BF528 Individual Project

Transcriptional Profile of Mammalian Cardiac Regeneration with mRNA-Seq

Benyu Zhou

Introduction

This project includes the programming and analyst part of Project 2, which replicates the in vivo part of the study of O'Meara, C.C. et al., *Transcriptional Reversion of Cardiac Myocyte Fate During Mammalian Cardiac Regeneration [1],* to look at molecular roadblocks that could prevent regeneration in an adult mammalian heart. To be specific, genes that differentially expressed between the P0 (postnatal day 0) and Ad (Adult) stage was investigated in the project.

The bioinformatics methods performed in this project include aligning and QA using tophat[2] and RseQC[3], quantifying gene expression with cufflinks, and identifying differentially expressed genes associated with myocyte differentiation using R and DAVID Functional Annotation. These steps are the key parts in the overall study which processed the raw data and ran through the whole analysis pipeline. The result of the project identifies a common set of differentially expressed genes during in vivo CM maturation.

Methods

The P0_1 and P0_2 FASTQ files obtained by previous lava-lamp data curator was aligned to mm9 reference using TopHat (v2.1.1)[2], a fast splice junction mapper for RNA-Seq reads. TopHat[2] required bowtie2, Boost, and Samtools modules to run on the terminal. A .bam file with successfully aligned pairs in the returning results was what to be used in the following steps. The BAM file was then indexed, and quality control metrics were then retrieved using Samtools flagstas [4] and RseQC Utilities (v3.0.0)[3]. This was done to ensure that there were no errors in alignment and mapping, as well as to guarantee the integrity of the original dataset.

Next, Cufflinks (v2.2.1)[2] was used to count how many reads mapped to annotated regions. The input files included mice genome annotation file (mm9.gtf), mm9 reference genome (mm9.fa) as well as the indexed BAM file produced in the earlier step. A gene tracking file listing FPKM values for all genes was produced. FPKM values are used to quantify gene expression, since in RNA-seq, the relative expression of the transcript is proportional to the number of cDNA. A graphical representation of the distribution of log10 FPKM values was then created, genes that had FPKM value smaller than 0.01 was filtered out. (Figure 3). Afterwards, Cuffdiff (v2.2.1)[2] was run to identify differentially expressed genes between P0 and Ad.

Onwards, the differential expression analysis results generated by Cuffdiff was read into R to better identify significant results and visualize them. A table with the top ten differentially expressed genes, with their FPKM values, log fold change, p and q-values was generated (table 1). Two histograms of the log2.fold_change, one for all the genes and the other for the significant genes only, were generated (Figure 4 and 5) genes were then additionally divided into Up-regulated and Down-regulated genes and saved as two separate files.

Finally, the up- and down-regulated gene sets were then grouped into functionally related clusters via the DAVID[5] software. Then the DAVID results were compared to the David results obtained by O'Meara, C.C. et al.

Results

The QC on the FastQ files using flagstats of samtools reported 49706999 total reads, 100% pass rate, 100% mapping rate, 71.09% properly (29422646 in numbers) paired reads. The bam stats of RseQc returned 2899954 non-unique reads (mapq <mapq_cut) and the percentage is 5.8%. The integrity of the fastq file and the quality of the alignment were assured.

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 1), as well as a slight 3’ bias as the peaks are between 60-80, closer to the 3’ end. (Figure 2).

The cufflinks returned 37469 genes. After filtering out FPKM <0.01, which are extremely low expressed genes that had very non-significant reads, 16337 genes remained. Looking at the FPKM distribution of the 16337 genes (Figure.3), it is clear that FPKM mostly distributed between 0 and 100.

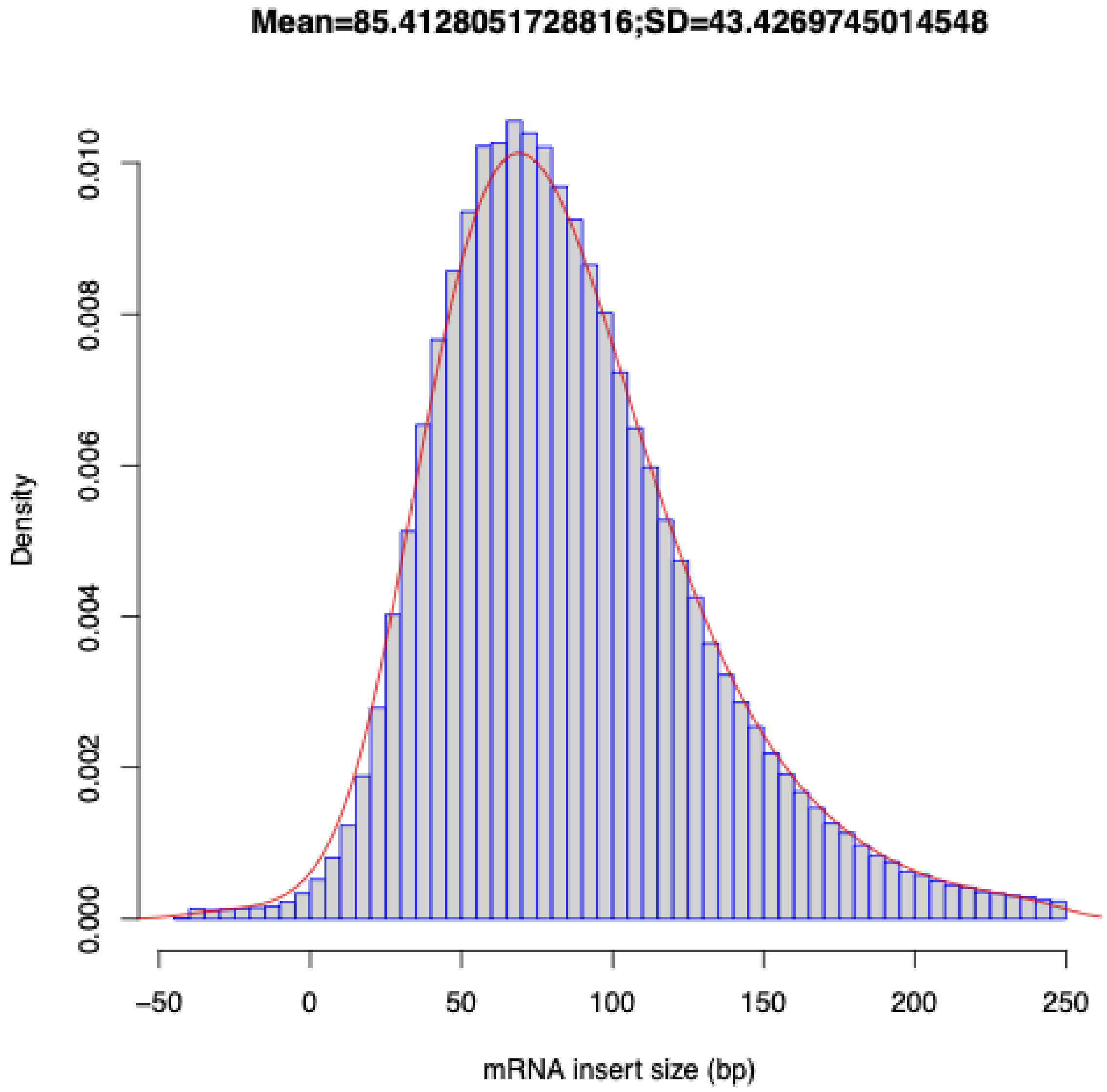

Figure 1 This histogram displays the distribution of mRNA insert sizes in terms of base pairs(bp). The distribution shown here has a mean of insert size of 85.4 bps, with a standard deviation of 43.42 bps.

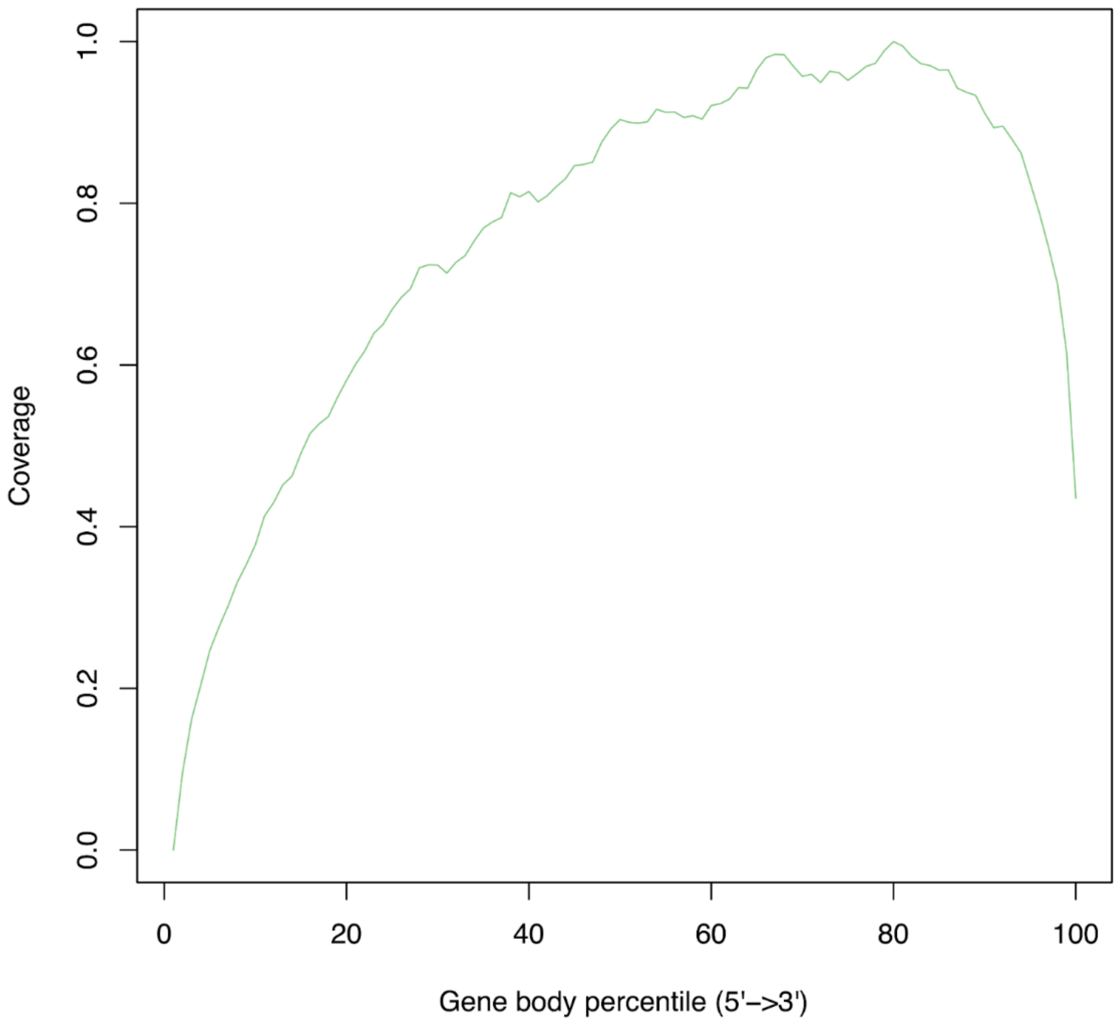

Figure 2. The Gene Body Coverage graph which determines whether there is 5' to 3' coverage bias, displaying coverage of reads between each respective end of the gene.

The cuffdiff DE analysis returned 36329 genes, of which 2139 genes were significant. Table 1 displayed the top 10 DE geens with FPKM values, log2 fold change, p-value and the adjusted q-value. Plekhb2, Mrpl30, Coq10b, Aox1, Ndufb3, Sp100, Cxcr7, Lrrfip1, Ramp1, Gpc1 were the top 10 differentially expressed genes found. The first log2 Fold distribution histogram (Fig. 4) measures the frequency of both significant and non-significant genes while the second histogram (Fig. 5) only includes significant DE genes. The distribution measures the gene expression change that occurs from the P0 to the adult mice (Ad) stage.

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. Top 10 differentially expressed genes with their FPKM data, Log2 Fold Change, p-value, and q-value, ranked by q-value

In Figure 4, it is clear that there is a tall, outstanding peak that occurs around 0. This distinct peak shows that the majority of the expressed genes were not significant, as was reported above that only 2139 out of 36329 genes were significant.

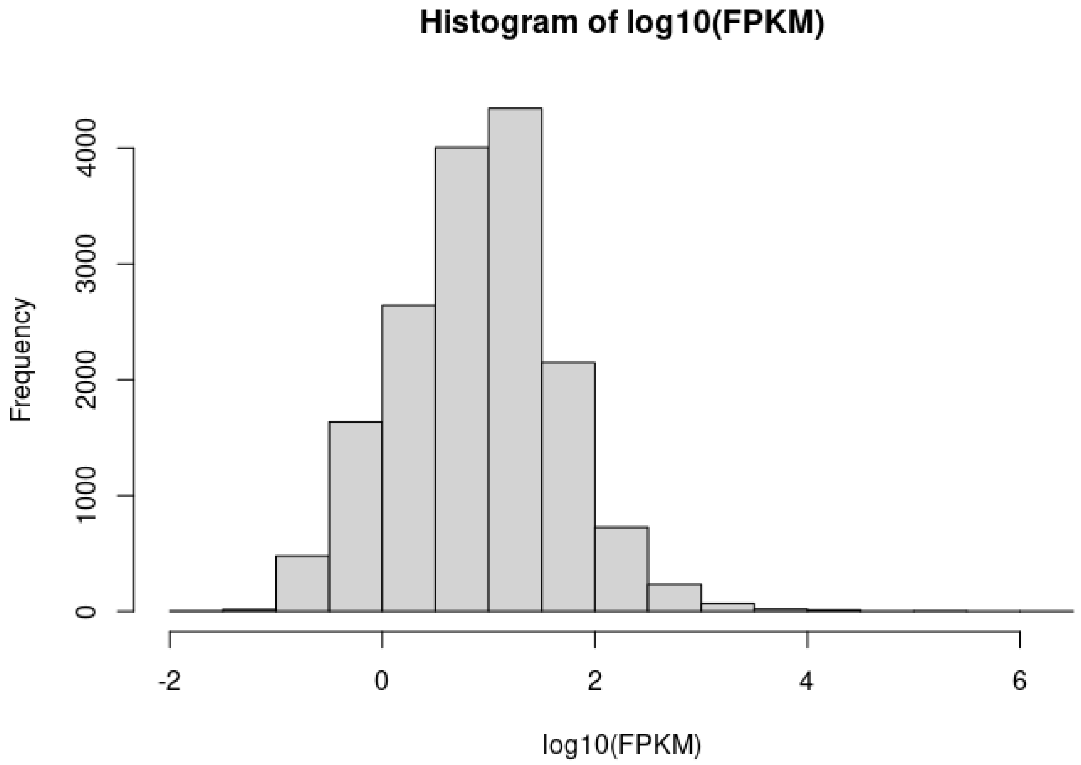

Figure 3. Histogram that illustrates the distribution of Log10 FPKM values that are >0.01.

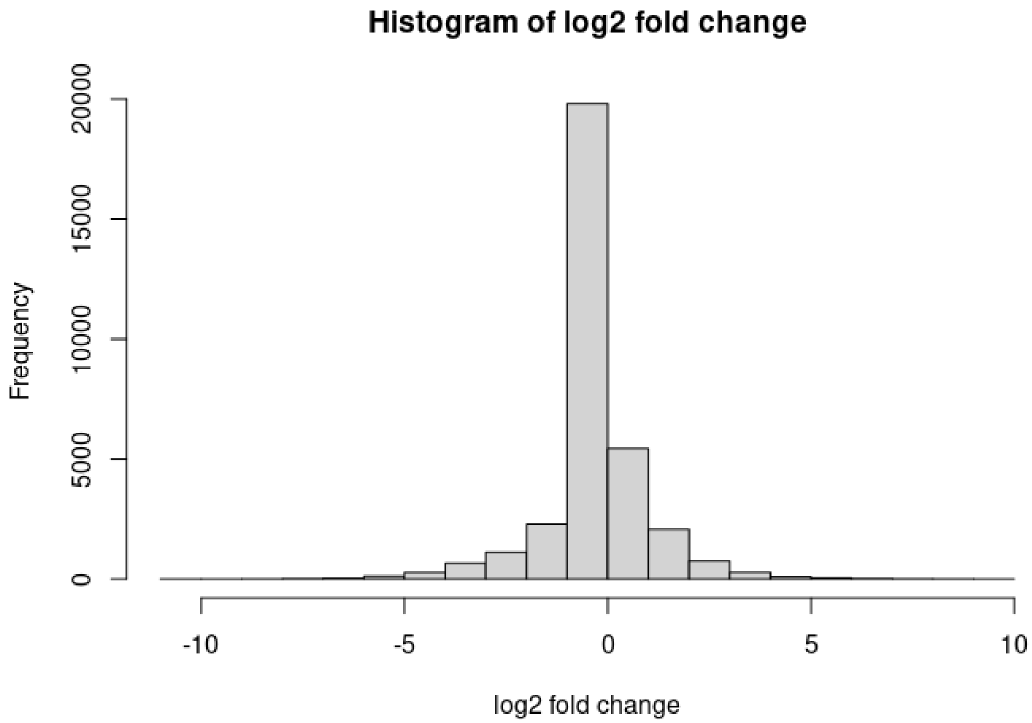

Figure 4. Histogram that illustrates the log2 fold change distribution of the total differentially expressed genes

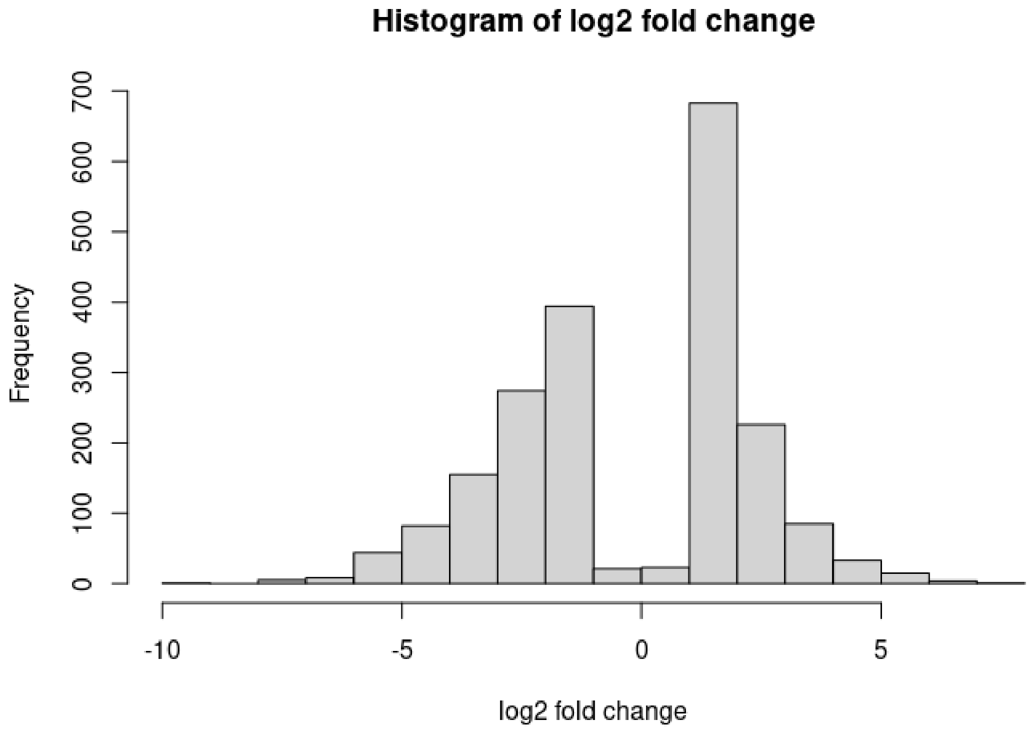

Figure 5. Histogram that illustrates the log2 fold change distribution of the significantly differentially expressed genes

Figure 5 shows no peaks near 0, as all the genes included in this histogram were significant. The genes with log2 fold change below 0 are down-regulated genes, while on the other hand, the genes with log2 fold change above 0 are the up-regulated genes. In total, there were 1084 Up-Regulated genes and 1055 Down-Regulated genes, and 36,329 total genes (Table 2).

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	23.8	GO:0043436 ~ mitochondrion	YES
Annotation Cluster 2	22.27	GO:0055114 ~ obsolete oxidation-reduction process	YES
Annotation Cluster 3	18.33	GO:0006629 ~ lipid metabolic process	YES
Annotation Cluster 4	8.82	GO:0030964 ~ NADH dehydrogenase	YES
Annotation Cluster 5	8.65	GO:0051186 ~ obsolete cofactor metabolic process GO:0006732 ~ obsolete coenzyme metabolic process	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.

The top five up-regulated genes clusters generated by DAVID were identified to all have the GO terms related to metabolism process. (Table 3). The top five down regulated genes clusters generated by DAVID were identified to all have the GO terms related to early life cell growth and development (proliferation, morphogenesis and embryo, organ development) (table 4). All the five top up-regulated genes clusters overlapped with the DAVID results of O'Meara, C.C. et al. in a broader sense since they are all related to metabolism. However, the paper includes more terms other than metabolism. None of the down-regulated genes clusters identified in this project overlapped with the DAVID 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	11.91	GO:0008283~ cell population proliferation	NO
Annotation Cluster 2	10.71	GO:0007010 ~ cytoskeleton organization	NO
Annotation Cluster 3	10.3	GO:0009887 ~ animal organ morphogenesis	NO
Annotation Cluster 4	9.72	GO:0009790 ~ embryo development	NO
Annotation Cluster 5	9.65	GO:0051276 ~ chromosome organization	NO

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.

Discussion

This project got the same results as the lava-lamp project 2 except the DAVID results. The alignment results and QA are the same as the group project and both looked good. Still the David results of down-regulated genes clusters are still very different from the results of results of O'Meara, C.C. et al. This should be the project only deals with in vivo mice mRNA data. Nevertheless, the 5 down-regulated clusters are still related to the early life cell development and regeneration, which makes sense as the potential of CM regeneration was lost in mice’s adult stage.

This project successfully aligns the in vivo P0 mRNA sequencing data and identified the differentially expressed genes between the P0 and AD stage. It could be an indication that the myocytes revert the transcriptional phenotype to a less differentiated state during regeneration.

Reference

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. 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). https://doi.org/10.1038/nprot.2012.016
3. Wang, L., Wang, S., & Li, W. (2012). RSeQC: quality control of RNA-seq experiments. Bioinformatics (Oxford, England), 28(16), 2184–2185. http://doi.org/10.1093/bioinformatics/bts356
4. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H, Twelve years of SAMtools and BCFtools, GigaScience (2021) 10(2) giab008 [33590861]
5. B.T. Sherman, M. Hao, J. Qiu, X. Jiao, M.W. Baseler, H.C. Lane, T. Imamichi and W. Chang. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). Nucleic Acids Research. 23 March 2022. doi:10.1093/nar/gkac194.[PubMed]