Result Replication: Post-mortem molecular profiling of three psychiatric disorders

Abstract

As genetic technologies improve, researchers are able to gain more insight into many diseases and disorders with previously unknown causes. Psychiatric disorders are of particular interest due to the high morbidity rates and the number of disorders that exist which are all clinically distinct, despite sharing many symptoms. While we know that there is a significant genetic etiology based on twin studies, the precise molecular mechanisms are hard to pin down due to the number of genes involved. Previous research has been done in areas of the brain such as the superior temporal gyrus and hippocampus but we can hypothesize potential significant findings in other areas of the brain based on how mental disorders affect behaviors associated with different brain regions. In this study, we look into gene expression based on RNA sequence reads from the anterior cingulate cortex, dorsolateral prefrontal cortex, and nucleus accumbens regions of the brains in post mortem sample from patients diagnosed with Schizophrenia, Bipolar Disorder, and Major Depressive Disorder, creating custom pipelines to clean, align, and run differential expression analysis on RNA samples. We used these results to create graphs which identified a significant number of differentially expressed genes in the AnCg region of SZ and BPD patients as well as found strong correlations between genes of SZ and BPD patients.

Introduction

Background

One of the largest causes of morbidity and mortality annually in the United States is suicide, with close to 90% of people who commit suicide being clinically diagnosed with a psychiatric disorder. Three of the most prominent disorders that affect the population are Schizophrenia (SZ), Bipolar Disease (BP), and Major Depressive Disorder (MDD).

Schizophrenia is an illness that affects a person's mood, feeling, and behavior as well as often causing psychosis in which patients lose touch with reality. While Schizophrenia can affect younger children, typically patients are diagnosed between late teen years to early thirties. In addition to general negative effects on feelings and behavior, symptoms also include hallucinations or delusions, loss of motivation or pleasure, and difficulty in paying attention or decision making. While there are treatments and medications that exist mainly to help suppress psychotic symptoms, because the etiology of the disease is not fully understood, there is

currently no cure. Based on various genetic studies, it is known that certain genes may be associated with an increased risk of the disorder but more work needs to be done in that area before using genes to predict the onset or help with a cure. [1]

Bipolar Disorder is an illness that affects mood, energy, and behavior, often experiencing various "mood episodes" of uncharacteristically intense emotions. There are three types of Bipolar Disorder characterized by different severities of manic episodes but in all three types, patients still exhibit patterns of manic and depressive episodes of varying lengths of duration. During manic episodes, patients can also exhibit psychotic symptoms such as hallucinations which can lead to a misdiagnosis of SZ in some cases. Like SZ, Bipolar Disorder is diagnosed during teen years to early adulthood and while symptoms vary over time, lifetime treatment is usually required. Genetic research has also been done on this disorder to confirm the effect of certain genes on risk.[2]

Unlike SZ and BPD which affect only a small subset of the population, Major Depressive Disorder is much more common yet still has a severe impact on mood, thinking, and behavior. Depression has a wide range of symptoms and the combination of said symptoms as well as circumstances result in the various forms of depression such as Persistent, Postpartum, or Seasonal affective disorder. While people with depression can experience similar extreme symptoms of psychosis as seen with SZ and BPD, Depression is more typically associated with low mood, fatigue, and trouble sleeping. Again, research has indicated a genetic connection with the disease supported by family history increasing risk, and genetic studies. However, circumstances and life events have also seen to be major causes of the disorders, unlike SZ and BPD. [3]

While each of these three disorders are clinically distinct, because of the overlapping symptoms and similar effects on mood and behavior across each disorder, it is highly possible that there are shared genetic causes. This idea has been backed by genome-wide association studies (GWAS), which is used to associate genetic variations with diseases. However, while a GWAS can point to associations in genes, other methods must be used to further specify and characterize genetic etiology in order to better understand underlying molecular mechanisms behind the diseases with the hope of creating better treatments for these types of diseases. [4]

One of the ways to dive deeper into the effect of gene expression is RNA sequencing (RNA-Seq), which provides the ability of transcriptome profiling and quantifying gene expression level. Compared to other transcriptomic methods, RNA-Seq offers high-throughput, low-cost sequencing based on less amount of RNA samples. In addition, RNA-Seq does not rely on existing genetic sequences, thus having the potential to analyze complex transcriptomes and reveal variations within the transcriptome. Therefore, employing RNA-Seq in mental health

research may reveal the commonalities across diseases and differences between patients and controls. [5]

Given that previous studies have found distinct gene expressions in superior temporal gyrus and hippocampus among SZ and BPD patients [6], this study dedicates to discovering the most significant disease-related differences in gene expressions among SZ, BP, and MDD by analyzing RNA-Seq data, laying the foundation for potential therapeutic directions, which can benefit the welfare of millions of patients and their families.

Dataset

The data was collected from post-mortem tissue in three areas of the brain: the anterior cingulate cortex (AnCg), the dorsolateral prefrontal cortex (DLPFC), and the nucleus accumbens (nAcc), which were chosen due to their associations with behaviors affected by said mental disorders. The source of data is the National Center for Biotechnology Information (NCBI), which contains four groups of samples, SZ, BPD, MDD, and control(CTL). Each of the groups has 24 individuals (96 in total)[9]. Some individuals have more than one sample taken, resulting in 94 MDD, 87 BPD, 87 CTL, and 84 SZ samples (352 in total). The sample counts of each disorder-brain region combinations are listed in table 1.

	AnCg	DLPFC	nAcc
Major Depression	30	31	33
Bipolar Disorder	28	28	31
Schizophrenia	28	29	27
Control	30	30	27

Table 1. The samples counts for each disorder-brain region combinations

Methods

Data Preparation

To check the quality of the raw fastq files, we randomly selected 50 samples (1/7 of the data) and ran FastQC[7]. Then we checked the Adapter content test result generated by FastQC. No adapters were found and there were no other major issues that would require dropping any sample. Because of this, we hypothesized that this dataset is already been cleaned by removing the adapters from the sequence though this information is not stated explicitly in the dataset

description. Therefore, we decided to forego using cutadapt which would have taken too long for it to likely not remove many if any adapters.

After the initial quality check step, we fed all the fastq files into Kallisto, a program for quantifying the abundance of transcripts from RNA-Seq data efficiently. Compared to other alignment and quantification tools like STAR, Kallisto only performs pseudo-alignment on the sequencing data and estimates gene counts via bootstrap. We set the number of bootstrap iterations to 100 in order to ensure more accurate estimates.

Data Cleaning

After running Kallisto on all the samples, we first put all of the estimated counts into a gene matrix with columns representing samples and rows representing genes using the pandas package from python. First, the X and Y chromosomes were removed from the matrix since they are not the genes of interest and may bring confounding effects to the subsequent analysis steps. The gene counts were then rounded to the nearest integer and the elements that did not have enough counts across samples were excluded based on a minimum value of 3000. Any gene that had 0 matches in a sample were assigned a pseudo-count 1 to avoid calculation error in downstream analysis. The genes with insufficient expression level were also removed by DESeq2's default independent filtering function.

Data Analysis

We used DESeq2[10], a differential gene expression package, to analyze genes differentially expressed between control and each disorder in each brain region. All the parameters of the DESeq remained to be the default value except that we chose the likelihood ratio test (LRT) as our hypothesis test method. The expression level of a gene is defined by the adjusted p-value given by the result of DESeq2. A gene is considered differentially expressed if the adjusted p-value is less than 0.05. The adjusted p-value is employed since it provides more accurate results. Given the large number of sequences we have in the dataset, normal p-values can be falsely significant, reflecting the result in an imprecise manner[11].

Here, we selected the age of the subject, brain pH, post-mortem interval (PMI) as our covariates to eliminate the difference caused by biological features of the sample that is irrelevant to this study.

After selecting the significantly expressed genes (padj < 0.05) for each disorder-brain region combination respectively, we used those sets of genes to perform hierarchical clustering on each disorder-control pair from the original dataset to see if those sets of genes are representative of a disorder. Before clustering, we first re-fit only the significantly expressed gene sets into the DESeq model with same covariances stated above, then normalized the data using the variance stabilizing transformation (VST) function in DESeq2, which makes the gene count matrix to

have homogeneous variance around the means. Then the euclidean distances between the samples were used in hierarchical clustering. The clustering process is achieved using hclust function in R with ward.D2 method.

Results

PCA

One of the first things we did with the gene count matrices was to perform PCA on the samples in order to get a preliminary look at any groupings or clustering between either the experimental groups or brain region (Fig. 1). Based on a plot of the first and second principal components (which explained 43% and 12% of the variance respectively) we were only able to see a distinct separation between subcortical nAcc samples and the AnCg and DLPFC samples. We did not find a significant enough distinction between any of the samples based on experimental group in either component.

Distribution of Adjusted P-Value for Differentially Expressed Genes

The primary results of DESeq2 we chose to use were the adjusted p-value of each gene in each sample with which we determined a gene to be differentially expressed based on a significance threshold of 0.05. Using this, we were able to create tables and graphs in order to examine and analyze any important relationships across brain samples and experimental groups. We generated 4 plots which looked at different summary statistics explaining the significance of differentially expressed genes. In our first graph we plotted a histogram of differential expression between cases and control groups in order to visualize the proportion and distribution of adjusted p-value (Fig. 2). From this we can see the largest number of differentially expressed genes between patients with BPD in the regions of DLPFC and nAcc. In fact, we found 4732 differentially expressed genes in the DLPFC and 3740 in the nAcc**.** In addition, we found a significant number of differentially expressed genes in patients with SZ in the AnCg region, although to a lesser extent then previously mentioned with 3075 differentially expressed genes.

Spearman Correlation of log2 fold gene expression changes

In our second plot, we calculated pairwise Spearman correlations of $\log_2$ fold gene expression changes between disorders (Fig. 3). The biggest strongest correlation was found between SZ and BPD especially in samples from all three regions of the brain. These results support current knowledge about the similarities between these two diseases in terms of how they affect behavior.

Differentially Expressed Genes Between disorders

To further investigate common genetic connections between disorders, we also created a venn diagram for each brain region with counts for the shared differentially expressed genes (Fig. 4). Again we saw a significant similarity between patients with SZ and BPD in the AnCg with 1345 shared differentially expressed genes compared to the 220 between SZ and MDD and 476 between BPD and MDD. Similar results were not found in the other regions of the brain with more shared differentially expressed genes found between BPD and MDD in each of the other two brain regions. Like in the rest of the plots, these values were based on an adjusted p-value less than 0.05.

Hierarchical Clustering

Our final plot contained hierarchical clustering of differentially expressed genes of SZ and control patients from AnCg samples (Fig. 5). We created the same plots on all other combinations of disorder and brain region but none of the other clusters separated groups as well as this. While not perfect, we can see a clear grouping from the first branches with the right branch classifying only SZ patients. The left side had a moderate accuracy with 30/43 labeled as Control which was still much greater than in any other clusters seen in the other brain region/disorder combinations. This indicates a fairly strong level of expression changes between SZ and the control group.

Figure 1

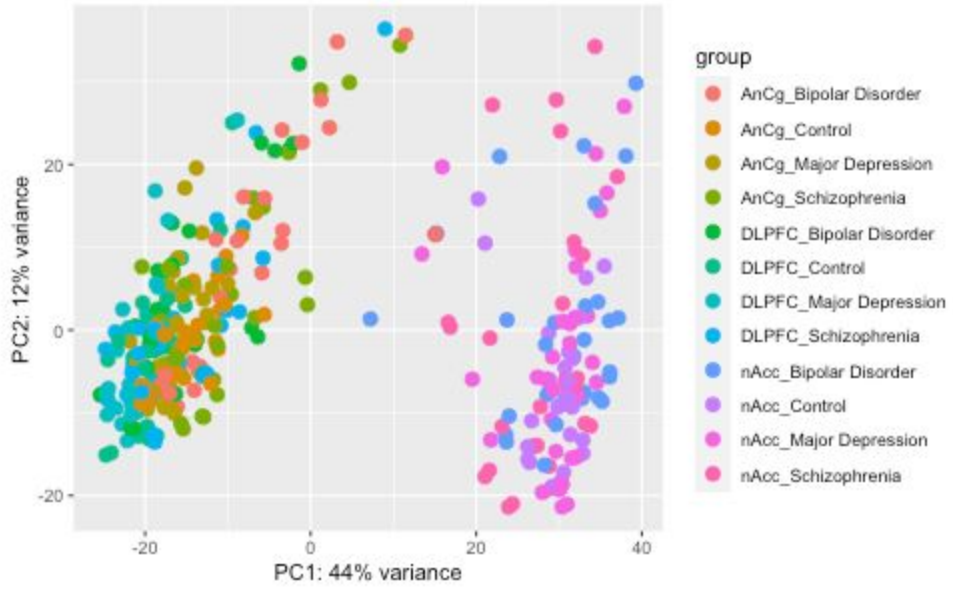

Figure 1. Scatter plot of first two principal components of all genes across all samples colored by both condition and brain region from which the sample was taken. See legend.

Figure 2

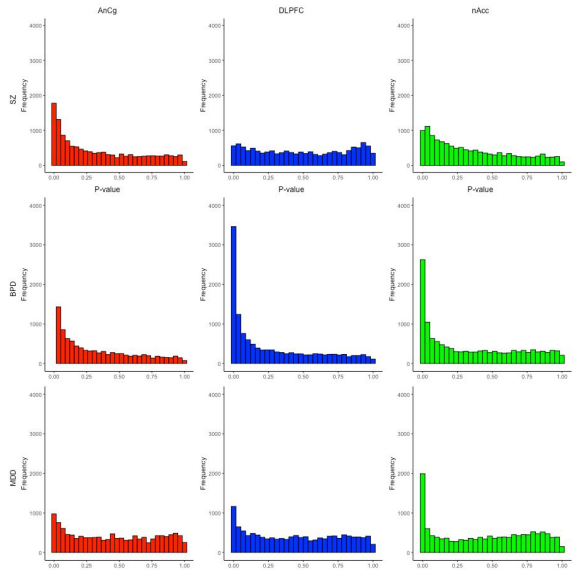

Figure 2. Histogram of the differential expressions (DESeq2 adjusted p-value) of each disorder-control pair for AnCg (red), DLPFC (blue), and nAcc (green). The genes with DESeq2 base mean less than 10 were excluded.

Figure 3

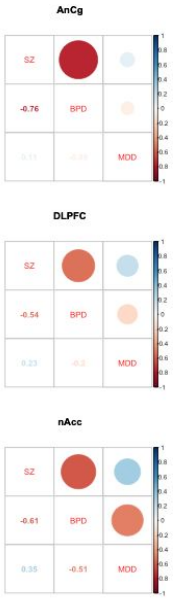

Figure 3. Pairwise Spearman correlations of the log2 fold change in gene expression between each disorder and control in AnCg, DLPFC, and nAcc. The size and the color of the circle in the upper-right part of each plot represent the corresponding value in the lower-left part of the plot.

Figure 4

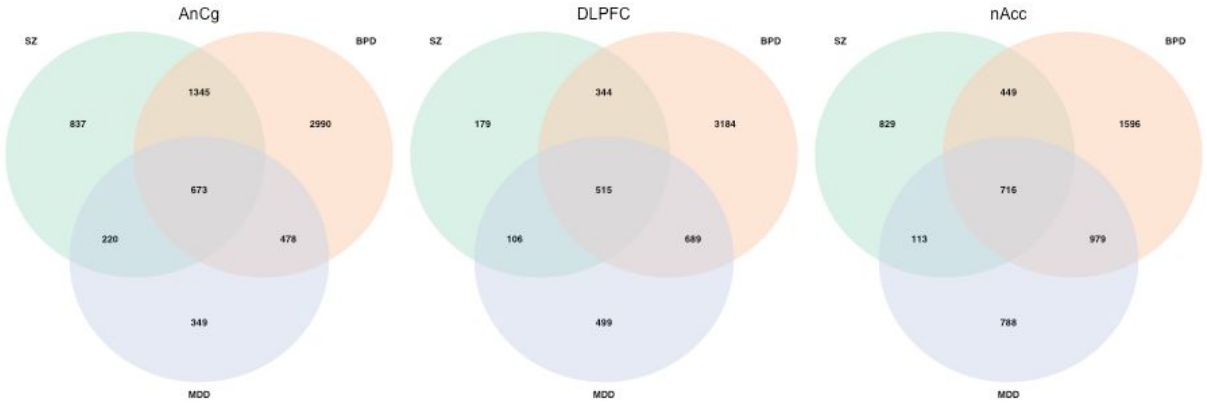

Figure 4. Venn diagrams showing the overlap of differentially expressed genes among SZ(green), BPD(orange), and MDD(violet) in each brain region.

Figure 5

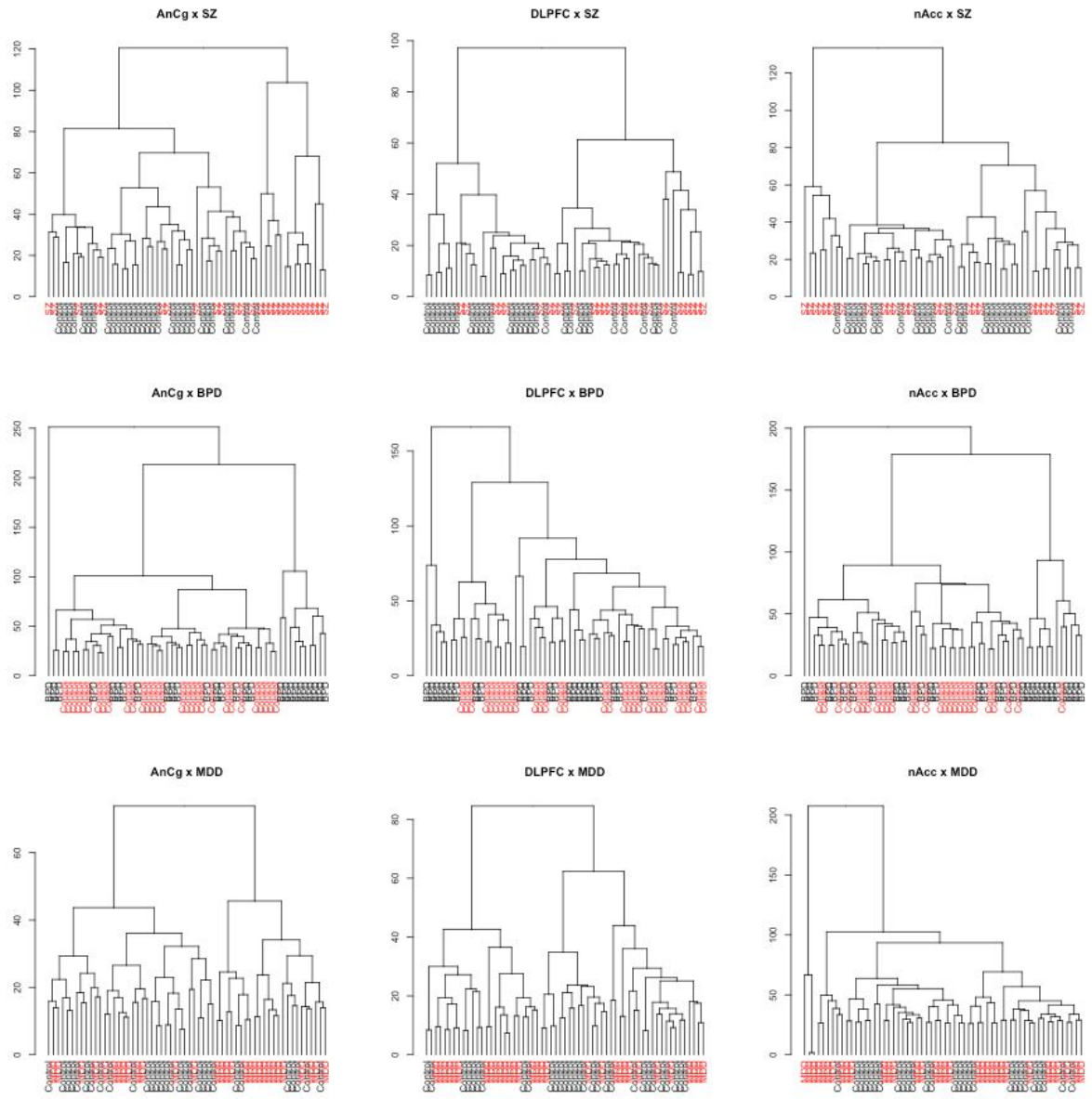

Figure 5. Hierarchical clustering results of SZ (first row, black = control, red = disorder), BPD (second row, black = disorder, red = control), and MDD (third row, black = control, red = disorder) in each brain region.

Discussion

Since we chose to use Kallisto instead of STAR in the alignment phase, we could not acquire the same quality metrics, especially the percentage of reads uniquely aligned (PRUA). This was certainly a factor in not producing the same results from PCA since we could not normalize the

data based on PRUA like the researchers did in the initial paper. We were still able to identify samples between brain regions but did not find as much significance in separation between conditions.

Overall while some of our results were similar to the original research[12] our findings saw much more significance in differential expression of genes in BPD patients particularly in Figures 2 and 4. Throughout the process, we had to make many different decisions that surely had an impact on the end results. As previously mentioned, one of the biggest changes we made was to use Kallisto instead of STAR in the alignment phase. While Kallisto was preferred due to its higher speed using pseudo alignment, we do not know exactly how different the estimated counts produced were than if we had chosen to run STAR, which uses a more traditional alignment technique before generating counts. We also chose not to run quality control using Picard post alignment due to lack of time although in the original paper it did not specify if any samples were removed as a result of this process. Another difference that could have been a large factor in the difference in identified differentially expressed genes was the lack of PRUA as a covariate. This too was because Kallisto does not provide this metadata unlike STAR. Not only was this covariate missing but in the paper they stated that they "corrected for PRUA by computing residuals to a linear model regressing PRUA on normalized gene expression level" which we were of course unable to do. Finally, the results may have been altered by a slight decrease in total samples provided to us (346 vs 352) as well as having to guess on a cutoff for the sum of genes across samples greater than 3000 which was mentioned in the paper but not specified. In the future we would test the results against other software such as STAR or sleuth for the differential analysis phase to see how the analysis and classification of differentially expressed genes would change.

Despite the problems and differences mentioned above, many results were still significant, mainly in Figures 2 and 3 which showed a strong correlation between gene expression in SZ and BPD patients. In both analyses, the similarities between SZ and BPD were strongest in the AnCg with the largest pearson correlation and the largest number of shared genes differentially expressed across all samples. One perplexing difference, however, was that the correlation between the samples in Figure two were almost all negative which would typically indicate that a gene is up-regulated in one condition and down-regulated in another. In the original paper, while the correlations were of similar magnitude, they all had a positive correlation and we were unable to figure out the cause behind this difference. Figure 5 also showed a moderately successful result with a clear distinction between clusters although to a lesser extent than the original, possibly due to the reasons mentioned above.

References

1. https://www.nimh.nih.gov/health/topics/schizophrenia/index.shtml
2. https://www.nimh.nih.gov/health/topics/bipolar-disorder/index.shtml

3. https://www.nimh.nih.gov/health/topics/depression/index.shtml
4. https://www.vox.com/science-and-health/2018/8/23/17527708/genetics-genome-sequencing-gwas-polygenic-risk-score
5. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10, 57–63 (2009). https://doi.org/10.1038/nrg2484
6. Hwang, Y., Kim, J., Shin, JY. et al. Gene expression profiling by mRNA sequencing reveals increased expression of immune/inflammation-related genes in the hippocampus of individuals with schizophrenia. Transl Psychiatry 3, e321 (2013). https://doi.org/10.1038/tp.2013.94
7. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
8. https://pachterlab.github.io/kallisto/manual
9. https://www.ncbi.nlm.nih.gov/sra?term=SRP073813
10. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15, 550 (2014). https://doi.org/10.1186/s13059-014-0550-8
11. Noble, W. How does multiple testing correction work?. Nat Biotechnol 27, 1135–1137 (2009). https://doi.org/10.1038/nbt1209-1135
12. 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