Differential Gene Expression Analysis of Mental Disorders Using Post-mortem Brain Tissues

By: Joseph Bui, Justin Lu

Abstract

In the past, biologists have attempted to use genome-wide association studies (GWAS) to determine if there are certain genetic factors underlying the progress of different human conditions, from people's heights to development of certain disorders. The studies focused on finding locations on the genome associated with a disease or behavior, and tracing the pathways to find places to introduce cures. However, GWAS does not specifically search for the genes that help to explain a disorder. Rather, they simply find certain base pairs that seem to relate with the disorder or trait that the researchers are studying. In our experiment, we aimed to pinpoint the specific genes that, when they are expressed in abnormally high or low amounts, contribute to someone developing a mental disorder. We worked with RNA-sequenced data from post mortem brain samples of individuals affected by schizophrenia, bipolar disorder, and depression, in order to identify the specific genes that were expressed differently in people with those disorders versus people who were healthy.

Background

Many psychiatric disorders have been studied rigorously throughout history as they have contributed remarkably to human pain, suffering, and mortality. Among these disorders, schizophrenia (SZ), bipolar disorder (BPD), and major depressive disorder (MDD) have some of the most complex etiologies[5], which require further investigation in order to determine their causes and implement preventative measures and solutions for cures. All of these psychiatric disorders have been known to cause irregularities in patients' everyday lives and activities. Schizophrenia is a mental illness that affects how a person thinks, feels, and behaves, which results in a loss of a sense of reality[1]. Bipolar disorder is a mental disorder that causes unusual shifts in moods, energy, and activity levels[2]. Depression is another mental disorder that is diagnosed in patients who have serious mood shift symptoms that affect how they feel and think[3]. These three disorders are multigenic diseases[5] that share many symptoms such as psychosis, suicidal ideation, sleep disturbances, and cognitive deficits, suggesting a shared set of genetic causes. Specifically, this means that because they have similarities with regard to

symptoms, the same genetic variations may possibly be found on the genomes of the individuals who suffer from these disorders.

In order to study the common causes of these disorders, many researchers in the past have used genome-wide association studies (GWAS[4]) to analyze unique locations on the genome that are associated with or may lead to a disease or condition. The results of GWAS tells us that because these psychiatric disorders show similar symptoms, there are likely to be similarities in the genome of the patients, specifically certain base pairs. These nucleotide variations are referred to as single-nucleotide polymorphisms (or SNPs). GWAS studies, however, merely look for SNPs that may be correlated with a trait or disorder, but do not actually find genes that directly cause that trait or disorder. Studies like GWAS did not focus on trying to differentiate those suffering from the disorders in terms of genes, especially not at the genetic level (with regards to mRNA). As a result, it became necessary to undergo further exploration of the underlying genetic causes of these disorders, like the extensive experiments done in our study. Our goal was to try and uncover genes that differed in expression trends when someone was afflicted with the mental disorders, and understand why those genes could have led to the development of the disorders. With this information and more, researchers and medical professionals can find dependable cures or treatments for these mental illnesses. Because current therapies for these patients only treat a subset of their symptoms, it is more beneficial for researchers to pinpoint the underlying pathologies of these disorders to come up with more universal cures, relieving stress and improving care for these patients around the world.

Introduction

In our replication study, the researchers focused on tissues from the anterior cingulate cortex (AnCg), dorsolateral prefrontal cortex (DLPFC), and nucleus accumbens (nAcc) regions of the brain, as these regions of the brain are associated with the behaviors known to be altered by psychiatric disorders[5]. Researchers performed RNA sequencing (referred to as RNA-seq) on the genetic data from dissected post-mortem brain tissues in 3 groups of 24 patients each with schizophrenia, bipolar disorder, major depressive disorder, as well as 24 healthy individuals that acted as the control group. RNA-seq is a technique for transcriptome profiling, specifically to quantify the levels of each transcript (mRNAs, small RNA, etc.) in a cell, which together are responsible for dictating the transcriptional structure of genes. Using these data, we can quantify the changing expression of each transcript (or gene) during development of a disease or

physiological condition. From our results, we were able to find some similarities between schizophrenia and bipolar disorder with regards to the genes that are abnormally expressed under those conditions, and we were also able to identify some of those genes for further research.

Methods

Patient sample collection and preparation[5]

Our brain tissue samples of the 4 groups were collected by the Brain Donor Program at the University of California, Irvine, Department of Psychiatry and Human Behavior. Coronal slices of the dissected brain samples were frozen in aluminum plates at -120℃. To ensure all factors of the samples were considered, psychological autopsies were performed. For instance, the families of the deceased individuals that provided the samples were interviewed, and the collection and analyses of medical and psychiatric records were observed. As for the control group, samples were selected based on a lack of severe psychiatric disturbances and mental illnesses within their first-degree relatives.

RNA-Seq and Data Processing/Quality Control

For the core of our data, we obtained .fastq files extracted from RNA-seq experiments by the Pritzker Neuropsychiatric Disorders Research Consortium[6], which provided a pair of SRRXXXXXXX_1.fastq and SRRXXXXXXX_2.fastq files for each sample (with unique sequences of numbers signified by X's), containing RNA-seq reads of a sample. (Each sample here corresponds to one test subject's brain sample). Using a program called FastQC[7], we performed a quality check on these high-throughput RNA-seq reads. After using FastQC, we can determine what sequences to discard or which reads should be removed based on whether the reads are good or bad quality. In our case, to evaluate quality, we focused on the 'pass/fail' score in the 'Basic Summary' section of each of the HTML outputs from FastQC. If a .fastq sequencing file passed this check, then we deemed it to be "good quality". This step is necessary because during the library preparation of RNA-seq, sequences can become contaminated with 3' adapters that are not relevant to our analysis. Therefore, reads need to be cleaned by cutting and discarding these adapters, leaving only the important parts of each read. For cleaning these adapters, we planned to use the tool cutadapt[8]. However, based on the results we received from FastQC, all the .fastq files in our dataset passed the check, so we can confidently proceed with analysis without cutadapt. This could have been the case because the researchers who processed this data possibly already cleaned the adapters from the reads.

Alignment

Once we confirmed that our data is acceptable for analysis based on our outputs from FastQC, we began to align our reads with a reference genome file using a software called Kallisto[9]. Kallisto uses *pseudo-alignment* to align reads with a provided reference genome, which results in faster computations of quantifying expression of the genes that we are focusing on. Specifically, we use the paired-end mode of Kallisto, since each of our samples is provided in pairs of .fastq files. Using the "abundance" data table from each Kallisto output for a sample, which provides the estimated counts of each gene for the respective sample, we created a gene matrix where each row represents a gene and each column represents a sample; the values in the matrix are concatenated from the counts column of each abundance file. In addition, we noticed that there were samples that included multiple runs (2-3), which usually means that the first one or two sequencing runs produced less than ideal sequencing data. As a result, the researchers who performed the RNA sequencing possibly ran it a few more times to gain better reads. Originally, we decided to keep the last sequencing read of the "repeated" samples because, intuitively, the most recent run should be the most ideal. However, this turned out to be a poor decision, as we are unaware of how the researchers decided to name the runs, so we could not make that assumption. As a result, when we created our gene count matrix, we summed the counts of the multiple runs and saved the value under the first SRR run. Then, we dropped the repeated runs, thereby "combining" the results of RNA-seq for the cases where there were multiple runs on the same sample. After creating our gene matrix using the quantified expression values, we filtered out genes that had accession numbers that started with 'NR', as these accession numbers stand for genes that encode non-coding RNA, which we are not focusing on in this study. We also filtered out any genes quantified on the X and Y chromosomes because these chromosomes differentiate between biological gender, which could be a confounding variable causing different gene expression patterns. This resulted in about 17,000 genes for our final differential gene expression analysis.

In our next step, we separated the gene count matrix into nine (3 disorders by 3 brain regions) separate datasets for comparing the different disorders within the different brain regions. For example, in the AnCg control versus schizophrenia data section, the first half of the columns represented the SZ samples, and the second half represented the CTL samples, but all the samples were from the AnCg. The rows of the data were the genes we were quantifying and the

values were the counts of each of the genes. Also, we created another nine data tables that contained information about each of the deceased subjects of our comparisons in row format, namely the age at which he/she died, the post mortem interval (PMI, the amount of time between his/her death and when the sample was collected), brain pH, and the disorder he/she was afflicted by (or if he/she was healthy, then control group). The ordering of the rows of these information datasets was aligned with the ordering of the columns in the count matrices. Some of the data regarding the sample information contained empty values in the pH column, which was handled by imputing the empty values with the mean of the pH of that table. These data tables are important for differential gene expression analysis because the count matrices and sample information are required by DESeq2[10] for comparison of each of the two groups in the next step.

Differential expression analysis and normalization

To study the gene expression changes between samples, we used the R-based package, DESeq2 using default parameters, but with the Likelihood-Ratio Test (LRT) for hypothesis testing. We performed differential gene expression analyses for each comparison of the three different disorders and control groups for the three different brain regions, resulting in nine different runs of DESeq2. Using the datasets created in the processing mentioned above, we used each pair of gene counts and information about the samples as input for the nine DESeq2 Dataset objects. However, before we ran the DESeq command, we found that we had to filter out some genes in order to hone in on the most "useful" genes. We did this by normalizing the full count matrix (with all the genes and all the samples) using the varianceStabilizingTransformation (vst) function implemented in DESeq2. We then computed the L2 (Euclidean) norm of the normalized counts of each gene as a measure of the gene count's variance across the samples. The L2 norm is calculated by taking a square root of the sum of the squared values of a vector, and was used as a single metric of how spread-out the counts were for each gene. The reasoning here was that if the counts had greater variance, it would provide more information to us because it varies across samples. If a gene's counts were constant across samples, then it would probably not be an important gene to study because its count value seldom makes big changes. So, we sorted the genes by this variance measure, then extracted the top 8000 genes with the highest variance for further analysis. Lastly, we used the original (un-normalized) counts for those 8000 genes for running DESeq, since the DESeq command already implicitly normalizes the counts. For all of the comparisons, we used the design formula `~ age + PMI + brain pH + disorder` for our full

model, and `~ age + PMI + brain pH` for the reduced model. Furthermore, we adjusted the results by calculating our p-values against a significance level of 0.05. The Likelihood-Ratio Test computes a p-value for a hypothesis test with the following null and alternative hypotheses:

H_0: There is no statistically significant difference in log fold change (gene expression) between disorder and control groups for a gene.

H_a: There is a statistically significant difference in log fold change (gene expression) between disorder and control group for a gene.

Results

To compare the differential gene expression between each disorder and control within the three brain regions, we plotted histograms (Figure 1) to visualize the p-values from each comparison and observed which comparisons yielded the most significant values. DESeq2 provided us with a p-value for each gene in our analysis. Based on our significance value of 0.05, p-values that were less than that threshold corresponded with genes that were found to be differentially expressed. In other words, we would reject the null hypothesis that there was no change in gene expression between disorder and control. In order to detect the upregulated and downregulated genes between groups, we plotted a correlation diagram (Figure 2) using the Spearman method to visualize the correlation of the $\log_2$ fold changes between disorders and control. Next, we wanted to examine if any of the disorders had any overlapping genes in the AnCg brain region, so we generated a Venn diagram (Figure 3) to compare the differentially expressed genes (p-value <0.05) of the three disorders. For our last visualization for analysis, we plotted a heatmap (Figure 4) of the AnCg region for SZ and CTL patients using hierarchical clustering to investigate the genes that were expressed lowly or highly in the different clusters. We multiplied the counts for each gene by their corresponding $\log_2$ fold change values to get a sense of this.

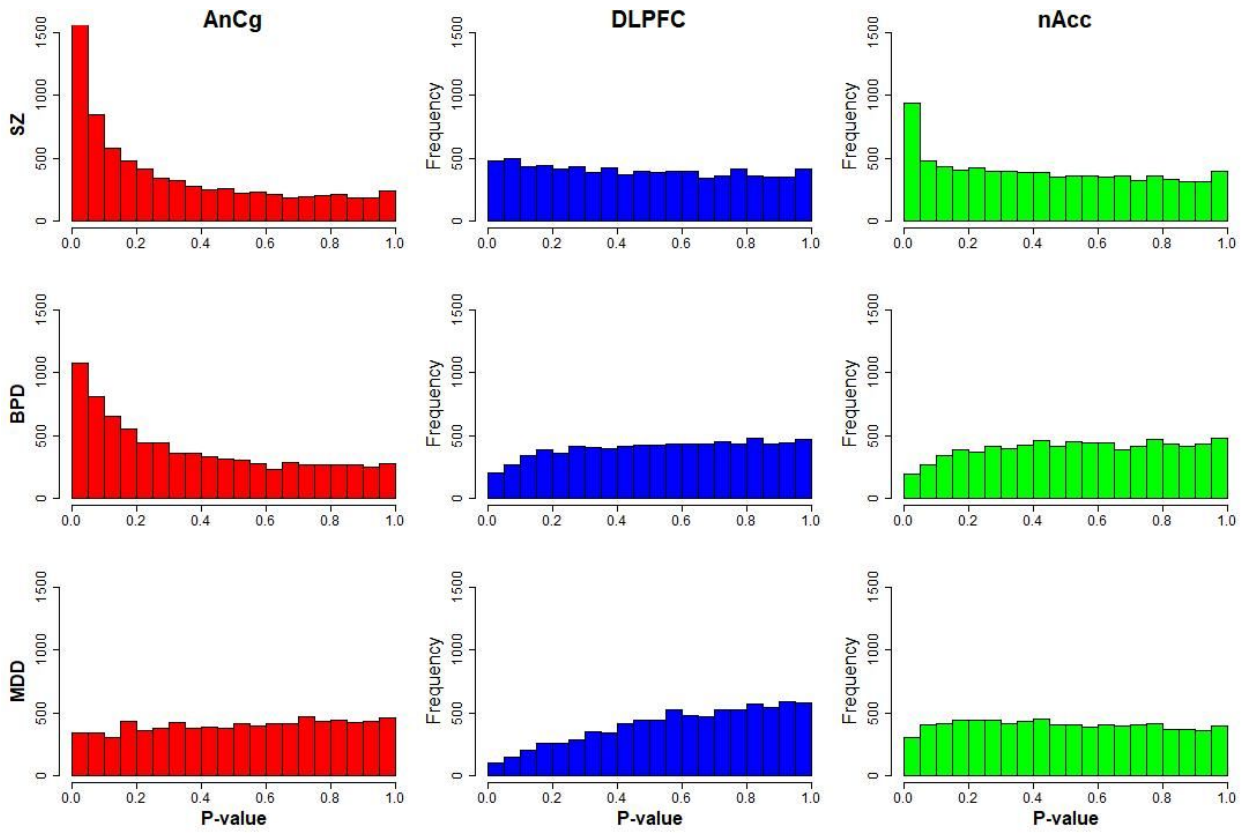

Figure 1 The histograms above show the comparison between case versus control differential expression p-values from the DESeq2 results. Red represents AnCg, blue represents DLPFC, and green represents the nAcc brain regions. The disorders are ordered from top to bottom as schizophrenia, bipolar disorder, and depression.

Based on the histograms in the AnCg region for SZ and BPD against controls from Figure 1, there is a large amount of statistically significant p-values (<0.05), which means that there is a significant difference in gene expression between the disorder and control groups in that brain region. In other words, there are many genes that are differentially expressed between the disorder groups and control in the AnCg. Additionally, the nAcc brain region for SZ against the control group produced a histogram with a great amount of significant p-values (<0.05) as well. This demonstrates that transcriptional changes were most pronounced in the AnCg region for SZ and BPD against the control group; it is also pronounced in the nAcc region for SZ against the control group. In contrast, the AnCg comparison regarding MDD versus control, DLPFC comparisons for all disorders, and nAcc comparison for BPD and MDD versus control did not display a substantial amount of significant p-values. This means that transcriptional

changes are less concentrated in these areas, and therefore there does not appear to be differences in gene expression in those areas between disorder and control groups.

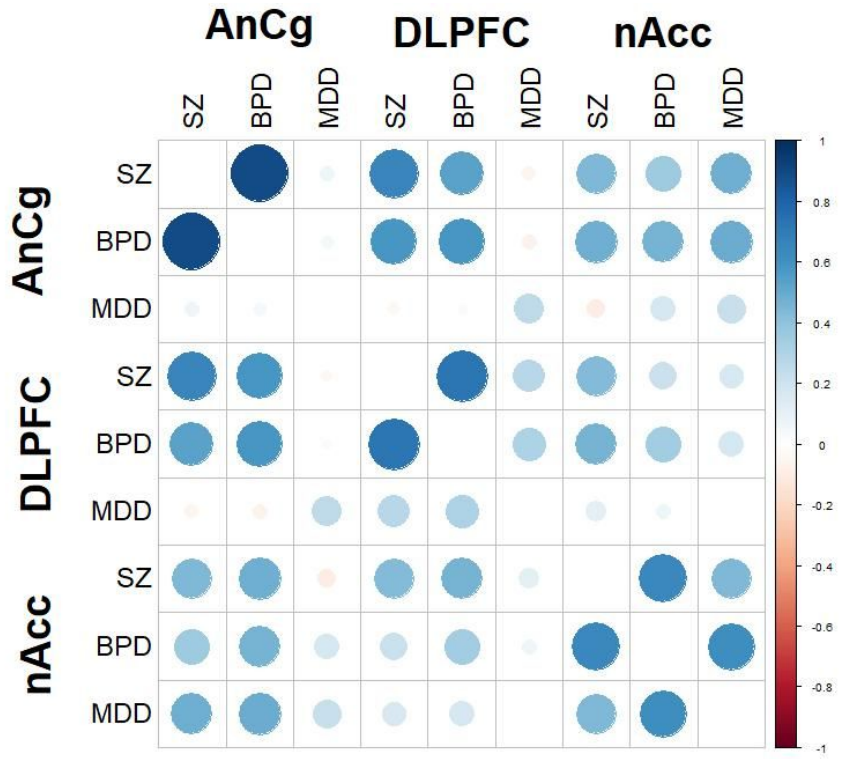

Figure 2 We performed pairwise Spearman correlation of the $\log_2$ fold changes between each disorder and control group in each region of the brain. The color and shape of the circles reflect the strength of the Spearman correlation.

In the pairwise Spearman correlation visualization in Figure 2, the numbers plotted are the correlations of $\log_2$ fold change between disorders and regions. The value of $\log_2$ fold change, calculated by DESeq2, quantifies the difference in gene expression between the disorder and control group for each comparison. The Spearman plot here shows the correlation between the differential gene expression between the disorders. The blue circle means that the two groups are positively correlated in terms of $\log_2$ fold change (both are expressed at the same level, high or low), and the shade of blue shows the extent to which those two groups are correlated. Correlated expression values means that the same genes are being expressed in similar amounts, suggesting broader similarities between the disorders. The red-shaded circles mean that those two groups are negatively correlated, exhibiting that the genes were upregulated (highly expressed) in one group and downregulated (lowly expressed) in the other group; so, the

expression patterns here show opposite trends. The size of each circle represents the correlation of the two groups, in that the larger the circle is, the more positively correlated the fold changes are and the smaller the circle is, the less positively correlated the fold changes are. The same applies to when the circles are red; the bigger the red circles, the more negatively correlated. As we can plainly see from the correlations, schizophrenia and bipolar display very similar gene expression patterns in all brain regions. In the nAcc, bipolar disorder and major depressive disorder also share similar gene expression patterns. This means that individuals afflicted by the two disorders share many abnormally expressed genes in those brain regions. This can be helpful down the line because certain cures or preventative measures could be applied to people with both disorders as they share these contributing genes.

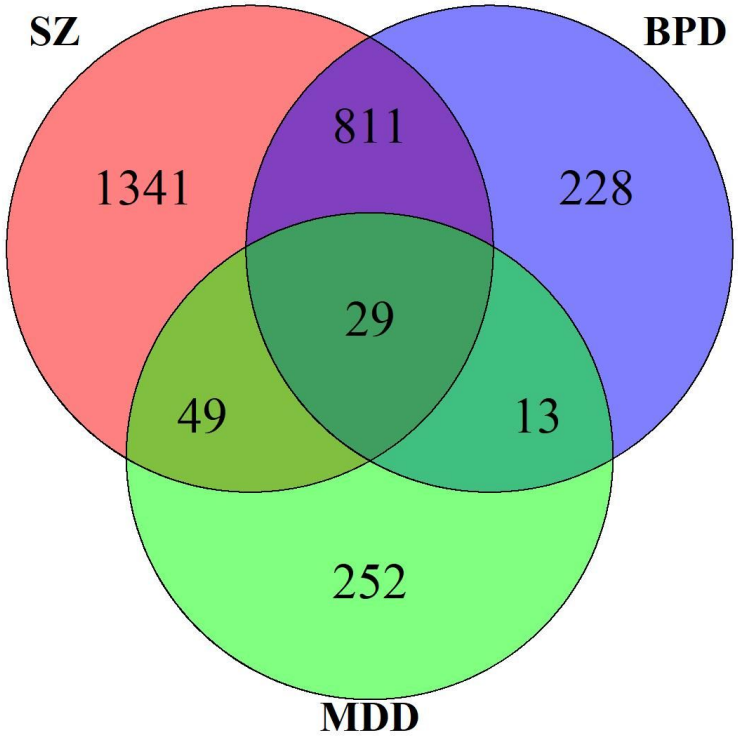

Figure 3 The Venn diagram shows the number of overlapping genes that were differentially expressed in the AnCg brain tissue between the disorders. Schizophrenia represents red, bipolar disorder as blue, and major depression as green.

As shown in the Venn diagram from Figure 3, SZ and BPD contained the most number of overlapping genes that were differentially expressed, which correlates with our findings in Figures 1 and 2 that SZ and BPD showed a considerable amount of genes that were commonly

expressed. Due to the fact that SZ and BPD portrayed notable transcriptional changes in the AnCg and contained the most number of overlapping abnormally expressed genes, we can infer that SZ and BPD disorders share a common underlying genetic signature. On the contrary, MDD and BPD did not showcase a large amount of similarly expressed genes, which means that these two disorders are uncorrelated. Likewise, there is no correlation between MDD and SZ because it did not encompass a huge amount of overlapped genes. Therefore, we can assume that MDD does not share underlying pathology with SZ and BPD within the genomes of these patients. Out of the total 8000 genes that were analyzed, there were only 29 genes that were overlapping, which means that there is not much overlap between all three disorders in terms of gene expression.

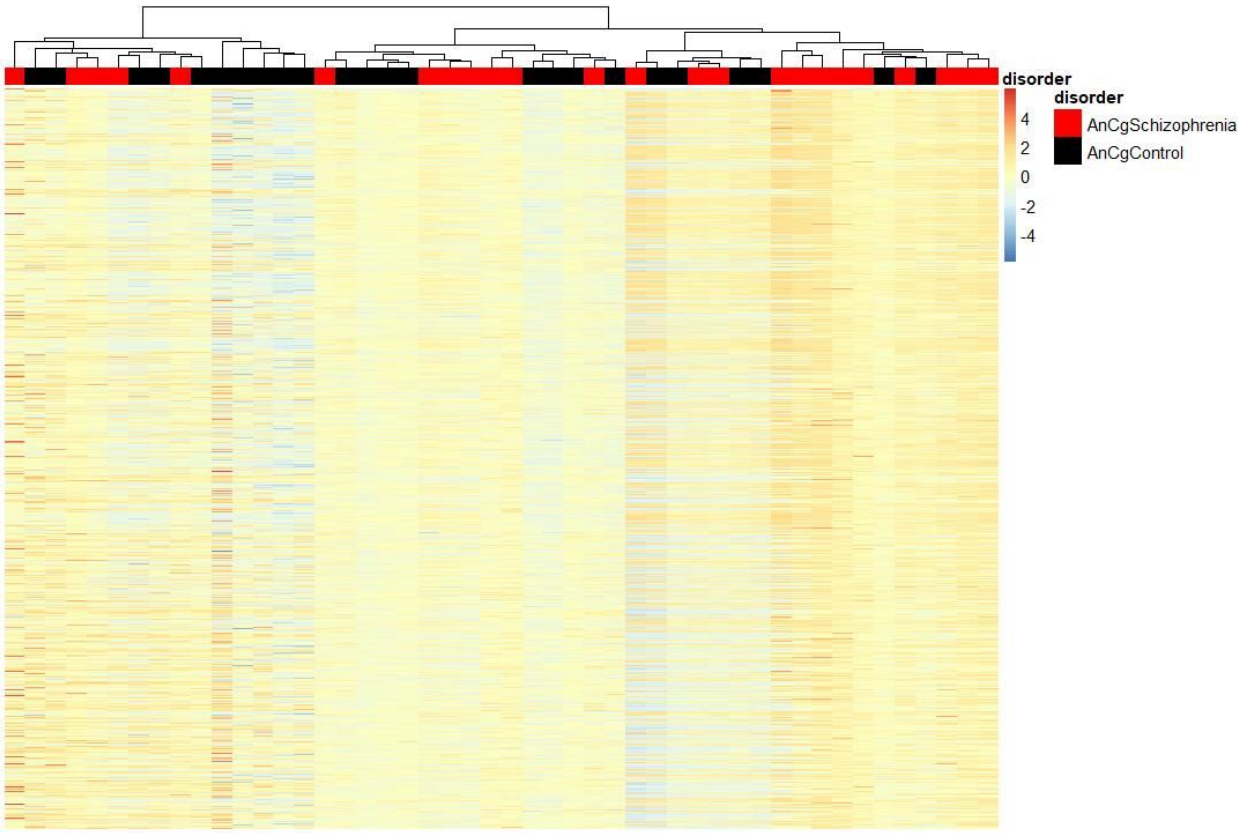

Figure 4 The diagram above portrays a heat map using hierarchical clustering of both 24 SZ and 24 CTL in the AnCg region using variance-stabilized transformed counts multiplied by $\log_2$ fold change of 8000 genes differentially expressed between the two groups. The bluer pixels represent genes that were lowly expressed and the more yellow/orange pixels represent genes that were highly expressed. In the dendrogram, red represents SZ patients in AnCg and black represents CTL patients in AnCg.

The heat map in figure 4 displays a visualization of the distribution of lowly and highly expressed genes in the SZ and CTL samples based on a hierarchical clustering of the two groups. As we can see, in the majority of the AnCg schizophrenia clusters, there is a significant amount of highly expressed genes (represented by darker orange/red pixels). As for the AnCg control clusters, there seems to be more lowly expressed genes (blue pixels) compared to expression in schizophrenia subjects. These genetic differences may be able to explain why patients with schizophrenia experience their symptoms, possibly because some proteins or transcription factors are being produced in excessive amounts.

According to the results of the paper, the transcription factor *EGR1* was significantly down-regulated in SZ patients. Our analysis showed that *EGR1* was one of the more down-regulated genes, but it was not the most down-regulated. Based on our results from DESeq2 output, sorting the genes by lowest $\log_2$ fold change (genes that were lowly expressed), the gene that is responsible for encoding neuronal cell adhesion molecules (*NRCAM*, accession number NM_001037132.4) was one of the most down-regulated for schizophrenia patients in the AnCg. *NRCAM*[11] is an immunoglobulin-like neuronal surface glycoprotein that binds to other cell adhesive proteins during neuronal growth. This typically plays a role in synaptic plasticity, learning, and memory, all of which are affected in schizophrenia patients. One of the most down-regulated genes for bipolar disorder based on the comparison in the AnCg was the gene for encoding kinesin family member 1B (*KIF1B*, accession number NM_001365951.3). *KIF1B*[12] is a variation of a protein that acts as "cargo trains" for cells; specifically in neurons, *KIF1B* transports materials necessary for the transmission of nerve impulses. These proteins are extremely important for brain function, so it would make sense that patients afflicted by SZ and BPD are not producing enough of those proteins. We could not seem to find any up-regulated genes that were correlated with symptoms of mental illness or the brain. However, we did note that down-regulated genes are possibly more important because the absence of important proteins would gravely affect normal bodily functions.

Discussion

The majority of the methods discussed above were influenced by the replication paper, but there were minor adjustments made throughout our experiment that will be discussed here. We specifically made decisions about what we should do when the replication paper was particularly vague about how they performed certain tasks. For checking the quality of our raw

data, we used the FastQC software to perform quality checks of our reads, and decided to label reads as "good quality" if they received a 'pass' score from the Basic Summary section of each FastQC report. As previously mentioned, all of the reads were considered "good quality", which means that there were no adapters identified. In this part, the paper just mentioned that they performed quality checks, but did not specify what metrics they used to gauge "quality", or even what tool(s) they used. This goes back to our suspicion that the researchers probably already cleaned the data before analysis, so they did not elaborate on their quality control of the raw data.

After confirming that the reads of the samples are satisfactory, the researchers in our replication paper aligned and counted the reads using a software called STAR and a reference genome for mapping what genes to count. Similarly, we used the same reference genome to map and count our reads; however, we used a different software for this task, Kallisto. The two softwares differ in terms of functionality: STAR[13] is an alignment tool that aligns reads to the reference genome to determine where in the genome each sequencing read came from, but it does so "non-contiguously". On the other hand, Kallisto does not perform the traditional alignment method. Instead, it pseudo-aligns reads to the reference genome, which produces a list of transcriptomes that are compatible with each read, resulting in faster computations of counts[14]. Due to this difference, it is conceivable that our gene matrix has different counts compared to the gene matrix produced by the researchers using STAR. Because the researchers did not provide their final gene matrix, we cannot make conclusions about how our gene matrices differ. Another discrepancy could be when producing our gene matrix from the Kallisto results of each sample, we explained that we summed the gene counts of samples that were run multiple times, because it was unclear how the researchers dealt with this. However, we did ensure that each of our comparisons had the same number of samples as the paper's comparisons. For example, based on Table S3 from Additional File 2 in the replication paper[5], there were 24 schizophrenia samples from the AnCg and 24 control samples from the AnCg, which matched with our data cleaning results.

To examine gene expression changes, we followed the same methods that the paper employed, which included using DESeq2 with the default settings and the Likelihood-Ratio Test for hypothesis testing. Before performing differential gene analysis for the replication paper, researchers normalized their data by performing a log-like normalization using DESeq2's varianceStabilizingTransformation function and corrected for PRUA based on residuals to a

linear model that regresses PRUA on a normalized gene expression level, and removed non-convergent genes from the data[5]. This was their way of filtering out genes that were not useful for analysis. In our case, we normalized the data using the same varianceStabilizingTransformation function from DESeq2, but computed the L2 norm of the normalized counts of the genes, sorted the genes by norm in descending order, and kept the top 8000 genes for our analysis. We filtered out genes in a different way to keep only the genes that had high-variance counts, because we believed they would be the most informative. We also could not employ their filtering method because we did not have access to PRUA values. The differing results in Figure 1 could be explained by this decision, as we recognized that the paper contained higher frequencies in their histograms. For example, in their AnCg bipolar disorder versus control group histogram, the frequency for low p-value (high differential expression) genes reached about 1500, while our frequency for the same calculation only reached a little over 1000. Furthermore, our system of filtering out genes could also explain the difference in Figure 3, where we have significantly less overlapping genes between the disorders. Although we have less overlaps, the trends of the values seem to be similar.

Finally, the covariates of our full model for DESeq2 were age, brain pH, and disorder, and did not include PRUA (percentage of reads uniquely aligned). Researchers were able to include PRUA as part of their model because STAR calculates it as part of its outputs, whereas Kallisto does not return PRUA in its process. This could also explain why our results may differ because PRUA accounts for changes in length of the sequences and sequence depth. Because we did not provide PRUA as part of our design model, our results may not be as accurate as the replication paper since we performed differential gene expression analysis without taking into consideration the different lengths of reads between sequences.

Conclusion

Our research provided insightful information into the underlying pathologies of these three psychiatric disorders. Based on our analysis from our decisions and methods, we discovered that there is a sizable amount of differentially expressed genes in the SZ and BPD patients individually compared with the control group in the AnCg brain region, as well as a significant amount of differentially expressed genes in SZ patients compared to control in the nAcc brain region. To strengthen these findings, our analysis showed that SZ and BPD have a notable number of overlapping differentially expressed genes in all brain areas. As a result, it is

plausible to conclude that the people with SZ and the people with BPD have key similarities within their genetic data that causes their respective diseases to arise. Furthermore, we identified down-regulated genes in both schizophrenia and bipolar patients that seem to be involved in brain-specific functions. Using these results, professionals could then focus their studies on these specific genes in order to uncover cures or preventative measures that would aid the livelihood of patients who suffer from SZ and BPD.

As mentioned earlier, PRUA was not included as part of our model when performing our analysis because of software limitations. Consequently, it is difficult to identify how the counts of the gene matrices differed in the two analyses for this step, so this could be studied further in future projects. We could also compare the discrepancies between Kallisto and STAR and if they possibly produce different results. In addition to observing differing results from other tools in future studies, this study could be expanded more by investigating how genes are differentially expressed in other brain regions, or in other disorders such as Alzheimer's. Finally, a future potential study could be to incorporate more of the subjects' backgrounds, like smoking history or childhood trauma, into our analyses, which delves into the field of epigenetics and the effect of external factors on our genetics. It is important to investigate these discrepancies as they could potentially play a role in bettering the lives of those affected by mental disorders.

Works Cited/References

1. "Schizophrenia." National Institute of Mental Health, U.S. Department of Health and Human Services, May 2020, www.nimh.nih.gov/health/topics/schizophrenia/index.shtml.
2. "Bipolar Disorder." *National Institute of Mental Health*, U.S. Department of Health and Human Services, Jan. 2020, www.nimh.nih.gov/health/topics/bipolar-disorder/index.shtml.
3. "Depression." *National Institute of Mental Health*, U.S. Department of Health and Human Services, Feb. 2018, www.nimh.nih.gov/health/topics/depression/index.shtml.
4. Resnick, Brian. "How Scientists Are Learning to Predict Your Future with Your Genes." *Vox*, Vox Media, 23 Aug. 2018, www.vox.com/science-and-health/2018/8/23/17527708/genetics-genome-sequencing-gwas-polygenic-risk-score.
5. CB. Caldwell, II. Gottesman, et al. "Post-Mortem Molecular Profiling of Three Psychiatric Disorders." *Genome Medicine*, BioMed Central, 1 Jan. 1990, genomemedicine.biomedcentral.com/articles/10.1186/s13073-017-0458-5.
6. Pritzker Neuropsychiatric Disorders Research Consortium homepage - https://pritzkerneuropsych.org/www/
7. *Babraham Bioinformatics - FastQC A Quality Control Tool for High Throughput Sequence Data*, www.bioinformatics.babraham.ac.uk/projects/fastqc/.
8. Martin, Marcel. "Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads." *EMBnet.journal*, journal.embnet.org/index.php/embnetjournal/article/view/200/479.
9. Pachter, Lior, et al. "Near-Optimal Probabilistic RNA-Seq Quantification." *Nature Biotechnology,* vol. 34, 9 Aug. 2016, pp. 525–527, https://www.nature.com/articles/nbt.3519.
10. I. Lönnstedt, T. Speed, et al. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." *Genome Biology*, BioMed Central, 1 Jan. 1970, genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8.
11. Weledji, Elroy P, and Jules C Assob. "The Ubiquitous Neural Cell Adhesion Molecule (N-CAM)." *Annals of Medicine and Surgery (2012)*, Elsevier, 23 July 2014, www.ncbi.nlm.nih.gov/pmc/articles/PMC4284440/.

12. “KIF1B Gene: MedlinePlus Genetics.” *MedlinePlus*, U.S. National Library of Medicine, 18 Aug. 2020, medlineplus.gov/genetics/gene/kif1b/.
13. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. “STAR: ultrafast universal RNA-seq aligner.” *Bioinformatics*. 2013;29:15–21, https://academic.oup.com/bioinformatics/article/29/1/15/272537.
14. Du, Yuheng, et al. “Evaluation of STAR and Kallisto on Single Cell RNA-Seq Data Alignment.” *G3 (Bethesda, Md.)*, Genetics Society of America, 4 May 2020, www.ncbi.nlm.nih.gov/pmc/articles/PMC7202009/.