Using information theory to investigate developmental changes that can lead to depression.

By: Nathan Richman

Introduction

Major Depressive Disorder (MDD) is a mental health disorder that causes a continued feeling of helplessness, sadness, and loss of interest in doing daily activities [1]. This disorder affects millions of people nationwide and has its causes as both environmental and genetic [2]. . It is known that many genes affect MDD, and recent studies have undertaken GWAS studies to determine some of these genes [3,4]. Additionally, the Brainspan atlas of the developing human brain has identified several genes that are associated with MDD through RNA arrays [2]. While some of these genes have been identified, GWAS is limited by ungenotyped causal SNPs. Brainspan provides whole transcriptome expression levels for 42 individuals ranging from 8 weeks to 40 years old. I propose to use this data to confirm the list of genes that are associated with MDD, and identify any other genes that may be related. I will use mutual information to investigate the relationship in expression levels with age, and brain location. This project will shed light on the clustering of genes that are associated with MDD during development, as well as possibly suggest other genes which are associated with MDD. If the genes cluster together at certain ages, it could suggest there is a larger signal they are responding to, such as a pioneer transcription factor. Additionally, it would be interesting to investigate the changes in these genes in an age when MDD is most likely to present.

Resources

For this project, I will need access to the data from Brainspan, as well as a computer to develop and run the program. Brainspan provides their data for free and available online here. For the purposes of this project, I can use my own machine for development and running, as well as the server provided by the biostatistics department.

Methods

For the data, I will use “RNA-Seq Gencode v3c summarized to genes”, as well as the subset of these that Brainspan supplies for the genes associated with MDD. From this data I will calculate mutual information between expression level and age, as well as between expression level and brain location, and finally I will implement a method of calculating multivariate mutual information between expression level, age, and brain location [5]. Once I have implemented these, I will observe how genes in the full data set cluster compared to genes that Brainspan has associated with depression and comment on the clustering using Gene Ontology [6,7]. I will also visualize clustering using a stochastic clustering approach called TSNE, which has already been implemented [8]. Finally, I will highlight certain genes over time and comment on the trend and what it could mean in terms of normal depression onset age, as well as whether my analysis suggests these genes are being controlled by a broader signal.

Anticipated Results

I expect to be able to identify at least the same genes that Brainspan has listed to be associated with MDD, however this list is most likely generated from an outside source. One major limitation is that Brainspan did not collect sample metadata on whether the individual had MDD, so a major assumption I am making as that the genes associated with depression still cluster together in individuals without depression, however I believe this is a valid assumption. I hope to be able to identify other genes that may be associated with depression and to validate that they make sense through the use of

Gene Ontology. Additionally, I expect to see an interesting trend samples taken from ~18-25yo which reflect that they are the ages when MDD would most likely present.

Resources

In this project, I will use TSNE which is already implemented [8], as well as Gene Ontology, which is also implemented [6,7].

References

[1] Mayo Clinic, “Depression (major depressive disorder),” Mayo Foundation for Medical Education and Research (MFMER), 2020.

[2] Brainspan, “Developmental Transcriptome,” Allen Institute for Brain Science, 2020.

[3] N. Cai et al., “Minimal phenotyping yields GWAS hits of reduced specificity for major depression,” bioRxiv, [preprint], 2020.

[4] A. K. Malhotra, “The pharmacogenetics of depression: Enter the GWAS,” Am. J. Psychiatry, vol. 167, no. 5, pp. 493–495, 2010.

[5] T. H. Pham, T. B. Ho, Q. D. Nguyen, D. H. Tran, and V. H. Nguyen, “Multivariate mutual information measures for discovering biological networks,” IEEE Conf. Res. Innov. Vis. Futur., 2012.

[6] Ashburner et al., “Gene ontology: tool for the unifiation of biology,” Nat Genet., vol. 25, no. 1, 2000.

[7] The Gene Ontology Consortium, “The Gene Ontology Resource: 20 years and still Going strong,” Nucleic Acids Res., vol. 47, no. D1, p. D330-D338, 2019.

[8] Pedregosa et al., “Scikit-learn: Machine Learning in Python,” JMLR, p. 2825-2830, 2011.