Transcriptome Analysis of Post-Mortem Brain Tissue for Schizophrenia

Richard Qian, John Aitken, Bishoy Pramanik

Abstract

Schizophrenia is a complex psychiatric disorder known to have multiple genetic and environmental factors that contribute to its expression. To study genetic factors behind schizophrenia, we reference a previous study (Ramaker et al., 2017) that sequenced the RNA transcriptomes of different *post-mortem* brain tissues of multiple psychiatric disorders, including schizophrenia. With sequenced transcriptomes, we can analyze gene expression differences between schizophrenia and control patients. Specifically, we analyzed the post-mortem tissues from the anterior cingulate cortex. We performed a differential analysis, as well as an unsupervised and supervised classification analysis to gain insight into the complex etiology of schizophrenia.

We present a list of 33 differentially expressed genes between schizophrenia and control patients. Additionally, our clustering analyses were able to conclude that there is a statistically significant difference in gene expression in the anterior cingulate cortex, confirming the findings of the original study. These results demonstrate quantifiable differences in the neurology behind schizophrenia at the gene expression level. Past research has associated the anterior cingulate cortex with the same functions that schizophrenia affects (Weissman), indicating that a further pathway analysis with the genes involved may reveal more about the biological mechanisms involved in the phenotype expression of schizophrenic traits. Understanding these mechanisms can provide valuable insight into the psychiatric treatment of schizophrenia.

Introduction

We reference a previous study conducted by Ramaker et al., 2017 which provided gene expression data from RNA-sequenced transcriptomes of tissue from three brain regions (the anterior cingulate cortex, dorsolateral prefrontal cortex, and nucleus accumbens) of bipolar disorder, major depression, schizophrenia, and control subjects. Our initial question was more exploratory. We asked if there are any differences in gene expression between the four psychiatric disorder groups and any of the three tissue types. With a rich dataset provided, we could explore many different comparisons (e.g. male versus female bipolar disorder patients or depression vs control expression in the nucleus accumbens).

Our team ultimately decided to test and explore a main result from the original study:

> *The most significant disease-related differences were in the anterior cingulate cortex of schizophrenia samples compared to controls. (Ramaker et al., 2017, para. 3)*

With this result, we pose the following question. Can we confirm the results of the original study and find differences in transcriptomes of the anterior cingulate cortex (AnCg) between patients diagnosed with schizophrenia (SZ) and control (CTL) patients? Additionally, what other differences can we discover? Our goal is to answer whether there is a significant quantifiable difference in gene expression between the two groups of patients. If there are, we attempt to explain the mechanisms behind those differences.

It is important to observe that schizophrenia like many psychiatric disorders is very complex in nature and is not only impacted by genetics. Our study only analyzes gene expression data and accompanying metadata from the sourced dataset. This is important to note because there may be several factors that influence the development of psychiatric disorders that are not accounted for in our study. But revealing differences in gene expression could assist in understanding biological mechanisms behind schizophrenia and better treatment.

Methods

Our aim in this study is to find whether there is a difference in gene expression between the control and schizophrenia patients in sequenced anterior cingulate cortex tissue. We perform the following methods to find the answer.

All the methods below are found in scripts in our code repository on Github: https://github.com/rqian239/bioinformatics-project.

Data Filtering

Before conducting subsequent methods, the original expression data must be filtered so that only SZ and CTL of the AnCg are included. The genes of the original matrix, which are represented by Ensembl IDs, must also be mapped to HUGO IDs. Any genes that do not map to a valid HUGO ID are discarded from the resulting dataset.

The original gene expression matrix contained expression levels of 43363 genes in 335 samples. After filtering, we result in a dataset with 55 samples and 29404 HUGO genes.

PCA and UMAP

For an initial analysis of our data, we wanted to visualize our samples and discover any patterns that emerge when plotting the data. Due to the sheer amount of genes (i.e. attributes) in our matrix, we employ Principle Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP) for dimensionality reduction.

The prcomp() R function was used to conduct PCA. The umap() function from the umap package was used.

Differential Analysis

We conducted differential analysis with the DESeq2 R package to extract a list of differentially expressed genes between the SZ and CTL groups. DESeq2 was also the package used in the original study. The DESeq2 code in our project is based on a refine.bio reference found here.

In this method, we input the filtered gene expression data and accompanying metadata to label which samples are SZ and which are CTL. We define a counts cutoff of 91.67 to include genes that only have an aggregate count higher than this value. This cutoff was defined to maintain the same ratio of counts to number of samples as found in the refine.bio reference (10:6 ratio).

Differential analysis will find a list of genes considered “significantly differentially expressed” between the SZ and CTL groups of AnCg tissue. The definition for a gene that is “significantly differentially expressed” is a gene with a log2 fold change magnitude of more than 1 and an adjusted p-value of less than 0.01.

A volcano plot was created to visualize the results of differential expression on all of the genes. With our list of differentially expressed genes, we form a heatmap along with our 55 samples to visualize expression levels of each gene.

Enrichment Analysis

Now that we have a list of differentially expressed genes, we perform several enrichment analysis methods to associate those genes with biological pathways. This allows us to study differential expression results on a pathway-basis instead of a gene-basis and provides insight into biological mechanisms affected by gene expression. We perform three methods of enrichment analysis: clustProfiler, topGO, and gProfiler.

Unsupervised (Clustering) Analysis

To conduct unsupervised analysis, we take the top-5000 most variable genes and apply clustering methods to observe if SZ and CTL will cluster together. Gene expression values (counts in the matrix) are the attributes used for clustering and membership assignment. Three methods were conducted: k-means clustering, PAM clustering, and hierarchical clustering.

K-means Clustering

K-means clustering separates the 55 samples into k different clusters; each sample or observation is a member of the cluster with the nearest cluster mean. In other words, the distance between a sample and a cluster mean is minimized when deciding cluster membership.

Conducting k-means clustering involves selecting an appropriate k value (the number of clusters being created). To do so, we can create an elbow plot. Our heuristic is the Within Sum of Squares (within_ss), which measures variance within a cluster. To find an “optimal” k, we find the point where adding more clusters results in diminishing returns (i.e. smaller reduction in within_ss). Any additional clusters beyond k would also begin to overfit the data.

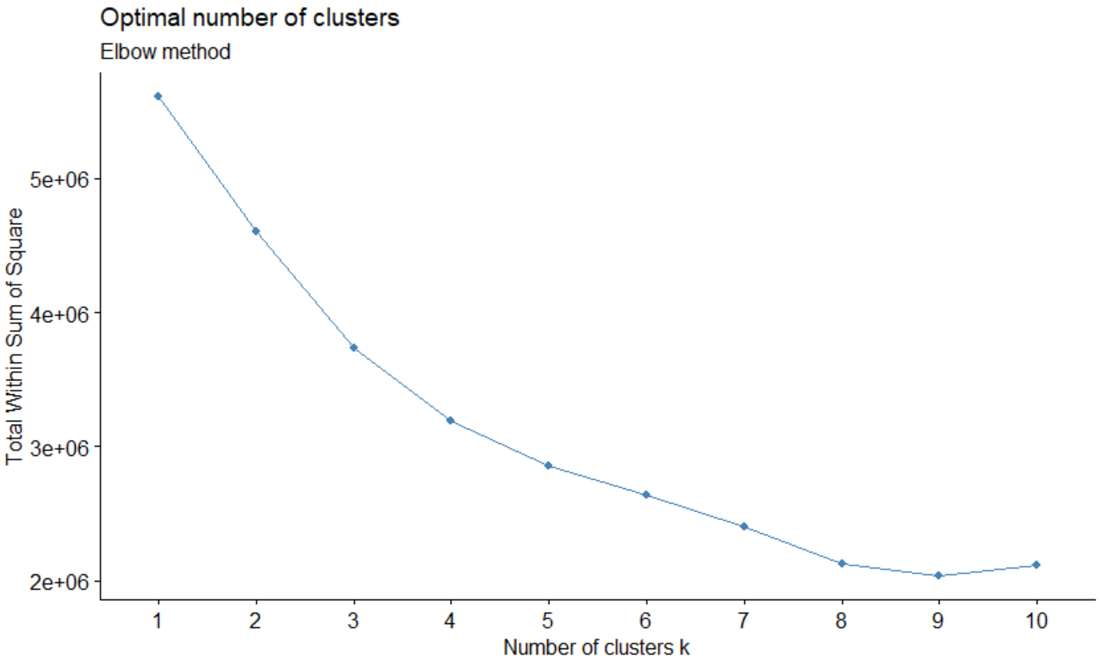


Figure 1: Elbow plot for selecting an appropriate k value for k-means clustering

Given the elbow plot, we can deduce that k=5 is a good choice. We will conduct k-means clustering on the top 5000 most variable genes and set k equal to 5.

The kmeans() R function with an nstart set to 55 was used to conduct k-means clustering.

PAM Clustering

PAM clustering was carried out using the "cluster" package's *pam()* function on the gene data. PAM or Partitioning Around Medoids separates given data into *k* clusters, using euclidean distance as a measure of similarity. In contrast to K-Means clustering, however, this method represents clusters using medoids rather than centroids. Following the methodology for our K-Means analysis, PAM clustering was carried out with *k* = 3, 5, and 7 on the top 5000, 1000, 100, and 10 most variable genes in the data set.

Hierarchical Clustering

Using the hclust() method built into R, the hierarchical clustering chooses the k-value based on its own analysis using a dissimilarity matrix. In our case, this was generated using Euclidean distance, and the clustering agglomeration method chosen was the

Ward’s Criterion algorithm. This clustering algorithm works by minimizing variance within a cluster, then pairs with nearest clusters.

Predictive Modeling (Supervised Analysis)

We applied three supervised classification techniques on the top-5000 most variable genes of our expression matrix: Support Vector Machine, K-Nearest Neighbors, and the Naive Bayes. All predictive modeling was tested following 5-fold cross validation.

Support Vector Machine

Support Vector Machines (SVMs) are commonly used supervised models for classification. In our analysis, SVM will classify samples as either SZ or CTL. We will also extract “gene signatures” or genes that are most influential in classification. Selection of the kernel function is crucial to ensure we are using the correct SVM method for our given data. Our implementation utilizes the e1071 package, a common R package for SVM. The library contains four main kernel functions: linear, polynomial, radial (RBF), and sigmoid.

For our purposes, we must extract the gene signatures of the model: the genes that influence the model most significantly when performing the classification. Out of the four main kernels, linear is the most straightforward for extracting gene signatures. Linear SVM yields a weight vector which contains the weight coefficients of all the attributes (genes in our case). Therefore, we can rank the magnitudes of these coefficients to extract gene signatures.

Before proceeding, we must show that a linear kernel is an appropriate kernel to select. Below is a table showing five SVM trials of each kernel and their AUC (Area Under the Receiver Operating Characteristic Curve) values which serves as a metric for the classification performance.

Kernel Function	Average AUC 1	Average AUC 2	Average AUC 3	Average AUC 4	Average AUC 5	Averages
Linear	0.8375	0.7544	0.8146	0.7267	0.85	0.7966
Polynomial	0.846	0.8709	0.8253	0.7887	0.8304	0.8323
Radial	0.8027	0.8029	0.792	0.8136	0.7767	0.79758
Sigmoid	0.714	0.7064	0.7733	0.73356	0.7884	0.7431

Figure 2: Table of results from five SVM trials of different kernel functions

Polynomial SVM had the highest AUC with 5-fold cross validation at around 0.83. However, linear SVM is also an acceptable method with an AUC around 0.8. This is preferred as Linear SVM allows for a straightforward extraction of gene signatures, as opposed to much more complex analysis if we were to use the other kernel functions.

For our study, we will proceed with linear SVM.

K Nearest Neighbors

K Nearest Neighbors (KNN) is a supervised machine learning algorithm we implemented using the *mlr3* library in R. The gene expression data was transformed as necessary for usage with the "classiff.kknn" learner and ran on initially the 5000 most variable genes, then the top 1000, 100, and 10.

Naive Bayes

One approach used in our predictive modeling analysis was to create a Naive Bayes model for classification. The Naive Bayes model functions on the principle of conditional probability. In our case, there are two categories: control and schizophrenia. During training, the model goes through the expression of genes for each subject and calculates the conditional probabilities associated. Using these probabilities, the Naive Bayes model determines the most likely classification during testing.

Results

We are able to support the result of the original study that found differences in the anterior cingulate cortex of schizophrenia samples compared to controls. We present that several data analysis methods were able to distinguish between CTL and SZ samples of the AnCg with an acceptable performance. We also present a list of 33 differentially expressed genes between the two groups. By providing differentially expressed genes and showing that unsupervised and supervised methods are able to group SZ and CTL samples, we demonstrate that there are quantifiable differences in expression between the two groups. However, it is important to note that not all methods were able to show clear distinctions between the two groups. It is also important to note that SZ samples had more variability in expression and that a small subgroup of SZ samples had more dramatic expression differences while many SZ subjects observed similar expression levels as CTL subjects.

Although we present a list of differentially expressed genes, our list differs from the one obtained in the original study. The original study defined an adjusted p-value cutoff of 0.05 and obtained 87 genes. In our analysis, we obtained 33 genes for a 0.01 cutoff and a list of 48 genes for a 0.05 cutoff. The differences may involve implementation specific details like the minimum counts cutoff or the combination of methods to derive differentially expressed genes in the original study. However, the difference in results is not understood.

Another “unexpected” result is that we were unable to yield significant gene ontology terms or biological pathways. In our list of 33 differentially expressed genes, our enrichment analysis did not yield significant results for biological pathways these genes are involved in. This may be due to errors in our implementation or incomplete pathway information in the packages used. We find this a considerable limitation of our study as we are unable to explain “how” expression differences between SZ and CTL patients impact biological pathways. So in summary, we are able to conclude that there are indeed expression differences between the two groups and that we present a list of 33 (0.01 adj p-value cutoff) and 48 (0.05 adj p-value cutoff) differentially expressed genes but are unable to identify the main pathways altered between the two.

Below we discuss the results of each method in more detail.

Differential Expression

PCA and UMAP

PCA and UMAP were conducted to view any initial separation of the two experimental groups.

For PCA, we plotted along PC1 and PC2, which captured 16.76% and 11.56% of variability in the data respectfully. PC1 and PC2 did not capture a significant portion of the data’s variability but the plot serves as a starting point to observe any separation of the two groups. Below is a bar plot showing the percentage of variance accounted for by each PC.

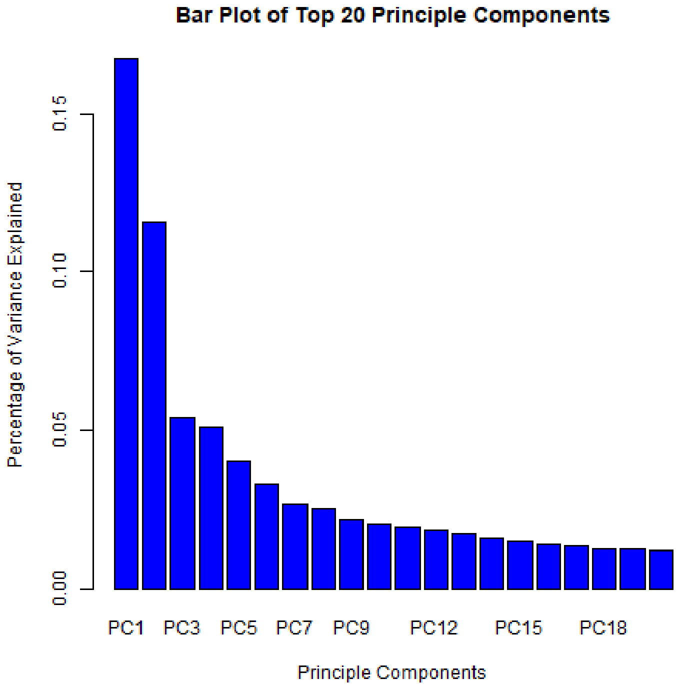


Figure 3: Barplot to represent the percentage of variance explained by each principal component

Plotting the subjects along the Principal Component 1 and 2 axes does not show a dramatic separation of CTL (red dots) and SZ (blue dots) samples but there are some takeaways. Along PC1, we observe that SZ observations show much more variance. As we increase along the PC1 axis, we observe more SZ samples. Variance is also higher for SZ samples along PC2 but is less dramatic. There is little clear separation of all SZ and CTL samples, but we observe a subgroup of SZ samples separate from the rest of the observations.

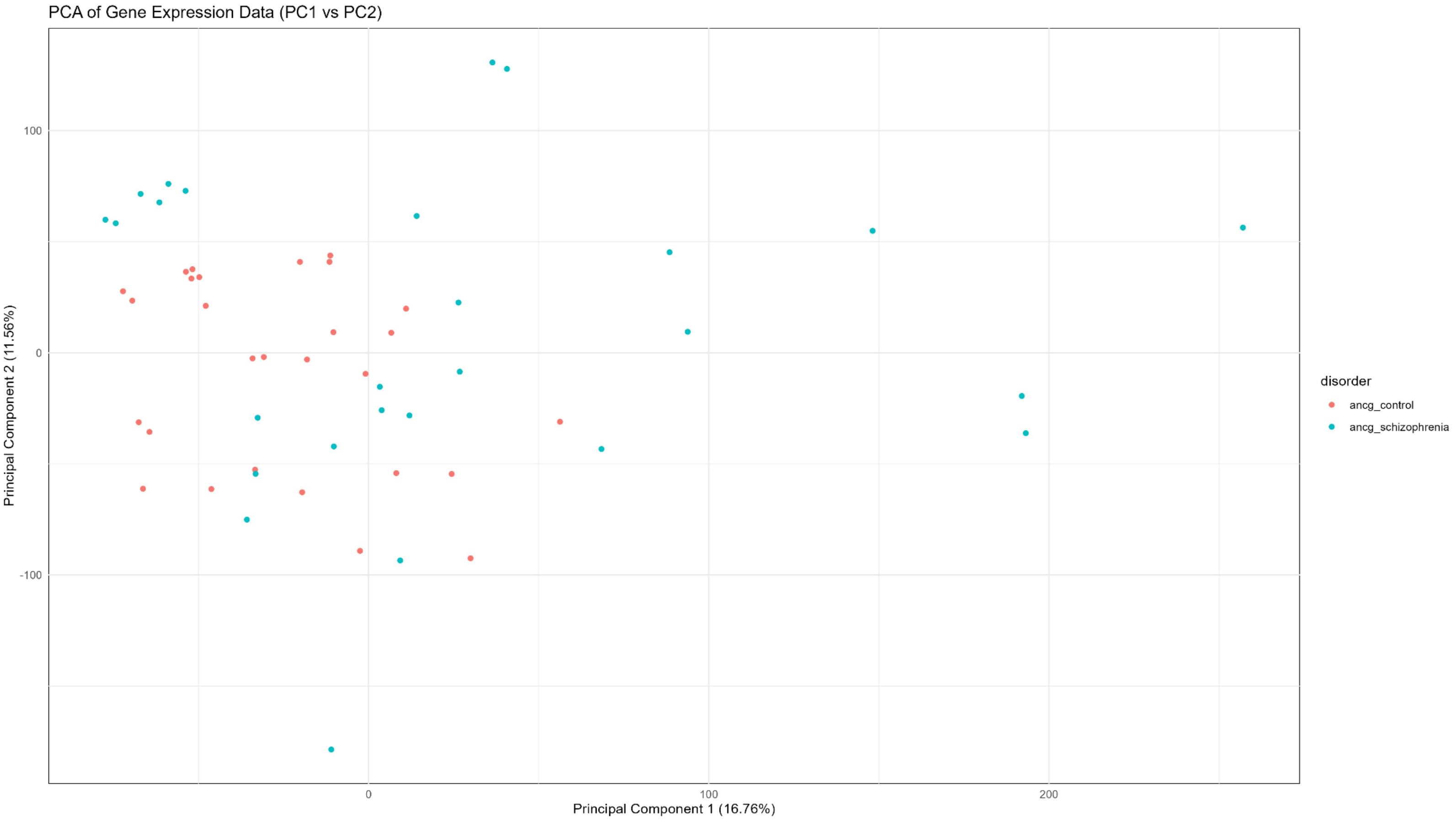


Figure 4: PCA Plot comparing PC1 (x-axis) and PC2 (y-axis) values of CTL and SZ samples

To further explore dimensionality reduction, we have also conducted the uniform manifold approximation and projection (UMAP) plot. The nature of UMAP allows for a higher dimensional exploration of the data to be transformed and reduced so we can visualize the data on a plot.

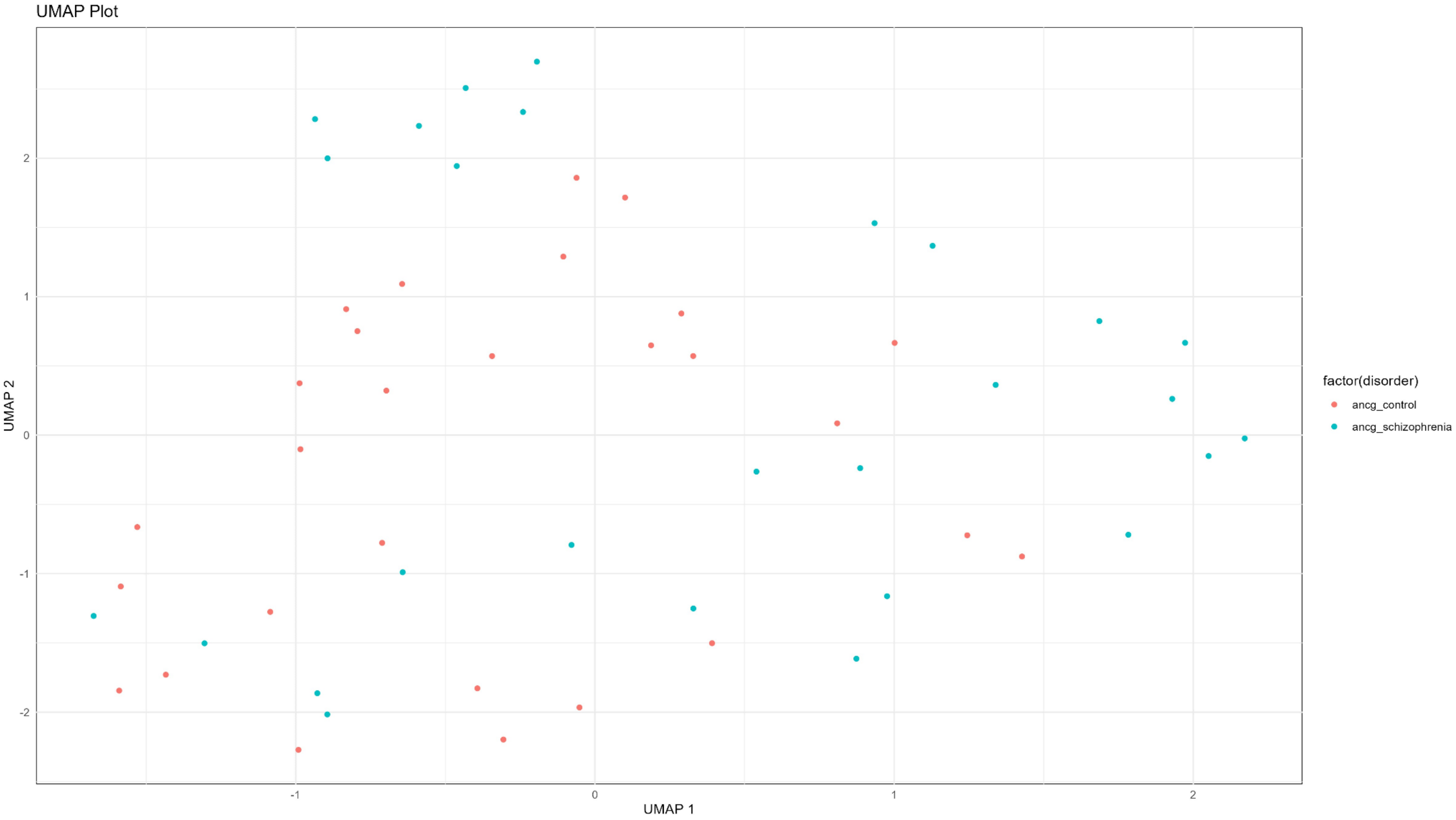


Figure 5: UMAP Plot comparing UMAP1 (x-axis) and UMAP2 (y-axis) values of CTL and SZ samples

Similar to our PCA plot, we do not observe a dramatic separation of SZ (blue dots) and CTL (red dots). It is interesting to note that again the SZ are more variable along both axes and that a subgroup of SZ are more extreme along UMAP1.

Differential Expression and Enrichment Analysis

Differential expression with DESeq2 resulted in a list of 33 differentially expressed genes with a 0.01 adjusted p-value cutoff. The genes are as follows:

> *SERPINA3, HS3ST3B1, IGHG4, MIR126, COQ7-DT, CHI3L2, STC1, CP, TMPRSS3, TMEM225, BAG3, GLI2, NPC1L1, LINC00397, C10orf105, MUC1, SLC14A1, DAPK1-IT1, HIF3A, TNFRSF10D, LINC01676, C9orf131, TEKT4,*

PAPLN, FAM225B, LINC02864, LOC105375166, LOC107984827, ITGB4, CYLC2, HCG22, TSSK2

A list of 48 genes can be derived with a 0.05 adjusted p-value cutoff (which was the threshold of the original study). The genes are as follows:

IGHG3, IGKC, FGA, SERPINA3, CSF3, HS3ST3B1, IGHG4, MIR126, COQ7-DT, CHI3L2, STC1, CCDC194, SLC34A1, TNFRSF6B, CP, TMPRSS3, SOCS3, TMEM225, MIA, TBC1D3E, RGR, BAG3, GLI2, NPC1L1, IL21, SPEM3, LINC00397, C10orf105, MUC1, SLC14A1, DAPK1-IT1, HIF3A, TNFRSF10D, LINC01676, C9orf131, TEKT4, PAPLN, IL1RL1, FAM225B, LINC02864, LOC105375166, CT45A3, LOC107984827, ITGB4, CYLC2, HCG22, CCL4, TSSK2

Using the differential expression results, we can create a volcano plot that shows log2 fold change along the x-axis and adjusted p-value along the y-axis.

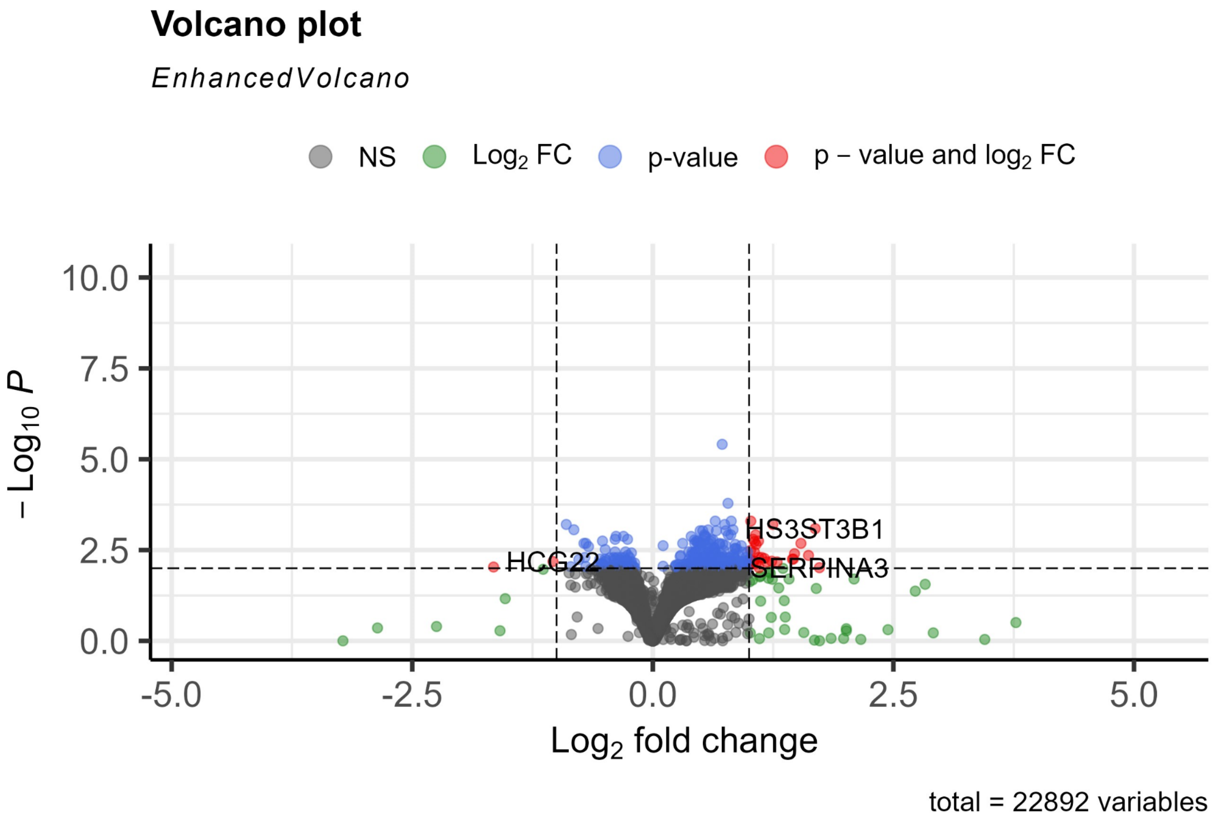


Figure 6: A volcano plot of the differential expression results. Differential expressed genes will have a log2 fold change magnitude > 1 and an adjusted p-value < 0.01. These thresholds are represented by dotted lines in the volcano plot.

We define a cutoff of 0.01 for the adjusted p-value. The volcano plot visualizes the “significantly differentially expressed” genes as red dots.

We proceed with creating a heatmap with our gene list.

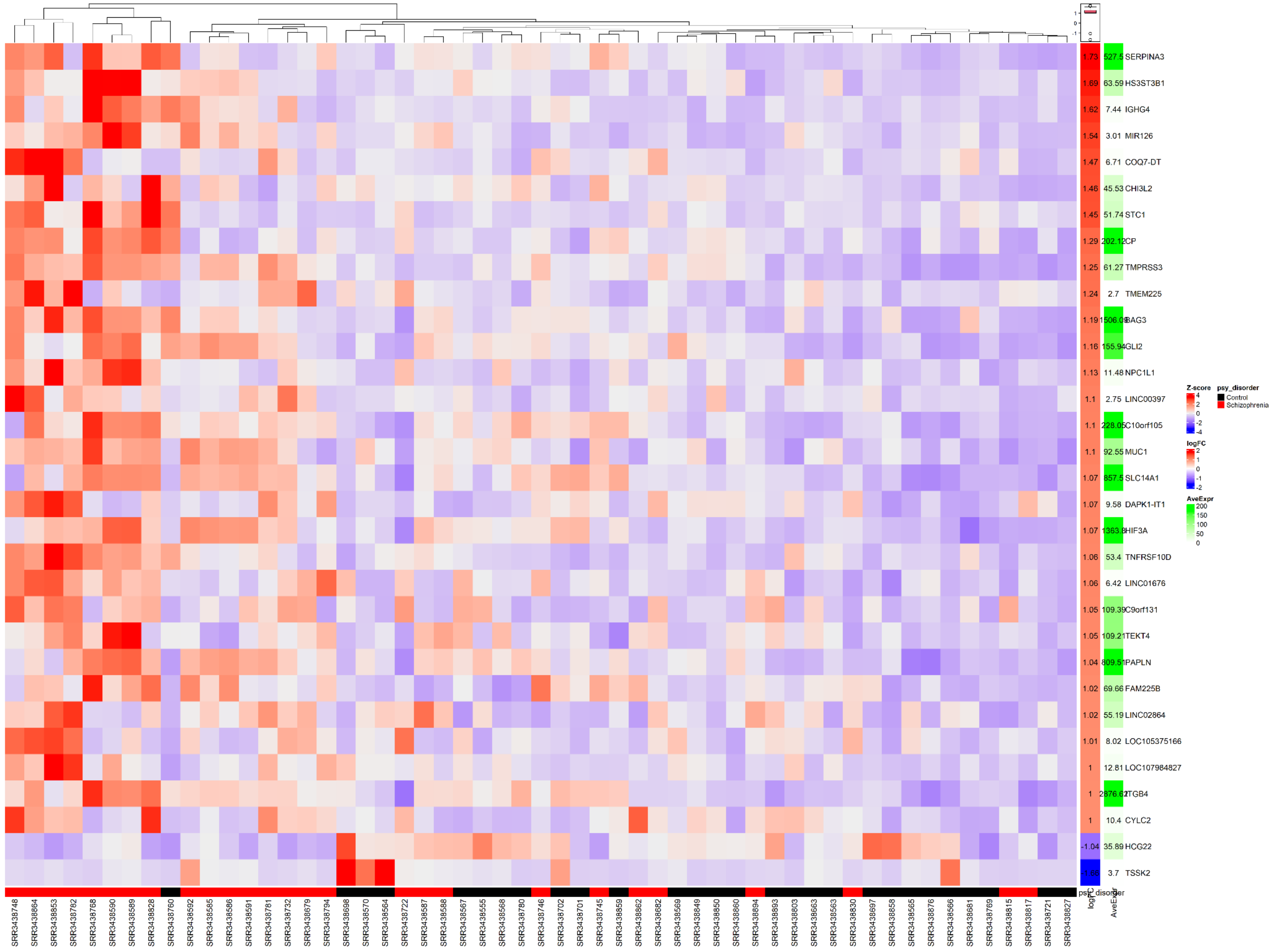


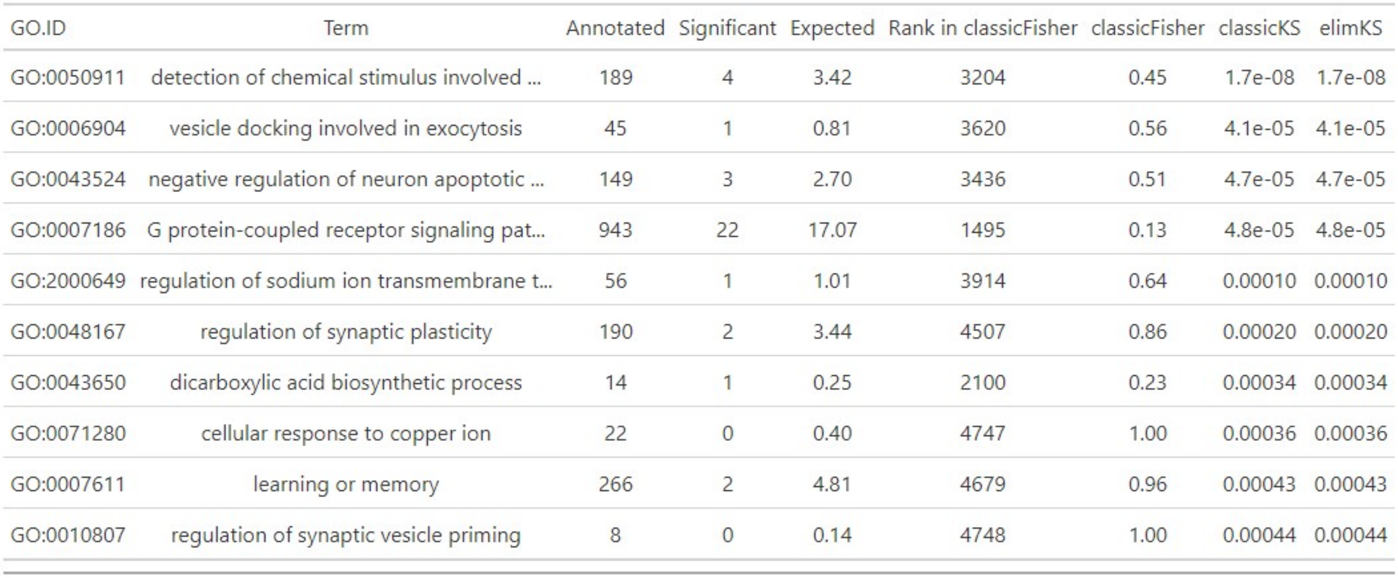
Figure 7: Heatmap of our 55 SZ and CTL samples with the extracted differentially expressed genes

A heatmap, as shown in the figure above, visualizes the differences in expression of the samples.

The columns of the heatmap are each of the 55 samples of our dataset and the rows are each of the differentially expressed genes. Red cells indicate a gene that is up-regulated in SZ samples while blue cells are genes that are down-regulated in SZ samples. The genes are ordered by log2 fold change which is annotated by the vertical red and blue bar. Genes that display a higher positive log2 fold change are the upper rows while the negative log2 fold change genes are the lower rows. There is another vertical annotation bar that quantifies the expression levels or the counts of the gene read. Finally, there is a horizontal red and black annotation bar by the sample labels. This labels a sample as either SZ (red) or CTL (black). The samples are clustered with an accompanying dendrogram above.

Observing our heatmap, we see a cluster of SZ samples toward the left that considerably up-regulate the list of genes. This result continues to support the idea that a subset of SZ samples exemplifies a dramatic difference in gene expression compared to the rest of the subjects. SZ samples tend to occupy the left of the heatmap while CTL samples occupy the right. We see that these genes tend to be up-regulated in SZ samples while down-regulated in CTL samples.

Continuing with enrichment analysis, it was noted previously that we were unable to extract any meaningful biological pathways. For running clustProfiler and gProfiler, no GO terms or pathways met the 0.05 adjusted p-value threshold to be considered statistically significant. As for topGO analysis, we extracted the following table of the top 10 GO terms, however they have high classic Fischer score p-values and are not considered statistically significant either.



GO.ID	Term	Annotated	Significant	Expected	Rank in classicFisher	classicFisher	classicKS	elimKS
GO:0050911	detection of chemical stimulus involved ...	189	4	3.42	3204	0.45	1.7e-08	1.7e-08
GO:0006904	vesicle docking involved in exocytosis	45	1	0.81	3620	0.56	4.1e-05	4.1e-05
GO:0043524	negative regulation of neuron apoptotic ...	149	3	2.70	3436	0.51	4.7e-05	4.7e-05
GO:0007186	G protein-coupled receptor signaling pat...	943	22	17.07	1495	0.13	4.8e-05	4.8e-05
GO:2000649	regulation of sodium ion transmembrane t...	56	1	1.01	3914	0.64	0.00010	0.00010
GO:0048167	regulation of synaptic plasticity	190	2	3.44	4507	0.86	0.00020	0.00020
GO:0043650	dicarboxylic acid biosynthetic process	14	1	0.25	2100	0.23	0.00034	0.00034
GO:0071280	cellular response to copper ion	22	0	0.40	4747	1.00	0.00036	0.00036
GO:0007611	learning or memory	266	2	4.81	4679	0.96	0.00043	0.00043
GO:0010807	regulation of synaptic vesicle priming	8	0	0.14	4748	1.00	0.00044	0.00044

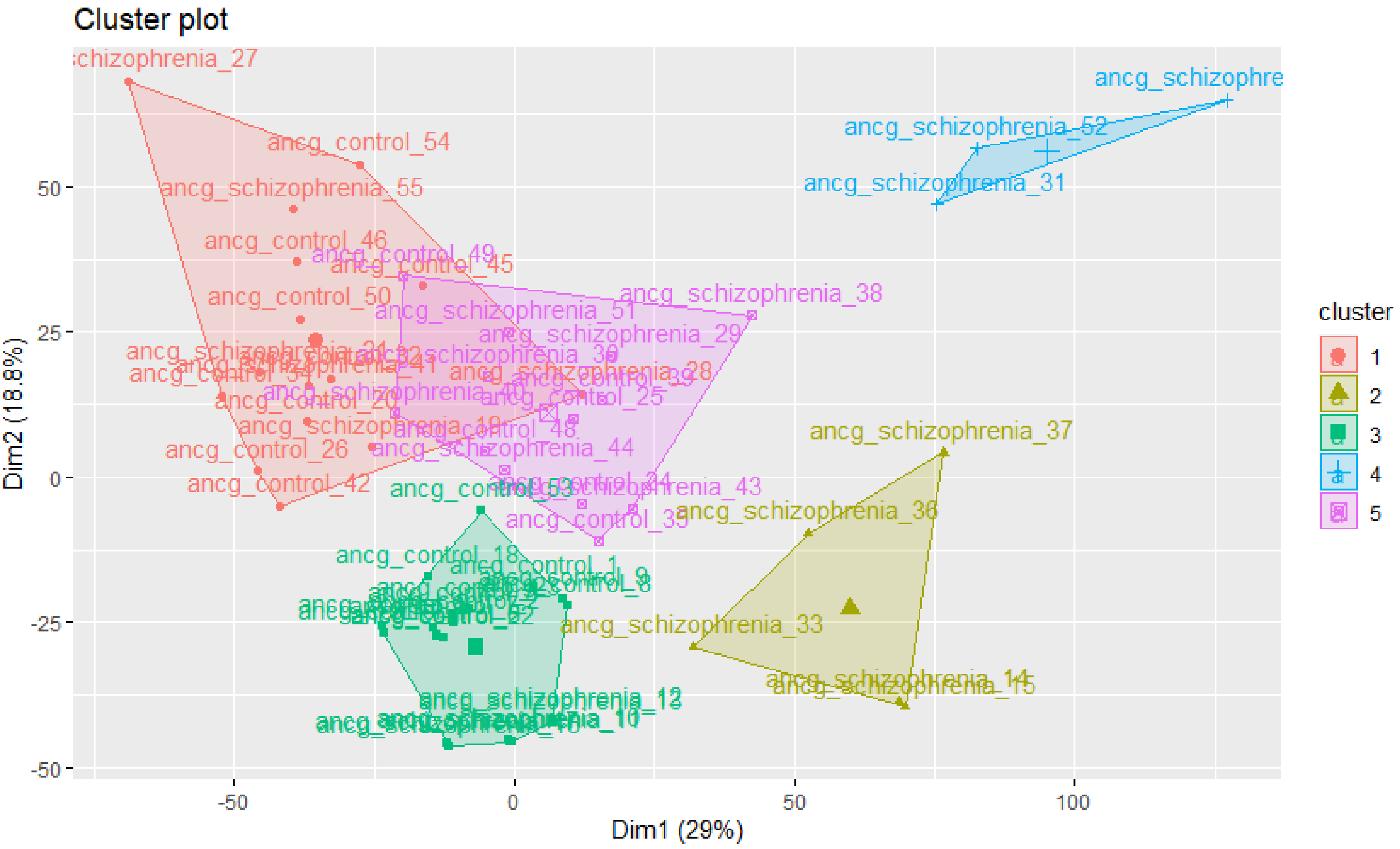
Figure 8: Table of top 10 GO terms from topGO analysis

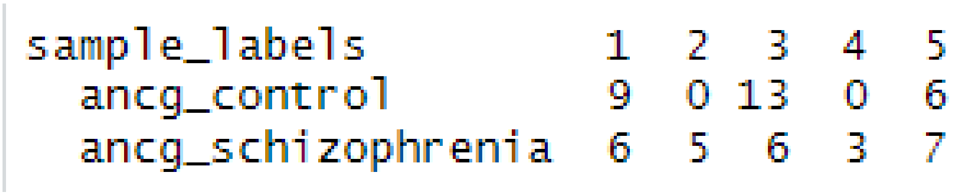
We find this as a considerable limitation in our study. We suspect incorrect implementations of enrichment analysis as a main reason for yielding insignificant results. However, we are confident in our list of differentially expressed genes. Therefore, we encourage further and “better” gene set enrichment analysis using our list of genes provided above.

Clustering

K-means Clustering

As noted in the Methods section, we selected an appropriate k to be 5. With k=5, we achieve **between_SS / total_SS = 49.1%**, which is a metric of how “good” the clustering explained differences between groups. With a value of 49.1%, the clustering conducted showed that much of the variance of our data is not explained by separation of clusters. Below is a plot of the k-means clustering and table showing cluster membership.



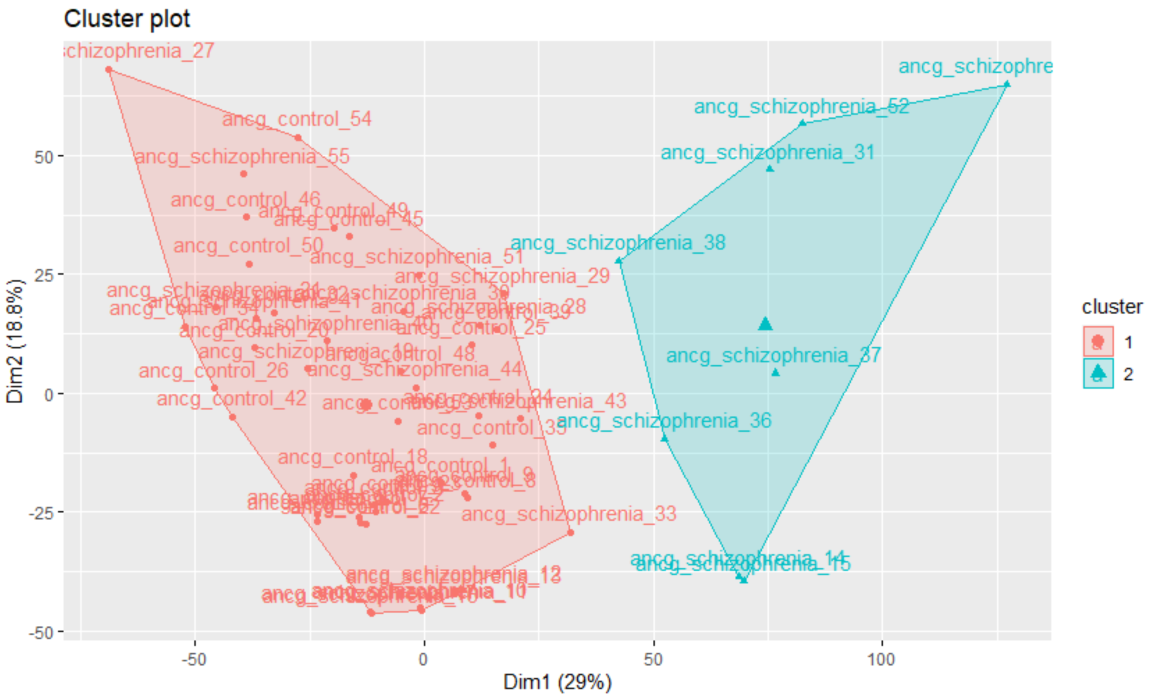


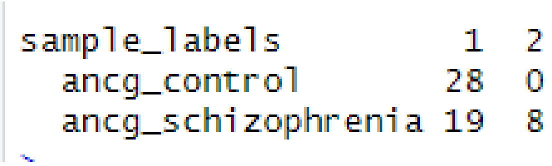
```
sample_labels          1   2   3   4   5
  ancg_control         9   0  13   0   6
  ancg_schizophrenia   6   5   6   3   7
```

Figure 9: Cluster plot created by conducting k-means clustering (k=5) and a table showing cluster membership by each group.

From our plot, we can see two groups (2 and 4) that contain only schizophrenia samples. Groups 1 and 3 have more control samples while group 5 contains almost equal amounts of both samples. There is not a strong separation of the two experimental groups. However, it is again interesting to observe a small subset of SZ samples that do exemplify a significant separation from the rest of the subjects. Here, these SZ samples cluster into groups 2 and 4.

When we set the k=2, we observe a similar result.





```
sample_labels        1   2
  ancg_control      28   0
  ancg_schizophrenia 19  8
```

Figure 10: Cluster plot created with k-means clustering (k=2) and a table showing cluster membership

We see that a group of SZ samples cluster strongly together as opposed to other SZ samples.

PAM Clustering

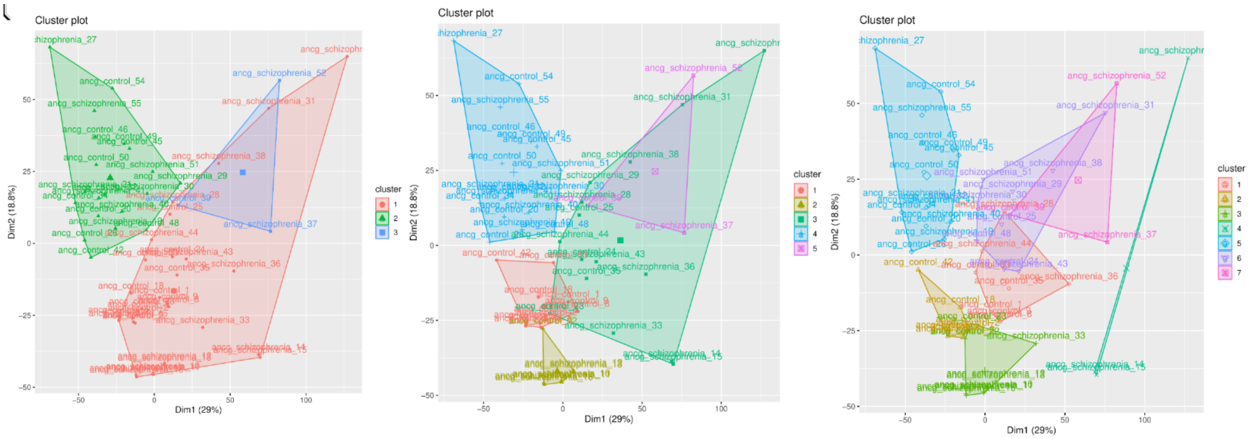


Figure 11: PAM cluster plots for k = 3, 5, and 7 respectively (from left to right).

The *k* = 3 plot has large overlap between clusters 1 and 3, and due to the low number of clusters the shapes are very large. The *k* = 7 plot creates a remarkably long and narrow cluster in group 7. The *k* = 5 plot creates the most compact cluster among all 3 plots in group 2. In keeping consistent with the K-means clustering procedure, *k* = 5 cluster analysis was proceeded with.

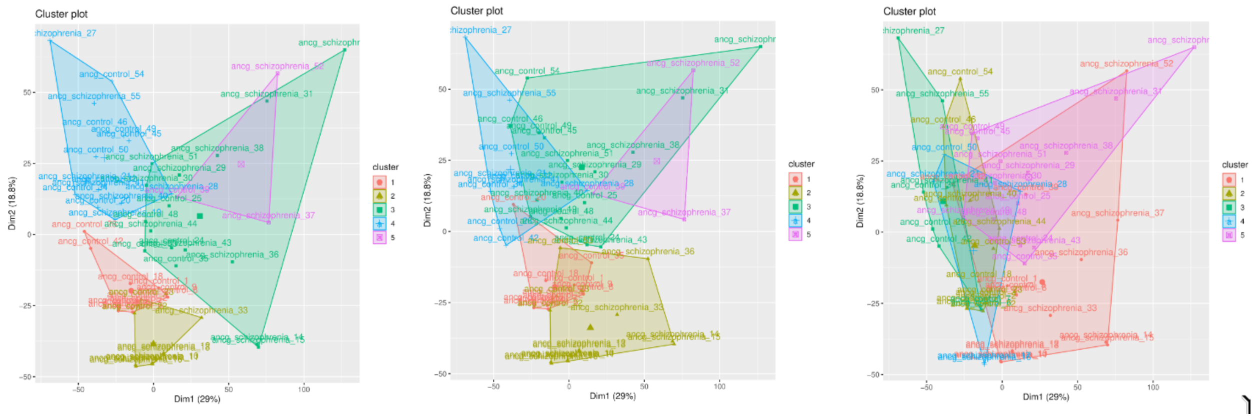


Figure 12: PAM cluster plots for k = 5 on the 1000, 100, and 10 most variable genes respectively (from left to right)

It appears that the number of genes in the dataset is associated with smaller cluster sizes and less overlap between clusters. The cluster plot with 10 genes shows overlap among all cluster groups, while plots for 100 and 1000 genes are less muddled.

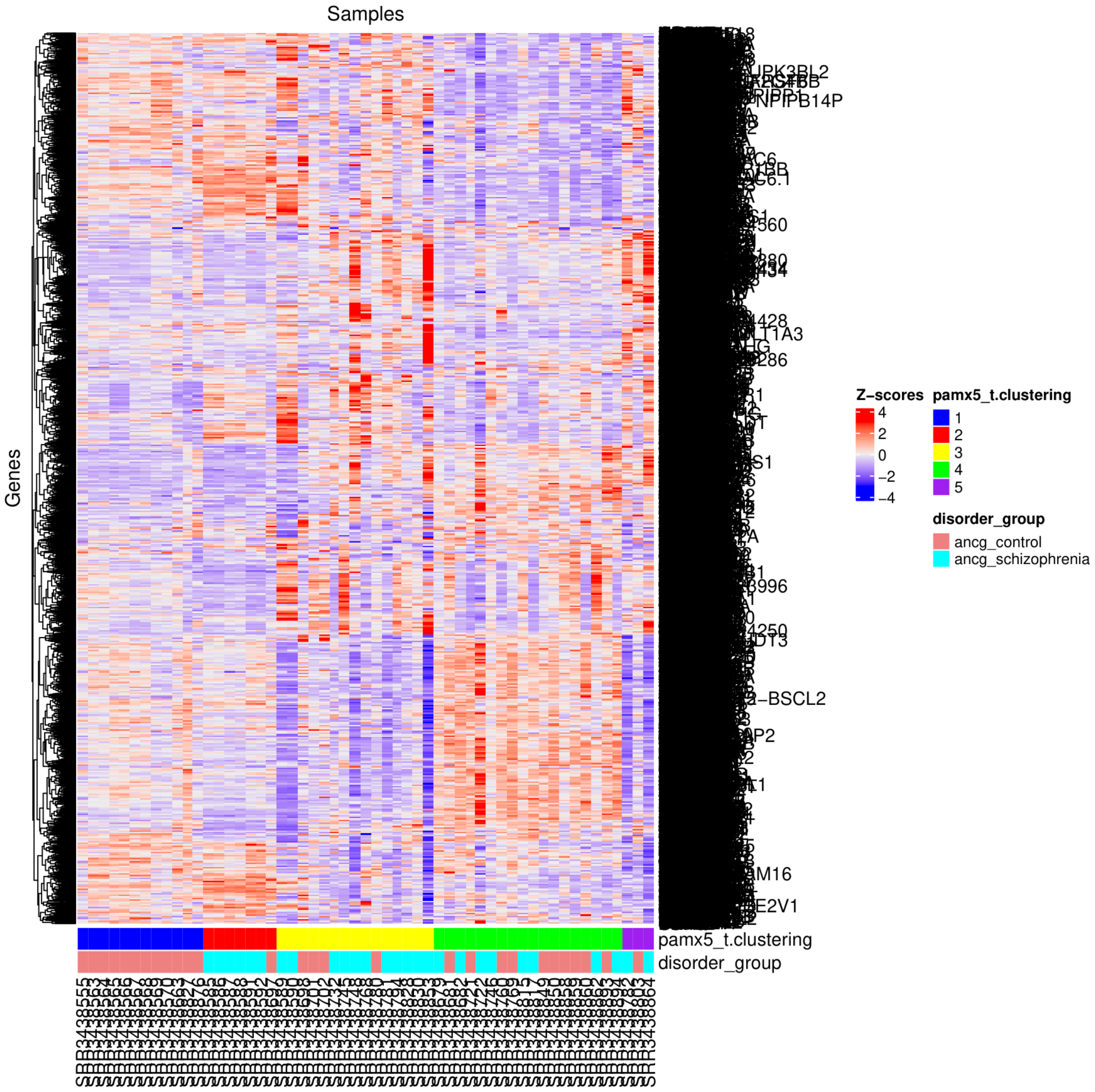


Figure 13: Heatmap plotting gene expression with PAM clustering annotation

Hierarchical Clustering

The figure below shows the hierarchical clustering analysis conducted on 10,000 genes.

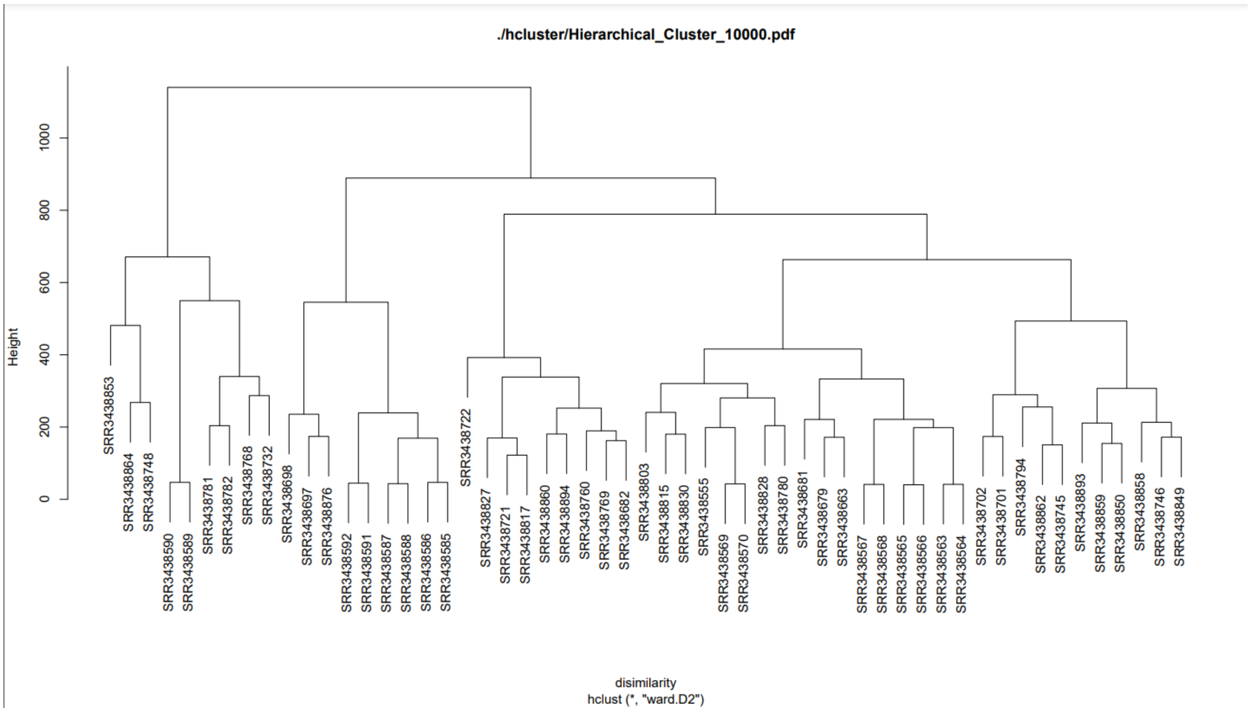


Figure 14: Dendrogram of hierarchical clustering result conducted on the 10,000 most variable genes

Additionally, the heatmap below visualizes the differential expression of the most variant genes on the Y-Axis. On the X-Axis, the categorical labels for each hierarchical cluster are overlaid with the control/schizophrenia labels.

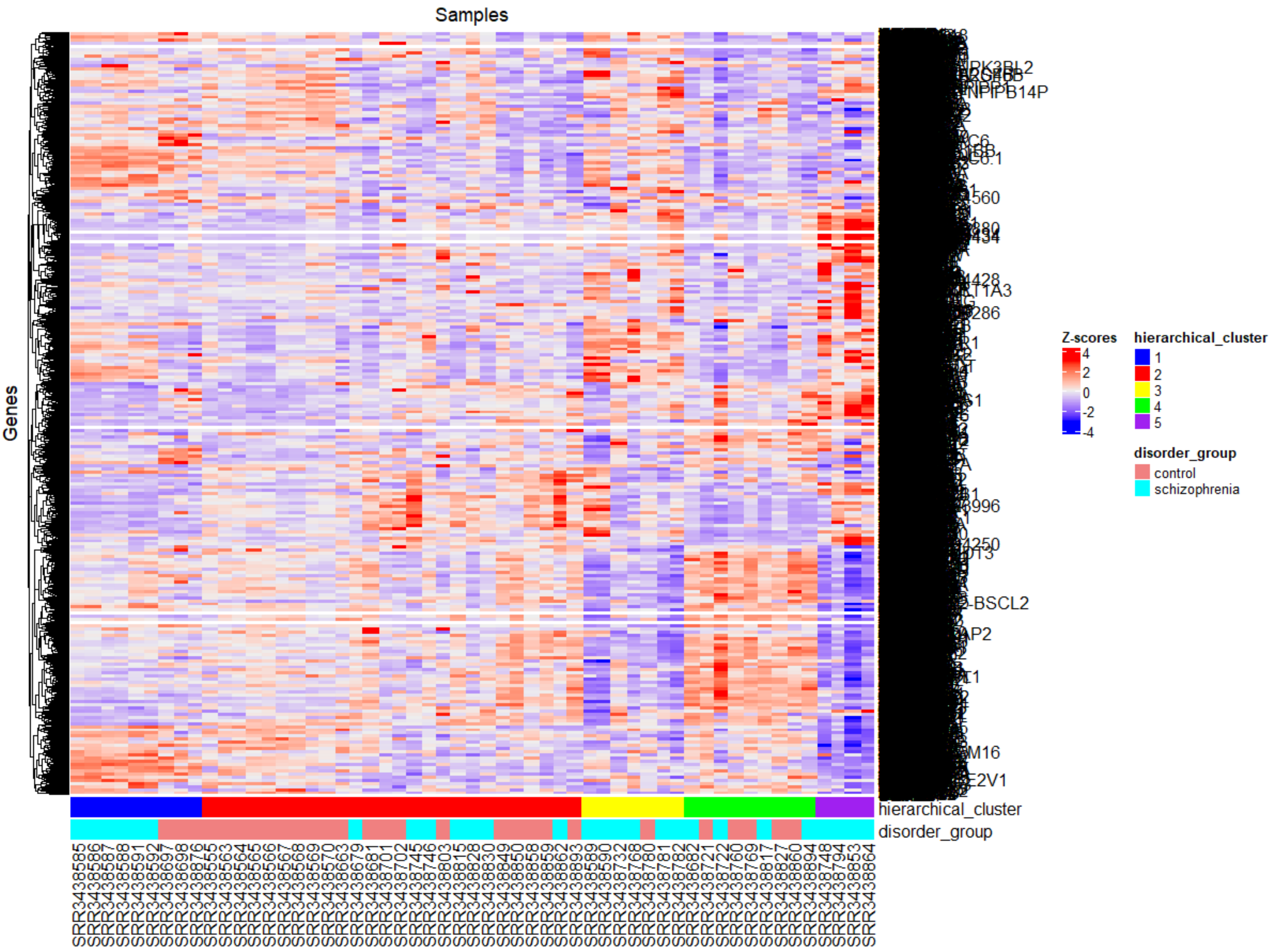


Figure 15: Heatmap plotting gene expression with a hierarchical cluster annotation

Alluvial Plot

Results for three clustering methods can be combined into a single visualization via an alluvial plot. The alluvial plot below displays cluster membership of each sample across the three methods.

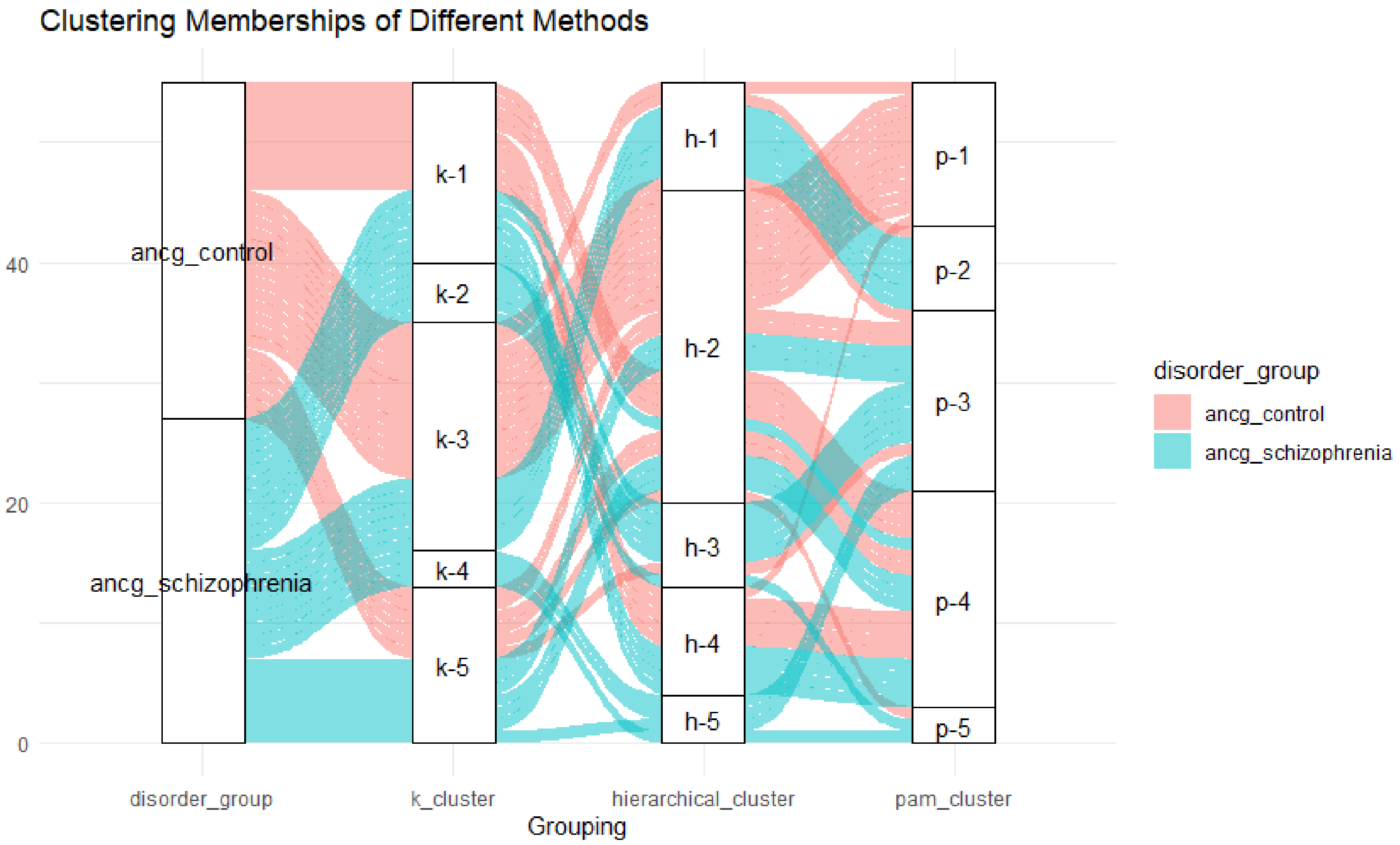


Figure 16: Alluvial Plot displaying cluster memberships of each sample. Control subjects are in red and schizophrenia subjects are in blue

From this plot, we can observe some interesting patterns. We can observe that all of cluster k-4, which contains only schizophrenia samples, are members of h-5. All k-2 samples, which are also only schizophrenia samples, are members of h-3 This may indicate a strong clustering of those schizophrenia samples as it has been identified by two methods.

We can also see that in many cases the different clustering methods identified similar clusters. A majority of the control samples of k-3 are members of h-2 and p-1. All the schizophrenia samples of k-3 are members of h-1 and p-2. This indicates that across three clustering methods, the same samples were clustered together in those aforementioned groups.

Chi Squared Test of Independence

With these three clustering methods, we tested whether cluster membership and disorder (SZ or CTL) were independent. Simply, a chi-squared test of independence was conducted using the cluster memberships assigned by each method. Below are the p-values from conducting a chi-squared test of independence for each method, and then an adjusted p-value for multiple hypothesis testing.

Cluster Method	Unadjusted P-Value	Adjusted P-Value
K-Means	0.02398	0.071940
PAM	0.0006611	0.001833
Hierarchical	0.00661	0.019830

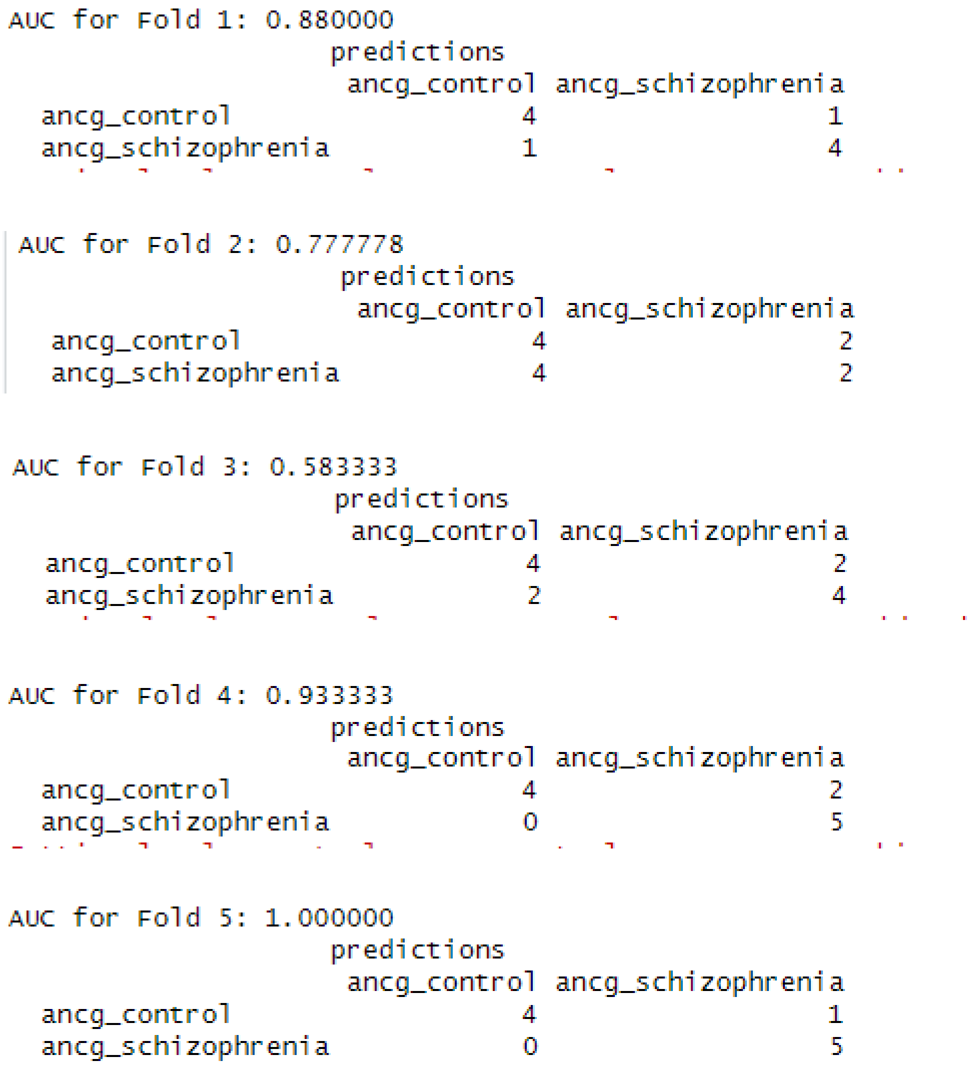
P-values were adjusted using the p.adjust() function in R with the Bonferroni method. The adjusted p-values demonstrated a correlation between cluster membership and psychiatric disorders for PAM and hierarchical clustering methods; however, k-means clustering was found to be not statistically significant. The adjusted p-value for k-means suggests that we are unable to reject that cluster group and psychiatric disorder are independent for that specific method. For the other two methods, The results from PAM and hierarchical clustering suggest psychiatric disorder and cluster membership are **not independent.**

Supervised Analysis (Predictive Modeling)

Linear SVM

Linear SVM models were trained to classify models as either CTL or SZ based on the 5000 most variable genes of our dataset. 5-fold cross validation is employed to train and evaluate the model. Due to fold-selection being inherently random, five interactions of linear SVM were conducted. Here are the five iterations of model training/validation, their respective confusion matrices, and their AUC.

In the confusion matrices, the columns correspond to predictions while the rows are the actual labels.



```
AUC for Fold 1: 0.880000
                     predictions
                      ancg_control ancg_schizophrenia
  ancg_control                   4                  1
  ancg_schizophrenia             1                  4
```

```
AUC for Fold 2: 0.777778
                     predictions
                      ancg_control ancg_schizophrenia
  ancg_control                   4                  2
  ancg_schizophrenia             4                  2
```

```
AUC for Fold 3: 0.583333
                     predictions
                      ancg_control ancg_schizophrenia
  ancg_control                   4                  2
  ancg_schizophrenia             2                  4
```

```
AUC for Fold 4: 0.933333
                     predictions
                      ancg_control ancg_schizophrenia
  ancg_control                   4                  2
  ancg_schizophrenia             0                  5
```

```
AUC for Fold 5: 1.000000
                     predictions
                      ancg_control ancg_schizophrenia
  ancg_control                   4                  1
  ancg_schizophrenia             0                  5
```

Figure 17: Confusion matrices of five different linear SVM trials

The **average AUC is 0.8349**, suggesting that around 83.49% of the time the model is able to correctly label a sample with its disorder.

Below are all the ROC curves in one plot - the y-axis (sensitivity) is the true positive rate while the x-axis (1-specificity) represents the false positive rate. Values toward the top left indicate a high-scoring model - being able to predict with a high true positive rate and low false positive rate.

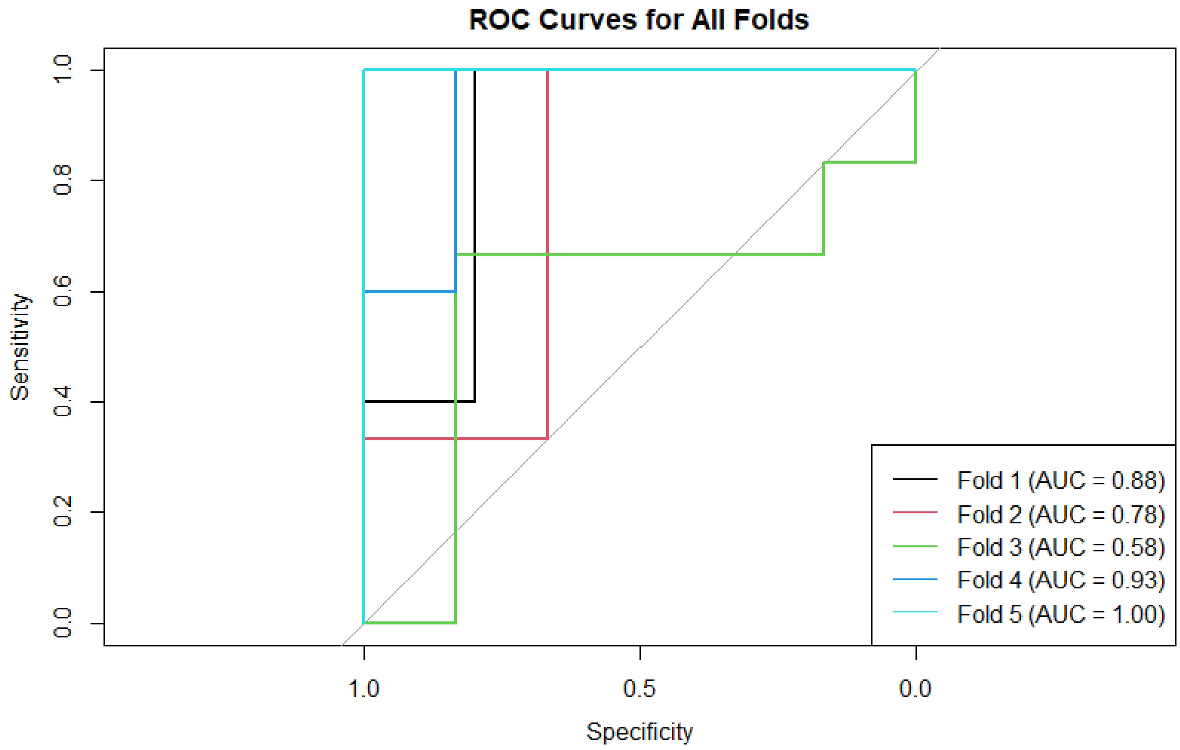
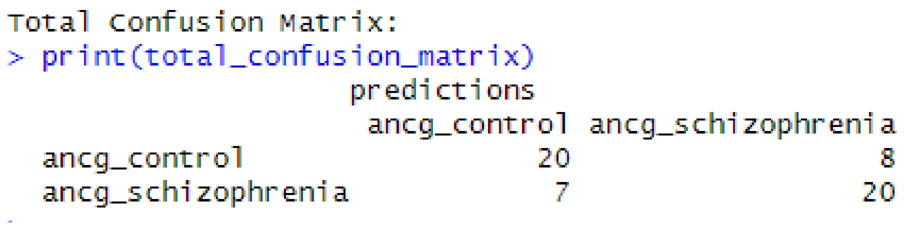


Figure 18: ROC curves of each of the five linear SVM trials

All the confusion matrices can be combined into one to nicely summarize the model predictions.



```
Total Confusion Matrix:
> print(total_confusion_matrix)
                   predictions
                    ancg_control ancg_schizophrenia
  ancg_control                20                  8
  ancg_schizophrenia           7                 20
```

Figure 19: Combined confusion matrix

With this confusion matrix, we can summarize the performance of the model using the following metrics: precision, recall, and f-1 score. Positive means predicting schizophrenia patients while negative corresponds to control predictions.

F1 score, like AUC, is a good summary metric that evaluates the performance of a model because it talkies precision and recall into account (which are often inversely related).

Positive (Schizophrenia) Precision: **0.7142**
Positive (Schizophrenia) Recall: **0.7407**
Positive (Schizophrenia) F1: **0.7273**

Negative (Control) Precision: **0.7407**
Negative (Control) Recall: **0.7143**
Negative (Control) F1: **0.7273**

With an F1 score of about 0.73, the linear SVM model **performs adequately to predict whether a sample is control or schizophrenia.**

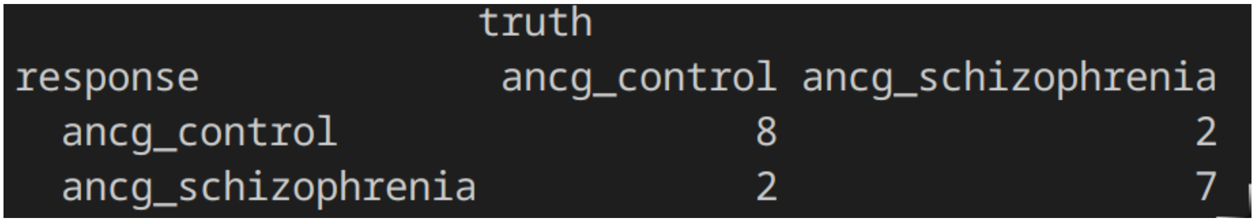
Naive Bayes

AUC was calculated using a 5-Fold cross validation through tidymodels. Data was passed through the vfold_cv() function, with *v = 5*. The classifiers were identified using all genes as features other than the class variable itself, labeling which patients were control or schizophrenic. Using the fit_resamples() method, the data was fit to the Naive Bayes model and evaluated for its AUC accuracy.

Number of Genes Included	Average AUC value across 5-Fold Cross Validation
10,000	0.777
5,000	0.788
1,000	0.746
100	0.730
10	0.717

K Nearest Neighbors

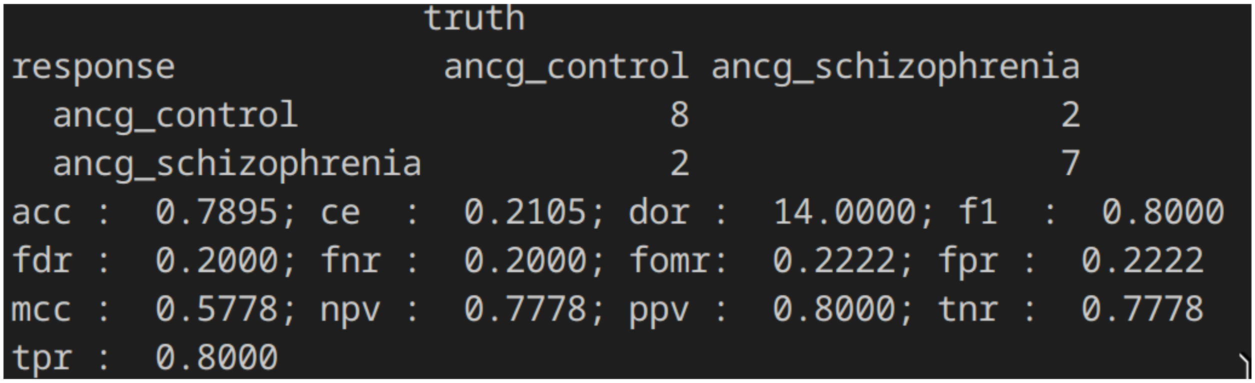
The confusion matrix that resulted from the test set (used for evaluating performance) on the K Nearest Neighbors learner model for the top 5000 most variable genes is shown below:



```
                    truth
response             ancg_control ancg_schizophrenia
  ancg_control                  8                  2
  ancg_schizophrenia            2                  7
```

Figure 20: Confusion matrix resultant from test set using KNN on 5000 most variable genes

This confusion table demonstrates the accuracy of the training set when compared to true results for the 5000 gene model on a subset of the samples.



```
                     truth
response              ancg_control ancg_schizophrenia
  ancg_control                   8                  2
  ancg_schizophrenia             2                  7
acc :  0.7895; ce  :  0.2105; dor :  14.0000; f1  :  0.8000
fdr :  0.2000; fnr :  0.2000; fomr:  0.2222; fpr :  0.2222
mcc :  0.5778; npv :  0.7778; ppv :  0.8000; tnr :  0.7778
tpr :  0.8000
```

Figure 21: Table demonstrating several common confusion matrix statistics.

Supervised analysis showed that several methods are able to classify SZ and CTL samples with adequate performance, indicating quantifiable expression differences between the two experimental groups.

The ACC measurement was used for determining efficacy of the mode across different numbers of genes. The table below outlines ACC scores for the top 5000, 1000, 100, and 10 genes.

Number of Genes	classif.acc
5000	0.7894737
1000	0.7368421
100	0.8421053
10	0.7368421

The 200 most important gene signatures were extracted from the task object via the "information_gains" filter in *mlr3*. The following table demonstrates how the number of intersecting gene signatures differs with the number of genes provided.

	5000	1000	100	10
5000				
1000	50			
100	7	19		
10	1	3	10	

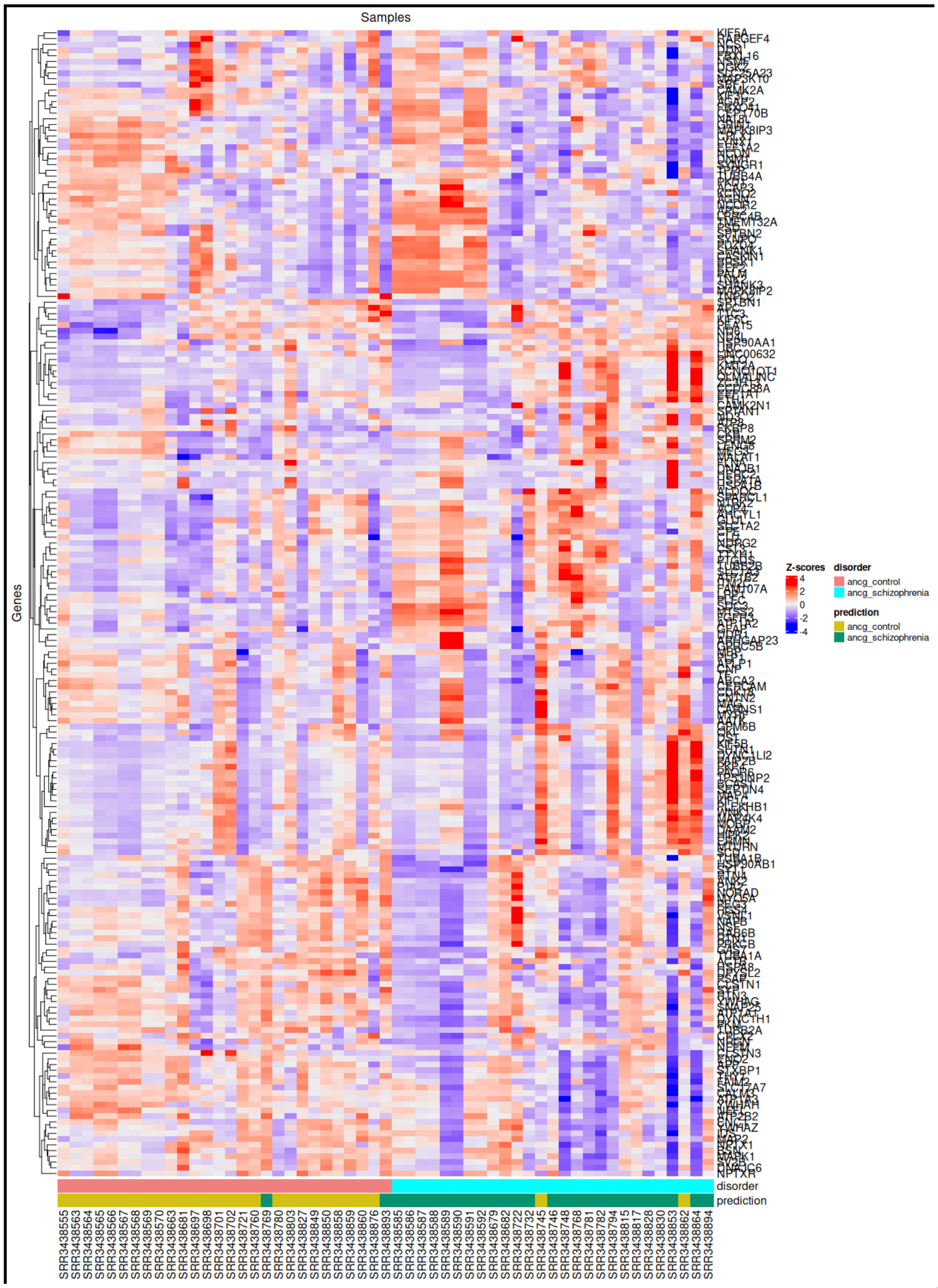


Figure 22: Heatmap plotting gene expression with KNN annotations

This heatmap displays the 200 most important genes as discriminated by the "information_gains" filter in *mlr3* and how they are differentially expressed across control and disorder patients. The heatmap also demonstrates how the k-neighbors model classified the samples for visual comparison. Similar differential expression patterns can be noted from previous analyses, notably the increase in differentially upregulated and downregulated genes in schizophrenic patients (i.e. more intense saturation of the red or blue). The disorder and prediction ribbons show remarkable overlap and demonstrate the accuracy of the k-neighbors algorithm in differentiating between disorder and control patients. It is interesting to note that there are more false negatives than false positives (where positive is disorder and negative is control).

Conclusion

Our study sought to identify differences in the gene expression levels of patients with schizophrenia.Our differential analysis was able to identify 33 genes that were statistically significant at a padj < 0.01 and 48 genes at padj < 0.05, which differed from the findings of the original study (Ramaker et al., 2017). Our PAM and hierarchical clustering analyses were able to confirm a statistically significant difference in the expression of genes between schizophrenia and control patients, as determined by our chi-squared test. Additionally, we were able to create predictive models and obtain gene signatures from our support vector machine and k-nearest neighbor models. These models classified CTL and SZ samples with an adequate performance.

Weaknesses of our analyses were the enrichment analysis and our lack of subject data. Our enrichment analysis under the clustProfiler and gProfiler yielded no statistically significant results, making it difficult to draw conclusions. We believe an incorrect implementation of enrichment analysis influenced the results.

Although our clustering and predictive modeling analyses were able to conclude that there was a statistically significant difference in the expression of genes in schizophrenic patients, our analysis does little in the way of explaining the biological mechanisms behind this. However, we were able, again, to demonstrate statistically significant and quantifiable differences in gene expression of control and schizophrenia patients, which we sought to test. In the future, our group would have liked to conduct a more in-depth pathway analysis using tools like Reactome. An exploration into the functional networks of the genes could provide more insight into the etiology of schizophrenia and how to treat it.

References

Ramaker, R.C., Bowling, K.M., Lasseigne, B.N. et al. Post-mortem molecular profiling of three psychiatric disorders. Genome Med 9, 72 (2017). https://doi.org/10.1186/s13073-017-0458-5

H. Wickham. ggplot2: Elegant Graphics for Data Analysis.Springer-Verlag New York, 2016.

Kuhn et al., (2020). Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles.https://www.tidymodels.org

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Weissman DH, Gopalakrishnan A, Hazlett CJ, Woldorff MG. Dorsal anterior cingulate cortex resolves conflict from distracting stimuli by boosting attention toward relevant events. Cereb Cortex. 2005 Feb;15(2):229-37. doi: 10.1093/cercor/bhh125. Epub 2004 Jul 6. PMID: 15238434.

Link to dataset:
https://www.refine.bio/experiments/SRP073813/rna-sequencing-of-human-post-mortem-brain-tissues

Further code references can be found in our repository at
https://github.com/rqian239/bioinformatics-project