Quantification of Cocaine-Induced Cell-Type Specific Gene Expression Changes

Xiaoke Zou

Li Shen, Department of Neuroscience, Mount Sinai Hospital

COLUMBIA MAILMAN SCHOOL OF PUBLIC HEALTH BIOSTATISTICS

Mount Sinai Hospital

BACKGROUND

- Next-generation sequencing (NGS), also known as high-throughput sequencing, is the catch-all term used to describe a number of different modern sequencing technologies. These technologies allow for sequencing of DNA and RNA much more quickly and cheaply than the previously used Sanger sequencing, and as such revolutionized the study of genomics and molecular biology.(1)
- Genes encode proteins and proteins dictate cell function. Therefore, the thousands of genes expressed in a particular cell determine what that cell can do. Moreover, each step in the flow of information from DNA to RNA to protein provides the cell with a potential control point for self-regulating its functions by adjusting the amount and type of proteins it manufactures.(2)

ROLE & LEARNING OBJECTIVES

- The goal of this project is to quantify cocaine-induced cell-type specific gene expression changes.
- My role was constructing code for overall analysis, including data entry, data quality control, statistical analysis and results visualization

METHODS

Quality Control

After the expression matrix was obtained, the genes that were not expressed in at least two cells were removed. Then the control features (genes), the mitochondrial genes, were defined. The total number of RNA molecules detected, the total number of unique genes detected in each sample, and the percent of mitochondrial genes in each sample were calculated.

Clustering and Determination of Cell Types

After the data was cleaned and tidy, SC3 clustering were performed to identify the cells types.(3) Silhouette Index were used to determine the K numbers of clusters. The confirmed clusters were then cross overlapped with Gokce 2016 top gene list to identify the cell types.(4)

Biological Analysis

For each cell type, the hurdle model was applied to the data to determine the log fold change between treatment cocaine and control group, and the between stimulation cocaine and saline group. The library of MAST in R were used. The cutoff of false discovery rate (FDR < 0.05) and the absolute value of log fold change (<0.06) were set to determine the significant differential list.

Rank – Rank Hypergeometric Overlap (RRHO) graphical maps

Rank–rank Hypergeometric Overlap is a threshold-free algorithm to measure the statistical significance of the number of overlapping genes, by stepping through two gene lists ranked by the degree of differential expression observed in two cell types.(5) The output is a graphical map that shows the strength, pattern and bounds of correlation between two cell type.

Gene Ontology

Gene Ontology (GO) term enrichment analysis was applied for interpreting sets of genes making use of the Gene Ontology system of classification, in which genes are assigned to a set of predefined bins depending on their functional characteristics. The differential list of each term was upload to the Database for Annotation, Visualization and Integrated Discovery (David) to determine the GO term for three cell types(6). The cutoff of raw p value (<0.1) was set to determine the significant GO terms. A chord plot of the top five significant GO terms with the related genes were made to visualize the result.(7)

STATISTICAL ANALYSIS & RESULTS

Quality Control

The original raw dataset contained 25399 genes and 13910 cells in total. After removing the genes that were not expressed in any cell, there were 15370 genes left. The cells that had a total count of RNA molecules smaller than 2000, a total number of unique genes detected smaller than 1000, and the percent of mitochondria genes larger than 15% were removed. Also, all the ribosomal protein related genes and mitochondrial genes were removed from the features. The cleaned expression matrix had 9212 genes and 512 cells.

Clustering

The SC3 algorithm were performed. A total number of five clusters were determined with a silhouette index of 0.64. Four from the five clusters can be identified with a certain cell type by the Gokce 2016 top gene list: Astrocytes, Microglia, Oligo, and Neurons.

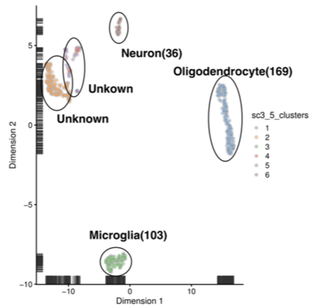

Figure 1: The UMAP map of the six clusters determined by SC3, including Microglia, Neurons, Oligo and unknown clusters. The number of cells for each cell type were also shown.

Differential Analysis

For each cell type, the hurdle model was applied to the data by using the library of MAST in R(6). The expression counts were assumed to be related with the stimulation stage, treatment stage, and their interaction. The cellular detection rate was adjusted for as a source of nuisance variation(6). The log fold changes of treatment, stimulation and interaction term were calculated for each gene of each cell type. The cutoff of false discovery rate (FDR < 0.05) and the absolute value of log fold change (>0.6) were set to determine the significant differential list.

$$\log(exp\ counts) = cngeneson + stimulation + treatment + treatment * stimulation$$

Rank – Rank Hypergeometric Overlap (RRHO) graphical maps

The rank-rank overlap analysis showed that there was strong correlation at the down - down regulated differential genes between oligodendrocyte and microglia for the stimulation and treatment term. Furthermore, for the interaction term there exist a strong consistency at the up – up regulated differential genes. There is no significant consistency showing between oligodendrocyte and microglia with neuron.

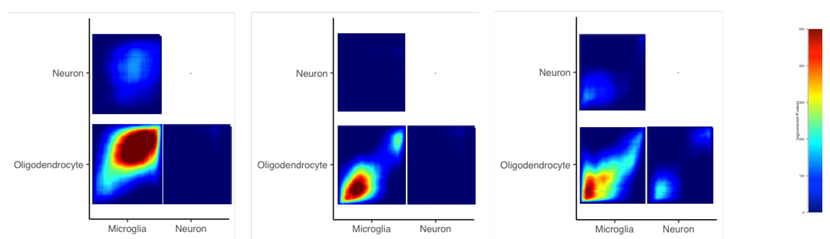

Figure 2: The RRHO plot of interaction, stimulation treatment between Microglia, Neuron, and Oligodendrocyte pair wisely.

STATISTICAL ANALYSIS & RESULTS

Gene Ontology

For Microglia, the top five gene ontology terms include: mRNA processing, regulation of proteasomal protein catabolic process, mRNA polyadenylation, RNA splicing and positive regulation. For Neuron, cerebral cortex development, hippocampus development, positive regulation of synapse assembly, RNA processing and positive regulation of proteasomal protein catabolic process. For Oligodendrocyte, GO enrichment analysis showed that nucleosome assembly, aging, cellular response to DNA damage stimulus, substantianigra development and cell–cell adhesion are the top five significant terms.

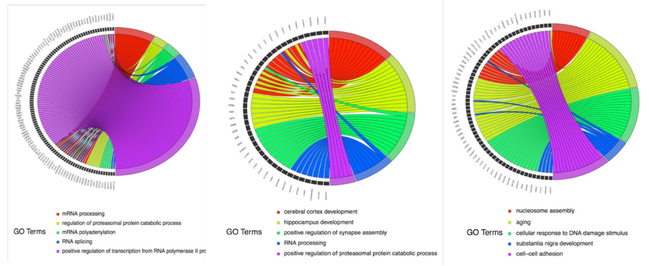

Figure 3: The top five significant GO terms of interaction term for Microglia, Neuron and Oligodendrocyte.

DISCUSSION

- The overall quality of data is not perfect. The overall effective usage of the cells and genes only accounts for 36% for genes and 4% for cells.
- The cell types determination is well-defined. Three types of cells were clearly clustered.
- The Neurons were mostly significant expressed between different cocaine induction comparing to Microglia and Oligodendrocyte.
- The downstream protein types related to significantly expressed genes were clearly shown in Gene Ontology plots.

REFERENCES

1. What is Next-Generation DNA Sequencing? (2019, August 2). Retrieved from https://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course/what-you-will-learn/what-next-generation-dna-
2. (n.d.). Retrieved from https://www.nature.com/scitable/topicpage/gene-expression-14121669/
3. Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., … Hemberg, M. (2017). SC3: consensus clustering of single-cell RNA-seq data. *Nature Methods*, *14*(5), 483–486. doi: 10.1038/nmeth.4236
4. Stanley, G., Gokce, O., Treutlein, B., Sudhof, T. C., & Quake, S. (2016). Cellular Taxonomy of the Mouse Striatum as Revealed by Single Cell RNA Sequencing. *Biophysical Journal*, *110*(3). doi: 10.1016/j.bpj.2015.11.1723
5. Plaisier, S. B., Taschereau, R., Wong, J. A., & Graeber, T. G. (2010). Rank–rank hypergeometric overlap: identification of statistically significant overlap between gene-expression signatures. *Nucleic Acids Research*, *38*(17). doi: 10.1093/nar/gkq636
6. Quackenbush, J., Wheeler, Chappey, C., Madden, Schuler, Tatusova, … Lempicki. (1970, January 1). DAVID: Database for Annotation, Visualization, and Integrated Discovery. Retrieved from https://genomebiology.biomedcentral.com/articles/10.1186/gb-2003-4-9-r60
7. Consortium, G. O. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, *32*(90001). doi: 10.1093/nar/gkh036