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

Konrad Thorner, Aishwarya Deengar, Jia Liu, Morgan Rozman

INTRODUCTION

It is well established that the mammalian cardiac myocytes exit the cell cycle and differentiate giving them limited capacity to regenerate in the event of injury. However, neonatal mice retain the ability to regenerate their hearts in case of an injury until the first week of life. Furthermore, the cells known to drive self-renewal are myocytes themselves as opposed to stem cells or progenitor cells present there.

In order to identify the genes and gene networks responsible, and if myocytes can thus be reverted to a “less differentiated” state, O’Meara et al. employed a mouse model and RNA sequencing to look at gene expression changes in this critical window of time [1]. Their focus was to obtain and analyze transcriptional data to find the regulators of such an event, which can then be further screened and confirmed *in vitro.* It is clear that in identifying the mechanisms by which myocytes undergo cell cycle activity during regeneration is crucial for shedding light on the molecular roadblocks which prevent regeneration in an adult human heart.

The purpose of our study is to recreate the results from O’Meara et al. In particular, we determine the number and type of differentially expressed genes between adult mice and mice at postnatal day 0. To do so, we first examine the original study’s data to ensure it is of good quality. We use alignment techniques to prepare the raw data for analysis, which then allows us to determine differential gene expression.

DATA

In the original study, the methodology began with extracting total RNA from either isolated mouse cardiomyocytes (CM) or ventricular myocardium (VM) samples using Trizol, followed by the steps of library preparation. The samples fall into one of four categories: those that matured normally *in vivo*, those that came from sham or resected mice, adult CMs that were explanted and cultured, and those that differentiated *in vitro* [1]. We are interested in the *in vivo* samples, which are summarized in Table 1.

Sample name	Description	Read (in millions)
P0_1	0 day postnatal ventricular myocardium	21.6
P0_2		25.7
P4_1	4 day postnatal ventricular myocardium	22.1
P4_2		24.4
P7_1	7 day postnatal ventricular myocardium	16.1
P7_2		16.9
Ad_1	Adult ventricular myocardium	24.2
Ad_2		21.9

Table 1. Summary of all RNA-seq samples used from O'Meara et al., taken from the Gene Expression Omnibus accession for the paper. Each sample type has two replicates, which themselves consist of at least two pooled heart ventricles.

One of these, that comes from the VM of postnatal day 0 mice, is called P0_1 and was processed by our group. P0 is the first timepoint, but there are also samples from P4, P7, and adult mice being assessed. Collectively they are used to trace the unique transcriptional changes of CM differentiation relative to each other, from the first week of life to after.

RNA-seq libraries were prepared and run on Illumina HiSeq 2000 and 2500 instruments, generating paired-end reads of 40 base pairs. Quality control can then be performed using the FastQC package [2] . Both reads have no bases flagged as poor quality, GC content of 49%, and no overrepresented sequences. However, FastQC gives a warning for per base sequence content. There is a nucleotide bias in approximately the first 15 bases (Figure 1), but this is a well-documented consequence of the random priming during library preparation [3]. It cannot be eliminated, but it should also not affect downstream analysis.

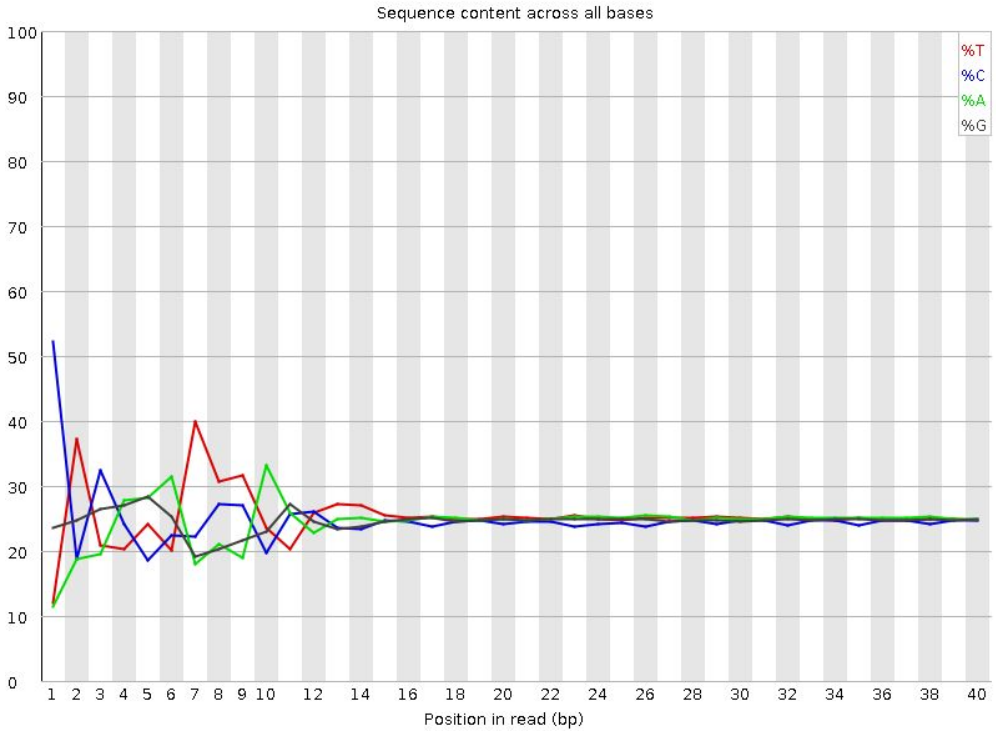

Figure 1: The FastQC per base sequence content plot. Visualizes the average proportion of nucleotides at each position of the read for all reads. Deviations from an equal proportion of A with T and G with C can indicate poor quality.

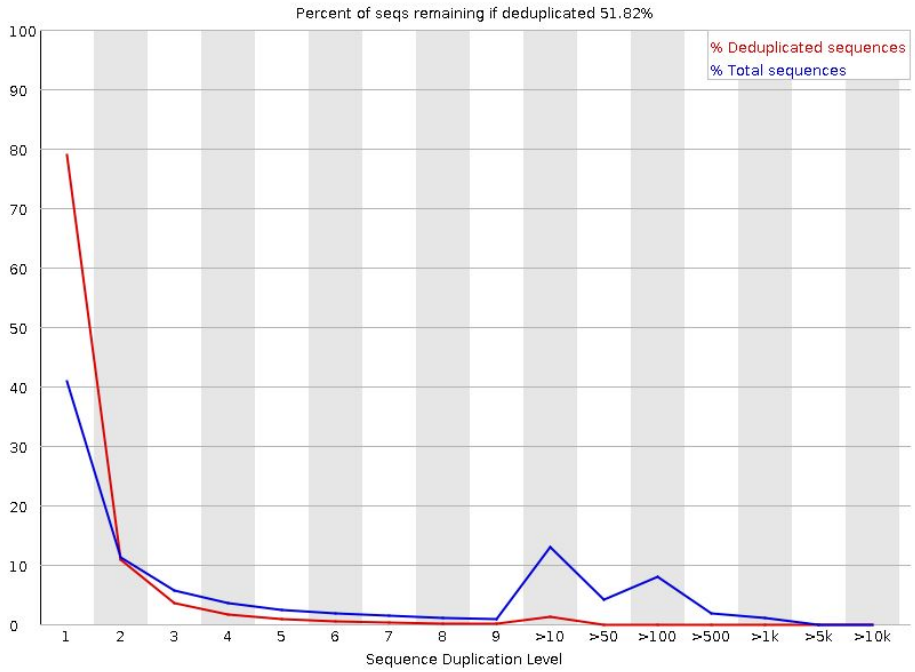

Figure 2: The FastQC sequence duplication levels plot. Displays the percentage of the library that was sequenced a given number of times. Lower levels indicate less redundancy and thus a more diverse library.

Additionally, it is apparent from Figure 2 that many reads are duplicated anywhere from 10 to 500 times, with percent of sequences remaining after deduplication being approximately 52%. Multiple reads mapping to the same region is expected in the context of RNA-seq experiments, since the abundance of different mRNAs will vary. Highly expressed transcripts from highly expressed genes would be sequenced many times, which would directly explain these peaks. Overall the levels are still low, suggesting good diversity in the library, and no correction is needed.

METHODS

We first align the paired-end reads from the P0 sample to the mouse reference genome, mm9 [4]. Alignment was performed using TopHat [5] with the same options as specified in the original paper's methods. These options include an expected inner distance between mate pairs of 200, a segment length of 20, allowance of up to 1 mismatch in each independent sequence alignment, and only aligning reads across junctions known and supplied by the reference. We chose not to exclude the multi mapped reads for simplicity, but it is important to note that including multi mapped reads can have an impact on abundance estimation. The summary statistics shown were gathered using the RSeQC package [6], which assesses quality of RNA-seq data. All reads in our data set passed mapping quality control metrics in RSeQC, as well as in SAMtools [7].

Inner distance is defined as the distance in base pairs between read 1 and read 2 in a paired-end alignment. We calculated the mean inner distance, or insert size, of alignments using RSeQC. We calculated coverage over the gene body using RSeQC.

We then report the expression values, normalized for sequencing depth and length, using FPKM (Fragments Per Kilobase Million). We performed this normalization using Cufflinks [8], which was chosen because it is the same program used in the original paper. We visualized distribution of FPKM with and without nearly-zero values, but we included them in further analysis because a count of zero can be significant when looking into differential expression.

To compare differential expression across age groups and to see the effects of maturation on expression, we used Cuffdiff, a program within Cufflinks, on the P0 and Ad samples found in Table 1.

Lastly, the significant gene set data is subsetted to obtain two tables containing upregulated genes and downregulated genes, that is further clustered by function. This is done with DAVID, or Database for Annotation, Visualization and Integrated Discovery Analysis, a functional classification tool that scans and groups genes into according to gene ontology (GO) annotations. [9].

RESULTS

From our initial TopHat analysis, we obtained all alignments, or hits, to the reference genome. Summary statistics on the accepted hits show that all 49,706,999 reads were mapped to the reference genome (Table 2). We found 77.4% of the reads were unique, and 5.8% were multi mapped.

Read Mapping Overview	
Mapped	49,706,999 (100.0%)
Unaligned	0 (0.0%)
Unique	38,489,380 (77.4%)
Multi Mapped	2,899,954 (5.8%)
Total Reads	49,706,999

Table 2. Summary statistics were generated using RSeQC and shown out of the entire read set: number of reads (% total reads). All reads were mapped to the mm9 reference genome, and we found 77.4% of the total reads were unique. All reads pass the quality control thresholds intrinsic in SAMtools and RSeQC.

The mean inner distance size of our reads is 85 base pairs with a standard deviation of 43 base pairs (Figure 3). The curve also has at the left tail end reads with sizes as small as -50, which indicates they overlap. The distribution of inner distance appears to be approximately normal, and we find the distribution to be acceptable for further analysis.

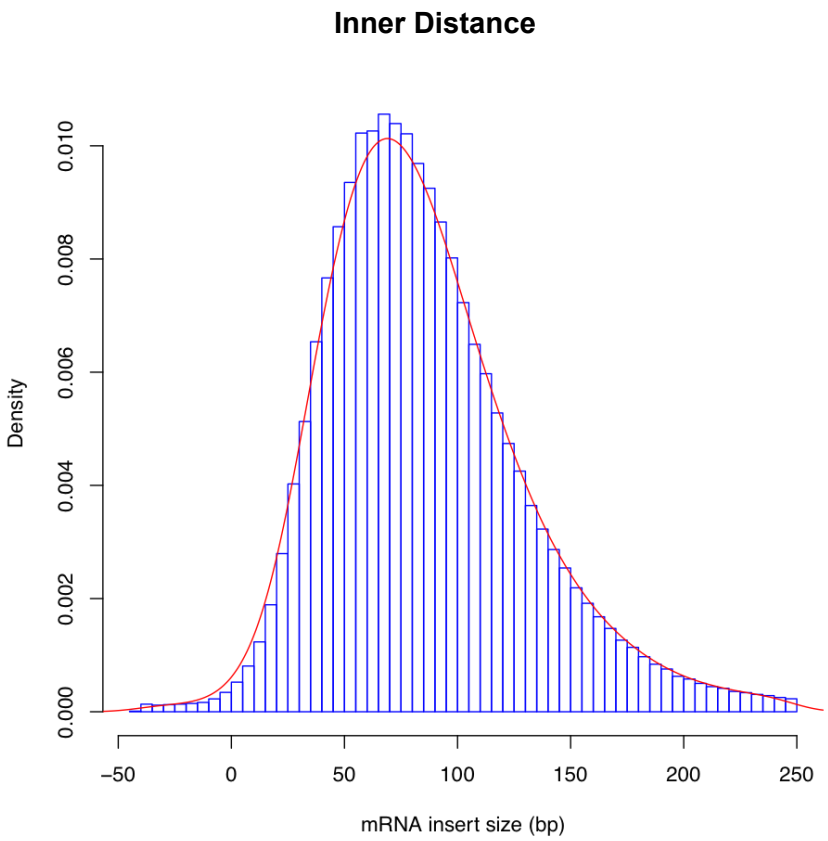

Figure 3: The inner distance, or insert size, of reads was found using the RSeQC package. Density is the proportion of alignments with a given mRNA insert size. The mean inner distance is 85 base pairs with a standard deviation of 43 base pairs.

Based on the gene body coverage results (Figure 4), we find that there is a 3' bias in our sample because there is less coverage near the 5' end of the gene body, with more coverage near the 3' end. We do not think that this bias is concerning, and conclude that the data is fit for further analysis.

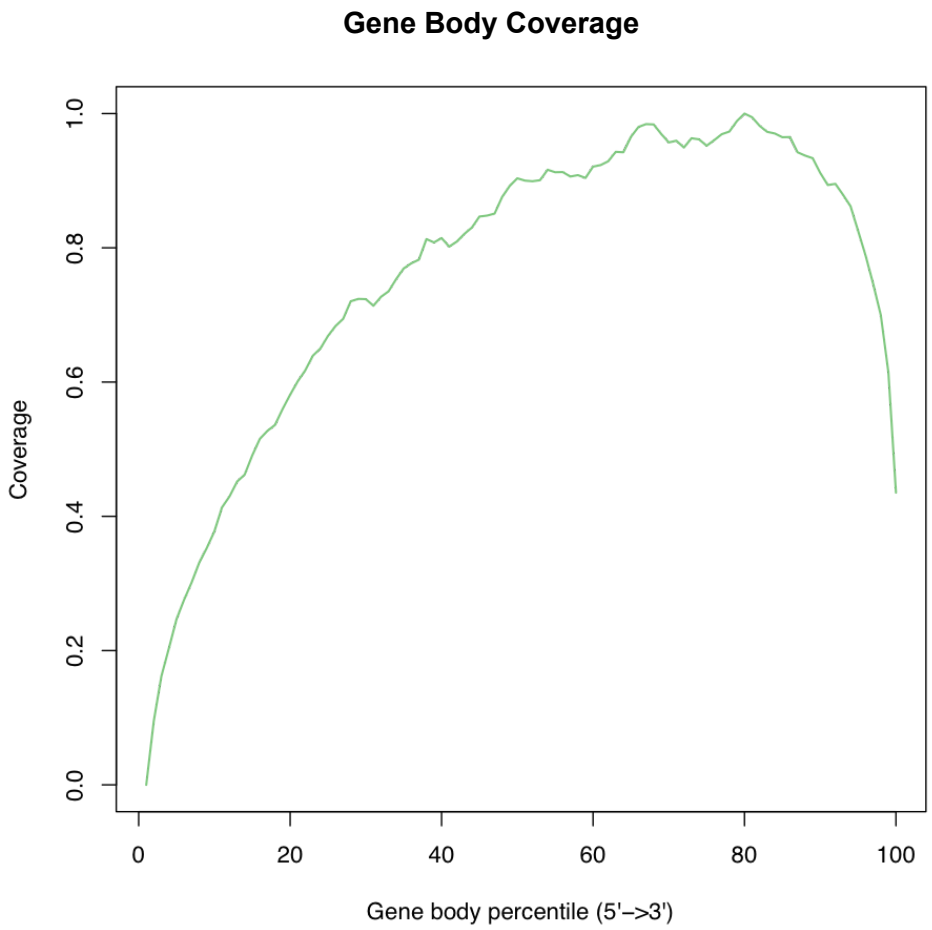

Figure 4: Coverage by percentile of the gene body from 5' to 3' region. Because we see greater coverage at the 3' end, there is 3' bias in our sample.

We found that 56% of the genes had an FPKM count of less than 10^{-4} (Table 3). We chose to visualize the distribution of FPKM with and without these nearly-zero values (Figure 5).

FPKM Distribution	
Total Genes	37,469
FPKM <= 0.0001	21,109 (56%)
FPKM > 0.0001	16,360 (44%)

Table 3. Distribution of FPKM (fragments per kilobase million). We found 56% of the genes had a count of nearly zero.

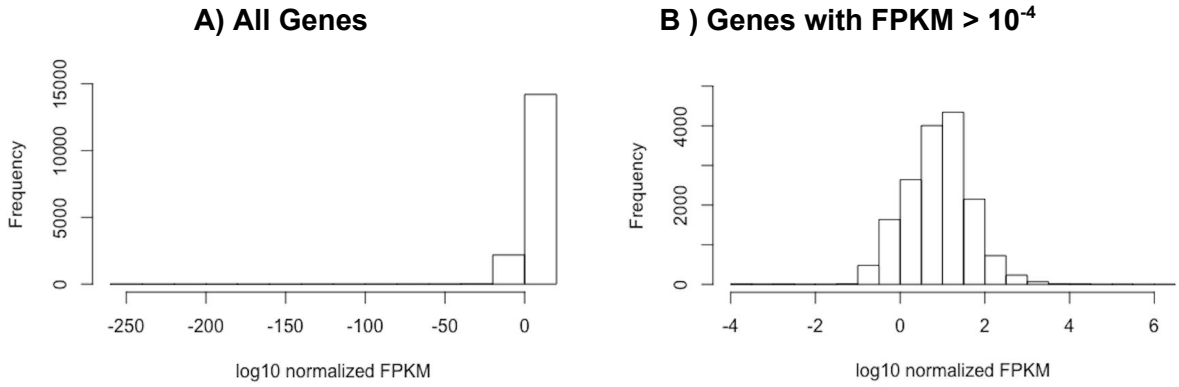

Figure 5. Log10 normalized FPKM counts for the read alignments per gene. Frequency refers to the number of genes. **(A)** All genes, including those with a very small FPKM. Most genes fall near log10(FPKM) = 0, but there are many genes that have an FPKM very near zero. **(B)** Only genes with an FPKM > 0.0001. Most genes have an FPKM of 1-100.

The differentially expressed genes obtained were sorted by q value in ascending order and the top ten genes were reported in Table 4 below. The columns in this table include the gene name, the FPKM values for postnatal day 0 (P0) versus adult, log2.fold change values (or the log-ratio of a gene's expression values in the two different conditions), the p value, and the q value.

Gene	P0 stage	Ad stage	Log2.fold change	p value	q value
Plekhb2	22.57	73.57	1.70	5.00E-05	0.0010
Mrpl30	46.45	133.04	1.51	5.00E-05	0.0010
Coq10b	11.058	53.30	2.27	5.00E-05	0.0010
Aox1	1.19	7.10	2.58	5.00E-05	0.0010
Ndufb3	100.60	265.24	1.40	5.00E-05	0.0010
Sp100	2.13	100.87	5.56	5.00E-05	0.0010
Cxcr7	4.96	32.28	2.70	5.00E-05	0.0010
Lrrfip1	118.99	24.64	-2.27	5.00E-05	0.0010
Ramp1	13.21	0.69	-4.26	5.00E-05	0.0010
Gpc1	51.21	185.33	1.85	5.00E-05	0.0010

Table 4. Top 10 differentially expressed genes as determined by q value following processing with cuffdiff. 'P0' stage is the postnatal day 0 gene expression data while 'Ad' is the gene expression data at the adult stage.

Upon sorting in ascending order, it was observed that the top 10 genes had the same p value and q value. There wasn't any clear discrepancy in the data since the values did increase as we moved down the list. Genes that were reported in this table have functions ranging from metabolism and cellular trafficking to their roles in events like pathogen infection. Genes like CXCR7, SP100, LRRFIP1 are associated with pathogen infection. COQ10b, AOX1 and NDUFB3 are genes which regulate metabolism and regulate the electron transport chain. PLEKHB2 regulates endosome recycling while RAMP1 and Gpc1 are membrane associated proteins which regulate trafficking of molecular components in the cell. MRPL3 is a mitochondrial protein associated with protein synthesis.

A histogram of the log2.fold change against frequency was generated using all the genes and is shown in Figure 6.

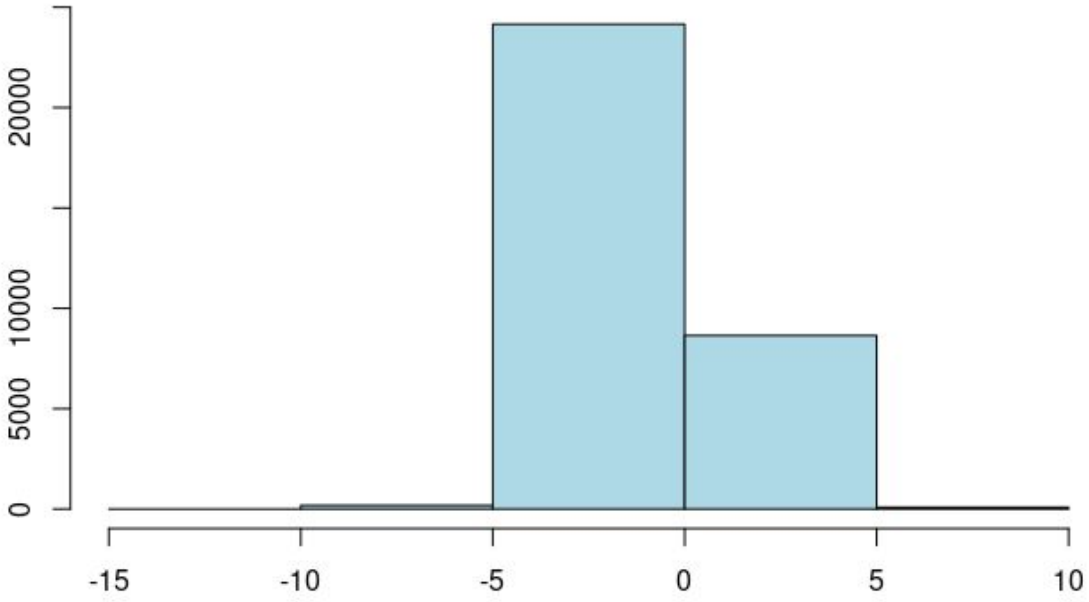

Figure 6(A)

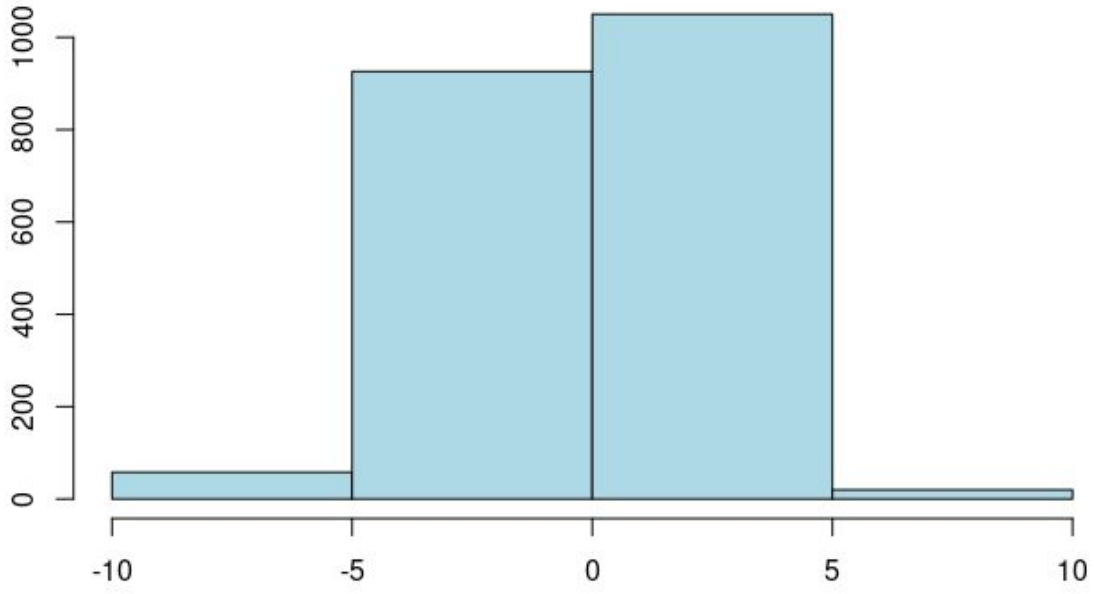

Figure 6(B)

Figure 6. (A) Log2.fold change vs frequency histogram for all genes. **(B)** Log2.fold change vs frequency histogram for the genes with significant values designated as 'Yes'.

Figure 6(A) shows the histogram generated using the log2 fold change values of all the genes. Figure 6(B) meanwhile is generated with genes having significance designated as 'Yes'. This means that the biological variability that might exist in the gene expression data is accounted for and was determined using the variance of the read counts. Upon studying the histograms, it was observed that there is a very high frequency of genes with a negative log-ratio of the two conditions when all genes were used, but there is a positive log ratio and slightly lesser frequency of genes with negative log ratio when significant genes were used. Furthermore, out of all significant genes, 1,055 were downregulated and 1,084 were upregulated (detected at $p < 0.01$).

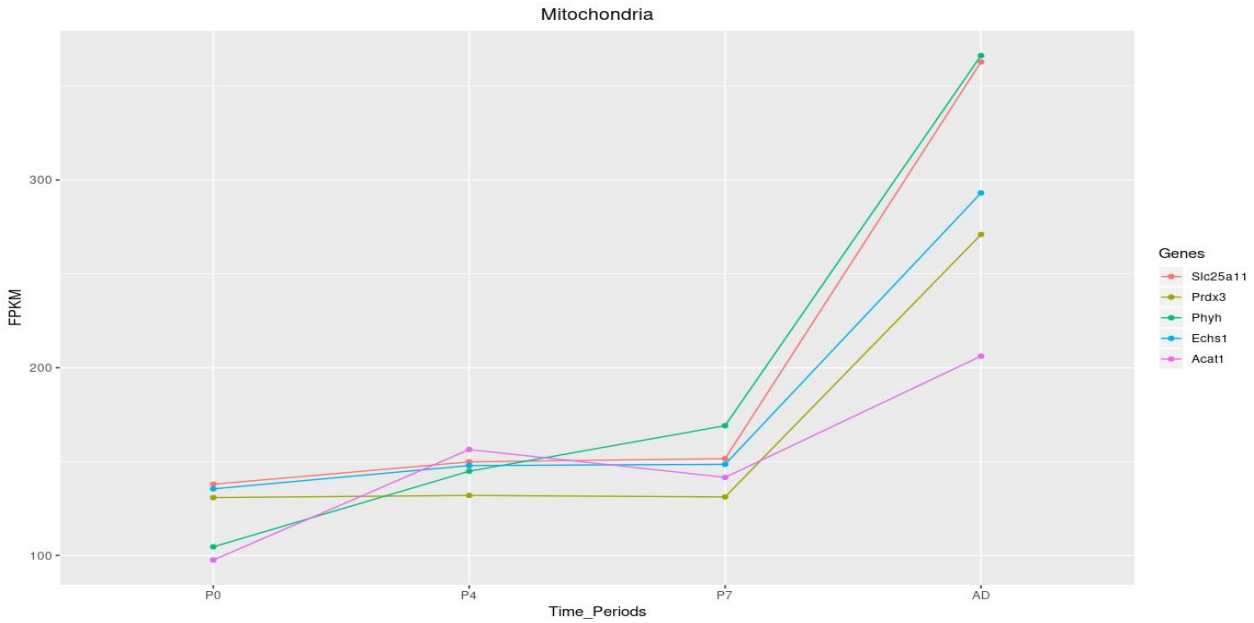

Figure 7(A)

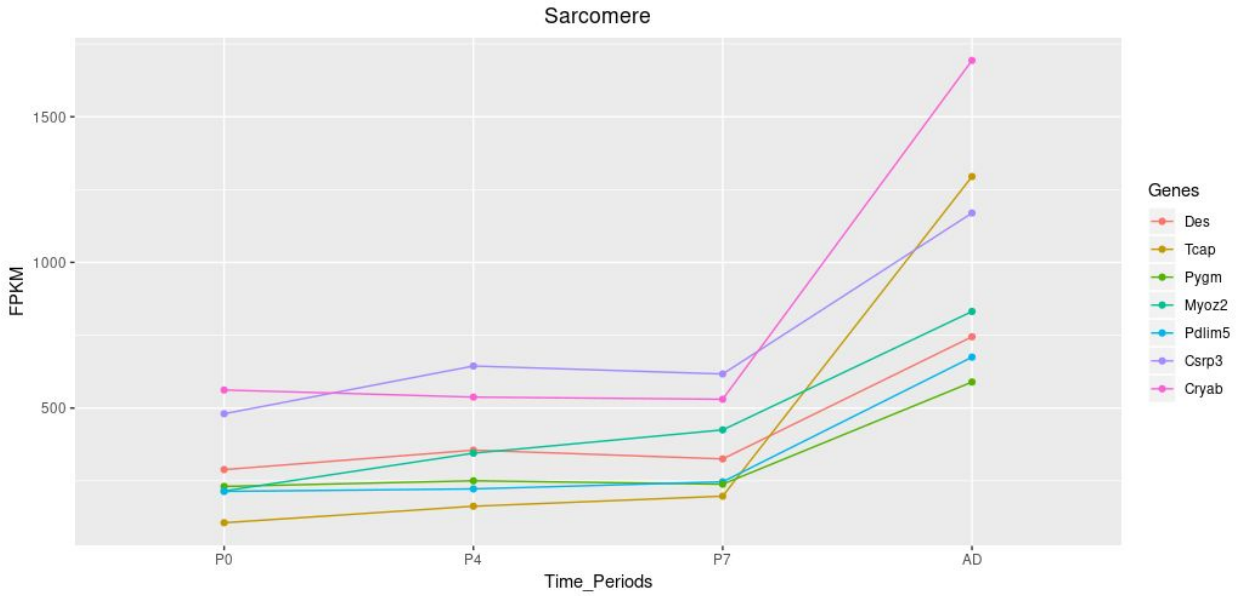

Figure 7(B)

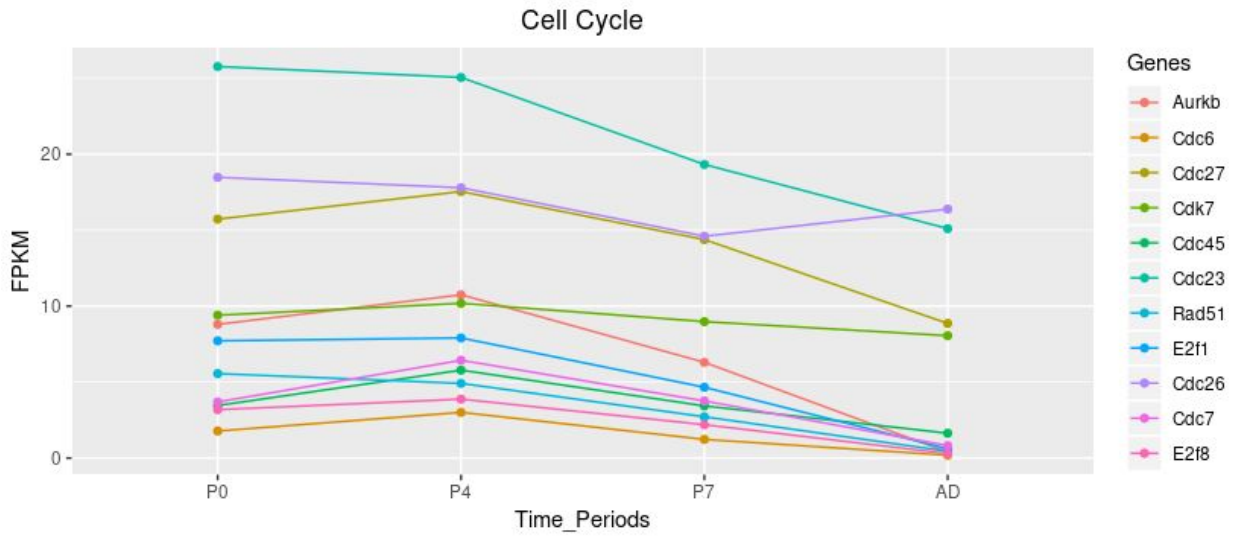

Figure 7(C)

Figure 7. FPKM values of representative mitochondrial, sarcomere and cell cycle genes significantly differentially expressed during P0, P4, P7 and adult (AD). **(A)** FPKM values of representative mitochondrial genes across different time periods. **(B)** FPKM values of representative sarcomere genes across different time periods. **(C)** FPKM values of representative cell cycle genes across different time periods.

In Figure 7, FPKM values are plotted for each set of genes at each timepoint, with results appearing very similar to those of the paper. When samples are in time periods P0, P4 and P7, FPKM has a trend of slow rise in mitochondrial and sarcomere FPKM, while in the adult period, the mitochondrial and sarcomere FPKM rise quickly. In cell cycle annotated genes, the opposite occurred. When samples are in time periods P0, P4 and P7, FPKM slowly decreases, and in the adult period, the cell cycle FPKM drops quickly. This result proves that cell cycle genes have a higher expression level in early life stages, which drops quickly in older life stages. At the same time, as samples get older, mitochondrial and sarcomere genes are expressed at a higher level.

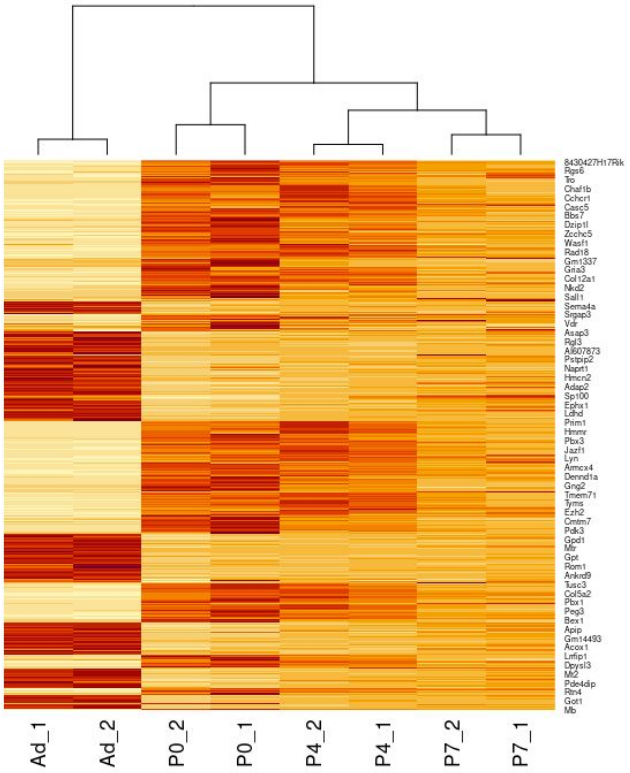

Figure 8. Top 1000 genes found to be differentially expressed between P0 and Ad. The clustered heatmap shows relation between 8 samples. (Ad_1 and Ad_2 are replications from the same time period, as are P0_1 and P0_2, P4_1 and P4_2, P7_1 and P7_2.)

Next, the heatmap in Figure 8 shows the gene expression pattern for each sample to detect more general trends. The distribution of clusters confirms the accuracy of sample data preprocessing, and replicates closely match one another. The cluster of the top 1000 differentially expressed genes is particularly concentrated, which may be the reason why the gene expression patterns are so clear.

Annotation cluster	Representative annotation terms	p value	Enrichment score
1 Cellular Component	Extracellular Space	8.4E-2	1.38
	Extracellular Region Part	1.7E-1	
	Extracellular Region	4.9E-1	
2 Molecular Function	Positive regulation of cell proliferation	2.1E-2	1.54
	Regulation of cell proliferation	6.8E-2	
	Transitional Metal Ion Binding	4.8E-1	
	Metal Ion Binding	5.5E-1	
	Cation Binding	5.6E-1	
	Ion Binding	5.7E-1	
3 Cellular Component	Intracellular Organelle Lumen	3.0E-1	1.92
	Organelle Lumen	3.0E-1	
	Membrane-enclosed Lumen	3.1E-1	

Table 5. Summary of the most upregulated gene functions, found to cluster into 12 groups by DAVID analysis

Annotation Cluster	Representative annotation terms	p value	Enrichment Score
1 Molecular Function	Metal Ion Binding	5.0E-2	0.78
	Cation Binding	5.2E-2	
	Ion Binding	5.5E-2	

Table 6. Summary of the most downregulated gene functions, found to cluster into 3 groups by DAVID analysis.

Downregulated Gene Pathway Table

Downregulated	**Enrichment Score: 1.3**	
Term	P-Value	Genes
*Metal ion binding	0.0503	B4GALT2, PDXP, CDHR4, HIST3H2A, ACSL3, CBX6
*Cation binding	0.0522	B4GALT2, PDXP, CDHR4, HIST3H2A, ACSL3, CBX6
*Ion binding	0.0549	B4GALT2, PDXP, CDHR4, HIST3H2A, ACSL3, CBX6

Upregulated Gene Pathway Table

Upregulated	**Enrichment Score: 7.2E-1**	
Term	P-Value	Genes
*Extracellular space	0.0839	IFNAR2, CNTF, ANG

Upregulated	**Enrichment Score: 6.5E-1**	
Term	P-Value	Genes
Positive regulation of cell Proliferation	0.0211	CNTF, ANG, DNAJA3
Regulation of cell proliferation	0.0679	CNTF, ANG, DNAJA3

Table 7. Comparison of DAVID analysis with O'Meara et al. results. Significance of p-value is 0.1. Downregulated gene pathways are in one cluster. Upregulated gene pathways are divided into two clusters. Term is the annotation for the biological pathway and * shows overlap with the results reported in the paper.

Finally, in Tables 5 through 7, we perform DAVID analysis and compare the results obtained with those reported in the paper. 12 functions were upregulated and 3 functions were downregulated. In addition, 9 terms for upregulated genes and 2 terms for the downregulated genes were not clustered. We then set the significance as 0.1 and filtered out all pathways with

p value greater than 0.1. Downregulated gene functions were wholly in agreement with the original study, but for upregulated gene functions, we find two new biological pathways: positive regulation of cell proliferation and regulation of cell proliferation.

DISCUSSION

When first assessing the raw data, nucleotide bias and sequence duplication were ruled out as issues as they are inherent to RNA-seq. We did however discover a 3' bias in the gene body coverage of our data. We speculate that this bias could be an artifact of the RNA-seq methods used by the authors of the original paper.

The significant genes were sorted to derive the upregulated and downregulated genes, and DAVID Analysis was performed. We found an enrichment of cell proliferation genes not found in the original study. However, the most prominent terms of that study, such as the mitochondria and sarcomere (with genes Fhl2, TCap, etc) were found to be upregulated but not as significantly as the authors reported. Terms we did identify, referring to organelles and ion-binding, may be linked to these.

While the DAVID analysis did not exactly replicate the author's results, there were also terms that failed to cluster. This could be due to an error in our analysis, or the result of using only the P0 and adult samples, which would not take into account more intermediate transcriptional states. Furthermore, the authors have done numerous experiments including *in vitro, in vivo* and using *explants* to validate their study. Upon comparison it can be concluded that our results using two samples, though close, might not portray exactly what the authors could.

The authors also highlighted the importance of STAT3 and STAT6 transcription factors in promoting heart regeneration as it mediated IL13 signalling which in turn is known to induce myocyte cell cycle entry [10]. Our list of upregulated genes included STAT3 which confirms their hypothesis of the existence of such a relationship.

In upregulated gene function, we find two new pathways, positive regulation of cell proliferation and regulation of cell proliferation. However, the enrichment score of these two pathways is low(0.65). So, more evidence is needed.

Comparing the FPKM values of representative sarcomere, mitochondrial, and cell cycle genes significantly differentially expressed in samples, we draw the conclusion that cardiac regeneration is not a stochastic loss of the mature cell state, but rather a direct transcriptional reversion of the differentiation process.

CONCLUSION

The analysis of the gene expression data from O'Meara et al. was in agreement in many respects, especially how the FPKMs vary temporally for many of the genes being investigated. But it also differed in some areas, possibly due to the more limited set of data, in particular when identifying functional enrichment.

REFERENCES

1. O'Meara et al. Transcriptional Reversion of Cardiac Myocyte Fate During Mammalian Cardiac Regeneration. Circ Res. Feb 2015.
2. Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc
3. Hansen K, Brenner S, Dudoit S. Biases in Illumina Transcriptome Sequencing Caused by Random Hexamer Priming. Nucleic Acid Res. Jul 2010.
4. Church DM, Goodstadt L, Hillier LW, Zody MC, Goldstein S, She X, Bult CJ, Agarwala R, Cherry JL, DiCuccio M. Lineage-specific biology revealed by a finished genome assembly of the mouse. PLoS Biol. 2009.
5. Trapnell C, Pachter L, Salzberg SL. Tophat: Discovering splice junctions with rna-seq. Bioinformatics. 2009.
6. Wang L, Wang S, Li W. RSeQC: quality control of RNA-seq experiments. Bioinformatics. 2012.
7. Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009.
8. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL and Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012.
9. Huang DW, Sherman BT, Tan Q, et al. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol*. 2007;8(9):R183. doi:10.1186/gb-2007-8-9-r183
10. Zhang C, Li Y, Wu Y, Wang L, Wang X, Du J. Interleukin-6/signal transducer and activator of transcription 3 (STAT3) pathway is essential for macrophage infiltration and myoblast proliferation during muscle regeneration. *J Biol Chem*. 2013;288(3):1489–1499. doi:10.1074/jbc.M112.419788