Data analysis of single cell RNA sequencing for neuropsychiatric disorders

Authors: Katharina Fijan, Darvesh Gorhe, Ju Hyun Jeon
Mentors: Professor Bin Xu & Professor Xiaofu He

Data Science Institute
COLUMBIA UNIVERSITY

Data Science Capstone Project with Professor Xu

Abstract

Biological data analysis methods have become increasingly important due to single-cell RNA sequencing (scRNAseq) technology. scRNAseq produces vast amounts of data, requiring new tools and techniques to process this data. This project seeks to automate an scRNAseq analysis pipeline and understand cell-cell communication and spatial transcriptomics to study miniature brain organoids from stem cells in different experimental conditions to uncover implications for neuropsychiatric diseases.

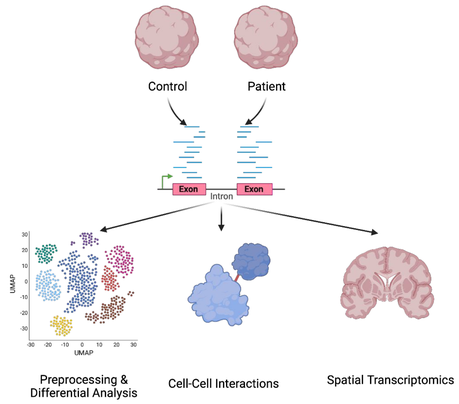

Figure 1. Overview of project and scope

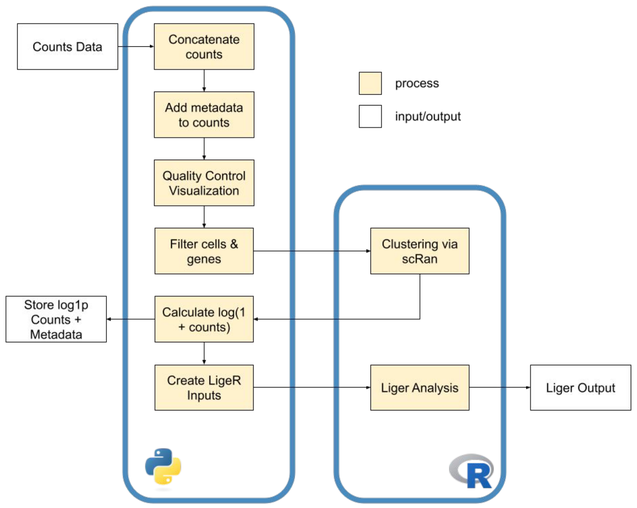

Figure 2. Preprocessing steps overview

Data Preprocessing

In the preprocessing step, we take the raw counts data from an aligner and manipulate it to our desired level of quality. The aligner takes short-read sequences and determines if it is from a known gene from a reference genome. The data is then augmented with metadata and quality control figures are generated for users to visually inspect. Finally, clustering via SCRAN and experimental conditions are compared via LIGER. The outputs are stored for further analysis downstream.

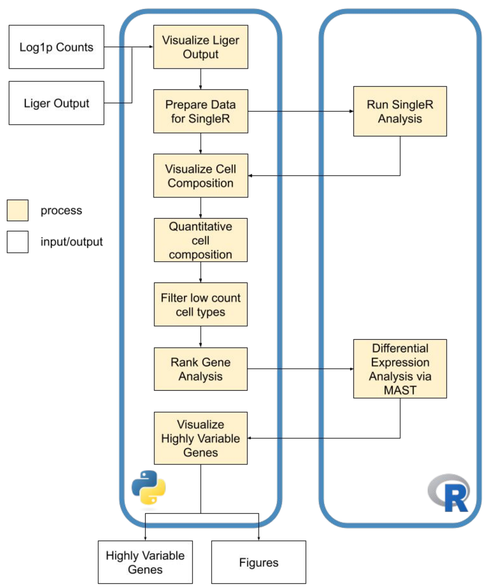

Figure 3. Main pipeline steps overview

Main Pipeline

This portion of the pipeline takes the outputs of the preprocessing step and performs further analyses to inspect the data and find genes of interest. SingleR uses known pure cell data sets to determine an unbiased cell-types. Quantitative cell composition refers to creating figures quantifying metrics based on a reference gene. Rank gene analysis is then performed to group genes via the Benjamini-Hochberg method. Highly variable genes are identified via MAST and visualized in Python after which the output and figures are stored on disk.

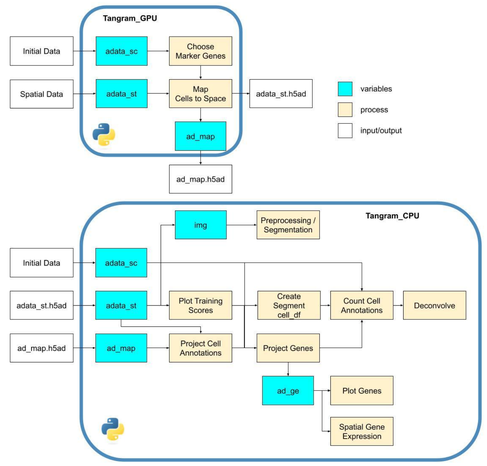

Figure 5. Spatial Transcriptomics Pipeline

Spatial Transcriptomics

Spatial transcriptomics maps single-cell data to pre-existing spatial maps. In this pipeline, we map our data to human brain data. After some preprocessing, we get spatial probability maps and spatial gene expression where we can compare control and mutated organoids. Deconvolution and Segmentation are performed to clarify our data and generate more specific images with our interest.

Cell Cell Interactions: Preliminary Results

Understanding if and how certain cell types communicate with one another can provide key insights into how an organ system functions or fails to function in the case where disease or mutation is present. Based on preliminary findings from organoid data generated by Professor Xu's lab, we see a number of differences in the cell-cell interaction counts between the control organoid and the diseased organoid, meaning the disease caused a shift in the way the cells organize and communicate with one another in the brain. Additional analysis is underway to determine significance and biological meaning.

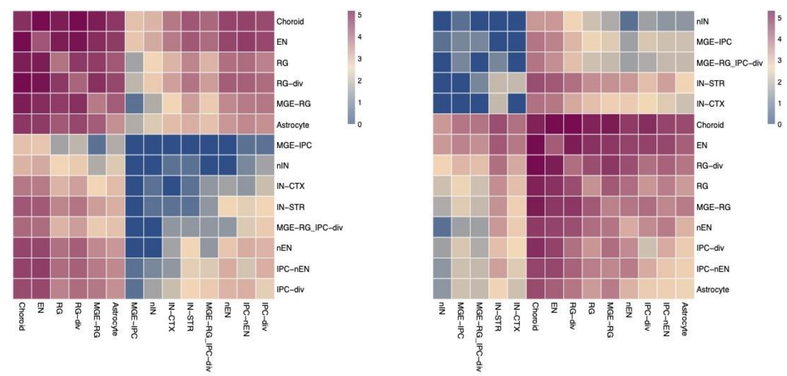

Figure 4. Control CCI log counts (left) and Mutant CCI log counts (right)

Acknowledgments

We'd like to thank Dr. Xu & Dr. He for their support and guidance during our Capstone Project. As well as Vivian Zhang and Katie Kim's assistance throughout the semester.

References

Biancalani T., Scalia G. et al. - Deep learning and alignment of spatially-resolved whole transcriptomes of single cells in the mouse brain with Tangram Nature Methods 18, 1352–1362 (2021).

Efremova, M., Vento-Tormo, M., Teichmann, S.A. et al. CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. Nat Protoc 15, 1484–1506 (2020).