Final Project BF528 - Ryan Yordanoff

Project 3 Biologist

Introduction

Microarray and high-throughput RNA-sequencing both aim to measure abundance of RNA molecules. While both technologies aim to attain the same biological etiology, the technologies are fundamentally different. The gene expression technologies require rigorous testing to merit use in academic research or clinical settings.

To investigate the difference in the technologies the Wang et al. (2014) designed studies to investigate concordance between the aforementioned technologies using 27 unique toxicology treatments from a range of known modes of action (MOA).

This analysis aims to discover enriched pathways in 3 MOA groups; AFL (DNA_Damage), MIC (orphan nuclear hormone receptors (CAR/PXR) toxicity), and PIR (peroxisome proliferator-activated receptor alpha (PPARA)). The analysis also aims to show clustering of normalized gene counts from the 3 MOA groups.

Methods

Differential expression results from each of the 3 MOA groups were mapped to gene symbols for both microarray and RNA-seq data. Genes differentially expressed that were shared were merged together. The top 1000 genes sorted by adjusted p-value, for each MOA, were utilized in enrichment analysis. Enrichment analysis was performed using DAVID Functional Annotation Tool (2021 update). For each MOA gene sets were converted to *Mus musculus* gene ID's using the DAVID gene id conversion tool. Gene sets were annotated using the functional annotation clustering tool. Using Medium stringency the top 5 GO biological processes were recorded for each MOA group.

Previously computed normalized expression data of all 3 MOA, generated by my group, were filtered using row variance ranking. The top 80th percentile of rows sorted by variance were utilized to create a heatmap clustering. Pheatmap version 1.0.12 was used with euclidean clustering distance to generate an annotated heatmap.

Results

Resulting enrichment analysis yielded mostly unique Gene Ontology (GO) term pathway enrichment. Mitochondrion was found in all MOA, while Endoplasmic reticulum was found in AFL and PIR. All other GO terms were unique to respective MOA groups (Table 1).

AFL TOP 5 GO TERMS	MIC TOP 5 GO TERMS	PIR TOP 5 GO TERMS
Cell cycle	**Ribosome biogenesis**	**Mitochondrion**
Endoplasmic reticulum	**Proteasome complex**	**Lipid metabolic process**
Mitochondrion	**RNA binding**	**Peroxisome**
Kinetochore	**Mitochondrion**	**Endoplasmic reticulum**
DNA replication	**Proteasome accessory complex**	**Fatty acid degradation**

Table 1. GO terms for each MOA generated by DAVID Enrichment Analysis

Clustered heatmap of normalized and filtered gene counts showed distinct clustering for each MOA group. Clustering shows a high level of intra-similarity between each MOA (Figure 1).

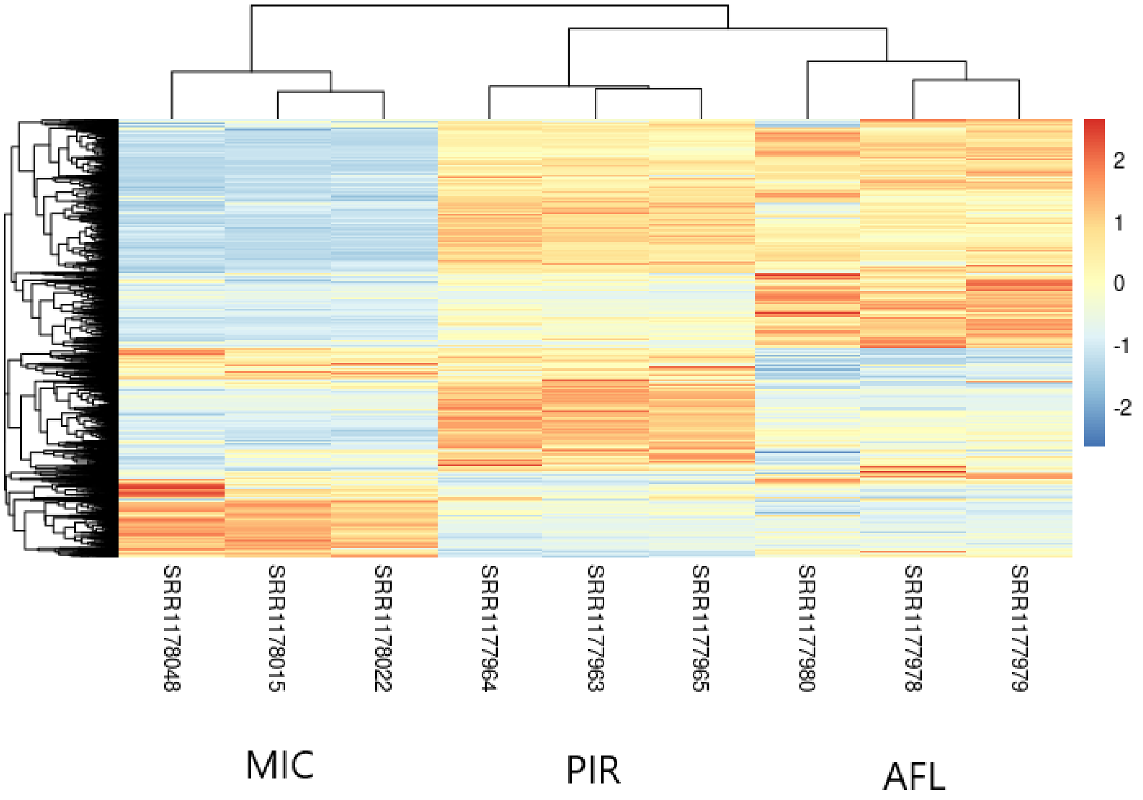


Figure 1. Heatmap clustering of 3 MOA groups from 9 Samples

Discussion

The goals of this analysis were to replicate the work by Wang et al. (2014). Comparing the output table to Wang et al. (2014) Supplementary Table 4. for MOA chemical groups that are shared by both RNA-seq and microarray I was unable to replicate the authors results. This may be due to the ambiguity of enrichment methods. My results however did show distinct enrichment patterns.

The heatmap was able to successfully cluster and resolve into 3 distinct groups using euclidean clustering. Samples showed a high level of similarity shown in the heatmap.

References

Charles Wang, Binsheng Gong et al. 2014 Comprehensive study design reveals treatment- and transcript abundance–dependent concordance between RNA-seq and microarray data. doi:10.1038/nbt.3001

Project 1 Biologist

Introduction

Marisa et al. (2013) aimed to elucidate gene expression profiles for several tumor subtypes using gene enrichment analysis. Gene expression profile discovery can give insight into the underlying unique biological process for each subtype. The goal of this analysis is to replicate the authors Fisher's exact test, by utilizing three different human pathway genesets.

Methods

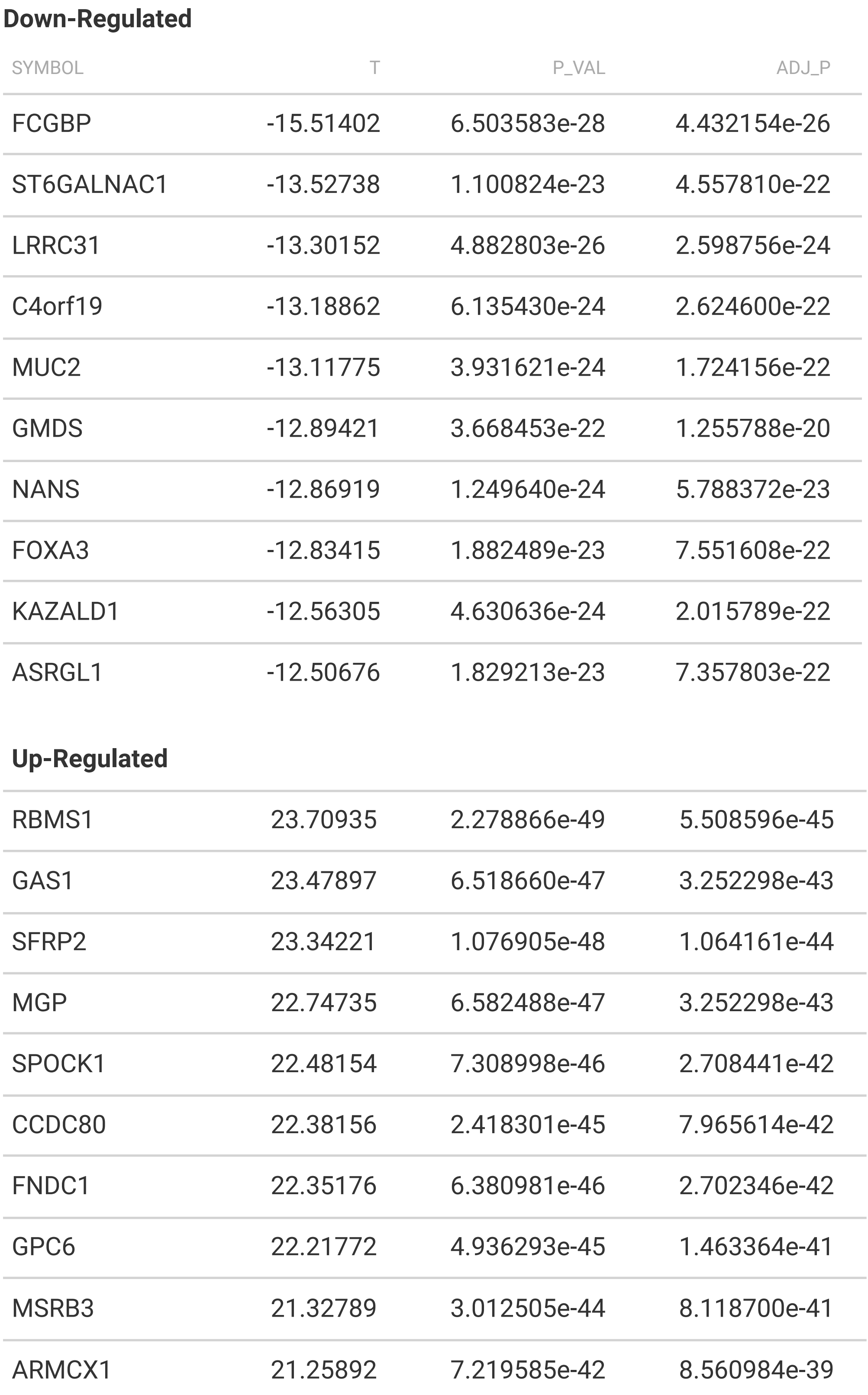
Differential expression results were pulled from previous group members data and probeset IDs were matched to gene symbols, tabulated from the R package hgu133plus2.db (version 3.13.0) using the first match in the case of duplicates. Three genesets were obtained from MsigDB: Hallmark, KEGG, and GO (version v7.5.1). The top one thousand up regulated genes and top 1000 down regulated genes were made into a table. The top ten of each were recorded.

Genesets were read in using GSEABase (1.58.0). Contingency tables were greeted with the aid of GSEABase and a Fisher's exact test was performed, using default parameters, on both up and down regulated genes. FDR was utilized to create a p-value-adjusted column for each row of the Fischer test results. The top three resulting gene sets, ranked by p-value-adjusted, were recorded for each pathway type.

Results

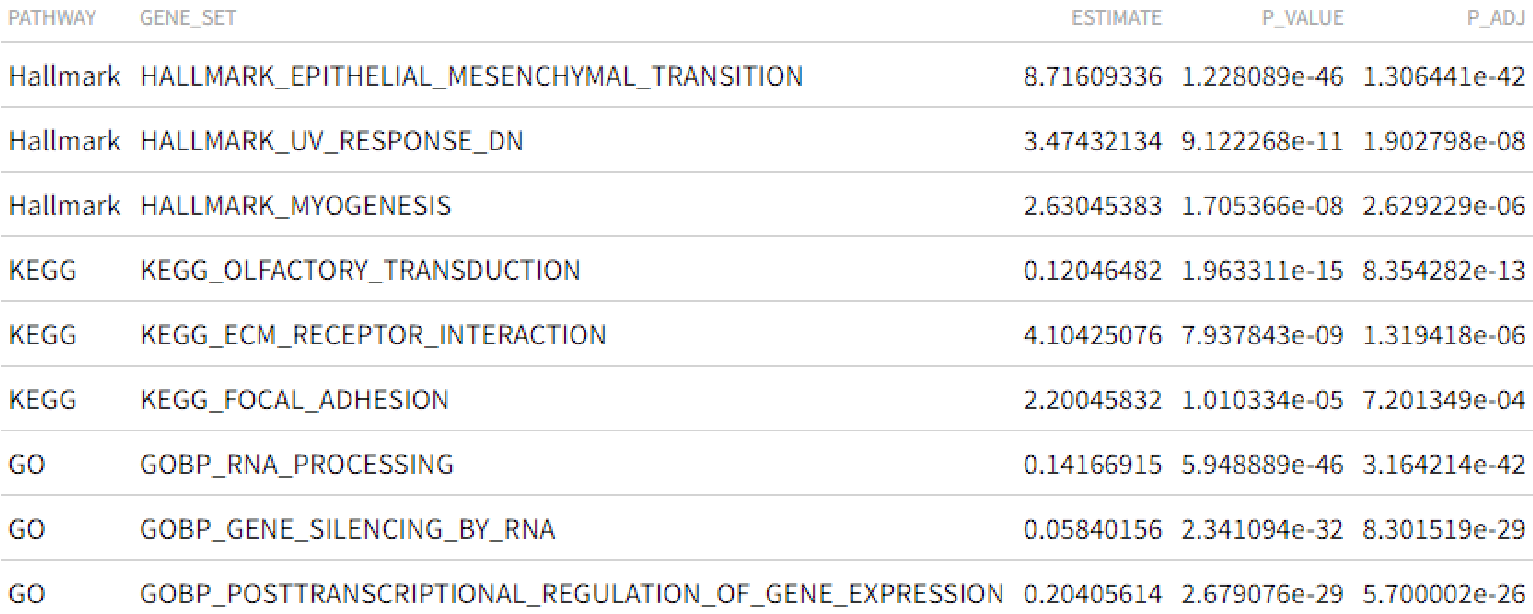
The most significantly down-regulated gene was FCGBP (t = -15.51, adjusted-pval = 6.5e-26). The most significantly up-regulated gene was RBMS1 (t = 23.71, adjusted-pval = 2.28e-45). Up-regulated genes in the top ten had higher significance and t statistic as a whole, compared to down-regulated genes (Table 1).

The hallmark pathways contained 50 genesets, the KEGG pathways contained 186 genesets, and the GO pathways contained 10,402 genesets. After performing Fisher's exact test 446 enriched genes were found to be padj < 0.05. Epithelial Mesenchymal Transition was the most significantly enriched hallmark gene set (p-adj = 1.31e-42). Olfactory Transduction was the most significantly enriched KEGG geneset (p-adj = 8.35e-13). RNA Processing was the most significantly enriched GO geneset (p-adj = 3.16e-42) (Table 2).



SYMBOL	T	P_VAL	ADJ_P
Down-Regulated			
FCGBP	-15.51402	6.503583e-28	4.432154e-26
ST6GALNAC1	-13.52738	1.100824e-23	4.557810e-22
LRRC31	-13.30152	4.882803e-26	2.598756e-24
C4orf19	-13.18862	6.135430e-24	2.624600e-22
MUC2	-13.11775	3.931621e-24	1.724156e-22
GMDS	-12.89421	3.668453e-22	1.255788e-20
NANS	-12.86919	1.249640e-24	5.788372e-23
FOXA3	-12.83415	1.882489e-23	7.551608e-22
KAZALD1	-12.56305	4.630636e-24	2.015789e-22
ASRGL1	-12.50676	1.829213e-23	7.357803e-22
Up-Regulated			
RBMS1	23.70935	2.278866e-49	5.508596e-45
GAS1	23.47897	6.518660e-47	3.252298e-43
SFRP2	23.34221	1.076905e-48	1.064161e-44
MGP	22.74735	6.582488e-47	3.252298e-43
SPOCK1	22.48154	7.308998e-46	2.708441e-42
CCDC80	22.38156	2.418301e-45	7.965614e-42
FNDC1	22.35176	6.380981e-46	2.702346e-42
GPC6	22.21772	4.936293e-45	1.463364e-41
MSRB3	21.32789	3.012505e-44	8.118700e-41
ARMCX1	21.25892	7.219585e-42	8.560984e-39

Table 1. Top 10 Up-Regulated and Down-Regulated genes



PATHWAY	GENE_SET	ESTIMATE	P_VALUE	P_ADJ
Hallmark	HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	8.71609336	1.228089e-46	1.306441e-42
Hallmark	HALLMARK_UV_RESPONSE_DN	3.47432134	9.122268e-11	1.902798e-08
Hallmark	HALLMARK_MYOGENESIS	2.63045383	1.705366e-08	2.629229e-06
KEGG	KEGG_OLFACTORY_TRANSDUCTION	0.12046482	1.963311e-15	8.354282e-13
KEGG	KEGG_ECM_RECEPTOR_INTERACTION	4.10425076	7.937843e-09	1.319418e-06
KEGG	KEGG_FOCAL_ADHESION	2.20045832	1.010334e-05	7.201349e-04
GO	GOBP_RNA_PROCESSING	0.14166915	5.948889e-46	3.164214e-42
GO	GOBP_GENE_SILENCING_BY_RNA	0.05840156	2.341094e-32	8.301519e-29
GO	GOBP_POSTTRANSCRIPTIONAL_REGULATION_OF_GENE_EXPRESSION	0.20405614	2.679076e-29	5.700002e-26

Table 2. Top Three Enriched Gene Sets for Each Pathway Obtained from MSigDB

Discussion

The resulting tables aimed to reproduce the analysis of Marisa et al. (2013). The analysis performed is similar in methodology and appeared to somewhat align with my results. From the KEGG geneset ECM Receptor Interaction was found to be in concordance with Marisa et al.'s C4 cluster. Hallmark epithelial to mesenchymal transition was found to be in the papers C2 and C4 cluster. GO RNA Processing from my analysis appears to roughly match the Immune system KEGG Antigen Processing found in cluster C2. Marisa et al did not appear to utilize results from Hallmark pathways, but concordance can be found when comparing different pathways to analogous pathway genesets.

References

Marisa L, de Reynie`s A, Duval A, Selves J, Gaub MP, et al. (2013) Gene Expression Classification of Colon Cancer into Molecular Subtypes: Characterization, Validation, and Prognostic Value. PLoS Med 10(5): e1001453. doi:10.1371/journal.pmed.1001453