        CBIO311/CBIO310 Advanced               programming and data analysis

                   Information Technology and Computer Science (ITCs)

    "BIOINFORMATICS INSIGHTS INTO     ALZHEIMER'S DISEASE AND        SCHIZOPHRENIA\*

|  |  |
| --- | --- |
| Jana elhadad | 211002077 |
| Alaa ashraf |  |
| Malak Ibrahim | 221001266 |
| Ahd Adel elshanhoury |  |

     1-Introduction

Alzheimer’s disease is a progressive neurodegenerative disease that leads to progressive cognitive decline. It is a common cause of dementia and particularly affects memory and cognitive abilities. Alzheimer’s usually develops slowly, with short-term memory, such as forgetting recent events or people’s names. As the disease progresses, individuals have difficulty performing daily tasks such as driving or managing finances. At advanced stages, patients may struggle to recognize family members or understand their surroundings. Researchers believe that Alzheimer's disease is caused by an accumulation of abnormal proteins, such as amyloid beta and tau, in the brain, which cause brain cell damage and death. Alzheimer's disease is associated with changes in the brain, such as muscle spasms, shrinkage in the areas of the brain responsible for memory and cognitive function and although many factors contribute to the development of the disease, genetic factors, factors with environmental influences and age being some of the most important factors that increase the terror. Schizophrenia, on the other hand, is a serious mental illness that affects a person's thinking, feelings, and behavior. Individuals with schizophrenia have difficulty distinguishing between reality and hallucinations, leading to delusions (seeing or hearing non-existent objects) and delusions (false, irrational beliefs) Other symptoms include being absorbed isolation, lack of emotional expression, and difficulty in making day-to-day decisions. Schizophreniausually appears in late adolescence or adulthood, and its exact cause is unknown. However, a combination of genetic, environmental and brain chemical factors are thought to contribute to the disorder. Schizophrenia is also associated with an imbalance of chemicals such as dopamine in the brain, which plays an important role in the regulation of thought and attention. Although Alzheimer's disease is primarily associated with cognitive decline and memory loss, dementia is characterized by severe psychosocial symptoms, including mood disorders and emotional issues and when diagnosed early and intervention critical in managing both situations.In the case of Alzheimer's disease, there are medications aimed at slowing the progression of the disease, physical therapy, cognitive exercises and other supportive therapies Antipsychotics and mood in therapy treats depression, which can help patients manage symptoms and improve their quality of life.Increasing awareness of both conditions is important, as early diagnosis and consistent support can help patients and their families cope with the day-to-day challengesAlthough Alzheimer's disease is primarily associated with cognitive decline and memory loss, dementia is characterized by severe psychosocial symptoms, including mood disorders and emotional issues and when diagnosed early and intervention critical in managing both situations.In the case of Alzheimer's disease, there are medications aimed at slowing the progression of the disease, physical therapy, cognitive exercises and other supportive therapies Antipsychotics and mood in therapy treats depression, which can help patients manage symptoms and improve their quality of life. Increasing awareness of both conditions is important, as early diagnosis and consistent support can help patients and their families cope with the day-to-day challenges. Both Alzheimer’s disease and dementia are complex diseases with multiple genetic factors. For example, the APOE gene, which is known to increase Alzheimer's risk, may also play a role in dementia, particularly the COMT gene, which is associated with cognitive impairment and dopamine metabolism, affecting cognitive function and contribute to Alzheimer's and dementia symptoms The study of has been conducted on dementia but also thought to play a role in Alzheimer's disease by synaptic health and cellular signaling affecting GRM3, which affects glutamate signaling, is associated with cognitive dysfunction in both conditions, while BDNF is a major contributor to both neurodegeneration and memory Alzheimer's and dementia Po even if it is regulated It is known, as additionally as a contributor to renal disease, the metabolism of the MTHFR gene involved in folate and risk to high levels associated with Alzheimer's disease by elevated homocysteine and dementia by impaired neurodevelopment Synaptic proteins including synapsin and synaptophysin Play an important role in both diseases through neurotransmitter release and synaptic plasticity monitoring Finally, inflammation pathways involving IL-6 and TNF-. α is particularly common in both diseases, . a shared genetic contribution to neurodegeneration in Alzheimer's disease and psychiatric traits in dementia highlight, and demonstrate, the complex interplay of biological pathways underlying both disorders that they can be targeted for further research and treatment strategies

                                                                          Methods of programming

**3.1. Data Collection and Preprocessing**

Gene Expression Omnibus (GEO) databases were used to gather datasets for schizophrenia (SCZ) and Alzheimer's disease (AD). GSE48350 was specifically chosen for AD because it contained transcriptome information from samples of brain tissue, whereas GSE263180 was chosen for SCZ because it concentrated on neural gene expression patterns.  
The number of control and illness samples chosen from each dataset was the same for consistency. Technical artifacts and outliers were eliminated from the raw data by applying quality control procedures. To establish comparability and account for batch effects, log-transformed values were used to normalize the expression data. The study employed GEO2R, an online tool offered by GEO that uses statistical techniques and linear modeling to identify differentially expressed genes (DEGs), to perform differential expression analysis (DEA). Based on corrected p-values (< 0.05), DEGs were chosen.

3.2. Similarity analysis and the identification of shared genes  
Lists of DEGs from each dataset were compared using R in order to find common molecular signatures between AD and SCZ. The genes that intersected were found to indicate possible common regulatory networks or biological pathways in both illnesses. In order to evaluate the overlap of DEGs between the datasets, similarity analysis was quantified using statistical metrics.  
By grouping comparable genes according to their co-expression patterns and functional functions in both disorders, functional similarity was further examined. This stage shed light on molecular processes that might be involved in the pathophysiological parallels between AD and SCZ that have been noted.

3.3. Analysis of Functional Annotation and Enrichment  
To functionally annotate shared genes, ShinyGO, a comprehensive bioinformatics platform, was used. Gene Ontology (GO) analysis was used to categorize genes according to cellular components, molecular roles, and biological processes. The overrepresented GO keywords and pathways found by enrichment analysis highlighted important processes in both illnesses.  
Furthermore, phylogenetic tree construction tools were made available by ShinyGO, enabling evolutionary insights on shared genes. In addition to highlighting their possible importance in neurodegenerative and mental illnesses, our investigations demonstrated whether shared genes are conserved across species.

3.4. Analysis of Pathways  
The Genomes (KEGG) database was used to conduct pathway enrichment analysis in order to learn more about the biological pathways connected to common genes. The molecular networks and interactions that may underlie both AD and SCZ were clarified by KEGG analysis, which revealed pathways that were strongly enriched with the shared genes.  
Prioritization of key pathways was determined by biological relevance and statistical significance (adjusted p-value < 0.05). Because they are known to be involved in the pathophysiology of both illnesses, pathways pertaining to inflammation, synaptic plasticity, neurodevelopment, and oxidative stress were of special interest.

3.5. Validation in Statistical and Bioinformatics  
False discovery rate (FDR) corrections for multiple comparisons were among the several criteria used in the statistical validation process to guarantee the dependability of the results. Consistency was confirmed by cross validating the results of the GEO2R and KEGG studies using different bioinformatics tools.  
The common genes, functional annotations, and pathway correlations between AD and SCZ that were found were made sure to be reliable, repeatable, and biologically significant by this methodical methodology.

**3.Result:**

3.1

**This image is a volcano plot, a common graphical representation in bioinformatics and genomics science, which is usually the case for the comparative analysis of gene expression. So how do we read such a plot?**

**1. Axes:**

**On the x-axis, the log2 fold change (Log2FC) for gene expression is displayed, It illustrates the extent to which one measurement of gene expression, say healthy samples against schizophrenia samples, is after some given amount of time increased or decreased. If the measure is positive, it indicates upregulation; whereas, negative measures imply downregulation.**

**On the y-axis, the -log10 p-value, which presents the significance levels of each of the changes. The higher the value, the more significant the output will be.**

**2. Dotted Lines:**

**A horizontal dashed line represents the threshold of significance (for example, a p value cutoff). Any point above this line is categorized as statistically significant.**

**Vertical dashed lines in graphs exhibit the boundaries for fold-change which is normally set at a threshold of ±1 or two-fold change.**

**3. Color Gradient:**

**The color scale depicts levels of significance ranging from red denoting the most significant genes to blue where the genes are not that significant.**

**4. Data Points:**

**A gene is represented by a dot which in turn, is included in a specific large number of populations. Consider a situation where substantial changes in the gene expression occur. Such genes will most likely be located on the uppermost areas of the right and left side (red Dots).**

**So this plot has been classified as that for schizophrenia which suggests that it portrays genes expressed differently on patients of schizophrenia in relation to control sample.**

 3.2

**The image depicts the boxplot for data obtained from the dataset GSE263180. It involves the following:**

**1. Objective of Boxplot:**

**A boxplot is a type of graphical portrayal that shows the central tendency and the dispersion of values among different samples or conditions. Each sample is diagrammatically shown as a box which is positioned vertically along the x axis with the label.**

**2. Axes:**

**The x axis shows sample IDs (for example: GSM187219 up to GSM187244) which probably represent unique biological samples in a gene expression experiment.**

**Each of the y axis depicts log10 transformed raw counts (the count was increased by 1 to eliminate log of 0). This is a familiar standardization procedure to enable the proper visualization of the values of gene expressions.**

**3. Features of Each Box:**

**The box shows the interquartile range (IQR) (25th to 75th percentiles) of the data.**

**The horizontal line inside the box represents the median expression value for that sample.**

**‘Whiskers’ which are lines that extend from the edges of a box show the spread of the data except the extreme ones.**

**4. Interpretation:**

**This boxplot most probably ensures the quality or the uniformity of the data set by comparing averaged data for each sample.**

**Confined medians and IQRs are definite signs of better normalization which means the data is cleaned up for further processing.**

**This plot shows the work of the RNA-seq or microarray, and it helps ensure that the samples distributions are in line with expectation after some preprocessing.**

 3.3

**This image is a volcano plot for the case of a sample and control gene expression dataset. It is mostly utilized in bioinformatics and gene expression analysis. The objective is to locate a statistical significance and the extent to which the change has occurred in a given dataset. Here we have provided the description for the plot.**

**Title**

**As statistic with respect to log2 fold change, control sample legid is provided in the title which states that the there are two groups being compared (control vs Sample)**

**Axes**

**X axis, log2 fold change (log2FC): This measures the amount a gene's expression varies between two groups. For instance, Onthis metric, Both log2FC x and the positive axis upregulates expression when there have been changes on the gne group between the sample and control.**

**However, the sample group is lower, while a decrease in the x value and the left side of the log2 side indicates that?.**

**Y axis, log10 negative of Padj: This graphically represents to an extent an adjusted p value which is Padj although on a minus log to the base 10 scale.**

**Interpreting it, the greater the parameter the gene has witnessed a more statistically significant alteration.**

**Data Points**

**Red dots, log2fc >= 0 and log2fc <= 1 and Padj < .05 .**

**Blue dots , log2 fc <= - 1 and Padj < 0.05.**

**Black dots, Padj greater than equal to 0.05 and less than 1.**

**Threshold**

**More domains are represented as Padj and highlighted as techniques with drastic evidence depicting changes at P values lesser or equal to 0.05. Conseqently genes that change in such a manner are represented in either of two colors red and blue depending on the color in which the change occurred.**

**In this regard, this volcano plot plays a crucial role in determining the genes which are differentially expressed between the control and sample.**

 3.4

**This is the MA plot for RNA sequencing and gene expression analysis, showing the relationship between the expression level and fold change of genes between two groups. The explanation of the key elements is done here.**

**Title:**

**The title itself, "GSE171110: control vs Sample," tells that there is a comparison to be made between a control group and a sample group.**

**Axes:**

**x-axis (log10(mean of normalized counts)):**

**The log10 average expression of a gene. Genes that have a higher average expression are plotted more to the right. y-axis (log2FoldChange): log2 fold change (log2FC) is a measure of how much the expression of a gene has changed between the control and sample groups. A positive value means upregulation (gene is expressed more in the sample group). A negative value means downregulation (gene is expressed less in the sample group). Data Points**

**Red dots: genes with significant upregulation, log2FC > 0 and Padj < 0.05. Blue dots: genes with significant downregulation, log2FC < 0 and Padj < 0.05. Black dots: genes that are not significant, Padj ≥ 0.05. Threshold:**

**Genes with Padj < 0.05 (statistically significant) are colored in either red or blue depending on the direction of change.**

**Black horizontal line at log2FC = 0 corresponds to no change in expression.**

**What the Plot Shows:**

**Lowly expressed genes are scattered on the y-axis because their fold change is not very reliable, due to statistical noise.**

**Highly expressed genes lie close to the horizontal line, reflecting fewer significant changes.**

**Red and blue points highlight genes whose expression is found to be differentially expressed with statistical significance.**

**This plot provides a better way of visualizing expression changes that are meaningful to the genes in question by taking into consideration the level of expression**

**3.5**

**This figure represents a Venn Diagram-like summary for differential gene expression analysis using DESeq2. The details are described below.**

**Title:**

**The title itself, "GSE171110: DESeq2, Padj < 0.05," has already suggested that the analysis has been done in the dataset GSE171110 by using the DESeq2 tool, choosing an adjusted p-value threshold (Padj) of less than 0.05 for considering the genes as differentially expressed.**

**Circle Content:**

**control vs Sample: In the circle marked "control vs Sample, the number contained is 8536 and represents the genes that had a Padj <0.05 hence statistically significant between the comparison groups of control versus sample. Overall Genes**

**Total: 17582**

**These are the overall number of genes analyzed through this study. Of this total number of genes analyzed, 8536 have met the significance threshold**

 3.6

**This is a \*\*boxplot\*\* of normalized count data across the dataset \*\*GSE171110\*\*, comparing two groups: \*\*control\*\* and \*\*Sample\*\*. This plot summarizes several key features:**

**### Key features:**

**1. \*\*X-axis\*\*:**

**- Shows sample identifiers, like GSM5218990, GSM5218991, etc., which likely represent different experiments or biological samples across this dataset.**

**- Green="control" samples, purple="Sample" group.**

**2. \*\*Y-axis\*\*:**

**- Plots the normalized counts on a \*\*log10 scale\*\* to depict the magnitude of data values post-normalization; for instance, gene expression counts.**

**3. \*\*Boxplot elements\*\*:**

**- Each box plot represents the distribution of the normalized data for a sample.**

**- \*\*Box\*\*: Represents the interquartile range, IQR, between 25th and 75th percentile.**

**- \*\*Horizontal line inside the box\*\*: Median of data.**

**- \*\*Whiskers\*\*: Extend to range of data points to a value 1.5 times the IQR.**

**- \*\*Dots (if any)\*\*: Outliers falling out of the whisker range.**

**4. \*\*Comparison\*\*:**

**- Both groups appear to have distributions of normalized counts which are similar, with overlapping medians and ranges to a large extent. This would suggest that normalization was effective and comparable across the two groups.**

**### Interpretation:**

**- The plot ensures that, upon normalization, the data across samples are uniformly distributed.**

**- If there were any marked differences in box heights, this may suggest variation in the quality of the data or even biological differences.**

 3.7

**The presented image shows a UMAP (Uniform Manifold Approximation and Projection) plot with label GSE171110 with the number of neighbors (‘nbrs’) equal to 9.**

**### Key features of the plot:**

**1. \*\*Groups\*\*:**

**- It can be discerned that there are two definite clusters:**

**- \*\*Cluster of Control\*\*: Dark green dots.**

**- \*\*Cluster of Sample\*\*: Purple dots.**

**- All these groups seem to have distinct distributions or characteristics as the separation among them is well marked.**

**2. \*\*Axes\*\*:**

**- X and Y are the umap dimensions that are low dimensional representation of tissues embedded in higher dimensions. These dimensions are not actual biology per se but are quantitative ways of differentiating the groups.**

**3. \*\*Interpretation\*\*:**

**- The divergence among the clusters suggests that there is sufficient profiling data to differentiate the control from the sample group.**

**- This is probably due to segmentation or several dimensions put together to be examined in the sample.**

**4. \*\*Data Source\*\*:**

**- The label “GSE171110” seems to be the target of some study or dataset from some repository(lowly GEO). Could have been pertaining biological or experimental data.**

4.Discussion: Differential Gene Expression Analysis in Schizophrenia

Gene expression studies are animportant way to understand the molecular basis of complex diseaseslike schizophrenia. Gene expressioncomparisons between groups, forexample, patients versus healthy controls, will be carried out. The volcano plot, boxplot, and Venn diagram are a few of the graphical displays that offer complementary views on the data to help in filtering, visualization, and interpretation of changes in gene expression.

4.1Volcano Plot: Identification of Significant Gene Expression Changes

The volcano plot is a very commonpractice for visualizing differential expression of genes. On such a plot, the X-axis is the log2 fold change, describing how much a gene has changed its expression between groups; a positive value represents up-regulation-a gene is more active in schizophrenia-while a negative value implies down-regulation, or less active. Along the Y-axis is plotted the -log10 p-value, showing the statistical significance of the change. The greaterthe number, the larger the statistical significance.

The design of the plot enables us to quickly identify genes that havesubstantial expression changes withstrong statistical significance. Red dots highlight highly significant, upregulated genes, and blue dots show downregulated genes. Black dots represent genes that are not significantly changed between the groups. Dotted lines in the plotrepresent significance thresholds and fold-change limits that further help in filtering the data to easily focus on the most biologically relevant genes.

This tool is very useful when dealingwith big datasets to find main genesthat could be related to schizophrenia. The analysis could be automated using Python libraries such as pandas for manipulation and matplotlib for plotting, which would enable the researcher to focus on genes showingmeaningful differences in expression.

4.2Boxplot: Data Quality and Normalization Check

Boxplots bear important informationfor quality checking and normalization in gene expression studies. A boxplotvisualizes the distribution of gene expression for the biological replicate individual samples; the X-axis shouldreflect the sample ids, while on the y-axis will fall the log-transformed value for gene expressions.

Boxplot provides an immediate display for quality checking in RNA-seq or microarray data analysis. The box depicts the IQR-the middle 50% of data-whereas the whiskers extend to the range of the data without outliers. The median line inside the box givesthe central tendency of gene expression. This plot helps evaluatewhether the data has been properly normalized; if the data across samples is consistent in distributions with narrow IQRs and medians, then the normalization went well.

This boxplot from the analysis usingthe GSE263180 dataset shows no major outliers, and the geneexpression is similarly distributedacross samples; this would suggest that the pre-processing steps were applied successfully. This is animportant precursor to more complex analyses in that it ensures the data is prepared for further exploration and interpretation.

4.3Venn Diagram: Overlap of Significant Genes

The above Venn diagram gives an intuitive overview of the number of genes that were differentially expressed between the sample and control with significance. Here, it canbe seen from the diagram that, from a total of 17,582 genes, 8,536 genes have a p-value less than 0.05. This shows the importance of filtering out those genes which are not significant, therefore leaving only those genes thathave some meaningful expression change and should be considered for further study.

The Venn diagram summarizes the results and provides an idea about thenumber of genes that are differentially regulated between the groups andhence a starting point for further biological interpretation.

4.4MA Plot: Expression Levels and Fold Changes

Complementing the volcano plot, the MA plot offers extra information onthe gene expression changes. The plot here combines mean expression withlog2 fold change to provide a more detailed overview of gene expression. It helps in identifying genes highlyexpressed that generally show more stable and reliable fold changescompared to the lowly expressed genes that may have a lot of noise. MA plot, therefore, permits a combinedvisualization of the magnitude of gene expression changes and their reliability across the dataset.

4.5Conclusion

Among them, the volcano plot would include the boxplot, Venn diagram, MA plot, and other ways to analyzedifferential gene expressions. Each figure gives different insights: avolcano plot for highlighting the significant alterations in genes, aboxplot that assured data quality and normalization, a Venn diagram topresent the summary statistics of the amount of significant genes, an MA plot on the dependability of changes in expressions. These bioinformatic tools allow researchers to sift out and prioritize genes that showedmeaningful, statistically significant change and further help in deducingthe molecular mechanisms involved in schizophrenia and other complex diseases.