META-ANALYSIS OF GENE EXPRESSION DATA FOR ALZHEIMER'S DISEASE

Project report submitted in partial fulfillment ofthe
requirement for the degree of
Bachelor of Technology
In
BIOINFORMATICS
By

ARTHZA SHUKLA (191906)
ARIN KAUSHAL (191905)

UNDER THE SUPERVISION OFDR.
TIRATHA RAJ SINGH

Department of Biotechnology and Bioinformatics

Jaypee University of Information Technology,
Wakhnaghat, 173234, Himachal Pradesh, IND

CANDIDATE'S DECLARATION

We hereby declare that the work presented in this report entitled Meta-Analysis of Gene Expression Data for Alzheimer's Disease in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in **Bioinformatics** submitted in the Department of Biotechnology/Bioinformatics**,** Jaypee University of Information Technology Wakhnaghat is an authentic record of my own work carried out over a period from January 2023 to May 2023 under the supervision of **Dr. Tiratha Raj Singh** Associate Professor Dept. of BT & BI.

The matter embodied in the report has not been submitted for the award of any other degree or diploma.

Arthza Shukla, 191906

Arin Kaushal, 191905

This is to certify that the above statement made by the candidate is true to the best of my knowledge.

Dr Tiratha Raj Singh

Professor

BT &BI Dated:

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

PLAGIARISM VERIFICATION REPORT

Date: .....

Type of Document (Tick): | Ph.D Thesis | M.Tech Dissertation/ Report | B.Tech Project Report | Paper |

Name:_____ Department:_____ Enrolment No _____

Contact No._____ E-mail._____

Name of the Supervisor: _____

Title of the Thesis/Dissertation/Project Report/Paper (In Capital letters): _____

_____

_____

UNDERTAKING

I undertake that I am aware of the plagiarism related norms/ regulations, if I found guilty of any plagiarism and copyright violations in the above thesis/report even after award of degree, the University reserves the rights to withdraw/revoke my degree/report. Kindly allow me to avail Plagiarism verification report for the document mentioned above.

Complete Thesis/Report Pages Detail:

- Total No. of Pages. =
- Total No. of Preliminary pages. =
- Total No. of pages accommodate bibliography/references. =

(Signature of Student)

FOR DEPARTMENT USE

We have checked the thesis/report as per norms and found **Similarity Index** at.....(%). Therefore, we are forwarding the complete thesis/report for final plagiarism check. The plagiarism verification report may be handed over to the candidate.

(Signature of Guide/Supervisor) Signature of HOD

FOR LRC USE

The above document was scanned for plagiarism check. The outcome of the same is reported below:

Copy Received on	Excluded	Similarity Index (%)	Generated Plagiarism Report Details (Title, Abstract & Chapters)	
	• All Preliminary Pages • Bibliography/Images/Quotes • 14 Words String		Word Counts	
Report Generated on			Character Counts	
		Submission ID	Total Pages Scanned	
			File Size	

Checked by
Name & Signature Librarian

Please send your complete thesis/report in (PDF) with Title Page, Abstract and Chapters in (Word File) through the supervisor at plagcheck.juit@gmail.com

ACKNOWLEDGEMENT

Firstly, we express my heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the project work successfully.

we are really grateful and wish our profound my indebtedness to Supervisor Dr. Tiratha Raj Singh, Associate Professor, Department of BT/BI, Jaypee University of Information Technology, Wakhnaghat. Deep Knowledge & keen interest of our supervisor in the field of "Bioinformatics" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

We would like to express my heartiest gratitude to Dr. Tiratha Raj Singh, Department of BT& BI, for his kind help to finish our project.

We would also generously welcome each one of those individuals who have helped us straightforwardly or in a roundabout way in making this project a win. In this unique situation, we might want to thank the various staff individuals, both educating and non-instructing, which have developed their convenient help and facilitated our undertaking.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

Arthza Shukla (191906)

Arin Kaushal (191905)

TABLE OF CONTENT

LIST OF ABBREVIATIONS

AD	ALZHEIMER'S DISEASE
NFT	NEUROFIBRILLARY TANGLE
MPSS	MASSIVELY PARALLEL SIGNATURE SEQUENCING
MRI	MAGNETIC RESONANCE IMAGING
CT	COMPUTERIZED TOMOGRAPHY
PET	POSITRON EMISSION TOMOGRAPHY
FDG	FLUORODEOXYGLUCOSE
RT-PCR	REAL TIME QUANTITATIVE
MMSE	MINI MENTAL STATUS EXAMINATION
SAM	SIGNIFICANCE ANALYSIS OF MICROARRAYS
MD	MEAN DIFFERENCE
Accord.	ACCORDINGLY,
QQ	QUANTILE QUANTILE

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Background: Around 47 million individuals worldwide suffer from dementia, which is a serious public health issue. A more severe kind of dementia that usually affects elderly people is Alzheimer's disease dementia (ADD). As the world population ages, dementia will become more common. The pathophysiology of Alzheimer's disease is still largely unclear.

Therefore, we sought to glean new information from existing data.

Methods: To discover potential biomarkers that may be beneficial for differentiating between upregulated and downregulated gene expression of dementia, gene expression data from diverse dataset studies were evaluated using a variety of software tools.

Results: We performed a meta-analysis using the GEO2R and WB-DEGS software, and we selected the top 250 genes based on p values. We found 16 up-regulated genes and 8 downregulated genes in ADD, and the most significant genes are those with the lowest P values. Additionally, we examined it using sam plots, umap, volcano plots, and boxplots; from there, we concluded that it might be beneficial as a potential biomarker.

Conclusion: These findings offer a better understanding of the variance in gene expressions seen in AD patients when comparing them across several tools, and they may prove helpful in the search for diagnostic and possible biomarkers.

keywords: Alzheimer’s disease; dementia, gene expression, meta-analysis, biomarkers.

CHAPTER 1 :
INTRODUCTION

1.1 INTRODUCTION

In 1906, AD was diagnosed by Alois Alzheimer, a neurologist from Germany, who conducted research on the case of a middle-aged individual named Auguste D. He performed an autopsy after she passed away and discovered senile plaques and neurofibrillary tangles. AD, often known as AD, is a type of dementia that impairs thinking and reasoning skills. As neurons are damaged and die across the brain, the connections between neural networks are broken, which obliterates memory. The brain's major regions start to shrink. When AD is at its most advanced stage, there is a significant loss of brain volume, known as brain atrophy.

Alzheimer's is also known as Senile Dementia.

Amyloid plaques and neurofibrillary tangles, which are the result of severe brain degeneration brought on by AD, are distinctive features of the human brain. The clustering of a beta-amyloid remnant fragment, which is toxic to neurons and interferes with cell-to-cell contact, is a major cause of neurodegenerative diseases. The aberrant build-up of the tau protein, which gathers inside neurons, is known as neurofibrillary tangles.

In addition to these, other key features of Alzheimer's include persistent inflammations, vascular issues, and cell death.

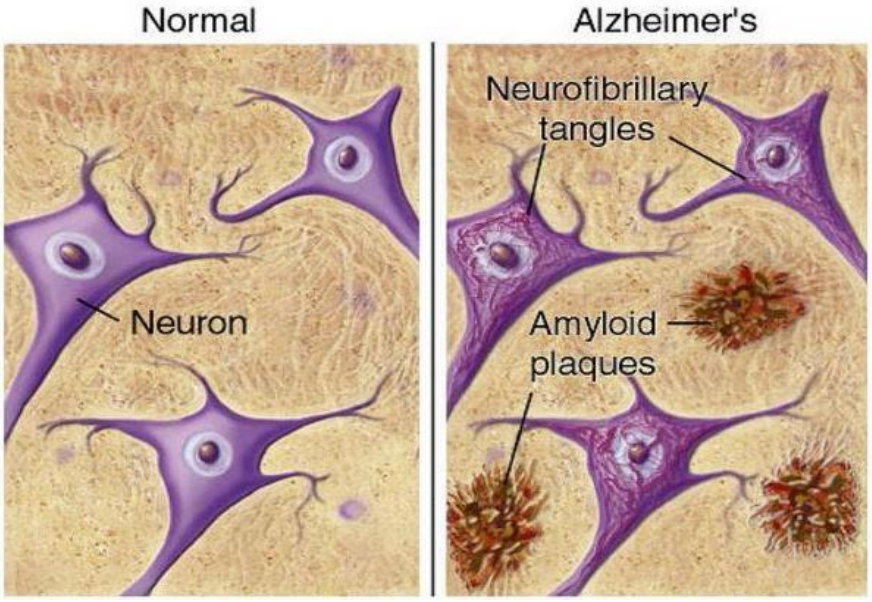

FIG1.1: Neurofibrillary Tangles and Amyloid Plaques [21]

1.1.1 SYMPTOMS

The primary sign of AD is memory loss. Early warning indications include having trouble recalling previous conversations or occurrences. Memory deficits worsen as the disease advances, and new symptoms appear.

A person who has AD may initially be aware of having trouble organizing their thoughts and remembering things. It's possible that a friend or family member would notice the symptoms getting worse first.

AD-related brain abnormalities cause increasing difficulties with:

1. Memory
2. Reasoning and Thinking
3. Making decisions and judgments
4. Preparing for and carrying out routine duties
5. Personality and behavioural changes: depression, mood swings etc.
6. Retained abilities
7. Other routine daily activities.

1.1.2 DIAGNOSIS

Previously, AD could only be definitively recognized when a patient died and the typical plaques and tangles were observed when the brain was inspected under a microscope. AD during a person's lifetime can now be identified with greater accuracy by doctors and researchers. Biomarkers can detect the presence of plaques and tangles by measuring the levels of amyloid beta and tau proteins in plasma and cerebral spinal fluid, as well as by employing specific types of PET scans.

A diagnostic work-up would most likely include the tests listed below:

A Neurological and Physical Examination: Reflexes, muscle strength and size, the ability to get out of a chair and walk around the room, hearing and vision, coordination, etc.

Lab tests :

a) Mental status and neuropsychological testing

b) Brain imaging

a) MRI

b) CT

c) PET: Fluorodeoxyglucose (FDG), Amyloid PET imaging, Tau PET imaging

1.1.3 TREATMENT:

Memory issues and other cognitive abnormalities may temporarily be helped by current Alzheimer's treatments. Presently, two types of medications are used to treat cognitive symptoms:

1. Cholinesterase inhibitors
2. Memantine (Namenda)
3. Antibody Adumanucab

To help manage the behavioural signs of AD, other drugs, such as antidepressants may occasionally be used.
Establishing a secure and encouraging environment.

1.1.4 PREVENTION:

Currently, there is no identified cure for AD, and there is also no definitive method to stop or slow down the progression of the condition. A healthy lifestyle, though, can reduce your risk. Lowering the chance of developing cardiovascular disease. Maintaining a healthy mind and social life.

1.1.5 CAUSES:

Current hypotheses suggest that the accumulation of abnormal proteins inside and around brain cells is what causes AD. One of the implicated proteins, amyloid protein, accumulates in plaques around brain cells. Another protein is known as tau, and accumulation of it cause tangles in brain cells.

Even while the exact cause is uncertain, scientists now understand that this process begins years before symptoms appear. As brain cells are destroyed, the chemical messengers (also referred to as neurotransmitters) that are needed to communicate or send signals between brain cells decrease. One neurotransmitter, acetylcholine, has particularly low levels in the brains of people with AD. Over time, the brain loses function in several areas. Frequently, memory-related areas are the first to suffer damage.

More uncommon forms of AD impact various areas of the brain. The initial symptoms might be problems with vision or language rather than memory problems. While the exact cause of AD remains unknown, there are several actions you can take to reduce your risk of developing the illness.

1. Ageing
2. Family background
3. Down's syndrome
4. Head injuries
5. Cardiovascular disease

1.2 Factors of AD Include:

1. Deposits of extracellular beta-amyloid (senile plaques)

2. Inside-the-cell neurofibrillary tangles (paired helical filaments)

The accumulation of beta amyloid and neurofibrillary tangles results in the loss of neurons and synapses, which causes a massive atrophy in the brain regions affected. The actuals mechanism which beta amyloid and neurofibrillary tangles causes deposition is still unknown, but it typically starts at the temporal lobe. AD patients' brains exhibit an immunological response; this inflammation can be considered the disease's primary pathologic trait. In AD the tau neurofibrillary tangles and the beta amyloid deposits have prion like properties, in prion diseases the normal surface of protein in brain also called prion protein becomes misfold into a pathogenic form, this leads to other prion proteins to misfold in a similar way which leads to increase in the abnormal proteins and increase in the damage to the brain.

Most cases of Alzheimer occur in late onset (65 years). The risk to develop this disease can only be predicted by age, however many of these cases occurs at an early age (65 years) and are mostly related to the gene mutations. Genetic mutations in affected patients lead to an alteration in the processing of the amyloid precursor protein, resulting in the accumulation of neurofibrillary tangles in the brain. Senile plaques primarily consist of beta amyloid, which includes dendritic processes. Additionally, beta amyloid can disrupt kinase activities, leading to the hyperphosphorylation of tau protein and subsequent formation of neurofibrillary tangles.

Risk components like smoking, diabetes & hypertension can raise the chance of developing AD, its suggested that treatment of these risk factors as soon as possible can reduce the risk of impairment at older age, The relationships with other factors like low levels of hormones and exposure to the metals with AD is not yet proven.

1.3 CURRENT THERAPEUTICS:

TRAMIPROSATE

Tramiprosate was the first drug to reach the clinical trial. It binds to the monomeric Beta. Amyloid and reduces the neurotoxicity while increasing clearance from the brain.

In Phase 2 trials, it was demonstrated that the treatment effectively reduces the levels of Beta Amyloid 42 in patients diagnosed with AD.

RAGE INHIBITOR

Amyloid is also seen to bind to the receptors on cell surfaces at Blood Brain Barrier. This banding can lead to the neuron's death. Recent laboratory studies showed that blocking of this binding can reduce the neurotoxicity and accumulation of the amyloid. It is investigated in a clinical study (Phase 2) for its potential AD therapy.

METHYLENE BLUE

It is a widely used dye from pre-historical times, Recent evidences show that it interferes with the tau aggregation. Ongoing research is currently investigating its potential as a therapeutic intervention for AD. The Phase 2 trial has successfully been completed and Phase 3 trials are now being planned.

NAP

A peptide derived from a neurotrophic protein called NAP. Animal studies showed that NAP reduces phosphorylation.

GAMMA-SECRETASE INHIBITORS

This enzyme complex plays a crucial role in cleaving the amyloid precursor protein at the end of the Beta Amyloid sequence. Its activity is necessary for the generation of Beta-Amyloid in the brain.

DIMEBON

This antihistamine, which was investigated in Russia, showed potential as a treatment for AD. In an in-vitro trial, the medication demonstrated improvement in behavioural and cognitive assessments for patients with AD. Also, the benefits of the drug treatment were increased greatly after extension of almost six months which lead to the possibility of the disease modifying effect. A Phase 3 trial is now being conducted.

In recent analysis some of them had a rather high incidences of side effects like vomiting, weight loss, abdominal pain, nausea and syncope.

Dizziness and headache were increased in cases with higher doses, although these side effects can be seen during the titration phase and are tried to be removed till the maintenance phase.

1.4 META ANALYSIS:

Meta analysis is a statistical technique that integrates the information from various studies to identify similar findings and detect overall trends. Sole purpose of such analyses is to look for individual as well as common factors involved in the regulation of disease or disease related conditions.

1.4.1 Why meta-analysis?

It offers more reliable outcomes that can aid in researchers' understanding of an effect's size.

1.4.2 Meta Analysis types:

Three fixed – effect methods
(Mantel-Haenszel , Peto and Inverse Variance)
One random effect procedure
(Der Simonion and laird inverse variance)

1.5 GENE EXPRESSION:

The method by which our DNA's instructions are translated into useful products, such as a protein.

1.5.1 GENE EXPRESSION ANALYSIS:

Figuring out the gene expression pattern at the level of genetic transcription, either in a specific cell or under specific circumstances.

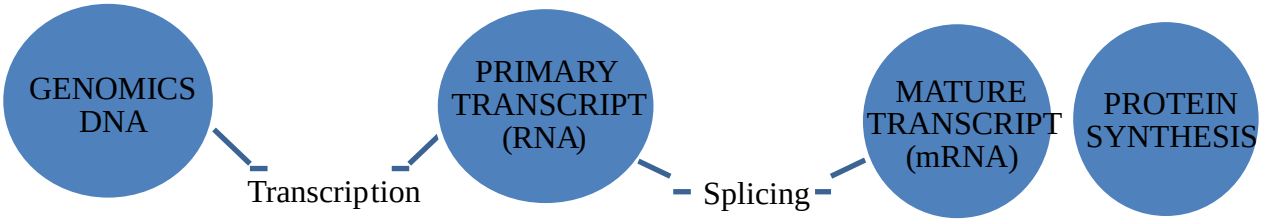

FIG 1.2: Gene Expression

1.6 GENE EXPRESSION OMNIBUS (GEO):

Biotechnology National Center for Information oversees the Gene Expression Omnibus repository, which is used for RNA methylation and gene expression profiling. These high throughput screening genomics data are derived from experimental microarray or RNA-Seq data.

It is a public functional genomics data repository that accepts data contributions that comply with Minimum Information About a Microarray Experiment standard.

Technologies:

1. RT-PCR
2. Microarray In Situ Hybridization
3. MPSS

1.7 TRANSCRIPTOMIC DATA:

It is an analysis of all transcriptional activity or a specific subset of RNA transcripts in a particular sample.

1.8 ANALYSIS OF MICROARRAY DATA:

There are many kinds of experiments that can use microarrays, including genotyping, epigenetics, translation profiling, and gene expression profiling (GEP).

GEP:

Extraction of features quality assurance

Normalization

Analysis of differential expression Results interpretation:

biological data entry into a public database

CHAPTER 2 :
LITERATURE SURVEY

Biological and biomedical literature was searched from PubMed about AD and other dementia. We have worked and explored several tools for analysing our data :

- WB-DEGS
- GEO2R
- MULTIPLE ARRAY VIEWER
- EXPANDER
- DAVID
- NASQAR
- WGCNA

2.1 WB-DEGS:

The free R / Bioconductor programme includes an application called WB-DEGS, which stands for Within and Between Group Comparisons for Differentially Expressed Gene Selection. It tries to reduce background noise and enhance the specification of the selection process. WB-DEGS can precisely pre-process, visualize, and choose genes so as to reduce the false positive rate. It also uses some traditional gene selection techniques. Complete flowchart is shown in figure 2.

Here are some of the packages required-

install.packages ("shiny")

install.packages ("VennDiagram")

biocLite ("affy")

biocLite ("affy PLM")

biocLite ("limma")

biocLite ("siggens")

biocLite ("twilight")

biocLite ("genefilter")

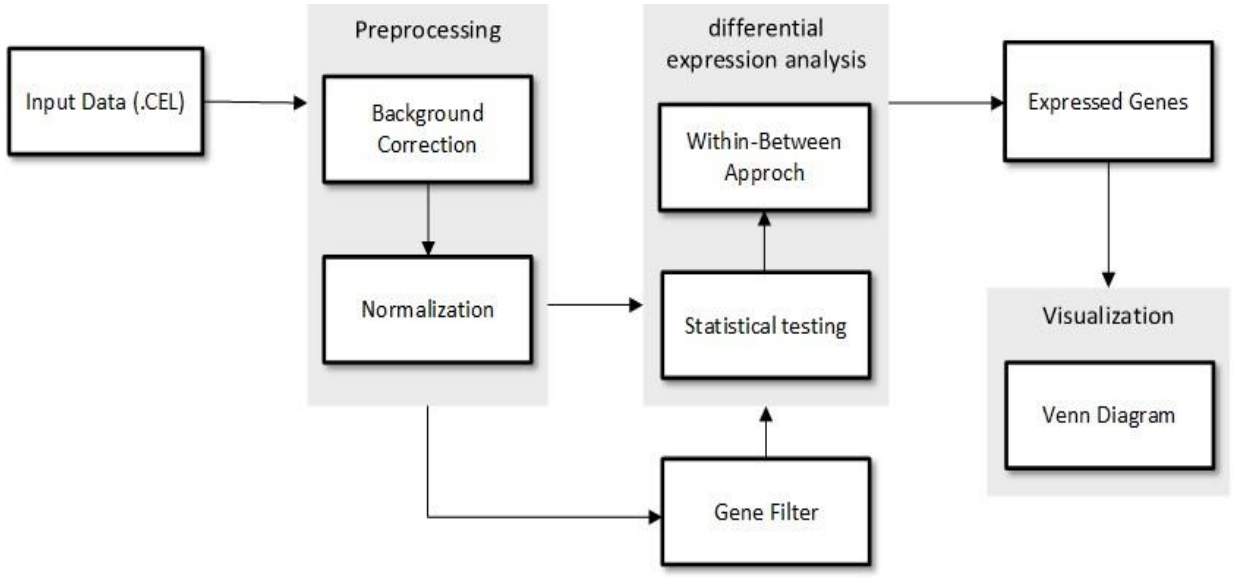

FIG 2.1: Working of WB_DEGS [22]

2.2 GEO2R:

The dynamic web application GEO2R is utilized to identify genes that exhibit differential expression under various experimental conditions. Users can compare two or more sets of samples from a GEO Series using this application. The results are presented in the form of a table that ranks genes based on their relevance and a series of visual plots that aid in the assessment of differentially expressed genes and the overall quality of the dataset. The R program GEO query and limma are employed in this process.

The GEO2R project within Bioconductor compares the data tables provided by the original submitters. Bioconductor, an open-source software project based on the R programming language, offers tools for analysing high-throughput genomic data. The GEO query R package parses GEO data and creates R data structures compatible with other R tools. The widely-used limma (Linear Models for Microarray Analysis) R package aids in identifying differentially expressed genes, supporting multiple-testing adjustments to mitigate false positives. It accommodates diverse experimental designs and data sources, making it a versatile choice for gene expression analysis. Together, these tools enable efficient analysis of high-throughput genomic data from GEO, assisting researchers in identifying significant gene expression patterns and gaining biological insights. The limma (Linear Models for Microarray Analysis) R package is widely recognized as a highly utilized statistical approach for detecting differentially expressed genes.

It offers support for diverse experimental designs and various types of data. The package incorporates multiple-testing corrections to P-values, addressing the potential issue of false positives. To facilitate user-friendly execution of R statistical analysis, GEO2R provides an interface that does not require prior command-line experience. Similar to other Dataset analysis tools within GEO, GEO2R directly analyses the Series Matrix data file instead of relying on curated Datasets.

As a result, more GEO data may be examined more quickly. GEO2R can analyse virtually any GEO Series, regardless of the data type or quality, therefore the user must be aware of its limitations and caveats. GEO2R analyses the Series Matrix data file directly as opposed to using curated Datasets like GEO's other Dataset studies, allowing for a quicker evaluation of a bigger volume of GEO data.

2.3 Multiple Array Viewer:

Software called Multiple Array Viewer (MAV) is used to analyse and display gene expression data from microarray research. Users can input and prepare unprocessed microarray data, run statistical analyses, and create visualisations to analyse the data.

The procedures for utilizing MAV to examine and display microarray data are as follows:

1) **Importing Raw Microarray Data and Normalising It:** MAV users can import microarray data in a variety of formats, such as CEL, TXT, and CSV. Once the data has been imported, it can be normalised using a variety of normalisation techniques, such as quantile normalisation or robust multiarray average (RMA), to account for systematic biases and variations in sample handling or preparation.

2) **Quality Control:** After normalisation, MAV offers tools for quality control analysis to spot outliers or underwhelming arrays that might influence subsequent analyses. Users can run a number of quality control checks, such as box plots, PCA, and visual evaluation of the raw data, to find technical variability and batch effects.

3) **Statistical Analysis:** In order to statistically analyse microarray data, MAV offers a number of methods, such as clustering, gene set enrichment analysis, and differential gene expression analysis. Users can compare the levels of gene expression between various ailments or therapies to find genes that are differentially expressed.

4) **Visualisation:** For examining and analysing microarray data, MAV offers a variety of visualisation tools. These consist of gene ontology (GO) enrichment analysis, PCA plots, volcano plots, and heatmap visualisation. Users can alter visualisations, and the outcomes can be exported for added analysis or dissemination.

In conclusion, MAV offers a user-friendly platform for the analysis and visualisation of gene expression data from microarray experiments, enabling researchers to better understand the underlying biology and find novel biomarkers or therapeutic targets.

2.4 Study on GSE185909

Researchers from Rush University Medical Centre in Chicago, USA, collected gene expression data from post-mortem brain tissues of people with AD and healthy controls for their study, " Examining the co-expression network in the frontal cortex during AD progression sheds light on molecular changes in the disease. It identifies key genes and pathways, aiding in therapeutic target discovery and biomarker identification". The dataset contained samples from people with dementia at various phases of the disease, from preclinical to severe dementia. Using microarray technology, the study examined the expression of more than 20,000 genes in each sample.

Weighted gene co-expression network analysis (WGCNA) was utilised by the researchers to find gene networks that were co-expressed throughout the various disease stages. To find gene modules that were significantly linked to the development of disease, they also ran statistical tests. The study also employed bioinformatics tools for hub gene discovery and pathway enrichment analysis for the detection of biological pathways that were overrepresented in the identified gene modules.

ANOVA, Student's t-test, and Pearson correlation coefficient were among the statistical methods employed in the study to analyse the data. In order to account for the false discovery rate, the statistical significance of the findings was assessed using multiple testing correction techniques such the Benjamini-Hochberg process.

The study's analysis of gene expression data and identification of co-expression networks linked to the development of AD involved a combination of statistical and bioinformatics methods. The discoveries shed light on the disease's molecular alterations and point to prospective areas of investigation and therapeutic focus for the future.

2.5 Study on GSE1297

Gene microarrays provide insights into the challenging pathophysiology of early-stage AD, distinguishing it from normal aging by examining concurrent biological pathway activity.

However, a lack of statistical power, a high rate of false positives or false negatives, and a lack of certainty on the relevance to functional endpoints frequently make it difficult to interpret microarray results. Here, using 31 different microarrays, we examined the hippocampus gene expression of nine healthy controls and 22 AD patients with differing degrees of severity. The expression of each gene was compared to MMSE and NFT scores in all 31 participants, irrespective of diagnosis. Comprehensive investigations showed a significant transcriptional response, with numerous genes associated with AD markers. However, a connection between several hundred of these genes and AD indicators was observed only in individuals with early-stage AD (MMSE > 20) and control subjects.

In the biological process categories linked to the genes with incipient AD-correlated expression, it was statistically found (using the EASE program) that many transcriptions factor/signalling genes controlling proliferation and differentiation, such as tumour suppressors, oligodendrocyte growth factors, and protein kinase A modulators, were upregulated. Additionally, adhesion, apoptosis, lipid metabolism, and the earliest stages of inflammation were regulated up, while protein folding, metabolism, and transport, as well as several energy metabolism and signalling pathways, were regulated down. The results propose a distinct AD etiology model in which genomic regulation leads to an up-regulation of tumor suppressor-mediated differentiation and involution processes. This mechanism is believed to drive the progression of the disease along myelinated axons.

CHAPTER 3 :
MATERIALS & METHODS

Research methodology helps to study and discuss the methods adopted in selecting the various samples. It also helps in selecting various statistical tool and techniques that help in analysing the data. Research methodology is the scientific and systematic method that helps in solving various research problems.

3.1 Data Collection

Data selection is defined as a process of collecting data for research purposes by using various resources. The data for the present study has been collected through databases provided online by Gene Expression Omnibus (GEO).

3.2 Data Analysis

The data for the present study has been analysed using a variety of software which uses mathematical tools like background correction methods, normalization, mean and standard deviation.

Background correction method

For microarray data, background correction is a crucial pre-processing step that makes an effort to correct the data for the ambient intensity around each feature. The "normexp" method simulates the observed pixel intensities as the sum of two random variables, one of which is exponentially distributed and represents background noise, and the other of which is normally distributed.

RMA

Robust Multichip Average is a method for obtaining the background corrected and normalized gene expression from the microarray data.

Normalization

A database's data is organized by normalization. Building tables and connecting them in line with rules designed to protect the data and boost the database's flexibility by removing duplication and inconsistent dependencies are necessary to do this. It is essential to get rid of biases in microarray data for accurate analysis. It is the process of adjusting microarray data for effects that result from technological variation as opposed to biological variations between the RNA samples or between the printed probes.

Quantiles

A nonlinear modification known as "quantile normalization" substitutes each feature value (row) with the average of the features across all samples of the same rank or quantile.

Lowess

Two-color data are combined using Lowess normalization, which also applies a smoothing correction to get rid of this fluctuation. The key advantage of using LOWESS for microarray normalisation is that it is resistant to extremely extreme outliers, and the cost function used in this way further limits the impact of such really extreme points in the regression.

Statistical Analysis

To identify patterns and trends, data is gathered and evaluated for statistical analysis. It is a component of data analytics. Statistical analysis can be used for tasks like compiling research interpretations, statistical modelling, or developing surveys and studies. It might be helpful for meta-analyses that deal with big amounts of data.

SAM

Also known as Significance Analysis of Microarray, it's used to obtain a non-parametric score for each gene using SAM, which is basically based on repeated measurements. However, it might lose some of its ability to accurately identify differentially expressed genes (DEGs) that defy homogeneity in general statistical tests.

P-Value

The p value is a number, obtained from a statistical test, that describes how likely you are to have found a certain collection of observations.

Benjamíni- Hochberg Procedure

A strong tool for lowering the false discovery rate is the Benjamini-Hochberg Procedure.

Limma precision weights (vooma)

The vooma technique, which also determines accuracy weights for each observation and approximates the mean-variance relationship of the log-counts, provides input to the limma empirical Bayes analysis pipeline. As a result, RNA-seq analysts now have access to a wide range of techniques developed for microarrays.

CHAPTER 4:
RESULTS

DATASET GSE1297

The top 250 genes according to P-value are listed in a table as the results are displayed in the browser. The most significant genes are those whose P-values are the lowest. To view the gene expression profile graph for a particular gene, click on a row.

Visualization:

To assist users in further exploring differentially expressed genes and evaluating dataset quality, several graphical plots produced.

4.1 Volcano Plot

To visualize differentially expressed genes, a volcano graphic compares stats significances (-log10 P val.) to the size of the change (log2 fold changes).

By default, when there are more than two groups of samples, there are shown an equal number of contrasts for each group, and they are compared to one another in the order in which they were formed. It is shown in fig 4.3.

4.2 Mean-Difference Plot

To visualize differentially expressed genes, a MD plot compares log 2-fold changes to average log 2 expression levels. Mean difference plots are employed to present the test results for a single contrast in a study. In cases where the study involves more than two sample groups, separate plots are created for each contrast to visually represent the results. Accord., when there are more than two groups of samples, there are shown an equal number of contrasts for each group, and they are compared to one another in the order in which they were formed. It is shown in fig 4.5.

4.3 UMAP

A dimensionality reduction method called uniform manifold approximation and projection (UMAP) can be used to visualize the relationships between Samples. The graphic shows how many nearest neighbours were considered in the calculation. It is possible to create this figure without selecting a sample group. Shown in fig 4.6.

4.4 Venn Diagram

The browsing and downloading functionality allows users to explore and obtain the substantial gene overlap between multiple contrasts. The tool can accommodate up to 5 contrasts for data plotting purposes. In cases where there are more than five defined groups, the default behaviour is to present contrasts between the most and least expressed genes. This feature provides a comprehensive view of the gene expression patterns and facilitates the identification of significant differences among the defined groups.

4.5 Boxplot

The tool determines value distribution in selected samples and color-codes them based on categories. If you want to know whether the Samples you have chosen are appropriate for a differential expression study, viewing the distribution can be helpful. In general, median-centre values indicate that data are cross-comparable and standardized. The graphic showcases the data post log transformation and normalization, if any. It is shown in fig 4.7.

4.6 Expression Density

The tool helps determine the distribution of values in selected samples, with samples being coloured based on groupings. It serves as a complement to the boxplot for verifying data normalization before conducting differential expression analysis. The plot visualizes the data after applying log transformation and normalization, if applicable. Shown in fig 4.8.

4.7 Adjusted P-Value Histogram

Use to see how the P-values in the analysis's findings were distributed. The P-value was calculated using all of the chosen contrasts and is identical to that in the table of the top differentially expressed genes. Plot shows P-value distribution for all genes, overcoming table size limitation (250 entries).

4.8 Moderated T-Statistic (Q-Q Plot)

The plot compares quantiles of a data sample to the quantiles of a fictitious Student's t distribution generated by the limma test. This assessment helps evaluate the accuracy of the limma test results. A perfectly straight line in the plot indicates that the computed values for the moderated t-statistic align with the expected distribution. Complete graph shown in fig 4.9.

4.9 Mean-Variance Trend

The image is employed to evaluate the relationship between the mean and variance of expression data after fitting a linear model. It helps determine the level of variability in the data. By examining this graph, one can assess whether it is advisable to utilize precision weights, which account for the mean-variance trend. When a pronounced mean-variance trend is observed, incorporating precision weights enhances the accuracy of test results.

Group selection is not necessary for the plot. A gene is represented by each point. Fig is on 4.10.

IN GEO2R

Sam Plot Results for Between Group –

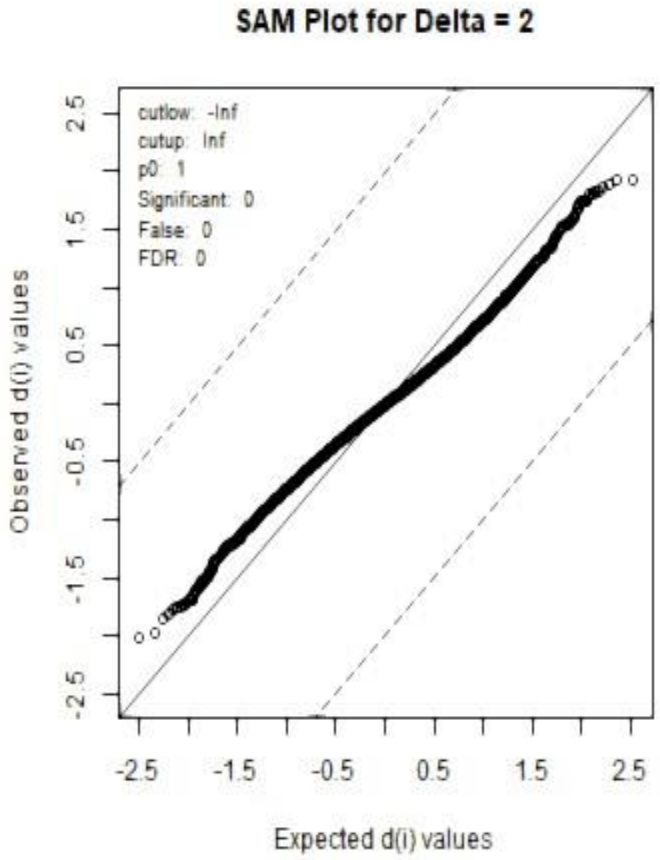

FIG 4.1: Control vs ALL

Figure 4.1 displays important genes determined using observed and expected di values between the Control group and all other groups.

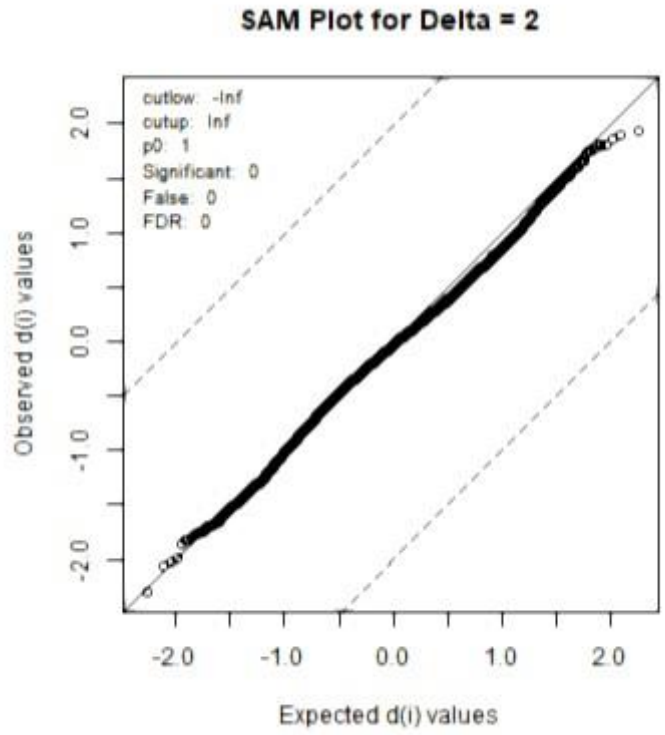

FIG 4.2: Control VS Moderate

Figure 4.2 displays important genes determined using observed and expected di values between the Control group and Moderate group.

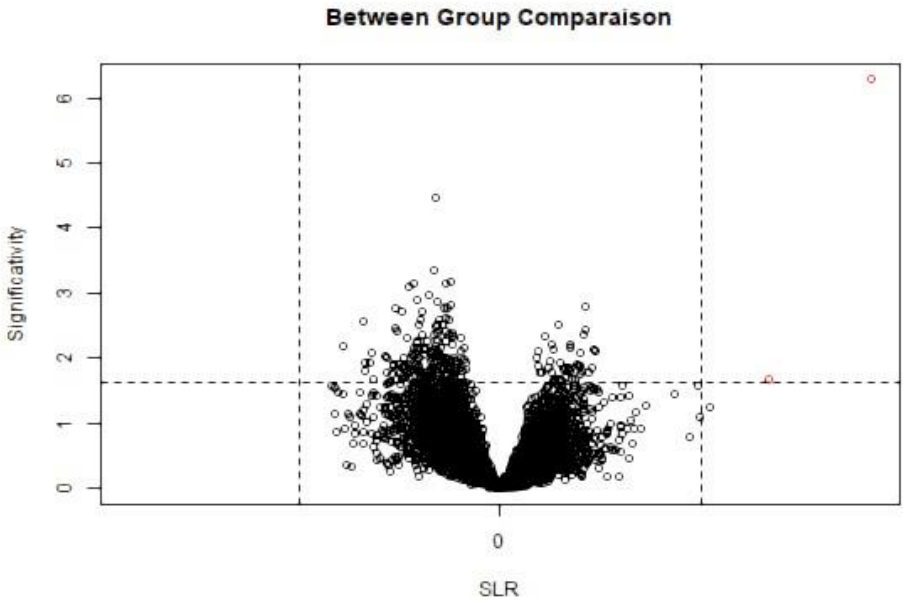

FIG 4.3: Volcano plot for between groups

Control vs All

The volcano plot in figure 4.3 compares the control group to all other groups. Important genes are displayed above the threshold line.

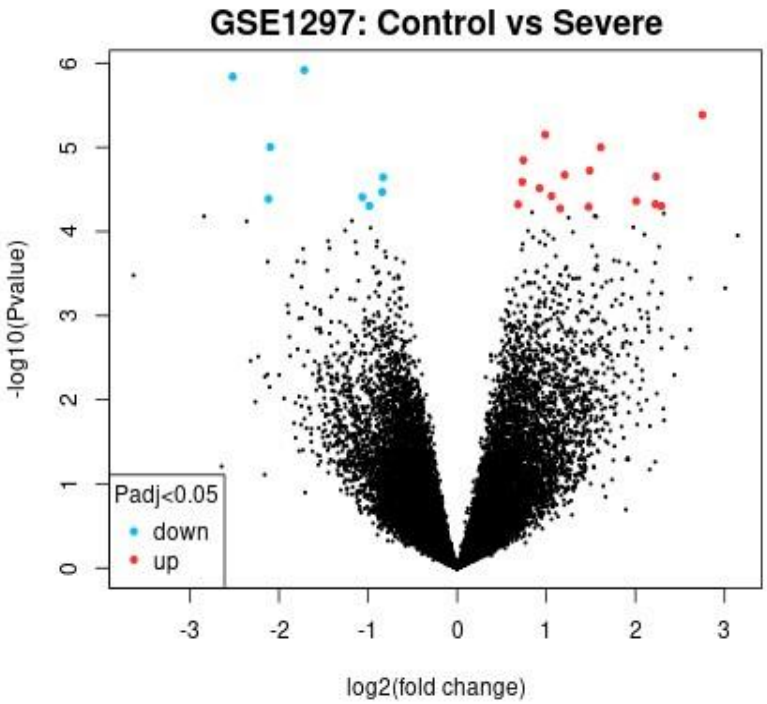

FIG 4.4: Volcano Plot for Control vs Severe

The important genes are plotted between p value and fold change in this, with up-regulated genes represented in red and down-regulated genes shown in blue.

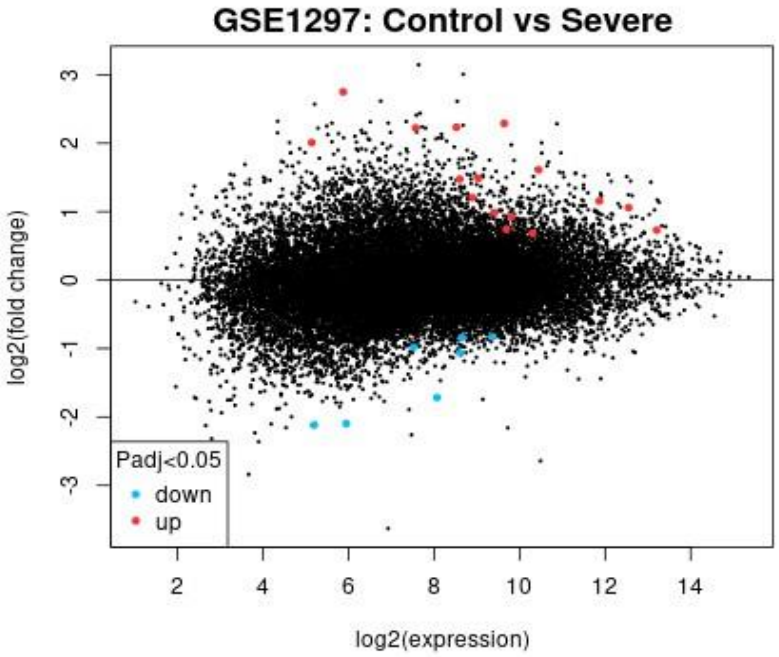

FIG 4.5: Mean Difference Plot for Control vs Severe

demonstrates a mean difference plot against a log 2-fold change and a log 2 expression. over expression showed by red color and under by blue.

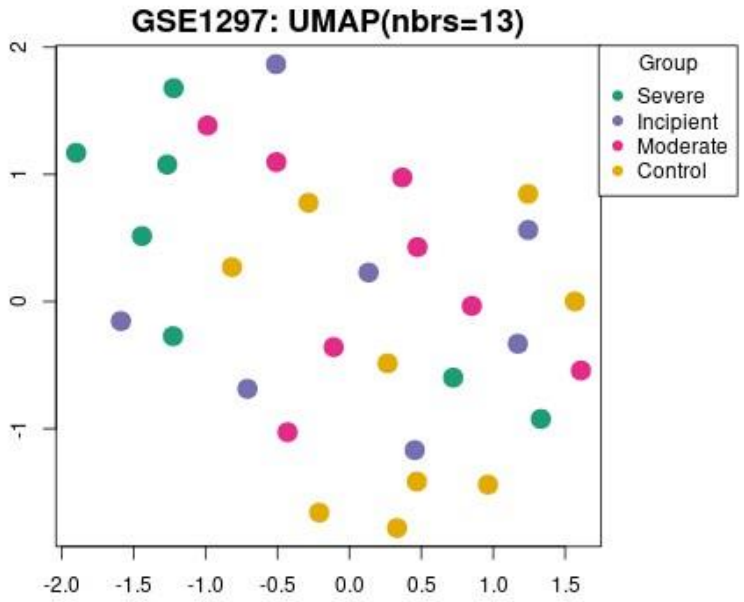

FIG 4.6: UMAP Plot for Between Groups

This figure groups clusters.

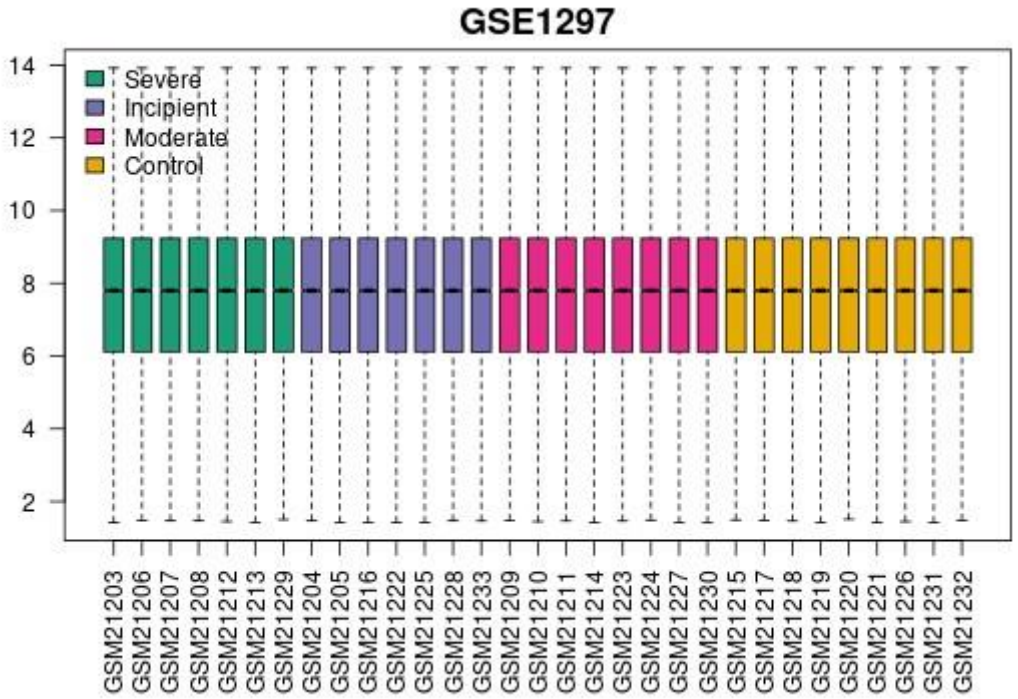

FIG 4.7: Boxplot for between groups

Through a Boxplot, figure 4.7 illustrates similarities between several groupings.

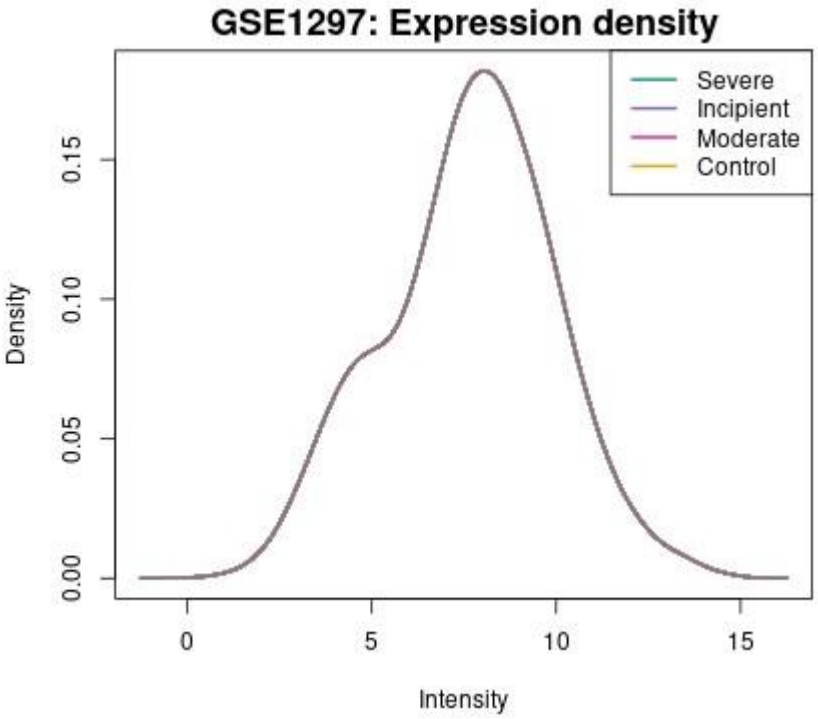

FIG 4.8: Expression Density Plot for Between Groups

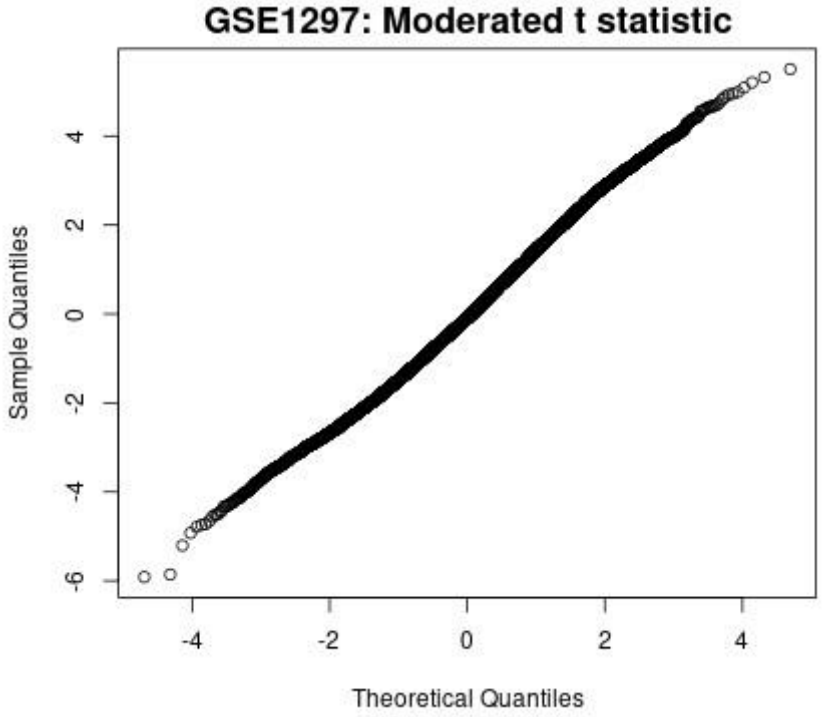

FIG 4.9: T-Statistic Quantile-Quantile Plot

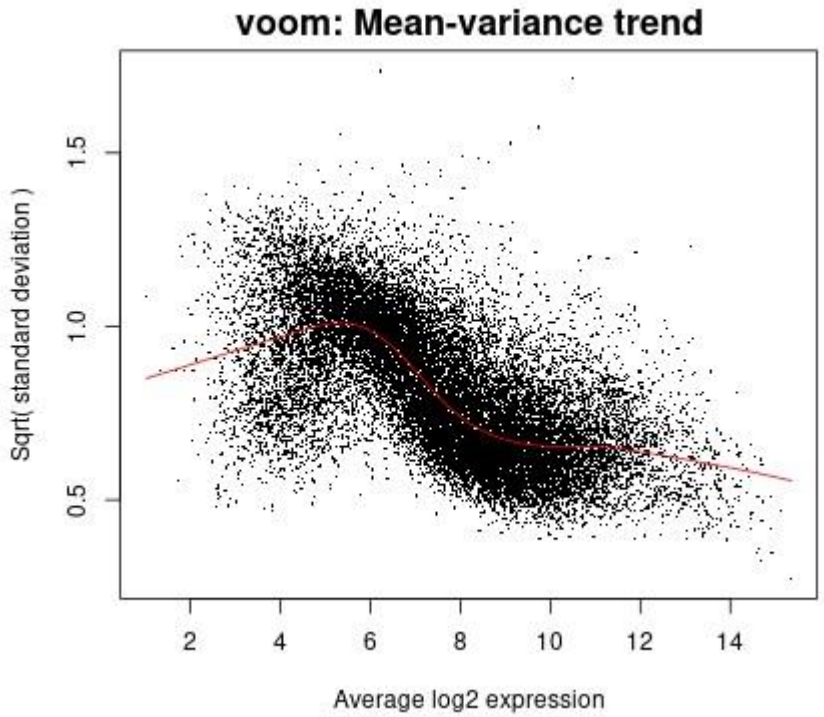

FIG 4.10: Mean Variance Trend Plot

Gene expression is displayed between standard deviation and log 2 expression in this figure.

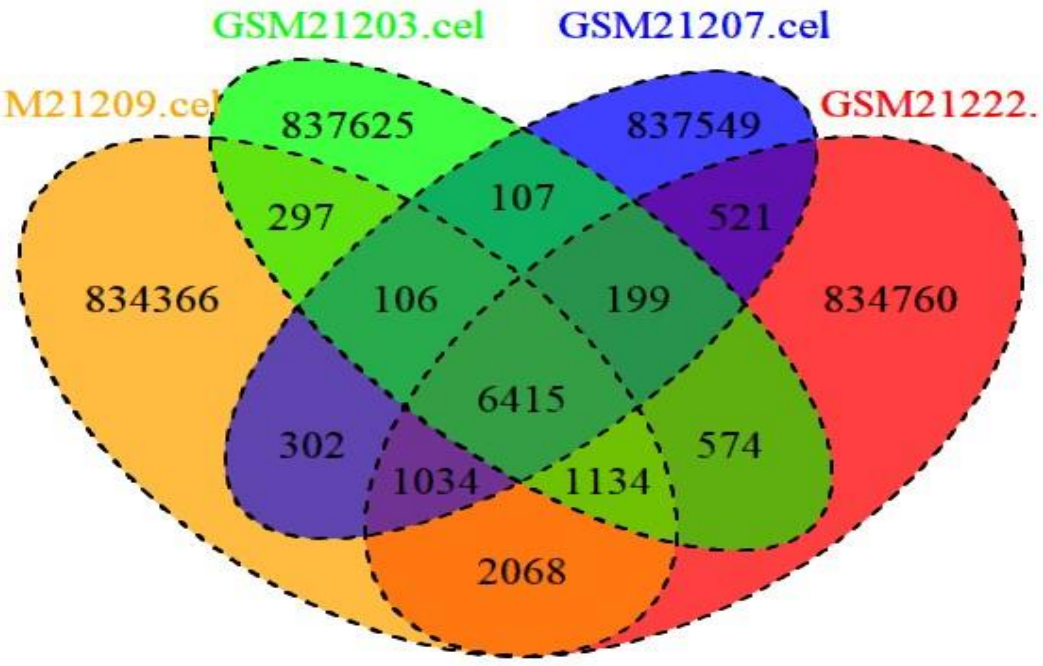

FIG 4.11: Venn Diagram Showing Gene Similarities Between Various Cases

The research revealed the following genes that have varied levels of expression. From most significant to least significant.

TABLE 4.1 (Under Expressed Genes)

Genes Found Under-Expressed

DOWN ID	GENE TITLE	GENE SYMBOL
215320_At	SPATA31 Subfamily C Member 2	SPATA31C2
206278_At	Platelet Activating Factor Receptor	PTAFR
217367_S_At	Zinc Fingers and Homeoboxes 3	ZHX3
219746_At	Double PHD Fingers 3	DPF3
208112_X_At	EH Domain Containing 1	EHD1
214927_At	Integrin Subunit Beta Like 1	ITGBL1

TABLE 4.2 (Highly Expressed Genes)

Genes Found Highly Expressed

GENE ID	GENE TITLE	GENE SYMBOL
203224_At	Riboflavin Kinase	Rfk
203340_S_At	Solute Carrier Family 25 Member 12	Slc25a12
203527_S_At	Apc, Wnt Signaling Pathway Regulator	Apc
206384_At	Calcium Voltage-Gated Channel Auxiliary Subunit Gamma 3	Cacng3
209853_S_At	Proteasome Activator Subunit 3	Psme3
211496_S_At	Phosducin	Pdc
218888_S_At	Neuropilin And Tolloid Like 2	Neto2
203894_At	Tubulin Gamma 2	Tubg 2
217733_S_At	Thymosin Beta 10	Tmsb10
218050_At	Ubiquitin Fold Modifier 1	Ufm1
210517_S_At	A-Kinase Anchoring Protein 12	Akap12
202457_S_At	Protein Phosphatase 3 Catalytic Subunit Alpha	Ppp3ca
213222_At	Phospholipase C Beta 1	Plcb1
204055_S_At	Ctage Family Member 5, Er Export Factor	Ctage5

GENE EXPRESSION

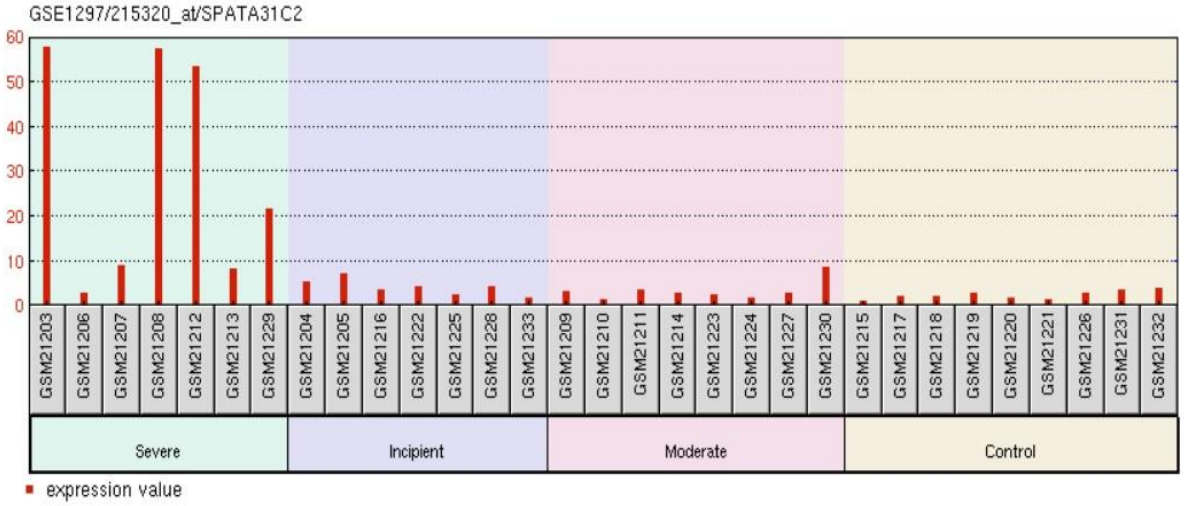

FIG 4.12: 215320_AT (GEO2R)

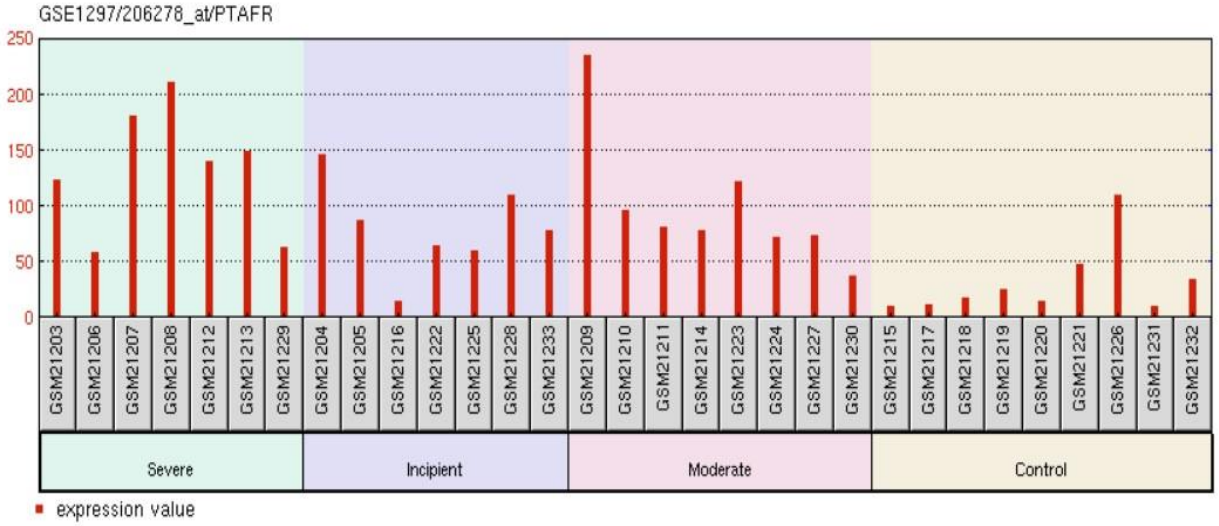

FIG 4.13: 206278_AT (GEO2R)

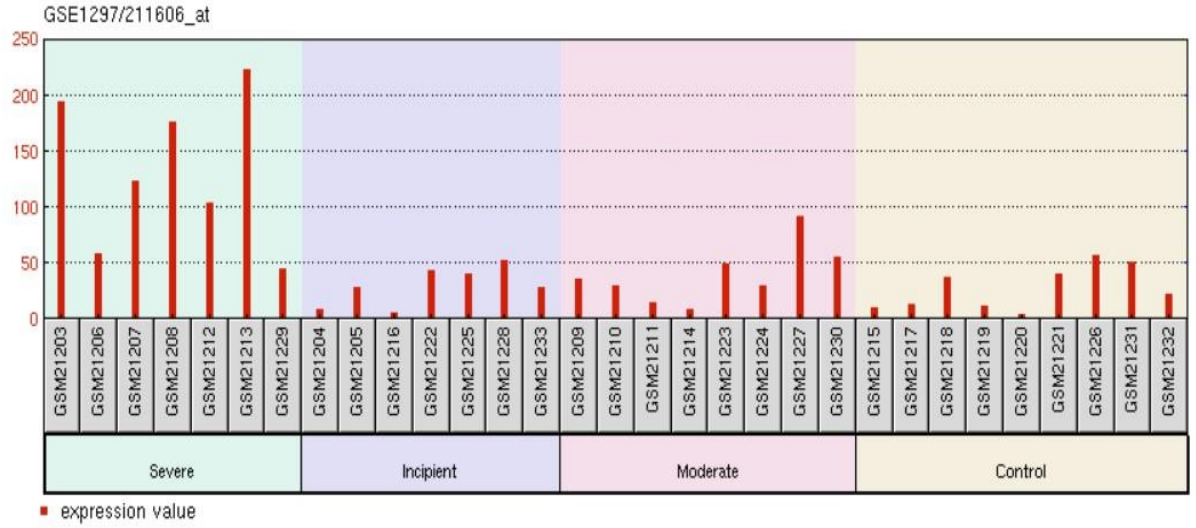

FIG 4.14: 211606_AT (GEO2R)

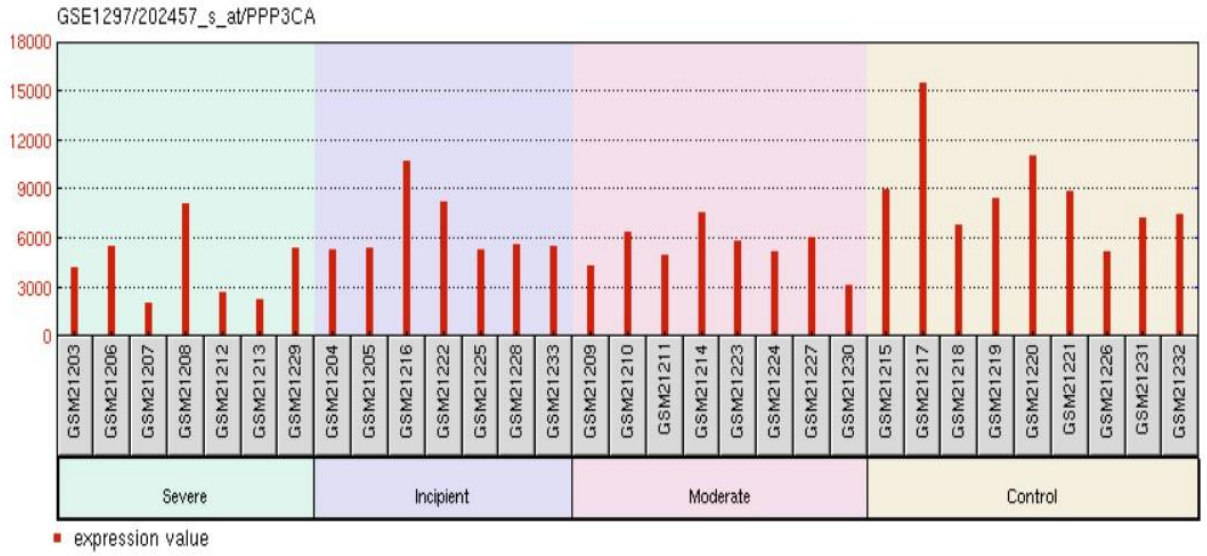

FIG 4.15: 202457_S_AT (GEO2R)

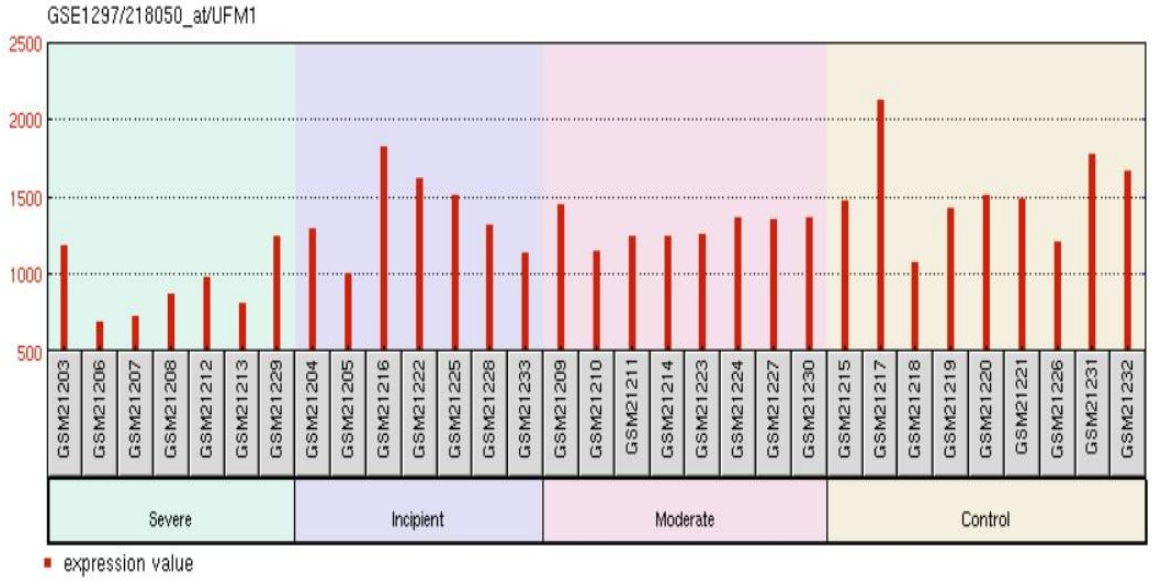

FIG 4.16: 218050_AT (GEO2R)

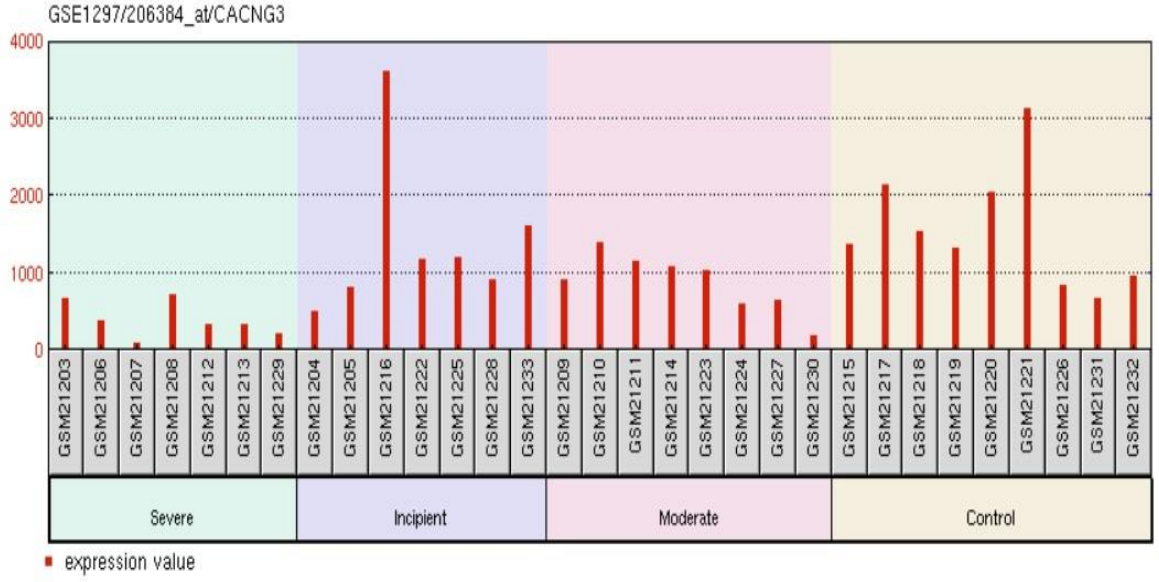

FIG 4.17: 206384_AT (GEO2R)

DATASET GSE185909

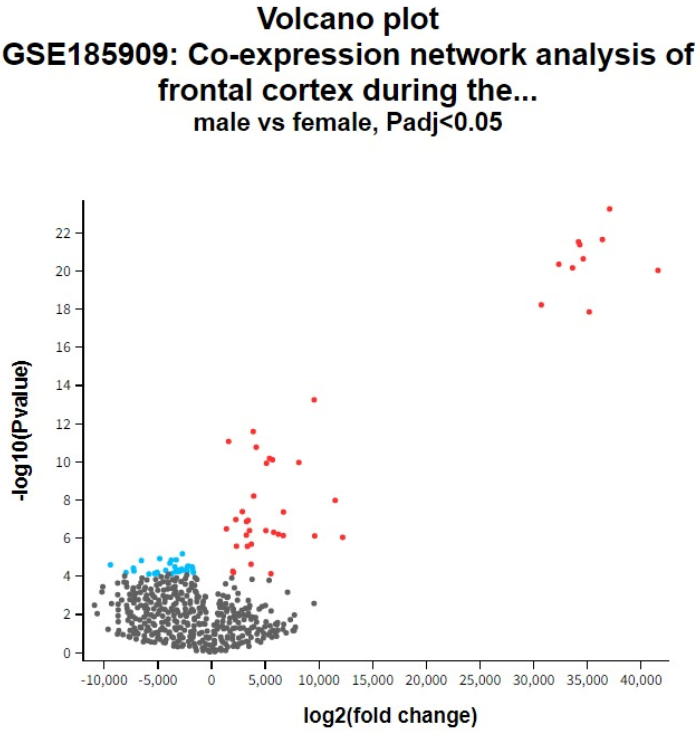

FIG 4.18: Volcano Plot for Male Vs Female

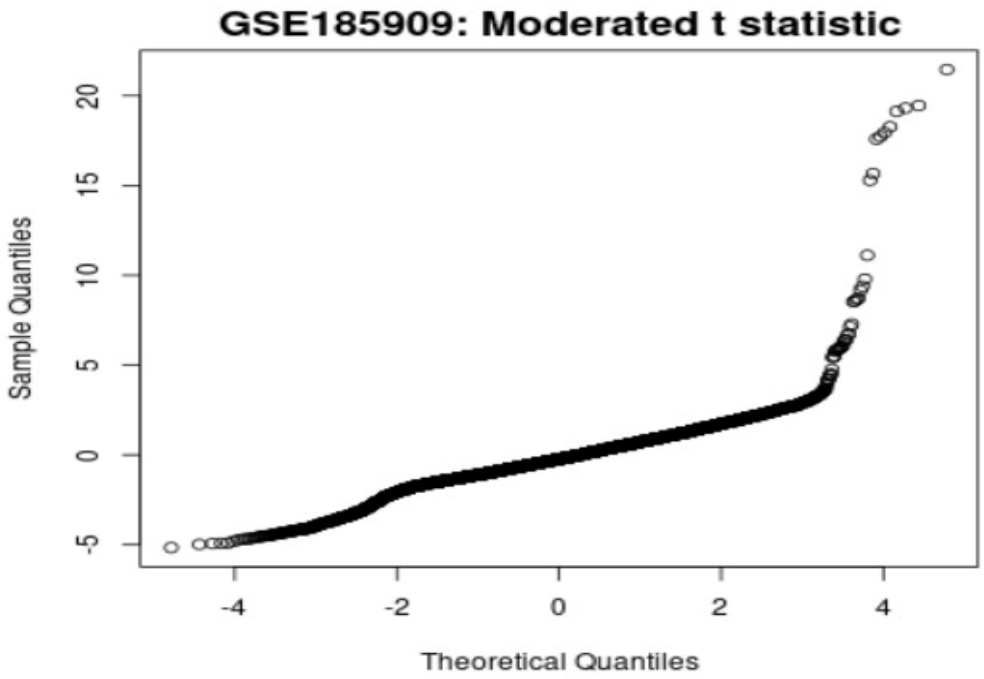

FIG 4.19: T- Statistics Quantile Quantile Plot

Plot between sample quantiles and theoretical quantiles.

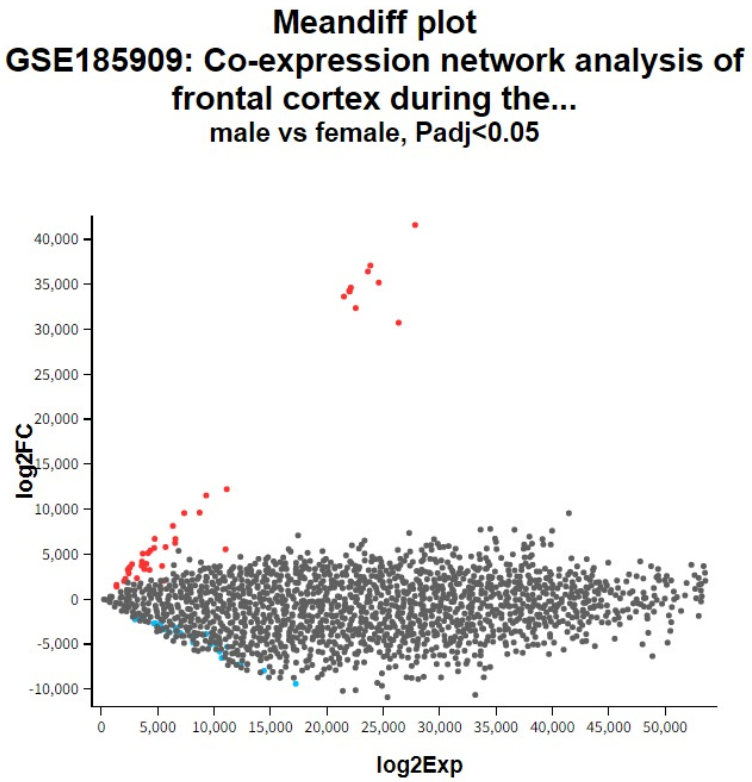

FIG 4.20: Meandiff Plot for Male Vs Female

This Figure shows Meandiff Plot between log 2 exp and log 2FC

TABLE 4.3 (UNDER EXPRESSED GENES)

GENES FOUND UNDER EXPRESSED

GENEID	GENE TITLE	GENE SYMBOL
AK130194	Riken cDNA 1810025E23	1810025E23Rik
BC012573	ATP synthase, H+ transporting, mitochondrial Fo complex subunit G2	ATP5L2
BC016770	Maestro heat-like repeat family member 6	MROH6
BC017166	Solute carrier family 25-member 43	SLC25A43
BC021719	Rho guanine nucleotide exchange factor 37	ARHGEF37
BC069784	Chromosome 12 open reading frame 75	C12orf75
BC101572	Leucine-rich repeat-containing protein 41	LRRC41
BC104930	Chromosome 9 open reading frame 16	C9orf16

TABLE 4.4 (HIGHLY EXPRESSED GENES)

GENES FOUND HIGHLY EXPRESSED

GENE ID	GENE TITLE	GENE SYMBOL
AF332237	Centrosomal protein 55	CEP55
AF332236	Sperm-associated antigen 9	SPAG9
AF332240	Centromere protein F	CENPF
AF332242	Hyaluronan mediated motility receptor	HMMR
AF517635	Solute carrier family 25-member 36	SLC25A36
AK026367	Phospholipid phosphatase 6	Plpp6
AK056903	Coiled-coil domain containing 88B	Ccdc88b
AK093413	Riken cDNA 1300002E11 gene	1300002E11Rik
AY130858	KIAA1671 gene	KIAA1671
BC026077	Nucleolar protein 9	NOL9
BC026100	Ubiquitin protein ligase E3 component n-recognin 5	UBR5

IN MULTIPLE ARRAY VIEWER

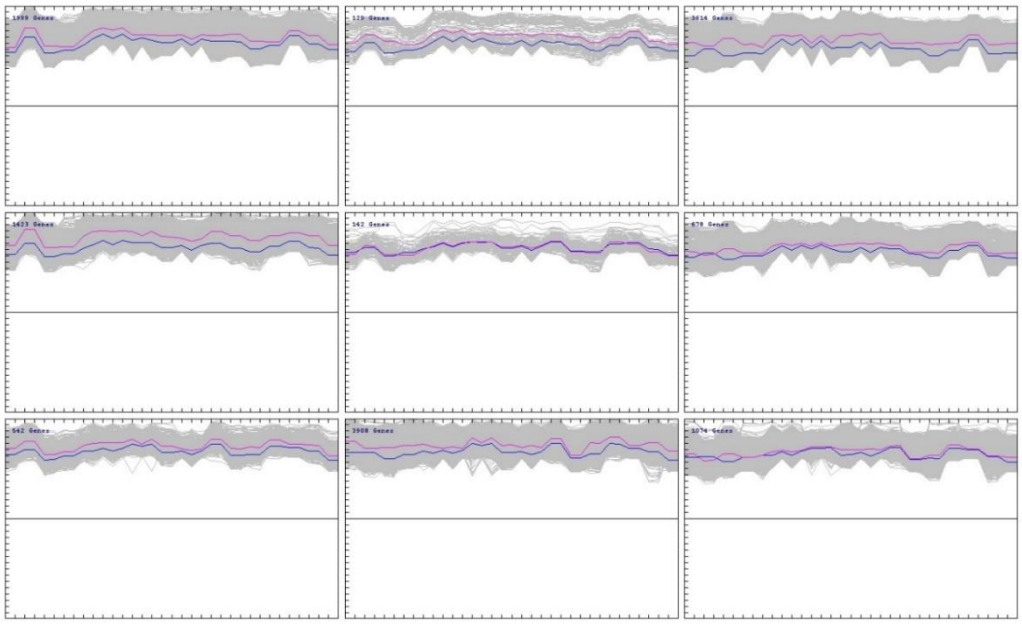

FIG 4.21: SOM Plot All Clusters

Figure 4.21 displays the expression density for various genes.

FIG 4.22: U Matrix Distance

In this the genes were clustered using u matrix distance.

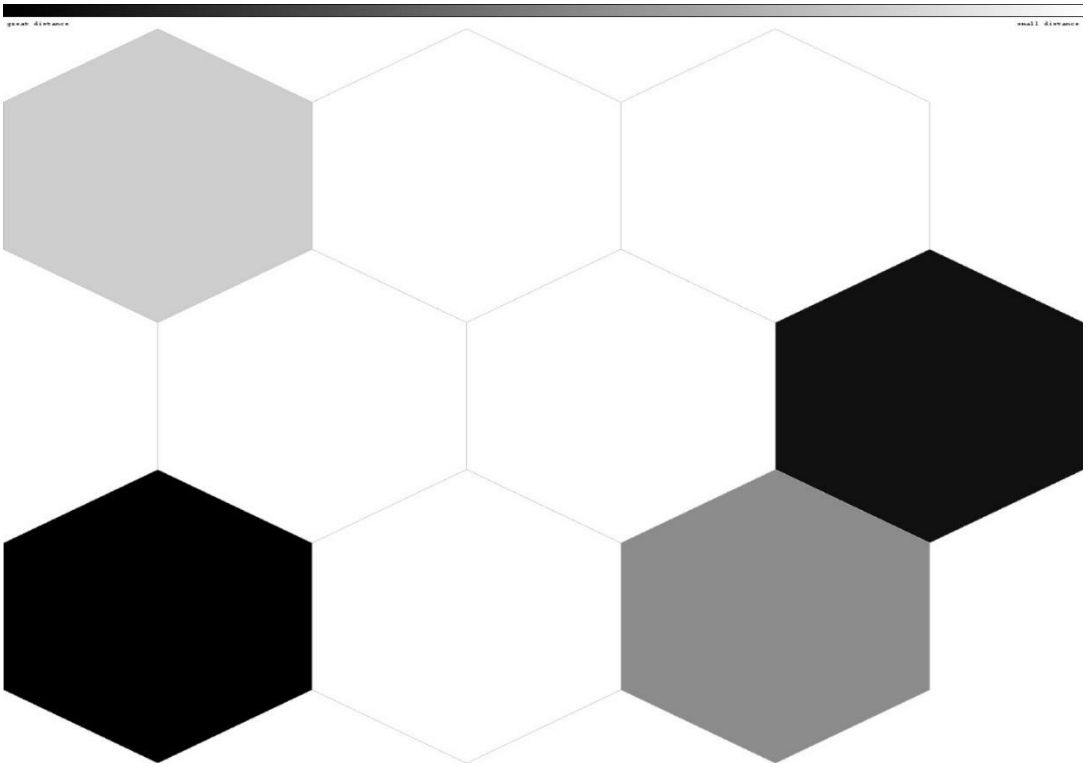

FIG 4.23: U Matrix Color

The genes were clustered using u matrix distance in coloured form.

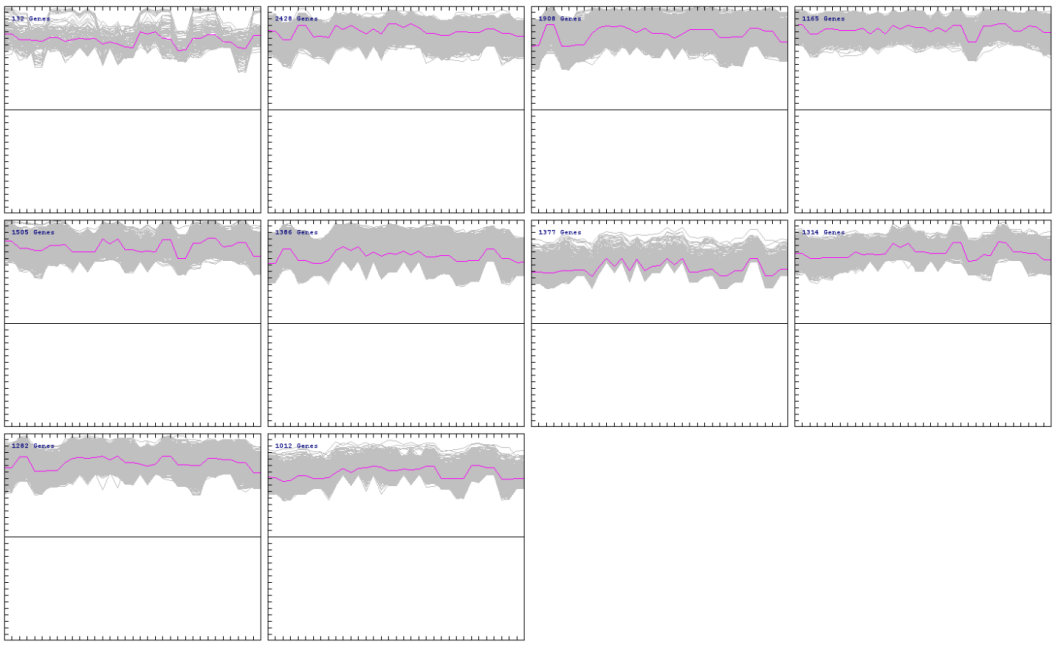

Fig 4.24: KMC Support All Clusters

This figure shows expression density of Different genes.

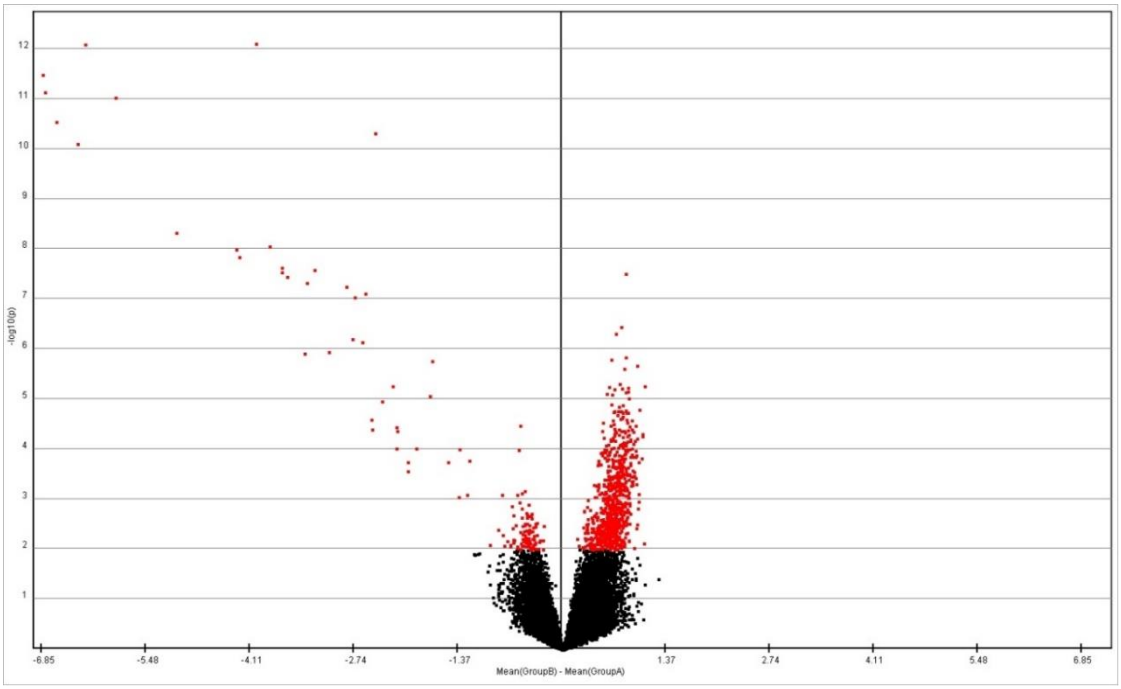

FIG 4.25: Volcano Plot

Is a plot between p value and difference of mean of both groups significant genes are shown in red colour.

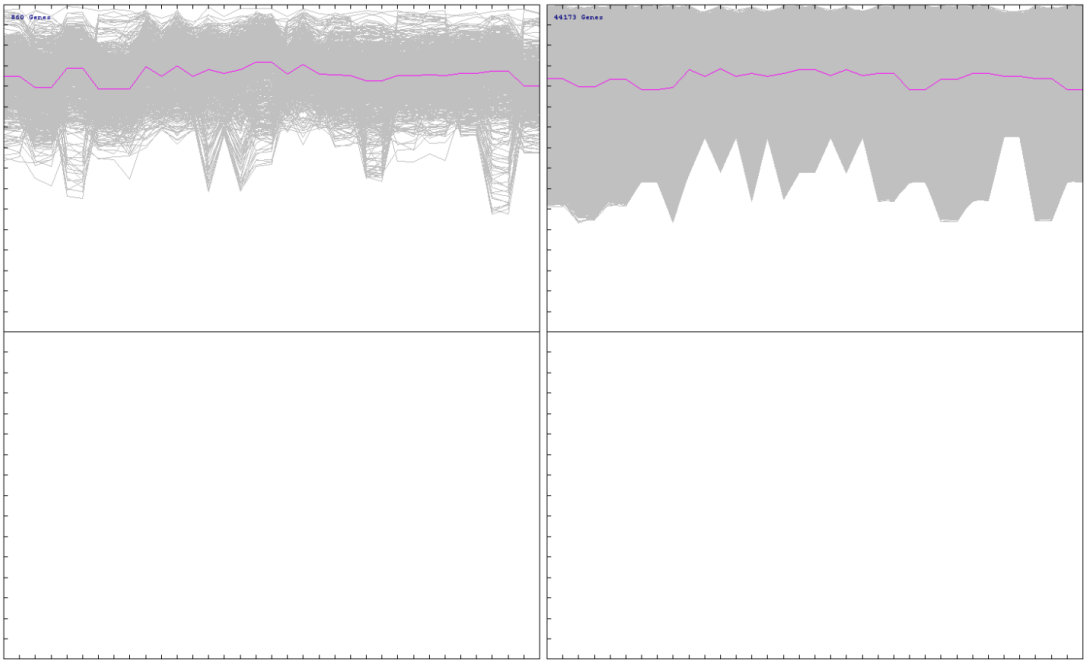

FIG 4.26: Significant and Non-Significant Genes

demonstrates how the expression density of genes with and without significance.

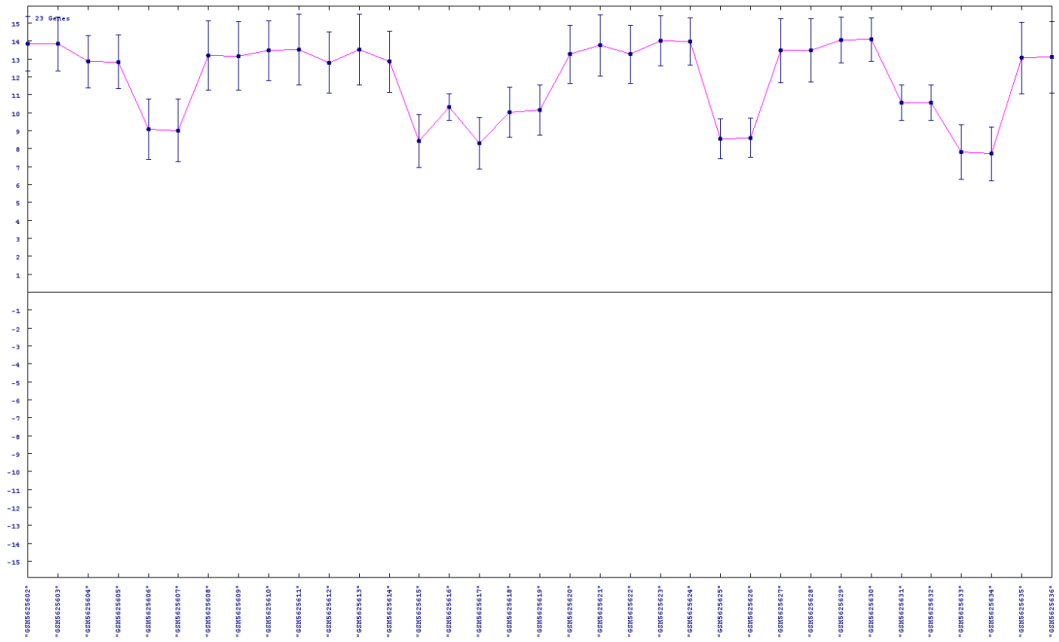

FIG 4.27: Significant Genes

This shows expression density of genes of different groups.

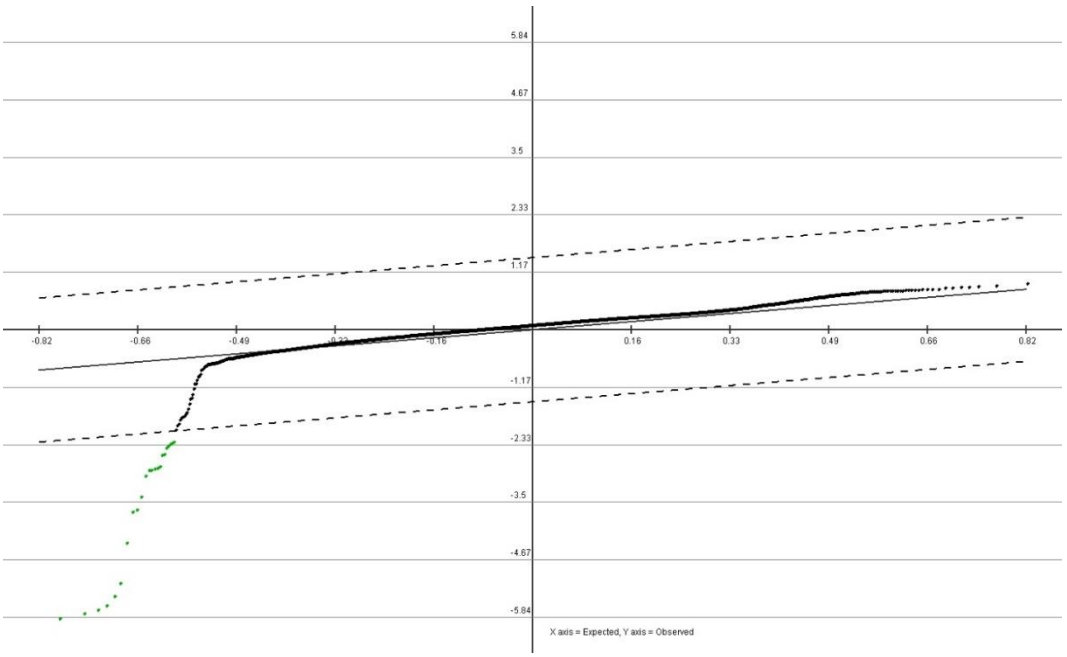

FIG 4.28: SAM Plot

This shows significant genes the significant genes are marked by green colour

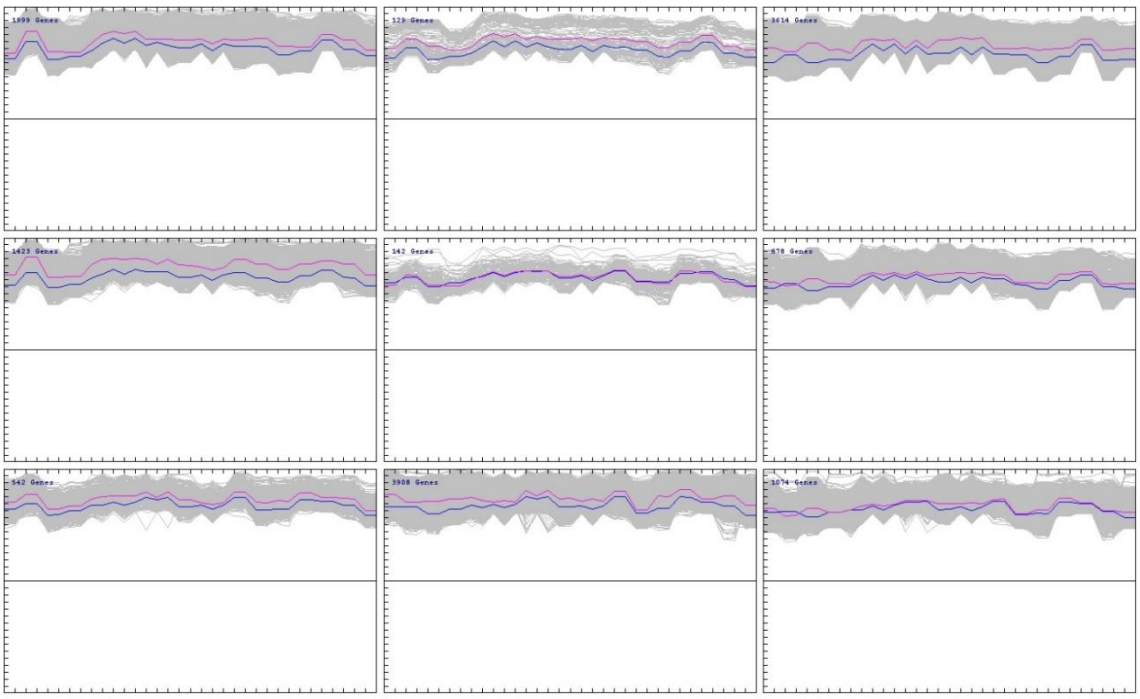

FIG 4.29: Clusters

This shows expression of genes arranged in clusters.

CHAPTER 5 :
CONCLUSION

5.1 CONCLUSION

Understanding the pathophysiology of early-stage AD (AD) has been challenging due to the complexity of the disease and the overlap of its early-stage markers with those of normal aging. However, gene microarrays offer valuable tools to tackle this complexity by providing comprehensive insights into the simultaneous activity of multiple biological pathways. By utilizing gene microarrays, researchers can gain a better understanding of the complex molecular processes associated with early-stage AD, enabling a deeper exploration of the disease's mechanisms and its differentiation from normal aging. However, a lack of statistical power, a high rate of false positives or false negatives, and a lack of certainty on the relevance to functional endpoints frequently make it difficult to interpret microarray results. Here, we studied the hippocampal gene expression of 22 AD patients with varying degrees of severity and nine healthy controls using 31 different microarrays. After that, we evaluated the correlation between each gene's expression using a variety of techniques, selecting the top 250 genes based on their p values and analysing 16 upregulated and 8 downregulated genes using a variety of parameters that may help us find potential biomarkers also to find any inaccuracies or additional data to compare, we would like to collect more data, compare it using other software programmes, and then compare all the other results for future prospects.

5.2 LIMITATIONS

Constraints in meta-analysis of Alzheimer's gene expression data: dataset heterogeneity, sample size, data pre-processing variation, publication bias, data quality assessment, and limited raw data availability:

1. **Data heterogeneity:** The experimental settings, sample sizes, and platforms used to obtain the gene expression data from various studies may differ. Because of the potential for data heterogeneity, it can be challenging to compare the findings of different studies.

2. **limited sample sizes**: The power of meta-analysis can be affected by the large number of research on AD that have tiny sample sizes. The data may become more variable as a result, making it more challenging to identify actual group differences.

3. **Variability in disease stage and severity**: AD is a complicated and heterogeneous condition that progresses through many stages and exhibits varying degrees of disease severity. It may be difficult to compare gene expression data between research due to this diversity.

4. **Lack of standardised analysis techniques**: There is disagreement over the most effective strategies for examining gene expression data in AD. Different research may employ various statistical, bioinformatic, and normalisation techniques, which may cause results to vary.

5. **Potential confounding factors**: Age, sex, medication use, and comorbidities are only a few of the many confounding variables that may affect gene expression in AD. These variables might not always be taken into consideration in individual research, which can result in biases in meta-analysis.

6. **Publication bias**: those with noteworthy findings are more likely to be published, whereas those with unfavourable findings may be less likely to do so. This might cause the genuine effect size to be overestimated in meta-analyses, which would be biased.

7. **Issues with reproducibility**: When utilising several platforms or normalisation techniques, gene expression data can be complex and challenging to repeat. This can make it difficult to replicate findings across research, which could result in inconsistencies in meta-analysis.

Gene expression meta-analysis in AD is a valuable method to discover promising biomarkers and pathways associated with the disease. When assessing the outcomes of these analyses, it is crucial to keep in mind their constraints and potential biases.

5.3 FUTURE PROSPECTS

Meta-analysis of gene expression data for AD has the potential to significantly advance our understanding of the condition and the underlying molecular pathways in the future. Some prospective future developments in this area are listed below:

1. **Integration with other omics data**: Gene expression data can be combined with other omics information, such as proteomics and metabolomics, to provide a more thorough knowledge of the illness. Important biomarkers and pathways connected to AD can be found using this information.

2. **Application of machine learning strategies**: Gene expression data can be used to identify predictive models for AD by applying machine learning strategies. By enabling early diagnosis and therapy, this can assist in identifying early biomarkers of the disease.

3. **Utilisation of longitudinal research**: Longitudinal studies can be used to monitor changes in gene expression over time in AD sufferers. This can aid in identifying important biological processes linked to disease progression and offer information on prospective treatment targets.

4. **Validation of findings in animal models**: Mechanistic insights into the disease can be gained through using animal models to validate results from gene expression meta-analysis. This may make it easier to find prospective drug development targets.

5. **Finding novel treatment targets**: Utilizing gene expression meta-analysis can aid in the identification of novel therapeutic targets for AD. Researchers can find novel drug targets and create fresh approaches to treating diseases by comprehending the molecular systems that underlie them.

Gene expression meta-analysis in AD offers promising prospects for future advancements in research. The use of longitudinal research, the integration with other omics data, the application of machine learning approaches, the validation in animal models, and the discovery of novel therapeutic targets are all areas that can assist expand our understanding of the illness and pave the path for new treatments.

REFERENCES

[1] E. M. Blalock, J. W. Geddes, K. C. Chen, N. M. Porter, W. R. Markesbery, And
P. W. Landfield, "Incipient AD: Microarray Correlation Analyses Reveal Major Transcriptional And Tumor Suppressor Responses," *Proc. Natl. Acad. Sci. U. S. A.*, Vol. 101, No. 7, Pp. 2173–2178, 2004.

[2] "GEO2R - GEO - NCBI," *Nih.Gov*. [Online]. Available: Https://Www.Ncbi.Nlm.Nih.Gov/Geo/Geo2r/. [Accessed: 28-Nov-2022].

[3] H. Patel, R. J. B. Dobson, And S. J. Newhouse, "A Meta-Analysis Of AD Brain Transcriptomic Data," *J. Alzheimers. Dis.*, Vol. 68, No. 4, Pp. 1635– 1656, 2019.

[4] "Bioinfoindia.Org," *Bioinfoindia.Org*. [Online]. Available: Https://Www.Bioinfoindia.Org/. [Accessed: 28-Nov-2022].

[5] "{{Ngmeta['Og:Title']}}," *Bio.Tools*. [Online]. Available: Https://Bio.Tools/Nasqar. [Accessed: 28-Nov-2022].

[6] "WGCNA: R Package For Performing Weighted Gene Co-Expression Network Analysis," *Ucla.Edu*. [Online]. Available: Https://Horvath.Genetics.Ucla.Edu/Html/Coexpressionnetwork /Rpackages/WGCN A/. [Accessed: 28-Nov-2022].

[7] *Rcpsych.Ac.Uk*. [Online]. Available: Https://Www.Rcpsych.Ac.Uk/Docs/Defaultsource/Other-Files/The-Mmse-Old-Age-Psychiatrist-3-1-14.Doc. [Accessed: 28Nov-2022].

[8] "Bgee," *Expasy.Org*. [Online]. Available: Https://Www.Expasy.Org/Resources/Bgee. [Accessed: 28-Nov-2022].

[9] Genecards Human Gene Database, “RALYL Gene - Genecards,” *Genecards.Org*. [Online]. Available: Https://Www.Genecards.Org/Cgibin/Carddisp.Pl?Gene=RALYL. [Accessed: 28-Nov-2022].

[10] “Home - GEO - NCBI,” *Nih.Gov*. [Online]. Available: Https://Www.Ncbi.Nlm.Nih.Gov/Geo/. [Accessed: 28-Nov-2022].

[11] “GEO Accession Viewer,” *Nih.Gov*. [Online]. Available: Https://Www.Ncbi.Nlm.Nih.Gov/Geo/Query/Acc.Cgi?Acc=Gse138261. [Accessed: 28-Nov-2022].

[12] A. Nitsche*et Al.*, “Alzheimer-Related Genes Show Accelerated Evolution,” *Mol. Psychiatry*, Vol. 26, No. 10, Pp. 5790–5796, 2021.

[13] “GEO Accession Viewer,” *Nih.Gov*. [Online]. Available: Https://Www.Ncbi.Nlm.Nih.Gov/Geo/Query/Acc.Cgi?Acc=Gse1297. [Accessed: 28Nov-2022].

[14] T. Baj And R. Seth, “Role Of Curcumin In Regulation Of TNF-A Mediated Brain Inflammatory Responses,” *Recent Pat. Inflamm. Allergy Drug Discov.*, Vol. 12, No. 1, Pp. 69–77, 2018.

[15] R. Shukla And T. R. Singh, “Identification Of Small Molecules Against Cyclin Dependent Kinase-5 Using Chemoinformatics Approach For AD And Other Tauopathies,” *J. Biomol. Struct. Dyn.*, Vol. 40, No. 6, Pp. 2815–2827, 2022.

[16] A. Kumar, A. Bansal, And T. R. Singh, “ABCD: AD Biomarkers Comprehensive Database,” *3 Biotech*, Vol. 9, No. 10, P. 351, 2019.

[17] R. Shukla And T. R. Singh, “Virtual Screening, Pharmacokinetics, Molecular Dynamics And Binding Free Energy Analysis For Small Natural Molecules Against Cyclin-Dependent Kinase 5 For AD,” *J. Biomol. Struct. Dyn.*, Vol. 38, No. 1, Pp. 248–262, 2020.

[18] A. Kumar And T. R. Singh, “Analysis For Biological Network Properties Of AD Associated Gene Set By Enrichment And Topological Examinations,” *Int. J. Bioinform. Res. Appl.*, Vol. 13, No. 3, P. 214, 2017.

[19] A. Kumar *Et Al.*, “Computational And In-Vitro Validation Of Natural Molecules As Potential Acetylcholinesterase Inhibitors And Neuroprotective Agents,” *Curr. Alzheimer Res.*, Vol. 16, No. 2, Pp. 116–127, 2019.

[20] *Researchgate.Net*. [Online]. Available: Https://Www.Researchgate.Net/Profile/Rohit-Shukla9/Publication/357205040_Identification_Of_Natural_Multitarget_Inhibitors_For_Alzheimer's_Disease_Using_Polypharmacology_Approach/Links/61c1a143abfb4634cb321bc7/Identification-Of-Natural-Multi-Targetinhibitors-For-Alzheimers-Disease-Using-Polypharmacologyapproach.Pdf#Page=82. [Accessed: 05-Dec-2022].

[21] (Amyloid Plaques and Neurofibrillary Tangles | Bright focus Foundation, 2015)

[22] (User Guide - WB-DEGS, N.D.)