Genomic Analysis on Schizophrenia

By Richard Clark, Leon Kwan, Doron Lisiansky

## Abstract

Schizophrenia is a brain disorder for which there is presently no cure. However, research is being conducted to discover the origins of schizophrenia, which are based on genetics, environmental factors, and changed brain chemistry. In terms of genetics, it's unclear which genes are linked to the development of schizophrenia. Unsupervised clustering algorithms based on genetic information are used to differentiate people with schizophrenia in this paper. Using gene ontology to cluster a sample of people with and without schizophrenia, we discovered particular genes. Using hierarchical and k means clustering, these approaches successfully differentiated between controls and schizophrenic patients. Our discoveries can be utilized to design drugs that target the genes we discovered.

## Introduction

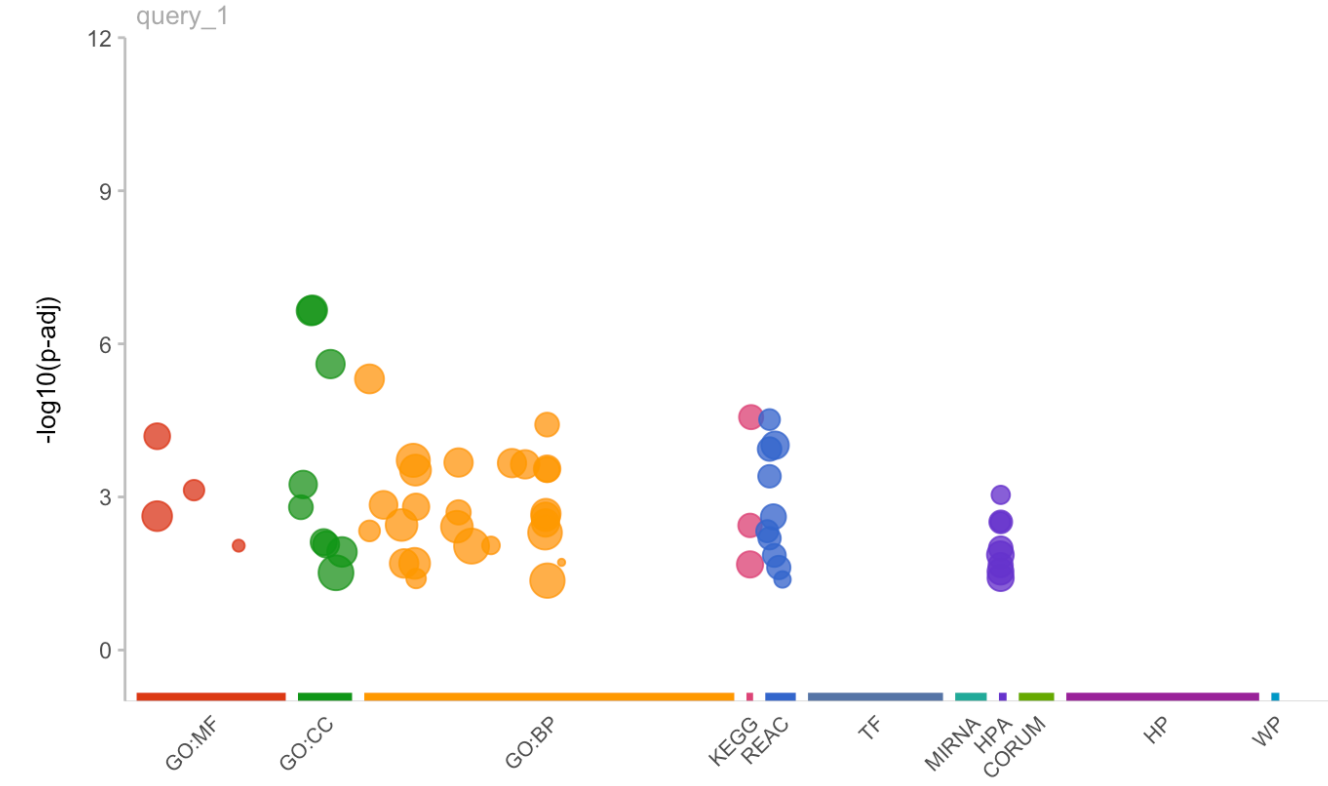
This analysis seeks to answer the question on which genes can be targeted by drugs to relieve schizophrenia. The [data](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE119290) used is from an expressed based-drug screening of neural progenitor cells from individuals with schizophrenia. In this dataset, there are 44 samples in this data set, and the expression matrix created is 26,364 x 44. With 44 samples, there are 26,364 different genes. The bulk of ranges are modest, with a few exceptions having relatively vast ranges. Although certain genes have wide expression ranges, the majority of genes are not. Using differential analysis on the samples from the schizophrenia (SZ) and control groups, a list of significantly differentially expressed genes are extracted. The genes are then run on three different enrichment analyses: gene and disease ontologies from clustProfiler, and gene ontology from gprofiler2. To show possible methods of identifying schizophrenia patients, three unsupervised clustering methods were modeled: KNN clustering, hierarchical clustering, and consensus clustering with PAM.

## Methods

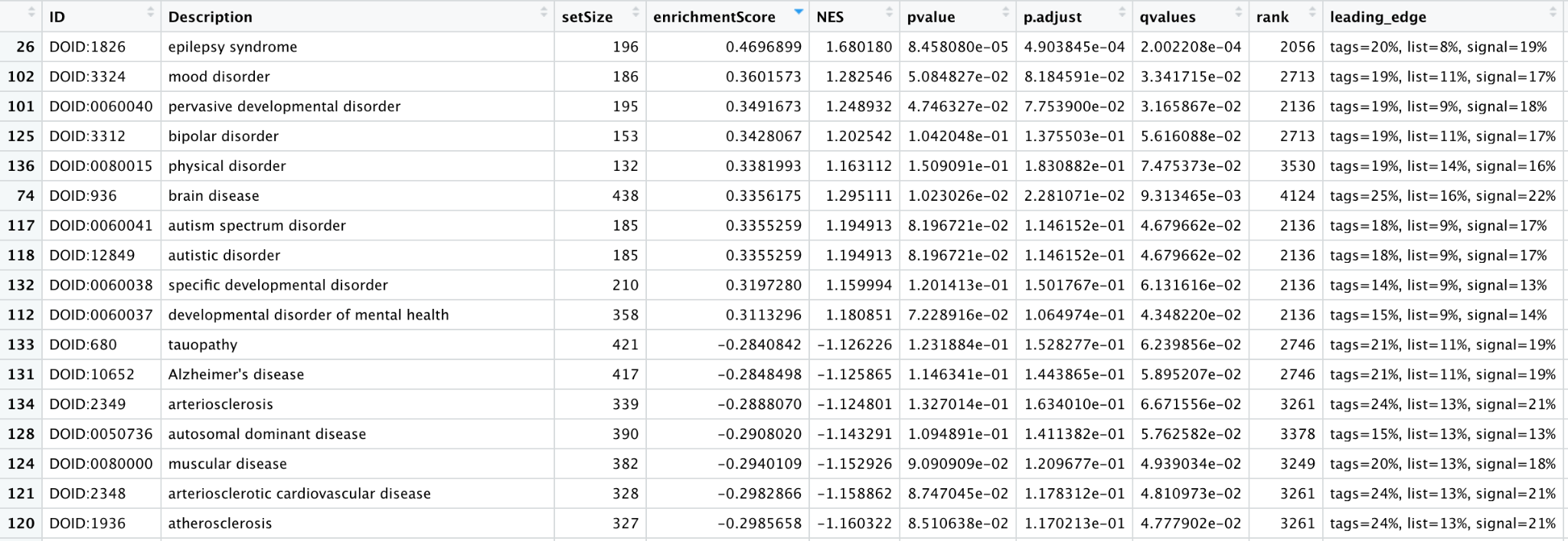
All code and analysis is for the next methods available on [github](https://github.com/r-clarke/bioinformatics).

### Enrichment analysis:

To add enrichment analysis, the first method was gprofiler paired with gene ontology. The gost tool was used to do this enrichment analysis after extracting the top 100 differentially expressed genes. The function allows gene lists to be functionally profiled. To discover overrepresentation of functions from Gene Ontology, the function uses statistical enrichment analysis. The functional terms are grouped and color-coded according to data sources on the x-axis, and they are positioned according to a fixed "source order." The order is set up such that words in the source hierarchy that are near to each other in the Manhattan plot are also close to each other in the Manhattan plot. The modified p-values are displayed on the y-axis in negative log10 scale. Every circle represents a single phrase, and the size of the circle corresponds to the term size, i.e., larger terms have larger circles.



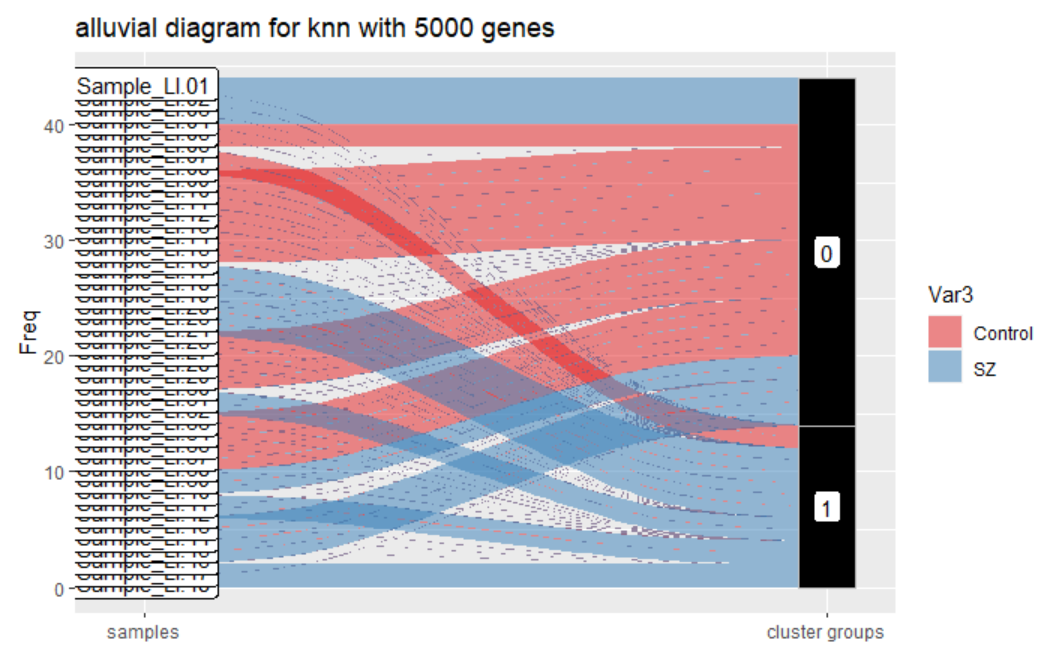
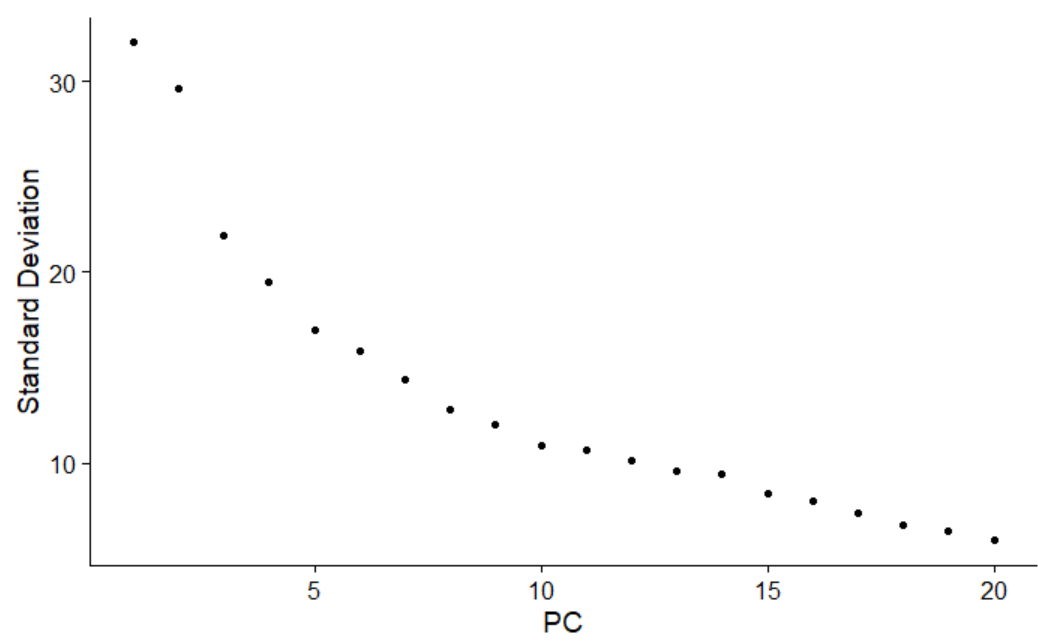
The next method and ontology used for enrichment analysis is clustProfiler and disease ontology. ClustProfiler uses methods to analyze and visualize functional profiles of genomic coordinates, gene and gene clusters. After extracting a list of differentially expressed genes from the data being used, clustProfiler was used to run enrichment analysis from the Disease Ontology. The table outputted below did not contain any specific indication of schizophrenia. When ordering the table based on p-value, some related disorders to schizophrenia include dementia and bipolar disorder which are psychological disorders, however not schizophrenia specifically. There can be many reasons for this lack of findings, however the most possible might be that schizophrenia is not in the database used in the disease ontology for this enrichment.



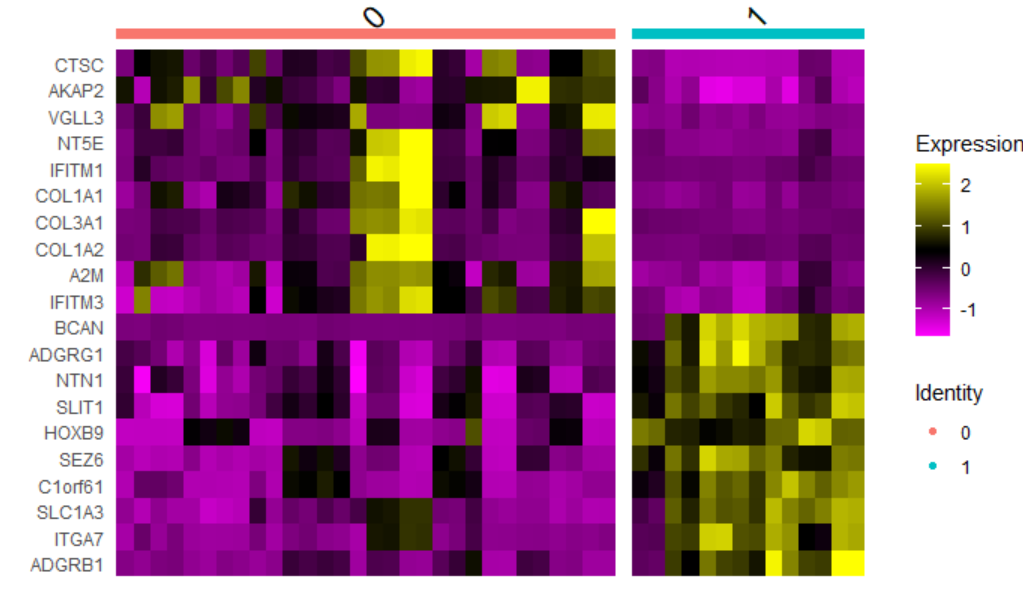
The last method and ontology pair we used continued clustProfiler as our method with gene ontology. As expected, the gene sets from the gene ontology are clearly related to schizophrenia, a neurodevelopmental disorder. The genes with the highest enrichment scores all relate directly with significant differences with genes for the synapse, an integral part of neuron interaction for brain function. For example, a decrease in postsynaptic density, listed fourth on the table, dendritic spine density, also listed in the table, has been linked to SZ in postmortem brain tissue.

### Unsupervised analysis/clustering algorithms:

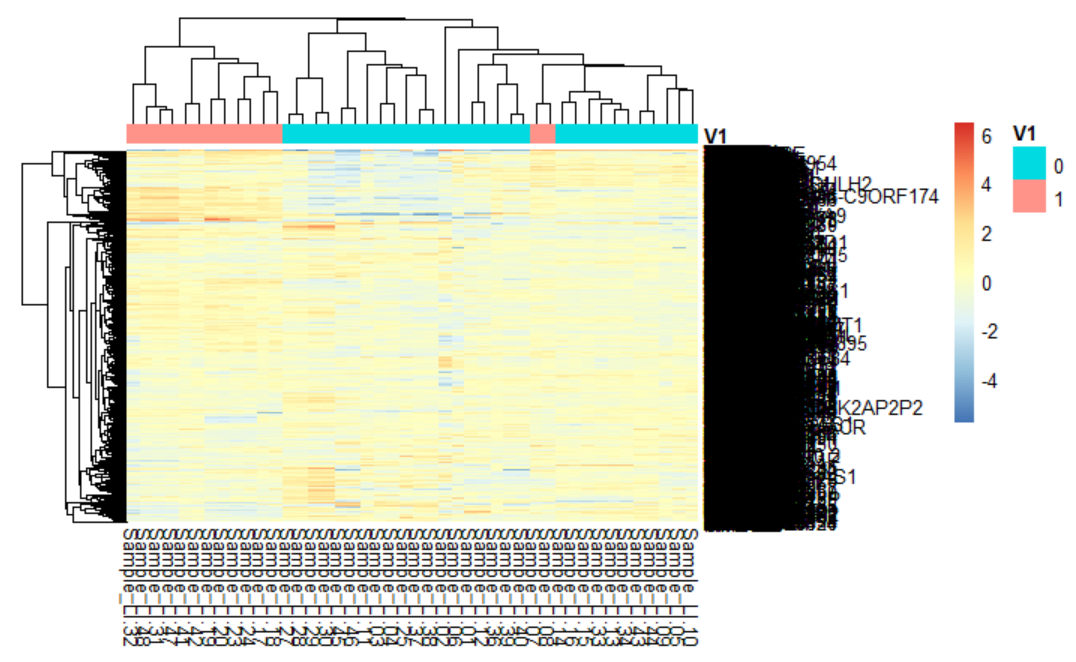
The first method used was KNN. To address the substantial technical noise in any single feature for scRNA-seq data, we employed the Seurat package, which groups the cells based on their PCA scores. The dataset's top principal components show a strong compression. We used the Elbow plot approach to determine how many components to include: a ranking of principal components based on the percentage of variation explained by each one. We are not picking k in the knn method. Changing the resolution option changes the number of clusters; according to what we've read, resolutions of 0.4-1.2 produce decent results for single-cell datasets of roughly 3K cells. We ran knn with the top 5000 variable genes, and the alluvial plot yielded two groups. In comparison to the genuine groupings, we can observe that the clustering was excellent (sz, control)

The Elbow plot for top 5000 variable genes: 

Using the Seurat::DoHeatmap method, a heatmap plot was created. The genes can also be clustered in this manner. Because clustering is only meaningful for genes that truly convey a signal, only a subset of the most highly variable genes is generally clustered. For the sake of illustration, we've chosen the 20 genes with the most variation among samples. Colored bars at the top of the heatmap indicate treatment status and cell line information.

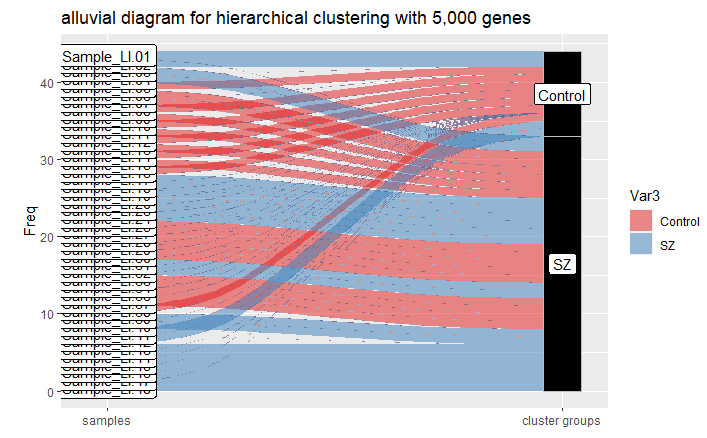


We construct the distance matrix and grouping in this figure. The dendrogram on the side of the sample distance heatmap gives us a hierarchical grouping of the samples, and we are picking all 5000 genes with the largest variation across samples. Colored bars at the top of the heatmap indicate treatment status and cell line information.

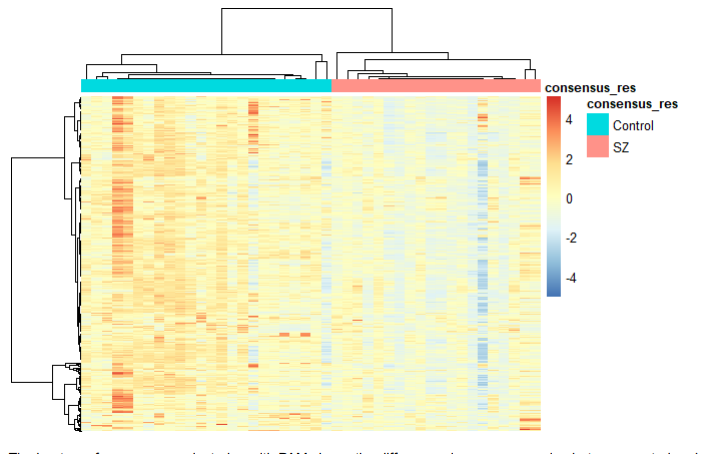


After running the knn clustering technique on 5000 genes and creating heatmaps, we ran the same procedure on 10/100/1000/10000 genes to compare the results. When comparing the results of running the clustering approach with 10, 100, 1000, and ten thousand genes, all of the findings indicate two-thirds clusters. Even though each number of genes returns the same number of clusters, the results are slightly different, as can be seen in the alluvial plots.

The next clustering method used is hierarchical clustering. To use hierarchical clustering, we first selected the top 10, 100, 1000, 5000, and 10000 most variable genes from the data to be clustered. Hierarchical clustering requires a matrix that stores the distances that each of the genes have from each other, thus we first scaled the most variable genes so that they would reflect accurate data. After scaling the variable genes, we used a distance function that creates a distance matrix required for hierarchical clustering. Now with all the parameters to run the hierarchical clustering method, we simply plug in the data variables and clustering settings into the hclust function and plot the result to show the created dendrogram. To show actual clustering with a heatmap, we had to choose a k value (the number of clusters). To do this, we started with 1 and incremented the k value by 1 until an accurate clustering had been formed, resulting in a k value of 2. After selecting a k value, we enter all the parameters into the cutree function and use the output to create the alluvial diagram and clustered dendrograms for the heatmaps.



For our final clustering method, we used Consensus Clustering with PAM for the 5000 most variable genes first identified clustering into two had the highest stability. This finding is consistent with the labeled data. The resulting heat map shows how the patients with SZ were identified and separated from the control.



When clustering by 100, 1000, 10000 highest variable genes, there was little change in accuracy between the clustering. This suggests that this model is unreliable for identifying patients with SZ.

## Results

Using the information collected from running enrichment analysis and clustering methods, we are able to aggregate all of the results to determine if the outcome is significant. Results from the enrichment analysis show that using clustProfiler, some of the identified genes are consistent with the experimental data and other literature regarding postsynaptic density and dendritic spine density. Using clustProfiler with disease ontology did not identify any relation to schizophrenia or other brain disorders. The results from the clustering methods reveal that consensus clustering and hierarchical clustering were able to identify patients with schizophrenia with no statistical significance compared to the labeled data, while KNN was not able to discriminate between the two groups accurately. This means that overall, the data is not statistically different from the sample label.

Looking back at the project as a whole, some weaknesses that we found include missing documentation regarding sample groups and lack of significance coming from clustProfiler with disease ontology. When trying to differentiate each sample to identify if they were in the control group or the schizophrenia group, the groups were not well documented resulting in unnecessary work being spent on identifying the groups before analysis can be done. Looking closer at the weakness regarding using clustProfiler with disease ontology, this enrichment analysis method had hopes of being able to identify any genes relating to schizophrenia but ultimately made no recognition of schizophrenia or other related brain disorders.

Regarding any bioethic issues that might arise from completing research on this topic, we have noted that there may be discrimination when identifying certain genes as associated with schizophrenia. This means that although someone might have similar genes to someone with schizophrenia, that does not necessarily mean that they also have schizophrenia. Although we considered this bioethical issue, due to the nature of the data there is not much that can be done regarding how society may discriminate against any genes that are identified to be associated with schizophrenia.

Looking forward into future work that can be done with this topic, we would like to expand the groups identified to also incorporate which drugs were taken by patients to see how their genes compare with individuals without schizophrenia. With this in mind, a new scientific question to be asked would be as follows: Using RNA-sequencing data from a drug screening of neural progenitor cells from individuals with schizophrenia and from the drugs found to be more effective against schizophrenia, what are their effects? Do they show the same treatment results for all patients, or is there a difference due to ethnicity, age, and gender?

## Conclusion

Furthermore, we were able to discover particular genes associated with schizophrenia using enrichment analysis methods. With the addition of clustering methods, these approaches also successfully differentiated between control and schizophrenic groups in the sample data. All of these analysis tools and methods ultimately help us answer our original question: Using RNA-sequencing data from a drug screening of neural progenitor cells from individuals with schizophrenia, which kind of drugs will be more effective? We discovered certain genes that might be future therapeutic targets using enrichment analysis and unsupervised clustering on people with and without schizophrenia. Genes like GPR12 and SLC6A11 were among them.

In this project, all analysis methods use groups divided by schizophrenia and control patients, but can be expanded with a division into different types of drugs. If we were to do work on this project again, we would improve our analysis and research by trying to focus more on the differences in the drugs given to patients, resulting in different results for different drug types and cells. We would use PCA and other analysis methods to find which pharmaceuticals are similar in effect to minimize the number of distinct drugs, then repeat the procedures from the assignments to undertake an extensive study on the groups we found.

### References

Amber Berdenis van Berlekom, Cita H Muflihah, Gijsje J L J Snijders, Harold D MacGillavry, Jinte Middeldorp, Elly M Hol, René S Kahn, Lot D de Witte, Synapse Pathology in Schizophrenia: A Meta-analysis of Postsynaptic Elements in Postmortem Brain Studies, Schizophrenia Bulletin, Volume 46, Issue 2, March 2020, Pages 374–386, https://doi.org/10.1093/schbul/sbz060

Consortium CaSWGotPG. Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. Nat. Genet., 49, 27–35 (2017).

Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci.

Carpenter WT, Koenig JI. The evolution of drug development in schizophrenia: past issues and future opportunities.

Ruderfer DM, et al. Polygenic overlap between schizophrenia risk and antipsychotic response: a genomic medicine approach.

Goncalves VF, et al. A hypothesis-driven association study of 28 nuclear-encoded mitochondrial genes with antipsychotic-induced weight gain in schizophrenia.

Higashi K, et al. Medication adherence in schizophrenia: factors influencing adherence and consequences of nonadherence, a systematic literature review.

Brennand KJ, et al. Modelling schizophrenia using human induced pluripotent stem cells. Nature.

Topol A, et al. Dysregulation of miRNA-9 in a subset of schizophrenia patient-derived neural progenitor cells.

Hao D, et al. MicroRNA profiling of neurons generated using induced pluripotent stem cells derived from patients with schizophrenia and schizoaffective disorder,

Hoffman GE, et al. Transcriptional signatures of schizophrenia in hiPSC-derived NPCs and neurons are concordant with post-mortem adult brains.

Fromer M, et al. Gene expression elucidates functional impact of polygenic risk for schizophrenia. Nat. Neurosci.

Du J, et al. KEGG-PATH: Kyoto encyclopedia of genes and genomes-based pathway analysis using a path analysis model.