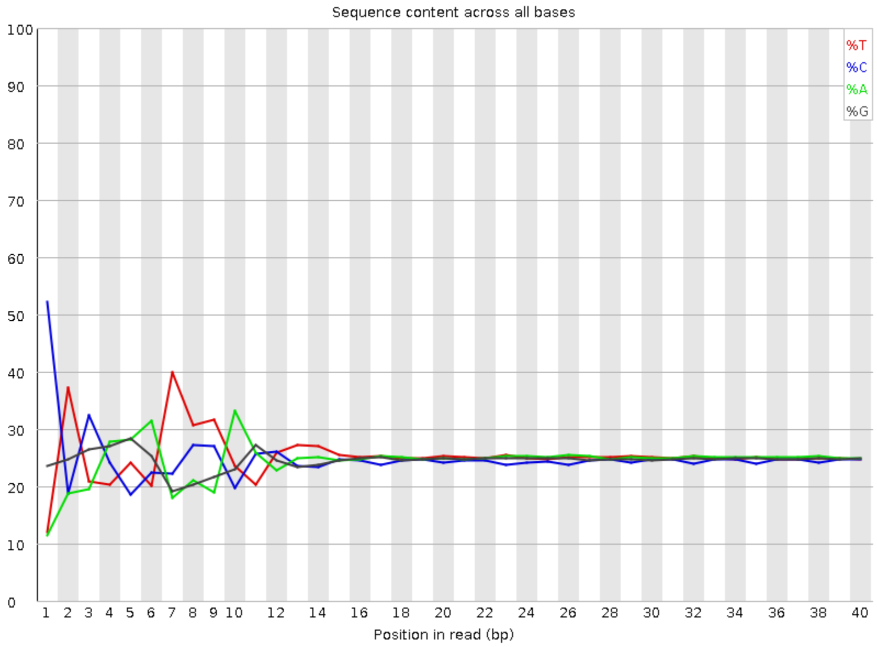


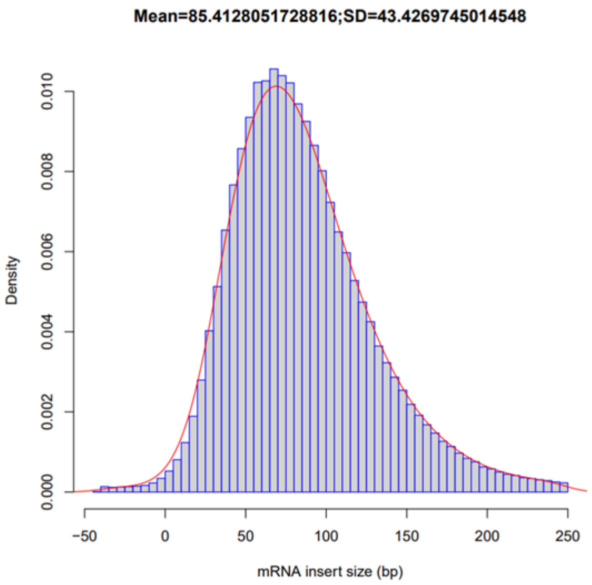
Figure 2 Per base sequence content of P0 (Read1)

Plots out the proportion of each base position in a file for which each of the four normal DNA bases has been called.

Aligning & Quantifying gene expression

We then performed sequence alignment and quantified gene expression using TopHat and RSeQC respectively._Only the P0_1 sample was considered. There were in total 49,706,999 reads, from which overall 95.9% reads map to the mm9 mouse reference genome. More specifically, 96.8% of forward reads mapped to the genome, from which 7% reads were multiple mappings, and 95.1% of reverse reads mapped to the genome, from which 7% reads are multiple mappings.

We can observe that the mean of fragment insert size from our RNA library was around 85 (Figure 3), together with high read coverage on the gene body (Figure 4), indicated that library preparation is robust enough for the downstream analysis.



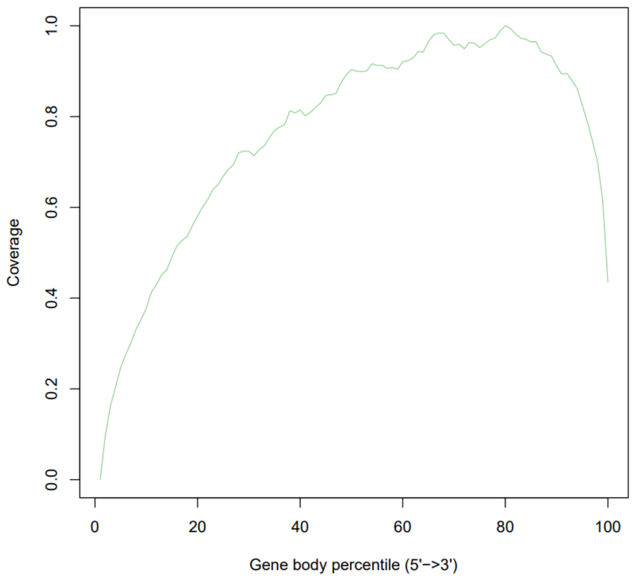


Figure 3. RNA-seq reads coverage over the gene body. Figure 4. Inner distance between read pairs.

Based on the quantification of gene expression, in total 16,453 genes remained after removing ones having FPKM of zero value. The distribution of expression level provides us insight into details into expression profiles (FIgure 5).

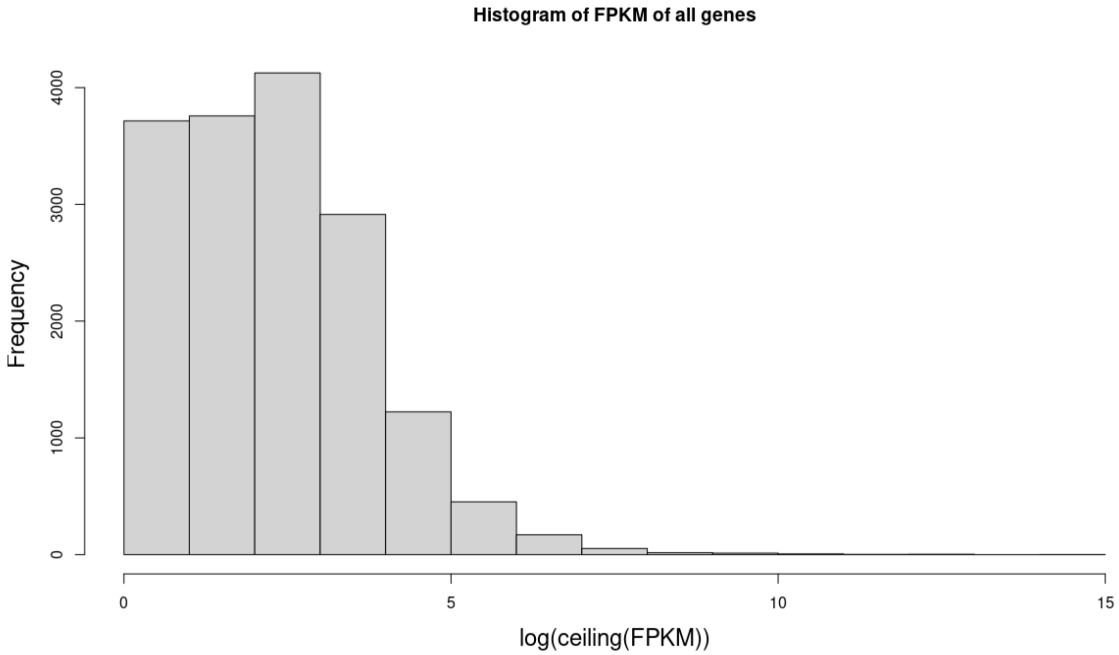
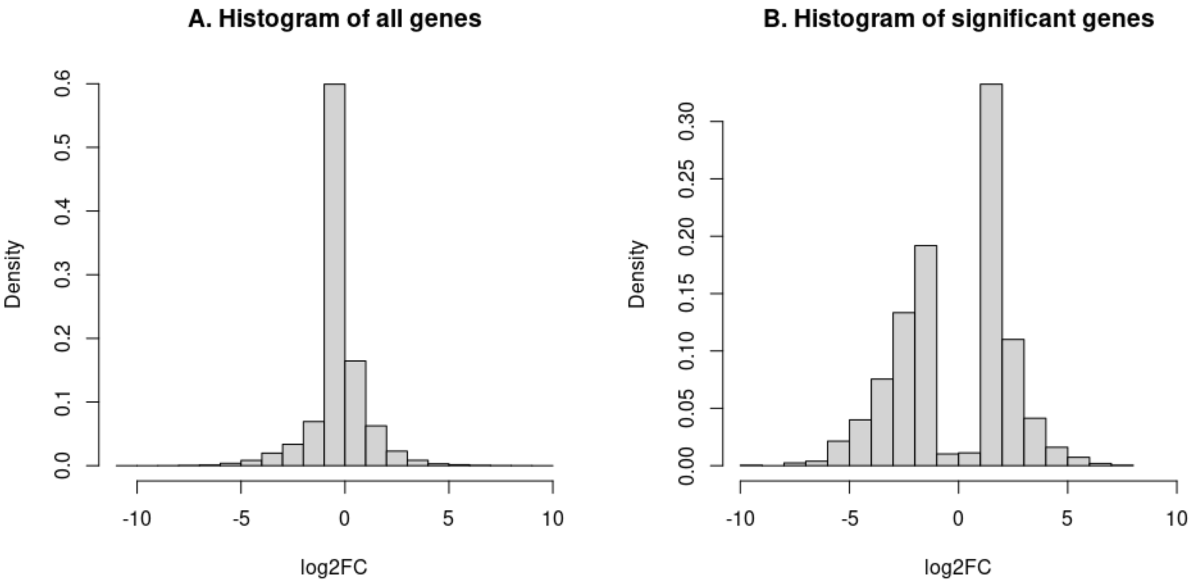


Figure 5. The histogram of the FPKM value of all genes
The x-axis indicates the log10 of ceiling of FPKM value.

For the visualization of the landscape of differential expression profiles, we highlighted the top ten genes differentially expressed in adults versus postnatal day zero samples (Table 1). In addition, we plotted two histograms that describe the distribution of log2 Fold Change of all genes and differentially expressed genes, respectively (Figure 6). More specifically, there were in total 1084 up-regulated genes (592; the cutoff of 0.01), and 1055 down-regulated genes (525; the cutoff of 0.01) under the threshold of 0.05.

Table 1. Top ten differentially expressed genes under the comparison of Adult Versus Postnatal Day 0 sample.

Gene	FPKM_1	FPKM_2	log2(fold change)	p_value	q_value
Plekhb2	22.5679	73.5683	1.70481	5.00E-05	0.00106929
Mrpl30	46.4547	133.038	1.51794	5.00E-05	0.00106929
Coq10b	11.0583	53.3	2.26901	5.00E-05	1.07E-03
Aox1	1.18858	7.09136	2.57682	5.00E-05	1.07E-03
Ndufb3	100.609	265.235	1.39851	5.00E-05	1.07E-03
Sp100	2.13489	100.869	5.56218	5.00E-05	1.07E-03
Cxcr7	4.95844	32.2753	2.70247	5.00E-05	1.07E-03
Lrrfip1	118.997	24.6402	-2.27184	5.00E-05	1.07E-03
Ramp1	13.2076	0.691287	-4.25594	5.00E-05	1.07E-03
Gpc1	51.2062	185.329	1.8557	5.00E-05	1.07E-03


A. Histogram of all genes

B. Histogram of significant genes

Figure 6. (A) The histogram of log2FC of all genes (B) The histogram of log2FC of differentially expressed genes

In order to acquire a view of transcriptional signature behind this, functional annotation clustering analysis revealed that the majority of up-regulated genes were involved in mitochondria, metabolism, sarcomere and respiration related processes, while down-regulated genes played significant roles in cell proliferation, cardiovascular system development, extracellular matrix and embryo development within the cells (Table 2).

Table 2. Top four functional enrichment clusters in the comparison of Adult versus Postnatal Day 0 sample

Enrichment Term*	Score
Up-regulated	
Mitochondria	13.72
Metabolism	12.08
Sarcomere	7.84
Respiration	7.45
Down-regulated	
Cell proliferation	7.42
Cardiovascular system development	7.39
Extracellular matrix	6.99
Embryo development	5.76

*: Only representative terms for each cluster are shown here.

To demonstrate the signatures of representative sarcomere, mitochondrial, and cell cycle genes significantly differentially expressed during in vivo maturation, we highlighted the expression profiles across four different stages (Figure 7). Together with GO enrichment analysis compared with the literature (Table 3), we observed that the majority of enriched terms reported by the previous study can be found in our analysis.

We further explored the expression of the most top 1000 differentially expressed genes found in postnatal day 0 (P0) versus Adult (Ad) mice across all different stages, and patterns in the heatmap potentially suggest that genes grouped by clusters undergone different biological processes from stage to stage within the cells (Figure 8), as indicated by the results of enrichment analysis as well.

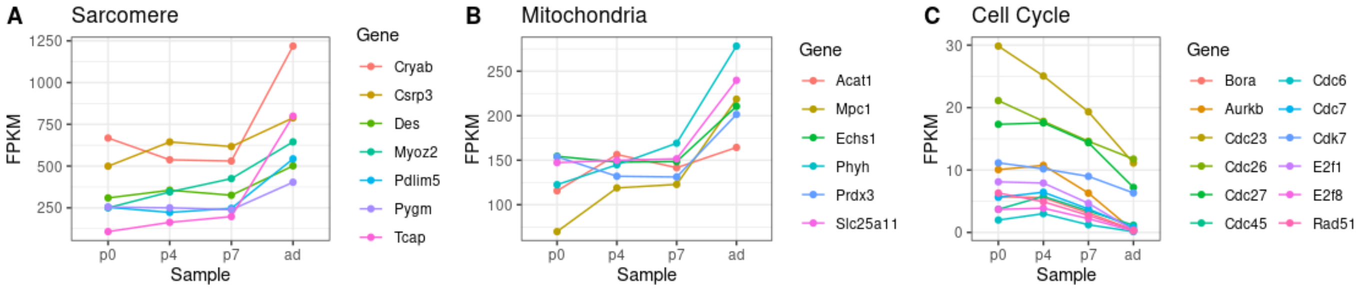


Figure 7. In vivo maturation model indicates a core transcriptional signature
FPKM values of representative Sarcomere, Mitochondria and Cell Cycle genes differentially expressed during in vivo maturation.

Table 3. The comparison results of enrichment analysis

Enrichment Term	**Score (Our study)**	**Score (O'Meara et al's)**	**Overlapped**
Up-regulated			
Mitochondria	13.72	14..35	Yes
Sarcomere	7.84	8.5	Yes
Respiration/Metabolism	7.45	4.98	Yes
Glycolysis	2.77	4.39	Yes
Down-regulated			
Non-membrane bound organelle	4.74	88.91	Yes
Nuclear Lumen	5.49	88.91	Yes
RNA processing	3.34	59.78	Yes
Cell Cycle	5.12	59.78	Yes
DNA repair	3.89	59.78	Yes

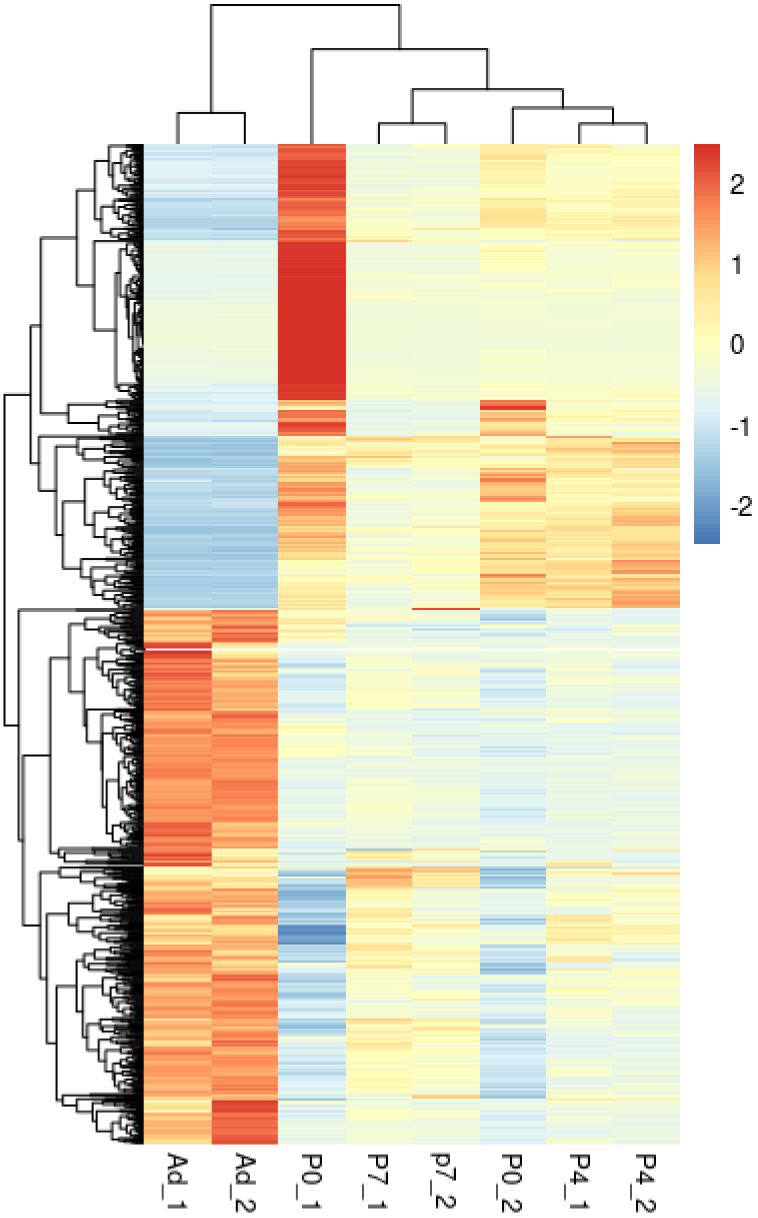


Figure 8. Hierarchy clustering of the top 1000 differentially expressed genes over the course of in vivo maturation.

Discussion

It was evident from the FastQC reports that the overall quality of sequencing reads was decent, except that the distribution of per base sequence content at the 5’ end was noisy. (Figure 1, 2). Possible explanations for this include improper operation during library preparation, and overrepresented sequences, both of which can lead to biased composition and potentially have an impact on downstream analysis.

The distribution of fold change was used to describe differential expression profiles in Adult versus Postnatal Day 0 samples (Figure 6). As we expected, differentially expressed genes that fall in the intervals between low fold changes are significantly depleted compared with the results of all genes, suggesting that we successfully reproduced the differential expression analysis of the original study. However, the number of differentially expressed genes reported by the previous study was different from our results, potentially due to the use of different statistical threshold values.

After tracking genes related with sarcomere, mitochondrial, and cell cycle processes (Figure 7), we demonstrated that our results were highly consistent with O’Meara et al’s study[2]. Genes critical to sarcomere and mitochondrial assembly showed significant increases in expression, while genes involved in cell cycle showed decreased expression over the course of differentiation, reflecting the transcriptional signatures of cardiac myocyte differentiation. However, there was also some inconsistency between our result and the original authors’. For instance, Mpc1 and Bora, which can not be found in our result, were replaced with their synonyms. This was potentially due to the use of discordant reference annotation in the upstream analysis.

Clustering results revealed that the Postnatal Day 0 group were not well aggregated together, suggesting that one of replicates suffered from aberrant gene expression (Figure 8). Adult samples showed several distinct expression patterns different from other Postnatal Day groups, demonstrating a significant transcriptional response during Adult differentiation.

Conclusion

In this study, we successfully processed the raw data, followed by performing the differential expression analysis. Compared with the results from the previous study, our functional analysis recovered insights into the transcriptional signature of cardiac myocyte differentiation.

References

1. Porrello ER, Mahmoud AI, Simpson E, Johnson BA, Grinsfelder D, Canseco D, Mammen PP, Rothermel BA, Olson EN, Sadek HA. Regulation of neonatal and adult mammalian heart regeneration by the miR-15 family.Proc Natl Acad Sci U S A. 2013; 110:187–192. doi: 10.1073/pnas.1208863110.
2. 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.
3. SRA-Tools. (n.d.). Retrieved March 17, 2021, from Github.io website: http://ncbi.github.io/sra-tools/
4. FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2015), "FastQC," https://qubeshub.org/resources/fastqc
5. Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, 25(9), 1105-1111.
6. Wang, L., Wang, S., & Li, W. (2012). RSeQC: quality control of RNA-seq experiments. Bioinformatics, 28(16), 2184-2185.