***Results***

As we have previously established, there is a lack of knowledge and understanding of BD which results in an insufficient diagnosis and treatment. We are hoping that we could uncover some of the mysteries of this disorder using machine learning algorithms – whether it be identifying genes which are associated specifically with BD, enriched pathways which are affected by BD, new ways to classify BD into subtypes based on biological differences etc.

We used RNA-seq gene expression data from E-GEOD-78936 (Hu et al., 2016) and E-GEOD-53239 (Akula et al., 2014) to compare samples of different brain areas from BD patients, schizophrenia (SZ) patients and healthy individuals.

First, we aim to identify genes that are differentially expressed in BD patients, SZ patients and healthy individuals. We performed the differential expression analysis using DESeq2 (Love et al., 2014) by using the raw count data and corresponding metadata regarding the diagnosis and brain area of each sample.

We have found that there is only one gene which is significantly highly expressed in BD compared to SZ, between BD and healthy people, we have found 4 significantly highly expressed genes (figure 1). In both those comparisons, we have identified only the gene MTND6P4 as a common significant gene.

In addition, we have examined the difference between the highly expressed genes in different areas of the brain in BD patients relative to healthy individuals. As the plots show, there are different genes which are highly expressed in different areas of the brain and can indicate the presence of BD (figure 2).

Those two findings show us very clearly that biomarkers for BD do exist and that they are not that rare or even hard to find. Unfortunately, the research literature does not support our claims, nor it disagree with it but rather it lacks any references regarding those genes in the context of BD. It is possible that this missing validation is caused due to insufficient research of the topic as we have mentioned before.

Using the DESeq2 (Love et al., 2014) we have also performed PCA which is a method of visualizing high-dimensional data in a more simplistic and easier to conceive way. We have plotted three PCAs: one is classified based on the diagnosis of each sample, the second is classified based on the brain area which is the source of the samples and the last one is based on both the diagnosis and the area of the samples. The PCA plots have showed complete chaos which actually validates that the normalization of the two distinct datasets we have based our study on did not separate them into two clusters.

|  |  |
| --- | --- |
|  |  |

Figure 1: two volcano plots that display the genes that were most differentially expressed in BD patients relative to healthy individuals (on the left) and SZ patients (on the right).

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

Figure 2: six volcano plots that display the genes that were most differentially expressed in different areas of the brain relative to each other.

After we have completed the differential expression analysis, we chose three of the highly expressed genes that we have found in the first section of the analysis (MTND6P4, LINC0234, MT1X) and using a violin plot, we compared the expression levels of the genes in the three populations we are dealing with (BD patients, SZ patients and healthy patients). The plots looked a bit odd because of a small number of samples which had extremely high counts, so we had to cutoff the plots in order to see them properly. As expected, the plots of BD and SZ were very similar to each other while very distinct from the healthy individuals (figure 3). It seems that the MTND6P4 is not actually differentially expressed between the BD and SZ patients but rather that the difference has occurred as a result of the rogue samples we have removed. If that's so, then unfortunately we have failed to find any gene that acts as a biomarker to distinguish between BD patients and SZ patients.

|  |  |
| --- | --- |
| Before cutoff | After cutoff |
|  |  |
|  |  |
|  |  |

Figure 3: in each column there are three violin plots that correspond to the gene expression of three significant genes (top: MTND6P4, middle: LINC0234, bottom: MT1X).