CGS4144 Final Project Report

By: Shlok Nangia, Daniel Cox, Srikar Chandaluri

Abstract:

Propensity for certain brain disorders can in part be explained by various genetic factors. Twin studies have shown upwards of 50% heritability for major depression disorder, and the same is true of schizophrenia, while type I bipolar disorder has been shown in family studies to have a strong genetic component. Though different studies have uncovered correlations between certain brain disorders, and certain genomic regions, the exact genetic links between these are still unknown.

In this study, we endeavored through comparison with a control group to discover which gene expressions might be strongly linked with schizophrenia, bipolar disorder, and major depression. Our studies of differentially expressed genes through differential analysis, clustering, predictive machine learning, and gene enrichment analysis have shown that most of the genetic differences occur in the cellular function, cellular component, and biological processing genes in the DNA of people with these mental disorders when compared to the control group. Most notable were results related to protein binding, which provided the most significant findings in our analysis.

In addition, clustering results revealed that pinning down the top ten most differentially expressed genes doesn't show the whole picture. Instead, it appeared that an approach accounting for many more genes interacting with each other was more productive. Also, results from supervised analysis showed great variance across methods for the number of significant genes, further clouding clear selection.

We think that though this particular examination of the data has not provided the sort of positive results needed to give a firm answer to the question that we posed, it can serve as a starting point for further research that might shed light on areas that were overlooked in our analyses.

Introduction:

In this project the dataset that we used consisted of RNA-seq profiling data from different parts of post-mortem brain tissues. The parts of the brains that were profiled included the anterior cingulate cortex (AnCg), nucleus accumbens (nAcc), and the dorsolateral prefrontal cortex (DLPFC) (4), these groups are often associated with mood alterations cognition, impulse control, motivation, reward, and pleasure. This is what we explored in our data as we compared the genetic composition of different groups which include: control (no mental illness), bipolar disorder, major depression, and schizophrenia.

Using this dataset our team developed the question "What are the respective specific genetic differences between patients with bipolar disorder, schizophrenia, and major depression and the control group?" The way our group approached this question was through a variety of steps. The first thing we did was look at one section of the brain to compare each of the separate groups' expression data. The section of the brain that we decided upon was the nAcc section due to its role in regulating motivation, aversion, and reward (5) which heavily attributes to the mental illnesses that we are comparing to the control.

After selecting which part of the brain's data we want to explore the next thing we did to help answer this question was transforming the data into HUGO genes from Entrez IDs, and log scaled the data so we could perform an enrichment analysis on the expression data to see which genes were differentially expressed. We did this through creating PCA plots, volcano plots, heatmaps, gprofiler plots, and a clusterProfiler plot to see which types of genes were differentially expressed between the groups.

Next we ran clustering techniques on the log scaled expression matrix to see how each of the samples were differently grouped based on their expression data. The Final thing we did to answer this question was perform supervised machine learning algorithms, and extract the most significant genes based on the algorithms we chose which included: Random Forest, Logistic Regression, and Support Vector Machine. All of these steps that we took helped us in answering our question on what the genetic differences were of the groups.

Methods:

Differential expression:

Schizophrenia vs. Control Bipolar vs. Control Major Depression vs. Control

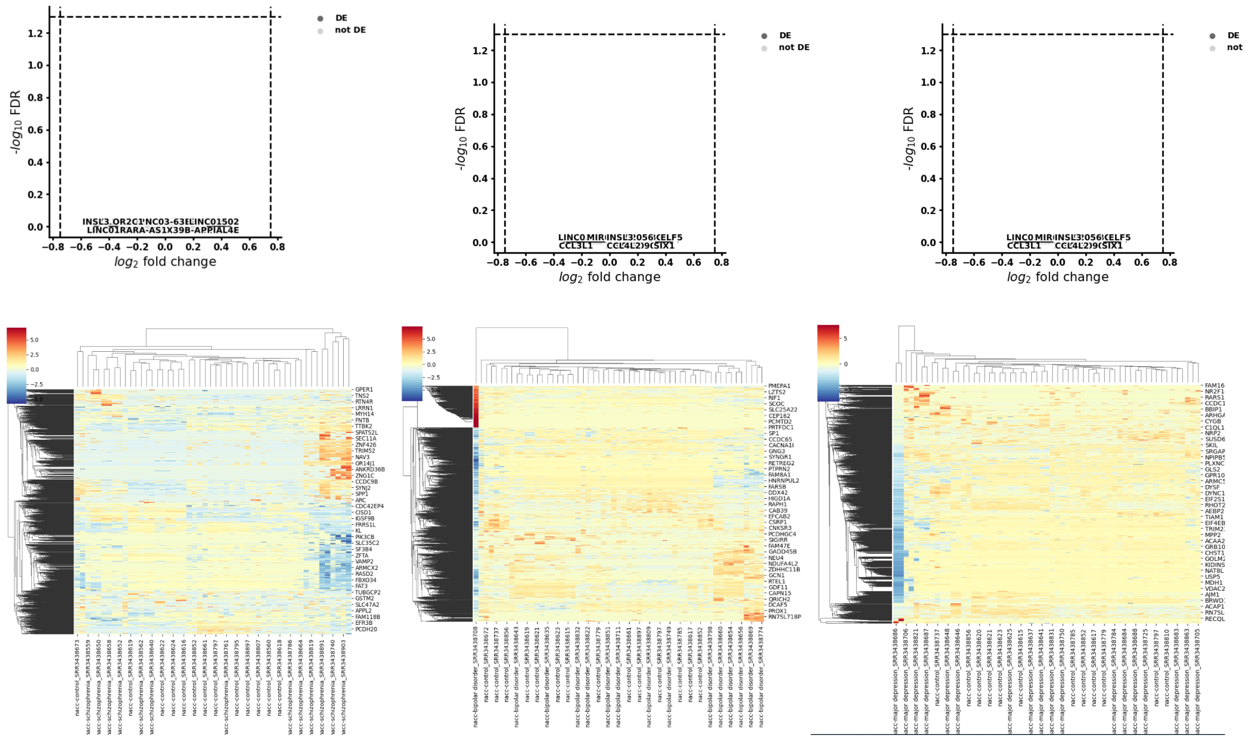


Figure 1: Volcano plots and their respective heatmaps

For analyzing the differentially expressed genes, we used a volcano plot and a heatmap to visualize the groups of genes that were differentially expressed for each condition compared to the control group. Using DESeq2, differential analysis was performed on the gene expression data and the differentially expressed genes were extracted to create the plots above. The common result from these plots is that there are no obvious significantly expressed genes that lead to their respective symptoms as seen by the emptiness of the volcano plots.

From the heatmaps in figure 1 it is clear that there may be some small groups of up regulated and down regulated genes, but overall these conditions seem to be the result of the interactions between many different genes but on a small scale that isn't noticeable. These results are also supported by PCA and t-SNE in figure 2 that show little variation in expression data. This doesn't outright prove that there isn't a genetic component to these conditions, but that it is more complex than we thought.

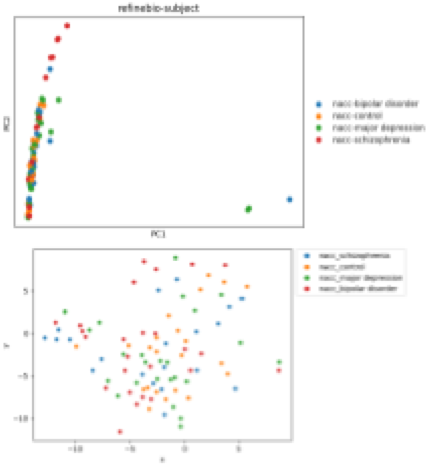
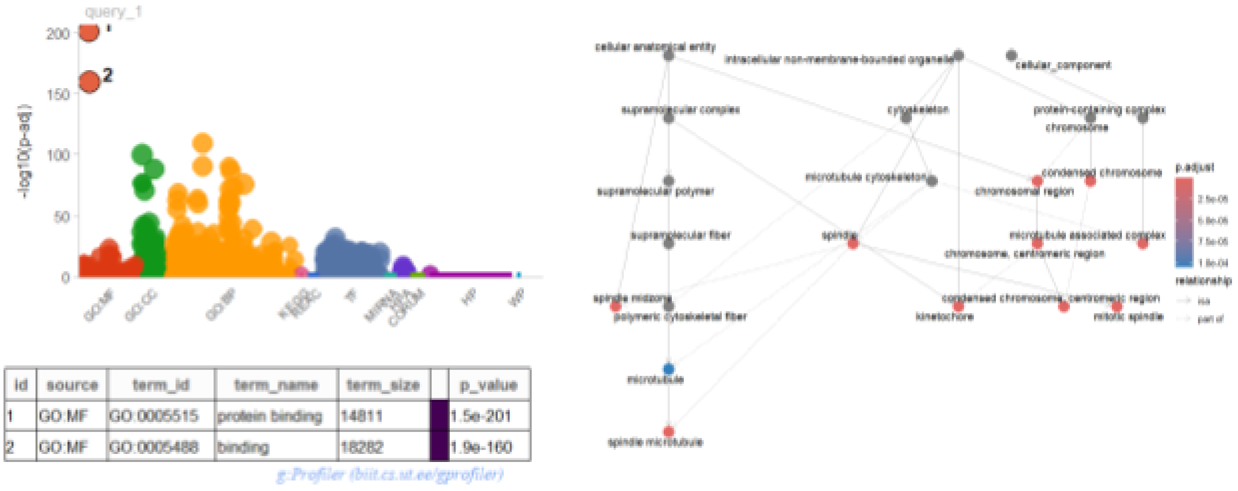


Figure 2: PCA & t-SNE plots

Gene Set Enrichment Analysis:



id	source	term_id	term_name	term_size		p_value
1	GO:MF	GO:0005515	protein binding	14811		1.5e-201
2	GO:MF	GO:0005488	binding	18282		1.9e-160

g:Profiler (biit.cs.ut.ee/gprofiler)

Figure 3: g-Profiler & clusterProfiler plot

In the enrichment analysis that we ran for the project we made a gProfiler2 plot and a clusterProfiler plot to see which genes were differentially expressed across all the groups that were provided to us. One interesting thing gained from this analysis was that the common thing that was causing problems in these groups was the chromosome, and proteins.

The control group seemed to have no problem with protein binding, and its chromosomes while the other groups in the project seemed to have that problem which could be one of the reasons why these mental illnesses occur in the first place because of problems in the cell like these in the cell which cascade and cause bigger problems in the person.

Clustering:

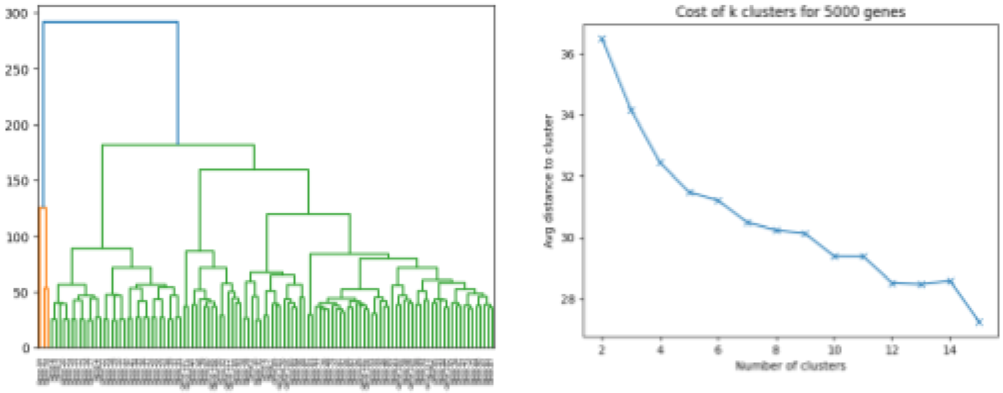


Figure 4: H-clust & k-means clustering plots

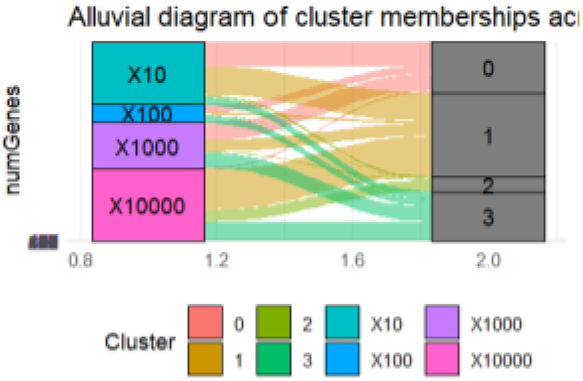
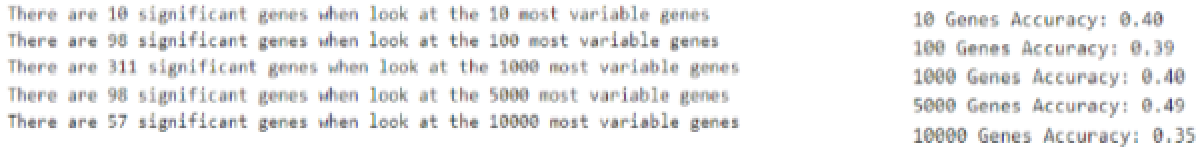


Figure 5: PAM clustering alluvial Plot

For the Clustering, we utilized k-means, hierarchical, and PAM clustering methods to group data into 4 different clusters meaning k=4. For the k-means approach, multiple values of k were tested and clusters of size 4-6 seemed to work best and even in the original data there were 4 groups, hence why k=4 was chosen. Each clustering method was tested on different sized subsets of the top most variable genes.

One thing that was common across all 3 clustering methods was that as the number of variable genes in the subset increased, cluster membership significantly changed, becoming more varied with some groups increasing or decreasing in size. This might be attributed to the fact that the factors causing these mental ailments are a result of complex relationships across many genes which was captured as more variable genes were considered for clustering, explaining why smaller subsets of variable genes had much more imbalance across clusters.

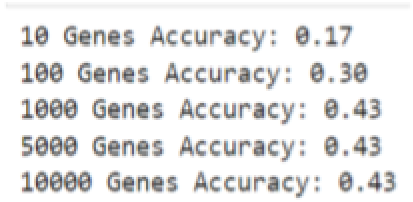
Predictive modeling:



```
There are 10 significant genes when look at the 10 most variable genes
There are 98 significant genes when look at the 100 most variable genes
There are 311 significant genes when look at the 1000 most variable genes
There are 98 significant genes when look at the 5000 most variable genes
There are 57 significant genes when look at the 10000 most variable genes
```

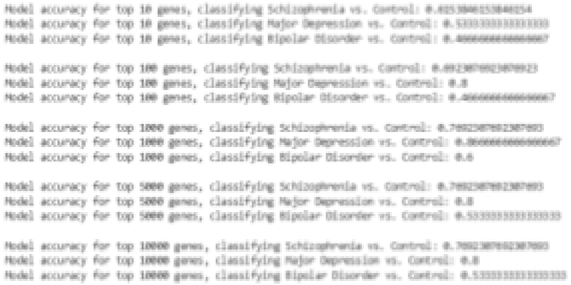
```
10 Genes Accuracy: 0.40
100 Genes Accuracy: 0.39
1000 Genes Accuracy: 0.40
5000 Genes Accuracy: 0.49
10000 Genes Accuracy: 0.35
```

Figure 6: Random Forest Data



```
10 Genes Accuracy: 0.17
100 Genes Accuracy: 0.30
1000 Genes Accuracy: 0.43
5000 Genes Accuracy: 0.43
10000 Genes Accuracy: 0.43
```

Figure 7: Logistic Regression model accuracy



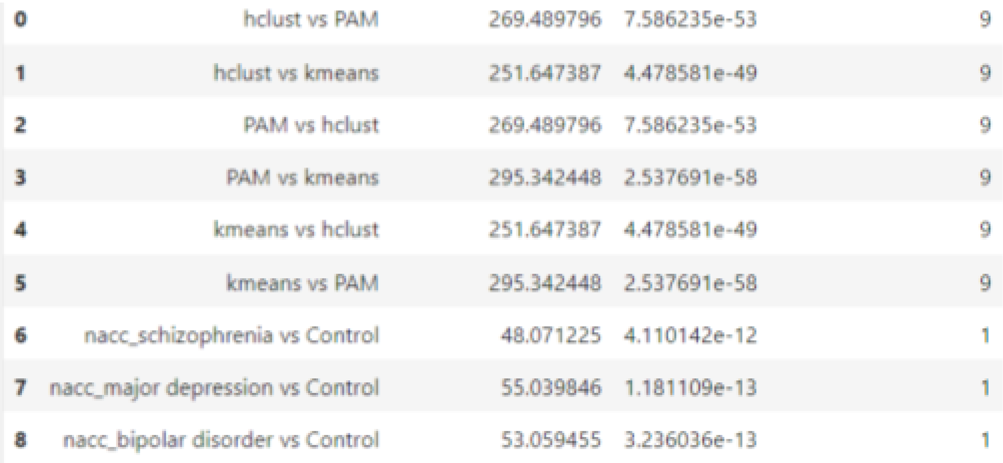
```
Model accuracy for top 10 genes, classifying Schizophrenia vs. Control: 0.6153846153846154
Model accuracy for top 10 genes, classifying Major Depression vs. Control: 0.5333333333333333
Model accuracy for top 10 genes, classifying Bipolar Disorder vs. Control: 0.4666666666666667

Model accuracy for top 100 genes, classifying Schizophrenia vs. Control: 0.6923076923076923
Model accuracy for top 100 genes, classifying Major Depression vs. Control: 0.8
Model accuracy for top 100 genes, classifying Bipolar Disorder vs. Control: 0.4666666666666667

Model accuracy for top 1000 genes, classifying Schizophrenia vs. Control: 0.7692307692307693
Model accuracy for top 1000 genes, classifying Major Depression vs. Control: 0.8666666666666667
Model accuracy for top 1000 genes, classifying Bipolar Disorder vs. Control: 0.6

Model accuracy for top 5000 genes, classifying Schizophrenia vs. Control: 0.7692307692307693
Model accuracy for top 5000 genes, classifying Major Depression vs. Control: 0.8
Model accuracy for top 5000 genes, classifying Bipolar Disorder vs. Control: 0.5333333333333333

Model accuracy for top 10000 genes, classifying Schizophrenia vs. Control: 0.7692307692307693
Model accuracy for top 10000 genes, classifying Major Depression vs. Control: 0.8
Model accuracy for top 10000 genes, classifying Bipolar Disorder vs. Control: 0.5333333333333333
```

Figure 8: SVM model accuracy

For predictive modeling, we used methods such as Support Vector Machines, Logistic Regression, and Random Forest to predict the groups of patients based on gene expression data. Each modeling method was run using different sized subsets of variable genes, measuring accuracy and significant genes across the different groups. For the SVM's since they work as a binary classifier, 3 models were used for each subset to represent all groups. A common result from all the clustering results was the more genes added to the training model the more accurate it becomes.

Furthermore all of them shared a common accuracy of around 40-50%, with the SVM being slightly higher most likely due to its nature. The significant genes belonging to each model also showed interesting results. For all the predictive methods it seemed as if there were very few genes in common across the different subsets of variable genes with the most being in common among the larger subsets of 1000, 5000, and 10000. This seems to support the idea that these conditions are a result of the interactions between many genes which are more prevalent in larger subsets, leading to a generally improved accuracy as the size of the subset increases.

Statistics:



0	hclust vs PAM	269.489796	7.586235e-53	9
1	hclust vs kmeans	251.647387	4.478581e-49	9
2	PAM vs hclust	269.489796	7.586235e-53	9
3	PAM vs kmeans	295.342448	2.537691e-58	9
4	kmeans vs hclust	251.647387	4.478581e-49	9
5	kmeans vs PAM	295.342448	2.537691e-58	9
6	nacc_schizophrenia vs Control	48.071225	4.110142e-12	1
7	nacc_major depression vs Control	55.039846	1.181109e-13	1
8	nacc_bipolar disorder vs Control	53.059455	3.236036e-13	1

Figure 9: Cluster techniques & groups Chi sq. test

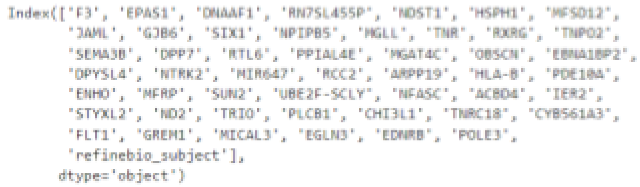
In the statistics table comparing the different clustering methods, and the groups vs. the control it is trying to determine how statistically significant the data. What we gathered from analyzing this graph is that the clustering methods reached different conclusions on each of the groups that were grouped in each of the clustering methods. Furthermore when comparing the different groups in the data compared to the control we noticed how different the two groups were based on the Adjs. p-value of comparing each of the groups

Github:

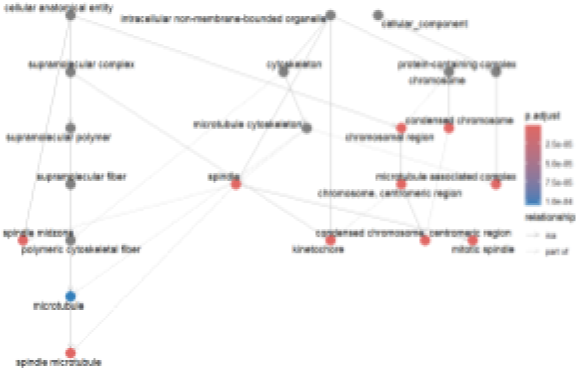
A link to all of our code is available in our github at the following line:
https://github.com/ShlokNangia24/CGS4144-Project

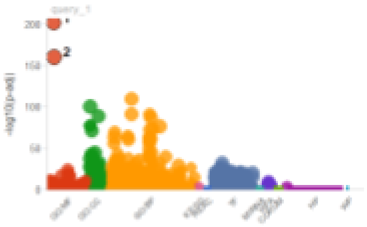
Results:

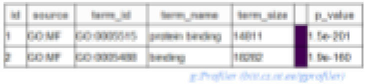
In our analysis we were able to answer the following question we asked at the beginning of the project “What are the respective specific genetic differences between patients with bipolar disorder, schizophrenia, and major depression and the control group?” This was primarily done through the analysis we ran in our predictive modeling & GSEA techniques we performed in this project. This can be seen in the plots below where we figure out which genes attribute the most to finding the difference between the four groups after performing an intersection on the important features in all three of the modeling techniques that we used which can be seen in the Fig 10. Additionally we were able to identify what these genes did in the body while doing the GSEA which can also be seen in the Fig 10.



```
Index(['F3', 'EPAS1', 'DNAAF1', 'RN7SL455P', 'NDST1', 'HSPH1', 'MFSD12',
       'JAML', 'GJB6', 'SIX1', 'NPIPB5', 'MGLL', 'TNR', 'RXRG', 'TNPO2',
       'SEMA3B', 'DPP7', 'RTL6', 'PPIAL4E', 'MGAT4C', 'OBSCN', 'EBNA1BP2',
       'DPYSL4', 'NTRK2', 'MIR647', 'RCC2', 'ARPP19', 'HLA-B', 'PDE10A',
       'ENHO', 'MFRP', 'SUN2', 'UBE2F-SCLY', 'NFASC', 'ACBD4', 'IER2',
       'STYXL2', 'ND2', 'TRIO', 'PLCB1', 'CHI3L1', 'TNRC18', 'CYB561A3',
       'FLT1', 'GREM1', 'MICAL3', 'EGLN3', 'EDNRB', 'POLE3',
       'refinebio_subject'],
      dtype='object')
```







id	source	term_id	term_name	term_size		p_value
1	GO:MF	GO:0005515	protein binding	14811	■	1.5e-201
2	GO:MF	GO:0005488	binding	18262	■	1.9e-160

g:Profiler (biit.cs.ut.ee/gprofiler)

Fig 10: significant gene expression info

Conclusion:

The main objective in this study was to discover specific genetic differences between the disease groups when compared against the controls, and in that endeavor, we met with partial success, especially when the results from the GSEA is taken into account. As for interpreting the somewhat lackluster results in other parts of the analysis, we can offer some insight to go towards explaining why that occurred as it did. When it comes to the genetic risk factors for diseases such as those in this study, the general understanding from the literature is that rather than any single or small group of genes, it is a complex of many genes, and possibly noncoding genes that were not included in the dataset we were analyzing, that leads to a higher risk of developing these diseases.

That noncoding segments of the genome could play a significant role in calculating the genetic risk is a point that aligns with our results from enrichment analysis, as it is known that those elements regulate when, where, and to what degree protein-coding genes are transcribed. This line also agrees with results taken from alluvial plots of our clustering data, where we reached similar conclusions regarding a greater proportion of cluster membership changes as we increased the subset size, viz., many genes acting in concert explaining cluster membership. Indeed, inasmuch could also be said with respect to the results from our supervised analysis, where modeling became more accurate with larger subsets too. As such, data from a whole-genome study may have been a more fertile field for this study had that option been available, since just analyzing the transcriptome as we have here paints a very unfinished picture.

Despite the successes of the GSEA, and the predictive modeling some weaknesses prevalent in our project can be seen in the Differential Expression in which the volcano plot found no significant genes. This could in part be due to the fact that we were comparing all the groups to one another and not only to the control.

This is something that we would like to improve upon with further work on this project. Instead of doing the differential analysis comparing all of the groups to one another we would like to separate the groups and perform the analysis compared only to the control, and not any of the other groups in the data. Additionally, we would improve it by perhaps choosing a larger data set as the dataset we worked with was quite small considering the number of patients in each group. One Additional analysis that we would have liked to do is seeing how accurately we could build a model that could classify a person, and their gene expression data, as this would have more of a use case in helping to diagnose patients and also figure out what features such a model uses to predict one way or another.

References

Dvorák, M., Urbánek, P., Bartůněk, P., Paces, V., Vlach, J., Pecenka, V., Arnold, L., Trávnicek, M., & Ríman, J. (1989, July 25). Transcription of the chicken myb proto-oncogene starts within a CPG island. Nucleic Acids Research. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC318186/

Levinson, D. F., & Nichols, W. E. (n.d.). Major Depression and Genetics. Stanford Medicine. Retrieved from https://med.stanford.edu/depressiongenetics/mddandgenes.html

NHS. (2023, April 13). Causes - Schizophrenia. Retrieved from https://www.nhs.uk/mental-health/conditions/schizophrenia/causes/

Ramaker, R. C., Bowling, K. M., Lasseigne, B. N., Hagenauer, M. H., Hardigan, A. A., Davis, N. S., ... & Myers, R. M. (2017). Post-mortem molecular profiling of three psychiatric disorders. Genome Medicine, 9, Article 72. https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-017-0458-5

Salgado, S., & Kaplitt, M. G. (2015). The Nucleus Accumbens: A Comprehensive Review. Neuroscience & Biobehavioral Reviews, 93(2), 75-93. https://doi.org/10.1159/000368279

Escamilla, M. A., & Zavala, J. M. (2008). Genetics of bipolar disorder. *Dialogues in clinical neuroscience*, *10*(2), 141–152. https://doi.org/10.31887/DCNS.2008.10.2/maescamilla