**BF528 Individual Project**

Daniel Golden

Project 2: Analyst role

**Introduction**

O’Meara et al.’s 2015 paper “Transcriptional Reversion of Cardiac Myocyte Fate During Mammalian Cardiac Regeneration” is aimed at characterizing the transcriptional dynamics of cardiac myocytes in adult and neonatal mice, as well as how those dynamics may be affected by the wound healing process. The project associated with this paper involved relatively routine programming and data analysis tasks, including the use of the gene functional annotation tool of the Database for Annotation, Visualization, and Integrated Discovery (DAVID). The analyst role from this project was selected for two major reasons: a desire to refresh my familiarity with DAVID, and a belief that the relative ease and brevity of this role would counterbalance the relative challenge and complexity of my other chosen role. In short, the role of the analyst in this project was to (1) execute routine analyses of the results of differential gene expression analysis on data from the original paper using common R packages for plotting and data frame manipulation, (2) identify which genes were differentially expressed in the provided data at a significant level of confidence and which of those genes were differentially expressed to the greatest extent, and (3) generate tables of functional annotations for those genes. This role related crucially to the paper in that it constituted the key source of meaningful results. The identification and functional annotation of the genes with the expression levels that were most increased or decreased during the timeline of the experiment provided a list of genes likely to be affected by, or responsible for, the biological mechanisms underlying cardiac myocyte function during cardiac regeneration.

**Methods**

The instructor-provided dataset of differential gene expression analysis results between timepoints P4 and P7 was read into Rstudio as a data frame. This included 36,336 rows of results. The data frame was sorted by q-value in ascending order, and then subset to include only those rows with results deemed statistically significant by the creator of the example dataset, narrowing the number of results down to 459. Two histograms of log2 fold change values (figures 1A and 1B) for the dataset were generated: one including the log2 fold change values for all genes in the original data frame, and another including only the log2 fold change values associated with the statistically significant results in the subset data frame. The top ten most differentially expressed genes with statistical significance in the dataset were identified by generating an additional subset of the statistically significant data frame, composed of the ten rows with the highest absolute values for log2 fold change. A data table of these results was generated, including relevant statistics (table 1).

The data frame of statistically significant results was divided into two data frames, one including only the results with positive log2 fold change values, and the other including only the results with negative log2 fold change values. The results in these data frames represented, respectively, 243 putatively upregulated genes and the 216 putatively downregulated genes. The names of the genes in these data frames were extracted and written out to a pair of files, which were then uploaded for functional annotation via DAVID Functional Annotation Clustering. Data tables were generated containing the top results from these analyses (tables 2 and 3).

**Results**

As shown in figure 1, the exclusion of results that did not reach statistical significance markedly decreased the relative proportion of results with a low absolute value of log2 fold change, specifically values closest to zero. This is unsurprising, as the threshold for significance is dependent partly on sample size and partly on the degree of change in measured gene expression.

Chart

Description automatically generated

**Figure 1A.** Notably, the histogram of log2 fold change values for both statistically significant and statistically insignificant results was dominated by a single peak at a log2 fold change value around zero.

Chart, histogram

Description automatically generated

**Figure 1B.** In contrast with the results displayed in figure 1A, the histogram of log2 fold change values for significant results showed peaks at log2 fold change values of around 1.0 and -1.0 but lacked any results with log2 fold change values close to zero.

The top ten most differentially expressed genes based on the absolute value of log2 fold change included a combination of genes with both positive and negative log2 fold change values, representing, respectively, genes highly expressed or lowly expressed at timepoint P7 as compared to timepoint P4. Enriched genes at timepoint P7 included Egr1, Nr4a1, Fos, Hmcn2, Apba3, and Stub1. Genes expressed more highly at timepoint P4 included Rabgttb, Cdh3, Chpf2, and Eif2s3y. Also noteworthy were the differences in the scale of measured expression from gene to gene. Independent of log2 fold change values, certain genes had higher overall expression in both timepoints than others. Cdh3, for example, has only a small fraction of the overall measured expression of Stub1 at both timepoints.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Gene Name** | **P4 Expression** | **P7 Expression** | **Log2 Fold Change** | **p-value** | **q-value** |
| Rabggtb | 285.758 | 55.4545 | -2.36542 | 5.0E-5 | 0.0033028 |
| Cdh3 | 1.52594 | 0.284666 | -2.42235 | 5.0E-5 | 0.0033028 |
| Egr1 | 4.93748 | 28.2367 | 2.51573 | 5.0E-5 | 0.0033028 |
| Nr4a1 | 2.69619 | 16.7728 | 2.63713 | 5.0E-5 | 0.0033028 |
| Fos | 1.13541 | 7.78304 | 2.77712 | 5.0E-5 | 0.0033028 |
| Hmcn2 | 0.544197 | 4.04103 | 2.89252 | 5.0E-5 | 0.0033028 |
| Apba3 | 14.8716 | 275.216 | 4.20993 | 5.0E-5 | 0.0033028 |
| Chpf2 | 85.5794 | 4.56199 | -4.22953 | 5.0E-5 | 0.0033028 |
| Stub1 | 66.7248 | 1818.67 | 4.76852 | 5.0E-5 | 0.0033028 |
| Eif2s3y | 14.4372 | 0.137104 | -6.71838 | 0.00015 | 0.0081068 |

**Table 1.** Six genes (Egr1, Nr4a1, Fos, Hmcn2, Apba3, and Stub1) had positive log2 fold change values among the top ten, while four genes (Rabggtb, Cdh3, Chpf2, and Eif2s3y) had negative values in this group.

With the use of the Functional Annotation Clustering tool associated with DAVID, tables of results were generated for the 243 upregulated and 216 downregulated genes identified from the instructor-provided dataset. As shown below in table 2, the top four clusters of results included functional annotations associating upregulated genes with cellular responses to various chemical stimuli, including lipids and various organic compounds, among others. Likewise, as shown in table 3, the results from the most enriched cluster for the downregulated genes exhibited functional annotations for cytoskeletal structures, mostly microtubules and tubulin, the primary molecular component of microtubules.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Annotation Cluster** | **GO term** | **Functional Annotation** | **Count** | **p-value** | **Benjamini** |
| Cluster 1, Enrichment: 6.64 | GOTERM\_BP\_FAT | response to oxygen-containing compound | 35 | 1.6E-10 | 5.4E-7 |
|  | GOTERM\_BP\_FAT | cellular response to chemical stimulus | 39 | 8.8E-7 | 9.7E-4 |
|  | GOTERM\_BP\_FAT | cellular response to oxygen-containing compound | 20 | 8.4E-5 | 9.7E-3 |
| Cluster 2, Enrichment: 6.14 | GOTERM\_CC\_FAT | extracellular space | 34 | 4.7E-8 | 1.5E-5 |
|  | GOTERM\_CC\_FAT | extracellular region part | 34 | 1.2E-6 | 1.9E-4 |
|  | GOTERM\_CC\_FAT | extracellular region | 37 | 6.6E-6 | 7.0E-4 |
| Cluster 3, Enrichment: 6.03 | GOTERM\_BP\_FAT | response to oxygen-containing compound | 35 | 1.6E-10 | 5.4E-7 |
|  | GOTERM\_BP\_FAT | response to lipid | 18 | 6.5E-5 | 9.0E-3 |
|  | GOTERM\_BP\_FAT | response to organic cyclic compound | 18 | 7.6E-5 | 9.3E-3 |
| Cluster 4, Enrichment: 5.46 | GOTERM\_BP\_FAT | response to organic substance | 45 | 1.6E-8 | 2.6E-5 |
|  | GOTERM\_BP\_FAT | cellular response to chemical stimulus | 39 | 8.8E-7 | 9.7E-4 |
|  | GOTERM\_BP\_FAT | cellular response to organic substance | 34 | 1.4E-6 | 1.2E-3 |
|  | GOTERM\_BP\_FAT | response to endogenous stimulus | 24 | 4.1E-5 | 6.7E-3 |
|  | GOTERM\_BP\_FAT | cellular response to endogenous stimulus | 18 | 6.6E-4 | 3.9E-2 |

**Table 2.** Top four clusters of DAVID Functional Annotation Clustering tool results for list of upregulated genes in instructor-provided dataset.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Annotation Cluster** | **GO term** | **Functional Annotation** | **Count** | **p-value** | **Benjamini** |
| Cluster 1, Enrichment: 4.96 | GOTERM\_CC\_FAT | microtubule cytoskeleton | 28 | 3.2E-10 | 9.8E-8 |
|  | GOTERM\_BP\_FAT | microtubule cytoskeleton organization | 18 | 6.7E-10 | 1.0E-7 |
|  | GOTERM\_BP\_FAT | microtubule-based process | 18 | 6.1E-8 | 3.4E-6 |
|  | GOTERM\_CC\_FAT | microtubule | 15 | 6.5E-8 | 6.8E-6 |
|  | GOTERM\_BP\_FAT | cytoskeleton organization | 20 | 7.6E-6 | 2.7E-4 |
|  | GOTERM\_CC\_FAT | supramolecular fiber | 18 | 1.4E-5 | 5.0E-4 |
|  | GOTERM\_CC\_FAT | supramolecular polymer | 18 | 1.6E-5 | 5.0E-4 |
|  | GOTERM\_CC\_FAT | supramolecular complex | 18 | 1.6E-5 | 5.0E-4 |
|  | GOTERM\_CC\_FAT | polymeric cytoskeletal fiber | 15 | 4.5E-5 | 9.4E-4 |
|  | GOTERM\_MF\_FAT | tubulin binding | 8 | 1.3E-3 | 1.8E-2 |
|  | GOTERM\_MF\_FAT | microtubule binding | 6 | 6.4E-3 | 7.0E-2 |
|  | GOTERM\_MF\_FAT | cytoskeletal protein binding | 10 | 3.1E-2 | 2.1E-1 |
|  | GOTERM\_BP\_FAT | microtubule bundle formation | 3 | 1.1E-1 | 9.2E-1 |

**Table 3.** Top cluster of DAVID Functional Annotation Clustering tool results for list of downregulated genes in instructor-provided dataset.

**Discussion**

The goals of my analysis were to isolate statistically significant results of the differential gene expression analysis in the instructor-provided dataset, to identify the most differentially expressed genes in that dataset, and to generate functional annotations for genes associated with statistically significant positive and negative log2 fold change values using the Functional Annotation Clustering tool from DAVID. These goals were met successfully. From a dataset of 36,336 results, of which 459 were statistically significant, 243 significantly upregulated genes and 216 significantly downregulated genes were functionally annotated, and the top 10 most differentially expressed genes were identified, with relevant data presented in data table 1.

References:

1. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, <https://doi.org/10.21105/joss.01686>
2. Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2022). dplyr: A Grammar of Data Manipulation. R package version 1.0.8. https://CRAN.R-project.org/package=dplyr
3. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
4. O'Meara, C. C., Wamstad, J. A., Gladstone, R. A., Fomovsky, G. M., Butty, V. L., Shrikumar, A., Gannon, J. B., Boyer, L. A., & Lee, R. T. (2015). Transcriptional reversion of cardiac myocyte fate during mammalian cardiac regeneration. Circulation research, 116(5), 804–815. https://doi.org/10.1161/CIRCRESAHA.116.304269