BF528 Individual Project Xiaojing Peng

Project 2 Section 6

Introduction

High-throughput sequencing technologies provide researchers a strong tool to understand and analyze genome-wide gene expression dataset, this project aims to explore the differences in the transcriptional profiles of neonatal and adult mice using mRNA abundance data from Transcriptional Profile of Mammalian Cardiac Regeneration. Since I was the programmer of project2, I take the *cuffdiff* output generated in project 2 section 5 and use it to replicate section 6, which is to identify differentially expressed genes association with myocyte differentiation and the biological interpretation.

O'Meara, et al. has generated the RNA seq data from cells collected at the embryonic stem cell and CM stages, and it is shown that within the first week of life, and into adulthood, mammalian CMs undergo a maturation process, therefore, the original paper has identified the number of common or unique differentially expressed genes during in vivo and in vitro CM differentiation. Here, we will examine the cuffdiff output file, which contains the differential expression statistics comparing P0 versus Adult, to interpret the results.

Methods

We start from the gene_exp.diff file, which contains the differential expression statistics comparing the two conditions (postanal day 0 and adult). By loading into R studio, we first sort the data in the increasing order of q-values and then generate a top table which contains the top ten differentially expressed genes. Then we subset the original data frame with genes which has a value equal to "yes" in the last column and compared the histogram of log fold change between the original table and subset table. Furthermore, we create two separate data frames with only the up- and down-regulated genes based on the positive or negative values of log 2 fold change.

Gene enrichment is analyzed using the Database for Annotation, Visualization, and Integrated Discovery (DAVID) tool on the differentially expressed genes determined earlier. The functional Clustering function uses a Kappa statistic score to measure relationships among the annotation terms based on the degrees of their co-association genes and a novel fuzzy clustering algorithm to group the similar, redundant, and heterogeneous annotation contents from the same or different resources into annotation groups.

The output from DAVID functional annotation clustering is used to organize the enriched gene sets into functionally related clusters. Mus musculus is the species of interest and GOTERM_BP_FAT, GOTERM_MF_FAT, and GOTERM_CC_FAT are selected as the gene ontology groups of interest in both sets of analyses.

Results

There are 36,329 genes in total, and the table containing the top ten differentially expressed genes in the comparison of Adult vs. P0, with the corresponding gene name, FKPM values, log2 foldchange, p-value and q-value is shown below. All the p-value are less than 0.05, suggesting gens are significantly differential expressed.

gene	value_1	value_2	log2.fold_change.	p_value	q_value
Adhfe1	12.7138	25.7368	1.01743	5.00E-05	0.00032056
Tmem70	36.9448	80.9577	1.1318	5.00E-05	0.00032056
Gsta3	0.412723	6.76999	4.03591	5.00E-05	0.00032056
Lmbrd1	6.58776	12.6753	0.944164	5.00E-05	0.00032056
Dst	19.721	51.5969	1.38755	5.00E-05	0.00032056
Plekhb2	25.8529	68.601	1.4079	5.00E-05	0.00032056
Cox5b	505.414	881.798	0.802984	5.00E-05	0.00032056
Mrpl30	56.205	124.292	1.14497	5.00E-05	0.00032056
Tmem182	46.2243	103.517	1.16314	5.00E-05	0.00032056
Nck2	12.174	6.29987	-0.950409	5.00E-05	0.00032056

Table 1. Top 10 differentially expressed genes generated from cuffdiff summary statistics file

The histogram of log2 foldchange for all genes is generated and we used 60 as the number of the breaks to control the number of bars in the plot. It can be shown that there is a roughly normal distribution as the histogram is relative symmetric, with the peak in the value of 0 log fold change. While in the histogram for genes that are only differentially expressed, it can be seen that the peak has been removed, which is as expected as genes with a log 2 fold change around 0 can be considered as insignificant.

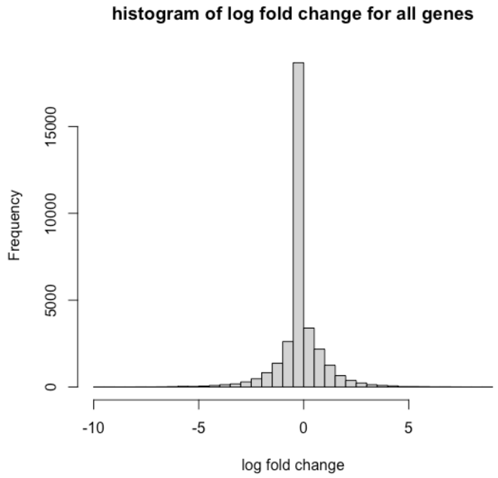

Figure1. Histogram of the log2.foldchange for all genes.

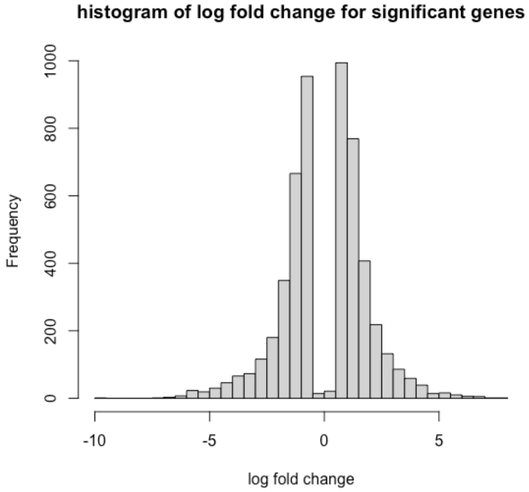

Figure2. Histogram of the log2.foldchange for only significantly differential expressed genes.

We subset the dataset that contains genes that are significantly differential expressed, and there are 5427 genes left after selecting. From then on, we write out the up-regulated genes with a positive value of log2 fold change and down-regulated genes with a negative value of log2 fold change. There are 2597 down-regulated genes and 2830 up-regulated genes.

We then upload the up- and down- regulated gene list to DAVID for gene set enrichment analysis. Mus Musculus is used as the identifier, and we choose GOTERM_BP_FAT, GOTERM_MF_FAT, and GOTERM_CC_FAT as the gene ontology background. There are 805 clusters found for up-regulated gene, and the top cluster has an overall enrichment score of 37.28. The higher the enrichment score, the more enriched. There are 706 clusters found for down-regulated gene, and the top cluster has an overall enrichment score of 37.38. Tables summarizing the top 5 cluster results from DAVID functional annotation clustering are shown below.

Annotation Cluster	Term	Enrichment Score	Count	P-value
Cluster1	Mitochondrion*	37.28	554	7.0E-92
Cluster2	Organophosphate metabolic process	34.82	282	1.5E-47
Cluster3	Carboxylic acid metabolic process	28.72	243	9.0E-36
Cluster4	Monocarboxylic acid metabolic process	18.37	166	1.0E-23
Cluster5	Intracellular signal transduction*	13.73	458	1.4E-25

Table2. DAVID functional annotation clustering top 5 clusters for upregulated genes

Annotation Cluster	Term	Enrichment Score	Count	P-value
Cluster1	Chromosome Organization*	37.38	320	3.7E-56
Cluster2	Nucleic acid binding*	37.35	627	2.4E-41
Cluster3	Regulation of nucleobase-containing compound metabolic process*	28.65	693	1.5E-37
Cluster4	Regulation of cellular component organization	23.88	507	6.6E-37
Cluster5	Chromosome organization*	23.68	320	3.7E-56

Table3. DAVID functional annotation clustering top 5 clusters for down-regulated genes

Discussions

Comparing the number of differentially expressed genes agree with O'Meara, et al, there are 2409 up-regulated and 7570 down-regulated genes discovered in their research, while we find 2597 down-regulated genes and 2830 up-regulated genes. Although our results do not align with the original paper, this should not be a concern as we only compared two of the sample groups and there are inconsistencies existing in the sample size. We are able to validate the DAVID functional annotation clustering results from the original paper, as most of the gene enrichment terms reported from our analysis are similar to those in O'Meara, et al.

Project 2 Section 7

Introduction

The authors compared the FPKM (Fragments Per Kilobase of exon per Million fragments) values of representative sarcomere, mitochondrial, and cell cycle genes significantly differentially expressed during in vitro differentiation and in vivo maturation at different time points (postanal day 0, 4, and 7 and by isoflurane overdose at 8 to 10 weeks of age) to draw biological interpretation of the experiment. We will replicate this step, compare and explore the content of the FPKM expression matrices for biological patterns.

Methods

We use the provided FPKM files of all samples, including Ad_1, Ad_2, P0_2, P4_1, P4_2, P7_1, and P7_2, except for P0_1, for which we use the output from our own test. By selecting the gene name, tracking id, FPKM from each of the sample file, we merge all 8 samples together into one data frame by gene name, and there are 34079 genes in total after merging. Then we take the mean FPKM for each related sample group as the final FPKM values that will be used in the line charts. We select Pdlim5, Pygm, Myoz2, Des, Csrp3, Tcap, and Cryab as sample genes of sarcomere. Prdx3, Acat1, Echs1, Slc25a11, and Phyh are sample genes of mitochondrial, and Cdc7, E2f8, Cdk7, Cdc26, Cdc6, Cdc27, E2f1, and Cdc45 are cell cycle sample genes.

We then subset the dataset and created a FPKM matrix of all 8 samples by selecting only the FPKM columns from each of the tracking tables into a single data frame. Also, we select genes that are only significantly differentially expressed between P0 and Ad from the gene_exp.diff file in 5.4, and then order the dataset by p-values. Therefore, genes that are most significant will be in the head of the data frame, and we picked top 1000 genes. Lastly, we subset the FPKM dataset by top 1000 gene names and created the heatmap, and there are 983 genes in total that were used for the hierarchical clustering.

Results

The line charts for FPKM values of representative genes in Sarcomere, Mitochondrial, and Cell Cycle are shown in Figure 3, Figure 4, and Figure 5. By comparing the trends that we obtain from our own data with those in O'Meara, et al. paper, for sarcomere, our results are similar to the original paper, as they both share the same direction of changes in FPKM and also the magnitude of effect for the P0 and Ad samples is similar. While for mitochondria, we don't have Mpc1 in the dataset, and Prdx3 shows a decrease from P0 to P4, which is different from that in the original paper. For the cell cycle, gene Bora is missing in our dataset, while the rest of the genes share a similar pattern and magnitude of effect for the P0 and Ad.

Comparing the results that we obtain from DAVID in 6.7, we can see most top gene enrichment terms are similar to what the original paper reports. The overlapping terms has been added with an asterisk in Table 2 and Table 3, indicating the common biological processes. Enrichment terms that are found to be common in upregulated genes are "Mitochondrion" and "Intracellular signal transduction", and in down-regulated genes, the terms are "Chromosome Organization", "Nucleic acid binding", "Regulation of nucleobase-containing compound metabolic process" and "Chromosome Organization".

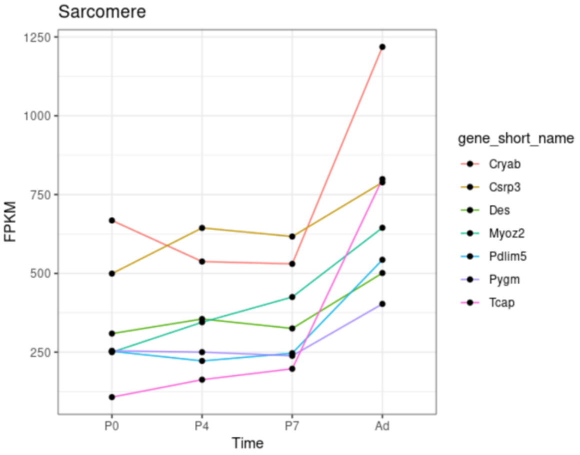

Figure 3. FPKM (Fragments Per Kilobase of exon per Million fragments) values of representative sarcomere genes significantly differentially expressed during in vitro differentiation and in vivo maturation.

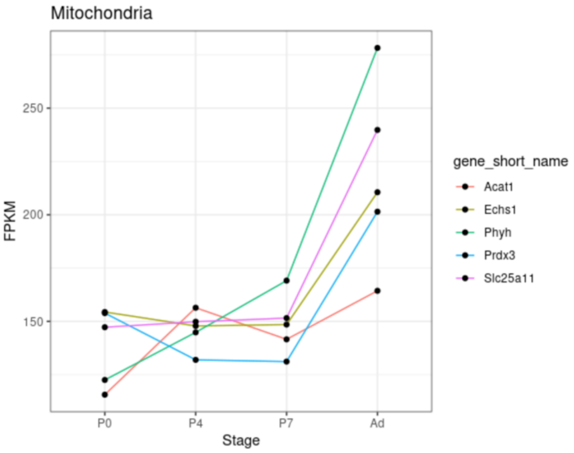

Figure 4. FPKM (Fragments Per Kilobase of exon per Million fragments) values of representative mitochondria genes significantly differentially expressed during in vitro differentiation and in vivo maturation.

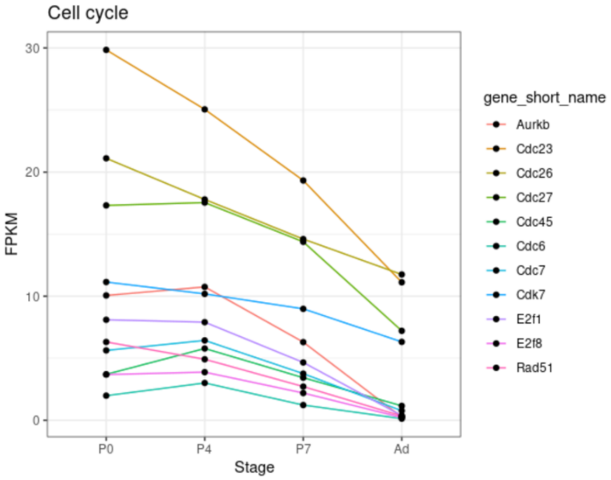

Figure 5. FPKM (Fragments Per Kilobase of exon per Million fragments) values of representative cell cycle genes significantly differentially expressed during in vitro differentiation and in vivo maturation

To compare changes in gene expression pattern of different sample groups, we perform hierarchical clustering and compare to Figure 2A in the original paper. Genes are in rows and Samples are in columns. The heatmap suggests that some genes show darker red, indicating they are differentially expressed in P0_1, P0_2, Ad_1 and Ad_2 sample groups comparing to other sample groups and there is an obvious transition of gene expression levels from P0 to Ad.

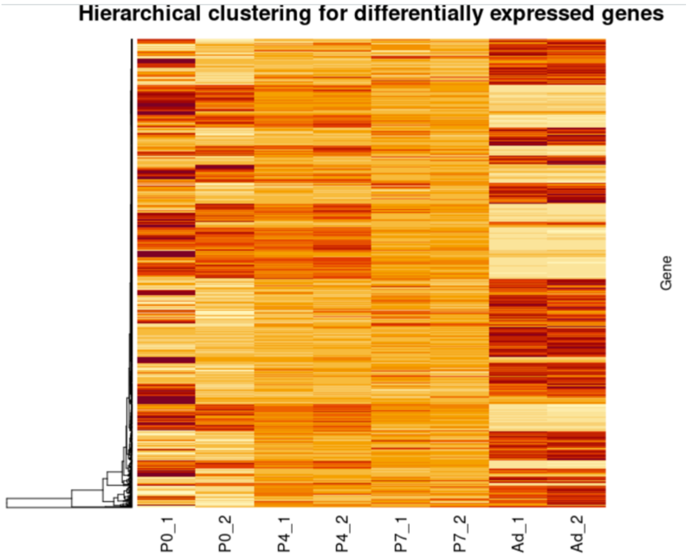

Figure 6. The clustered heatmap of FPKM values using the top 1000 differentially expressed genes found in P0 vs Ad analysis

Discussions

The line charts of differential expression of representative genes in the sarcomere, mitochondrial and cell cycle compartments show that most of the genes shares the same direction of changes in FPKM and a similar magnitude of effect for the P0 and Ad samples. Although we missed two genes, “Mpc1” and “Bora”, we are able to replicate most of the results.

Hierarchical clustering allows us to investigate the process of *in vivo* maturation of cells and there are obvious differences in the levels of gene expression among different time points. There is an obvious transition of gene expression levels from P0 to Ad, as some genes were upregulated in P0 and then downregulated in Ad, vice versa.

Overall, we are able to replicate and validate partial results from original paper. The difference in sample size is one of the main reasons, but also deep understanding of the quality control steps while cleaning the dataset, data processing and also study design of O’Meara, et al. can better help us to replicate the analysis.