Using information theory to investigate Major Depressive Disorder progress report.

By: Nathan Richman

Proposal Updates

My original proposal addressed primarily developmental changes in Major Depressive Disorder associated genes in healthy brains. I hypothesized that these genes may be controlled by a larger factor, such as a pioneer transcription factor, and if so, this cluster of genes could show a telltale change in healthy individuals around the time MDD would show up in a diseased individual. Thus my project would only investigate healthy subjects. However, now I have expanded to healthy and diseased subjects. The bulk of the project will be similar, however I now have new data and have adjusted my pipeline.

I will use data from two studies that did post-mortem whole-transcriptome sequencing on control and diseased subjects. One study dataset (NCBI GSE101521, 59 samples) recorded subject age and performed RNA-seq on tissue from the dorsal prefrontal cortex (Brodmann area 9), while the other (GSE80655, 139 samples) recorded age and tissue. I will still use mutual information between expression level and disease condition, as well as the multivariate mutual information between expression level, age, and disease condition. I will also use hierarchical clustering to see how genes in the dataset change over age, and visualize the age progression of these genes using the diffusionmap dimensionality reduction algorithm. Finally, I will use GSEA to look at enriched gene sets in the control and MDD populations.

Data

The intersection of genes from both GSE datasets contains expression levels for 55,584 unique genes across a total of 198 samples. While both of these datasets are normalized, the variations in how the transcripts were extracted could lead to a large variation in expression values, so the two datasets will be considered and analyzed separately, using the same 55,584 unique genes that are the intersection of the two. Additionally, since the data available is for Brodmann Area 9 in the first dataset, the dorsolateral prefrontal cortex (DLPFC) region will be used in the second since it is the most similar region.

Progress

So far I have downloaded both datasets (GSE101521, GSE80655). I have created metadata files for both datasets and filtered out samples that didn't meet requirements (GSE80655 also includes biopolar and schizophrenia). GSE101521 came already normalized, but GSE80655 came in raw read counts. I downloaded R and the DESeq2 package for R and used it to generate normalized counts for GSE80655 in the same format as GSE101521 [1]. I have also downloaded GSEA from the broad institute and run it on the GSE101521 dataset [2]. I have installed a python diffusionmap package (PyDiffMap) and started using it on the first dataset. I have also tried hierarchical clustering using scipy on smaller datasets, but have not yet tried it on my large datasets [3].

Challenges

The biggest challenge so far is integrating the two different datasets. While they have both been normalized using the same ratio-of-classes normalization method implemented in DESeq2, the values for the same genes are significantly different between the two sets, possibly due to differences in extraction protocols. While I might be able to correct for this if I had raw read counts for the first dataset, I don't believe there is a good way of correcting this without running all my analyses separately. Additionally, running GSEA on the first dataset using the hallmark gene set did not give results I expected. In fact no significant gene sets were differentially upregulated in the MDD group, but 42/50 were differentially upregulated in the control group. This both makes me think I might be running the GSEA wrong, or that there could be some confounding factor such as the post-mortem interval that might lead to changing gene expression in the control samples.

Related work:

1. Ramaker et al. “Post-mortem molecular profiling of three psychiatric disorders,” *Genome Medicine*, vol. 9, no. 72, 2017.
2. Krebs et al., “Whole blood transcriptome analysis in bipolar disorder reveals strong lithium effect,” *Psychological Medicine*, pg. 1-12, 2019.
3. K. C, Vadodaria et al., “Seratonin-induced hyperactivity in SSRI-resistant major depressive disorder patient-derived neurons,” *Molecular Psychiatry*, vol. 24, no. 6, pg. 795-807, 2019.

References:

[1] M.I. Love, W. Huber, S. Anders, “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.” *Genome Biology*, vol. 15, pg. 550, 2014.
[2] A. Subramanian, et al., “Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles,” *Proc. Nat. Acad. Sci.*, vol. 102, pg. 15545-15550, 2005.
[3] T. E. Oliphant, “Python for Scientific Computing”, *Comp. Sci. Eng.*, vol. 9, pg. 10-20, 2007.